Philips Journal of Research

Volume 33

1 9 7 8

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1978.
Articles or illustrations reproduced, in whole or in part, must be accompanied by full acknowledgement of the source: Philips Journal of Research.

CONTENTS

PHILIPS JOURNAL OF RESEARCH, VOL. 33

8-3-48=52

Philips Journal of Research

Vol. 33 Nos. 1/2 1978

PHILIPS

Philips Journal of Research

Philips Journal of Research, edited by Philips Research Laboratories, Eindhoven, The Netherlands, is a bimonthly publication containing papers on research carried out in the various Philips laboratories. Volumes 1 to 32 have appeared under the name of Philips Research Reports.

Annual subscription price for Volume 33 is Dfl. 60, payable in advance. Payments should be made only after receipt of an invoice. Correspondence should be addressed to: Philips Journal of Research, Philips Research Laboratories, Building WBp, Room No. 42, Eindhoven, The Netherlands.

Cover design based on a visual representation of the sound pressure associated with the spoken word "Philips".

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1978. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

PHILIPS JOURNAL OF RESEARCH - Vol. 33 Nos. 1/2 - pp. 1-116 - 1978

CONTENTS

Date of publication: May 31, 1978.

Philips J. Res. Vol., 33, 1-19, 1978 *R 977*

THE REACTION OF BROMINE AND OXYGEN WITH A TUNGSTEN SURFACE

PART I: MEASUREMENT OF THE RATE OF REACTION IN A GAS-FLOW EXPERIMENT

by G. ROUWELER and B. J. de MAAGT

Abstract

The ablation rate of polycrystalline tungsten in bromine and oxygen was measured in a gas-flow experiment with argon as a carrier gas at atmospheric pressure. The bromine and oxygen partial pressures ranged between 3×10^{-3} and 3×10^{3} Pa, and between 6×10^{-2} and 30 Pa, respectively, and the temperatures ranged between 975 and 1448 K. Two types of reaction were observed, the "W–Br" and the "W–O" reaction. The rate of the former reaction (in mol s^{-1} cm^{-2}) was found to obey the empiric relation: $j_{W-Br} = 4.7\times10^{-13}\,p_{Br} + 4.0\times10^{-15}\,p_{Br}{}^{2}$, while the rate of the latter could be expressed as $j_{W-O} = k_{W-O}\,p_{O_2}{}^{f}\,p_{Br_2}{}^{g}$, k_{W-O}, f and g being temperature-dependent parameters. It is proposed that the rate-limiting step in the W–Br and in the W–O reaction is the formation of a tungsten–bromine surface complex by an Eley-Rideal mechanism, and the formation of a tungsten–oxygen surface complex by a Langmuir–Hinselwood mechanism, respectively.

1. Introduction

In the literature, studies on the reactivity of tungsten–halogen(–oxygen) systems at high temperatures have appeared from several different disciplines. They can be grouped into publications on halogen–metal reactions [1–11]), halogen metallurgy [12]), the purification of metals by vapour transport [13]) and on the working principles of lamps with tungsten transport [14–18]). The last of these especially are of importance in halogen incandescent lamps. Here tungsten is transported from regions of a relatively high tungsten vapour solubility towards regions of low solubility [19,20]). Through careful selection of the gas-phase components in these lamps the wall is kept clear of tungsten, with only small deterioration effects along the tungsten filament. In the theoretical treatment of tungsten-transport rates the Langmuir model is generally adopted, which states that there is a well-mixed zone of uniform chemical composition and temperature at the wall. Around the filament a stagnant layer of gas is thought to be present where tungsten transport takes place by vapour-phase diffusion, assuming local thermodynamic equilibrium (LTE). The model more or less adequately explains or predicts lamp results [19-21]. However, every now and then, the opinion is expressed that chemical kinetics in the "low" temperature parts of the lamps, even during continuous operation, might be the dominant mechanism in tungsten transport [22,23]. For this reason we studied the ablation rate of polycrystalline tungsten in a reactive gas flow of argon with bromine and oxygen (which is a common contaminant in lamps) at temperatures between 975 and 1448 K. Two distinct types of reaction were observed, one involving only bromine and the other one depending on the action of both oxygen and bromine. These reactions are interpreted with the aid of reaction-rate theory as an interaction of bromine and oxygen respectively with a chemisorption layer of bromine on tungsten [24]. Furthermore it was shown that the rate of reaction at the tungsten surface may explain the velocity of the process of tail-erosion in halogen incandescent lamps.

2. Experiments

2.1. *System description*

The measurements on the reaction of tungsten with bromine and oxygen were performed in an open system, see fig. 1. Tungsten probes were present in the form of two straight wires, in most cases of normal lamp quality (W-d). The contaminant concentration in this material, as for probes of slightly different composition (W-zg), which were also tested in the experiments, is given in table I.

TABLE I

Contaminant composition in percents by weight determined by spectroscopic analysis

element	W-d (wt%)	W-zg (wt%)
Mo	0.0033	0.01
Si	0.012	0.004
Al	0.0019	0.002
Fe	0.0027	0.002
Cu	0.0005	0.0002
Mn	$\leqslant$ 0.0001	0.0001
K	0.008	0.0019

The experimental variables were gas composition, temperature and the rate of gas flow. The carrier gas argon had a lamp-production quality and was

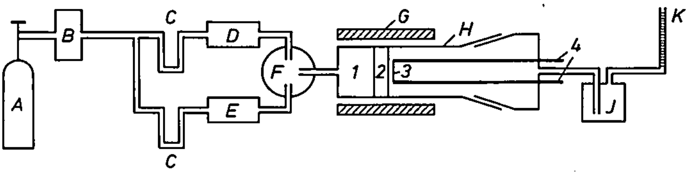


Fig. 1. Schematic description of the experimental set-up during measurement of the reaction rate. A: argon cylinder; B: gas purification; C: flow meters; D: oxygen generator/sensor; E: bromine generator; F: gas mixing vessel; G: electric oven; H: reaction tube with silica heat exchanger (1), silica flow channel (2), tungsten probes (3), and tungsten supports (4); J: bromine absorption vessel; K: soap-film flow meter.

purified further by a Messer–Griesheim adsorption cell up to an oxygen and water mole fraction level of less than 0.5×10^{-6}. Bromine was obtained by electrolysis of a melt of silver bromide (UCB, analytical quality). Oxygen was also generated by electrolysis in a zirconia solid–electrolyte cell. The quantities of gas thus generated were in accordance with Faraday's law, as was checked experimentally by titration. A small correction factor had to be applied in the case of bromine, due to electronic conduction in the melt, and to the uptake of Br_2 in the melt followed by reaction at the silver cathode. A check was carried out to ensure that no oxygen was generated from oxygen-containing contaminants which might have been present in the AgBr melt. The zirconia cell had two pairs of platinum electrodes. The downstream one was used to measure oxygen partial pressure by the application of Nernst's law. Now the quality of the argon could be checked by adding or withdrawing small quantities of oxygen at the upstream electrodes, so that a kind of titration curve was obtained. Small traces of oxygen, whether chemically bonded or not, were still present in mole fractions up to a maximum of 0.2×10^{-6} in O_2 equivalents; reducing components were present in mole fractions up to a maximum of 0.5×10^{-6} H_2 equivalents.

After the gas flow containing Br_2 and O_2 had been mixed, the gas was led into the reaction tube, which was surrounded by an electric oven. In the tube a silica heat exchanger and a flow channel were placed. The latter consisted of a set of narrowly spaced parallel silica plates. In this way a gas flow with a constant velocity along the probes was obtained. At the location of the probes it was possible to keep the temperature constant to within 0.3 K. The probes consisted of two tungsten wires — length 22 mm, initial diameter 49.6 μm — which were attached to two pairs of heavy tungsten supports. Before an experiment was started the probes were pretreated by exposing them overnight at 1273 K to an argon flow containing an O_2 mole fraction of 0.3×10^{-6}.

The rate of tungsten removal, j_W (mol cm^{-2} s^{-1}), was determined by continuously registering the voltage across the wires, through which a constant direct current of 10 mA was led. Assuming that the wire retains its cylindrical geometry one derives

$$j_W = (d_W\, r_0/M_W\, t)\, [1-(V_0/V_t)^{\frac{1}{2}}]. \tag{1}$$

Here M_W and d_W are the molecular mass and density of tungsten respectively, t is the reaction time, r_0 the initial wire radius, V the voltage. The value of j_W was obtained from a least-squares evaluation of $(V_0/V_t)^{\frac{1}{2}}$ vs t plots.

Each experiment was continued under stationary conditions until the wire diameter was reduced by at least 0.1 μm. The overall diameter reduction was never allowed to exceed 15 μm. In this way an accuracy of ca 10% in the value of j_W could be obtained. The tungsten reaction products condensed in the cooler parts of the reaction tube, Br_2 was removed from the main gas by bubbling it twice through an NaOH solution. The gas-flow rate was measured by a soap-film flow meter.

2.2. *Mass-transfer limited ablation*

In a reaction between gaseous components and a solid, whereby volatile reaction products are formed, one generally distinguishes between three consecutive steps:

(1) transport of reagents towards the surface,

(2) surface reaction (sub-division: adsorption, reactions in the adsorbed state, desorption),

(3) transport of reaction products towards the gas phase.

In most cases one of these steps is relatively slow, so the overall process rate is governed by this rate-limiting step. In this section we will focus our attention on mass transfer, namely the first and last step, which can be dealt with using the same model. The mass transport can be calculated from the formal analogy between mass and heat transport from and towards solid objects in convective systems [25,26]). In the undisturbed gas flow, as well as at the surface, chemical equilibrium is assumed; in the boundary layer no chemical reactions are thought to occur. We apply this model to our system in which an isothermal homogeneous transverse gas flow is directed to a tungsten cylinder. The molar mass transport j_i, for species i, across the layer per cm^2 of cylinder area is then given by

$$j_i = k_{M,i}\, (p_{i,s} - p_{i,\sim})/P. \tag{2}$$

Here $k_{M,i}$ is the mass-transfer coefficient, p partial pressure, P total pressure and '$\sim$' and 's' refer to conditions in the undisturbed gas flow and at the surface respectively. The value of $k_{M,i}$ can be found from the Reynolds number in a Chilton–Colburn diagram [26]). In our case, with relatively small flow rates, we had to extrapolate the characteristic in the diagram to small Reynolds numbers, which led to an approximate value for $k_{\mathrm{M,i}}$:

$$k_{\mathrm{M,i}} = 0.56 \frac{P}{RT} D_{\mathrm{i}}^{\frac{2}{3}} \left(\frac{v_{\mathrm{g}}}{2r}\right)^{\frac{1}{2}} \left(\frac{d_{\mathrm{g}}}{\mu_{\mathrm{g}}}\right)^{\frac{1}{6}}. \tag{3}$$

Here v_{g} is the velocity of the undisturbed gas flow, D_{i} the binary diffusion coefficient of species i in the main gas, and μ_{g} the viscosity of the gas. The other symbols have their usual meaning.

From chemical equilibrium calculations in the W–Br_2–(O_2) system [27]), at the temperature and pressure conditions prevailing in our gas-flow experiments, it is shown that practically all bromine and oxygen is bonded in WBr_4 (and WO_2Br_2). It can then be shown that step 1, the supply of reagents, is the rate-limiting step, while $p_{\mathrm{WBr_4,s}}$ and $p_{\mathrm{WO_2Br_2,s}}$ adapt themselves to satisfy the mass balance.

In order to make an approximate test of the validity of eqs (2) and (3) we varied v_{g}, the quantity in $k_{\mathrm{M,i}}$ which is varied most easily; some typical graphs are shown in fig. 2. At small flow rates here oxygen transport is the rate-limiting step, while at higher flow rates tungsten ablation is limited by surface-reaction rates. From these graphs and numerous other experiments not shown here, it is concluded that the model gives the right order of magnitude for the mass transport. A thorough test of eqs (2) and (3) in the range exclusively determined by mass transport was not attempted.

2.3. *Reaction-limited ablation*

In studying surface reactivity (see sec. 2.2, step 2), one has to realize that mass

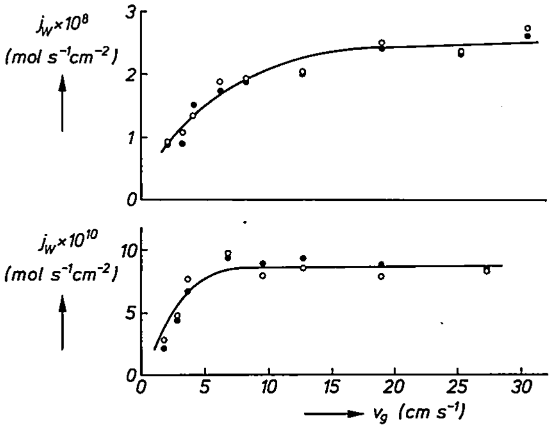


Fig. 2. Tungsten reaction rate j_{W} as a function of linear gas-flow velocity at a temperature of 1166 K and a Br_2 partial pressure of 1.9 Pa. The oxygen pressure is 8 and 1 Pa for the upper and lower curve respectively.

transport (step 1, 3) always is in series with surface reaction. So only at infinitely large values of $k_{M,i}$ is the value of $p_{i,s}$ equal to $p_{i,\sim}$. And because $p_{i,\sim}$ is the quantity known from the boundary conditions, one should extrapolate the experimental values of $j_W(v_g)$ to infinitely large flow-rate velocity to obtain the correct relation between reaction rate and gas composition. However, in practice this applies only to experiments where the oxygen supply is the critical mass-transport step. In that situation a combination of relatively high surface-reaction rates and small values of $p_{O_2,\sim}$ necessitates high values of k_{M,O_2} in order to have enough mass-transport capacity (see eq. (2)). This effect, combined with the small dependence of $k_{M,i}$ on temperature, prevented us from carrying through our reaction-rate experiments in the W–Br_2–O_2 system at higher temperatures, because of the experimental limitation in the gas flow and thus in $k_{M,i}$ in our experimental set-up. In the evaluation of the experiments on the kinetics of the W–Br_2–O_2 reaction the reaction rate is defined as the value of the ablation rate obtained at gas-flow velocities where $\partial j_W/\partial v_g = 0$. This criterion was checked in every measurement.

2.4. *Measurements and results*

We determined experimentally the rate of reaction of tungsten in bromine and in bromine–oxygen mixtures with the aid of two straight tungsten wires placed in a laminar gas flow. This probe-geometry is advantageous from the view of mass transport. Thus we were able to carry out our experiments in a larger range of temperatures and of reactant concentrations than Zubler [14,15]) and Goddard and Pett [18]) in their work. These authors published experiments on the reactivity of tungsten in bromine–oxygen mixtures and in bromine respectively, using flat tungsten plates as test probes. There is also a difference in the measuring method. By our choice of probe-geometry we had to drop the method of monitoring weight change. We chose the method of monitoring the electrical resistance and calculating the weight change from that. Because of the strong temperature dependence of the resistivity of tungsten, the method required an oven stability of 0.3 K during an experiment. The overall weight change of the tungsten wire after a series of experiments proved to be identical with the one calculated from our resistance measurements.

One of the major complications in our analysis of the results was the apparent inconstant behaviour of the reaction rate. Some examples of this phenomenon are shown in fig. 3. Typical is the steep fall-off of the curve during the first few µm of the ablation process and the gradual increase of the rate afterwards. An explanation for this effect is the enlargement of the micro-surface to geometric-surface ratio during progressive ablation, a phenomenon which is clear from

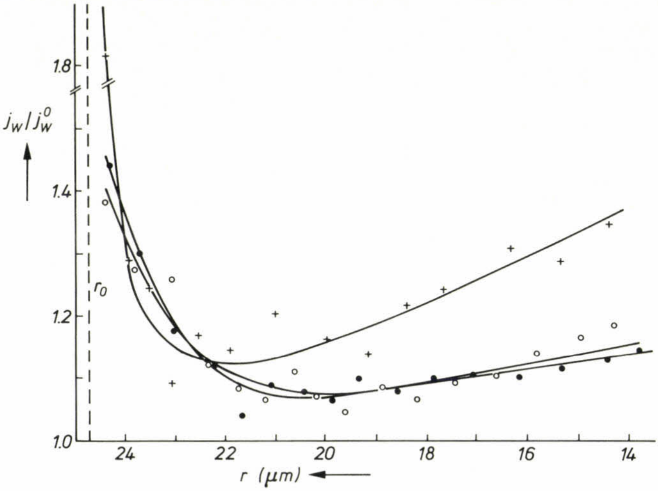


Fig. 3. Relative change $j_W/j_W{}^0$ of tungsten reaction rate as a function of wire radius r for "lamp-quality" material W-d (●, o) and material of slightly different composition W-zg (+). The values of $j_W{}^0$ are obtained by extrapolation of the reaction rate from low values of r to the initial values (r_0 = 24.8 μm). The values of $j_W{}^0$ are 3.9×10^{-9} mol s^{-1} cm^{-2} (●), 3.85×10^{-9} mol s^{-1} cm^{-2} (o), and 3.5×10^{-9} mol s^{-1} cm^{-2} (+). Conditions during reaction: T = 1282 K, Br_2 pressure 3.6 Pa and oxygen pressure 2.5 Pa.

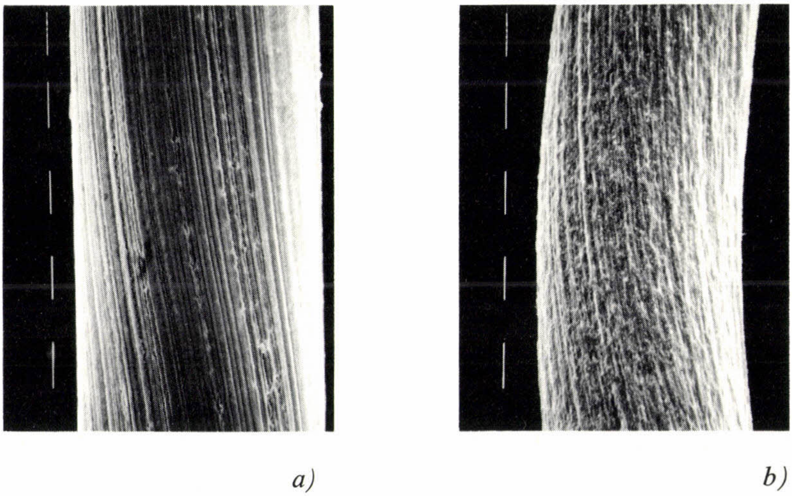


Fig. 4. SEM photographs of tungsten-d wire as received (*a*) and after etching it to a diameter of 40 μm (*b*). Each mark line is about 10 μm.

SEM-inspection, see fig. 4. The coarsening of the surface may be attributed to preferential attack of the tungsten crystals at the surface which have a favourable orientation. This view is supported by some preliminary experiments on the reaction of single crystals with different crystal-axis orientation. Rods with their axis parallel to the ⟨100⟩, ⟨110⟩ and ⟨111⟩ orientations of the crystal are ablated at 1287 K, $p_{Br_2} = 590$ Pa and $p_{O_2} = 1.3$ Pa at a ratio of 1.2 : 0.8 : 1.3 with respect to polycrystalline wire.

Moreover, when recrystallized wires were ablated, strong facet building occurred, developing notably the {100} and {111} faces. Now the initially fast ablation rate of the wire, which consists of a "cable" of elongated tungsten crystals with a diameter of about 0.5 μm, is attributed to selective ablation of relatively reactive crystals. This initial process leaves a less reactive surface but a microsurface expanding with respect to the geometric surface because of the uniform distribution of crystals of different specific reactivity in the tungsten bulk. This leads to an ostensible increase in ablation rate because this quantity is related to the geometric surface. Since the effect under discussion is inherent to reaction-rate measurements on polycrystalline tungsten, we accepted the spread caused by this phenomenon, using results obtained during ablation between radii of 23 and 17 μm. In principle it would have been possible to extrapolate the measuring points to zero ablation (see fig. 3 for instance), so eliminating the effect of surface coarsening. This would, however, have extended measuring time unduly. So our results have to be interpreted as results obtained in the measurement of the reaction rate in mol s^{-1} per cm^2 of geometric surface of polycrystalline tungsten with a stationary surface distribution of tungsten crystals and a slowly expanding micro-surface.

The effect discussed above may cause an inaccuracy of ca 25% in the experimental results, but this is only of minor importance since the reaction rate changes several orders of magnitude in our experiments. Also, this effect will be left out further on in the discussion, when comparing experiments of the authors mentioned above with our data.

The reaction rate was measured as a function of the bromine content in the gas phase at five temperatures and is shown in fig. 5. The reaction rate in the bromine–oxygen system was measured at three temperatures, as is shown in figs 6*a*, 6*b* and 6*c*.

No essential difference between W-d and W-zg material was found. On the abcissa of both figures the Br_2 partial pressure is plotted, which was calculated from the electrolysis current, the gas-flow rate and the dissociation constant at the prevailing temperature [28]). In the same manner the O_2 partial pressure is calculated, but here dissociation can be neglected. In all measuring points the result of at least three separate experiments — and thus of six probes —

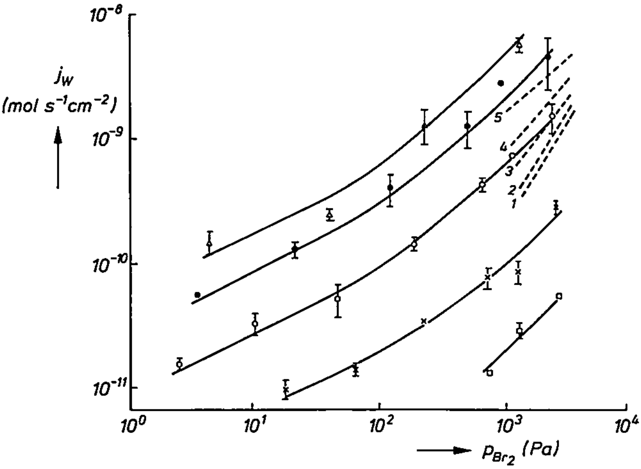


Fig. 5. Tungsten reaction rate j_W as a function of partial Br_2 pressure, measured at 975 K (□), 1067 K (×), 1166 K (o), 1282 K (•), and 1448 K (△). The broken lines represent the measurements of Goddard and Pett (ref. 18) at 1063 K (1), 1083 K (2), 1113 K (3), 1163 K (4), and 1213 K (5).

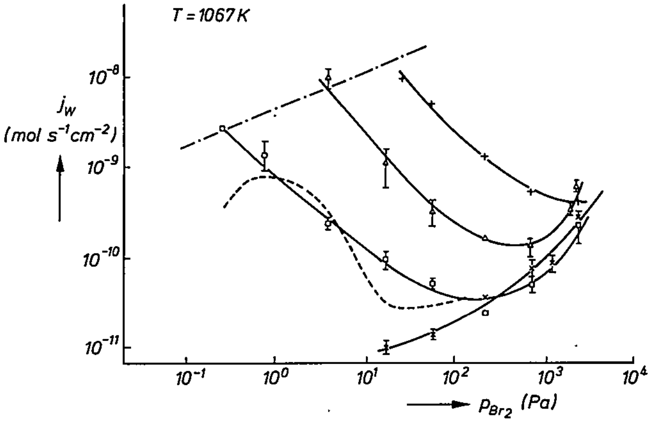


Fig. 6*a*. Tungsten reaction rate j_W as a function of partial Br_2 pressure at 1067 K at various oxygen pressures: $p_{O_2} = 0$ (×), 1 Pa (□), 8 Pa (△), and 30 Pa (+). The broken curve represents the measurements of Zubler (ref. 15) at 1073 K in helium and argon as well, at an oxygen pressure of 1.15 Pa. Left of the dash-dotted line oxide formation takes place.

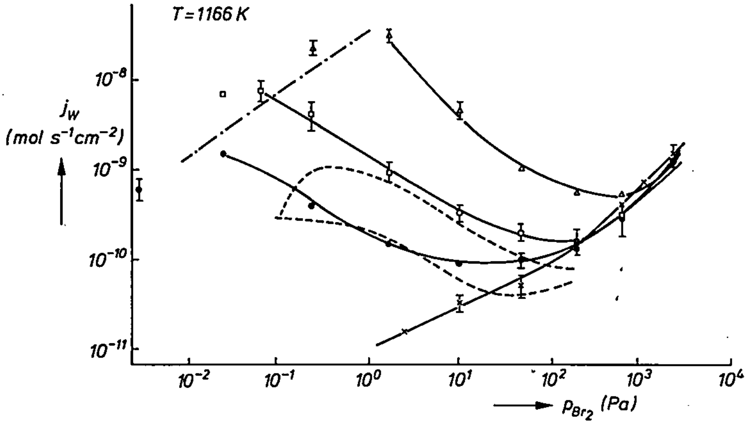


Fig. 6*b*. Tungsten reaction rate j_W as a function of partial Br_2 pressure at 1166 K at various oxygen pressures: $p_{O_2} = 0$ (×), 0.25 Pa (●), 1 Pa (□), and 8 Pa (△). The broken curves are measurements of Zubler (ref. 15) at 1173 K in helium. The lower one represents results at an oxygen pressure of 0.27 Pa, the upper one at 1.15 Pa. Left of the dash-dotted line oxide formation takes place.

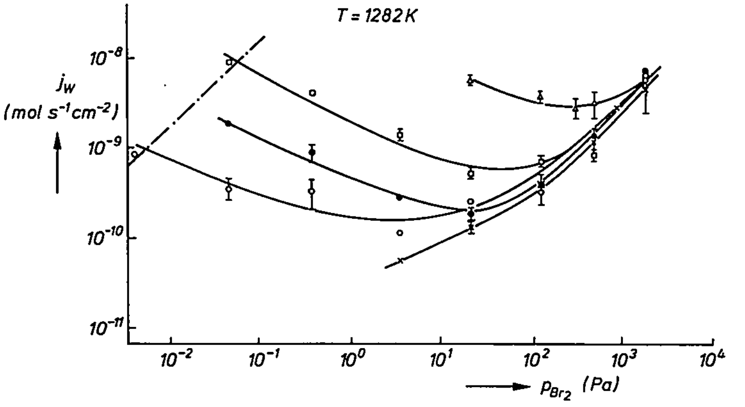


Fig. 6*c*. Tungsten reaction rate j_W as a function of partial Br_2 pressure at 1282 K at various oxygen pressures: $p_{O_2} = 0$ (×), 0.06 Pa (o), 0.25 Pa (●), 1 Pa (□), and 8 Pa (△). Left of the dash-dotted line oxide formation takes place.

is combined. The onset of oxide formation at our probes is indicated by a dash-dotted line.

In fig. 5 the measurements recorded by Goddard and Pett [18]) are also included *). Goddard and Pett's data on pure bromine have the same order of

*) The reaction rates in their publication have to be corrected by a factor of 10^8, as can be shown from a recalculation of the results in their fig. 2.

magnitude as our results but they show a different temperature dependence. (We did not include their experiments in argon because we believe their values are low due to the small gas-flow rates in their experiment leading to mass-transport-controlled ablation rates).

In figs *6a* and *6b* Zubler's data are also shown [15]). However, there is a slight discrepancy in temperature between his measurements and ours. Zubler's results, obtained in argon and helium as carrier gases, compare favourably with ours, as can be seen from figs *6a* and *6b*. The decreasing rate in his measurements at increasing O : Br ratio might be due to insufficient mass-transport capacity. Then tungsten oxides might start to condense whilst in our experiments condensation starts at higher O : Br ratios.

SEM inspection after ablation revealed that the tungsten wires had retained their cylindrical form and their metallic appearance. However, the wire surface was roughened considerably, showing clearly the distinct cablelike arrangement of tungsten crystals.

Incidentally, a spherical etch pit was observed. Oxidation phenomena were found in cases where a relatively high oxygen to bromine ratio was sustained during the experiments. These phenomena ranged from needle-like crystals in pits, via veils extending to some distance around the wire, to surfaces fully blanketed with oxide nodules. Measurements where these non-volatile oxidation products were observed will be excluded from the discussion.

Several attempts were made to characterize the reaction products. By X-ray diffraction the presence of WO_2Br_2 as a dominating species was demonstrated in the material condensed at the cooler parts of the reaction tube. Chemical analysis of the same samples resulted in W : Br ratios in the range of 1/2 to 1/2.8, presumably due to mixtures of WO_2Br_2 and WBr_4. However, the samples contained the products of several experiments and could be collected only at a location far from the reactive surface, so one could not be certain that there was a direct relation between the sample and the actual reaction products at the surface of the probe. For this reason no attempts were made to find a better definition of the reaction products.

3. Discussion

In this paragraph we will critically evaluate our experimental results by comparison with theory and experiments found in literature on similar systems. Eventually conclusions are drawn from literature data combined with our own findings.

3.1. *Gas-flow experiments*

As can be seen from fig. 6 the overall reaction in the W–Br_2–O_2 system has

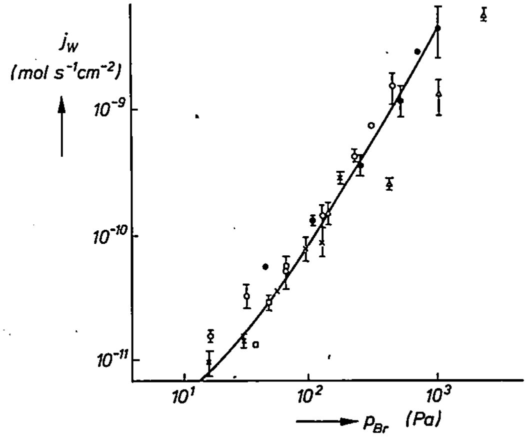


Fig. 7. Tungsten reaction rate j_W as a function of atomic bromine pressure at 975 K (□), 1067 K (×), 1166 K (o), 1282 K (●), and 1448 K (△). The drawn curve obeys eq. (4).

to be described by at least two different types of reaction. The first one, which we will call the W–Br reaction, is found in its pure form in the W–Br_2 system (fig. 5). A specific characteristic of this reaction is that the reaction rate is nearly temperature-independent when atomic bromine is taken as the active species. In this case we prefer to present the data of the W–Br reaction in this form (see fig. 7), rather than giving Arrhenius plots at constant molecular bromine partial pressure. From this effect it is tempting to take the bromine dissociation as the rate-determining step, but it is known that the dissociation energy, being 191 kJ/mol, is larger than the activation energy for dissociation, being 134 kJ [29]), which excludes this possibility here. From a least-squares treatment of the data in the temperature traject between 975 and 1282 K, an empirical relation for the ablation rate, j_{W-Br} (mol s^{-1} cm^{-2}), is found (see also curve drawn in fig. 7):

$$j_{W-Br} = k_1\, p_{Br} + k_2\, p_{Br}{}^2, \tag{4}$$

the value of the reaction constants k_1 and k_2 being respectively 4.7×10^{-13} mol s^{-1} cm^{-2} Pa^{-1} and 4.0×10^{-15} mol s^{-1} cm^{-2} Pa^{-2}. At 1448 K the stability of the reaction product WBr_4 has already diminished considerably and the bromine adsorption also diminishes [24]), so at this temperature the reaction path may have been changed.

In the W–Br_2–O_2 system a second reaction is found — the W–O reaction — whose rate is more or less proportional to the oxygen partial pressure and which, in contrast to the W–Br reaction, decreases with increasing bromine partial pressure (see table II and fig. 6). A similar effect is found by Rosner and Allendorf[6]) in the W–Cl_2–O_2 system. The empirical relation for the ablation rate in the W–O system is

$$j_{\mathrm{W-O}} = k_{\mathrm{W-O}}\, p_{\mathrm{O_2}}{}^{f}\, p_{\mathrm{Br_2}}{}^{g}. \tag{5}$$

TABLE II

Partial pressure dependence of the reaction rate (mol s^{-1} cm^{-2}) on partial pressure (Pa) as calculated from a least-squares treatment of the data in the W–O reaction

T (K)	f	g	$10^{10}\, k_{\mathrm{W-O}}$
1067	1.55	−0.92	9.8
1166	1.20	−0.61	12
1282	1.05	−0.40	23

An explanation for the reactive behaviour is attempted with the aid of reaction-rate theory (see ref. 30 for example). Here the rate of a single reaction is determined by the rate of decomposition of the so-called activated complex $X^{\neq}$. In a chain of reactions the overall rate is determined by the rate of decomposition of the critical activated complex $X^{\neq}{}_{\mathrm{crit}}$ in the rate determining step

$$j = \nu\, [X^{\neq}{}_{\mathrm{crit}}]. \tag{6}$$

Here ν is the frequency of the vibration mode by which the critical activated complex dissociates. A consequence of the theory is that all reagents and reaction products in the steps following the formation of the critical activated complex are in chemical equilibrium. The same is true for all components involved in the reactions preceding the decomposition of that complex, including the critical complex itself. We apply this reasoning to the reactions in the W–Br_2–O_2 system. As a generalization, such a reaction can be represented by

$$q\mathrm{S} + a\mathrm{Br_2} + b\mathrm{O_2} \rightleftarrows \ldots \rightleftarrows \mathrm{S}_q\, \mathrm{Br}_{2a}\, \mathrm{O}_{2b}{}^{\neq} \rightleftarrows \ldots \rightleftarrows \text{reaction products.} \tag{7}$$

Here S symbolizes a free tungsten atom at the surface. The reaction rate in mol s^{-1} cm^{-2} then is given by

$$j = \frac{kT}{hN_{\mathrm{A}}}\, \frac{F^{*}{}_{\mathrm{X}^{\neq}}}{F_{\mathrm{S}}{}^{q}\, F_{\mathrm{Br_2}}{}^{a}\, F_{\mathrm{O_2}}{}^{b}}\, c^{q}\, n_{\mathrm{Br_2}}{}^{a}\, n_{\mathrm{O_2}}{}^{b} \exp\left(-E_{\mathrm{a}}/RT\right). \tag{8}$$

Here k is Boltzmann's constant, N_A is Avogadro's number, h is Planck's constant, F is the partition function, c and n are the particle densities at the surface and in the gas phase respectively and E_a is the energy difference between the activated complex and the initial reagents. The asterisk in the partition function of the activated complex means, that the term belonging to the vibration mode by which the complex dissociates, is taken from the partition function of the complex.

First we will try to find a definition for the stoichiometry of the activated complex in the W–Br as well as in the W–O reaction. The next step will be to test the numerical value of eq. (8) against experiment for the former reaction.

It was concluded [24]) that in the W–Br system there is a quasi-equilibrium between the atomic bromine in the adsorption layer and atomic bromine in the gas phase. Since in the gas-flow experiment atomic and molecular bromine are in equilibrium too, one can formulate the adsorption equilibrium in Langmuir's adsorption model, taking atomic bromine as the adsorbing species, as

$$\frac{F_{\mathrm{SBr}}}{F_{\mathrm{Br_2}}{}^{\frac{1}{2}}\, F_{\mathrm{S}}} \exp\left(-E_{\mathrm{SBr}}/RT\right) = \frac{c_{\mathrm{SBr}}}{c_{\mathrm{S}}\, n_{\mathrm{Br_2}}{}^{\frac{1}{2}}}\,. \tag{9}$$

The left-hand side of the equation is the statistic thermodynamic equivalent for the equilibrium constant and E_{SBr} is the adsorption energy per mole atomic bromine bonded at the surface (this bonding energy was determined in the LEIS experiment [24])). Since the bromine particle densities in the gas-flow experiment are far larger than in the LEIS experiments one may assume that the first monolayer is nearly completely filled with bromine. So $c_{\mathrm{SBr}} = C$, where C is the monolayer capacity. Eliminating $c_{\mathrm{S}}/F_{\mathrm{S}}$ from eqs (9) and (8) results in

$$j = \frac{kT}{hN_{\mathrm{A}}}\, \frac{F^{*}{}_{\mathrm{X}}{}^{\neq}}{F_{\mathrm{Br_2}}{}^{(a-q/2)}\, F_{\mathrm{O_2}}{}^{b}\, F_{\mathrm{SBr}}{}^{q}}\, C^{q}\, n_{\mathrm{Br_2}}{}^{(a-q/2)}\, n_{\mathrm{O_2}}{}^{b} \exp\left[(qE_{\mathrm{SBr}} - E_{\mathrm{a}})/RT)\right]. \tag{10}$$

In the W–Br reaction the value of b is of course zero. Also it seems improbable that more than one tungsten atom participates in the activated complex because bromine is incapable of bridging; this makes $q = 1$. From pressure dependence of the ablation rate (see eq. (4)) it follows that the value of $(a - q/2)$ in eq. (10) lies in the range between 0.5 and 1. If we interpret eq. (4) as an indication for two mechanisms in the W–Br reaction, we can conclude that the critical activated complex is $\mathrm{SBr_2}^{\neq}$ at lower and $\mathrm{SBr_3}^{\neq}$ at higher atomic bromine partial pressures. Since, as stated above, the surface is completely covered with chemisorbed bromine, the critical step in the reaction must be the addition of a bromine atom direct from the gas phase to the tungsten–bromine surface complexes. This type of reaction is known as the Eley–Rideal reaction [30]).

In comparing the reaction rate as predicted by eq. (10) with the experimental findings (eq. (4)), we confine ourselves only to the lower partial pressure regime, because there the extrapolation of the conclusions drawn in the "vacuum" experiments [24]), which are hidden in eq. (10), is most reliable. So we take only the first term in eq. (4). Since in this region $SBr_2^{\neq}$ is the critical activated complex and thus $a = 1$ (see eq. (7)), we arrive, after substituting the dissociation equilibrium for Br_2 in eq. (10), at the expression

$$j = \frac{kT}{hN_A}\,\frac{F^{*}{}_{X}{}^{\neq}}{F_{Br}\,F_{SBr}}\,C\,n_{Br}\exp\left[(E_{SBr} + \tfrac{1}{2}E_d - E_a)/RT\right]. \tag{11}$$

Here E_d is the dissociation energy of molecular bromine. The term $(E_{SBr} + \frac{1}{2}E_d - E_a)$ has to be interpreted as the activation energy necessary for the formation of the $SBr_2^{\neq}$ complex from atomic bromine (Br) and bromine chemisorbed to tungsten (SBr); in fact the equation describes the rate of that reaction. The values of F in eq. (11) can be estimated. In the surface complexes only vibration makes a contribution, which is rather small at the temperatures of the experiment, so we take them to be unity. The F value for a mono-atomic gas is $(2\pi mkT)^{\frac{3}{2}}/h^3$, where m is the atomic mass. Substitution of these values in eq. (11) gives

$$j = \left\{\frac{h^2\,C}{2\pi m_{Br}\,kT}\exp\left[(E_{SBr} + \tfrac{1}{2}E_d - E_a)/RT\right]\right\}\frac{n_{Br}}{N_A}\left(\frac{kT}{2\pi m_{Br}}\right)^{\frac{1}{2}}. \tag{12}$$

In this formulation the reaction rate is represented as the product of the sticking coefficient (the term between braces) and the molar collision frequency. From fig. 7 and the empiric relation eq. (4) derived from it, it is known that the term in the exponent has to be practically equal to zero. In this form eq. (12) can be equated to eq. (4) here ignoring the last term. Taking $C = 1.9\times10^{15}$ atoms per cm^2 (ref. 24) and $\bar{T} = 1166$ K, the theoretical value of k_1 is 9.2×10^{-10} mol s^{-1} cm^{-2} Pa^{-1}, which is about three orders of magnitude larger than the experimental one. In reaction kinetics this can still be considered as a reasonable agreement between theory and experiment. Also it can be seen from eq. (12) that in the presentation of the data in fig. 7 the non-exponential temperature dependence of the reaction rate, being $T^{-\frac{1}{2}}$, is ignored. In conclusion, it can be stated that the reactivity of the W–Br reaction in the lower Br_2 pressure regime can be explained satisfactorily from the principles found in the "vacuum" experiments.

Now we will deal with the W–O reaction. Here too we start with the findings from the LEIS experiment which indicate that at "low" temperatures and high Br : O ratios in the gas phase the chemisorbed monolayer is filled with bromine. Certainly in the range where the W–Br reaction rate is equal or larger than the W–O rate, this condition is fulfilled. Departing from that situation in the direction of decreasing Br : O ratios in the gas phase, somewhere before surface oxidation starts, oxygen will appear in significant concentrations in the adsorption layer. However, because, as can be seen from fig. 6, the rate characteristic does not change very much over most of the range, we assume that in the W–O reaction too the tungsten surface is almost completely covered with a chemisorbed monolayer of bromine.

This means that, under this assumption, eqs (9) and (10) are still valid. First we will try to find the structure of the activated complex. Since complexes containing more than two tungsten surface atoms seem highly improbable, we get $q = 1$ or $q = 2$ (see eq. (7)). From table II we learn that the partial pressure dependence of j is a function of temperature. Since $b = f$ and $(a - q/2) = g$, one can use the approach that at 1067 K, $b = 1.5$ and $(a - q/2) = -1$, and at 1282 K $b = 1$ and $(a - q/2) = -0.5$. This gives $S_2O_3^{\neq}$ at 1067 K and $SO_2^{\neq}$ or $S_2BrO_2^{\neq}$ at 1282 K as the structure of the critical activated complex. To be consistent with this finding we drop $S_2BrO_2^{\neq}$ rather arbitrarily. This leads to the somewhat speculative conclusion that in the W–O reaction the critical step is the formation of an oxygen adsorption complex. At lower temperatures in this activated complex three oxygen atoms are bonded to two tungsten substrate atoms, at higher temperatures two to one. These tungsten to oxygen ratios are quite normal in surface complexes [31]). The mechanism then is of the Langmuir–Hinselwood type [30]) in which the reagent, oxygen, first enters the few positions left open by bromine in the chemisorption layer in a dissociative adsorption process, the next step being diffusion towards the point where the active complex is formed. Since the pressure dependence of the W–O reaction is also temperature-dependent, within the scope of our experiments no statement can be given on the value of the activation energy or the theoretical value of the reaction constant in eq. (10). Although the model given is simpler than the one given by Zubler [15]), who needs a two-layer adsorption model, from the generally accepted theories on reaction rate and on adsorption we can, at least give a quälitative explanation of the facts observed in experiments. We believe that a second physical adsorption layer can be present under the experimental conditions, but that it is of little importance because of its small binding energy.

Adsorption and desorption in that layer can therefore be considered as an equilibrium process, being one of the steps preceding the rate-limiting step.

3.2. *Kinetics in halogen incandescent lamps*

In this section we will confine ourselves to the consequences of the experimental findings discussed above for the interpretation of tungsten-transport phenomena in halogen incandescent lamps. We will evaluate whether the physical process of normal diffusion under LTE conditions or the chemical reaction at the tungsten surface is the rate determining step for tungsten transport along the tail ends. Our approach will be rather academic from the production point of view, since we will consider the case of a lamp filled with bromine and oxygen as active gases. In the lower temperature regime then WBr_4 and WO_2Br_2 are the most important tungsten-carrying species. These species dissociate in the temperature range of 1000–1500 and 1600–2100 K respectively. We will consider axial mass transport along the tungsten tail end and apply Fick's law to the WBr_4 transport (WO_2Br_2 is left out of the discussion because no data are available on the reaction rate in the W–O_2–Br_2 system at temperatures in the dissociation range):

$$j_{\mathrm{WBr_4}} = -\frac{\bar{D}_{\mathrm{WBr_4}}}{R\bar{T}} \frac{\partial p_{\mathrm{WBr_4}}}{\partial z}. \tag{13}$$

Here $\bar{D}_i$ is the binary diffusion coefficient of the species i in the inert gas (in the calculation Ar at 4×10^5 Pa) at the mean temperature, $\bar{T}$ (1250 K) and the z axis is taken parallel to the axis of the tail end. In this order of magnitude approach we take a two-step model with solid tungsten and WBr_4 as the only relevant species at 1000 K, and solid tungsten and Br at 1500 K. The mass balance for bromine is then given by

$$j_{\mathrm{Br}} + 4\,j_{\mathrm{WBr_4}} = 0. \tag{14}$$

This equation can be rewritten by the substitution of eq. (13) in its integrated form

$$\frac{\bar{D}_{\mathrm{Br}}}{R\bar{T}} \frac{p_{\mathrm{Br}}(1500)}{\Delta z} - \frac{4\bar{D}_{\mathrm{WBr_4}}}{R\bar{T}} \frac{p_{\mathrm{WBr_4}}(1000)}{\Delta z} = 0. \tag{15}$$

Here Δz is the length of the dissociation range (0.1 cm). Now by substitution of this result in the integrated form of eq. (13) one derives

$$j_{\mathrm{WBr_4}} = \frac{\bar{D}_{\mathrm{Br}}}{R\bar{T}} \frac{p_{\mathrm{Br}}(1500)}{4\,\Delta z}. \tag{16}$$

The numerical result of this tungsten transport equation, taking $\bar{D}_{\mathrm{Br}}$ as

1.4 cm^2 s^{-1}, is given by

$$j_W = j_{WBr_4} = 3.4\times 10^{-10}\, p_{Br}\,(1500). \qquad (17)$$

From the value of j_W calculated from this equation it can be seen that the transport rate in the LTE model far exceeds the chemical reaction rate (see eq. (4) and fig. 7). So from the considerations given above it can be concluded that under more or less "normal" lamp situations the chemical reaction at the surface is the rate-determining step in tail erosion. Therefore the "classic" LTE model, at least in the situation sketched above, does not apply and has to be changed for a more sophisticated model in which chemical reaction rates are taken into account.

Acknowledgement

We thank Mr J. H. Dettingmeijer, Mr W. A. Klessens and Mr F. G. P. Sools for their efforts in clearing our way by starting the gas-flow experiment and for stimulating discussions.

Philips Lighting Division *Eindhoven, February 1978*

REFERENCES

[1] M. J. Duell and R. L. Moss, Trans. Far. Soc. **61**, 2262, 1965.
[2] B. McCarrol, in J. Wiley (ed.), Proc. 4th Int. Mat. Symp. Berkeley 1968, 61-1.
[3] B. McCarrol, J. Appl. Phys. **40**, 1, 1969.
[4] D. L. Fehrs and R. E. Stickney, Surf. Sci. **17**, 298, 1969.
[5] C. W. Jowett and B. J. Hopkins, Surf. Sci. **22**, 392, 1970.
[6] D. E. Rosner and H. D. Allendorf, AIAA J. **5**, 1489, 1967.
[7] D. E. Rosner and H. D. Allendorf, J. Phys. Chem. **75**, 308, 1971.
[8] P. C. Abott and R. E. Stickney, J. Phys. Chem. **76**, 2930, 1972.
[9] N. R. Avery, Surf. Sci. **43**, 101, 1974.
[10] N. R. Avery, Proc. 2nd Int. Conf. Solid Surf. 1974, Japan, J. Appl. Phys. Suppl. 2, Part 2, 193, 1974.
[11] K. Faulian and E. Bauer, Phys. Lett. **54A**, 313, 1975.
[12] A. Landsberg, J. Electrochem. Soc. **118**, 1331, 1971.
[13] H. Schaeffer, Z. Anorg. Allg. Chem. **400**, 242, 1973.
[14] E. G. Zubler, J. Phys. Chem. **74**, 2479, 1970.
[15] E. G. Zubler, J. Phys. Chem. **76**, 320, 1972.
[16] E. G. Zubler, J. Phys. Chem. **79**, 1703, 1975.
[17] J. D. McKinley, in J. W. Mitchell et al. (eds), Reactivity of solids, Wiley, N.Y., 1969.
[18] V. W. Goddard and C. Pett, J. Chem. Soc. Dalton **9**, 767, 1975.
[19] J. H. Dettingmeijer, G. Dittmer, A. Klopfer and J. Schröder, Philips Techn. T. **35**, 315, 1975.
[20] W. J. van den Hoek and G. Rouweler, Philips Res. Repts **31**, 23, 1976.
[21] F. J. Harvey, Metal. Trans. A. **7A**, 1167, 1976.
[22] S. K. Gupta, Lecture at the High Temperature Metal Halide Chemistry Symposium, Atlanta, 1977.
[23] J. F. Waymouth, Lecture at the High Temperature Metal Halide Chemistry Symposium, Atlanta, 1977.
[24] H. H. Brongersma, G. C. J. van der Ligt and G. Rouweler, to be published in Philips J. Res.

[25] R. W. Bartlett, J. Electrochem. Soc.: Electrochem. Sci. **114**, 547, 1967.
[26] R. B. Bird, W. E. Stewart and E. N. Lightfoot, Transport phenomena, Chs 13, 21, J. Wiley and Sons, N.Y., 1960.
[27] G. M. Neumann, Z. Metallkunde **64**, 26, 1973.
[28] D. R. Stull and H. Prophet, JANAF Thermochemical Tables, 2nd edn., NSRDS-NBS 37, Ed. Nat. Stand. Ref. Data Ser., Nat. Bur. Stand., 1971.
[29] C. H. Banford and C. F. H. Tipper, Comprehensive Chemical Kinetics, Vol. 13, Elsevier, 1976.
[30] V. Ponec, Z. Knor and S. Cerny, Adsorption on solids, Butterworths, London, 1974.
[31] T. E. Madey, J. J. Czyzewski and J. T. Yates, Surf. Sci. **49**, 765, 1975.

MORPHOLOGICAL STABILITY ANALYSIS OF GROWTH FROM THE VAPOUR

by C. H. J. van den BREKEL

Abstract

Morphological stability of vapour growth processes has been studied by means of a first-order perturbation analysis. A criterion for absolutely stable growth is given. From this criterion, experimental data can be analysed and a conclusive test of the stability theory can be made.

1. Introduction

Studies on the morphological stability of chemical vapour-deposition (CVD) processes have been published recently [1,2]. It was found that the growth of CVD layers is essentially unstable. Relevant time constants, however, are often much longer than practical times of experiments, which explains why an apparent stability is observed.

The analysis given in a previous study (ref. 2) rested upon the assumption that the temperature of the perturbed interface during deposition is uniform. This assumption was reasoned from three arguments: (1) the relatively high heat conductivity of the solid, (2) the fact that the reaction heat flux is small as compared to the heat flux in the system and (3) the fact that the radiation of heat levels out possible differences in temperature. The two latter premises, however, do not apply for CVD processes carried out at low interface temperatures. In that case the constant interface temperature assumption may certainly not be made a priori.

In this paper we therefore develop a first-order perturbation stability analysis in which the coupling between heat and mass transport is taken into account. The result obtained will be compared with the stability analysis of the solidification of binary alloys developed by Mullins and Sekerka [3], hereafter referred to as MS treatment.

1.1. *First-order perturbation analysis*

Mass transport of reactive compound AB in vapour growth processes is assumed to take place by diffusion over a boundary layer of thickness δ and by mass transfer at the substrate interface [4,5]. This boundary layer coincides with the thermal and the stagnant momentum boundary layers [6]. Further, the same assumptions as made in a previous paper [2]) are made here: viz. no homogeneous gas-phase decomposition of compound AB, negligibly small interface velocity as compared to the diffusion rate, the surface parameters are isotropic and the surface reaction rate is proportional to the supersaturation.

To avoid unnecessarily complicated expressions the diffusivity is assumed to be independent of the gas-phase temperature. As shown before [2]), even for relatively large thermal gradients in the gas phase, such an assumption can be made without changing the accuracy of the result by more than a factor of three.

We consider a rectangular coordinate system XZ, where the X axis coincides with the gas–solid interface, the positive Z axis pointing into the gas phase. The stagnant boundary layer extends from $z = 0$ to $z = \delta$. The substrate of thickness d is positioned between $z = 0$ and $z = -d$.

As an arbitrarily perturbed interface can be Fourier analysed, it is sufficient to consider a sinusoidally perturbed interface $z(t) = \varepsilon(t) \sin \omega x$, where the parameter t relates to time, ε and ω are the amplitude and spatial frequency respectively.

In order to calculate how the amplitude of the perturbations $\varepsilon(t)$ varies with time, we must find the velocity of each element of the interface in terms of the local thermal and diffusion gradients. This requires the determination of the thermal and diffusion fields.

The steady-state differential equations for mass transport and heat transport read

$$\nabla^2 c = 0, \tag{1a}$$

$$\nabla^2 T_g = 0, \tag{1b}$$

$$\nabla^2 T_s = 0, \tag{1c}$$

where c is the concentration of reactive compound AB and T_g and T_s the temperatures in the gas phase and solid phase respectively. The set of Laplace equations (1a) to (1c) must be solved subject to following boundary conditions:

(1) conservation of mass at the interface [1]):

$$k_D (c - c_{eq}{}^r) = -\boldsymbol{J} \cdot \boldsymbol{n}, \tag{2a}$$

where k_D is the mass-transfer coefficient, $c_{eq}{}^r$ the equilibrium concentration at a curved interface with radius of curvature r, $\boldsymbol{J}$ the mass flux and $\boldsymbol{n}$ the unit normal vector which points into the gas phase;

(2) conservation of heat at the interface:

$$\frac{\Delta H}{N} J \cdot n = (-K_g \nabla T_g + K_s \nabla T_s) \cdot n + a\sigma T^4, \qquad (2b)$$

where ΔH is the reaction enthalpy per mol ($\Delta H < 0$ for an exothermal reaction), N the Avogadro number, K the heat conductivity, the subscripts g and s refering to gas and solid phase, a the emission coefficient, $\sigma = 1.35 \times 10^{-12}$ cal/K^4 s cm^2 the Stefan–Boltzmann radiation constant and T the interface temperature;

(3) continuity of temperature at the interface

$$T_s\,(x, \varepsilon \sin \omega x) = T_g\,(x, \varepsilon \sin \omega x); \qquad (2c)$$

(4) at the upper edge of the stagnant boundary layer we require

$$T_g\,(x, \delta) = T_b \qquad (2d)$$

and

$$c\,(x, \delta) = c_b, \qquad (2e)$$

where T_b and c_b are the temperature and concentration of the bulk gas;

(5) at the back-side of the substrate the temperature equals T_1

$$T_s\,(x, -d) = T_1. \qquad (2f)$$

When $\varepsilon(t) \ll 2\pi/\omega$, the solutions of the Laplace equations (1a) to (1c) can be written as the superposition of the unperturbed solution and a perturbation term. This yields

$$c(x, z) = c_0 + G_c\, z + (A_c \exp \omega z + B_c \exp -\omega z)\, \varepsilon \sin \omega x, \qquad (3a)$$

$$T_g(x, z) = T_0 + G_g\, z + (A_g \exp \omega z + B_g \exp -\omega z)\, \varepsilon \sin \omega x, \qquad (3b)$$

$$T_s(x, z) = T_0 + G_s\, z + (A_s \exp \omega z + B_s \exp -\omega z)\, \varepsilon \sin \omega x, \qquad (3c)$$

where c_0 and T_0 are the interface concentration and temperature respectively at a planar interface, G_c the concentration gradient, G_g the thermal gradient in the gas phase, G_s the thermal gradient in the solid phase at an unperturbed interface and A_c, A_g, A_s, B_c, B_g and B_s are constants. The unperturbed solutions are linear functions of z. The perturbation terms are proportional to the surface perturbation itself. They are the first-order corrections to the unperturbed fields corresponding to the surface harmonic. The form of eqs (3) may be regarded as justified when self-consistent values are obtained for the sets of coefficients A and B. By inspection it can be verified that (3a) to (3c) are solutions of (1a) to (1b) indeed.

The interface temperature is modulated (c.f. expressions (3b) and (3c)); the modulation (ΔT) is proportional to the interface perturbation:

$$\Delta T = T(x, \varepsilon \sin \omega x) - T_0 = (G_s + A_s + B_s)\, \varepsilon \sin \omega x. \tag{4}$$

This means that the temperature dependence of k_D and $c_{eq}{}^r$ must be taken into account. The mass transfer coefficient k_D is the rate constant of the heterogeneous decomposition reaction and varies exponentially with the interface temperature T

$$k_D = k_0 \exp\left(-\frac{\Delta E}{RT}\right), \tag{5}$$

where k_0 is a pre-exponential factor, ΔE the apparent activation energy and R the gas constant. Expanding expression (5) in a Taylor series around T_0 and truncating after the first-order term in ε yields

$$k_D(T) = k_D(T_0)\left(1 + \frac{\Delta E}{R{T_0}^2}\, \Delta T\right). \tag{6}$$

The equilibrium constant K_p of the decomposition reaction

$$AB_g \rightleftarrows A_s + B_g \tag{7}$$

(subscripts g and s refer to volatile and solid molecules respectively) is given by

$$K_p = \frac{c_{eq}{}^{B}}{c_{eq}{}^{AB}} = \exp\left(-\frac{\Delta G}{RT_0}\right), \tag{8}$$

where $\Delta G = \Delta H - T_0\, \Delta S$ is the change of the Gibbs free energy and ΔS the entropy change involved with reaction (7). From eq. (8) it follows for a slightly varying interface temperature that in first-order approximation the equilibrium concentration of AB at the interface is given by

$$c_{eq}{}^{AB}(T) = c_{eq}{}^{AB}(T_0)\left(1 - \frac{\Delta H}{R{T_0}^2}\, \Delta T\right). \tag{9}$$

Expression (9) has been derived under the assumption that ΔH and ΔS are temperature-independent within the considered temperature range. The equilibrium concentration $c_{eq}{}^{r}$ at a curved interface depends not only on the temperature but also on the interface curvature as given by the Gibbs–Thomson equation [7]); hence (the index AB is now omitted)

$$c_{eq}{}^{r} = c_{eq}{}^{\infty}\left(1 - \frac{\Gamma}{r}\right)\left(1 - \frac{\Delta H}{RT_0{}^2}\,\Delta T\right), \tag{10}$$

where $c_{eq}{}^{\infty}$ is the equilibrium concentration at a planar interface at temperature T_0 and Γ the capillarity constant.

The six constants A_c, A_g, A_s, B_c, B_g and B_s appearing in eqs (3) can now be determined using boundary conditions (2a) to (2f), and eqs (4), (6) and (10). The local flux can then be calculated from the known concentration field eq. (3a). The relation

$$V = -\frac{M}{\varrho N}\,J, \tag{11}$$

where V is the local advance rate of the interface, M the molecular weight and ϱ the specific density of the solid material, combined with

$$V = V_0 + \dot{\varepsilon}\sin\omega x, \tag{12}$$

where V_0 is the constant unperturbed growth rate and $\dot{\varepsilon} = \mathrm{d}\varepsilon/\mathrm{d}t$, yields

$$\frac{\dot{\varepsilon}}{\varepsilon} = V_0\,\omega\left[\frac{(G_c - c_{eq}{}^{\infty}\,\Gamma\,\omega^2)\,K + l\,(K_s'\,G_s + K_g'\,G_g)}{K\,(\tanh\omega\delta + \omega\delta/Nu)\,G_c + l\,(K_g\,G_g - K_s\,G_s - \sigma a T_0{}^4)}\right], \tag{13}$$

where

$$K = K_g' + K_s' + K_r, \tag{14a}$$

$$K_s' = K_s/\tanh\omega d, \quad K_g' = K_g/\tanh\omega\delta, \quad K_r = \frac{4aT_0{}^3}{\omega}, \tag{14b}$$

$$l = c_{eq}{}^{\infty}\,\frac{\Delta H}{RT_0{}^2} + (c_0 - c_{eq}{}^{\infty})\,\frac{\Delta E}{RT_0{}^2}, \tag{15}$$

and

$$Nu = k_D\,\delta/D. \tag{16}$$

Integration of (13) gives *)

$$\varepsilon(t) = \varepsilon(0) \exp [f(\omega)\, t], \tag{17}$$

where $\varepsilon(0)$ is the value of ε at time $t = 0$ and $f(\omega)$ the stability function as explicitely given by the right-hand side of eq. (13).

From eq. (17) it follows that Fourier components whose corresponding $f(\omega)$ is positive will increase exponentially with time, while those whose corresponding $f(\omega)$ is negative will decay exponentially with time. We will therefore evaluate the frequency dependence of the stability function in the next section.

2. Discussion

First it should be noted that the present model is only realistic when $\delta \gg 2\pi/\omega$, since otherwise the upper boundary edge will be perturbed (see appendix of ref. 1). This means that our result eq. (13) is valid when $\omega\delta \gg 2\pi$, hence $\tanh \omega\delta \approx 1$ in the relevant frequency range. Because the boundary-layer thickness is of the order of millimeters, the relevant frequency range is $\omega > 100\ \mathrm{cm}^{-1}$. This means that for not too thin substrates (e.g. $d \geqslant 250$ μm) also the term $\tanh \omega d$ can be approximated by unity. Using the definitions for the conductivity-weighted temperature gradients

$$G \equiv \frac{2\, K_g\, G_g}{K}, \tag{18a}$$

$$G' \equiv \frac{2\, K_s\, G_s}{K}, \tag{18b}$$

and

$$G'' \equiv \frac{2\, a\sigma T_0{}^4}{K}, \tag{18c}$$

expression (13) is simplified into

$$\frac{\dot{\varepsilon}}{\varepsilon} = V_0\, \omega \left[\frac{2\, G_c - 2\, c_{eq}{}^{\infty}\, \Gamma\, \omega^2 + l\,(G + G')}{2\,(1 + \omega\delta/Nu)\, G_c + l\,(G - G' - G'')}\right]. \tag{19}$$

The effect of heat radiation on $\dot{\varepsilon}/\varepsilon$, which is absent in the MS treatment, is evidenced in two ways: viz. by the factors K_r and G''. A finite value of the parameter K_r, which depends on the spatial frequency of the considered per-

*) d is supposed to be so large that $\tanh \omega d \approx 1$.

turbation, contributes to the decrease of the absolute value of the conductivity-weighted gradients. For most CVD processes, however, this effect is negligibly small, since K_r is only of importance in K at exceptionally high interface temperatures. The parameter K_r becomes of the order of K_s, which is the main term in K, when $T_0^3 \geqslant \omega K_s/4a\sigma$. So, with $K_s = 5 \times 10^{-2}$ cal/K cm^2 s and $a = 0.5$ we obtain for the frequency range of interest ($\omega > 10^2$ cm^{-1}) $T_0 \geqslant 12000$ K.

The parameter G'' is of the order of G at substrate temperatures roughly above 900 K (for H_2, $K_g \approx 5 \times 10^{-4}$ cal/K cm^2 s and $G_g \approx 10^3$ K/cm). The effect of G'' on stability depends on the sign of l, but since G'' only figures in the denominator of eq. (19) the effect is small.

Apart from the radiation term, the expression of eq. (19) for infinite Nu number (diffusion-limited growth) shows a striking resemblance with the $\dot{\delta}/\delta$ expression as derived in the MS treatment (see formula 15.43 in ref. 3). Comparison reveals that the MS parameters m and T_M play the same role and have the same dimensions as $1/l$ and c_{eq}/l, respectively, for vapour growth; l and c_{eq}^∞ contain thermodynamical and experimental parameters of the process. The characteristic process parameter l (eq. (15)) is the key in the translation of the thermal gradient effect into the mass gradient effect. The parameter l consists of an equilibrium term, containing the reaction enthalpy difference, and a kinetic term, containing the activation energy of the process. Both terms are the product of the relevant interface concentration and a Van 't Hoff constant; the reciprocal terms are analogous to such expressions as the ebullioscopic constant. The Van 't Hoff constants express the temperature dependence of the chemical equilibrium $\partial \log K_p/\partial T$ and of the reaction rate $\partial \log W/\partial T$ respectively. The second term of (15) only contributes to l when the deposition process is kinetically controlled, because the supersaturation $c_0 - c_{eq}$ vanishes at large Nu numbers. For diffusion-limited growth the equilibrium term remains; it is the CVD analog of the liquidus slope m in the MS treatment.

In order to study the frequency dependence of $\dot{\varepsilon}/\varepsilon$ we first notice that, as shown in the appendix, the denominator of $\dot{\varepsilon}/\varepsilon$, is positive, as it should be for physical reasons. For, a negative denominator reverses the sign of the capillarity term, which implies that under such circumstances capillarity would favour unstable growth.

Equation (19) shows that $\dot{\varepsilon}/\varepsilon$ is composed of three terms: a positive term proportional to G_c, which represents the mass-gradient effect favouring growth of the perturbation; a negative term proportional to Γ, which represents the capillarity effect favouring decay of the perturbation; and a term proportional to $l\,(G + G')$, which represents the effect of heat transport.

The mass-gradient effect is present when Γ and $G + G'$ vanish. If the deposition process is diffusion-limited ($Nu > 1$) the iso-concentration lines in the gas phase follow the interface shape, but they are bunched together above the protrusions and rarified above the recessions [4]). This results in growth of the perturbation amplitude due to preferential supply at protrusions.

If the deposition process is surface-controlled ($Nu < 1$), the curvature of the iso-concentration lines is reversed. This pattern just compensates for the effect of interface geometry on mass transport, in that the interface flux at protrusions equals the flux in the deep valleys. In first approximation, therefore, the amplitude of the harmonic remains unaffected during growth, which also follows from the value of the stability function, since G_c approximates zero for vanishing Nu.

The capillarity effect influences the local equilibrium concentration; it decreases the supersaturation at protrusions, whereas it increases the supersaturation in the valleys, which results in a decreasing perturbation amplitude. This stabilizing effect originates from the surface tension, which tries to minimize the interface surface.

The size of the thermal gradient term depends on the nature of the surface process ΔH and ΔE and on the experimental conditions Nu, among them the sign of $G + G'$. The sign of the latter parameter can be chosen arbitrarily by the operator.

Again, two growth modes can be distinguished: diffusion-limited and surface-controlled growth. Let us first consider a diffusion-limited process, hence $l = c_{eq}\,\Delta H/RT_0{}^2$. When $G + G'$ is negative (substrate hot with respect to the gas phase), the protrusions of the surface harmonic are slightly cooler than the valleys. The Van 't Hoff theorem states that, when $\Delta H > 0$ the equilibrium concentration in the valleys is slightly lower than at the protrusion. The surface amplitude will decay because of the lower growth rate at the protrusions as compared to that in the valleys. This conclusion is consistent with the negative sign of $l\,(G + G')$ in eq. (19), which indicates a stabilizing effect.

When on the other hand the decomposition reaction is exothermal ($\Delta H < 0$) and $G + G'$ is still negative, the protrusions will show enhanced growth, because the supersaturation at the cooler peaks is greater than at the hotter valleys. This destabilizing effect is evidenced by the positive sign of the $l\,(G + G')$ term in this case. By the same reasoning a positive $G + G'$ predicts stability for endothermal reactions and instability for exothermal reactions.

We will now briefly discuss the case where the deposition process is not purely diffusion-limited. From eq. (15) it follows that, because ΔE is positive, l is positive in the case of an endothermal reaction, whereas in the case of an exothermal reaction the sign of l depends on the supersaturation, thus on the value of Nu. In the latter case no general conclusions on stability can be drawn.

The sign of $\dot{\varepsilon}/\varepsilon$ is therefore determined by the sign of the numerator of (20). Inspection reveals that $\dot{\varepsilon}/\varepsilon$ is negative for any frequency, hence the deposition process is absolutely stable when

$$2\,G_c < -l\,(G + G'). \tag{20}$$

Criterion (20) is similar to the modified constitutional super cooling stability criterion as found in the MS treatment. It should be noted that $G + G'$ in the MS treatment is a positive term, whereas in vapour growth processes it may either be positive or negative.

Condition (20) can in principle be satisfied by reducing G_c sufficiently, provided the sign of $l\,(G + G')$ is negative. Reduction of G_c is realized by either lowering c_b or by decreasing T_0, hence the Nu number. Both methods affect G_c and l in a complex way. It is therefore interesting to investigate whether the criterion is fulfilled in an important deposition process: the deposition of Si from $SiHCl_3$ in a hydrogen atmosphere.

2.1. *Chemical vapour deposition of silicon*

Since the deposition process of Si at high temperatures ($T_0 > 1300$ K) is diffusion-limited, l may be calculated from thermodynamical data. Table I gives values of ΔH, partial pressure and l for $T_0 = 1400$–1600 K. The sum of the conductivity-weighted thermal gradients in this temperature range has been estimated with the aid of data given by Bloem [8]) to be $G + G' \approx -200$ K/cm. For $c_b = 0.1$ vol.%, $\delta = 0.3$ cm and $T_b = 1000$ K we obtain $G_c \approx 2.5\times 10^{16}$ cm^{-4}. Thus, for absolute stability l must be larger than 2.5×10^{14} cm^{-3} K^{-2}.

The calculated l values (table I) are seen to be smaller by seven orders of magnitude, which shows that absolute stability is impossible under these con-

TABLE I

Reaction enthalpies of the decomposition of $SiHCl_3$, partial vapour pressures of $SiHCl_3$ and l values calculated using eq. (15)

T (K)	ΔH *) (cal/mol)	P_{eq} **) (dyne/cm^2)	l (cm^{-3} K^{-1})
1400	43815	2.2×10^{-4}	1.3×10^{7}
1500	43651	7.0×10^{-5}	3.4×10^{6}
1600	43490	2.6×10^{-5}	1.0×10^{6}

*) Calculated from Janaf Tables [10]).
**) Data given by P. van der Putte [11]), based on Cl/H = 1.5×10^{-2}.

ditions. This is in agreement with experimental observations [1]. Moreover, it is clear that the stabilizing thermal gradient effect, though present, is negligibly small as compared to the destabilizing mass-gradient effect. This shows that $\dot{\varepsilon}/\varepsilon$ is described accurately enough by the previously given stability function expression [1].

At lower surface temperatures, where the deposition process is kinetically controlled, l cannot be approximated by the first term of eq. (15) alone; upon decreasing T_0 the second term becomes important. The activation energy ΔE of the deposition process of Si is found to be 40 kcal [9], which roughly equals the reaction enthalpy of the reduction of $SiHCl_3$. Thus, for vanishing Nu, ($c_0 \approx c_b$), l may be approximated by

$$l \approx c_b \frac{\Delta E}{R{T_0}^2} \tag{21}$$

which, with $c_b = 0.1$ vol.% and $T_0 = 1000$ K, yields $l = 1.5 \times 10^{14}$ cm^{-3} K^{-1}. The term $G + G'$ at this temperature is estimated [8] to be $G + G' = -60$ K/cm. The measurements reported by Kroon [9] give a growth rate $V = 0.004$ μm/min at 700 °C, hence with $D = 1$ cm^2/s and using eq. (11) we find $G_c = 3.3 \times 10^{14}$ cm^{-4}, which shows that under the given conditions the process could be absolutely stable, since the stability condition is satisfied.

Due to the use of expression (21) the calculated value of l is a rough approximation, so it is not certain that the conclusion above is correct for $T_0 = 1000$ K. $T_0 = 1000$ K. Nevertheless, it will be clear that absolute stability can be realized at sufficiently low temperatures, because G_c decreases exponentially with T_0, while l is proportional to $1/{T_0}^2$.

In a previous paper [1] based on a simpler theory the absence of unstable growth in experiments in this temperature range was attributed to an apparent stability. This conclusion was justified by the calculated relaxation time, which for $Nu = 0.1$ was found to be 3×10^5 s $= 83$ h. A decisive experiment on absolute stability should therefore last for more than four days.

Philips Research Laboratories *Eindhoven, January 1978*

Appendix

The denominator of the stability function contains a number of material constants and process parameters. Since they can have different values, it cannot be reasoned as such that the denominator is always positive. Substitution of reasonable values, however, shows that in general this is the case. With the aid of the zero-order boundary condition (2) and using the definition of l we obtain for the denominator of eq. (20)

$$\left\{\frac{\Delta H}{N}\,\frac{D}{K}\left[c_{\mathrm{eq}}{}^{\infty}\,\frac{\Delta H}{RT_0{}^2} + (c_0 - c_{\mathrm{eq}}{}^{\infty})\,\frac{\Delta E}{RT_0{}^2}\right] + 2\left(1 + \frac{\omega\delta}{Nu}\right)\right\} G_{\mathrm{c}}. \qquad \text{(A1)}$$

The term containing the supersaturation $c_0 - c_{\mathrm{eq}}{}^{\infty}$ depends on the Nu number and can be rewritten using (2a), (11) and (16). After rearrangement expression (A1) is then converted into

$$2\left[A + 1 + (B + \omega)\,\frac{\delta}{Nu}\right] G_{\mathrm{c}}, \qquad \text{(A2)}$$

where

$$A = \frac{c_{\mathrm{eq}}{}^{\infty}}{2}\,\frac{D}{NK}\,\frac{\Delta H^2}{RT_0{}^2} \quad \text{and} \quad B = \frac{V\varrho}{2\,MK}\,\frac{\Delta H\,\Delta E}{RT_0{}^2}\,.$$

The term A of expression (A2) is usually small with respect to unity. This may be shown by substitution of typical values (referring to the deposition of Si from $\mathrm{SiHCl_3}$): $D = 3$ cm^2/s, $K = 5.5\times10^{-2}$ cal/s K cm^2, $\Delta H = 4\times10^4$ cal, which at $T_0 \geqslant 500$ K and $c_{\mathrm{eq}} = 0.1$ vol.% $= 10^{16}$ cm^{-3}, yields $A \leqslant 1.6\times10^{-3}$.

Expression (A2) is positive under diffusion-limited growth conditions ($Nu \rightarrow \infty$ and $G_{\mathrm{c}} > 0$), where the last terms vanish. For finite values of Nu the term $(B + \omega)$ must be considered in detail, because B is positive or negative, depending on whether the surface reaction is endothermal ($\Delta H > 0$) or exothermal ($\Delta H < 0$).

It is clear that the factor B and thus expression (A2) is positive when ΔH is positive. Substitution of typical values for a case where ΔH is negative, e.g. $\Delta H = -4\times10^4$ cal and $V = 1$ μm/min, $M = 30$, $\varrho = 2$ g/cm^3, $\Delta E = 4\times10^4$ cal/mol, yields $B = -2\times10^4/T_0{}^2$. This shows that even at room temperature the factor B may be ignored with respect to ω, since the relevant frequency range for CVD processes [1]) comprises $\omega > 10^2$ cm^{-1}. In conclusion we may say that the denominator of expression (19) is positive.

REFERENCES

[1]) C. H. J. van den Brekel and A. K. Jansen, J. Crystal Growth **43**, 364, 1978.
[2]) A. K. Jansen and C. H. J. van den Brekel, J. Crystal Growth **43**, 371, 1978.
[3]) R. F. Sekerka in P. Hartman (ed.), Crystal Growth, North-Holland Publ. Co., 1973.
[4]) C. H. J. van den Brekel, Philips Res. Repts **32**, 118, 1977.
[5]) C. H. J. van den Brekel and J. Bloem, Philips Res. Repts **32**, 134, 1977.
[6]) F. C. Eversteijn, P. J. W. Severin, C. H. J. van den Brekel and H. L. Peek, J. Electrochem. Soc. **117**, 925, 1970.
[7]) W. W. Mullins, J. Appl. Phys. **30**, 77, 1959.
[8]) J. Bloem, A. H. Goemans, J. Appl. Phys. **43**, 1281, 1972.
[9]) F. C. Eversteijn, Philips Res. Repts **29**, 45, 1974.
[10]) Janaf Thermochemical Tables, 2nd ed., NSRDS-NBS 37, 1971.
[11]) P. van der Putte, private communication.

OPTIMIZATION OF MULTIVALUED DECISION ALGORITHMS

by M. DAVIO and A. THAYSE

Abstract

One studies the possibility of evaluating the local values of a discrete function by means of a specific type of algorithm called multivalued decision algorithm. The reason for this choice is its wide range of applications that encompasses both hardware and software problems. Possible applications of the multivalued decision algorithm are: combinatorial synthesis of discrete functions by means of multiplexers, sequential synthesis of discrete functions using multivalued ROM's and multiplexers, transformation and optimization of microprograms. The synthesis of Boolean functions constitutes an important particular case of the theory presented in this paper.

0. General definitions and notations

Let $\mathbf{x} = (x_{n-1}, \ldots, x_1, x_0)$ be a set of n variables; a discrete function $f(\mathbf{x})$ is a mapping $f\colon S^n \rightarrow S$ with $S = \{0, 1, \ldots, m-1\}$.

The lattice exponentiation is defined as follows:

$$x_i^{(C_i)} = m-1 \quad \text{iff } x_i \in C_i \quad \text{with } C_i \subseteq \{0, 1, \ldots, m-1\} = S,$$
$$x_i^{(C_i)} = 0 \quad \text{otherwise.}$$

For x_i a Boolean function: x_i holds for $x_i^{(1)}$ and $\bar{x}_i$ holds for $x_i^{(0)}$.

The following operation symbols are used:

$\vee$ or $\bigvee$ stands for the disjunction, i.e.: $a \vee b = \max(a, b)$; $a, b \in S$;

$\bigwedge$ or absence of symbol stands for the conjunction, i.e.: $a \wedge b = \min(a, b)$;

A cube function is a discrete function of the form

$$l \wedge \bigwedge_{i=0}^{n-1} x_i^{(C_i)}, \qquad l \in \{0, 1, \ldots, m-1\}, \qquad C_i \subseteq S \; \forall \, i.$$

An implicant of a discrete function f is a cube function smaller than f; a prime implicant of f is an implicant which is not smaller than any other implicant.

1. Introduction

We call *discrete function of n variables* any mapping

$$f: \{0, 1, \ldots, m-1\}^n \rightarrow \{0, 1, \ldots, m-1\}. \tag{1}$$

This function may be denoted as $f(x_{n-1}, \ldots, x_1, x_0)$ or, in short, $f(\mathbf{x})$. Each of the variables x_i ($i = 0, 1, \ldots, n-1$) takes its value in the set $\{0, 1, \ldots, m-1\}$. The most important particular kind of discrete functions are the *Boolean functions*, i.e. the mappings

$$f: \{0, 1\}^n \rightarrow \{0, 1\}$$

which were extensively studied in the literature during these last twenty years.

Various classical methods are available for realizing discrete functions by a combinational circuit: one could e.g. use a two-level circuit of AND-, OR-gates. These gates realize the operations of logical conjunction and disjunction respectively; consider AND- and OR-gates with e.g. p inputs: $(a_0, a_1, \ldots, a_{p-1})$, $a_i \in \{0, 1, \ldots, m-1\}$ $\forall$ *i*. These gates realize the following discrete functions respectively:

$$\text{AND-gate:} \quad \bigwedge_{i=0}^{p-1} a_i = \min \{a_0, a_1, \ldots, a_{p-1}\}$$

$$\text{OR-gate:} \quad \bigvee_{i=0}^{p-1} a_i = \max \{a_0, a_1, \ldots, a_{p-1}\}$$

The synthesis goal generally consists in obtaining an acceptable compromise between various design parameters such as e.g. the hardware cost and the response time. The essential features of the combinational synthesis are high speed and high cost and, for specific applications this compromise may happen to be unacceptable. Hence the need to investigate alternatives to the combinational approach.

In the present paper, we study the possibility of evaluating the local values of a discrete function by means of a specific type of algorithm, to be defined formally below, and called *multivalued decision algorithm*. The reason for this choice is its wide range of applications that encompasses both hardware and software problems. This point will be discussed with more details in what follows.

Definition 1. A multivalued decision algorithm is a loopfree algorithm containing only a finite number of labelled instructions of the type

$$N_j \, x_i \, s_0 \, s_1 \ldots s_k \ldots s_{m-1}, \tag{2}$$

where

(a) N_j is an instruction label,

(b) x_i is a variable name,

(c) the symbols s_k denote statements of one of the following two types:

— go to label N_l (non terminal statement),

— the value of the function is h (terminal statement).

The instruction (2) is interpreted as: if the variable x_i has the value k, then execute the statement s_k. It thus corresponds to the *switch* instruction in ALGOL and to the *case* instruction in PASCAL. If the variable x_i is binary, instruction (2) reduces to a classical *if* ... *then* ... *else* instruction. Furthermore, in the case of Boolean functions a multivalued decision algorithm is called a *binary decision algorithm.*

A multivalued decision algorithm is displayed as a labelled graph: we associate a node marked (j, x_i) to each instruction (2) and a node marked h $(h \in \{0, 1, \ldots, m-1\})$ to each terminal statement. From a node (j, x_i) corresponding to the instruction (2) we draw m edges marked $0, 1, \ldots, m-1$ to the nodes corresponding to the statements $s_0, s_1, \ldots, s_{m-1}$ respectively, if some of these statements are such that e.g. $s_k = s_l = \ldots = s_m$ we draw from (j, x_i) an edge labelled $k, l, \ldots, m$ to the node corresponding to the statement s_k. In conclusion, if one takes into account the fact that some of the statements may be identical, from each node (j, x_i) one draws at most m edges.

The *in-degree* of a node is the number of edges ending in this node; a multiple decision algorithm will be called a *multiple decision tree* if the in-degree of all its nodes (j, x_i) is at most 1.

Example 1. The multiple decision algorithm given by table I

TABLE I

N	x	s_0	s_1	s_2
1	x_2	N_2	N_3	N_2
2	x_1	N_4	0	0
3	x_1	N_4	N_4	2
4	x_0	0	1	2

is represented graphically in fig. 1. An easy computation shows that this algorithm defines the discrete function $f(x_2, x_1, x_0)$:

$$f: \{0, 1, 2\}^3 \rightarrow \{0, 1, 2\}$$

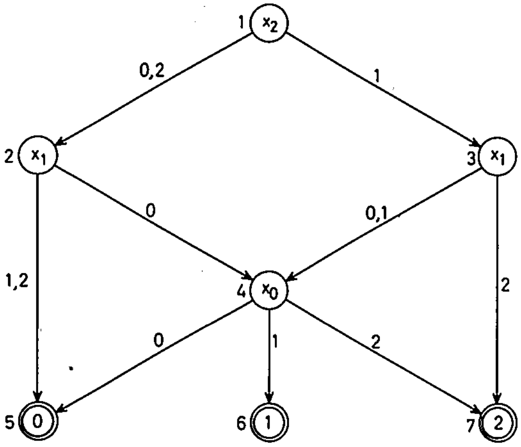


Fig. 1.

which is also described in the table II herebelow.

TABLE II

x_1				
2	0	0	0	
1	0	0	0	
0	0	1	2	
	0	1	2	x_0

$x_2 = 0$

x_1				
2	2	2	2	
1	0	1	2	
0	0	1	2	
	0	1	2	x_0

$x_2 = 1$

x_1				
2	0	0	0	
1	0	0	0	
0	0	1	2	
	0	1	2	x_0

$x_2 = 2$

It is clear from the above example that every multivalued decision algorithm completely defines a unique discrete function. Conversely, given a discrete function, their exist numbers of multivalued decision algorithms that compute that function. That observation immediately raises various optimization problems. In order to define appropriate optimization criteria, we first review some possible applications of the problem at hand.

1.1. *Combinatorial synthesis of a discrete function*

The general problem is to synthesize a discrete function with a minimum number of combinational *multiplexers.* A multiplexer is a combinational network made up of m data inputs a_i and one control input $x \in \{0, 1, \ldots, m-1\}$ realizing the following output function z:

$$z = \bigvee_{i=0}^{m-1} a_i\, x^{(i)}. \tag{3}$$

Since $a_i\, x^{(i)} = a_i$ for $x = i$,
$a_i\, x^{(i)} = 0$ for $x \neq i$,

it is clear from the comparison between (2) and (3) that any instruction N_j may be realized by a multiplexer; indeed the instruction (2) may also be written in the form

$$N_j = \bigvee_{k=0}^{m-1} s_k\, x_i^{(k)}. \tag{4}$$

This point is also illustrated in fig. 2 for the function of example 1. In this application, the hardware cost is proportional to the number of multiplexers, i.e. to the number of instructions in the multiple decision algorithm while the response time is proportional to the *logical depth*, i.e. to the maximum number of multiplexers encountered when passing from one input to the output. In

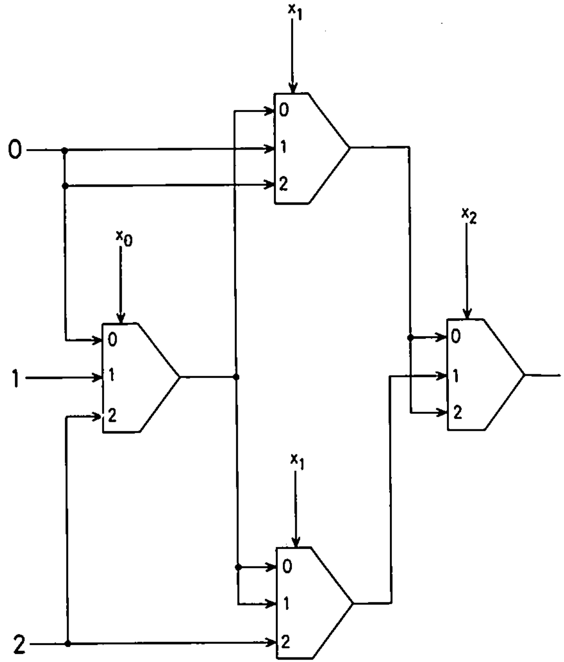


Fig. 2.

the graph model, the logical depth is represented by the maximum distance between the initial node and any one of the terminal nodes.

1.2. *Sequential synthesis of discrete functions*

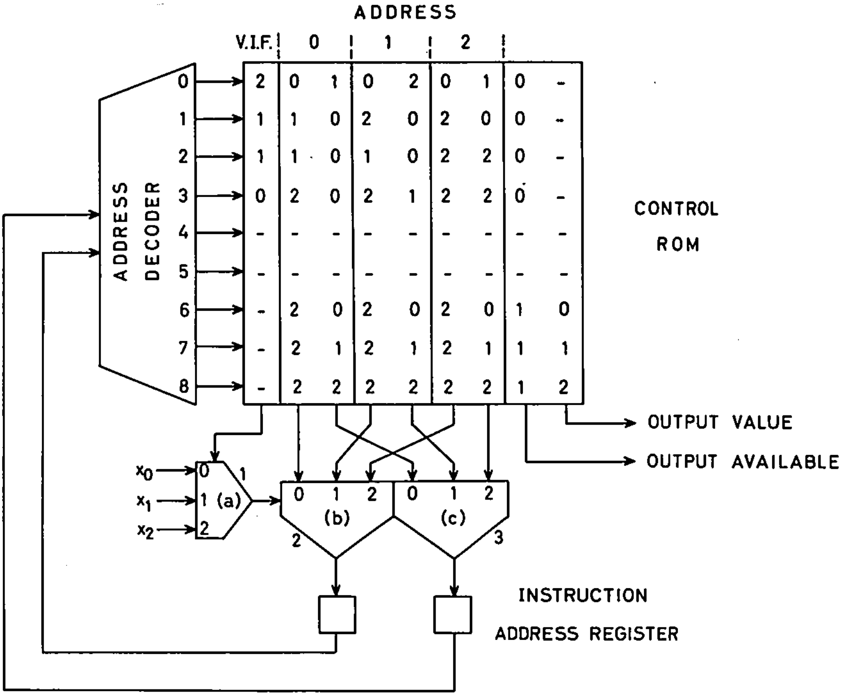
The problem of investigating binary decision algorithms seems to have first been tackled by C. Lee [1]; more recently R. Boute [2]) observed that, as any algorithm, a binary decision algorithm may be implemented as a sequential machine and, more precisely by microprogrammed-like architectures. In particular as the automaton associated with a binary decision algorithm is a condition automaton in the sense of Ito [3]), it is obviously possible to subdivide the control store word in four fields: a variable identification field (V.I.F.), two address fields and an output field (see e.g. Davio and Thayse [4]) and below). The generalization of this synthesis method to the case of multivalued decision algorithms is straightforward and will be illustrated herebelow by the function of example 1.

First let us slightly modify the algorithm in order to avoid the presence in the same column of the decision algorithm of two kinds of statements i.e. the "go to label N_k" statement and the "terminal statement". It can easily be checked that the multiple decision algorithm of example 1 may be transformed in the following one (see table III).

TABLE III

N	x	s_0	s_1	s_2	O.A.	O.V.
1	2	N_2	N_3	N_2	0	–
2	1	N_4	N_5	N_5	0	–
3	1	N_4	N_4	N_7	0	–
4	0	N_5	N_6	N_7	0	–
5	–	N_5	N_5	N_5	1	0
6	–	N_6	N_6	N_6	1	1
7	–	N_7	N_7	N_7	1	2

The presence of a "1" in the column O.A. (Output Available) indicates that the computation is completed and that the value taken by f at the point (x_2, x_1, x_0) is available in column O.V. (Output Value). The implementation of this algorithm by means of ternary Read-Only-Memories (generally referred to as ROM's), address decoders, multiplexers and registers is only a matter of encoding of the labels $N_j, j = 1$ to 7. The following encoding has been chosen in fig. 3.



	V.I.F.	0		1		2			
0	2	0	1	0	2	0	1	0	-
1	1	1	0	2	0	2	0	0	-
2	1	1	0	1	0	2	2	0	-
3	0	2	0	2	1	2	2	0	-
4	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-
6	-	2	0	2	0	2	0	1	0
7	-	2	1	2	1	2	1	1	1
8	-	2	2	2	2	2	2	1	2

Fig. 3.

N_1	N_2	N_3	N_4	N_5	N_6	N_7
00	01	02	10	20	21	22

The behaviour of the circuit may be briefly explained as follows:
— The address presently contained in the *instruction address register* selects one of the rows of the *control ROM* by means of the *address decoder.*
— The variable identification field (V.I.F.) of the selected row selects one of the variables x_0, x_1 or x_2 by means of the multiplexer *a.*
— The value of the selected input variable selects one of the three address fields and determines the next instruction address by means of the multiplexers *b* and *c*; this new address is stored in the instruction address register.
— Once the computation is completed, a "1" appears on the *output available wire* and $f(x_2, x_1, x_0)$ is available on the *output value wire.* (The symbol "–" in fig. 3 stands for the don't care condition.)

In the general case of a multivalued decision algorithm computing a *n*-variable *m*-valued discrete function in *p* steps, a row of the control memory must contain the following informations:

V.I.F. (1 out of n)	address 0 (1 out of p)	— — — —	address ($m-1$) (1 out of p)	O.A.	O.V.

If we use an m-ary ROM for storing those informations, the number of cells per row is

$$\lceil \log_m n \rceil + m \lceil \log_m p \rceil + 2.$$

Hence, the total number of cells in the control store must be greater than or equal to

$$p\,(2 + \lceil \log_m n \rceil + m \lceil \log_m p \rceil).$$

The number p of instructions is again the most important factor in the hardware cost evaluation.

An important feature of this type of synthesis is that the response time is data dependent. It becomes thus interesting to have an estimation of the *average processing time*. If we assume that all the variables in example 1 take the values 0, 1 and 2 with probabilities equal to $\frac{1}{3}$, the average processing time expressed in clock periods is

$$\tfrac{12}{27} \times 3 + \tfrac{15}{27} \times 2 \cong 2.44$$

since 12 of the 27 cases are processed in 3 time units and 15 in 2 time units.

1.3. *Application to microprogrammed systems*

It has already been mentioned that multivalued decision algorithms reduce to condition automata. Clearly enough, the problem of obtaining e.g. a binary decision algorithm for computing a Boolean function plays an essential role in transforming an arbitrary microprogram with the various consequences this fact may imply on the architecture of the control part of the system (see e.g. Davio [5])). The sequential evaluation of discrete functions thus appears as a typical case of process sequentialization and thus of discussion of cost versus time compromise.

In conclusion to the above discussion, we note that three possible optimization criteria have been defined:

(a) the *number of instructions* in a multivalued decision algorithm $P(f)$;
(b) the *maximum processing time* $\tau_{\mathrm{M}}(f)$;
(c) the *average processing time* $\tilde{\tau}(f)$.

The problem of minimizing the average processing time for the particular case of Boolean functions has recently been tackled by Breitbart and Reiter [6]) and Perl and Breitbart [7]). Note also that the optimization of binary decision algorithms is related to that of transforming decision tables into computer programs and to the problem of organizing files as binary search trees. Survey papers on these topics have been presented by Pooch [8]) and Nievergelt [10]) respectively.

The present paper attempts to present for the first time results and applications for the multivalued decision algorithms (it is understood that all the results in the literature quoted hereabove are exclusively concerned with binary decision algorithms). Moreover, these results give raise to new and very competitive methods for Boolean functions.

The various optimization algorithms that will be developed are gathered in secs 3 and 4 while some preliminary optimization theorems are first briefly stated in sec. 2.

The algorithms presented in sec. 3 are referred to as *Leave-to-Root* (or L–R) algorithms; their construction starts from the leaves i.e. the nodes corresponding to the terminal statements of the decision algorithm and ends at its root i.e. the node corresponding to the initial instruction N_1 of the decision algorithm. It will be seen that the number of operations to be performed for computing optimal L–R algorithms is generally high. These algorithms are, however, easily mechanisable and use only elementary types of operations.

The algorithms presented in sec. 4 are referred to as *Root-to-Leave* (or R–L) algorithms since they start from the root of the decision tree and come to its leaves. The number of operations to be performed for computing R–L algorithms is generally much smaller than for L–R algorithms; these operations are, however, more complex and hence less mechanizable.

Section 5 is devoted to a short review of some other optimization algorithms.

It should finally be noted that the present paper is as self-consistent as possible.

In order to illustrate the concepts that will be introduced in the course of this paper the running example 1 will be continued throughout the following sections. Moreover, due to the importance of the binary case (Boolean functions) and to the significant simplifications to which it leads, the example 2 introduced herebelow will also be fully developed further on.

Example 2 (Boolean function). Consider the binary decision algorithm and its equivalent modified form given in the tables IVA and IVB respectively.
To the table IVA correspond the graph of fig. 4 and the network of fig. 5 made up of multiplexers.

Consider table IVB and let us choose the following encoding for the binary variables x_i and for the instruction N_j respectively

TABLE IVA

N	x	s_0	s_1
1	x_2	N_2	N_5
2	x_0	N_3	N_4
3	x_1	1	0
4	x_3	1	0
5	x_0	1	0

TABLE IVB

N	x	s_0	s_1	O.R.	O.V.
1	x_2	N_2	N_5	0	–
2	x_0	N_3	N_4	0	–
3	x_1	N_6	N_7	0	–
4	x_3	N_6	N_7	0	–
5	x_0	N_6	N_7	0	–
6	–	N_6	N_6	1	1
7	–	N_7	N_7	1	0

$$\frac{x_0 \quad x_1 \quad x_2 \quad x_3}{00 \quad 01 \quad 10 \quad 11},$$

$$\frac{N_1 \quad N_2 \quad N_3 \quad N_4 \quad N_5 \quad N_6 \quad N_7}{000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 110 \quad 111}$$

This encoding leads to the network of fig. 6.

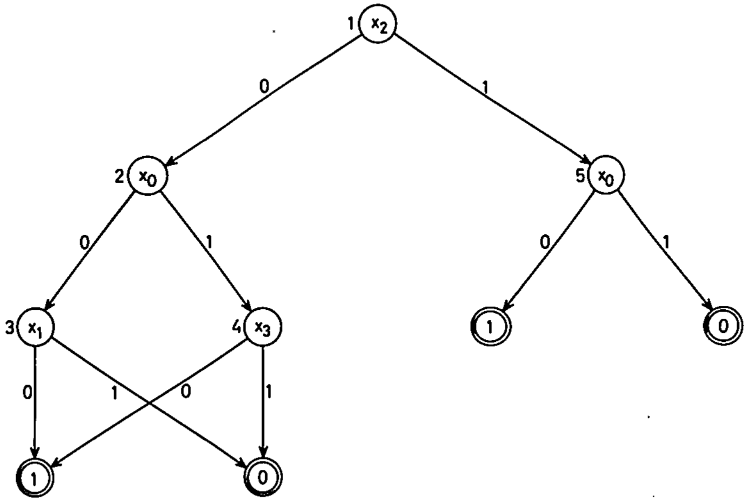


Fig. 4.

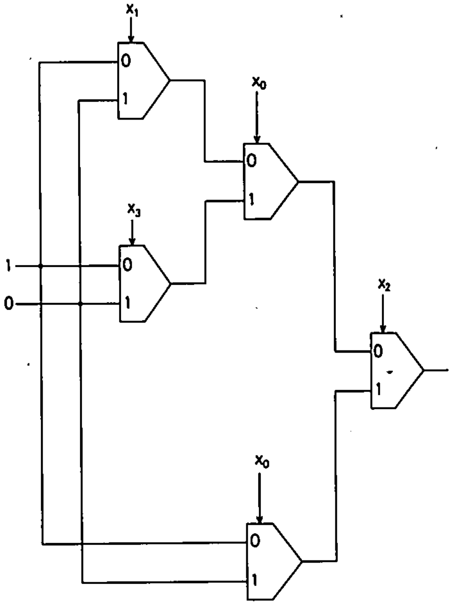
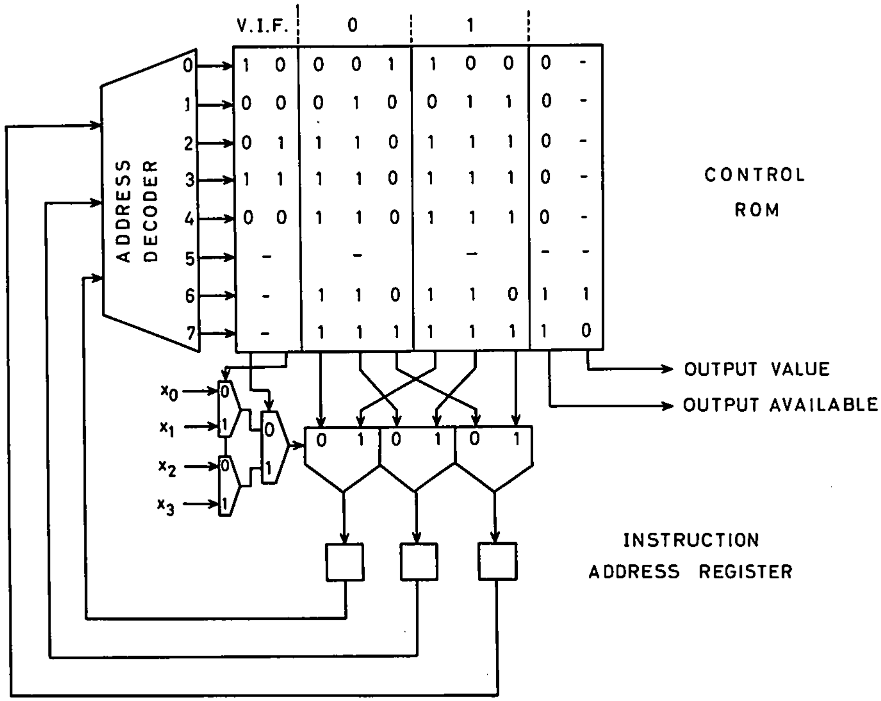


Fig. 5.



	V.I.F.		0			1				
0	1	0	0	0	1	1	0	0	0	-
1	0	0	0	1	0	0	1	1	0	-
2	0	1	1	1	0	1	1	1	0	-
3	1	1	1	1	0	1	1	1	0	-
4	0	0	1	1	0	1	1	1	0	-
5	–		–			–			-	-
6	-		1	1	0	1	1	0	1	1
7	-		1	1	1	1	1	1	1	0

Fig. 6.

Remark. The comparison between figs 5 and 6 leads to the ascertaining that the cost of the sequential realization (fig. 6 contains six multiplexers, ROM and registers) is much higher than the cost of the combinational realization (fig. 5 contains only five multiplexers).

2. General optimization theorems

Let C denote one of the three considered optimization criteria: $P(f)$, $\tau_M(f)$ and $\tilde{\tau}(f)$. We say that a multivalued decision algorithm is *C-optimal* if no multivalued decision algorithm computing f yields a lower value of C. Moreover, we speak of *τ-optimality* when a statement applies both to maximum and to average processing times.

THEOREM 1. *There always exists a τ-optimal multivalued decision tree.*
Proof. Given an optimal multivalued decision algorithm computing f, it is always possible to generate a multivalued decision tree computing f and having an identical processing time by duplicating the shared subroutines. □

Comment. If the optimization goal is τ-optimality one may thus restrict one's enumeration to multivalued decision trees. This statement is not true for P-optimality where the efficient use of subroutines is clearly essential. According to that remark, we shall say that a multivalued decision tree is *P_t-optimal* if no multivalued decision tree computing the same function yields a lower value of P; thus a P_t-optimal multivalued decision tree is not necessarily a P-optimal multivalued decision algorithm.

A multivalued decision algorithm is *simple* if none of the paths relating its initial node to a terminal node contains two occurrences of the same variable.

THEOREM 2. *There always exist a τ-optimal and a P_t-optimal simple multivalued decision type.*
Proof. On any path a node labelled with an occurrence of an already met variable may be deleted together with one of the subtrees it precedes. This decreases $\tilde{\tau}$ and P and does not increase τ_M. □

The statement of theorem 2 does not apply to multivalued decision algorithms in general and there may thus exist non-simple P-optimal multivalued decision algorithms. An example of such an algorithm for a Boolean function is the following one.

Example 3

TABLE V

N	x	s_0	s_1
1	x_2	N_2	N_3
2	x_1	0	N_4
3	x_0	N_4	0
4	x_3	N_5	N_6
5	x_4	0	1
6	x_1	0	1

It may be shown by exhaustive enumeration that no simple binary algorithm has fewer than 7 instructions; the above non simple algorithm computes the Boolean function given by the Karnaugh map of table VI.

TABLE VI

<table>
<tr><td></td><td></td><td colspan="2">x_0</td><td></td><td></td><td colspan="2">x_0</td><td></td><td></td></tr>
<tr><td></td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td><td></td></tr>
<tr><td rowspan="2">x_2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td><td></td></tr>
<tr><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td rowspan="2">x_3</td></tr>
<tr><td></td><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td colspan="2">x_1</td><td></td><td></td><td colspan="2">x_1</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="4">x_4</td><td></td></tr>
</table>

According to that observation, we shall say that a simple multivalued decision algorithm is P_s-*optimal* if no simple multivalued decision algorithm computing the same function yields a lower value of P.

A multivalued decision tree is *complete* if it is simple and if every path from the initial node to any terminal node has length n (it is recalled that n is the number of variables of $f(\mathbf{x})$). Clearly any simple multivalued decision tree may be augmented to become a complete tree. Similarly, a simple multivalued decision algorithm is easily transformed into a complete tree by performing successively subroutine duplication and augmentation. Reversing these processes, we obtain

THEOREM 3. (a) *Any simple multivalued decision tree may be obtained from a complete multivalued decision tree by relabelling with the constant h non-terminal nodes having all their successors (at most m) labelled h and by deleting these successors.*

(b) *Any simple multivalued decision algorithm may be obtained from a simple multivalued decision tree by subroutine merging.*

The above theorems immediately yield a brute force approach for obtaining τ-optimal, P_t-optimal and P_s-optimal multivalued decision algorithms. For τ-optimal and P_t-optimal trees, the method consists in enumerating all the complete trees and in performing on each of them the collapsing operation described in theorem 3(a). For obtaining P_s-optimal algorithms, one should furthermore perform on the obtained trees all the possible subroutine mergings.

3. Leave-to-Root algorithms

3.1. *Optimization of the maximal processing time $\tau_M(f)$*

3.1.1. Preliminary definitions and basic algorithm

We first define a generalized version of the *consensus operation.*

Definition 2. Let $C = \bigwedge_{l=0}^{n-1} x_l^{(C_l)}$ and $C' = \bigwedge_{l=0}^{n-1} x_l^{(C_l')}$ be two cubes. The *consensus* of C and of C' with respect to the variable x_k or *kth consensus* of C and of C' is the cube denoted $C \underset{k}{*} C'$ and defined by

$$C \underset{k}{*} C' = x_k^{(C_k \cup C_k')} \wedge \bigwedge_{l \neq k} x_l^{(C_l \cap C_l')}. \tag{5}$$

It could easily be verified (see also e.g. Davio and Bioul [11]) that the kth consensus is a commutative and associate operation on the set of cubes and that any consensus of two cubes is included in their disjunction.

Definition 3.

(a) The *T-terms of order* 0 or *T^0-terms* of the discrete function $f(\mathbf{x})$ will be denoted $T_i^0(f)$ with i a counting index; they are all the cubes covering the subdomains where the function f uniformly takes the same value h ($h \in \{0, 1, \ldots, m-1\}$).

(b) The *T-terms of order q* ($q \geqslant 1$) or *T^q-terms* of the discrete function $f(\mathbf{x})$ will be denoted $T_i^q(f)$ and are defined as follows:

$$T_i^q(f) = T_j^{q-1}(f) \underset{k}{*} T_{l_0}^{p_0}(f) \underset{k}{*} T_{l_1}^{p_1}(f) \underset{k}{*} \ldots \underset{k}{*} T_{l_{s-1}}^{p_{s-1}}(f),$$

at most m cubes, $p_e \leqslant q-1$, $0 \leqslant e \leqslant s \leqslant m-1$; $T_i^q(f)$ is a T^q-term if

(a) the $s+1$ exponents C_e $(0 \leqslant e \leqslant s+1 \leqslant m)$ of $x_k^{(C_e)}$ that appear in the $s+1$ cubes $T(f)$ of the right member of the above relation constitute a partition of the set $\{0, 1, \ldots, m-1\}$;

(b) $T_l^q(f) \neq T_l^r(f)$, $r \leqslant q-1$ and $\forall\, l$.

Stated otherwise, the consensus between a T^{q-1}-term and at most $m-1$ T^{p_e}-terms $(p_e \leqslant q-1\ \forall\, e)$ produces a T^q-term iff (a) the result of the consensus operation is independent of x_k and (b) the result is not a T^r-term $(r \leqslant q-1)$.

Definition 4. A T-term of order q will said to be a *prime T^q-term* iff it is not contained in another T^q-term.

Example 1 (continued, see page 33). Consider again the example 1 together with the table VII; the prime T^0-terms of the function of table I are numbered from 1 to 8. Its prime T^1-terms, prime T^2-terms and prime T^3-terms are numbered from 9 to 17, from 18 to 23 and from 24 to 26 respectively. The following information is gathered in the column "consensus operation": e.g. for the term numbered 17 one has: $(x_0 : 6, 5, 2)$ which means that the term 17 has been obtained by performing the consensus operation between the terms numbered 6, 5 and 2 and with respect to the variable x_0. The letters $x_0^{(0)}$, $x_0^{(1)}$ and $x_0^{(2)}$ are present in the terms 6, 5 and 2 respectively. In the same way the information $(x_2 : 14, 11, 14)$ for the term 21 means that this last has been obtained by consensus operation between the terms 11 and 14 and with respect to the variable x_2; the letters $x_2^{(0,2)}$ and $x_2^{(1)}$ are present in the terms 14 and 11 respectively.

The algorithm, for optimizing the maximum processing time in decision trees, that will be developed is grounded on the following theorem.

THEOREM 4. *To each T^q-term $\bigwedge_l x_l^{(e_l)}$ that has been obtained by the consensus operation (see the notation of table I): $(x_j : N_{j_1}, N_{j_2}, \ldots, N_{j_l})$ can be associated a multivalued decision tree having a maximum processing time of q, starting with x_j as decision variable and describing that part of f limited to the subdomain characterized by the equation*

$$\bigwedge_l x_l^{(e_l)} = m-1. \tag{6}$$

Proof. Consider the T^q-term (6) as the root-node of the decision tree; from this node are issued at most m branches (according to the number of the different N_{j_e}'s) associated to the values $0, 1, \ldots, m-1$ of the decision variable x_k and leading to the nodes numbered $N_{j_1}, N_{j_2}, \ldots, N_{j_l}$ respectively. These

TABLE VII

q	N	T^q-terms	consensus operation	weights	associated function
0	1	$x_0^{(2)}x_1^{(0)}$	—	0,0	2
0	2	$x_0^{(2)}x_2^{(1)}$	—	0,0	2
0	3	$x_1^{(2)}x_2^{(1)}$	—	0,0	2
0	4	$x_0^{(1)}x_1^{(0)}$	—	0,0	1
0	5	$x_0^{(1)}x_2^{(1)}x_1^{(0,1)}$	—	0,0	1
0	6	$x_0^{(0)}x_1^{(0,1)}$	—	0,0	0
0	7	$x_1^{(1,2)}x_2^{(0,2)}$	—	0,0	0
0	8	$x_0^{(0)}x_2^{(0,2)}$	—	0,0	0
1	9	$x_1^{(0)}$	$x_0: 6,4,1$	1,1	$1x_0^{(1)} \vee 2x_0^{(2)}$
1	10	$x_0^{(2)}x_1^{(1,2)}$	$x_2: 7,2,7$	1,1	$2x_2^{(1)}$
1	11	$x_0^{(1)}x_2^{(1)}$	$x_1: 5,5,3$	1,1	$1x_1^{(0,1)} \vee 2x_1^{(2)}$
1	12	$x_0^{(0)}x_2^{(1)}$	$x_1: 6,6,3$	1,1	$2x_1^{(2)}$
1	13	$x_1^{(2)}$	$x_2: 7,3,7$	1,1	$2x_2^{(1)}$
1	14	$x_0^{(1)}x_2^{(0,2)}$	$x_1: 4,7,7$	1,1	$1x_1^{(0)}$
1	15	$x_0^{(1)}x_2^{(1)}$	$x_1: 5,5,3$	1,1	$1x_1^{(0,1)} \vee 2x_1^{(2)}$
1	16	$x_0^{(1)}x_1^{(1)}$	$x_2: 7,5,7$	1,1	$1x_2^{(1)}$
1	17	$x_1^{(0,1)}x_2^{(1)}$	$x_0: 6,5,2$	1,1	$1x_0^{(1)} \vee 2x_0^{(2)}$
2	18	$x_1^{(1)}$	$x_2: 7,17,7$	$2,\frac{4}{3}$	$1x_2^{(1)}x_0^{(1)} \vee 2x_2^{(1)}x_0^{(2)}$
2	19	$x_0^{(0)}$	$x_2: 8,12,8$	$2,\frac{4}{3}$	$2x_2^{(1)}$
2	20	$x_0^{(2)}$	$x_1: 1,10,10$	$2,\frac{5}{3}$	$2x_1^{(0)} \vee 2x_1^{(1,2)}x_2^{(1)}$
2	21	$x_0^{(1)}$	$x_2: 14,11,14$	3,2	$1x_2^{(0,2)}x_1^{(0)} \vee 1x_2^{(1)}x_1^{(0,1)} \vee 2x_2^{(1)}x_1^{(2)}$
2	22	$x_2^{(0,2)}$	$x_1: 9,7,7$	$2,\frac{4}{3}$	$1x_0^{(1)}x_1^{(0)} \vee 2x_0^{(2)}x_1^{(0)}$
2	23	$x_2^{(1)}$	$x_1: 17,17,3$	$2,\frac{5}{3}$	$1x_0^{(1)}x_1^{(0,1)} \vee 2x_0^{(2)}x_1^{(0,1)} \vee 2x_1^{(2)}$
3	24	$m-1$	$x_1: 9,18,13$	$5,\frac{19}{9}$	f
3	25	$m-1$	$x_0: 19,21,20$	$8,\frac{8}{3}$	f
3	26	$m-1$	$x_2: 22,23,22$	$5,\frac{22}{9}$	f

nodes are in turn each associated with T^{p_e}-terms ($e \leqslant m$, $0 \leqslant p_e \leqslant q-1 \; \forall \, e$); these T^{p_e}-terms are each considered as nodes of the multivalued decision tree and the process is continued iteratively until arriving at the terminal nodes which correspond to the T^0-terms. Clearly the multivalued decision tree so constructed has a maximum processing time q and computes that part of f corresponding to the domain characterized by the equation (6). □

Based on the above observations the following algorithm may now be stated.

Algorithm 1.

Starting from the prime T^0-terms, compute the prime T^q-terms until a level q has been obtained such that at least one T^q-term is equal to $m-1$. To this T^q-term corresponds a multivalued decision tree having a maximum processing time of value q; this value is minimal since the T^q-term is prime.

Example 1 (continued, see page 33). Consider the table VII; to the prime T^3-terms 24, 25 and 26 correspond the optimal decision trees of figs 7*b*, *a* and *c* respectively. They all have a maximal processing time of 3.

3.1.2. The particular case of Boolean functions

For the Boolean functions the definition 3 of T-term reduces to the following one.

Definition 3 (*bis*).

(a) The T-terms of order 0 or T^0-terms of the Boolean function $f(\mathbf{x})$ will be denoted $T_i^0(f)$ with i a counting index; they are all the implicants of the functions $f(\mathbf{x})$ and $\bar{f}(\mathbf{x})$.

(b) The T-terms of order q ($q \geqslant 1$) or T^q-terms of the Boolean function $f(\mathbf{x})$ will be denoted $T_i^q(f)$, they are defined as follows:

$$\begin{aligned} &T_i^q(f) = T_j^{q-1}(f) * T_k^p(f), \quad p \leqslant q-1, \\ \text{if} \quad &T_i^q(f) \neq T_l^r(f), \qquad r \leqslant q-1 \quad \text{and} \quad \forall \, l, \end{aligned} \tag{7}$$

where $*$ means the consensus operation.

Stated otherwise the consensus between a T^{q-1}-term and a T^p-term ($p \leqslant q-1$) produces a T^q-term iff the result of the consensus operation is not a T^r-term ($r \leqslant q-1$).

Let us recall (see e.g. refs 12, 13) that the simple Boolean derivative of f with respect to the variable x_i is defined as follows:

$$\frac{\partial f}{\partial x_i} = f(x_i = 0) \oplus f(x_i = 1). \tag{8}$$

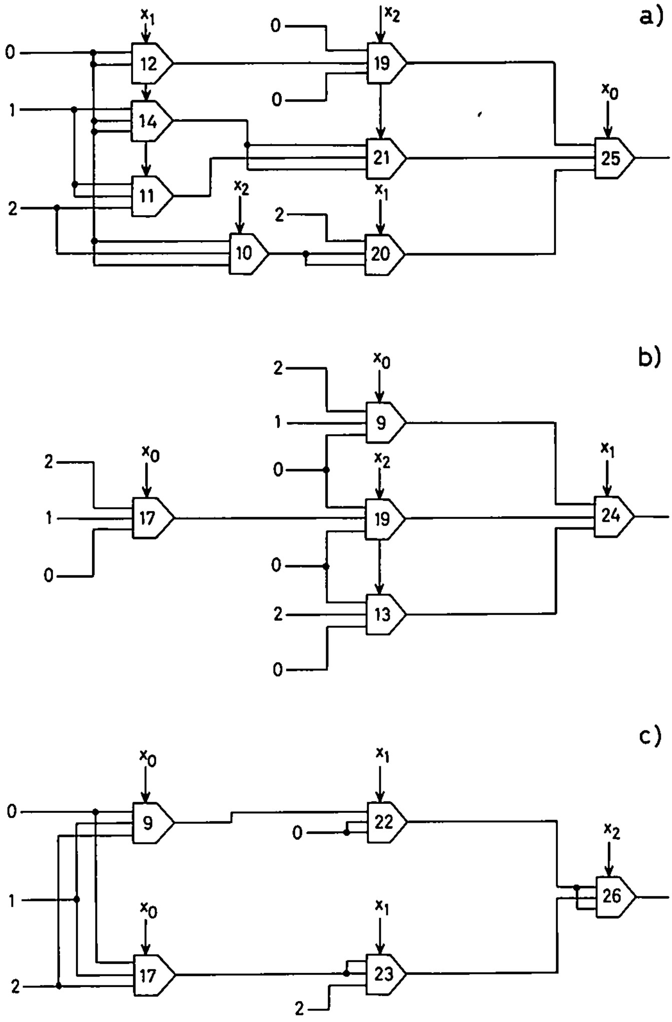


Fig. 7.

The T^1-terms are connected to the simple derivatives of f by the following theorem.

THEOREM 5. *The T^1-terms are the prime implicants of the functions*

$$\frac{\partial f}{\partial x_i} f(x_i = 0) \quad \text{and} \quad \frac{\partial f}{\partial x_i} f(x_i = 1).$$

Proof.

$$\begin{aligned}\frac{\partial f}{\partial x_i} f(x_i = 0) &= f(x_i = 0)\bar{f}(x_i = 1),\\ \frac{\partial f}{\partial x_i} f(x_i = 1) &= \bar{f}(x_i = 0) f(x_i = 1).\end{aligned} \tag{9}$$

Since $f(x_i = 0)\bar{f}(x_i = 1)$ and $\bar{f}(x_i = 0) f(x_i = 1)$ are the functions obtained by performing the consensus operation between f and $\bar{f}$ the theorem is proved. From theorem 4 one immediately deduces that the prime T^1-terms are the prime implicants of the functions

$$\frac{\partial f}{\partial x_i} f(x_i = 0), \quad \frac{\partial f}{\partial x_i} f(x_i = 1), \quad i = 0, 1, \ldots, n-1. \tag{10}$$

The prime T^1-terms are thus each contained in a prime implicant of at least one simple derivative. □

Example 2 (continued, see page 39). Consider again the example 2 together with the table VIII; this table has the same meaning as table VII. It contains the prime T-terms of the considered Boolean function plus some T-terms (not prime) that are added for reasons that will appear in the next section (the not prime T-terms are numbered N_j^* instead of N_j); to the prime T-terms 30 and 31 correspond the optimal binary decision trees of figs 8*a* and 8*b* respectively. Besides this the binary decision tree of fig. 8*a* corresponds to the algorithm in table V. The nodes of the decision trees of fig. 8 have been numbered in correspondence with the T-terms of table VIII that generate the decision trees.

TABLE VIII

q	N	T^q-terms	consensus operation	weights	associated function
0	1	$\bar{x}_0\,\bar{x}_1$	—	0,0	1
0	2	$\bar{x}_0\,x_2$	—	0,0	1
0	3	$x_0\,\bar{x}_2\,\bar{x}_3$	—	0,0	1
0	4	$\bar{x}_1\,\bar{x}_2\,\bar{x}_3$	—	0,0	1
0	5	$x_0\,x_2$	—	0,0	0
0	6	$x_0\,x_3$	—	0,0	0
0	7	$x_1\,\bar{x}_2\,x_3$	—	0,0	0

to be continued

TABLE VIII (continued)

q	N	T^q-terms	consensus operation	weights	associated function
0	8	$\bar{x}_0\,x_1\,\bar{x}_2$	—	0,0	0
1	9	x_2	$x_0 : 2,5$	1,1	$\bar{x}_0$
1	10	$\bar{x}_1\,x_3$	$x_0 : 1,6$	1,1	$\bar{x}_0$
1	11	$\bar{x}_0\,\bar{x}_2$	$x_1 : 1,8$	1,1	$\bar{x}_1$
1	12	$\bar{x}_0\,x_1$	$x_2 : 8,2$	1,1	x_2
1	13	$x_0\,\bar{x}_3$	$x_2 : 3,5$	1,1	$\bar{x}_2$
1	14	$x_0\,\bar{x}_2$	$x_3 : 3,6$	1,1	$\bar{x}_3$
1	15	$x_1\,\bar{x}_2\,\bar{x}_3$	$x_0 : 3,8$	1,1	x_0
2	16	$\bar{x}_0$	$x_2 : 11,2$	$2,\frac{3}{2}$	$\bar{x}_2\,\bar{x}_1 \vee x_2$
2	17*	$\bar{x}_2\,\bar{x}_3$	$x_0 : 11,3$	$2,\frac{3}{2}$	$\bar{x}_0\,\bar{x}_1 \vee x_0$
2	18	$\bar{x}_1\,\bar{x}_3$	$x_2 : 4,9$	$2,\frac{3}{2}$	$\bar{x}_2 \vee x_2\,\bar{x}_0$
2	19*	$\bar{x}_1\,\bar{x}_2$	$x_3 : 4,10$	$2,\frac{3}{2}$	$x_3\,\bar{x}_0$
2	20	x_0	$x_2 : 14,5$	$2,\frac{3}{2}$	$\bar{x}_2\,\bar{x}_3$
2	21*	$\bar{x}_2\,x_3$	$x_0 : 11,6$	$2,\frac{3}{2}$	$\bar{x}_0\,\bar{x}_1$
2	22	$x_1\,x_3$	$x_0 : 12,6$	$2,\frac{3}{2}$	$\bar{x}_0\,x_2$
2	23*	$x_1\,\bar{x}_2$	$x_3 : 15,7$	$2,\frac{3}{2}$	$\bar{x}_3\,x_0$
2	24	$x_1\,\bar{x}_3$	$x_2 : 15,9$	3,2	$\bar{x}_2\,x_0 \vee x_2\,\bar{x}_0$
2	25	$\bar{x}_2$	$x_0 : 11,14$	3,2	$\bar{x}_0\,\bar{x}_1 \vee x_0\,\bar{x}_3$
3	26	x_3	$x_0 : 16,6$	$3,\frac{7}{4}$	$\bar{x}_0\,\bar{x}_2\,\bar{x}_1 \vee \bar{x}_0\,x_2$
3	27	$\bar{x}_1$	$x_0 : 1,20$	$3,\frac{7}{4}$	$\bar{x}_0 \vee x_0\,\bar{x}_2\,\bar{x}_3$
3	28	$\bar{x}_3$	$x_0 : 16,13$	$4,\frac{9}{4}$	$\bar{x}_0\,\bar{x}_2\,\bar{x}_1 \vee \bar{x}_0\,x_2 \vee x_0\,\bar{x}_2$
3	29	x_1	$x_0 : 12,20$	$4,\frac{9}{4}$	$\bar{x}_0\,x_2 \vee x_0\,\bar{x}_2\,\bar{x}_3$
3	30	1	$x_2 : 25,9$	$5,\frac{5}{2}$	$\bar{x}_2\,\bar{x}_0\,\bar{x}_1 \vee \bar{x}_2\,x_0\,\bar{x}_3 \vee x_2\,\bar{x}_0$
3	31	1	$x_0 : 16,20$	$5,\frac{5}{2}$	$\bar{x}_0\,\bar{x}_2\,\bar{x}_1 \vee \bar{x}_0\,x_2 \vee x_0\,\bar{x}_2\,\bar{x}_3$
4	32*	1	$x_3 : 28,26$	8,3	$\bar{x}_3\,\bar{x}_0\,\bar{x}_2\,\bar{x}_1 \vee \bar{x}_3\,\bar{x}_0\,x_2 \vee \bar{x}_3\,x_0\,\bar{x}_2$ $\vee\, x_3\,\bar{x}_0\,\bar{x}_2\,\bar{x}_1 \vee x_3\,\bar{x}_0\,x_2$
4	33*	1	$x_1 : 27,29$	8,3	$\bar{x}_1\,\bar{x}_0 \vee \bar{x}_1\,x_0\,\bar{x}_2\,\bar{x}_3 \vee x_1\,\bar{x}_0\,x_2 \vee x_1\,x_0\,\bar{x}_2\,\bar{x}_3$

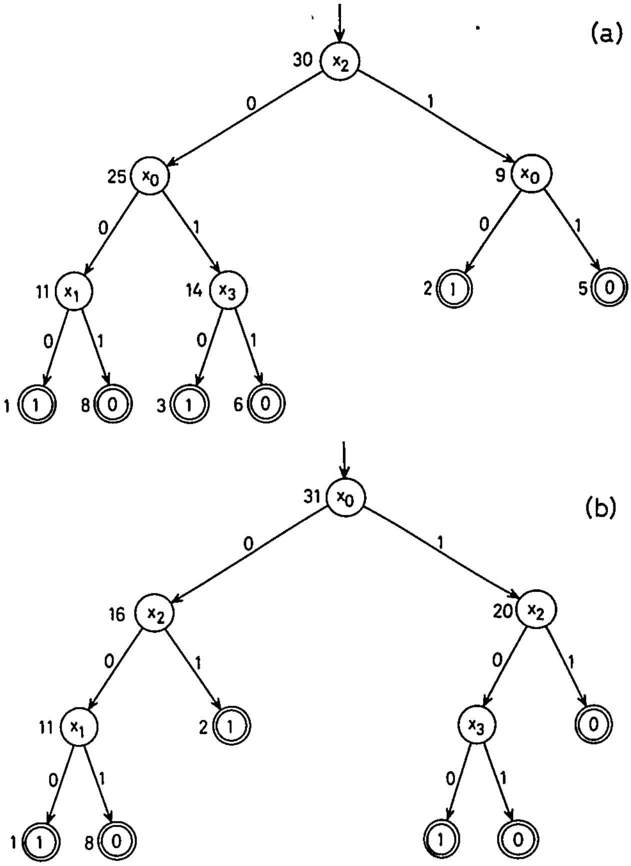


Fig. 8. Optimal binary decision trees for the example 2.

3.2. *The weighted T-terms and the optimization of the number of instructions and of the average processing time in decision trees*

It has been seen that the concept of prime T-term was the adequate mathematical tool for the optimization of maximal processing times in multivalued decision trees. The concept of weighted T-term that will be introduced hereto below will be used in the optimization of the number of instructions and of the average processing time in multivalued decision trees.

Definition 5. A *weighted T-term* is a pair

$$\{\alpha, c(\mathbf{x})\},$$

where α is a nonnegative rational number called the *weight* of the T-term and where $c(\mathbf{x})$ is a cube.

The weighted T-terms of the discrete function $f(\mathbf{x})$ are defined iteratively as follows.

(a) The weighted T-terms of order 0 or weighted T^0-terms of the discrete function $f(\mathbf{x})$ are the pairs

$$\{m_0, \mathfrak{T}_i{}^0(f)\} = \{0, T_i{}^0(f)\}. \tag{11}$$

(b) The weighted T-terms of order q or weighted T^q-terms of the discrete function $f(\mathbf{x})$ are the pairs

$$\{m_q, \mathfrak{T}_i{}^q(f)\}.$$

They are defined as follows: let

$$\{m_{q-1}, \mathfrak{T}_j{}^{q-1}(f)\},\ \{m_{l_0}, \mathfrak{T}_{l_0}{}^{p_0}(f)\},\ \ldots,\ \{m_{l_{s-1}}, \mathfrak{T}_{l_{s-1}}{}^{p_{s-1}}(f)\}$$
$$p_e \leqslant q - 1\ \forall\, e, \quad s \leqslant m - 1$$

be at most m weighted T-terms. Then

$$\{m_q, \mathfrak{T}_i{}^q(f)\} = \{\Phi(m_{q-1}, m_{l_0}, \ldots, m_{l_s}),\ \mathfrak{T}_j{}^{q-1}(f) \underset{k}{*} \mathfrak{T}_{l_0}{}^{p_0}(f) \underset{k}{*} \ldots \underset{k}{*} \mathfrak{T}_{l_s}{}^{p_s}(f)\} \tag{12}$$

(with Φ a function to be defined further on) is a weighted T^q-term if

(a) the $s + 1$ exponents C_e $(0 \leqslant e \leqslant s + 1 \leqslant m)$ of $x_k{}^{(C_e)}$ that appear in the $s + 1$ cubes $\mathfrak{T}(f)$ of the right member of (12) constitute a partition of the set $\{0, 1, \ldots, m - 1\}$;

(b) either $\mathfrak{T}_i{}^q(f) \neq \mathfrak{T}_l{}^r(f)$, $r \leqslant q - 1$ and $\forall\, l$
or $\mathfrak{T}_i{}^q(f) = \mathfrak{T}_l{}^r(f)$ for some $r \leqslant q - 1$ and some l
and $m_r > m_q$ for the weighted T^r-term $\{m_r, \mathfrak{T}_l{}^r(f)\}$.

Definition 6. A prime weighted T^q-term, is a weighted T^q-term $\{m, c(\mathbf{x})\}$ such that for any other weight T^q-term $\{m', c'(\mathbf{x})\}$ one has either

$$c(\mathbf{x}) \nsubseteq c'(\mathbf{x})$$

or if $c(\mathbf{x}) \supseteq c'(\mathbf{x})$ then $m < m'$.

Two types of functions $\Phi\,(m_{q-1}, m_{l_0}, \ldots, m_{l_{(k-1)}})$will now be defined:

$$\Phi_0(m_{q-1}, m_{l_0}, \ldots, m_{l_{(k-1)}}) = m_{q-1} + m_{l_0} + \ldots + m_{l_{(k-1)}} + 1, \tag{13}$$

$$\Phi_1(m_{q-1}, m_{l_0}, \ldots, m_{l_{(k-1)}}) = \frac{1}{m}(m_{q-1} + m_{l_0} + \ldots + m_{l_{(k-1)}}) + 1. \tag{14}$$

Example 1 (continued, see page 33). The weights of the T^q-terms of table VII and with respect to the laws Φ_0 and Φ_1 have been gathered in the column "weights".

The algorithm quoted in this section is grounded on the following theorem.

THEOREM 6. *To each weighted T^q-term $\bigwedge_i x_i^{(e_i)}$ with weights m_0 and m_1 (with respect to the laws Φ_0 and Φ_1 respectively) can be associated a multivalued decision tree having a maximum processing time q, a number m_0 of instructions, an average processing time m_1 and describing that part of f limited ot the subdomain characterized by the equation*

$$\bigwedge_i x_i^{(e_i)} = m - 1.$$

Proof. It can easily be verified that the laws Φ_0 and Φ_1 compute the number of instructions and the average processing time of the considered multivalued decision tree respectively; the proof then derives from theorem 4. □

Algorithm 2

Compute the prime weighted T-terms of the function f; from the list of prime weighted T-terms equal to $m-1$ select those having a minimal weight m_0 (resp. m_1). To those weighted T-terms correspond binary decision trees having a minimum number of m_0 instructions (resp. a minimum average processing time of value m_1).

Example 1 (continued, see page 33). Consider the multivalued decision trees corresponding to the T^3-terms 24, 25 and 26 of table I; they have a number of instructions of 5, 8 and 5 respectively and an average processing time of 19/9, 8/3 and 22/9 respectively. This can also be verified by considering the three networks of fig. 7.

Example 2 (continued, see page 39). Consider the table VIII; the multivalued decision trees corresponding to the T^3-terms 30 and 31 are both optimal with respect to the number of instructions and with respect to the average processing time; this appears also in the decision trees of fig. 8. The terms numbered N_j* are prime weighted T-terms that are not prime T-terms.

3.3. *Reduction of the number of instructions in multivalued decision algorithms*

Until now we have only considered minimization problems in the restricted frame of multivalued decision trees. Minimizing the number of instructions in multivalued decision algorithms is a much more complex problem.

We have seen in the preceding sections that minimization problems in decision trees may be solved (at least formally) by using the concept of T^q-term casually associated with a weight; each T^q-term covers a subdomain of the function domain as characterized by the eq. (6). If one wants to deal with the minimization of the number of instructions in decision algorithms, one has not only to consider the function subdomains but also the subfunctions that are realized in these subdomains: indeed, the minimization of the number of instructions requires that instructions should be merged, i.e. that identical subfunctions should be recognized. This will appear more clearly in the course of example 1 that will be continued further on in this section.

To each T^q-term will be associated a discrete function in the following way.

Definition 7.

(a) To a T^0-term: $\bigwedge_l x_l^{(e_l)}$ which covers a subdomain where f takes uniformly the value h will be associated the function h. The pairs $\{T^0$-term, associated function$\}$ will be written

$$\{\bigwedge_l x_l^{(e_l)}, h\}$$

(b) Let $\{T_j^{q-1}(f), g_j\}$, $\{T_{l_0}^{p_0}(f), g_{l_0}\}$, ..., $\{T_{l_s}^{p_s}(f), g_{l_s}\}$ be T-terms with their associated function; if

$$T_i^q(f) = T_j^{q-1}(f) \underset{k}{*} T_{l_0}^{p_0}(f) \underset{k}{*} \dots \underset{k}{*} T_{l_s}^{p_s}(f)$$

is a T^q-term and if one assumes that $x_k^{(C_j)}$ is present in $T_j^{q-1}(f)$ and $x_k^{(C_e)}$ is present in $T_{l_e}^{p_e}(f)\ \forall\, e$, the function associated to $T_i^q(f)$ is

$$g_i = g_j\, x_k^{(C_j)} \vee g_{l_0}\, x_k^{(C_{l_0})} \vee \dots \vee g_{l_s}\, x_k^{(C_{l_s})}. \tag{15}$$

The reduction of the number of instructions in multivalued decision algorithms will be obtained by *merging* some instructions in these trees; the merging of instructions will be defined as follows.

Consider two instructions with their associated function, i.e.:

— instruction N_j: $\{T_j^q(f), g_j\}$,

— instruction N_i: $\{T_i^p(f), g_i\}$.

Definition 8. Two instructions N_j and N_i may be merged if and only if $g_j = g_i$; the merged instruction N_{ij} is then

$$\text{instruction } N_{ij}\text{: } \{T_j^q(f) \vee T_i^p(f), g_i\}. \tag{16}$$

The merging of more than two instructions may be performed iteratively. The following theorem 7 constitutes a quite straightforward consequence of the above definition 8.

THEOREM 7. *In a multivalued decision algorithm, the merging of instructions transforms the primitive algorithm in an equivalent one (that is an algorithm computing the same discrete function) having a smaller number of instructions.*

It should also be noted that the merged instruction N_{ij} represented in (16) describes that part of f limited to the subdomain characterized by the equation

$$T_j{}^q(f) \vee T_i{}^p(f) = m-1. \tag{17}$$

It could evidently be possible to build straightforward algorithms for the minimization of the number of instructions in multivalued decision algorithms by using instead of T^q-terms or of weighted T^q-terms, the pair $\{T^q$-term, associated discrete function$\}$. *Prime pairs* could then be defined in the same way as prime T^q-terms and as prime weighted T^q-terms. However, the number of pairs that must be taken into account renders algorithms similar as those developed in secs 3.1 and 3.2 practically unusable. Another approach that might be used to reduce (but not to minimize) the number of instructions in decision algorithms is e.g. to evaluate systematically all the discrete functions associated with prime T^q-terms or with prime weighted T^q-terms. Either the algorithm of sec. 3.1, or the algorithm of sec. 3.2 is then used for satisfying the corresponding minimization criterion and the associated discrete functions are then used in an "a posteriori" treatment in order to reduce the number of instructions. It can easily be verified that the merging of instructions does not affect the minimal character of the maximal processing time or of the average processing time.

Example 1 (continued, see page 33). Consider the functions associated with the prime T^q-terms and that are gathered in the last column of table VII. Since the functions associated with the instructions 9 and 17 are the same, these instructions may be merged and the network of fig. 7*c* (with five multiplexers) is transformed into that of fig. 2 having only four multiplexers.

4. Root-to-Leave algorithms

4.1. *Discrete polynomials attached to a multivalued decision algorithm*

In the present section we first develop formal techniques for associating with each node of a multivalued decision algorithm two sets of products called *input* and *output set* respectively. The input set $I(k)$ attached to a node N_k describes the various way for passing from the initial node to the node N_k (hence, if the algorithm is a tree, $I(k)$ reduces to a single product term). The output set $O(k)$ attached to the node N_k represents the function realized when choosing N_k as initial node.

The computation methods proposed here aim at producing unique expressions both for the input sets and for the output sets. The characteristic properties of these expressions will then be studied and exploited for the research of optimal multivalued decision trees.

Computation of the input sets

We set $I(1) = \varnothing$. Assume then that the node N_k has the predecessors N_{k1}, $N_{k2}, \ldots, N_{kp}$, that these nodes are labelled $x_{k1}, x_{k2}, \ldots, x_{kp}$, respectively and that the transition $N_{kl} \rightarrow N_k$ takes place for $x_{kl} = C_{kl}$ ($C_{kl} \subset \{0, 1, \ldots, m-1\}$). Then

$$I(k) = \bigcup_{l=1}^{p} I(k_l) \times x_{kl}^{(C_{kl})}, \tag{18}$$

where "×" denotes the external product of the set I by the variable x. Formula (18) yields a recursive computation of the input sets.

Computation of the output sets

Let T represent the set of terminal nodes and define the subsets T_j of T ($j = 0, 1, \ldots, m-1$) by

$$T_j = \{k \mid k \in T \quad \text{and} \quad f(k) = j\}.$$

We then set $O(k) = j$ if $k \in T_j$.
The output set attached with the node N_k is then computed according to

$$O(k) = \bigcup_{l} x^{(C_{kl})}\, O(k_l)$$

where x is the variable attached to N_k, $O(k_l)$ are the output sets attached to the successor nodes of N_k which are connected to N_k through arcs labelled C_{kl} respectively.

Note also that the use of set theoretical operations prevents one from carrying out logic simplifications.

Example 1 (continued, see page 33). Consider the algorithm of fig. 1. The corresponding input and output sets are represented herebelow.

$$N_1 \begin{cases} I(1) = \varnothing \\ O(1) = \{2x_0^{(2)} x_1^{(0,1)} x_2^{(1)}, 2x_1^{(2)} x_2^{(1)}, 2x_0^{(2)} x_1^{(0)} x_2^{(0,2)}, 1x_0^{(1)} x_1^{(0,1)} x_2^{(1)}, \\ \qquad 1x_0^{(1)} x_1^{(0)} x_2^{(0,2)}\} \end{cases}$$

$$N_2 \begin{cases} I(2) = \{x_2^{(0,2)}\} \\ O(2) = \{2x_0^{(2)} x_1^{(0)}, 1x_0^{(1)} x_1^{(0)}\} \end{cases}$$

$$N_3 \begin{cases} I(3) = \{x_2^{(1)}\} \\ O(3) = \{2x_1^{(2)}, 2x_0^{(2)} x_1^{(0,1)}, 1x_0^{(1)} x_1^{(0,1)}\} \end{cases}$$

$$N_4 \begin{cases} I(4) = \{x_2^{(0,2)}\, x_0^{(0)},\, x_2^{(1)}\, x_0^{(0,1)}\} \\ O(4) = \{1x_0^{(1)},\, 2x_0^{(2)}\} \end{cases}$$

$$N_5 \begin{cases} I(5) = \{x_2^{(0,2)}\, x_1^{(1,2)},\, x_2^{(0,2)}\, x_1^{(0)}\, x_0^{(0)},\, x_2^{(1)}\, x_1^{(0,1)}\, x_0^{(0)}\} \\ O(5) = \{0\} \end{cases}$$

$$N_6 \begin{cases} I(6) = \{x_2^{(0,2)}\, x_1^{(0)}\, x_0^{(1)},\, x_2^{(1)}\, x_1^{(0,1)}\, x_0^{(1)}\} \\ O(6) = \{1\} \end{cases}$$

$$N_7 \begin{cases} I(7) = \{x_2^{(0,2)}\, x_1^{(0)}\, x_0^{(2)},\, x_2^{(1)}\, x_1^{(0,1)}\, x_0^{(2)},\, x_2^{(1)}\, x_1^{(2)}\} \\ O(7) = \{2\} \end{cases}$$

We now define a *standard set of products* as a set of products satisfying the two following rules.

(a) There exists a variable x_i that appears as first letter of every product of the set under the form $x_i^{(C_{ik})}$, with $C_{ik} \cap C_{ik'} = \varnothing \forall\, k, k'$, where k and k' are subscripts associated to any pair of cubes of the set.

(b) If there exists p different C_{ik}'s, the p sets of products obtained by grouping the coefficients of $x_i^{(C_{ik})}$ $(k = 0, 1, \ldots, p-1)$ are themselves standard sets.

It will be shown that standard sets play a central role in deriving multivalued decision algorithms. The key result in this respect will appear hereunder as theorem 10. We first derive without proof some quite obvious properties of the standard sets.

THEOREM 8. *The product terms in a standard set are pairwise disjoint.*

THEOREM 9.

(a) *The input sets $I(k)$ and output sets $O(k)$ associated with a multivalued decision algorithm are standard sets.*

(b) *With a standard set representing a discrete function $f(\mathrm{x})$ it is possible to associate a multivalued decision algorithm computing that function.*

Note also that the input and output sets attached to the various nodes satisfy the equation

$$O(1) = \bigcup_{\substack{k \in T_i \\ i \geqslant 1}} I(k) = f(\mathrm{x}) \tag{19}$$

The above theorems exhibit a strong relationship between multivalued decision algorithms and standard sets. As a matter of fact, that correspondence becomes a bijection if one restricts oneself to the case of binary decision trees. It is thus interesting to trace out the specific properties of standard sets associated with optimal decision trees.

Before starting the description of the computational methods, we note that the process of associating an algorithm with a standard set naturally yields (at least), a simple algorithm. This is why, in the remaining part of this paper, we consider only simple decision algorithms and thus implicitely refer to τ-optimality and P_s-optimality.

4.2. *Minimal simple decision algorithms*

A standard set is irreducible if it does not contain any pair of terms of the form $\{mx^{(C_0)}, mx^{(C_1)}, \ldots, mx^{(C_{p-1})}\}$ with $\cup\, C_i = \{0, 1, \ldots, m-1\}$ (taking into account the ordering of the literals). One immediately observes that the replacement of the p-tuple $\{mx^{(C_0)}, mx^{(C_1)}, \ldots, mx^{(C_{p-1})}\}$ by the single term m does not modify the standard character. Furthermore:

THEOREM 10. *There always exists a C-optimal decision algorithm to which is associated an irreducible standard set* $O(1)$.
Proof. Consider a C-optimal decision algorithm with which is associated a standard set containing the p terms $\{mx^{(C_0)}, mx^{(C_1)}, \ldots, mx^{(C_{p-1})}\}$. The replacement of that pair of terms by the single term m:
(a) reduces the instruction number by $p-1$ units;
(b) does not increase the maximum processing time;
(c) reduces the average processing time. □

Consider now a set of product terms representing a discrete function. A substitution–elimination operation on this set consists in a sequence of the two following steps:
(a) replacement of the variable x_i by the constant h ($h \in \{0, 1, \ldots, m-1\}$) wherever it appears;
(b) suppression in the obtained list of all the terms contained in other terms.
Let $(\mathbf{x}_1, \mathbf{x}_0)$ be a partition of the set and let $\mathbf{e}_0$ be a fixed value of $\mathbf{x}_0$

$$(\mathbf{e}_0 \in \{0, 1, \ldots, m-1\}^p \quad \text{if} \quad \mathbf{x}_0 = \{x_0, x_1, \ldots, x_{p-1}\}).$$

LEMMA 1. *The representation of the function $f(\mathbf{x}_1, \mathbf{e}_0)$ as a set of terms obtained from the list of all the prime implicants of $f(\mathbf{x}_1, \mathbf{x}_0)$ by a series of substitution–elimination operations is the list of all the prime implicants of $f(\mathbf{x}_1, \mathbf{e}_0)$.*
Proof. If $p(\mathbf{x})$ is a prime implicant of f, then $p(\mathbf{x}_1, \mathbf{e}_0)$ is either a prime implicant of $f(\mathbf{x}_1, \mathbf{e}_0)$, or is included in $p'(\mathbf{x}_1, \mathbf{e}_0)$ where $p'(\mathbf{x})$ is another prime implicant of f; this derives immediately from the concept of prime implicant of a function. If $p(\mathbf{x}_1, \mathbf{e}_0)$ is a prime implicant of $f(\mathbf{x}_1, \mathbf{e}_0)$, it is included in an implicant of f and thus in a prime implicant of f. □

As an immediate consequence of that lemma we obtain:

THEOREM 11. *The trees obtained by applying substitution–elimination operations to the set of all the prime implicants of a function $f(\mathbf{x})$ are associated with irreducible standard sets. Furthermore, any irreducible standard set representing $f(\mathbf{x})$ is accessible by this process for an appropriate ordering of the substituted variables.*

The application of the tree-enumeration method to the list of all the prime implicants of a function could thus yield *C*-optimal algorithms. The reduction of the amount of computations is due to the fact that one only enumerates a restricted amount of incomplete trees instead of enumerating the whole set of complete trees. A further reduction is made possible by the following theorem 12.

THEOREM 12. *If the set of all the prime implicants of a function $f(\mathbf{x})$ may be transformed into a standard set by an appropriate ordering of the variables, then the associated decision algorithm is C-optimal.*

Proof. From the hypothesis, it immediately follows that

(a) the prime implicants are pairwise disjoint,

(b) the prime implicants do not form any consensus term: indeed, the prime implicant containing a consensus term would be disjoint of the consensus forming implicants.

Hence, all the prime implicants are essential and their set is thus the only irreducible standard set. The proof then follows from theorem 10. □

Algorithm 3.

Perform on the set of all the prime implicants of the function and in every possible way the substitution–elimination operations until they result in a family of standard sets.

Example 1 (continued, see page 33). The prime implicants of the function of example 1 are given in (20); the substitution elimination operations are performed in the scheme herebelow.

$$f = \{2x_0^{(2)}\,x_1^{(0)}, 2x_0^{(2)}\,x_2^{(1)}, 2x_1^{(2)}\,x_2^{(1)}, 1x_0^{(1,2)}\,x_1^{(0)}, 1x_0^{(1,2)}\,x_2^{(1)}\}. \tag{20}$$

$$(1)\left\{\begin{array}{ll}
x_0 = 0; & \{2x_1{}^{(2)}x_2{}^{(1)}\} \\
x_0 = 1; & (4)\left\{\begin{array}{l}
\{2x_1{}^{(2)}x_2{}^{(1)}, 1x_1{}^{(0)}, 1x_2{}^{(1)}\}\ (6)\left\{\begin{array}{l} x_1 = 0; \{1\} \\ x_1 = 1; \{1x_2{}^{(1)}\} \\ x_1 = 2; \{2x_2{}^{(1)}\} \end{array}\right. \\
(7)\left\{\begin{array}{l} x_2 = 0; \{1x_1{}^{(0)}\} \\ x_2 = 1; \{2x_1{}^{(2)} \vee 1\}\ (10)\left\{\begin{array}{l} x_1 = 0; \{1\} \\ x_1 = 1; \{1\} \\ x_1 = 2; \{2\} \end{array}\right. \\ x_2 = 2; \{1x_1{}^{(0)}\} \end{array}\right.
\end{array}\right. \\
x_0 = 2; & (5)\left\{\begin{array}{l}
\{2x_1{}^{(0)}, 2x_2{}^{(1)}\}\ (8)\left\{\begin{array}{l} x_1 = 0; \{2\} \\ x_1 = 1; \{2x_2{}^{(1)}\} \\ x_1 = 2; \{2x_2{}^{(1)}\} \end{array}\right. \\
(9)\left\{\begin{array}{l} x_2 = 0; \{2x_1{}^{(0)}\} \\ x_2 = 1; \{2\} \\ x_2 = 2; \{2x_1{}^{(2)}\} \end{array}\right.
\end{array}\right.
\end{array}\right.$$

$$(2)\left\{\begin{array}{ll}
x_1 = 0; & \{2x_0{}^{(2)}, 1x_0{}^{(1,2)}\}\ (11)\left\{\begin{array}{l} x_0 = 0; \{0\} \\ x_0 = 1; \{1\} \\ x_0 = 2; \{2\} \end{array}\right. \\
x_1 = 1; & (12)\left\{\begin{array}{l}
\{2x_0{}^{(2)}x_2{}^{(1)}, 1x_0{}^{(1,2)}x_2{}^{(1)}\}\ (13)\left\{\begin{array}{l} x_0 = 0; \{0\} \\ x_0 = 1; \{1x_2{}^{(1)}\} \\ x_0 = 2; \{2x_2{}^{(1)}\} \end{array}\right. \\
(14)\left\{\begin{array}{l} x_2 = 0; \{0\} \\ x_2 = 1; \{2x_0{}^{(2)}, 1x_0{}^{(1,2)}\}\ (15)\left\{\begin{array}{l} x_0 = 0; \{0\} \\ x_0 = 1; \{1\} \\ x_0 = 2; \{2\} \end{array}\right. \\ x_2 = 2; \{0\} \end{array}\right.
\end{array}\right. \\
x_1 = 2; & \{2x_2{}^{(1)}\}
\end{array}\right.$$

$$
(3)\left\{\begin{array}{l}
x_2=0;\ (16)\left\{\begin{array}{l}
\{2x_0{}^{(2)}x_1{}^{(0)},1x_0{}^{(1,2)}x_1{}^{(0)}\}\ (18)\left\{\begin{array}{l} x_0=0;\ \{0\}\\ x_0=1;\ \{1x_1{}^{(0)}\}\\ x_0=2;\ \{2x_1{}^{(0)}\}\end{array}\right.\\[2ex]
(19)\left\{\begin{array}{l} x_1=0;\{2x_0{}^{(2)}\vee 1x_0{}^{(1,2)}\}\ (22)\left\{\begin{array}{l}x_0=0;\{0\}\\ x_0=1;\{1\}\\ x_0=2;\{2\}\end{array}\right.\\ x_1=1;\ \{0\}\\ x_1=2;\ \{0\}\end{array}\right.
\end{array}\right.\\[4ex]
x_2=1;\ (17)\left\{\begin{array}{l}
\{2x_0{}^{(2)},2x_1{}^{(2)},1x_0{}^{(1,2)}\}\ (20)\left\{\begin{array}{l} x_0=0;\ \{2x_1{}^{(2)}\}\\ x_0=1;\ \{1\vee 2x_1{}^{(2)}\}\ (23)\left\{\begin{array}{l}x_1=0;\{1\}\\ x_1=1;\{1\}\\ x_1=2;\{2\}\end{array}\right.\\ x_0=2;\ \{2\}\end{array}\right.\\[2ex]
(21)\left\{\begin{array}{l} x_1=0;\{2x_0{}^{(2)},1x_0{}^{(1,2)}\}\ (24)\left\{\begin{array}{l}x_0=0;\{0\}\\ x_0=1;\{1\}\\ x_0=2;\{2\}\end{array}\right.\\ x_1=1;\{2x_0{}^{(2)},1x_0{}^{(1,2)}\}\ (\text{see } x_1=0)\\ x_1=2;\ \{2\}\end{array}\right.
\end{array}\right.\\[4ex]
x_2=2;\ \ \{2x_0{}^{(2)}x_1{}^{(0)},1x_0{}^{(1,2)}x_1{}^{(0)}\}\ \ (\text{see } x_2=0)
\end{array}\right.
$$

This scheme may be understood as follows (see also fig. 9). The substitution–elimination procedure may be started by any one of the variables x_0, x_1 and x_2; if x_0 is chosen e.g., the node 1 is considered (see the above scheme and fig. 9). For $x_0 = 0$ one obtains a standard set (denoted by s in fig. 9) and the computation ends; for $x_0 = 1$ or 2 the obtained set is not standard (see the nodes 4 and 5 respectively) and one has to continue the substitution–elimination operation with respect to the variables x_1 and x_2. From the nodes 4,5 one reaches the nodes 6,8 or 6,9 or 7,8 or 7,9 according to the couple of variables chosen i.e.: x_1, x_1 or x_1, x_2 or x_2, x_1 or x_2, x_2 respectively. The substitution–elimination operations end as soon as standard sets have been obtained.

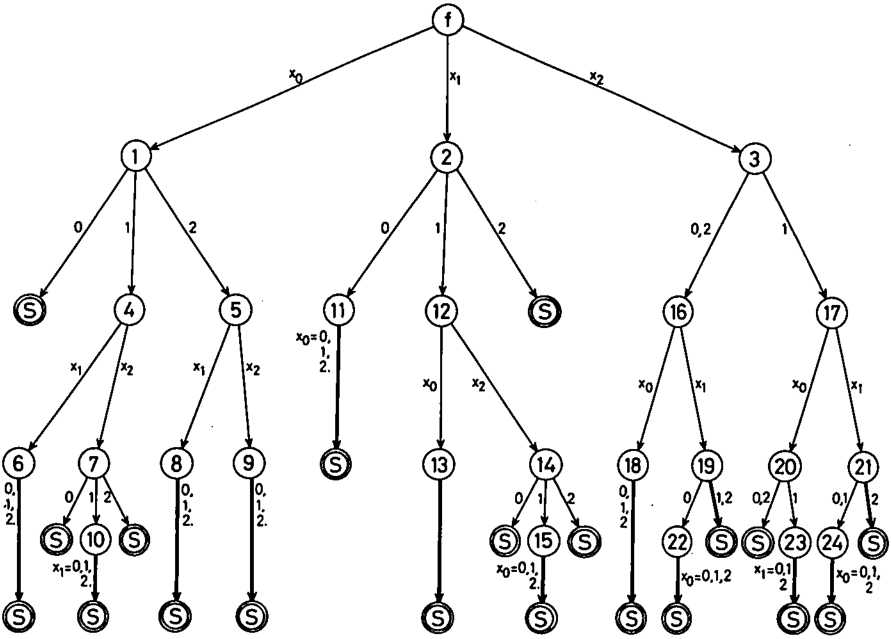


Fig. 9.

Example 2 (continued, see page 39). Set of the prime implicants of *f*:

$$\{\bar{x}_0\,\bar{x}_1,\ \bar{x}_0\,x_2,\ x_0\,\bar{x}_2\,\bar{x}_3,\ \bar{x}_1\,\bar{x}_2\,\bar{x}_3\}.$$

Substitution–elimination operations on this set:

$$(1)\left\{\begin{array}{ll} x_0=1; & \{\bar{x}_2\,\bar{x}_3\}\\[2ex] x_0=0; & \left\{\begin{array}{ll}\{\bar{x}_1, x_2\} & \begin{cases}x_2=1;\ \{1\}\\ x_2=0;\ \{\bar{x}_1\}\end{cases}\\[3ex] & \begin{cases}x_1=1;\ \{x_2\}\\ x_1=0;\ \{1\}\end{cases}\end{array}\right.\end{array}\right.$$

$$(2)\left\{\begin{array}{ll} x_1=1; & \{\bar{x}_0\,x_2, \bar{x}_0\,\bar{x}_2\,\bar{x}_3\}\\[2ex] x_1=0; & \left\{\begin{array}{ll}\{\bar{x}_0, \bar{x}_2\,\bar{x}_3\} & \begin{cases}x_3=1;\ \{\bar{x}_0\}\\ x_3=0;\ \{\bar{x}_2\}\end{cases}\\[3ex] & \begin{cases}x_2=1;\ \{\bar{x}_0\}\\ x_2=0;\ \{\bar{x}_3\}\end{cases}\\[3ex] & \begin{cases}x_0=1;\ \{x_2\,\bar{x}_3\}\\ x_0=0;\ \{1\}\end{cases}\end{array}\right.\end{array}\right.$$

$$(3)\quad \left\{\begin{array}{ll}
x_2 = 1; & \{\bar{x}_0\} \\
x_2 = 0; & \left\{\begin{array}{ll}
\{\bar{x}_0\,\bar{x}_1, x_0\,\bar{x}_3, \bar{x}_1\,\bar{x}_3\} &
\begin{array}{l}
\left\{\begin{array}{ll}
x_3 = 1; \{\bar{x}_0\,\bar{x}_1\} & \\
x_3 = 0; \{x_0, \bar{x}_1\} & \begin{array}{l}
\left\{\begin{array}{l} x_1 = 1; \{x_0\} \\ x_1 = 0\,; \{1\} \end{array}\right. \\
\left\{\begin{array}{l} x_0 = 1; \{1\} \\ x_0 = 0; \{\bar{x}_1\} \end{array}\right.
\end{array}
\end{array}\right. \\
\left\{\begin{array}{ll}
x_1 = 1; \{x_0\,\bar{x}_3\} & \\
x_1 = 0; \{\bar{x}_0, \bar{x}_3\} & \begin{array}{l}
\left\{\begin{array}{l} x_3 = 1; \{\bar{x}_0\} \\ x_3 = 0; \{1\} \end{array}\right. \\
\left\{\begin{array}{l} x_0 = 1; \{\bar{x}_3\} \\ x_0 = 0; \{1\} \end{array}\right.
\end{array}
\end{array}\right. \\
\left\{\begin{array}{l} x_0 = 1; \{\bar{x}_3\} \\ x_0 = 0; \{\bar{x}_1\} \end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right.$$

$$(4)\quad \left\{\begin{array}{ll}
x_3 = 1; & \left\{\begin{array}{ll}
\{\bar{x}_0\,\bar{x}_1, \bar{x}_0\,x_2\} &
\begin{array}{l}
\left\{\begin{array}{l} x_2 = 1; \{\bar{x}_0\} \\ x_2 = 0; \{\bar{x}_0\,\bar{x}_1\} \end{array}\right. \\
\left\{\begin{array}{l} x_1 = 1; \{\bar{x}_0\,x_2\} \\ x_1 = 0; \{\bar{x}_0\} \end{array}\right. \\
\left\{\begin{array}{ll}
x_0 = 1; \{0\} & \\
x_0 = 0; & \left\{\begin{array}{ll}
\{\bar{x}_1, x_2\} & \begin{array}{l}
\left\{\begin{array}{l} x_2 = 1; \{1\} \\ x_2 = 0; \{\bar{x}_1\} \end{array}\right. \\
\left\{\begin{array}{l} x_1 = 1; \{x_2\} \\ x_1 = 0; \{1\} \end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right.
\end{array}
\end{array}\right. \\
x_3 = 0; & \left\{\begin{array}{ll}
\begin{array}{l}\{\bar{x}_0\,\bar{x}_1, \bar{x}_0\,x_2, \\ x_0\,\bar{x}_2, \bar{x}_1\,\bar{x}_2\}\end{array} &
\begin{array}{l}
\left\{\begin{array}{ll}
x_2 = 1; & \{\bar{x}_0, x_0\,\bar{x}_1\} \\
x_2 = 0; & \left\{\begin{array}{ll}
\{x_0, \bar{x}_1\} & \begin{array}{l}
\left\{\begin{array}{l} x_1 = 1; \{x_0\} \\ x_1 = 0; \{1\} \end{array}\right. \\
\left\{\begin{array}{l} x_0 = 1; \{1\} \\ x_0 = 0; \{\bar{x}_1\} \end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right. \\
\left\{\begin{array}{ll}
x_1 = 1; & \{\bar{x}_0\,x_2, x_0\,\bar{x}_2\} \\
x_1 = 0; & \left\{\begin{array}{ll}
\{\bar{x}_0, \bar{x}_2\} & \begin{array}{l}
\left\{\begin{array}{l} x_2 = 1; \{\bar{x}_0\} \\ x_2 = 0; \{1\} \end{array}\right. \\
\left\{\begin{array}{l} x_0 = 1; \{\bar{x}_2\} \\ x_0 = 0; \{1\} \end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right. \\
\left\{\begin{array}{ll}
x_0 = 1; & \{\bar{x}_2\} \\
x_0 = 0: & \left\{\begin{array}{ll}
\{\bar{x}_1, \bar{x}_2\} & \begin{array}{l}
\left\{\begin{array}{l} x_2 = 1; \{\bar{x}_1\} \\ x_2 = 0; \{1\} \end{array}\right. \\
\left\{\begin{array}{l} x_1 = 1; \{\bar{x}_2\} \\ x_1 = 0; \{1\} \end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right.
\end{array}
\end{array}\right.
\end{array}\right.$$

5. Conclusions

The problem of optimizing multivalued decision algorithms is a typical example of a non-classical discrete optimization problem. It shares with these problems a high algorithmic complexity. Further investigations could be devoted to improve the presently available algorithms, to obtain efficient heuristics or to study functional properties, such as the degenerescence or the decomposability, the discovery of which has an impact on the computation method and on its result.

The problem is, however, very attractive, since it is a core problem that can be extended in various directions:

(a) synthesis of incompletely specified functions;

(b) simultaneous synthesis of a number of functions;

(c) use of multiplexers with more than one control input;

(d) general transformations of arbitrary microprograms into condition microprograms.

Several extensions and possible simplifications relative to the material presented in this paper may be found in refs 4, 14 and 15. The problem of evaluating sequentially the local value of a Boolean function has been tackled by various authors: besides the already quoted references, we should mention the work of Kuntzman [16]) who presents an algorithm for minimizing the number of literals in a standard set (lexicographical polynomial). That algorithm is of the leave-to-root type but is closer to exhaustion than the algorithm proposed in section 4.2. More recently Mange and Sanchez [15]), suggested an original method for achieving P_t-optimality: in essence, that method is inspired by an algorithm devised by Meisel [17]) to minimize the number of states in incompletely specified automata: it is merely an attempt to cover the Boolean cube by sets of pairwise disjoint T^0-cubes. The idea of covering the cube has also been exploited by Roegiers [18]) who exploited T^1-terms.

MBLE Research Laboratory — *Brussels, October 1977*

REFERENCES

[1]) C. Lee, Bell Syst. Tech. J. **38**, 985-999, 1959.
[2]) R. Boute, Euromicro Newsletter 3, 16-22, 1976.
[3]) T. Ito, Sigmicro Newsletter 4, 5-17, 1973.
[4]) M. Davio and A. Thayse, Sequential evaluation of Boolean functions, MBLE Internal Report R341, March 1977.
[5]) M. Davio, Hardware implementation of algorithmic computations, MBLE Internal Report R333, July 1976.
[6]) Y. Breitbart and A. Reiter, Acta Informatica 4, 107-117, 1975.
[7]) J. Perl and Y. Breitbart, Inf. Sci. **11**, 1-12, 1976.
[8]) U. Pooch, ACM Computing Surveys **6**, 125-151, 1974.

[9] P. Spira, IEEE Trans. Comput. **C-20**, 104-105, 1971.
[10] J. Nievergelt, ACM Computing Surveys **6**, 195-206, 1974.
[11] M. Davio and G. Bioul, Philips Res. Repts **25**, 370-388, 1970.
[12] F. Sellers, M. Hsiao and L. Bearnson, IEEE Trans. Comput. **C-17**, 676-683, 1968.
[13] A. Thayse and M. Davio, IEEE Trans. Comput. **C-22**, 409-420, 1973.
[14] A. Thayse, Optimization of binary decision algorithms, MBLE Internal Report R348, May 1977.
[15] D. Mange and E. Sanchez, Synthèse des fonctions logiques avec des multiplexers, paper submitted for publication to Digital Processes.
[16] J. Kuntzmann, Algèbre de Boole, Dunod, Paris, 1965.
[17] W. Meisel, IEEE Trans. Electr. Comput. **EC-16**, 508-509, 1967.
[18] D. Roegiers, private communication.

Philips J. Res., Vol. 33, 66-77, 1978 *R 980*

WORST-CASE ANALYSIS OF ALGORITHMS

A. Some g.c.d. algorithms

by C. van TRIGT

Abstract

A new method for obtaining the so-called worst-case behaviour of algorithms is presented. It directly associates with the algorithm a natural model in which the behaviour of the algorithm is analysed. The method is applied to some g.c.d. algorithms whose worst-case behaviour has been obtained in the past from the theory of continued fractions. The present method is shown to be effective and to yield the correct answers.

1. Introduction

Suppose we have an algorithm that terminates in a finite number of steps. Suppose furthermore that we have a finite set of inputs and a parameter $N \gg 1$, such that the size of the set increases with increasing N. We define the function which assigns to every input the number of steps in which the algorithm terminates on this input, and are interested in the maximum of the function on the set (in general some function of N) when $N \rightarrow \infty$. It is a rough measure of the efficiency of the algorithm. Though the problem is easy to state, it is unexpectedly difficult to answer. For most algorithms [1]) only "big-Oh" formulas are known, i.e. one can prove that the maximum is $O(\ln N)$, $O(N)$, $O(2^N)$ etc. for $N \rightarrow \infty$. The results are usually obtained by rough estimates. This is an unsatisfactory situation. We present here a new method by which the worst-case behaviour can be precisely calculated. The basic idea is this. Suppose we have a finite set of inputs on which the algorithm terminates. Suppose furthermore that the algorithm proceeds by repeatedly executing elementary steps where executing an elementary step has the following effect on an input: as a result either the final result is obtained (the step is the one by which the algorithm terminates) or a new (uniquely determined) input is obtained belonging to the set to which input a following step is then applied. As an example, consider the Euclidean g.c.d. algorithm with set of inputs (m,n), $1 \leqslant n \leqslant m \leqslant N$, $N \gg 1$. An elementary step changes (m,n) into $(n,m \bmod n)$. The step is terminating if $m \bmod n = 0$. If not, a new step is executed with input $(n,m \bmod n)$ which clearly belongs to the original set.

Usually, the algorithmic determination of a quantity takes place by letting the algorithm operate on *one* input and executing *all* the steps needed to obtain a result. The number of steps executed is in general a very *irregular function of the input* and calculating the maximum of this function is a rather hopeless affair.

Consider the, in some sense converse, situation in which as inputs *all* the members of the set are used but only *one* step of the algorithm is executed. The resulting set of new inputs (for the following step) is a subset of the previous set and usually the inclusion is proper, i.e. the set of inputs contracts under application of one step of the algorithm. It will be demonstrated that this contraction is a *regular function of the number of steps*, at least for the examples treated here. There is hope that this is a rather general rule. We wish to calculate the least number of steps $k + 1$ say such that after application of $k + 1$ steps the resulting space has become empty. Because the contraction is a regular function of the number of steps, this is now feasible. It is clear that if the entire set of inputs has become empty under application of $k + 1$ steps of the algorithm, no separate input can need more than $k + 1$ steps. The maximum number of steps is therefore equal or less than $k + 1$. In the examples treated here we easily find an input which needs exactly $k + 1$ steps so that the upper bound is attained and cannot be improved.

We shall apply the idea here to three g.c.d. algorithms. One of them is the classical Euclidean algorithm. It is an old result due to Lamé [2]) that the maximum number of steps is asymptotically

$$k + 1 = [\log (N \sqrt{5})/\log \{(1 + \sqrt{5})/2\}] - 2,$$

where $[a]$ denotes the entier of a. The result is obtained from the theory of continued fractions. In section 2 it will be rederived by the present method. The contracted spaces are a function of N and the Fibonacci numbers F_j, where j is the number of steps and $F_j = \{\phi^j - (1 - \phi)^j\}/\sqrt{5}$, $\phi = (1 + \sqrt{5})/2$. In section 3 we briefly deal with two other g.c.d. algorithms, one of which is the so-called least remainder algorithm. In this case the contracted spaces are a function of N and $a_j = \{(1 + \sqrt{2})^j + (1 - \sqrt{2})^j\}/2$. The maximum number of steps $k + 1$ is asymptotically $k + 1 = [\ln (2N)/\ln (1 + \sqrt{2})]$ for $N \to \infty$. This is faster than the Euclidean algorithm by almost a factor of two.

2. The Euclidean algorithm

The g.c.d. algorithm we shall investigate assigns to any ordered pair (m,n) with $1 \leqslant n \leqslant m \leqslant N$ its g.c.d. by repeated application of the following process: change (m,n) into $(n, m \bmod n)$ where $m \bmod n = m - [m/n]\, n$, $[m/n]$

entier of m/n; if $m \bmod n = 0$ stop, the g.c.d. is n; if $m \bmod n > 0$ repeat the process.

Applying this process once is called a step. More precisely let us define $t(m,n)$ as the number of steps used to calculate the g.c.d. of (m,n). We then have the following recursion relations:

$$\begin{aligned} &\text{(a) } t(m,n) = 1 + t(n, m \bmod n) \quad m \geqslant n \geqslant 1,\\ &\text{(b) } t(m,0) = 0,\\ &\text{(c) } t(m,m) = 1. \end{aligned} \tag{1}$$

The first rule is clear; the second states that the algorithm has ended one step earlier (with g.c.d. m); the third rule is a consequence of rules (a) and (b).

The mean number of steps when the pairs (m,n) are chosen within the triangle $1 \leqslant n \leqslant m \leqslant N$ with a certain probability $f(m,n)$ is defined as

$$\bar{t}_f = Q^{-1} \sum_{m=1}^{N} \sum_{n=1}^{m} t(m,n) f(m,n),$$

$$Q = \sum_{m=1}^{N} \sum_{n=1}^{m} f(m,n). \tag{2}$$

If $f(m,n) = 1$ everywhere then $Q = (N+1)N/2$, the number of lattice points within the triangle $1 \leqslant n \leqslant m \leqslant N$. It is clear that the knowledge of $\bar{t}_f$ for all possible functions f, is equivalent to knowing $t(m,n)$ for all pairs (m,n). For, trivially, if we take $f(m,n) = 1$ for a certain (m_0,n_0) and $f(m,n) = 0$ for all $(m,n) \neq (m_0,n_0)$, then $\bar{t}_f = t(m_0,n_0)$. It is therefore convenient to view f as some kind of test function, by means of which we can infer something about the behaviour of $t(m,n)$. We introduce (1) into (2) and obtain

$$\begin{aligned} \bar{t}_f &= Q^{-1} \sum_{m=1}^{N} \sum_{n_0=1}^{m} \{1 + t(n_0, m \bmod n_0)\} f(m,n_0)\\ &= 1 + Q^{-1} \sum_{n_0=1}^{N-1} \sum_{m=n_0+1}^{N} t(n_0, m \bmod n_0) f(m,n_0)\\ &= 1 + Q^{-1} \sum_{n_0=1}^{N-1} \sum_{m=n_0+1}^{N} \sum_{n_1=0}^{n_0-1} t(n_0,n_1)\, \delta(n_1, m \bmod n_0) f(m,n_0)\\ &= 1 + Q^{-1} \sum_{n_0=1}^{N-1} \sum_{n_1=0}^{n_0-1} t(n_0,n_1) f_0(n_1,n_0)\\ &= 1 + Q^{-1} \sum_{n_0=2}^{N-1} \sum_{n_1=1}^{n_0-1} t(n_0,n_1) f_0(n_1,n_0) \end{aligned} \tag{3}$$

with

$$f_0(n_1,n_0) = \sum_{m=n_0+1}^{N} \delta(n_1,m \bmod n_0) f(m,n_0). \tag{4}$$

The Kronecker δ-function is defined as

$$\begin{aligned} \delta(n,m \bmod n_0) &= 1 \quad \text{if } n_1 = m \bmod n_0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The meaning of the equations (3) and (4) is the following: when the g.c.d. algorithm is applied once to all pairs (m,n_0), then these pairs are mapped on other pairs (n_0,n_1). If the pairs (m,n_0) are originally present with probability distribution $f(m,n_0)$, then the (n_0,n_1) are present with distribution $f_0(n_1,n_0)$ given by (4). Apart from the fact that the summation in (2) includes the diagonal $n = m$, equations (2) and (3) are formally the same. The procedure can therefore be repeated and one easily proves by induction the following formula

$$\begin{aligned} \bar{t}_f = 1 + Q^{-1} \sum_{j=0}^{k} \left\{ \sum_{n_j=2}^{N-j-1} \sum_{n_{j+1}=1}^{n_j-1} f_j(n_{j+1},n_j) \right\} \\ + Q^{-1} \sum_{n_{k+1}=2}^{N-k+2} \sum_{n_{k+2}=1}^{n_{k-1}-1} t(n_{k+1},n_{k+2}) f_{k+1}(n_{k+2},n_{k+1}) \end{aligned}$$

with

$$f_j(n_{j+1},n_j) = \sum_{n_{j-1}=n_j+1}^{N-j} \delta(n_{j+1},n_{j-1} \bmod n_j) f_{j-1}(n_j,n_{j-1}). \tag{5}$$

The function $f_j(n_{j+1},n_j)$ describes how the original distribution function $f(m,n_0)$ is changed under $(j+1)$ applications of the g.c.d. algorithm.

We shall now prove that the f_j are different from zero only on a nested sequence of subspaces of the initial space of input variables, the triangle $1 \leqslant n \leqslant m \leqslant N$. More precisely, let us define the Fibonacci numbers as follows $F_0 = 0$, $F_1 = 1$, $F_{n+2} = F_{n+1} + F_n$, $n \geqslant 0$. It will be shown by induction for $j \geqslant 0$ that $f_j(n_{j+1},n_j) = 0$ in

$$\{1 \leqslant n_{j+1} < n_j \leqslant N-j-1\} \cap \left\{ \frac{F_{j+2}}{F_{j+1}} \left(\frac{N}{F_{j+2}} - n_j \right) < n_{j+1} \right\} \tag{6}$$

independent of the initial distribution function $f(m,n_0)$. The meaning of the result is that when the g.c.d. algorithm is applied j times to the space of input variables, this space is contracted to the part of it to the left of and including the line $N/F_{j+1} - F_{j+2}\, n_j/F_{j+1} = n_{j+1}$. (See fig. 1.)

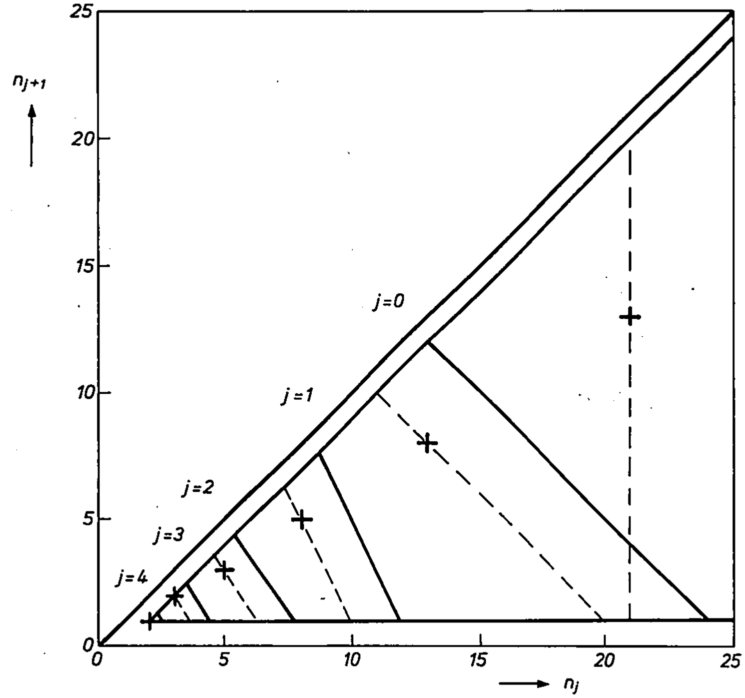


Fig. 1. Contraction of the space of input variables. The line segments (———) marked $j = 0, 1, 2$ are the pieces of the lines $N - F_{j+2}\, n_j = F_{j+1}\, n_{j+1}$ which satisfy $1 \leqslant n_{j+1} < n_j$ with $N = 25$. The segments (- - - - -) satisfy the same equations for $N = 21$, a Fibonacci number. The triangles to the left and including the segments are the domains outside of which the functions f_j vanish, cf. eq. (6).

Basically, the proof only exploits the fact that the Kronecker $\delta(n_{j+1},n_{j-1} \bmod n_j$ in (5) is different from zero only at a few points in the range of summation, namely at $n_{j-1} = n_{j+1} + \lambda n_j$, $\lambda = 1, \ldots, [(N - j - n_{j+1})/n_j]$, where $[\alpha]$ denotes the entier of α. It will be shown that this simple property induces a contraction of the space of input variables under application of the algorithm. Since the property of, let us call it "being sparsely different from zero" is a basic one for algorithms, the point of view may turn out to be useful in the analysis of more difficult cases (the simple linear relation $n_{j-1} = n_{j+1} + \lambda n_j$ of course being replaced by more difficult ones).

We shall need one auxiliary lemma concerning the Fibonacci numbers viz.

$$F_N = F_{j+1}\, F_{N-j} + F_j\, F_{N-j-1}, \quad j \geqslant 1. \tag{7}$$

The basis $j = 1$ is obvious. Let (7) be true for j. We prove it for $j + 1$ as follows.

$$\begin{aligned}F_{j+2}\,F_{N-j-1}+F_{j+1}\,F_{N-j-2}&\\ &=(F_{j+1}+F_j)(F_{N-j}-F_{N-j-2})+F_{j+1}\,F_{N-j-2}\\ &=F_{j+1}\,F_{N-j}+F_j\,F_{N-j-1}=F_N.\end{aligned}$$

We now prove (6) by induction. For $j=0$ we have from (4)

$$f_0(n_1,n_0)=\sum_{m=n_0+1}^{N}\delta(n_1,m \bmod n_0)\,f(m,n_0)$$

to be considered in the domain $1\leqslant n_1<n_0\leqslant N-1$, see (5). Informally, when m runs through n_0+1, n_0+2, ..., $2n_0-1$, $2n_0$, $m \bmod n_0$ runs through the values $1, 2, \ldots, n_0-1, 0$. Now suppose that $2n_0>N$ then $m \bmod n_0$ does not take all of the values $1, 2, \ldots, n_0-1$. We can choose n_1 such that the equality $n_1=m \bmod n_0$ cannot occur. Consequently, the Kronecker-delta will then be zero in the entire range of summation and thus $f_0(n_1,n_0)$ will also be zero. The formal proof runs as follows. Let $2n_0>N$. In the range of summation we have under this condition

$$n_0+1\leqslant m\leqslant N$$

$$1+\frac{1}{n_0}\leqslant\frac{m}{n_0}\leqslant\frac{N}{n_0}<2$$

and thus $[m/n_0]=1$. Furthermore

$$m \bmod n_0=m-\left[\frac{m}{n_0}\right]n_0=m-n_0.$$

Consequently, if $2n_0>N$ and $n_0+1\leqslant m\leqslant N$, we have

$$1\leqslant m \bmod n_0\leqslant N-n_0.$$

The Kronecker $\delta(n_1,m \bmod n_0)$ is different from zero only if $n_1=m \bmod n_0$ and if $N-n_0<n_1$ this equality cannot occur. Hence, $\delta(n_1,m \bmod n_0)$ is then equal to zero in the entire range of summation and $f_0(n_1,n_0)$ is also zero. Thus $f_0(n_1,n_0)$ is zero in

$$\{1\leqslant n_1<n_0\leqslant N-1\}\cap\{2n_0>N\}\cap\{N-n_0<n_1\}.$$

But since the relations $n_0>n_1$ and $n_1\geqslant N-n_0$ imply $2n_0>N$ we have that $f_0(n_1,n_0)$ is zero in

$$\{1\leqslant n_1<n_0\leqslant N-1\}\cap\{N-n_0<n_1\}.$$

This is the case $j=0$ of (6). We now assume (6) for j and prove it for $j+1$.

Apart from a few technicalities, this proof is the same as for the case $j = 0$ so that this part (until eq. (9)) may be omitted at first reading.

By equation (5) we have

$$f_{j+1}(n_{j+2},n_{j+1}) = \sum_{n_j=n_{j+1}+1}^{N-j-1} \delta(n_{j+2},n_j \bmod n_{j+1}) f_j(n_{j+1},n_j) \tag{8}$$

and by induction hypothesis $f_j(n_{j+1},n_j) = 0$ in

$$\{1 \leqslant n_{j+1} < n_j \leqslant N-j-1\} \cap \left\{\frac{F_{j+2}}{F_{j+1}}\left(\frac{N}{F_{j+2}} - n_j\right) < n_{j+1}\right\}.$$

We may assume $j \leqslant N-3$ (otherwise the range of summation is void and the sum zero) and show that for $0 \leqslant j \leqslant N-3$

$$\frac{N}{F_{j+2}} - \frac{F_{j+1}}{F_{j+2}} n_{j+1} \leqslant N-j-1.$$

The proof is again by induction. The basis $j = 0$ is obvious. Assuming the result for j, we have for $j+1$

$$\frac{N}{F_{j+3}} - \frac{F_{j+2}}{F_{j+3}} n_{j+2} \leqslant \frac{F_{j+2}}{F_{j+3}}\left(\frac{N}{F_{j+2}} - 1\right)$$

$$\leqslant \frac{F_{j+2}}{F_{j+3}}\left(N-j-2+\frac{F_{j+1}}{F_{j+2}} n_{j+1}\right) \leqslant \frac{F_{j+2}}{F_{j+3}}\left(1+\frac{F_{j+1}}{F_{j+2}}\right)(N-j-2)$$

$$= N-j-2$$

Consequently, the upper limit $N-j-1$ in the sum may be replaced by $\leqslant (N - F_{j+1}\, n_{j+1})/F_{j+2}$.

Let $2n_{j+1} > (N - F_{j+1}/n_{j+1})/F_{j+2}$. By the same argument as in the case $j = 0$ we have $[n_j/n_{j+1}] = 1$ and $n_j \bmod n_{j+1} = n_j - [n_j/n_{j+1}]\, n_{j+1} = n_j - n_{j+1}$. In the summation in eq. (8) $n_j \bmod n_{j+1}$ takes the values $n_j \bmod n_{j+1} = 1, \ldots, (N - F_{j+1}\, n_{j+1})/F_{j+2} - n_{j+1}$.

If $n_{j+2} > (N - F_{j+1}\, n_{j+1})/F_{j+2} - n_{j+1}$, or by application of the recursion relation for the Fibonacci numbers, if $(N - F_{j+3}\, n_{j+1})/F_{j+2} < n_{j+2}$ the equality $n_{j+2} = n_j \bmod n_{j+1}$ cannot occur so that $\delta(n_{j+2}, n_j \bmod n_{j+1})$ is zero in the entire range of summation in (8). Hence, $f_{j+1}(n_{j+2},n_{j+1})$ is zero in

$$\{1 \leqslant n_{j+2} < n_{j+1} \leqslant N-j-2\} \cap \left\{2n_{j+1} > \frac{F_{j+1}}{F_{j+2}}\left(\frac{N}{F_{j+1}} - n_{j+1}\right)\right\}$$

$$\cap \left\{\frac{F_{j+3}}{F_{j+2}}\left(\frac{N}{F_{j+3}} - n_{j+1}\right) < n_{j+2}\right\}. \tag{9}$$

However, the second inequality in eq. (9) is satisfied because of the third and the first ($n_{j+2} < n_{j+1}$) and may be omitted. The proof of (6) by induction has been completed.

There is a second, more intuitive proof of the same result, showing clearly that the contraction of the space of input variables under application of the g.c.d. algorithm is associated with the fact that $\delta(n_{j+1}, n_{j-1} \bmod n_j)$ in (5) is "sparsely different from zero" as mentioned earlier. Consider

$$f_{j+1}(n_{j+2}, n_{j+1}) = \sum_{n_j} \delta(n_{j+2}, n_j \bmod n_{j+1})\, f_j(n_{j+1}, n_j),$$

where $n_j = n_{j+1} + 1, \ldots, \leqslant (N - F_{j+1}\, n_{j+1})/F_{j+2}$.

The Kronecker $\delta(n_{j+2}, n_j \bmod n_{j+1})$ is different from zero in

$$1 \leqslant n_{j+2} < n_j \leqslant N - j - 2 \quad \text{for} \quad n_j = n_{j+2} + \lambda n_{j+1}$$

where λ takes the values

$$1, 2, \ldots, \leqslant \left(\frac{N - F_{j+1}\, n_{j+1}}{F_{j+2}} - n_{j+2}\right) / n_{j+1}$$

but the range of summation is void if

$$\frac{N - F_{j+1}\, n_{j+1}}{F_{j+2}} - n_{j+2} < n_{j+1}.$$

By the recursion relation for the Fibonacci numbers this is equivalent to

$$\frac{F_{j+3}}{F_{j+2}} \left(\frac{N}{F_{j+3}} - n_{j+1}\right) < n_{j+2}$$

so that again the last induction step has been proved, namely that $f_{j+1}(n_{j+2}, n_{j+1})$ is zero in

$$\{1 \leqslant n_{j+2} < n_{j+1} \leqslant N - j - 2\} \cap \left\{\frac{F_{j+3}}{F_{j+2}} \left(\frac{N}{F_{j+3}} - n_{j+1}\right) < n_{j+2}\right\}.$$

Let us now derive Lamé's result concerning the maximum number of steps in which the Euclidean algorithm terminates. The algorithm terminates after a certain number of steps which is determined by the requirement that the space where $f_j(n_{j+1}, n_j)$ is different from zero, is empty. In other words, the space where $f_j(n_{j+1}, n_j)$ is zero i.e. the space

$$\{1 \leqslant n_{j+1} < n_j \leqslant N - j - 1\} \cap \left\{\frac{F_{j+2}}{F_{j+1}} \left(\frac{N}{F_{j+2}} - n_j\right) < n_{j+1}\right\}.$$

should contain the entire triangle $1 \leqslant n_{j+1} < n_j \leqslant N - j - 1$. Therefore, for all n_j, n_{j+1} in the triangle we should have $N < F_{j+2}\, n_j + F_{j+1}\, n_{j+1}$ and this condition is satisfied if it is satisfied for $n_j = 2$, $n_{j+1} = 1$. By means of the recursion relation for the Fibonacci numbers, this reduces to $N < F_{j+4}$. We have the asymptotic relation for $j \gg 1$, $\sqrt{5}\, F_{j+4} \sim \phi^{j+4}$, $\phi = (\sqrt{5} + 1)/2$ (ref. 2), and thus the Euclidean algorithm terminates for all j satisfying

$$j + 4 > \log (N \sqrt{5})/\log \phi. \tag{10}$$

Now observe that the function $f_j(n_{j+1}, n_j)$ is obtained after $j + 1$ steps and that we want the least j, call it k, satisfying (10). We then obtain Lamé's result that the Euclidean algorithm terminates in at most $k + 1$ steps with

$$k + 1 \sim [\log (N \sqrt{5})/\log \phi] - 2, \qquad N \to \infty. \tag{11}$$

This does not exclude a priori that the algorithm actually terminates in less steps. We have proved that $f_j(n_{j+1}, n_j)$ is zero in the space defined by eq. (6) but this does not mean that $f_j(n_{j+1}, n_j)$ cannot be zero in a larger space. If this would be true, then the contraction would be faster and the algorithm would terminate in less steps.

We shall now show that in general the space where $f_j(n_{j+1}, n_j)$ is zero is not larger than as defined by eq. (6). Consider the input pair (F_N, F_{N-1}) with $F_N = N$. After j applications of the g.c.d. algorithm the pair has become (F_{N-j}, F_{N-j-1}). From (7) we have for $j \geqslant 1$

$$\frac{F_{j+1}}{F_j}\left(\frac{F_N}{F_{j+1}} - F_{N-j}\right) = F_{N-j-1}$$

and this identity expresses precisely that the pair (F_{N-j}, F_{N-j-1}) is situated on the boundary of the region where $f_{j-1}(n_j, n_{j-1})$ is not zero, see (6). Hence, the region where $f_j(n_{j+1}, n_j)$ is zero, is in general not larger than is indicated in (6), and as a consequence (11) cannot in general be improved. This completes the discussion of the Euclidean algorithm.

3. The least remainder and slow g.c.d. algorithms

The least remainder algorithm calculates the g.c.d. of two integers m and n by repeatedly transforming the pair (m,n) into $(n, m \bmod n)$ if $m \bmod n < n/2$ and into $(n, n - m \bmod n)$ if $m \bmod n \geqslant n/2$. It is verified that this is equivalent with transforming the pair (m,n) into

$$(n, \left|m \bmod n - [\tfrac{1}{2} + (m \bmod n)/n]\, n\right|) = (n, \left|m - [\tfrac{1}{2} + m/n]\, n\right|).$$

Of course we have

$$0 \leqslant \left|m - [\tfrac{1}{2} + m/n]\, n\right| \leqslant n/2. \tag{12}$$

The analysis of the least remainder algorithm is essentially the same as for the Euclidean algorithm. The number of steps used for calculating the g.c.d. is again defined to be $t(m,n)$. The recursion relations are (cf. eq. (1))

$$\begin{aligned}&\text{(a)}\ \ t(m,n) = 1 + t(n, \left|m - [\tfrac{1}{2} + m/n]\, n\right|),\\ &\text{(b)}\ \ t(m,0) = 0.\end{aligned} \tag{13}$$

The mean number of steps is defined as in eq. (2). By the methods of sec. 2 one proves by induction

$$t_f = 1 + Q^{-1}\sum_{j=0}^{k}\left\{\sum_{n_{j+1}=1}^{(N-1)\,2^{-j-1}}\ \sum_{n_j=2n_{j+1}}^{(N-1)\,2^{-j}} f_j(n_{j+1},n_j)\right\}$$

$$+\ Q^{-1}\sum_{n_{k+2}=1}^{(N-1)\,2^{-k-2}}\ \sum_{n_{k+1}=2n_{k+2}}^{(N-1)\,2^{-k-1}} t(n_{k+1},n_{k+2})\,f_{k+1}(n_{k+2},n_{k+1}), \tag{14}$$

where $f_j(n_{j+1},n_j)$ for $j \neq 0$ is defined in terms of $f_{j-1}(n_j,n_{j-1})$ according to the following equation

$$f_j(n_{j+1},n_j) = \sum_{n_{j-1}=2n_j}^{(N-1)\,2^{-j+1}} \delta(n_{j+1}, \left|n_{j-1} - [\tfrac{1}{2} + n_{j-1}/n_j]\, n_j\right|)\, f_{j-1}(n_j,n_{j-1}) \tag{15}$$

and for $j = 0$ according to

$$f_0(n_1,n_0) = \sum_{m=n_0+1}^{N} \delta(n_1, \left|m - [\tfrac{1}{2} + m/n_0]\, n_0\right|)\, f(m,n_0).$$

It is understood in eqs (14) and (15) that the summations extend over all integers smaller or equal to $(N-1)\,2^{-j-1}$, $(N-1)\,2^{-j}$, $(N-1)\,2^{-j+1}$.

We shall now prove that $f_0(n_1,n_0)$ is zero in the domain

$$\{1 \leqslant n_1 \leqslant n_0/2 \leqslant (N-1)/2\} \cap \{N - n_0 < n_1\}. \tag{16}$$

Let $N < 3n_0/2$. In the range of summation $n_0 + 1 \leqslant m \leqslant N$ we have as a consequence of this inequality $1/n_0 + 3/2 \leqslant m/n_0 + \frac{1}{2} < 2$ and thus $[m/n_0 + \frac{1}{2}] = 1$. Hence we also have

$$1 \leqslant \left|m - [m/n_0 + \tfrac{1}{2}]\, n_0\right| \leqslant N - n_0.$$

If we also impose the condition $N - n_0 < n_1$, then in the second equation of (15) the δ-function is always zero in the entire range of summation and thus $f_0(n_1,n_0)$ is zero in the domain

$$\{1 \leqslant n_1 \leqslant n_0/2 \leqslant (N-1)/2\} \cap \{n_0 > 2N/3\} \cap \{N - n_0 < n_1\}.$$

However, the first and the third inequalities imply the second, which is therefore superfluous and may be omitted. Equation (16) has been established. It is left to the reader to prove by the same methods that $f_j(n_{j+1},n_j)$ is zero in the domain

$$\left\{1 \leqslant n_{j+1} \leqslant \frac{n_j}{2} \leqslant \frac{N-1}{2^{j+1}}\right\} \cap \{N - a_{j+1}n_j < a_j n_{j+1}\}, \tag{17}$$

where the numbers a_j satisfy the recursion relation $a_{j+1} = 2a_j + a_{j-1}$ with boundary conditions $a_0 = 1$, $a_1 = 1$. The solution of this recursion relation is $a_j = \{(1 + \sqrt{2})^j + (1 - \sqrt{2})^j\}/2$. Since $|1 - \sqrt{2}| < 1$ we also have the asymptotic relation $a_j \sim (1 + \sqrt{2})^j/2$. We now calculate the least number of steps in which the algorithm terminates. As in the foregoing section this number is determined by the requirement that the space where the f_j are different from zero is empty. We must have that for all n_{j+1}, n_j in the triangle

$$1 \leqslant n_{j+1} \leqslant n_j/2 \leqslant (N-1)/2^{j+1}$$

the inequality $N < a_{j+1}\, n_j + a_j\, n_{j+1}$ is satisfied. This is the case, if it satisfied for $n_j = 2$, $n_{j+1} = 1$ and we obtain that the functions f_j are identically zero for all j satisfying

$$N < 2a_{j+1} + a_j = a_{j+2}.$$

If N is large we may use the asymptotic relation for the a_j to obtain

$$j + 2 > \ln(2N)/\ln(1 + \sqrt{2}). \tag{18}$$

The function f_j is obtained after executing $j + 1$ steps (recall that f_0 is obtained after executing 1 step), is identically zero for any j satisfying (18), or, in other words, the space of inputs is empty for all j satisfying (18). The least number of steps, $k + 1$, say, in which the least remainder algorithm terminates, is then the least $j + 1$ which satisfies eq. (18). We therefore have

$$k + 1 \sim [\ln(2N)/\ln(1 + \sqrt{2})], \qquad N \to \infty. \tag{19}$$

This estimate cannot be improved. This is easily proved if one observes that if N is such that $N > 2$ and $N = a_N + a_{N-1}$ for some a_N, then the algorithm transforms the pair (N,a_N) into (a_N,a_{N-1}), (a_{N-1},a_{N-2}) etc. By induction one proves

$$N = a_{j+1}\, a_{N-j} + a_j\, a_{N-j-1}.$$

This equality tells us that the pairs (a_{N-j},a_{N-j-1}) are precisely situated on the boundary of the regions where the functions f_j are not zero cf. equation (17)). The contraction of the space of input variables is therefore in general not faster than as indicated by (17).

It is left to the reader to prove by the methods expounded above that the so-called slow g.c.d. algorithm terminates in at most N steps and that this upper bound is actually attained. The slow g.c.d. finds the g.c.d. of two integers (m,n), $1 \leqslant n \leqslant m \leqslant N$ as follows: (1) change the pair (m,n) into $(m-n,n)$ if $m-n \geqslant n$ or into $(n,m-n)$ if $n > m-n$; (2) if the new pair (m,n) has $n = 0$, stop, the g.c.d. is n, otherwise repeat the process. The application of the procedure is called a step. The result follows directy when it has been proved that the functions $f_j(n_{j+1},n_j)$ vanish within the regions

$$\{1 \leqslant n_{j+1} \leqslant n_j \leqslant N - j - 1\} \cap \{n_{j+1} > N - (j+1)\, n_j\}$$

and that this estimate cannot be improved.

4. Discussion

It has been shown that the space of input variables for a number of g.c.d. algorithms contracts upon repeated application of the steps of the algorithm and that this qualitative idea can be worked out smoothly in these cases, resulting in a complete quantitative picture of the contraction, from which the worst-case behaviour is readily obtained. Of course, the g.c.d. algorithms dealt with are extremely simple and it is not sure that methods that turn out to be effective for obtaining the worst-case behaviour here, will also be effective in the analysis of more difficult algorithms. On the other hand, it is observed that the basic equations (2) to (5) are essentially general and that the special character of the algorithm is only used for the precise calculation of the contraction, so that there is hope that also less simple algorithms can be dealt with.

We have also tried to calculate the mean number of steps needed by the Euclidean algorithm when the input variables are present with equal probability, i.e. $\bar{t}_f$ in (2) with $f(m,n) = 1$ in $1 \leqslant n \leqslant m \leqslant N$, a famous open problem, cf. ref. 2. This calculation requires more information concerning the distribution functions $f_j(n_{j+1},n_j)$ in (5) than that contained in eq. (6). It is easy to show that $f_0(n_1,n_0) = [(N-n_1)/n_0]$ but our attempts to calculate the f_j for $j > 0$ when $N \gg 1$ have not met with success. We therefore leave the problem open.

Philips Research Laboratories *Eindhoven, January 1978*

REFERENCES

1) A. Aho, J. Hopcroft and J. D. Ullman, The design and analysis of computer algorithms, Addison Wesley Publ. Co., Reading, 1975.
2) D. E. Knuth, The art of computer programming Vol. II, Addison Wesley Publ. Co., Reading, 1975.

ON THE TRANSPOSITION OF LINEAR TIME-VARYING DISCRETE-TIME NETWORKS AND ITS APPLICATION TO MULTIRATE DIGITAL SYSTEMS

by T. A. C. M. CLAASEN and W. F. G. MECKLENBRÄUKER

Abstract

Time-varying discrete-time networks are considered and their description by means of a transmission function is given. Such a description can be applied to discrete-time networks which contain e.g. modulators and subsystems operating at different sampling rates. Two forms of Tellegen's theorem are derived for these networks. Each of these forms suggests a definition of transposition, called hermitian transpose and generalized transpose respectively. The generalized transpose can be seen as a generalization of the transposition concept defined for time-invariant networks which it includes as a special case. For networks with real parameters the two transposition concepts are the same, but hermitian transposition has certain advantages for systems with complex parameters. A transposition theorem is discussed that relates the transmission function of either form of transpose network to that of the original network. As an application of this theorem a sensitivity analysis is given. Finally an extension of the foregoing theory is discussed for networks containing both analogue and digital parts.

1. Introduction

The complexity of a digital system for signal processing depends inter alia on the number of arithmetical operations that must be performed per unit of time. This number is proportional to the sampling rate at which the system operates. One of the aims in the design of a digital system is therefore to set the sampling rate at its lowest possible value. On the other hand it is known that a digital signal with a sampling rate $f_s = 1/T$ can only uniquely represent frequencies up to $f_s/2$ so that f_s must be higher than twice the highest frequency occurring in the signal. If the whole digital system operates at the same sampling rate this rate is determined by the highest frequency component that will ever be present in the system.

A more economical use of the arithmetical units can often be made by using different sampling rates for different parts of the system. Each sampling rate can then be adapted to the spectral content of the signals to be processed in the corresponding subsystem. Several such multi-rate processing systems have recently been proposed in the literature [1-7]). The increase or decrease of the sampling rate that is necessary to interconnect the various subsystems can be implemented very easily if the sampling rates are related by integer factors [1]). Introduction of the sampling rate increase (SRI) or sampling rate decrease (SRD) does not affect the linearity but makes the system time-variant as will be shown in sec. 2. It is therefore clear that many implementations of digital signal-processing schemes are time-varying discrete-time systems.

In this paper a description of linear discrete-time systems is given that takes into account these time variations. In section 2 the concepts of impulse response and transmission function are introduced. In section 3 two forms of Tellegen's theorem for these linear time-varying discrete-time systems are derived. Each of these forms suggests a definition of transposition that will be called hermitian transpose and generalized transpose respectively. For systems with real parameters the two forms of transposition are the same. The generalized transpose generalizes the concept of transposition as defined for time-invariant systems [8]), which it contains as a special case *). The hermitian transpose has the advantage that it yields the same result when applied to a network with complex parameters or to the real implementation of it. These forms of transposition are discussed in sec. 4, resulting in a transposition theorem that relates the transmission function of both forms of transpose networks to that of the original network.

Transposition of time-invariant networks leaves the transmission function the same and thus offers an alternative implementation of a transmission function [8]). Transposition when applied to time-varying systems yields in general a different transmission function, as can be expected from the fact that the input and output sampling rates of a system and its transpose need not be the same. The transpose system implements what can be called the complementary operation of that performed by the original system. Due to this property, transposition naturally arises in system analysis and synthesis and may lead to efficient designs of systems performing such complementary operations once the implementation of the original operation has been found. As an example, it is shown in sec. 5 that transposition of a decimator leads to an interpolator and of a modulator to a demodulator and vice versa.

Tellegen's theorem can also be used to obtain expressions for the sensitivity of the transmission function of a system to changes in the parameters of the network. These expressions will be derived in sec. 6. Finally in sec. 7 it will be shown that the concepts introduced before can be extended to incorporate networks containing both analogue and digital elements.

*) This is the reason that we have preferred to speak of transposition rather than of duality, which is more customary in control theory [9]), or of adjointness, which is used in mathematics [10]).

2. Description of linear time-varying discrete-time systems

Every linear discrete-time system with one input and one output can be described by an impulse response $h(n,m)$, which is the response of the system to the input signal $x(n) = u(n - m)$, where $u(n)$ is the unit sample sequence *):

$$u(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0. \end{cases} \tag{1}$$

The output $y(n)$ for an arbitrary input signal $x(n)$ is given by

$$y(n) = \sum_{m=-\infty}^{\infty} h(n,m)\, x(m). \tag{2}$$

For time-invariant systems $h(n,m)$ depends only on the difference $n - m$ and thus takes the simpler form

$$h(n,m) = \tilde{h}(n - m), \tag{3}$$

which makes (2) a discrete convolution. A frequency domain description of a linear discrete-time system can be obtained by means of the Fourier transform for discrete-time signals:

$$X(\theta) = \sum_{n=-\infty}^{\infty} x(n) \exp(-\mathrm{j}n\theta) \tag{4}$$

with inverse transform

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\theta) \exp(\mathrm{j}n\theta)\, \mathrm{d}\theta. \tag{5}$$

In these expressions θ is a relative frequency which is related to the actual frequency ω by

$$\theta = \omega T, \tag{6}$$

where T is the sampling period of $x(n)$.

The transmission function $H(\theta,\xi)$ of a system with impulse response $h(n,m)$ is defined by **)

*) In contrast to the conventional notation we use $u(n)$ for the unit sample sequence and reserve the symbol δ for the Dirac function, which will be used later.

**) The transmission function so defined is the Fourier transform of the frequency-response function that usually is used in the analysis of linear time-varying systems [11,12]. It is the discrete-time analogon of the bi-frequency system function introduced by Zadeh [13].

$$H(\theta,\xi) \overset{\Delta}{=} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h(n,m) \exp\,[-\mathrm{j}(n\theta - m\xi)]. \tag{7}$$

For a system with a real impulse response it follows from (7) that

$$H(\theta,\xi) = H^*(-\theta,-\xi), \tag{8}$$

where the asterisk denotes complex conjugation.

With the above definition of the transmission function the relation between output spectrum $Y(\theta)$ and input spectrum $X(\theta)$ takes the form

$$Y(\theta) = \int_{-\pi}^{\pi} H(\theta,\xi)\, X(\xi)\, \mathrm{d}\xi. \tag{9}$$

In general the sampling period of $y(n)$ may differ from that of $x(n)$ and will be denoted by T_2 and T_1 respectively. From the frequency relation specified by eq. (6) it then follows that $X(\omega T_1)$ is the value of the input spectrum at frequency ω and $Y(\Omega T_2)$ the value of the output spectrum at frequency Ω. In a linear system these values are linearly related, and eq. (9) states that the proportionality factor is precisely $H(\Omega T_2,\omega T_1)$.

As an example the decrease in sampling rate by an integer factor N will be considered, which has the input–output relation [1])

$$y(n) = x(nN) \qquad \forall\, n. \tag{10}$$

Comparison of (10) with (2) yields

$$h(n,m) = u(nN - m). \tag{11}$$

Since $h(n,m)$ given by eq. (11) is not of the form of eq. (3) it must be concluded that the sampling rate decrease is a time-varying element. Its transmission function can be obtained from eq. (7):

$$H(\theta,\xi) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \exp\,[-\mathrm{j}n(\theta - N\xi)]. \tag{12}$$

The right-hand side can be rewritten using the identity [14])

$$\sum_{n=-\infty}^{\infty} \exp\,(-\mathrm{j}n\theta) = 2\pi \sum_{k=-\infty}^{\infty} \delta(\theta - 2k\pi). \tag{13}$$

This expression is a Dirac pulse-train with period 2π and will frequently occur in our analysis. Therefore we introduce the following function which apart from a scale factor is equal to the "shah"-function used by Bracewell [15]):

$$\sqcup\!\sqcup\,(\theta) \overset{\Delta}{=} \sum_{k=-\infty}^{\infty} \delta(\theta - 2\pi k). \tag{14}$$

Similarly as $\delta(\omega)$ in the case of continuous-time signals, $\sqcup\!\sqcup\,(\theta)$ occurs in the spectral analysis of discrete-time signals, taking account of the periodicity of the spectra of these signals. Manipulation with this function is very similar to that with the δ function.

From (14) and (12) it follows that

$$H(\theta,\xi) = \sqcup\!\sqcup\,(\theta - N\xi). \tag{15}$$

This leads to a relation between input and output spectrum of the form

$$Y(\theta) = \frac{1}{N} \sum_{k=0}^{N-1} X\left(\frac{\theta - 2\pi k}{N}\right) \tag{16}$$

using the fact that $\sqcup\!\sqcup\,(\theta - N\xi) = 0$ if $N\xi \neq \theta - 2\pi k$. In terms of the actual frequencies the expression is

$$Y(\omega T_2) = \frac{1}{N} \sum_{k=0}^{N-1} X\left(\omega T_1 - k\,\frac{2\pi}{N}\right). \tag{17}$$

This is illustrated in fig. 1 for $N = 3$. Here, and in all subsequent examples, only the fundamental interval $-\pi \leqslant \theta \leqslant \pi$ of each of the spectra is depicted, and since $T_2 = NT_1$ the corresponding lengths of the frequency intervals are different for the two spectra. The various relations and the symbols used for

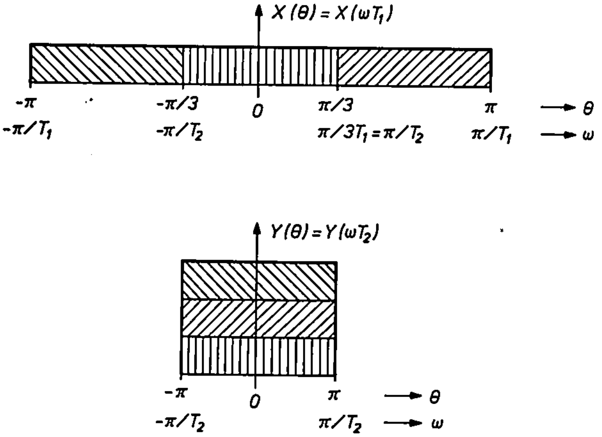
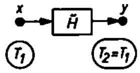
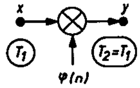
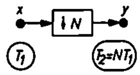
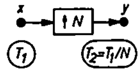


Fig. 1. Input spectrum and output spectrum of a sampling rate decrease for $N = 3$.

this SRD and for several other elements are summarized in table I, where $\tilde{H}(\theta)$ and $\Phi(\theta)$ are the Fourier transforms of $\tilde{h}(n)$ and $\varphi(n)$ respectively.

TABLE I

Time-domain and frequency-domain description of elements of linear time-varying discrete-time systems

operation	symbol	time-domain description		frequency-domain description	
		imp. resp. $h(n,m)$	input - output	transm. funct. $H(\theta,\xi)$	input - output
time-invariant		$\tilde{h}(n-m)$	$y(n)=\tilde{h}(n)*x(n)=\sum_{m=-\infty}^{\infty}\tilde{h}(n-m)\,x(m)$	$\tilde{H}(\theta)\,\text{Ш}(\theta-\xi)$	$Y(\theta)=\tilde{H}(\theta)X(\theta)$
modulation		$\varphi(n)\,u(n-m)$	$y(n)=x(n)\varphi(n)$	$\frac{1}{2\pi}\Phi(\theta-\xi)$	$Y(\theta)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi(\theta-\xi)X(\xi)d\xi$
● cosine		$\cos(n\theta_c)\,u(n-m)$	$y(n)=x(n)\cos(n\theta_c)$	$\frac{1}{2}\text{Ш}(\theta-\xi-\theta_c)+\frac{1}{2}\text{Ш}(\theta-\xi+\theta_c)$	$Y(\theta)=\frac{1}{2}X(\theta-\theta_c)+\frac{1}{2}X(\theta+\theta_c)$
● sine		$\sin(n\theta_c)\,u(n-m)$	$y(n)=x(n)\sin(n\theta_c)$	$\frac{1}{2j}\text{Ш}(\theta-\xi-\theta_c)-\frac{1}{2j}\text{Ш}(\theta-\xi+\theta_c)$	$Y(\theta)=\frac{1}{2j}X(\theta-\theta_c)-\frac{1}{2j}X(\theta+\theta_c)$
sampling rate decrease (SRD)		$u(nN-m)$	$y(n)=x(nN)$	$\text{Ш}(\theta-N\xi)$	$Y(\theta)=\frac{1}{N}\sum_{k=0}^{N-1}X(\frac{\theta-2\pi k}{N})$
sampling rate increase (SRI)		$u(n-mN)$	$y(n)=\begin{cases}x(n/N) & n=0,\pm N,\dots\\ 0 & \text{elsewhere}\end{cases}$	$\text{Ш}(N\theta-\xi)$	$Y(\theta)=X(N\theta)$

The transmission function as introduced in eq. (7) can be used in much the same way as is conventionally done for time-invariant networks. For example the transmission function H of a cascade of two systems H_1 and H_2 as shown in fig. 2 can be expressed in terms of the individual transmission functions according to

$$H(\theta,\xi)=\int_{-\pi}^{\pi}H_2(\theta,\eta)\,H_1(\eta,\xi)\,\mathrm{d}\eta. \tag{18}$$

From eq. (18) it can be seen that the order in which the transmission functions

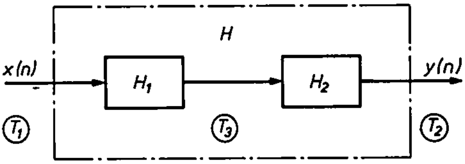


Fig. 2. Cascade of two discrete-time systems.

occur is of importance. An interchange of the two is in general not possible. Since expressions of the form (18) will frequently occur, the following shorthand notation is introduced:

$$H_2(\theta,\cdot)\bullet H_1(\cdot,\xi)\overset{\triangle}{=}\int_{-\pi}^{\pi}H_2(\theta,\eta)\,H_1(\eta,\xi)\,\mathrm{d}\eta. \tag{19}$$

Without ambiguity this notation can be extended to situations where the functions depend only on one variable. For example eq. (9) in this notation reads

$$Y(\theta)=H(\theta,\cdot)\bullet X(\cdot). \tag{9'}$$

3. Tellegen's theorem for time-varying networks

What nowadays is referred to as Tellegen's theorem [16,17]) is actually a very general network principle. The theorem is derived starting from an identity, that, as indicated by Penfield et al. [17]), may be expressed in various different forms. This degree of freedom makes the theorem very powerful since it allows its formulation to be adapted to particular classes of networks or to the problems under investigation. Fettweis [18]) has shown that to signal-flow networks the difference form of Tellegen's theorem is applicable. Also in this formulation there still remains a large amount of freedom in the precise form, a fact that can be used to advantage. In this paper we will give two formulations of the difference form of Tellegen's theorem. These forms are derived with the intent to generalize the concept of transposition to linear time-varying systems. Two different definitions of transposition will result as is discussed in sec. 4, which together with Tellegen's theorem lead to a transposition theorem applicable to arbitrary linear discrete-time signal-flow networks.

To this end let us consider such a network S having a certain topology. It consists of a set of I nodes connected by oriented branches. To each node i there corresponds a node variable w_i. Two types of signals are distinguished entering each node: x_i representing source variables and v_{ij} representing the output signal of the branch connecting node j to node i. This is illustrated in fig. 3, where output signals y_i are also indicated in the way proposed by Fettweis [18]). For each of the nodes the following equation holds

$$w_i(n)=x_i(n)+\sum_{j=1}^{I}v_{ij}(n)\qquad i=1,\ldots,I. \tag{20}$$

It should be recalled that different sampling periods are allowed in various parts of the system, and, to deal with the most general situation, sampling periods T_i will be associated with each of the nodes as indicated in fig. 3. Of course the trivial assumption has been made that all signals entering a

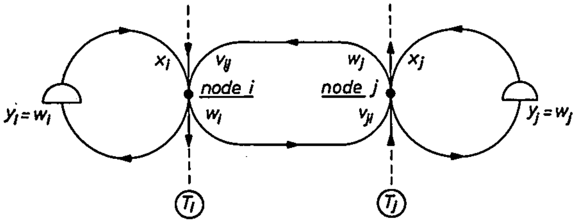


Fig. 3. Flow-graph representation of a discrete-time network. Shown are two nodes with their interconnections and sources.

specific node have the same sampling period. The Fourier transform of eq. (20) gives

$$W_i(\theta) = X_i(\theta) + \sum_{j=1}^{I} V_{ij}(\theta) \qquad i = 1, \ldots, I. \tag{21}$$

Tellegen's theorem relates variables of two different networks S and S' having the same topology. The variables in S' will be denoted by primed symbols and satisfy relations similar to (20) and (21).

The forms of Tellegen's theorem that we aim at can be derived from the following two identities.

$$\int_{-\pi}^{\pi} \sum_{i=1}^{I} [W_i(\theta)\; W_i'^*(-\theta) - W_i'^*(-\theta)\; W_i(\theta)]\, \mathrm{d}\theta = 0, \tag{22}$$

and

$$\int_{-\pi}^{\pi} \sum_{i=1}^{I} [W_i(\theta)\; W_i'(\theta) - W_i'(\theta)\; W_i(\theta)]\, \mathrm{d}\theta = 0. \tag{23}$$

First it can be remarked that if all $w_i'(n)$, the node variables in S', are real then (22) and (23) are the same, and thus differences can only be expected in networks with complex signals. Indeed in sec. 4 it will be shown that two different forms of transposition theorem result for networks with complex parameters from the two different forms of Tellegen's theorem. Secondly, a comparison with the derivation of Tellegen's theorem, as given by Fettweis, reveals that both (22) and (23) differ from Fettweis' formulation in that both expressions are integrated over a fundamental interval of θ. In this way account is taken of the fact that in a time-varying system frequency components of a signal at a certain node may be transferred to other frequencies during the transmission from one node to the other. It also makes the derivation more "symmetrical" in the sense that a similar derivation in the time domain is possible after applying Parseval's equality to (22) or (23), but this time domain form will not be given here. Only the derivation of the first form will be given explicitly since that of the second form follows the same lines.

From eq. (21) applied once to W_i and once to W_i' it follows that

$$\begin{aligned}&\int_{-\pi}^{\pi}\sum_{i=1}^{I}\sum_{j=1}^{I}[W_i(\theta)\,V_{ij}'^*(-\theta)-W_i'^*(-\theta)\,V_{ij}(\theta)]\,\mathrm{d}\theta\\&\quad+\int_{-\pi}^{\pi}\sum_{i=1}^{I}[W_i(\theta)\,X_i'^*(-\theta)-W_i'^*(-\theta)\,X_i(\theta)]\,\mathrm{d}\theta=0.\end{aligned}\tag{24}$$

This is a form of Tellegen's theorem that holds for any discrete-time network, whether linear or not. If the network is linear, then in accordance with sec. 2 impulse responses $f_{ij}(n,m)$ with corresponding transmittances $F_{ij}(\theta,\xi)$ may be associated with the branches in S such that

$$V_{ij}(\theta)=F_{ij}(\theta,\cdot)\bullet W_j(\cdot),\qquad i,j=1,\ldots,I.$$

A similar relation holds for the primed variables. Using these relations eq. (24) can be rewritten as

$$\begin{aligned}&\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{I}\sum_{j=1}^{I}W_i(\theta)\,F_{ij}'^*(-\theta,\xi)\,W_j'^*(\xi)\,\mathrm{d}\theta\,\mathrm{d}\xi\\&\quad-\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{I}\sum_{j=1}^{I}W_i'^*(-\theta)\,F_{ij}(\theta,\xi)\,W_j(\xi)\,\mathrm{d}\theta\,\mathrm{d}\xi\\&\quad+\int_{-\pi}^{\pi}\sum_{i=1}^{I}[W_i(\theta)\,X_i'^*(-\theta)-W_i'^*(-\theta)\,X_i(\theta)]\,\mathrm{d}\theta=0.\end{aligned}\tag{25}$$

Now the order of the summations in the first double sum can be reversed and the integration variables θ and ξ replaced by ξ and $-\theta$ respectively to yield the desired result. With this derivation and the analogous derivation starting from eq. (23) the following theorem has been proved.

THEOREM 1 (Tellegen's theorem). In every two linear discrete-time networks S and S' with the same topology, the spectra of the signals satisfy the relations

$$\begin{aligned}&\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{I}\sum_{j=1}^{I}W_i'^*(-\theta)\,W_j(\xi)\,[F_{ji}'^*(-\xi,-\theta)-F_{ij}(\theta,\xi)]\,\mathrm{d}\theta\,\mathrm{d}\xi\\&\quad+\int_{-\pi}^{\pi}\sum_{i=1}^{I}[W_i(\theta)\,X_i'^*(-\theta)-W_i'^*(-\theta)\,X_i(\theta)]\,\mathrm{d}\theta=0\end{aligned}\tag{26}$$

and

$$\begin{aligned}&\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{I}\sum_{j=1}^{I}W_i'(\theta)\,W_j(\xi)\,[F_{ji}'(\xi,\theta)-F_{ij}(\theta,\xi)]\,\mathrm{d}\theta\,\mathrm{d}\xi\\&\quad+\int_{-\pi}^{\pi}\sum_{i=1}^{I}[W_i(\theta)\,X_i'(\theta)-W_i'(\theta)\,X_i(\theta)]\,\mathrm{d}\theta=0.\end{aligned}\tag{27}$$

4. Transposition of linear time-varying networks

Transposition or flow-graph reversal is a well-known procedure for giving time-invariant networks a different structure while leaving the transmission between input and output unchanged [8,18]. Invariance of the transmission function for flow-graph reversal cannot be expected for time-varying systems since input and output may operate at different sampling rates. A high input rate and low output rate will become a low input rate and high output rate after flow-graph reversal and vice versa. Therefore a different definition of transposition is required. Two such definitions are suggested by the two forms of Tellegen's theorem derived in sec. 3. They will be denoted by hermitian transpose and generalized transpose respectively.

Let S be a linear discrete-time network with I nodes and branch transmittances $\{F_{ij}(\theta,\xi)\}_{i,j=1}^{I}$ where F_{ij} is the transmittance of the branch that connects node j to node i.

Definition 1. The hermitian transpose of S is a linear discrete-time network S^{H} with the same topology as S and in which node j is connected to node i by a branch with transmittance

$$F_{ij}^{\mathrm{H}}(\theta,\xi) = F_{ji}^{*}(-\xi,-\theta) \qquad i,j = 1, \ldots, I \tag{28}$$

with corresponding impulse response

$$f_{ij}^{\mathrm{H}}(n,m) = f_{ji}^{*}(-m,-n) \qquad i,j = 1, \ldots, I. \tag{29}$$

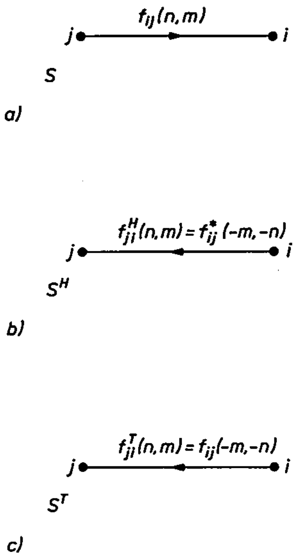
Definition 2. The generalized transpose of S is a linear discrete-time network S^{T} with the same topology as S and in which node j is connected to node i by a branch with transmittance

$$F_{ij}^{\mathrm{T}}(\theta,\xi) = F_{ji}(\xi,\theta) \qquad i,j = 1, \ldots, I \tag{30}$$

with corresponding impulse response

$$f_{ij}^{\mathrm{T}}(n,m) = f_{ji}(-m,-n) \qquad i,j = 1, \ldots, I. \tag{31}$$

It can be seen from eqs (29) and (31) that both forms of transposition preserve causality. From eq. (8) it follows that in the case of networks where f_{ij} is real for all i and j the two definitions coincide and thus $S^{\mathrm{H}} = S^{\mathrm{T}}$, but in the case of systems with complex parameters they generally differ. The hermitian transpose then has an important advantage over the generalized transpose. To see this, consider the network S of which the branch connecting node j to node i is depicted in fig. 4*a*. In the hermitian transpose network S^{H} and the generalized


$f_{ij}(n,m)$
j
i
S
a)
$f_{ji}^H(n,m)=f_{ij}^*(-m,-n)$
j
i
S^H
b)
$f_{ji}^T(n,m)=f_{ij}(-m,-n)$
j
i
S^T
c)

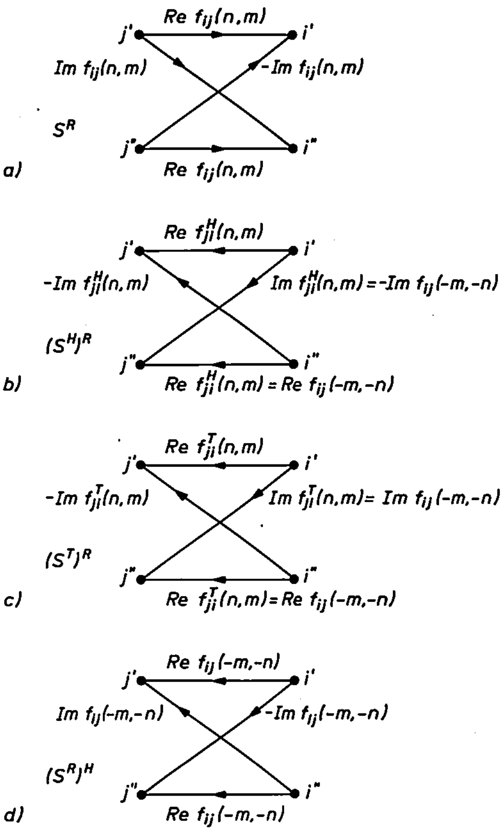

a)
$\mathrm{Re}\ f_{ij}(n,m)$
j'
i'
$\mathrm{Im}\ f_{ij}(n,m)$
$-\mathrm{Im}\ f_{ij}(n,m)$
S^R
j''
i''
$\mathrm{Re}\ f_{ij}(n,m)$
b)
$\mathrm{Re}\ f_{ji}^H(n,m)$
j'
i'
$-\mathrm{Im}\ f_{ji}^H(n,m)$
$\mathrm{Im}\ f_{ji}^H(n,m)=-\mathrm{Im}\ f_{ij}(-m,-n)$
$(S^H)^R$
j''
i''
$\mathrm{Re}\ f_{ji}^H(n,m)=\mathrm{Re}\ f_{ij}(-m,-n)$
c)
$\mathrm{Re}\ f_{ji}^T(n,m)$
j'
i'
$-\mathrm{Im}\ f_{ji}^T(n,m)$
$\mathrm{Im}\ f_{ji}^T(n,m)=\mathrm{Im}\ f_{ij}(-m,-n)$
$(S^T)^R$
j''
i''
$\mathrm{Re}\ f_{ji}^T(n,m)=\mathrm{Re}\ f_{ij}(-m,-n)$
d)
$\mathrm{Re}\ f_{ij}(-m,-n)$
j'
i'
$\mathrm{Im}\ f_{ij}(-m,-n)$
$-\mathrm{Im}\ f_{ij}(-m,-n)$
$(S^R)^H$
j''
i''
$\mathrm{Re}\ f_{ij}(-m,-n)$

Fig. 4

Fig. 5

Fig. 4. Flow-graph representation of connections between two nodes in various networks; (*a*) original, (*b*) hermitian transpose, (*c*) generalized transpose.

Fig. 5. Flow-graph of networks implementing the complex transmittances of fig. 4; (*a*) original, (*b*) hermitian transpose, (*c*) generalized transpose, (*d*) transpose of the network of fig. 5*a*.

transpose network S^T node i is connected to node j as indicated in figs 4*b* and 4*c* respectively. A practical realization of S will be a system S^R such that to every node i in S there correspond two nodes i' and i'' in S^R with node variables

$$w_{i'}^{R}(n) = \mathrm{Re}\ w_i(n), \tag{32}$$

$$w_{i''}^{R}(n) = \mathrm{Im}\ w_i(n). \tag{33}$$

The transmissions from nodes j' and j'' to nodes i' and i'' are characterized by the impulse responses

$$f_{i'j'}{}^{\mathrm{R}}(n,m) = \quad \mathrm{Re}\, f_{ij}(n,m), \tag{34}$$

$$f_{i''j'}{}^{\mathrm{R}}(n,m) = \quad \mathrm{Im}\, f_{ij}(n,m), \tag{35}$$

$$f_{i'j''}{}^{\mathrm{R}}(n,m) = -\mathrm{Im}\, f_{ij}(n,m), \tag{36}$$

$$f_{i''j''}{}^{\mathrm{R}}(n,m) = \quad \mathrm{Re}\, f_{ij}(n,m), \tag{37}$$

as shown in fig. 5*a*. The networks S^{H} and S^{T} will similarly have associated with them a network with real parameters that implement the complex transmissions as shown in figs 5*b* and 5*c*. These networks will be indicated by $(S^{\mathrm{H}})^{\mathrm{R}}$ and $(S^{\mathrm{T}})^{\mathrm{R}}$ respectively. Since the network S^{R} is also a linear discrete-time network, transposition can be applied to it. Clearly, since S^{R} has only real parameters, its hermitian and generalized transposes will be the same and can be indicated by $(S^{\mathrm{R}})^{\mathrm{H}}$ or $(S^{\mathrm{R}})^{\mathrm{T}}$ as desired. The nodes i', i'' and j', j'' of this network and the corresponding connections are shown in fig. 5*d*. It can be seen by comparison that

$$(S^{\mathrm{R}})^{\mathrm{H}} = (S^{\mathrm{H}})^{\mathrm{R}} \tag{38}$$

but

$$(S^{\mathrm{R}})^{\mathrm{T}} \neq (S^{\mathrm{T}})^{\mathrm{R}}, \tag{39}$$

which means that different systems result when generalized transposition is applied to a network with complex parameters, or to its practical realization.

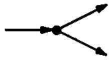
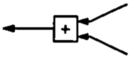
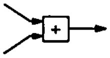
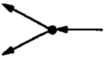
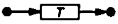
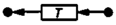
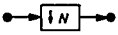
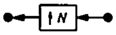
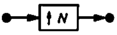
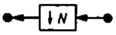
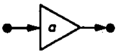
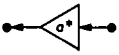
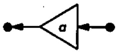
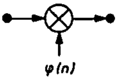
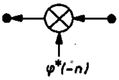
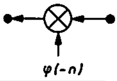
Applying hermitian or generalized transposition to a discrete-time system first of all implies flow-graph reversal, as indicated by the reversal of indices in eqs (28) to (31). This means that branch points in the system become summation points and vice versa. Besides this flow reversal the elements must be replaced by elements having the specified transmittance. From inspection of table I it can be seen, for example, that an SRD must be replaced by an SRI and vice versa. Table II summarizes the necessary replacements. From the definition of S^{H} and S^{T} it follows that $(S^{\mathrm{H}})^{\mathrm{H}} = (S^{\mathrm{T}})^{\mathrm{T}} = S$, which means that if an element must be replaced by an other upon transposition then after transposition this latter element must be replaced by the first one, so that two such elements are always mutually transposed.

The usefulness of the definitions given above becomes clear when we apply Tellegen's theorem to the network and its transpose. Applying the first form (eq. (26)) to S and S^{H} we see that the double sum vanishes and thus

$$\int_{-\pi}^{\pi} \sum_{i=1}^{I} [W_i(\theta)\, X_i^{\mathrm{H}*}(-\theta) - W_i^{\mathrm{H}*}(-\theta)\, X_i(\theta)]\, \mathrm{d}\theta = 0. \tag{40}$$

TABLE II

Elements in the original discrete-time system S and the corresponding elements in the transposed systems S^H and S^T

original	hermitian transpose	generalized transpose
		
		
		
		
		
		
		

A similar form holds for S and S^T. Equation (40) is a generalization of the interreciprocity relation as defined in ref. 18.

Now assume that S is excited by a single input $x_a(n)$ incident on node a, and consider as output the signal $y_b(n) = w_b(n)$ on node b. The transmission from input to output is characterized by a transmission function $H_{ba}(\theta,\xi)$ such that according to eq. (9′)

$$Y_b(\theta) = H_{ba}(\theta, \cdot) \bullet X_a(\cdot). \qquad (41)$$

Due to the reversal of the signal flow the hermitian transpose system S^H will have a single input $x_b{}^H(n)$ incident on node b and output $y_a{}^H(n) = w_a{}^H(n)$. The spectra of these signals are related by

$$Y_a{}^H(\theta) = H_{ab}{}^H(\theta, \cdot) \bullet X_b{}^H(\cdot). \qquad (42)$$

Inserting eqs (41) and (42) in the interreciprocity relation (eq. (40)) and observing that $X_i(\theta) = 0, i \neq a$, and $X_i{}^H(\theta) = 0$, $i \neq b$, the first part of the transposition theorem follows immediately. The second part results from a similar reasoning but applied to S and S^T.

THEOREM 2 (Transposition theorem). If a linear discrete-time system S realizes a transmission function $H_{ba}(\theta,\xi)$ between input node a and output node b then

(1) its hermitian transpose system S^{H} realizes a transmission function $H_{ab}{}^{\mathrm{H}}(\theta,\xi)$ between input node b and output node a given by

$$H_{ab}{}^{\mathrm{H}}(\theta,\xi) = H_{ba}{}^{*}(-\xi,-\theta), \tag{43a}$$

(2) its generalized transpose system S^{T} realizes a transmission function $H_{ab}{}^{\mathrm{T}}(\theta,\xi)$ between input node b and output node a given by

$$H_{ab}{}^{\mathrm{T}}(\theta,\xi) = H_{ba}(\xi,\theta). \tag{43b}$$

The implications of these properties will be clarified by means of some examples in sec. 5, but it may be noted here that in general the transmission function will not be invariant upon either type of transposition. Therefore transposition does not merely provide an alternative implementation of a certain transmission function, as in the time-invariant case, but rather it yields an implementation of a system that performs a complementary operation.

A property of both types of transposition is that it changes neither the number of multipliers nor the rate at which these multipliers operate. This observation leads to the following corollary, which clearly shows the impact of these forms of transposition on a hardware implementation.

Corollary. If a linear discrete-time system S that realizes the transmission function $H(\theta,\xi)$ is optimized with respect to multiplication rate, then

(1) S^{H} is an optimal realization with respect to multiplication rate of the transmission function

$$H^{\mathrm{H}}(\theta,\xi) = H^{*}(-\xi,-\theta).$$

(2) S^{T} is an optimal realization with respect to multiplication rate of the transmission function

$$H^{\mathrm{T}}(\theta,\xi) = H(\xi,\theta).$$

5. Applications

To clarify the concepts of transposition introduced in sec. 4 a number of examples will be given. We start with an implementation of a decimation-in-time Fast Fourier Transform (FFT) algorithm, of which an 8-point version is shown in fig. 6. Since this system is time-invariant the generalized transposition

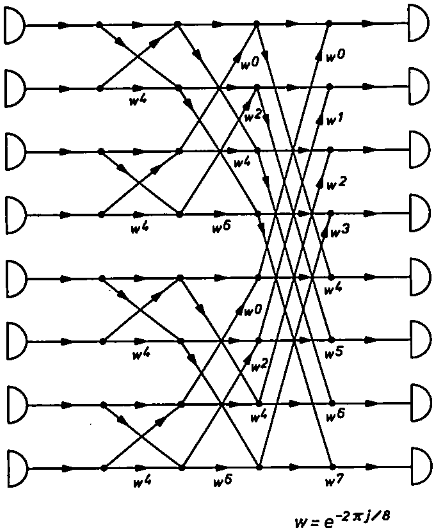


Fig. 6. Flow-graph of a decimation-in-time FFT algorithm for 8 points.

coincides with the conventional transposition and, as is well known, leads to a decimation-in-frequency FFT algorithm (ref. 8, sec. 6.3.2). Since the implementation of the FFT in fig. 6 has complex parameters, the hermitian transpose will be different from the generalized transpose. From the transposition theorem it follows that the hermitian transpose implements the inverse discrete Fourier transform with a decimation-in-frequency FFT algorithm.

The system in the next example has real parameters. Without ambiguity we then use the term transposition. Figure 7*a* depicts an implementation of a decimator [3]) derived from a FIR filter with linear phase and sampling rate decrease. Use is made of the symmetry of the impulse response to reduce the multiplication rate. The transmission function of the system is

$$H_{ba}(\theta,\xi) = \tilde{H}(\xi) \sqcup\!\sqcup (\theta - N\xi), \tag{44}$$

where

$$\tilde{H}(\xi) = \left(h_0 + 2 \sum_{k=1}^{M} h_k \cos k\xi\right) \exp(-\mathrm{j}M\xi).$$

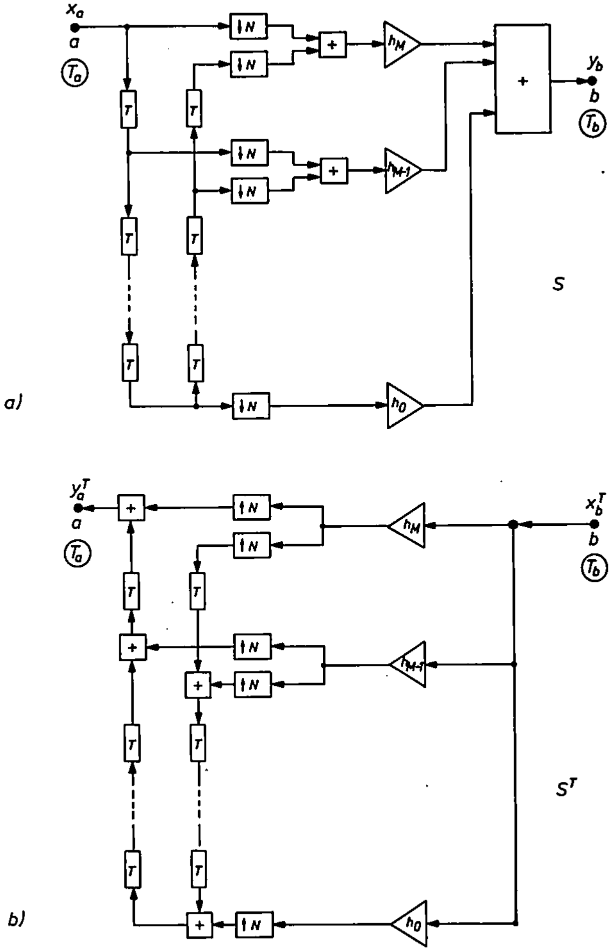


Fig. 7. (*a*) Implementation of a linear phase FIR filter for sampling rate reduction (decimator). (*b*) Transpose of the decimator. This structure realizes an interpolator.

If $X_a(\theta)$ is the input spectrum the output spectrum equals

$$Y_b(\theta) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{H}\left(\frac{\theta - 2\pi k}{N}\right) X_a\left(\frac{\theta - 2\pi k}{N}\right). \tag{45}$$

In such a decimator $\tilde{H}$ has a low-pass characteristic with cut-off frequency at $\omega = \pi/T_b$. A schematic representation of $\tilde{H}(\theta)$ and the spectra is given in figs 8*a*, *b* and *c*. The transpose of this decimator is shown in fig. 7*b* and according to the transposition theorem it has the transmission function

$$H_{ab}{}^{\mathrm{T}}(\theta,\xi) = \tilde{H}(\theta) ⧢ (\xi - N\theta). \tag{46}$$

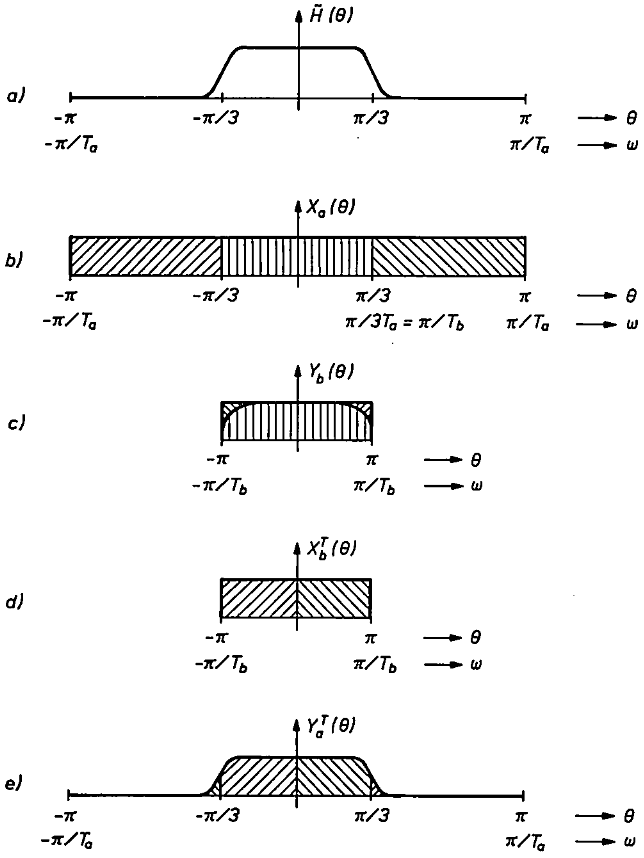


Fig. 8. Spectra of various signals of systems in figs 7*a* and *b*; (*a*) characteristic of the function $\tilde{H}$, (*b*) input spectrum of decimator, (*c*) output spectrum of decimator, (*d*) input spectrum of interpolator, (*e*) output spectrum of interpolator.

Excitation of this transpose system with an input signal with spectrum $X_b^{\mathrm{T}}(\theta)$ gives the output spectrum

$$Y_a^{\mathrm{T}}(\theta) = \tilde{H}(\theta)\, X_b^{\mathrm{T}}(N\theta) \qquad (47)$$

as shown in figs 8*d* and 8*e*. It can be concluded that this transpose system is an interpolator [2–5]. In fact this system is a particular implementation of an interpolating FIR filter that was previously proposed by Bellanger and Bonnerot [19]. It is important to note that use is again made of the symmetry of the impulse response of the FIR filter to reduce the number of multiplications, in contrast to conventional implementations of an interpolator [3–6]. This is an immediate consequence of the corollary in sec. 4.

The following example concerns a Weaver single-sideband modulator shown in fig. 9*a*. In this system $\tilde{H}$ is a low-pass filter with cut-off frequency at $\pi/2T_a$,

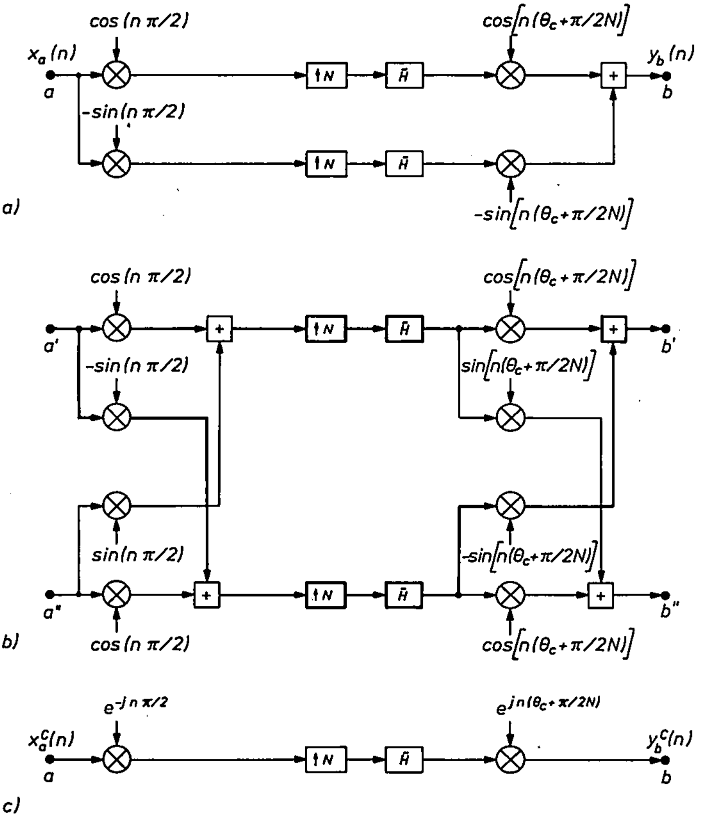


Fig. 9. (*a*) Weaver single-sideband modulator. (*b*) Modulator of fig. 9*a* extended with additional input and output. (*c*) Complex representation of the modulator of fig. 9*b*.

as shown in fig. 10*a*. For the input spectrum in fig. 10*b* the output spectrum of fig. 10*c* results. Extending this modulator as shown in fig. 9*b* we obtain a real implementation of the complex system shown in fig. 9*c*. Of course for real input signals the real part of the output signal of the complex modulator is the same as the output of the modulator in fig. 9*a*. The transmission function of the complex modulator equals

$$H_{ba}{}^{C}(\theta,\xi) = \tilde{H}(\theta - \theta_c - \pi/2N)\ \text{Ш}\ [\xi - N(\theta - \theta_c)] \tag{48}$$

and a sketch of the output spectrum $Y_b{}^{C}(\theta)$ is given in fig. 10*d*, assuming the input spectrum of fig. 10*b*. Due to the form of the output spectrum the complex system may be called an upper sideband modulator. The hermitian transpose of this system is obtained by flow-graph reversal and changing the elements as prescribed by table II. In this case all elements remain the same except for the SRI, which is replaced by an SRD. The transmission function of the hermitian transpose system is

$$H_{ab}{}^{CH}(\theta,\xi) = \tilde{H}(\xi + \theta_c + \pi/2N)\ \text{Ш}\ [\theta - N(\xi + \theta_c)]. \tag{49}$$

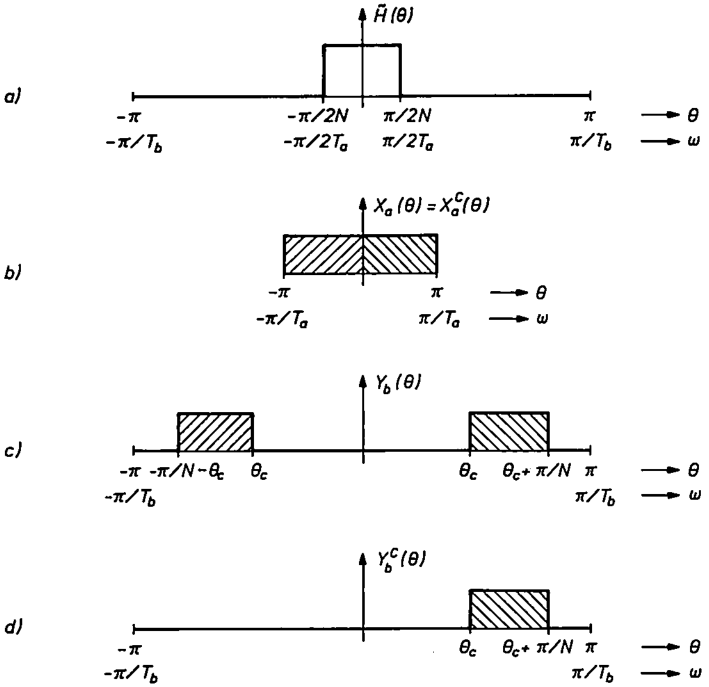


Fig. 10. (*a*) Transmission function of filter $\tilde{H}$ in fig. 9. (*b*) Input spectrum of the modulators in fig. 9. (*c*) Output spectrum of the Weaver modulator of fig. 9*a*. (*d*) Output spectrum of the complex modulator of fig. 9*c*.

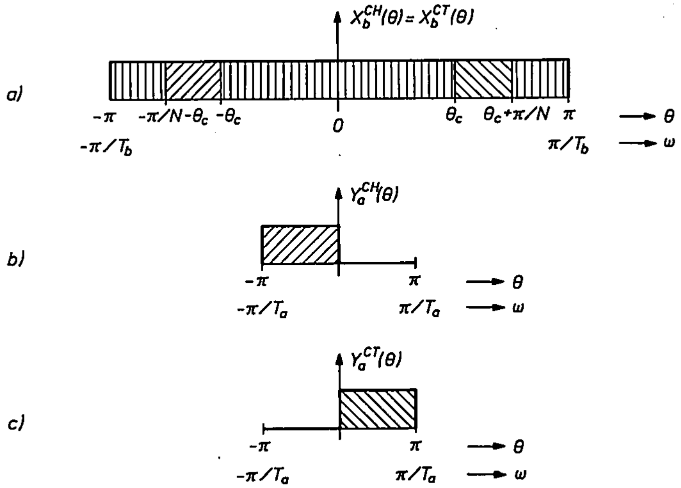


Fig. 11. (*a*) Input spectrum of the transposes of the modulators of fig. 9. (*b*) Output spectrum of the hermitian transpose of the modulator of fig. 9*c*. (*c*) Output spectrum of the generalized transpose of the modulator of fig. 9*c*.

For the input spectrum $X_b^{CH}(\theta)$ of fig. 11*a* the output spectrum is sketched in fig. 11*b*, and the hermitian transpose can be seen to be a lower sideband demodulator. In accordance with the discussion in sec. 4, the transpose of the modulator of fig. 9*b* will be a real implementation of the hermitian transpose of the complex modulator, and when only real parts are considered the transpose of the Weaver modulator of fig. 9*a* results and is a single-sideband demodulator. If we apply generalized transposition to the complex modulator of fig. 9*c*, then not only must we replace the SRI by an SRD but we must also replace *n* by—*n* in the modulation function. This yields the transmission function

$$H_{ab}^{CT}(\theta,\xi) = \tilde{H}(\xi - \theta_c - \pi/2N) \sqcup\!\sqcup\, [\theta - N(\xi - \theta_c)]. \qquad (50)$$

The output spectrum of this system is shown in fig. 11*c* and it can be concluded that the generalized transpose of an upper sideband modulator is an upper sideband demodulator.

As a final example the TDM–FDM translator discussed in ref. 7 can be mentioned, which in its most general form contains complex-valued signal processing operations. The structure of the corresponding FDM–TDM translator is found by applying hermitian transposition to the TDM–FDM system. Generalized transposition too will yield an FDM–TDM translator but, for a complex system, will have a slightly different implementation.

6. Sensitivity analysis

An important application of Tellegen's theorem is the derivation of formulae for the sensitivity of transmission functions to changes of system parameters [8,18]. Such formulae are important for determining the influence of parameter quantization [8]) on the system characteristics. Moreover, as indicated by Jackson [20]) there is a close relation between the coefficient sensitivity of a network and its roundoff noise resulting from signal quantization.

Completely analogous to the method given by Fettweis, a sensitivity formula for time-varying discrete-systems can be derived from any of the two forms of Tellegen's theorem as stated in sec. 3. Here only the result of the derivation will be given.

Consider a system S as depicted in fig. 12*a*. The input and output nodes are labeled a and b respectively, and the transmission function between these nodes is $H_{ba}(\theta,\xi)$. We assume a transmittance $F_{ji}(\theta,\xi)$ in the branch connecting node i to node j, and want to determine the influence of changes of F_{ji} on H_{ba}. To this end we introduce a system S' which is identical to S except for the branch that connects node i to node j, which has a transmittance $F_{ji}(\theta,\xi) + \Delta F_{ji}(\theta,\xi)$. Denoting the transmittances in S' by primed variables, the following expression for the variation of H_{ba} can be derived:

$$\begin{aligned}\Delta H_{ba}(\theta,\xi) &\overset{\Delta}{=} H_{ba}'(\theta,\xi) - H_{ba}(\theta,\xi) \\ &= H_{bj}(\theta,\cdot) \bullet \Delta F_{ji}(\cdot,\cdot) \bullet H_{ia}'(\cdot,\xi),\end{aligned} \tag{51}$$

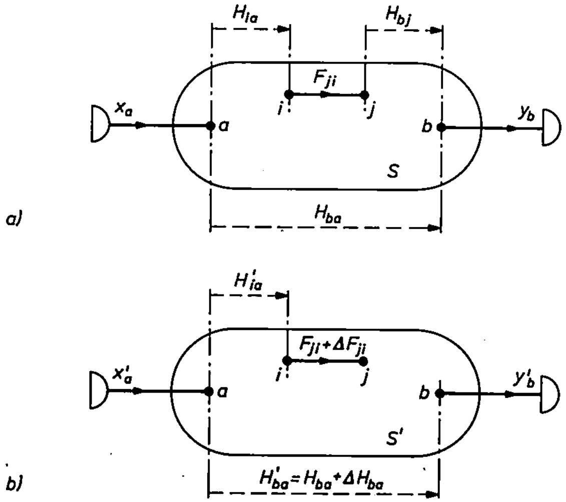


Fig. 12. (*a*) Discrete-time system S realizing the transmission function H_{ba}. (*b*) Discrete-time system S' obtained from S by perturbing the transmittance F_{ji} by ΔF_{ji}.

where the notation introduced in sec. 2 is used. Equation (51) describes a cascade of three subsystems as shown in fig. 13*a*. Since H_{ia}' too is a transmission function of the perturbed system S', we may likewise write

$$H_{ia}'(\theta,\xi) = H_{ia}(\theta,\xi) + \Delta H_{ia}(\theta,\xi), \tag{52}$$

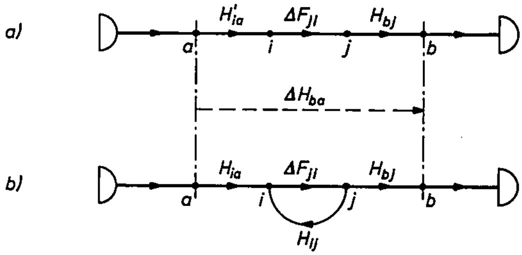


Fig. 13. (*a*) Flow-graph of the cascade that realizes the variation ΔH_{ba} of the transmission function H_{ba} due to a change ΔF_{ji} in F_{ji} in the system S. (*b*) Alternative network that realizes ΔH_{ba} and only contains transmission functions of S.

where $\Delta H_{ia}(\theta,\xi)$ is given by (51) but with index b replaced by i. It therefore follows that

$$H_{ia}'(\theta,\xi) = H_{ia}(\theta,\xi) + H_{ij}(\theta,\cdot\,) \bullet \Delta F_{ji}(\,\cdot\,,\cdot\,) \bullet H_{ia}'(\,\cdot\,,\xi) \tag{53}$$

and thus $H_{ia}'(\theta,\xi)$ is the solution of this integral equation. The interpretation is that $H_{ia}'(\theta,\xi)$ may be constructed by a cascade of H_{ia} and a feedback loop with transmission function ΔF_{ji} and H_{ij} as shown in fig. 13*b*. Apart from ΔF_{ji} this latter system only contains transmission functions of the unperturbed system S. If each of the transmission functions in eq. (53) describes a time-invariant system, then the cascade operation becomes a simple multiplication and the well-known relation for large scale variations [8,21]) follows immediately *). Such an explicit relation does not exist in the general time-varying case, but repeated substitution of (53) into (51) leads to the Neumann series

$$\begin{aligned}&\Delta H_{ba}(\theta,\xi)\\ &\quad= H_{bj}(\theta,\cdot\,)\bullet [\Delta F_{ji}(\,\cdot\,,\cdot\,) + \Delta F_{ji}(\,\cdot\,,\cdot\,)\bullet H_{ij}(\,\cdot\,,\cdot\,)\bullet \Delta F_{ji}(\,\cdot\,,\cdot\,) + \ldots]\bullet H_{ia}(\,\cdot\,,\xi).\end{aligned} \tag{54}$$

The terms in the brackets are related to network sensitivities of increasing order [21]). In particular the first term

$$H_{bj}(\theta,\cdot\,) \bullet \Delta F_{ji}(\,\cdot\,,\cdot\,) \bullet H_{ia}(\,\cdot\,,\xi) \tag{55}$$

*) Such a simplification is also possible if both ΔF_{ji} and H_{ij} are transmission functions corresponding to time-invariant impulse responses.

gives the first-order variation of H_{ba} to changes in the transmittance F_{ji}. For the specific case that only a constant multiplier with coefficient λ_{ji} connects node i to node j this expression yields the sensitivity

$$\frac{\partial H_{ba}(\theta,\xi)}{\partial \lambda_{ji}} = H_{bj}(\theta,\cdot\,) \bullet H_{ia}(\,\cdot\,,\xi) \tag{56}$$

which very much resembles the familiar relation for time-invariant systems [8,18].

Finally, in eq. (56) the transmission function $H_{bj}(\theta,\xi)$ may be replaced by $H_{jb}{}^{\mathrm{T}}(\xi,\theta)$, which is the transmission from node b to node j in the generalized transpose system S^{T}. With this modification the sensitivity of H_{ba} with respect to all network coefficients can be determined by analysing once the original network (to obtain all $H_{ia}(\theta,\xi)$) and once its transpose (to obtain all $H_{jb}{}^{\mathrm{T}}(\xi,\theta)$) [18].

7. Extension to systems with continuous-time and discrete-time signals

The foregoing discussion can easily be extended to networks that contain both continuous-time and discrete-time signals. In such systems we must allow elements with analogue inputs and digital outputs or vice versa. Such elements have impulse responses of the form $h_{\mathrm{da}}(n,\tau)$ and $h_{\mathrm{ad}}(t,m)$ with corresponding input–output relations

$$y_{\mathrm{d}}(n) = \int_{-\infty}^{\infty} h_{\mathrm{da}}(n,\tau)\, x_{\mathrm{a}}(\tau)\, \mathrm{d}\tau \tag{57}$$

and

$$y_{\mathrm{a}}(t) = \sum_{m=-\infty}^{\infty} h_{\mathrm{ad}}(t,m)\, x_{\mathrm{d}}(m) \tag{58}$$

respectively. Natural candidates for such elements are the ideal A/D converter (in which amplitude quantization effects are disregarded) and the idealized D/A converter (that produces weighted δ functions), which have the impulse responses $\delta(nT-\tau)$ and $\delta(t-mT)$ respectively, where T is the sampling period of the devices. The transmission functions of these hybrid elements are now defined by

$$H_{\mathrm{da}}(\theta,\Omega) \overset{\triangle}{=} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} h_{\mathrm{da}}(n,\tau) \exp\left[-\mathrm{j}(n\theta - \Omega\tau)\right] \mathrm{d}\tau \tag{59}$$

for analogue input and digital output, and

$$H_{\mathrm{ad}}(\omega,\xi) \overset{\triangle}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_{\mathrm{ad}}(t,m) \exp\left[-\mathrm{j}(\omega t - m\xi)\right] \mathrm{d}t \tag{60}$$

for digital input and analogue output. Applying these definitions we find that the transmission functions of the A/D converter and the D/A converter are given by $\sqcup\!\sqcup\,(\theta - \Omega T)$ and $\sqcup\!\sqcup\,(\omega T - \xi)$ respectively. The corresponding spectral relations can be obtained from

$$Y_{\mathrm{d}}(\theta) = \int_{-\infty}^{\infty} H_{\mathrm{da}}(\theta,\Omega)\, X_{\mathrm{a}}(\Omega)\, \mathrm{d}\Omega \tag{61}$$

and

$$Y_{\mathrm{a}}(\omega) = \int_{-\pi}^{\pi} H_{\mathrm{ad}}(\omega,\xi)\, X_{\mathrm{d}}(\xi)\, \mathrm{d}\xi. \tag{62}$$

With the transmission functions thus defined it is easy to modify Tellegen's theorem and the transposition theorem in such a way that hybrid systems containing both continuous-time and discrete-time subsystems can also be dealt with. It then follows that the A/D converter and the D/A converter are mutually transposed.

8. Conclusions

Two forms of Tellegen's theorem have been derived that are applicable to any pair of linear discrete-time networks with the same topology. Both forms of this theorem suggested a definition of transposition, which were called hermitian transposition and generalized transposition respectively, thus generalizing the transposition concept that hitherto only applied to time-invariant systems.

Next a transposition theorem was given that relates the transmission function of a linear network to that of its hermitian or generalized transpose. In contrast to the time-invariant case where the transpose has the same transmission function as the original network, the transposes in the time-varying case realize functions that in a sense are complementary to that of the original network. As examples it was shown that the hermitian transpose of an FFT implementation performs the inverse transform, transposition of a decimator yields an interpolator and a modulator became a demodulator after transposition. Therefore transposition offers a simple and effective way to derive implementations of systems that realize such a complementary operation once the implementation of the original operation is known. In particular when the original system has been optimized with respect to multiplication rate the transpose will automatically be optimal in this sense too.

Sensitivity formulae were derived that make it possible to compute the influence of small- and large-scale changes of network elements on the transmission function of a system. Finally an extension was discussed to networks that consist of both continuous-time and discrete-time parts.

Acknowledgement

We are grateful to M. Bellanger and G. Bonnerot for kindly sending us a preprint of their paper [19]).

Philips Research Laboratories *Eindhoven, February 1978*

REFERENCES

[1]) R. W. Schafer and L. R. Rabiner, Proc. IEEE **61**, 692-702, 1973.

[2]) M. G. Bellanger, J. L. Daguet and G. Lepagnol, IEEE Trans. **ASSP-22**, 231-235, 1974.

[3]) R. E. Crochiere and L. R. Rabiner, IEEE Trans. **ASSP-23**, 444-456, 1975.

[4]) L. R. Rabiner and R. E. Crochiere, IEEE Trans. **ASSP-23**, 457-464, 1975.

[5]) R. E. Crochiere and L. R. Rabiner, IEEE Trans. **ASSP-24**, 296-311, 1976.

[6]) M. G. Bellanger, IEEE Trans. **ASSP-25**, 344-346, 1977.

[7]) T. A. C. M. Claasen and W. F. G. Mecklenbräuker, IEEE Trans. **CAS-25**, No. 5, 1978.

[8]) A. V. Oppenheim and R. W. Schafer, Digital signal processing, Prentice Hall, Englewood Cliffs, 1975.

[9]) H. Kwakernaak and R. Sivon, Linear optimal control systems, Wiley-Interscience, New York, 1972, p. 465.

[10]) P. M. Morse and H. Feshbach, Methods of theoretical physics, McGraw-Hill, 1953.

[11]) L. A. Zadeh, Proc. IRE **49**, 1488-1503, 1961.

[12]) H. D'Angelo, Linear time-varying systems: Analysis and synthesis, Allyn and Bacon, Boston, 1970, ch. 9.

[13]) L. A. Zadeh, Proc. IRE **32**, 291-299, 1950.

[14]) A. Papoulis, The Fourier integral and its applications, McGraw-Hill, N.Y., 1962, p. 44.

[15]) R. Bracewell, The Fourier transform and its applications, McGraw-Hill, N.Y., 1965, p. 77.

[16]) B. D. H. Tellegen, Philips Res. Repts **7**, 259-269, 1952.

[17]) P. Penfield, Jr, R. Spence and S. Duinker, Tellegen's theorem and electrical networks, The MIT Press, Cambridge, Mass., 1970.

[18]) A. Fettweis, Arch. El. Übertr. **25**, 557-561, 1971.

[19]) M. G. Bellanger and G. Bonnerot, IEEE Trans. **ASSP-26**, 50-55, 1978.

[20]) L. B. Jackson, IEEE Trans. **CAS-23**, 481-485, 1976.

[21]) R. E. Crochiere and A. V. Oppenheim, Proc. IEEE **63**, 581-595, 1975.

N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Philips J. Res., Vol. 33, 103-112, 1978 *R 982*

FIRST-ORDER CORRECTION TO "WEAK-GUIDANCE" APPROXIMATION IN FIBRE OPTICS THEORY

by D. L. A. TJADEN

Abstract

The paper considers mode propagation in a straight optical fibre of arbitrary cross-section index profile, in the limit of small index contrast. A perturbation analysis is applied to the vector wave equation which governs the problem, with the zero-order limit corresponding to the generally assumed scalar approximation. Simple expressions are obtained for the first-order term in an asymptotic power series expansion of the mode propagation constants, in terms of the index contrast parameter. The effect is discussed of the various possible symmetries of the index profile.

1. Introduction

The theoretical analysis of mode propagation in optical fibres is much simplified by application of the "weak-guidance" approximation introduced by Snyder [1]) and by Gloge [2]) which was later recognized by Arnaud [3]) to be equivalent to the "scalar" or "Helmholtz" approximation of wave optics. Inherent in this approximation, which uses the assumption that spatial variations of the refractive index are small, is a seeming degeneracy of the modal solutions. This degeneracy is twofold in the general case but may be of a more complicated nature due to symmetries of the configuration. Owing to the fact that in most practical cases the index difference between core and cladding is about 1 %, the results thus obtained are sufficiently accurate for many purposes. In certain situations, however, one is particularly interested in the differences between propagation constants or group velocities of the modes which constitute a degenerate group in the limit of the scalar approximation. We give an example of such a problem in a subsequent paper [4]). In these cases a more accurate approach is needed, avoiding, however, the substantial efforts which are generally required by an exact solution of the electromagnetic problem.

In this paper we shall derive expressions for the first-order term in an asymptotic power series expansion of the mode propagation constants, in terms of the index contrast parameter. This is done for the general case of an arbitrary (though two-dimensional) index profile, by applying a perturbation method to the two-dimensional vector wave equation which governs the problem. Next, the effect is discussed of the various possible symmetries of the index profile.

2. Basic concepts

We assume an isotropic, linear, nonconducting medium in which we introduce dimensionless Cartesian coordinates (x, y, z) which are normalized by some characteristic length a of the configuration. The magnetic permeability is assumed to have its vacuum value μ_0, whereas the electric permittivity ε is a function of x and y only. We consider a wave travelling in the positive z direction with propagation constant β and angular frequency ω. The complex electromagnetic field vectors thus have the form

$$\begin{aligned} \boldsymbol{E}(x, y, z, t) &= \boldsymbol{e}(x, y) \exp(-\mathrm{i}\omega t + \mathrm{i}a\beta z), \\ \boldsymbol{H}(x, y, z, t) &= \boldsymbol{h}(x, y) \exp(-\mathrm{i}\omega t + \mathrm{i}a\beta z), \end{aligned} \tag{1}$$

for the electric field and the magnetic field respectively. We introduce unit vectors $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ along the x, y, and z axis respectively and put

$$\nabla_t \equiv \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y}, \qquad \boldsymbol{e}_t \equiv \mathbf{i}\, e_x + \mathbf{j}\, e_y, \qquad \text{and} \qquad \boldsymbol{h}_t \equiv \mathbf{i}\, h_x + \mathbf{j}\, h_y.$$

According to Maxwell's equations we then have

$$\begin{aligned} \nabla_t \times \boldsymbol{e}_t &= \mathrm{i}\omega\mu_0 a\, \mathbf{k}\, h_z, \qquad \nabla_t \times \boldsymbol{h}_t = -\,\mathrm{i}\omega\varepsilon a\, \mathbf{k}\, e_z, \\ -\,\mathbf{k}\times\nabla_t\, e_z + \mathrm{i}a\beta\, \mathbf{k}\times\boldsymbol{e}_t &= \mathrm{i}\omega\mu_0 a\, \boldsymbol{h}_t, \\ -\,\mathbf{k}\times\nabla_t\, h_z + \mathrm{i}a\beta\, \mathbf{k}\times\boldsymbol{h}_t &= -\,\mathrm{i}\omega\varepsilon a\, \boldsymbol{e}_t, \end{aligned} \tag{2}$$

from which it is easily derived that $\boldsymbol{e}_t$ should satisfy the vector wave equation

$$\nabla_t^{\,2}\, \boldsymbol{e}_t + a^2(\omega^2\mu_0\,\varepsilon - \beta^2)\, \boldsymbol{e}_t + \nabla_t\left(\frac{\nabla_t\,\varepsilon}{\varepsilon}\cdot \boldsymbol{e}_t\right) = \mathbf{0}. \tag{3}$$

Here, as usual, $\nabla_t^{\,2} \equiv \nabla_t\,\nabla_t\,\cdot - \nabla_t\times\nabla_t\times$.

The other field components follow from $\boldsymbol{e}_t$ by

$$e_z = -\frac{1}{\mathrm{i}a\beta\varepsilon}\,\nabla_t\cdot(\varepsilon\boldsymbol{e}_t), \tag{4}$$

$$\boldsymbol{h}_t = \frac{\beta}{\omega\mu_0}\,\mathbf{k}\times\left\{\boldsymbol{e}_t - \frac{1}{a^2\,\beta^2}\,\nabla_t\left[\frac{1}{\varepsilon}\,\nabla_t\cdot(\varepsilon\boldsymbol{e}_t)\right]\right\}, \tag{5}$$

and

$$h_z = \frac{1}{\mathrm{i}a\omega\mu_0}\,\mathbf{k}\cdot(\nabla_t\times\boldsymbol{e}_t). \tag{6}$$

We now assume that $\varepsilon(x, y)$ can be written in the form

$$\varepsilon(x, y) = \varepsilon_1 [1 + \delta f(x, y)]. \tag{7}$$

in which $\delta \ll 1$, $\max\limits_{x,y} \{f(x, y)\} = 1$ and $f(x, y) \to 0$ if $r \equiv (x^2 + y^2)^{\frac{1}{2}} \to \infty$.

In accordance with the usual notation for fibres with circular symmetry and core radius a we define the normalized frequency parameter v by

$$v = a\omega\, (\mu_0\, \varepsilon_1\, \delta)^{\frac{1}{2}}, \tag{8}$$

and furthermore we put

$$\lambda = a^2\, (\beta^2 - \omega^2\, \mu_0\, \varepsilon_1). \tag{9}$$

Note that $\lambda = w^2$ in the usual notation. Unlike most authors we define the relative contrast parameter δ with reference to the cladding permittivity. From (3) and (7) to (9) the equation for e_t becomes

$$(\nabla_t{}^2 + v^2 f)\, e_t + \delta \nabla_t \left(\frac{\nabla_t f}{1 + \delta f} \cdot e_t\right) = \lambda e_t. \tag{10}$$

Bound modes, if they exist, correspond to regular solutions of (10), vanishing at infinity and belonging to real positive eigenvalues λ.

Now we treat δ as a variable (at v fixed) and consider the asymptotic behaviour of such solutions for $\delta \to 0$ by proposing the "Ansatz"

$$\begin{aligned} e_t &\sim e_0 + \delta e_1 + \delta^2\, e_2 + \dots \quad , \\ \lambda &\sim \lambda_0 + \delta \lambda_1 + \delta^2\, \lambda_2 + \dots \quad . \end{aligned} \tag{11}$$

Upon insertion of (11) in (10) we have from the zero-order terms

$$(\nabla_t{}^2 + v^2 f)\, e_0 = \lambda_0\, e_0, \tag{12}$$

whereas the first-order terms give

$$(\nabla_t{}^2 + v^2 f - \lambda_0)\, e_1 = \lambda_1\, e_0 - \nabla_t\, (\nabla_t f \cdot e_0). \tag{13}$$

Equation (12) corresponds to the usual scalar approximation; both Cartesian components of e_0 satisfy the scalar wave equation

$$[\nabla_t{}^2 + v^2 f(x, y)]\, \psi = \lambda_0\, \psi. \tag{14}$$

Due to symmetries of $f(x, y)$, solutions of (14) may be degenerate. This complicates the further analysis and, in the next section, we will at first exclude such cases.

3. General case, no symmetry

Let ψ be a nondegenerate, quadratically integrable solution of (14) belonging to the eigenvalue λ_0. We may take ψ real and normalized by

$$\iint \psi^2 \, \mathrm{d}\sigma = 1, \tag{15}$$

where the integration extends over the (x, y) plane.

The corresponding solution of (12) can be expressed as

$$e_0 = E_0 \, (\mathbf{i} \cos \theta + \mathbf{j} \sin \theta) \, \psi. \tag{16}$$

Here E_0 is an arbitrary constant with the proper physical dimension, whereas θ represents the polarization angle (thus far unknown) of the modal solution under consideration.

Separation of the first-order equation (13) into Cartesian components and insertion of (16) gives the two equations

$$\begin{aligned}
&(\nabla_t^2 + v^2 f - \lambda_0) \, e_{1x} \\
&\quad = E_0 \left\{ \left[\lambda_1 \, \psi - \frac{\partial}{\partial x} \left(\psi \frac{\partial f}{\partial x} \right) \right] \cos \theta - \frac{\partial}{\partial x} \left(\psi \frac{\partial f}{\partial y} \right) \sin \theta \right\}, \\
&(\nabla_t^2 + v^2 f - \lambda_0) \, e_{1y} \\
&\quad = E_0 \left\{ - \frac{\partial}{\partial y} \left(\psi \frac{\partial f}{\partial x} \right) \cos \theta + \left[\lambda_1 \, \psi - \frac{\partial}{\partial y} \left(\psi \frac{\partial f}{\partial y} \right) \right] \sin \theta \right\}.
\end{aligned} \tag{17}$$

We multiply both sides of these equations by ψ and integrate over the (x, y) plane. Noting the self-adjointness of the left-hand operator and using (14) and (15) we find after some partial integrations

$$\begin{aligned}
\lambda_1 \cos \theta &= A_{11} \cos \theta + A_{12} \sin \theta, \\
\lambda_1 \sin \theta &= A_{21} \cos \theta + A_{22} \sin \theta,
\end{aligned} \tag{18}$$

in which

$$\begin{aligned}
A_{11} &= \tfrac{1}{2} \iint (\psi^2)_{xx} f \, \mathrm{d}\sigma, \\
A_{12} &= A_{21} = \tfrac{1}{2} \iint (\psi^2)_{xy} f \, \mathrm{d}\sigma, \\
A_{22} &= \tfrac{1}{2} \iint (\psi^2)_{yy} f \, \mathrm{d}\sigma.
\end{aligned} \tag{19}$$

Here the subscripts x and y denote partial differentiations.

We thus find two solutions, λ_{11} and λ_{12} say, for λ_1, given by

$$\begin{matrix}\lambda_{11} \\ \lambda_{12}\end{matrix} = \tfrac{1}{2} \{ A_{11} + A_{22} \pm [(A_{11} - A_{22})^2 + 4 A_{12}{}^2]^{\frac{1}{2}} \}. \tag{20}$$

Unless $\lambda_{11} = \lambda_{12}$, which happens if $A_{11} = A_{22}$ and $A_{12} = 0$, the degeneracy of the zero-order solution is thus removed. The corresponding polarization directions are mutually perpendicular and given by

$$\tan\theta_{1,2} = \{A_{22} - A_{11} \pm [(A_{22} - A_{11})^2 + 4A_{12}{}^2]^{\frac{1}{2}}\}/2A_{12}. \qquad (21)$$

With the help of (4) to (6) the lowest-order contributions to the modal field distributions can be summarized as follows

$$\begin{aligned} e_x &= E_0\,\psi\cos\theta + O(\delta),\\ e_y &= E_0\,\psi\sin\theta + O(\delta), \qquad (22)\\ e_z &= \mathrm{i}E_0\,v^{-1}\,\delta^{\frac{1}{2}}\,(\psi_x\cos\theta + \psi_y\sin\theta) + O(\delta^{\frac{3}{2}}),\end{aligned}$$

and

$$\begin{aligned} h_x &= -\,(\varepsilon_1/\mu_0)^{\frac{1}{2}}\,E_0\,\psi\sin\theta + O(\delta),\\ h_y &= (\varepsilon_1/\mu_0)^{\frac{1}{2}}\,E_0\,\psi\cos\theta + O(\delta), \qquad (23)\\ h_z &= \mathrm{i}\,(\varepsilon_1/\mu_0)^{\frac{1}{2}}\,E_0\,v^{-1}\,\delta^{\frac{1}{2}}\,(\psi_y\cos\theta - \psi_x\sin\theta) + O(\delta^{\frac{3}{2}}).\end{aligned}$$

Thus far we have tacitly assumed certain smoothness properties of $\varepsilon(x, y)$, i.e. differentiability of f and of ∇f. Discontinuity surfaces, as occur in most practical models of index profiles, require application of the usual boundary conditions (continuity of the tangential components of $\boldsymbol{e}$ and $\boldsymbol{h}$). In terms of $\boldsymbol{e}_t$ these lead to the requirement that $\boldsymbol{n}\times\boldsymbol{e}_t$, $\boldsymbol{n}\cdot(\varepsilon\boldsymbol{e}_t)$, $\nabla_t\times\boldsymbol{e}_t$ and $(\nabla_t + \nabla_t\varepsilon/\varepsilon)\cdot\boldsymbol{e}_t$ all pass continuously through the surface (here $\boldsymbol{n}$ is the unit normal). For $\boldsymbol{e}_0$ this means, from (7) and (11), that $\boldsymbol{e}_0$ as well as its derivatives $\nabla_t\cdot\boldsymbol{e}_0$ and $\nabla_t\times\boldsymbol{e}_0$ are continuous, which finally leads to the requirement of continuity of ψ and $\nabla_t\,\psi$. Consideration of the first-order terms in the boundary conditions then leads to the formulation of prescribed jumps at the surface of both $\boldsymbol{n}\cdot\boldsymbol{e}_1$ and $\nabla_t\cdot\boldsymbol{e}_1$. Our final result, however, as expressed by (19) to (23), remains unchanged.

Equivalent forms of (19) obtained by partial integration can be more practical, in particular if ψ is known as the result of a numerical computation. In the case of a step-index fibre ($f = 1$ in the core region and $f = 0$ elsewhere) one obtains in this way

$$\begin{aligned} A_{11} &= \tfrac{1}{2}\oint(\psi^2)_x\,(\mathbf{i}\cdot\boldsymbol{n})\,\mathrm{d}s,\\ A_{12} &= \tfrac{1}{2}\oint(\psi^2)_y\,(\mathbf{i}\cdot\boldsymbol{n})\,\mathrm{d}s = \tfrac{1}{2}\oint(\psi^2)_x\,(\boldsymbol{j}\cdot\boldsymbol{n})\,\mathrm{d}s, \qquad (24)\\ A_{22} &= \tfrac{1}{2}\oint(\psi^2)_y\,(\mathbf{j}\cdot\boldsymbol{n})\,\mathrm{d}s,\end{aligned}$$

in which the integration is along the core-cladding boundary in the (x, y) plane. The unit normal $\boldsymbol{n}$ points outward from the core region.

Agreement with the usual notation is obtained by putting $w = \lambda^{\frac{1}{2}}$, $u = (v^2 - \lambda)^{\frac{1}{2}}$. Then we have

$$u \sim u_0 + \delta u_1 + O(\delta^2), \qquad (25)$$

in which

$$u_0 = (v^2 - \lambda_0)^{\frac{1}{2}}, \qquad u_1 = -\frac{\lambda_1}{2u_0}. \tag{26}$$

4. Effects of symmetry

In general the mode polarization directions as given by (21) are different for each mode pair. If, however, the index profile has a plane of mirror symmetry such that, for a suitable choice of coordinate axes, $f(x, y) = f(-x, y)$ we find that $A_{12} = 0$ for all modes. These are either x- or y-polarized with λ_1 given by $\lambda_1 = A_{11}$ or $\lambda_1 = A_{22}$ respectively. The same applies, of course, if in addition $f(x, y) = f(x, -y)$.

A more complicated situation exists if $f(x, y)$ has n-fold rotation symmetry with $n \geqslant 3$. We define the rotation operator R by

$$Rf(x, y) = f\left(x\cos\frac{2\pi}{n} - y\sin\frac{2\pi}{n},\; x\sin\frac{2\pi}{n} + y\cos\frac{2\pi}{n}\right). \tag{27}$$

When R is to be applied to a vector function of position like $\boldsymbol{e}_t$ it must be noted that a rotation of its direction is involved. In particular we have

$$R\mathbf{i} = \mathbf{i}\cos\frac{2\pi}{n} + \mathbf{j}\sin\frac{2\pi}{n},$$

and

$$R\mathbf{j} = -\,\mathbf{i}\sin\frac{2\pi}{n} + \mathbf{j}\cos\frac{2\pi}{n}. \tag{28}$$

If f is invariant under the operations of the plane rotation group (C_n) generated by R (thus $Rf = f$) then the equations (10) for $\boldsymbol{e}_t$ and (14) for the Cartesian components of $\boldsymbol{e}_0$ are also invariant. If ψ is a solution of (14) then $R\psi$ is also a solution. Similarly if $\boldsymbol{e}_t$ satisfies (10) then so too does $R\boldsymbol{e}_t$.

We must note at this stage that we assume all solutions considered here to be real. This is no loss of generality because the real and imaginary parts of complex solutions are solutions as well.

Then, in general, solutions of (10) and (14) must conform to one of the real irreducible matrix representations of C_n. We shall label the various possible transformation rules as $T_+, T_-, T_1, T_2, \ldots, T_{k_n}$, in which

$$k_n = \begin{cases} (n-1)/2 & (n \text{ odd}) \\ (n-2)/2 & (n \text{ even}). \end{cases} \tag{29}$$

Non-degenerate solutions ψ of (14) satisfy either

$$T_+: \qquad R\psi = \psi, \tag{30}$$

or, if n is even,

$$T_-: \qquad R\psi = -\psi. \tag{31}$$

Doubly degenerate solutions of (14) are linear combinations of basic solutions (ψ_1, ψ_2) satisfying

$$T_k: \qquad \begin{pmatrix} R\psi_1 \\ R\psi_2 \end{pmatrix} = \begin{pmatrix} \cos \dfrac{2k\pi}{n} & \sin \dfrac{2k\pi}{n} \\ -\sin \dfrac{2k\pi}{n} & \cos \dfrac{2k\pi}{n} \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \qquad (32)$$

in which $1 \leqslant k \leqslant k_n$.

An exactly similar statement applies to solutions $\boldsymbol{e}_t$ of (10), and thus for their zero-order limit $\boldsymbol{e}_0$. These considerations are of much help in determining the ways in which solutions of (14) may occur as Cartesian components of $\boldsymbol{e}_0$. The situation is simplest if ψ is nondegenerate, transforming according to either T_+ or T_-. The corresponding solutions for $\boldsymbol{e}_t$ (and thus for $\boldsymbol{e}_0$) are doubly degenerate and $\boldsymbol{e}_0$ is of the form $\boldsymbol{a}\psi$ in which $\boldsymbol{a}$ is a constant vector in the (x, y) plane. If ψ transforms according to T_+ then $\boldsymbol{e}_t$ transforms according to T_1, as is seen if $(\mathbf{i}\psi, \mathbf{j}\psi)$ is chosen as a basis for $\boldsymbol{e}_0$. Similarly, if n is even and ψ transforms according to T_-, $\boldsymbol{e}_t$ transforms according to $T_{\frac{1}{2}n-1}$. The basis for $\boldsymbol{e}_0$ is then $(\mathbf{i}\psi, -\mathbf{j}\psi)$. In these cases the theory of the foregoing section fully applies with, however, $A_{11} = A_{22}$ and $A_{12} = A_{21} = 0$ in (18). It follows that

$$\lambda_1 = \tfrac{1}{4} \iint f \nabla_t^2 \psi^2 \, \mathrm{d}\sigma. \qquad (33)$$

Let us now assume that the function pair (ψ_1, ψ_2) forms a solution basis for (14), transforming according to T_k. The functions ψ_1 and ψ_2 are orthogonal and can be normalized such that

$$\iint \psi_1^2 \, \mathrm{d}\sigma = \iint \psi_2^2 \, \mathrm{d}\sigma = 1. \qquad (34)$$

Now let

$$\begin{aligned} \boldsymbol{e}_\mathrm{a} &= \mathbf{i}\psi_1 + \mathbf{j}\psi_2, & \boldsymbol{e}_\mathrm{b} &= \mathbf{i}\psi_2 - \mathbf{j}\psi_1, \\ \boldsymbol{e}_\mathrm{c} &= \mathbf{i}\psi_1 - \mathbf{j}\psi_2, & \boldsymbol{e}_\mathrm{d} &= \mathbf{i}\psi_2 + \mathbf{j}\psi_1. \end{aligned} \qquad (35)$$

A simple calculation shows that, for $k > 1$, the function pair $(\boldsymbol{e}_\mathrm{a}, \boldsymbol{e}_\mathrm{b})$ transforms according to T_{k-1}, whereas for $k = 1$ both $\boldsymbol{e}_\mathrm{a}$ and $\boldsymbol{e}_\mathrm{b}$ transform according to T_+. On the other hand, for $k < k_n$, the function pair $(\boldsymbol{e}_\mathrm{c}, \boldsymbol{e}_\mathrm{d})$ transforms according to T_{k+1}. If $k = k_n$ and n is odd then $(\boldsymbol{e}_\mathrm{c}, -\boldsymbol{e}_\mathrm{d})$ transforms according to T_{k_n}. Finally, if $k = k_n$ and n is even, both $\boldsymbol{e}_\mathrm{c}$ and $\boldsymbol{e}_\mathrm{d}$ transform according to T_-.

This means that $\boldsymbol{e}_0$ must be either a linear combination of $\boldsymbol{e}_\mathrm{a}$ and $\boldsymbol{e}_\mathrm{b}$, or a linear combination of $\boldsymbol{e}_\mathrm{c}$ and $\boldsymbol{e}_\mathrm{d}$. Depending on the transformation rule for $(\boldsymbol{e}_\mathrm{a}, \boldsymbol{e}_\mathrm{b})$, respectively $(\boldsymbol{e}_\mathrm{c}, \boldsymbol{e}_\mathrm{d})$, the corresponding solutions for $\boldsymbol{e}_0$ (and thus for $\boldsymbol{e}_t$) are either non-degenerate or doubly degenerate.

We now return to eq. (13) in which we substitute

$$e_0 = E_0\,(e_a \cos\theta + e_b \sin\theta). \tag{36}$$

Scalar multiplication by e_a and by e_b respectively, followed by integration over the (x, y) plane leads to a pair of equations similar to (18) in which now, however,

$$\begin{aligned}
A_{11} &= \iint [\tfrac{1}{2}(\psi_1{}^2)_{xx} + \tfrac{1}{2}(\psi_2{}^2)_{yy} + (\psi_1\,\psi_2)_{xy} + \psi_{1x}\,\psi_{2y} - \psi_{2x}\,\psi_{1y}] \\
&\qquad \times f(x, y)\,\mathrm{d}\sigma, \\
A_{22} &= \iint [\tfrac{1}{2}(\psi_1{}^2)_{yy} + \tfrac{1}{2}(\psi_2{}^2)_{xx} - (\psi_1\,\psi_2)_{xy} + \psi_{1x}\,\psi_{2y} - \psi_{2x}\,\psi_{1y}] \\
&\qquad \times f(x, y)\,\mathrm{d}\sigma,
\end{aligned}$$

and (37)

$$A_{12} = A_{21} = \tfrac{1}{2}\iint [(\psi_1\,\psi_2)_{xx} - (\psi_1\,\psi_2)_{yy} - (\psi_1{}^2 - \psi_2{}^2)_{xy}]\,f(x, y)\,\mathrm{d}\sigma.$$

If $k = 1$, as we have seen, there are two non-degenerate solutions for e_0 and the corresponding values of λ_1 and $\tan\theta$ are given by (20) and (21) with (37). If, in addition to the rotation symmetry $f(x, y)$ has reflection symmetry, this case is further simplified by choosing one of the n mirror axes as the y axis and choosing for ψ_1 and ψ_2 even and odd functions of x respectively. Then we have $A_{12} = 0$ and find $e_0 = E_0\,e_a$ and $e_0 = E_0\,e_b$ as solutions with $\lambda_1 = A_{11}$ and $\lambda_1 = A_{22}$ respectively.

If $k > 1$, e_0 is doubly degenerate and consequently it must hold that $A_{12} = A_{21} = 0$, and $A_{11} = A_{22}$, Then λ_1 is given by

$$\lambda_1 = \iint [\tfrac{1}{4}\,\nabla_t{}^2\,(\psi_1{}^2 + \psi_2{}^2) + \psi_{1x}\,\psi_{2y} - \psi_{2x}\,\psi_{1y}]\,f\,\mathrm{d}\sigma. \tag{38}$$

A completely similar treatment may be given assuming

$$e_0 = E_0\,(e_c \cos\theta + e_d \sin\theta). \tag{39}$$

Then the case where n is even and $k = k_n$ leads to non-degenerate solutions for e_0, otherwise e_0 is doubly degenerate. The expressions corresponding to (37) and (38) are simply found by replacing $\partial/\partial y$ by $-\,\partial/\partial y$ in (37) and (38).

After the foregoing analysis, a treatment of the practically most important case of circular symmetry is a relatively simple matter. We introduce polar coordinates (r, φ) in the (x, y) plane according to $x = r\cos\varphi$, $y = r\sin\varphi$, and assume $f = f(r)$. Solutions of eq. (14) are all of the form

$$\psi = \phi(r)\begin{cases}\cos m\varphi \\ \sin m\varphi\end{cases} \qquad (m = 0,\ 1,\ \ldots) \tag{40}$$

in which $\phi(r)$ is a solution of

$$\phi'' + \frac{1}{r}\phi' + \left[v^2 f(r) - \frac{m^2}{r^2}\right]\phi = \lambda_0\,\phi \tag{41}$$

subject to the normalization condition

$$\int_0^\infty \phi^2\, r\, \mathrm{d}r = \begin{cases} \dfrac{1}{2\pi} & (m = 0) \\ \dfrac{1}{\pi} & (m > 0). \end{cases} \tag{42}$$

If $m = 0$, ψ transforms according to T_+ and if $m > 0$ the pair of solutions $(\phi \cos m\varphi, \phi \sin m\varphi)$ transforms according to T_m. Application of the preceding theory then immediately leads to the well-established structure of the modal system of the circular dielectric waveguide and to the way in which the vaiious electromagnetic propagation modes (denoted as HE-, EH-, TM-, and TE-modes) are related to the LP-modes of the scalar approximation [1-3,5,6]. The first-order perturbation coefficients λ_1 for the various modes in their usual designations are summarized in table I.

TABLE I

m	mode	λ_1	u_1 for step-index fiber
0	HE_{1n}	$\pi \int_0^\infty (\phi\phi' r)' f\,\mathrm{d}r$	$\dfrac{u_0\, w_0}{2v^2}\,\dfrac{K_0}{K_1}$
1	TE_n	0	0
	TM_n	$\pi \int_0^\infty (\phi\phi' r + \phi^2)' f\,\mathrm{d}r$	$\dfrac{u_0\, w_0}{v^2}\,\dfrac{K_1}{K_2}$
	HE_{2n}	$\frac{1}{2}\pi \int_0^\infty (\phi\phi' r - \phi^2)' f\,\mathrm{d}r$	$\dfrac{u_0\, w_0}{2v^2}\,\dfrac{K_1}{K_0}$
>1	$\mathrm{EH}_{m-1,n}$	$\frac{1}{2}\pi \int_0^\infty (\phi\phi' r + m\phi^2)' f\,\mathrm{d}r$	$\dfrac{u_0\, w_0}{2v^2}\,\dfrac{K_m}{K_{m+1}}$
	$\mathrm{HE}_{m+1,n}$	$\frac{1}{2}\pi \int_0^\infty (\phi\phi' r - m\phi^2)' f\,\mathrm{d}r$	$\dfrac{u_0\, w_0}{2v^2}\,\dfrac{K_m}{K_{m-1}}$

As a check on these results we applied them to the analytically known modal solutions for the circular step-index fibre for which

$$f(r) = \begin{cases} 1 & (r < 1) \\ 0 & (r > 1) \end{cases} \tag{43}$$

These results, expressed as u_1 according to (26), are also given in table I. They are in full agreement with those following from the exact characteristic equation [5]) of the circular step-index fibre. In this case the zero-order characteristic equation, satisfied by u_0 and w_0, is

$$\frac{J_m(u_0)}{u_0\, J_{m-1}(u_0)} + \frac{K_m(w_0)}{w_0\, K_{m-1}(w_0)} = 0, \qquad {u_0}^2 + {w_0}^2 = v^2, \tag{44}$$

in which J_m and K_m denote Bessel functions and modified Bessel functions respectively. In table I we have abbreviated $K_m(w_0) = K_m$.

5. Final remarks

Except in the case of three- or more-fold rotational symmetry. all modes are linearly polarized in the zero-order limit. On the other hand, most modes of rotationally symmetric configurations have a rather more complicated structure and lead to expressions for the coefficient λ_1 which are different from those found for the general case. The question of how the patterns of these modes — and their propagation constants — behave with small disturbances of the symmetry, deserves further attention and we shall treat this problem in a separate paper [7]).

The effects of symmetry, apart from the circular case, are well demonstrated by the results of Goell [8]) who computed numerically propagation constants and mode patterns for a step-index fibre with rectangular core cross-section. His results include the case of square symmetry and his findings concerning the corresponding mode degeneracies fully conform with those which would follow from our analysis.

Philips Research Laboratories *Eindhoven, March 1978*

REFERENCES

[1]) A. W. Snyder, IEEE Trans. Microwave Theory Tech. **MTT-17**, 1130-1138, 1969.
[2]) D. Gloge, Applied Optics **10**, 2252-2258, 1971.
[3]) J. A. Arnaud, Bell Syst. Tech. J. **53**, 675-696, 1974.
[4]) D. L. A. Tjaden, Birefringence in single-mode optical fibers due to core ellipticity, to be published.
[5]) E. Snitzer, J. Opt. Soc. Am. **51**, 491-498, 1961.
[6]) C. N. Kurtz, J. Opt. Soc. Am. **65**, 1235-1240, 1975.
[7]) D. L. A. Tjaden, Philips J. Res., to be published.
[8]) J. E. Goell, Bell Syst. Tech. J. **48**, 2133-2160, 1969.

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*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 Avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 5100 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).

L. K. H. van Beek: Special properties of physical development processes.
Photogr. Sci. Engng. **20,** 88-91, 1976. *E*

C. Belin: On the growth of large single crystals of calcite by travelling solvent zone melting.
J. Crystal Growth **34,** 341-344, 1976. *L*

C. Belouet, M. Monnier, E. Dunia and J. F. Petroff: X-ray topographic study of dislocations in $KH_{2(1-x)}D_{2x}PO_4$ single crystals.
Mat. Res. Bull. **11,** 903-910, 1976. *L*

F. Berz and H. K. Kuiken: Theory of lifetime measurements with the scanning electron microscope: steady state.
Solid-State Electronics **19,** 437-445, 1976. *R*

K. Bethe: Novel precision thermistor from intrinsic germanium.
IEEE Trans. **IECI-23,** 420-424, 1976. *H*

R. N. Bhargava, P. M. Harnack, S. P. Herko, P. C. Mürau and R. J. Seymour: Thermally stimulated current measurements in Cu and O-doped GaP.
J. Luminescence **12/13,** 515-519, 1976. *N*

G. W. Blackmore, J. B. Clegg, J. S. Hislop and J. B. Mullin: Concentrations of carbon and oxygen in indium phosphide and gallium arsenide crystals grown by the LEC technique.
J. Electronic Mat. **5,** 401-413, 1976. *R*

R. D. Boehnke: Anwendung und Kontrolle des Ionenätzens in Hochfreqenz-Zerstäubungsanlagen.
Vakuum-Technik **25,** 195-199, 1976. *H*

J. van den Boomgaard, A. M. J. G. van Run and J. van Suchtelen: Magnetoelectricity in piezoelectric-magnetostrictive composites.
Ferroelectrics **10,** 295-298, 1976. *E*

M. R. Boudry, J. A. Morice and E. J. Millett: A versatile multi-access computing system for laboratory instrumentation.
On-line computing in the laboratory, R. A. Rosner, B. K. Penny and P. N. Clout(eds), publ. Advance Publ., London, 1976, pp. 353-361. *R*

P. W. J. M. Boumans: Corrections for spectral interferences in optical emission spectrometry with special reference to the RF inductively coupled plasma.
Spectrochim. Acta **31B,** 147-152, 1976. *E*

A. L. J. Burgmans and J. P. Woerdman: Selective reflection from sodium vapour at low densities.
J. Physique **37**, 677-681, 1976. *E*

A. Bril and A. W. de Jager-Veenis: Quantum efficiency standard for ultraviolet and visible excitation.
J. Electrochem. Soc. **123**, 396-398, 1976. *E*

H. H. Brongersma, N. Hazewindus, J. M. van Nieuwland, A. M. M. Otten and A. J. Smets: Angular-dependent Ne^+-ion scattering from a solid Au target.
J. Vac. Sci. Technol. **13**, 670-675, 1976. *E*

E. Bruninx: X-ray fluorescence analysis by means of crystal dispersion and a position-sensitive counter.
Spectrochim. Acta **31B**, 221-223, 1976. *E*

K. H. J. Buschow and M. Brouha: Narrow Bloch walls in the RCo$_5$-type rare-earth—cobalt compounds.
J. Appl. Phys. **47**, 1653-1656, 1976. *E*

K. H. J. Buschow and A. M. van Diepen: Effect of hydrogen absorption on the magnetic properties of YFe_2 and $GdFe_2$.
Solid State Comm. **19**, 79-81, 1976. *E*

K. H. J. Buschow: Hydrogen absorption and its effect on the magnetic properties of rare-earth iron intermetallics.
Solid State Comm. **19**, 421-423, 1976. *E*

K. L. Bye: Structural dependence of the electro-optic properties of some PLZT ceramics.
Ferroelectrics **12**, 221-223, 1976. *R*

T. A. C. M. Claasen, W. F. C. Mecklenbrauker and J. B. H. Peek: A survey of quantization and overflow effects in recursive digital filters.
Proc. 1976 IEEE Int. Symp. on Circuits and systems, Munich, pp. 621-624. *E*

P.-J. Courtois and H. Vantilborgh: A decomposable model of program paging behaviour.
Acta Informatica **6**, 251-275, 1976. *B*

J. P. M. Damen: Measurement of growth rate and nucleation temperature of flux-grown $YbFeO_3$.
J. Crystal Growth **33**, 266-270, 1976. *E*

P. A. Devijver: Entropie quadratique et reconnaissance des formes.
Computer oriented learning processes, J. C. Simon (ed.), publ. Noordhoff, Leiden, 1976 (NATO ASI Ser. E 14), pp. 257-278. *B*

J. A. W. van der Does de Bye: Dependence of recombination in p-type GaP(Zn,O) on dopant concentrations.
J. Electrochem. Soc. **123**, 544-551, 1976. *E*

E. Dormann and K. H. J. Buschow: A comparative study of the hyperfine fields in ferromagnetic GdRh and GdZn.
J. Appl. Phys. **47**, 1662-1667, 1976. *E*

H. Durand: Transformer le rayonnement solaire en électricité?
Revue Andès No. 19, 24-27, 1976. *L*

J. B. A. A. Elemans and K. H. J. Buschow: Comment on the crystal and magnetic structures of $CaCu_5$-type compounds of Th with Fe, Co, and Ni.
Phys. Stat. Sol. (a) **34**, 355-359, 1976. *E*

J. B. A. A. Elemans, P. C. M. Gubbens and K. H. J. Buschow: On the interpretation of Mössbauer-effect and diffraction measurements on rare earth - iron compounds of the type R_2Fe_{17}.
J. Less-Common Met. **44**, 51-62, 1976. *E*

H. A. van Essen: I^2L and its application in a digital data transmitter.
Proc. 1976 IEEE Int. Symp. on Circuits and systems, Munich, pp. 735-738. *E*

V. H. C. M. Evers: Operational experience with the ANS on-board computer.
J. Brit. Interplanetary Soc. **29**, 417-427, 1976. *E*

G. Frens: Measurements with colloids.
Progr. Colloid & Polymer Sci. **59**, 27-32, 1976. *E*

R. C. French: Radio propagation in London at 462 MHz.
Radio and Electronic Engr. **46**, 333-336, 1976. *R*

C. T. Foxon, B. A. Joyce and S. Holloway: Instrument response function of a quadrupole mass spectrometer used in time-of-flight measurements.
Int. J. Mass Spectrom. Ion Phys. **21**, 241-255, 1976. *R*

M. Gleria and R. Memming: Novel luminescence generation by electron transfer from semiconductor electrodes to ruthenium-bipyridil complexes.
Z. phys. Chemie neue Folge **101**, 171-179, 1976. *H*

J.-M. Goethals: Nonlinear codes defined by quadratic forms over GF(2).
Information and Control **31**, 43-74, 1976. *B*

P. C. M. Gubbens and K. H. J. Buschow: Mössbauer effect study of $Tm_2Fe_{17-x}Co_x$ and $Tm_2Fe_{17-x}Ni_x$ compounds.
Phys. Stat. Sol. (a) **34**, 729-735, 1976. *E*

J. Hasker: A new class of efficient low-pressure gas discharges with high radiation output per unit volume.
Appl. Phys. Letters **28**, 586-588, 1976. *E*

E. E. Havinga: Why are the close-packed structures of the noble metals cubic?
Physica **82B**, 277-287, 1976. *E*

B. K. Herbert: A circuit for stabilizing the electron current to the anode of a hot-filament device.
Vacuum **26**, 363-369, 1976. *R*

D. Kasperkovitz and R. J. M. Verbeek: A low-power circuit block for digital telephone exchanges.
Microelectronics and Reliability **15**, 163-170, 1976. *E*

W. L. Konijnendijk (Philips Lighting Division, Eindhoven) **and J. M. Stevels** (Eindhoven University of Technology): Density and refractive index of borosilicate glasses in relation to their structure.
Verres Réfract. **30**, 223-225, 1976.

H. K. Kuiken: Theory of lifetime measurements with the scanning electron microscope: transient analysis.
Solid-State Electronics **19**, 447-450, 1976. *R*

D. J. Kroon: Automatisch meten van luchtverontreiniging.
Extern **5**, 113-131, 1976. *E*

D. Kuppers, J. Koenings and H. Wilson: Codeposition of glassy silica and germania inside a tube by plasma-activated CVD.
J. Electrochem. Soc. **123**, 1079-1083, 1976. *A*

R. Metselaar, J. P. M. Damen, P. K. Larsen and M. A. H. Huyberts: Investigation of colour centres in gadolinium gallium garnet crystals.
Phys. Stat. Sol. (a) **34**, 665-670, 1976. *E*

R. Metselaar and P. K. Larsen: Diffusion of oxygen vacancies in yttrium iron garnet investigated by dynamic conductivity measurements.
J. Phys. Chem. Solids **37**, 599-605, 1976. *E*

P. L. A. Chr. M. van der Meer, L. J. Giling (both with University of Nijmegen) **and S. G. Kroon** (Philips Semiconductor Development Laboratory, Nijmegen): The emission coefficient of silicon coated with Si_3N_4 or SiO_2 layers.
J. Appl. Phys. **47**, 652-655, 1976.

A. E. Morgan and H. W. Werner: Quantitative analysis of low alloy steels by secondary ion mass spectrometry.
Anal. Chem. **48**, 699-708, 1976. *E*

W. J. Oosterkamp: Benefit/risk comparisons in diagnostic radiology.
Medicamundi **21**, 2-6, 1976. *E*

A. van Oostrom: Application of AES to the study of selective sputtering of thin films.
J. Vac. Sci. Technol. **13**, 224-227, 1976. *E*

J. J. Opstelten and L. B. Beijer (Philips Lighting Division, Eindhoven): Specification of colour rendering properties of light sources for colour television.
Lighting Res. Technol. **8**, 89-102, 1976.

J. A. Pals: On the detectibility of spin-triplet pairing superconductivity with the aid of the Josephson effect.
Physics Letters **56A**, 414-416, 1976. *E*

H. L. Peek: Twin-layer PCCD performance for different doping levels of the surface layer.
IEEE J. **SC-11**, 167-170, 1976. *E*

W. Puschert and H. Scholz: k-ray spectra detected with HgI_2 at room temperature.
Appl. Phys. Letters **28**, 357-359, 1976. *A*

H. Rau: Homogeneity range of high temperature $Ni_{3\pm x}S_2$.
J. Phys. Chem. Solids **37**, 929-930, 1976. *A*

G. Renelt: Easily decodable runlength code (e.d.r.c.) for source encoding of black-and-white facsimile pictures.
Electronics Letters **12**, 633-634, 1976. *H*

A. Rijbroek (Philips' Telecommunicatie Industrie B.V., Huizen): Design approaches for a PCM-codec per channel.
Proc. 1976 IEEE Int. Symp. on Circuits and Systems, Munich, pp. 587-590.

F. G. Rudenauer, W. Steiger and H. W. Werner: On the use of the Saha-Eggert equation for quantitative SIMS analysis using argon primary ions.
Surface Sci. **54**, 553-560, 1976. *E*

J. M. S. Schofield: The physics of gas discharge cells within d.c. memory display panels.
4th Int. Conf. on Gas discharges, Swansea, 1976 (IEE Conf. Publn No. 143), pp. 397-400. *R*

M. F. H. Schuurmans: Spectral narrowing of selective reflection.
J. Physique **37**, 469-485, 1976. *E*

G. Simpson and E. T. Keve: Anomalous ferroelectric behaviour in PLZT.
Ferroelectrics **12**, 229-231, 1976. *R*

J. L. Sommerdijk and A. Bril: On the position of the $^5D_0$ level of Eu^{3+} in $AMgF_3$ ($A = K$, Rb, Cs).
J. Luminescence **12/13**, 669-673, 1976. *E*

A. L. N. Stevels: Recent developments in the application of phosphors.
J. Luminescence **12/13**, 97-107, 1976. *E*

J. B. Theeten and F. Hottier: On the role of chlorine in the vapour phase epitaxy of (100)GaAs as evidenced by LEED and RHEED.
Surface Sci. **58**, 583-589, 1976. *L*

J. van der Veen: Liquid crystalline isothiocyanates.
J. Physique **37**, C3/13-15, 1976 (Colloque C3). *E*

C. H. F. Velzel: A general theory of the aberrations of diffraction gratings and gratinglike optical instruments.
J. Opt. Soc. Amer. **66**, 346-353, 1976. *E*

J. O. Voorman: The adaptive gyrator.
Proc. 1976 IEEE Int. Symp. on Circuits and systems, Munich, pp. 34-37. *E*

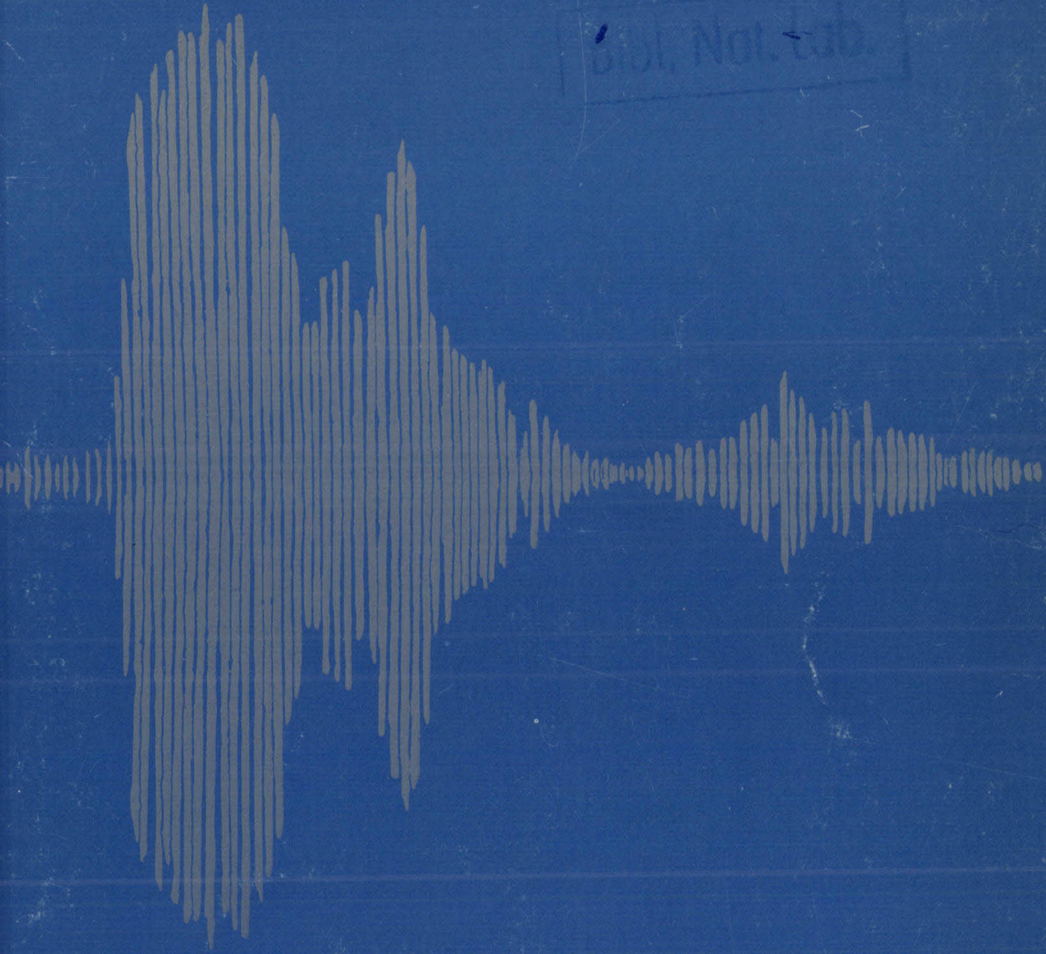
SR. 3-40=33

hilips Journal of Research

Vol. 33 Nos. 3/4 1978

12 MAART 1979

Bibl. Nat.Lab.



HILIPS

Philips Journal of Research

Philips Journal of Research, edited by Philips Research Laboratories, Eindhoven, The Netherlands, is a bimonthly publication containing papers on research carried out in the various Philips laboratories. Volumes 1 to 32 have appeared under the name of Philips Research Reports.

Annual subscription price for Volume 33 is Dfl. 60, payable in advance. Payments should be made only after receipt of an invoice. Correspondence should be addressed to: Philips Journal of Research, Philips Research Laboratories, Building WBp, Room No. 42, Eindhoven, The Netherlands.

Cover design based on a visual representation of the sound pressure associated with the spoken word "Philips".

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1978. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

12 MAART 1979

PHILIPS JOURNAL OF RESEARCH - Vol. 33 Nos. 3/4 - pp. 117-210 - 1978

CONTENTS

Date of publication: January 8, 1979

LINE EMISSION OF $LiBaAlF_6$: Eu^{2+}

by J. L. SOMMERDIJK, P. VRIES and A. BRIL

Abstract

The luminescence of $LiBaAlF_6:Eu^{2+}$ consists of band emission due to a $4f^6\,5d \rightarrow 4f^7$ transition and line emission due to a $4f^7 \rightarrow 4f^7$ transition. Both emissions are located in the ultraviolet region around 360 nm. The energy difference between the emission levels and the transition probabilities have been arrived at by measuring the line-band intensity ratio and the decay time of the luminescence as a function of temperature. The occurrence of line emission is compared with that of other Eu^{2+}-activated Ba or Sr compounds.

1. Introduction

The luminescence of Eu^{2+} mostly consists of a $4f^6\,5d \rightarrow 4f^7$ broad band transition [1]. Some Eu^{2+}-activated lattices are also observed to show line emission, which is ascribed to a transition from the lowest excited $4f^7$ level ($^6P_{7/2}$) to the $4f^7$ ground level ($^8S_{7/2}$). In reference 2 a summary is given of Eu^{2+}-activated phosphors showing line emission. This summary includes the lattices $SrAlF_5$, $BaAlF_5$, $BaMg(SO_4)_2$, SrF(Cl, Br), BaF(Cl, Br), $SrBe_2Si_2O_7$ and $SrAl_{12}O_{19}$. Line emission has also been reported for $BaBe_2Si_2O_7$ (ref. 3), $LiBaF_3$ (ref. 4–6), $AMgF_3$ (A = Na, K, Rb, Cs) (ref. 7–9), $BaCaLu_2F_{10}$ (ref. 10), BaY_2F_8, $SrSiF_6$ and $BaSiF_6$ (ref. 11). In all these lattices the $^6P_{7/2}$ level of Eu^{2+} is situated below the lowest $4f^6$ 5d level. The relative intensity of the line and band emissions depends on the transition probabilities (W_f and W_d) corresponding to these emissions, on the energy difference (Δ) between the two emission levels and on the temperature.

In a study of the luminescence of $LiMAlF_6:Eu^{2+}$ (M = Ca, Sr, Ba), it was found that the compound with M = Ba also belongs to the class of line-emitting Eu^{2+} phosphors, whereas the compounds with M = Sr or Ca do not [12]. In the work reported here we examine the parameters governing the line and band emissions (W_f, W_d, Δ) by measuring the line-band intensity ratio and the decay time of the luminescence of $LiBaAlF_6:Eu^{2+}$ as a function of temperature between liquid nitrogen temperature and 500 K. The results are briefly discussed. In addition, we discuss the occurrence of line emission in relation to Eu^{2+} compounds which we have studied earlier, namely $SrFCl:Eu^{2+}$ and $BaFCl:Eu^{2+}$ (ref. 13), $SrBe_2Si_2O_7$ and $BaBe_2\,Si_2\,O_7:Eu^{2+}$ (ref. 3), and $SrAl_{12}O_{19}:Eu^{2+}$ and $BaAl_{12}O_{19}:Eu^{2+}$ (ref. 14).

2. Experimental

The samples were prepared by J. G. Verlijsdonk of Philips Lighting Division. The optical measurements were carried out in our laboratories. The determination of the spectral power distributions of the emissions (S.P.D.) and of the decay times has been described earlier [15]). For the present work the following changes were made in the experimental arrangement.

S.P.D. The photomultiplier tube was replaced by an EMI 9659 QA with an extended red response. The spectra were measured in the second grating order (blaze 500 nm, 25000 grooves/inch), as indicated in the previous description. The resolving power used was sufficient to give the real width of the Eu^{2+} lines.

Decay. The "TRW Nanosource" instrument was equipped with a deuterium lamp. The shortest pulse was 4 ns (repetition frequency 2000 cps); the pulse time could be extended in steps up to 100 ns by changing the capacitance (length of cable) of the D_2 lamp ignition circuit.

With the aid of a boxcar integrator (presentation of the decay curve on a recorder instead of an oscilloscope), the noise level of the photomultiplier output was considerably improved compared with the direct oscilloscope display.

Temperature dependence. The luminescence output as a function of temperature was determined with a liquid nitrogen cold finger type cryostat with a heating element on which the phosphor was applied in a very thin layer.

3. Results and discussion

Figure 1 shows the emission spectra of $LiBaAlF_6:Eu^{2+}$ for three different

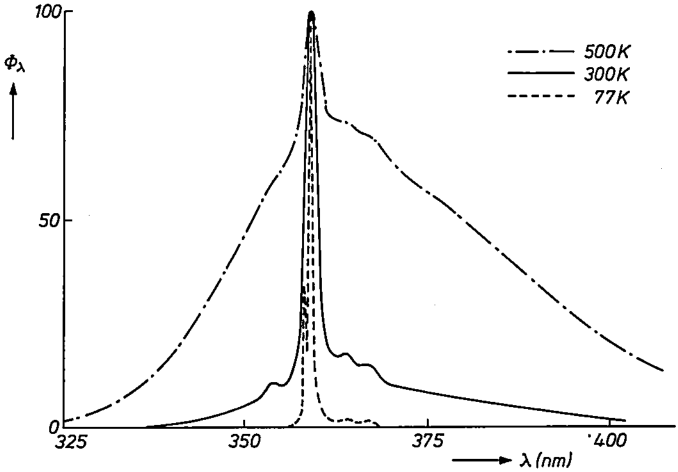


Fig. 1. Spectral energy distribution of the luminescence under 250–270 nm excitation of $LiBaAlF_6:Eu^{2+}$ at three different temperatures. Φ_λ denotes the spectral radiant power in arbitrary units. The maximum of each spectrum is set equal to 100.

temperatures. It is seen that, as usual, with lower temperatures the line emission increases at the expense of the band emission. The total quantum efficiency of the luminescence remains practically constant ($\cong$ 60%) when going from liquid nitrogen temperature to 500 K. At temperatures above 500 K the luminescence becomes weaker due to thermal quenching.

The occurrence of line emission in LiBaAlF$_6$:Eu^{2+} indicates that the lowest excited 4f^7 level (^{6}P$_{7/2}$) in this compound is situated below the lowest 4f^6 5d level. Upon excitation with short-wave ultraviolet radiation, the Eu^{2+} ions are excited to one of the 4f^6 5d levels after which relaxation occurs to the lowest 4f^6 5d level. From here Eu^{2+} can return to the ground state with a transition probability W_d, resulting in band emission. Another possibility is that Eu^{2+} relaxes further to the ^{6}P$_{7/2}$ level with a probability W_{df}. A transition from this level to the ground level (probability W_f) results in line emission. The temperature dependence of the line-band intensity ratio R can be written as

$$R = \frac{(W_f/W_d)\exp(\Delta/kT)}{1 + (W_f/W_{df})\exp(\Delta/kT)} \qquad (1)$$

in which Δ is the energy difference between the lowest 4f^6 5d level and the ^{6}P$_{7/2}$ level. It is assumed here that the degeneracies of the two emission levels do not differ much. At sufficiently low temperatures ($kT \ll \Delta$) eq. (1) reduces to $R \cong W_{df}/W_d$. By substitution of the R value measured at 77 K, we obtain $W_{df} \cong 1.4\ W_d$. On the other hand, the value of W_d is much larger than that of W_f, since the band emission corresponds to a parity-allowed transition, whereas the line emission does not. This implies that $W_f/W_{df} \ll 1$, and therefore at high temperatures ($kT \gg \Delta$) the second term in the denominator becomes negligible, so that eq. (1) reduces to

$$R \cong (W_f/W_d)\exp(\Delta/kT). \qquad (2)$$

Considering now fig. 2 we see indeed a straight line at temperatures $\geqslant$ 300 K when the experimental value of log R is plotted against T^{-1}. From the slope of this line we derive a value of Δ, giving $\Delta = 0.15$ eV.

The decay curves of the Eu^{2+} luminescence were measured over one decade and were found to be exponential in this region. Figure 3 shows the value of the reciprocal of the decay time (τ^{-1}) for various temperatures between 77 K and 500 K. The value of τ^{-1} gives the radiative probability of the luminescence. The temperature dependence of τ^{-1} can be described by (see, for example, ref. 16)

$$\tau^{-1} = \frac{W_f + W_d\exp(-\Delta/kT)}{1 + \exp(-\Delta/kT)}. \qquad (3)$$

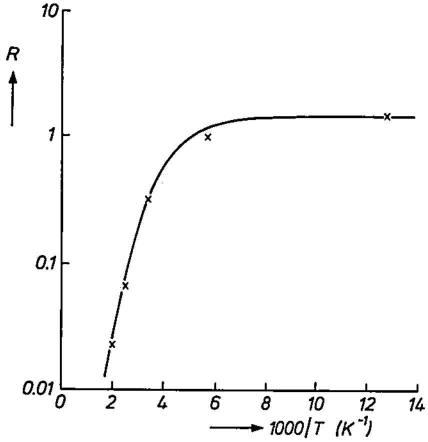


Fig. 2. Line/band intensity ratio R (logarithmic scale) versus T^{-1} for the luminescence of $LiBaAlF_6:Eu^{2+}$;
×: determined from the measured luminescence spectra,
drawn curve: calculated from eq. (1) with
$\Delta = 0.15$ eV, $W_f = 5\times10^2\ s^{-1}$, $W_d = 5\times10^5\ s^{-1}$ and $W_{df} = 7\times10^5\ s^{-1}$.

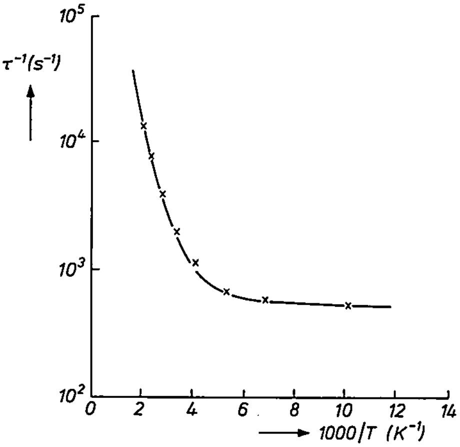


Fig. 3. Reciprocal of the decay time τ (logarithmic scale) versus T^{-1} for the luminescence of $LiBaAlF_6:Eu^{2+}$;
×: determined from the measured decay curves of the luminescence,
drawn curve: calculated from eq. (3) with
$\Delta = 0.15$ eV, $W_f = 5\times10^2\ s^{-1}$ and $W_d = 5\times10^5\ s^{-1}$.

At very low temperatures ($kT \ll \Delta$) the second terms in both the numerator and the denominator can be disregarded, so that $\tau^{-1} \simeq W_f$. Inspection of fig. 3 shows that this situation is practically reached at 77 K, giving $W_f \simeq 5\times10^2\ s^{-1}$.

When the temperature is increased, first the second term in the numerator of eq. (3) begins to contribute, since $W_d/W_f \gg 1$. In a certain temperature region its contribution becomes significantly larger than that of W_f, whereas the second term in the denominator of eq. (3) is still negligible. Then the radiative probability can be approximated by

$$\tau^{-1} \cong W_d \exp(-\Delta/kT). \tag{4}$$

This results in a straight line when $\log \tau^{-1}$ is plotted against T^{-1}. Figure 3 shows that this situation occurs between $T \cong 350$ K and $T \cong 500$ K. The value of Δ derived from the slope of the straight line amounts to 0.15 eV, in agreement with the value obtained from the $\log R$ vs T^{-1} plot (fig. 2). At even higher temperatures the second term in the denominator of eq. (3) begins to contribute as well, so that then τ^{-1} is approximately

$$\tau^{-1} \cong \frac{W_d \exp(-\Delta/kT)}{1 + \exp(-\Delta/kT)}. \tag{5}$$

Unfortunately, this dependence cannot be verified experimentally owing to the strong quenching of the Eu^{2+} luminescence in this region, so that eq. (3), based on a constant total quantum efficiency, no longer holds. In order to obtain an estimate for W_d, we can extrapolate the straight line observed in fig. 3 towards $T^{-1} = 0$, giving $W_d \cong 5 \times 10^5$ s^{-1}.

Figures 2 and 3 show also the calculated temperature dependence of R and τ^{-1} between 77 K and 500 K. The calculations are based on eqs (1) and (3), in which the parameters have the following values: $\Delta = 0.15$ eV, $W_f = 5 \times 10^2$ s^{-1}, $W_d = 5 \times 10^5$ s^{-1} and $W_{df} = 7 \times 10^5$ s^{-1}. It is seen that the experimental data are quite well accounted for. This confirms the assumption that the degeneracies of the 4f^7 and 4f^6 5d emission levels do not differ much.

When the Ba^{2+} ions in LiBaAlF$_6$:Eu^{2+} are replaced by Sr^{2+} ions, line emission is no longer observed. The same holds when the Ba^{2+} ions are replaced by Ca^{2+} ions. It is interesting to compare these results with data we obtained earlier on some Eu^{2+}-activated phosphors [3,13,14]). In table I we have summarized the occurrence of line emission and the value of Δ for a number of Sr and Ba compounds activated with Eu^{2+}. The compounds SrFCl:Eu^{2+}, BaFCl:Eu^{2+}, SrBe$_2$Si$_2$O$_7$:Eu^{2+} and BaBe$_2$Si$_2$O$_7$:Eu^{2+} all show line emission. The value of Δ is higher for the Sr than for the corresponding Ba compounds. For SrAl$_{12}$O$_{19}$:Eu^{2+} and BaAl$_{12}$O$_{19}$:Eu^{2+} the Sr compound shows line emission whereas the Ba compound does not. This is attributable to the relatively strong crystal field in the case of BaAl$_{12}$O$_{19}$:Eu^{2+} (ref. 17), so that the lowest 4f^6 5d level lies at a too low energy, i.e. below the ^{6}P$_{7/2}$ level. Considering the foregoing three types of lattices, we see that Eu^{2+} line emission

TABLE I

Occurrence of line emission, and the energy difference Δ between the lowest $4f^6$ 5d level and the $^6P_{7/2}$ level of Eu^{2+} for some Eu^{2+}-activated Sr and Ba compounds

lattice	line emission	Δ (eV)	ref.
SrFCl	+	0.06	13
BaFCl	+	0.05	13
$SrBe_2Si_2O_7$	+	0.15	3
$BaBe_2Si_2O_7$	+	0.09	3
$SrAl_{12}O_{19}$	+	0.06	14
$BaAl_{12}O_{19}$	—	a	14
$LiSrAlF_6$	—	a	b
$LiBaAlF_6$	+	0.15	b

a: Lowest $4f^6$ 5d level below the $^6P_{7/2}$ level.
b: This work.

is more favoured in Sr than in Ba compounds. The present results for $LiBaAlF_6$:Eu^{2+} and $LiSrAlF_6$:Eu^{2+} show a different behaviour. This can be explained by considering the crystal structure of the two compounds. In $LiBaAlF_6$:Eu^{2+}, which has the same crystal structure as $LiBaCrF_6$ (ref. 18), the Ba^{2+} and Eu^{2+} ions are surrounded by twelve F^- ions at a relatively large distance. In $LiSrAlF_6$:Eu^{2+} on the other hand, the Sr^{2+} and Eu^{2+} ions are surrounded by six F^- ions at a relatively short distance (ref. 19). Hence the crystal field at the Eu^{2+} ions is much stronger in $LiSrAlF_6$:Eu^{2+} than in $LiBaAlF_6$:Eu^{2+}. As has been pointed out earlier [3]), Eu^{2+} ions in a strong crystal field cannot show line emission since the lowest $4f^6$ 5d level of these ions is situated below the $^6P_{7/2}$ level. Apparently this situation exists in the case of $LiSrAlF_6$:Eu^{2+} In the isostructural compound $LiCaAlF_6$:Eu^{2+} no line emission is found either. The crystal field at the Eu^{2+} ions is even stronger since the Ca^{2+} ions are smaller than the Sr^{2+} ions, resulting in a shorter Eu^{2+}–F^--distance.

Acknowledgement

The authors are indebted to J. G. Verlijsdonk for preparing the samples and putting them at our disposal and to Dr A. T. Vink for valuable discussions.

Philips Research Laboratories *Eindhoven, May 1978*

REFERENCES

[1] G. Blasse and A. Bril, Philips Tech. Rev. **31**, 304, 1970.
[2] G. Blasse, Structure and Bonding **26**, 43, 1976.
[3] J. M. P. J. Verstegen and J. L. Sommerdijk, J. Luminescence **9**, 297, 1974.
[4] N. S. Al'tshuler, S. L. Korableva, L. D. Livanova and A. L. Stolov, Soviet Phys. Solid State **15**, 2155, 1974.
[5] B. Tanguy, P. Merle, P. Pezat and C. Fouassier, Mat.·Res. Bull. **9**, 831, 1974.
[6] J. L. Sommerdijk, J. M. P. J. Verstegen and A. Bril, J. Luminescence **10**, 411, 1975.
[7] S. N. Bodrug, E. G. Valyashko, V. N. Mednikova, D. T. Sviridov and R. K. Sviridov, Opt. Spectr. **34**, 176, 1973.
[8] N. S. Al'tshuler, L. D. Livanova and A. L. Stolov, Opt. Spectr. **36**, 72, 1974.
[9] J. L. Sommerdijk and A. Bril, J. Luminescence **11**, 363, 1976.
[10] P. Valon, J. C. Cousseins, A. Védrine, J. C. Gâcon, G. Boulon and F. K. Fong, Mat. Res. Bull. **11**, 43, 1976.
[11] C. Fouassier, B. Latourrette, J. Portier and P. Hagenmuller, Mat. Res. Bull. **11**, 933, 1976.
[12] J. G. Verlijsdonk, private communication.
[13] J. L. Sommerdijk, J. M. P. J. Verstegen and A. Bril, J. Luminescence **8**, 502, 1974.
[14] J. M. P. J. Verstegen, J. L. Sommerdijk and A. Bril, J. Luminescence **9**, 420, 1974.
[15] A. Bril, G. Blasse and J. A. de Poorter, J. Electrochem. Soc. **117**, 346, 1970.
[16] B. di Bartolo, Optical interactions in solids, Wiley, New York, 1968, ch. 18.
[17] A. L. N. Stevels and A. D. M. Schrama-de Pauw, J. Electrochem. Soc. **123**, 691, 1976.
[18] W. Viebahn and D. Babel, Z. Anorg. Allg. Chem. **406**, 38, 1975.
[19] W. Viebahn, Z. Anorg. Allg. Chem. **386**, 355, 1971.

ON THE VACUUM-ULTRAVIOLET EXCITATION SPECTRA AND QUANTUM EFFICIENCIES OF LUMINESCENT POWDERS, MAINLY RARE-EARTH ACTIVATED BORATES

by A. W. VEENIS and A. BRIL

Abstract

Quantum efficiencies of phosphors in the vacuum-ultraviolet region down to 110 nm have been determined. Special attention has been paid to materials which can be used as the red component for gas-discharge display panels. The orange-red luminescent Y and Gd borates activated with Eu^{3+} and Eu^{3+} – Bi^{3+} show high quantum efficiencies in the excitation region between 160 and 170 nm.

1. Introduction

In gas-discharge display panels [1,2,3,4]) the luminescent phosphor layers will be excited by vacuum-ultraviolet (VUV) radiation with wavelengths of 150–170 nm from Xe, and Xe, Kr discharges. From the many properties and requirements of the phosphor layer, e.g. light output, decay time, colour rendering, easy applicability, we considered the light output of the emitted radiation under VUV excitation down to a wavelength of 110 nm.

Suitable blue and green luminescent phosphors for this purpose are already known, e.g. (Ca,Mg)SiO_3–Ti or Eu^{2+}-activated phosphors (blue) and Zn_2SiO_4–Mn or $CeMgAl_{11}O_{19}$–Tb (green), see fig. 1 and refs 5, 6, 7 and 8; up to now there is no satisfactory red one. In this paper special attention will be

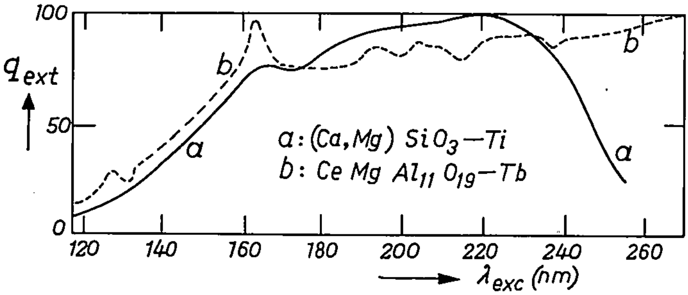


Fig. 1. Excitation spectrum of the blue luminescent (Ca, Mg) SiO_3–Ti, (quantum efficiency $q = 75\%$ at $\lambda_{exc} = 220$ nm) and the green luminescent $CeMgAl_{11}O_{19}$ – Tb (quantum efficiency $q = 80\%$ at $\lambda_{exc} = 260$ nm); q_{ext} denotes the external quantum efficiency (quantum output) arbitrarily fixed at 100 for the maximum of each curve.

paid to phosphors of the type MBO_3 (M = Gd, Sc or Y), activated with Eu^{3+} as potential red components. We found that many Eu^{3+}- and Tb^{3+}- activated borates of the MBO_3 composition have a reasonably high light output when their luminescence is excited with radiation of λ = 150–170 nm. The properties of $GdBO_3$–Eu^{3+} in the middle UV region (200–300 nm) were discussed earlier [9]). In ref. 10 special attention was paid to the energy transfer in YBO_3–Bi^{3+}, Eu^{3+}.

Finally we give here the definition of some of the quantities used in the paper.

— The external quantum efficiency q_{ext} is the ratio of the number of photons emitted N_{em} to the number of photons N_{exc} impinging on the luminescent material: $q_{ext} = N_{em}/N_{exc}$.

— The light output L, as far as used in this and in our previous papers, refers to the relative emitted luminescent flux expressed in quanta. It is therefore proportional to q_{ext} for constant excitation density as a function of wavelength.

— The quantum efficiency q is defined as the ratio of the number of emitted quanta N_{em} to the number of absorbed quanta N_{abs}: $q = N_{em}/N_{abs}$. This means that the quantum efficiency q is obtained by correcting the external quantum efficiency for the absorption factor α, where $\alpha = N_{abs}/N_{exc}$. Thus we find $q = \alpha q_{ext}$.

— The absorption factor of the powder materials is found from the diffuse reflectance r: $\alpha = 1 - r$.

— The excitation spectrum is defined as the luminescence efficiency as a function of wavelength of the exciting radiation. In this paper all excitation spectra given are external excitation spectra, i.e. the spectra refer to the external quantum efficiency.

The definitions given here are in accordance with the concepts of the C.I.E. committee (Commission Internationale de l'Eclairage) dealing with these matters.

2. Instrumental

Two vacuum-ultraviolet instruments were at out disposal.

(a) We started measurements for the wavelength region 160 nm $< \lambda <$ 300 nm using the McPherson 235 monochromator with a Seya–Namioka concave grating mounting and a 50 cm focal length (see fig. 2*a* and ref. 11, p. 69). The focussing of the entrance and exit slits in this instrument is optimal at all wavelengths when the grating is rotated, because of the angle of 70° 15′ between the optical axes connecting the slits and the centre of the grating. The light source is a Cathodeon D_2-lamp with suprasil window. The space between window and entrance slit of the monochromator is argon-flushed. The mono-

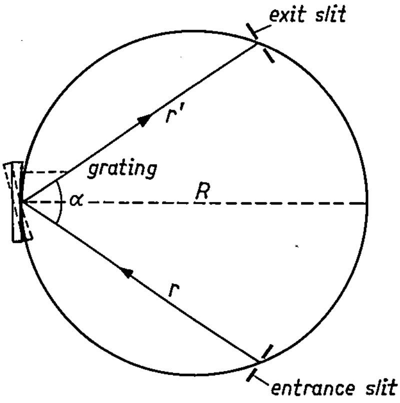


Fig. 2*a*. Schematic diagram of the McPherson 235 Seya–Namioka monochromator. Focal distance 50 cm; $\alpha = 70° 15'$. The grating rotates with centre G see fig. 2*b*; $r = r' = 40.8$ cm.

chromator itself and the sample chamber are evacuated to about 10^{-5} Torr by a fast turbopumping system. Normal oil pumps cannot be used owing to possible back diffusion of oil-vapour, which contaminates the grating (the oil layer absorbs the VUV radiation). The (visible) luminescence of the phosphors is detected as a function of the exciting wavelength, by an EMI 9558, S20-cathode, photomultiplier, which is cooled to about −30 °C. Noise is then reduced to a level of about 10^{-11} A. As the signal will be about 10^{-9} A, dc amplification is used with the aid of a Keithley 610 C amplifier. To remove the influence of the spectral characteristics of the lamp and the monochromator transmission, the spectra of the sample and the sodium salicylate are both determined. Sodium salicylate is assumed to have a constant quantum efficiency and absorption over the whole UV and VUV excitation range (see ref. 11, p. 216 and ref. 12). By dividing the sample spectrum values by the corresponding ones of sodiumsalicylate, we obtain the relative quantum output of the sample, i.e. the external quantum efficiency as a function of the exciting wavelength.

(b) The second instrument available was a McPherson 225, 1 m normal-incidence monochromator. It is not evacuated, but He-flushed (see fig. 2*b* and ref. 11, p. 64).

The throughput of this instrument is higher than that of the Seya–Namioka type. Because of the nearly normal incidence there is much less astigmatism. By mechanical means care is taken that entrance slit, exit slit and centre of the concave grating remain on the Rowland circle to which the grating is tangent. In this way focussing is maintained. The light source here is a home-made Hunterlamp (for a description see ref. 11, pp 237–239), slowly flushed with hydrogen (pressure ≅ 1 Torr) so that the H_2-line spectrum and its con-

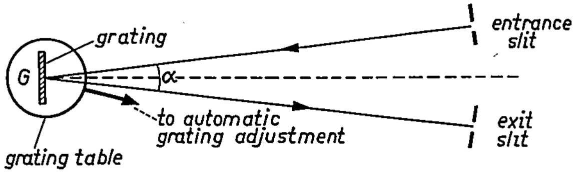


Fig. 2*b*. Schematic diagram of the McPherson 225 normal-incidence monochromator. Focal distance 1 m; $\alpha = 15°$. The grating is positioned for every λ with the aid of an empirically determined cam.

tinuum can be used down to $\lambda = 110$ nm, the cut-off of the LiF windows. A differential pumping system is available for the case that we want to use shorter wavelengths *).

At the exit slit an oscillating mirror directs the beam to a pair of photomultipliers (Philips 150 AVP with an S11 photocathode). The photomultiplier in the reference beam is coated with sodium-salicylate. In the sample beam the phosphor to be measured is placed in a thin layer in front of the photomultiplier. With the aid of a logarithmic amplifier (McPherson type 782) the sample spectrum data are divided by those of the sodium-salicylate spectrum of the reference channel. In this way the relative excitation spectrum appears directly on the recorder in a logartihmic scale. (When transmissions are measured, the absorbance is thus directly plotted as a function of wavelengths). To improve the sensitivity of the system, especially for red phosphors, we can replace the sample photomultiplier by one with an S20 photocathode.

When the relative external excitation spectra of the quantum efficiencies have been obtained in the way described, the question which then arises is how to convert them into absolute data.

This can generally be carried out by determining the absolute value of the quantum efficiency with excitation in one of the following regions.

(1) Excitation in the λ = 250–270 nm region or at the λ = 254 nm mercury vapour discharge line.

(2) Excitation at λ = 229 nm with the aid of a Cd vapour discharge. These techniques have been described earlier [7,12,13]).

Reflectances for $\lambda < 200$ nm were not measured. For quantum efficiency calculations in the cases considered here we assumed a value of 10%.

*) The VUV facilities described here were set up initially by J. H. Haanstra and A. J. J. van Dijsseldonk of our laboratories. The Hunter-lamp was made after an original design of ESTEC in Noordwijk (Netherlands). The He-flushing system and some of the accessories were designed and constructed in cooperation with our Vacuum Department under J. W. van Loenhout.

3. Results and discussion

Comparing the orange-red luminescent Eu^{3+}- and the green luminescent Tb^{3+}-activated Y, Gd and Sc borates, we see that they all show a peak in the excitation spectrum near $\lambda = 160$ nm (see figs 3 and 6), the top of the peak only differing slightly. The Eu^{3+}-activated borates in particular show high quantum efficiencies (see table I).

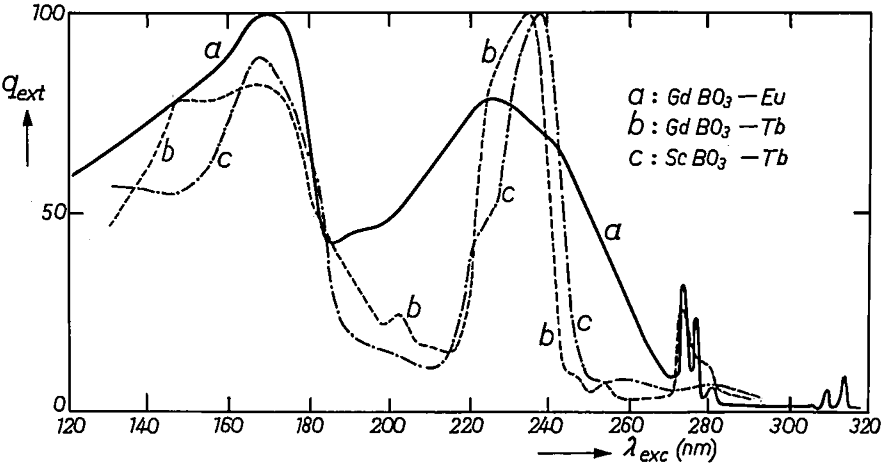


Fig. 3. Excitation spectra of $GdBO_3$–Eu(*a*), $GdBO_3$–Tb(*b*) and $ScBO_3$–Tb(*c*). See also subscript of fig. 1 and table I.

Because peaks near $\lambda = 160$ nm are present both in Eu^{3+}- and in Tb^{3+}-activated borates, the absorption in that region is most probably due to absorption in the host lattice, i.e. in the BO_3 complex.

With Bi as a coactivator in, for example, YBO_3–Eu and $GdBO_3$–Eu the excitation spectra of the quantum efficiencies are much flatter, no minimum appears at $\lambda_{exc} = 190$ nm and 260 nm (see fig. 4). Bi^{3+}–Eu^{3+} phosphors show additional absorption in that region, leading to higher external quantum efficiency (see fig. 5).

The excitation spectrum of $CeBO_3$–Tb^{3+} (fig. 4) is different from that of the other borates. The absorption is principally due to the Ce^{3+} ions, because the excitation spectrum is resembling that of unactivated $CeBO_3$. The latter is nearly flat between the wavelengths $\lambda_{exc} = 250$ nm and $\lambda_{exc} = 150$ nm (no minimum near $\lambda_{exc} = 210$ nm).

Previously [14,15]) we have discussed that the RE luminescence can originate from absorption as follows.

(1) In the 4f levels of the activator ions (absorption for $\lambda > 260$ nm). Examples of these narrow absorptions can be found for instance in ref. 9.

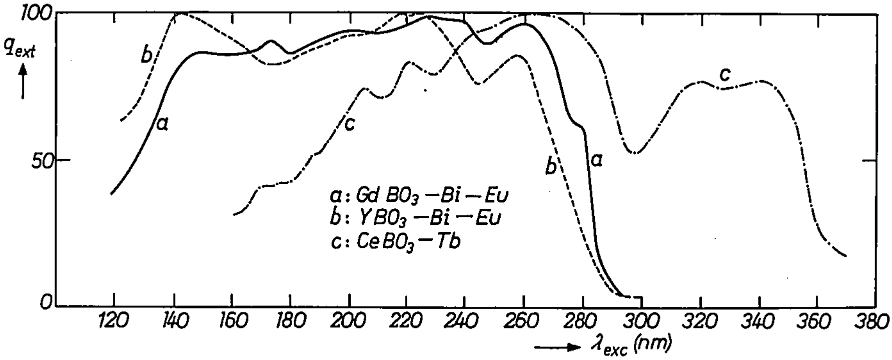


Fig. 4. Excitation spectra of $GdBO_3$-Bi-Eu (*a*), YBO_3-Bi-Eu (*b*) and $CeBO_3$-Tb (*c*). See also caption of fig. 1 and table I.

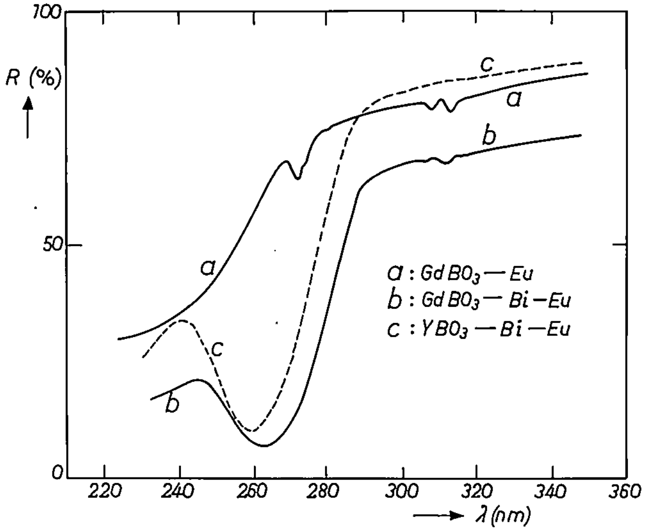


Fig. 5. Reflectance spectra of $GdBO_3$-Eu (*a*), $GdBO_3$-Bi-Eu (*b*) and YBO_3-Bi-Eu (*c*).

(2) In the 4f–5d levels of the activator ions. Because now the 5d level is involved, the absorption due to these levels is more dependent on the host lattices. For Tb^{3+} they are generally located between $\lambda \cong 180$ nm and 250 nm and are relatively narrow, but broader than those due to the 4f levels (see figs 3 and 6).

(3) In charge-transfer levels, due to mixing of the 4f with ligand wave functions. For the Eu^{3+}–O^{2-} state the absorption is near $\lambda = 250$ nm; for Tb^{3+} it is near 160 nm. When we compare YBO_3–Eu^{3+} with $GdBO_3$–Tb^{3+} and $ScBO_3$–Tb^{3+} given in fig. 3, we see in the Eu^{3+} phosphor a rather broad

excitation band with maximum at $\lambda_{exc} \cong 230$ nm, while in the Tb phosphors the bands in that region are much narrower. This can be explained by assuming that in the case of Eu^{3+} a charge-transfer band is involved in the excitation process, whereas the band in the Tb^{3+} phosphor is due to 4f–5d transitions.

(4) In the host lattice, e.g. CeF_3–Tb^{3+}. In this phosphor the absorption is determined mainly by the Ce^{3+} ion and not by the activator.

The fact that the efficiencies of the Y, Gd and Sc borates with Eu^{3+} and Tb^{3+} are generally high, means that there is an efficient transfer through the host lattice, where the exciting energy is absorbed, to the activator centre. This is in agreement with the reasonably high cathode-ray (CR) radiant efficiency of e.g. YBO_3–Tb $ScBO_3$–Tb and $ScBO_3$–Eu. With CR excitation most of the primary electrons are absorbed far from the centres, so that a good energy transfer is necessary. To obtain high efficiencies it is moreover necessary that hardly any radiationless processes should occur in the centres. For $GdBO_3$ and $ScBO_3$, for instance, this can probably be understood because of the fact that they possess or nearly possess a centre of symmetry. In that case there is not much interaction with the host lattice and as a consequence there is little chance that the excitation energy will be lost by radiationless processes.

In the excitation spectra of the Gd phosphors the transfer of energy absorbed in Gd^{3+} to the Eu^{3+} or Tb^{3+} centres can be observed: Eu^{3+} and Tb^{3+} luminescence is found near $\lambda_{exc} = 311$ nm, $\lambda_{exc} = 270$ nm and $\lambda_{exc} = 250$ nm due to absorption in the ^6I and ^6D levels of Gd^{3+} (see ref. 16 and figs 3, 4 and 5). For YBO_3–Eu^{3+}–Bi^{3+} the quantum efficiency at $\lambda_{exc} \cong 260$ nm is found to be $q = 85\%$ (diffuse reflectance $r = 16\%$). We also carried out measurements for $\lambda_{exc} = 229$ nm, giving $q = 95\%$ ($r = 26\%$).

With the aid of the relative excitation spectrum of the light output (external quantum efficiency) we can derive the efficiency at wavelengths between 150 and 170 nm (fig. 4). When we assume that the reflectance is 10% in that region we find at $\lambda_{exc} = 160$ nm a quantum efficiency of 70%. The other Eu^{3+}-activated phosphors have also high efficiencies. There are thus excellent phosphors for excitation in the region $\lambda_{exc} = 150$–170 nm as far as efficiency and external quantum efficiency is concerned.

The results of the various phosphors have been summarized in table I. For a discussion of the accuracy see ref. 7.

The absolute efficiencies for a phosphor like $YAl_3(BO_3)_4$–Tb^{3+} have to be determined in a different way. The relative excitation spectrum is given in fig. 6. It can be seen that the efficiencies at the wavelength $\lambda_{exc} = 254$ nm and $\lambda_{exc} = 229$ nm are very low. In this case it is therefore better to choose a wavelength near $\lambda_{exc} = 160$ nm to determine the absolute efficiency; this can be

TABLE I

Absolute efficiencies

phosphor	λ_{exc} = 250–270 nm				λ_{exc} = 229 nm				maximum between λ_{exc} = 150 nm and λ_{exc} = 170 nm	
	r	η	q	q_{ext}	r	η	q	q_{ext}	q_{ext}	q^*
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
YBO_3–Bi–Eu	16	35	85	75	26	35	95	70	70	75
$GdBO_3$–Eu	66	25	60	20	32	25	70	50	65	70
$GdBO_3$–Bi–Eu	8	35	80	75	16	30	85	70	65	70
$ScBO_3$–Tb					28	20	45	35	25	30
$GdBO_3$–Tb					26	15	40	30		
$CeBO_3$–1 % Tb	5	17	40						15	15

r = reflection
η = radiant efficiency
q = quantum efficiency
q_{ext} = external quantum efficiency
q^* = quantum efficiency, assuming a reflection of r = 10%.

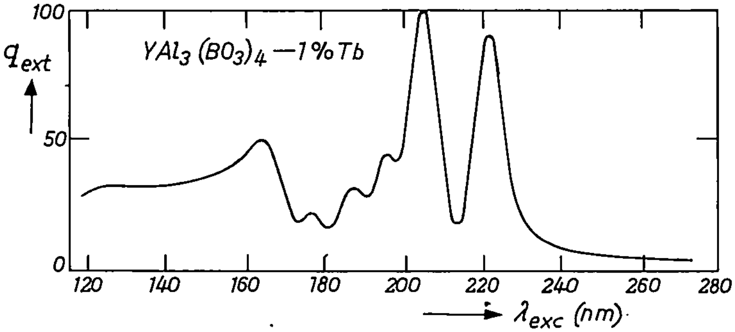


Fig. 6. Excitation spectrum of $YAl_3(BO_3)_4$–Tb. See also the caption of fig. 1 and table I.

carried out with the aid of sodium salicylate as a reference [12]).

Measurements of excitation spectra with VUV radiation have also been carried out by Koïke, Kojima, Toyonaga, Kagami and Nagatani [2]) for a number of phosphors with a view to their applicability in gas-discharge panels. Their results on $GdBO_3$–Eu and $GdBO_3$–Tb are in good agreement with ours. They find an absolute efficiency of $GdBO_3$–Eu^{3+} which is 40% higher than for Y_2O_3–Eu^{3+}. We find a gain of even 65%.

Excitation spectra measurements of Eu- and Eu–Bi-activated $GdBO_3$ and YBO_3 have been performed by G. Tsujimoto [3]. The difference between the singly activated Eu^{3+} and doubly-activated Eu^{3+}–Bi^{3+} phosphors is in agreement with our measurement. The quantum efficiency of the Eu phosphor is high near 160 and 230 nm, and very low at about 190 nm, the quantum efficiency of the Eu^{3+}–Bi^{3+} phosphor is high and nearly constant between 160 and 230 nm.

Acknowledgement

We are much indebted to W. L. Konijnendijk and to J. M. P. J. Verstegen of our Lighting Division for preparing the Eu- and Bi–Eu-activated Y and Gd borates.

Further our thanks are due to R. B. Testelmans-Engelen for carrying out the absolute efficiency measurements at the 260 nm and 229 nm excitation.

Philips Research Laboratories *Eindhoven, July 1978*

REFERENCES

[1] C. J. Gerritsma and G. H. F. de Vries, private communication.
[2] J. Koïke, T. Kojima, T. Toyonaga, H. Takahashi, T. Kagami and T. Nagatani, 161 Meeting of Investigations on Luminescent Substances, May, 1976.
[3] Y. Tsujimoto, Japanese patent application No. 49-141193, 1974.
[4] M. Fukushima, S. Marayama, T. Kaji and S. Mikoshiba, IEEE Trans. El. Dev. **ED-22**, 1975.
[5] J. D. Kingsley and G. W. Ludwig, J. Electrochem. Soc. **117**, 353, 1970.
[6] A. W. Veenis and A. Bril, J. Electrochem. Soc. **123**, 396, 1976.
[7] A. Bril and W. Hoekstra, Philips Res. Repts **19**, 296, 1964.
[8] A. L. W. Stevels, P. Vries, A. W. Veenis and A. T. Vink, The Electrochemical Society Spring Meeting 1978, Seattle, Extended abstracts, Paper No. 337.
[9] A. Bril and W. L. Wanmaker, J. Electrochem. Soc. **111**, 1363, 1964.
[10] G. Blasse and A. Bril, J. Chem. Phys. **47**, 1920, 1967.
[11] J. A. R. Samson, Techniques of vacuum ultraviolet spectroscopy, J. Wiley and Sons, New York, 1967.
[12] A. Bril and A. W. Veenis, J. Res. Nat. Bur. Stand. **80A**, 401, 1976.
[13] G. Blasse and A. Bril, J. Lum. **3**, 109, 1970.
[14] A. W. Veenis and A. Bril, Proc. 5th Internat. Conf. on Vacuum-ultraviolet Radiation Physics, Montpellier, 1977, paper no. 40.
[15] G. Blasse and A. Bril, Philips Res. Repts **22**, 481, 1967.
[16] A. Bril and W. L. Wanmaker, J. Chem. Phys. **43**, 2559, 1965.

THERMOLUMINESCENCE OF UV IRRADIATED CsI:Na

by A. L. N. STEVELS

Abstract

Thermoluminescence spectra of CsI:Na irradiated at 77 K show a peak at 124–127 K which is linked to Na-related luminescence centres. Peaks at 157–162 K, 184–193 K and 220–226 K have no connection with the presence of Na.

1. Introduction; initial experiments

As a part of our studies on the luminescence of CsI:Na [1,2]), we have investigated the thermoluminescence of CsI:Na samples which were irradiated for 20 min at 77 K by a D_2 lamp. The spectrum of this light source extends from about 3 to 8 eV, a large part of the emitted light being concentrated at energies above 5.8 eV, which is the band gap energy of CsI.

The thermoluminescence — mainly blue — was detected by a Philips UVP type photomultiplier; the heating rate after irradiation was 5 °C/min.

Three types of samples were studied: powder samples prepared in the way described in ref. 2 and evaporated layers deposited either on a relatively cold ($\cong$ 50 °C) or a relatively warm ($\cong$ 300 °C) substrate. The evaporation procedure has been described earlier [1]). Before the measurements each sample was heated at 400 °C in a stream of N_2 in order to enhance the blue luminescence of the Na-related centres, see ref. 2.

Figure 1 shows the shape of the thermoluminescence spectra of CsI doped with 0.04 at % Na. The curves have been normalized by taking the strong peak at 124–127 K equal to 100. On an absolute scale the measured intensities of these peaks shown are comparable. Due to different path geometries of the generated light, however, no quantitative comparison is possible.

Like the intensity of the blue luminescence of CsI:Na [2]), the intensity of the thermoluminescence peak at 124–127 K decreases after storage of freshly heated samples in air. The glow peaks at higher temperature (see below) have intensities which were observed not to be influenced by the ageing process. This finding suggests that the thermoluminescence peak at 124–127 K is in some way linked with Na-related luminescence centres. We will now discuss two experiments which give further support for this conjecture.

2. Relation of the 124–127 K peak with the Na content of the sample

In figure 2 plots are given of the measured peak area of the 124–127 K ther-

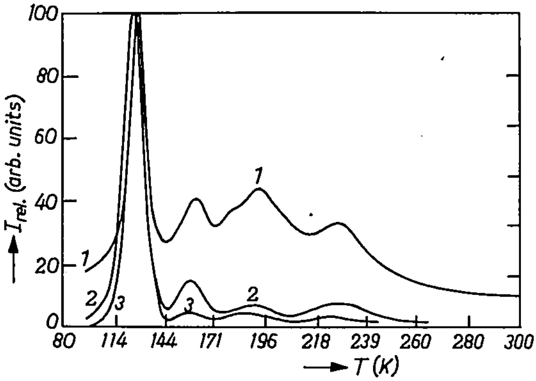


Fig. 1. Normalized thermoluminescence spectra of CsI:0.04 at % Na samples. Irradiation: 20 min D_2 lamp at 77 K.
Curve 1: powder sample, curve 2: vapour-deposited layer (substrate temperature ca 50 °C), curve 3: vapour-deposited layer (substrate temperature ca 300 °C).

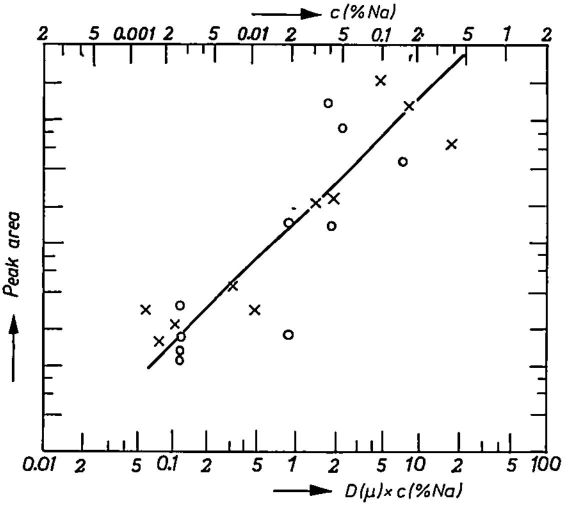


Fig. 2. Area of 124–127 K glow peak vs sodium concentration (% Na, open circles) or vs total sodium content [$D(\mu) \times$ % Na, crosses] for vapour-deposited CsI:Na layers (substrate temperature ca 300 °C). Irradiation: 20 min D_2 lamp at 77 K.

moluminescence. Since there is appreciable host-lattice absorption above 5.8 eV [2]), the penetration depth of the radiation giving rise to the thermoluminescence is not known. Therefore we have taken two parameters as abcissae: the Na concentration (the case relevant for strong host-lattice absorption, data represented by open circles) and the total Na content of the sample (the case of weak host-lattice absorption, data represented by crosses).

In both cases a linear relationship with the peak area is suggested to exist. In figure 2 data are presented for samples which were deposited on a relatively warm substrate. A similar result was also found for cold-deposited layers. The strength of the UV absorption band of the blue-emitting centres in CsI:Na is linearly related to the Na content [2]). Since at 300 K the absorbing and the emitting state of the blue centres are not identical [2]), there might be a relationship between the absorbing state at 300 K and the thermoluminescent trap discussed here.

In reference 2 it was also reported that when CsI:Na phosphors are heated in iodine vapour the blue luminescence of the Na-related centres is quenched. How such a treatment affects thermoluminescence is shown in fig. 3. In a CsI:0.04% Na cold-deposited layer the 124–127 K glow peak is reduced to 40% of its original height when the sample is heated in N_2 containing 20 Torr I_2 vapour at 400 °C.

Panova and Shiran [3]) have reported a strong thermoluminescence peak at 134 K in CsI:Na irradiated either with UV or with X-rays. This band was said to increase linearly with increasing Na concentration in the crystals and was suggested to be due to the release of electrons from trapping levels like Na^0 centres. This peak seems to be the same one as observed by us at 124–127 K. Since no experimental details were given in ref. 3, we cannot offer a definite explanation for the difference in temperature. Probably it is caused by different heating rates. As regards the assignment of the thermoluminescence centre as being a Na^0 centre, we note that Monnier [4]) has recently calculated that the

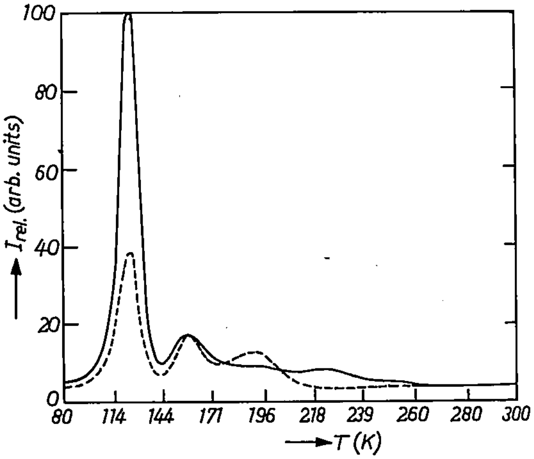


Fig. 3. Thermoluminescence spectrum of a vapour-deposited (substrate temperature ca 50 °C) CsI layer containing 0.04 at % Na, drawn line. The dashed line represents the thermoluminescence after heating the sample in 20 Torr I_2 (+N_2) at 400 °C. Irradiation: 20 min D_2 lamp at 77 K.

binding energy for an electron trapped on a substitutional Na ion in CsI is 0.45 eV. Below we will derive for the trap depth an energy of 0.23 eV, which is lower by a factor of about two.

3. Decay; trap depth

In connection with the foregoing it seems useful to mention that measurements by Van der Does de Bye and by Bril [5]) showed that the decay of the blue Na-related luminescence has three characteristic temperature regions. Above about 160 K approximately exponential decays were observed with a strongly temperature-dependent decay time. Below about 130 K the luminescence decay is basically non-exponential. Between about 130 K and 160 K a mixed decay characteristic has been observed. It is therefore very likely that the luminescence mechanism at room temperature is at least partly different from that at, for instance, 77 K. One of the possible intermediate states at low temperatures seems to be inactivated thermally above about 124–127 K.

The trap depth ΔE of the thermoluminescence was determined by methods reviewed by Kivits and Hagebeuk [6]). We use below their notation. As a function of the retrapping ratio δ (= ratio of the capture cross-section of the traps and that of the luminescent centres), ΔE has been calculated from the peak temperature $T_{\text{th,max}}$ (= 125 K) and the peak halfwidth ω (= 14 K). The part of ω below $T_{\text{th,max}}$ (= λ) has been taken to be 7.5 K, that above $T_{\text{th,max}}$ (= σ) = 6.5 K.

Assuming that $\delta = 0.01$, ΔE is found to be 0.23 ± 0.01 eV for the methods analysed [6]) to be the reliable ones. For the less reliable methods, deviations in ΔE were found which were in good agreement with those predicted by Kivits and Hagebeuk. If is assumed to be equal to 1, no equivocal figures for ΔE were obtained even for those calculation methods which were said in ref. 6 to be correct approaches. In last-named cases ΔE varied between 0.28 and 0.35 eV. For higher values of δ, the scatter in the ΔE figures is even larger.

Above-mentioned results suggest that most probably $\Delta E \simeq 0.23$ eV and that the capture cross-section of the Na-related thermoluminescent centre in CsI:Na is low compared with that of the luminescent centres. As already mentioned, its trap depth of about 0.23 eV is much smaller than the calculated binding energy of an electron trapped on a substitutional Na ion in CsI. Hsu and Bates [7]) calculated that the energy difference between excitons bound to F centres (β excitons) and free excitons is 0.2–0.4 eV depending on the exact configuration. The detrapping energy for excitons bound near anion vacancies was found to be less than 0.15 eV. Our trap depth of about 0.23 eV would therefore suggest that the thermoluminescence involves the detrapping of excitons bound to F centres. Since such centres are not localized near Na, the relation between peak area of the thermoluminescence and the Na content of the sample has to be an indirect one. This is not unreasonable since the traps involve iodine vacancies which in turn arise on adding NaI to CsI, see ref. 2.

4. Thermoluminescence above 150 K

Let us now say a few words on the thermoluminescence peaks in CsI:Na observed above 150 K. In figure 1 we see that these occur at 157–162 K, 184–193 K and 220–226 K. Curve 1 shows, additionally, a shoulder around 180 K. None of these peaks was found to be susceptible to ageing so that it is unlikely that any of them is connected with the blue Na-related centres. Panova and Shiran [3]) report that glow peaks at 163 K, 185 K and 215 K were also found in unactivated CsI and that, both in Na-activated and unactivated CsI, the emitted light has a spectral distribution with strong peaks between 470 and 560 nm. It might be that the three above-mentioned thermoluminescence peaks are associated with the, Na-independent, yellow-emitting centres in CsI (see ref. 2). The presence of these centres is dependent on the starting material and the preparation conditions; it may be that surface states are involved [2]). The present thermoluminescence experiments give additional qualitative evidence for the latter correlation in so far as for powders the glow peaks above 150 K are relatively strong and those for evaporated layers are relatively weak. For warm-deposited layers, which have the best defined and the most regular surfaces [1]), the high temperature glow peaks are the weakest. On the other hand, fig. 3 suggests that after heating in iodine vapour, the 157–162 K peak remains, the 184–193 K peak increases, whereas the 220–226 peak almost disappears, so that in addition to common features, the glow peaks also show differences in behaviour.

5. Samples with very low Na content

Finally we wish to draw attention to the fact that the thermoluminescence spectrum of unintentionally doped CsI, which in our case still contained $5\times10^{-5}\,\%$ Na, shows two additional peaks when compared to the samples containing 0.04 % Na, see fig. 4. In the spectrum of a thick cold-deposited layer, a peak at 87 K and another at 110 K are found to precede the one at 124 K and those above 150 K. These peaks were found to decrease with increasing Na content and to have disappeared in samples containing more than about 0.02 at % Na. The same concentration limit was found for the occurrence of a UV luminescence band in CsI with low Na content. The 87 K peak is also known from the work of Panova and Shiran [3]) and from that of Sidler et al. [8]). In contradistinction to our result, the intensity of the peak was said in ref. 3 to increase with increasing sodium content. Sidler et al. [8]) give no data on a Na

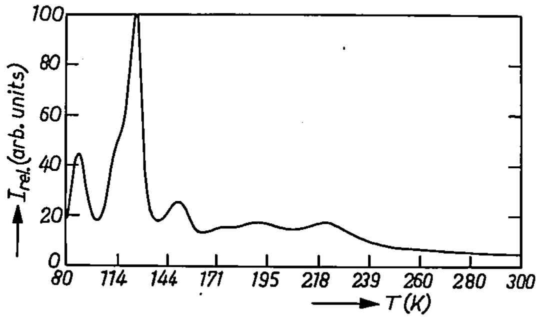


Fig. 4. Thermoluminescence spectrum of a vapour-deposited (substrate temperature ca 50 °C) CsI layer without intentional Na dope. Irradiation: 20 min D_2 lamp at 77 K.

dependence but present evidence that the band is due to migration of V_k centres.

Acknowledgement

The author wishes to thank Mrs A. D. M. Schrama-de Pauw for her contributions to this investigation.

Philips Research Laboratories *Eindhoven, February 1978*

REFERENCES

[1] A. L. N. Stevels and A. D. M. Schrama- de Pauw, Philips Res. Repts **29**, 340, 1974 and **29**, 353, 1974.
[2] A. L. N. Stevels and A. D. M. Schrama- de Pauw, Philips Res. Repts **31**, 1, 1976.
[3] A. N. Panova and N. V. Shiran, Iz. Akad. Nauk, SSSR Ser. Fiz. **35**, 1348, 1971 or Bull. Acad. Sci. USSR Phys. Ser. **35**, 1232, 1971.
[4] R. Monnier, Solid State Comm. **19**, 681, 1976.
[5] J. A. W. van der Does de Bye, unpublished results and A. Bril, unpublished results.
[6] P. Kivits and H. J. L. Hagebeuk, J. Luminescence **15**, 1, 1977.
[7] O. L. Hsu and C. W. Bates Jr., Phys. Rev. **B15**, 5821, 1977.
[8] T. Sidler, J. P. Pellaux, A. Nouailhat and M. A. Aergerter, Solid State Comm. **13**, 479, 1973.

CAPACITANCES OF CIRCUITS IN A TIGHT SCREENED PAIR OR QUAD

by V. BELEVITCH

Abstract

In a *tight* cable the insulated wires touch each other and touch the screen. The electrostatic capacitances per unit length are computed for circuits (side, phantom . . .) in a single tight pair and a single tight quad. Two values of the relative dielectric constant are considered for the insulation: $\varepsilon = 1$ (air) and $\varepsilon = 2.3$ (polythene). The apparent relative dielectric constant is evaluated in the second case. The screen is perfectly conducting and not internally insulated. It is proved theoretically and verified numerically that, for a vanishing insulation thickness, the apparent relative dielectric constant does not tend to 1 (as might be naively expected), but to $\sqrt{\varepsilon}$.

1. Introduction

We consider the screened pair of fig. 1 and the screened quad of fig. 2. The wires have radius a without insulation and b with insulation. The interaxial distance of a pair is D, and the screen of radius a_0 is not internally insulated. Because of the tight geometry, we have

$$a_0 = D; \qquad b = D/2 \tag{1}$$

for the pair, and

$$a_0 = (1 + 1/\sqrt{2})\, D/2; \qquad b = D/(2\sqrt{2}) \tag{2}$$

for the quad. Two values of the relative dielectric constant are considered for the insulation: $\varepsilon = 1$ (air) and $\varepsilon = 2.3$ (polythene). The parameter

$$\eta = \frac{\varepsilon - 1}{\varepsilon + 1} \tag{3}$$

used in the computation is thus 0 in the homogeneous case and 0.39394 in the non-homogeneous case.

The capacitances are computed by solving numerically the set of linear equations deduced from the boundary conditions on the multipole expansions with undetermined coefficients for the potential. The equations are taken from our previous publications [1,2]) and are briefly recalled in sec. 2. The results are

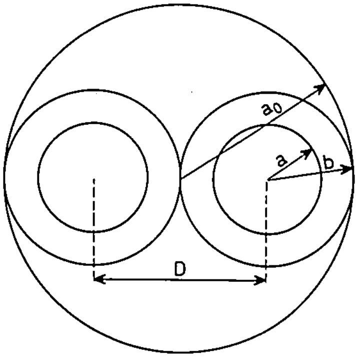


Fig. 1. Cross-section of the tight pair.

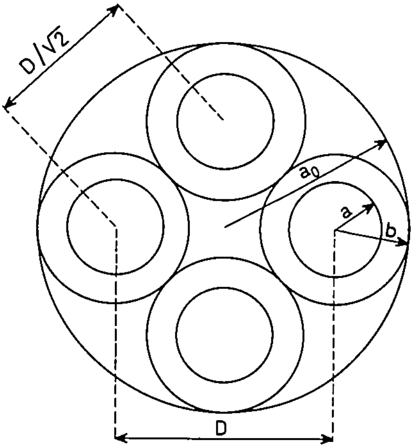


Fig. 2. Cross-section of the tight quad.

described in sec. 3. The case of vanishing insulation thickness, where

$$\frac{b-a}{a} \ll 1 \tag{4}$$

is treated in sec. 4.

2. Theory

We compute the normalized inverse capacitance

$$H = 2\pi\varepsilon_0 K \tag{5}$$

(where $K = 1/C$) as a sum

$$H = A + B + P, \tag{6}$$

where A is the value for thin wires with homogeneous insulation, B the correction for thin wires with non-homogeneous insulation, and P the proximity correction for thick wires. In terms of the normalized variables

$$\delta = \frac{a}{D}; \qquad u = \left(\frac{D}{2a_0}\right)^2; \qquad t = \frac{a}{b} \tag{7}$$

the expressions of A and B are given in table I. The values of A are taken from Kaden [3]) and the value of B is elementary. The value of P corresponds to the last column of table I of ref. 1, where z (called there Z_{prox}) is

$$z = \sum_{n=1}^{\infty} x_n \, y_n \tag{8}$$

and results from the solution of the infinite linear system

$$\sum_{m=1}^{\infty} K_{nm} \, x_m - \frac{\lambda_n \, x_n}{\delta^{2n}} = y_n \tag{9}$$

of ref. 1 with the values of K_{nm} and y_n of table I of ref. 1, except for the side mode of the quad where the linear system has a double size and results from (34). The parameter λ_n is

$$\lambda_n = n \, \frac{1 + \eta t^{2n}}{1 + \eta / t^{2n}} \tag{10}$$

TABLE I

	A	B	P
sym. pair	$2 \ln \dfrac{1 - u}{(1 + u)\,\delta}$	$\dfrac{4\eta}{1 + \eta} \ln t$	$2z$
asym. pair	$\dfrac{1}{2} \ln \dfrac{1 - u^2}{4u\delta}$	$\dfrac{\eta}{1 + \eta} \ln t$	$\dfrac{z}{2}$
side quad	$2 \ln \dfrac{1 - u}{(1 + u)\,\delta}$	$\dfrac{4n}{1 + n} \ln t$	$2z$
pantom quad	$\ln \dfrac{1 - u^2}{2\delta\,(1 + u^2)}$	$\dfrac{2\eta}{1 + \eta} \ln t$	z
asym. quad	$\dfrac{1}{4} \ln \dfrac{1 - u^4}{8\delta u^2}$	$\dfrac{\eta}{2(1 + \eta)} \ln t$	$\dfrac{z}{4}$

in accordance with (17) of ref. 2. By (1) and (2) we have

$$u = \tfrac{1}{4}; \qquad t = 2\delta \tag{11}$$

for the tight pair, and

$$u = (2 - \sqrt{2})^2; \qquad t = 2\delta\sqrt{2} \tag{12}$$

for the tight quad.

3. Results

For the tight pair, the insulation vanishes for $\delta = \frac{1}{2}$. The capacitances were tabulated for δ varying from 0.05 to 0.45 in steps of 0.05. The infinite linear system was truncated to dimensions 15 and 20, and a comparison of the results shows that 4-digit accuracy was obtained. For the homogeneous case, the value of H is shown in fig. 3 and decreases to zero for $\delta = 0.5$ as expected. For the non-homogeneous case, instead of giving the value of H, we have plotted the ratio $H_{\mathrm{hom}}/H_{\mathrm{non\ h.}}$ defining the apparent relative dielectric constant shown on fig. 4. For δ approaching 0.5, the apparent constant tends to $\sqrt{\varepsilon} = 1.5166$, in accordance with the theory of sec. 4.

For the tight quad, the insulation vanishes for $\delta = 1/(2\sqrt{2}) = 0.35355$. The results were tabulated for δ varying from 0.05 to 0.35 in steps of 0.05. The accuracy remains 4 digits up to 0.3 but drops to a few percents at $\delta = 0.35$. The results are fig. 5 and fig. 6.

4. Infinitely thin insulation

For the homogeneous pair without screen the rigorous value of H is

$$H = 2 \operatorname{arccosh} (D/2a). \tag{13}$$

For $D/2a$ tending to 1, hence for vanishing H, it results from the first two terms of the Taylor expansion of cosh that one has

$$H = 2\left(\frac{D}{a} - 2\right)^{\frac{1}{2}}. \tag{14}$$

For the excentric coaxial of fig. 7, the rigorous value of H is [4])

$$H = \operatorname{arccosh} \frac{a^2 + {a_0}^2 - (D/2)^2}{2a\,a_0}. \tag{15}$$

When the conductors almost touch each other, i.e. for

$$a_0 \cong a + D/2$$

one obtains similarly

$$H = \left[2\,\frac{a_0 - a - D/2}{a\,(1 + 2a/D)}\right]^{\frac{1}{2}}. \tag{16}$$

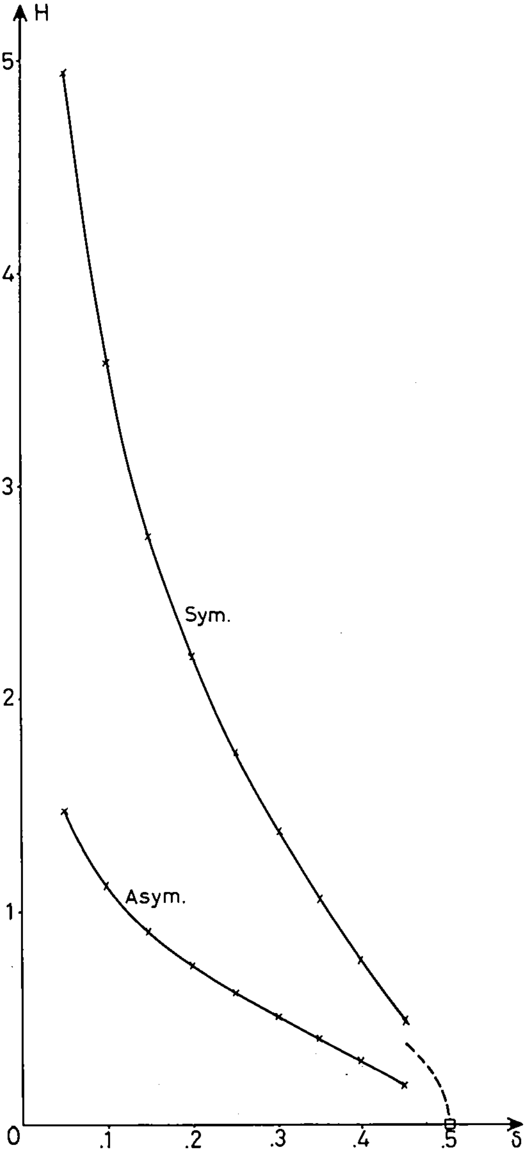


Fig. 3. Normalized inverse capacitances of the homogeneous pair.

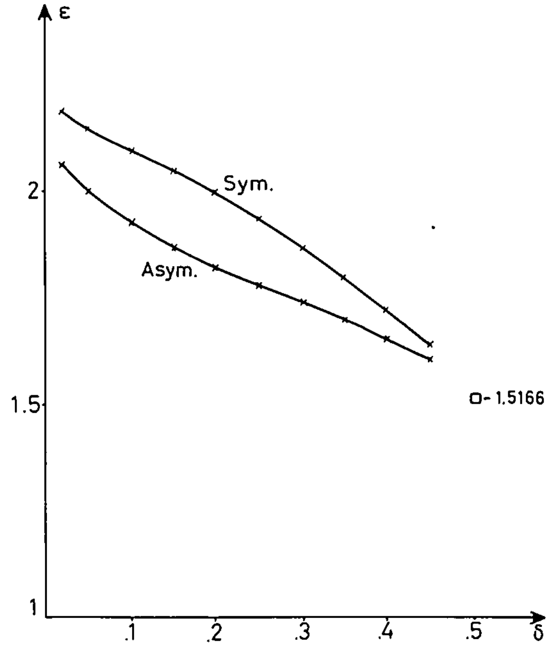


Fig. 4. Apparent relative dielectric constants of the non-homogeneous pair.

For the pair, call C_2 the direct capacitance between wires, and C_1 the direct capacitance between one wire and the screen, so that one has (A = asymmetric, S = side)

$$C_A = 2C_1; \qquad C_S = C_2 + \frac{C_1}{2}. \tag{17}$$

For the inverse values, H_2 is simply (14), whereas H_1 is (16) with $a_0 = D$ and with $a/D = \frac{1}{2}$ in the denominator. One thus has

$$H_2 = 2\left(\frac{D}{a} - 2\right)^{\frac{1}{2}}; \qquad H_1 = \left(\frac{D}{2a} - 1\right)^{\frac{1}{2}} = \frac{H_2}{2\sqrt{2}}. \tag{18}$$

From (17) and (18) it results that, for thin insulation, one has

$$\frac{H_A}{H_S} = \frac{C_S}{C_A} = \frac{C_2}{2C_1} + \frac{1}{4} = \frac{H_1}{2H_2} + \frac{1}{4} = \frac{1}{4}\left(\frac{1}{\sqrt{2}} + 1\right) = 0.4268 \tag{19}$$

with

$$H_S = 0.8284\,(1/\delta - 2)^{\frac{1}{2}}.$$

The corresponding approximation is plotted in dotted line in fig. 3.

For a non-homogeneous case with thin isulation, it results from (22) of ref. 2 that one can treat the problem as homogeneous with $\varepsilon = 1$ provided one replaces the wire radius a by the modified radius

$$a' = \frac{a}{\varepsilon} + \left(1 - \frac{1}{\varepsilon}\right) b. \tag{20}$$

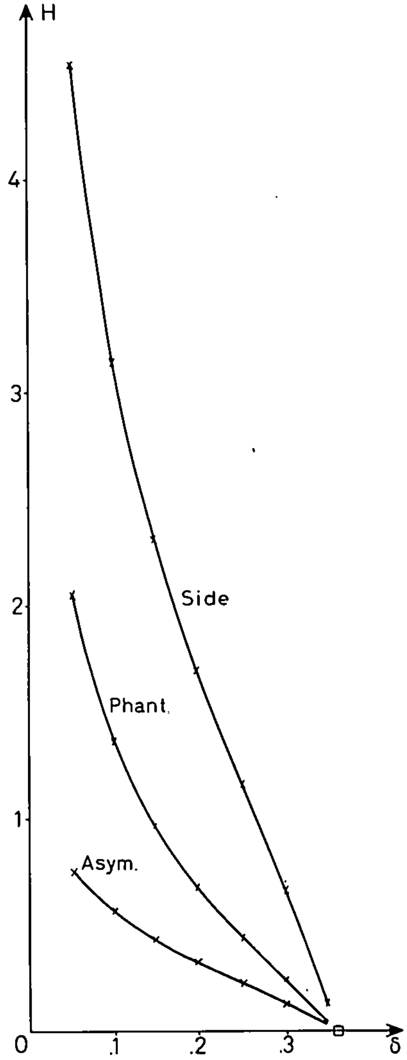


Fig. 5. Normalized inverse capacitances of the homogeneous quad.

For the pair, the expression $(D/2a)$–1 appearing in H_1 must thus be replaced by

$$\frac{D}{2a'} - 1 = \frac{D/2a - 1}{1 + (\varepsilon - 1)\, D/2a}. \tag{21}$$

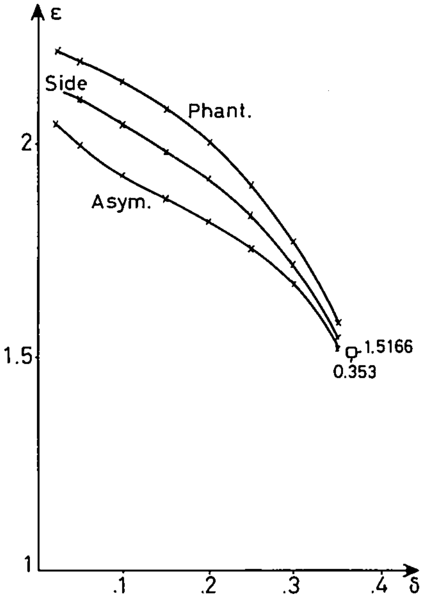


Fig. 6. Apparent relative dielectric constants of the non-homogeneous quad.

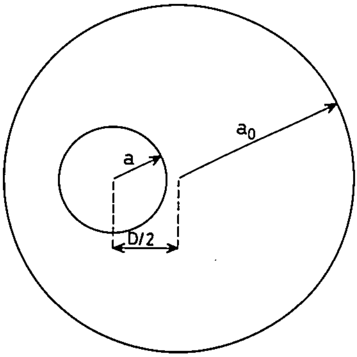


Fig. 7. Excentric coaxial.

For $a \cong D/2$, this becomes

$$\frac{D}{2a'} - 1 \cong \frac{1}{\varepsilon}\left(\frac{D}{2a} - 1\right) \tag{22}$$

so that all H are divided, and all C multiplied, by $\sqrt{\varepsilon}$. It has been verified from the results of figs 3 and 4, that the ratio H_A/H_S indeed tends to (19) when δ tends to 0.5, both for the homogeneous and non-homogeneous cases.

For the screened quad, we call C_1 the capacitance between one wire and the screen, C_2 the capacitance between adjacent and C_3 between opposite wires. We use the classical formulae (F = phantom)

$$C_A = 4C_1, \tag{23}$$

$$C_F = C_1 + 4C_2, \tag{24}$$

$$C_S = \tfrac{1}{2}\, C_1 + C_2 + C_3. \tag{25}$$

For thin insulation, C_1 and C_2 become infinite, whereas C_3 stays finite and can be neglected in (25). We thus have

$$\frac{C_F}{C_A} = \frac{C_2}{C_1} + \frac{1}{4}\,, \tag{26}$$

$$\frac{C_S}{C_A} = \frac{C_2}{4C_1} + \frac{1}{8}. \tag{27}$$

For adjacent wires, H_2 is (14) with D replaced by $D/\sqrt{2}$, hence

$$H_2 = 2\left(\frac{D}{a\sqrt{2}} - 2\right)^{\frac{1}{2}} \tag{28}$$

For H_1, we use (16) with the value (2) of a_0 and with $a/D = 1/(2\sqrt{2})$ in the denominator. This yields

$$H_1 = \left[\frac{D/(a\sqrt{2}) - 2}{1 + 1/\sqrt{2}}\right]^{\frac{1}{2}} = \frac{H_2}{2\,(1 + 1/\sqrt{2})^{\frac{1}{2}}} = 0.3827\, H_2. \tag{29}$$

From (29) and (26–27) it results that

$$\frac{C_F}{C_A} = 0.6327; \qquad \frac{C_S}{C_A} = 0.2207. \tag{30}$$

In the non-homogeneous case, (20) produces

$$\frac{D}{2a'\sqrt{2}} - 1 \cong \frac{1}{\varepsilon}\left(\frac{D}{2a\sqrt{2}} - 1\right)$$

with the same consequences as for the pair.

For a quad without screen, one has $C_1 = 0$ in (24–25) and C_3 stays finite and is negligible. One thus has

$$\frac{C_S}{C_1} = \frac{1}{4} \tag{31}$$

and this can be verified in the results of table V of ref. 5. We take this opportunity to correct a misprint in the third column of that table where the entry 0.2335 for $\delta^2 = 0.115$ should read 0.2235.

One might naively expect that the apparent relative dielectric constant should tend to 1 for a vanishing insulation, and the fact that the actual value is $\sqrt{\varepsilon}$ requires a physical explanation. The dominant components of the field are dipoles at the points of contact of the wires with each other and with the screen. In the small regions around these points of contact, which consist almost exclusively of insulating material, the field is very large; the increase of field intensity thus counteracts the vanishing insulation thickness.

Acknowledgement

The writer is grateful to R. R. Wilson (NKF Kabel, Delft) and to G. C. Groenendaal (Philips Res. Labs, Eindhoven) for many discussions, and to J. Praet who programmed all computations.

MBLE Research Laboratory — *Brussels, August 1978*

REFERENCES

[1] G. C. Groenendaal, R. R. Wilson and V. Belevitch, Philips Res. Repts **32**, 412-428, 1977.
[2] V. Belevitch, G. C. Groenendaal and R. R. Wilson, Proc. Intern. Wire and Cable Symp., Cherry Hill, N.J., 1977, pp. 338-342.
[3] H. Kaden, Wirbelströme und Schirmung in der Nachrichtentechnik, 2. Aufl., Springer, Berlin, 1959, pp. 159-163.
[4] E. Hallen, Electromagnetic theory, Chapman & Hall, London, 1962, p. 55.
[5] V. Belevitch, Philips Res. Repts, **32**, 96-117, 1977.

Philips J. Res. 33, 149-185, 1978 *R 987*

GAS BUBBLE FORMATION AHEAD OF A SOLIDIFICATION FRONT

by J. van den BOOMGAARD

Abstract

During the in situ growth of a magnetoelectric composite material with the aid of the E.F.G. (Edge defined Film-fed Growth) technique of Labelle and Mlavski [1,2,3,4]), we investigated the manner in which the properties of the material were influenced by the growth rate and the partial pressure of O_2 in an ambient consisting of $N_2 + O_2$ with a total pressure of 1 atmosphere. It was observed that, within the range of the growth rates used, gas bubbles were formed near the solidification front *below* and not *above* a certain growth rate which depended on the applied partial pressure of oxygen. When the same material was solidified under comparable circumstances in a Bridgman set-up, gas bubbles were formed up to higher growth rates than in the E.F.G. set-up. This cannot be explained by the existing theory which shows that bubble formation is suppressed at low growth rates because then the volatile solute has time to evaporate. Using a simple nucleation model it is shown that gas bubble formation can also be suppressed by using high growth rates so that bubbles do not get time to be formed. Under normal circumstances, these rates are so high that no reversible growth can be expected. A multichannel body, placed just ahead of the solidification front, reduces this rate to reasonable values. The use of such a body also prevents the nucleation of solid phases ahead of the solidification front.

1. Introduction

If a liquid contains a volatile component, e.g. P in Si or As in GaAs, this component will be expelled into the liquid during solidification if its distribution coefficient over the solid and liquid phase k is less than 1. As a consequence a concentration gradient of the volatile component is built up ahead of the solid–liquid interface during solidification until a steady state is reached in which this gradient no longer changes with time, which is only possible if the partial pressure of the volatile component in the gas phase is kept constant during the solidification process.

If the total pressure of the gas phase is also kept constant, by using a mixture of the volatile component and an insoluble gas, three regions may be distinguished in the liquid ahead of the solidification front in the steady state.

(a) At a large distance from the solidification front the liquid is in equilibrium with the applied partial pressure of the volatile component P_p, and has the equilibrium concentration C_{eq}.

(b) At intermediate distances the concentration is higher than C_{eq} but lower than C_b, the concentration whose equilibrium pressure is equal to the applied total gas pressure P_t plus the pressure due to the liquid column above the position where the concentration equals C_b. In this region the volatile component evaporates from the melt at the liquid–vapour interface; the average concentration of that component therefore decreases, but no formation of gas bubbles occurs.

(c) At small distances from the solid–liquid interface the concentration may be higher than C_b, and consequently bubbles may nucleate.

At low growth rates, R, the time during which the solid grows, is long and an appreciable amount of the solute may evaporate, resulting in a decrease of the concentration in region b and c. As a consequence the length of region c decreases and may even vanish below a certain growth rate, so that no gas bubble are formed.

This is in accordance with the theoretical considerations of Wilcox and Kuo [5], who concluded that the tendency for gas bubbles to form increases with increasing growth rate if the other quantities such as P_p and P_t are kept constant. In a group of solidification experiments at different rates R and at different values of P_p at a fixed value of P_t, we found on the contrary that gas bubbles were formed below a certain value of R and not above it. This value of R was higher, the higher the applied partial pressure P_p. The experiments are described in section 2.

In an attempt to explain these unexpected results two model processes are considered in section 3, where the probability of gas bubble formation is calculated as a function of the growth rate (sub-sections 3.2 and 3.3). In section 4 some general considerations about the two processes are given. The theoretical results are compared with the experimental results in section 5 and in the last section some other consequences are discussed.

2. Experimental

We investigated the influence of the growth rate and the partial pressure of O_2 at a constant total pressure of $N_2 + O_2$ on the properties of magnetoelectric composite material grown in situ and consisting of the phases; $BaTiO_3$, $\{(Co_2TiO_4)_{1-x}(CoFe_2O_4)_x\}$ and $BaFe_{12-2y}Co_yTi_yO_{19}$ (refs 1 and 2). $BaTiO_3$ and $\{(Co_2TiO_4)_{0.53}(CoFe_2O_4)_{0.47}\}$ were mixed in the mole ratio 61:39. To this mixture 10.5 wt % $BaFe_3Co_{4.5}Ti_{4.5}O_{19}$ was added. This mixture was melted by RF heating in a Pt 30% Rh crucible and bars of the composite material were grown by means of the E.F.G. method reported by Labelle and Mlavski [1,2,3,4]. The crucible contained a multichannel capillary and was covered by a lid, both of the same material (see fig. 1). The length of the

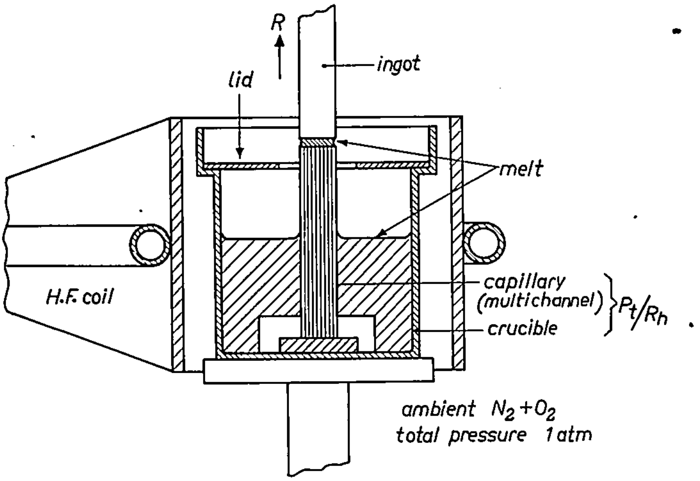


Fig. 1. Sketch of crucible with multichannel capillary used in the E.F.G. technique for growing magnetoelectric composite materials.

capillary was approximately 40 mm, its diameter 4.5 mm and the ratio between the total area of the capillary cylinder and the sum of the areas of the capillary channels in a transverse section was about 15. Before starting the growth of a bar the liquid was kept at the melting temperature for a time long enough to achieve equilibrium with the applied oxygen pressure. A platinum wire was then lowered until it was wetted by the liquid. Thereafter it was moved upward at a constant rate R and the temperature was regulated such that only a very thin liquid film was present between the growing bar and the top of the capillary cylinder. This temperature was controlled with an optical system within a few tenths of a degree. At the very flat solidification front three solid phases grew from the liquid. Bars of a constant diameter were grown at different rates in different nitrogen–oxygen mixtures with a fixed total pressure of 1 atmosphere. Observations were made of the growth rates at which gas bubbles were formed. The experimental results are given in fig. 2.

Liquids of the same composition solidified with the Bridgman technique in platinum tubes in air (N_2 + 20% O_2) exhibited strong gas bubble formation at a rate of 5 mm/min; in the E.F.G. technique the boundary between bubble formation and no bubble formation is already amply passed at this rate.

As the difference in this behaviour is to be sought in the use of the multichannel cylinder, we will discuss the bubble nucleation during directional solidification under different conditions.

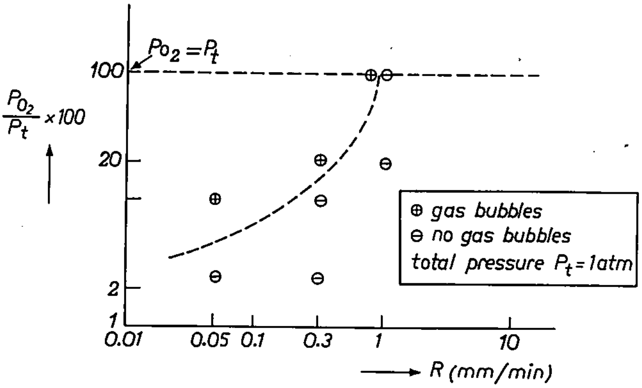


Fig. 2. The regions in which gas bubbles are formed and in which they are not formed during solidification of a magnetoelectric composite material of a fixed average composition, in a $P_{O_2}/P_t \times 100$ vs R plot on a logarithmic scale. P_t = 1 atm $O_2 + N_2$.

3. Theoretical

The shape of the concentration gradient, present ahead of the solidification front during directional solidification, is dependent on the value of the distribution coefficient k, on the applied partial pressure of the volatile component, on the growth rate and, if gas bubbles are formed, also on the applied total pressure. If a multichannel cylinder having the same diameter as the growing bar is placed just ahead of the solidification front, the shape of the concentration gradient will be influenced by it. Once all the conditions have been chosen, the solidification process reaches a steady state after some time, provided the length of the liquid bar is sufficient. In that case gas bubbles may nucleate in region c, if present, in which the concentration of the volatile component is so high that the associated partial pressure exceeds the applied total pressure. If we know the nucleation rate as a function of the concentration, we can calculate the length or the volume of the bar that can be solidified before a bubble has nucleated. Conversely, if a certain volume of a given cross-section has to be solidified without bubbles being formed, the values of the parameters at which this is possible can be calculated.

In section 3.1 calculations are given that demonstrate the influence of a multichannel cylinder on the concentration gradient of non-volatile solute.

In section 3.2 a model process is considered in which a volatile solute is present and the liquid outside the multichannel cylinder is in contact with the gas phase (case (1)).

In section 3.3 a model process is considered in which the liquid is not in contact with the gas phase except for its extreme part (case (2)).

3.1. *The influence of a multichannel cylinder ahead of the solidification front on the concentration gradient of a non-volatile solute*

A liquid containing a non-volatile solute moves at a rate R through a semi-infinite porous tube of a material which is insoluble in the liquid; the liquid cannot pass through the tube wall. (Although it is not important in the process described here whether this tube is porous or not, we will use the same set-up as in the process described in section 3.2 where it is essential for the tube wall to be porous.) Placed on top of the tube is a multichannel cylinder made of an insoluble material (see fig. 3). It has a length L and an outer diameter $2r_1$ equal to the inner diameter of the tube. The cylinder contains a number of channels n each with a diameter $2r_2$. The cylinder and the tube are at the same temperature T_2. At the top of the cylinder a solid bar grows from the liquid. This bar is also cylindrical, with a diameter $2r_1$, and is kept at a uniform temperature T_1. The melting point of the solid lies between T_1 and T_2. The solid moves upward at a rate R andT_2 is regulated in such a way that the solid–liquid interface is at an infinitesimal distance from the top of the cylinder. The liquid moves at the same rate R through the tube. Inside the channels, however, it flows at a rate αR, where $\alpha = r_1{}^2/nr_2{}^2$. The concentration gradient ahead of the

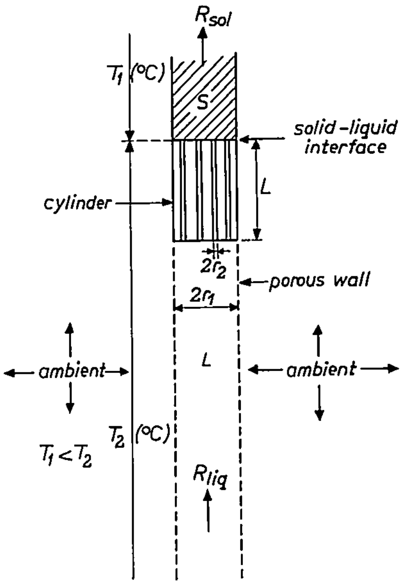


Fig. 3. Schematic drawing of the solidification process also used in case (1).

solidification front in the steady state is given by the following relations:

$$\frac{\partial C_x}{\partial t} = D\frac{\partial^2 C_x}{\partial x^2} + \alpha R\frac{\partial C_x}{\partial x} = 0 \qquad \text{for} \qquad 0 \leqslant x \leqslant L, \quad \text{(I)}$$

and

$$\frac{\partial C_x}{\partial t} = D\frac{\partial^2 C_x}{\partial x^2} + R\frac{\partial C_x}{\partial x} = 0 \qquad \text{for} \qquad L \leqslant x \leqslant \infty, \quad \text{(II)}$$

where x is the distance to the solidification front, C_x the concentration of the solute at x, L the length of the multichannel tube and D the diffusion coefficient in the liquid.

If we start with a homogeneous liquid, in which the concentration of the solute is given by C_w, the weighed in concentration, the boundary conditions are given by $C_x \rightarrow C_w$ for $x \rightarrow \infty$ in (II) and $C_0 = C_w/k$ for $x = 0$ in (I). C_L (C_x for $x = L$) in (I) equals C_L in (II) and the flux of the solute in each transverse section equals RC_w, independent of the position. (Inside the multichannel the flux inside each channel equals αRC_w but mean flux equals RC_w.) Combination of these conditions leads to

$$C_x = \frac{1-k}{k}\, C_w \exp\,(-\,\alpha Rx/D) + C_w \qquad \text{for } 0 \leqslant x \leqslant L \quad (1)$$

and

$$C_x = \frac{1-k}{k}\, C_w \exp\left\{-\frac{R}{D}\left[x + (\alpha - 1)\,L\right]\right\} + C_w \qquad \text{for } L \leqslant x \leqslant \infty. \quad (2)$$

For $L \gg D/\alpha R$ the whole concentration gradient is described by (1) and for $L = 0$ by (2). In intermediate cases both equations have to be used.

The effect of the cylinder in the liquid is that inside it the concentration gradient is compressed to the solidification front by a factor α.

The three possibilities for L are shown in fig. 4.

3.2. *Bubble formation in case (1)*

In case (1) the liquid contains a volatile solute. The liquid is solidified in the set-up given in fig. 3 and the experiment is carried out in an ambient consisting of the volatile solute with a fixed partial pressure and an inert gas. The total pressure of the gas phase is kept constant at a desired value. The wall of the porous tube is assumed to be completely permeable for the gas phase, the wall of the cylinder completely impermeable. Evaporation of solute from the liquid

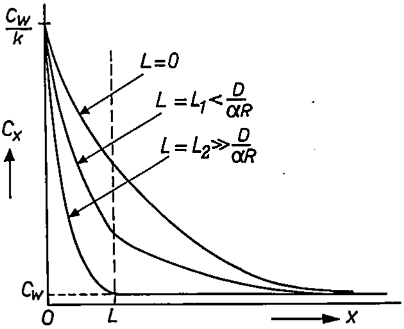


Fig. 4. Concentration gradients of a non-volatile solute ahead of the solidification front at three different lengths of the multichannel cylinder; x is the distance to the solidification front.

near the solid–liquid interface is assumed to be negligible because of the infinitesimal distance between the top of the cylinder and this interface. As already remarked in section 1, bubbles can only be formed in the region where $C_x > C_b$. In order to calculate the number of bubbles formed in the steady state per unit time we will use a somewhat oversimplified model. We assume that

(1) in the solid, the liquid and the gas the volatile component is present in the same form, i.e. in each phase as atoms or as molecules of the same dimensions, and the equilibrium concentration of this component in the liquid is proportional to its equilibrium pressure; the concentration of the volatile component in the solid phase once being formed remains unchanged,

(2) radial concentration gradients are negligible,

(3) the number of bubbles that would be formed if the concentration gradient of the volatile component ahead of the solidification front were not influenced by bubble formation is a measure of the probability of bubble formation during the solidification process,

(4) the bubble nucleation is homogeneous.

Assumption (2) seems a little odd. However, if a multichannel cylinder is used and we are interested in the conditions that no bubbles are formed, which will appear to be either at high or at low growth rates, this assumption is practically correct. If no multichannel is used it is less correct, but for a small radius of the solidifying bar it will be not too bad.

In the steady state the concentration of the volatile component C_x is independent of time and thus $\partial C_x/\partial t = 0$. The change of the concentration with time at a certain value of x is a result of three processes: diffusion, flow and evaporation. In the differential equations the first two processes give rise to the same terms as in the eqs (I) and (II).

The third process gives rise to an evaporation term

$$\left(\frac{\partial C_x}{\partial t}\right)_{\text{evap}}$$

By virtue of the second and third assumptions the number of atoms or molecules n passing per unit time through the cylindrical surface of a segment of a bar with a radius r_1 and a length $\mathrm{d}x$ may be put proportional to the difference between the actual concentration and the equilibrium concentration of the solute and to S', the surface area of that cylinder segment in contact with the gas phase:

$$\frac{\partial n}{\partial t} = -K'\,(C_x - C_{\text{eq}})\,S', \tag{3a}$$

where K' is a positive proportionality constant. As $n = C_x\,\mathrm{d}V$ where $\mathrm{d}V$ is the volume of the segment, and $\mathrm{d}V = \pi r_1{}^2\,\mathrm{d}x$, while $S' = 2\pi r_1\,\mathrm{d}x$, we may write

$$\left(\frac{\partial C_x}{\partial t}\right)_{\text{evap}} = -\frac{K}{r_1}\,(C_x - C_{\text{eq}}), \tag{3b}$$

where $K = 2K'$, which has the dimension cm/s. Because no evaporation can take place when $0 \leqslant x \leqslant L$, i.e. inside the multichannel cylinder the evaporation term is equal to zero, the differential equation governing the concentration gradient in this region is the same as eq. (I).

Outside the multichannel cylinder, however, the evaporation term has to be taken into account and there the differential equation becomes

$$\frac{\partial C_x}{\partial t} = D\,\frac{\partial^2 C_x}{\partial x^2} + R\,\frac{\partial C_x}{\partial x} - \frac{K}{r_1}\,(C_x - C_{\text{eq}}) = 0. \tag{4}$$

The boundary conditions for these equations are the following.

(1) For $x \to \infty$, $C_x \to C_{\text{eq}}$ in (4).

(2) C_L must have the same value in (I) and (4).

(3) The flow of solute towards the solidification front inside the cylinder ($0 \leqslant x \leqslant L$) is independent of x and equal to $-kRC_0$, where C_0 is the concentration at the solidification front. Outside the cylinder ($x \geqslant L$) it is dependent on x. It is equal to $-RC_{\text{eq}}$ for $x = \infty$ and to $-kRC_0$ for $x = L$. The solutions of (I) and (4) are given by

$$C_x = Q\left[\frac{k}{1-k}\,Q\exp\left(\frac{\alpha RL}{D}\right) + 1\right]^{-1} \times \left\{\exp\left[\frac{-\alpha R}{D}\,(x-L)\right] + \frac{k}{1-k}\exp\left(\frac{\alpha RL}{D}\right)\right\} C_{\text{eq}} \tag{5a}$$

for $0 \leqslant x \leqslant L$

and

$$C_x = (Q-1)\left\{\left[\frac{k}{1-k}\,Q\exp(\alpha RL/D)+1\right]^{-1}\right.$$

$$\left.\times\ \exp\left[-\frac{K}{r_1\,R}(Q-1)(x-L)\right]+1\right\}C_{\text{eq}}, \qquad (5b)$$

for $x > L$,

where

$$Q = \frac{R}{2KD}\left[r_1\,R + (r_1^2\,R^2 + 4r_1\,KD)^{\frac{1}{2}}\right] + 1. \qquad (6)$$

From $R \geqslant 0$ it follows that $Q \geqslant 1$.
Further

$$C_0 = \frac{QC_{\text{eq}}}{kQ + (1-k)\exp(-\alpha RL/D)} \qquad (7)$$

and

$$C_L = \frac{[(1-k)\exp(-\alpha RL/D) + k]\,Q}{kQ + (1-k)\exp(-\alpha RL/D)}\,C_{\text{eq}}. \qquad (8)$$

Thus the relative difference between C_0 and C_L is given by

$$\frac{C_0 - C_L}{C_0} = (1-k)\left[1 - \exp(-\alpha RL/D)\right] \qquad (9)$$

and this difference increases with increasing values of α, R and L from zero for $\alpha RL = 0$ to $(1-k)$ for $\alpha RL = \infty$. Equations (*5a*) and (*5b*) describe the steady-state concentration gradient ahead of the solidification front. With the aid of these equations the number of bubbles that may be formed per unit time can be calculated if the nucleation rate is known. This nucleation rate J is defined as

$$J = \frac{\mathrm{d}^2N}{\mathrm{d}V\,\mathrm{d}t}, \qquad (10)$$

where V is the volume, t is the time and N the number of bubbles *). In the same way as for any form of homogeneous nucleation, this rate may be expressed by

*) This notation deviates from the normal practice in which N is put equal to the number of nuclei in unit volume and thus $J = \mathrm{d}N/\mathrm{d}t$. The reason for this deviation is that we want to calculate the number of bubbles formed in a certain time in an inhomogeneous liquid. Therefore we consider an infinitesimal volume $\mathrm{d}V$ in which an infinitesimal number of bubbles $\mathrm{d}N$ are formed. The number of bubbles in unit volume is given by $\mathrm{d}N/\mathrm{d}V$ in that case and hence the nucleation rate is $J = \mathrm{d}^2N/\mathrm{d}V\,\mathrm{d}t$.

$$J = A \exp\left(-\frac{\Delta F}{k^* T}\right), \tag{11}$$

where ΔF is the free energy of formation of a critical embryo, k^* is Boltzmann's constant and T the absolute temperature [5]). A is a constant comprising such quantities as the accommodation coefficient, the molecular concentration in the liquid and the interface energy of the gas bubble [6]).

Equation (11) can also be written as

$$J = A \exp\left[-\frac{B}{k^*T\,(P_g - P_h)}\right], \tag{12}$$

where B is a constant containing the interface energy of the gas bubble and other quantities, P_g the gas pressure inside the bubble and P_h the hydrostatic pressure inside the critical embryo [6,7,8]).

If the temperature of the melt is homogeneous and constant, J is a function of P_g and P_h only, and as P_g is a function of the concentration, J is dependent only on C_x and P_h. In the hypothetical set-up given in fig. 3 the value of P_h is dependent only on the distance to the solidification front and on the total gas pressure P_t i.e. it changes with x, not with time.

Gas bubbles may only be formed in the region in which $C_x > C_b$, the concentration whose equilibrium pressure is equal to P_h. The length of the region in which bubbles may be formed is therefore given by x_b, the distance to the solidification front, where $C_x = C_b$. As the length of the region in which C_x differs from the equilibrium concentration C_{eq} is of the order of D/R, x_b will be of the order of only a few millimeters. As a consequence the contribution of the liquid column to P_h is small and may be neglected if the value of P_t is not too small (e.g. of the order of one atmosphere). Therefore P_h may be considered as a constant for a given value of P_t to a good approximation and so too, therefore, may C_b. For fixed values of P_t, T and P_p (the applied partial pressure of the volatile component) the nucleation rate is a function of C_x only

$$J_{P_t,P_p,T} = f(C_x) \tag{13}$$

and $f(C_x) = 0$ for $C_x \leqslant C_b$ (i.e. $P_g < P_h \cong P_t$).

The total number of bubbles that may nucleate per unit time during steady-state solidification is given by

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \int_0^{V_b} J\,\mathrm{d}V, \tag{14}$$

where V_b is the volume of the liquid between the solidification front and a plane parallel to the front at x_b. Denoting the area of a transverse section through the solidifying liquid for $x > L$ by S, eq. (14) can be written as

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha}\int_0^{x_b} J\,\mathrm{d}x \qquad (15a)$$

if $x_b \leqslant L$

and as

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha}\int_0^{L} J\,\mathrm{d}x + S\int_L^{x_b} J\,\mathrm{d}x \qquad (15b)$$

if $x_b > L.$

With eqs (5a) and (5b) x can be expressed in C_x which results in

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha^2 R}\int_{C_b}^{C_0} DJ\left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ+(1-k)\exp\left(-\alpha RL/D\right)}\right]^{-1}\mathrm{d}C_x \qquad (16a)$$

if $C_L \leqslant C_b$ (corresponding to $X_b \leqslant L$)
and

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha^2 R}\int_{C_L}^{C_0} DJ\left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ+(1-k)\exp\left(-\alpha RL/D\right)}\right]^{-1}\mathrm{d}C_x$$

$$+\frac{S}{R}\{2\,[1+(1+4KD/r_1\,R^2)^{\frac{1}{2}}]^{-1}\}\int_{C_b}^{C_L}\frac{DJ}{C_x - C_{\mathrm{eq}}}\,\mathrm{d}C_x \qquad (16b)$$

if $C_b < C_L$ (corresponding to $x_b > L$). The integrals in (16a) and (16b) are not easy to solve because of J. However, they are sufficient for our purpose. As the concentration gradient ahead of the solidification front is not dependent on time in the steady state, because the growth rate is kept constant during the solidification process, the value of $\partial N/\partial t$ is also independent of time and therefore

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \left(\frac{\Delta N}{\Delta t}\right)_{P_t,P_p,T}.$$

The number of bubbles that may be expected in a time Δt during steady-state growth is given by

$$\Delta N_{P_t,P_p,T} = \frac{S\,\Delta t}{\alpha^2 R}\int_{C_b}^{C_0} DJ\left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ+(1-k)\exp\left(-\alpha RL/D\right)}\right]^{-1}\mathrm{d}C_x \qquad (17a)$$

if $C_L \leqslant C_b$

and by

$$\Delta N_{P_t,P_p,T} = \frac{S\,\Delta t}{\alpha^2\,R} \int_{C_L}^{C_0} DJ \left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ + (1-k)\exp(-\alpha RL/D)} \right]^{-1} \mathrm{d}C_x$$

$$+ \frac{S\,\Delta t}{R} \{2\,[1 + (1 + 4KD/r_1\,R^2)^{\frac{1}{2}}]^{-1}\} \int_{C_b}^{C_L} \frac{DJ}{C_x - C_{\mathrm{eq}}} \mathrm{d}C_x \qquad (17b)$$

if $C_b < C_L$.

Thus, if we want to grow a bar with a given length a and a homogeneous composition without bubbles being formed, ΔN must be smaller than 1 in the time $\Delta t = a/R$ during which the bar solidifies in the steady state. Because $aS = V_a$, i.e. the volume of the bar solidified in a time Δt, the conditions for bubble-free growth at a constant P_t, P_p and T are given by

$$\frac{V_a}{\alpha^2\,R^2} \int_{C_b}^{C_0} DJ \left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ + (1-k)\exp(-\alpha RL/D)} \right]^{-1} \mathrm{d}C_x < 1 \qquad (18a)$$

if $C_L \leqslant C_b$

and by

$$\frac{V_a}{\alpha^2\,R^2} \int_{C_L}^{C_0} DJ \left[C_x - \frac{kQC_{\mathrm{eq}}}{kQ + (1-k\exp(-\alpha RL/D)} \right]^{-1} \mathrm{d}C_x$$

$$+ \frac{V_a}{R^2} \{2\,[1 + (1 + 4KD/r_1\,R^2)^{\frac{1}{2}}]^{-1}\} \int_{C_b}^{C_L} \frac{DJ}{C_x - C_{\mathrm{eq}}} \mathrm{d}C_x < 1 \qquad (18b)$$

if $C_b < C_L$.

From these two equations the rates at which no bubbles are formed can be calculated as functions of the parameters V_a and L at constant values of P_p, P_t and T, because in these circumstances J is always the same function of C_x and K, k, D, C_b and C_{eq} are constant. If, however, the influence of the parameters P_p and P_t is investigated one must realize that the melting temperature of the solid changes with the values of these parameters, mainly with P_p, and so too, therefore, do the values of K, k, D, C_b and C_{eq}.

Therefore, in our model the temperature of the melt is fixed at a constant value of T_2 and the temperature of the growing solid T_1 is adapted to the changed circumstances in such a way that the solidification front is at an infinitesimal distance from the top of the multichannel cylinder. In this way the values of D and K are independent of P_t and P_p, while C_b depends only on P_t and C_{eq} depends only on P_p. As the distribution coefficient is determined by the temperature at the solidification front, it changes with P_p, but this variation will be neglected in our considerations.

According to eq (18a) and (18b) the conditions in which ΔN is less than 1 depend on the value of C_b with respect to C_0 and C_L. These latter two quantities can be expressed in C_{eq} by (7) and (8) respectively and their difference by (9). The value of C_b, which is determined by P_t, is always greater than that of C_{eq}, because even when $P_t = P_p$ the pressure of the liquid has to be taken into account. Three regions of values of C_b with respect to C_0 and C_L can be distinguished by the conditions
$C_0 > C_L > C_b\ (> C_{eq})$, $C_0 > C_b > C_L\ (> C_{eq})$ and $C_b > C_0 > C_L\ (> C_{eq})$.
In the first and second regions of values of C_b the conditions in which eqs (18b) and (18a) hold are valid. In the conditions of the third region no bubbles are formed at all during solidification of a bar of any desired dimension. According to (7) the condition $C_b \geqslant C_0$ can be written as

$$C_{eq} \leqslant \left[k + \frac{1-k}{Q} \exp\left(-\alpha RL/D\right) \right] C_b. \tag{19}$$

Thus, once C_b and C_{eq} are fixed by P_t and P_p, the value of R below which no bubbles are formed can be expressed in the other parameters (k, K, D, etc.). This value of R can be calculated from the equation obtained from (19) by making the right-hand term equal to the left-hand term and combining it with (6), which leads to

$$\left\{ \frac{R}{2KD} \left[r_1\, R + (r_1^{\,2}\, R^2 + 4 r_1\, KD)^{\frac{1}{2}} \right] + 1 \right\} \exp\left(\alpha RL/D\right) = \frac{(1-k)\, C_b}{C_{eq} - kC_b}. \tag{20}$$

Although R cannot be written as an explicit function of C_b and C_{eq} and of the other parameters, some conclusions can be drawn.

As negative values of R do not have physical meaning and as α, L, D, K, r_1 and k (< 1) are positive quantities, the value of the root of eq. (20), given by the symbol $R_{b1}{}^{(\alpha L)}$, which is the growth rate below which no bubbles are formed, is a monotonically decreasing function of αL. It decreases from $R_{b1}{}^{(0)}$, its value in a solidification process without the use of a multichannel cylinder, to 0 for cylinder of infinite length.
Thus

$$0 \leqslant R_{b1}{}^{(\alpha L)} \leqslant R_{b1}{}^{(0)}. \tag{21}$$

Insertion of $L = 0$ in (20) yields

$$R_{\mathrm{b}1}{}^{(0)} = \frac{C_{\mathrm{b}} - C_{\mathrm{eq}}}{C_{\mathrm{eq}} - kC_{\mathrm{b}}} \left[\frac{KD\,(C_{\mathrm{eq}} - kC_{\mathrm{b}})}{r_1\,(1-k)\,C_{\mathrm{b}}} \right]^{\frac{1}{2}} \tag{22}$$

The value of $R_{\mathrm{b}1}{}^{(\alpha L)}$ varies from 0 to ∞ if C_{b} is varied from C_{eq} to C_{eq}/k, as is easily seen from (20). Disregarding the liquid pressure this means that P_{t} is varied at a constant value of P_{p} from P_{p} to the equilibrium pressure of the volatile component at a concentration C_{eq}/k, which is the highest concentration that occurs at the solidification front. At still higher values of P_{t} eq. (20) has no real root, which means that no bubbles are formed at any rate if $C_{\mathrm{b}} \geqslant C_{\mathrm{eq}}/k$.

In a plot of log ΔN vs log R at constant P_{t} and P_{p} the line $\log R = \log R_{\mathrm{b}1}{}^{(\alpha L)}$ is an asymptote to that curve. The log ΔN vs log R curve has another asymptote to which this curve approaches at high values of R. This asymptote will be treated in the next two sub-sections.

3.2.1. Bubble formation and bubble-free growth if $L = 0$

If $L = 0$ the difference between C_0 and C_L vanishes and eq. (17b) in which $S\,\Delta t$ is replaced by V_a/R simplifies to

$$\log \Delta N = \log V_a - 2 \log R + \log 2 - \log\,[1 + (1 + 4KD/r_1\,R^2)^{\frac{1}{2}}] + \log \int_{C_{\mathrm{b}}}^{C_0} \frac{DJ}{C_x - C_{\mathrm{eq}}}\,\mathrm{d}C_x. \tag{23}$$

The value of C_0 becomes C_{eq}/k if

$$R^2 \gg \frac{(1-2k)^2}{k\,(1-k)} \frac{KD}{r_1}$$

(see (6) and (7)) and the term $\log\,[1 + (1 + 4KD/r\,R^2)^{\frac{1}{2}}]$ becomes log 2 if $R^2 \gg 4\,KD/r_1$. Thus for values of k, $\frac{1}{2} - \frac{1}{4}\sqrt{2} < k < \frac{1}{2} + \frac{1}{4}\sqrt{2}$, it is sufficient that $R^2 \gg 4KD/r_1$ and for values of $k < \frac{1}{2} - \frac{1}{4}\sqrt{2}$ or $> \frac{1}{2} + \frac{1}{4}\sqrt{2}$ it is sufficient that

$$R^2 \gg \frac{(1-2k)^2}{k\,(1-k)} \frac{KD}{r_1}$$

to simplify (23) to

$$\log \Delta N = \log V_a - 2 \log R + \log \int_{C_{\mathrm{b}}}^{C_{\mathrm{eq}}/k} \frac{DJ}{C_x - C_{\mathrm{eq}}}\,\mathrm{d}C_x. \tag{24}$$

Because the boundaries of the integral are independent of C_x and thus of R, eq. (24) can be written as

$$\log \Delta N = E - 2 \log R, \tag{25}$$

where

$$E = \log V_a + \log \int\limits_{C_b}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, \mathrm{d}C_x = \text{constant.}$$

As k in most cases will be smaller than $\frac{1}{2} - \frac{1}{4}\sqrt{2}$ (∞ 0.15), we will use the condition

$$R^2 \gg \frac{(1-2k)^2}{k\,(1-k)} \frac{KD}{r_1}$$

in order to estimate the growth rate above which eq. (25) will hold, D is of the order of 10^{-5} cm^2/s, $r_1 \cong \frac{1}{4}$ cm and K will have a value $\cong 10^{-2}$ cm/s, and thus, for $k = 10^{-3}$, $R^2 \gg 4 \times 10^{-4}$ cm^2/s^2. At growth rates $R > 2 \times 10^{-2}$ cm/s (72 cm/h) eq. (25) will hold to a good approximation and this equation represents the second asymptote to the $\log \Delta N$ vs $\log R$ curve.

Because

$$\log\,[1 + (1 + 4KD/r_1\, R^2)^{\frac{1}{2}}] > \log 2$$

and

$$\log \int\limits_{C_b}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, \mathrm{d}C_x > \log \int\limits_{C_b}^{C_0} \frac{DJ}{C_x - C_{eq}} \, \mathrm{d}C_x$$

the value of $\log \Delta N$ calculated from eq. (24) for each chosen value of R is always greater than the value of ΔN calculated from (23) for the same value of R. This means that the $\log \Delta N$ vs $\log R$ curve given by (23) is situated to the left of the asymptote given by (24). As a consequence the $\log \Delta N$ vs $\log R$ curve is situated between the asymptotes given by (22) and (24). This is shown in fig. 5.

The asymptotes intersect the straight line $\log \Delta N = 0$ at the points $\log R = \log R_{b1}^{(0)}$ and $\log \mathrm{R} = \log R_{b2}^{(0)}$,

$$\log R_{b_1}^{(0)} = \log (C_b - C_{eq}) - \tfrac{1}{2} \log (C_{eq} - kC_b) - \tfrac{1}{2} \log (1-k)\, C_b \tag{26}$$

$$+ \tfrac{1}{2} \log K + \tfrac{1}{2} \log D - \tfrac{1}{2} \log r_1, \qquad \text{(see (22))}$$

and

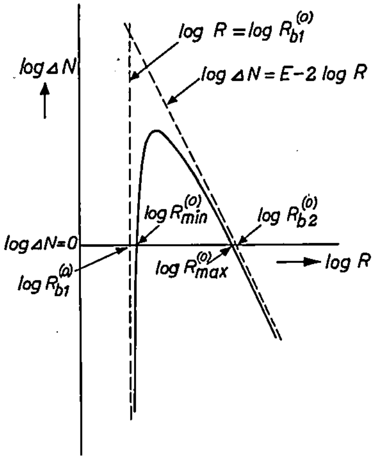


Fig. 5. The curve log ΔN vs log R with its two asymptotes (dashed lines) for the case $L = 0$ at constant values of P_p and P_t.

$$\log R_{\mathrm{b}2}{}^{(0)} = \tfrac{1}{2} \log E \left(= \tfrac{1}{2} \log V_a \int\limits_{C_\mathrm{b}}^{C_\mathrm{eq}/k} \frac{DJ}{C_x - C_\mathrm{eq}} \, \mathrm{d}C_x \right). \tag{27}$$

(see (23))

Between these two values of R, ΔN moves through a maximum value. The intersection points of the curve with $\log \Delta N = 0$ are given by $\log R = \log R_{\mathrm{min}}{}^{(0)}$ to the right of $\log R_{\mathrm{b}1}{}^{(0)}$ and by $\log R = \log R_{\mathrm{max}}{}^{(0)}$ to the left of $\log R_{\mathrm{b}2}{}^{(0)}$. At growth rates $R_{\mathrm{min}}{}^{(0)} < R < R_{\mathrm{max}}{}^{(0)}$ bubbles are formed during the growth of a solid bar of desired dimensions.

Unfortunately it is not possible to solve (23) for $\log \Delta N = 0$ and therefore the dependence of $R_{\mathrm{max}}{}^{(0)}$ and $R_{\mathrm{min}}{}^{(0)}$ on the different parameters cannot be determined exactly. For our purpose, however, it is sufficient to determine the position of the intersection points of both asymptotes with the line $\log \Delta N = 0$ as a function of these parameters. The dependence of $\log R_{\mathrm{b}1}{}^{(0)}$ on these quantities is given by eq. (26) and that of $\log R_{\mathrm{b}2}{}^{(0)}$ by eq. (27), containing the integral

$$\int\limits_{C_\mathrm{b}}^{C_\mathrm{eq}/k} \frac{DJ}{C_x - C_\mathrm{eq}} \, \mathrm{d}C_x,$$

which cannot be solved analytically. The upper boundary of this integral depends only on P_p and k and its lower boundary only on P_t. This makes it possible to verify how its value depends on these three quantities. As D, J, and $C_x - C_\mathrm{eq}$

are positive quantities, the value of the integrant is also a positive quantity. As a consequence the integral always has a positive value which increases with increasing value of D en P_p and with decreasing value of k and P_t. With the aid of these properties of the integral in (27), conclusions about the behaviour of $R_{b1}^{(0)}$ and $R_{b2}^{(0)}$ as a function of the different parameters can be drawn from (26) and (27). The parameter r_1 will be left out of consideration because one of the assumptions in our model was that no radial concentration gradients occur. The larger the value of r_1 the less this condition is fulfilled. Therefore we leave r_1 constant at a reasonably low value (0.1–0.5 cm).

In table I, part 1, the influence of the different parameters on $R_{b1}^{(0)}$, $R_{b2}^{(0)}$ and $R_{b1}^{(0)}$–$R_{b2}^{(0)}$ is given, assuming that the others are kept constant. Only the

TABLE I

Dependence of $R_{b1}^{(0)}$ and $R_{b2}^{(0)}$ and of $R_{b1}^{(\alpha L)}$ and $R_{b2}^{(\alpha L)}$ on the different parameters for a given solidified volume V_a and a given nucleation rate J.

part	increasing value of	$R_{b1}^{(0)}$ and $R_{b1}^{(\alpha L)}$	$R_{b2}^{(0)}$ and $R_{b2}^{(\alpha L)}$	$R_{b2}^{(0)}$–$R_{b1}^{(0)}$ $R_{b2}^{(\alpha L)}$–$R_{b1}^{(\alpha L)}$
1	P_t (C_b)	increasing	decreasing	decreasing
	P_p (C_{eq})	increasing	increasing	increasing
	K	increasing	remain unchanged	decreasing
	D	increasing	increasing	either decreasing, increasing, or constant
	k	increasing	decreasing	decreasing
2	L	decreasing	remain unchanged above a certain small value of L	increasing
	α	decreasing	decreasing	either decreasing, increasing or constant

influence of increasing values of these parameters is indicated. In the opposite case this influence will be reversed.

For a chosen solidification reaction only P_p (i.e. C_{eq}) and P_t (i.e. C_b) can be varied, because the other quantities (k, K, D, etc.), are fixed. Therefore we will consider the dependence of $R_{b1}{}^{(0)}$ and $R_{b2}{}^{(0)}$ on these two parameters in somewhat more detail.

In the first place we consider the variation of $R_{b1}{}^{(0)}$ and $R_{b2}{}^{(0)}$ with P_p at a constant P_t. The value of P_p can only be varied from 0 to P_t; this means that C_{eq} can be varied from 0 to $C_{eq\ max}$, which is somewhat smaller than the constant C_b, fixed by P_t, because of the presence of a liquid column, as already remarked. According to (28) log $R_{b1}{}^{(0)}$ varies from $+\infty$ to $-\infty$ if C_{eq} is varied from kC_b to $C_b(k < 1)$. According to (27) log $R_{b2}{}^{(0)}$ varies from $-\infty$ to $+\infty$ if C_{eq} is varied from kC_b to $+\infty$. However, because C_{eq} can never pass the value of $C_{eq\ max}$, the values of $R_{b1}{}^{(0)}$ and log $R_{b2}{}^{(0)}$ at values of $C_{eq} > C_{eq\ max}$ have no physical meaning.

In figure 6 the variation of log $R_{b1}{}^{(0)}$ and log $R_{b2}{}^{(0)}$ with log P_p at a constant P_t are given by the curves I and II respectively, assuming that P_p is proportional to C_{eq} and P_t to C_b. The value of log $R_{min}{}^{(0)}$ approaches asymptotically to curve I with increasing value of log P_p, and the value of log $R_{max}{}^{(0)}$ approaches asymptotically to curve II with increasing value of log P_p. The values of log $R_{min}{}^{(0)}$ and log $R_{max}{}^{(0)}$ are situated on the same curve III and bubbles are formed only at values of P_p and R that give rise to a point above curve III (shaded area). At values of P_p lower than $P_{p\ min}$, no bubbles are formed at any rate.

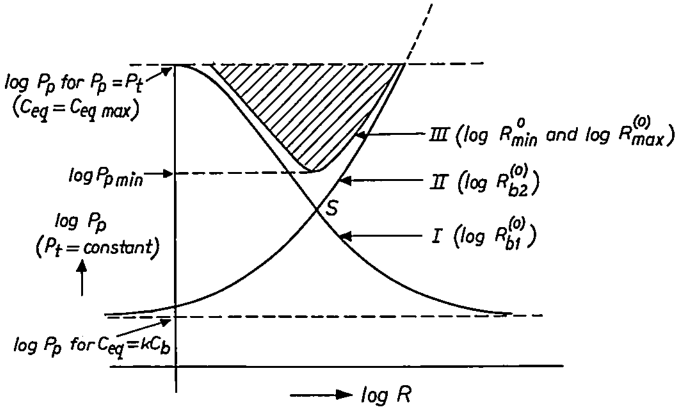


Fig. 6. Variation of log $R_{b1}{}^{(0)}$ and log $R_{b2}{}^{(0)}$ with log P_p at a constant P_t (curves I and II). Curve III gives the continuous curve for log $R_{min}{}^{(0)}$ and log $R_{max}{}^{(0)}$. Bubble formation is only possible in the shaded region.

From the data in table I and from fig. 6 it follows that the value of $P_{p\,min}$ is low in the case of solidifications with low values of k and K.

The whole curve III cannot always be realized. It may be that above and/or below a certain value of P_p unwanted phases solidify from the liquid, and there may be no value of P_p below which the desired phase can be grown from the liquid at any rate without bubble formation.

In the second place we consider the process in which P_p is kept constant and P_t is varied, i.e. C_{eq} is kept constant and C_b is varied. The value of P_t can be varied from P_p to $+\infty$. This means that C_b varies from $C_{b\,min}$ to $+\infty$. $C_{b\,min}$ is somewhat larger than C_{eq} because of the pressure of the liquid column, as already stated.

In figure 7 the curves of log $R_{b1}^{(0)}$ (curve I) and of log $R_{b2}^{(0)}$ (curve II) are sketched in a log P_t vs log R plot at constant P_p. It is seen that $R_{min}^{(0)}$ and $R_{max}^{(0)}$ are situated on curve III, which approaches asymptotically to the curves I and II. Bubble formation can only take place at values of P_t and R that give rise to points below curve III (shaded area). For the same reason as in the case of variation of P_t at a constant P_t, the region in which bubble formation occurs is greater the smaller the values of k and K. $P_{t\,max}$ is the maximum value of P_t at which bubble formation can be expected.

The total curve III can be realized for all values of P_p that give rise to the desired solid phase growing from the liquid. If this phase is only in equilibrium with the liquid for $P_{p1} \leqslant P_p \leqslant P_{p2}$, no type III curves are found for $P_p \leqslant P_{p1}$ or $P_p \geqslant P_{p2}$, as is easily understood. The same remark as made before about the appearance of unwanted phases applies in this case as well.

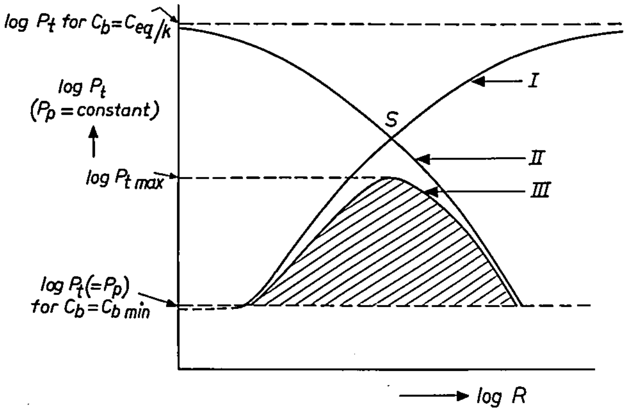


Fig. 7. Variation of log $R_{b1}^{(0)}$ and log $R_{b2}^{(0)}$ with log P_t at a constant P_p. (curves I and II). Curve III gives the continuous curve for log $R_{min}^{(0)}$ and log $R_{max}^{(0)}$. Bubble formation is only possible in the shaded region.

3.2.3. Bubble formation and bubble-free growth if $L > 0$

The introduction of a multichannel cylinder ahead of the solidification front changes the log ΔN vs log R curve at a constant P_t, P_p, and T. We have already shown that the rate below which no bubbles are formed $R_{b1}^{(\alpha L)}$ is smaller than $R_{b1}^{(0)}$ and it decreases with increasing value of αL. The other asymptote to this curve is found for $R \to \infty$. From (7) and (8) it follows that in this case $C_0 \to C_{eq}/k$ and $C_L \to C_{eq}$. This means that $C_b > C_L$, and thus (17a) simplifies to

$$\Delta N_{P_t,P_p,T} = \frac{V_a}{\alpha^2 R^2} \int_{C_b}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \tag{28}$$

or, written in the simplified form,

$$\log \Delta N = E - 2 \log \alpha - 2 \log R, \tag{29}$$

where E has the same value as in (25).

The value of R for which (28) may be used in order to calculate ΔN with sufficient accuracy can be estimated in the following way. If we want the difference between C_{eq}/k and C_0 to be less than 1%, it follows from (7) that R must fulfil the condition

$$R > 2.3 \frac{D}{\alpha L} \left(\log \frac{1-k}{kQ} + 2 \right). \tag{30}$$

For the same condition the coefficient of C_{eq} in (17a) deviates less than 1% from 1, and thus if eq. (30) is fulfilled eq. (28) holds to a good approximation. However, eq. (30) is not an explicit expression of R because Q contains R. Nevertheless it is possible to estimate the numerical value of R above which (28) is valid because R increases with decreasing value of Q. The lowest value of Q is found for $K = \infty$, i.e. outside the multichannel capillary the liquid is always in equilibrium with the ambient. Thus, if Q is put equal to 1 for a given set of values of D, α, L and k, and the rate above which eq. (28) holds is calculated, this rate is an upper limit and its real value will be lower for $K \neq \infty$.

Inserting unfavourable values in (30), i.e. low values of αL and k, and taking the value of $D = 10^{-5}$ cm²/s, a relatively high value of R will be found. For instance when $\alpha = 1$, $L = 0.1$ cm and $k = 10^{-5}$, it is found that eq. (28) may be used for $R \geqslant 1.6 \times 10^{-3}$ cm/s (= 5.8 cm/hr). For more realistic values ($\alpha = 10$, $L = 1$ cm, $k = 10^{-4}$) the situation given by (28) is already obtained for $R > 1.4 \times 10^{-5}$ cm/s (= 5×10^{-2} cm/hr).

In conclusion it may be said that the situation described by (28) is already reached at fairly low rates under reasonable conditions.

Since, at the values of R given by eq. (30), C_L differs less than 1 % from C_{eq} and since $C_{\mathrm{b}} > C_{\mathrm{eq}}$, even if $P_{\mathrm{t}} = P_{\mathrm{p}}$, it may be concluded that the values of R for which ΔN must be calculated with (17a) are extremely low.

In figure 8 both asymptotes and the $\log \Delta N$ vs $\log R$ curve are sketched (drawn lines and curve) and compared with the situation of $L = 0$ (dashed lines and curve). From what has been said above, it is easily seen that the asymptote to which the curve approaches for large values of R has been moved parallel to lower values of R by an amount $\log \alpha$, owing to the introduction of the multichannel cylinder irrespective of the value of L, provided L is not too small ($\geqslant 0.1$ cm). The other asymptote, $\log R = \log R_{\mathrm{b}1}{}^{(\alpha L)}$, has been moved parallel to lower values of R by this introduction. The amount of the displacement, $\log \beta$, is dependent on α as well as on L. Therefore, if a multichannel cylinder with a length L' and a given α is replaced by a cylinder with a length $L'' > L'$ but with the same α, only the asymptote $\log R = \log R_{\mathrm{b}1}{}^{(\alpha L)}$ will move parallel to a lower value of R, and the other will remain in the same position. For $L \to \infty$ the asymptote $\log R_{\mathrm{b}1}{}^{(\alpha L)}$ moves to $\log R_{\mathrm{b}1}{}^{(\alpha L)} = -\infty$. The curve $\log \Delta N$ vs $\log R$ degenerates in this case into the two lines, $\log R = -\infty$ and $\log \Delta N = E - 2 \log \alpha - 2 \log R$. This means that, if a long multichannel cylinder is used, bubble formation only can be avoided by using high growth rates. The parameters listed in table I of part 1 have the same influence on $R_{\mathrm{b}1}{}^{(\alpha L)}$ and $R_{\mathrm{b}2}{}^{(\alpha L)}$ as they have on $R_{\mathrm{b}1}{}^{(0)}$ and $R_{\mathrm{b}2}{}^{(0)}$. The influence of and L on $R_{\mathrm{b}1}{}^{(\alpha L)}$ and $R_{\mathrm{b}2}{}^{(\alpha L)}$ is indicated in part 2 of table I.

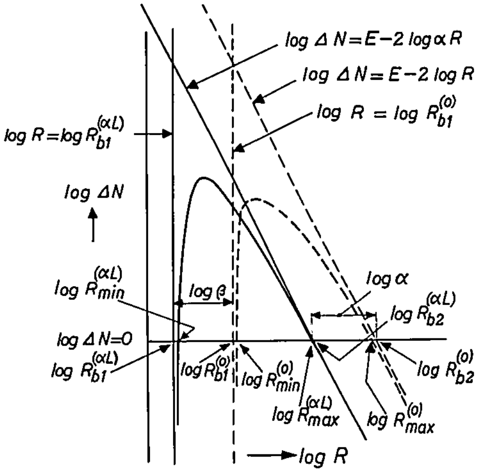


Fig. 8. The curve of $\log \Delta N$ vs $\log R$ at constant values of P_{p} and P_{t} with its two asymptotes (drawn lines and curve) for the case $L \neq 0$ compared with the case $L = 0$ (dotted lines and curve).

The curve $\log \Delta N$ vs $\log R$ intersects the line $\log \Delta N = 0$ twice, viz. at $\log R = \log R_{\mathrm{min}}^{(\alpha L)} > \log R_{\mathrm{b1}}^{(\alpha L)}$ and at $\log R = \log R_{\mathrm{max}}^{(\alpha L)} < \log R_{\mathrm{b2}}^{(\alpha L)}$. Just as in the case of $L = 0$, we can draw the curves of $\log R_{\mathrm{b1}}^{(\alpha L)}$, $\log R_{\mathrm{b2}}^{(\alpha L)}$, $\log R_{\mathrm{min}}^{(\alpha L)}$ and $\log R_{\mathrm{max}}^{(\alpha L)}$ in a $\log P_{\mathrm{p}}$ vs $\log R$ plot at a constant P_{t}, keeping the other parameters constant in accordance with the same rules. In figure 9 the curves I^a, II^a, and III^a (dotted curves) give the situation for $L = 0$ as in fig. 6. The curves I^b, II^b, and III^b give the situation for $L = L_{\mathrm{b}}$ and $\alpha = \alpha_1$, while the curves I^c, II^c, and III^c show the situation for $L = L_{\mathrm{c}}$ ($> L_{\mathrm{b}}$) and the same value $\alpha = \alpha_1$. Bubbles are formed at values of R and P_{p} that give rise to points situated between curve III and $\log P_{\mathrm{p}} = \log P_{\mathrm{t}}$. For the situation $L = L_{\mathrm{c}}$ and $\alpha = \alpha_1$ this only takes place in the shaded region. The larger the value of L the more the minimum in curve III shifts to lower values of $\log R$ and $\log P_{\mathrm{p}}$ and the more it becomes flatter.

Because of the degeneration at $L = \infty$, mentioned above, bubble formation takes place in this case in all situations that give rise to a point above curve II in the $\log P_{\mathrm{p}}$ vs $\log R$ plot at a given P_{t} and $T (= T_1)$, i.e. there is no value of R below which no bubbles are formed for values of P_{p} giving rise to $kC_{\mathrm{b}} < C_{\mathrm{eq}} < C_{\mathrm{b}}$. In practice this situation is already reached when L is a few cm, because the values of R below which no bubbles are formed at this condition are impractically low.

All that has been said about fig. 6, as for instance that only a part of curve III can be realized holds in this case as well. It is also possible to draw the $\log P_{\mathrm{t}}$ vs $\log R$ curves for the case $L \neq 0$ at a constant value of P_{p}, as is done for the case $L = 0$ in fig. 8. Here too, there is a displacement of the curve, giving the

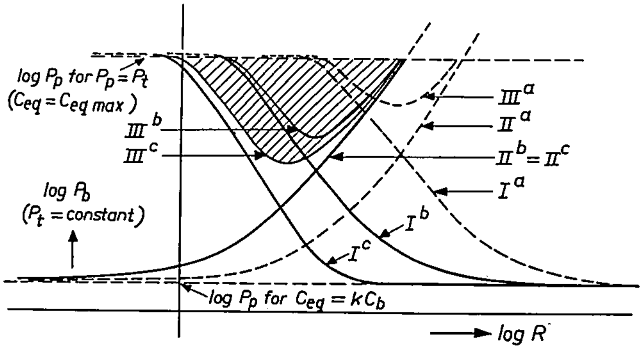


Fig. 9. Variation of $\log R_{\mathrm{b1}}$ and $\log R_{\mathrm{b2}}$ with $\log P_{\mathrm{p}}$ at constant P_{t}.
Curves I^a and II^b for $L = 0$ ($R_{\mathrm{b1}}^{(0)}$ and $R_{\mathrm{b2}}^{(0)}$).
$$\left.\begin{array}{l}\text{Curves I}^a\text{ and II}^b\text{ for } L = L_{\mathrm{b}} \text{ and } \alpha = \alpha_1 \\ \text{Curves I}^c\text{ and II}^c\text{ for } L = L_{\mathrm{c}}\ (> L_{\mathrm{b}}) \text{ and } \alpha = \alpha_1\end{array}\right\}\ (R_{\mathrm{b1}}^{(\alpha L)} \text{ and } R_{\mathrm{b2}}^{(\alpha L)}).$$
Curves III^a, III^b and III^c give the corresponding variation of R_{min} and R_{max} for these cases.
In the situation $L = L_{\mathrm{c}}$ and $\alpha = \alpha_1$ gas bubbles can only be expected within the shaded area.

value of $R_{\min}^{(\alpha L)}$ and $R_{\max}^{(\alpha L)}$ with respect to the situation $L = 0$, but no new insight is obtained from these curves.

Summarizing, it can be said that the introduction of multichannel cylinder into the liquid ahead of the solidification front results in a decrease of the value of $R_{\min}^{(0)}$, the rate below which no bubbles are formed, and of the value of $R_{\max}^{(0)}$, the rate above which no bubbles are formed. The maximum value of the factor by which $R_{\max}^{(0)}$ is decreased is α, and the range of values of R in which this maximum decrease is obtained extends to lower values of R the larger the value of L.

3.3. *Bubble formation and bubble-free growth in case (2)*

In this case too, the liquid contains a volatile solute. The liquid is solidified in the set-up shown in fig. 10 and experiments are carried out in an ambient consisting of the volatile solute with a desired partial pressure and an inert gas. The total gas pressure is kept constant at a desired value. The liquid is present in a tube closed at the bottom. This tube is filled to a length l with the solidifying

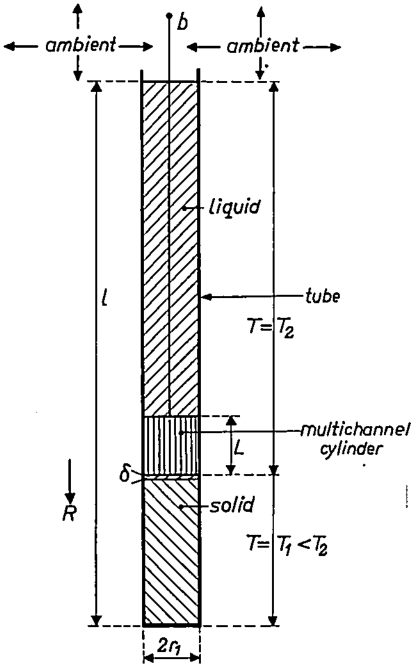


Fig. 10. Bridgman type of solidification with a multichannel cylinder just ahead of the solidification front. The Bridgman tube moves downward at a rate R. The cylinder is kept at a fixed position by means of a rod b. (For details see text.)

liquid. The wall of the tube does not react with the liquid and it is impermeable for the gas phase. The internal diameter of the tube is $2r_1$. In the tube a multichannel cylinder is placed at a position fixed by means of a thin rod b. Its length is L and it contains n channels with a diameter $2r_2$. Its lower end is at a distance δ from the solidification front, δ is supposed to be infinitesimal. The liquid and the cylinder are kept at a uniform temperature T_2 above the solidification temperature of the liquid and the solid is kept at a temperature T_1 which is below that temperature. The tube moves downward at a rate R; the solidification front, however, remains at the position fixed by the multichannel cylinder. The distribution coefficient of the solute k is smaller than 1.

It is easily seen that, if the solute were not volatile and the length of the liquid column l were large enough to allow a steady state to be reached, the concentration gradient ahead of the solidification front would be given by the same equations as in section 3.1, because the presence of the wall of the tube has no influence in this case.

However, because the solute is volatile, we have to take the reaction between the gas phase and the liquid into account and therefore we have to make some assumptions. These are the same as made for the preceding case (section 3.2) except that the second assumption can be omitted, because in this case no radial concentration gradients can be present.

If the liquid is brought into equilibrium with the ambient before the solidification reaction is started, the concentration of the solute in the liquid is given by C_{eq}. During solidification a concentration gradient is built up ahead of the solidification front. If the length l of the liquid column is great enough to allow a steady state to be reached, the concentration at the liquid–gas interface remains C_{eq}. This concentration becomes higher than C_{eq} if the distance between the solidification front and this interface has become of the order of D/R or of $D/\alpha R$, depending on the presence of a multichannel cylinder. In that case the solidification process leaves the steady state. During the steady state the concentration gradient ahead of the solidification process is given by the same equations as given in section 3.1, when C_{w} is replaced by C_{eq}. The number of bubbles that may be formed per unit time is therefore given by

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha}\int_0^L J\mathrm{d}x + S\int_L^{x_b} J\mathrm{d}x \qquad (31\text{a})$$

if $L \leqslant x_b$

or by

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha}\int_0^{x_b} J\mathrm{d}x \qquad (31\text{b})$$

if $L \geqslant x_b$.

In these equations the symbols have the same meaning as in (15a) and (15b). Equations (31a) and (31b) can be rewritten as

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha^2 R} \int_{C_L}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x + \frac{S}{R} \int_{C_b}^{C_L} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x \tag{32a}$$

if $C_b \leqslant C_L$

or

$$\left(\frac{\partial N}{\partial t}\right)_{P_t,P_p,T} = \frac{S}{\alpha^2 R} \int_{C_b}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x \tag{32b}$$

if $C_L \leqslant C_b$.

These expressions are the same as (16a) and (16b) for $Q \to \infty$ $(K \to 0)$.

For the calculation of the total number of bubbles that might be formed during solidification of a bar of given length a and diameter $2r_1$ under steady-state conditions, these equations have to be integrated over the time needed to solidify this length. This time is equal to a/R. However, unlike the case treated in section 3.2, C_b may not be considered as a constant because during solidification the length of the liquid column contributing to the value of P_h and thus to that of C_b decreases linearly with time. This is because it decreases linearly with the fraction of solidified material from $C_b^{(t1)}$ the value of C_b at t_1 at the moment the observation is started to $C_b^{(t2)}$ the value at t_2 the moment the desired length has solidified: $t_2 - t_1 = a/R$ and thus

$$\Delta N_{P_p,P_t,T} = \frac{V_a}{\alpha^2 R^2} \int_{C_L}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x + \\ + \int_{t_1}^{t_1+a/R} \frac{S}{R} \left(\int_{C_b^{(t)}}^{C_L} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x \right) \mathrm{d}t, \tag{33a}$$

where $C_b \leqslant C_L$ during the whole steady-state process and V_a is the solidified volume

$$\Delta N_{P_t,P_p,T} = \int_{t_1}^{t_1+a/R} \frac{S}{\alpha^2 R} \left(\int_{C_b^{(t)}}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, \mathrm{d}C_x \right) \mathrm{d}t, \tag{33b}$$

where $C_L \leqslant C_b^{(t)}$ during the whole steady-state process. If during this process the situation $C_b^{(t)} \leqslant C_L$ changes into $C_b^{(t)} > C_L$, the equation becomes more intricate. For $L = 0$ both equations simplify to

$$\Delta N_{P_t,P_p,T} = \int\limits_{t_1}^{t_1+\alpha/R} \frac{S}{R}\left(\int\limits_{C_b^{(t)}}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \mathrm{d}C_x \right) \mathrm{d}t. \tag{34}$$

Equations (33a), (33b) and (34) cannot be solved. Nevertheless some conclusions can be drawn from them. Because we are interested in a bar of given dimensions solidified under steady-state conditions, and since the upper limit of the first integral in (33a) and that of the integral in (33b) are always equal to C_{eq}/k, there is no growth rate in the steady state below which no bubbles are formed. Of course it is possible to use such low rates that the concentration at the solid–gas interface exceeds C_{eq}, i.e. the solute evaporates to the gas phase at this position. As a consequence the upper limit of the mentioned integral may no longer be equal to C_{eq}/k but to $C_0 < C_{eq}/k$. At a certain rate it may happen that every value of $C_b^{(t)}$ during the solidification process is equal to or larger than C_0. Below that rate, which will be extremely low in most cases, no bubbles will be formed but no steady state will be reached either. Thus in a log ΔN vs log R plot at constant values of all the parameters there is no asymptote $\log R = \log R_{b1}^{(\alpha L)}$ $(\neq -\infty)$ in the steady-state situation, because there is no (positive) value of R below which no bubbles are formed.

For values of R for which $C_L < C_b''$, which is the value of $C_b^{(t)}$ at the moment the bar of desired dimensions has solidified, the log ΔN vs log R curve is completely described by eq. (33b) and in the case $L = 0$ by (34). In order to obtain an impression of the shape of the log ΔN vs log R curves at given values of all the parameters, we determine $\Delta N'$, the number of bubbles that would be formed during the solidification of a bar with desired dimensions if $C_b^{(t)} = C_b'$, i.e. the highest value of $C_b^{(t)}$ at the beginning of the steady-state process, and $\Delta N''$, which is the number of bubbles that would be formed during the process if $C_b^{(t)} = C_b''$ i.e. the lowest value of $C_b^{(t)}$ reached at the moment. the bar has solidified to its desired length. It is easily seen that the real value of ΔN is situated between these two values.

From (33a) and (33b) it follows that

$$\Delta N'_{P_t,P_p,T} = \frac{V_a}{\alpha^2 R^2} \int\limits_{C_L}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \mathrm{d}C_x + \frac{V_a}{R^2} \int\limits_{C_b'}^{C_L} \frac{DJ}{C_x - C_{eq}} \mathrm{d}C_x \tag{35a}$$

if $C_b' < C_L$

or

$$\Delta N'_{P_t,P_p,T} = \frac{V_a}{\alpha^2 R^2} \int_{C_b'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (35b)$$

$$\text{if } C_b' \geqslant C_L$$

and

$$\Delta N''_{P_t,P_p,T} = \frac{V_a}{\alpha^2 R^2} \int_{C_L}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x + \frac{V_a}{R^2} \int_{C_b''}^{C_L} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (36a)$$

$$\text{if } C_b'' < C_L$$

or

$$\Delta N''_{P_t,P_p,T} = \frac{V_a}{\alpha^2 R^2} \int_{C_b''}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (36b)$$

$$\text{if } C_b'' \geqslant C_L.$$

For $L = 0$ eq. (35a) gives rise to

$$\Delta N'_{P_t,P_p,T} = \frac{V_a}{R^2} \int_{C_b'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (37)$$

and eq. (36a) to

$$\Delta N''_{P_t,P_p,T} = \frac{V_a}{R^2} \int_{C_b''}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (38)$$

First we will consider the case $L = 0$. In a $\log \Delta N$ vs $\log R$ plot eqs (37) and (38) give rise to two parallel lines (see fig. 11 curves I′ and I″) whose equations are given by

$$\log \Delta N' = \log V_a - 2 \log R + \log \int_{C_b'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (39)$$

and

$$\log \Delta N'' = \log V_a - 2 \log R + \log . \int_{C_b''}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \qquad (40)$$

The intersection points of these curves with the $\log R$ axis ($\log \Delta N' = 0$,

$\log \Delta N'' = 0$) are given by

$$\log R_{b_2}{}^{(0)\prime} = \frac{1}{2}\left(\log \int\limits_{C_b{}'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, dC_x + \log V_a\right) \tag{41}$$

and

$$\log R_{b_2}{}^{(0)\prime\prime} = \frac{1}{2}\left(\log \int\limits_{C_b{}''}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, dC_x + \log V_a\right). \tag{42}$$

Combination of (41) and (42) gives rise to

$$\log R_{b_2}{}^{(0)\prime\prime} - \log R_{b_2}{}^{(0)\prime}$$

$$= \frac{1}{2}\log\left[1 + \left(\int\limits_{C_b{}''}^{C_b{}'} \frac{DJ}{C_x - C_{eq}}\, dC_x\right)\left(\int\limits_{C_b{}'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}}\, dC_x\right)^{-1}\right]. \tag{43}$$

The distance d between the curves I′ and II″ in fig. 11 is given by

$$\mathrm{d} = \tfrac{1}{2}\sqrt{2}\,(\log R_{b2}{}^{(0)\prime\prime} - \log R_{b2}{}^{(0)\prime}). \tag{44}$$

The intersection point of the real $\log \Delta N$ vs $\log R$ curve with $\log \Delta N = 0$ is given by $\log R_{b2}{}^{(0)}$, for which it must hold that

$$\log R_{b2}{}^{(0)\prime} \leqslant \log R_{b2}{}^{(0)} \leqslant \log R_{b2}{}^{(0)\prime\prime}.$$

This real curve is indicated by curve I in fig. 11 (no conclusions may be drawn from this curve either about its position or about its curvature).
In order to see in which region of a $\log P_p$ vs $\log R$ diagram we may expect bubbles at constant values of the other parameters, we must also keep $C_b{}'$ and $C_b{}''$ constant. This is only possible if the liquid column ahead of the solidification front has the same value in every experiment at the moment the observation of the bubble formation is started after the steady state has been reached. A second condition is that after the desired length a has been solidified, the solidification reaction must still be in the steady state. Under these conditions $C_b{}'$ and $C_b{}''$ and thus

$$\int\limits_{C_b{}''}^{C_b{}'} \frac{DJ}{C_x - C_{eq}}\, dC_x = F$$

are constant and

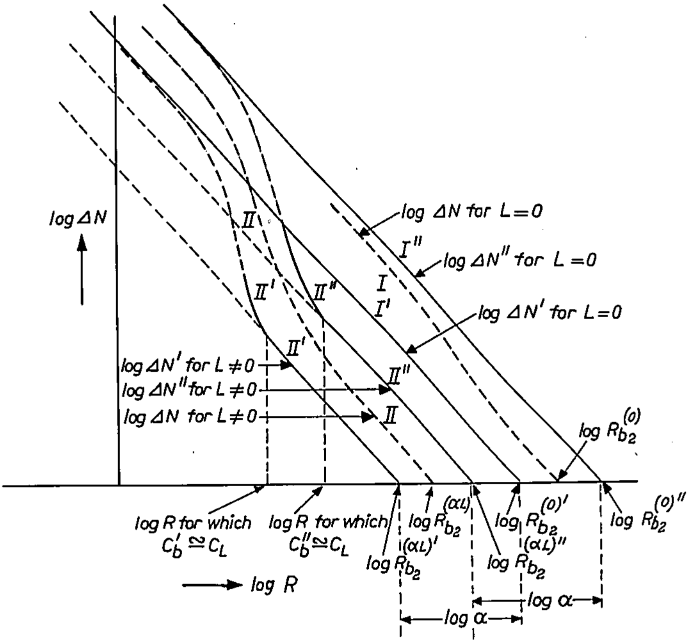


Fig. 11. Variation of log $\Delta N''$, log ΔN and log $\Delta N'$ at constant P_p, P_t and T with log R for the case $L = 0$ (curves I′, I and I″ respectively) and for the case $L \neq 0$ (curves II′, II and II″ respectively) in a Bridgman–type solidification process.

$$\log R_{\mathrm{b}2}{}^{(0)''} - \log R_{\mathrm{b}2}{}^{(0)'} = \tfrac{1}{2} \log \left\{ 1 + F \left[\int\limits_{C_{\mathrm{b}'}}^{C_{\mathrm{eq}}/k} DJ/(C_x - C_{\mathrm{eq}})\, \mathrm{d}C_x \right]^{-1} \right\}. \qquad (45)$$

For high values of

$$\int\limits_{C_{\mathrm{b}}'}^{C_{\mathrm{eq}}/k} \frac{DJ}{C_x - C_{\mathrm{eq}}}\, \mathrm{d}C_x$$

with respect to F, i.e. for high values of P_p, eq. (45) simplifies to $\log R_{\mathrm{b}2}{}^{(0)''} = \log R_{\mathrm{b}2}{}^{(0)'}$. With increasing value of P_p the curve of $\log R_{\mathrm{b}2}{}^{(0)'}$ and that of $\log R_{\mathrm{b}2}{}^{(0)''}$, I′ and I″ in fig. 12, approach each other asymptotically in a $\log P_\mathrm{p}$ vs $\log R$ plot at a constant P_t, but they lose their physical meaning when $P_\mathrm{p} > P_\mathrm{t}$. With decreasing value of P_p the right hand term in (45) increases and becomes infinite at the value of P_p at which $C_{\mathrm{eq}} = kC_{\mathrm{b}}'$. Because $R_{\mathrm{b}2}{}^{(0)'} \leqslant R_{\mathrm{b}2}{}^{(0)} \leqslant R_{\mathrm{b}2}{}^{(0)''}$, the curve of $\log R_{\mathrm{b}2}{}^{(0)}$ moves asymptotically to the curves of $\log R_{\mathrm{b}2}{}^{(0)'}$ and $\log R_{\mathrm{b}2}{}^{(0)''}$ with increasing value of P_p.

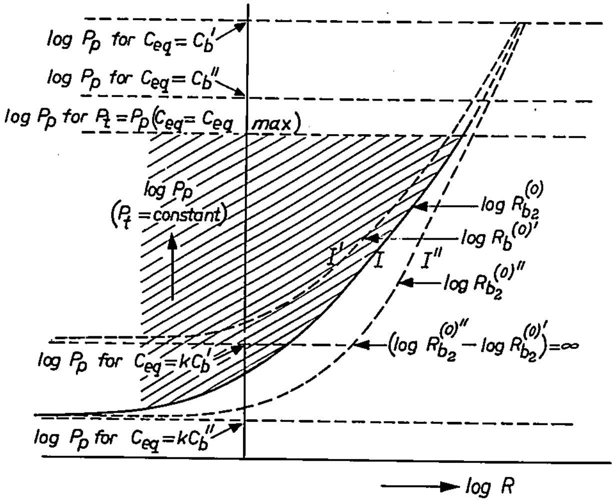

log P_p for $C_{eq}=C_b'$
log P_p for $C_{eq}=C_b''$
log P_p for $P_t=P_p(C_{eq}=C_{eq}$ max)
log P_p
(P_t = constant)
log $R_{b2}^{(0)}$
log $R_b^{(0)'}$
I'
I
I''
log $R_{b2}^{(0)''}$
(log $R_{b2}^{(0)''}$ − log $R_{b2}^{(0)'}$) = ∞
log P_p for $C_{eq}=kC_b'$
log P_p for $C_{eq}=kC_b''$
log R

Fig. 12. Variation of log $R_{b2}{}^{(0)\prime}$ (curve I′), log $R_{b2}{}^{(0)}$ (curve I), and of $R_{b2}{}^{(0)\prime\prime}$ (curve I″) with log R in a log P_p vs log R diagram at constant P_t.

In order to check the shape of the log $R_{b2}{}^{(0)}$ curve in the plot we look at eq. (34). It is easily seen that this equation can only be obeyed for $\Delta N = 1$ if

$$\int_{C_b{}^{(t)}}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, \mathrm{d}C_x > 0,$$

which means that $C_{eq}/k > C_b{}^{(t)}$, independent of the value of t, and thus $C_{eq}/k > C_b''$ because for $C_b{}^{(t)} = C_b''$ the integral has its largest value. As consequence the curve of log $R_{b2}{}^{(0)}$ moves to log $R = -\infty$ if P_p moves to the value at which $C_{eq} = kC_b''$. At lower value of P_p there is no rate at which bubbles are formed.

With the aid of these data the shape of the log $R_{b2}{}^{(0)}$ curve has been drawn in fig. 12, curve I, which is the boundary between the region in which bubbles may be expected (hatched region) and the region in which they may not be expected. It is easily understood that curve I loses its physical meaning at very low values of R, because in that case the constancy of C_b' and C_b'' cannot be maintained as D/R may become larger than the length of the liquid column left after the bar has reached its desired dimensions.

If the value of L for $L > 0$ is sufficiently large (see section 3.2.3) and if the growth rates are high enough, the curve log ΔN vs log R must be situated between the curves

$$\log \Delta N' = \log V_a - 2 \log \alpha - 2 \log R + \log \int_{C_b'}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x \quad (46)$$

(see (35b))

and

$$\log \Delta N'' = \log V_a - 2 \log \alpha - 2 \log R + \log \int_{C_b''}^{C_{eq}/k} \frac{DJ}{C_x - C_{eq}} \, dC_x, \quad (47)$$

(see (36b))

which are straight lines. At low growth rates, however, the distance to the solidification front where $C_x = C_b''$ becomes larger than L, and at the value of $\log R$ belonging to this situation, $\log \Delta N''$ leaves the straight line given by (47). At still lower values of R the distance to solidification front at which $C_x = C_b'$ will also become larger than L, and at the value of $\log R$ belonging to this situation, $\log \Delta N'$ leaves the straight line given by (46). It is easily seen that for $\log R \rightarrow -\infty$, $\log \Delta N'$ approaches to (39) and $\log \Delta N''$ to (40). As the real $\log \Delta N$ vs $\log R$ curve following from (33a) and (33b) must be situated between the curves of $\log \Delta N'$ and $\log \Delta N''$, the situation becomes as sketched in fig. 11, where curve II ($\log \Delta N$) lies between curve II′ ($\log \Delta N'$), and II″ ($\log \Delta N''$). The values of $\log R$ at which the curves II′ and II″ practically coincide with the curve I′ and I″ respectively will be situated at such low values of R that hardly any steady-state solidification can be expected with given values of C_b' and C_b''.

The curves II′ and II″ intersect the line $\log \Delta N = 0$ at the points $\log R_{b2}^{(\alpha L)'}$ and $\log R_{b2}^{(\alpha L)''}$. The intersection point of curve II with this line, $\log R_{b2}^{(\alpha L)}$, is situated between these points. The dependence of $\log R_{b2}^{(\alpha L)'}$, $\log R_{b2}^{(\alpha L)''}$ and $\log R_{b2}^{(\alpha L)}$ on P_p at a constant P_t follows the same rules as those applicable to the corresponding quantities for $L = 0$, and the situation becomes as sketched in fig. 13. The curve of $\log R_{b2}^{(\alpha L)}$ (curve II) has shifted to lower values of R compared with the curve of $\log R_{b2}^{(0)}$ (curve I). For high values of R this shift is practically equal to $\log \alpha$, for low values of R it is smaller. Bubble formation can only be expected within the hatched region, which is very similar to that in case (1) for $L = \infty$. For the same reasons as given for curve I in fig. 12, curve II loses its physical meaning at very low values of R.

It is of course also possible to construct the region in which bubble formation may be expected in a $\log P_t$ vs $\log R$ diagram at a constant P_p, which can be compared with the result given in fig. 8 for case (1). In this case, however, C_b' and C_b'' cannot be kept constant because they depend on P_t, but there is a

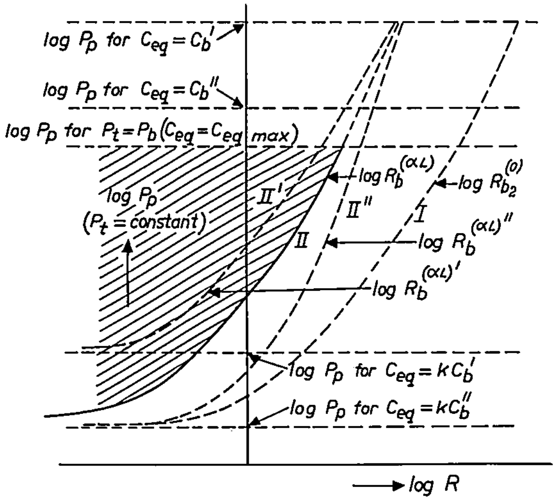


Fig. 13. Variation of log $R_{b2}^{(\alpha L)'}$ (curve II′), log $R_{b2}^{(\alpha L)}$ (curve II), and log $R_{b2}^{(\alpha L)''}$ (curve II″) with log R in a log P_p vs log R diagram at a constant P_t. The curves are compared with the curve of the variation of log $R_{b2}^{(0)}$ with log R (curve I).

simple relationship between these quantities and P_t, because of assumption (1) in section 3.2.

4. General considerations of the two model processes

From the model process described in section 3.2 it is clear that there are several ways in which it is possible to avoid the formation of gas bubbles during solidification of a bar of desired dimensions under steady-state conditions.

(a) The values of P_t and P_p can be chosen such that $P_p < P_{p\,min}$ and there is no rate at which bubbles are formed (e.q. take care that $C_b \geqslant C_{eq}/k$). It may be, however, that the desired phase is not formed under these conditions or that this phase is formed with an undesired concentration of the volatile component. Moreover, this choice of P_t and P_p might involve impractical experimental conditions. In such cases it is necessary to use another possibility.

(b) Low growth rates may be used. These may be so low, however, that a bar of desired dimensions and composition cannot be grown in a reasonable time without the use of a multichannel cylinder. The use of a multichannel cylinder reduces these rates to even more inconvenient values.

(c) High growth rates may be used. If no multichannel cylinder is used, these rates may be so high that no reversible solidification of the liquid with a planar solidification front is possible. The use of such a cylinder can reduce rates to values at which such a solidification is possible, depending on the value of α.

The model process described in section 3.3 has an additional parameter, the length of the liquid column ahead of the solidification front. For large lengths the value of C_b'' at a given P_t may become so high that $C_b'' \geqslant C_{eq/k}$ and no bubbles will be formed at any rate under steady-state conditions.

In this process bubbles cannot be avoided by using very low rates, but it is possible at high rates, which can be reduced by a factor α if a multichannel cylinder is used.

5. Comparison of experimental results and theory

Although the experimental set-up described in section 2 and given in fig. 1 deviates markedly from that of the model process described in section 3.2 (see fig. 3), the difference in the experimental conditions is rather small. Because the multichannel capillary used has a length $L = 40$ mm and the liquid in the crucible was brought into equilibrium before the solidification reaction started, C_L may be put equal to C_{eq} at the growth rates used and thus $C_b > C_L$. The liquid layer on top of that capillary does not have infinitesimal thickness, but it is very thin (< 0.1 mm). Because the capillary is made of metal, the temperature differences along it are small, which means that the larger part of the temperature gradient is present in the liquid layer. This causes the supersaturation at the solidification front to be lower than it would be if the layer were of infinitesimal thickness. Together with the large ratio between the radius and the thickness of this layer, this has the consequence of making the evaporation of O_2 from this layer negligible. Inside the capillary there are no radial gradients.

There are, however, significant deviations from the model process. In the first place, oxygen is present in the gas phase as O_2 and it is not likely that it will dissolve in the liquid phase as O_2. In the second place it is not possible to define a simple distribution coefficient k for the concentration of oxygen in the liquid and solid phase in equilibrium with each other, particularly since in our experiments three solid phases grow at the same time from the liquid phase. This means that we have to describe the distribution of oxygen over the liquid and a heterogeneous solid by at least three solid–liquid equilibria. If, however, the surplus of oxygen over the stoichiometric composition in each of the solid phases in equilibrium with the liquid is small, each of these equilibria may be described by a single distribution coefficient [9,10]. This gives rise to an average distribution coefficient $\bar{k}$, which governs the expulsion of oxygen at the solidification front into the liquid phase. This average distribution coefficient is the ratio of the average surplus of oxygen in the heterogeneous solid over the average content of the stoichiometric solid phases to the surplus of oxygen in the liquid over the same average oxygen content of the stoichiometric solid phases. The concentration of this surplus of oxygen in the liquid is proportional to the square root of P_{O_2} in this case, provided the applied P_{O_2} is large with respect to the P_{O_2} of each of the stoichiometric solid phases at the solidification temperature of the liquid, and provided the oxygen is present in the melt as atoms. If these conditions are fulfilled the boundary curve between the region within which bubbles may be expected and the region within which they are not be expected, given in a $\log P_p$ vs $\log R$ plot at a constant P_t, has the same shape as that in fig. 9. The difference is that k must be replaced by $\bar{k}$ and that $\log P_p$ is not proportional to C_{eq} but to C_{eq}^2, which is only a scale factor in the logarithmic plot, and this is not an essential difference. If they are not fulfilled we have to account for a value of $\bar{k}$ and a scale factor for the conversion of $\log C_{eq}$ into $\log P_p$ that vary with R. But this will not affect the general shape of the boundary curve.

Because the value of L is 4 cm and α is approximately 15 in our experiments, $D/\alpha R \cong 1.7 \times 10^{-7}$ cm/s, which means that even at the lowest value of R used in our experiments the boundary curve coincides with the $\log R_{b2}{}^{(\alpha L)}$ curve.

The experimental curve of fig. 2 is qualitatively in good agreement with the theoretical curve for large values of L (curve II^b in fig. 9) but the displacement of this curve with respect to the case where no multichannel cylinder is used is not proved, and we cannot do the same experiment without a multichannel cylinder. The Bridgman experiment given in section 2, however, provides evidence on this aspect. This experiment was carried out without a multichannel cylinder ($L = 0$). Therefore we have to look at the curve of $\log R_{b2}{}^{(0)}$ in a $\log P_p$ vs $\log R$ diagram at a constant total pressure, which is situated between the curves of $\log R_{b2}{}^{(0)\prime}$ and of $\log R_{b2}{}^{(0)\prime\prime}$. Because the length of the Bridgman tubes used was relatively small, the contribution of the liquid column to the value of C_b' and C_b'' is small compared with that contributed by P_t, which means that, under the conditions of the Bridgman experiment in section 2,

$$\frac{C_b' - C_b''}{C_b} \ll 1.$$

As also P_{O_2} (0.2 atm) is not too small compared with P_t (1 atm) the value of $R_{b2}{}^{(0)}$ at these values of P_p and P_t will be given to a good approximation by (27), and therefore the value of $R_{b2}{}^{(\alpha L)}$ of the E.F.G. experiment in section 2 at the same values of P_p and P_t will be about a factor α ($\cong$ 15) lower than $R_{b2}{}^{(0)}$ of the Bridgman experiment.

Thus, in order to grow the same material at a given P_p and P_t in a Bridgman tube without bubble formation, the growth rate must be a factor of 15 higher than in the E.F.G. technique.

From fig. 2 it is seen that in air the growth rate above which no bubbles are formed with the second technique is about 0.5 mm/min, and thus with the Bridgman technique this rate would become 7.5 mm/min. As we used a rate of 5 mm/min bubbles must be formed during growth, which is in accordance with our observation.

6. Other consequences of the use of a multichannel cylinder

The introduction of a multichannel cylinder ahead of the solidification front has some other consequences, besides the decrease of the growth rate above which no bubbles are formed.

(a) The maximum value of the concentration at the solid–liquid interface C_{eq}/k is already reached at values of R that are roughly a factor α lower than in the case where no cylinder is used. If the right value of α is chosen, reversible growth of the solid with a higher content of the volatile solute than is in equilibrium with the applied partial pressure is possible at moderate values of R. The virtual saturation pressure at the solidification front, i.e. the equilibrium pressure that would give rise to the same concentration at that location under equilibrium conditions, may be even higher than the applied total pressure. This suggests that it might be possible, as already indicated in section 4(a), to grow a solid bar of a composition which, under equilibrium conditions, could only be grown in unfavourable experimental circumstances. In the case of the experimental example in section 2 the equilibrium pressure of O_2 at the growth temperature (ca. 1350 °C) must be high, > 1 atm, because the electrical resistance of the material must be high. Because the RF-heated crucible is the heat source it has a higher temperature than the capillary, but at this higher temperature the rhodium may oxidize to a volatile oxide which condenses on the cooler places, i.e. also on the growing bar, which results in a contamination of that bar and damages the crucible. This oxidation certainly takes place in an ambient of N_2 containing more than 10 vol % O_2 at a total pressure of 1 atm. This leads to the conflicting requirements that the P_{O_2} must be > 1 atm in order to obtain the desired electrical resistance but < 0.1 atm in order to prevent oxidation of the Rh. The use of a multichannel cylinder allows these requirements to be met because of the virtual P_{O_2} at the solidification front, which is already high enough if an ambient of 1 atm N_2 + 2.5 vol % O_2 is used, in which the oxidation of Rh is negligible. Similar circumstances may be encountered in other solidification reactions.

(b) If the cylinder is made of a material that is a good heat conductor the top will be an isothermal plane. If the temperature is regulated such that the solidification front is very close to the top, it will have the same shape as the top. In such a way very planar solidification fronts may be obtained, as is the case in the example treated in section 2.

(c) During the solidification of a solid phase containing a solute with a distribution coefficient $k < 1$, a concentration gradient is built up ahead of the solidification front which may give rise to a supersaturated region in the liquid owing to constitutional supercooling [11]). As the nucleation of a solid phase is governed by the same laws as the nucleation of gas bubbles, the nucleation probability of a solid phase in the supersaturated region during steady-state solidification will also depend on the growth rate; in other words, above a certain rate no nucleation will take place ahead of the solid liquid interface. This rate will be reduced by the use of a multichannel cylinder. This holds not only for solidification of a single-phase material containing a solute but also for the solidification of more than one phase at the same time if, under steady-state conditions, the average composition of the solid growing from the liquid differs from the composition of that liquid [12]). If the top of the cylinder is very flat and the liquid layer on it is very thin, the isothermal surfaces are planar perpendicular to the axis of the cylinder and thus the heat flow is parallel to that axis and so too is the growth direction. This means that after the solidification reaction has started only those grains survive during further solidification whose preferential growth direction is parallel to the axis of the growing bar. The others grow out of the bar (competitive growth). This holds even if the solidification front is not planar because of dentritic growth; the tips of these dentrites will be situated in an isothermal plane. As a result, at rates above which no solid nuclei are formed ahead of the solidification front only the preferred crystallographic axes are present in the solid bar. In the example, treated in section 2 the $\langle 100 \rangle$ directions of the perovskite phase ($BaTiO_3$) and the spinel phase $\{(CeFe_2O_4)_x(Co_2TiO_4)_{1-x}\}$ and the $\langle 205 \rangle$ direction of the magnetoplumbite phase ($BaFe_3Co_{4.5}Ti_{4.5}O_{19}$) are exactly parallel to each other and to the growth direction over the whole surface of a transverse section, even if the spinel phase has grown dentritically.

Conclusions

According to the model used in this paper there are three ways to prevent the formation of gas bubbles ahead of a solidification front in a liquid containing a volatile component.

Firstly the value of the partial pressure of the volatile component in the gas phase P_p, and that of the total pressure (volatile component + inert gas) P_t, may be chosen such, that the equilibrium pressure of the volatile component at the concentration just ahead of the solidification front is lower that the total pressure, independent of the growth rate R.

Secondly the growth rate may be chosen so low that the equilibrium pressure of the volatile component at the concentration at the solidification front is lower than the total pressure.

These two possibilities are generally known.

There is a third possibility. At high growth rates the concentration gradient of the volatile component becomes steep. Consequently the region ahead of the solidification front in which the equilibrium pressure of the volatile component is higher than the total pressure becomes so narrow that above a certain rate gas bubbles do not have time to form. This effect can be enhanced by the use of a multichannel cylinder placed ahead of the solidification front, which reduces the value of the growth rate above which no gas bubbles are formed. Depending on the system chosen and the material of the cylinder, other advantages may arise from the use of such a cylinder, such as low value of P_p, a planar solidification front and prevention of the nucleation of solid phases ahead of this front.

Acknowledgement

The author is indebted to J. Liebe for his assistance in the experiments and to R. M. van Essen for valuable discussions about the mathematics of the subject.

Philips Research Laboratories *Eindhoven, March 1977*

REFERENCES

[1] J. van den Boomgaard, D. R. Terrell, R. A. J. Born and H. F. J. I. Giller, J. Mat. Sci. **9**, 1705, 1974.

[2] J. van den Boomgaard and A. M. J. G. van Run, Proceedings of the Conference on In Situ Composites II, Sept. 25, 1975, Bolton Landing, Lake George, N.Y.

[3] H. E. Labelle and A. J. Mlavski, Mat. Res. Bull. **6**, 581, 1971.

[4] F. H. Cocks, J. T. A. Pollock and J. S. Barley, Proceedings of the Conference of In Situ Composites, Sept. 5–8, 1972, Lakeville, Connecticut, Vol. 1, pp. 141-152.

[5] W. R. Wilcox and V. H. S. Kuo, J. Cryst. Growth **19**, 221-228, 1973.

[6] J. P. Hirth, G. M. Pound and G. R. St. Pierre, Met. Trans. I, 939, 1970.

[7] J. P. Hirth and G. M. Pound, in Progr. in Mat. Sci. Vol. 11, Condensation and Evaporation, MacMillan, New York, 1963. ch.F.

[8] L. D. Landau and E. M. Lifshitz, Statistical Physics, 2nd ed., Addison-Wesley, Reading, Mass., 1969, pp. 460-462.

[9] J. van den Boomgaard, Philips Res. Repts **11**, 27-44, 1956.

[10] J. van den Boomgaard, Philips Res. Repts **11**, 95-102, 1956.

[11] W. A. Tiller, K. A. Jackson, J. W. Rutten and B. Chalmers, Acta Met. **1**, 428, 1953.

[12] J. van den Boomgaard, Met. Trans. **4**, 1485, 1973.

Philips J. Res. ***33****, 186-202, 1978* *R 988*

AN EMPIRICAL FORMULA FOR THE CALCULATION OF LATTICE CONSTANTS OF OXIDE GARNETS BASED ON SUBSTITUTED YTTRIUM- AND GADOLINIUM-IRON GARNETS

by B. STROCKA, P. HOLST*) and W. TOLKSDORF

Abstract

An empirical formula is proposed which allows the calculation of lattice constants of cubic oxide garnets by means of the compositional parameters, the distribution of the cations on the different sets of lattice sites and the corresponding cation radii. The formula is developed assuming merely cation size effects on lattice constants for cation radii in the ranges 0.97 Å $< r^{\mathrm{VIII}} <$ 1.14 Å for cations on dodecahedral lattice sites, 0.54 Å $< r^{\mathrm{VI}} <$ 0.79 Å for Fe^{3+}-substituting cations on octahedral lattice sites and 0.28 Å $< r^{\mathrm{IV}} <$ 0.49 Å for Fe^{3+}-substituting cations on tetrahedral lattice sites. Own lattice constant data and lattice constants from the literature of different series of substituted garnets in the yttrium–iron-garnet (YIG) system and the gadolinium–iron-garnet (GdIG) system have been used for the derivation of the formula. The radii of the cations on dodecahedral sites and of the Fe^{3+}-substituting cations on octahedral and tetrahedral sites have been calculated by further empirical relations for the garnet system. The agreement of measured with calculated lattice constants is about 0.01 Å for the investigated garnet compositions.

1. Introduction

The crystallographic, magnetic and optical properties of garnets are extensively dependent on their compositions. Since the lattice constants, are affected by the composition they can be used as an indication or as reference of many other physical parameters. This might be the reason that many proposals have been made for the calculation of garnet lattice constants.

A linear relationship for the determination of lattice constants of several silicate garnets in dependence of the radii of the divalent and trivalent cations has been developed by McConnel [1]) using Ahrens' radii [2]) for sixfold coordination. Novak et al. [3]) derived an equation by regression analysis from the lattice constants of numerous silicate garnets which relates the lattice constants to the mean radius of the dodecahedral cation and the mean radius of the octahedral cation using the effective radii of Shannon and Prewitt [4]). The model of cluster

*) Present address: Philips GmbH - EWI, Kassel, West-Germany

components (MCC) has been applied by Talanov et al. [5-7] to calculate the lattice parameters of garnet solid solutions. The formulas given by Suchow et al. [8] are based on the lattice constants of basic garnet compositions using Ahrens' [2] radii. In their recent publications [9-11] the Ahrens' radii have been replaced by the "IR" radii of Shannon and Prewitt [4,12] and the revised data of Shannon [13], respectively, taking into account the coordination numbers. Assuming a linear relation between lattice parameter and compositional parameter lattice constant calculations have been performed by Glass et al. [14] for the estimation of lattice expansion of yttrium–iron-garnet LPE films caused by lead incorporation. Similar calculations have been applied for the explanation of lattice constant variations in gadolinium–gallium garnets [15-16].

Some equations of this work have been already published [17]. The equations have been used successfully for lattice parameter calculations of epitaxial garnet films [18] and garnet substrate materials [19,20], respectively, or for the confirmation of Pb valence in iron garnets [21]. All equations presented in this paper have been developed by means of the cation radii published by Shannon and Prewitt in 1969 and 1970 [4,12], although some of them have been revised by Shannon in 1976 [13]. On account of more available lattice constant data the formerly used cation radii in refs 18 and 20 have been slightly varied in some cases.

The lattice constants published by Winkler et al. [22] of polycrystalline garnet materials which have been prepared in our laboratory and the corresponding algebraic equations for the dependence of the lattice constants on the compositional parameters are an essential part of this publication.

We postulate that lattice constants are only ascribed to cation sizes, cation distribution and composition excluding any variations from cubic symmetry by site ordering [23,24].

2. Experimental

The preparation of the samples certainly can influence the lattice constant [17] and other properties of the material [25] since impurities, vacancies, valence state, site distribution etc. can be different.

Polycrystalline materials for our measurements were prepared according to standard ceramic techniques [22,25]. For data of materials from other sources the references are quoted in table I.

Single crystals used in this work, were grown from $PbO/PbF_2/B_2O_3$ high-temperature solutions by slow cooling [26,27]. The main impurities are Pb^{2+} and F^- which are in the order of 0.01 per formula unit [25,27]. Single crystals of gallium and aluminium garnets were also prepared by the Czochralski method [28].

The lattice constants of the single crystal materials of CaGe-, CaSn-, Al- and Sc-substituted YIG and the Ga-substituted GdIG single crystals were

TABLE I

Coefficients C_1 to C_3 of the equation $a = C_1 + C_2\,u + C_3\,u^2$

system	coefficients			range	ref.
	C_1	C_2	C_3		
CaV–YIG	12.37494	+ 0.08823	— 0.04270	$0 \leqslant x \leqslant 0.5$	22
CaV–YIG *)	12.37636	+ 0.05623	+ 0.00176	$0 \leqslant x \leqslant 1.5$	32
CaGe–YIG	12.37632	+ 0.00257	— 0.00645	$0 \leqslant x \leqslant 2.2$	22
CaGe–YIG *)	12.38071	— 0.01779	— 0.00129	$0 \leqslant x \leqslant 3.0$	42
CaSi–YIG	12.37726	— 0.06776	— 0.01370	$0.1 \leqslant x \leqslant 1.5$	22
CaZr–YIG	12.37646	+ 0.16021		$0.1 \leqslant x \leqslant 2.0$	22
CaSn–YIG	12.37465	+ 0.13184		$0.1 \leqslant x \leqslant 1.0$	22
CaSn–YIG *)	12.37771	+ 0.12308		$0 \leqslant x \leqslant 2.0$	33
CaTi–YIG	12.37509	+ 0.03087		$0.1 \leqslant x \leqslant 1.0$	22
Al–YIG	12.37573	— 0.06502	— 0.00175	$0.33 \leqslant x \leqslant 5.0$	34, 35
Al–YIG *)	12.37318	— 0.07168	+ 0.00081	$0 \leqslant x \leqslant 4.01$	36, 45, 46
Ga–YIG	12.37816	— 0.01707	— 0.00054	$0.2 \leqslant x \leqslant 5.0$	34, 37, this work
Ga–YIG [1,*]	12.37782	— 0.01506	— 0.00106	$0 \leqslant x \leqslant 5.0$	this work
CaSb–YIG	12.37603	+ 0.13586		$0.5 \leqslant x \leqslant 1.5$	38
In–YIG	12.37756	+ 0.12012		$0.05 \leqslant x \leqslant 0.8$	22, 34, 39, 40
Sc–YIG	12.37540	+ 0.08156		$0.25 \leqslant x \leqslant 1.5$	34, 35
La–YIG *)	12.37532	+ 0.12149		$0.1 \leqslant x \leqslant 0.6$	22
CaTh–YIG *)	12.37676	+ 0.12942		$0.1 \leqslant x \leqslant 1.0$	22
Bi–YIG *)	12.37560	+ 0.08280		$0 \leqslant x \leqslant 1.0$	43
Al–GdIG	12.47208	— 0.06447	— 0.00150	$0 \leqslant x \leqslant 5.0$	41, 47, 48
CaSi–GdIG	12.47151	— 0.09225	— 0.01627	$0 \leqslant x \leqslant 3.0$	41, 47, 49
GaGe–GdIG	12.47077	— 0.02186	— 0.00933	$0 \leqslant x \leqslant 3.0$	41, 47, 50
Sc–GdIG	12.47025	+ 0.08157		$0 \leqslant x \leqslant 1.5$	41, 47
Ga–GdIG [2]	12.47315	— 0.01048	— 0.00177	$0 \leqslant x \leqslant 5.0$	this work
CaZr–GdIG	12.47182	+ 0.12491		$0 \leqslant x \leqslant 2.0$	41
Bi–GdIG *)	12.47200	+ 0.05071		$0 \leqslant x \leqslant 1.4$	44

[1] only single crystals, lattice constant determinations by Bond's method [29]
[2] only single crystals, lattice constant determinations by film-method [22]
*) neglected in derivations of eqs (2) - (14).

determined after pulverization by the earlier described method [22], whereas Syton *) polished cuts of YIG and Ga-substituted yttrium–iron-garnet single crystals have been measured by the method described by Bond [29] using a double-crystal diffractometer [30]. The double-crystal diffraction measurements were made with a Siemens–Omega goniometer using Cu $K\alpha_1$ radiation and a (111) cut of a highly polished dislocation free gadolinium–gallium-garnet single crystal as monochromator crystal in symmetric (888) Bragg position. The accuracy of these measurements using the half-peak midchord method [31] is about $\Delta a/a = 2\times 10^{-5}$.

*) Product of Monsanto

3. Results

Similar to the formulas which describe the lattice constants in dependence on the compositional parameters x and y as they are given by Winkler et al. [22]) for various polycrystalline yttrium–iron-garnet solid solutions the algebraic expression

$$a = C_1 + C_2\,u + C_3\,u^2 \qquad (1)$$

has been applied on merely one compositional parameter u of lattice constant measurements of various substituted single crystalline and polycrystalline garnets with $R = Y^{3+}$, Gd^{3+} of the general formulas

$R_3Fe_{5-u}M_uO_{12}$ $(M = Al^{3+}, Ga^{3+}, In^{3+}, Sc^{3+})$,
$R_{3-u}Ca_uFe_{5-u}M_uO_{12}$ $(M = Ge^{4+}, Si^{4+}, Zr^{4+}, Sn^{4+}, Ti^{4+})$,
$R_{3-2u}Ca_{2u}Fe_{5-u}M_uO_{12}$ $(M = V^{5+}, Sb^{5+})$ and
$R_{3-u}R'_uFe_5O_{12}$ $(R' = Ca^{2+}_{0.5}Th^{4+}_{0.5}, Bi^{3+}, La^{3+})$.

Coefficient C_1 is identical with the lattice constant of unsubstituted yttrium–iron garnet (YIG) or gadolinium–iron garnet (GdIG), respectively. Coefficients C_1 to C_3 of eq. (1) have been determined by the least-squares method for numerous lattice constants of various substituted garnets in the YIG-system and the GdIG-system, respectively, which have been measured by the described methods [22,30]) and completed by lattice constant data from the literature [32–50]).

A compilation of the calculated coefficients C_1 to C_3 and those found by Winkler et al. [22]) is given in table I. Since the lattice constant data which belong to the sets of coefficients marked with an asterisk in table I were available after completion of the following calculations they are not taken into account for the computations. These data are tabulated in table I for the purpose of comparison to our results or completion of table I. With the exception of the coefficients C_1 to C_3 for Ga-substituted GdIG all other data under consideration in the following calculations are based on ceramic materials. The dependences of the lattice constants a for these data on the compositional parameter u are presented graphically in figs 1 and 2.

Optimal fitting by linear expressions for the lattice constant as function of the compositional parameter u have been found for

CaZr–YIG, CaSn–YIG, CaTi–YIG, CaSb–YIG, In–YIG, Sc–YIG, Bi–YIG

and

Sc–GdIG, CaZr–GdIG, Bi–GdIG, respectively,

whereas second-degree approximations have been determined for

CaV-YIG, CaGe-YIG, CaSi-YIG, Al-YIG, Ga-YIG

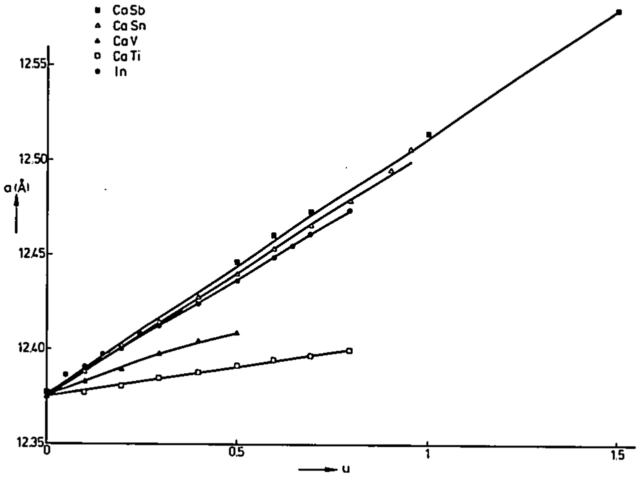


Fig. 1. Lattice constants *a* of substituted YIG vs compositional parameters *u*.

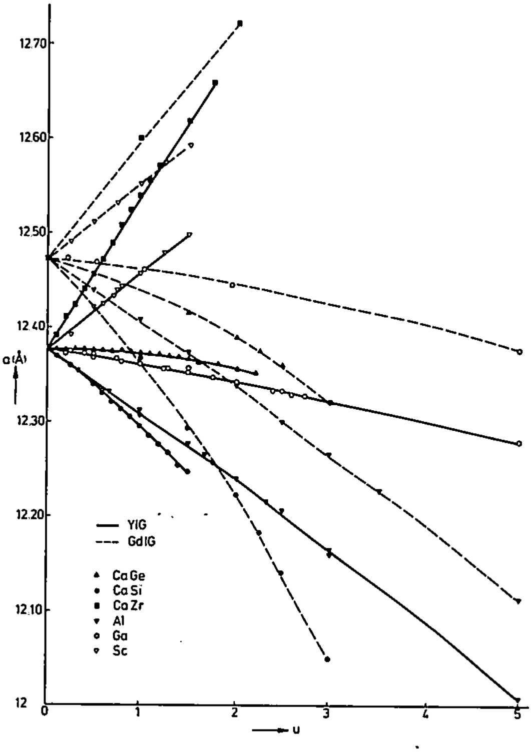


Fig. 2. Lattice constants *a* of substituted YIG and GdIG vs compositional parameters *u*.

and

CaGe-GdIG, CaSi-GdIG, Al-GdIG, Ga-GdIG, respectively.

Referring to the work of Geller [51]) the substituting cations occupy the following lattice sites in the garnet lattice.

c-sites:	Bi^{3+}, Ca^{2+}, Y^{3+}, Gd^{3+}
a-sites:	Sc^{3+}, In^{3+} (preferentially), Ti^{4+} (preferentially), Zr^{4+}, Sb^{5+}, Sn^{4+} (preferentially)
d-sites:	Si^{4+}, Ge^{4+} (preferentially), V^{5+}
a- and d-sites:	Al^{3+}, Ga^{3+}, Fe^{3+}.

Assuming the different slopes $\mathrm{d}a/\mathrm{d}u$ of the fitting curves to be caused by the size of the substituting cation relative to the size of Fe^{3+} on octahedral sites and tetrahedral sites, respectively, the cation radii given by Shannon [13]) — see table III — will lead to the following sequences of $\mathrm{d}a/\mathrm{d}u$ for the investigated garnet compositions.

a-sites: $(\mathrm{d}a/\mathrm{d}u)_{In^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Sc^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Zr^{4+}} > (\mathrm{d}a/\mathrm{d}u)_{Sn^{4+}}$
$> (\mathrm{d}a/\mathrm{d}u)_{Fe^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Ga^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Ti^{4+}} > (\mathrm{d}a/\mathrm{d}u)_{Sb^{5+}}$
$> (\mathrm{d}a/\mathrm{d}u)_{Al^{3+}}$

and

d-sites: $(\mathrm{d}a/\mathrm{d}u)_{Fe^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Ga^{3+}} > (\mathrm{d}a/\mathrm{d}u)_{Al^{3+}} \cong (\mathrm{d}a/\mathrm{d}u)_{Ge^{4+}}$
$> (\mathrm{d}a/\mathrm{d}u)_{V^{5+}} > (\mathrm{d}a/\mathrm{d}u)_{Si^{4+}}$.

The quantitative results of the differentiated expressions of table I are plotted for pure tetrahedral and octahedral Fe^{3+}-substitution for the YIG- and GdIG-system in fig. 3. The derivatives $(\mathrm{d}a/\mathrm{d}u)_{M,u=0}$ are ascribed to uniform substitution of Fe^{3+} by the cation M on tetrahedral or octahedral lattice sites, respectively. In Ga- and Al-substituted YIG and GdIG $(\mathrm{d}a/\mathrm{d}u)_{M,u=0}$ gives the lattice constant variation for substitution on tetrahedral sites whereas $(\mathrm{d}a/\mathrm{d}u)_{M,u=5}$ refers to pure octahedral substitution. If the variation of the lattice constants with composition $\mathrm{d}a/\mathrm{d}u$ is assumed to be additive in the case of simultaneous substitution on dodecahedral and octahedral or tetrahedral lattice sites, respectively, the amount $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}}$ in the case of Ca^{2+}M-substitution ($M = V^{5+}$, Ge^{4+}, Si^{4+}, Zr^{4+}, Sn^{4+}, Ti^{4+}, Sb^{5+}) to $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}M}$ can be estimated by means of the equations in table I and the cation radii given by Shannon [13]). Since Ge^{4+} and Al^{3+} are of the same size in tetrahedral coordination ($r^{IV}_{Ge^{4+}} = r^{IV}_{Al^{3+}} = 0.390$ Å) and Sc^{3+} and Zr^{4+} are of similar size in octahedral coordination ($r^{VI}_{Sc^{3+}} = 0.745$ Å, $r^{VI}_{Zr^{4+}} = 0.720$ Å) the amount $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}}$ to

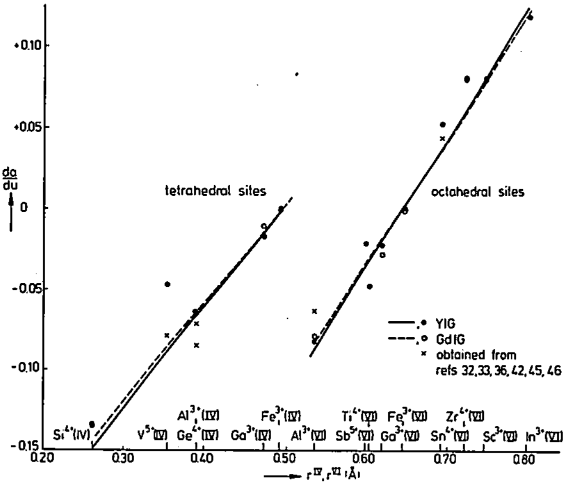


Fig. 3. Slopes da/du of substituted YIG and GdIG vs radii of Fe^{3+}-substituting cation for octahedral and tetrahedral lattice site occupation (coordination numbers in parentheses).

$(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}Ge^{4+}}$ and to $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}Zr^{4+}}$, respectively, may be estimated by the following equations.

Tetrahedral site occupation:

$$(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}} = (\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}Ge^{4+}} - (\mathrm{d}a/\mathrm{d}u)_{Al^{3+}} \tag{2}$$

Octahedral site occupation:

$$(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}} = (\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}Zr^{4+}} - (\mathrm{d}a/\mathrm{d}u)_{Sc^{3+}}. \tag{3}$$

The estimated values of $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}}$ which are assumed to be valid for all other Ca^{2+}M-substitutions are listed in table II. Linear expressions have been found by least-squares method for $(\mathrm{d}a/\mathrm{d}u)_{M}$ in dependence of the radii r_{M}^{IV} and

TABLE II

Increment $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}}$ to slope $(\mathrm{d}a/\mathrm{d}u)_{Ca^{2+}M}$

system	octahedral sites	tetrahedral sites
YIG	+ 0.0787	+ 0.0676
GdIG	+ 0.0433	+ 0.0426

$r_{\mathrm{M}}^{\mathrm{VI}}$ for the Fe^{3+}-substituting cation M in the YIG- and GdIG-system, respectively.

Tetrahedral site occupation:

YIG: $$(\mathrm{d}a/\mathrm{d}u)_{\mathrm{M}} = -0.317 + 0.644\, r_{\mathrm{M}}^{\mathrm{IV}} \tag{4}$$

GdIG: $$(\mathrm{d}a/\mathrm{d}u)_{\mathrm{M}} = -0.306 + 0.623\, r_{\mathrm{M}}^{\mathrm{IV}} \tag{5}$$

Octahedral site occupation:

YIG: $$(\mathrm{d}a/\mathrm{d}u)_{\mathrm{M}} = -0.515 + 0.802\, r_{\mathrm{M}}^{\mathrm{VI}} \tag{6}$$

GdIG: $$(\mathrm{d}a/\mathrm{d}u)_{\mathrm{M}} = -0.501 + 0.781\, r_{\mathrm{M}}^{\mathrm{VI}}. \tag{7}$$

The values $(\mathrm{d}a/\mathrm{d}u)_{\mathrm{M}}$ of the expressions $a = f(u)$ given in table I which are marked with an asterisk are plotted for the purpose of comparison in fig. 3 besides the data which we have used for our calculations. They are not included in the computations. Using the lattice constant data of various ceramic rare-earth–iron, –gallium, and –aluminium garnets from references [47–48,51–55]) and the effective radii formerly given by Shannon and Prewitt [4,12]) we have determined by least-squares method linear expressions for the lattice constants in dependence on the radii of the rare-earth cation R for dodecahedral site occupation:

$\mathrm{R_3Fe_5O_{12}}$: $$a = 10.2090 + 2.1322\, r_{\mathrm{R}}^{\mathrm{VIII}} \tag{8}$$

$\mathrm{R_3Ga_5O_{12}}$: $$a = 10.0419 + 2.2008\, r_{\mathrm{R}}^{\mathrm{VIII}} \tag{9}$$

$\mathrm{R_3Al_5O_{12}}$: $$a = 9.7153 + 2.2577\, r_{\mathrm{R}}^{\mathrm{VIII}}. \tag{10}$$

By means of eqs (4)–(7) and (8)–(10) the radii of the Fe^{3+}-substituting cations and the rare-earths radii have been calculated, respectively. The mean

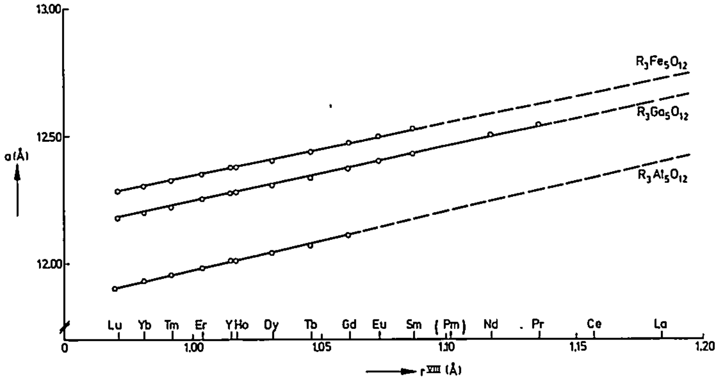


Fig. 4. Lattice constants a of garnets $\mathrm{R_3M_5O_{12}}$ (R = rare earths; M = Fe, Ga, Al) vs rare earths radii for dodecahedral site occupation.

TABLE III

Comparison between calculated radii and radii given by Shannon [13])

ion	dodecahedral {c} sites			octahedral [a] sites			tetrahedral (d) sites		
	r_{calc} (Å)	r(ref. 13) (Å)	Δr (10^{-3} Å)	r_{calc} (Å)	r(ref. 13) (Å)	Δr (10^{-3} Å)	r_{calc} (Å)	r(ref. 13) (Å)	Δr (10^{-3} Å)
Ca^{2+}	1.124 $^{b)}$	1.12 *)	+ 4						
Sr^{2+}	1.240	1.26	—20						
Y^{3+}	1.016 $^{a)}$	1.019 *)	— 3						
La^{3+}	1.190	1.16	+ 30						
Pr^{3+}	1.137	1.126	+ 11						
Nd^{3+}	1.120	1.109 *)	+ 11						
Sm^{3+}	1.087 $^{a)}$	1.079	+ 8						
Eu^{3+}	1.073 $^{a)}$	1.066	+ 7						
Gd^{3+}	1.061 $^{a)}$	1.053	+ 8						
Tb^{3+}	1.044 $^{a)}$	1.040	+ 4						
Dy^{3+}	1.030 $^{a)}$	1.027	+ 3						
Ho^{3+}	1.017 $^{a)}$	1.015	+ 2						
Er^{3+}	1.004 $^{a)}$	1.004	± 0						
Tm^{3+}	0.991 $^{a)}$	0.994	— 3						
Yb^{3+}	0.982 $^{a)}$	0.985	— 3						
Lu^{3+}	0.972 $^{a)}$	0.977	— 5						
Bi^{3+}	1.132 $^{a)}$	1.17	—38						
Th^{4+}	1.095	1.05	+ 45						
Al^{3+}				0.539 $^{b)}$	0.535 *)	+ 4	0.390 $^{b)}$	0.39 *)	± 0
Sc^{3+}				0.745 $^{b)}$	0.745 *)	± 0			
Fe^{3+}				0.642$^{b)}$	0.645 *)	— 3	0.492 $^{b)}$	0.49 *)	+ 2
Ga^{3+}				0.610 $^{b)}$	0.62 *)	—10	0.470 $^{b)}$	0.47 *)	± 0
In^{3+}				0.792	0.800 *)	— 8			
Si^{4+}					0.400 *)		0.279 $^{b)}$	0.26 *)	+ 19
Ti^{4+}				0.582	0.605 *)	—23			
Ge^{4+}					0.53 *)		0.390 $^{b)}$	0.39 *)	± 0
Zr^{4+}				0.745 $^{b)}$	0.72 *)	+ 25			
Sn^{4+}				0.708 $^{b)}$	0.69 *)	+ 18			
V^{5+}					0.54		0.370	0.355 *)	+ 15
Sb^{5+}				0.615	0.60 *)	+ 15			

$\Delta r = r_{calc} - r$ (ref. 13),
*) error $\leqslant$ 0.01 Å from ref. 13,
[a]) error $\leqslant$ 0.001 Å,
[b]) error $\leqslant$ 0.005 Å.

values are listed in table III and compared to the radii given by Shannon [13]). The errors of 0.001 Å (a) and 0.005 Å (b) represent the maximum deviations of r_{calc} when eqs (4) and (6), (5) and (7), and (8)–(10) have been used.

In correspondence to references 56–57 we assume that the average radii for the substitution on a certain lattice site is a function of the mole fraction and postulate that the substituting cations are randomly distributed in a given set of lattice sites. For garnets of the general formula $\{C_{3-x}C'_x\}(A_{2-y}A'_y)(D_{3-z}D'_z)O_{12}$

the average cation radii for dodecahedral site occupation r^{VIII}, for octahedral site occupation r^{VI}, and for tetrahedral site occupation r^{IV} can then be calculated by (11)-(13):

$$r^{\text{VIII}} = r_{\text{C}}^{\text{VIII}} + \frac{x}{3}(r_{\text{C}'}^{\text{VIII}} - r_{\text{C}}^{\text{VIII}}) , \tag{11}$$

$$r^{\text{VI}} = r_{\text{A}}^{\text{VI}} + \frac{y}{2}(r_{\text{A}'}^{\text{VI}} - r_{\text{A}}^{\text{VI}}) \tag{12}$$

$$r^{\text{IV}} = r_{\text{D}}^{\text{IV}} + \frac{z}{3}(r_{\text{D}'}^{\text{IV}} - r_{\text{D}}^{\text{IV}}) \tag{13}$$

If substitution occurs simultaneously on different lattice sites we propose for the dependence of the lattice constant a on the average cation radii r^{VIII}, r^{VI} and r^{IV} and on the compositional parameters x, y, and z, respectively, now considering site occupation, eq. (14):

$$a = b_1 + b_2\, r^{\text{VIII}} + b_3\, r^{\text{VI}} + b_4\, r^{\text{IV}} + b_5\, r^{\text{VIII}}\, r^{\text{VI}} + b_6\, r^{\text{VIII}}\, r^{\text{IV}}, \tag{14}$$

Coefficients b_1 to b_6 have been determined by means of Shannon's and Prewitt's radii [4]) for Y^{3+} and Gd^{3+} in eightfold coordination, for Fe^{3+}, Ga^{3+}, and Al^{3+} in sixfold and fourfold coordination using eqs (4) - (10) and eq. (14). The best fit of calculated lattice constants with measured lattice constants has been found by the following set of coefficients b_1 to b_6:

$$\begin{aligned} b_1 &= +7.02954 \\ b_2 &= +3.31277 \\ b_3 &= +2.49398 \\ b_4 &= +3.34124 \\ b_5 &= —0.87758 \\ b_6 &= —1.38777. \end{aligned}$$

4. Discussion

From the increments $(\mathrm{d}a/\mathrm{d}u)_{\text{Ca}^{2+}}$ to the slope of the corresponding lattice constant-compositional parameter curves as they are given in table II it is obvious that in the case of simultaneous substitution in YIG and GdIG the influence on lattice constant variation increases as well for octahedral site occupation as for tetrahedral site occupation with increasing size difference of the cations in dodecahedral sites which amounts to $r_{\text{Ca}^{2+}}^{\text{VIII}} - r_{\text{Y}^{3+}}^{\text{VIII}} = 0.10$ Å and $r_{\text{Ca}^{2+}}^{\text{VIII}} - r_{\text{Gd}^{3+}}^{\text{VIII}} = 0.07$ Å in these cases. This difference in lattice expansion leads to the intersection of the extrapolated curves which are plotted in fig. 2 in the cases of CaGe–YIG and -GdIG, CaSi–YIG and -GdIG, and CaZr–YIG and -GdIG. The same effect can be seen for pure substitution on dodecahedral sites by Bi^{3+} from the coefficients C_2 (table I) for Bi–YIG and Bi–GdIG.

The agreement of the cation radii recently given by Shannon [13]) and the calculated radii r_{calc} computed by means of (4) – (10) is mostly better than 0.01 Å, see table III. Since the reliability of the most reliable cation radii is no more than 0.01 Å [13]) the tabulated values have to be regarded as calculation parameters to fit eqs (4) – (10). The agreement between both series of cation radii is remarkably good, although it is known that distortion can occur in the garnet lattice and that the degree of distortion is depending on composition [58,59]). Both factors are affecting mean interatomic distances [13]). The ion radii recently published by Shannon [13]) are favoured in table III, e.g. the radii $r^{VIII} = 1.019$ Å for Y^{3+} and $r^{VIII} = 1.015$ Å for Ho^{3+} suit better with measured lattice constants of 12.376 Å [47]) for $Y_3Fe_5O_{12}$ and 12.375 Å [47]) for $Ho_3Fe_5O_{12}$ than those published in 1969 (ref. 4).

The sensitivity of (14) to cation radii variation may be represented by the calculated lattice constants $a = 12.3721$ Å and 12.3804 Å for $Y_3Fe_5O_{12}$, if different radii for Y^{3+} for eightfold coordination $r^{VIII} = 1.015$ Å [4]) and $r^{VIII} = 1.019$ Å [13]) are used ($r^{VI}_{Fe^{3+}} = 0.642$ Å; $r^{IV}_{Fe^{3+}} = 0.492$ Å).

The capability of (14) may be demonstrated by a comparison of systematically chosen lattice constants measured by Winkler et al. [22]) for polycrystalline substituted yttrium–iron garnets with calculated lattice constants using Shannon's radii [13]) and calculated lattice constants by means of the calculated radii determined by ourselves, respectively (see table IV).

TABLE IV

Deviation of calculated lattice constants from observed lattice constants of some polycrystalline substituted yttrium–iron garnets of ref. 22 (standard deviation $\sigma \cong 0.0003$ Å).

composition	x	y	a_{obs} (ref. 22) (Å)	Δa [a]) (10^{-4} Å)	Δa [b]) (10^{-4} Å)
$\{Y_3\}[Fe_2](Fe_3)O_{12}$	0	0	12.3742 ...12.3812	± 0 ...+ 70	— 72 ...— 2
$\{Y_{3-2x}Ca_{2x}\}[Fe_{2-y}In_y](Fe_{3-x}V_x)O_{12}$	0.5	0	12.4085	— 19	— 1
	0	0.3	12.4125	+ 22	— 61
	0.7	0.3	12.4573	— 31	+ 9
	0	0.4	12.4265	+ 42	— 45
	0.8	0.4	12.4764	— 32	+ 23
	0	0.5	12.4383	+ 40	— 51
	0.8	0.5	12.4899	— 13	+ 37
	0	0.6	12.4496	+ 33	— 61
	0.8	0.6	12.5001	— 28	+ 19
	0	0.7	12.4608	+ 25	— 73
	0.8	0.7	12.5152	+ 7	+ 50

to be continued

TABLE IV (continued)

composition	x	y	a_{obs} (ref. 22) (Å)	Δa [a]) (10^{-4} Å)	Δa [b]) (10^{-4} Å)
$\{Y_{3-x}Ca_x\}[Fe_{2-y}In_y](Fe_{3-x}Ge_x)O_{12}$	1.0	0	12.3721	−125	−162
	2.2	0	12.3520	−497	−487
	0	0.3	12.4125	+ 22	− 61
	1.0	0.3	12.4115	− 85	−133
	1.4	0.3	12.4070	−178	−211
	0	0.4	12.4265	+ 42	− 45
	1.0	0.4	12.4240	− 78	−129
	1.4	0.4	12.4194	−171	−208
	0	0.5	12.4383	+ 40	− 51
	1.0	0.5	12.4378	− 58	−119
	1.4	0.5	12.4327	−155	−196
	0	0.6	12.4496	+ 33	− 61
	1.0	0.6	12.4518	− 35	− 95
	1.4	0.6	12.4461	−138	−183
	0	0.7	12.4608	+ 25	− 73
	1.0	0.7	12.4633	− 38	−101
	1.4	0.7	12.4588	−128	−176
$\{Y_{3-x}Ca_x\}[Fe_{2-y}In_y](Fe_{3-x}Si_x)O_{12}$	1.0	0	12.2947	−203	−121
	1.5	0	12.2466	−415	−256
	0	0.4	12.4265	+ 42	− 45
	0.6	0.4	12.3829	− 25	− 19
	0	0.5	12.4383	+ 40	− 51
	0.6	0.5	12.3962	− 10	− 9
	0	0.6	12.4496	+ 33	− 61
	0.6	0.6	12.4117	+ 26	− 24
$\{Y_{3-x}Ca_x\}[Fe_{2-x}Zr_x](Fe_3)O_{12}$	1.0	—	12.5386	+ 91	+288
	1.5	—	12.6182	+123	+450
$\{Y_{3-x}Sr_x\}[Fe_{2-x}Zr_x](Fe_3)O_{12}$	0.4	—	12.4680	− 4	− 33
$\{Y_{3-x}Ca_x\}[Fe_{2-x}Sn_x](Fe_3)O_{12}$	0.5	—	12.4401	+ 25	+ 61
	1.0	—	12.5065	+ 61	+202
$\{Y_{3-x}Ca_x\}[Fe_{2-x}Ti_x](Fe_3)O_{12}$	0.4	—	12.3875	+ 26	− 90
	0.8	—	12.3996	+ 37	−122
$\{Y_{3-x}La_x\}[Fe_2](Fe_3)O_{12}$	0.4	—	12.4225	− 2	+ 23
$\{Y_{3-2x}Ca_xTh_x\}[Fe_{2-y}In_y](Fe_3)O_{12}$	0.5	0	12.4420	+ 34	+152
	0.9	0	12.4926	+ 24	+294
	0.3	0.3	12.4527	+ 42	+ 71
	0.3	0.7	12.4978	+ 17	+ 30

$\Delta a = a_{obs} - a_{calc}$,
[a]) radii of present investigation are used,
[b]) radii of ref. 13 are used.

The wide range of the observed lattice constant a from 12.3742 Å to 12.3812 Å nominally unsubstituted, polycrystalline yttrium–iron garnet is probably caused by variation of the Y : Fe ratios and different divalent iron content due to small, not reproducible variations of the preparation conditions [60,61]).

The mean deviations from measured data of ref. 22 are $\Delta a \cong 0.01$ Å for a certain substitution in both cases. The degree of deviation depends as well on the substituting cations as on the compositional parameters x and y of ref. 22 and cannot be simply explained by the uncertainty of a special cation radius.

TABLE V

Lattice constants of garnet single crystals

composition	a_{obs} (Å)	a_{calc} (Å)	Δa (10^{-4} Å)	a_{calc}(cor.) (Å)	Δa (10^{-4} Å)
$Y_{2.850}Ca_{0.150}Fe_{4.845}Ge_{0.155}O_{12}$	12.3792(6)[1]	12.3758	+ 34		
$Y_{2.660}Ca_{0.340}Fe_{4.660}Ge_{0.340}O_{12}$	12.3781(4)[1]	12.3786	— 5		
$Y_3Fe_{4.77}Sc_{0.23}O_{12}$	12.3966(6)[1]	12.3932	+ 34		
$Y_3Fe_{4.59}Sc_{0.41}O_{12}$	12.4104(5)[1]	12.4081	+ 23		
$Y_3Fe_{4.38}Sc_{0.62}O_{12}$	12.4255(5)[1]	12.4254	+ 1		
$Y_3Fe_{4.20}Sc_{0.80}O_{12}$	12.4400(4)[1]	12.4402	— 2		
$Y_3Fe_{4.885}Ga_{0.115}O_{12}$	12.3757(2)[*,1]	12.3726	+ 31	12.3726[a]	+ 31
$Y_3Fe_{4.750}Ga_{0.250}O_{12}$	12.3720(2)[*,1]	12.3707	+ 13	12.3707[a]	+ 13
$Y_3Fe_{4.610}Ga_{0.390}O_{12}$	12.3725(2)[*,1]	12.3687	+ 38	12.3686[a]	+ 39
$Y_3Fe_{3.850}Ga_{1.150}O_{12}$	12.3599(2)[*,1]	12.3579	+ 20	12.3571[a]	+ 28
$Y_3Fe_{2.580}Ga_{2.420}O_{12}$	12.3355(2)[*,1]	12.3399	— 44	12.3324[b]	+ 31
$Y_3Fe_{2.490}Ga_{2.510}O_{12}$	12.3330(2)[*,1]	12.3387	— 57	12.3303[b]	+ 27
$Y_3Fe_{2.320}Ga_{2.680}O_{12}$	12.3295(2)[*,1]	12.3363	— 68	12.3266[b]	+ 29
$Y_3Ga_5O_{12}$	12.2761(2)[*,1]	12.2805	— 44		
	12.2758(2)[1]	12.2805	— 47		
$Y_3Fe_{4.64}Al_{0.36}O_{12}$	12.3552(4)[1]	12.3506	+ 46	12.3497[c]	+ 55
$Y_3Fe_{4.36}Al_{0.64}O_{12}$	12.3369(3)[1]	12.3322	+ 47	12.3304[c]	+ 65
$Y_3Fe_{4.03}Al_{0.97}O_{12}$	12.3088(2)[1]	12.3105	— 17	12.3072[c]	+ 16
$Y_3Fe_{3.51}Al_{1.49}O_{12}$	12.2780(2)[1]	12.2764	+ 16	12.2703[c]	+ 77
$Y_3Al_5O_{12}$	12.0067(2)[1]	12.0122	— 55		
	12.0073(6)[2]	12.0122	— 49		
$Y_{2.81}Ca_{0.19}Fe_{4.78}Sn_{0.22}O_{12}$	12.3964(3)[1]	12.4006	— 42		
$Y_{2.55}Ca_{0.45}Fe_{4.39}Sn_{0.61}O_{12}$	12.4183(3)[1]	12.4412	—229		
$Gd_{3.0}Fe_{4.78}Ga_{0.22}O_{12}$	12.4728(4)[1]	12.4643	+ 85		
$Gd_{3.0}Fe_{4.47}Ga_{0.53}O_{12}$	12.4679(3)[1]	12.4601	+ 78		
$Gd_{3.0}Fe_{3.96}Ga_{1.04}O_{12}$	12.4601(3)[1]	12.4533	+ 68		
$Gd_{3.0}Fe_{3.05}Ga_{1.95}O_{12}$	12.4455(3)[1]	12.4406	+ 49		
$Gd_3Ga_5O_{12}$	12.3759(3)[1]	12.3761	— 2		
$Gd_{3.03}Ga_{4.97}O_{12}$	12.3814(2)[2]	12.3838[d]	— 24		
$Y_3Fe_5O_{12}$	12.3777(2)[1]	12.3742	+ 30		
	12.3769(2)[*,1]	12.3742	+ 27		
	12.3796(2)[*,1]	12.3742	+ 54		
$Gd_3Fe_5O_{12}$	12.4737(4)[1]	12.4672	+ 65		
$Dy_3Al_5O_{12}$	12.0420(2)[1]	12.0444	— 24		
$Sm_3Ga_5O_{12}$	12.4336(3)[1]	12.4313	+ 23		
	12.4367(3)[2]	12.4313	+ 54		
$Gd_{2.36}Tb_{0.59}Eu_{0.05}Fe_5O_{12}$	12.4696(2)[1]	12.4607	+ 89		
$Tb_{0.93}Er_{2.07}Ga_{1.13}Fe_{3.87}O_{12}$	12.3606(2)[1]	12.3590	+ 16		
$Gd_{2.35}Yb_{0.69}Fe_{4.96}O_{12}$	12.4405(5)[1]	12.4390	+ 15		
$Nd_3Ga_5O_{12}$	12.5082(2)[2]	12.5015	+ 67		

$\Delta a = a_{obs} - a_{calc}$,

[*] Lattice constant determination by Bond's method [29],
[1] Flux-grown,
[2] Czochralski-grown,
[a] Using distribution coefficients of ref. 50,
[b] Using distribution coefficients of ref. 37,
[c] Using distribution coefficients of ref. 66,
[d] 0.03 formular units of Gd^{3+} on octahedral sites,
Standard deviations σ are given in parentheses (e.g. a_{obs} = 12.3792 (6) Å means a_{obs} = 12.3792 ± 0.0006 Å).

With the exception of CaSn-substituted YIG similar consistency of measured lattice constants with calculated lattice constants using self-determined cation radii and chemical analysis have been found for some recently prepared single crystal garnet materials (see table V). The discrepancy in the case of CaSn–YIG may be explained by non-consideration of charge compensation or the uncertainty of Sn valence, although the assumption of Fe^{2+} for charge neutralization of excess Sn^{4+} does not give significantly better results.

As it is known from the literature [17]) differences of the lattice constants between single crystal garnets and ceramic samples can be expected, since the cation distributions [62,63]), the stoichiometry [56,57]) or the impurity concentrations [64]) can be different depending on the growth technique for the garnet single crystals.

The recorded variation of the lattice parameter in fluxgrown YIG which covers the range from 12.3769 Å to 12.3796 Å is probably due to variations of the impurity concentration including those within the growth bands since (110) cuts have been measured in which the growth bands lie parallel but in statistic distance to the surface. Similar lattice constant variations have been found by Isherwood [65]) in YIG crystals grown in a flux consisting of PbO–PbF_2 and B_2O_3.

The difference between the lattice constants of Czochralski-grown and flux-grown gadolinium–gallium garnet is mainly attributed to the excess of Gd^{3+} in Czochralski-grown samples which occupies octahedral lattice sites [57]) and may be confirmed by the calculated value (see table V).

For the calculation of lattice constants of single crystalline Ga–YIG, Al–YIG and Ga–GdIG the distribution coefficients for Ga^{3+} and Al^{3+} as they have been published for Ga–YIG [37,62,66,67]) and Al–YIG [68]), respectively, have been neglected in column 3, table V. In column 5 the lattice constants a_{calc}(cor.) of the Ga–YIG single crystals are calculated under consideration of the distribution coefficients using the values of ref. 62 as they are determined for single crystals in the range $0.115 \leqslant u \leqslant 1.15$ and the values of ref. 37 as they are measured on polycrystalline samples in the range $2.42 \leqslant u \leqslant 2.68$. In the case of Al–YIG the values measured for polycrystalline materials [68]) are taken into account. Although the Ga-distribution coefficients are not identical in polycrystalline and single crystalline material [62]) the agreement between measured lattice constants and calculated lattice constants is slightly improved for a higher degree of substitution. It should be noted that the effect of lattice site distribution on the lattice constant is in these cases too small to give distinct changes of the lattice constants and therefore calculation of distribution coefficients would not be accurate enough to give a reliable information about site occupancy, although in principal (14) would give different results depending on site distribution. This shortcoming may be partly caused by the uncertainty of measured lattice constants and by the uncertainty of site distribution, since the actual cation distribution can be different in polycrystalline and single crystalline material [62]) and is determined as it has been recently shown [69]) by the thermal history of the crystals during growth.

Another application of (14) can be the estimation of the most probable valency stage of certain cation species as it has been shown successfully by the determination of the Pb valence in iron garnets [21]).

Also in the case of multiple substitution in a certain set of lattice sites (14) yields to good agreement between measured and calculated lattice constants [19,20]). If the general formula for garnets changes into

$$\{C_{3-x'-x''\cdots-x^n}C'_{x'}C''_{x''}\ldots C^n_{x^n}\}\,[A_{2-y'-y''\cdots-y^n}A'_{y'}A''_{y''}\ldots A^n_{y^n}]$$
$$(D_{3-z'-z''\cdots-z^n}D'_{z'}D''_{z''}\ldots D^n_{z^n})O_{12}$$

(11)–(13) are substituted by

$$r^{\mathrm{VIII}} = r^{\mathrm{VIII}}_{\mathrm{C}} + \frac{x'}{3}(r^{\mathrm{VIII}}_{\mathrm{C'}} - r^{\mathrm{VIII}}_{\mathrm{C}}) + \frac{x''}{3}(r^{\mathrm{VIII}}_{\mathrm{C''}} - r^{\mathrm{VIII}}_{\mathrm{C}}) + \ldots \frac{x^n}{3}(r^{\mathrm{VIII}}_{\mathrm{C}n} - r^{\mathrm{VIII}}_{\mathrm{C}})$$

$$r^{\mathrm{VI}} = r^{\mathrm{VI}}_{\mathrm{A}} + \frac{y'}{2}(r^{\mathrm{VI}}_{\mathrm{A'}} - r^{\mathrm{VI}}_{\mathrm{A}}) + \frac{y''}{2}(r^{\mathrm{VI}}_{\mathrm{A''}} - r^{\mathrm{VI}}_{\mathrm{A}}) + \ldots \frac{y^n}{2}(r^{\mathrm{VI}}_{\mathrm{A}n} - r^{\mathrm{VI}}_{\mathrm{A}}) \tag{16}$$

$$r^{\mathrm{IV}} = r^{\mathrm{IV}}_{\mathrm{D}} + \frac{z'}{3}(r^{\mathrm{IV}}_{\mathrm{D'}} - r^{\mathrm{IV}}_{\mathrm{D}}) + \frac{z''}{3}(r^{\mathrm{IV}}_{\mathrm{D''}} - r^{\mathrm{IV}}_{\mathrm{D}}) + \ldots \frac{z^n}{3}(r^{\mathrm{IV}}_{\mathrm{D}n} - r^{\mathrm{IV}}_{\mathrm{D}}). \tag{17}$$

It is probable that the agreement between observed lattice constants and those calculated by means of formula (14) using (15)–(17) will decrease with complexity of garnet composition.

It may be noticed that our model considers lattice constant variation only by the uniform lattice deformation due to cation size. Changes of cubic symmetry and ionic state of the garnet lattice with substituting species or amount of substitution are excluded. Another shortcoming of the given formulas may be the fact that the lattice variation is merely attributed to the substituting cation, but they do not explain the simultaneous distortion of the other coordination polyhedra, which results from the variation of the maximum amount for a certain substitution by simultaneous occupation of other lattice sites [58,59]).

If the coefficients of a polynomial as they are given in table I are known for a certain substitution the agreement between lattice constants calculated with such a polynomial and observed data will be mostly better than by calculation using (14). Lattice constant determination by means of (14) will be advantageous if such a polynomial is not known and lattice constant variation by a certain substitution is of interest.

Acknowledgement

The authors would like to thank all their colleagues who have contributed to the work described here, in particular D. Obertop for writing a computer program for the lattice parameter calculations and for the X-ray diffraction measurements of the polycrystalline garnet materials.

Philips GmbH Forschungslaboratorium, Hamburg *Hamburg, July 1978*

Note added in proof

Since the completion of this manuscript two further papers concerning the calculation of the lattice parameters of garnets have become available; Brice et al. [70]) suggest in the case of multiple substitution a linear dependence of the lattice constants of LPE films upon the amounts of the substituent ions and Dukhovskaya et al. [71]) propose a formula which links the lattice constants with the average cation–anion distances.

REFERENCES

[1]) D. McConnell, Bull. Soc. franc. Minér Crist. **89**, 14, 1966.
[2]) L. H. Ahrens, Geochim. Cosmochim. Acta **2**, 155, 1952.
[3]) G. A. Novak and G. V. Gibbs, Am. Mineralogist **56**, 791, 1971.
[4]) R. D. Shannon and C. T. Prewitt, Acta Cryst. **B 25**, Part 5, 925, 1969.
[5]) V. M. Talanov, Yu. P. Vorob'ev, A. N. Men' and V. M. Levchenko, Soviet Phys. J. **17**, 6, 1974.
[6]) V. M. Talanov, Yu. P. Vorob'ev and A. N. Men', J. Phys. Chem. Solids **36**, 641, 1975.
[7]) V. M. Talanov, Yu. P. Vorob'ev, T. I. Selivanova, V. M. Levchenko and A. N. Men', Inorganic Materials **11**, 4, 1975.
[8]) L. Suchow, M. Kokta and V. J. Flynn, J. Solid State Chem. **2**, 137, 1970.
[9]) L. Suchow and M. Kokta, J. Solid State Chem. **5**, 85, 1972.
[10]) L. Suchow and M. Kokta, J. Solid State Chem. **5**, 329, 1972.
[11]) R. Mondegarian, M. Kokta and L. Suchow, J. Solid State Chem. **18**, 369, 1976.
[12]) R. D. Shannon and C. T. Prewitt, Acta Cryst. **B 26**, Part 7, 1046, 1970.
[13]) R. D. Shannon, Acta Cryst. **A 32**, 751, 1976.
[14]) H. L. Glass and M. T. Elliott, J. Cryst. Growth **27**, 253, 1974.
[15]) M. Allibert, C. Chatillon, J. Mareschal and F. Lissalde, J. Cryst. Growth **23**, 289, 1974.
[16]) H. Makino, S. Nakamura and K. Matsumi, J. Appl. Phys. **15**, 415, 1976.
[17]) Landolt-Börnstein, Numerical date and functional relationships in science and technology, New Series, III 12a, Springer-Verlag, Berlin–Heidelberg–New York, 1978.
[18]) W. Tolksdorf, G. Bartels, P. Holst and W. T. Stacy, J. Cryst. Growth **26**, 122, 1974.
[19]) D. Mateika, J. Herrnring, R. Rath and Ch. Rusche, J. Cryst. Growth **30**, 311, 1975.
[20]) D. Mateika and Ch. Rusche, J. Cryst. Growth **42**, 440, 1977.
[21]) G. B. Scott and J. L. Page, J. Appl. Phys. **48**, 1342, 1977.
[22]) G. Winkler, P. Hansen and P. Holst, Philips Res. Repts **27**, 151, 1972.
[23]) M. W. Muller, Phys. Stat. Sol. (b) **83**, 177, 1977.
[24]) M. W. Muller, Phys. Stat. Sol. (b) **86**, 345, 1978.
[25]) W. Tolksdorf in A. Paoletti (ed.), Proc. Internat. School of Physics 'Enrico Fermi', Course LXX (1977), North-Holland, Publ. Co., Amsterdam, in press.

[26] W. Tolksdorf and F. Welz, J. Cryst. Growth **35**, 285, 1976.
[27] W. Tolksdorf and F. Welz, Crystals: growth, properties and applications, Vol. 1, Springer-Verlag, Berlin - Heidelberg - New York, 1978, in press.
[28] D. Mateika, P. Flisikowski, H. Kohler and R. Kilian, J. Cryst. Growth **41**, 262, 1977.
[29] W. L. Bond, Acta Cryst. **13**, 814, 1960.
[30] H. Hashizume and K. Kohra, Oyo Buturi **41**, 1240, 1972.
[31] R. L. Barns, Adv. X-ray Anal. **15**, 330, 1972.
[32] S. Geller, G. P. Espinosa, H. J. Williams, R. C. Sherwood and E. A. Nesbitt, J. Appl. Phys. **35**, 570, 1964.
[33] S. Geller, H. J. Williams, R. C. Sherwood and G. P. Espinosa, J. Phys. Chem. Solids **26**, 443, 1965.
[34] M. A. Gilleo and S. Geller, Phys. Rev. **110**, 73, 1958.
[35] S. Geller, H. J. Williams, G. P. Espinosa and R. C. Sherwood, Bell Syst. Tech. J. **43**, 565, 1964.
[36] P. Fischer, W. Hälg, P. Roggwiller and E. R. Czerlinsky, Solid State Commun. **16**, 987, 1975.
[37] S. Geller, J. A. Cape, G. P. Espinosa and D. H. Leslie, Phys. Rev. **148**, 522, 1966.
[38] S. Geller, H. J. Williams, G. P. Espinosa and R. C. Sherwood, J. Appl. Phys. **35**, 542, 1964.
[39] M. M. Schieber, Experimental magnetochemistry, nonmetallic magnetic materials, North-Holland Publ. Co., Amsterdam, 1967.
[40] E. E. Anderson, J. R. Cunningham Jr., G. E. McDuffie Jr. and R. F. Stander, J. Phys. Soc. Japan **17**, Suppl. B-I, 365, 1962.
[41] S. Geller, H. J. Williams, R. C. Sherwood and G. P. Espinosa, J. Appl. Phys. **36**, 88, 1965.
[42] M. Shimada, S. Kume and M. Koizumi, J. Phys. Chem. Solids **31**, 2165, 1970.
[43] S. Geller, H. J. Williams, G. P. Espinosa, R. C. Sherwood and M. A. Gilleo, Appl. Phys. Lett. **3**, 21, 1963.
[44] H. Takeuchi, Japan J. Appl. Phys. **14**, 1903, 1975.
[45] F. Bertaut and F. Forrat, Compt. Rend. **244**, 96, 1957.
[46] S. Geller and M. Gilleo, Acta Cryst. **10**, 239, 1957.
[47] G. P. Espinosa, J. Chem. Phys. **37**, 2344, 1962.
[48] F. Euler and J. A. Bruce, Acta Cryst. **19**, 971, 1965.
[49] B. J. Skinner, Am. Mineralogist **41**, 428, 1956.
[50] S. Geller, C. E. Miller and R. G. Treuting, Acta Cryst. **13**, 179, 1960.
[51] S. Geller, Z. Krist. **125**, 1, 1967.
[52] R. W. G. Wyckoff, Crystal structures Vol. 3, 2nd ed., New York, 1965.
[53] S. Geller, G. P. Espinosa and P. B. Crandall, J. Appl. Cryst. **2**, 86, 1969.
[54] C. B. Rubenstein and R. L. Barns, Am. Mineralogist **49**, 1489, 1964.
[55] C. B. Rubenstein and R. L. Barns, Am. Mineralogist **50**, 782, 1965.
[56] C. D. Brandle and R. L. Barns, J. Cryst. Growth **20**, 1, 1973.
[57] C. D. Brandle and R. L. Barns, J. Cryst. Growth **26**, 169, 1974.
[58] E. L. Dukhovskaya, A. P. Erastova, B. E. Rubinshtein, Yu. G. Saksonov and L. A. Vorob'eva, Inorg. Mater. USSR (english transl.) **9**, 1074, 1973.
[59] M. Kokta, J. Solid State Chem. **8**, 39, 1973.
[60] H. J. van Hook, J. Am. Ceram. Soc. **44**, 208, 1961.
[61] A. E. Paladino and E. A. Maguire, J. Am. Ceram. Soc. **53**, 98, 1970.
[62] P. Hansen, P. Röschmann and W. Tolksdorf, J. Appl. Phys. **45**, 2728, 1974.
[63] P. Görnert and C. G. D'Ambly, Phys. Stat. Sol. (a) **29**, 95, 1975.
[64] G. A. Slack and D. W. Oliver, Phys. Rev. **B4**, 592, 1971.
[65] B. J. Isherwood, J. Appl. Cryst. **1**, 299, 1968.
[66] R. L. Streever and G. A. Uriano, Phys. Rev. **139**, A305, 1965.
[67] P. Fischer, W. Hälg, E. Stoll and A. Segmüller, Acta Cryst. **21**, 765, 1966.
[68] E. R. Czerlinsky and R. A. McMillan, Phys. Stat. Sol. **41**, 333, 1970.
[69] P. Röschmann, W. Tolksdorf and F. Welz, Proc. IEEE **Mag-14**, Intermag 1978, in press.
[70] J. C. Brice, J. M. Robertson, W. T. Stacy and J. C. Verplanke, J. Cryst. Growth **30**, 66, 1975.
[71] E. L. Dukhovskaya and Yu. G. Saksonov, Sov. Phys. Crystallogr. **22**, 622, 1977.

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*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 Avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 5100 Aachen, West-Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, West-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).

E. Arnold and J. M. Shannon: Anomalous Hall effect and carrier transport in bandtails at the Si–SiO$_2$ interface.
Solid State Comm. **18**, 1153-1156, 1976. *R*

H. M. J. M. van Ass, R. G. Gossink and P. J. W. Severin: Preparation of graded-index optical glass fibres in the alkali germanosilicate system.
Electronics Letters **12**, 369-370, 1976. *E*

E. F. Assmus, Jr., J.-M. Goethals and H. F. Mattson, Jr.: Generalized t-designs and majority decoding of linear codes.
Information and Control **32**, 43-60, 1976. *B*

W. J. Bartels, L. Blok and C. W. Th. Bulle: X-ray topography and diode efficiency of vapour grown GaAs$_{1-x}$P$_x$ layers.
J. Crystal Growth **34**, 181-188, 1976. *E*

H. Baudry: Le contrôle rhéologique des pâtes pour déterminer leur aptitude à la sérigraphie.
Electronique & Microél. ind. No. 226, 38-41, 1976. *L*

C. Belouet: Review of European proposals for vapour growth.
Material sciences in space, Proc. 2nd Eur. Symp., Frascati, 1976, pp. 245-248. *L*

C. Belouet: Fluid-dynamics problems connected with crystal growth.
Material sciences in space, Proc. 2nd Eur. Symp., Frascati, 1976, pp. 283-289. *L*

U. Bergmann, K. Klose and L. Spiess: Focus path of a laser beam deflected by a prismatic polygen mirror: its calculation and optimization.
Appl. Optics **15**, 3084-3088, 1976. *H*

J. W. M. Biesterbos, M. Brouha and A. G. Dirks: Pressure dependence of magnetic properties of amorphous RE-TM thin films.
AIP Conf. Proc. **29**, 184-185, 1976. *E*

J. W. M. Biesterbos and A. G. Dirks: Amorfe metallische systemen.
Polytechn. T. Werktuigbouw **31**, 145-150, 1976. *E*

J. van den Boomgaard, R. A. J. Born, G. Frens and J. W. A. Nelissen: The λ-R relationship in the directional decomposition of ε-FeAl.
J. Crystal Growth **35**, 311-314, 1976. *E*

J. van den Boomgaard and A. M. J. G. van Run: Poling of a ferroelectric medium by means of a built-in space charge field, with special reference to sintered magnetoelectric composites.
Solid State Comm. **19**, 405-407, 1976. *E*

A. J. van Bommel and J. E. Crombeen: LEED, Auger electron spectroscopy (AES) and photoemission studies of the adsorption of cesium on the epitaxially grown GaAs(110) surface.
Surface Sci. **57**, 109-117, 1976. *E*

A. J. van Bommel and J. E. Crombeen: Experimental determination of the correlation between the LEED pattern and the Ga-As bond vectors in the surface of GaAs(001).
Surface Sci. **57**, 437-440, 1976. *E*

M. R. Boudry: A simple conversion formula for type 'T' (copper-constantan) thermocouple readings.
J. Physics E 9, 1064-1065, 1976. *R*

P. W. J. M. Boumans and F. J. de Boer: Studies of a radio frequency inductively coupled argon plasma for optical emission spectrometry, III. Interference effects under compromise conditions for simultaneous multi-element analysis.
Spectrochim. Acta **31B**, 355-375, 1976. *E*

P. W. J. M. Boumans, G. H. van Gool and J. A. J. Jansen: A computerised programmable monochromator for flexible multi-element analysis with special reference to the inductively coupled plasma.
Analyst **101**, 585-587, 1976. *E*

D. J. Breed and R. P. Kramer: Stable and unstable surface state charge in thermally oxidized silicon.
Solid-State Electronics **19**, 897-907, 1976. *E*

F. J. A. den Broeder: Metallische glazen: een nieuwe materiaalklasse.
Chem. Weekbl. Mag. 1976, m 324-326 (juni). *E*

F. J. A. den Broeder and H. Zijlstra: Relation between coercive force and microstructure of sintered $SmCo_5$ permanent magnets.
J. Appl. Phys. **47**, 2688-2695, 1976. *E*

T. M. Bruton: The growth of 'thick' single crystals of $Bi_4Ti_3O_{12}$ from bismuth borate solutions.
J. Crystal Growth **36**, 36-40, 1976. *R*

K. H. J. Buschow, J. H. N. van Vucht and W. W. van den Hoogenhof: Note on the crystal structure of the ternary rare earth-3*d* transition metal compounds of the type RT_4Al_8.
J. Less-Common Met. **50**, 145-150, 1976. *E*

F. M. A. Carpay: Reply to the comment of J. D. Livingston on 'The λ-R relationship in the directional decomposition of ε-FeAl' by J. van den Boomgaard, R. A. J. Born, G. Frens and J. W. A. Nelissen.
J. Crystal Growth **35**, 316-317, 1976. *E*

V. Chalmeton, C. Patanchon and C. Mesnage: Radioscopie industrielle télévisée à haute énergie appliquée aux moteurs à carburant solide.
8th World Conf. on Nondestructive Testing, Cannes, 1976, paper 3E 13, 8 pp. *L*

P.-J. Courtois and G. Louchard: Approximation of eigencharacteristics in nearly-completely decomposable stochastic systems.
Stoch. Proc. Appl. **4**, 283-296, 1976. *B*

J. E. Curran, R. V. Jeanes and H. Sewell: A technology of thin-film hybrid microwave circuits.
IEEE Trans. **PHP-12**, 304-309, 1976. *R*

H. T. van Dam: The conductance of heptyl viologen dibromide in water and methanol.
J. Electrochem. Soc. **123**, 1181-1184, 1976. *E*

M. Delfino: Solution crystal growth of ionic salts by electrolytic solvent decomposition.
J. Crystal Growth **32**, 378-380, 1976. *N*

P. A. Devijver: Error and reject tradeoff for nearest neighbor decision rules.
Preprints NATO ASI Signal Processing, Portovenere, 1976, Part 2, pp. 22/1-22/24. *B*

A. M. van Diepen: The B-site Mössbauer linewidth in Fe_3O_4.
Physics Letters **57A**, 354-356, 1976. *E*

N.V. PHILIPS' GLOEILAMPENFABRIEKEN

H. Dimigen, H. Lüthje, H. Hübsch and U. Convertini: Influence of mask materials on ion-etched structures.
J. Vac. Sci. Technol. **13**, 976-980, 1976. *H*

J. A. W. van der Does de Bye and L. Blok: Room temperature minority carrier lifetime and efficiency of p-type $GaAs_{1-x}P_x$.
J. Luminescence **14**, 101-113, 1976. *E*

H. Dötsch: Stability and dynamics of microwave generated ring domains.
AIP Conf. Proc. **29**, 78-83, 1976. *H*

G. P. Edwards, T. Preston, L. D. J. Eggermont and M. H. H. Höfelt: System and hardware design considerations for a single-channel analogue-to-PCM via -HIDM encoder.
1976 Int. Zürich Seminar on Digital Communications, pp. B3.1-B3.5. *E*

J. M. L. Engels and A. J. M. van Kimmenade: The mobility of excess electrons in liquid methane.
Chem. Phys. Letters **42**, 250-252, 1976. *E*

D. den Engelsen: Optical anisotropy in ordered systems of lipids.
Surface Sci. **56**, 272-280, 1976. *E*

L. J. M. Esser: Peristaltic charge coupled devices: what is special about the peristaltic mechanism.
Solid state imaging, P. G. Jespers, F. van de Wiele and M. H. White (eds), publ. Noordhoff, Leiden, 1976, pp. 343-425. *E*

W. G. Essers: New process combines plasma with GMA welding.
Welding J. **55**, 394-400, 1976. *E*

E. Fabre: MIS silicon solar cells.
Appl. Phys. Letters **29**, 607-610, 1976. *L*

A. Farrayre, J. Grau, B. Kramer and J. Magarshack: Design integrated amps with hi-lo Impatt diodes.
Microwaves **15**, No. 10, pp. 42, 44, 46 & 47, 1976. *L*

H. Figiel, A. Oppelt, E. Dormann and K. H. J. Buschow: Transferred hyperfine fields at the Y sites in Y-Co compounds.
Phys. Stat. Sol. (a) **36**, 275-283, 1976. *E*

S. R. Fletcher, E. T. Keve and A. C. Skapski: Structural studies of triglycine sulphate: Part I. Low radiation dose (Structure A), Part II. After X-irradiation/field treatment (Structure B).
Ferroelectrics **14**, 775-787, 789-799, 1976. *R*

R. C. French: Error performance of p.s.k. and f.f.s.k. subcarrier data demodulators.
Radio and electronic Engr. **46**, 543-458, 1976. *R*

A. D. Giles and F. F. Westendorp: The effect of cobalt substitutions on some properties of manganese zinc ferrites.
J. Physics D **9**, 2117-2122, 1976. *R*

J. J. Goedbloed and J. Joosten: Responsivity of avalanche photodiodes in the presence of multiple reflections.
Electronics Letters **12**, 363-364, 1976. *E*

H. C. de Graaff and J. W. Slotboom: Some aspects of LEC transistor behaviour.
Solid-State Electronics **19**, 809-814, 1976. *E*

G. Groh: Future aspects of tomography from a physicist's point of view.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Genoa, 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 71-75, 1976. *H*

P. C. M. Gubbens, A. M. van der Kraan and K. H. J. Buschow: Relation between anisotropy and crystal structure in rare-earth (3*d*) transition compounds (R_2M_{17}).
Solid State Comm. **19**, 355-356, 1976. *E*

G. J. van Gurp, D. Sigurd and W. F. van der Weg: Tungsten as a marker in thin-film diffusion studies.
Appl. Phys. Letters **29**, 159-161, 1976. *E*

T. K. Halstead, N. A. Abood and K. H. J. Buschow: Study of the diffusion of hydrogen in $LaNi_{5+x}H_6$ compounds by ^{1}H NMR relaxation.
Solid State Comm. **19**, 425-428, 1976. *E*

J. C. M. Henning and J. H. den Boef: Strain-modulated electron spin resonance of Co^{2+} in MgO: a comparison of extensional and flexural modes.
Phys. Rev. B **14**, 26-34, 1976. *E*

A. van Herk and D. L. A. Tjaden: The magnetic field near the side edge of narrow magnetic recording heads.
Proc. Conf. on Video and Data Recording, Birmingham, 1976, pp. 223-225. *E*

H. Hieber: Aging properties of gold layers with different adhesion layers.
Thin Solid Films **37**, 335-343, 1976. *H*

I. D. Higgins: Easy and accurate method for the characterisation of dielectric materials at X-band.
Electronics Letters **12**, 573, 1976. *R*

I. D. Higgins: Performance of self-oscillating GaAs m.e.s.f.e.t. mixers at X-band
Electronics Letters **12**, 605-606, 1976. *R*

B. Hill and K. P. Schmidt: The realization and technology of interfaces between optics and electronics in holographic memories.
Laser 75 Opto-electronics, Proc. Conf. Munich, 1975, pp. 255-259, 1976. *H*

W. Hoekstra and V. H. C. Evers: Een geautomatiseerd microfiche-archiefsysteem voor persoonlijk gebruik.
Informatie **18**, 494-501, 1976. *E*

B. Hoekstra, J. M. Robertson and G. Bartels: Variations of magnetic anisotropy within epitaxial films of $Y_{2.85}La_{0.15}Fe_{3.75}Ga_{1.25}O_{12}$ obtained from spin wave resonance.
AIP Conf. Proc. **29**, 111-112, 1976. *E, H*

L. Hollan: Evolution des techniques d'élaboration de GaAs épitaxial pour les dispositifs hyperfréquences.
Le Vide **31**, 131-137, 1976. *L*

L. Hollan, J. M. Durand and R. Cadoret: Influence of the growth parameters in GaAs vapor phase epitaxy.
J. Electrochem. Soc. **124**, 135-139, 1977. *L*

E. P. Honig and B. R. de Koning: Ellipsometric investigation of the skeletonization process of Langmuir–Blodgett films.
Surface Sci. **56**, 454-461, 1976. *E*

L. P. J. Hoogeveen, F. W. Willmott and R. J. Dolphin: An automatic monitor for the analysis of organochlorine compounds in milk.
Z. Anal. Chemie **282**, 401-406, 1976. *E, R*

A. Humbert, L. Hollan and D. Bois: Influence of the growth conditions on the incorporation of deep levels in VPE GaAs.
J. Appl. Phys. **47**, 4137-4144, 1976. *L*

H. Ihrig: On the polaron nature of the charge transport in $BaTiO_3$.
J. Physics C **9**, 3469-3474, 1976. *A*

A. W. de Jager-Veenis and A. Bril: Vacuum ultraviolet excitation spectra of phosphors for use in gas discharge display panels.
J. Electrochem. Soc. **123**, 1253-1254, 1976. *E*

L. Jeunhomme, A. Cozannet, R. Bouillie and J. P. Hazan: Mesure des caractéristiques de transmission de conducteurs optiques.
Onde Electr. **56**, 564-571, 1976. *L*

J. L. Kirk and L. E. Cross, J. P. Dougherty: Pressure and temperature dependence of the dielectric properties and phase transitions of the ferroelectric $Pb_5Ge_3O_{11}$.
Ferroelectrics **11**, 439-443, 1976. *N*

E. Klotz and H. Weiss: Short-time tomosynthesis.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Genoa, 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 65-70, 1976. *H*

A. J. R. de Kock: Characterization and elimination of defects in silicon.
Festkörperprobleme **16**, 179-193, 1976. *E*

A. J. R. de Kock: Silicium-monokristallen zonder imperfecties.
Chem. Weekbl. Mag. 1976, m 329-330 (juni). *E*

J. Köhler and B. Schiek: FM-Rauschmeßplatz für Mikrowellen-Oszillatoren.
Mikrowellen-Magazin 4/76, pp. 276, 278 & 281-282, 1976. *H*

W. L. Konijnendijk (Philips Lighting Division) **and J. M. Stevels** (Eindhoven, University of Technology)**:** The linear expansion of borosilicate glasses in relation to their structure.
Verres Réfract. **30**, 371-377, 1976.

W. L. Konijnendijk and J. M. Stevels: Raman scattering measurements of silicate glasses and compounds.
J. Non-Cryst. Solids **21**, 447-453, 1976. *E*

E. Kooi, J. G. van Lierop and J. A. Appels: Formation of silicon nitride at a Si-SiO_2 interface during local oxidation of silicon and during heat-treatment of oxidized silicon in NH_3 gas.
J. Electrochem. Soc. **123**, 1117-1120, 1976. *E*

A. M. van der Kraan, J. N. J. van der Velden, J. H. F. van Apeldoorn, P. C. M. Gubbens and K. H. J. Buschow: Mössbauer effect study and magnetization measurements of RFe_3 intermetallic compounds with R = Sm, Gd, Tb, Dy, Ho, and Th.
Phys. Stat. Sol. (a) **35**, 137-151, 1976. *E*

B. M. Kramer, A. C. Derycke, A. Farrayre and C. F. Masse: High-efficiency frequency multiplication with GaAs avalanche diodes.
IEEE Trans. **MTT-24**, 861-863, 1976. *L*

J. van Laar and A. Huijser: Contact potential differences for III-V compound surfaces.
J. Vac. Sci. Technol. **13**, 769-772, 1976. *E*

P. K. Larsen and R. Metselaar: Electrical properties of yttrium iron garnet at high temperatures.
Phys. Rev. B **14**, 2520-2527, 1976. *E*

P. K. Larsen, R. Metselaar and B. Feuerbacher: UV photoemission studies of yttrium iron garnet.
AIP Conf. Proc. **29**, 668-669, 1976. *E*

P. K. Larsen & J. M. Robertson: Changes in optical absorption in iron garnet films due to impurity incorporation.
Appl. Phys. **11**, 259-263, 1976. *E*

R. E. van de Leest: Solid-state ion-selective electrodes based on thin ion-selective layers deposited on ionic conductors.
Analyst **101**, 433-438, 1976. *E*

P. E. J. Legierse (Philips Philite- and Metaalwarenfabrieken Eindhoven)**:** Fotochemische metaalbewerking.
Metaalbewerking **42**, 301-306, 1976.

F. J. M. J. Maessen, J. W. Elgersma and P. W. J. M. Boumans: A systematic and rigorous statistical approach for establishing the accuracy of analytical results and its application to a comparison of alternative d.c. arc procedures for trace analysis of geological materials.
Spectrochim. Acta **31B**, 179-199, 1976. *E*

H. H. van Mal, K. H. J. Buschow and A. R. Miedema: Hydrogen absorption of rare-earth (3*d*) transition intermetallic compounds.
J. Less-Common Met. **49**, 473-475, 1976. *E*

W. F. G. Mecklenbräuker (I, II), R. M. Mersereau (I, II) and T. F. Quatieri, Jr. (I) McClellan transformations for two-dimensional digital filtering: I. Design, II. Implementation.
IEEE Trans. **CAS-23**, 405-414, 414-422, 1976. *E*

F. Meyer: Ellipsometric studies of adsorption reactions on clean surfaces.
Surface Sci. **56**, 37-48, 1976. *E*

D. Meyer-Ebrecht: Trends in der elektronischen Röntgenbilderzeugung und -verarbeitung.
Röntgenstrahlen **35**, 2-10, 1976. *H*

D. Meyer-Ebrecht, J. Dittrich and J. Guldberg: Tomosynthesis.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Geona, 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 58-64, 1976. *H*

A. R. Miedema: On the heat of formation of plutonium alloys.
5th Int. Conf. on Plutonium and other Actinides, 1975, Baden-Baden, pp. 3-20, 1976. *E*

A. R. Miedema, K. H. J. Buschow and H. H. van Mal: Which intermetallic compounds of transition metals form stable hydrides?
J. Less-Common Met. **49**, 463-472, 1976. *E*

E. J. Millett: Digital techniques in laboratory automation.
J. Physics E **9**, 794-802, 1976. *R*

A. Mircea, A. Mitonneau, L. Hollan and A. Brière: Out-diffusion of deep electron traps in epitaxial GaAs.
Appl. Phys. **11**, 153-158, 1976. *L*

A. Mircea, A. Mitonneau and J. Vannimenus: Temperature dependence of ionization energies of deep bound states in semiconductors.
J. Physique Lettres **38**, L 41-43, 1977. *L*

A. Molenaar, G. H. C. Heynen and J. E. A. M. van den Meerakker: Physical development by copper complexes using ferrous-ferric ions as a redox system.
Photogr. Sci. Engng. **20**, 135-139, 1976. *E*

A. E. Morgan and H. W. Werner: On the abundance of molecular ions in secondary ion mass spectrometry.
Appl. Phys. **11**, 193-195, 1976. *E*

B. J. Mulder: Preparation of ultra-thin mica windows.
J. Physics E **9**, 724-725, 1976. *E*

C. Mulder and H. E. J. Wulms: High speed integrated injection logic (I^2L).
IEEE J. **SC-11**, 379-385, 1976. *E*

J. H. Neave and B. A. Joyce: The origin of spurious peaks in mass spectra.
J. Physics D **9**, 2195-2200, 1976. *R*

A. G. van Nie: Electroless NiP processing for hybrid integrated circuits.
Microelectronics and Reliability **15**, 221-226, 1976. *E*

A. Oppelt and K. H. J. Buschow: NMR investigation of the hyperfine interactions in $Y(Fe_{1-x}A_x)_2$ (A = Al, Co, Pt).
Phys. Rev. B **13**, 4698-4704, 1976. *E*

J. A. Pals and L. H. J. Graat: The simultaneous occurrence of Josephson effects and series resistance in Nb-U_6Fe point contacts.
Physics Letters **56A**, 487-488, 1976. *E*

J. B. H. Peek: Het Nederlands URSI-comité.
T. Ned. Elektronica- en Radiogen. **41**, 61-63, 1976. *E*

H. Rau: Range of homogeneity and defect energetics in $Co_{1-x}S$.
J. Phys. Chem. Solids **37**, 931-934, 1976. *A*

E. D. Roberts: Electron-sensitive film-forming materials and their uses in semiconductor technology.
Vacuum **26**, 459-467, 1976. *R*

J. M. Robertson: Improvement of lead-free flux systems for the growth of bismuth-substituted substituted iron garnet films by liquid phase epitaxy.
J. Electrochem. Soc. **123**, 1248-1249,1976. *E*

P. J. Roksnoer, W. J. Bartels and C. W. T. Bulle: Effect of low cooling rates on swirls and striations in dislocation-free silicon crystals.
J. Crystal Growth **35**, 245-248, 1976. *E*

P. Röschmann: Two-magnon-scattering contributions to FMR linebroadening in polycrystalline garnets.
AIP Conf. Proc. **34**, 253-258, 1976. *H*

C. W. J. Schiepers (Institute for Perception Research, Eindhoven): Global attributes in visual word recognition: Part 1. Length perception of letter strings, Part 2. The contribution of word length.
Vision Res. **16**, 1343-1349 & 1445-1454, 1976.

W. Schilz, R. Jacobson and B. Schiek: Mikrowellen-Entfernungsmeßsystem mit ± 2,5 mm Genauigkeit.
Mikrowellen-Magazin 2/76, pp. 102-107, 1976. *H*

H. Scholz: Crystal growth by dynamic gradient reversal techniques, III. Pure thermodynamic derivation of growth conditions.
Solid State Comm. **20**, 447-448, 1976. *A*

H. Schomberg: Parallel solution of a nonlinear elliptic boundary value problem.
Simulation of systems, L. Dekker (ed.), North-Holland Publ. Co., Amsterdam, 1976, pp. 461-470. *H*

J. Schröder: Thermal energy storage and control.
Trans. ASME B (J. Engng. Ind.) **97**, 893-896, 1975. *A*

J. Schröder: Thermal energy storage using fluorides of alkali and alkaline earth metals.
Proc. Symp. on Energy Storage, publ. Electrochem. Soc., pp. 206-220, 1976. *A*

J. W. Slotboom and H. C. de Graaff: Measurements of bandgap narrowing in Si bipolar transistors.
Solid-State Electronics **19**, 857-862, 1976. *E*

F. W. Smith: Surface-acoustic-wave parametric amplifier.
Electronics Letters **12**, 545-546, 1976. *R*

J. L. Sommerdijk, A. C. van Amstel and F. M. J. H. Hoex-Strik: On the luminescence of β-Ga_2O_3:Dy^{3+}.
J. Luminescence **11**, 433-436, 1976. *E*

J. L. Sommerdijk and A. Bril: Divalent europium luminescence in perovskite-like alkaline - earth alkaline fluorides.
J. Luminescence **11**, 363-367, 1976. *E*

J. L. Sommerdijk, J. A. W. van der Does de Bye and P. H. J. M. Verberne: Decay of the Ce^{3+} luminescence of $LaMgAl_{11}O_{19}$:Ce^{3+} and of $CeMgAl_{11}O_{19}$ activated with Tb^{3+} or Eu^{3+}.
J. Luminescence **14**, 91-99, 1976. *E*

W. T. Stacy and W. Guse: X-ray topographic study of Czochralski grown mullite.
J. Crystal Growth **35**, 153-158, 1976. *E*

A. L. N. Stevels: Luminescentie van europium(II)-geactiveerde aluminaten.
Chem. Weekbl. Mag. 1976, m 331-332 (juni). *E*

A. L. N. Stevels and A. D. M. Schrama-de Pauw: Eu^{2+} luminescence in hexagonal aluminates containing large divalent or trivalent cations.
J. Electrochem. Soc. **123**, 691-697, 1976. *E*

A. L. N. Stevels and A. D. M. Schrama-de Pauw: Theoretical and experimental efficiencies of X-rays screens.
J. Electrochem. Soc. **123**, 886-888, 1976. *E*

A. L. N. Stevels and A. D. M. Schrama-de Pauw: Effects of defects on the quantum efficiency of Eu^{2+}-doped aluminates with the magnetoplumbite-type crystal structure.
J. Luminescence **14**, 147-152, 1976. *E*

A. L. N. Stevels and A. D. M. Schrama-de Pauw: Luminescence spectra of non-stoichiometric aluminates doped with Eu^{2+} ions.
J. Luminescence **14**, 153-157, 1976. *E*

J. B. Theeten: Les ions lents: un outil (presque) idéal pour l'analyse des surfaces.
La Recherche **7**, 770-772, 1976. *L*

J. B. Theeten, F. Hottier and H. Paradan: Appareillage et méthodologie d'étude des surfaces de GaAs en cours de croissance en épitaxie phase vapeur.
Rev. Phys. Appl. **11**, 587-595, 1976. *L*

R. Tijburg: Advances in etching of semiconductor devices.
Phys. in Technol. **7**, 202-207, 1976. *E*

R. P. Tijburg and T. van Dongen: Selective etching of III-V compounds with redox systems.
J. Electrochem. Soc. **123**, 687-691, 1976. *E*

J. C. Tranchart, A. Farrayre and L. Hollan: Electrochimie du GaAs: étude et applications.
Proc. Coll. Matériaux et Technologie pour la Micro-Electronique, tendances actuelles, Montpellier, 1976 (Suppl. Le Vide No. 183), pp. 77-94. *L*

L. Vriens and M. Adriaanzs: Near-resonant light scattering and fluorescence in a dense hightemperture plasma.
Appl. Phys. **11**, 253-257, 1976. *E*

J. H. N. van Vucht and K. H. J. Buschow: Note on the occurrence of intermetallic compounds in the lithium-palladium system.
J. Less-Common Met. **48**, 345-347, 1976. *E*

W. Wagner: Reconstruction of object layers from their X-ray projections: a simulation study.
Computer Graph. and Image Proc. **5**, 470-483, 1976. *H*

H. Weiss: Use and abuse of the modulation transfer function.
7th L.H. Gray Conf.: Medical images, Leeds, 1976, pp. 161-172. *H*

H. W. Werner and N. Warmoltz: The influence of selective sputtering on surface composition.
Surface Sci. **57**, 706-714, 1976. *E*

H. W. Werner and A. E. Morgan: Charging of insulators by ion bombardment and its minimization for secondary ion mass spectrometry (SIMS) measurements.
J. Appl. Phys. **47**, 1232-1242, 1976. *E*

K. R. Whight, P. Blood and K. H. Nicholas: Implanted high value resistors.
Solid-State Electronics **19**, 1021-1027, 1976. *R*

P. Wiedijk (Philips Lighting Division, Eindhoven)**:** High-temperature extraction of gas from ceramics.
Anal. Chem. **48**, 1095-1096, 1976.

C. E. C. Wood: Molecular beam epitaxial GaAs layers for MESFET's.
Appl. Phys. Letters **29**, 746-748, 1976. *R*

P. Zandveld: Some properties of ion-implanted *p-n* junctions in silicon.
Solid-State Electronics **19**, 659-667, 1976. *E*

H. Zijlstra: Permanent magnets.
Phys. in Technol. **7**, 98-107, 1976. *E*

H. Zijlstra: Permanente magneten.
Natuur en Techniek **44**, 362-379, 1976. *E*

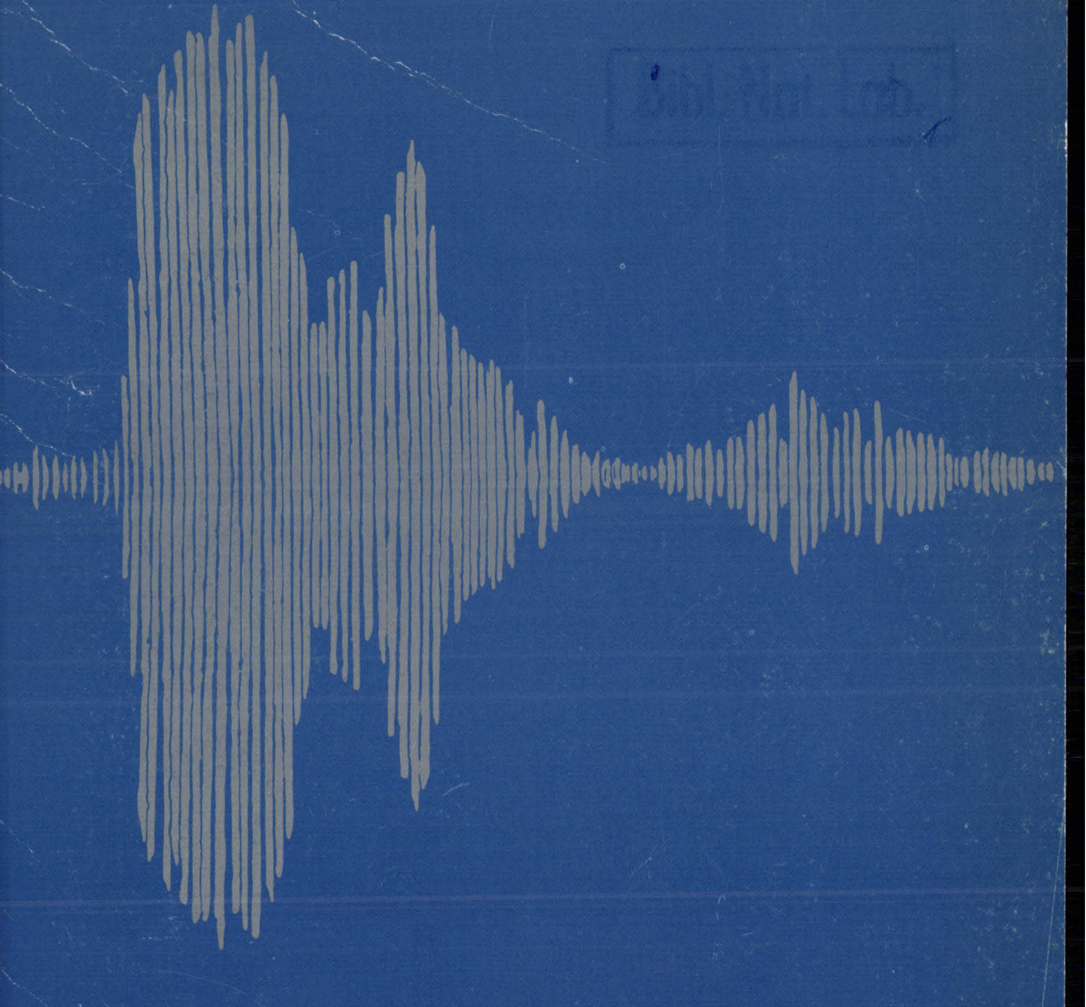
D. J. Zwanenburg and Th. A. M. M. Maas: The reactions between 3′-methyl-6-nitrospiro[2*H*-1-benzopyran-2,2′-benzothiazolines] and 1,1-bis[4-(dimethylamino)phenyl]-ethylene and the reactions of salicylaldehydes with 1,1-bis[4-(dimethylamino)phenyl]propene. A striking effect of an extra methyl group; photochromic chromenes.
Recueil Trav. Chim. Pays-Bas **95**, 97-98, 1976.

D. J. Zwanenburg and W. A. P. Reynen: An improved synthesis of salicylaldehydes. No influence of steric hindrance.
Synthesis, 1976, pp. 624-625, *E*

hilips Journal of Research

Vol. 33 Nos. 5/6 1978

Bibl. Nat. Lab.



HILIPS

Philips Journal of Research

Philips Journal of Research, edited by Philips Research Laboratories, Eindhoven, The Netherlands, is a bimonthly publication containing papers on research carried out in the various Philips laboratories. Volumes 1 to 32 have appeared under the name of Philips Research Reports.

Annual subscription price for Volume 33 is Dfl. 60, payable in advance. Payments should be made only after receipt of an invoice. Correspondence should be addressed to: Philips Journal of Research, Philips Research Laboratories, Building WBp, Room No. 42, Eindhoven, The Netherlands.

Cover design based on a visual representation of the sound-wave associated with the spoken word "Philips".

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1978. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

9R. 3-98-39

17 SEP. 1979

PHILIPS JOURNAL OF RESEARCH - Vol. 33 Nos. 5/6 - pp. 211-296 - 1978

Date of publication: March 28, 1979

FAST SWITCHABLE MAGNETO-OPTIC MEMORY - DISPLAY COMPONENTS

by B. HILL and K. P. SCHMIDT

Abstract

On the basis of magneto-optic iron-garnet memory films, a new technology for spatial light modulation and memory-display techniques is presented. Direct electronic control via a thin-film network allows fast switching of a pattern of domains in the magneto-optic film at random access. Due to the high Faraday rotation exhibited, the pattern of domains can be viewed in transmission using polarization optics. The main field of application of the new technique will be found in non-mechanical optical printing and data output terminals.

1. Introduction

Magneto-optic storage materials for the application in optical memories of the discrete bit-type are already subject of research and development since many years [1-11].

Digital information can be stored non-volatile and at high bit density by switching thermomagnetically the directions of magnetization of a pattern of domains in a thin ferrimagnetic film. The addressing of discrete bits in the memory is accomplished with help of a focussed laser beam deflected or scanned across the film. Reading of information is mostly done by using the Faraday effect by which domains with different directions of magnetization become visible in polarization optics.

A magneto-optic memory material being already far advanced is the bismuth-substituted iron-garnet film [9-12,14,16]. Recent improvement of the Faraday rotation has made possible the direct visualization of a stored magnetization pattern, thus allowing the development of fast switchable light-modulation lines and memory-display components, based on the spatial modulation of light by a pattern of domains in transmission.

A related development using magnetic bubbles has already been published in refs 15 and 16. This work resulted in the fabrication of a bubble display with 32 × 32 picture elements. Due to the "shift-register" nature of the bubble display the time required for writing an image is, however, rather long. This problem has been overcome by the new addressing technology discussed in this paper, employing thin-film structure for fast and direct electronic switching of magneto-optic memory cells at random access.

The basic principle of this technology is sketched in the first chapter followed by the ideas how to realize integrated linear and *x–y*-addressed light-modulation and display components. Finally, an experimental 256-bit light-modulation line is presented and essential features are discussed.

2. Principle of a magneto-optic display

The best magneto-optic material presently being available for display purposes is the bismuth-substituted iron-garnet film [9–12,14–16]. This film consists of a single crystalline layer of gadolinium–iron garnet with certain amounts of bismuth and gallium substituents ($Gd_{3-x}Bi_xFe_{5-y}Ga_yO_{12}$). The film is grown epitaxially on a single crystalline substrate made from gadolinium–gallium garnet with magnesium and zirconium as substituents ($Gd_3Ga_{4.0}Mg_{0.5}Zr_{0.5}O_{12}$). The typical thickness of the substrate and the film are 500 and 3–5 μm respectively. The film is ferrimagnetic and it possesses a uniaxial magnetic anisotropy whose easy axis is perpendicular to the plane of the film. The anisotropy brings about the desired magnetic bistability; a binary information stored in the film is associated with the two antiparallel directions of magnetization. An image is, thus, represented by a two-dimensional pattern of the directions of magnetization of domains.

For reasons of the geometrical stability of a pattern of domains the iron-garnet film is structured. Part of the film is removed by an etching process in such a way that only islands of ferrimagnetic material remains as shown in figs 1*a* and 1*b*. In each island one domain is locked in, unable to move or couple to the neighboured islands. Thus, an island is the basic memory cell of

250

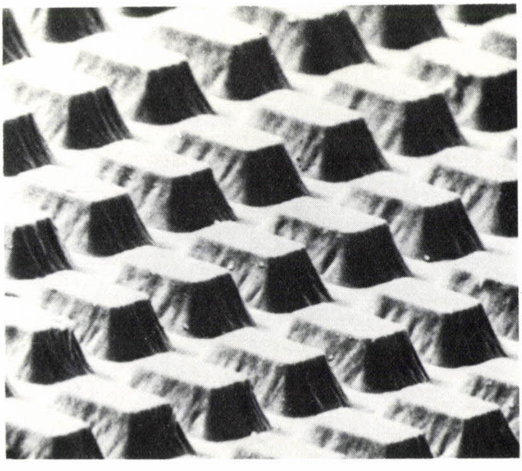


Fig. 1*a*. Iron-garnet memory film as seen in an electron scanning microscope.

250%

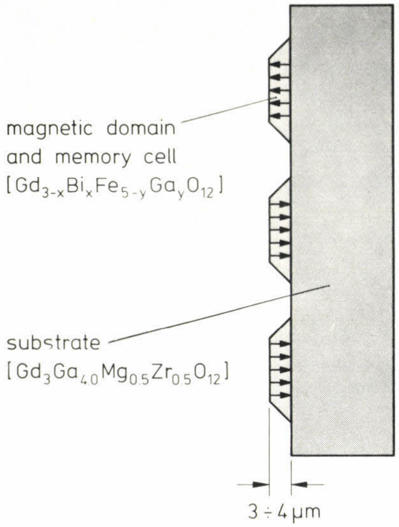

magnetic domain
and memory cell
$[Gd_{3-x}Bi_xFe_{5-y}Ga_yO_{12}]$
substrate
$[Gd_3Ga_{4.0}Mg_{0.5}Zr_{0.5}O_{12}]$
3 ÷ 4 μm

Fig. 1*b*. Magneto-optic iron-garnet memory film.

the film able to store two information states according to a digital image point. In experimental devices the typical size of an island is $10 \times 10\,\mu m^2$–$100 \times 100\,\mu m^2$.

Read-out of an image

Direct viewing of the pattern of domains is based on the Faraday effect in the material. When looking along the normal to the film the plane of polarization of linearly polarized light transmitted by the film is rotated either to the left or to the right in dependance on the direction of magnetization present. The transmitted light oscillating in one of the possible planes of polarization at the output is then blocked by an analyzer installed behind the film as shown in fig. 2. Light with the plane of polarization rotated to the other side is accordingly transmitted more or less, depending on the angle of the Faraday rotation. Hence, a pattern of domains becomes visible in simple polarization optics. There it appears like an array of dark and bright spots as demonstrated by the picture in fig. 3 that has been taken from an iron-garnet wafer in a polarization microscope.

3. Local switching of domains

The switching of domains with a given magnetization requires the application of an external magnetic field (fig. 4). However, to be able to write a pattern of domains with different directions of magnetization, it becomes necessary to localize the action of the external magnetic field, applied to the whole film, to

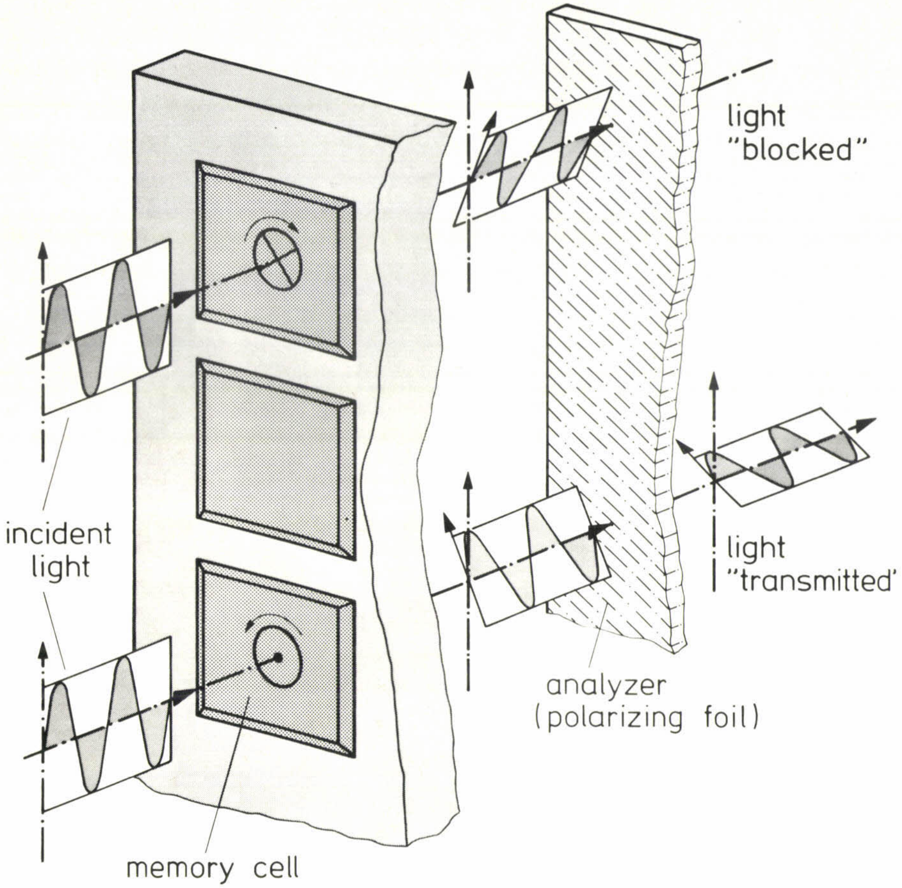


Fig. 2. Viewing of magnetic domains using the Faraday effect and polarization optics.

150%

only one of the domains (one memory cell). This is achieved by local heating via the thin-film structure and associated thermomagnetic changes. The ferrimagnetism results from the magnetic moments of three different sublattices in the material. These three magnetic moments depend differently upon temperature. The typical dependence of the net magnetization resulting from the superposition of the sublattice magnetizations upon temperature is shown in fig. 5. At a certain temperature, called the compensation temperature, the sublattice magnetizations cancel each other with the result that the net magnetization of the film vanishes. At this temperature, it is not possible to change the pattern of domains by any external magnetic field for the reason that the field cannot couple to the compensated spins of the atoms.

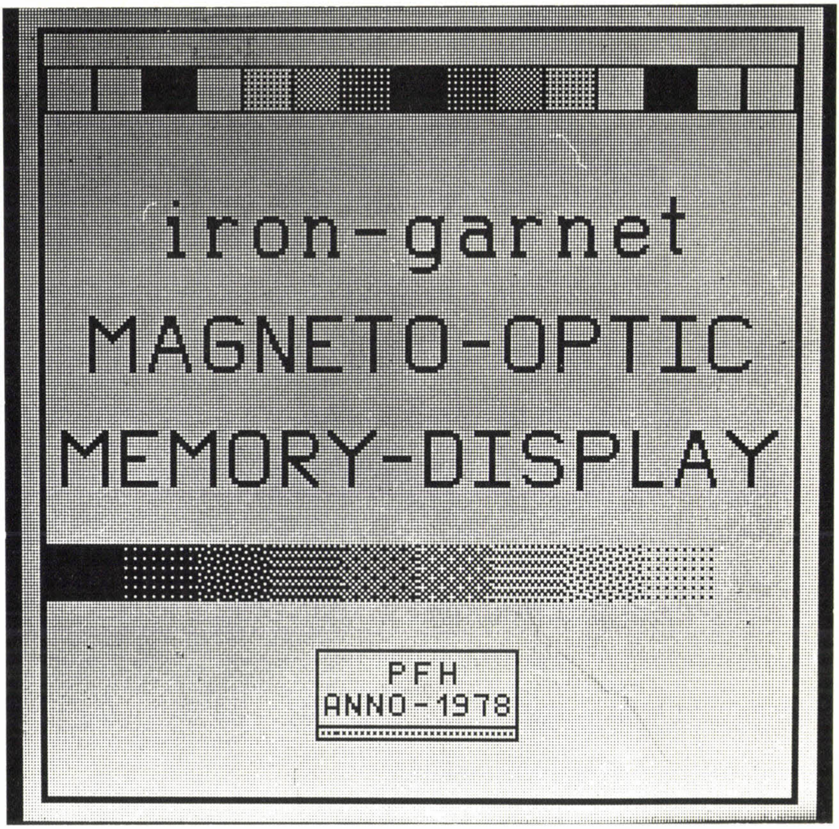


Fig. 3. Magneto-optic iron-garnet wafer viewed in polarization optics (256×256 memory cells, pitch: 70 μm).

50%

The compensation point is located at room temperature in experimental devices which is achieved by proper molecular engineering of the composition of the material. At room temperature, any pattern of domains in the film is, therefore, absolutely stable and insensitive to external magnetic fields; it is stored in a nonvolatile way.

A first condition to achieve switching of the direction of magnetization under the action of an external field is, therefore, the setting-up of a deviation of the temperature from the temperature of the compensation point in order to achieve a net magnetization. However, the existence of a net magnetization is not yet sufficient to really get switching at acceptable values of the external field strength. This is, because the uniaxial anisotropy of the film forcing the magnetizations to align perpendicularly to the plane of the film is too high. This

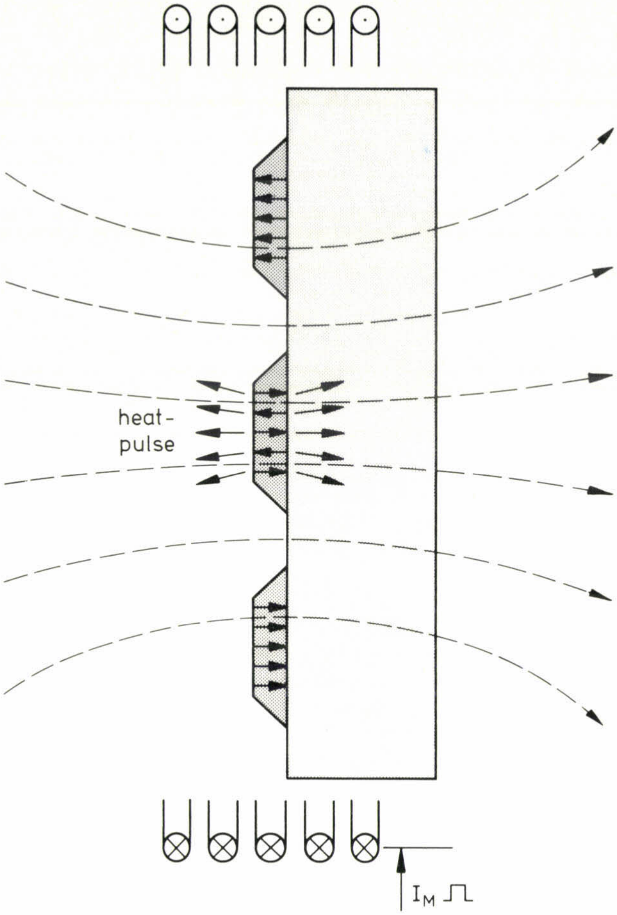


Fig. 4. Switching of the magnetization of a memory cell by applying a heat pulse and an external magnetic field.

anisotropy must be reduced to allow reasonable switching fields. Fortunately, it has been found that the anisotropy can be drastically reduced by inducing a local stress in a domain. This occurs by rapidly heating-up a domain with respect to its surrounding, i.e. by setting-up a temperature gradient.

Accordingly, switching of the direction of magnetization of a domain is achieved in two steps:

(1) applying an external magnetic field of the proper directions and,

(2) inducing a temperature increase and temperature gradient by a local heat pulse.

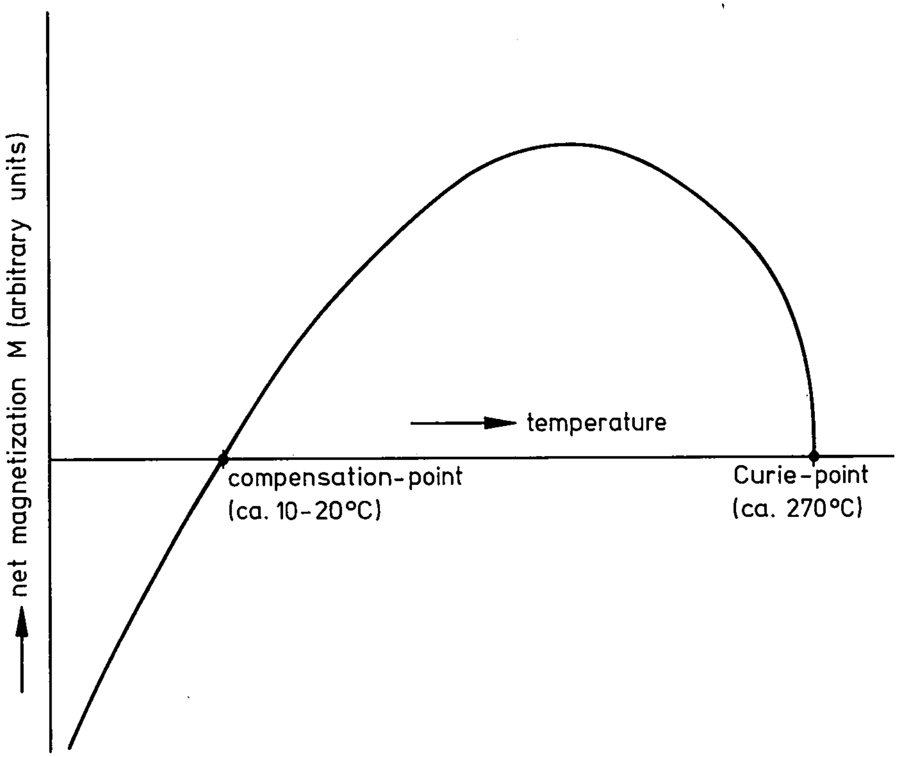


Fig. 5. Net magnetization of the iron-garnet memory film as function of the temperature.

Other domains not being heated, remain unaffected by the external magnetic field.

In experimental devices, a magnetic field of ca. 100 Oe has to be applied together with a heat pulse that raises the temperature of a domain to be switched to about 20–30 °C above the compensation temperature. Typically, the heat pulse is applied for ca. 10 μs. The switching can be repeated at any time without deteriorating the film.

4. An electronic switching technique

There are various methods to apply a local heat pulse in the magneto-optic film switching.

A first one is based on absorption of light from a laser beam focussed on the film [6,9,11]). A second more sensitive method uses ohmic heat produced locally in a photoconductive layer which is deposited on the magneto-optic film and exposed by a light spot [13,14]). Both these methods are applied in magneto-optic memories for digital data storage.

For fast light switching in display components another method has been developed, based on the generation of ohmic heat in electronically addressed thin-film resistances. A sketch of such magneto-optic light-modulation cell realized in a thin-film technology is shown in fig. 6. The island of magneto-optic material is covered by a resistive layer. This layer is connected to an electronic currrent source via metallic thin-film conductors. If a current pulse

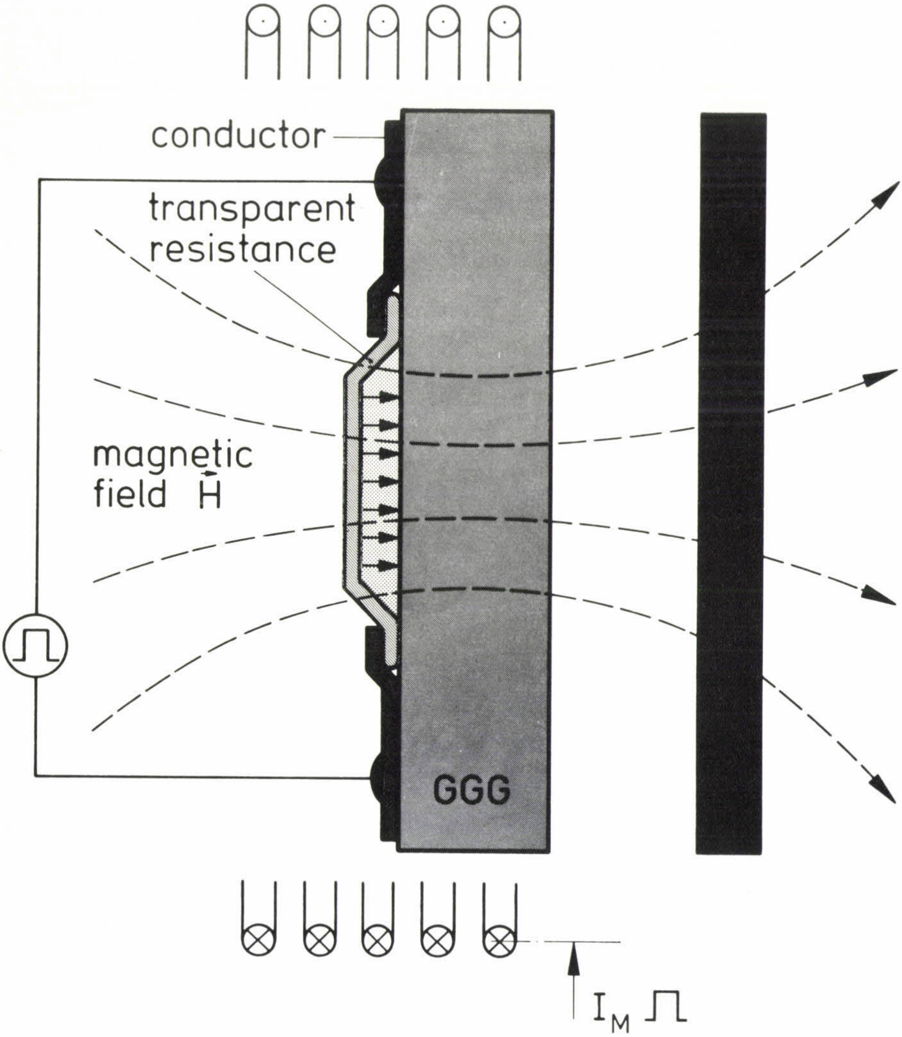


Fig. 6. Electronic switching via a transparent resistance layer.

is applied, then, ohmic heat is produced in the layer giving rise to a temperature increase in the adjacent memory cell. According to thermal diffusion, the switching process takes about 1–10 μs, depending on the thickness of the magneto-optic film.

For the viewing of the direction of magnetization of the memory cell in transmission, the resistance layer has to be transparent, for example, by using a tin–indium-oxyde layer.

5. Linear- and x–y-addressed arrays

A large number of electronically switchable magneto-optic memory cells can be integrated on a single substrate including the electronic wiring. The example of the concept for a linear light modulation device is sketched in fig. 7. A line of electronically switchable magneto-optic memory cells is assumed to be connected in parallel to an electronic drive circuit, that consists of a shift register with sequential input and parallel output and as much buffer- and drive-stages as there are memory cells on the chip.

The digital information of a line is read into the shift register sequentially and afterwards to buffer- and drive-stages in parallel. The drive switches are clocked in synchronism to produce the current pulses for heating-up the memory cells simultaneously.

Any transparency pattern of the memory cells is set-up in two steps. One of the possible modes of operation starts by switching all the cells to the non-transparent state. Then, after having reversed the externally applied magnetic field and after having loaded the buffer stages with the respective information, the magnetization of those cells is being reversed that shall become transparent. Another method is to switch with two opposite patterns of the information. This requires more expenditure of the electronic pattern generation, but, on the other hand, requires less drive power since the total number of magnetic switching events is smaller.

A two-dimensional array of electronically switchable memory cells based on a rather simple x–y-addressing technique is sketched in fig. 8. Though the x–y-addressing network is a linear one (only connections to resistances are made when looking from the electronic circuit) selective switching of a cell at the crosspoint of a line electrode and column electrode can be achieved. This is due to the fact that the switching characteristic of the magnetization exhibits a well-defined threshold. For a given external magnetic field, the threshold is determined by the heating current or vice versa. Since crosstalk currents in an x–y-addressed net can always be kept below the threshold current, they will not effect memory cells aside the one at an addressed crosspoint.

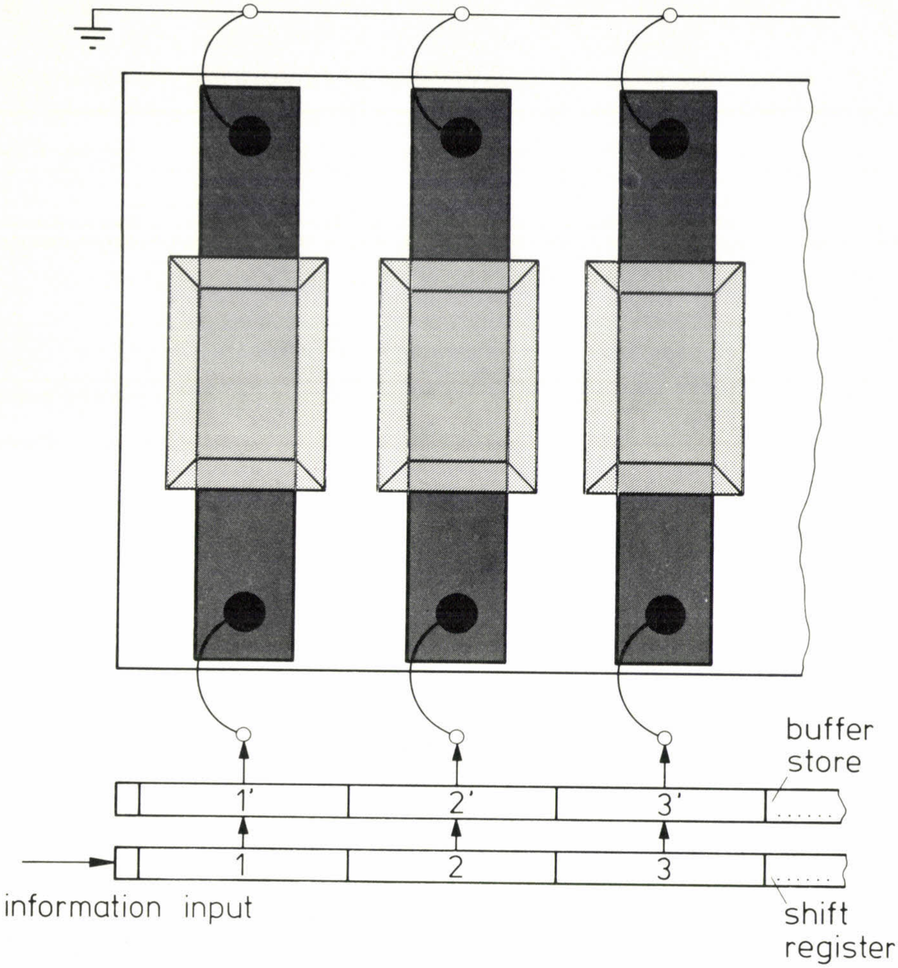


Fig. 7. Iron-garnet light-modulation line with parallel electronic control.

6. An experimental 256-bit modulation line

A line of 256 memory cells switched by direct electronic control has been developed. The memory cells with the size of 50×50 μm^2 are arranged at a pitch of 60 μm. An enlarged view on the component and the memory cells is shown in fig. 9. Each cell is partly covered by a tin–indium-oxide resistance layer connected to 4 integrated electronic circuits which are cemented on the same substrate. The connections are being made by thin-film conductors and ultrasound bonding. In fig. 10 a view on the component with the integrated

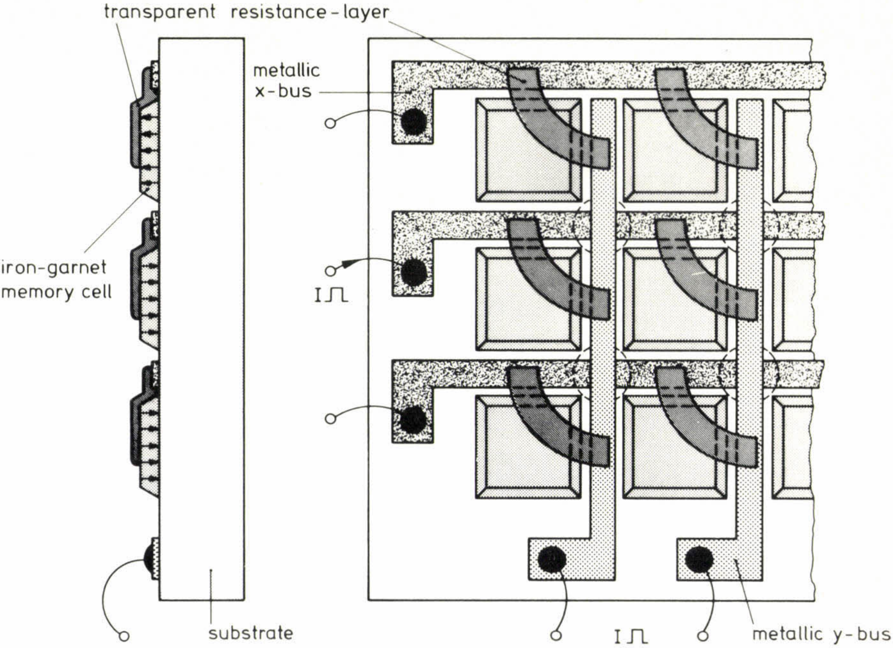


Fig. 8. The *x*-*y*-addressed iron-garnet light-modulation array.

circuits is shown. The integrated circuits consist of a 64-bit shift register each and 64 buffer stores and drive stages at the output respectively. The information can be written-in electronically at a speed of up to 20 Mbit/s. The optical display is possible every 10–20 μs. A switching event with the light which is transmitted by a memory cell measured by a photodiode is displayed in fig. 11. As has been pointed out in the last chapter, there is a threshold for either the heating current at a given magnetic field or for the magnetic field at a given heating current to achieve switching. For this experiment, a constant heating current has been chosen. Therefore, the memory-cell switches from one state to the other immediately after the magnetic field applied via an external magnetic coil reaches its threshold value. The switching requires the application of an electronic pulse of 15–50 mW depending on the composition of the material. The drive voltage has been chosen ca. 10 V.

The optical efficiency in transmission depends on the Faraday rotation and absorption within the material. For the spectral range of 550 nm, the overall optical efficiency reaches 10 %, the Faraday rotation being 26° in the 4-μm thick iron-garnet layer. The corresponding optical contrast ratio between dark and bright elements comes up to ca. 60.

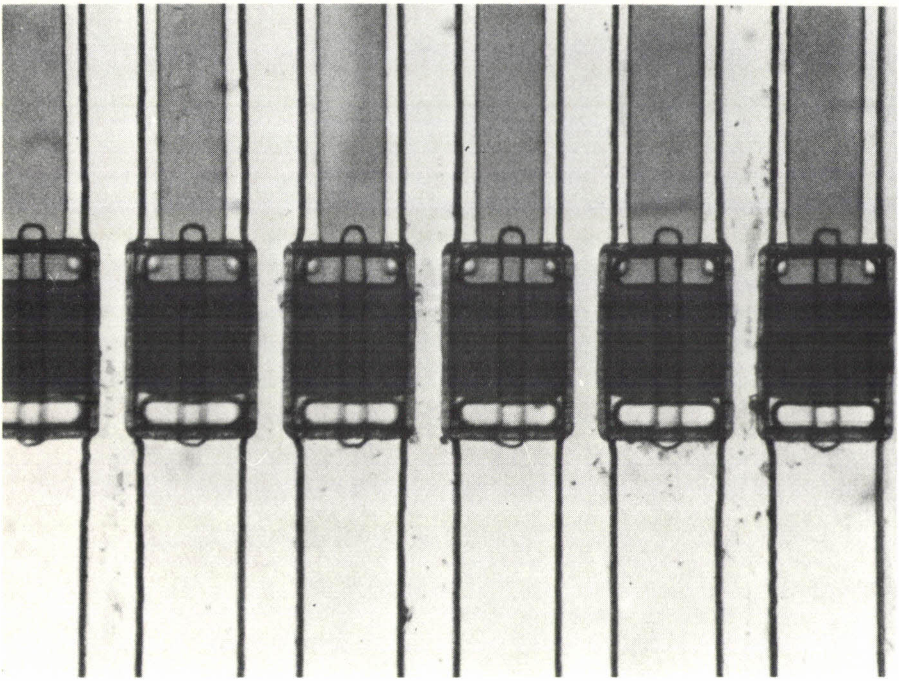


Fig. 9. Enlarged view on a line of memory cells covered by stripes of thin-film resistances connected to metallic conductors.

7. Fields of application and conclusions

The technique of switching a pattern of domains in a magneto-optic storage film thermomagnetically by direct electronic control via an evaporated resistance network has proved to be suitable for the implementation of integrated light-modulation lines or *x–y*-addressed arrays of high optical resolution. On the basis of parallel electronic control, information can be optically displayed at a rate of up to 20 Mbit/s. Fatigue problems have never been observed even not at high cycle frequencies. The overall optical efficiency for polarized light with presently available materials comes up to 10% at the wavelength of $\lambda = 550$ nm, a value already acceptable for many applications. Further improvements are subject of research. The new technology has been demonstrated with an experimental 256-element light-modulation line driven by integrated circuits. An *x–y*-addressed array is under development.

An important application for a modulation line is the field of optical printing. There are already optical printers on the market using mechanically deflected

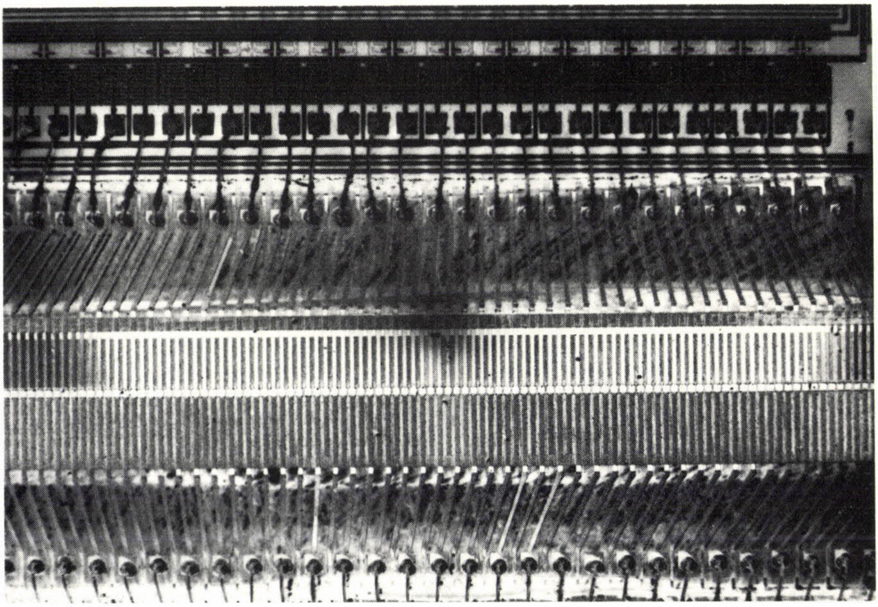


Fig. 10. View on a part of the 256-bit modulation line with integrated circuits on the garnet substrate.

laser beams for line exposure [18,19]. This rather expensive and voluminous exposure technique could be replaced by a high-resolution integrated iron-garnet line with a simple lamp for illumination and an objective lens for the projection on the light sensitive medium (either photographic material, dry-silver paper or a photoconductor drum depending on the type of printer). An essential advantage aside of the cost aspects would be the perfect geometrical stability of the point raster produced along the line and the high life time that can be expected.

Two-dimensional *x–y*-addressed arrays could find their main field for application in data output terminals with inherent memory. The transmission image of the iron-garnet component can either be projected on a screen using slide-projector-like technique or, with an electroluminescent layer behind, viewed directly. This latter technology could lead to a high-resolution flat-pocket data display. Other professional applications are found in the field of optical filtering (controllable spatial filter) or page composing in holographic storage devices [20].

A grey scale can also be produced by switching each memory cell at higher frequency with controlled on–off ratio. Another possibility would be, to attach

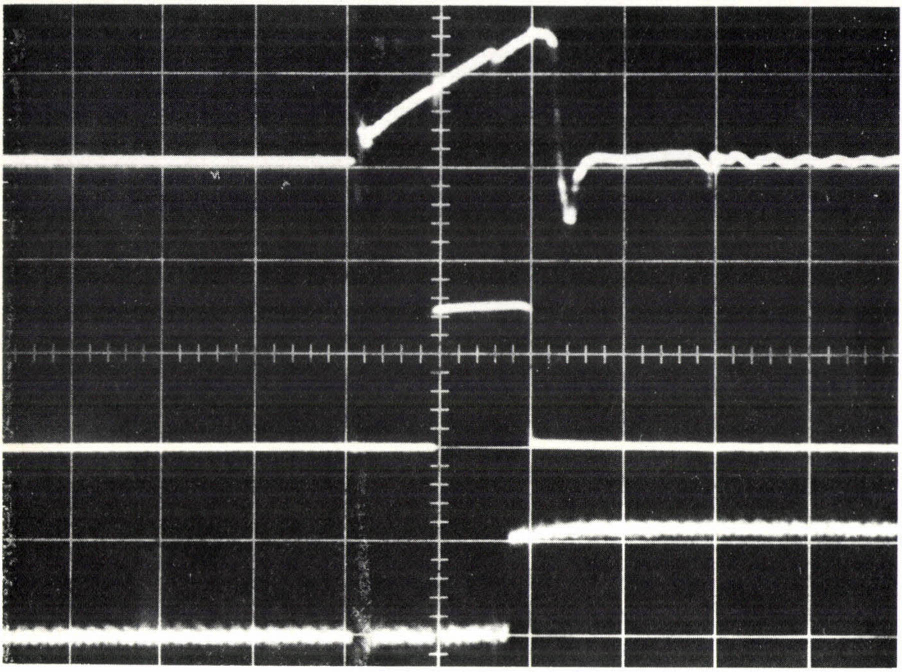


Fig. 11. Switching of an iron-garnet memory cell. Upper curve: magnetic field current, center curve: heating current, lower curve: light signal, horizontal: 20 μs/div.

a number of memory cells to one image point to control the average transmission by point-density modulation and viewing this beyond the resolution limit. This method is demonstrated in fig. 3 by bars with different densities of the "black" memory cells.

Acknowledgement

The authors are gratefully indepted to W. Tolksdorf, D. Mateika and G. Bartels for growing the garnet materials, P. Hansen and K. Witter for materials research, J. Lehmann and G. Much for thin-film technology, D. Riekmann and Lullies for electronic control circuits and R. M. G. Wijnhoven, B. J. Mulder and J. C. M. van den Akker (Philips Research Laboratories, Eindhoven) for fabricating the integrated drive circuits.

The described work was sponsored by the German Federal Ministry for Research and Technology (BMFT) under grant no. NT 0821 0. Only the authors are responsible for the contents of this publication.

Philips GmbH Forschungslaboratorium, Hamburg *Hamburg, August 1978*

REFERENCES

[1] L. Mayer, J. Appl. Phys. **29**, 1003, 1958.
[2] J. I. Dillon, jr., J. Phys. Radium **20**, 374, 1959.
[3] R. P. Hunt, IEEE Trans. **MAG-5**, 700, 1969.
[4] J.-P. Krumme, J. Verweel, J. Haberkamp, W. Tolksdorf, G. Bartels and G. P. Espinosa, Appl. Phys. Lett. **20**, 451, 1972.
[5] R. L. Aagard, T. C. Lee and D. Chen, Appl. Optics **11**, 2133, 1972.
[6] O. N. Tufte and D. Chen, IEE Spectrum **10**, 26, 1973.
[7] B. R. Brown, Appl. Optics **13**, 761, 1974.
[8] J.-P. Krumme and H. J. Schmitt, IEEE Trans. **Mag.-11**, 1097, 1975.
[9] B. Hill, J.-P. Krumme, G. Much, R. Pepperl, J. Schmidt, K. P. Schmidt, K. Witter and H. Heitmann, Appl. Optics **14**, 2607, 1975.
[10] W. Tolksdorf, IEEE Trans. **MAG-11**, 1074, 1975.
[11] J.-P. Krumme, P. Hansen and K. Witter, J. Appl. Phys. **47**, 3681, 1976.
[12] D. F. Buhrer, J. Appl. Physics **40**, 4500, 1969.
[13] J.-P. Krumme, B. Hill, J. Krüger and K. Witter, J. Appl. Phys. **46**, 2733, 1975.
[14] H. Heitmann, B. Hill, J.-P. Krumme and K. Witter, Philips Tech. Rev. **37**, 197, 1977.
[15] D. E. Lacklison, G. B. Scott, R. F. Pearson and J. C. Page, IEEE Trans. **MAG-11**, 1118, 1975.
[16] G. B. Scott and D. E. Lacklison, IEEE Trans. **MAG-12**, 292, 1976.
[17] D. E. Lacklison, G. B. Scott, A. D. Giles, J. A. Clarke, R. F. Pearson and J. L. Page, IEEE Trans. **MAG-13**, 973, 1977.
[18] P. Graf, J. Photographic Science **25**, 186, 1977.
[19] W. Meye, J. Photographic Science **25**, 183, 1977.
[20] B. Hill, Appl. Optics **11**, 182, 1972.

DIGITAL SIGNAL PROCESSING AND LSI IN MODEMS FOR DATA TRANSMISSION*)

by L. E. ZEGERS and N. A. M. VERHOECKX

Abstract

Digital signal processing and LSI are steadily being introduced in the field of data transmission. Still there is a noticeable discrepancy between existing design possibilities and actual practical application. The backgrounds of this contrast are briefly considered in this paper. A short survey on data modems is presented showing the wide diversification that has to be coped with. The main part of the paper is devoted to a few major developments that have taken place in the last five years. Particular attention is paid to the related work that has been carried out at Philips Research Laboratories.

1. Introduction

The data transmission field provides an excellent stimulus for the investigation of digital signal processing as well as for the inherent development of the art of LSI circuit design. Quite advanced data transmission equipment using integrated digital processors for the implementation of various electronic functions, such as filtering and modulation, is already commercially available. Nevertheless, there is still a noticeable discrepancy between existing design possibilities for advanced LSI circuits and actual practical application.

In the present paper, the backgrounds of this contrast between ability and application will briefly be considered. A general survey on data modems will be presented and the possible role of digital signal processing in coping with the wide diversification of modems will be outlined.

The main subject of the paper will be the presentation of a few major developments in the data transmission field over a period of the past five years. This will show the fast progress which has been made in the application of LSI circuits. Particular attention will be paid to the philosophy of the work carried out at Philips Research Laboratories.

2. Ability in LSI circuit design and application considerations in data transmission

2.1. *General considerations*

In the past few years considerable progress has been made in digital signal

*) Paper based on a lecture given by the first author at a meeting of the Nachrichtentechnische Gesellschaft (NTG) on VLSI in Berlin, January, 18–20, 1978.

processing and LSI circuit realization in general. Technologies have been gradually developed to offer continuously increasing processing speeds from digital LSI circuits of growing complexity with production yields that meet economical standards. In close interaction with system research this has stimulated the conception of programmable digital sub-systems of various kinds. In data transmission equipment, in particular in those devices (usually called modems) that connect a data terminal to a transmission line, this has led to the possibility of fully integrated digital filters and modulators and other special purpose processors.

On the other hand, the actual application of these possibilities has remained subject to relevant necessity as determined by system constraints. To some extent prestige arguments may also have some influence on application decisions.

The situation which has to be envisaged by the modem manufacturer is as follows. The variety of standardized and non-standardized modulation techniques is quite substantial while system requirements, e.g. with regard to performance, reliability, costs, flexibility, power consumption, maintenance, diagnostics etc., have an important impact on design. The electronic functions within a modem which can be integrated represent a relatively small percentage of overall costs (table I).

In view of the mentioned diversification and taking account of the given system constraints, the modem manufacturer will therefore in general try to define functional electronic sub-systems which can be used in multitude over the full range of his equipment. Monolithic integration of such units may then be economical and meet other system requirements.

TABLE I

Indication of cost of electronic functions (such as coding, decoding, modulation, demodulation, filtering, equalization, synchronization and echo cancellation) that can be performed by integrated circuits as a percentage of overall cost for different types of modems.

type of modem	percentage of cost (%)
modems for 0.2–4.8 kbit/s	< 20
modems for 4.8–9.6 kbit/s	< 30–40
baseband modems	< 40–50

2.2. *A survey of modem diversification*

A great variety of data modems exists for transmission rates ranging, roughly speaking, from 10^2 bits per second up to 10^5 bits per second (table II). For the lower rates (up to 1.2 kbit/s) asynchronous frequency shift keying (FSK) is most commonly used. In the region between 2.4 and 9.6 kbit/s several modulation schemes (phase modulation and combined phase and amplitude modulation) with different carrier frequencies have been standardized by the CCITT. For wideband circuits a separate class of modems is generally used. More recently some other types of coding are being promoted in different countries for application to baseband transmission in data networks.

TABLE II

A survey of modem diversification

type of modem	bit rate (kbit/s)	symbol rate (symbols/s)	modulation/ coding scheme	carrier frequency (kHz)	CCITT recommendation
voice-band (asynchronous)	0.2–0.3	200–300	FSK	0.98–1.18 1.65–1.85	V21
	0.6	600	FSK	1.3 –1.7	V23
	1.2	1200	FSK	1.3 –2.1	
voice-band (synchronous)	2.4	1200	PSK (4φ)	1.8	V26
	4.8	1600	PSK (8φ)	1.8	V27
	9.6	2400	mod. 4φ–4A	1.7	V29
wideband (60–108 kHz)	48 (40.8)		VSB	100	V35
	48–72		SSB	100	V36
baseband (2 or 4 wire)	1.2–19.2		biphase, AMI, delay modulation		

2.3. *Possibilities for digital signal processing*

This diversification of modems offers some interesting possibilities for the application of digital signal processing. First of all this is due to the inherent flexibility of programmable digital circuits.

Other important aspects of digital processing are the ease with which signals can be stored, the well-defined accuracy with which operations are performed and the possibility of introducing new functions. For different reasons (e.g. power consumption and physical dimensions) digital circuits must be implemented by means of monolithic integration, but at the same time this is also a very attractive feature. Of course certain limitations are set by the allowable circuit complexity and the available technology. The technology dictates, for example, the maximum obtainable density of integrated elements, the power consumption and the yield. For each particular application the right choice has to be made between special purpose LSI and general purpose LSI, the right balance has to be found between applicable hardware and software and appropriate testing facilities have to be provided. The allied considerations are of such complexity that they cannot suitably be handled from a technical point of view only; commercial and other aspects have to be taken into account right from the beginning as well.

3. A schematic overview of an evolving development

After the foregoing practical considerations made from an application point of view, we will now turn our attention to a schematic overview of the evolving development in the field of data modems. We will do this by means of examples taken from the research program of the Philips Research Laboratories. In this paper we will restrict ourselves to essentials. More detailed information can be found in the references cited.

3.1. *Digital filters of the first generation*

As a first example we mention the very basic function of filtering. In 1973 we started in our laboratory to study the feasibility of a universal finite impulse response (FIR) digital filter suitable for integration. A straightforward translation of known elementary principles into a single LSI module was out of the question because of the need for either a great number of multipliers or a very high internal operating frequency. However, we solved this problem by cascading a transversal filter and a recursive filter (fig. 1). In this way the required impulse response was obtained by means of a differential approximation. By additionally constraining the values of the coefficients of both filters to integral powers of two, the multiplications could be greatly simplified. We have called this type of filter a "difference routing digital filter" (DRDF, ref. 1). For the determination of the filter coefficients a simple design procedure can be used, which guarantees stability of the composite filter and finite duration of the impulse response. In figure 1 the basic block diagram of the DRDF with parallel structure is shown. The main components are the memory stages for the signals, the coefficient multipliers and the two adders. The required coefficient memory is not explicitly shown. In figure 2 the more practical realization with a sequential transversal structure is represented. The n-bit input signals, which may be obtained from an analog-to-digital converter are stored in n recirculating shift registers.

From figure 2 the term "routing circuit", with which the coefficient multiplier is denoted, becomes more obvious: with the previously mentioned solution of differential approximation of the impulse response, and with the additional constraint that the coefficients shall be integral powers of two, the process of multiplication reduces to shifting and/or inverting. Each multiplication by an integral power of two requires only one shift of one of the stored n-bit input signals over a particular number of binary positions. The properties of the

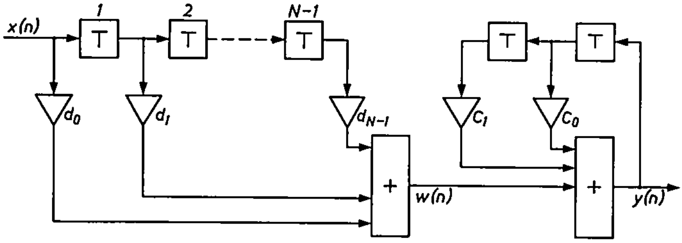


Fig. 1. Basic diagram of a difference routing digital filter (DRDF); the coefficients in both the transversal part (d_0, d_1, . . ., d_{N-1}) and the recursive part (c_0, c_1) are zero or integral powers of two. The transversal part is shown here in a parallel structure.

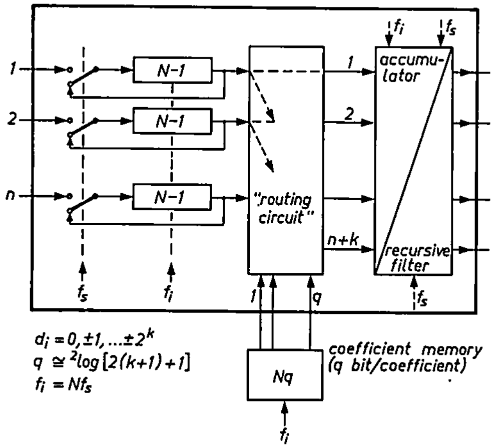


Fig. 2. Block diagram of difference routing digital filter with sequential transversal structure.

DRDF can be summarized as follows:

- digital filter suitable for realization as one LSI integrated circuit (ca 10 mm^2 in I^{2}L technology) due to simplified multiplications;
- low-pass, band-pass and high-pass filters are feasible given a correct choice of the recursive filter part;
- simple design procedure in the time domain;
- attenuation and accuracy are well balanced with the requirements to be met in data transmission;
- typical stand-alone filter with desirable versatility.

This last property, however, also implies the disadvantage that the DRDF does not lend itself to further simplification as is possible, for example, in interpolating digital filters. Nor does it lend itself to further overall optimization of modem design, as obtainable, for example, by combination of different modem functions.

3.2. *Integrated data transmitter*

The previous research topic and the resulting conclusions paved the way to the problem we tackled in 1974: the realization of an integrated linear modulation stage (ref. 2). This functional unit could serve as an almost fully integrated data transmitter for different types of linear modulation. In general such a modulation stage comprises a premodulation filter, a modulator and a postmodulation filter (fig. 3*a*). Realization as a digital integrated circuit became feasible by virtual elimination of the digital modulator and postmodulation filter (fig. 3*b*). To this end the premodulation filter was designed as an inter-

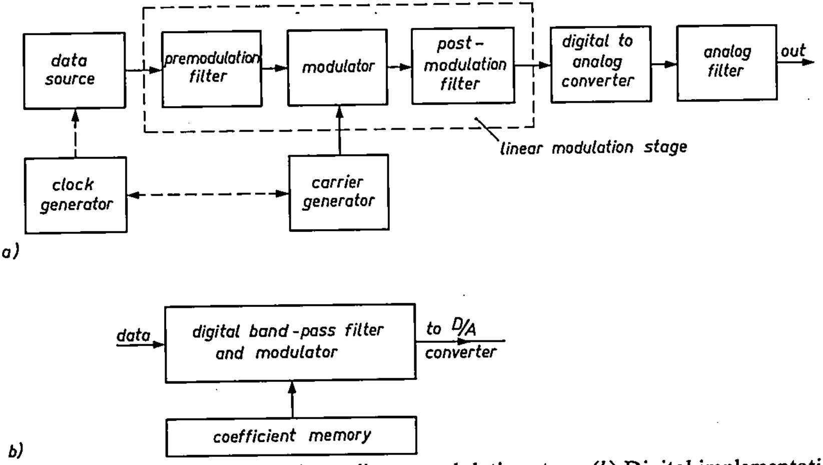


Fig. 3. (*a*) Modem transmitter based on a linear modulation stage. (*b*) Digital implementation of a linear modulation stage, which has now been reduced to a digital band-pass filter and a controlled inverter that functions as modulator. The filter characteristics are determined by the coefficients which are stored in a digital memory.

polating digital band-pass filter, in which use was made of the periodic spectral properties of digital signals. The desired frequency characteristic of this band-pass filter is programmed by means of an external coefficient memory.

More details are given in figs 4*a* and 4*b*. Digital signals with a sampling frequency $f_d = \omega_d/2\pi$ have a spectrum that is periodic with ω_d. Each half period of the spectrum contains all information of the digital signal. Therefore a band-pass filter situated at $K\omega_d/2$, where K is an arbitrary integer, can in principle be used to obtain a band-limited signal that can be shifted to a prescribed frequency position in the telephone channel by subsequent modulation of a carrier. Because the digital filter also has a periodic frequency characteristic one has to take care that the sampling frequency of this filter ($\omega_h/2\pi$) is a multiple of the sampling frequency of the input signal ($\omega_d/2\pi$). This procedure, which causes the spectral periodicity of the filter characteristic to be larger than that of the signal to be filtered, is called interpolation. Figure 4*a* shows that in the band-pass filter a sampling rate increase factor or interpolation factor of M is applied. At the output of the filter the sampling frequency can be reduced again by a factor L. This reduction is important in order to achieve a

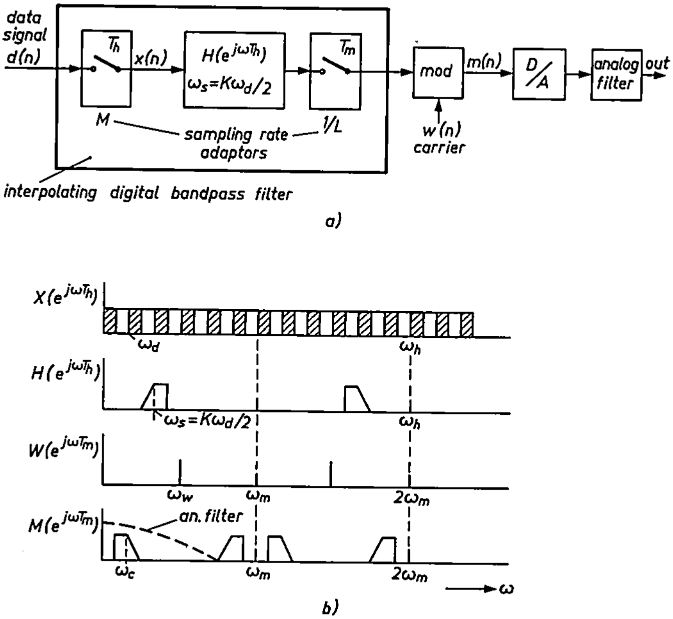


Fig. 4. (*a*) More detailed block diagram of digital linear modulation stage. (*b*) Schematic spectral representation (for $K = 4$, $L = 2$ and $M = 12$) of the main digital signals occurring in the linear modulation stage. The spectra $X(\exp j\omega T_h)$, $W(\exp j\omega T_m)$ and $M(\exp j\omega T_m)$ correspond to the digital signals $x(n)$, $w(n)$ and $m(n)$ respectively; $H(\exp j\omega T_h)$ is the transfer function of the digital band-pass filter. In the actual design K = variable, $L = 4$ and $M = 23$.

simple harmonic relation between sampling frequency and carrier frequency, by which the modulator reduces to a simple controlled inverter.

Summarizing we can state that there exist some particularly favourable combinations of values for the factors M, L and K, given by

$$\omega_c = \left(\frac{M}{L} - K\right)\frac{\omega_d}{2},$$

where ω_c is the required carrier frequency of the modulated signal. For these combinations the operation of modulation can be performed by an invertor circuit and the postmodulation filtering can be realized with a simple RC network. It can be shown that the introduction of the interpolation with a factor M does not necessarily require an increase of the number of input signal memory stages by the same factor.

At our laboratories we managed to implement the modulation stage as a single LSI module in I^2L technology. In table III the main parameters of the module are tabulated, and a photograph of the integrated circuit is shown in fig. 5. The whole circuit comprises about 800 logical gates. The IC measures 3.6×3.7 mm^2, and a yield of more than 20% was obtained. The total power consumption is typically only 10 mW. Figure 6 shows a photograph of a vestigial side band (VSB) data transmitter built with the integrated linear modulation stage. All filter coefficients are stored in an external ROM. The output signals are converted to analog by an 8-bit digital-to-analog converter. For application to other modulation schemes our modulation stage has to be supplemented with a small amount of additional random logic, because its design has been optimized for VSB modulation.

TABLE III

Main parameters of the integrated digital linear modulation stage

technology	integrated injection logic (I^2L)
area	3.6×3.7 mm^2
number of logic gates	811
number of pins	32
power consumption	10 mW
supply voltage	5 V
output format	8 bit parallel
external clock (nominal)	$\cong$ 220 kHz
external clock (maximum)	550 kHz
coefficient memory	external ROM
compatibility	TTL compatible

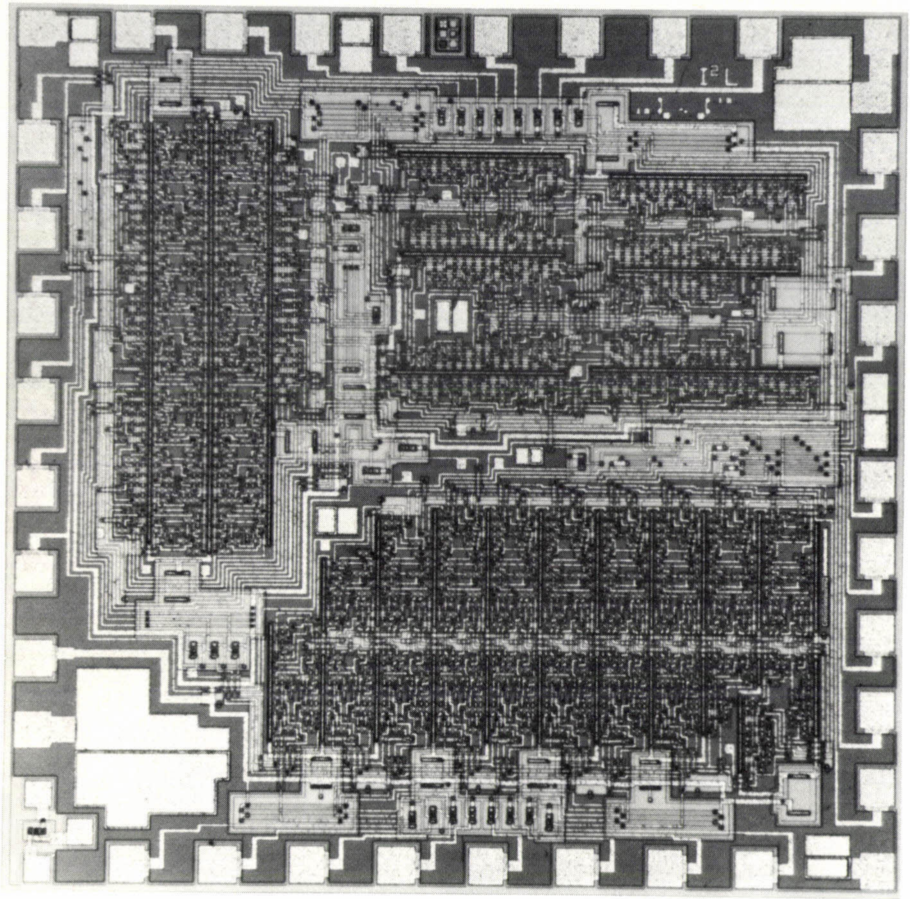


Fig. 5. Photograph of the integrated digital linear modulation stage.

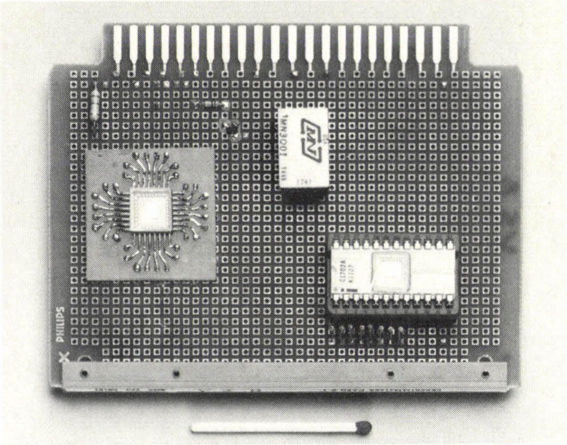


Fig. 6. Vestigial side band (VSB) data transmitter built with the integrated digital linear modulation stage. The additional components are an external ROM for the filter coefficients and an 8-bit D/A-convertor for the output signal.

3.3. *Microprocessor modems*

In 1975 a general discussion arose about the manifold possibilities for the application of microprocessors. In our laboratory this trend led to the following question: can we implement all standardized synchronous modulation schemes with general-purpose commercially available microprocessors both for the transmitter side and for the receiver side? The solution was found in the choice of versatile basic schemes, which were applicable to all usual types of synchronous modulation (ref. 3). We have implemented these basic schemes, taking some additional special measures. In the transmitter, which is based on an orthogonal structure (fig. 7*a*), we could apply a simplified multiplication algorithm by observing that for all modulation schemes the number of possible amplitude values of the input signals x and y is very restricted. Additionally the required interpolation in the digital filtering is not accompanied by a prohibitive increase in the number of multiplications if the right procedure is pursued.

In the receiver (fig. 7*b*), which likewise has an orthogonal structure, the usual low-pass filters at the outputs can be avoided by applying dual phase compensation to the demodulated data signals. This approach also allowed for a relatively low sampling frequency in the demodulators and at the output of both input band-pass filters. Moreover, the algorithms for the two band-pass filters could be combined to a large extent by their derivation as transformations of one low-pass filter. To this end the central frequency of the band-pass filters has to have a certain fixed relation to the input sampling frequency of the receiver.

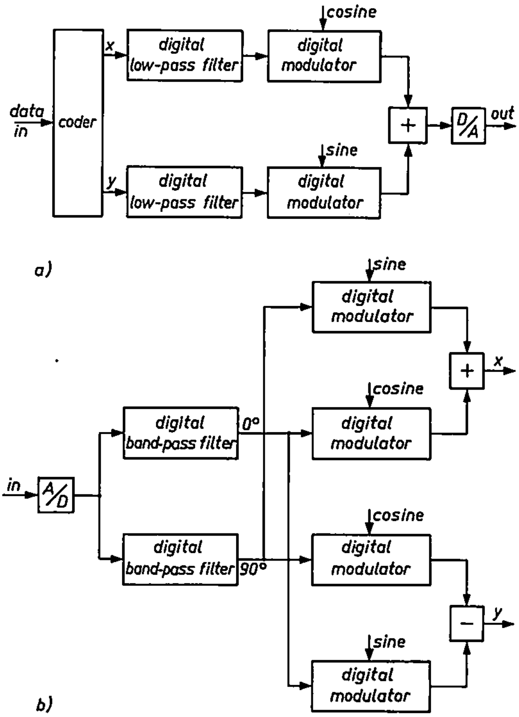


Fig. 7. Versatile digital data transmitter (*a*) and receiver (*b*) suitable for implementation with a microprocessor (basic schemes).

In table IV a survey is shown of the linear modulation schemes which have been implemented with one microprocessor as the transmitter and one microprocessor as the receiver (Signetics 3000). Simplified flow diagrams (fig. 8) indicate the number of micro-instructions per cycle time used in this bit-slice type of microprocessor. The transmitter has a cycle time of about 70 μs (14.4 kHz) which corresponds to 460 micro-instructions, 331 of which are actually used. The great majority (265) of these are necessary for the operations of filtering and modulation. The receiver has a cycle time of 416 μs (2.4 kHz) corresponding to a maximum of 2760 micro-instructions. Here filtering and demodulation require 2290 from the total number of 2367 micro-instructions actually used. So almost all micro-instructions are devoted to multiplication operations. Although the original problem had been satisfactorily solved, one of the important resulting conclusions was that the common type of microprocessor by itself is not very suited to this kind of real time signal processing. The combination of a fast hardware multiplier, when available, with a relatively simple

TABLE IV

Survey of linear modulation schemes, which have been implemented with one microprocessor as the transmitter and one microprocessor as the receiver (Signetics 3000)

type of modulation	bit rate (bit/s)	symbol rate (symbol/s)	carrier frequency (Hz)	number of signal points (x, y)
4φ(A)	2400	1200	1800	4
4φ(B)	2400	1200	1800	2×4
8φ	4800	1600	1800	8
mod. 4φ-4A	9600	2400	1700	16
4×4 AM	9600	2400	1700	16
4 level VSB	4800	2400	2100	2×4

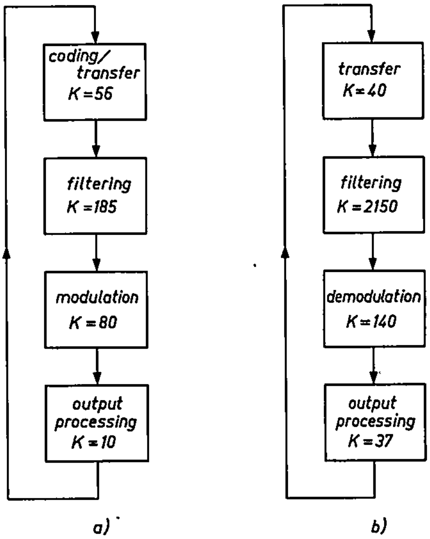


Fig. 8. Simplified flow diagrams of the programs executed by the microprocessor data transmitter (*a*) and the microprocessor data receiver (*b*). The cycle time for the transmitter program is 70 μs and for the receiver program 416 μs; K = number of microinstructions.

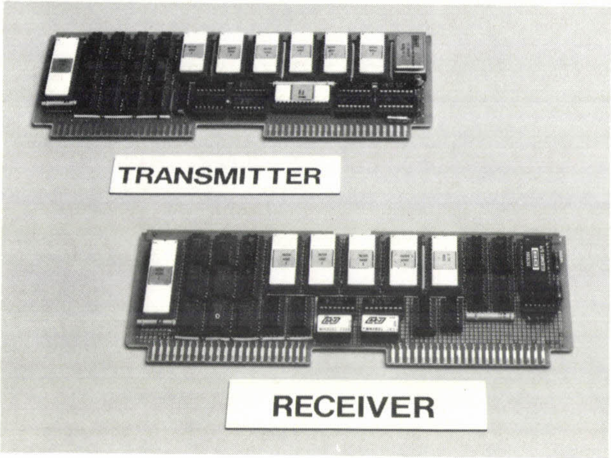


Fig. 9. Photograph of microprocessor (Signetics 3000) implemented data transmitter and data receiver.

microprocessor control unit seemed to be a better choice for the future. By yielding this insight the research work just described has actually stimulated the development of integrated hardware multipliers and their application in modems. This will be underlined by the example given in the next section.

Figure 9 shows the realized microprocessor transmitter and receiver. Although the universal implementation of microprocessor modems is quite attractive, the resulting circuitry is somewhat more complicated than the dedicated special purpose realization of any single modulation scheme. However, with respect to cost of development the microprocessor solution is at an advantage.

3.4. *An adaptive equalizer with general purpose LSI*

For several years now automatic equalizers for data transmission have been studied extensively. Several schemes have been developed and applied. Recently, at our laboratory, we made it our task to design an adaptive fast equalizer on the basis of general purpose LSI-modules. The main incentive was the availability of an attractive, integrated multiplier meant for the calculation of the inner product of two vectors, called a product accumulator. As a starting point for solving the problem we chose the concept of a frequency domain equalizer of the preset type (ref. 4). According to this concept first the discrete Fourier transform (DFT) of a received test signal is calculated. Next the coefficients of a transversal equalizer filter can be computed. The required LSI modules consist

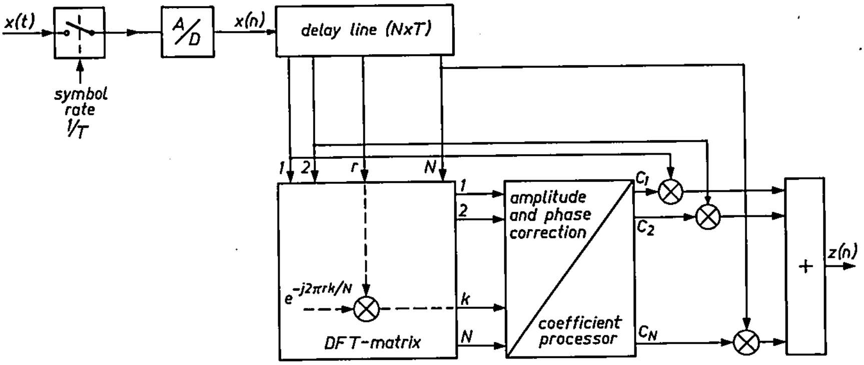


Fig. 10. Block diagram of the automatic digital DFT preset equalizer for data transmission.

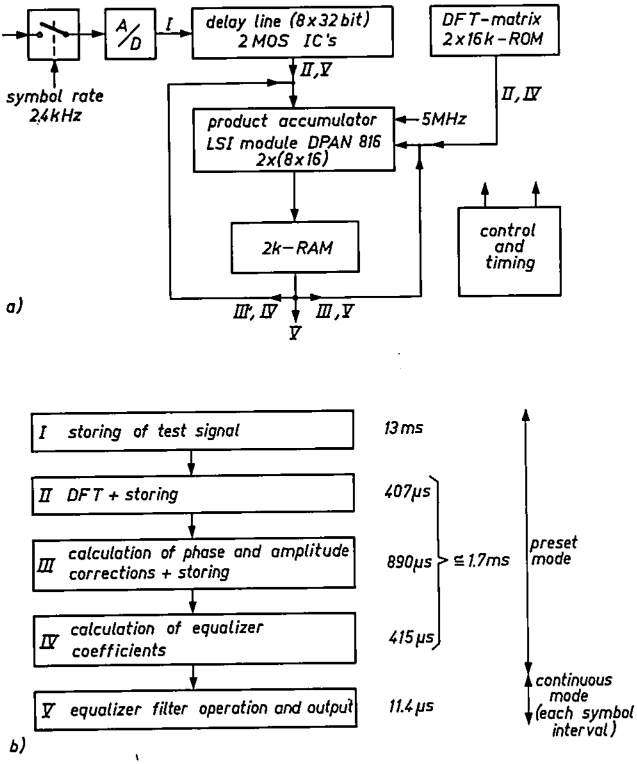

A/D
I
delay line (8x32bit) 2 MOS IC's
DFT-matrix 2x16k-ROM
symbol rate 2.4kHz
II,V
II,IV
product accumulator LSI module DPAN 816 2x(8x16)
5MHz
2k-RAM
control and timing
a)
III,IV
V
III,V
I storing of test signal
13ms
II DFT + storing
407μs
III calculation of phase and amplitude corrections + storing
890μs
≦ 1.7ms
preset mode
IV calculation of equalizer coefficients
415μs
V equalizer filter operation and output
11.4μs
continuous mode (each symbol interval)
b)

Fig. 11. Automatic digital equalizer. Hardware (*a*) and simplified flow diagram (*b*) for a symbol rate of 2.4 kHz and $N = 32$. The roman numbers refer to subsequent states in the flow diagram.

of a fast product accumulator and memories only. In figure 10 the main operations are indicated. The incoming data signal is sampled, the samples are stored in a shift register and converted into N complex Fourier coefficients by means of a DFT matrix. Because the test signal is known in advance, an amplitude and phase correction for each Fourier coefficient can be calculated. The transformation from Fourier coefficients into filter coefficients $c_1, c_2, \ldots, c_N$ of an equivalent transversal equalizer filter is performed by a coefficient processor. The heart of the equalizer hardware (fig. 11*a*) is a twofold product accumulator (DPAN 816), which performs 8 × 16 bit multiplications. A number of successive operation cycles of this product accumulator are used to perform a 32 point DFT, to calculate amplitude and phase corrections and to determine the coefficient values of the corresponding equalizing filter. The parameters of the DFT matrix are stored in two ROMs, whereas intermediate results are written into a RAM. The delay line for the storage of the delayed input data samples requires 2 MOS-IC's.

The simplified flow diagram of fig. 11*b* shows that in the preset mode of operation only 1.7 ms are needed for adjusting the equalizer, apart from the initial 13 ms which are necessary to receive the 32 data samples of the test signal. In the continuous operation mode the calculations in the equalizer filter after the reception of each new input data sample take only 11.4 μs. This implies that during each symbol interval so much time is left that the equalizer can easily be made adaptive, if required. More details of the specific product accumulator will be given in the last example of this paper.

At this point it might be tempting to conclude that with a relatively small number of universal modules quite attractive designs can be made. However, it would be incorrect to generalize. In many cases it is worthwhile or even necessary to resort to special purpose modules. We will elucidate this with the next example.

3.5. *An integrated baseband echo canceller*

After thorough preparatory studies in our laboratories we decided to realize an adaptive integrated echo canceller for full-duplex baseband data transmission on two-wire circuits. The function of an echo canceller is shown in fig. 12 (ref. 5). The transmission of data from a local transmitter and the simultaneous reception of the data from a remote transmitter is possible by application of a so-called hybrid. This hybrid and also the remote one are in general not optimally terminated. Therefore the received data signal is perturbed by local leakage and remotely generated reflections of the own transmitted signal. Due to the correlation between these disturbances and the transmitted data it is possible to discriminate them from the received data and to compensate them

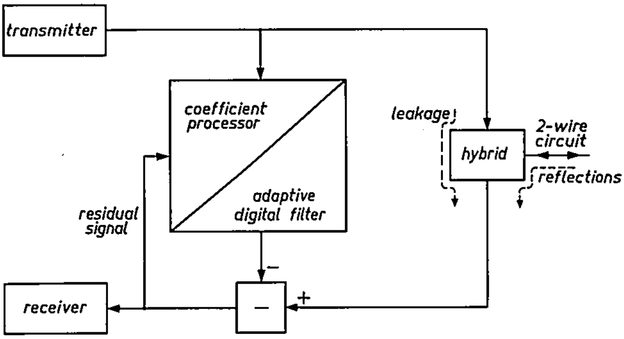


Fig. 12. Adaptive echo canceller for a 2-wire full-duplex data transmission system (basic principle).

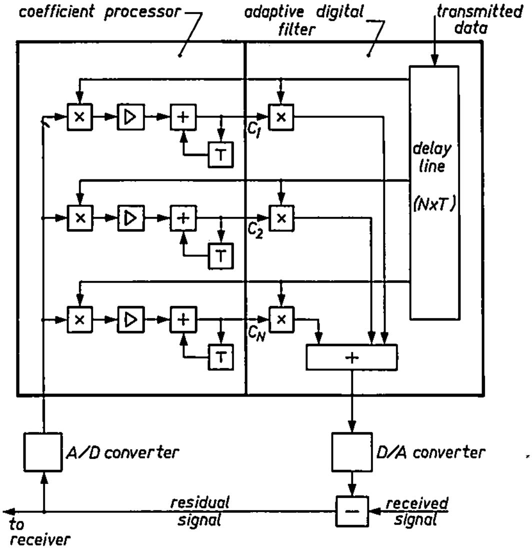


Fig. 13. Block diagram of a digitally implemented adaptive echo canceller.

by means of an adaptive filter and a subtractor. To this end a suitable adaptation algorithm, by which the coefficients of the adaptive filter are continuously updated, has to be available. The algorithm minimizes in an iterative way the correlation between the transmitted data and the residual echo signal obtained after compensation.

The final design of the echo canceller was made on the basis of extensive logical simulations, from which the values of the different parameters could be determined, and a number of experiments on several test models. The circuit comprises 64 filter coefficients and constitutes an extendible module. The design also takes careful account of the constraints imposed by the applied technology.

Figure 13 shows the block diagram of the echo canceller, mainly consisting of a coefficient processor and an adaptive digital filter. The adaptation algorithm is the so-called "stochastic approximation" gradient algorithm, which implies that during each sampling interval each coefficient c_l of the adaptive filter is updated in proportion to the correlation between the transmitted data and the current residual signal. By changing the proportionality factor, a compromise can be found between the speed of convergence and the value of the residual signal after convergence.

As already mentioned, the practical realization was influenced by the applied technology. For example, the maximum allowable "acccumulator width" was limited to 4 bits by the complexity of the required carry-look-ahead circuits. Therefore the echo canceller was split into 4-bit wide logical slices. The chip contains 5 of such slices. The transfer of carries between adjacent slices occurs in successive clock intervals. This does not increase processing time, because the various slices have been made to operate in a pipelined fashion.

TABLE V

Main features and parameters of the integrated adaptive digital echo canceller

general data	
coding schemes	biphase, Miller, AMI
maximum bit rate	19.2 kbit/s
maximum echo length	64 T
residual echo/received signal	$\leqslant -20$ dB
dynamic range of received signal	40 dB
logic design data	
number of coefficients	64
wordlength of coefficients	18 bit
accumulator width	19 bit
logic structure	4-bit slices (5 ×)
LSI-data	
technology	4-phase-dynamic-NMOS
internal clock frequency (max.)	2.5 MHz
area	38 mm^2
complexity	ca. 12,000 MOS transistors
power consumption (max.)	250 mW
yield	20%
supply voltage	12 V

Table V gives the main features and parameters of an integrated echo canceller which will soon become available. The general data indicate that after cancellation a ratio of received signal to residual echo of at least 20 dB over a dynamic range of 40 dB is obtainable. Different coding schemes at different bit rates are applicable. Under the heading "logic design data" some results of the previously mentioned logical simulations and practical experiments are summarized. Finally some data of the LSI module are listed. We have used 4-phase-dynamic-NMOS technology. The area will be approximately 38 mm^2 containing 12,000 MOS transistors. The expected yield is 20%. The first samples of this example of very large scale integration are due to be delivered in the middle of 1978.

3.6. *A fast, integrated product accumulator*

The last circuit that we shall describe here is also an example of very large scale integration. This example is an answer to the already mentioned need for a fast integrated product accumulator (ref. 6). The basic idea which underlies this circuit is that in the calculation of a sum of products $\Sigma\, a_i\, b_i$ $(i = 1, 2, .., N)$ the individual products $a_i\, b_i$ and the subtotals during accumulation are of no interest. Hence it is wasteful to ripple out carries after each single multiplication and addition. A high throughput is obtained by implementation of a parallel array multiplier in synchronous logic. This allows the same array to be clocked as a two-dimensional pipelined sum- and carry-save accumulator, while in each clock period a new pair of operands $(a_i\, b_i)$ is taken in and processed. The total processing time for the calculation of the inner product of two N-vectors with elements (operands) of 8 and 16 bits respectively is $N + 8 + 16$ clock periods; the latter 24 periods are used for ripple-out to empty the array.

Two identical product accumulators, which are meant for alternating operation, are combined into one VLSI circuit, the DPAN 816 (fig. 14). It contains two 8 × 16 bit accumulating array multipliers. During the rippling-out of one array, the other array processes the next pair of input vectors. In this way repetitive multiplication of N-vectors is possible at a rate

$$\begin{aligned} \tfrac{1}{2}(N + 24) \times 0.2\ \mu\text{s}, &\quad \text{for} \quad N \leqslant 24 \\ N \times 0.2\ \mu\text{s}, &\quad \text{for} \quad N > 24. \end{aligned}$$

Several DPAN 816 circuits can be combined in a modular way, if higher precision for the input operands is required.

Table VI summarizes the main features of the integrated product accumulator. The number representation used is the two's complement. The integrated circuit measures 33 mm^2. On this area approximately 15,000 MOS-transistors are realized in 4-phase-dynamic-NMOS technology. Previous experience in-

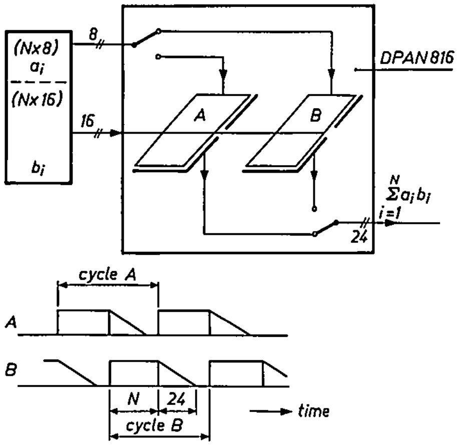


Fig. 14. Integrated dual product accumulator DPAN 816.

TABLE VI

Main features of the integrated dual product accumulator DPAN 816

General data	
function	product accumulation $\sum_{i=1}^{N} a_i b_i$
number representation	two's complement
input word length ($a_i \times b_i$)	8×16 bit
output word length	24 bit
dual circuit	overlapping alternating operation
multi-precision (parallel)	$8p \times 16q$ bits ($p \times q$ chips)
multi-precision (byte-serial)	$8p \times 16$ bits (1 chip)
LSI-data	
technology	4-phase-dynamic-NMOS,
internal clock frequency	5 MHz
area	33 mm^2
complexity	ca. 15,000 MOS transistors (3100 logic gates)
yield	20%
power consumption	50 mW/MHz (250 mW max.)
supply voltage	12 V
clock generator	external
pins	42
interface	24 bits I/O data bus, tri-state, TTL compatible

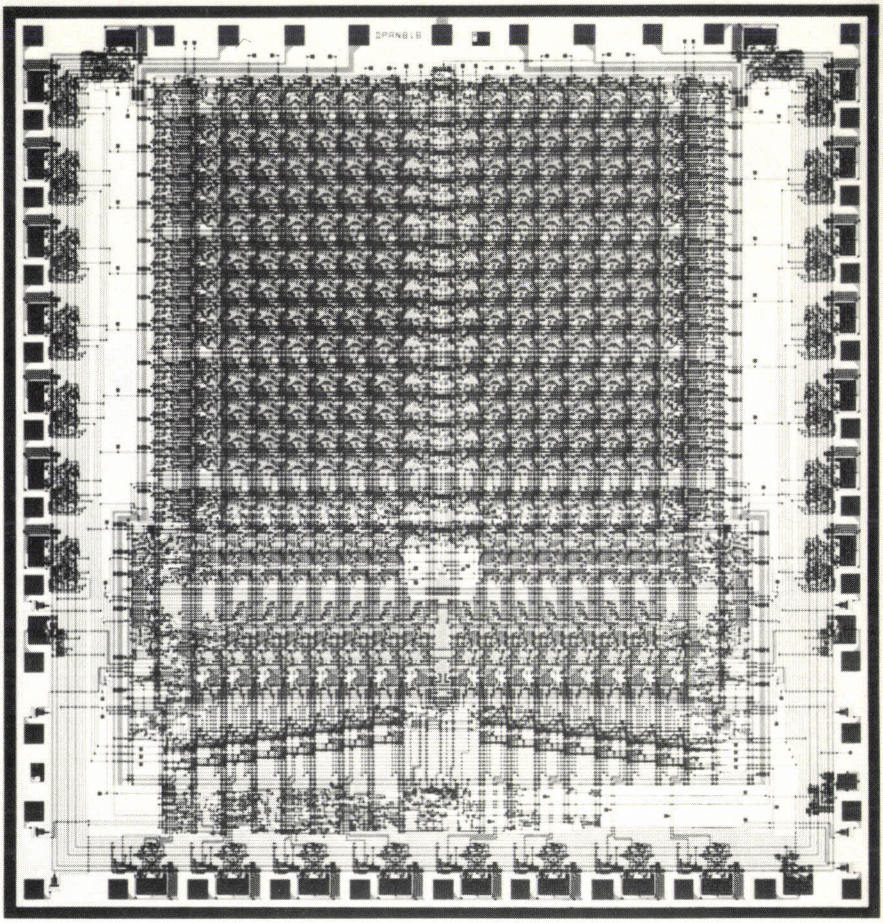


Fig. 15. Photograph of DPAN 816.

dicate an expected yield of 20%. At the maximum clock rate of 5 MHz the power consumption amounts to 250 mW.

In figure 15 a photograph is shown of the DPAN 816. Both array multipliers can easily be recognized, separated by clock lines and control logic. At the edges TTL-compatible three-state interfaces are visible. This LSI circuit will become available in the course of 1978.

4. Conclusions

— Since modems are subsystems in transmission networks, LSI design for modems is not only influenced by inherent contraints like price and versatility, but also by such basic design factors as reliability and diagnostics.

— The deliberate application of LSI modules in data modems is increasing steadily.

— The programmability of digital LSI modules enables the realization of many, in some cases new, electronic functions with relatively few "universal" components.

— Programmable special purpose LSI modules offer a more functional and compact implementation and are sometimes unavoidable anyway.

— During the last few years great advances, both theoretical and practical, have been made. The state of the art will allow the realization of VLSI in the very near future.

— It may be observed, however, that the actual appearance of VLSI strongly relies on an intensive interaction between basic possibilities and practical considerations, in which the system constraints determine the temporary balance.

5. Acknowledgement

The work described in this paper is the result of cooperation between a number of people in our Research Laboratories, most of whom are authors in one or more of the references cited.

Additionally, the authors would like to express their appreciation to Dr W. F. G. Mecklenbräuker for his contribution to the preparation of the original lecture on which this paper has been based.

Philips Research Laboratories *Eindhoven, March 1978*

REFERENCES

[1] P. J. van Gerwen, W. F. G. Mecklenbräuker, N. A. M. Verhoeckx, F. A. M. Snijders and H. A. van Essen, IEEE Trans. on Comm. **COM-23**, 222-234, 1975.

[2] F. A. M. Snijders, N. A. M. Verhoeckx, H. A. van Essen and P. J. van Gerwen, IEEE Trans. on Comm. **COM-23**, 1259-1270, 1975.

[3] P. J. van Gerwen, N. A. M. Verhoeckx, H. A. van Essen and F. A. M. Snijders, IEEE Trans. on Comm. **COM-25**, 238-250, 1977.

[4] F. de Jager and M. Christiaens, Philips Tech. Rev. **37**, 10-24, 1977.

[5] H. C. van den Elzen, P. J. van Gerwen and W. A. M. Snijders, Echo cancellation in a 2-wire full-duplex data transmission system with bipolar encoding, Conference paper NTC 76, Dallas, pp. 8.4.1-8.4.6

[6] N. F. Benschop and L. C. M. Pfennings, A pipelined array product accumulator in dynamic NMOS for efficient signal processing, Proceedings ESSCIRC, Sept. 18-21, 1978, Amsterdam.

SINGLE - CHANNEL ERROR - CORRECTING CONVOLUTIONAL CODES

by J.-M. GOETHALS

Abstract

In some applications information is to be transmitted (or stored) along n parallel channels (or storage media). In this paper we present a method of protecting the information from errors due to occasional or permanent failure in one of the channels (or storage media) by use of convolutional encoding.

1. Introduction

We consider a system designed for the transmission (or storage) of binary information along n independent but synchronous parallel channels (or storage media), and we want the system to be protected against occasional or permanent failure in one of the channels. We assume such a failure has the effect of producing an arbitrary binary sequence at the output of the corresponding channel, independently of the information being transmitted. Thus, we may consider an error sequence of binary digits (modulo 2) has been added to the sequence actually transmitted along the channel in failure. Of course, any binary single-error-correcting block code of length n could be used to encode the information, but this will require at least $\log_2 (n + 1)$ parity check digits per codeword of n digits. In this paper, we shall show that it is sufficient to use two parity check digits per word of n digits, provided that these parity checks involve at least $\log_2 (n - 1)$ consecutive words. In sec. 2, we biiefly describe a coding method which was proposed by Th. Krol [1]), and which uses a q-ary single-error-correcting code, where q is any power of 2 which is greater than or equal to $n - 1$. In sec. 3, we propose the use of a special type of convolutional code for which this application seems to be new.

2. q-ary single-error-correcting block codes

For a given n, let ν be any integer greater than or equal to $\log_2 (n - 1)$, let GF(q) denote the Galois field with $q = 2^\nu$ elements, and let α be a primitive element in GF(q). Then, the q elements $0, 1, \alpha, \alpha^2, \alpha^3, \ldots, \alpha^{q-2}$ are all distinct elements of GF(q), and the ν elements $1, \alpha, \alpha^2, \ldots, \alpha^{\nu-1}$ form a basis of GF(q) viewed as a vector space over the binary field GF(2). Thus, any element of GF(q) can be uniquely expressed as a linear combination

$$a_0 + a_1\,\alpha + a_2\,\alpha^2 + \ldots + a_{\nu-1}\,\alpha^{\nu-1}$$

of these basis elements with binary coefficients a_i. This enables us to identify the 2^ν elements of GF(q) with the binary ν-tuples $(a_0, a_1, \ldots, a_{\nu-1})$ of their coefficients with respect to that basis. Thus, we may consider that ν consecutive binary digits do represent in a unique way an element of GF(q). Similarly, n-dimensional column vectors with components in GF(q) can be identified with $n \times \nu$ arrays of binary elements.

Now, let H be the $2 \times n$ matrix with elements in GF(q) defined by

$$H := \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 & 1 & 0 \\ 1 & \alpha & \alpha^2 & \ldots & \alpha^{n-3} & 0 & 1 \end{bmatrix}. \tag{1}$$

The set of n-dimensional vectors u,

$$u^{\mathrm{T}} := (u_0, u_1, u_2, \ldots, u_{n-3}, p, r), \tag{2}$$

satisfying $Hu = 0$, that is

$$p = \sum_{i=0}^{n-3} u_i, \quad r = \sum_{i=0}^{n-3} u_i\,\alpha^i, \tag{3}$$

over GF(q), is a linear code which is capable of correcting any single error, cf. ref. 2, p. 404, for example. As proposed by Krol [1]), this type of code can be used for our purpose as follows.

Let us consider ν consecutive digits to be transmitted along each of the n parallel channels. The n ν-tuples thus defined may be viewed as the components of an n-dimensional vector u over GF(q). Let the first $n-2$ components u_i, $i = 0, 1, \ldots, n-3$, be chosen arbitrarily and let the last two components be calculated and denoted as in (3). This defines the encoding process. Note that the eqs (3) remain linear when applied to the binary ν-tuples representing the field elements u_i, p and r. Let u_i, p' and r' denote the field elements corresponding to the binary ν-tuples obtained after transmission along the n channels, and let P and R be the field elements defined by

$$P = p' + \sum_{i=0}^{n-3} u'_i, \quad R = r' + \sum_{i=0}^{n-3} u'_i\,\alpha^i. \tag{4}$$

If no channel is in failure during transmission, we have $u'_i = u_i$, $p' = p$, and $r' = r$, whence $P = R = 0$, by (3).

If the channel carrying r, but no other, is in failure, we have $u'_i = u_i$, $p' = p$, and $r' = r + e$, say. Then we have by (3) and (4) $P = 0$, $R = e$.

Similarly, if the channel carrying p, but no other, is in failure, we have, with $e = p' + p$, $P = e$, $R = 0$.

Finally, let us assume that the channel carrying u_j, but no other, is in failure, and let $e = u'_j + u_j$. Then we have $P = e$, $R = e\alpha^j$. Clearly, the information to be transmitted is entirely contained in the set of u_i, $i = 0, 1, \ldots, n-3$, and all we have to do is to determine these from the received components u'_i, p' and r'. Assuming no more than one channel is in failure, we deduce from the above that

(i) if either P or R is equal to zero, we have $u'_i = u_i$ for $i = 0, 1, \ldots, n-3$, and no action has to be taken to recover the information;

(ii) if both P and R are nonzero, we have $P^{-1} R = \alpha^j$, $u_j = u'_j + P$, and $u_i = u'_i$ for $i \neq j$; thus, we may determine j by calculating $P^{-1} R$, and u_j by adding P to u'_j. Hence, in all cases, provided that no more than one channel is in failure during the transmission of ν consecutive digits, we can recover the information with a transmission rate of $(n-2)/n$. For more details concerning implementation of the method, we refer to ref. 2. Let us quote at this point that the above method requires calculating the check digits r for ν words (of n digits) at a time, thus requiring a buffer for encoding a continuous flow of information. In the method to be discussed in the next section, the two parity check digits to be attached to a word of n digits are calculated at the time they are to be transmitted; thus, no buffer is needed.

3. Convolutional encoding

3.1. *Convolutional encoder – General definitions*

An (n, k) binary convolutional encoder is a k-input n-output constant linear causal finite-state sequential machine, cf. ref. 3. With such a machine, k information digits are used to produce n binary digits at the output every unit of time, which is what we need for n parallel channels. It is therefore natural to investigate this type of encoder for our problem. Sequences of binary digits appearing at the input or output of such a machine are generally represented by means of polynomials in the delay operator D, like

$$x(D) = x_0 + x_1 D + x_2 D^2 + \ldots,$$

where x_i denotes the value occurring in the sequence at the ith unit of time. With this notation, let $x_j(D)$, $j = 1, 2, \ldots, k$, be the input sequences of an (n, k) convolutional encoder. Then, the output sequences $y_i(D)$, $i = 1, 2, \ldots, n$, can be expressed as

$$y_i(D) = \sum_{j=1}^{k} g_{ij}(D)\, x_j(D), \tag{5}$$

where the $g_{ij}(D)$ are polynomials of finite degree in the delay operator and characterize the encoder. The encoder is *systematic* if the first k outputs are identical to the inputs, that is

$$y_i(D) = x_i(D) \quad \text{for } i = 1, 2, \ldots, k.$$

For more details concerning the general properties of convolutional encoders and their realization, we refer to ref. 3.

3.2. *Single - channel error - correcting convolutional encoders*

3.2.1. Here we shall consider a special type of systematic $(n, n-2)$ encoder for which the last two outputs are given by

$$p(D) = \sum_{i=1}^{n-2} x_i(D), \tag{6}$$

and

$$r(D) = \sum_{i=1}^{n-2} g_i(D)\, x_i(D), \tag{7}$$

respectively, where we assume the $n-2$ polynomials $g_i(D)$ to be nonzero and distinct from one another. Let ν be the maximal degree of the polynomials $g_i(D)$, and let these polynomials be expressed as

$$g_i(D) = \sum_{j=0}^{\nu} g_{ij}\, D^j \quad \text{for } i = 1, 2, \ldots, n-2\ . \tag{8}$$

Then, such an encoder can be realized as indicated in fig. 1 by means of a ν-stage shift-register. The first $n-2$ outputs $y_1, y_2, \ldots, y_{n-2}$, are identical to the inputs; the $(n-1)$th output $y_{n-1} = p$ is given as the modulo-two sum of all inputs, hence only depends on the present inputs; the last output $y_n = r$ depends on the present and ν previous inputs and is realized by the ν-stage shift-register. The input at the jth stage of this shift register is given by the

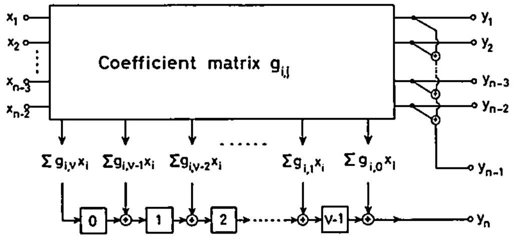


Fig. 1. An $(n, n-2)$ convolutional encoder.

modulo-two sum of the output of the previous stage and of the linear combination

$$\sum_{i=1}^{n-2} g_{i,\nu-j}\, x_i$$

of the inputs, for $j = 0, 1, 2, \ldots, \nu$. Since this contributes to the output with a delay of $\nu - j$ units of time, it is easily seen that the shift register will produce at its output the sequence $r(D)$ given by (7). Note that the assumption that the $n - 2$ polynomials $g_i(D)$ are distinct nonzero polynomials of degree ν or less forces us to choose ν sufficiently large, so that

$$2^{\nu+1} - 1 \geqslant n - 2$$

holds. This shows that the nth output of the encoder depends on at least $\log_2 (n - 1)$ successive inputs.

3.2.2. Let us now show how this type of coding method can be used for our purposes. We assume the n sequences produced at the output of the encoder are sent on n parallel binary channels and we denote by $y'_i(D)$, $i = 1, 2, \ldots, n$, the sequences obtained at the output of these n channels. We further assume at most one channel is in failure during transmission and we denote by $e(D)$ the error sequence that is added to the transmitted sequence on that channel. At the output, one calculates the *syndrome sequences*

$$P(D) = y'_{n-1}(D) + \sum_{i=1}^{n-2} y_i'(D), \tag{10}$$

$$R(D) = y'_n(D) + \sum_{i=1}^{n-2} g_i(D)\, y'_i(D), \tag{11}$$

by re-encoding the first $n - 2$ received sequences and adding the received check sequences $y'_{n-1} = p'$ and $y'_n = r'$. Assuming the jth channel, and no other, was in failure during transmission, we have

$$y'_j(D) = y_j(D) + e(D),$$

and

$$y'_i(D) = y_i(D) \qquad \text{for } i \neq j.$$

Hence, for $j = 1$, or 2, or $\ldots$, or $n - 2$, we have

$$P(D) = e(D), \quad R(D) = g_j(D)\, e(D), \tag{12}$$

and for $j = n - 1$ or n, we have

$$P(D) = e(D), \quad R(D) = 0, \tag{13}$$

or

$$P(D) = 0, \quad R(D) = e(D), \quad \text{respectively.} \tag{14}$$

Hence, if the polynomials $g_j(D)$ are all distinct and nonzero of degree ν, these cases can be distinguished from one another by observing the syndrome sequences during $\nu + 1$ consecutive units of time. Thus, it should be possible to decide on which channel the error sequence $e(D)$ was actually produced. This could be done, for example, by comparing $R(D)$ with each of the sequences

$$g_i(D)\,P(D), \quad i = 1, 2, \ldots, n - 2,$$

during $(\nu + 1)$ consecutive units of time, after having detected the presence of a failure in any one of the first $n - 1$ channels, which is indicated by the presence of a nonzero digit in the syndrome sequence $P(D)$. Of course, in the absence of error, we have $P(D) = R(D) = 0$. Once the erroneous channel has been identified, the information sequence can be recovered by adding the error sequence given by $P(D)$ to its output. This can be done with a delay of ν units of time. Note that, for any finite-length message, one should always add ν more check digits of the parity sequence $r(D)$. Hence, for a message of $M\,(n - 2)$ information digits, say, we have $2M + \nu$ parity digits. For M sufficiently large, the transmission rate is approximately $(n - 2)/n$.

Acknowledgements

The author is indebted to Th. Krol (Philips Res. Labs, Eindhoven) for communicating his ideas on this problem. Thanks are also due to P. Piret for some discussions on the subject of this paper.

MBLE Research Laboratory *Brussels, February 1978*

REFERENCES

[1] Th. Krol, private communication.
[2] D. C. Bossen, IBM J. Res. Develop. **14**, 402-408, 1970.
[3] G. D. Forney, IEEE Trans. Information Theory **IT-16**, 720-738, 1970.

BIREFRINGENCE IN SINGLE-MODE OPTICAL FIBRES DUE TO CORE ELLIPTICITY

by D. L. A. TJADEN

Abstract

The splitting-up of the degenerate HE_{11} fundamental mode of a step-index optical fibre, due to a slight elliptical deformation of the core cross-section, is considered. A first-order perturbation analysis is applied with respect to both the ellipticity parameter and the index contrast parameter. The results, which differ from those earlier given in the literature, show that the group delay difference between both modes equals zero at a value of the reduced frequency parameter $v \cong 2.478$.

1. Introduction

One possible cause of group dispersion in single-mode optical fibres is splitting-up of the degenerate HE_{11} ground mode due to deviations of the circular symmetry of the configuration. Particularly relevant in this respect are more-or-less elliptical deformations of the core cross-section.

Rigorous analysis of the modes of a dielectric step-index waveguide with elliptical core cross-section has been the subject of various papers [1-3], all using expansions in terms of Mathieu functions. Of these only Yeh [3]) gives actually numerical results, obtained by extremely involved computational efforts. These results, however, do not cover the parameter range of present interest (small ellipticity, small index difference between core and cladding).

In later papers by Schlosser [4,5]), Yeh [6]), and Dyott and Stern [7]) various approximative approaches to the problem were attempted. The obvious inconsistency of these results, which were sometimes obtained by quite involved calculations, induced the present author to carry out a new analysis, which is the subject of this paper.

Throughout the paper we use dimensionless Cartesian coordinates (x, y, z) reduced by the average core radius a. The z axis coincides with the fibre axis. In addition, we introduce polar coordinates in the (x, y) plane according to $x = r \cos \varphi$, $y = r \sin \varphi$.

The elliptical core boundary, with semi-axes $a(1 + q)$ and $a(1 - q)$ respectively, is represented as

$$r = r_c(\varphi) \equiv 1 + q \cos 2\varphi + O(q^2), \qquad q \to 0, \tag{1}$$

and we restrict our analysis to first-order terms in power series expansions with respect to q.

We will use a harmonic time factor $\exp(-\mathrm{i}\omega t)$ and consider waves propagating in the positive z direction with propagation constant β, thus giving rise to a z dependence of the electromagnetic field according to a factor $\exp(\mathrm{i}a\beta z)$. The magnetic permeability is μ_0 and the dielectric permittivities of the core and the cladding are denoted by ε_2 and ε_1 respectively. We put

$$\delta = (\varepsilon_2 - \varepsilon_1)/\varepsilon_1, \tag{2}$$

and, in accordance with the usual notation, introduce reduced parameters u, v and w by

$$\begin{aligned} v &= a\omega\,(\mu_0\,\varepsilon_1\,\delta)^{\frac{1}{2}}, \\ u &= a\,(\omega^2\,\mu_0\,\varepsilon_2 - \beta^2)^{\frac{1}{2}}, \\ \text{and} \qquad & \\ w &= a\,(\beta^2 - \omega^2\,\mu_0\,\varepsilon_1)^{\frac{1}{2}}, \end{aligned} \tag{3}$$

satisfying

$$u^2 + w^2 = v^2. \tag{4}$$

We introduce unit vectors **i**, **j** and **k** along the x, y and z axis, respectively. Furthermore we put

$$\nabla_t = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y}$$

and

$$E_t = \mathbf{i}\,E_x + \mathbf{j}\,E_y = e_t(x, y)\exp(\mathrm{i}a\beta z - \mathrm{i}\omega t).$$

Then e_t should satisfy the vector wave equation

$$\begin{aligned} \nabla_t^2\,e_t + u^2\,e_t &= \mathbf{0}, \qquad (r < r_c(\varphi)) \\ \nabla_t^2\,e_t - w^2\,e_t &= \mathbf{0}, \qquad (r > r_c(\varphi)) \end{aligned} \tag{5}$$

and vanish for $r \to \infty$, with the requirement of continuity at $r = r_c(\varphi)$ of $\mathbf{n}\times e_t$, $\mathbf{n}\cdot(\varepsilon e_t)$, $\nabla_t\times e_t$, and $(\nabla_t + \nabla_t\,\varepsilon/\varepsilon)\cdot e_t$. Here $\mathbf{n}$ is the unit normal to the core-cladding interface, whereas ∇_t^2 represents the two-dimensional vector Laplacian.

For all practical fibres $\delta \ll 1$ and it is well known [8]) that the solutions and corresponding eigenvalues are then well represented by those of the so-called scalar (or "weak-guidance") approximation according to which all Cartesian transverse field components are proportional to a scalar function ψ satisfying

$$\begin{aligned} \nabla_t^2\,\psi + u_0^2\,\psi &= 0, \qquad (r < r_c(\varphi)) \\ \nabla_t^2\,\psi - w_0^2\,\psi &= 0, \qquad (r > r_c(\varphi)) \end{aligned} \tag{6}$$

with

$$u_0^2 + w_0^2 = v^2. \tag{7}$$

At the core boundary $r = r_c(\varphi)$ both ψ and $\nabla_t \psi$ must be continuous. In the limit of this scalar approximation the field solutions are linearly polarized with an arbitrary polarization direction and as such this approximation is too crude for our purpose. As we have shown in a previous paper [9]) its solution, however, easily enables us to find the first-order term in a power series expansion

$$u = u_0 + \delta u_1 + \ldots \tag{8}$$

of the normalized propagation constant u. Our analysis will thus consist of two steps. First (in sec. 2) we will solve the scalar problem, correct to the first order in q. Next (in sec. 3), we will derive an expression for u_1 in (8), finally leading to a term of the order $q\delta$ in a double power series expansion of u with respect to q and δ.

In the appendix we will merely state the result of a rather involved calculation starting directly from (5) with the exact form of the boundary conditions. This result, valid for general values of δ, enables us to check our outcomes with those of Yeh [3]).

2. Scalar approximation

We attempt solutions of (6) in the form

$$\psi = \begin{cases} \sum_{k=0}^{\infty} q^k \, \alpha_k(u_0 \, r, \varphi), & (r < r_c) \\ \sum_{k=0}^{\infty} q^k \, \gamma_k(w_0 \, r, \varphi), & (r > r_c) \end{cases} \tag{9}$$

furthermore assuming u_0 and w_0 to depend on q according to

$$\begin{aligned} u_0 &= U + q u_{01} + \ldots, \\ w_0 &= W + q w_{01} + \ldots. \end{aligned} \tag{10}$$

By virtue of (8) we have

$$U^2 + W^2 = v^2, \quad \text{and} \quad U u_{01} + W w_{01} = 0. \tag{11}$$

The functions $\alpha_k(\varrho, \varphi)$ are solutions of

$$\frac{\partial^2 \alpha_k}{\partial \varrho^2} + \frac{1}{\varrho} \frac{\partial \alpha_k}{\partial \varrho} + \frac{1}{\varrho^2} \frac{\partial^2 \alpha_k}{\partial \varphi^2} + \alpha_k = 0. \tag{12}$$

As they are regular at $\varrho = 0$ they are linear combinations of terms $J_m(\varrho) \cos m\varphi$ and $J_m(\varrho) \sin m\varphi$ ($m = 0, 1, 2, \ldots$). Similarly the functions $\gamma_k(\varrho, \varphi)$ are linear combinations of terms $K_m(\varrho) \cos m\varphi$ and $K_m(\varrho) \sin m\varphi$.

According to the boundary conditions at $r = r_c(\varphi)$ we have

$$\sum_{k=0}^{\infty} q^k \left[\alpha_k \left(u_0\, r_c(\varphi), \varphi\right) - \gamma_k \left(w_0\, r_c(\varphi), \varphi\right)\right] = 0 \tag{13}$$

and

$$\sum_{k=0}^{\infty} q^k \left[u_0\, \alpha_k' \left(u_0\, r_c(\varphi), \varphi\right) - w_0\, \gamma_k' \left(w_0\, r_c(\varphi), \varphi\right)\right] = 0, \tag{14}$$

in which the primes denote differentiations with respect to ϱ. We now insert (1) and (10), expand the functions α_k and γ_k in Taylor series about respectively $\varrho = U$ and $\varrho = W$, and collect terms with equal powers of q. From the zero-order terms we have

$$\begin{aligned} &\alpha_0(U, \varphi) - \gamma_0(W, \varphi) = 0 \\ \text{and} \\ &U\alpha_0'(U, \varphi) - W\gamma_0'(W, \varphi) = 0, \end{aligned} \tag{15}$$

thus leading to the familiar solutions for the circular fibre. As we are interested in the behaviour of the fundamental mode we take the solutions

$$\begin{aligned} &\alpha_0(\varrho, \varphi) = B_0\, J_0(\varrho)/J_0(U) \\ \text{and} \\ &\gamma_0(\varrho, \varphi) = B_0\, K_0(\varrho)/K_0(W), \end{aligned} \tag{16}$$

in which the constant B_0 is to be determined later by a normalization condition and in which U and W are subject to (11) and to

$$\frac{J_0(U)}{UJ_1(U)} = \frac{K_0(W)}{WK_1(W)}. \tag{17}$$

Using (16), (17), and some familiar Bessel function identities we find from the first-order terms in the expansions of (13) and (14)

$$\begin{aligned} \alpha_1(U, \varphi) - \gamma_1(W, \varphi) &= B_0\, (u_{01}\, J_1/J_0 - w_{01}\, K_1/K_0), \\ U\alpha_1'(U, \varphi) - W\gamma_1'(W, \varphi) &= B_0\, v^2 \cos 2\varphi, \end{aligned} \tag{18}$$

in which we abbreviated $J_m(U) = J_m$, $K_m(W) = K_m$.

From these equations together with (11) and (17) it is not difficult to derive that

$$u_{01} = w_{01} = 0, \tag{19}$$

$$\alpha_1(\varrho, \varphi) = B_1 \frac{J_0(\varrho)}{J_0} + B_0 \frac{UWK_2\, J_2(\varrho) \cos 2\varphi}{2J_1\, K_1}, \tag{20}$$

and

$$\gamma_1(\varrho, \varphi) = B_1 \frac{K_0(\varrho)}{K_0} + B_0 \frac{UWJ_2\, K_2(\varrho) \cos 2\varphi}{2J_1\, K_1}, \tag{21}$$

where B_1 is a constant.
Substitution of (10), (16) and (19) to (21) in (9) gives

$$\psi(r, \varphi) = (B_0 + B_1\, q) \frac{J_0(Ur)}{J_0} + B_0\, q \frac{UWK_2\, J_2(Ur) \cos 2\varphi}{2J_1\, K_1} + O(q^2), \qquad (r < r_c(\varphi)) \tag{22}$$

$$\psi(r, \varphi) = (B_0 + B_1\, q) \frac{K_0(Wr)}{K_0} + B_0\, q \frac{UWJ_2\, K_2(Wr) \cos 2\varphi}{2J_1\, K_1} + O(q^2), \qquad (r > r_c(\varphi)) \tag{23}$$

whereas

$$u_0 = U + O(q^2), \qquad w_0 = W + O(q^2). \tag{24}$$

We will normalize $\psi(r, \varphi)$, thus fixing the values of the constants B_0 and B_1, by requiring

$$\int_0^{2\pi} \mathrm{d}\varphi \int_0^{\infty} \psi^2(r, \varphi)\, r\, \mathrm{d}r = 1 + O(q^2). \tag{25}$$

We find readily that $B_1 = 0$, together with

$$2\pi\, B_0{}^2 \left\{ \int_0^1 J_0{}^2(Ur)/J_0{}^2\, r\, \mathrm{d}r + \int_1^{\infty} K_0{}^2(Wr)/K_0{}^2\, r\, \mathrm{d}r \right\} = 1. \tag{26}$$

These integrals are easily evaluated and, requiring $B_0 > 0$, we find that

$$B_0 = \frac{1}{\sqrt{\pi}} \frac{UK_0}{vK_1}. \tag{27}$$

3. First-order term in δ

Adopting the notation introduced by Yeh [3]) we denote the modes in which the fundamental (HE_{11}) mode of a circular fiber is split due to an elliptical deformation as the ${}_eHE_{11}$-mode and the ${}_oHE_{11}$-mode. The subscripts e and o refer to the even and odd symmetries of the electric field with respect to the long axis.

If $q > 0$ the ellipse's long axis is along the x axis. Then, for the ${}_{\mathrm{e}}\mathrm{HE}_{11}$-mode, the coefficient u_1 in (8) is given by [9])

$$u_1 = -\frac{1}{4u_0}\iint \frac{\partial^2\psi^2}{\partial x^2}\,\mathrm{d}\sigma, \tag{28}$$

where the integration extends over the core region $r < r_{\mathrm{c}}(\varphi)$ in the (x, y) plane. With $\partial^2 \psi^2/\partial x^2 \equiv \Phi(r, \varphi)$ we have

$$u_1 = -\frac{1}{4U}\int_0^{2\pi} \mathrm{d}\varphi \left[\int_0^1 \Phi(r, \varphi)\, r\, \mathrm{d}r + \Phi\,(1, \varphi) q \cos 2\varphi\right] + O(q^2). \tag{29}$$

We write, for $r < r_{\mathrm{c}}(\varphi)$,

$$\psi^2 = P(r) + qQ(r)\cos 2\varphi + O(q^2), \tag{30}$$

in which, according to (22) and (27)

$$P(r) = \frac{1}{\pi}\,\frac{W^2}{v^2\,J_1{}^2}\,J_0{}^2(Ur)$$

and

$$Q(r) = \frac{1}{\pi}\,\frac{U\,W^3 J_0 K_2}{v^2\,J_1{}^3\,K_1}\,J_0(Ur)\,J_2(Ur). \tag{31}$$

Then

$$\Phi(r, \varphi) = \left(\cos\varphi\frac{\partial}{\partial r} - \frac{\sin\varphi}{r}\,\frac{\partial}{\partial\varphi}\right)^2 [P(r) + qQ(r)\cos 2\varphi] + O(q^2). \tag{32}$$

Substitution of (32) in (29) gives after some calculations

$$u_1 = -\frac{\pi}{8U}\{2P'(1) + q\,[P''(1) - P'(1) + Q'(1) + 2Q(1) - 2Q(0)]\} + O(q^2). \tag{33}$$

Finally, substitution of (31) in (33) gives, with the help of some Bessel function relations,

$$u_1 = \frac{UW^2}{4v^2}\{2F - q\,[1 + (U^2 - W^2)\,F^2 + U^2\,W^2\,F^3]\} + O(q^2), \tag{34}$$

where, in accordance with (17),

$$F = \frac{K_0}{WK_1} = \frac{J_0}{UJ_1}. \tag{35}$$

A similar treatment of the $_{\mathrm{o}}\mathrm{HE}_{11}$-mode would, at first sight, require evaluation of (28) with $\partial^2\psi^2/\partial x^2$ replaced by $\partial^2\psi^2/\partial y^2$. It is easily seen, however, that the correct result is simply found by replacing q by $-q$ in (34).

4. Phase and group retardation

From (3) we have for the propagation constant β

$$\beta = \frac{\omega n}{c}\left(1 - \frac{\delta}{1+\delta}\frac{u^2}{v^2}\right)^{\frac{1}{2}} \tag{36}$$

where $n = (\varepsilon_2/\varepsilon_0)^{\frac{1}{2}}$ is the core refraction index and $c = (\mu_0\,\varepsilon_0)^{-\frac{1}{2}}$. Expansion in powers of δ gives with (8)

$$\beta = \frac{\omega n}{c}\left[1 - \delta\frac{{u_0}^2}{2v^2} + \delta^2\left(\frac{{u_0}^2}{2v^2} - \frac{{u_0}^4}{8v^4} - \frac{u_0 u_1}{v^2}\right) + O(\delta^3)\right]. \tag{37}$$

Denoting the propagation constants for the $_{\mathrm{e}}\mathrm{HE}_{11}$- and $_{\mathrm{o}}\mathrm{HE}_{11}$-modes by β_{e} and β_{o} respectively, we find from (34) for the leading term in an expansion of $\beta_{\mathrm{e}} - \beta_{\mathrm{o}}$

$$\Delta\beta \equiv \beta_{\mathrm{e}} - \beta_{\mathrm{o}} \cong \frac{\omega n}{c}\,\delta^2\,q\,G_{\mathrm{p}}(v), \tag{38}$$

where

$$G_{\mathrm{p}}(v) = \frac{U^2\,W^2}{2v^4}\,[1 + (U^2 - W^2)\,F^2 + U^2\,W^2\,F^3]. \tag{39}$$

The group delay difference per unit length is found as

$$\frac{\mathrm{d}\Delta\beta}{\mathrm{d}\omega} = \frac{n}{c}\,\delta^2\,q\,G(v), \tag{40}$$

where

$$G(v) = \frac{\mathrm{d}}{\mathrm{d}v}\,[vG_{\mathrm{p}}(v)]. \tag{41}$$

From (11) and (35) it may be derived that

$$\begin{gathered}
\frac{\mathrm{d}U}{\mathrm{d}v} = \frac{U}{v}(1 - W^2\,F^2), \qquad \frac{\mathrm{d}W}{\mathrm{d}v} = \frac{W}{v}(1 + U^2\,F^2) \\
\text{and} \\
\frac{\mathrm{d}F}{\mathrm{d}v} = -\frac{1}{v}(1 - W^2\,F^2)\,(1 + U^2\,F^2).
\end{gathered} \tag{42}$$

With the help of these relations it is found that

$$G(v) = \frac{U^2 W^2}{2v^4}\{1 - 2(U^2 - W^2) F + (5U^2 - 5W^2 - 3U^2 W^2) F^2$$
$$+ (13U^2 W^2 - 2v^4) F^3 + [2v^4 - 12U^2 W^2 - 3U^2 W^2 (U^2 - W^2)] F^4$$
$$+ 6U^2 W^2 (U^2 - W^2) F^5 + 3U^4 W^4 F^6\}. \tag{43}$$

Both functions $G_p(v)$ and $G(v)$ are shown in fig. 1. For $v < 0.5$ they are practically zero. In the region of interest, about $v = 2$, $G_p(v) \cong 0.22$ whereas $G(v)$ decreases and crosses zero at $v \cong 2.478$, which is just outside the region of single-mode operation ($v < 2.4$).

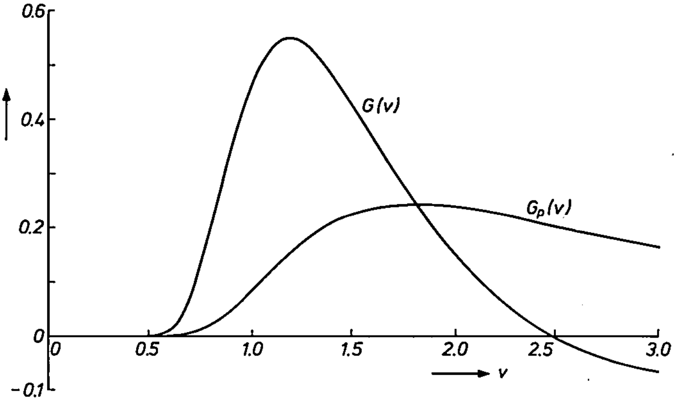


Fig. 1. Normalized phase (G_p) and group (G) delay vs normalized frequency parameter v.

5. Final remarks

Our result (38) has the same form as that obtained by Schlosser [5]), but for $G_p(v)$ he finds (in our notation)

$$G_p(v) = \frac{U^4 W^2 (1 + W^2 F)}{2v^6 J_1{}^2}. \tag{44}$$

His earlier calculation [4]), upon which this result is based, follows rather different lines from ours and involves certain approximations the ultimate effect of which can not easily be estimated. In a very recent paper by Snyder and Young [10]) a result is given which in our notation would imply that

$$G_p(v) = \frac{U^2 W^2}{2v^4}(1 + 2W^2 F^2 - U^2 W^2 F^3). \tag{45}$$

Here the approach seems to be rather similar to ours but unfortunately no details of the calculation are given.

Assuming $n = 1.5$, $\delta = 0.006$, and $q = 0.1$ (i.e. an axial ratio of the core cross-section of 1.22) we find for $v = 2$ a group delay difference of 2.9 ps/km. At $v = 2.3$ we find a value of 0.8 ps/km only, whereas both refs 5 and 10 would predict a value of 4.7 ps/km at $v = 2.3$.

Our expressions hold for all HE_{1n}-modes, provided that the proper roots of (17) are substituted. Similar expressions could be obtained for the EH_{1n}-modes. A rather different type of analysis is required for the TE_n, TM_n and HE_{2n} mode group of elliptically deformed fibres [10,11]).

Philips Research Laboratories *Eindhoven, October 1978*

Appendix

A perturbation analysis with respect to q, similar to that of sec. 2, can be applied directly to the vector wave equation (5). A characteristic equation is thus obtained for the HE_{1n}-modes and the EH_{1n}-modes, correct to the first order in q. If we put

$$\eta_1 = \frac{J_0(u)}{uJ_1(u)} - \frac{1}{u^2} \quad \text{and} \quad \eta_2 = -\frac{K_0(w)}{wK_1(w)} - \frac{1}{w^2}, \tag{46}$$

we find that

$$(\eta_1 + \eta_2)\,[(1+\delta)\eta_1 + \eta_2] - \left(\frac{1}{u^2} + \frac{1}{w^2}\right)\left(\frac{1+\delta}{u^2} + \frac{1}{w^2}\right)$$
$$\pm \tfrac{1}{2}q\delta\left[\eta_1 + \eta_2 - \left(\frac{\eta_1}{w^2} - \frac{\eta_2}{u^2}\right)(1 - u^2\,w^2\,\eta_1\,\eta_2)\right] = 0, \tag{47}$$

where the upper and lower signs refer to the even and odd modes respectively. As it should, eq. (34) follows again from this result for $\delta \to 0$.

Equation (47) enables us to make a comparison with the results obtained by Yeh [3]) which are for $\delta = 1.5$. For this purpose we choose his figures 2 and 6 which show, for the ${}_e\mathrm{HE}_{11}$-mode and the ${}_o\mathrm{HE}_{11}$-mode respectively, the relationship between $(1 + q)\,u$ and $(1 + q)\,w$, for various values of the ellipticity. In Yeh's paper the latter is expressed by a parameter $\xi_0 = -\frac{1}{2}\ln q$. The smallest nonzero value of q considered is $q \cong 0.1353$, corresponding to $\xi_0 = 1$. In figures 2 and 3 a few of Yeh's curves are reproduced, together with the corresponding curves obtained by numerical solution of (47). For $q = 0.1353$ the agreement is still reasonable.

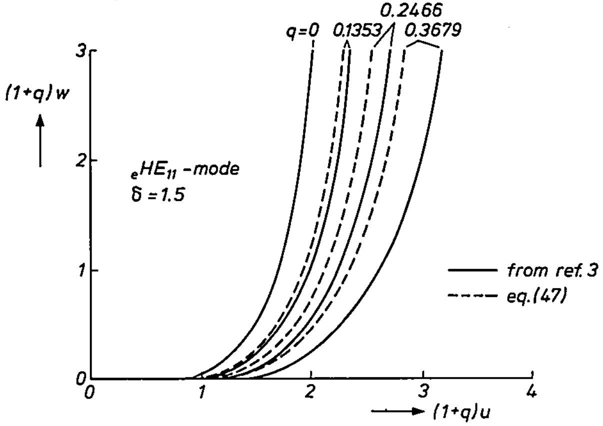


Fig. 2. Comparison with Yeh's results [3]) for $_e$HE$_{11}$-mode.

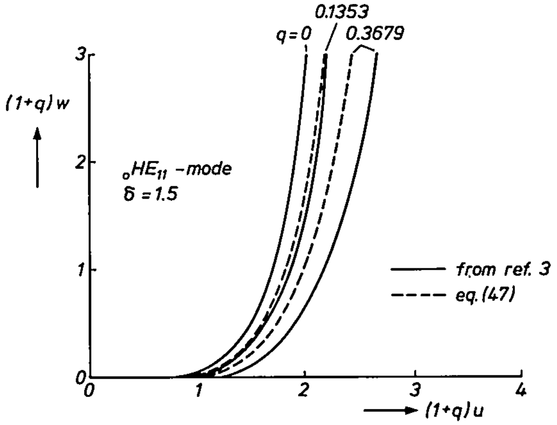


Fig. 3. Comparison with Yeh's results [3]) for $_o$HE$_{11}$-mode.

REFERENCES

[1]) L. A. Lyubimov, G. I. Veselov and N. A. Bei, Radio Engng and Electr. Phys. **6**, 1668-1677, 1961.
[2]) G. Piefke, A.E.Ü. **18**, 4-8, 1964.
[3]) C. Yeh, J. Appl. Phys. **33**, 3235-3243, 1962.
[4]) W. Schlosser, A.E.Ü. **19**, 1-8, 1965.
[5]) W. O. Schlosser, Bell Syst. Tech. J. **51**, 487-492, 1972.
[6]) C. Yeh, Opt. and Quantum Electronics **8**, 43-47, 1976.
[7]) R. B. Dyott and J. R. Stern, Electronics Letters **7**, 82-84, 1971.
[8]) A. W. Snyder, IEEE Trans. Microwave Theory Tech. **MTT-17**, 1130-1138, 1969.
[9]) D. L. A. Tjaden, Philips J. Res. **33**, 103-112, 1978.
[10]) A. W. Snyder and W. R. Young, J. Opt. Soc. Am. **68**, 297-309, 1978.
[11]) D. L. A. Tjaden, Philips J. Res., to be published.

Philips J. Res. 33, 264-280, 1978 *R 993*

THE COPRECIPITATION OF Cu, Ag, Au; Ge, Sn, Pb; Mn AND Ca WITH FERRIC HYDROXIDE

by E. BRUNINX

Abstract

A procedure has been established in order to distinguish between real coprecipitation — by means of ion exchange in the precipitate — and purely mechanical collection phenomena. The simple ion-exchange reaction model is replaced by a model taking into account the hydrolysis of the elements. In the appendix a survey is given of coprecipitation data for other elements.

1. Introduction

In a previous publication [1]) data were presented on the coprecipitation behaviour of zinc, cadmium and mercury, together with a brief survey of the more important mechanisms of coprecipitation. From these results it appeared that coprecipitation phenomena could not be described in an unambiguous way by means of adsorption isotherms (Langmuir, Freundlich). It was also shown that ion exchange reactions are quite evident in the coprecipitation of elements with ferric hydroxide, and in the present paper further evidence is presented in support of this mechanism. From the experimental data it is also possible to make a distinction between "ion exchange-coprecipitation" and "occlusion-coprecipitation", i.e. chemical reactions, such as precipitation followed by collection of the precipitate in the ferric-hydroxide phase. As in the previous work, the emphasis was laid upon the measurement of a fixed number of variables for as many elements as possible, rather than an exhaustive study of a single element.

2. Method of precipitation

The technique used in all experiments was the same as described in the earlier paper [1]). We shall only recall that the distribution coefficient D is defined as follows.

$$D = \frac{Me_{\text{ppt}}}{Me_{\text{soln}}},$$

where Me_{ppt} is the amount of metal in the precipitate and

Me_{soln} is the amount of metal in the whole solution after precipitation.

All data are represented graphically as $\log_{10} D$. Where necessary the standard deviation of $\log_{10} D$ (due to counting statistics) has been indicated in the graphs. For all other values of $\log_{10} D$ values without error bars the standard deviation due to the counting statistics is less than 0.005 log units.

Transformation of D into D_v, taking into account the mass of the ferric hydroxide and the volume of the solution, is as follows.

$$D_v = \frac{\text{amount of element/g Fe}}{\text{amount of element/ml solution}} = D \frac{\text{volume solution}}{\text{weight Fe (g)}}.$$

The volume of the solution after precipitation was 42.5–43 ml in nearly all experiments and the amount of iron was 4×10^{-4} g. It follows that

$$D_v = D \frac{43}{4 \times 10^{-4}} = 1.08\, D \times 10^5$$

or on the $\log_{10}$ scale

$$\log_{10} D_v = \log_{10} D + 5.03.$$

3. Cu, Ag, Au

3.1. *General*

All the earlier work done on copper coprecipitation with ferric hydroxide has been connected with the search for optimal conditions in gravimetric analysis [2,3,4]). In most experiments rather high copper and ammonia concentrations were used. The influence of varying concentrations upon the coprecipitation of copper was examined by Novikov et al.[5]).

Silver coprecipitation was studied at very low concentrations by Dyck [6]) in the absence of any ion that might form precipitates with silver. The pH varied between 4 and 8 and the silver was added after the formation of the ferric hydroxide. More limited studies were made by Upor [7]) and Kepak [8]). Only one publication briefly mentioned the coprecipitation of trace quantities of gold [9]).

3.2. *Cu*

The measured values of $\log_{10} D$ as a function of pH, total copper concentration, iron concentration and volume are plotted in figs 1, 2, 3 and 4. As in the case of the earlier studied elements (zinc, cadmium, mercury) the coprecipitation yield rises with increasing pH (see fig. 1) according eq. (1).

$$\log_{10} D = n\,\text{pH} + \log_{10} K + n \log_{10} [\text{FeH}]. \qquad (1)$$

In this equation FeH stands for the precipitated ferric hydroxide; K is the equilibrium constant for the reaction between metal ion and ferric hydroxide; n is the number of hydrogen atoms exchanged. See also ref. 1.

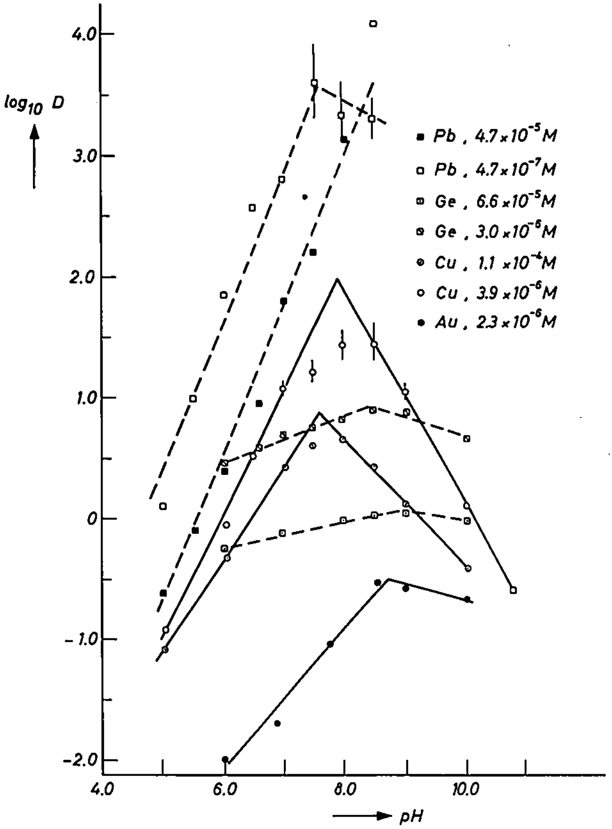


Fig. 1. Log_{10} D for Pb, Ge, Cu and Au in KNO_3 (10^{-2} M) as a function of pH; $[\text{Fe}]_T = 1.73 \times 10^{-4}$ M. The numbers following the chemical symbols indicate the concentration before precipitation.

Figure 5 gives a plot of the percentage distribution of the different copper species present at varying pH. The percentage distribution of the different species present in solution was computed by means of a computer program similar to the one described in ref. 10. The values of the overall equilibrium constants β were taken from refs 11, 12 and 13. The hydroxy complexes can be disregarded since their contribution is less than 1% at the most. In the region $7.5 < \text{pH} < 8.5$ only copper-ammine and copper–hydroxy–ammine complexes are present and no free copper ions.

It is not likely that the copper ammine complex will split off ammine groups in order to be bound to the ferric hydroxide. We have observed earlier that zinc coprecipitates just as well with NaOH as with NH_4OH. We therefore believe

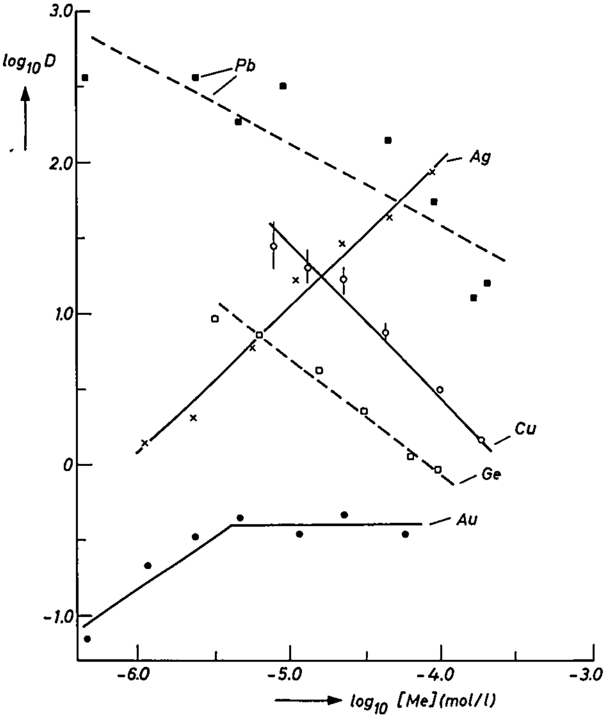


Fig. 2. $\log_{10} D$ for Pb, Ge, Cu, Au and Ag in KNO_3 (10^{-2} M) at pH = 8.5 and varying metal concentration (before precipitation); $[Fe]_T = 1.73 \times 10^{-4}$ M.

that the copper is bound to the ferric hydroxide by means of the OH groups (or H_2O molecules) attached to the copper complex. When the copper is completely coordinated with NH_3 groups the coprecipitation yield drops. A similar trend was observed for zinc and cadmium.

3.3. *Ag*

The coprecipitation of silver as a function of pH shows two tendencies (no results for silver are shown in figs 1, 3 and 4).

(a) With NH_4OH as reagent, the coprecipitation yield decreases continually with increasing pH. This is due to the formation of silver–ammine complexes.

(b) If KOH or NaOH is employed, the coprecipitation yield appears to be more or less independent of the pH: chloride and (or) oxide precipitation will in all likelyhood play the most important role.

Increasing the amount of silver and keeping the iron concentration constant (see fig. 2) results in an increase of the coprecipitation yield, and not in a

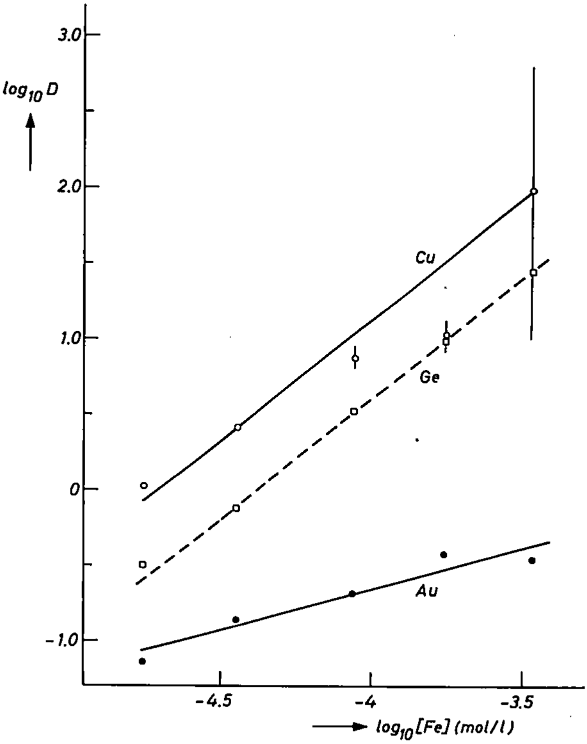


Fig. 3. $\text{Log}_{10}\, D$ for Cu, Ge and Au in KNO_3 (10^{-2} M) for varying initial iron concentration. The initial metal concentrations are (M): $[\text{Cu}]_T = 3.9 \times 10^{-6}$, $[\text{Ge}]_T = 3 \times 10^{-6}$, $[\text{Au}]_T = 2.3 \times 10^{-6}$.

decrease as was observed for e.g. copper (see fig. 2) or zinc, cadmium and mercury, see ref. 1.

This can be explained qualitatively as follows. Consider a chemical reaction

$$\text{A} + \text{B} \rightarrow \text{AB}$$

with $[\text{A}]_{\text{tot}} = [\text{A}] + [\text{AB}] = C_0$ (originally present). To this we add B as reagent with a concentration xC_0 or

$$[\text{B}]_{\text{tot}} = [\text{B}] + [\text{AB}] = xC_0.$$

With increasing B increasing quantities of AB will be formed and according to the magnitude of the equilibrium constant K the curve will have a shape as shown in fig. 6. Detailed calculations can be found in ref. 14. In these experiments it was also observed that the amount of silver in the aqueous phase (after coprecipitation) remained constant. This corresponds to the equilibrium

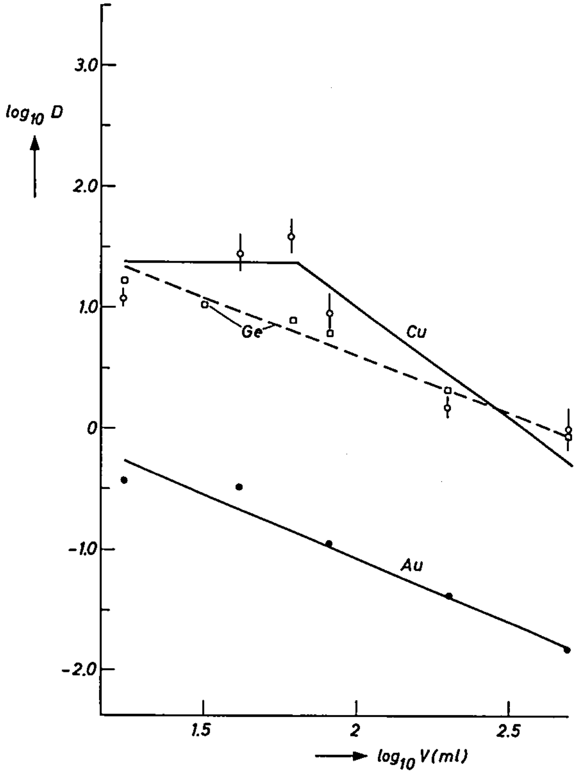


Fig. 4. Log_{10} D for Cu, Ge and Au in KNO_3 (10^{-2} M) at pH 8.5 and varying initial volume. The amount of metal added is Cu = 0.32 μmol, Ge = 0.13 μmol, Au = 0.1 μmol, Fe = 7.16 μmol.

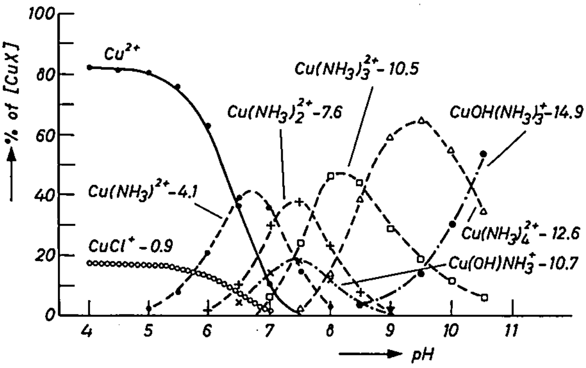


Fig. 5. Distribution of different species — CuX — at $[Cu]_T = 7.8 \times 10^{-6}$, $[Cl]_T = 0.0269$, $[NH_3]_T = 0.042$. The numbers following the chemical symbol are $\log \beta$ for the reaction in question; β is the overall equilibrium constant.

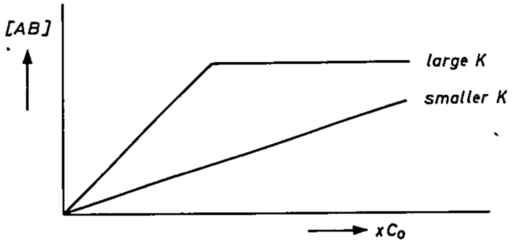


Fig. 6. Schematic variation of concentration of AB in the reaction A + B → AB for a small and a large value of the equilibrium constant.

between a precipitate and its saturated solution whereby — in this particular case — the silver precipitate is collected in the ferric-hydroxide phase. To sum up, real coprecipitation will only occur in the absence of chloride and ammonium ions and at very low silver concentrations as was realized by Dyck [6]).

3.4. *Au*

The distribution ratio *D* for gold is low under all circumstances: for analytical purposes coprecipitation is thus a useless method.

The general pattern in figs 1, 3 and 4 presents no marked departure from the usual trend we observed with zinc, cadmium and mercury.

Increasing the gold concentrations (fig. 2) reveals the same behaviour as for silver. Gold forms mainly complexes with Cl ($AuCl_4^-$, $\beta = 26$–30) and NH_3 ($Au(NH_3)_4^-$, $\beta = 30$). Computation of the concentration of the different species shows that Au^{3+} is very small at all pH values. The concentration of the chloride or ammine complexes is difficult to calculate due to the uncertainty of the β-$AuCl_4^-$ value. For β-$AuCl_4^- = 30$, $AuCl_4$ is the dominant species between pH 5 and 9; if β-$AuCl_4^- = 26$ then $Au(NH_3)_4^-$ is the dominating species: precipitation of $Au(OH)_3$ cannot then be ruled out entirely. This may perhaps explain the shape of the gold data in fig. 2.

4. Ge, Sn, Pb

4.1. *General*

Most of the earlier work on germanium indicates that the coprecipitation is nearly complete in the pH range between 7–9 (see refs 15, 16, 17 and 18). In a few of these studied it cannot be excluded that iron–germanium compounds were formed, due to the high concentrations employed.

No published data on tin coprecipitation could be found. Koch [19]), Mizuike [20]) and West [21]) simply mention, without any further specification, that coprecipitation occurs.

Some data on the coprecipitation of lead — mainly for analytical purposes — can be found in refs 22, 23 and 24.

4.2. *Ge*

Results are shown in figs 1, 2, 3 and 4. As can be seen from fig. 1, the value of n (eq. (1)) is much less than for all other elements. In the pH range 6–11 only two germanium compounds are of interest: $Ge(OH)_4$ and $GeO(OH)_3^-$. Formation of cationic species such as $Ge(OH)_y^{(4-y)+}$ can be excluded under the actual conditions. There is also some evidence for formation of iron germanates such as $Fe(Ge(OH)_6)$ (see ref. 25).

When we assume that coprecipitation with iron hydroxide occurs mainly by means of an ion exchange mechanism, then it follows that for a given amount of iron, a limited number of sites is available for coprecipitation to occur. Thus if two elements compete for a limited number of sites, then the more strongly bound (coprecipitated) element will preferentially occupy the available sites and decrease the coprecipitation yield of the other element. Table I shows the distribution ratio of germanium with increasing quantities of lead (strongly coprecipitated, see fig. 1).

TABLE I

Distribution ratio of Ge, at two concentrations, for increasing amounts of Pb

μg Pb added	D_{Ge} for 18.8 μg Ge	D_{Ge} for 187.5 μg Ge
0	7.6	1.0
20	8.1	1.1
40	6.6	1.2
200	15.2	1.6
400	17.4	2.2
800	29.7	3.5

We notice that the coprecipitation yield of germanium increases rather than decreases. This phenomenon cannot be explained by an ion exchange mechanism and the formation of chemical compounds such as $3PbGeO_3.2H_2O$, as reported by Pugh [26]), followed by occlusion in the ferric hydroxide, cannot be excluded.

4.3. *Sn*

Coprecipitation of tin at varying pH yielded practically constant distribution ratios. Since tin is uniquely present as Sn^{4+} and because the solubility product of its hydroxide is extremely small (10^{-56}), no real coprecipitation will occur, but rather occlusion. The occlusion is complete and thus offers a possibility for analytical applications.

4.4. *Pb*

Lead is the element with the highest coprecipitation yield known. The measurement of very high distribution ratios is somewhat uncertain due to the very low count rates in the aqueous phase. For this reason the dependence of D as a function of the iron concentration and the volume were not measured.

Lead forms only hydroxy complexes; the formation of $PbCl_2$ is not possible under the actual conditions and the formation of solid $Pb(OH)_2$ can also be excluded.

Figure 7 shows the calculated distribution of the predominant species at different pH values. Species such as $Pb(OH)_3{}^-$, $Pb_3(OH)_4{}^{2+}$ were taken into consideration but as their concentration becomes only important above pH = 10 they have been omitted in fig. 7.

The calculated percentage distribution is quite dependent upon the choice of the equilibrium constant and, for a given set of β values, also upon the total lead concentration.

Comparing figs 1 and 7 it is evident that the increase in coprecipitation yield coincides again with the rise of the $PbOH^+$ concentration.

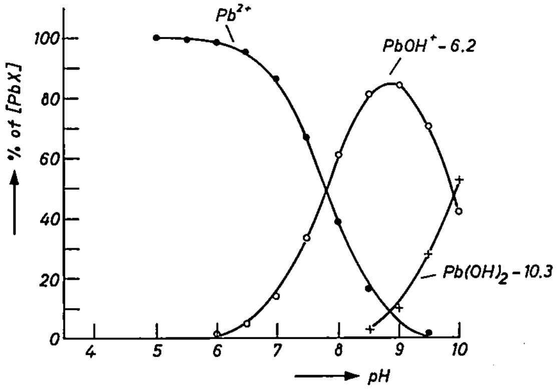


Fig. 7. Distribution of different species — PbX — at $[Pb]_T = 4.7 \times 10^{-7}$. The numbers following the chemical symbols are log β for the reaction in question; β is the overall equilibrium constant.

5. Mn, Ca

5.1. *General*

A fairly extensive study of the manganese coprecipitation with ferric hydroxide was made by Azzam [27]). Somewhat unfortunately, he used rather high manganese concentrations, so that many coprecipitation effects were obscured by precipitation reactions. Additional results can be found in refs 4, 28 and 29. No data on calcium could be found.

5.2. *Mn*

The results are summarized in figs 8, 9, 10 and 11 where the same trend is observed as for all the other coprecipitating elements. However, the increase of D with rising pH starts at a much higher value (pH = 7) than with most other elements. This behaviour correlates again very well with the onset of the $Mn_2(OH)_3^+$ and $MnOH^+$ concentrations, as can be seen in fig. 12.

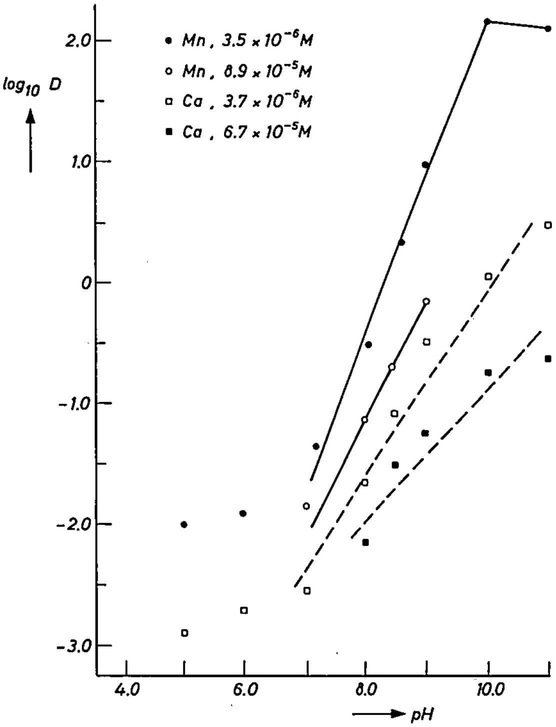


Fig. 8. $\log_{10} D$ for Mn and Ca in KNO_3 (10^{-2} M) as a function of pH; $[Fe]_T = 1.73 \times 10^{-4}$ M. The numbers following the chemical symbols indicate the concentration before precipitation.

As for all other coprecipitating elements the distribution ratio increases with larger iron concentrations both at low and high manganese concentration. The two deviating points in fig. 10 at high manganese concentration are due to the number of sites available being too small.

The dependence of D upon volume is not very clearly marked owing to the scatter of the points: no cause could be found for this large scatter. The results indicate a trend comparable to that of other elements.

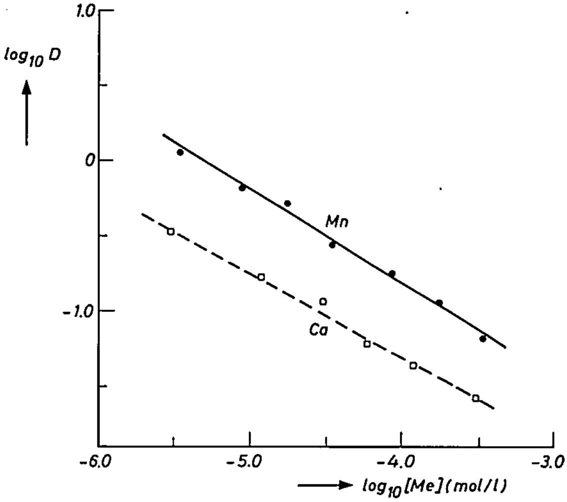


Fig. 9. $\text{Log}_{10}\, D$ for Mn and Ca in KNO_3 (10^{-2} M) at pH = 8.5 and varying metal concentration (before precipitation); $[Fe]_T = 1.73 \times 10^{-4}$ M.

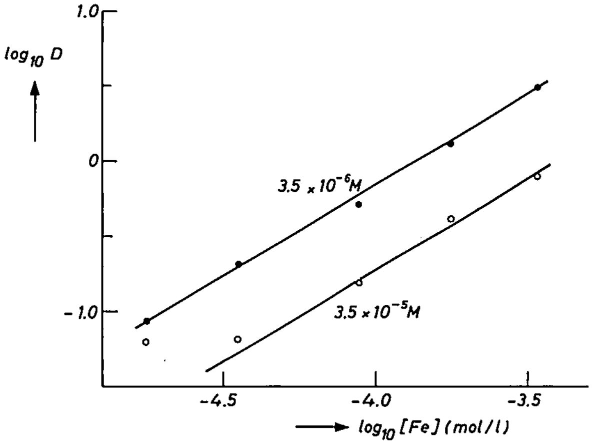


Fig. 10. $\text{Log}_{10}\, D$ for Mn in KNO_3 (10^{-2} M) for varying initial iron concentration at pH = 8.5. The initial metal concentrations (M) have been indicated.

Increasing the amount of manganese in the original solution decreases the value of D, as for all other coprecipitating elements. The results of Azzam et al.[27]) indicate precisely the opposite trend. This is due to their high manganese concentration. Under these circumstances precipitation (and occlusion) of manganese occurs and the same explanation as in sec. 3.3 can be given.

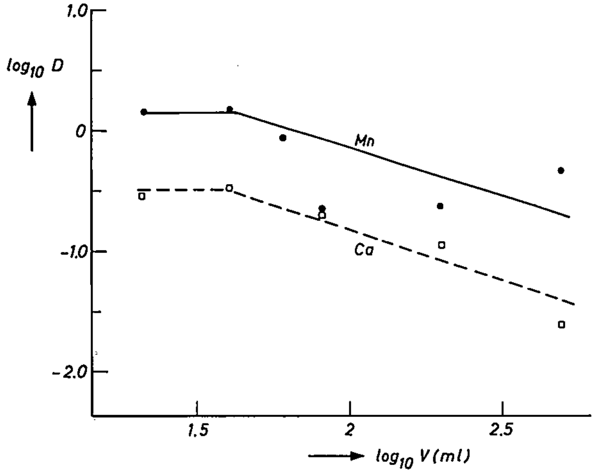


Fig. 11. Log_{10} D for Mn and Ca in KNO_3 (10^{-2} M) at pH = 8.5 and varying initial volume. The amount of metal added is Mn = 0.15 μmol, Ca = 0.13 μmol, Fe = 7.16 μmol.

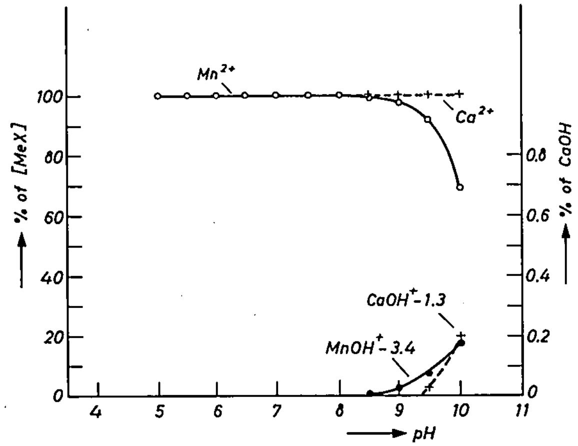


Fig. 12. Distribution of different species — MeX — at $[Mn]_T = 3.5 \times 10^{-6}$ M and $[Ca]_T = 3.1 \times 10^{-6}$ M. The numbers following the chemical symbols are $\log \beta$ for the reaction in question; β is the overall equilibrium constant. Note the different scale for $CaOH^+$.

Finally, manganese was coprecipitated in the presence of increasing quantities of lead. The results in table II show a decrease of D at the higher lead concentration. Here we have a pure coprecipitation mechanism whereby the more strongly bound lead displaces the more weakly bound manganese from the available number of sites, while in the case of Ge/Pb (table I) a chemical compound was formed.

TABLE II

Distribution ratio of Mn, at two concentrations, for increasing amounts of Pb

μg Pb added	D for 20 μg Mn	D for 200 μg Mn
0	0.6	0.1
20	0.5	0.1
40	0.6	0.1
200	0.3	0.06
400	0.2	0.07
800	0.1	0.03

5.3. *Ca*

The coprecipitation of calcium was examined at varying pH, varying calcium concentration and varying volume. The results are shown in figs 8, 9 and 11. The values of D are all extremely low and the increase starts off at a much higher pH value than for all other elements. This can be correlated with the formation of the hydroxy complex — $CaOH^+$ — with a rather small equilibrium constant. The concentration of this complex only becomes noticeable above pH = 9.5.

6. Conclusions

(1) Amongst all the elements studied so far silver, gold and tin do not coprecipitate via an ion exchange mechanism as given by eq. (1); these elements are precipitated and collected in the ferric-hydroxide phase. This can best be observed by studying the variation of D as a function of the total metal concentration. For silver only coprecipitation via ion exchange is possible at lower metal concentrations and in the absence of any precipitating reagent.

(2) For copper, manganese and calcium the distribution ratio remains constant up to a certain volume (in which coprecipitation takes place). Thereafter the distribution ratio drops. Although the imprecision of the data is fairly large,

the crossover point appears to be the same as observed earlier for zinc, cadmium and mercury [1]. The reason for this behaviour still is not clear.

(3) For the elements copper, calcium, manganese, lead and germanium and those in the previous work (zinc, cadmium) it appeared that the rise of the coprecipitation yield coincided with the disappearance of the free metal ion. Simultaneously the formation of hydroxy complexes became more important.

In figure 13 we have tried to correlate the value of $\log_{10} D$ with the percentage of hydroxy complexes present at pH = 8.5. The concentrations of all elements were roughly comparable. The value for the germanium hydroxy complex was taken from ref. 30. For zinc, copper and cadmium it is known that these elements form a series of complexes with one, two etc. NH_3 groups. As long as these elements are not completely coordinated by the NH_3 groups, the remaining space is taken up by water molecules. In the construction of fig. 13 we therefore took the sum of all three complexes $Me(NH_3)_{1,2,3}$ as a measure of the species that could be bound to the ferric hydroxide.

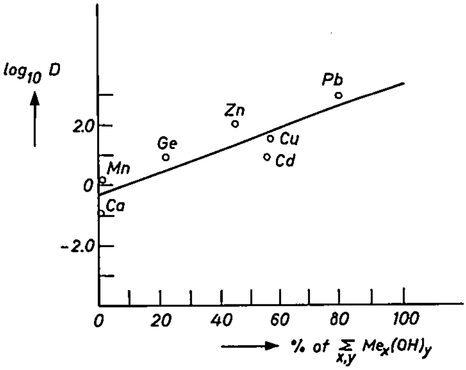


Fig. 13. Correlation between the distribution ratio and the percentage of an element present in any OH⁻ containing complex.

The graph demonstrates that with these somewhat crude assumptions, there is a fairly good correlation between the observed distribution ratio and the percentage of an element in its hydroxy complex (correlation coefficient 0.9).

It is therefore appropriate to replace eq. (1) by either of the following reactions (FeOOH stands for the ferric hydroxide):

(a) formation of surface hydroxy complex

$$FeOOH + Me^{2+} \rightarrow FeOOH.Me^{2+}$$
$$FeOOH.Me^{2+} \rightarrow FeOOH.MeOH^{+} + H^{+}$$

(b) hydrolysis followed by adsorption,

$$Me^{2+} + H_2O \rightarrow MeOH^{+} + H^{+}$$
$$FeOOH + MeOH \rightarrow FeOOH.MeOH^{+}.$$

Both models are equivalent and have been proposed by James et al.[31]). A more detailed discussion of these two mechanisms can be found in their paper.

Philips Research Laboratories *Eindhoven, October 1978*

Appendix

In this survey we have summarized the available data scattered in various publications on the coprecipitation with ferric hydroxyde. They are classified in two large groups and indicated with the letter M or P; M stands for articles whereby the emphasis is lead upon the study of coprecipitation mechanism; P indicates work where only a few variables such as pH etc. were varied. During recent years many Russian articles have been published which were not available to us. The number in the table indicates the reference.

Ag:	7P; 6M; 8P;
As:	32P; 33M; 34P; 35P;
Au:	9P;
Ba:	36P; 37P; 38M; 39M; 40M;
Cd:	41M; 22P;
Ca:	42P;
Co:	7P; 43M; 44M; 39M; 28P; 45P; 46P;
Cr:	47P;
Cu:	2P; 4P; 5M;
Eu:	28P;
Ga:	48P; 49P;
Ge:	49P; 17M; 16M; 18M;
Hg:	50M; 51M;
I:	52P; 46P; 53P;
Ir:	9P;
La:	46P;
Mn:	28P; 27P; 4P;
Mo:	47P; 54P; 55P;
Nb:	28P; 56P;
Ni:	4P;
P:	57M; 58M;
Pb:	22P; 23P;
Pa:	28P;
Pd:	9P; 59P;
Pm:	42P;
Pt:	9P;
Rh:	59P; 9P;
Ru:	28P; 52P; 59P;
S:	58M;
Sb:	34P;
Se:	60P; 49P;
Sr:	38M; 61M; 62M; 63P; 64P;
Th:	65M;
Tl:	66P; 22P;
U:	35M;
V:	47P; 60P;
W:	47P;
Y:	61M; 28P; 68M; 56P;
Zn:	7P; 22P; 2P; 4P; 49P; 69P;
Zr:	56P; 28P.

REFERENCES

[1] E. Bruninx, Philips Res. Repts **30**, 177-191, 1975.

[2] T. Tarantelli, G. Piantoni and G. Saini, Atti. Accad. Sci. Tor., Class. Sci. Fis. Mat-Nat. **92**, 576-589, 1957.

[3] For a summary see: I. Kolthoff, E. Sandell, E. Meehan and S. Bruckenstein, Quantitative chemical analysis, 4e ed., McMillan Cy, London, 1969.

[4] S. Rubel and M. Lugowska, Chem. Anal. **20**, 293-302, 1975.

[5] A. Novikov, A. Shaffert and E. Shckekaturova, J. Anal. Chem. USSR **32**, 872, 1977. (English translation.)

[6] W. Dyck, Can. J. Chem. **46**, 1441-1444, 1968.

[7] E. Upor, A. Ronai and M. Görbicz, Acad. Scient. Hungar. **61**, 1-11, 1969.

[8] F. Kepak and I. Nova, Radiochem. Radioanal. Lett. **22**, 361-366, 1975.

[9] Beyermann, Z. Anal. Chem. **200**, 183-197, 1964.

[10] D. Perrin and I. Sayce, Talanta **14**, 833-842, 1967.

[11] G. Charlot, Les réactions chimiques en solution, Masson, 1969.

[12] A. Ringbom, Complexation in analytical chemistry, Interscience, 1963.

[13] L. Sillen, Stability constants of metal ion complexes, Suppl. No. 1, The Chemical Society, 1971.

[14] G. Charlot and R. Gauguin, Les méthodes d'analyse des reactions en solution, Masson et cie, 1951, p. 36 ff.

[15] A. Novikov and Shekotyva, Radiokhimia **14**, 152-153, 1972.

[16] I. Tananaev and M. Shpirt, Russ. J. Inorg. Chem. **7**, 221-222, 1962.

[17] G. Agarkova and L. Aksenova, Russ. J. Inorg. Chem. **14**, 837-838, 1969.

[18] X. Kuus, Zhurn. Anal. Khim **26**, 166-170, 1971.

[19] O. Koch and G. Koch, Handbuch der Spurenanalyse, Springer Verlag, Berlin, 1974, p. 359.

[20] A. Mizuike in R. Bunshah (ed.), Techniques of metals research, Vol. III, part I, ch. 2, p. 225, Wiley, 1970.

[21] T. S. West, Anal. Chim. Acta **25**, 405-421, 1971.

[22] R. Gadde and H. Laitinen, Anal. Chem. **46**, 2022-2036, 1974.

[23] R. Gadde and H. Laitinen, Envir. Lett. **5**, 223-235, 1973.

[24] E. Bruninx and E. van Meyl, Anal. Chim. Acta **80**, 85-95, 1975.

[25] F. Cotton and G. Wilkinson, Advanced inorganic chemistry, Interscience Publ., New York, 1972, p. 475.

[26] W. Pugh, J. Chem. Soc. **128**, 2828-2832, 1926.

[27] A. Azzam, A. Elatrash and N. Ghattas, J. Radional. Chem. **2**, 255-262, 1969.

[28] P. Strohal and D. Höthig-Hus, Mikrochim. Acta 1974, pp. 899-907.

[29] E. Rona, D. Hood, L. Muse and B. Buglio, Limn. Oceanogr. **7**, 201-206, 1962.

[30] C. F. Baes and R. E. Mesmer, The hydrolysis of cations, Wiley-Interscience, New York, 1976, p. 348.

[31] R. James, P. Stiglich and T. Healy, Faraday Discuss. of the Chemical Soc. **59**, 142-156, 1975.

[32] J. Gulledge and J. O'Connor, J. Am. Water Works Ass. 1973, pp. 548-552.

[33] J. Ferguson and M. Anderson, Chem. Wat. Supply, Treatm. Distr. Symp. 1973, ch. 7.

[34] T. Fujinaga, M. Koyama, K. Izutsu, S. Himeno and M. Kawashima, J. Chem. Soc. Jap., Chem. and Ind. Chem. 1974-8, pp. 1489-1493.

[35] A. Novikov and L. Gordeeva, Radiokhimia **14**, 14-20, 1972.

[36] M. Kurbatov, F. Yu and J. Kurbatov, J. Chem. Phys. **16**, 87-91, 1948.

[37] M. Kurbatov and J. Kurbatov, J. Am. Chem. Soc. **69**, 438-441, 1967.

[38] J. Kurbatov, J. Kulp and E. Mach, J. Am. Chem. Soc. **67**, 1923-1929, 1945.

[39] J. Duval and M. Kurbatov, J. Phys. Chem. **56**, 982-984, 1952.

[40] M. Kurbatov, J. Am. Chem. Soc. **71**, 858-863, 1949.

[41] H. Posselt and W. Weber, Chem. Wat. supply, Treatm. Distr. Symp. 1973, ch. 5.

[42] F. Kepak, M. Nuderoya and J. Kanka, J. Radioanal. Chem. **14**, 325-335, 1973.

[43] M. Kurbatov, G. Wood and J. Kurbatov, J. Phys. Chem. **55**, 1170-1182, 1951.

[44] M. Kurbatov and G. Wood, J. Phys. Chem. **56**, 698-701, 1952.

[45] E. Suzuki and N. Ikeda, Radioisot. (Jap.) **23**, 373-378, 1974.

[46] J. Sipalo-Zuljevic and R. Wolf, Mikrochim. Acta 1973, pp. 315-320.

[47] D. Kyriacou, Surf. Sci. **8**, 371-372, 1967.

[48] R. Rafaeloff, Radiochem. Radioanal. Lett. **9**, 373-379, 1972.
[49] A. Novikov and E. Tsjechotirova, Radiokhim. **14**, 152-154, 1972.
[50] R. Addis Lockwood and K. Chen, Envir. Lett. **6**, 151-166, 1974.
[51] S. Shimonura, Y. Nishihara, Y. Fukumoto and Y. Tanase, J. Hyg. Chem. **15**, 84-89, 1969.
[52] F. Kepak, Coll. Czech. Chem. Comm. **31**, 3500-3510, 1966.
[53] F. Kepak, Coll. Czech. Chem. Comm. **31**, 1493-1500, 1966.
[54] G. LeGendre and D. Runnels, Envir. Sci. Techn. **9**, 744-749, 1975.
[55] Y. Kim and H. Zeitlin, Anal. Chim. Act. **46**, 1-8, 1969.
[56] A. Novikov, E. Tsjechotirova and T. Zakrevskaya, Radiokhim. **13**, 728-733, 1971.
[57] Z. Kolarik and J. Krtil, Coll. Czech. Chem. Comm. **30**, 724-735, 1965.
[58] F. Kepak, Coll. Czech. Chem. Comm. **30**, 1464-1472, 1965.
[59] A. Novikov and S. Rustamov, Radiokhim. **13**, 134-137, 1971.
[60] S. Nagatsuku, Y. Tanizaki and Y. Okano, Radioisot. (Jap.) **23**, 27-33, 1974.
[61] Z. Kolarik and V. Kourim, Coll. Czech. Chem. Comm. **25**, 1000-1007, 1960.
[62] Z. Kolarik, Coll. Czech. Chem. Comm. **27**, 938-949, 1962.
[63] W. Schulze and M. Scheffler, Z. Anal. Chem. **226**, 395-401, 1967.
[64] W. Schulze and M. Scheffler, Z. Anal. Chem. **229**, 161-169, 1967.
[65] A. Novikov and L. Gordeva, Radiokhim **13**, 793-977, 1971.
[66] E. Kaneko, Bunseki Kagaku **25**, 299-301, 1976.
[67] Z. Kolarik and V. Kourim, Coll. Czech. Chem. Comm. **26**, 1082-1091, 1961.
[68] P. Benes and E. Vidova, Coll. Czech. Chem. Comm. **36**, 2032-2036, 1971.
[69] V. Rachinskii and L. Zkukova, Radiokhim **15**, 7-12, 1973.

EFFECTS OF TRICHLOROETHANE OXIDATION OF SILICON WAFERS ON SiO_2 AND Si PROPERTIES

by A. J. LINSSEN and H. L. PEEK

Abstract

The effects of oxidation of (100) Si in dry O_2 mixed with 1.1.1.-$C_2H_3Cl_3$ (C_{33}) at 1050 °C up to 1200 °C on SiO_2 and Si properties have been investigated. The addition of C_{33} tot the O_2 flow enhanced the oxidation rate with respect to the oxidation in dry oxygen. The chlorine was found to be located within the first 300 Å of the SiO_2 layer from the Si/SiO_2 interface. The Na and K contents of C_{33} oxide layers were higher than those of dry thermal oxide layers due to the transport of Na and K from the quartz tube wall to the Si slices by the HCl. Neutralization of mobile sodium was obtained if the SiO_2 surface appeared to be grainy. C_{33} oxide layers showed instabilities after the application of a positive and a negative field stress at 290 °C. A negative oxide charge and interface states were created in the upper half of the Si bandgap. The C_{33} oxidations were not observed to improve the fixed oxide charge, the interface state density or the minority-carrier bulk lifetime. Complete suppression of stacking fault generation was the predominant effect on Si bulk material.

1. Introduction

The quality of thermally grown SiO_2 layers determines to a great extent the performance of MOS devices and charge-coupled devices. It has been reported that the quality of thermally grown SiO_2 layers is improved and the minority-carrier generation lifetime is increased when O_2 mixed with HCl is used [1,2,3,4]. However, a grainy oxide surface due to the formation of a second phase within the SiO_2 layer near to the Si/SiO_2 interface [7]) and a negative bias instability [5,6]) have also been observed for certain HCl oxides. Good correlation exists between the formation of a second phase and the passivation of mobile sodium ions [7]).

For reasons of safety and because of the need of special installation requirements involved in the use of HCl, it has been proposed that C_{33} (1.1.1.-trichloroethane) would be a substitute for HCl [8,9]). The relative concentration of reaction products HCl, Cl_2, H_2O formed by heating up mixtures of O_2 and HCl or C_{33} are exactly the same provided that the HCl concentration is three times the C_{33} concentration [8]). As long as lean mixtures of C_{33} and O_2 are used to allow the formation of CO_2, the carbon atoms do not intervene.

It is the purpose of this paper to show that certain C_{33} oxide layers show a positive and negative bias instability. Neutralization of mobile sodium ions is achieved if the oxide surface is grainy. Minority-carrier lifetime is not improved by C_{33} oxidations, but the generation of stacking faults can be completely suppressed.

2. Sample preparation

Chemically polished dislocation-free silicon wafers, (100)-oriented, were prepared from phosphorus doped wafers (resistivity 1 Ωcm) grown by the floating zone method. These wafers were cleaned, rinsed and then oxidized in dry O_2 mixed with various concentrations of C_{33} at 1050 °C up to 1200 °C during oxidation times up to six hours. After metallization of the frontside with aluminium the oxide was removed from the backside in buffered HF. Circular MOS capacitor electrodes (area: 1 mm^2) were defined by photolithography and etching of the aluminium. After metallization of the backside a low-temperature anneal at 450 °C in wet N_2 during 30 minutes was performed.

3. Results and discussions

3.1. *The increase in oxidation rate*

The oxide thickness versus the oxidation time is plotted in fig. 1 for various

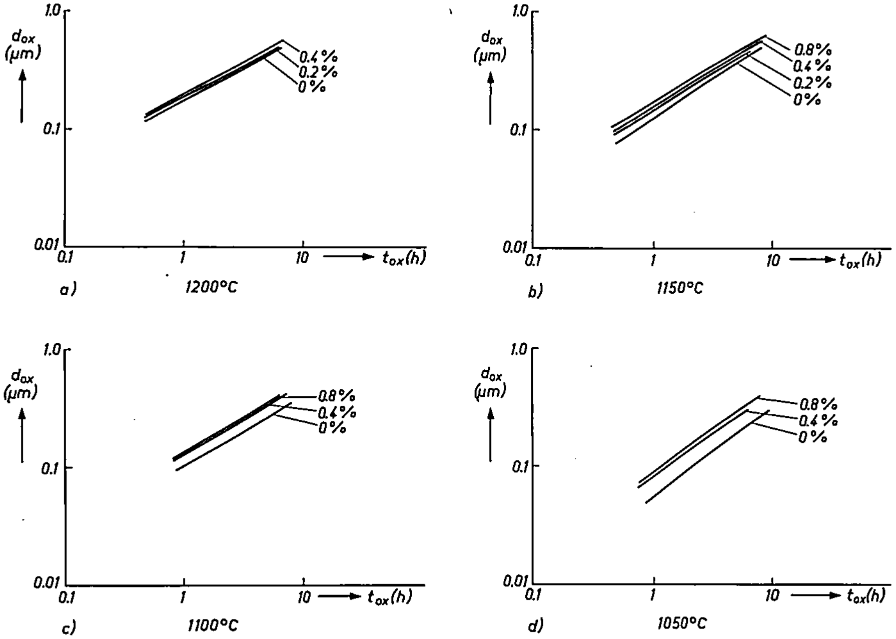


Fig. 1. Oxide thickness d_{ox} vs oxidation time t_{ox} for the oxidation of (100)-oriented n-type silicon in various O_2/C_{33} mixtures at (*a*) 1200 °C; (*b*) 1150 °C; (*c*) 1100 °C; (*d*) 1050 °C.

C_{33} concentrations and oxidation temperatures. The addition of C_{33} has the effect of increasing the oxidation rate relative to the oxidation in dry O_2. This enhanced oxidation rate can not be explained by the influence of water vapour formed during the combustion of C_{33}, because the effect of C_{33} addition is greater than would be expected in a comparable O_2/H_2O ambient [1]).

The Cl_2 and HCl formed will affect the oxidation rate, particularly the reaction velocity [10]). Furthermore a higher diffusivity of O_2 and H_2O in oxides grown in O_2 mixed with a chlorine compound will enhance the oxidation rate [10,11]). The SiO_2 films grown in C_{33} concentrations for longer times showed a grainy appearance when viewed through an optical microscope. This grainy appearance has also been observed for HCl oxides grown in large HCl concentrations of for longer oxidation times at lower HCl concentrations [6,7,10]). This grainy structure may be due to a second phase within the oxide near the Si/SiO_2 interface [7]) and a corrosion of the Si surface.

3.2. *Chlorine distribution*

The chlorine distribution in SiO_2 layers, grown at 1200 °C in dry O_2 mixed with 0.2 or 0.4% C_{33} up to six hours oxidation time, was analysed by Rutherford backscattering (RBS). To determine the location of the chlorine in the SiO_2 layers, thin layers were subsequently removed in buffered HF (see table I). The analysis showed large chlorine concentrations very near to the Si/SiO_2 interface and low concentrations ($\leqslant 6 \times 10^{18}$ cm^{-3}) in the rest of the SiO_2 layer. Almost all the chlorine was incorporated in the first 300 Å from the Si/SiO_2 interface, an effect which has also been observed for HCl oxides [6,12]). The profiles had the same shape regardless of the SiO_2 thickness and the C_{33} concentration. The total amount of chlorine incorporated in the SiO_2 layer was proportional to the C_{33} concentration and to the SiO_2 thickness.

Surface roughening of the SiO_2 layer was observed if more than 2.5×10^{15} chlorine atoms/cm^2 were incorporated in the SiO_2 layer. At the same amounts square-shaped etchpits were observed on the silicon surface after it had been etched by a preferential etch proposed by Jenkins [24]). Corrosion of the silicon surface was observed if about 5×10^{15} chlorine atoms/cm^2 were incorporated in the SiO_2 layer. The formation of a second phase in the SiO_2 layer near the Si/SiO_2 interface (which caused in the grainy oxide surface [7])) was observed before corrosion of the silicon surface. Probably the silicon surface is damaged by the formation of the second phase in the SiO_2 layer.

3.3. *Sodium content of the SiO_2 layer*

Oxidation in a chlorine containing ambient can lead to neutralization of mobile sodium ions [2,6]). The neutralization is attributed to the chlorine incor-

TABLE I

Oxidations of phosphorus doped Si wafers in O_2/C_{33} ambients at 1200 °C (1 Ωcm, (100) float zone)

C_{33} concentration (%)	oxidation time (h)	SiO_2 thickness (μm)	SiO_2 thinned down to	[Cl] (cm^{-2})
0.2	1	0.18	0.18 μm	7×10^{14}
			0.0085	4×10^{14}
			0	$< 7 \times 10^{13}$
0.2	2	0.28	0.28	13×10^{14}
			0.021	13×10^{14}
			0	$< 7 \times 10^{13}$
0.2	6	0.50	0.50	
			0.026	27×10^{14}
			0	$< 7 \times 10^{13}$
0.4	6	0.52	0.52	
			0.16	56×10^{14}
			0.12	56×10^{14}
			0.08	56×10^{14}
			0.038	56×10^{14}
			0	$< 7 \times 10^{13}$

porated in the oxide. The sodium content of C_{33} and dry thermal oxide layers grown on phosphorus doped wafers was determined by applying a triangular voltage sweep (TVS) and integrating the extra displacement current caused by the mobile ions [13]). At a temperature of 290 °C and a voltage sweep of 21 mV/s the mobility of sodium is sufficiently high to cross an oxide layer 0.5 μm thick within a few seconds at an electric field of 100 V/cm [14]). At this electric field potassium ions do not contribute to the displacement current since they are less mobile than sodium ions [14]). In figure 2 the mobile sodium content of oxide layers versus the oxidation time was plotted at various concentrations of C_{33} and oxidation temperatures. From these figures it can be seen that the mobile sodium content built in the SiO_2 layer drops after longer oxidation times for low C_{33} concentrations or after shorter oxidation times for higher C_{33} concentrations. When the SiO_2 layer is grainy a drop in the mobile sodium content is observed. Monkowski et al.[7]) observed neutralization of mobile sodium only if the SiO_2 surface was rough or grainy, i.e. when a second phase was formed in the oxide near the Si/SiO_2 interface.

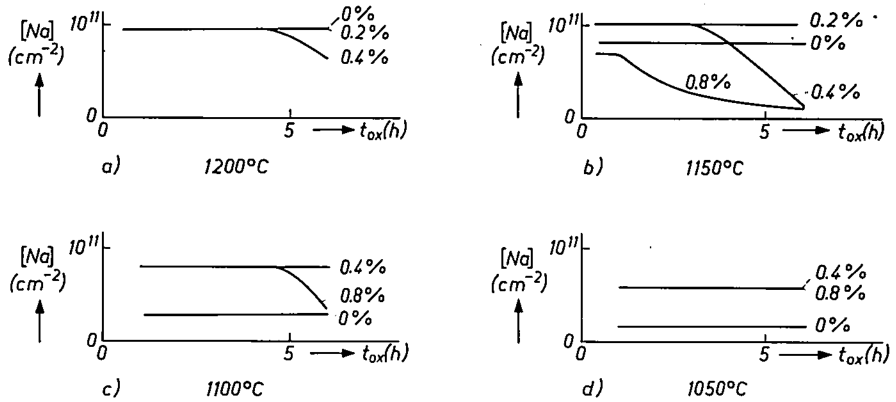


Fig. 2. [Na] vs oxidation time t_{ox} for the oxidation of (100) oriented *n*-type silicon in various O_2/C_{33} mixtures at (*a*) 1200 °C; (*b*) 1150 °C; (*c*) 1100 °C; (*d*) 1050 °C.

It has also been found that more than 10^{15} chlorine atoms/cm^2 must be incorporated in SiO_2 layers grown at 1200 °C in O_2/HCl ambient before neutralization of sodium is observed [12]). In our case a drop in the mobile sodium content of the SiO_2 layer was observed if more than 2×10^{15} chlorine atoms/cm^2 were incorporated in the SiO_2 layers grown at 1200 °C.

The mobile sodium content of dry thermal oxide layers was in most cases lower than with C_{33} oxides. During oxidation in O_2/C_{33} ambient there is a reaction between the formed HCl with the sodium and potassium in the quartz tube wall [15]). NaCl and KCl reaction products are evaporated and can be incorporated in the growing SiO_2 layer.

3.4. *Stability of the oxide layers*

In order to study the stability of C_{33} oxide layers positive and negative field stress measurements at 290 °C during 5 min were performed on MOS capacitors. The electric field in the oxide layer was 10^6 V/cm. From the shift in flat-band voltage V_{FB} before and after the field stress an "effective" oxide charge $Q^{\pm}{}_{eff}$ was calculated:

$$Q^{\pm}{}_{eff} = -C_{ox}\,\Delta V_{FB},$$

where C_{ox} is the oxide capacitance, while $Q^{+}{}_{eff}$ and $Q^{-}{}_{eff}$ are the effective oxide charges after positive and negative field stress, respectively. The measured shift of the flat-band voltage contains all effects that may occur during positive or negative field stress: displacement of mobile ions, trapping or detrapping of holes and/or electrons creation of interface states and fixed oxide charge.

In figures 3 and 4 the effective oxide charge versus the oxidation time at various concentrations of C_{33} and oxidation temperatures has been plotted for

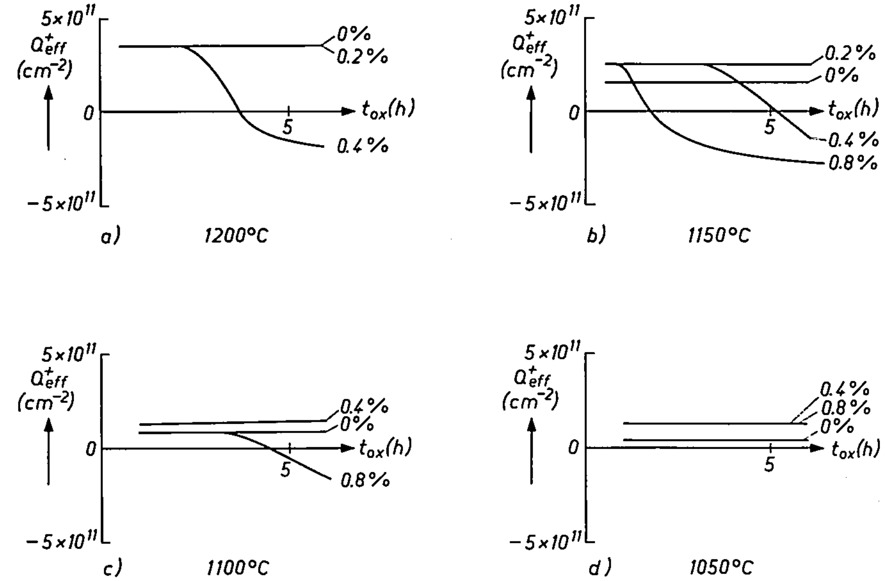


Fig. 3. Effective oxide charge $Q^{+}{}_{\text{eff}}$ after positive field stress (290 °C; 10^6 V/cm; 5 min) vs oxidation time t_{ox} for the oxidation of (100) oriented *n*-type silicon in various O_2/C_{33} mixtures at (*a*) 1200 °C; (*b*) 1150 °C; (*c*) 1100 °C; (*d*) 1050 °C.

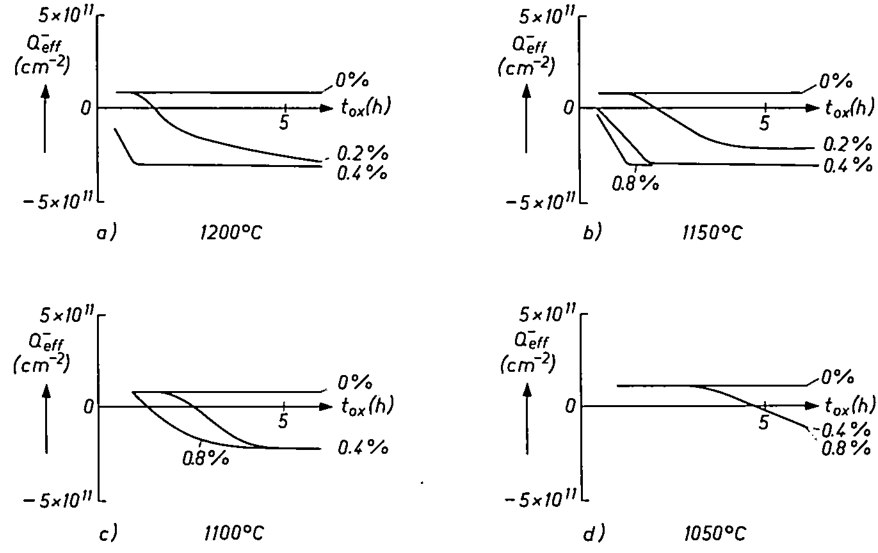


Fig. 4. Effective oxide charge $Q^{-}{}_{\text{eff}}$ after negative field stress (290 °C; 10^6 V/cm; 5 min) vs oxidation time t_{ox} for the oxidation of (100) oriented *n*-type silicon in various O_2/C_{33} mixtures at (*a*) 1200 °C; (*b*) 1150 °C; (*c*) 1100 °C; (*d*) 1050 °C.

positive and negative field stress, respectively. From TVS measurements at 400 °C on SiO_2 layers which did not show a grainy surface, we concluded that the effective oxide charge was completely determined by the transport of Na and K from the SiO_2/metal to the Si/SiO_2 interface. For SiO_2 layers with show a grainy structure the effective oxide charge dropped, becoming negative and saturated.

To explain the negative effective oxide charge we propose the following. The negative oxide charge was observed just when a second phase in the SiO_2 layer was formed. The mechanism is related to the structure at the second phase, which could act as electron traps. Due to the electric field and temperature an electron can be captured in these traps leading to negatively charged traps. The drop in the effective oxide charge can be explained by neutralization of mobile sodium and potassium ions and electron traps becoming negatively charged during stressing.

After negative field stress, dry thermal SiO_2 layers showed an increase in fixed oxide charge and in the number of interface states [5]). The effective oxide charge was about 10^{11} cm^{-2}. In the case of C_{33} oxide layers, interface states were created in the upper half of the silicon bandgap. (The *C–V* curves on *p*-type silicon showed no distortion, since interface states in the upper half of the silicon bandgap do not change their charged state. Only a shift to negative voltage was observed.) Hess [5]) has also observed the creation of interface states in the upper half of the silicon bandgap after negative field stress for HCl oxide layers. The negative field effect could be caused by breaking of Si–Si, Si–O and Si–Cl bonds near the Si/SiO_2 interface. The presence of chlorine in the interfacial region might be expected to promote bond breaking due to a straining of these bonds [5]). The effective oxide charge decreases with increasing chlorine concentration untill a second phase is formed in the SiO_2 layer. It saturates at about 3×10^{11} cm^{-2}.

The instabilities make C_{33} oxide layers less suited for application as a gate oxide, particularly if they are grown at temperature of 1100 °C and higher. At 1050 °C for a C_{33} concentration of 1 % and with an oxidation lasting less than three hours, the instabilities are small. Taking into account that the effect density of SiO_2 layers grown at 1000 °C in O_2 mixed with 1 % C_{33} is significantly reduced [8]) it may be advantageous to grow gate oxide layers in O_2/C_{33} ambients at temperatures lower than 1050 °C.

3.5. *Interface state density/fixed oxide charge*

The interface state density measured by the quasi-static method [16]), was less than 2×10^{10} cm^{-2}, except for samples that had a grainy oxide surface ($\geqslant 5 \times 10^{10}$ cm^{-2} eV^{-1}). C_{33} oxidations are not necessary to achieve a low interface density, which means having a clean interface with a small degree of disorder. Due to the cleaning effect of C_{33} on the furnace tube wall, the possible contamination by impurities of the oxide layer is minimized during oxidation. This could result in a cleaner interface and thus a lower interface state density. The fixed oxide charge of C_{33} oxide films was about 50% higher than the fixed oxide charge of oxides grown in dry O_2. An explanation for this may be the following. During the oxidation of silicon in O_2/HCl or O_2/C_{33} ambient non bridging, or terminally bonded, intermediates Si–Cl will be formed. These intermediates increase the fixed oxide charge [17]) and so a decrease of the fixed oxide charge may not be expected.

3.6. *Silicon bulk properties*

The effect of C_{33} oxidation on bulk silicon is to suppress completely the generation of stacking faults at 1150 °C and 1200 °C for C_{33} concentratons of 1% or lower. For larger oxidation of the silicon square-shaped etchpits on the silicon surface are observed. The thicknesses of the SiO_2 layers and the C_{33} concentrations at which complete suppression of stacking fault generation are obtained and square-shaped etchpits on the silicon surface are avoided, are given in table II for different oxidation temperatures. Only at 1200 °C there is a range of SiO_2 thicknesses and C_{33} concentrations where both requirements

TABLE II

The C_{33} concentration and oxide thickness (d_{SiO_2}) at which the generation of stacking faults (SF) is completely suppressed and no etchpits are revealed at different oxidation temperatures T_{ox}

T_{ox} (°C)	C_{33} concentration (%)	SF suppression for $d_{SiO_2} \geqslant$	no etchpits $d_{SiO_2} \leqslant$
1200	0.2	0.28 μm	0.42 μm
1200	0.4	0.14	0.21
1150	0.2	A *)	A *)
	0.4	0.37	0.37
	0.8	0.18	0.18
1100	0.4	A *)	A *)
	0.8	A *)	0.32
1050	0.4	A *)	A *)
	0.8	A *)	A *)

A *) Since only oxide thicknesses up to about 0.5 μm were grown, complete suppression of stacking fault generation and/or etchpits was not observed in these cases.

can be fulfilled. Complete suppression of stacking fault generation was obtained when more than 10^{15} chlorine atoms/cm^2 where incorporated in the SiO_2 layer, whereas to avoid a "damage" silicon surface less than 2×10^{15} chlorine atoms/cm^2 had to be incorporated. A correlation was observed between the presence of stacking faults in the depletion layer and the dark current and the number of dark-current spikes [9,18]). When C_{33} oxidations are used under conditions where no stacking faults are generated during oxidation, the number of dark-current spikes is reduced. The effect of C_{33} oxidations on the generation of bulk dark currents can be examined by studying the relaxation times of MOS capacitors pulsed in deep depletion. A Zerbst analysis was applied to the *c–t* curves to separate bulk minority-carrier generation and surface generation [19,20,21]).

Bulk generation lifetimes are 200–800 μs and the surface generation velocity is about 2 cm/s for a depleted surface and less than 0.01 cm/s for an inverted surface. An optimum C_{33} concentration could not be found. A strong decrease of the generation lifetime was observed after C_{33} oxidation which yielded a corroded silicon surface. Robinson et al.[8]) and Young [22]) did observe an increase in minority bulk lifetime after HCl oxidation. However, Green et al.[23]) found no marked effects on the metallic content of silicon wafers after a heat treatment in a mixture of HCl, H_2 and SiH_4. Also from our observations it can be concluded that the suppression of dark-current spikes due to the suppression of stacking fault generation seems to be the predominant mechanism of C_{33} oxidation and not the decrease in the metallic content of silicon wafers.

4. Conclusions

The addition of C_{33} to the O_2 flow resulted in a higher oxidation rate of (100) silicon than the oxidation rate in dry O_2. The chlorine in the SiO_2 layer is located in the first 300 Å from the Si/SiO_2 interface, as was determined by means of RBS and subsequent removal of thin layers of SiO_2 in buffered HF.

C_{33} oxide layers had more mobile sodium and potassium ions than dry oxygen layers due to the transport of Na and K from the quartz tube wall to the slices by the HCl. Neutralization of mobile Na was achieved when the SiO_2 surface was grainy and the silicon surface showed square-shaped etchpits after preferential etching of the silicon surface. An increase in interface state density was also observed. SiO_2 layers which showed a grainy surface showed a negative effective oxide charge after positive field stressing at 290 °C with an electric field of 10^6 V/cm. After negative field stress at 290 °C with an electric field of 10^6 V/cm, interface states were created in the upper half of the Si bandgap. It is suggested that chlorine in the SiO_2 layer and the breaking of Si–Cl bonds in the first 300 Å of the SiO_2 layer could account for these effects. The effective oxide charge after positive and negative field stress saturated when a second phase was formed very near to the Si/SiO_2 interface. C_{33} oxidations were not observed to have any influence on minority-carrier bulk lifetime and interface state density. Only at 1200 °C we were able to suppress the generation of stacking faults completely and achieve a silicon surface without square-shaped etchpits after preferential etching. When complete suppression of stacking faults generation at 1200 °C was achieved more than 10^{15} chlorine atoms/cm^2 were built in the SiO_2 layer. The instabilities observed make C_{33} oxide layers less suited for application as a gate oxide grown at a temperature of 1050 °C and higher. The advantages of using C_{33} oxidations are the complete suppression of stacking fault generation at 1200 °C, a reduction of SiO_2 defect density for oxide layers grown at 1000 °C with 1% C_{33}, and the continuous cleaning of the furnace tube from the metallic contaminants. Oxide films grown in dry oxygen in furnace tubes which were flushed with O_2/C_{33} mixture before oxidation, were found to have a lower Na and K content.

Acknowledgement

The authors wish to thank Mr Y. Tamminga for doing the RBS measurements.

Philips Research Laboratories *Eindhoven, August 1978*

REFERENCES

[1] R. J. Kriegler, Y. C. Cheng and D. R. Colton, J. Electrochem. Soc. **119**, 388, 1972.
[2] R. J. Kriegler, Proc. ISSCC Philadelphia, 1975, p. 56.
[3] C. M. Osburn, J. Electrochem. Soc. **121**, 809, 1974.
[4] D. H. Robinson and F. P. Heiman, J. Electrochem. Soc. **118**, 141, 1971.
[5] P. W. Hess, J. Electrochem. Soc. **124**, 740, 1977
[6] Y. J. van der Meulen, C. M. Osburn and J. F. Ziegler, J. Electrochem. Soc. **122**, 284, 1975.
[7] J. Monkowski, J. Stach and R. E. Tressler, in H. R. Huff and E. Sirtl (eds), Semicond. silicon 1977, Proc. Vol. 77-2, p. 324.
[8] E. J. Janssens, to be published in J. Electrochem. Soc.
[9] C. L. Claeys, G. J. Declerck, E. E. Laes and R. J. van Overstraeten, in H. R. Huff and E. Sirtl (eds), Semicond. silicon 1977, Proc. Vol. 77-2, p. 773.
[10] K. Hirabayashi and J. Iwanura, J. Electrochem. Soc. **120**, 1595, 1973.
[11] D. W. Hess and B. E. Deal, J. Electrochem. Soc. **124**, 735, 1977.
[12] A. Rohatgi, S. R. Butler and F. J. Feigl, Appl. Phys. Lett. **30**, 104, 1977.
[13] N. Y. Chou, J. Electrochem. Soc. **118**, 601, 1971.
[14] J. P. Stagg, Appl. Phys. Lett. **31**, 532, 1977.
[15] J. Mayo and W. H. Evans, J. Electrochem. Soc. **124**, 780, 1977.
[16] M. Kuhn, Solid State Electron. **18**, 873, 1970.
[17] S. I. Raider and A. Berman, J. Electrochem. Soc. **125**, 629, 1978.
[18] K. Tanikawa, Appl. Phys. Lett. **28**, 285, 1976.
[19] M. Zerbst, Z. Angew. Physik **22**, 30, 1966.
[20].D. K. Schröder and H. C. Nathanson, Solid State Electr. **13**, 577, 1970.
[21] D. K. Schröder and J. Guldberg, Solid State Electr. **14**, 1285, 1971.
[22] D. R. Young and C. M. Osburn, J. Electrochem. Soc. **120**, 1578, 1973.
[23] J. M. Green, C. M. Osburn and T. O. Sedgick, J. Electr. Mat. **3**, 579, 1974.
[24] M. W. Jenkins, J. Electrochem. Soc. **124**, 757, 1977.

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*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique appliquée. 3 Avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 5100 Aachen, West-Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, West-Germany
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).

P. Baudet, M. Parisot and R. Veilex: Amplificateur bas bruit à transistors F.E.T. au GaAs.
AGARD Conf. Proc. No. 197, pp. 3.1-3.16, 1977. *L*

C. I. M. Beenakker: A cavity for microwave-induced plasmas operated in helium and argon at atmospheric pressure.
Spectrochim. Acta **31B**, 483-486, 1976. *E*

G. Bergmann: Die Zündung von Gasentladungslampen.
Lichttechnik **28**, 493-496, 1976. *A*

H. Bouma and Ch. P. Legein (Institute for Perception Research, Eindhoven): Foveal and parafoveal recognition of letters and words by dyslexics and by average readers.
Neuropsychologia **15**, 69-80, 1977.

J. W. Broer: Let's become scientist-proof!
Proc. 23rd Int. Technical Communication Conf., Washington D.C., 1976, pp. 113-118. *E*

R. Brehm: Het ultrasoon boren van gaatjes met een diameter kleiner dan 200 μm.
Mikroniek **16**, 91-95, 1976. *E*

R. Brehm and M. Helmig: Het meten van hoge snelheden in een glazen smoor-ejecteur.
Mikroniek **16**, 138-141, 1976. *E*

K. H. J. Buschow: Note on the change in magnetic properties of $GdCo_2$ on hydrogen absorption.
J. Less-Common Met. **51**, 173-175, 1977. *E*

K. H. J. Buschow and F. J. van Steenwijk: Magnetic properties and Mössbauer effect of some Eu-Cd compounds.
Physica **85B**, 122-126, 1976. *E*

F. M. A. Carpay and A. L. Stuijts: Characterization of grain growth phenomena during sintering of single-phase ceramics.
Science of Ceramics **8**, 23-38, 1976. *E*

V. Chalmeton: Principaux éléments d'un système radiographique.
Acta Electronica **20**, 83-87, 1977. *L*

V. Chalmeton: Chaîne de radioscopie 400 kV avec intensificateur à galette de microcanaux.
Acta Electronica **20**, 53-64, 1977. *L*

T. A. C. M. Claasen, W. F. G. Mecklenbräuker and J. B. H. Peek: Effects of quantization and overflow in recursive digital filters.
IEEE Trans. **ASSP-24**, 517-529, 1976. *E*

T. A. C. M. Claasen and J. B. H. Peek: A digital receiver for tone detection applications.
IEEE Trans. **COM-24**, 1291-1300, 1976. *E*

J. E. Curran: Vacuum technologies applied to electronic component fabrication.
J. Vac. Sci. Technol. **14**, 108-113, 1977. *R*

N. H. Dekkers and H. de Lang: A detection method for producing phase and amplitude images simultaneously in a scanning transmission electron microscope.
Philips Tech. Rev. **37**, 1-9, 1977. *E*

N. H. Dekkers, H. de Lang and K. D. van der Mast: Field emission STEM on a Philips EM 400 with a new detection system for phase and amplitude contrast.
J. Microsc. Spectrosc. Electron. **1**, 511-512, 1976. *E*

P. Delsarte: Properties and applications of the recurrence
$F(i+1, k+1, n+1) = q^{k+1}F(i, k+1, n) - q^kF(i, k, n)$.
SIAM J. Appl. Math. **31**, 262-270, 1976. *B*

A. M. van Diepen and Th. J. A. Popma: Mössbauer effect and magnetic properties of an amorphous Fe_2O_3.
J. Physique **37**, C6/755-758, 1976. *E*

R. J. Dolphin and F. W. Willmott: Band-broadening effects in a column-switching system for HPLC.
J. Chromatogr. Sci. **14**, 584-588, 1976. *R*

J. W. F. Dorleijn and A. R. Miedema: The residual resistivities of dilute iron-based alloys in the two-current model.
J. Physics F **7**, L 23-25, 1977. *E*

J. W. F. Dorleijn and A. R. Miedema: The magnetic resistance anisotropy in nickel and iron based alloys.
AIP Conf. Proc. **34**, 50-54, 1976. *E*

H. Durand: Introduction *(to issue on Solar energy)*.
Acta Electronica **20**, 97-99, 1977. *(In English and in French.)* *L*

L.-E. Eriksson and H. C. van den Elzen: An equalizer structure with reduced sampling time reference sensitivity.
IEEE Trans. **COM-24**, 1337-1343, 1976. *E*

G. Frank, L. Brock and H. D. Bausen: The solubilities of Sn in In_2O_3 and of In in SnO_2 crystals grown from Sn-In melts.
J. Crystal Growth **36**, 179-180, 1976. *A*

J. A. Geurst: Canonical form of generalised two-fluid equations for helium II.
Physics Letters **59A**, 351-352, 1976. *E*

J.-M. Goethals: Codes as elements in a group algebra.
Proc. Int. Symp. on Mathematical Systems Theory, Udine, 1975, pp. 277-283, 1976. *B*

G. Groh: Tomosynthesis and coded aperture imaging: new approaches to three-dimensional imaging in diagnostic radiography.
Proc. Roy. Soc. London B **195**, 299-306, 1977. *H*

G. J. van Gurp: Electromigration in cobalt films.
Thin Solid Films **38**, 295-311, 1976. *E*

D. Hennings: The broadening of the Curie peak by lattice defects in ferroelectric lead titanate — diffuse phase transition.
Science of Ceramics **8**, 203-211, 1976. *A*

E. P. Honig and D. den Engelsen: Reflection from stratified anisotropic media: an alternative method.
Optica Acta **24**, 89-95, 1977. *E*

F. de Jager and M. Christiaens: A fast automatic equalizer for data links.
Philips Tech. Rev. **37**, 10-24, 1977. *E*

E. Kirchner and J. Leu (C. H. F. Müller, Hamburg): Technical components for the generation of X-rays.
Acta Electronica **20**, 25-32, 1977.

F. M. Klaassen: Physics of and models for I^2L.
Tech. Dig. 1976 Int. Electron Devices Meeting, Washington D.C., pp. 299-303. *E*

H. A. Klasens and J. Goossen: The iodide interference with silver chloride electrodes.
Anal. Chim. Acta **88**, 41-46, 1977. *E*

W. L. Konijnendijk: Structure of glasses in the systems CaO-Na_2O-B_2O_3 and MgO-Na_2O-B_2O_3 studied by Raman scattering.
Phys. Chem. Glasses **17**, 205-208, 1976. *E*

W. L. Konijnendijk (Philips Lighting Division, Eindhoven) **and J. M. Stevels** (Eindhoven University of Technology): Viscosity of borosilicate glasses in relation to their structure.
Verres Réfract. **30**, 821-826, 1976.

L. Koppens: The decomposition of organometallic precursors for ferrite powders.
Science of Ceramics **8**, 101-109, 1976. *E*

G. Kowalski: Reconstruction of objects from their projections. A simple reconstruction algorithm, theoretical and simulation studies.
EDV in Medizin und Biologie **8**, 1-8, 1977. *H*

J.-P. Krumme, H. Heitmann, D. Mateika and K. Witter: MOPS, a magneto-optic-photoconductor sandwich for optical information storage.
J. Appl. Phys. **48**, 366-368, 1977. *H*

W. Kühl and J. E. Schrijvers (Philips Elcoma Division, Eindhoven): Design aspects of X-ray image intensifiers.
Acta Electronica **20**, 41-51, 1977.

P. E. J. Legierse (Philips Philite- en Metaalwarenfabrieken, Eindhoven): Fotochemische metaalbewerking.
Polytechn. T. Werktuigbouw **32**, 27-35, 1977.

D. A. Lucas and R. P. Vincent: A precision approach monitor.
The future of aircraft all-weather operations?, Int. Conf. London, 1976 (IEE Conf. Publn No. 147), pp. 80-83. *R*

J. Magarshack: Device innovation in communication systems.
Microwave J. **20**, No. 2, p. 50, Feb. 1977. *L*

R. Memming: Electrochemical surface reactions on non-metals.
Nat. Bur. Stand. spec. Publn No. 455, 267-289, 1976. *H*

R. F. Mitchell: Basics of SAW frequency filter design: a review.
Wave Electronics **2**, 111-132, 1976. *R*

J. H. Neave and B. A. Joyce: Some comments on electron-beam-induced adsorption.
J. Physics D **10**, 243-248, 1977. *R*

Ngo-Tich-Phuoc, G. M. Martin, C. Belin and E. Fabre: Homogeneity along Cl-compensated THM grown CdTe ingot.
Rev. Phys. Appl. **12**, 195-198, 1977. *L*

K. H. Nicholas: Implantation damage in silicon devices.
J. Physics D **10**, 393-407, 1977. *R*

A. G. van Nie: Modulated carriers in nonlinear systems.
Int. J. Circuit Theory & Appl. **5**, 69-79, 1977. *E*

C. van Opdorp, C. Werkhoven and A. T. Vink: A method to determine bulk lifetime and diffusion coefficient of minority carriers; application to *n*-type LPE GaP.
Appl. Phys. Letters **30**, 40-42, 1977. *E*

G. Piétri: Industrial X-ray control.
Acta Electronica **20,** 8 *(in English),* 9 *(in French),* 1977. *L*

R. J. van de Plassche: Dynamic element matching for high-accuracy monolithic D/A converters.
IEEE J. **SC-11,** 795-800, 1976. *E*

A. Posthuma de Boer and A. J. Pennings: Polyethylene networks crosslinked in solution: preparation, elastic behavior, and oriented crystallization, I. Crosslinking in solution.
J. Polymer Sci. Pol. Phys. Edn **14,** 187-210, 1976. *E*

W. Schäfer, G. Will and K. H. J. Buschow: A neutron diffraction investigation on the magnetic properties of ErMg and DyMg.
J. Physics C **9,** L 657-661, 1976. *E*

G. B. Scott and J. L. Page: The absorption spectra of $Y_3Fe_5O_{12}$ and $Y_3Ga_5O_{12}$:Fe^{3+} to 5.5 eV.
Phys. Stat. Sol. (b) **79,** 203-213, 1977. *R*

M. Sintzoff: Composing specifications of information structures.
New directions in algorithmic languages 1975, in S. A. Schuman (ed.), publ. IRIA, Rocquencourt, 1976, pp. 207-216. *B*

J. W. Slotboom: Minority carrier injection into heavily doped silicon.
Solid-State Electronics **20,** 167-170, 1977. *E*

N. V. Smith, P. K. Larsen, M. M. Traum and S. Chiang: Miniature analyser and its use in angle-resolved photoemission using synchrotron radiation.
Proc. Int. Symp. on Photoemission, Noordwijk, 1976, pp. 119-123. *E*

W. T. Stacy, A. B. Voermans and H. Logmans: Increased domain wall velocities due to an orthorhombic anisotropy in garnet epitaxial films.
Appl. Phys. Letters **29,** 817-819, 1976. *E*

S. Strijbos: Friction between a powder compact and a metal wall.
Science of Ceramics **8,** 415-427, 1976. *E*

E. H. Stupp and A. Milch: Measurement of thickness and diffusion length of thin epitaxial layers of GaP.
J. Appl. Phys. **48,** 282-285, 1977. *N*

J.-B. Theeten: Les agrégats, entre l'atome et le solide.
La Recherche **8,** 170-171, 1977. *L*

H. J. Verbeek (Philips Elcoma Division, Eindhoven): A model to evaluate design and application of components regarding heat transfer during soldering.
DVS-Ber. **40,** 35-39, 1976.

H. Verweij and W. L. Konijnendijk: Structural units in K_2O-PbO-SiO_2 glasses by Raman spectroscopy.
J. Amer. Ceramic Soc. **59,** 517-521, 1976. *E*

H. W. Werner: Characterization of ceramics by means of modern thin film and surface analytical techniques.
Science of Ceramics **8,** 55-79, 1976. *E*

H. P. J. Wijn: Trends in the technology of magnetic devices.
Physics in industry in E. O'Mongain and C. P. O'Toole (eds), Proc. Int. Conf. Dublin, 1976, publ. Pergamon Press, Oxford, 1976, pp. 69-73. *E*

Y. T. Yeow, M. R. Boudry, D. R. Lamb and S. D. Brotherton: Sources of errors in quasistatic capacitance-voltage determination of the interface state density distribution in the MOS system.
J. Physics D **10,** 83-95, 1977. *R*

E. Zieler (C. H. F. Müller, Hamburg): Possibilities and limits of industrial radiography and radioscopy.
Acta Electronica **20,** 11-24, 1977.

AUTHOR INDEX

Author index