hilips Journal of Research

Vol. 39 Nos 1/2 1984

HILIPS

IPR2018-1557
HTC EX1023, Page 1

Philips Journal of Research

Philips Journal of Research, published by the 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 39 is Dfl. 75, 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.

Editorial Board

A. van Oostrom (General Editor) Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Y. Genin, Brussels, Belgium
R. Kersten, Aachen, Germany
R. Memming, Hamburg, Germany
R. F. Milsom, Redhill, Surrey, U.K.
W. A. Smith, Briarcliff Manor, N.Y., U.S.A.
W. T. Stacy, Sunnyvale, CA., U.S.A.
J. B. Theeten, Limeil-Brévannes, France

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

© Philips International B.V., Eindhoven, The Netherlands, 1984. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

IPR2018-1557
HTC EX1023, Page 2

N.V. PHILIPS' GLOEILAMPENFAB.

Philips Journal of Research

Volume 39

1984

© Philips International B.V., Eindhoven, The Netherlands, 1984.
Articles or illustrations reproduced, in whole or in part, must be accompanied by full acknowledgement of the source: Philips Journal of Research.

IPR2018-1557
HTC EX1023, Page 3

CONTENTS

PHILIPS JOURNAL OF RESEARCH, VOL. 39

IPR2018-1557
HTC EX1023, Page 4

IPR2018-1557
HTC EX1023, Page 5

Philips J. Res. **39**, *1-10, 1984* *R 1077*

A COHERENT DIGITAL DEMODULATOR FOR MINIMUM SHIFT KEY AND RELATED MODULATION SCHEMES

by R. J. MURRAY and R. W. GIBSON

Philips Research Laboratories, Redhill, Surrey, RH1 5HA, U.K.

Abstract

We show that in a coherent digital demodulator both the clock and carrier signals can be recovered from the hard-limited outputs of the two quadrature channels. Clock and carrier are recovered simultaneously thus permitting fast-acquisition direct-conversion radio receivers.
MSK, TFM, GMSK and similar schemes can be demodulated. Measured Bit–Error–Rates for TFM were within 0.5 dB of that obtained with a reference clock and carrier. Acquisition within 30 bits was achieved with a degradation in performance of less than 1 dB: shorter acquisition times are possible with some further loss of performance.

EECS numbers: 61, 64.

1. Introduction

The possibility of fully integrated radio receivers has renewed interest in direct demodulation (zero i.f.) techniques. At the same time applications are beginning to arise for receivers which are required to handle digital signals only. One class of digital signal particularly suited to radio is Minimum Shift Key (MSK) and its derivatives Tamed Frequency Modulation (TFM) and Gaussian Minimum Shift Key (GMSK)[1,2]). It may be observed that the coherent demodulator for this class of signals is almost identical in layout to a direct demodulation radio receiver. Hence the possibility arises of combining the two functions.

Figure 1 shows the form of a combined receiver and demodulator. The modulated r.f. signal is shifted down to baseband by a pair of quadrature mixers and then lowpass filtered. Since the lowpass filters provide the selectivity of the receiver it follows that both carrier and clock recovery must be performed *downstream* of these filters. Thus we have to recover both clock and carrier from the baseband signal; this is different from many versions of the coherent demodulator which recover the carrier by a phase locked loop at r.f. (or at a conventional non-zero i.f.). Also, since MSK is a constant envelope signal it is advantageous to amplify the filter outputs in hard limiting amplifiers, thus avoiding problems of a.g.c.

IPR2018-1557
HTC EX1023, Page 6

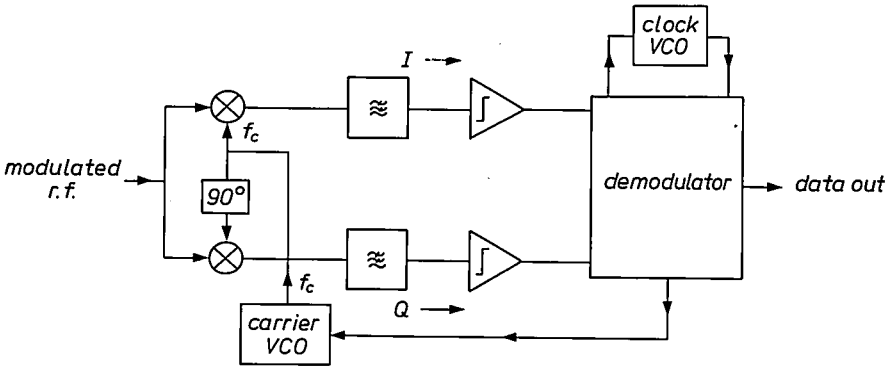


Fig. 1. Combined receiver and demodulator.

We will now proceed to show, both by theory and experimental results, that it is possible to recover both the clock and carrier signals solely from the hard limited baseband signals. Moreover the technique employed allows simultaneous acquisition of the clock and carrier and is thus suitable for systems requiring fast acquisition.

2. Principles of the clock and carrier recovery technique

Since we propose to recover both the clock and carrier solely from the hard limited outputs of the two quadrature channels the only information available to us is contained in the timing of the zero crossings. The technique will be explained with reference to fig. 2 in which the phase trajectories shown represent either MSK with no bandwidth restrictions (the sharp angles), or TFM, GMSK, etc. (the smoothed curve). The curve shows the phase of the modulated r.f. signal relative to its carrier. But the phase of the signal is transferred directly through the two quadrature mixers, subject only to two constant shifts separated by 90 degrees. We can therefore label the phase axis of the diagram with the equivalent phases at the outputs of the I and Q channels.

The zero crossings in the I and Q channels then correspond to the trajectory crossing the horizontal lines which are spaced at 90 degrees of phase. The *solid* lines represent the zero and 180 degree phases in the Q channel and the dashed lines the zero and 180 degree phases in the I channel. The *solid* vertical lines represent the instants at which the Q channel is expected to have a transition and the I channel to be at the centre of its eye. The dashed vertical lines correspond to the I channel transitions and the eye centres of the Q channel.

IPR2018-1557
HTC EX1023, Page 7

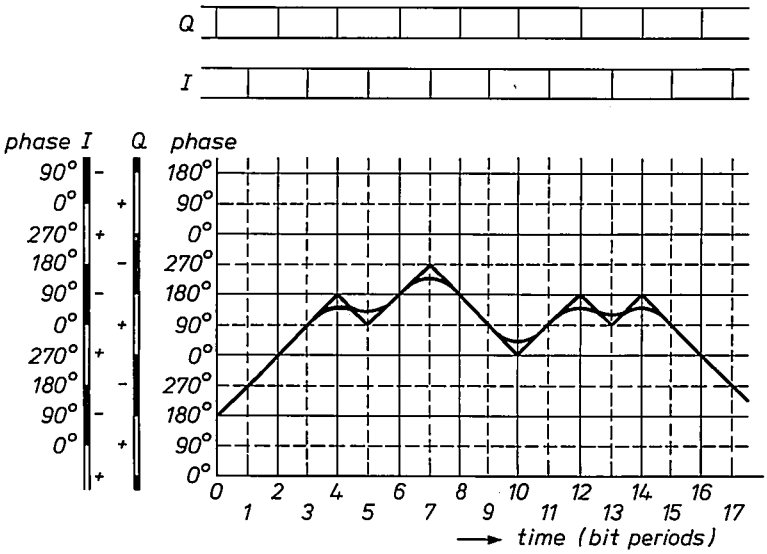


Fig. 2. Phase trajectories.

On this diagram an error in the carrier phase shows as a vertical shift of the trajectory. Similarly, an error in the clock phase shows as a horizontal shift. Figure 3 shows errors in both carrier and clock phase; the clock is running early by x and the carrier oscillator phase is early by y.

The error observed in the times of the zero crossings depends on the slope of the trajectory, i.e. on whether the instantaneous frequency is higher or lower than the carrier.

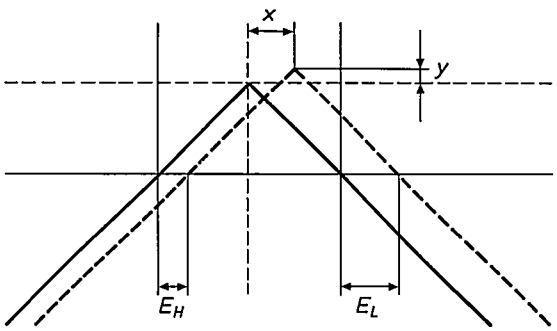


Fig. 3. Combined effect of clock and carrier phase errors.

IPR2018-1557
HTC EX1023, Page 8

From the diagram

$$E_H = x - y$$

and

$$E_L = x + y,$$

whence

$$x = \left(\frac{E_L + E_H}{2}\right)$$

and

$$y = \left(\frac{E_L - E_H}{2}\right).$$

Therefore, in principle, on a noise-free, unlimited-bandwidth, MSK signal we would only need one high frequency transition and one low frequency transition to be able to set the phase of both the carrier and clock oscillators. What is more, this can be done in two simple steps as follows:

Assume errors x and y as above.

If the first transition is due to a low frequency, then we measure $E_L = x + y$. Apply a phase correction of $(x+y)/2$ to both clock and carrier oscillators, any subsequent low frequency transition sees zero error but the next high frequency transition sees:

$$E_H = \left(\frac{x-y}{2}\right) - \left(\frac{y-x}{2}\right) = x - y.$$

A correction of $(x-y)/2$ is applied to the clock and $-(x-y)/2$ applied to the carrier, giving zero residual error to both.

In practice the signal is always bandlimited and the phase trajectory is not a set of straight lines and sharp angles. The simple two step correction given above is therefore not practical. Nevertheless, if we apply partial corrections the process will converge on the desired position. The optimum strategy will depend on what compromise between error performance and acquisition time we wish to adopt and also on the relative stabilities of the carrier and clock oscillators.

The corrections to the phase of the clock and carrier oscillators can be derived from the following rules:

(a) CLOCK

If a transition is early, advance the clock.

If a transition is late, retard the clock.

(b) CARRIER

In this case we use the same rule as above *or its inverse* according to whether the transition was due to a high or a low frequency. The rule is thus:

IPR2018-1557
HTC EX1023, Page 9

transition	*frequency*	*action*
late	low	advance
late	high	retard
early	low	retard
early	high	advance

We can determine whether a transition corresponds to a high or low frequency by observing the sign of the other channel at the moment the transition occurred.

Since in many cases the carrier phase will need more frequent correction than the clock, we adopted the following strategy:

For each transition, detected in either the I or Q outputs, determine whether the transition is early or late, then:

(i) always adjust the carrier phase so as to reduce the error,

(ii) only adjust the clock phase if the transition corresponds to a frequency *different* from that of the previous transition (*different* meaning above the carrier as compared to below the carrier or vice versa).

This rule ensures that a relatively stable clock oscillator will not be unduly disturbed by carrier phase fluctuations. There is scope for a fuller investigation as to how necessary or advantageous this approach is.

3. Circuit realisation

The clock and carrier synchronisation rules described above may be implemented by means of simple digital circuits.

The synchronising circuits are fed with the hard limited I and Q signals and from these produce separate control signals for the carrier and clock oscillators. The control signals are positive or negative pulses superimposed on the d.c. levels which control the frequencies of the clock and carrier oscillators. The effect of the pulses is to momentarily speed up or slow down the oscillator and thus *nudge* the phase of the oscillator in the required direction. Two forms of demodulator were investigated; a *constant nudge* demodulator (the correction pulses having fixed duration of 1 bit period) and a *proportional nudge* demodulator (the pulse duration being directly proportional to how early or late the transitions occur).

4. Experimental results

A proportional nudge demodulator circuit was constructed using 4000B series CMOS digital integrated circuits.

For experimental purposes a TFM signal was used with a bit rate of 69.27 kbps and carrier frequency 138.5 kHz. The TFM modulator incor-

IPR2018-1557
HTC EX1023, Page 10

porated a ROM and D/A converter [3]). The data was differentially encoded by taking the exclusive–OR of the current data bit and the previous data bit from the differentially encoded output.

The demodulator clock and carrier VCO's were each implemented using the VCO of a 4046B. The sensitivities of the oscillators could be adjusted to permit phase corrections per bit period of more than 60 degrees for both clock and carrier. We can then define *loop gain* in terms of:

$$\text{loop gain} = \frac{\text{phase nudge applied to oscillator}}{\text{timing (phase) error measured}}.$$

The gains of the carrier and clock loops were individually adjustable.

The I and Q channel arm filters used were low pass 4th order Butterworth filters. It should be emphasised that these filters are non-optimum and so cause intersymbol interference and have an excessive noise bandwidth and cause degradation of the demodulator performance. Two filter bandwidths were used:

(1) 25 kHz, which was found to give best results for the steady state bit error rate measurements,
(2) 30 kHz, which was the narrowest bandwidth to give well defined cross-overs in the I and Q channel eye diagrams, which is best for fast acquisition.

4.1. *Steady state bit error rates*

Measurements of bit error rate as a function of S/N ratio were carried out for many different loop gains (fig. 4). The channel filters were set to 25 kHz (= 0.36 fb). Also shown for comparison is the equivalent curve using a perfect clock and carrier (taken directly from the modulator, suitably delayed).

With perfect clock and carrier the measured bit error rate curve is 2.6 dB below MSK optimum at an error rate of 10^{-2} and 4.5 dB below MSK optimum at an error rate of 10^{-3}. The MSK optimum curve includes the effect of differential encoding. Some of the degradation is due to the use of non-optimum arm filters.

The recovered clock and carrier measurements show that for loop gains of 0.11 for carrier and 0.028 for clock the measured bit error rate performance is only slightly degraded compared to the perfect clock and carrier measurement (approximately 0.5 dB). As the loop gains are increased the measured bit error rate curve moves further away from that of the perfect clock and carrier. This is as expected since small corrections in the presence of noise cause only slight jitter in the recovered clock and carrier and so few errors, whereas large corrections will cause more jitter and hence more errors.

IPR2018-1557
HTC EX1023, Page 11

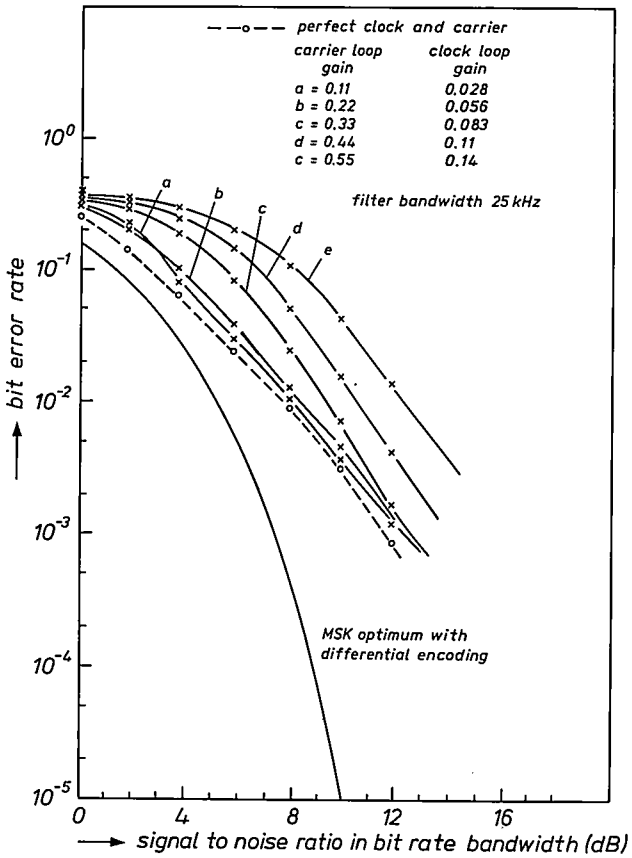


Fig. 4. Steady state BER curves. Filter bandwidth 25 kHz.

With a filter bandwidth of 30 kHz the steady state bit error rate performance is slightly degraded (less than 0.5 dB).

4.2. *Acquisition measurements*

It is difficult to precisely define the acquisition time of a demodulator, particularly in the presence of noise. We therefore measured the bit error rate of individual data bits in an acquisition pattern as a function of S/N ratio. The measurements were carried out for a number of combinations of clock and carrier loop gains. In all cases a bandwidth of 30 kHz was used for the I and Q channel arm filters. The acquisition pattern used was ....000111000111000.... after differential encoding. The appendix describes why this sequence was

IPR2018-1557
HTC EX1023, Page 12

chosen and also why the bit error rates for a *zero* are greater than for a *one*.

As an example of the measurements fig. 5 shows the measured bit error rates of the 16th bit (a one) for a variety of clock and carrier loop gains. As the loop gains are increased the degradation is initially reduced but eventually becomes more severe. The optimum loop gain is approximately 0.33 for carrier and 0.083 for clock, however the optimum is very broad and the case 0.22 for carrier and 0.11 for clock is also good. For the 30th and 31st bits similar behaviour is observed, in this case all of the curves are closer to the MSK optimum than for the 15th and 16th bits. Again a rather broad range of optimum loop gains is observed, the optimum being approximately 0.33 for the carrier and 0.083 for the clock. Figure 6 shows the measured error

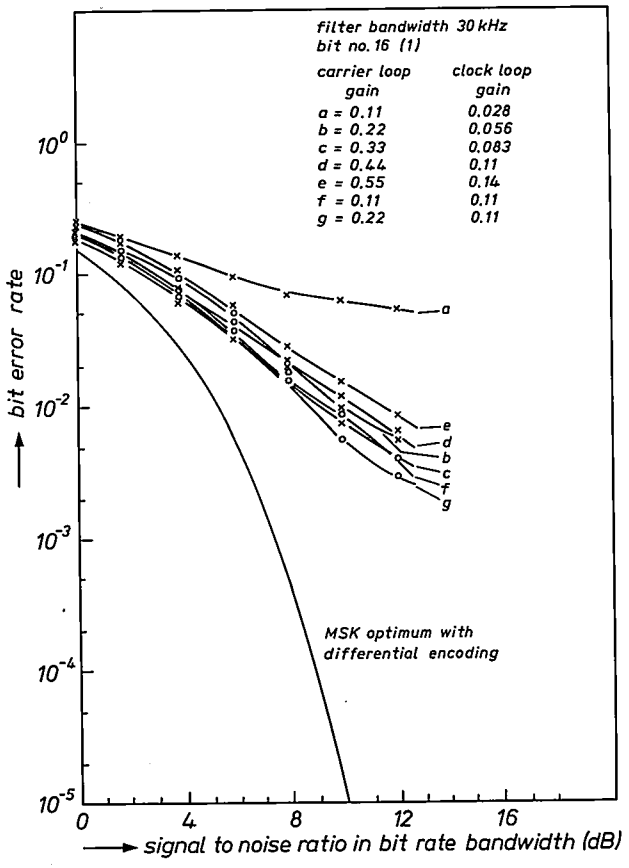


Fig. 5. Error rate of the 16th bit in the acquisition sequence.

IPR2018-1557
HTC EX1023, Page 13

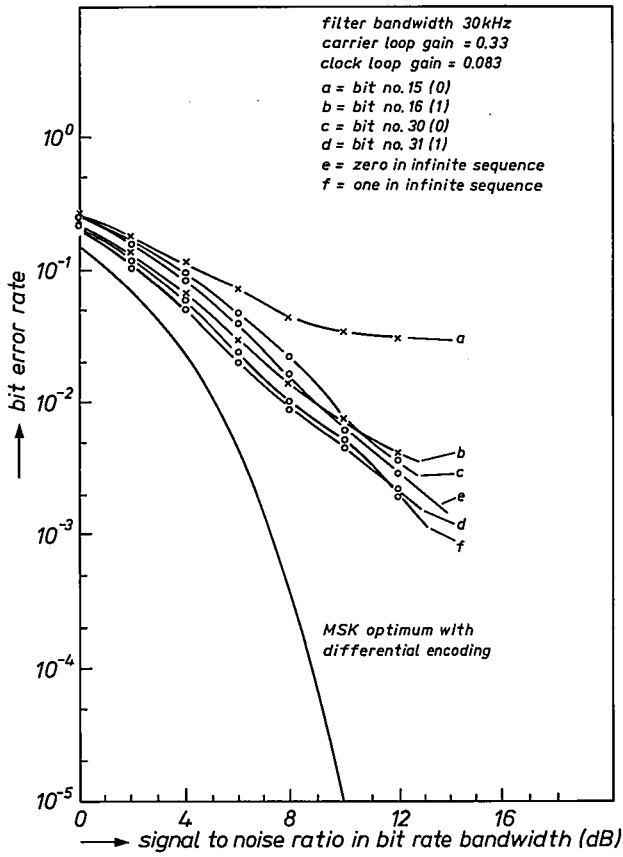


Fig. 6. Bit error rate for selected bits in acquisition sequence.

rates for the 15th, 16th, 30th and 31st bits for a carrier loop gain of 0.33 and a clock loop gain of 0.083.

5. Conclusion

We have described a coherent digital demodulator which extracts the necessary synchronisation information from the hard limited baseband signals. Experiments show that good steady state performance and moderately fast acquisition times (approximately 16–30 bits) can be achieved with relatively simple versions of the coherent demodulator. The theoretical limits of the technique have not been investigated.

Some modifications could be made to the demodulator. These include:

IPR2018-1557
HTC EX1023, Page 14

(a) the use of optimum filters in the I and Q channels,
(b) the use of a crystal derived clock, which could be adjusted by direct switching of the phase in a digital divider. This would allow virtually instantaneous phase correction.
(c) integration and filtering of the VCO control signals. This would convert the control loops from type 1 to type 2 or higher [4]).

APPENDIX

Acquisition patterns

The oscillator synchronisation information is derived from the crossovers of the hard limited baseband eye diagrams. All crossovers provide information which was used in adjusting the phase of the carrier oscillator but only some crossovers (denoted *different*) provide information to adjust the phase of the clock oscillator. For fast acquisition it is necessary to use an acquisition pattern with many transitions and preferably with mainly *different* transitions.

It should be noted that the sequence ....10101010101010.... (after differential encoding) is not always useful as an acquisition pattern with this type of demodulator. This is because the phase trajectory of the TFM or GMSK signal is substantially a constant 45 degrees and results in I and Q channel signals which have no transitions and so provide no synchronisation information.

Three fast-acquisition sequences have been considered, ...110011001100... (sequence 1), ...0001100011000... (sequence 2) and ...111000111000... (sequence 3). These sequences represent the data after differential encoding. For all three sequences the data eye opening corresponding to a *zero* is less than that corresponding to a *one*. So, for all three sequences the demodulator will, in the presence of noise, produce more errors for data zeros than for data ones. Sequence 1 does not give reliable fast acquisition because it allows an unstable false lock with no transitions in either channel. Sequences 2 and 3 do not permit a long term false lock and so give better acquisition. Experiment shows that sequence 3 results in better acquisition than either of sequences 1 or 2. For this reason sequence 3 was used for the detailed measurements described in sec. 4.

REFERENCES

[1]) F. de Jager and C. B. Dekker, IEEE Trans. Vol. **COM-26** (5), 534 (1978).
[2]) K. Murota and K. Hirade, IEEE Trans. Vol. **COM-29** (7), 1044 (1981).
[3]) K. S. Chung and L. E. Zegers, Philips J. Res. **37**, 165 (1982).
[4]) F. M. Gardiner, Phaselock Techniques, J. Wiley, New York 1979, Chapter 2.

IPR2018-1557
HTC EX1023, Page 15

Philips J. Res. **39**, *11-23, 1984* *R 1078*

IN QUEST OF THE SPIKE: ENERGY DEPENDENCE OF THE SPUTTERING YIELD OF ZINC BOMBARDED WITH NEON AND XENON IONS

by P. C. ZALM and L. J. BECKERS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Sputtering yields of Zn were determined for bombardment with Ne and Xe ions at normal incidence in the energy range 0.2–20 keV under ultrahigh vacuum conditions. Mass selection was employed and the energy spread was limited to a few electron volts. High fluences of the order of 3×10^{17} cm^{-2} were used to obtain stationary state yields. The yields were determined from accurate weighing of the samples prior to and after the irradiation stage. The measured yields obtained in low current density bombardments are in excellent agreement with predictions from Sigmund's linear cascade theory. There is no evidence of an anomalous, spike-like, cascade behaviour, in sharp contrast to what could be expected on the basis of experimental systematics or theoretical considerations. Yields measured for 20 keV Ne^+ bombardment of Zn at high flux show a sharp deviation from the Sigmund estimate. Tests have revealed this to be induced by beam heating of the target. The huge yields observed, however, cannot be attributed to a simple evaporation process but must be interpreted as tentative evidence of a so-called thermal spike.

PACS numbers: 79.20.Nc, 61.80.Jh, 34.90.+q.

1. Introduction, outline and theoretical preliminaries

Many observed regularities in the sputtering behaviour of amorphous elemental targets bombarded with atomic ions can satisfactorily be accounted for by the linear cascade theory as formulated by Sigmund [1]. In this model the penetrating projectile shares its energy with target atoms initially at rest in a series of binary collisions, a process in which fast recoils are created. These, in turn, set other target atoms in motion and a continuously increasing number of progressively slower atoms results until transferable energies are less than the displacement threshold. At this stage the elastic collision cascade is damped by energy dissipation through e.g. phonon-assisted processes. The sputtering yield for an ion with incident energy E_i is given by [1]

$$Y(E_i) = \Lambda \cdot F_D(E_i, x = 0), \tag{1}$$

where F_D stands for the amount of energy deposited at the surface ($x = 0$) in the form of target atom motion and Λ is a material parameter containing the

IPR2018-1557
HTC EX1023, Page 16

angle-, depth- and recoil-energy averaged, escape probability for target atoms. F_D is linearly proportional to the reduced nuclear stopping cross-section and Λ is inversely proportional to a surface escape barrier U_0 (usually taken as the sublimation energy).

The assumptions underlying the description leading to eq. (1) are known to break down in two cases:

i) for low-energy ($E_i \lesssim 1$ keV) light projectiles (e.g. H^+, He^+), where the energy transferred is insufficient to generate energetic recoils;
ii) for high-energy ($E_i \gtrsim 50$ keV) heavy ions, where the number of energetic recoils is so high that locally all atoms are instantaneously moving.

The latter situation, the so-called *spike* regime, has attracted considerable attention in the literature [2-9]). In particular the possible role of the target temperature (at $T \simeq \frac{3}{4} T_{melt}$) in invoking spikes [3]) is still very much a subject of debate [4,5]).

Nonlinear effects in heavy-ion sputtering have been studied extensively by Andersen and Bay [6]). They investigated in particular the projectile atomic number dependence ($Z_p \simeq 10 - 80$) of the sputtering yields of Si, Cu, Ag and Au targets bombarded with 45 keV ions. These authors also reported sputtering results with diatomic (homonuclear) molecular heavy ions at various energies.

A very systematic study of "collisionally induced spikes", involving both projectile atomic number and bombarding energy dependences, has been carried out by Thompson [7]). He collected sputtering yield data for Ag ($U_0 =$ 2.95 eV), Au (3.80) and Pt (5.85) bombarded with a variety of heavy ions ranging from P to Bi and clusters thereof at energies from 10 keV to 260 keV per atom. Some interesting regularities were observed. First, a sharp break away from the predicted eq. (1) was observed above a certain energy deposition density or critical yield. Above the critical yield the measured yields were found empirically to be more or less proportional to the third power of the energy deposition distribution function $F_D(E, 0)$. Secondly the critical yield, at which this anomalous behaviour was seen to occur, seemed to be linearly related to the sublimation energy of the target ($Y_{crit} \simeq 4.5\ U_0$ [eV]·atoms/ion).

In particular the latter conclusion, if not sheer coincidence or a measurement artefact, would allow for a guided search for spike phenomena at lower projectile energies. One only needs to select a target material with a low sublimation energy U_0, as this both enhances the expected yield according to eq. (1), since $\Lambda \propto 1/U_0$, *and* reduces the critical yield, since $Y_{crit} \propto U_0$. In principle frozen noble gases seem ideally suitable for such an experiment. Unfortunately there is a sharp distinction between the binding forces in metals and that in Van der Waals crystals, thus rendering any comparison dubious. A more feasible alternative for a target could be found in the group I (Na, K, ...)

IPR2018-1557
HTC EX1023, Page 17

and group II (Mg, Ca, ...) elements, but then the extreme chemical reactivity virtually prohibits their use as contamination problems will be severe and unavoidable. Yet another problem is that low-sublimation-energy materials and ultra high vacuum (UHV) conditions, required for good quality sputtering measurements, are hardly compatible. Moreover, one must be extremely careful to prevent thermal evaporation effects from obscuring the outcome of the experiment.

In this paper we present a dedicated search for spike effects in the bombardment of Zn ($U_0 = 1.35$ eV) with Ne^+ and Xe^+ ions at energies ranging from 0.2–20 keV. We shall see that zinc, when treated with caution, is a suitable target material for providing answers to some open questions in the assessment of phenomena accompanying or ascribed to spikes. Below we rehearse the arguments leading to the conclusion that zinc is a likely spike-prone candidate.

For numerical evaluation, eq. (1), in the case of perpendicular incidence, can be recast into the form [1])

$$Y(E_i) = C_{pt}\, S_n\,(E_i/E_{pt}), \tag{2}$$

where C_{pt} and E_{pt} are characteristic constants depending on projectile and target parameters (viz. atomic number, mass and U_0 in eV) given by the approximate value [10])

$$C_{pt} \simeq \frac{(Z_p\, Z_t)^{\frac{5}{6}}}{3\, U_0} \tag{3}$$

and the exact expression

$$E_{pt} = \frac{1}{32.5}\left(1 + \frac{M_p}{M_t}\right) Z_p\, Z_t\, (Z_p^{\frac{2}{3}} + Z_t^{\frac{2}{3}})^{\frac{1}{2}}\ \text{keV}. \tag{4}$$

The nuclear stopping cross-section $S_n(\varepsilon)$ is well approximated by [11])

$$S_n(\varepsilon) = \frac{\frac{1}{2}\ln(1 + \varepsilon)}{\varepsilon + \left(\frac{\varepsilon}{385}\right)^{\frac{3}{8}}}. \tag{5}$$

The sputtering yields for Zn predicted a priori from eqs (2–5) range from 3–10 and 3–29 atoms/ion for bombardment with Ne^+ or Xe^+ ions, respectively, in the energy range 0.2–20 keV. From the systematics observed by Thompson [7]) a nonlinear behaviour would be expected above a critical yield $Y_{crit} \simeq 6$. If this value indeed marked the onset of the above-mentioned third-power-of-F_D-like behaviour of the yield one could expect a yield of about 500–600 atoms/ion in the case of Zn bombarded with 20 keV Xe^+ ions, a huge effect

IPR2018-1557
HTC EX1023, Page 18

indeed. On the basis of theoretical considerations Sigmund and Claussen[2]) predicted that pronounced spike contributions can be expected only if the yield calculated through eq. (1) exceeds 10 atoms/ion.

Finally, Sigmund[9]) stressed the importance of two quantities in determining whether or not spike effects will occur and showed that two conditions have to be fulfilled simultaneously. The effective maximum energy density θ_0 in the central core of the spike must be larger than the sublimation energy U_0 and the time constant τ governing the quenching (exponential decay) of the spike must exceed τ_0, the slowing down time of the projectile. In fig. 1 a graphical representation of these requirements is given by depicting the energy dependence of θ_0 and τ for 1–100 keV Xe^+ bombardment of Zn (calculated with eqs (3), (4) and (9) of ref. 9, with $m = \frac{1}{3}$ applying in this case; as approximately[12]) $S_n(\varepsilon) \simeq \frac{5}{3}\sqrt{\varepsilon}$, the range $R(E_i) \propto \sqrt{E_i}$ and consequently $\tau_0 \propto R(E_i)/\sqrt{E_i}$ is roughly constant; a more elaborate estimate of τ_0 would show an increase from 2×10^{-13} sec to 3×10^{-13} sec over this energy range). It is clear from fig. 1

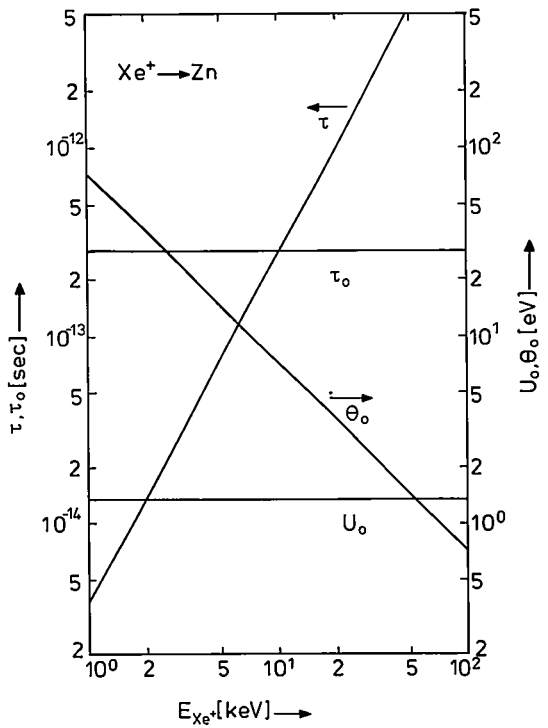


Fig. 1. Spike parameters for Xe^+ ions incident on Zn, estimated according to ref. 9, versus bombardment energy. θ_0 is the effective maximum energy density in the spike, τ the time constant for decay, and τ_0 the slowing down time of the projectile.

IPR2018-1557
HTC EX1023, Page 19

that spike effects are to be expected in the range of 10 keV ($\tau \geqslant \tau_0$) to 50 keV ($\theta_0 \leqslant U_0$). Similar calculations for Ne^+ bombardment of Zn reveal that "collisionally induced" spikes will certainly not be able to develop.

2. Experimental details

2.1. *Ion beam equipment characteristics*

Beams of single-charged, mass-selected, Ne^+ and Xe^+ ions at energies of 0.2–20 keV, with an energy spread limited to a few electron volts, were generated in our ion-beam apparatus discussed in detail elsewhere[13]). For convenience, some pertinent information will be given though.

The ion beam apparatus, suitable for perpendicular incidence only, operates in either of two modes, depending on whether bombarding energies from about 10–25 keV or below 2.5 keV are required. In the latter case the impinging ions, extracted at 15–20 keV, are decelerated at the very end of their trajectory by mounting the target in an electrostatic lens kept at a high positive potential. It was observed that by increasing the extraction (acceleration) voltage to 20 keV, rather than the 15 keV used earlier[10]), a laterally very uniform spot with a diameter of 10 mm resulted even with deceleration down to 200 eV. In this retardation mode the beam current, depending on the particular noble-gas ion species, ranged from 20–30 μA.

In the direct mode (no deceleration) the beam diameter was about 16 mm at currents of up to 60 μA. Such high currents were avoided, however, in order to prevent heating of the target, except in one particular series of experiments.

While in the retardation mode no secondary electrons will escape from the target, greatly facilitating total current integration, electron emission will take place in the direct mode. The target was therefore kept at a positive potential of 30–100 V to prevent electron liberation. Elsewhere[14]) we show that fairly accurate measurements of ion bombardment induced secondary electron emission can be made this way.

The background pressure prior to bombardment was always less than 2×10^{-7} Pa, the residual gas being hydrogen and water. During bombardment the pressure rose by at most two orders of magnitude, the residual-gas analyser spectra indicating only noble gas addition. In the course of the experiments, which started with the highest bombarding energies, the background pressure steadily improved due to the excellent gettering quality of the zinc sputter-deposited onto the wall of the vacuum chamber.

2.2. *Target preparation and yield determination*

Circular ($\varnothing = 18$ mm) targets were cut from 0.5 mm thick ultrapure (99.999%) rolled Zn platelets. The targets were etched in a HNO_3 solution,

IPR2018-1557
HTC EX1023, Page 20

rinsed in water and alcohol and finally baked at 400 K for 70 hr in a mild vacuum ($\sim 10^2$ Pa) environment to remove all residue. Subsequently the targets (on average 800 mg) were weighed with a sensitive Metler ME30 microbalance, with a claimed accuracy of 1 μg (and an observed 3 μg spread), and stored in mild vacuum for later usage.

It was decided to combine accurate weight loss and integrated current measurements to establish the yield. Trial dummy experiments revealed that a weight increase, saturating at about 10 μg, could occur during target transport, mounting and handling prior to and after the bombardment stage. Therefore it was deemed necessary to remove at least 0.5 mg of material via irradiation in order to obtain sufficiently reliable results. At the same time a total fluence of about 3×10^{17} cm^{-2} was considered necessary to ensure that the results reflected steady-state sputtering conditions and to guarantee that pronounced texturing (facet- and cone-generation) would not take place. Fortunately both requirements could simultaneously be fulfilled according to calculations with eqs (2–5), even in the absence of spike-induced yield enhancement. Except for the data points taken at 200 eV ion bombardment energy, always well over 1 mg of Zn was removed. Inspection of the samples irradiated with 10–20 keV ions with a scanning electron microscope revealed no significant cone formation.

A possible uptake of H_2O, CO_2, N_2 or O_2 from the air, which could be responsible for the observed $\lesssim 10$ μg weight increase, does not constitute a serious problem. Auger electron spectroscopy (AES) spectra, like those shown in fig. 2, taken in situ from samples after an ion fluence of the order of 10^{15} cm^{-2}, essentially showed a clean Zn target except for some trace contamination with oxygen (generally less than 5% of a monolayer after such low doses). Thus yields obtained with a high fluence will not be affected.

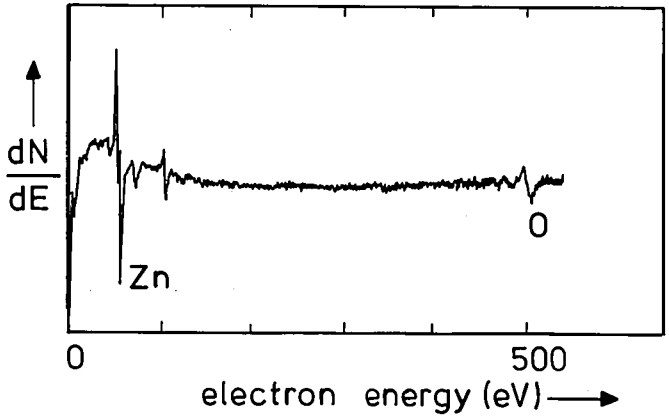


Fig. 2. Auger spectrum of a Zn target after mild sputter cleaning (0.2 keV Xe^+ ions, dose $\sim 10^{16}$ cm^{-2}). Trace oxygen contamination is visible.

IPR2018-1557
HTC EX1023, Page 21

3. Results and discussion

3.1. *Low-density data*

In order to avoid beam heating of the target the input power (beam current times projectile energy) was limited to a maximum of 0.25 W. Since the most interesting results were expected to occur with Xe^+ ions as a projectile (see sec. 1), the majority of the experiments involved Xe rather than Ne. The sputtering yields determined in this study for Zn, bombarded with Ne^+ and Xe^+ ions, are given in table I. The relative error is estimated to be 5%, but to this an absolute error in the yield of $\Delta Y = 0.2$ must be added statistically. This latter uncertainty is a direct consequence of the possible initial weight increment.

TABLE I

Measured sputtering yields in (atoms/ion) for Ne^+ and Xe^+ bombardment of Zn at normal incidence. An estimate of the experimental error is given in the main text.

E_i keV	0.2	0.5	1	2	10	15	20
Y_{Ne^+}	2.4			6.1	8.5		7.8
Y_{Xe^+}	2.7	5.3	8.0	10.2	23.5	25.5	27.9

In fig. 3 the yield data are plotted as a function of energy. Also shown are the curves predicted according to Sigmund's linear collision cascade theory, eqs (2, 4 and 5), with C_{pt} treated as an adjustable parameter to be fitted to the data. The reason for this is that there exists some (10–25%) ambiguity in the calculated values of C_{pt} [1,10,15]), which becomes more pronounced for lighter projectiles. In part this ambiguity arises from the fact that the experimental sputtering yield data refer to a steady-state saturation loading condition of the targets. For noble gas ions, with a maximum retention concentration of some 5%, the theoretically predicted energy dependence of the yield is hardly affected although there can be some influence on its total magnitude. Therefore a detailed comparison with theory can be made, in contrast to high-dose experiments with chemically active or metallic ions where trapping is a severe problem. Table 2 summarizes the parameters used for the curves plotted in fig. 3, along with some information regarding the quality of the fit.

It may be concluded from fig. 3 and table 2 that there is excellent agreement between linear cascade theory and experiment throughout the whole energy

IPR2018-1557
HTC EX1023, Page 22

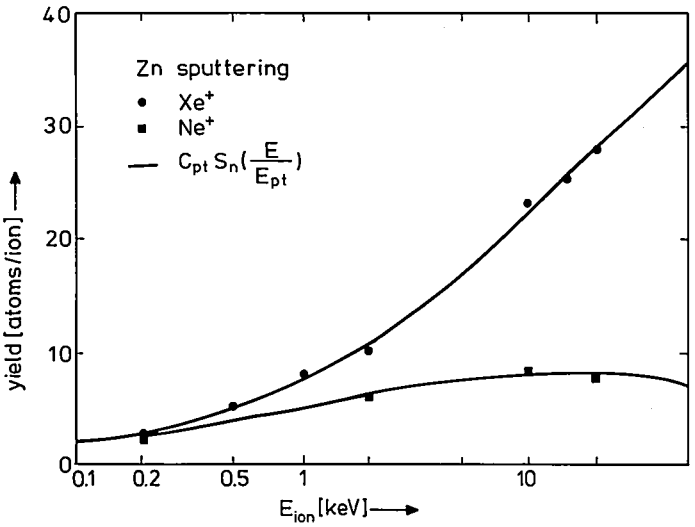


Fig. 3. Yield data for Ne^+ and Xe^+ bombardment of Zn as a function of energy. The estimates of Sigmund's linear cascade model $Y(E) = C_{pt}.S_n(E/E_{pt})$, where $S_n(\varepsilon)$ is the nuclear stopping cross-section taken from ref. 11 and E_{pt} (eq. 4) and C_{pt}, adjusted for optimal agreement, are characteristic constants depending on the projectile-target combination (see table II and the main text).

range investigated. The immediate and important conclusion is that the data deny the existence of any pronounced spike-effects, thereby refuting existing theories and observed regularities. A second interesting observation is that the theory apparently holds down to energies as low as 200 eV, a phenomenon observed earlier [10]) for noble-gas sputtering of Si and Mo. This could be somewhat surprising, as the validity of a transport approach (as used in the theory), which in the case of low energy cascades does not develop statistically, is

TABLE II

Parameters entering Sigmund's linear cascade model for the description of sputtering yield data for Ne^+ and Xe^+ bombardment of Zn. The root-mean-square deviation in the yield and the normalized χ^2 (with 3 and 6 degrees of freedom, respectively) are included as a measure of the quality of the agreement.

	E_{pt} [keV]	C_{pt} eq. (3)	C_{pt} fit	$\sqrt{\Delta Y^2}$	χ_n^2
Ne	46	28	23	0.3	1.0
Xe	733	117	115	0.5	0.9

IPR2018-1557
HTC EX1023, Page 23

questionable. Perhaps the breakdown of this assumption is not too important as its effects will, at least partially, be masked by another averaging mechanism, namely that experimentally determined yields are a product of many ($\sim 10^{17}$) individual cascades (see also ref. 16).

One note of caution seems appropriate at this stage. Formally the characteristic constant C_{pt} should contain a term that is weakly energy dependent. At higher energies, where the electronic stopping cross-section becomes comparable to the nuclear stopping contribution of the total energy deposition, this leads [17]) to a slow decrease of C_{pt}. By using C_{pt} as an adjustable parameter (although still keeping it within reasonable agreement with the value calculated through eq. 3) one could compensate for this effect while at the same time removing a possible discrepancy at lower energies. In our case the electronic stopping is negligible for Xe^+ ion bombardment of Zn (even at 20 keV it amounts to less than 7% of the total stopping, as calculated with the results from ref. 18), but not for Ne^+ ions (45% at 20 keV). Therefore only the Xe-on-Zn data fully validate the above remarks.

A few results on noble gas ion sputtering of Zn, under a variety of experimental conditions, have been reported earlier by several authors [19]). As in all cases (Ar^+: 1, 5, 12.5 keV; Kr^+: 39 keV, Xe^+: 45 keV) only a single ionic species at one bombardment energy was employed, no reliable systematics could be inferred from these data. It suffices to note here that the communicated yields agree to within some 15–20% with the predictions of eqs (2, 4 and 5) in accordance with our present observations.

3.2. *High-density data*

In an earlier series of measurements we studied the influence of the ion current density on the Zn sputtering yield when bombarded by 20 keV Ne^+ ions at beam currents of 10–60 μA. The irradiated target area was approximately 2 cm^2 (see also sec. 2.1). Visual inspection of the targets after exposure to the beam revealed texturing (macroscopic crystallite formation) above approximately 0.7 W input power. Temperature increase of the targets due to beam heating could only be measured directly with an optical pyrometer. Such an instrument, however, is notably inaccurate at temperatures below 600 K and hence could only provide reasonably reliable relative scaling. Thus an additional experiment for absolute calibration of the temperature was necessary.

The data obtained are depicted in fig. 4. A roughly exponential increase of the yield with ion flux was observed. The temperature of the target, as indicated by the pyrometer, is seen to increase rapidly with beam current but apparently levels off at higher input power at a saturation value of 470 K. In order to calibrate the temperature scale properly, we heated in situ a target

IPR2018-1557
HTC EX1023, Page 24

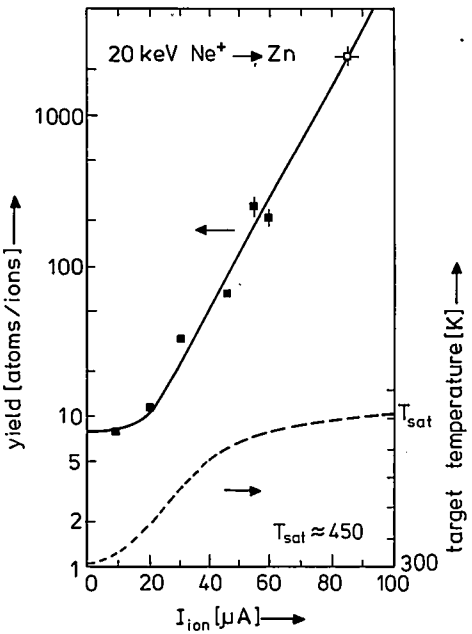


Fig. 4. Beam current dependence of the sputtering yield of Zn bombarded with 20 keV Ne^+ ions to a total dose of $3-4\times10^{17}$ (target area is 2.5 cm^2). Also included is the ion-beam-induced temperature increase of the targets as measured with an optical pyrometer. The trustworthiness of the temperature estimates is discussed in the main text.

mounted on an oven for about one hour to the same saturation value, read off the indication on the scale of the pyrometer, after short sputter-cleaning (1 keV Ne^+, dose 10^{15} cm^{-2}). A weight loss of about 1–2 μg/min was measured afterwards. For a clean target this corresponds to a temperature of 450 ± 15 K, in fair agreement with the pyrometer data. The background pressure during this test was sufficiently low (1×10^{-7} Pa) to ensure that contamination effects could not have had a serious influence on the rate of evaporation from the target (and hence on the temperature estimate).

Consequently, the dramatic yield increase with ion flux observed in fig. 4 cannot be explained with thermal evaporation effects alone. Furthermore the total fluence was always $3-4\times10^{17}$ (i.e. 17 μA·hr), hence the measurement time decreased with increasing beam current thus slightly reducing the evaporation contribution to the total yield. On the other hand, a current density increase alone cannot explain the observations. The data presented in sec. 3.1 show quite conclusively that local cascade predictions are obeyed for low ion current densities. Even at the maximum current density (corresponding to 2×10^{14} Ne^+ ions/cm^2 sec), however, there is virtually zero probability of cas-

IPR2018-1557
HTC EX1023, Page 25

cade overlap due to lack of correlation in space and time of ion arrival. Thus one is led to the conclusion that synergistic effects, of combined ion bombardment and temperature increase, could be responsible for the yield enhancement. To check this further, a Zn target was bombarded with a 12 μA 20 keV Ne^+ beam while being externally heated simultaneously to the saturation temperature. A huge yield of about 2500 atoms/ion was found. This point is indicated in fig. 4 with an open square, its position on the current axis being estimated by

$$I = 12\,\mu\text{A} + \frac{\text{applied power (1.5 W)}}{\text{acceleration voltage (20 kV)}}\,.$$

The huge yields themselves offer a nice, be it tentative, explanation for the observed "temperature saturation effect", as at such enormous particle ejection rates much of the input energy of the imping ion could leave the surface as kinetic energy carried away by the massive amount of sputtered atoms. For slow collisional, i.e. Sigmund-type sputtering the energy distribution of the sputtered atoms should have the form (see e.g. ref. 4) $\phi(E) \propto E/(E+U_0)^3$. If such a distribution were to hold for sputtered atoms with energies up to say 100 eV, one would expect the average energy $\langle E \rangle$ carried away by one sputtered atom to be about $\langle E \rangle \simeq 6U_0$. It is, however, highly probable that other processes, like e.g. prompt thermal sputtering (for a discussion see again ref. 4), contribute to the total yield as well. This would lower $\langle E \rangle$ substantially. Nevertheless we may expect input energy loss through sputtered atoms to be sizeable, the remainder perhaps being accounted for by conductive and radiative losses.

Synergistic effects of the type discussed above have been reported previously. In experiments where Ag, Au, Bi, Cu, Ge and Zr were bombarded with 45 keV Xe^+ ions, Nelson[3]) observed an anomalous increase in the sputtering yield if the target temperature exceeded about 75% of the melting temperature (for Zn $T_{melt} = 693$ K). Similar effects were reported for 5 keV Ar^+ bombardment of pyrolytic graphite at elevated temperatures[20]). For low-energy (17–180 eV) Ar^+ bombardment of solid and liquid In only a small (10%) difference in sputtering yield was observed[21]), indicating that the condition of the target alone is not sufficient to induce large abnormalities and that combined effects have to be assumed. Hofer and co-workers[5]) could not reproduce Nelson's findings for the particular case of Ag. Although still open to discussion, there has been a strong tendency to discard Nelson's data and ascribe his findings to thermal evaporation problems (see e.g. ref. 22). It is difficult, however, to envisage how thermal evaporation could have influenced his results for Bi, with its very low vapour pressure of about 10^{-8} Pa at $\frac{3}{4}\,T_{melt}$. Now our results for Zn too seem to confirm Nelson's earlier findings be it at a slightly lower temperature.

IPR2018-1557
HTC EX1023, Page 26

More elaborate experimental tests will be required to resolve existing discrepancies.

4. Conclusions

The low current-density sputtering yield data for Zn, bombarded with Ne^+ and Xe^+ ions in the energy range 0.2–20 keV, measured in this study are both qualitatively and quantitatively well described with Sigmund's linear collision cascade model [1]). No evidence has been found to support the existence of spike effects. This implies that the empirical rules for such phenomena which can be extracted from the systematics observed in Thompson's experiments [7]) cannot readily be extrapolated to other projectile/target combinations. Theoretical predictions [2,9]) for the onset of strongly nonlinear spike-like behaviour, which were tested against data for heavy ion projectiles on heavy atom targets, are not readily applicable to lighter mass targets. The fact that the most pronounced effects in Thompson's experimental work [7]) are encountered with polyatomic projectiles (clusters) could well point to another, as yet unknown, important mechanism governing spike effects.

The high current-density sputtering yield data, for Zn bombarded with 20 keV Ne^+ ions, presented in this paper indicate, although perhaps not entirely conclusively, that target temperature plays an important role in invoking spike effects, thereby confirming Nelson's original study [3]). From a theoretical point of view [22]), however, it is difficult to envisage how a small increase in vibrational amplitude of the target atoms, which would induce only minor changes in the physics of the collisional cascade accompanying the penetration of an energetic projectile, could have a major influence on the sputtering yield.

REFERENCES

[1]) P. Sigmund, Phys. Rev. **184**, 383 (1969).
[2]) P. Sigmund and C. Claussen, J. Appl. Phys. **52**, 990 (1981).
[3]) R. S. Nelson, Phil. Mag. **11**, 291 (1965).
[4]) R. Kelly, Surf. Sci. **90**, 280 (1979);
R. Kelly, Rad. Eff. **80**, 167 (1983).
[5]) K. Besocke, S. Berger, W. O. Hofer and U. Littmark, Rad. Eff. **66**, 35 (1982);
W. O. Hofer, K. Besocke and B. Stritzker, Appl. Phys. **A30**, 83 (1983).
[6]) H. H. Andersen and H. L. Bay, Rad. Eff. **13**, 67 (1972) and **19**, 139 (1973);
H. H. Andersen and H. L. Bay, J. Appl. Phys. **45**, 953 (1974) and **46**, 2416 (1975).
[7]) D. A. Thompson and S. S. Johar, Appl. Phys. Lett. **34**, 342 (1979);
S. S. Johar and D. A. Thompson, Surf. Sci. **90**, 319 (1979);
D. A. Thompson and S. S. Johar, Nucl. Instr. and Meth. **170**, 281 (1980);
D. A. Thompson, J. Appl. Phys. **52**, 982 (1981);
D. A. Thompson, Rad. Eff. **56**, 105 (1981).
[8]) V. I. Protasov and V. G. Chudinov, Rad. Eff. **66**, 1 (1982);
M. Szymónski and A. Poradzisz, Appl. Phys. **A28**, 175 (1975).
[9]) P. Sigmund, Appl. Phys. Lett. **25**, 169 (1974); ibid. **27**, 52 (1975).
[10]) P. C. Zalm, J. Appl. Phys. **54**, 2660 (1983).
[11]) W. D. Wilson, L. G. Haggmark and J. P. Biersack, Phys. Rev. **B15**, 2458 (1977).

IPR2018-1557
HTC EX1023, Page 27

[12]) J. P. Biersack and J. F. Ziegler, in Ion Implantation Techniques, edited by H. Ryssel and H. Glawischnig (Springer, Berlin, 1982), p. 157.
[13]) G. E. Thomas, L. J. Beckers, J. J. Vrakking and B. R. de Koning, J. Cryst. Growth **56**, 557 (1982).
[14]) P. C. Zalm and L. J. Beckers, to be published.
[15]) P. Blank and K. Wittmaack, J. Appl. Phys. **50**, 1519 (1979).
[16]) P. Sigmund, in: Proc. IV Int. Conf. on SIMS, Osaka, Japan (1983), Springer Series in Chem. Phys., to be published.
[17]) P. Sigmund, in Sputtering by Particle Bombardment I, edited by R. Behrisch (Springer, Berlin, 1981), p. 9.
[18]) D. J. Land and J. G. Brennan, At. Data and Nucl. Data Tables **22**, 236 (1978).
[19]) O. Almén and G. Bruce, Nucl. Instr. and Meth. **11**, 257 (1961);
H. Schirrwitz, Beitrage aus der Plasmaphysik **2**, 188 (1962);
A. Benninghoven, Z. Angew. Physik **27**, 51 (1961);
H. J. Smith, Rad. Eff. **18**, 55 (1973);
K. H. Krebs, Atomic and Molecular Data for Fusion, IAEA, Wien (1977), p. 185.
[20]) V. Philipps, K. Flaskamp and E. Vietzke, J. Nucl. Mat. **111** and **112**, 781 (1982).
[21]) B. L. Hurst and C. B. Cooper, J. Appl. Phys. **53**, 6372 (1983).
[22]) P. Sigmund and M. Szymonski, Appl. Phys. **A33**, 141 (1984).

IPR2018-1557
HTC EX1023, Page 28

Philips J. Res. ***39***, *24-50, 1984* *R 1079*

COMPARISON BETWEEN ATTENUATION CORRECTION METHODS IN TRANSAXIAL SINGLE PHOTON EMISSION TOMOGRAPHY

by J. PERGRALE, C. BERCHE, D. IACHETTI and G. NORMAND

Laboratoires d'Electronique et de Physique Appliquée, 94450 Limeil Brévannes, France

Abstract

After a short recall of the main used or known techniques of image reconstruction and of γ-rays attenuation correction in transaxial single photon emission tomography, a comparison is done between all these attenuation corrections. A classification is derived from the ability to recover the true activity values of a numerical phantom and from their robustness towards some simulated experimental conditions.

PACS numbers: 07.85.

1. Introduction

Conventional gammagraphy techniques currently used in nuclear medicine, produce two-dimensional images of a three-dimensional radioactivity distribution projected on the detector plane which is, in most cases, the large scintillation crystal plate of a gamma-camera.

In scintigraphic images thus obtained, the information on the location in depth of the radioactive fixation and on the value of the locally fixed radioactivity concentration are practically lost: these images are the result of a weighed superimposition of the planes of the three-dimensional activity distribution, parallel to the detector plane. The weighing is due to the attenuation (absorption and diffusion) of the emitted gamma-rays by biological tissue.

Due to the increasing interest of the nuclear medicine for quantitative studies, various methods have been suggested in order to avoid these drawbacks; most of them are computer-assisted and include the analysis of multi-incidence scintigrams. Transaxial single-photon emission-tomography is one of these methods. It uses a general process similar in its principle to the one developed in *X*-ray computerized transaxial tomography (for C.T. Scanner): slices of activity distribution, perpendicular to a longitudinal axis of the patient under examination, are calculated from informations included in the multiincidence scintigrams. The juxtaposition of such adjacent slices allows to reach the three-dimensional radioactive distribution; the scintigrams are obtained by rotating a detector (usually a gamma-camera) around this patient.

IPR2018-1557
HTC EX1023, Page 29

Reconstructing a two-dimensional (2D) parameter distribution corresponding to one slice from its one-dimensional (1D) projections, is a well-known problem which has already found solutions commonly used in *X*-ray transaxial tomography. But in emission tomography, due to the attenuation of the γ-rays by biological tissue, the scintigrams are not true projections of the radioactive distribution and the rough application of the known reconstruction methods to the scintigrams does not lead to the true activity values.

The introduction of attenuation correction in the reconstruction process particularizes the emission tomography. Only approximative solutions have been proposed to get rid of the attenuation effects, a practical rigorous method would suppose that a γ-ray attenuation coefficients map of the body at the γ-ray energy of the isotope which is used is available.

The introduction of the attenuation correction can be done in three different ways:

— Modification of the rough projections deduced from scintigrams. The classical reconstruction methods can then be applied.
— Modification of the image reconstruction methods themselves.
— Modification of the results obtained by applying the classical reconstruction methods to the rough projections, combined to an iterative process.

A description of the various reconstruction processes used in transaxial emission tomography will be given here. These processes include then a reconstruction of an object from its projections associated to an attenuation correction and are mainly characterized by the kind of this latter. Comparison will be done between the results obtained in the different cases by using phantoms simulating cross sections of the human body with radioactive area and according to the following criteria: image quality, ability of giving quantitative results, robustness of the various methods relatively to the experimental conditions and mainly to the "depth effect" which is proper to the data acquisition in gammagraphy and which is not taken into account in the hypotheses of the various methods.

2. Reconstruction of a plane object from its projections

Let us consider a plane object defined by a parameter distribution $f(\varrho,\varphi)$ in polar coordinates. This object can be limited by a contour outside with the function $f(\varrho,\varphi)$ is nill. A projection $p(r,\theta)$ of the object at the θ-angle, r being given, is defined by:

$$p(r,\theta) = \int_{Y=-\infty}^{Y=+\infty} f(\varrho,\varphi)\,\mathrm{d}Y \quad \text{with} \quad r = \varrho\cos(\varphi-\theta) \quad \text{and} \quad Y = \varrho\sin(\varphi-\theta).$$

IPR2018-1557
HTC EX1023, Page 30

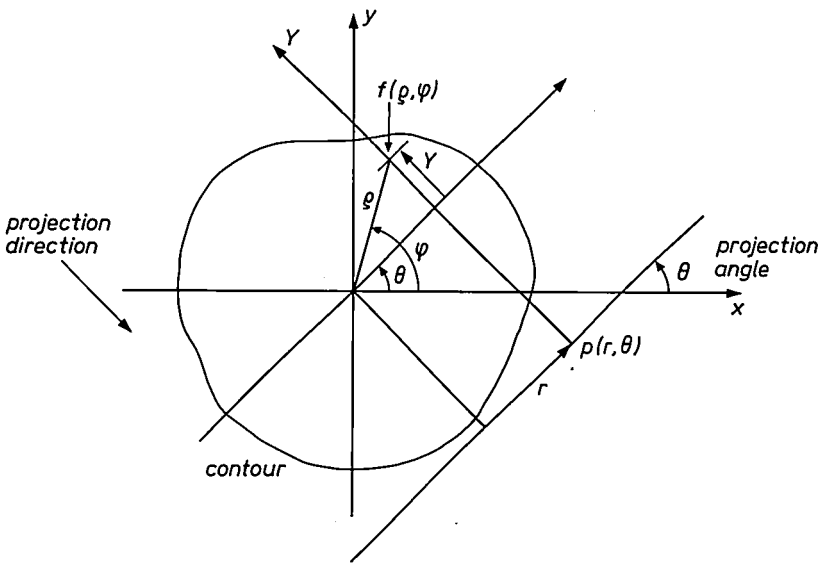


Fig. 1. Parameters used in the definition of a projection $p(r,\theta)$.

The transform $f(\varrho,\varphi) \rightarrow p(r,\theta)$ is called the Radon transform and the reconstruction of the $f(\varrho,\varphi)$ distribution knowing only its projections $p(r,\theta)$ is possible if a way of inverting practically this transform can be found.

Different ways to solve this problem have been proposed. Two of them are analytical ones, the filtered back-projection and the Fourier synthesis, the other ones lead to iterative algorithms. A brief recall of these techniques will be done hereafter.

2.1. *Filtered back-projection*

This technique has been proposed by Radon [1]), the basic principle is the following one: the image obtained by expanding the projections in the direction opposite to the one from which these projections have been acquired and by superimposing them, is related to the real object $f(\varrho,\varphi)$. It has been demonstrated that this image is the result of the convolution of the object $f(\varrho,\varphi)$ with the function $\left|\frac{1}{\varrho}\right|$ which acts as a blurring function. The restoration of the object consists then in deblurring this image by filtering it with a filter which compensates the effect of the convolution. The blurring effect corresponds to a filtering of the spatial frequencies repartition of the object with the function $\left|\frac{1}{R}\right|$ $\left(\text{the Fourier transform of } \left|\frac{1}{\varrho}\right|\right)$ in the spatial frequencies domain

IPR2018-1557
HTC EX1023, Page 31

(fig. 2*a*). The adequate compensating filter is then the function $|R|$ which is commonly called "ramp filter" (fig. 2*a*). In practical cases, the projections are sampled and digitized functions and Shannon's theorem allows in that case a limitation of the bandwidth of the useful spatial frequencies depending on the sampling step. For a sampling step a the maximum frequency available in the sampled projection is $\frac{1}{2a}$, called Nyquist frequency. So, practically the deblurring filter is the ramp filter multiplied by a rectangular window limited at the Nyquist frequency (fig. 2*c*).

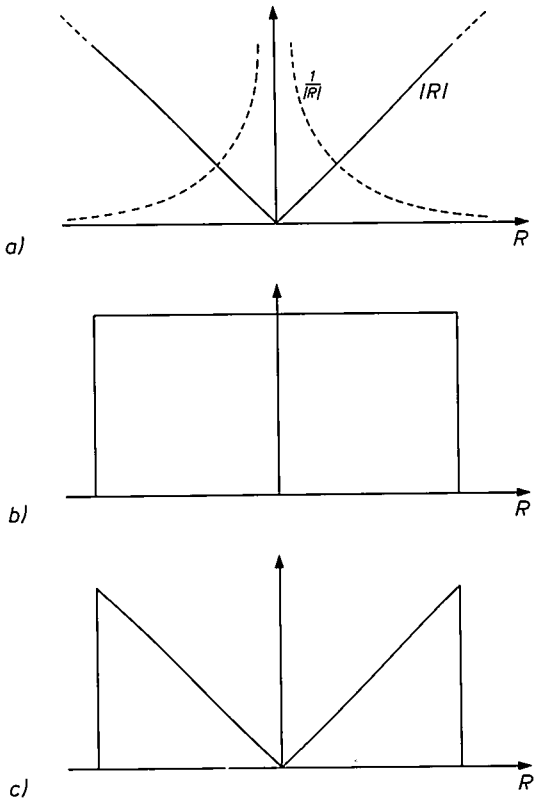


Fig. 2. *a*) Blurring function and compensating filter; *b*) Selecting window; *c*) Practical filter.

In fact, all the mathematical operations leading to the object $f(\varrho,\varphi)$ from the projections $p(r,\theta)$ are linear ones and it can be more convenient to filter first the projections (1D-filtering) and then to back-project the result.

It has to be noticed that the window associated to the ramp filter can be different from the rectangular one and its choice depends on the statistical

IPR2018-1557
HTC EX1023, Page 32

noise of the projections and on the a priori estimated spatial frequencies repartition of the object. The emission tomography software available for the present work includes 12 various filters, but in order to avoid any interference between the filter effect and the studied attenuation correction effect on the object reconstruction, only the ramp filter associated with a rectangular window at the Nyquist frequency has been used.

The back-projection technique is the most commonly used in transaxial tomography due to its implementation simplicity in the real domain.

2.2. *Fourier synthesis*

The Fourier synthesis is based on the "Central slice theorem": the 2D-Fourier transform of the object, along a radial line at the θ-angle in the Fourier domain, is equal to the 1D-Fourier transform of the projection of this object obtained at the same angle (see fig. 3).

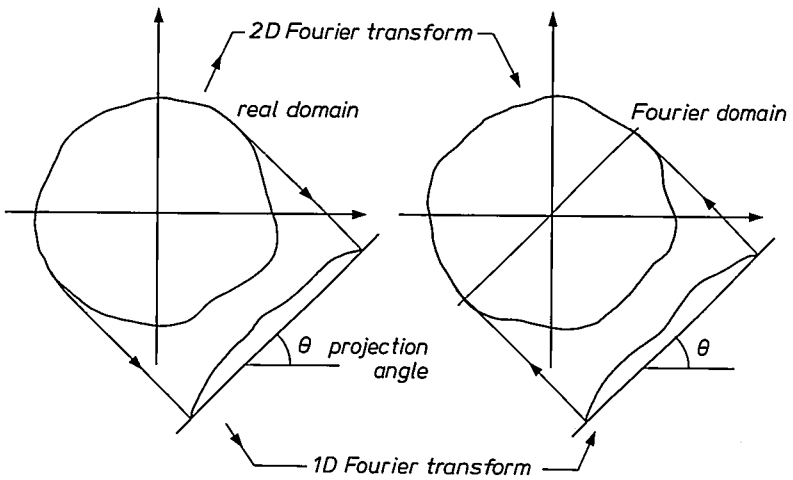


Fig. 3. "Central slice theorem" principle.

The principle of the reconstruction is the following one: the 1D-Fourier transforms of the projections are radially placed in a 2D-axis system, each of them at an angle equal to the angle of the corresponding projection. According to the central slice theorems, the so built 2D-function is the Fourier transform of the original object. By calculating the 2D-inverse Fourier transform, the object can be obtained. Practically, the algorithm used for calculating the Fourier transform are numerical ones and can be only applied on numerical functions defined in a cartesian coordinates system, so interpolation calcula-

IPR2018-1557
HTC EX1023, Page 33

tions are necessary in order to replace the polar coordinates system in which the Fourier transform of the object is built, by a cartesian coordinates system.

This reconstruction technique needs a lot of Fourier transform calculations and was until now not too much used. But, due to the increasing use of fast processors (Array processors) in tomographic equipments, the Fourier synthesis is now more and more employed. The final results are comparable to the ones obtained with the filtered back-projection technique.

2.3. *Iterative methods*

The general principle common to all the iterative reconstruction methods is to modify inside iteration loops an estimated object in order to make the projections calculated from this estimated object as close as possible to the measured ones. The estimated object is equal to the real one when all the projections calculated from it are equal to the measured ones.

As an example the so-called ART (Algebraic Reconstruction Technique) is very briefly described hereafter (see fig. 4): let us consider an estimated object supposed constant inside its contour. The projection calculated from this object at the θ-direction is different from the measured one at the same angle. The difference between the two projections is back-projected on the estimated object, in the same direction in such a way that, for this new estimated object, the calculated and measured projections at the θ-angle are equal. The same process is applied to the new estimated object for the following angle and so on.

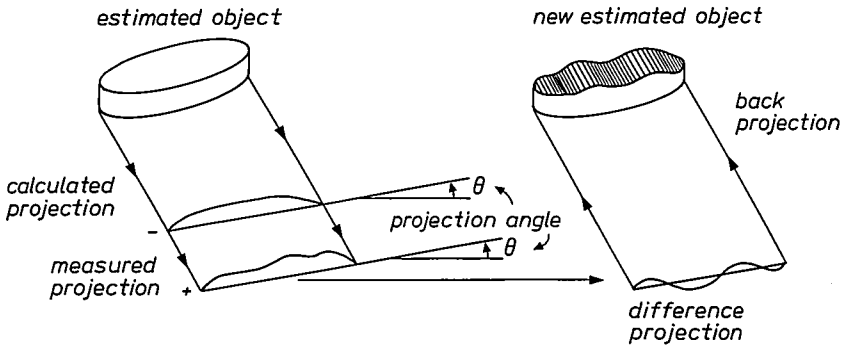


Fig. 4. ART principle.

The difference between the various iterative methods is mainly due to the way of modifying the estimated object and to the way of doing the first estimation and on the convergence criteria.

IPR2018-1557
HTC EX1023, Page 34

The main difficulties occuring in the iterative methods are to know if the used process is converging, if it is converging in all parts of the object, and in that case, when it can be stopped and according to what criteria. Some methods are very sensitive to the statistical noise, all are time-consuming. They are not very much used except in particular attenuation correction as it will be seen later.

3. Attenuation correction

The attenuation of the γ-rays by the biological tissue modifies the expression of the Radon transform given in sec. 2. If $f(\varrho,\varphi)$ is the radioactive concentration repartition defined inside the contour ($f(\varrho,\varphi) \equiv 0$ outside) and if $\mu(\varrho,\varphi)$ is the linear attenuation coefficient repartition for the γ-rays emitted by the isotope which is used ($\mu(\varrho,\varphi) \equiv 0$ outside the contour), the projection $p'(r,\theta)$ taking into account the attenuation (in fact corresponding to the measured one) is defined by

$$p'(r,\theta) = \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp\left(-\int_{-\infty}^{+Y} \mu(\varrho',\varphi')\,\mathrm{d}Y'\right) \mathrm{d}Y,$$

see fig. 5.

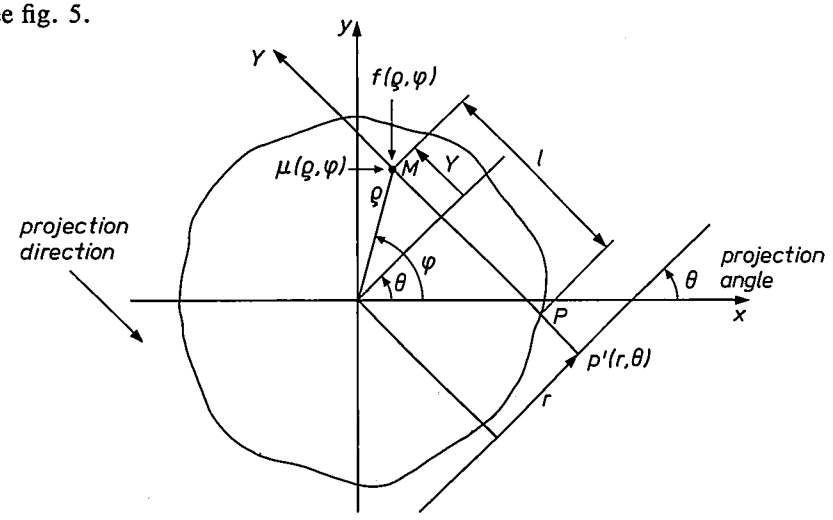


Fig. 5. Parameters used in the definition of an attenuated projection $p'(r,\theta)$.

The coefficient $\exp\left(-\int_{-\infty}^{+Y} \mu(\varrho',\varphi')\,\mathrm{d}Y'\right)$ is the probability for a γ quantum of not being absorbed or scattered along its trajectory between the considered emission point M and the point P of the contour in the projection direction.

IPR2018-1557
HTC EX1023, Page 35

The transform $f(\varrho,\varphi) \rightarrow p'(r,\theta)$ defines the attenuated Radon transform. The problem of its inversion has no analytical solution but iteratives solutions can be envisaged assuming that the $\mu(\varrho,\varphi)$ repartition is known. Practically and until now this last kind of solution is difficult to use because a preliminary examination has to be done on the patient in order to get on exact attenuation coefficient repartition (for instance by using transmission tomography with a monoenergetic source).

All the techniques studied here and used for getting rid of the attenuation effect assume that $\mu(\varrho,\varphi)$ is constant over all the object surface. In that case, $p'(r,\theta)$ becomes $p'(r,\theta) = \int_{-\infty}^{+\infty} \mu(\varrho,\varphi) \exp(-\mu l)\, dY$. Some of these techniques use other assumptions in order to simplify the calculations.

They can be classified in three parts: preprocessing techniques, intrinsic techniques and iterative techniques. The general principle of each studied technique will be briefly described hereafter.

3.1. *Preprocessing techniques*

In these methods the attenuation correction is realized by modifying the rough projections. The obtained modified projections are supposed to be the projections that would be obtained without attenuation on the object. The classical reconstruction methods like the filtered back-projection can then be applied.

3.1.1. Techniques using the geometrical mean of two opposite projections

The theory is due to Sorenson[2]). The principle can be briefly explained in two steps:

— Assuming that there is only a punctual source of activity A (fig. 6) in the object and a constant linear attenuation coefficient μ, the two opposite attenuated projections are given by $p'(r,\theta) = A \exp(-\mu l)$ and $p'(r,\theta+\pi) = A \exp(-\mu l')$. The combination

$$\sqrt{p'(r,\theta)\, p'(r,\theta+\pi)} \exp\left(\frac{\mu L}{2}\right)$$

is found to be equal to the value of the projection at the same angle but without attenuation.

— This principle can be applied to an extended source (fig. 7). Let us consider two opposite projections the detection line of which crosses the active area in a unique part (not fractioned) of length FL (F being the active fraction

IPR2018-1557
HTC EX1023, Page 36

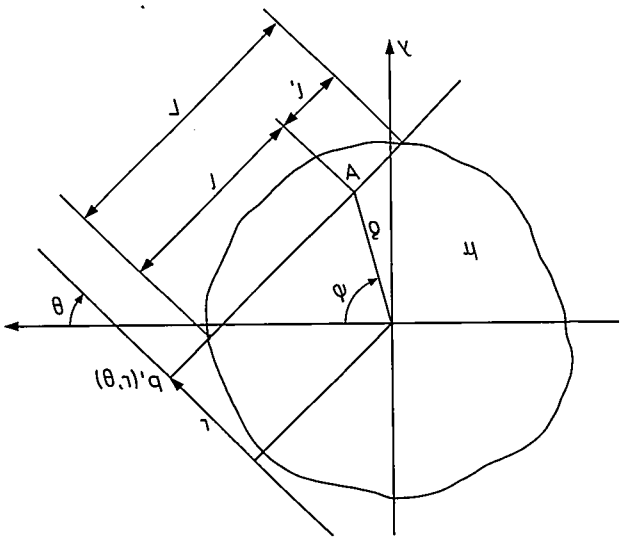


Fig. 6. Parameters used in the attenuation correction for one punctual source.

of L), part on which the activity concentration A is supposed to be constant. It can be demonstrated, always assuming μ = constant that the combination

$$\sqrt{p'(r,\theta) \cdot p'(r,\theta+\pi)}\ \frac{\frac{\mu FL}{2}\exp\left(\frac{\mu L}{2}\right)}{\sinh\left(\frac{\mu FL}{2}\right)}$$

is equal to FLA, the value of the projection at the same angle without attenuation.

Practically all the hypothesis are not satisfied but this projection combination is also applied to real cases. μ is estimated a priori (generally it is taken as being the linear attenuation coefficient of the water for the given γ-rays energy). L is deduced from the knowledge of the contour, which is either deduced from the measurement of the scattered γ-rays or approximated by an analytical curve like an ellipse. F is chosen a priori, having the same value for all the projections for all the angles. This last a priori choice is not so critical due to the fact that F is included in the function

$$\frac{\frac{\mu FL}{2}}{\sinh\left(\frac{\mu FL}{2}\right)}$$

which is slowly varying with F for current valves of μ and L. Some authors [3-6])

IPR2018-1557
HTC EX1023, Page 37

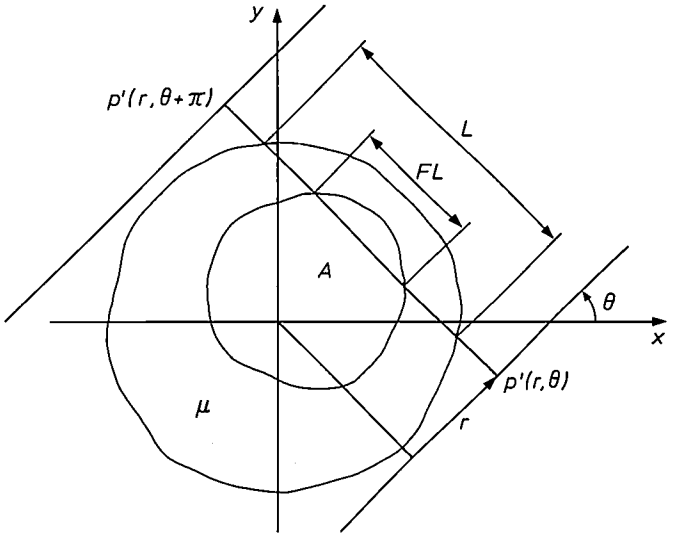


Fig. 7. Parameters used in the attenuation correction for a convex source with a constant activity A, obtained by calculating the geometrical mean of two opposite projections.

assume $F = 0.5$ or $F = 1$ for all the projections (the whole section is active) and the correcting coefficient

$$\frac{\frac{\mu FL}{2}}{\sinh\left(\frac{\mu FL}{2}\right)} \exp\left(\frac{\mu L}{2}\right)$$

is reduced in this last case to

$$\frac{\mu L}{1 - \exp(-\mu L)}.$$

Some authors assume the correcting coefficient is 1 and the present attenuation correction method is reduced in calculating new projections by doing only the geometrical mean of two opposite projections.

It has to be noticed that a good estimation of F can be done by doing a first reconstruction and by estimating from the so obtained image activity lengthes closer to the true ones. This process is time-consuming, because it needs two reconstructions, and then is not commonly used.

The present study includes three attenuation corrections techniques including the geometrical mean of two opposite projections:

— Simple geometrical mean.

— Geometrical mean and correcting coefficients calculated with $F = 0.5$.

— Geometrical mean and correcting coefficients calculated with $F = 1$.

IPR2018-1557
HTC EX1023, Page 38

3.1.2. Techniques using the arithmetical mean of two opposite projections

The hypotheses necessary for some of these methods are exactly the same as the ones used for the methods above described. The geometrical mean is replaced by the arithmetical mean and in this case, the correcting coefficient becomes

$$\frac{\frac{\mu FL}{2}\exp\left(\frac{\mu L}{2}\right)}{\sinh\left(\frac{\mu FL}{2}\right)\cosh(\mu Y_0)}$$

differing from the correcting factor used with the geometrical mean by the term $\cosh(\mu Y_0)$ at the denominator. Y_0 is the coordinate of the middle of the active part referred to the middle of the total attenuation part (see fig. 8).

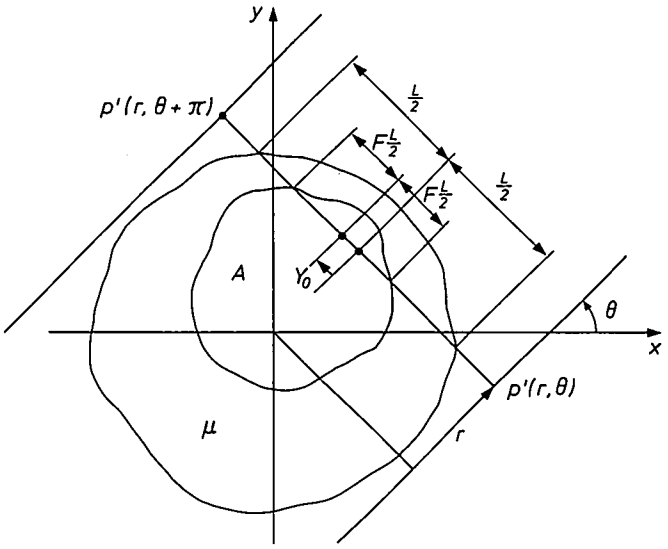


Fig. 8. Parameters used in the attenuation correction obtained by calculating the arithmetical mean of two opposite projections.

Y_0 has to be a priori estimated. Generally it is taken to be equal to zero (assuming that the active area has a center superposed to the center of the object). In that case $\cosh(\mu Y_0) = 1$ and the difference between this method and the previous one is only the replacement of the geometrical mean by the arithmetical one.

Three variants of the above described techniques including the arithmetical mean of two opposite projections have been tested in this study:

IPR2018-1557
HTC EX1023, Page 39

— Simple arithmetical mean.
— Arithmetical mean and correcting factors calculated with $F = 0.5$.
— Arithmetical mean and correcting factors calculated with $F = 1$.

Another approximate technique due to Kay[7]) including the arithmetical mean has also been tested. The principle is the following one:
If no hypothesis is done on the activity distribution and assuming μ is constant in the whole studied slice, an attenuated projection $p'(r,\theta)$ is given by

$$p'(r,\theta) = \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp(-\mu\, l)\, \mathrm{d}Y$$

along the Y axis, and

$$p'(r,\theta + \pi) = \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp(-\mu\, l')\, \mathrm{d}Y$$

(see fig. 6).
The arithmetical mean is then

$$\frac{p'(r,\theta) + p'(r,\theta + \pi)}{2} = \int_{-\infty}^{+\infty} f(\varrho,\varphi)\, \frac{\exp(-\mu\, l) + \exp(-\mu\, l')}{2}\, \mathrm{d}Y.$$

In this method, the main assumption in addition to μ constant is to say that the coefficient

$$C(l) = \frac{\exp(-\mu\, l) + \exp(-\mu\, l')}{2} = \frac{\exp(-\mu\, l) + \exp(-\mu(L - l)}{2}$$

is constant. The value of this constant is taken a priori to be equal to the mean value C_M between the maximum and the minimum of this function of l

$$C_M = \frac{1 + \exp(-\mu\, L) + 2 \exp\left(-\frac{\mu L}{2}\right)}{2},$$

in that case

$$\frac{p'(r,\theta) + p'(r,\theta + \pi)}{2} = C_M \int_{-\infty}^{+\infty} f(\varrho,\varphi)\, \mathrm{d}Y = C_M\, p(r,\theta)$$

and

$$p(r,\theta) = \frac{1}{C_M}\, \frac{p'(r,\theta) + p'(r,\theta + \pi)}{2}.$$

A classical reconstruction method like the filtered back-projection is then applied.

IPR2018-1557
HTC EX1023, Page 40

3.1.3. Bellini's technique

At the opposite of the other techniques, this technique due to Bellini [8]) needs only the hypothesis of a constant μ and is in this case a theoretically perfect attenuation correction method. It operates in the frequency domain by correcting the influence of the attenuation on the spatial frequency spectrum of the projections at a given angle. The principle is simple but the practical implementation needs some mathematical development, a summarized description of which is given thereafter.

PRINCIPLE: The starting point is the "Central slice theorem" (see 2.2).

In polar coordinates, the 1D-Fourier transform $TF_1[p'(r,\theta)]$ of the projection at the θ angle or the radial line of the object 2D-Fourier transform $TF_2[f(\varrho,\varphi)]$ at the same angle can be expressed by

$$F(R,\theta) = TF_1[p(r,\theta)] = TF_2[f(\varrho,\varphi)] =$$

$$\int_0^{\pi} \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp[-2i\pi\varrho R \cos(\varphi - \theta)]\, \varrho \, d\varrho \, d\varphi$$

(see fig. 9 for the variable definition).

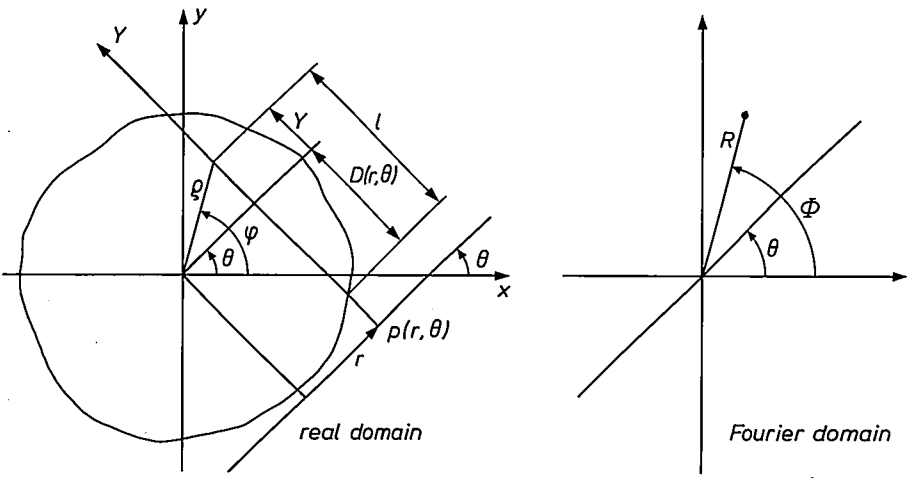


Fig. 9. Parameters used in Bellini's technique of attenuation correction.

The introduction of the attenuation modifies the expression of the projections which becomes

$$p'(r,\theta) = \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp(-\mu l)\, dY.$$

IPR2018-1557
HTC EX1023, Page 41

In order to simplify the ulterior calculations, this attenuated projection is multiplied by $\exp(\mu D)$ (see fig. 9) assuming that the body contour is known and becomes

$$p''(r,\theta) = \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp[-\mu \varrho \sin(\varrho - \theta)]\, dY.$$

The 1D-Fourier transform is depending on μ and becomes

$$G(R,\theta,\mu) = \int_0^{\pi} \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp\left[-2i\pi\varrho R\left(\cos(\varphi-\theta) - \frac{i\mu}{2\pi R}\sin(\varphi-\theta)\right)\right] \varrho\, d\varrho\, d\varphi,$$

which is not the 2D-Fourier transform of $f(\varrho,\varphi)$ because of the presence of the term $-\frac{i\mu}{2\pi R}\sin(\varphi-\theta)$ in the exponential argument.

The idea of Bellini is to do an adequate variable transformation on the Fourier variables (R,θ) in order to obtain again a 2D-Fourier transform of $f(\varrho,\varphi)$ and so to find $f(\varrho,\varphi)$, the unknown activity distribution, by including the imaginary term inside a new real term which eliminates this complex expression.

By identifying the above expression and the following one, that has to be obtained

$$F(\bar{R},\bar{\theta}) = \int_0^{\pi} \int_{-\infty}^{+\infty} f(\varrho,\varphi) \exp[-2i\pi \bar{R} \cos(\varphi - \theta)]\, \varrho\, d\varrho\, d\varphi,$$

and by assuming $\bar{\theta} = \theta - i\theta'$, we obtain easily

$$\bar{R} = \sqrt{R^2 + \frac{\mu^2}{4\pi^2}}$$

and

$$\bar{\theta} = \theta - i \operatorname{arcsinh}\left(\frac{\mu}{2\pi\bar{R}}\right)$$

with

$$\theta' = \operatorname{arcsinh}\left(\frac{\mu}{2\pi\bar{R}}\right).$$

The variable change $R \to \bar{R}$ is easily done by an interpolation. The change $\theta \to \bar{\theta}$ is more difficult to imagine but it can be realized after the following considerations:

— From the projections we obtain $G(R,\theta,\mu)$ but we wish to know $G(R,\theta,0)$ and we know that by replacing R and θ by $\bar{R}$ and $\bar{\theta}$ in the expression of $G(R,\theta,\mu)$, we get $G(\bar{R},\bar{\theta},\mu) = G(R,\theta,0)$ which is equal to $F(R,\theta)$.

— For R constant $G(\bar{R},\theta,\mu)$ is a periodic function $H(\theta)$ of θ having the period 2π. It possesses then a Fourier expansion

$$H(\theta) = \sum_{k=-\infty}^{+\infty} a_k \exp(i\,k\,\theta) \quad \text{with} \quad a_k = \frac{1}{2\pi}\int_0^{2\pi} H(\theta)\exp(i\,k\,\theta)\, d\theta,$$

in the same way $H(\bar{\theta})$ is also periodic because $\bar{\theta} = \theta - i\theta'$ with θ' constant for R constant so

IPR2018-1557
HTC EX1023, Page 42

$$H(\bar{\theta}) = \sum_{k=-\infty}^{+\infty} b_k \exp(\mathrm{i}\, k\, \theta),$$

but $H(\bar{\theta}) = H(\theta)$ for $\bar{\theta}$ real, so

$$b_k \equiv a_k \quad \text{and} \quad H(\bar{\theta}) = \sum_{k=-\infty}^{+\infty} a_k \exp(\mathrm{i}\, k\, \bar{\theta}),$$

then

$$H(\bar{\theta}) = \sum_{k=-\infty}^{+\infty} a_k \exp(-k\,\theta') \exp(\mathrm{i}\, k\, \theta) = \sum_{k=-\infty}^{+\infty} a'_k \exp(\mathrm{i}\, k\, \theta)$$

with

$$a'_k = a_k \exp(-k\theta').$$

In summary, knowing $G(\bar{R},\theta,\mu)$, its Fourier series development give the coefficients a_k. By multiplying them by $\exp(-k\,\theta')$, we obtain the coefficients a'_k of the series corresponding to $G(\bar{R},\bar{\theta},\mu)$ which allows to calculate the value of $G(\bar{R},\bar{\theta},\mu) = G(R,\theta,0)$. Practically, the coefficients a_k are calculated by doing the discrete Fourier transform of $G(\bar{R},\theta,\mu)$, the multiplication by $\exp(-k\,\theta')$ is in fact equivalent to a filtering, and the value of the new Fourier series $G(\bar{R},\bar{\theta},\mu)$ is obtained by calculating the reverse discrete Fourier transform of the filtered function.

In fact, in practical cases, the ratio between the maximum and minimum values of $\exp(-k\,\theta')$ can be very high, and it's preferable to work on the arithmetical mean of two opposite projections instead of working on single projection. It can be demonstrated that in this case the filtering is done with the function $\cosh(k\theta')$, more adapted to the numerical calculation than $\exp(-k\,\theta')$.

There is still a problem when values of $G(\bar{R},\bar{\theta},\mu)$ have to be calculated with $\bar{R}$ going to zero. It can be demonstrated that when $\bar{R} \to 0$ the function $G(\bar{R},\bar{\theta},\mu)$ tends towards

$$\frac{1}{2\pi} \int_0^{2\pi} G\left(\frac{\mu}{2\pi}, \theta, \mu\right) \mathrm{d}\theta$$

independent of θ.

After all these operations, $G(R,\theta,0)$ or $F(R,\theta)$ is known then, either a reverse 1D-Fourier transform is applied with R as variable giving so the corrected projections and the filtered back-projection can be applied in order to find the object, or the Fourier synthesis is applied by using all the $G(R,\theta,0)$ functions.

3.2. *Intrinsic techniques*

These techniques realize the attenuation correction during the process of image reconstruction itself. They lead to a weighed back-projection the compensating filter of which varies with the kind of weighing. Although two

IPR2018-1557
HTC EX1023, Page 43

methods have been studied in this work, only one will be reported here, the other one giving presently too much artefacts for being of any interest [9]).

3.2.1. Gustafson's technique

The principle [10]), based on intuitive considerations, does not bring a rigorous solution to the attenuation problem but gives rather good results as it will be shown later.

The arithmetical mean of two opposite projections with constant attenuation is expressed by

$$\frac{p'(r,\theta) + p'(r,\theta+\pi)}{2} = \int_{-\infty}^{+\infty} f(\varrho,\varphi)\left[\frac{\exp(-\mu\, l) + \exp(-\mu\, l')}{2}\right] \mathrm{d}Y$$

(see fig. 6 for parameters definition).

In order to compensate the effect of the coefficient

$$C(l,l') = \frac{\exp(-\mu\, l) + \exp(-\mu\, l')}{2},$$

during the reconstruction process, a weighed back-projection is realized on this arithmetical mean, the weighing factor being $1/C(l,l')$ for the corresponding point and the corresponding θ angle for each point (ϱ,φ) on the straight line defined by the equation $r = \varrho\cos(\theta - \varphi)$.

The filtering of the projections is still done with the standard ramp filter.

3.3. *Iterative techniques*

The general principle of these techniques is to modify during an iterative process, an estimate of the object in order to make the projections calculated from this estimated object (or reprojections) as close as possible to the measured ones.

3.3.1. Walters' technique

This technique follows rigorously the principle above described [11]):

a) From the measured projections a classical filtered back-projection is done giving the first estimate of the object $f_0(\varrho,\varphi)$.
b) This estimate is projected by simulating an acquisition process with a constant attenuation.
c) These calculated projections are substracted from the corresponding measured projections. Error projections $e(r,\theta)$ are then obtained.
d) The image reconstruction process is applied to these error projections leading to an error object $f_e(\varrho,\varphi)$.

IPR2018-1557
HTC EX1023, Page 44

e) The estimate $f_0(\varrho,\varphi)$ is corrected by the error object $f_e(\varrho,\varphi)$ in the following way,

$$f_1(\varrho,\varphi) = f_0(\varrho,\varphi) + \lambda f_e(\varrho,\varphi)$$

giving a new estimate $f_1(\varrho,\varphi)$ of the object. λ is a damping factor ($|\lambda| < 1$) limiting the divergence posibility of the process.

f) The process restarts to b), for an other iteration.

The main problem is to decide when the process has to be stopped. As some authors recommended and as it has been confirmed in our study, it is not useful to do more than one iteration, the improvement brought by additional iterations being not significant, our data being without noise (except the sampling noise) and with or without constant attenuation.

3.3.2. Chang's technique [12])

The general scheme of the Walters' technique is kept but in addition a first order attenuation correction is performed at the steps a) and d), the principle of which is the following one [7]).

Each point of the object $f(\varrho,\varphi)$ participates to the projection at the angle θ, weighed by the coefficient $\exp(-\mu l)$ (see fig. 5). If the object is reduced to just one point source, a back-projection gives at this point a reconstructed value equal to

$$A_r = \frac{1}{2\pi} \int_0^{2\pi} A \exp(-\mu l)\, d\theta,$$

A being the activity of the source and l being depending on θ,

$$A_r = A \frac{1}{2\pi} \int_0^{2\pi} \exp(-\mu l)\, d\theta.$$

An attenuation correction at this point can be done by dividing the reconstructed activity value by

$$\frac{1}{2\pi} \int_0^{2\pi} \exp(-\mu l)\, d\theta,$$

which is the mean value of global attenuation in all the directions for the considered point.

This principle is extended to a continuous source: each reconstructed activity of the points of the source is divided by the corresponding mean attenuation value. This technique needs then to built a correction matrix depending on the object contour. It has been concluded in our study, as some authors did, that the process can be stopped at the first estimate corrected at the first order [13–15]).

IPR2018-1557
HTC EX1023, Page 45

4. Quantitative results: Comparison *)

All the above presented attenuation correction techniques have been associated to the filtered back-projection method for the image reconstruction, and using the same ramp filter.

In the following parts they will be named by the corresponding author names when possible (Kay, Bellini, Gustafson, Walters, Chang) or by some abbreviation; GM, GM 0.5, GM 1 (for the simple geometrical mean, or the corrected geometrical mean with $F = 0.5$ or $F = 1$ (see part 3.1.1) and AM, AM 0.5, AM 1 (for the simple arithmetical mean or the corrected arithmetical mean with $F = 0.5$ and $F = 1$ (see part 3.1.2).

4.1. *Means and methods*

In order to separate the influence of the parameters which are implicated in the real clinical cases, and not taken into account in the hypothesis of each correction method, the comparison has been done on the reconstruction results obtained, from a simulated phantom: statistical noise and camera defects (inhomogeneity of the field of view) have not been taken into account, but the effects of a non uniform attenuating medium and of the variation of the spatial resolution in function of the depth of the source (depth effect due to the solid angle determined by the collimator geometry) have been studied. The way of evaluating the results has been chosen close to the one used in clinical environment and makes them by this fact directly interpretable for the same kind of clinical images.

4.4.1. Numerical phantom

On an IBM 4341 computer projections of a numerical phantom simulating a transverse section of the abdomen (liver and spleen) have been generated with an angular sampling of 128 projections for 360° and a lateral sampling of 128 elements, 3 mm wide each (field of view: 38.4 cm in diameter).

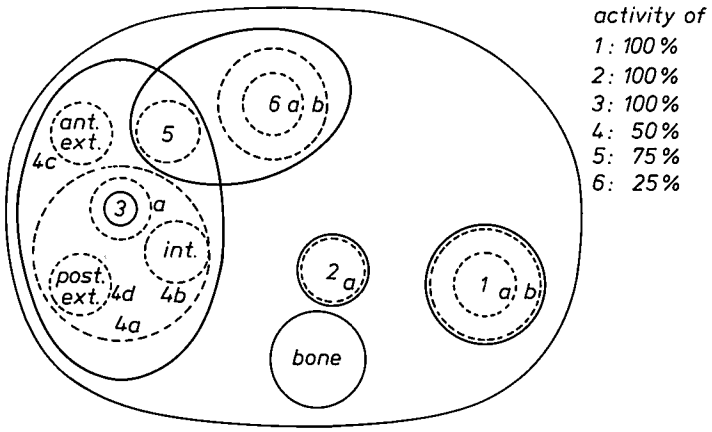
The phantom (fig. 10) includes 6 radioactive zones with various activity concentrations indicated on the scheme. It simulates an hepatosplenic disease.

The projections have been generated four times:

- With a constant auto-attenuation (the abdominal section has been modelled by an ellipse with a linear attenuation coefficient equal to 0.16 cm^{-1}.
- With a non constant auto-attenuation (simulation of the rachis by a 6 cm diameter cylinder with a linear attenuation coefficient equal to 0.27 cm^{-1}).
- Each of these two simulations have been made either without or with depth-effect.

*) This part has been reported at the IEEE Nuclear Science Symposium at San Francisco (October 19-21, 1983) and has been published in the corresponding "proceedings" [16].

IPR2018-1557
HTC EX1023, Page 46



	organ	diameter in mm of the regions of interest
1	spleen	1a 30 centered 1b 80 centered
2	vessel	2a 30 centered 2b max. of 2a
3	hyperfixation	3a 11 centered on 3 3b max. of 3a
4	right lobe	4a 111 4b 30 internal 4c 30 ant. ext. 4d 30 post. ext.
5	interlobar zone	5 30
6	left lobe	6a 30 6b 50

Fig. 10. Numerical phantom used for quantitative study.

For this last case, experimental point spread functions, measured with a Philips large-field scintillation camera provided with a standard high resolution collimator have been used: the functions have been obtained by using a point source located at various distances between 2 cm and 30 cm from the front of the collimator and the FWHM (Full Width at Half Maximum) of these point spread functions have been found varying respectively from 0.8 cm to 1.7 cm.

IPR2018-1557
HTC EX1023, Page 47

4.1.2. Estimates

The obtained images have been visualized on a SAVC*) Philips imaging console and the quantitative values have been estimated by using the software included in this console. A first qualitative image analysis gave the global characteristics of each technique and oriented us in the quantitative study.

The reconstruction artifacts (such as the ones due to the angular sampling) did not allow to use a pixel by pixel-quantitative estimation. We chose to estimate the value of the activity concentrations by taking the average of the values obtained in each circular region of interest (ROI) as shown in fig. 10. The distribution of these average values have been used for comparison. For the small zones (2 and 3) the maximum pixel values have also been considered.

Finally, 13 quantitative estimates for each image have been kept, 9 of them concerning ROI having a diameter strictly greater than 30 mm.

4.1.3. Statistical study

The estimates of different kinds (average or maximum values) led us to use non parametric techniques for the comparison of the results, such as the Wilcoxon test for paired series which does not need any hypothesis on the probability distributions under study. This test has mainly been used in order to show the influence of the rachis and of the depth effect in each attenuation correction method.

The average of all the estimates for each technique and the standard deviations of their distributions have also been calculated in order to give a global classification of the various methods.

4.2. *Results*

4.2.1. Qualitative analysis

The qualitative comparison of the 44 (11 correction techniques on the 4 variants of the phantom) reconstructed images allows to point out some characteristics common to the various correction techniques. They can be classified in two groups:

— Techniques with central under-compensation (see for example fig. 11*a*);
— Techniques with central over-compensation (see for example figs 11*b*, 11*c* and 11*d*).

The first group is related to the corrected arithmetical and geometrical means assuming the whole section is active (GM1, AM1) and to the Kay's and Walters' techniques. Kay's technique and the geometrical mean one show in addition anterior and posterior over-compensations. The effect of the rachis

*) SAVC: Stand Alone Viewing Console.

IPR2018-1557
HTC EX1023, Page 48

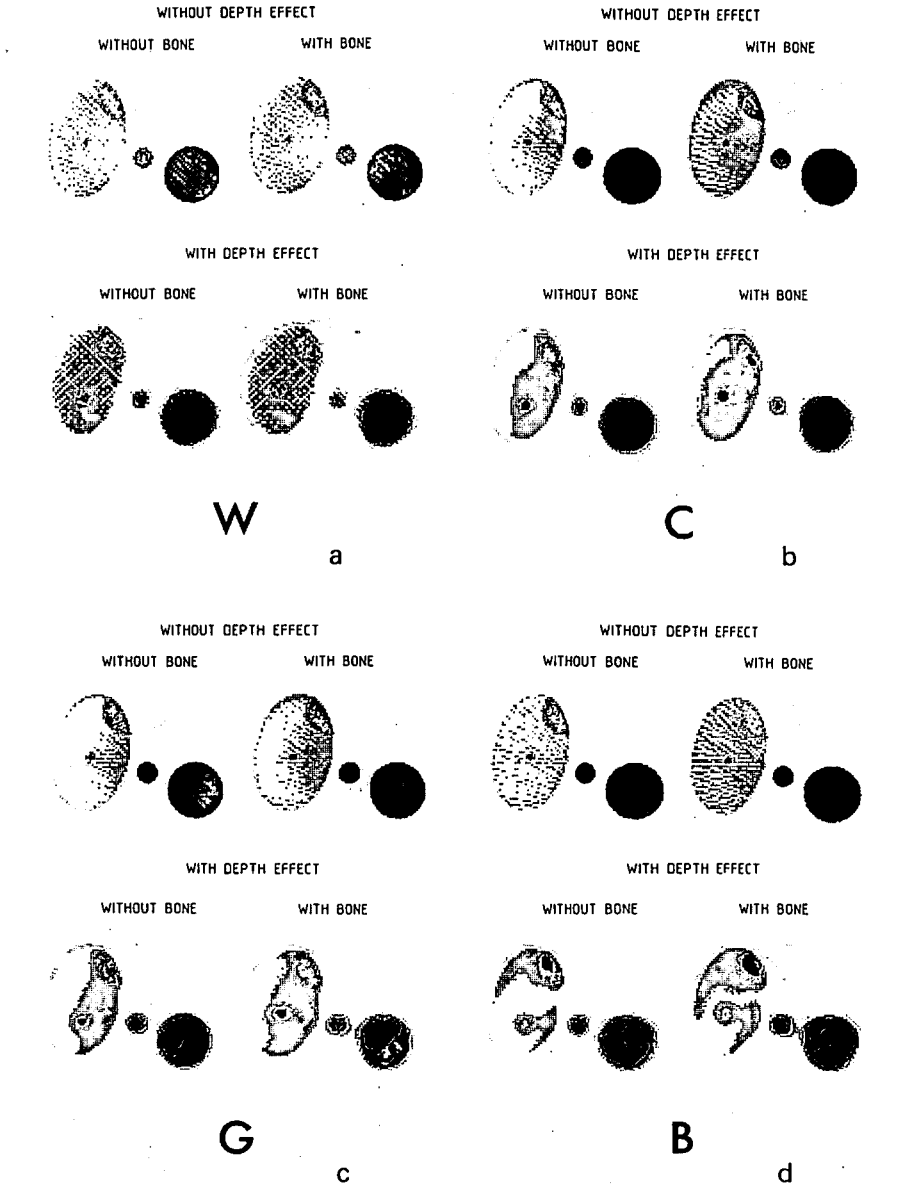


Fig. 11. Imaging results for 4 attenuation correction techniques: *a*) Walters; *b*) Chang; *c*) Gustafson; *d*) Bellini.

seems always practically negligible. Without depth effect, there is no image distorsion and in this case, Bellini's technique can only be considered as an exact one (fig. 11*d*). The taking into account of the depth effect implies two consequences:

IPR2018-1557
HTC EX1023, Page 49

— The parasitic ray structure due to the angular sampling disappears because of the spatial resolution loss, except for Walters' technique (fig. 11*a*).
— The images of the liver lobes are distorted except for the simple arithmetical mean technique, Chang's (fig. 11*b*), Kay's and Walters' techniques (fig. 11*a*).

In fact for all the techniques the depth effect seems to increase the defects already existing without it. In all the cases the ROI 3 (11 mm diameter) shows an underestimate of the activity concentration. This is probably due to the sampling and this led us to study the maxima for ROI having a diameter smaller than 30 mm (ROI 2 and 3).

4.2.2. Quantitative analysis: absolute results

The study of the estimates obtained from each technique (9 ROI × 4 variants of the phantom by excepting the ROI 2 and 3) leads us to suggest the following classification:

— Group 1: techniques giving estimates close to the true values (Bellini, Gustafson, Chang and Walters).
— Group 2: techniques giving an over-estimate of the true values (GM 0.5 and AM 0.5).
— Group 3: techniques giving an important under-estimate of the true values (simple arithmetical and geometrical mean).
— Group 4: techniques giving an intermediate under-estimate between the ones given by the groups 1 and 3: GM 1, AM 1, Kay).

These four groups can be represented in one drawing by reporting the average of all the estimates for each technique (fig. 12*a*). A priori, only the groups 1 and 2 seem clinically interesting for the direct evaluation of the activity concentrations. The study of the distributions of the estimates leads us to suggest another classification by using the standard deviations normalized to the mean of true values. Three main groups can be defined in this way (fig. 12*b*).

— Group A: techniques characterized by small standard deviations (about 10%) which mean a good reliability (Bellini, Gustafson and GM 0.5).
— Group C: techniques characterized by large standard deviations (more than 20%) (Kay, simple geometrical and arithmetical mean techniques).
— Group B: techniques characterized by intermediate standard deviations (the other ones).

Such classifications can be modified when considering the relative results and a technique which does not allow to reach the true activity concentrations could be satisfactory for the determination of the ratios between the various concentrations.

IPR2018-1557
HTC EX1023, Page 50

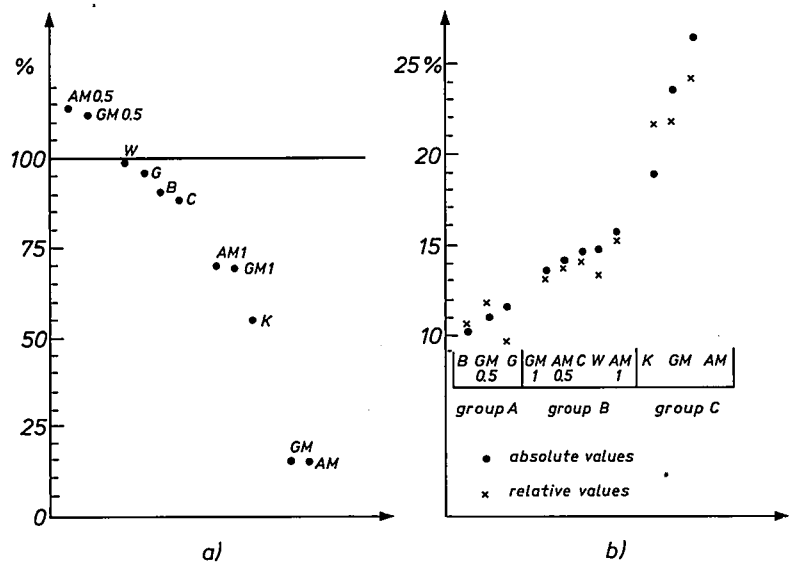


Fig. 12. *a*) Average values of all the estimates of the activity for each attenuation correction technique; *b*) normalized standard deviation of the distributions for each technique; AM: arithmetical mean; GM: geometrical mean; AM 1, AM 0.5, GM 1, GM 0.5: corrected arithmetical or geometrical mean: W: Walters; G: Gustafson; B: Bellini; C: Chang; K: Kay.

4.2.3. Quantitative analysis: relative results

The estimate of the activity concentration in the ROI 1*a* has been used as the normalization reference. We obtained in this way 11 relative estimated values (1*b* being excluded).

In that case the estimate of activity concentration values in ROI 4, 5 and 6 are close to the true relative values but more or less spread according to the technique which has been used. The estimates in ROI 2 and 3 are widely spread in all the cases because both averages and maxima have been used.

In the same way as for the absolute results and with the same limitations, we have calculated the relative standard deviations of the various estimates obtained for each technique. They have been found to be very close to the ones obtained for the absolute results (fig. 12*b*). The three previous groups remain identical. That confirms the interest for Bellini's and Gustafson's techniques.

It has to be noticed that the Sorenson's techniques (GM 1, GM 0.5, AM 1, AM 0.5) are strongly depending on the structure of the activity domain in the studied section. The results can then vary from one section to another one if the same hypothesis on the activity domain are kept. That is why we excluded them from the following study keeping only the techniques of Bellini, Gustafson, Chang and Walters.

IPR2018-1557
HTC EX1023, Page 51

4.2.4. Influence of the depth effect and of a non-homogeneous attenuating medium

The problem is to show the influence of the presence of the rachis and of the depth effect on the various activity concentration estimates obtained by using the four variants of the phantom.

A convenient way of representation is to draw for each of the 13 estimates the differences between the results obtained with and without the rachis in the phantom simulation. A curve showing the effect of the presence of the rachis is then obtained for each correction technique. By drawing the "difference curves" for phantom simulations with and without the depth effect, the influence of this latter can be shown (fig. 13).

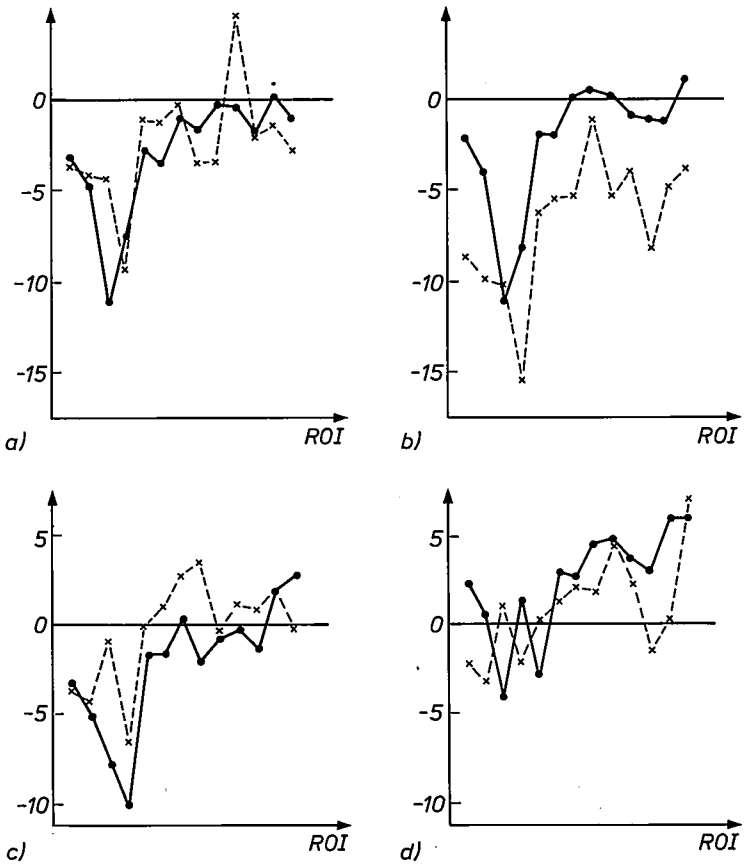


Fig. 13. "Difference curves" for 4 attenuation correction techniques: *a*) Walters; *b*) Chang; *c*) Gustafson; *d*) Bellini. Regions of interest (ROI): 1*a*, 1*b*, 2*a*, 2*b*, 3*a*, 3*b*, 4*a*, 4*b*, 4*c*, 4*d*, 5, 6*a*, 6*b*. Drawn line without depth effect, dotted line with depth effect.

IPR2018-1557
HTC EX1023, Page 52

It can be pointed out in this way that Walters' and Bellini's techniques do not show any significant differences when taking into account the depth effect (figs 13*a* and 13*d*) (the paired series are not significantly different), contrary to those of Gustafson and Chang (figs 13*b* and 13*c*).

It is useful to verify if there are significant variations for the small ROI 2 and 3 (diameters smaller than 30 mm). The activity concentration in the ROI 2 located in front of the rachis is under-estimated by about 10% on the "difference curves" (except in Bellini's technique) whereas the ROI 3 at the right lobe centre does not show any significant modification. The taking into account of the depth effect leads to a recovering of 5% in the estimate in the ROI 2 except with Chang's technique where a decrease of 7% is observed.

Globally the Bellini, Gustafson and Walters techniques seem to be the least sensitive ones to the presence of both the rachis and the depth effect.

It can be concluded that the Bellini and Gustafson techniques allow to obtain correct evaluations of the activity concentrations even for the ones located close to the rachis. The Walters technique, generally little sensitive to the bone and the depth effect presence, is not convenient for the estimate of the activity concentrations, close to the rachis. The Chang technique, sensitive to the bone presence for the concentrations close to the rachis is globally sensitive to the depth effect.

4.3. *Discussion*

The study reported here, concerning a phantom of the abdomen cannot be directly extended to the thorax, the pelvis or the brain. In order to be able to conclude in a general manner, it would be necessary to generate typical phantoms of these body regions and to do similar tests as the ones we did. Our results however remain valid for the abdomen which is an important object of medical examinations.

We chose variations of the activity concentrations from 1 to 4 corresponding to the main variations generally encountered in the abdominal pathology. The statistical fluctuations of clinical data could be of the same order as the lowest concentration value. The measured point spread functions used for the depth effect simulation are compatible with the ones obtained with most of the present tomographic systems. The gamma scattering effect has been partially taken into account by the fact that these point spread functions have been obtained in a scattering medium.

We did not study the influence of the statistical noise and of the non-uniformity of the camera field of view which can be analyzed independently of the auto-attenuation correction methods.

IPR2018-1557
HTC EX1023, Page 53

We used for the reconstruction the classical filtered back-projection with the ramp filter which only corrects the blurring introduced by this method. We did not introduce any additional spatial frequency window which might have screened the interesting results.

In the case of small active volumes we were obliged to use the maximum values because of the too poor estimate obtained from the average value in the corresponding ROI, as it is presently done in clinical examinations. Other kinds of abdominal phantoms would possibly modify slightly the conclusions but it can be expected that the studied parameters will remain in equivalent ratios for the various correction techniques.

The techniques using the corrected or non-corrected geometrical or arithmetical means have been rejected because they are connected too closely to the structure of the activity and the attenuation domains. However a study of Sorenson's technique with an adaptation of the correcting coefficient to the section structure would be interesting. The Walters' and Chang's techniques have been used with only one iteration and the estimated values were close to the true ones. The fact that these techniques do not take into account the bone in their principles do not allow to hope to recover the true value after more than one iteration. A map of the exact attenuation coefficients (obtained by means of a C.T. scanner for example) would help to reach more exact estimates.

The conclusion on the global sensitivity of the correction techniques to the presence of the bone and of the depth effect has to be weighed by an examination of their ability to estimate the absolute concentration values in the small active volumes. We saw that for the latter the use of a local maximum can recover more or less an acceptable estimate. In clinical environment the size of some active volumes is not known and it would be useful to estimate the activity concentration and the volume at the same time. A complete solution to this problem is not known but a better sampling can bring an improvement.

5. Conclusions

The increasing interest of nuclear medicine for quantitative studies in various domains of applications (psychiatry, immunology using labelled monoclonal antibodies, cardiology etc.) makes the choice of a good attenuation correction method of great importance. The techniques of attenuation correction which have been studied here associated to a classical image reconstruction method, are the most used or described ones in the literature. Some are interesting because of their principle (Bellini) or their generality (Walters). Their comparison, carried out on a numerical abdominal phantom, allowed to classify them in function of their ability to recover the true activity concentrations. The study of the influence of a non-homogeneous medium and of the

IPR2018-1557
HTC EX1023, Page 54

depth effect on reconstructions shows that some techniques allowing to obtain globally good absolute or relative estimates also have a good behaviour towards these two effects (Bellini, Gustafson, Walters). The other ones have to be rejected (simple or arithmetical mean of two opposite projections) or carefully used in quantitative studies.

Acknowledgement

The authors would like to thank the people who make the work reported here possible either by their active participation or their advices: Mrs. Leblanc, Pauvert, Decaesteker, Paulin and Jatteau.

REFERENCES

1) J. Radon, Berichte Saechriche Akademic der Wissenschaften **69**, 262 (1917).
2) J. A. Sorenson, Instrumentation in Nuclear Medicine, Vol. 2, 311 (Edited by G. J. Hine and J. A. Sorenson), Academic Press (1974).
3) W. I. Keyes, Brit. J. Radiol. **49**, 62 (1976).
4) T. F. Budinger and G. T. Gullberg, IEEE Trans. on Nucl. Sci. **NS 21**, 2 (1974).
5) T. F. Budinger and G. T. Gullberg, Reconstruction tomography in diagnostic radiology and nuclear medicine, 315 (Edited by Ter Pogossian et al.), University Park Press (1977).
6) T. F. Budinger et al., Image reconstruction from projections: implementation and applications (Edited by G. T. Herman), Topics in Applied Physics, Vol. 32, 147, Springer-Verlag, Berlin (1979).
7) D. B. Kay and J. W. Keyes, J. Nucl. Med. **16**, 540 (1975).
8) S. Bellini et al., IEEE Trans. on Acoust. Speech and Sign. Proc., ASSP **27**, 213 (1979).
9) O. J. Tretiak and P. Delanay, SIAM J. Appl. Math. **39**, 341 (1980).
10) D. E. Gustafson et al., Radiology **129**, 187 (1978).
11) T. E. Walters et al., J. Comp. Assi. Tomo. **5**, 89 (1981).
12) L. T. Chang, IEEE Trans. on Nucl. Sci. **NS 25**, 638 (1978).
13) R. J. Jaszczak et al., IEEE Trans. on Nucl. Sci. **NS 27**, 1137 (1980).
14) R. J. Jaszczak et al., Phys. Med. Biol. **24**, 1123 (1979).
15) R. J. Jaszczak et al., IEEE Trans. on Nucl. Sci. **NS 28**, 69 (1981).
16) J. Pergrale, C. Berche, D. Jacchetti, G. Normand and M. Jatteau, to be published in IEEE Trans. on Nucl. Sci.

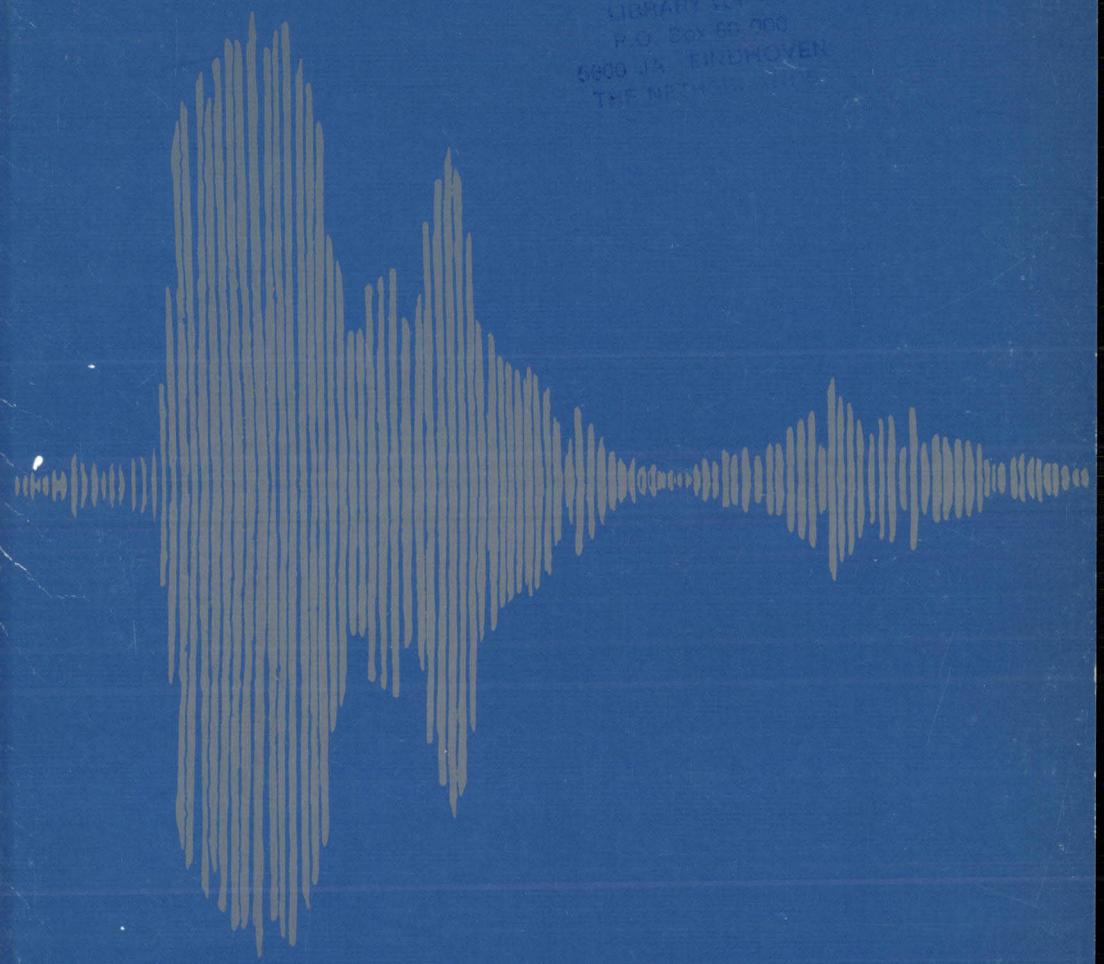
IPR2018-1557
HTC EX1023, Page 55

17 DEC 1984
3-48=27

Philips Journal of Research

Vol. 39 No. 3 1984

Nat. Philips Bedrijven B.V.
PHILIPS RESEARCH LABS.
LIBRARY NL - 1
P.O. Box 80 000
5600 JA EINDHOVEN
THE NETHERLANDS



PHILIPS

IPR2018-1557
HTC EX1023, Page 56

Philips Journal of Research

Philips Journal of Research, published by the 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 39 is Dfl. 75, 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.

Editorial Board

A. van Oostrom (General Editor) Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Y. Genin, Brussels, Belgium
R. Kersten, Aachen, Germany
R. Memming, Hamburg, Germany
R. F. Milsom, Redhill, Surrey, U.K.
W. A. Smith, Briarcliff Manor, N.Y., U.S.A.
W. T. Stacy, Sunnyvale, CA., U.S.A.
J. B. Theeten, Limeil-Brévannes, France

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

© Philips International B.V., Eindhoven, The Netherlands, 1984. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

IPR2018-1557
HTC EX1023, Page 57

Philips J. Res. **39**, 51-60, 1984 R 1080

AN EFFICIENT SILICON COLD CATHODE FOR HIGH CURRENT DENSITIES

I. Experimental data and main results

by G. G. P. VAN GORKOM and A. M. E. HOEBERECHTS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Experimental data and main results of measurements on silicon p–n emitter cold cathodes are described. It is shown that very shallow p–n junctions (depth below the surface $\approx$ 10 nm) biased in avalanche breakdown are able to emit electrons into vacuum with high efficiencies and current densities. After cesiation of the surface, efficiencies of about 2% have been reached and even higher after in situ (partial) cleaning of the surface. Current densities in excess of 1000 A/cm^2 have been measured on emitters with a diameter of 1 μm and about 250 A/cm^2 for emitters of 6 μm diameter. The effective electron temperature is found to be about 0.5 eV.

PACS numbers: 79.70.+q, 73.30.+y

1. Introduction

It has long been known that shallow p–n junctions in semiconductors are able to emit electrons into a vacuum when biased in avalanche breakdown. The first paper on this subject appeared in 1957 [1]), and for a number of years there was some research in this area, because it was thought that this phenomenon might lead to a new type of cold cathode. Most experiments were done on p–n junctions in silicon [2–4]), and at first rather rarely on other semiconductors [2]). However, at the end of the 1960's and the beginning of the 1970's, a lot of work has been done on p–n emitters made of SiC [5]).

As far as we know, these investigations did not result in commercially available devices employing this type of cold cathode. The reason for this is not quite clear. The early silicon devices suffered, probably, from too low efficiencies and low current densities. The SiC cathodes had some trouble with reproducibility and had rather uncomfortable geometries.

IPR2018-1557
HTC EX1023, Page 58

In most of the studies mentioned above, the p–n junction intersected the surface and the electrons emerged at the surface from a very small area around the intersecting line. Bartelink et al.[3]) studied electron emission from diodes with the junction parallel to the surface. A modern version of this experiment was performed by Shahriary et al.[6]). Besides electron emission into vacuum, electron injection in SiO_2 layers has also been studied in detail since 1970[7]), mainly because of its importance in memory devices and in device reliability problems. In these cases as well two geometries were studied, one with the junction intersecting the Si–SiO_2 interface[8]) and one with the junction parallel to the interface[9]). From refs 8 and 9 it can be concluded that high current densities and high injection efficiencies for electron emission in SiO_2 layers can be achieved. This stimulated us to look again at the electron emission into vacuum and to try to increase this to levels at which practical application would be possible, using the knowledge obtained from the oxide injection experiments and by using modern silicon technology. As we shall show in the following, we have come close to this goal indeed. Key factors for the success are the availability and use of low energy ion implantation to produce very shallow junctions and the notion that the dimensions of the emitters have to be small to get high current densities.

The important parameters of the p–n cold cathode are the efficiency of the emission process, the energy spread of the emitted electrons, the obtainable current density (sometimes called cathode load) and total currents. These will be described in a series of 3 papers. This first one deals with experimental data and with our most important results. In our second paper (which will be referred to as HG II) these results will be discussed in detail and the experiments compared with theory. Emission stability problems (caused by desorption and drift of the cesium atoms) will be the subject of the third paper (GH III).

In a previous paper[10]), we reported the energy distribution of the emitted electrons for both types of emitters, namely with the junction intersecting the surface (class I) and with the junction parallel to the surface (class II). It was found that the former (class I) devices produced very broad energy distributions, which is of course a distinct disadvantage if these emitters are to be used in, for instance, cathode ray tubes. Furthermore it was found that after depositing a monolayer of cesium on the surface of class I emitters (as is necessary for reaching sufficient emission currents, as will be discussed later) a stability problem was encountered, due to the fact that the cesium atoms (which are fractionally charged) were forced to drift in the very large electric field parallel to the surface. For these two reasons, the class I cathodes are unsuitable for almost all applications. In the following we shall therefore deal with the class II emitters only.

IPR2018-1557
HTC EX1023, Page 59

2. Experimental

2.1. *Silicon cold cathode preparation*

There are several ways of producing a p–n avalanche cathode within the silicon technology. Here we describe roughly a typical version in circular geometry. Other geometries and processing methods will be described in HG II. The measurements on emitters made by different methods and on different geometries all yielded similar results regarding efficiencies and energy spread.

The device structure is drawn schematically in fig. 1. The design is such that we create two types of diodes:

1) a diode which merely acts as a contact and a protection;
2) a small diode, the electron-emitting diode.

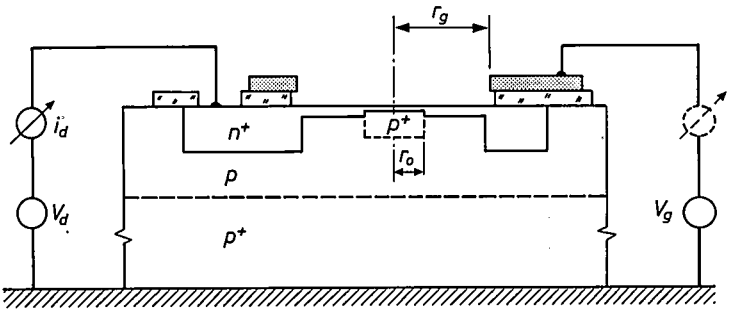


Fig. 1. Schematic drawing of the silicon p–n emitter cold cathode with circular geometry. The active p^+ area has a radius r_o in the range from 0.5 to 3 μm. The radius of the hole in the gate (r_g) is usually 5–10 μm.

The breakdown voltage of the contact diode, produced by n–type diffusion in a p–type substrate, is 4–20 volts higher than that of the active diode, depending on geometry and process conditions. By locally diffusing boron (in this description a circular area with radius r_o) we lower the breakdown voltage to about 5 volts. This provides a good compromise between the avalanche contribution to the diode current and the onset of the tunnelling current contribution. To reduce the series resistance an epitaxial layer on a heavily doped substrate can be used. Details are listed in table I.

The process starts with an oxidation of a ⟨100⟩ (epitaxial) substrate. The oxidation conditions are chosen such that the oxide can withstand a phosphorous deposition and diffusion. With a conventional photolithographic technique (using projection imaging of masks) and subsequent etching of the glass, a window is formed for a phosphorous deposition and diffusion. Within the n^+ area a circular oxide island is left with a diameter of about 20 microns. Using a second mask this island is removed. Within the (circular) p–type

IPR2018-1557
HTC EX1023, Page 60

TABLE 1

Substrate	*p*–type
thickness	~ 380 μm
resistivity	< 0.01 Ωcm
orientation	⟨100⟩
Epitaxial layer	*p*–type
thickness	~ 6 μm
concentration	2×10^{16} cm^{-3}
Diffusion contact	
impurity	phosphorus
surface concentration	$> 5 \times 10^{19}$ cm^{-3}
depth	~ 3 μm
Diffusion n-channel	
impurity	arsenic/antimony
surface concentration	2×10^{19} cm^{-3}
depth (at p^+)	10 nm
Dielectric layer	
material	oxide or oxide/nitride
thickness	~ 1 μm
Gate (grid)	
material	polysilicon/Al
thickness	~ 0.5 μm
Mounting	SOT 14
Bonding	
material	Al
wire	25–40 μm
method	ultrasonic

region just defined, a circular region is defined by developing a window in photolacquer. The active area (with radius r_o), is defined by boron implantation, cleaning and subsequent oxidation. On top of the oxide layer a conductive layer can be deposited. Because this work started from MOS transistor studies we call this layer *the gate*. The gate is structured in a subsequent lithographic step leaving the oxide layer covered for almost the complete chip. In operation, the gate can act as an anode or a grid or, when connected to the *n* layer, it serves simply to avoid charging of the dielectric layer by spurious

IPR2018-1557
HTC EX1023, Page 61

electrons. At the outer edge room is left for a contact window in the uncovered oxide area. This window is defined together with an area around the active diode. A final implantation of arsenic or antimony ions followed by annealing completes the high temperature part of the process. From fig. 1 it can be seen that this last formed shallow layer makes contact with the phosphorus-diffused area and that it defines the earlier mentioned *active area*. Several conditions of implantation doses and energies are tested and will be described in a subsequent paper. In general the conditions end up with the metallurgical junction at a depth around 10 nm.

The surface of the active diode is left with an unsaturated 'natural' oxide which is etched just before mounting in a UHV system or tube. If an HF based etch is used aluminium metallization patterns are severely attacked at steps. For ease of preparation we just mount a chip on a standard header and ultrasonically bond wires directly on silicon and, if present, on polysilicon. Etching of the thus completed chip gives no complications concerning reliability, even after prolonged heating of UHV equipment or tubes.

To avoid contamination by deposits when the devices are operated in air, testing is done only with the protective natural oxide layer present and to about 5 V reverse bias only. A typical characteristic of one of our diodes is given in fig. 2.

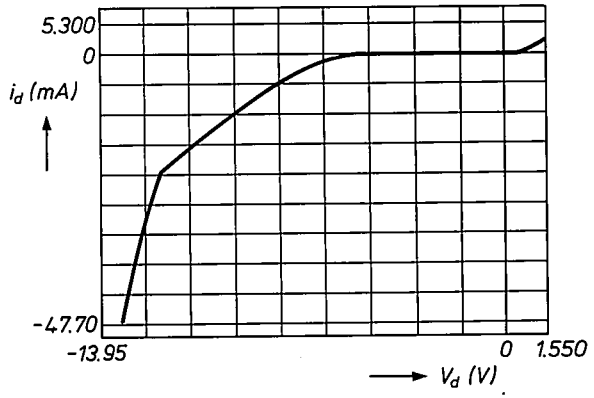


Fig. 2. *I–V* characteristic of a *p–n* emitter with $r_o = 3$ µm.

2.2. *Vacuum conditions*

Most experiments reported here were done in a getter-ion-pumped stainless steel UHV system. Pumping down from atmospheric pressure proceeds using an oil-free rotating pump and cryosorption pumps. The whole system is baked at a temperature of about 250 °C overnight. After cooling down an ultimate

IPR2018-1557
HTC EX1023, Page 62

pressure below 10^{-8} Pa is reached. This UHV system contains the spherical symmetric retarding field energy analyser, described in some detail in ref. 10, which is used to measure the total energy distribution [11]) of the emitted electrons. The resolution is about 0.1 eV.

A number of experiments were performed on emitters mounted in cathode ray tubes. These tubes were pumped using turbo-molecular pumps and were baked at 380 °C overnight. In most cases the emission efficiency was found to be somewhat lower in the tubes and the results showed a larger spread in the results, as compared with the UHV results. This could be due to the fact that in tubes it is more difficult to evaporate a clean monolayer of cesium on the surface than in a UHV system.

2.3. *Cesium deposition*

As will be discussed further in sec. 3, a monolayer of cesium on the surface is necessary to produce high emission efficiency. This is, of course, due to the lower work function of a cesiated surface. To get the best results it is important to evaporate a pure cesium layer. Unfortunately, the commonly used dispenser cesium sources (consisting of cesium chromate and a reducing agent) produce a lot of gas, including also oxidizing gases (O_2, H_2O, CO_2) which are harmful for the emission. It is therefore important to degas the cesium source thoroughly before evaporating the cesium. But even then, the release of cesium is still accompanied by some gas. So, in order to get the best results, the cesium has to be deposited on some intermediate surface and after that, when the pressure in the system is again low enough, it is re-evaporated and deposited on the *p–n* emitter. Such procedure is followed in our UHV system but in tubes this is clearly more difficult and this leads to a loss in performance of the cold cathode in such tubes.

3. Main results

We are interested in the following cathode properties: efficiency, energy distribution, current density (sometimes called cathode load) and total emission current.

3.1. *Efficiency*

The efficiency η of the *p–n* emitter is defined as

$$\eta = \frac{i_{vac}}{i_d + i_{vac}} \simeq \frac{i_{vac}}{i_d}, \qquad (1)$$

where i_{vac} is the emitted electron current and i_d the (non-emitted) diode current. As will be discussed in HG II, the value of η is determined by the electric

IPR2018-1557
HTC EX1023, Page 63

field in the depletion layer, the mean free path of the electrons within the silicon, the depth of the non-depleted part of the very shallow n^+ layer and the work function of the surface. The measured values for the devices as sketched in fig. 1 are:

$\eta = 1\text{–}5 \times 10^{-5}$ pure silicon

$\eta = 1\text{–}2 \times 10^{-2}$ silicon-cesium.

The gain by a factor of about 200–1000 when cesium is used is due to the lowering of the work function from ~ 4.7 eV (pure Si) to ~ 1.7 eV (Si–Cs). The spread in the results is due to residual contaminations on the surface e.g. oxygen and carbon, which apparently differ somewhat from one emitter to another.

In a separate experiment[12]), p–n emitters were mounted in a scanning Auger microscope (SAM 590 from Physical Electronics) enabling us to measure the elemental composition of the surface. It was found that at an efficiency of about 2%, oxygen was still present on the surface. This oxygen could be partly removed by using the electron beam (electron stimulated desorption). After this treatment an efficiency $\eta = 3.6\%$ was measured. An even higher efficiency was obtained by sputter-cleaning of the surface with 500 eV argon ions. This resulted in $\eta = 7.25\%$ (after re-deposition of cesium, but without a thermal anneal).

3.2. *Energy distribution*

Due to the fact that we use a spherical symmetric energy analyser we measure the so-called total energy distribution[11]). In ref. 10 we reported e.g. the energy distribution of a class II emitter which was produced using 25 keV arsenic ion implantation as the final step. The shape of the measured curve could be fitted well to a Maxwellian distribution with an electron temperature $kT_e = 0.18$ eV. A simpler method of deducing electron temperatures from energy distribution is by using the full width at half maximum ΔU of the distribution curve. The electron temperature then follows from[11]):

$$kT_e = \frac{\Delta U}{2.45} . \tag{2}$$

In the former case a $\Delta U = 0.50$ eV was found, leading to $kT_e = 0.20$ eV. This is somewhat higher than the value obtained from the fitting procedures, the difference is caused by the resolution of the energy analyser (~ 0.1 eV).

The energy distribution of p–n emitters with 5 keV arsenic implantation is, without cesium on the surface, found to be Maxwellian again with $\Delta U = 0.93$ eV. Without corrections for the resolution, this leads to $kT_e = 0.38$ eV. With cesium on the surface, the shape is not Maxwellian, see fig. 3. As will be

IPR2018-1557
HTC EX1023, Page 64

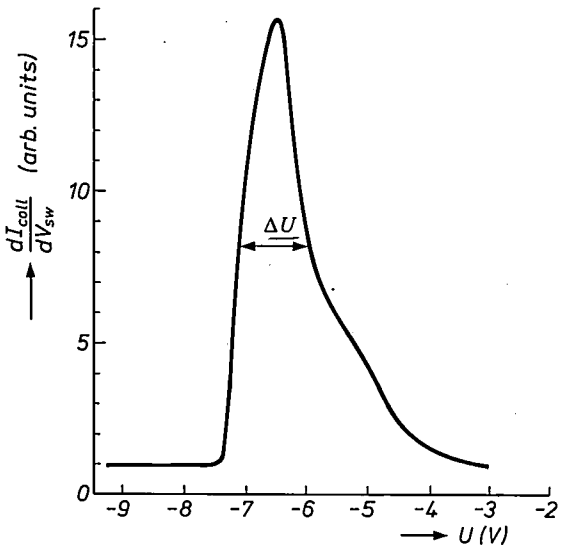


Fig. 3. Energy distribution of a cesiated p–n emitter. The full width at half maximum is ΔU.

discussed in HG II, this is probably caused by the structure in the density of states of the silicon conduction band. However, the deviation from the Maxwellian shape is not too large and so it seems reasonable to use eq. (2) again to find an effective electron temperature which, from the measured value $\Delta U = 1.20$ eV, results in $(kT_e)_{eff} = 0.49$ eV.

3.3. *Current density*

The current density into vacuum j_{vac} is simply given by

$$j_{vac} = \eta j_{d\perp}, \tag{3}$$

where $j_{d\perp}$ is the current density in the diode perpendicular to the surface, see fig. 4. To achieve high current densities (at a given value of η) this $j_{d\perp}$ should be as large as possible. However, the limiting factor is not $j_{d\perp}$ but $j_{d//}$ (fig. 4), because of the fact that the non-emitted diode current has to be squeezed into the very shallow n^+ channel. As the thickness of the channel must be of the order of 10 nm to obtain high efficiencies, it will be clear that less current crowding will occur with small emitters and so the highest current densities can be expected from the smallest emitters. This turns out to be the case. As will be discussed in detail in GH III, we have found empirically that, for stable emission (apart from instabilities due to poor vacuum conditions) $j_{d//}$ has to be restricted to

IPR2018-1557
HTC EX1023, Page 65

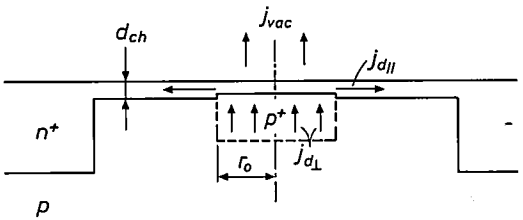


Fig. 4. The current density $j_{d//}$ in the very thin n^+ channel (~ 10 nm) is much larger than the current density $j_{d\perp}$ in the avalanche region. It is $j_{d//}$ that limits the maximum available vacuum current density j_{vac}.

$$j_{d//} \lesssim 2.5 \times 10^6 \text{ A/cm}^2. \tag{4}$$

From fig. 4 and assuming a homogeneous avalanche breakdown it can easily be seen that the relation between $j_{d\perp}$ and the maximum value of $j_{d//}$, $j_{d//}^m$ is given by

$$j_{d\perp} = \frac{2d_{ch}}{r_o} j_{d//}^m \tag{5}$$

and so, using $d_{ch} \approx 10^{-6}$ cm, eqs (3) and (4)

$$j_{vac} \lesssim \frac{5\eta}{r_o}, \tag{6}$$

where r_o has to be expressed in cm, j_{vac} is then given in A/cm^2. For an emitter with $r_o = 5 \times 10^{-5}$ cm this leads to $j_{vac} \lesssim \eta \cdot 10^5$ A/cm^2.

For $\eta \approx 1$–2% this results in $j_{vac} \simeq 1000$–2000 A/cm^2. In practice we have indeed achieved current densities in excess of 1000 A/cm^2 with good stability. For emitters with $r_o = 3 \times 10^{-4}$ cm, current densities $j_{vac} \simeq 250$ A/cm^2 have been obtained.

3.4. *Total emission currents*

In normal thermionic cathodes almost any total current can be obtained at a certain current density just by increasing the emitting area. For p–n emitters the case is more complex. The obtainable total vacuum current i_{vac} is (see eq. (1)) given by

$$i_{vac} = \eta\, i_d. \tag{7}$$

Having a certain value for η, the vacuum current can be increased by increasing the diode current i_d. This is, again, limited by the maximum permissible value of $j_{d//}$. The corresponding maximum diode current i_d^m is given by

$$i_d^m = 2\pi\, r_o\, d_{ch}\, j_{d//}^m. \tag{8}$$

Using $j_{d//} \lesssim 2.5 \times 10^6$ A/cm^2 and $d_{ch} \simeq 10^{-6}$ cm, we find from eq. (7) and (8)

$$i_{vac} \lesssim 5\pi\, \eta\, r_o \quad (r_o \text{ in cm, } i_d \text{ in amps}). \tag{9}$$

IPR2018-1557
HTC EX1023, Page 66

It follows that (if $\eta \simeq 1.5 \times 10^{-2}$) $i_{vac} \lesssim 10$ μA for emitters with $r_o = 5 \times 10^{-5}$ cm, and $i_{vac} \lesssim 60$ μA for emitters with $r_o = 3 \times 10^{-4}$ cm. These values are in agreement with the experiments.

If larger total currents are needed, r_o has to be increased (eq. (9)) but this leads to a decrease in the current density, eq. (6). At some point, as will be discussed further in HG II, it is better to leave the concept of a homogeneous circular emitter and to go over to concentric annular emitters or to some other geometry.

4. Conclusions

The conclusions that can be drawn from the measurements of the emission properties of silicon p–n junction emitters covered with a monolayer of cesium are

— efficiencies of max. ~2% are attainable,
— the energy spread is rather large and corresponds to an effective electron temperature of ~0.5 eV,
— very high current densities can be obtained, especially with small emitters,
— the total available emission current depends on the geometry and can be increased at the expense of current density.

Acknowledgement

The authors are indebted to a number of colleagues, especially to J. Verwey, P. Zalm, J. van Santen, J. van Laar and A. van Oostrom for stimulating discussions, to P. Drummen for manufacturing the silicon devices, to J. van Esdonk and co-workers for processing the tubes, and to P. Peters, P. Dessens and B. Cadier for technical assistance.

REFERENCES

[1]) J. A. Burton, Phys. Rev. **108**, 1342 (1957).
[2]) R. J. Hodgkinsson, Sol. St. Electr. **5**, 269 (1962).
[3]) D. J. Bartelink, J. L. Moll and N. I. Meyer, Phys. Rev. **130**, 972 (1963).
[4]) M. Waldner, J. Appl. Phys. **36**, 188 (1965).
[5]) See e.g. R. V. Bellau and A. E. Widdowson, J. Phys. D. **5**, 656 (1972).
[6]) I. Shahriary, J. R. Schwank and F. G. Allen, J. Appl. Phys. **50**, 1428 (1979).
[7]) For an excellent review on these matters, see: T. H. Ning, Sol. St. Electr. **21**, 273 (1978).
[8]) J. F. Verwey and B. J. de Maagt, Sol. St. Electr. **17**, 963 (1974).
[9]) T. H. Ning, C. M. Osborn and H. N. Ya, J. Appl. Phys. **48**, 286 (1977).
[10]) G. G. P. van Gorkom and A. M. E. Hoeberechts, J. Appl. Phys. **51**, 3780 (1980).
[11]) R. D. Young, Phys. Rev. **113**, 110 (1959).
[12]) A. van Oostrom, L. J. M. Augustus, G. G. P. van Gorkom and A. M. E. Hoeberechts, to be published.

IPR2018-1557
HTC EX1023, Page 67

Philips J. Res. 39, 61-76, 1984 *R 1081*

SECONDARY ELECTRON YIELDS FROM CLEAN POLYCRYSTALLINE METAL SURFACES BOMBARDED BY 5-20 keV HYDROGEN OR NOBLE GAS IONS

by P. C. ZALM and L. J. BECKERS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Ion-induced secondary electron yields for clean polycrystalline Al, Ti, Ni, Cu, Zn, Mo, Ag, Au and Pb bombarded at normal incidence with 5–20 keV H^+, H_2^+, He^+, Ne^+, Ar^+, Kr^+ or Xe^+ ions at current densities of about 10 $\mu A/cm^2$ have been measured. Ultra-high vacuum conditions ($P < 1 \times 10^{-7}$ Pa) and mass-selected particle beams were employed. Target cleaning was achieved by sputtering and monitored by Auger electron spectroscopy. A relatively simple electron yield determination method was used, employing direct current measurement combined with a variable positive target bias voltage. Good overall agreement was found with the scarce experimental data available for the same energy range. A confrontation with predictions of recent sophisticated theoretical work reveals that, even qualitatively, the features of the bombardment-energy dependence of the ion-induced secondary electron yields are not well understood.

PACS numbers: 79.20.Nc, 61.80.Jh, 34.90.+q

1. Introduction, outline and theoretical preliminaries

Bombardment of a solid surface with swift particles generally leads to electron emission. This phenomenon has consequences for e.g. ion current measurements, plasma-wall interactions etc. Also, total electron yield, energy distribution of the emitted electrons and their angular ejection pattern may reflect properties of detail processes of the particle-solid interaction. For these reasons secondary electron emission has for many years received considerable attention in the literature. Excellent recent reviews of both the experimental and theoretical status of this subject can be found in refs 1–3. The present study deals with the particular case of the total yield of secondary electrons emitted from elemental (polycrystalline metal) surfaces bombarded by energetic (positive) ions.

Conventionally, two independent mechanisms are distinguished for positive ion bombardment-induced secondary electron emission. *Potential emission* is associated with the neutralization of the impinging ion immediately prior to hitting the surface. The process can take place if the neutralization energy of

IPR2018-1557
HTC EX1023, Page 68

the ion, E_i, exceeds twice the work function, φ, of the solid. It has no kinetic threshold. Consequently, at low ion energies ($\lesssim$1 keV), potential electron emission will be fairly constant and dominant, while it is expected to decrease rapidly at higher projectile velocities. In a theoretical study of the Auger neutralization model, Kishinevskii[4]) arrived at the approximate estimate for the low-energy potential electron emission yield

$$\gamma_{PEE} \simeq \frac{0.2(0.8\,E_i - 2\varphi)}{\varepsilon_F}, \qquad (1)$$

where ε_F is the Fermi energy of the target electrons. Baragiola et al.[1]) proposed a slightly different semi-empirical prediction, which reads

$$\gamma_{PEE} \simeq 0.032(0.78\,E_i - 2\varphi). \qquad (1')$$

Both eqs (1) and (1′) agree reasonably well ($\sim$ 25%) with experimental data[5,6]), with ionization energies E_i, taken from ref. 7, work functions φ, taken from a recent compilation[8]), and Fermi energies ε_F, taken as described in ref. 9. The main ambiguity in applying eqs (1 and 1′) stems from the values of φ, which in a particular experiment may be influenced by trace contamination of the surface.

The second mechanism governing electron ejection is called *kinetic electron emission*. It is associated with the ionizing collisions taking place after the incident ion penetrates the surface. A four-step process governs the emission[2])
- ionization by the bombarding ions,
- secondary ionization by energetic recoils, electrons or photons,
- transport of liberated electrons to the surface,
- electron ejection through the surface barrier.

Kinetic emission depends on a large number of parameters
- energy E_p, atomic number Z_p, mass M_p and angle of incidence of the projectile;
- atomic number Z_t, mass M_t, work function and Fermi energy of the target as well as its surface structure and temperature and possible contamination effects.

Whereas the total electron yield for potential emission is fairly well understood, the same cannot be said for kinetic emission. By now, there is mutual agreement among theoreticians that for light projectiles (e.g. H^+, He^+) on heavy atom targets the electronic stopping, i.e. the energy loss through excitation of target electrons, dominates the electron ejection outside the potential emission regime. Likewise, for heavier projectiles at high energy ($\gtrsim$100 keV) one should expect a near proportionality between total electron yield and electronic stopping cross section ($\gamma_{KEE} \propto S_e(E_p)$), owing to the negligible influence

IPR2018-1557
HTC EX1023, Page 69

of the energetic recoil atoms, created through elastic collisions governed by a nuclear stopping cross-section $S_n(E_p)$, which is small compared to $S_e(E_p)$. The contribution from such energetic recoils has become clear from the theoretical work of Schou [10]). In analogy to Sigmund's successful theory for sputtering he solved a system of Boltzmann transport equations describing the ionization cascade in a random, semi-infinite monoatomic medium to arrive at

$$\gamma_{KEE} = \Lambda D(E_p, x = 0). \tag{2}$$

Here Λ is an assumed target material constant, which contains the energy-, depth- and angle-averaged escape probability for electrons liberated inside the target and encomprises the surface barrier. D stands for the mean energy deposited in electronic excitation at the surface ($x = 0$) by an incident ion with energy E_p. In a slightly more transparant form eq. (2) can be rewritten as

$$\gamma_{KEE} = \Lambda N \left[\beta_0 S_e(E_p) + \beta_r \frac{\eta(\mu E_p)}{\mu E_p} S_n(E_p)\right], \tag{3}$$

where N is the target density and $\mu = 4M_pM_t/(M_p + M_t)^2$ the maximum energy transfer factor. $S_e(E_p)$ and $S_n(E_p)$ are the projectiles electronic and nuclear stopping powers, respectively, evaluated at the surface. In our energy regime $S_e(E_p) = K_{pt}\sqrt{E_p}$ and $S_n(E_p) \simeq C_{pt}\sqrt{E_p}/(1 + 4.5\,\varepsilon)$ where the reduced energy ε is given by $\varepsilon = E_p/\Sigma_{pt}$, with

$$\Sigma_{pt} = \frac{1}{32.5}\left(1 + \frac{M_p}{M_t}\right) Z_p Z_t (Z_p^{\frac{2}{3}} + Z_t^{\frac{2}{3}})^{\frac{1}{2}} \text{ keV}.$$

Here K_{pt} and C_{pt} are characteristic constants depending on projectile and target atomic number and mass.

[Some numerical values: $K_{pt} \simeq 0.05$ Å^2 keV$^{\frac{1}{2}}$ for H^+ and He^+ ions and ranges from 0.06–0.18 Å^2 keV$^{\frac{1}{2}}$ for the heavier noble gas ions in Z_t = 13-82 materials; C_{pt} is around $0.3/Z_t$ Å^{2}keV$^{\frac{1}{2}}$ for H^+ and He^+ and varies from 0.6 Å^2 keV$^{\frac{1}{2}}$ for $Ne^+ \rightarrow$ Al to 1.3 Å^2 keV$^{\frac{1}{2}}$ for $Xe^+ \rightarrow$ Pb.]

β_0 ($\simeq 1$ for higher energies) and β_r($\simeq [M_t + M_p]/10M_p$ for $M_t/M_p \lesssim 10$ and levels off to around 2.0($\pm$ 0.5) for $M_t \gg M_p$) are dimensionless factors accounting for energy transport through the surface by backscattered primary ions, excited target electrons and swift recoils. These factors are weakly dependent on energy and the target/projectile atomic mass ratio. The factor $\eta(\mu E_p)/\mu E_p$, roughly equal to 0.3 $Z_t^{\frac{1}{6}}(\mu\varepsilon)^{0.2}$ (deduced from data in ref. 10), reflects the secondary electron generation by recoil atoms. It is an almost universal function, which depends only weakly on target parameters. Although eq. (3) is synthesized from a complete physical picture, and therefore most suited for prediction, it is not readily applicable to experiment because of the large number of functional forms that are involved and have to be evaluated.

IPR2018-1557
HTC EX1023, Page 70

In spite of the fact that the investigation of ion-induced secondary electron emission is an old subject, accurate experimental data are not abundant, owing to ill-defined measurement conditions, as for example poor vacuum. Moreover, the main body of high-quality data refers to the energy range of the potential emission regime. In this paper we report on measurements of total electron yields for 5–20 keV H_2^+, He^+, Ne^+, Ar^+, Kr^+ and Xe^+ ions incident on clean targets of Al, Ti, Ni, Cu, Zn, Mo, Ag, Au and Pb to offer a sizeable, self-contained, data set for future theoretical reference. The energy range is such that the potential electron emission mechanism will play a minor role, but that the contribution of energetic recoils to the total yield may be expected to be considerable if not dominant, for the heavier noble gas ions. Our data are compared with reliable results recently obtained elsewhere and confrontated with some qualitative predictions from theory.

2. Experimental procedure details

2.1. *Ion beam equipment and characteristics*

Beams of singly-charged, mass-selected H^+, H_2^+ or noble gas ions at energies of $E_p = 5$–20 keV were generated in the ion-beam apparatus discussed in detail elsewhere[11]). A short description of its main features pertinent to the appraisal of our results is given below.

The ions are formed in a hollow hot-cathode source in which a high-density plasma is sustained at a low discharge voltage ($\simeq 40$ V). The energy spread in the beam is thus limited to a few electronvolts. Through a three-stage differentially pumped focusing system, comprising two Einzel lenses with an $E \times B$ Wien-type mass filter between them (with a selectivity of about $M/\Delta M = 16$ at $E_p = 10$ keV), the beam impinges, at normal incidence, on the target in the main UHV chamber. The base pressure in this chamber is always kept below 10^{-7} Pa prior to bombardment. During irradiation it was found to increase to at the most 5×10^{-6} Pa. Residual gas analysis showed this to be solely due to leak-in from the source and reflection of (neutralized) projectiles from the target. A final beam-defining diaphragm of 6 mm diameter is located at 40 cm from the target and 125 cm from the source. It is kept at $+30$ V to prevent liberation of secondary electrons. The maximum attainable beam currents on the target are depicted in fig. 1*a* for the different ionic species used in this work. The almost circular beam had a diameter of about 15 mm for $E_p = 20$ keV, decreasing roughly linearly to 10 mm at $E_p = 5$ keV, owing to the improved focusing abilities of the ion optical system at lower projectile energies. For H_2^+, He^+ and Ne^+ ions the extraction (acceleration) voltage was limited to 20 keV as above this value an arc discharge was ignited between the

IPR2018-1557
HTC EX1023, Page 71

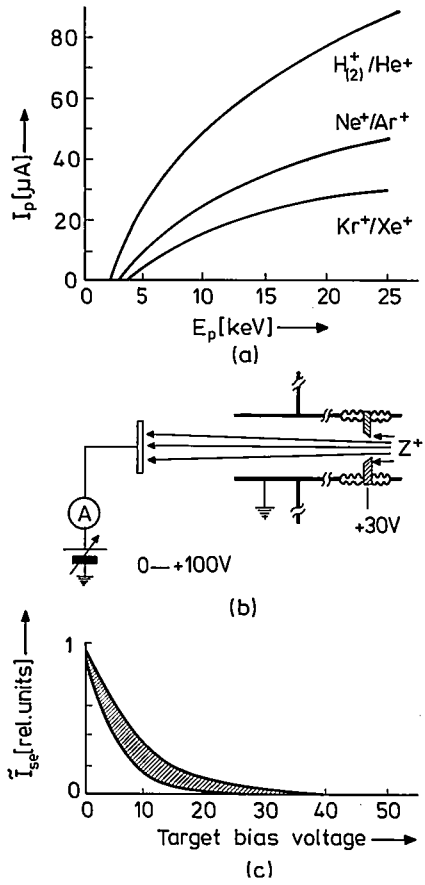


Fig. 1. *a*) Maximum attainable primary ion currents I_p as a function of extraction energy, measured with our beam apparatus. *b*) Sketch of the set-up for the electron yield determination. Details are discussed in the text. *c*) The relative secondary electron current as a function of target bias voltage ($\tilde{I}_{se} = [I(V_{bias}) - I(100)]/[I(0) - I(100)]$). All measurements, at 0(10)50 V, fall in the hatched area.

ion source and the gas supply vessel. For Ar^+, Kr^+ and Xe^+ ions 22.5 keV (and occasionally 25 keV) could be reached.

2.2. *Target preparation and handling*

Circular ($\varnothing = 18$ mm) targets were cut from 1–3 mm thick ultrapure (99.999%) optically flat metal platelets. The targets were etched in a mild acid and rinsed in chemically pure water and alcohol or hexane. Subsequently the targets were mounted on an oven, capable of heating to a temperature of ~ 800 °C, and inserted in the UHV chamber through a vacuum lock. Cleaning in situ could thus be achieved by annealing, by sputtering or both. Auger

IPR2018-1557
HTC EX1023, Page 72

Electron Spectroscopy (AES) was used to check the cleanliness of the targets on an atomic scale. In general, it was observed that annealing alone was insufficient and that sputtering to a total dose of about 1×10^{16} ions cm^{-2} with either Ar^+, Kr^+ or Xe^+ at 10–20 keV was needed to obtain atomically clean surfaces. In two cases trace impurity contamination of the order of 1% of a monolayer, viz. S in Ni and O in Mo, was present, even after sputter-cleaning. The invariance with dose of the total electron yield, measured simultaneously, provided in itself an accurate indication of the cleanliness of the surface.

No attempt was made to expel the implanted noble gas atoms from the target by a heat treatment in order to avoid outdiffusion of impurities from the bulk to the surface. It might be argued that loading of the target surface (i.e. to a depth of ~1 nm), to a steady-state concentration of the order of 5–10%, with noble gas atoms could influence the measured total secondary electron yield. We found, however, no differences for samples pre-saturated with the respective noble gas species. Moreover, for Au targets, where an annealing step was sufficient for cleaning, no measurable influence on the total fluence was observed within the experimental accuracy. Nevertheless it was decided always to employ a dose of at least 5×10^{16} ions cm^{-2} per ionic projectile *type* to ensure maximum reproducibility and steady-state conditions. This large dose prohibited the use of N^+ and N_2^+ ions as bombarding species because it appeared that in most cases, for example with Al and Cu targets, metal-nitride formation spoiled the experiment.

An 'in situ' Reflection High Energy Electron Diffraction (RHEED) facility was available to check that the targets were polycrystalline (i.e. showed no long-range ordering), even after prolonged exposure to the ion beam. As the fluences used were rather high, Scanning Electron Microscopy (SEM) was used, after completing the experiments on the most sputter-prone target materials (Zn, Pb), to make sure that pronounced bombardment-induced texturing (facet and/or cone formation) had not influenced the electron yield measurements.

2.3. *Yield determination method*

The set-up for the electron yield determination is shown schematically in fig. 1*b*. The beam hits the centre of the target to which a variable positive bias voltage of 0–100 V can be applied. Leakage currents caused by the bias voltage were normally negligibly small. At 0 V the total current measured is

$$I_0 = I_p + I_{se} + I_{si}, \tag{4}$$

where I_p is the primary beam current, I_{se} the current of secondary electrons emitted from the target and I_{si} the total current of (positive and/or negative) secondary heavy atomic and molecular ions ejected from the target during

IPR2018-1557
HTC EX1023, Page 73

bombardment and reflected primaries which have not undergone neutralization. At +100 V one measures approximately

$$I_{100} \simeq I_p + I_{si}, \tag{5}$$

as the vast majority of secondary electrons will return to the target, having energies well below 100 eV as can be seen in fig. 1*c*, and the secondary ion current will not be affected much. As this latter contribution is generally much less than 0.1% of the primary ion current, as is known from Secondary Ion Mass Spectrometry (SIMS) studies on atomically clean elemental targets, we find the total electron yield per incident projectile as

$$\gamma \simeq \frac{I_0}{I_{100}} - 1. \tag{6}$$

Formally the influence of other potentially feasible processes, that have so far been ignored in the evaluation of eq. (6), should be considered. Stray electrons, created as backscattered primaries or fast secondary ions hit the wall of the vacuum chamber, may be attracted to the target when it is biased at +100 V, and may even reach it incidentally in the absence of any bias. In the latter case the remedy is to apply a negative bias to the target [12]). No measurable effect was observed, however, so such a correction was not considered necessary. As the effective solid angle subtended by the target is small, the distance from the vacuum chamber wall being so large ($\gtrsim$15 cm), such ternary contributions could also be expected a priori to be negligible. A further conceivable source of error is the existence of fast neutrals in the beam. These do not show up in the measured primary ion current, but will nevertheless produce secondary electrons. The long flight path ($\simeq$ 1.5 m) strongly reduces the possibility of a sizeable neutral flux. In order to check this, a strong magnet was placed just in front of the beam-defining aperture to deflect the ion beam away from the target. No detectable secondary electron current was recorded. Finally the role of the beam current density was examined. No effect was found at any ion energy for fluxes in the range of about 1 $\mu A/cm^2$ to the maximum attainable current densities (see fig. 1*a*).

The overall relative uncertainty in the total electron yields measured in this way was estimated to be about 7%. Immediately after cleaning, measurements commenced at the highest energy and going down to the lowest and up and down once more. This resulted in three yield points per energy, which mutually agreed to within 3%, except for heavy noble gas ion bombardment of Zn and Pb where the spread turned out to be twice as high. In a few cases the experiments were repeated with another sample of the same metal. Agreement with earlier results of a similar kind was always observed. Occasional irreproducibility of the result led us to discard the entire measurement and after addi-

IPR2018-1557
HTC EX1023, Page 74

tional sputter-cleaning the measurement was performed again. As a consequence, the energy dependence of the electron yield for a particular projectile/target combination emerging from the sketched experimental procedure is expected to have a higher internal consistency than the accuracy with respect to a comparison with data sets obtained elsewhere.

3. Results and discussion

3.1. *Experimental data*

For brevity we restrict the presentation of the data obtained in this work largely to table I, for H^+ and H_2^+ ions, and table II, for the noble gas ions. Among the scarce results of older investigations available for comparison in our energy regime, the data by Carlston et al.[13]), for 1–10 keV Ne^+, Ar^+, Kr^+ and Xe^+ bombardment of Al, Cu, Ni, Mo and Ag single crystals at different surface orientations, provide the most complete and reliable data set. Their experiments were performed at 10^{-8} Torr with high beam current densities to keep the surface clean by sputtering. No additional chemical analysis of the surface composition was made. The close-packed (111) face of the fcc lattices is expected to resemble the polycrystalline situation best. Recent high-quality

TABLE I

Secondary electron yields (in electrons/ion) for H^+ and H_2^+ ions as obtained in the present work. Experimental errors are discussed in the text.

Target	Ion	E_i (keV)								
		5.0	6.25	7.5	8.75	10.0	12.5	15.0	17.5	20.0
Al	H^+	0.45	0.53	0.57	0.62	0.65	0.75	0.86	0.92	0.98
	H_2^+	0.50	0.59	0.69	0.71	0.76	0.84	0.95	1.02	1.07
Ti	H^+	0.45	0.48	0.55	0.57	0.60	0.67	0.70	0.77	0.83
	H_2^+	0.51	0.56	0.64	0.71	0.76	0.86	0.95	1.05	1.08
Ni	H^+	0.53	0.59	0.62	0.69	0.76	0.82	0.90	0.97	1.04
	H_2^+	0.68	0.76	0.86	0.93	0.98	1.09	1.17	1.26	1.35
Cu	H^+	0.63	0.67	0.74	0.79	0.81	0.87	0.91	0.99	1.05
	H_2^+	0.80	0.85	0.98	1.02	1.09	1.17	1.27	1.36	1.45
Mo	H^+	0.79	0.87	0.94	1.02	1.05	1.25	1.32	1.43	1.53
	H_2^+	1.00	1.12	1.21	1.35	1.46	1.63	1.80	1.95	2.10
Ag	H^+	0.86	1.00	1.10	1.19	1.29	1.43	1.52	1.61	1.69
	H_2^+	0.95	1.00	1.17	1.37	1.58	1.77	1.90	2.05	2.23
Au	H^+	0.79	0.94	1.09	1.15	1.20	1.32	1.46	1.50	1.60
	H_2^+	0.94	1.11	1.28	1.37	1.46	1.74	2.02	2.14	2.25
Pb	H^+	0.92	1.00	1.11	1.20	1.28	1.42	1.56	1.68	1.79
	H_2^+	1.00	1.16	1.37	1.46	1.61	1.79	2.00	2.20	2.38

IPR2018-1557
HTC EX1023, Page 75

TABLE II

Secondary electron yields (in electrons/ion) for the noble gas ions on Al, Ti, Ni, Cu, Zn, Mo, Ag, Au and Pb as obtained in the present work. Experimental errors are discussed in the text.

Target	Ion	E_i (keV)							
		5.0	7.5	10.0	12.5	15.0	17.5	20.0	22.5
Al	He^+	0.60	0.70	0.78	0.85	0.90	0.96	1.02	
	Ne^+	0.70	0.76	0.83	0.90	1.00	1.06	1.13	
	Ar^+	0.38	0.53	0.70	0.86	1.00	1.10	1.29	1.43
	Kr^+	0.20	0.26	0.35	0.40	0.49	0.55	0.63	0.70
	Xe^+	0.12	0.21	0.31	0.39	0.46	0.55	0.64	0.73
Ti	He^+	0.51	0.64	0.75	0.84	0.92	1.00	1.07	
	Ne^+	0.52	0.61	0.69	0.81	0.91	1.00	1.11	
	Ar^+	0.46	0.70	0.80	0.88	1.03	1.07	1.14	1.22
	Kr^+	0.16	0.29	0.41	0.61	0.82	0.95	1.11	1.28
	Xe^+	0.18	0.21	0.29	0.37	0.46	0.71	0.57	0.63
Ni	He^+	0.46	0.58	0.69	0.76	0.83	0.91	0.97	
	Ne^+	0.65	0.86	0.96	1.13	1.20	1.37	1.51	
	Ar^+	0.47	0.65	0.85	1.00	1.16	1.34	1.44	1.59
	Kr^+	0.36	0.51	0.66	0.83	0.96	1.12	1.27	1.42
	Xe^+	0.22	0.37	0.47	0.61	0.72	0.88	0.92	1.03
Cu	He^+	0.37	0.53	0.65	0.77	0.88	0.97	1.06	
	Ne^+	0.49	0.67	0.87	1.00	1.16	1.36	1.52	
	Ar^+	0.54	0.73	0.95	1.14	1.33	1.46	1.61	1.71
	Kr^+	0.46	0.63	0.83	1.05	1.24	1.47	1.66	1.86
	Xe^+	0.29	0.47	0.61	0.73	0.86	1.00	1.10	1.31
Zn	He^+	0.44	0.60	0.75	0.88	0.99	1.12	1.17	1.27
	Ne^+	0.53	0.75	0.93	1.12	1.27	1.49	1.69	
	Ar^+	0.42	0.63	0.79	0.91	1.07	1.12	1.31	1.45
	Kr^+	0.44	0.63	0.83	1.07	1.29	1.53	1.72	1.94
	Xe^+	0.18	0.27	0.36	0.45	0.55	0.64	0.72	0.81
Mo	He^+	0.65	0.80	1.03	1.14	1.28	1.39	1.50	
	Ne^+	0.62	0.82	0.95	1.08	1.20	1.32	1.42	
	Ar^+	0.47	0.58	0.77	0.92	1.06	1.21	1.36	1.50
	Kr^+	0.12	0.20	0.34	0.47	0.60	0.72	0.85	0.98
	Xe^+	0.15	0.26	0.38	0.51	0.62	0.74	0.85	0.97
Ag	He^+	0.60	0.83	1.09	1.24	1.47	1.60	1.70	1.79
	Ne^+	0.77	1.05	1.26	1.52	1.73	1.90	2.07	2.23
	Ar^+	0.30	0.42	0.57	0.67	0.79	0.91	1.04	1.16
	Kr^+	0.12	0.19	0.27	0.37	0.48	0.52	0.58	0.70
	Xe^+	0.24	0.37	0.56	0.68	0.89	1.04	1.22	1.39
Au	He^+	0.60	0.79	1.00	1.19	1.38	1.52	1.66	
	Ne^+	0.66	0.82	1.04	1.30	1.46	1.75	1.93	
	Ar^+	0.21	0.34	0.43	0.57	0.65	0.75	0.87	0.98
	Kr^+	0.14	0.17	0.20	0.28	0.31	0.37	0.42	0.46
	Xe^+	0.11	0.19	0.33	0.38	0.49	0.58	0.67	0.76
Pb	He^+	0.78	0.97	1.14	1.32	1.43	1.53	1.66	
	Ne^+	0.75	0.90	1.10	1.27	1.42	1.52	1.66	
	Ar^+	0.29	0.43	0.60	0.74	0.89	1.03	1.16	1.28
	Kr^+	0.09	0.15	0.20	0.28	0.35	0.45	0.52	0.60
	Xe^+	0.17	0.33	0.47	0.59	0.74	0.92	1.04	1.15

IPR2018-1557
HTC EX1023, Page 76

total electron yield results in the energy range discussed here stem almost exclusively from Baragiola and co-workers[1,14,15]). They made extensive studies of 2–50 keV light ion H_2^+, D_2^+ and He^+ bombardment induced secondary electron emission of Li, Al, Cr, Cu, Mo, Ag and Au[14]), and the projectile

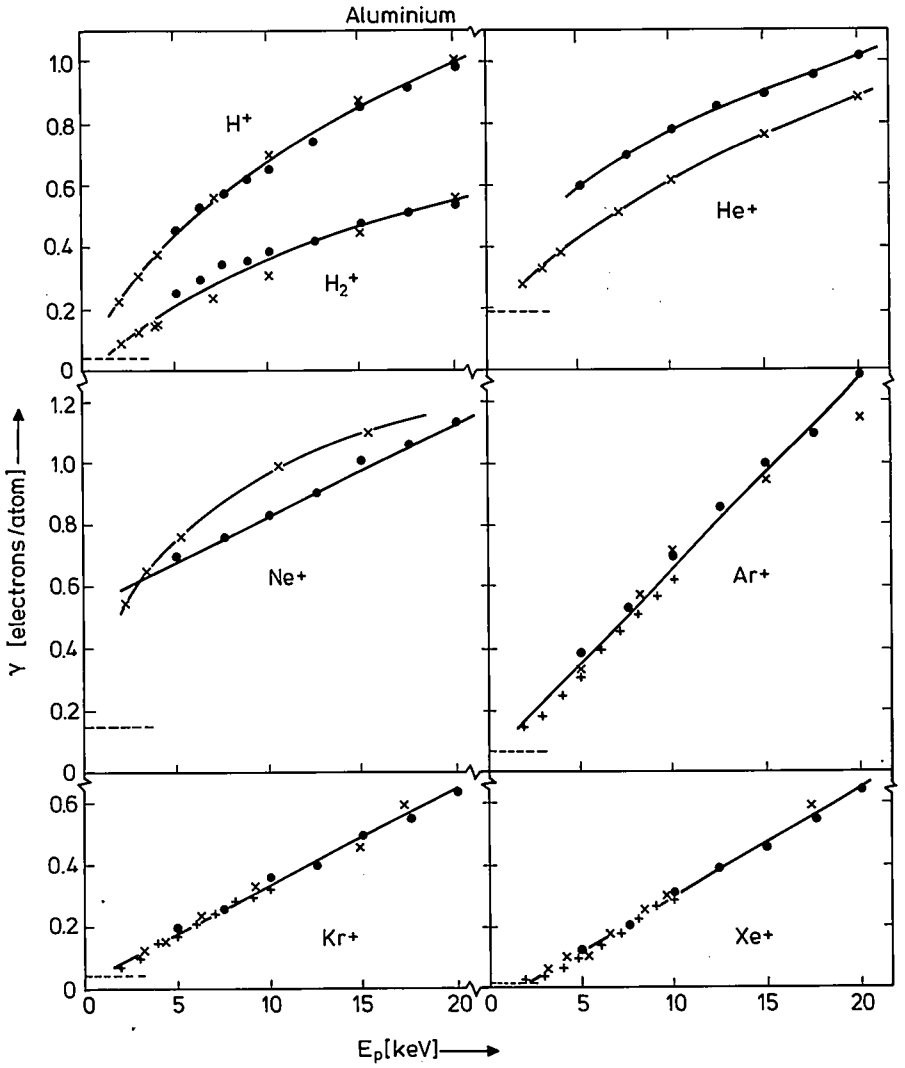


Fig. 2*a*). The energy dependence of the ion-bombardment-induced secondary electron yield for Al. Filled circles (•) this work; crosses (×) data from Baragiola and co-workers (refs 1, 14 and 15); plus signs (+) data from Carlston et al. (ref. 13) for the (111) crystal face. The curves drawn are intended only to guide the eye. The dashed line at low energy gives the potential electron emission contribution, calculated from eq. (1).

IPR2018-1557
HTC EX1023, Page 77

atomic number and energy dependence for Al targets[15]). These authors also reported some less complete results for noble gas ions on Cu, Mo, Ag and Au.

The energy dependences of the electron yields reported in this paper for Al and, as an example, for Cu targets, are shown in figs 2*a* and 2*b*, respectively, and compared with the above-mentioned earlier results. The calculated potential electron emission contribution is indicated to stress its relative (un)importance. We see on the whole excellent agreement among reported data sets. This proves that our relatively simple current measurement system is just as

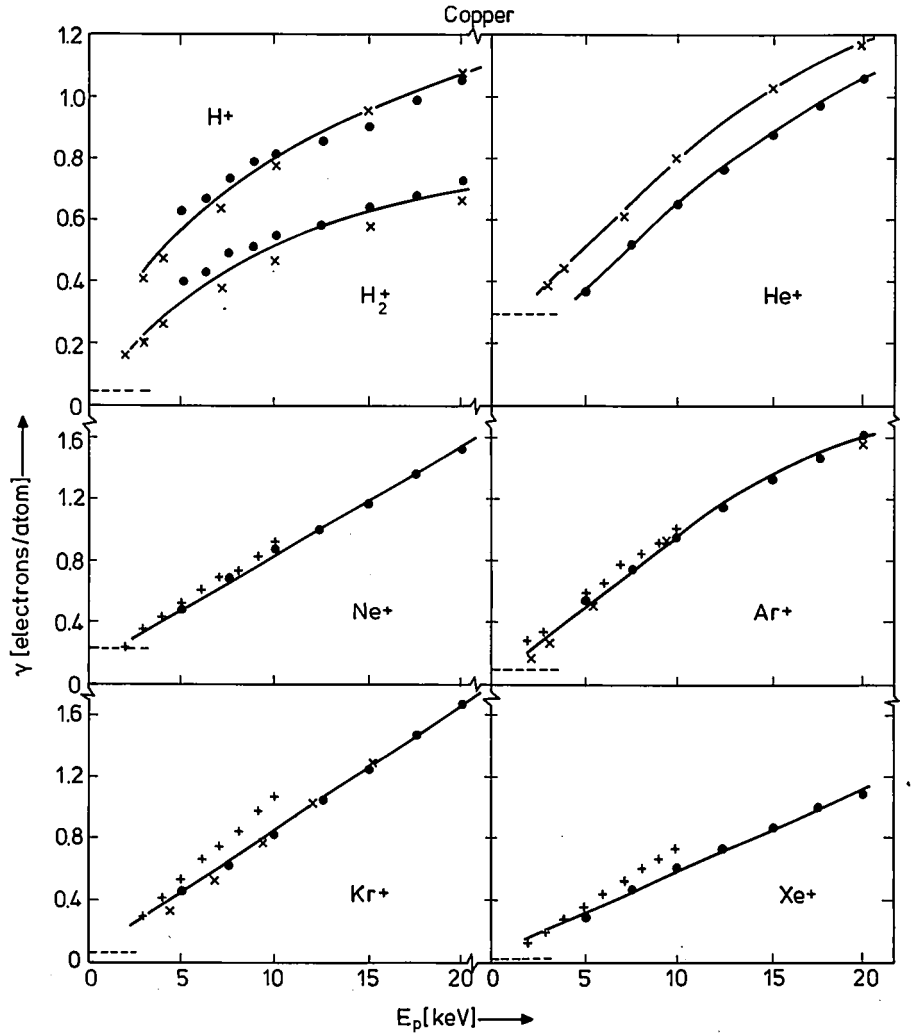


Fig. 2*b*. As fig. 2*a*), but now for Cu targets.

IPR2018-1557
HTC EX1023, Page 78

satisfactory as various more elaborate charge collection systems proposed[6,14,16]). Some additional remarks about the full comparison, which included elements not shown, are made below.

a) Our results agree with the data by Carlston et al.[13]) to within the experimental error, except in the case of Ni and Mo targets. For Ni, a difference in sulphur surface contamination (see also sec. 2), which could be higher in their experiments, is a likely source of this discrepancy.
b) While our H^+ and H_2^+ results for Al, Cu, Mo, Ag and Au confirm the results by Baragiola et al.[14]) the He^+ ion results are systematically about 10% lower except for Al, where they are 25% higher. The reason for this is not understood.
c) Comparing our Ar^+, Kr^+ and Xe^+ results on Cu, Mo, Ag and Au with those of Baragiola et al., perfect agreement is found except for Kr^+ ion bombardment of Au, where we find much higher yields at lower energy. Again no explanation can be given, but our results seem to be more regular.

No attempt has been made to construct figures displaying the functional dependence of the electron yields on the projectile or target atomic number. Such plots are rather sensitive to the bombarding energy at which the comparison is made. This obscures possible trends and systematics.

3.2. *Theoretical remarks*

The energy dependence of our electron yield results for light ion bombardment (H^+, H_2^+ and He^+) can be fitted very well with the expression $\gamma(E_p) = A_{pt}\sqrt{E_p} + B_{pt}$, with A_{pt} and B_{pt} as characteristic parameters for a certain projectile target combination. For the heavier noble gas ions the functional form $\gamma(E_p) = A_{pt}E_p + B_{pt}$ leads to the best results, except perhaps for the data for Ar^+ on Al and Ti. It should be noted that the division into two groups is not necessarily entirely meaningful as $\sqrt{E}$ can be approximated by a linear function of E to within 3% for E in the 5-20 keV range.

At first sight a square-root-like energy dependence of the light-ion electron yield looks promising from a theoretical point of view. In the energy regime studied the electronic stopping for the incident ion is proportional to velocity [$S_e(E_p) = K_{pt}\sqrt{E_p}$ see also discussion following eq. (3)], and yield contributions from recoiling target atoms will be negligible[2,10]). The 'correction factor' B_{pt} can be attributed to a combination of the potential electron emission contribution, discussed in sec. 1, and a term reflecting the threshold for direct knock-on electron liberation[14,15]), (i.e. $B \simeq \gamma_{PEE} - A\sqrt{E_{th}}$). Thus one would expect the ratio of the fitted constants A_{He^+}/A_{H^+} for a specific target to be equal to the ratio of the corresponding electron stopping-power prefactors (i.e. K_{He^+t}/K_{H^+t}), as any dependence on the angle-, energy- and depth-averaged escape probability

IPR2018-1557
HTC EX1023, Page 79

for liberated target electrons will drop out. Energy reflection is small and similar for H^+ and He^+ ions in the energy regime described here[17]), so it may be discarded in this discussion. Unfortunately this prediction was never fulfilled, even after a fit of the yield data to experimental stopping-power data[18]), which in particular for He^+ ions may deviate slightly from the $E^{\frac{1}{2}}$ dependence.

Our data substantiate the 'molecular' effect

$$R \equiv \frac{\gamma(H_2^+, E)}{2\gamma(H^+, E)} < 1, \tag{7}$$

discussed by Baragiola et al.[1,14]). We found on average $R = 0.90 \pm 0.03$, independent of energy, with the exception of $R(\mathrm{Al}) = 0.82$ and $R(\mathrm{Mo}) = 0.96$. If one assumes fragmentation of the impinging H_2 molecule in its two constituent H atoms, which each get equal energy, the theory of eq. (3) would predict $R = 1$. A slightly asymmetrical energy distribution over the two H atoms would not reduce the ratio R by more than 1–2% given a square-root-like energy dependence of the secondary electron yield. Even taking into account the difference in potential electron emission contributions for H_2^+ and two H^+ ions ($\gamma_{\mathrm{PEE}}(H_2^+) \simeq \gamma_{\mathrm{PEE}}(H^+) \simeq 0.05$, for the metals discussed here except Ni and Au where $\gamma_{\mathrm{PEE}} \simeq 0.01$; see also eqs (1, 1′), one cannot explain the observed discrepancy. Moreover with increasing energy the influence of the potential electron emission should diminish (i.e. $R \rightarrow 1$). As an explanation for the 'molecular' effect we therefore propose a higher particle reflection coefficient of (the fragments of) the H_2 dumbell molecule than for single H atoms. This suggestion is supported by the fact that for heavier molecular ions (e.g. N_2^+ and O_2^+) the ratio R differs less from unity.

As for the heavier noble gas ion results, we were unable to reconstruct the observed energy dependences for the various targets on the basis of Schou's theory (see sec. 1 and ref. 10). The experimentally determined variation with energy was always much stronger than could be accounted for by eq. (3). As a possible explanation we suggest that the factors β_0 and β_r are not as energy independent as was assumed. Also the averaged escape probability Λ is not independent of the energy spectrum of liberated target electrons, which in turn varies with projectile type and bombardment energy, so we assume the discrepancy can be resolved along these lines. Of course this further reduces the predictive power of eq. (3). A more qualitative, somewhat less stringent, test of Schou's theory can be carried out with our data for Cu and Zn. These two elements have nearly equal electronic stopping powers $S_e(E_p)$ for the projectiles studied here[18,19]). Moreover they have also almost equal mass and atomic number and consequently the factors $S_n(E_p)$, β_0, β_r and $\eta(\mu E_p)/\mu E_p$ will be approximately the same. Thus eq. (3) predicts that the ratio of the

IPR2018-1557
HTC EX1023, Page 80

secondary electron yields for Zn and Cu for a specific projectile would be a constant ($\gamma_{Zn}/\gamma_{Cu} = \Lambda_{Zn} N_{Zn}/\Lambda_{Cu} N_{Cu}$), independent of energy, as potential electron emission contributions are not materially different. This is excellently fulfilled by the experimental data from table II as shown in fig. 3. On the other

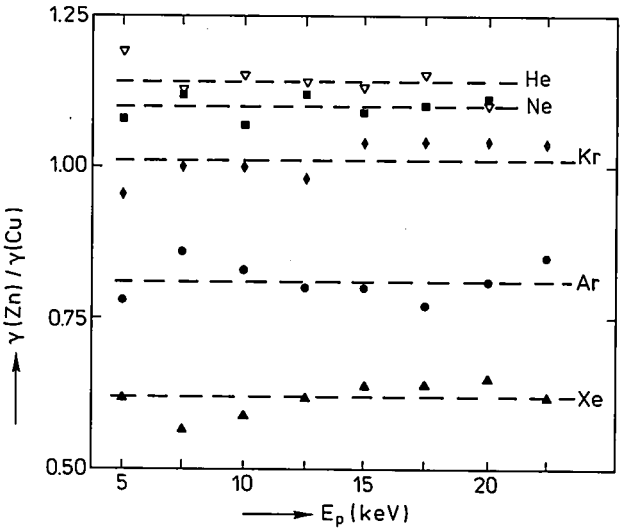


Fig. 3. The ratio of the ion induced secondary electron yields for Zn and Cu for different projectiles as a function of energy. According to the theory (see text) this ratio should be a constant independent of projectile type.

hand eq. (3) predicts that these constants are independent of projectile type, which is not true! Similar results emerge from the comparison of electron yields for the pair Ni/Cu, although only the data obtained with Ar^+, Kr^+ and Xe^+ ions can be used because for H^+, He^+ and Ne^+ either $S_e(E_p)$ or γ_{PEE} or both differ considerably. The ratios $\gamma_{Ni}/\gamma_{Cu} = 0.89, 0.78, 0.81$ for Ar^+, Kr^+ and Xe^+, respectively, differ much less than in the case of the Zn/Cu pair. The conclusion must be that the averaged electron escape probability can be decomposed in a universal (?) energy dependent part and a characteristic projectile/target factor. Presumably this latter quantity is at least strongly dependent on projectile properties.

Finally, one other observation is worth mentioning. According to calculations by Sigmund [20]), high-density atomic collision cascades can be expected for energetic (10–20 keV) heavy-ion bombardment of Zn. The occurrence of such, so-called, *spike effects* in Ag, Au and Pt have been correlated by Thompson [21]) with a pronounced nonlinearity of the sputtered particle yield. His data suggest, or are at least consistent with, an anomalous enhancement of the

IPR2018-1557
HTC EX1023, Page 81

elastic energy deposition. The latter quantity is normally proportional to $S_n(E_p)$. Thus, from Schou's theory one could also expect a steep increase in the ion-induced secondary electron yield for Zn bombarded by e.g. Xe^+ at the onset of the 'spike' regime (i.e. for $E_p \gtrsim 10$ keV). No such behaviour is found, however, in perfect agreement with the observations of Thum and Hofer[22]) who found no enhanced electron emission on stainless steel bombarded with V_n or $Nb_n (n = 1, 2, \ldots, 9)$ clusters at 12.5–25 keV. Nevertheless, there too the experimental conditions guaranteed very high-density atomic collision cascades. The conclusion must be that the contribution of elastic collision spikes to the secondary electron emission is negligible.

4. Conclusions

It has been shown that reliable ion induced secondary electron yields can be obtained from simple current measurements, provided adequate precautions have been taken to avoid target contamination. Conversely a decrease of the electron yield to its stationary value during sputtering may serve as a good indication that the ultimate level of cleanliness (or the steadystate contamination situation) for a particular target has been reached.

The almost 'universal' theoretical prediction that for very light ions the energy dependence of the electron yield is linearly proportional to the electronic stopping power is only seemingly fulfilled. A thorough confrontation with the data reveals a remnant lack of understanding of process details. For heavier ion bombardment with energies of the order of 10 keV, where the contribution of fast recoils will be appreciable, theoretical understanding of the experimental observations is even less complete. Some promising results have been discussed for the ratios of electron yield data for Zn and Cu, almost equal-mass targets with an identical stopping power behaviour, but even such a restricted test of the theory was not fully satisfactory. More experimental information on such pairs of 'similar' targets will be helpful to sort out the remaining problems.

REFERENCES

[1]) R. A. Baragiola, E. V. Alonso, J. Ferron and A. Oliva-Florio, Surface Sci. 90, 240 (1979).
[2]) P. Sigmund and S. Tougaard, in: Inelastic particle-surface collisions, eds: E. Taglauer and W. Heiland, Springer, Berlin, 1981; cf. p. 2.
[3]) K. H. Krebs, Vacuum 33, 555 (1983).
[4]) L. M. Kishinevskii, Rad. Eff. 19, 23 (1973).
[5]) H. D. Hagstrum, in: Inelastic ion-surface collisions, eds: N. H. Tolk et al., Academic Press, New York, 1977; cf. p. 1.
[6]) H. Oechsner, Phys. Rev. B17, 1052 (1977).
[7]) C. Kittel, Introduction to solid state physics, 5th ed., Wiley, New York, 1976.
[8]) J. Hölzl and F. K. Schulte, in: Solid surface physics, Springer, Berlin, 1979; cf. p. 1.

IPR2018-1557
HTC EX1023, Page 82

[9]) W. A. Harrison, Electronic structure and the properties of solids, W. H. Freeman and Co, San Francisco, 1980.
[10]) J. Schou, Phys. Rev. **B22**, 2141 (1980).
G. Holmén, B. Svensson, J. Schou and P. Sigmund, Phys. Rev. **B20**, 2247 (1979).
[11]) G. E. Thomas, L. J. Beckers, J. J. Vrakking and B. R. de Koning, J. Cryst. Growth **56**, 557 (1982).
[12]) E. Veje, Nucl. Instr. and Meth. **194**, 433 (1982).
[13]) C. E. Carlston, G. D. Magnuson, P. Mahadevan and D. E. Harrison, Phys. Rev. **139**, A729 (1965).
[14]) R. A. Baragiola, E. V. Alonso, O. Auciello, J. Ferrón, G. Lantschner and A. Oliva-Florio, Phys. Lett. **67A**, 211 (1978);
R. A. Baragiola, E. V. Alonso and A. Oliva-Florio, Phys. Rev. **B19**, 121 (1979);
J. Ferrón, E. V. Alonso, R. A. Bargiola and A. Olivia-Florio, J. Phys. D **14**, 1707 (1981).
[15]) E. V. Alonso, R. A. Baragiola, J. Ferrón, M. M. Jakas and A. Oliva-Florio, Phys. Rev. **B22**, 80 (1980).
[16]) G. D. Magnuson and C. E. Carlston, Phys. Rev. **129**, 2403 (1963).
[17]) O. S. Oen and M. T. Robinson, Nucl. Instr. and Meth. **132**, 647 (1976).
[18]) H. H. Andersen and J. F. Ziegler, Hydrogen stopping powers in all elements, Pergamon, New York, 1977 and
J. F. Ziegler, Helium stopping powers and ranges in all elemental matter, Pergamon, New York, 1977.
[19]) D. J. Land and J. G. Brennan, Atomic data and nucl. data tables **22**, 235 (1978).
[20]) P. Sigmund, Appl. Phys. Lett. **25**, 169 (1974).
[21]) D. A. Thompson, J. Appl. Phys. **52**, 982 (1981).
[22]) F. Thum and W. O. Hofer, Surface Sci. **90**, 331 (1979).

IPR2018-1557
HTC EX1023, Page 83

Philips J. Res. 39, 77-81, 1984 *R 1082*

THE CRYSTAL STRUCTURE OF LaNiSn

by J. L. C. DAAMS and K. H. J. BUSCHOW

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

The crystal structure of the compound LaNiSn has been determined; it is found to be homotypic with the orthorhombic oP12 (C23) structure. LaNiSn is isotypic with ε-TiNiSi and shows great resemblance with the structure of $CeCu_2$.

PACS numbers: 61.55.Hg.

1. Introduction

The growing interest being shown in the study of ternary intermetallic compounds is due to particular properties of ternary compounds not found in binary intermetallics. These properties comprise, for instance, the occurrence of superconductivity and magnetic ordering in one and the same compound [1]), the observation of huge values for the magneto-optical Kerr rotation [2]) and a pronounced first-order spinflop transition with a transition hysteresis of more than 30 k Oe [3]). Ternary compounds have also attracted much attention from the theoretical point of view [4,5]). The Cl_b-type Heusler compounds, in particular, were found to have unusual features in their band structure, leading to 100% spin polarization of the conduction electrons.

In a previous investigation [6]) the occurrence of Heusler-type compounds based on Sn was studied. It was found that the combination of Sn and 3d elements with rare earths did not lead to the formation of such ternary compounds, but that other ternary compounds were formed instead. In the present investigation we have studied the crystal structure of the compound LaNiSn.

2. Experimental procedures and results

The sample was prepared by arc melting in purified argon gas, followed by vacuum annealing for two weeks at 800 °C. After annealing the sample was pulverized and stress released by heating in vacuum at 500 °C for three hours. *X*-ray diagrams were obtained by means of a Philips PW 1700 diffraction

IPR2018-1557
HTC EX1023, Page 84

system. CuKα radiation was used in combination with a graphite monochromator.

We indexed the X-ray diffraction diagram of LaNiSn on the basis of a primitive orthorhombic unit cell with the lattice constants $a = 7.671$ Å, $b = 4.652$ Å and $c = 7.592$ Å. From the systematic extinctions we derived

TABLE I.

Observed and calculated diffraction angles θ (in degrees) and intensities I (in arbitrary units) of the intermetallic compound LaNiSn.

			observed		calculated	
h	k	l	θ	I	θ	I
1	0	1	8.21	2	8.21	5.4
0	1	1	11.20	4	11.20	8.7
2	0	0	11.61	0.5	11.59	0.6
0	0	2	11.74	1	11.71	4.4
1	1	1	12.63	11	12.63	11.9
2	0	1		0.5	13.00	0.1
1	0	2	13.09	15	13.08	14.4
2	1	0	15.09	13.5	15.09	13.1
2	1	1			16.22	75.4
1	1	2	16.24	128	16.29	52.8
2	0	2	16.60	50	16.59	46.3
3	0	1	18.53	50	18.53	44.3
1	0	3	18.70	18	18.69	14.3
2	1	2		0.5	19.27	0.0
0	2	0	19.34	30	19.34	40.8
0	1	3	20.27	41	20.27	39.3
3	1	1	21.00	2.5	21.00	1.3
1	2	1			21.14	2.0
1	1	3	21.12	1.5	21.15	0.0
3	0	2		0.5	21.30	0.0
2	0	3	21.38	3	21.39	3.3
2	2	0		0.5	22.78	1.7
0	2	2	22.88	1	22.85	3.0
3	1	2	23.53	6.5	23.53	6.1
2	2	1			23.60	0.1
2	1	3			23.61	7.6
1	2	2	23.64	11.5	23.65	4.5
4	0	0			23.68	0.9
0	0	4	23.96	1	23.94	0.0
4	0	1	24.47	10.5	24.47	2.4
1	0	4	24.67	1.5	24.71	1.3

IPR2018-1557
HTC EX1023, Page 85

Pnma (62) [7]) as the most probable space group. In table I the indexing together with the observed reflecting angles are compared with the calculated angles.

3. Structure determination

The crystal structure was determined by using the trial structure obtained by means of a cell-filling program [8]). After seven refinement cycles we reached convergence with an ultimate reliability factor $R = 13.92\%$ based on intensities. The atomic position parameters are:

4 Sn in 4 (*c*)	$x = 0.185$	$y = 0.750$	$z = 0.070$
4 Ni in 4 (*c*)	$x = 0.312$	$y = 0.750$	$z = 0.452$
4 La in 4 (*c*)	$x = 0.478$	$y = 0.250$	$z = 0.191$

The value of B occurring in the expression of the Debije-Waller temperature factor was found to be equal to $2.6 \cdot 10^{-16}$ cm^2 for the Sn sites, $4.4 \cdot 10^{-16}$ cm^2 for the Ni sites and $12.4 \cdot 10^{-16}$ cm^2 for the La sites. Table I also gives a comparison of the observed intensities and the calculated intensities. In table II the interatomic distances up to 4 Å are given. A schematic representation of the structure is given in fig. 1, where the viewing direction is almost perpendicular to the (100) plane.

TABLE II

Number of neighbours and interatomic distances in LaNiSn.

atom	neighbour	number	distance (Å)
Sn	Sn	2	3.825
Ni	Ni	2	3.781
Ni	Sn	2	2.494
		1	2.871
		1	3.055
La	La	2	3.737
		2	3.938
La	Ni	1	2.869
		1	3.154
		2	3.309
		2	3.626
La	Sn	1	3.137
		1	3.259
		2	3.350
		2	3.360

IPR2018-1557
HTC EX1023, Page 86

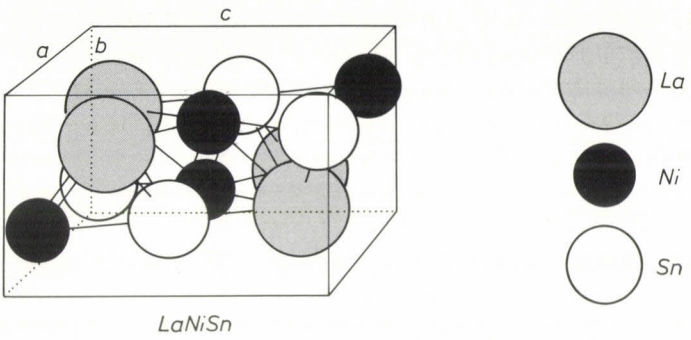


Fig. 1. Schematic representation of the crystal structure of LaNiSn.

4. Discussion

The formation of the compound LaNiSn is accompanied by a substantial contraction of the atomic volumes compared to the pure starting materials (36.9 Å^3, 10.9 Å^3 and 26.9 Å^3 for La, Ni and Sn respectively). From the lattice constants of LaNiSn given above one finds a volume per unit cell equal to 270.9 Å^3, which is more than 9% smaller than the value calculated on the basis of the atomic volumes of the elements.

The crystal structure of LaNiSn can be regarded as a distortion of the hexagonal Ni_2In structure [9]). As in the other members of the RNiSn compounds, the distortion is relatively strong, due to the large difference in metallic radii between the constituent metal atoms. Our values for the lattice constants are in good agreement with those given by Dwight [10]). Dwight reports that his value for *a* falls significantly below that expected on the basis of the linear relationship between the values of *a* and the radii of the *R* component in the remainder of the *R*NiSn compounds. By contrast, the *a* value found by us is in good agreement with the linear relationship mentioned. Closer inspection of Dwight's data showed that the values for *a* and *c* of LaNiSn are apparently interchanged in his plot, so that deviations from the linear behaviour remain restricted to the *b* and the *c* axis.

If one compares the atomic positions found in the course of the present investigation for LaNiSn with those of TiNiSi [11,12]) one notices substantial differences between the two compounds. Owing to this difference in the atomic position, a slight shift in coordination of the atoms in LaNiSn has taken place with respect to TiNiSi. In the latter compound the Ti, Ni and Si atoms are coordinated by 15, 12 and 9 atoms respectively. As can be derived from the data listed in table II, in LaNiSn the coordination of La, Ni and Sn corresponds to 16, 10 and 10, respectively. These changes in coordination are in

IPR2018-1557
HTC EX1023, Page 87

accordance with the metallic radii of the constituent elements. The La and Sn atoms are larger than the Ti and Si atoms, respectively. For the former atoms the coordination number has increased. By contrast, Ni in LaNiSn is surrounded by larger atoms than Ni in TiNiSi. For Ni therefore the coordination number has decreased. However there is a more striking resemblance between the LaNiSn structure and the structure of $CeCu_2$. $CeCu_2$ has a bodycentered orthorhombic unit cell with lattice constants $a = 4.425$ Å, $b = 7.057$ Å and $c = 7.475$ Å, space group 74 Imma [12-14]).

REFERENCES

1) J. R. Remeika, G. P. Espinosa, A. S. Cooper, H. Barz, J. M. Rowerll, D. B. McWan, J. H. Vandenberg, D. E. Moncton, Z. Fisk, L. D. Woolf, H. C. Hamaker, M. B. Maple, G. Shirane and W. Thomlinson, Solid State Commun. **34**, 923 (1980).
2) P. G. van Engen, K. H. J. Buschow, R. Jongebreur and M. Erman, Appl. Phys. Lett. **42**, 202 (1983).
3) T. T. M. Palstra, J. A. Mydosh, G. J. Nieuwenhuys, F. R. de Boer and K. H. J. Buschow (to be published).
4) J. Kübler, A. R. Williams and C. B. Sommers, Phys. Rev. (B) **28**, 1745 (1983).
5) R. A. de Groot, F. M. Mueller, P. G. van Engen and K. H. J. Buschow, Phys. Rev. Lett. **50**, 2024 (1983).
6) P. G. van Engen, K. H. J. Buschow and M. Erman, J. Magn. Mat. **30**, 374 (1983).
7) N. F. M. Henry and K. Lonsdale (eds), International tables for crystallography, Vol. I, Kynoch Press, Birmingham, 1952.
8) W. P. J. Fontein, J. M. M. Verbakel and J. H. N. van Vucht, Philips J. Res. **34**, 238 (1979).
9) W. B. Pearson, The crystal chemistry and physics of metals and alloys, John Wiley and Sons Inc., New York, 1972; cf. p. 532-534.
10) A. E. Dwight, J. Less-Common Met. **93**, 411 (1983)
11) C. B. Shoemaker and D. P. Shoemaker, Acta Cryst. **18**, 900 (1965).
12) W. B. Pearson, A handbook of lattice spacings and structures of metals and alloys, Vol. 2, Pergamon Press, 1965; cf. p. 1142-1143.
13) A. C. Larson and D. T. Cromer, Acta Cryst. **14**, 73 (1961).
14) E. Hovestreydt, N. Engel, K. Klepp, B. Chabot and E. Parthé, J. Less-Common Met. **85**, 247 (1982).

IPR2018-1557
HTC EX1023, Page 88

Philips J. Res. 39, 82-93, 1984 *R 1083*

MAGNETO-OPTICAL PROPERTIES OF RARE EARTH COBALT COMPOUNDS AND AMORPHOUS ALLOYS

by K. H. J. BUSCHOW and P. G. VAN ENGEN

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

The polar Kerr rotation spectra of Co-rich intermetallic compounds in which Co is combined with one of the rare earth elements La, Y or Gd were investigated at room temperature. The most prominent feature of the La–Co and Y–Co compounds is the reduction in Kerr intensity associated with the reduction in magnetization with increasing rare earth content. The spectra of the Gd–Co compounds and amorphous alloys consist of two separate contributions due to Gd and Co. The effect of surface oxidation is briefly discussed.

PACS numbers: 78.20.L.

1. Introduction

Amorphous thin films composed of Gd and either Co or Fe have attracted much interest owing to their favourable magneto-optical and thermomagnetic properties, making them good starting materials for erasable magneto-optical recording devices [1,2]). While the role of the Gd atoms is well established as regards the magnetic and thermomagnetic properties, it is still not quite clear in how far the Gd atoms contribute to the magneto-optical properties. In order to be able to evaluate the influence of the Gd atoms on the Kerr rotation spectra we investigated the magneto-optical properties of several Gd–Co intermetallic compounds and compared them with those of similar rare earth compounds in which the rare earth component consists of nonmagnetic La or Y instead of Gd. Conclusions arrived at from these investigations were then used to interpret the Kerr rotation spectra of amorphous Gd–Co alloys.

2. Experimental

The crystalline samples were prepared by means of arc melting, using an atmosphere of purified argon and starting from 99.9 pure metals. After arc

IPR2018-1557
HTC EX1023, Page 89

melting the samples were wrapped in Ta foil and vacuum annealed in the temperature range 900-1000 °C for about 10 days. X–ray diffraction was used to establish whether the annealing had resulted in single-phase samples. The amorphous $Gd_{1-x}Co_x$ films were prepared by means of two-source evaporation onto quartz substrates 1 cm in diameter. The film thickness was about 200 nm. Composition and homogeneity of the samples were controlled by independent monitoring of the evaporation rates of the two electron gun sources, using quasi-continuously operating feedback-stabilized Airo Temescal film deposition controllers (TDC-700-l). The base pressure of the high-vacuum system was equal to 5×10^{-8} Torr. Thickness calibration was made on pure crystalline films, using Talystep techniques.

The polar Kerr rotation ($2\varphi_K$) and the ellipticity ($2\varepsilon_K$) were measured at room temperature in the range 0.6–4.4 eV, with an accuracy better than 0.005 degrees over most of the energy range considered. A Faraday modulator was used for polarization modulation, allowing phase-sensitive detection of the optical output signal. The Kerr ellipticity was determined by using a number of quarter-wave plates, the retardation of which had been carefully measured as a function of the wavelength. The magneto-optical measurements were made on the film side of the samples prepared by vapour deposition. Measurements on the crystalline samples were made on flat portions, polished with diamond paste. The magnetization direction was kept perpendicular to the mirror surface during the magneto-optical measurements by means of an external field of 920 kA/m. The mean angle of incidence of the light beam on the sample surface was smaller than 4.5 °C.

3. Experimental results

Comparison of the magneto-optical spectra obtained on several crystalline La–Co samples with the spectrum of Co metal in fig. 1 shows that combining La with Co has a considerable effect on the Kerr spectra. In $LaCo_{13}$ the first $|2\varphi_K|$ peak near 1.3 eV has shifted to lower energies and increased in intensity relative to pure Co metal. The second peak near 3.8 eV has also shifted to lower energy, although here a reduction in intensity has occurred. In the spectra of the two compounds $LaCo_5$ and La_2Co_7 the most obvious change compared with Co metal is the strong decrease in intensity of the first peak. The decrease in intensity of the high-energy peak is much less strong.

Experimental results for the Y–Co compounds are shown in fig. 2. At Co concentrations lower than corresponding to Y_2Co_{17} there is a general broadening of the two peaks. In all compounds the intensity has decreased relative to Co metal, although the general shape of the spectra has remained more or less the same.

IPR2018-1557
HTC EX1023, Page 90

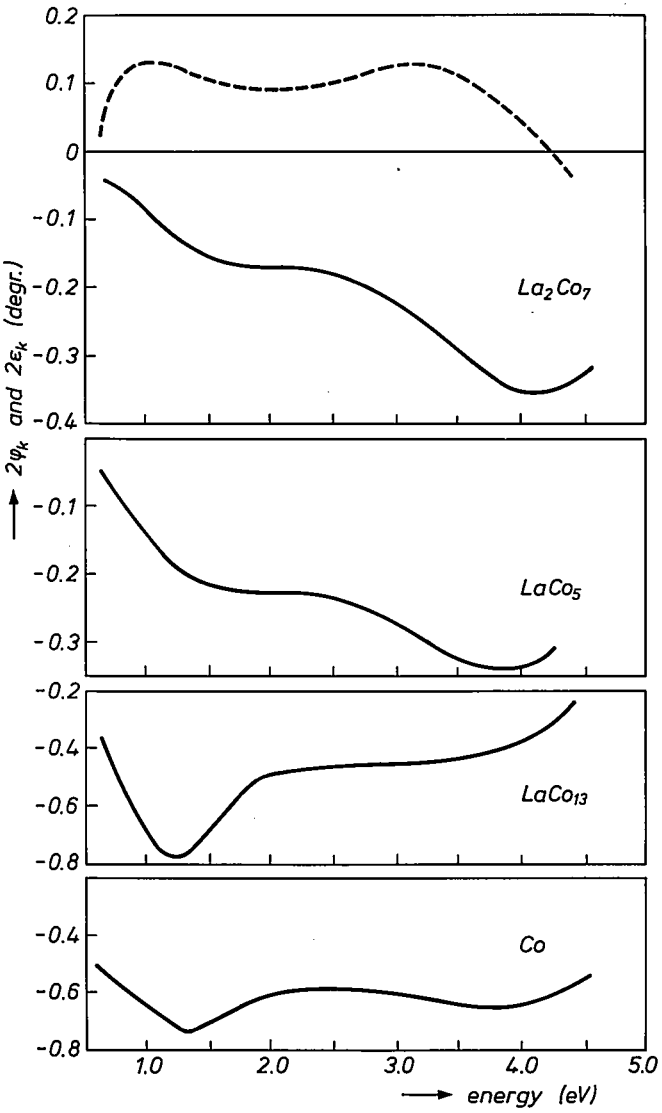


Fig. 1. Energy dependence of the Kerr rotation $2\varphi_K$ in various La–Co compounds. For the compound La_2Co_7 the energy dependence of the ellipticity is also given ($2\varepsilon_K$, broken line).

Kerr spectra of Gd–Co compounds are shown in fig. 3. It follows from the magnetic properties of these compounds that there is antiparallel coupling of the Co moments and the Gd moments. With decreasing Co concentration the relative strength of the Co sublattice magnetization decreases. At room tem-

IPR2018-1557
HTC EX1023, Page 91

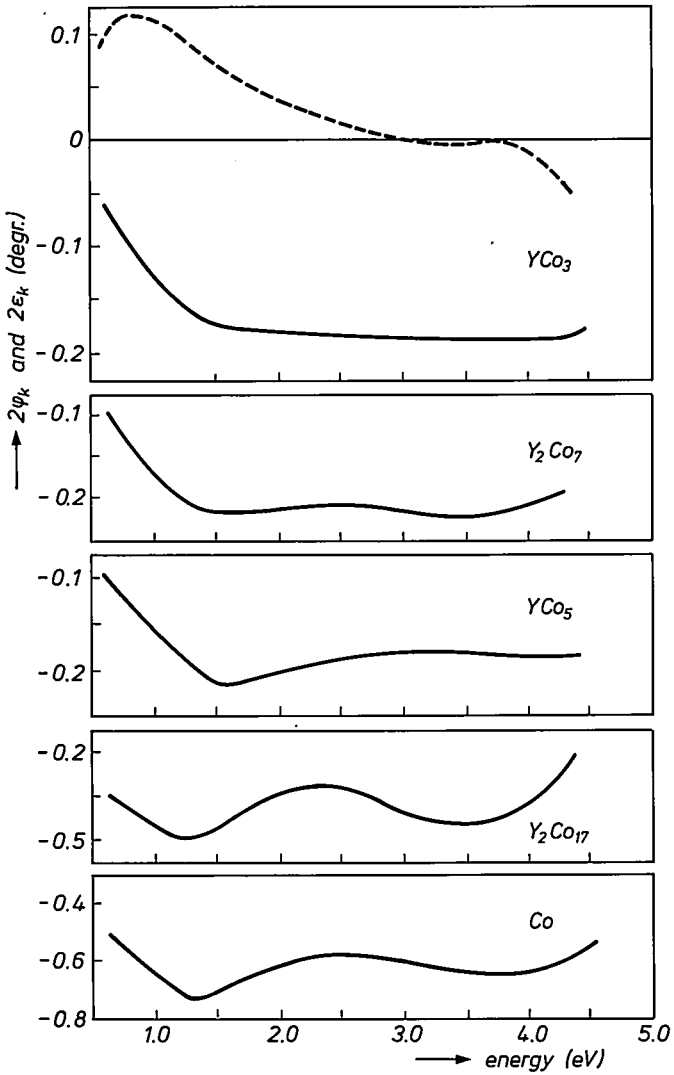


Fig. 2. Energy dependence of the Kerr rotation $2\varphi_K$ in various Y–Co compounds. For the compound YCo_3 the energy dependence of the ellipticity is also given ($2\varepsilon_K$, broken line).

perature the total magnetization in Gd_2Co_{17} and $GdCo_5$ is still dominated by the Co sublattice magnetization ($M_{tot} = M_{Co} - M_{Gd}$) so that the Co moments remain parallel to the external field. The situation has changed in $GdCo_3$ and $GdCo_2$. Here bulk magnetic measurements showed that the Gd sublattice magnetization dominates at room temperature ($M_{tot} = M_{Gd} - M_{Co}$), the Co

IPR2018-1557
HTC EX1023, Page 92

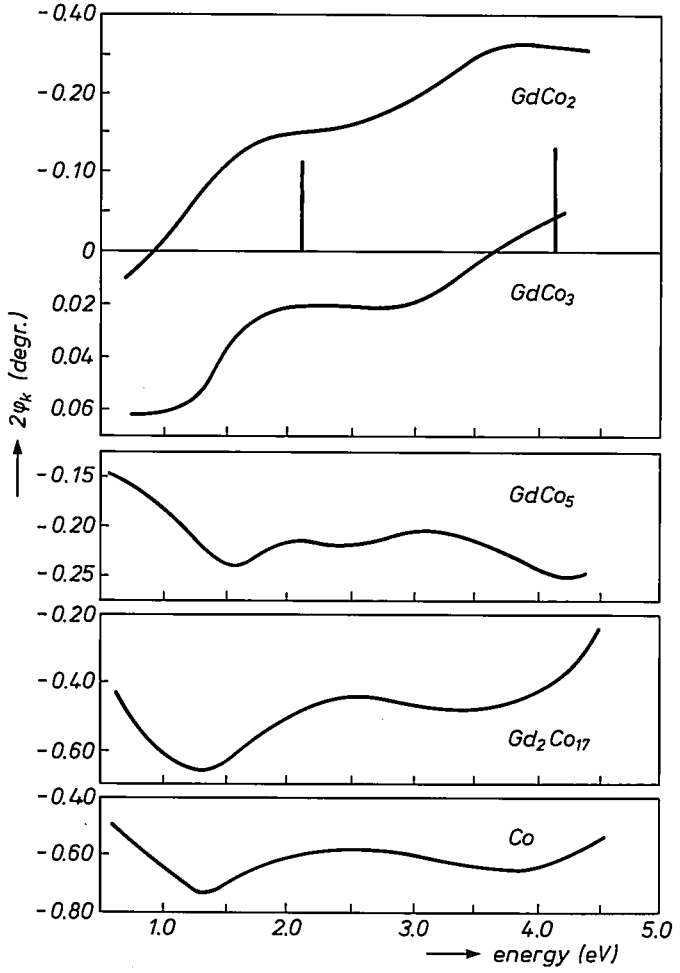


Fig. 3. Energy dependence of the Kerr rotation $2\varphi_K$ in various Gd–Co compounds. For the compounds $GdCo_3$ and $GdCo_2$ the negative (positive) values associated with the $2\varphi_K$ axis have been plotted in the upward (downward) direction. For all materials considered in this figure the Co moments are parallel to the external field and their $2\varphi_K$ contribution is in the downward direction. The Gd moments are antiparallel to the external field and the corresponding $2\varphi_K$ contribution is in the upward direction. The locations of the two main peaks observed in Gd metal by Erskine and Stern [8]) have been included schematically in the top part of the figure (in arbitrary units).

moment being oriented antiparallel to the external field. In order to facilitate a comparison of the spectra of the different compounds we have reversed the direction of the vertical axis in fig. 3. When plotted in this way, all the Co moments in the compounds considered have the same (parallel) orientation relative to the external field. It will be clear from the results shown that in-

IPR2018-1557
HTC EX1023, Page 93

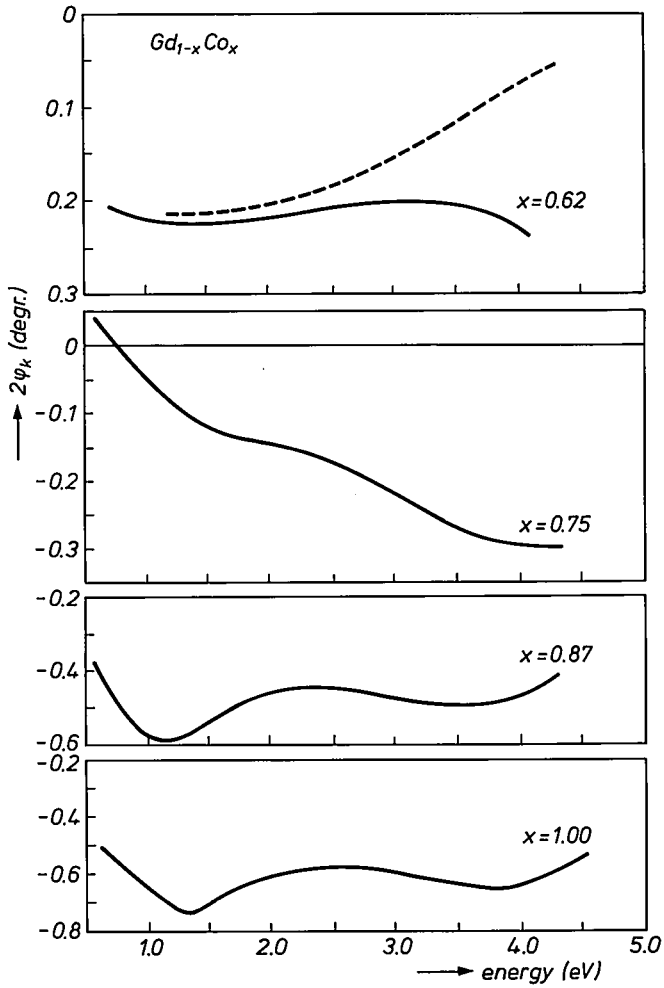


Fig. 4. Concentration dependence of the Kerr rotation $2\varphi_K$ observed in several amorphous $Gd_{1-x}Co_x$ alloys. (The contribution to $2\varphi_K$ of the Co atoms is in the downward direction in all cases; see fig. 3.) The broken line reflects the result of measurements obtained after keeping the sample in air for an extended period.

creasing Gd concentration eventually leads to a completely different type of Kerr spectra.

Kerr spectra obtained on thin vapour-deposited Gd-Co films are shown in fig. 4. The room temperature magnetization changes from Co-dominated to Gd-dominated at a concentration between 75 and 67 at% Co. This can be inferred, for instance, from the sign reversal of the field dependence of the

IPR2018-1557
HTC EX1023, Page 94

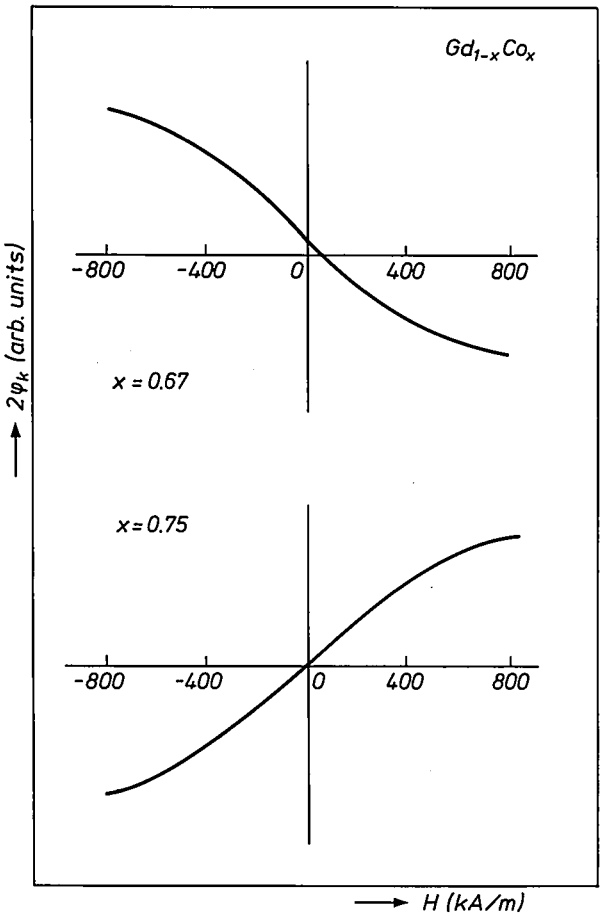


Fig. 5. Field dependence of the Kerr rotation in two amorphous $Gd_{1-x}Co_x$ alloys.

Kerr rotation, as is shown in fig. 5. For comparative purposes we have therefore again interchanged the directions of the vertical axes of the spectrum of the alloy richest in Gd (top part of fig. 4). Comparing the results of fig. 3 and fig. 4 one may notice that the change of the Kerr spectra with increasing Gd concentration is less marked in the case of the amorphous alloys.

In order to obtain an impression of the effect of changes in the magneto-optical properties of the thin films when kept in air we re-examined the Gd-richest film after a period of one week. The Gd-richest film was chosen because the tendency to deteriorate increases with rare earth concentration. The results of the re-examination are included as a broken line. A marked decrease

IPR2018-1557
HTC EX1023, Page 95

of the Kerr intensity occurs primarily at the high energy side of the spectrum while the changes at the low energy side are only very modest.

4. Discussion

4.1. *Comparison with literature data*

The magneto-optical properties of thin films of the compounds Nd_2Co_7, $NdCo_5$ and of Co metal were investigated by Stoffel in the limited energy range from 1.8 eV to 3.1 eV as long ago as 1968 [3]). Stoffel found that increasing rare earth content leads to a decrease of the longitudinal Kerr effect. These findings are in agreement with the results displayed in the various figures.

Katayama and Hasegawa measured the polar Kerr spectrum of $GdCo_5$ on a single crystal [4]). In the low energy range their spectrum is similar to that observed by us. But in the high energy range (4–5 eV) these authors observed a strong increase of $2\varphi_K$ reaching the value $2\varphi_K = 0.7$ degr at about 4.7 eV. Since Katayama and Hasegawa did not use a protective coating on their samples and performed their measurements in air (as we did) it cannot be excluded that the discrepancy with our result (see fig. 3) reflects a different degree of surface oxidation which becomes prominent at high photon energies in particular (see also the discussion in section 4.2). However, it is possible that the large $2\varphi_K$ value found by them corresponds to a particular crystallographic direction, since the hexagonal compound is strongly anisotropic.

Results obtained on amorphous Gd–Co films by various authors are reproduced in fig. 6 after adapting the sign convention of φ_K to that used in the present report. There is some ambiguity as to the sign convention used by Visnovsky et al.[5]). These authors claim to apply the same sign convention as that used by Kahn et al.[6]). After checking the numerous equations contained

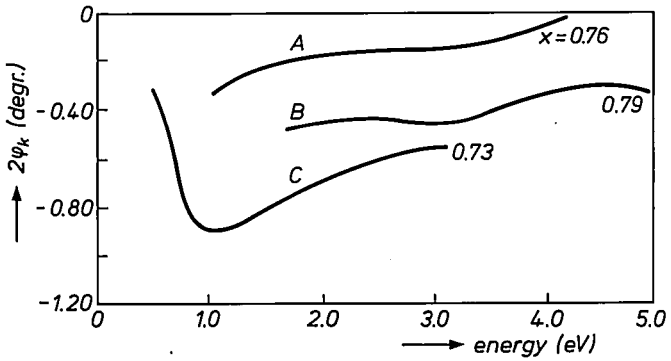


Fig. 6. Comparison of $2\varphi_K$ spectra of several amorphous $Gd_{1-x}Co_x$ alloys reported in the literature (A: ref. 5; B: ref. 4; C: ref. 2). For the sign convention of $2\varphi_K$, see text.

IPR2018-1557
HTC EX1023, Page 96

in the article by Kahn et al. we have the impression that their sign convention is similar to ours. However, from the discussion given by Visnovsky et al. it follows that $2\varphi_K$ for Gd and Co is positive in the energy range considered, making their sign convention opposite from that used here. The latter assumption regarding their sign convention was used in reproducing their data in fig. 6 (curve *A*).

A comparison of the spectra in fig. 6 with the results given in fig. 4 shows that good agreement with our data is found only with curve *B*, after taking into consideration that the concentration corresponding to curve *B* is slightly lower than that corresponding to the alloy with $x = 0.83$ in fig. 4. In part the discrepancy between our data and curves *A* and *C* may be due to the fact that all the films of fig. 6 were obtained by sputtering from targets of markedly lower Co concentration than the film composition stated in fig. 6, while a large bias voltage was applied. The latter fact, in particular, is known to lead to substantial changes of the physical properties in thin amorphous rare-earth transition metal films.

4.2. *Interpretation of the spectra*

The results shown in fig. 2 for various Y–Co compounds are rather well suited for evaluating the changes of the Co contribution to the Kerr spectra in compounds of Co with rare earth elements. Since Y has an atomic size and an electronegativity similar to those of the lanthanide elements (*R*) one may expect these changes to be the same in *R*–Co compounds and Y–Co compounds. This similarity in behaviour is reflected also in the similarity of structural and magnetic properties. The advantage of Y in the description of magneto-optical spectra is based on the fact that Y is nonmagnetic and magneto-optically rather inactive in view of the small spin-orbit splitting of this element. Accordingly, the results shown in fig. 2 behave more or less as one would have expected from the reduction in magnetization associated with the compound formation. There is some indication that the two $|2\varphi_K|$ maxima present in Co tend to approach each other in going from Co to YCo_3. Very likely this might be the result of a decrease in 3d band width due to an increasing mean Co–Co separation in the various compounds considered.

Inspection of the results presented in fig. 1 shows that the situation in the La–Co compounds is not that simple. Since $LaCo_{13}$ has a peculiar crystal structure we will focus our attention first on the compounds $LaCo_5$ and La_2Co_7. Comparison with Co and with the corresponding Y–Co compounds shows that the intensity of the low-energy $|2\varphi_K|$ maximum has decreased to about the same extent in both series. This part of the spectra therefore behaves in a normal way. However, it can be seen from fig. 1 that there is an ad-

IPR2018-1557
HTC EX1023, Page 97

ditional strong intensity in the 4 eV region of the spectra which is not expected to arise from the Co atoms (compare with Y_2Co_7 and YCo_5 in fig. 5). This additional contribution increases in going from $LaCo_5$ to La_2Co_7 and is ascribed to the La atoms. As the spin-orbit coupling of Y is lower than that of La, this additional contribution will not be present in the spectra of the Y compounds.

Alternatively one may present arguments showing that of all the materials considered in this paper the compounds $LaCo_5$ and La_2Co_7 are likely to give rise to some surface oxidation. As will be shown below, this will lead to an enhanced negative contribution at photon energies corresponding to the second $2\varphi_K$ peak of Co metal and be responsible for the additional intensity mentioned above.

The compound $LaCo_{13}$ shows a rather peculiar Kerr spectrum compared to Co and the other La–Co compounds, probably due to pecularities of its crystal structure (cubic $NaZn_{13}$ type). A schematic representation of this structure type is shown in fig. 7. There are two crystallographically inequivalent Co sites

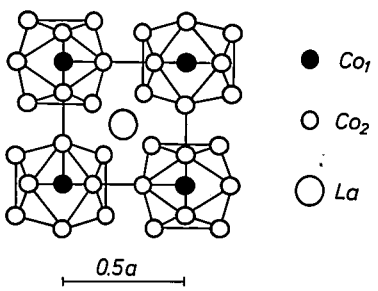


Fig. 7. Schematic representation of the crystal structure of $LaCo_{13}$.

designated by 8(*b*) and 96(*i*) in Wyckoff notation. The Co atoms at the 8(*b*) sites and the La atoms have an atomic arrangement corresponding to the well-known CsCl structure. Each of the Co_1 atoms is surrounded by an icosahedron comprising $12Co_2$ neighbours. These Co_2 icosahedra are packed in alternate orientations (see fig. 7). The shortest Co–Co separation is between the Co_1 and Co_2 atoms and equals 0.2386 nm. It is shorter than the nearest neighbour separation in Co metal. Erskine and Stern [7]) ascribe the first $|2\varphi_K|$ maximum near 1.2 eV in Co to d→p transitions, starting from the top of the majority band. The enhanced intensity of the Kerr spectrum in $LaCo_{13}$ near 1.2 eV could then be interpreted as meaning that the peculiar Co coordination in this material leads to a comparatively large density of states near the top of the 3d band.

IPR2018-1557
HTC EX1023, Page 98

The results given in fig. 3 are plotted in such a way that the Co moments are always parallel to the applied field and Co contributes to $2\varphi_K$ in the downward direction. By contrast, the Gd moments are antiparallel to the applied field and the contribution of the Gd atoms to $2\varphi_K$ is in the upward direction. Results obtained by Erskine and Stern [8]) in Gd metal have been included in the top part of fig. 3 by means of two vertical solid lines. The interpretation of the evolution of the Kerr spectra of the Gd–Co compounds appears then to be rather straightforward. With increasing Gd concentration the contribution due to Co gradually decreases while that due to Gd gradually increases. The compound Gd_2Co_{17} behaves somewhat differently in this respect. Here the second $2\varphi_K$ maximum appears to be slightly reduced. But in R_2Co_{17} a feature in atomic arrangement exists which is similar to that described already for $LaCo_{13}$. Two of the 17 Co atoms in R_2Co_{17} occupy positions known as dumb-bell pairs, characterized by a mutual Co–Co separation shorter than in Co metal. Compared to $LaCo_{13}$, however, the number of short Co–Co separations is relatively small. Consequently their effect, though still discernible in the spectra of Gd_2Co_{17} and Y_2Co_{17}, is much smaller than in $LaCo_{13}$.

Comparison of the results of figs 3 and 4 shows that the detrimental influence of the increasing Gd content on $|2\varphi_K|$ is far less strong in the amorphous alloys than in the crystalline materials. In the alloy with $x = 0.83$ the $|2\varphi_K|$ values are generally more than twice as large as in $GdCo_5$. This is mainly a consequence of the fact that the moment reduction due to the increasing rare earth content is less strong in amorphous materials. This also becomes apparent from a comparison of $GdCo_3$ with amorphous $Gd_{0.25}Co_{0.75}$. In the former material the Gd sublattice magnetization dominates at room temperature whereas in the latter there is still a dominant Co sublattice magnetization.

It is still unclear what causes the changes observed in relatively rare-earth rich alloys after the sample has been kept in air for some time. Most likely the rare earth at the surface will become preferentially oxidized, leaving Co metal in the form of small clusters of metal atoms. The magnetic atoms within each cluster will couple ferromagnetically but will be decoupled from the Gd atoms of the main sample. The Gd_2O_3 formed is paramagnetic at room temperature and does not contribute. The $2\varphi_K$ values shown for the alloy $x = 0.62$ in the top part of fig. 4 are positive. This is the result of the Co sublattice magnetization being oriented antiparallel to the field applied and to the magnetically dominant Gd sublattice magnetization, although the former still dominates the Kerr intensity. Prolonged oxidation and the concomitant formation of Co clusters will lead to an enhancement of negative contributions to $2\varphi_K$ since the Co clusters are decoupled from the Gd sublattice and will orient themselves parallel to the external field, i.e. antiparallel to the Co moments in the main

IPR2018-1557
HTC EX1023, Page 99

part of the sample. One expects therefore a general decrease of the positive $2\varphi_K$ contributions. However, the effective depth associated with the Kerr effect measurements is several tens of nm in the low energy region but only about 10 nm at the high energy side. A deterioration of only the topmost few nanometres of the film would then lead to no more than minor changes in the low energy region but to a substantial decrease of the positive $2\varphi_K$ contribution at higher photon energies, in agreement with the experiment. Moreover, the interference effects associated with the oxide layer may also change the spectrum. Here, too, the effects at high photon energies are greater than at low photon energies.

Conclusions

The study of the Kerr rotation spectra of several Y–Co compounds has shown that the effect of decreasing Co concentration is a general reduction of the Kerr intensity associated with the Co atoms, with no major changes in the shape of the spectra.

The spectra of the Gd–Co can be explained as originating from a superposition of contributions due to Co and Gd. In the pure metals the main Kerr intensity in the range 0.5–4.5 eV in Gd and Co has the same sign. In the compounds these two contributions have a dissimilar sign owing to the antiparallel coupling of the moments of Co and Gd.

In amorphous Gd–Co alloys the situation is similar to the one encountered in the crystalline materials, albeit the Co sublattice magnetization and the corresponding Kerr contribution decrease much more slowly with increasing Gd concentration. The effect of surface oxidation in alloys of relatively high Gd content affects primarily the high energy part of the spectrum. No major changes were observed in the range 1–2 eV corresponding to the wavelengths of practical interest, using He–Ne or GaAs lasers (λ = 633 nm and 830 nm, respectively).

REFERENCES

[1]) K. A. Schouhamer Immink and J. J. M. Braat, Proc. 73rd AES Convention, March 1983, Eindhoven.
[2]) Y. Togami, IEEE Trans. Magn. **MAG-18**, 1233 (1982).
[3]) A. M. Stoffel, J. Appl. Phys. **39**, 563 (1968).
[4]) T. Katayama and K. Hasegawa in "Rapidly Quenched Metals IV", The Japan Metals Society, Sendai 1981 (eds: T. Masumoto and K. Suzuki); cf. p. 915.
[5]) S. Visnovsky, B. Knappe, V. Prosser and H. R. Müller, Phys. Stat. Sol. (a) **38**, K53 (1976).
[6]) F. J. Kahn, P. S. Pershan and J. P. Remeika, Phys. Rev. **186**, 891 (1969).
[7]) J. L. Erskine and E. A. Stern, Phys. Rev. Lett. **30**, 1329 (1973).
[8]) J. L. Erskine and E. A. Stern, Phys. Rev. B **8**, 1239 (1973).

IPR2018-1557
HTC EX1023, Page 100

Philips J. Res. 39, 94-102, 1984 *R 1084*

APPLICATIONS OF A DIGITAL AUDIO-SIGNAL PROCESSOR IN T.V. SETS

by W. J. W. KITZEN and P. M. BOERS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Digital audio-signal processors are already being used in professional applications (mixing consoles, artificial reverberation etc.). In this paper some applications of such a device in a consumer product, e.g. a television set, are discussed. Apart from well-known functions such as tone, volume and balance control the following rather new functions can also be implemented using a digital processor:

— a pseudo-stereo circuit that can be used if the television set is equipped to reproduce stereophonic signals, but the program is transmitted in mono,
— a stereo base expanding circuit to be used when the two loudspeakers are closely spaced,
— a circuit for enhancing the impression of spaciousness,
— a volume control with combined loudness correction.

PACS number: 4388

1. Introduction

Until now, most of the applications of digital signal processors in the audio field have been professional applications. Because of their rather complex structures and the many components required, those processors were too expensive for consumer products. However VLSI now makes it possible for consumer products to take advantage of digital processing. The advantages are:

— There is no signal distortion and no processing noise is added if the number of bits that represents the signals in the processor is large enough.
— A perfect delay line can be implemented digitally which is impossible in an analog way. For many applications a delay line is needed, e.g. a spaciousness enhancing circuit.
— Quite complex processing can be implemented rather easily.
— The implementation of a function in a digital processor is in fact a program that consists of instructions, coefficients and (data) RAM addresses. By changing this program, a different function can be implemented, so many functions can be covered with a single processor.

In this paper, after a short discussion of the structure of a digital processor, we discuss four applications that can be implemented and used in television sets.

IPR2018-1557
HTC EX1023, Page 101

2. Structure of a digital processor

The basic structure of the digital processor used for our experiments is depicted in fig. 1. Via the I/O ports data from an external device such as an AD converter can be stored in the data RAM. The multiplier can multiply these

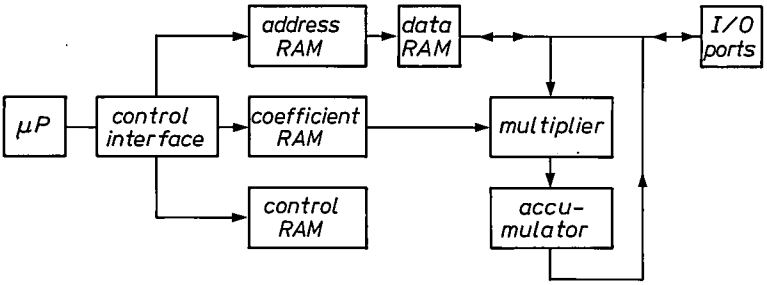


Fig. 1. Basic structure of digital processor.

data by coefficients from the coefficient RAM. The results of some multiplications can be added in the accumulator, which is again connected with the data RAM. Processed data are available for external devices such as a DA converter via the I/O ports. The processor is controlled by a program in the control, address and coefficient RAMs. Each instruction consists of a control word and if necessary a RAM address and a coefficient. A multiplication, for example, involves data on a particular RAM address, a coefficient in the coefficient RAM and a control word in the control RAM. Because we have used random access memories for coefficients as well as for addresses and control words, which can all be updated by a microprocessor via a control interface, the processor is very flexible. This has turned out to be very useful for experiments with different kinds of algorithms. Of course a digital processor has limitations. One of them is the size of the data RAM, which fixes the maximum delay time that can be implemented. The word-size of the coefficients determines the accuracy of the filters implemented, the word-size of the data RAM the processing noise. The complexity of the applications depends among other things on the speed of the processor, i.e. the speed of the RAMs, ROMs and the multiplier. By using pipe-lining techniques the processor is speeded up.

3. Applications

Apart from well-known functions such as tone, volume and balance control some rather new functions can also be implemented with a digital processor. In this paper we discuss four of them. The first function is a pseudo-stereo circuit. At present few television programs are transmitted with stereophonic

IPR2018-1557
HTC EX1023, Page 102

sound. With a pseudo-stereo circuit it is possible to convert a monophonic input signal into pseudo-stereophonic output signals.

Another function that can be used in a television receiver is a stereo base expanding circuit. In most television sets equipped to reproduce stereophonic signals the two loudspeakers are closely spaced, so the physical stereo base is rather small. By using a stereo base expanding circuit, the apparent stereo base can be widened.

We also discuss a spaciousness-enhancing circuit with which it is possible to influence the listener's sense of spaciousness.

Finally we describe a volume control with combined loudness correction.

3.1. *A pseudo-stereo circuit*

With monophonic reproduction the sound is localized in one direction. The objective of a pseudo-stereo circuit is that the sound should be perceived as if it arrives from many directions.

Several psycho-acoustical phenomena related to directional hearing can be explained if we assume that our hearing system performs a short-time cross-correlation process on the two ear-signals. A normalized binaural short-time cross-correlation function as defined by Blauert [1]) is:

$$\Phi_{(t,\tau)} = \frac{\int_{-\infty}^{t} l(\theta)\, r(\theta+\tau)\, G(t-\tau)\, \mathrm{d}\theta}{\sqrt{\int_{-\infty}^{t} l^2(\theta)\, G(t-\theta)\, \mathrm{d}\theta \int_{-\infty}^{t} r^2(\theta)\, G(t-\theta)\, \mathrm{d}\theta}}$$

where $l(t)$ is the left signal, $r(t)$ is the right signal and $G(t)$ is a weighting function:

$$G(t) = \exp(-t/\tau_e).$$

The time constant τ_e is, according to Blauert, certainly smaller than a few milliseconds. By determining the time τ for which the cross-correlation function is maximum, our hearing is able to localize the sound source. If the cross-correlation function is very small for all values of τ, then, according to this theory, the sound cannot be localized and the sound image is broadened or diffusely localized. So in order to realize pseudo-stereo the binaural cross-correlation function should be small for all values of τ. If a set-up with two loudspeakers is used, a prerequisite for a small binaural cross-correlation is a small cross-correlation between the two loudspeaker signals. This can be achieved by realizing two impulse responses (from input to the left loudspeaker and from input to the right loudspeaker) having little correlation. There are many practical ways of doing this. Because the localization of sounds is determined by

IPR2018-1557
HTC EX1023, Page 103

about the first 5 ms of the impulse responses, the actual impulse responses need not be longer than 5 ms.

Figure 2 shows a solution proposed by Lauridsen[2]) using two complementary comb filters. The left and right output signals in fig. 2 have little correlation since both phase and amplitude characteristics of the left and right

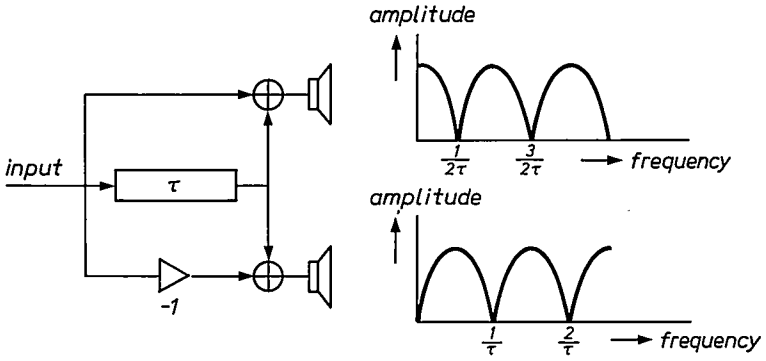


Fig. 2. Pseudo-stereo circuit (see ref. 2).

impulse responses are different. However, because the amplitude-frequency responses are not flat, these filters may introduce colouration.

A solution given by Schroeder[3]), where the filters in the left and right channels are both allpass filters but have different phase characteristics, is shown in fig. 3. Because of the different phase characteristics an envelope delay difference between left and right channel is introduced which is frequency-dependent. As explained by Schroeder, this results in localization of some frequency

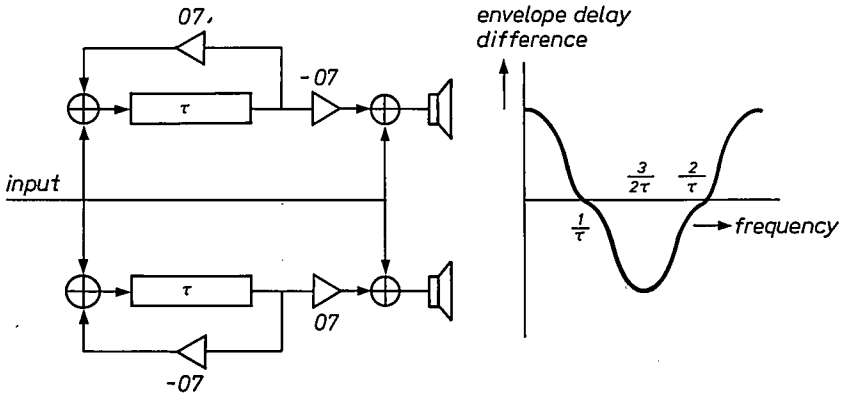


Fig. 3. Pseudo-stereo circuit (see ref. 3).

IPR2018-1557
HTC EX1023, Page 104

components at the position of the left loudspeaker and other components at the right loudspeaker. However, if the time delay τ is chosen too large, allpass filters may also introduce audible colouration. If we consider a short-time Fourier transform (integration over a finite time interval) of the impulse response, such a filter is not allpass at all, and, as we have seen before, the integration time of our hearing is only a few milliseconds.

3.2. *Stereo base expanding*

Stereo sound reproduction in a stereo set-up may create virtual sources at every position between the loudspeakers, i.e. the stereo base. If the loudspeakers are closely spaced, the small stereo base can be widened by applying delayed crosstalk in antiphase between the two channels, as has been explained recently [4]).

Suppose we apply a signal only at the right input (fig. 4). If no crosstalk is applied, a listener in the plane of symmetry localizes the source at the right

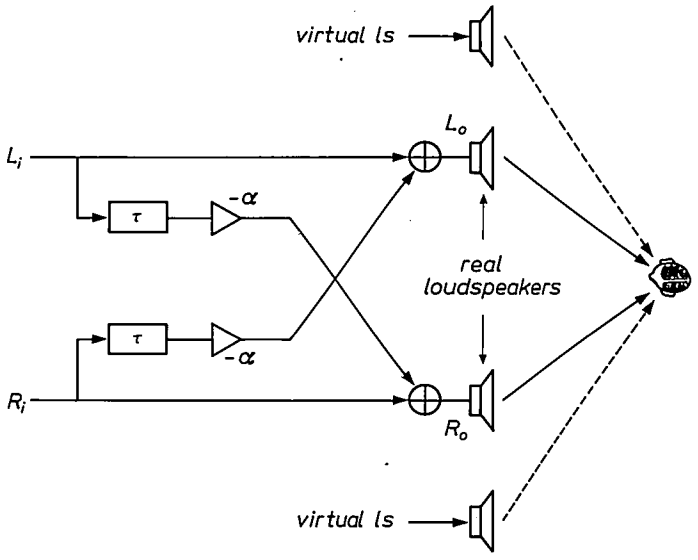


Fig. 4. Stereo base expanding circuit.

loudspeaker position. If undelayed crosstalk in antiphase ($\tau = 0$) is applied, a virtual source will shift to the right. This is caused by the increased interaural phase and group delays [4]). Thus, the right loudspeaker in fig. 4 is replaced by a virtual one shifted to the right and the left loudspeaker by a virtual one shifted

IPR2018-1557
HTC EX1023, Page 105

to the left, i.e. the stereo base has been widened. Adding a small delay of about 0.1 ms in the crosstalk circuit will make the effect more pronounced. This is brought about by interaural level differences which, for a listener in the plane of symmetry, reinforce the effect of the interaural time delay differences. A disadvantage of this crosstalk in fig. 4 is that, in the low frequency range, the input signals R_i and L_i (when they are completely correlated) will be attenuated, whereas when uncorrelated the signals will be amplified. In the case of correlated signals ($L_o = R_i$) the output signals are:

$$|L_o|^2 = |R_o|^2 = |L_i|^2 \times (1 - \alpha)^2$$

for

$$\omega\tau \ll \frac{\pi}{2}.$$

When L_i and R_i are uncorrelated and $|L_i|^2 = |R_i|^2$:

$$|L_o|^2 = |R_o|^2 = |L_i|^2 \times (1 + \alpha^2).$$

In the case of normal stereo signals, the left and right signals are highly correlated in the low-frequency range and uncorrelated at high frequencies. The attenuation of the low frequencies relative to the high frequencies is therefore:

$$\frac{(1 - \alpha)^2}{(1 + \alpha^2)}.$$

For $\alpha = 0.7$ this is about 12 dB. This can be corrected by using a bass-boost filter before the crosstalk processing.

3.3. *A spaciousness-enhancing circuit*

In a real room, e.g. a concert hall, the impression of spaciousness is caused by the early reflections via the side-walls that arrive within $t < 100$ ms [5]). Because of these reflections signals with little correlation reach the ears of a listener. To enhance the impression of spaciousness in an artificial way, therefore, delayed signals should be added to the original signals that are uncorrelated. The signals should be delayed because we do not want to influence the localization of the sounds. Here too, then, as in the pseudo-stereo circuit, we must realize two impulse responses that are uncorrelated, but in this case the impulse responses should be longer (about 50 ms). A circuit with which this is possible shown in fig. 5*a*. In order to produce many reflections, a recursive filter is used. Each channel consists of a delay line with a feedback loop. The

IPR2018-1557
HTC EX1023, Page 106

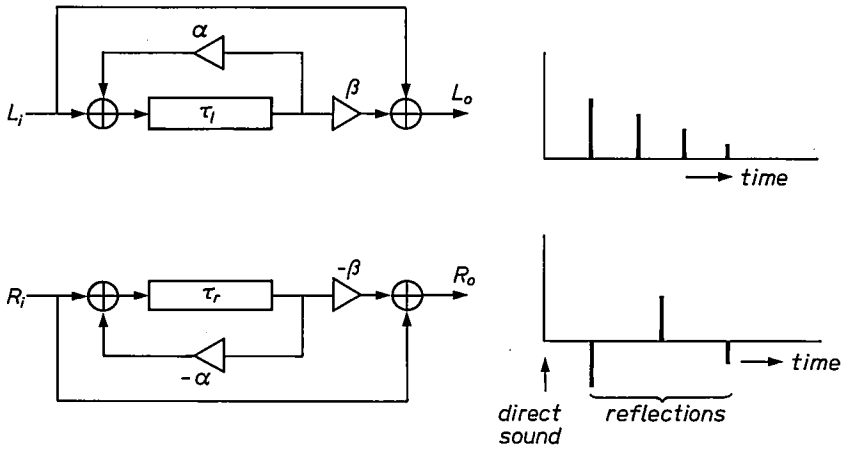


Fig. 5*a*. Spaciousness-enhancing circuit.

output of that structure is added to the original input signal. So the impulse responses from the input to the outputs consist of an impulse at $t = 0$, which represents the direct sound, and a large number of reflections. By choosing different delays and feedback factors in the left and right channels, the correlation of the impulse response of the left channel with the impulse response of the right channel is made small. By changing the amplitude of the uncorrelated signals to be added to the original signals (variable β), the impression of spaciousness can be adjusted. The comb filter effect in the circuit may produce colouration. This effect can be reduced by making the feedback factor α frequency dependent (fig. 5*b*). As a consequence the length of the impulse responses for frequencies higher than about 1500 Hz is reduced. However this

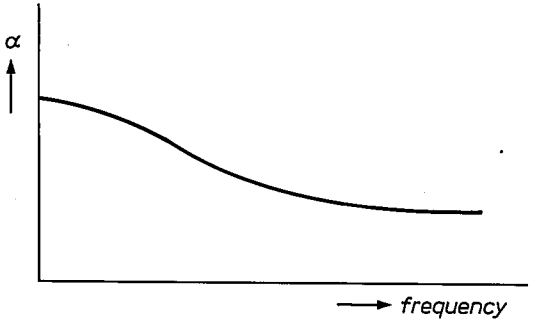


Fig. 5*b*. Frequency dependency of feedback factor α.

IPR2018-1557
HTC EX1023, Page 107

does not affect the impression of spaciousness because this impression, according to Barron, is mainly determined by the low frequencies [5]).

Because the localization of the sources depends only on the first arriving sound, i.e. the first 5 ms of the impulse responses, it is unaffected by the circuit when $\tau_{l,r} > 5$ ms. If the input signal is to be reproduced by two loudspeakers that are closely spaced, the decorrelated signals can also be reproduced via an expanded stereo circuit.

3.4. *Volume control combined with loudness-correction*

Loudness is the subjective quantity by which we measure the perceived sound pressure level. The way loudness increases with sound pressure level strongly depends on frequency for frequencies below about 400 Hz [6]). As follows from the equal loudness contours (fig. 6), the reproduction of music

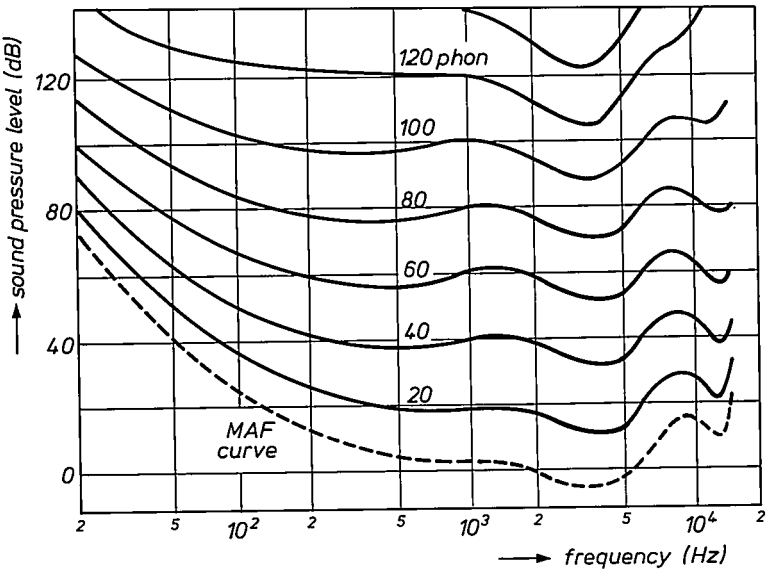


Fig. 6. Equal-loudness contours (see ref. 6).

at a level below the original level will lead to a decrease in loudness of the low frequencies relative to the high frequencies i.e. a change in timbre. To compensate for this effect a simple first-order filter can be implemented which enhances the low frequencies more the lower the setting of the volume control (fig. 7). Here the output level of 0 dB corresponds to the level in the studio mixing room. We assume that the timbre at this level is optimal without any correction.

IPR2018-1557
HTC EX1023, Page 108

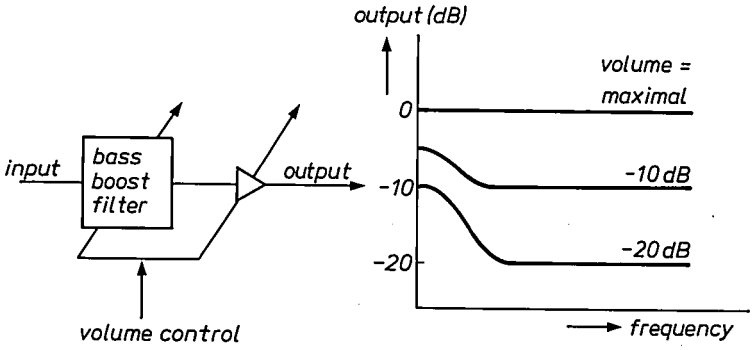


Fig. 7. Volume control combined with loudness-correction.

4. Conclusions

We have discussed some applications of a digital signal processor in a T.V. set. One of the advantages of our digital processor, because of its flexibility, is the ease with which one can experiment with different algorithms and compare their qualities. If a single digital processor is to do all the processing for tone and balance control as well as for pseudo-stereo, stereo base expanding, spaciousness and volume control with combined loudness correction, then its speed should be about 100 instructions/sample period. The wordlength of the data depends on the desired S/N ratio, the wordlength of the coefficients on the desired accuracy of the filters. With the VLSI technology of today it is possible to integrate such a processor on one single chip. Such a chip will constitute a powerfull building block which performs all kinds of processing, not only for applications in a television set but also for other sound reproducing systems.

REFERENCES

[1]) J. Blauert, Raumliches Horen, S. Hirzel Verlag, Stuttgart, 1974, English Edition: Spatial Hearing, MIT Press, Boston, Mass., 1983; cf. 278.
[2]) H. Lauridsen and F. Schleger, Gravesaner Blatter, H.5, 27 (1956).
[3]) M. R. Schroeder, J. Acoust. Soc. Am. 33, 1061 (1961).
[4]) P. M. Boers, AES 73rd Conv. Eindhoven, The Netherlands, No. 1967 (A5) (March 1983).
[5]) M. F. E. Barron, The effects of early reflections on the subjective quality in concert halls, Doctor's thesis, University of Southampton (1974).
[6]) D. W. Robinson and R. S. Dadson, Brit. J. Appl. Phys. 7, 156 (1956). ISO Recommendation R226-1961.

IPR2018-1557
HTC EX1023, Page 109

Philips J. Res. 39, 103-108, 1984 *R 1085*

PABLO A VERSATILE VLSI TECHNOLOGY

by H. G. R. MAAS and J. A. APPELS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

A method to make very narrow grooves in a substrate is described. The main feature on which the process is based is the lateral oxidation of a polysilicon layer, which is later on removed. The lateral oxide rim thus obtained in an early stage of the process eventually determines the width of the groove in the underlying material. The acronym PABLO (Perfect Alignment By Lateral Oxidation) has been chosen for this technology. Although the paper describes only the implementation of the PABLO process towards a self-aligned bipolar device structure, the technique is really versatile and flexible and can be used in almost any device where grooves are needed, such as non-overlapping submicron gap polysilicon gate structures in CCD applications.

PACS numbers: 85.40 C

1. Introduction

Progress in optical and electron beam lithography for VLSI purposes, with the aim of achieving smaller dimensions and higher densities per chip, is steadily being made. Nevertheless sophisticated self-aligning processes which make it possible to realize very small contacts or gates as well as submicron spacings between them, will remain very important in the coming years. Techniques such as wet chemical undercutting of silicon nitride or polysilicon layers are less suitable. It is true that an improvement can be obtained by a combination of boron implanted zones in polysilicon and KOH etching [1]). However, the process is then confined to wet chemical (under)etching while the locally heavily boron-doped polysilicon layer seriously reduces the number of applications.

It will be shown in this paper that the PABLO technique (Perfect Alignment By Lateral Oxidation) overcomes these hurdles and therefore is ideally suitable to obtain minimum dimensions (grooves) between polysilicon contacts or gates.

The main step in PABLO is the lateral oxidation of a polysilicon layer. The width of the oxide grown on the sidewall of the polysilicon eventually determines the width of the groove to be made in the underlying layer. Since oxida-

IPR2018-1557
HTC EX1023, Page 110

tion is a well controllable process, the groove can be dimensioned very reproducibly. Moreover, the groove width is independent of doping-type, impurity concentration or thickness of the polysilicon layer. With regard to circuit performance the process has the possibility for silicidation of the contacts and interconnection lines, also in a self-registered way. The method can elegantly be applied in a number of devices such as bipolar and MOS transistors with an extended drain. A very interesting application is a non-overlapping polysilicon gate structure with submicron gaps in densely packed CCDs. Still another application is to use the technique to make fairly deep and narrow isolation trenches in order to reduce the inactive circuit area.

2. Description of the technology

The basic PABLO principle is the lateral oxidation of a polysilicon layer with the aim to form a well-defined oxide stud, which eventually determines the width of the groove to be made in the underlying substrate. A lot of variations are possible in this process. The most important probably are the ones, where the ultimate structure ends up with regions at one side of the groove covered with silicon oxide, while the adjacent areas are covered with silicon nitride. In that case the nitride covered silicon can be opened selectively and this being self-aligned will lead to a higher packing density. Only one of these PABLO variations, called PABLO I, will be described here and will be explained on the basis of a bipolar transistor process. The process steps are described below and are depicted in fig. 1. Photographs of some process steps are shown in figs 2–4.

2.1. *PABLO I*

Like in conventional bipolar IC-technology, the process starts with making the buried layers, growth of the epitaxial layer and the definition of the isolation regions. Next the following steps are carried out:

fig. 1*a*. Deposition of four layers: Polysilicon layer P_1, Silicon nitride layer N_1, Polysilicon layer P_2, Silicon nitride layer N_2. Layer N_2 must be thicker than N_1. Next N_2 is etched at the regions where at the end of the process P_1 is unwanted, whereupon the now exposed parts of P_2 are converted into silicon oxide.

fig. 1*b*. N_2 and P_2 are etched away, except for the future emitter regions.

fig. 1*c*. At the edges of these emitter regions, the P_2 layer which is sandwiched between N_1 and N_2 is laterally oxidized.

fig. 1*d*. The exposed parts of N_1 and also the top of layer N_2 are etched away. Part of layer N_2 must be maintained to protect P_2 during further process steps.

IPR2018-1557
HTC EX1023, Page 111

fig. 1*e*. The lateral oxide formed in step *c* and the oxidized P_2 above the finally unwanted P_1 regions are etched away.

fig. 1*f*. The exposed parts of P_1 are superficially oxidized. At the same time a new narrow lateral oxide is formed at the edge of P_2.

fig. 1*g*. The exposed parts of N_1 are etched away, thus giving access to P_1. The external base is implanted through the oxide on P_1, P_2 still left on N_1 serves as a mask on the emitter regions.

fig. 1*h*. The exposed parts of P_1 are etched away throughout the thickness of the layer. This defines the outer dimensions of P_1 as well as the small grooves in it. Also the remaining P_2 is removed from N_1.

fig. 1*j*. The submicron groove of step *h* is oxidized and simultaneously the external base of step *g* is activated.

fig. 1*k*. N_1 is removed by deliberate overetching so that the emitter windows are opened and the narrow lateral oxide produced in step *f* falls off.

Next the active base is implanted and diffused, now followed by an emitter implantation and diffusion. Thereafter the process can be completed in the usual manner.

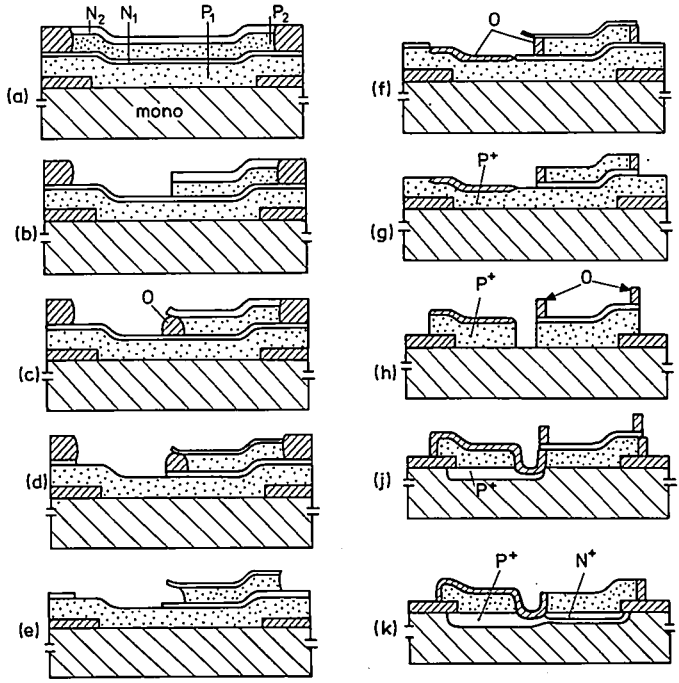


Fig. 1. Fabrication steps of PABLO I. P_1 and P_2 are polysilicon, N_1 and N_2 are silicon nitride. O is silicon oxide.

IPR2018-1557
HTC EX1023, Page 112

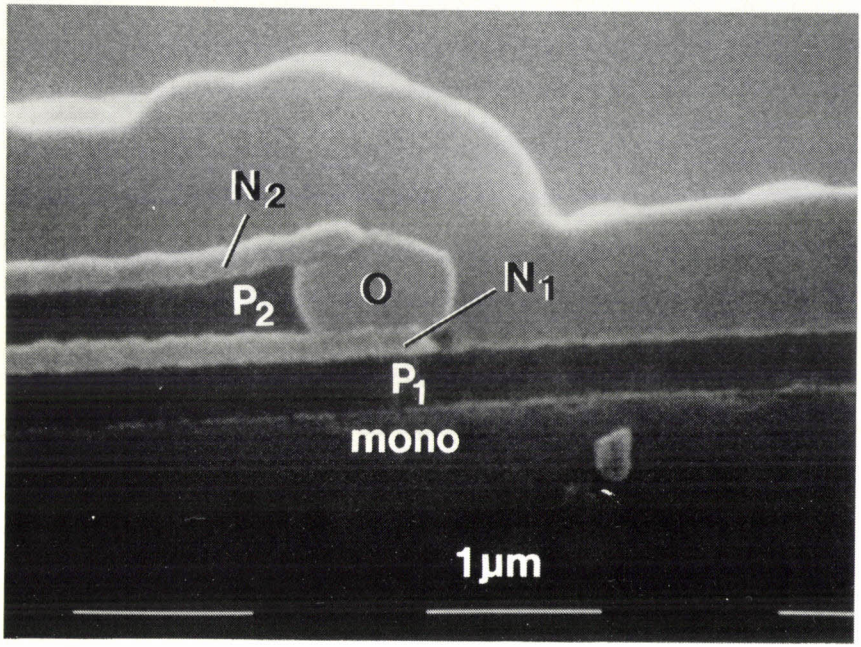


Fig. 2. SEM-photograph of a cross-section corresponding with step (*d*) in fig. 1.

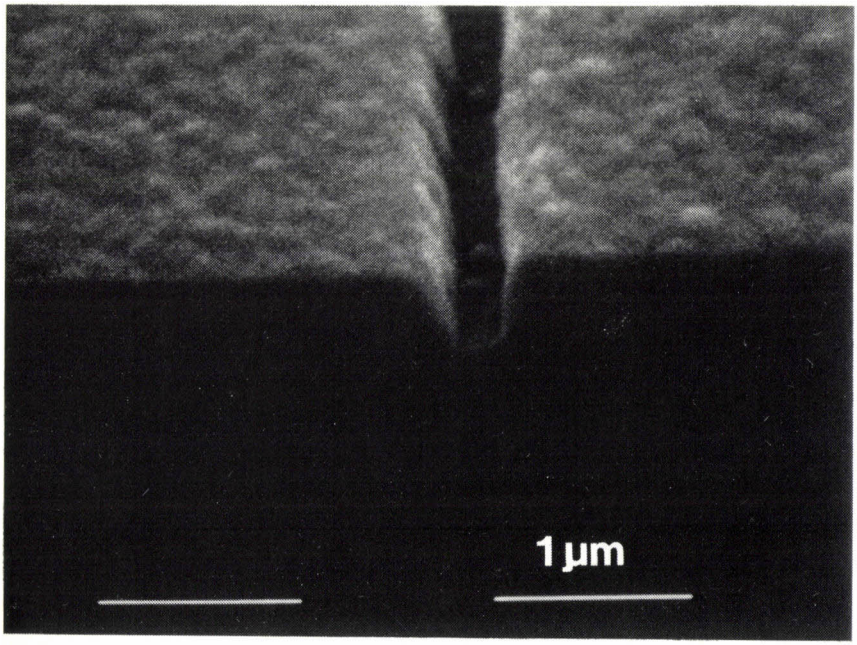


Fig. 3. SEM-photograph showing a bare groove.

IPR2018-1557
HTC EX1023, Page 113

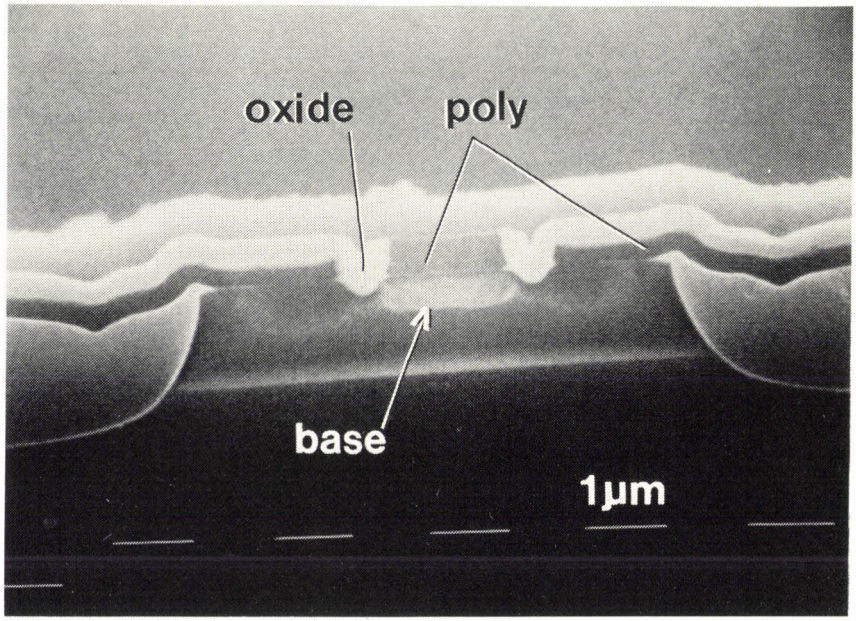


Fig. 4. SEM-photograph showing a cross-sectional view of a transistor with a submicron oxide separation between the emitter- and base polysilicon contacts.

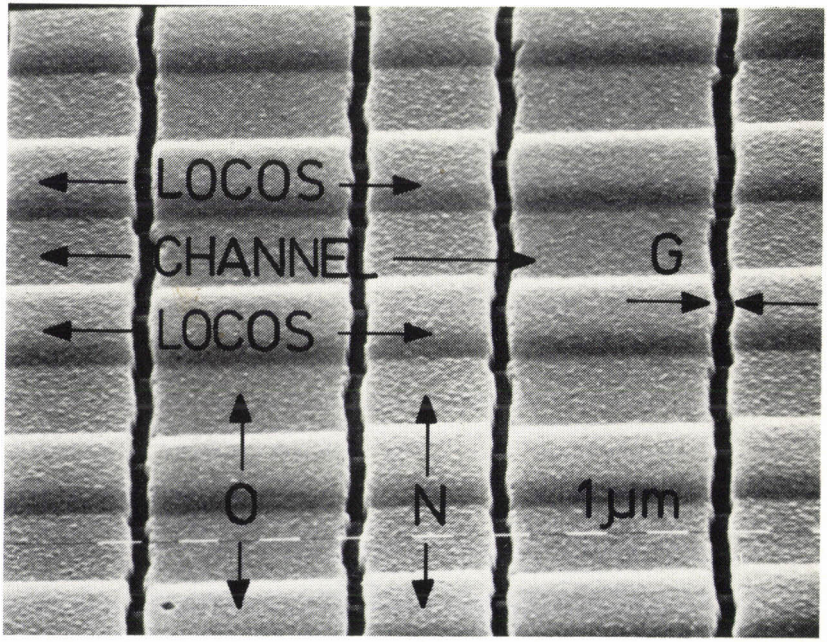


Fig. 5. SEM-photograph showing a top view of polysilicon electrodes in a CCD structure separated by a submicron groove (G). N = nitride covered poly electrode, O = oxide covered poly electrode. Magnification 5000×.

IPR2018-1557
HTC EX1023, Page 114

3. Summary and conclusions

A reliable and reproducible method for the fabrication of submicron grooves in substrates has been demonstrated. Basically the method consists of the lateral oxidation of a polysilicon layer. The width of the oxide stud grown at the sidewall of the polysilicon determines the width of the groove to be made in the substrate. In this paper PABLO was explained on the basis of a bipolar transistor process. However, the technique has a much wider field of applications. In general it can be used for making closely spaced (submicron) self-aligned electrodes, which e.g. can profitably be applied in the fabrication of very densely packed CCDs (fig. 5). Also MOS transistors with extended drains or submicron gates can be made with this technique. Still another application is the fabrication of deep isolation grooves or capacitors.

REFERENCE

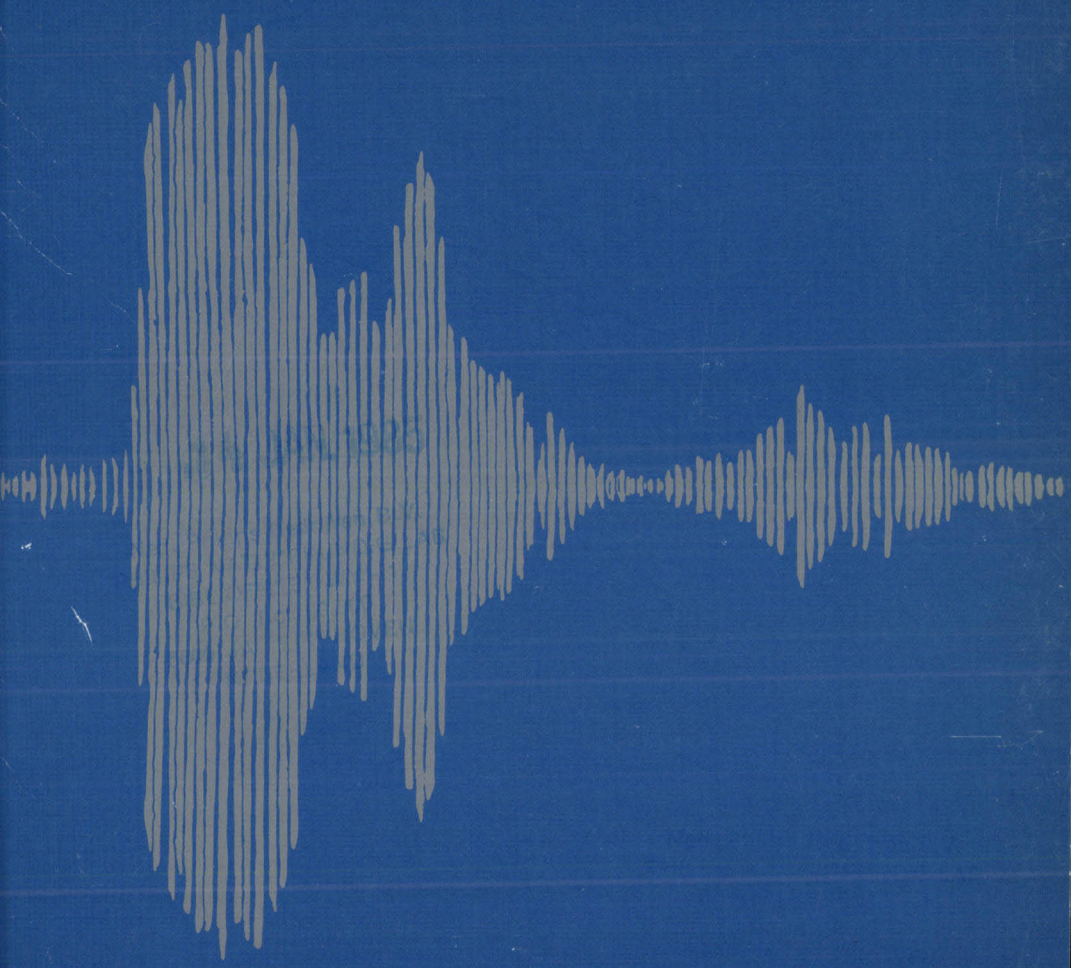
[1]) Tetsushi Sakai et al., ISSCC Dig. Tech. Papers, 216 (1981).

IPR2018-1557
HTC EX1023, Page 115

3-48=27

hilips Journal of Research

Vol. 39 Nos 4/5 1984



HILIPS

In honour of prof. V. Belevitch

IPR2018-1557
HTC EX1023, Page 116

Philips Journal of Research

Philips Journal of Research, published by the 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 39 is Dfl. 75, 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.

Editorial Board

A. van Oostrom (General Editor) Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Y. Genin, Brussels, Belgium
R. Kersten, Aachen, Germany
R. Memming, Hamburg, Germany
R. F. Milsom, Redhill, Surrey, U.K.
W. A. Smith, Briarcliff Manor, N.Y., U.S.A.
W. T. Stacy, Sunnyvale, CA., U.S.A.
J. B. Theeten, Limeil-Brévannes, France

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

© Philips International B.V., Eindhoven, The Netherlands, 1984. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

IPR2018-1557
HTC EX1023, Page 117

Professor V. Belevitch

IPR2018-1557
HTC EX1023, Page 118

PREFACE

Balthasar van der Pol, nestor of radio scientists, and I once had a somewhat facetious discussion of the use of mathematics. Van der Pol emphasized both the beauty and the power of analytical methods — which he himself handled with great virtuosity — whereas I stressed their limitations. "The capacity of a sphere", I said, "I can write down at once, that of an ellipsoid I can calculate with some difficulty, but calculating the capacity of a cube leads to unwieldy series. Yet the electrons manage to distribute themselves over the surface in a nanosecond or so. Calculating the impedance of a network with four meshes is already hard work; for a network with one hundred meshes it is almost impossible. The currents, however, find their way without difficulty. Mathematics is more often than not unable to deal with quite simple problems".

Now this situation has been radically changed by the computer. To solve 100 equations with 100 unknowns has become almost literally childs play. Differential equations are easily solved and non-linear problems, formerly unmanageable, can be dealt with. And this is only one aspect of the rapid evolution of information technology brought about by solid state electronics, the further development of which owes in turn much to computers.

The older methods of analysis have not become obsolete, on the contrary, in many cases their value has been enhanced by the computer, but there has also arisen a number of new disciplines for which "informatics" is a fairly satisfactory name and which is associated with such keywords as "digitalization" and "computer languages". This trend became evident in the fifties and has by now even been recognized in government circles.

When in 1963 Belevitch entered the Philips orbit and undertook to create at Brussels a research group dealing specifically with such subjects this was a most valuable addition to our research potential.

Belevitch combines specialized mathematical competence with a very broad range of interests, including general linguistics. This made him not only an excellent researcher and leader of research, but also a welcome participant in almost any kind of discussion. Incidentally, it also made him an excellent chairman of the jury of the Philips contest for young scientists and inventors, an activity near to my heart.

The papers in the present volume, dedicated to him on the occasion of his retirement, relate to many — although not to all — aspects of Belevitch's work and I am convinced that he will enjoy reading and criticizing them.

IPR2018-1557
HTC EX1023, Page 119

Let me finish by expressing the hope and expectation that in this case "retirement" will not mark the end, but rather the beginning of a new phase of his scientific activities.

H. B. G. Casimir

IPR2018-1557
HTC EX1023, Page 120

Philips J. Res. **39**, *111-120, 1984* R 1086

AN INTRODUCTION TO MULTILINEAR ALGEBRA AND SOME APPLICATIONS

by A. BLOKHUIS *) and J. J. SEIDEL **)

*) *Center of Mathematics and Computer Science, 1009 AB Amsterdam*
**) *Department of Mathematics, University of Technology, 5600 MB Eindhoven, The Netherlands*

Abstract

The present paper intends to give a quick and easy introduction to tensors, in particular to the exterior and the symmetric algebra of a vector space and the relations between them. The entries of the k–th exterior, and the k–th symmetric power of a matrix are expressed as determinants, and permanents, respectively. The generating polynomials of these powers are related by their traces. As applications we mention MacDonald's proof of MacMahon's master theorem, Bebiano's permanent expansion[1]) and the permanent version of the solution of Fredholm's integral equation, observed by Kershaw[2]) and by de Bruijn[3]).

Math. Rev.: 15A69

1. Symmetric functions[4])

The *elementary symmetric* polynomial of degree k in d variables, $x_1, \ldots, x_d$, is defined by

$$e_k(x) := \sum_{1 \leqslant j_1 < \ldots < j_k \leqslant d} x_{j_1} \ldots x_{j_k},$$

or, equivalently, by its generating function

$$E(t; x) := \sum_{k \geqslant 0} e_k(x) t^k = \prod_{i=1}^{d} (1 + x_i t).$$

The k–th *complete symmetric* polynomial is defined by

$$h_k(x) := \sum_{1 \leqslant j_1 \leqslant \ldots \leqslant j_k \leqslant d} x_{j_1} \ldots x_{j_k},$$

with generating function

$$H(t; x) := \sum_{k \geqslant 0} h_k(x) t^k = \prod_{i=1}^{d} \frac{1}{1 - x_i t}.$$

Theorem 1.1

$$E(-t; x)\, H(t; x) = 1.$$

Remark

$\{e_1, \ldots, e_d\}$ forms a basis for the graded ring S_d of symmetric polynomials

IPR2018-1557
HTC EX1023, Page 121

in d variables. Let j denote the all-one vector of size d, then

$$e_k(j) = \binom{d}{k} \quad \text{and} \quad h_k(j) = \binom{d+k-1}{k}.$$

2. Tensors[5-7])

We start with a quick definition. Let V be a (real) vector space of dimension d, with (positive definite) inner product $(\ ,\ )$ and orthonormal basis $e_1, \ldots, e_d$. We define tensor products by their components with respect to this basis:

$$x \otimes y \quad \text{by the } d^2 \text{ components } x_i y_j,$$
$$x \otimes y \otimes z \quad \text{by the } d^3 \text{ components } x_i y_j z_k.$$

The tensor product $V \otimes V$ is the linear space spanned by all tensor products $x \otimes y$; $x, y \in V$. The inner product $(x \otimes y, u \otimes v) := (x, u)(y, v)$ provides $V \otimes V$ with the orthonormal basis $\{e_i \otimes e_j \mid 1 \leqslant i, j \leqslant d\}$.

Similarly tensor products of k vectors, $x_1 \otimes \ldots \otimes x_k$ are defined by their d^k components $(x_1)_i (x_2)_j \ldots (x_k)_m$, with respect to the orthonormal basis

$$\{e_i \otimes e_j \otimes \ldots \otimes e_m \mid \quad 1 \leqslant i \leqslant j \leqslant \ldots \leqslant m \leqslant d\}$$

of $T_k(V) := \otimes^k V$, with inner product

$$(x_1 \otimes \ldots \otimes x_k, y_1 \otimes \ldots \otimes y_k) = \prod_{i=1}^{k} (x_i, y_i).$$

Clearly, $\dim T_k V = d^k$. The *tensor algebra* TV is the direct sum

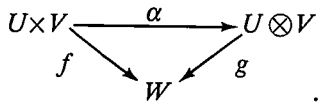
$$TV := T_0 V \oplus T_1 V \oplus \ldots = \sum_{k \geqslant 0}^{\oplus} T_k V,$$

where $T_0 V \simeq R$, $T_1 V \simeq V$. Its Poincaré polynomial (as a graded algebra) is

$$P_T(t) := \sum_{k \geqslant 0} (\dim T_k V)\, t^k = \frac{1}{1 - d\,t}.$$

Remark

We refer to the literature for details about the following coordinatefree definition. The tensor product of the vector spaces U and V is the pair of a vector space $U \otimes V$ and a bilinear map $\alpha : U \times V \to U \otimes V$ such that for all bilinear f there is a unique linear g such that the following diagram commutes



3. The exterior algebra ΛV[5,6])

$\Lambda_2 V$ is the linear subspace of $T_2 V$ spanned by the *skew* 2-tensors:

$$x \wedge y := \tfrac{1}{2}(x \otimes y - y \otimes x); \quad x, y \in V.$$

IPR2018-1557
HTC EX1023, Page 122

Clearly, $x \wedge x = 0$, $x \wedge y = -y \wedge x$ and $\dim \Lambda_2 V = \binom{d}{2}$.

The inner product inherited from $T_2 V$ becomes

$$(x \wedge y, u \wedge v) = \tfrac{1}{2}((x,u)(y,v) - (x,v)(y,u)) = \tfrac{1}{2} \det \begin{pmatrix} (x,u) & (x,v) \\ (y,u) & (y,v) \end{pmatrix}.$$

An orthonormal basis for $\Lambda_2 V$ is provided by $\{\sqrt{2}(e_i \wedge e_j) \mid 1 \leqslant i < j \leqslant d\}$.

Likewise, for $0 \leqslant k \leqslant d$ we define

$$x_1 \wedge \ldots \wedge x_k := \frac{1}{k!} \sum_{\sigma \in \mathcal{S}_k} (-1)^\sigma x_{\sigma 1} \otimes \ldots \otimes x_{\sigma k},$$

where $(-1)^\sigma$ is 1 (-1) if the permutation σ is even (odd), and γ_k is the symmetric group on k letters, and

$$\Lambda_k V := \langle x_1 \wedge \ldots \wedge x_k \mid x_i \in V \rangle_R,$$

with inner product

$$(x_1 \wedge \ldots \wedge x_k, y_1 \wedge \ldots \wedge y_k) = \frac{1}{k!} \det((x_i, y_j)).$$

Note that $\dim \Lambda_k V = \binom{d}{k} = e_k(j)$; an orthonormal basis for $\Lambda_k V$ is provided by

$$\{\sqrt{k!}\, e_{i_1} \wedge \ldots \wedge e_{i_k} \mid 1 \leqslant i_1 < \ldots < i_k \leqslant d\}.$$

The *exterior algebra* ΛV is the (finite) direct sum

$$\Lambda V = \Lambda_0 V \oplus \Lambda_1 V \oplus \ldots \oplus \Lambda_{d-1} V \oplus \Lambda_d V,$$

where $\Lambda_0 V \simeq \Lambda_d V \simeq R$ and $\Lambda_1 V \simeq \Lambda_{d-1} V \simeq V$. Its Poincaré polynomial is

$$p_\Lambda(t) = (1 + t)^d = E(t, j).$$

Remark

$x_1, \ldots, x_k$ are linearly dependent iff $x_1 \wedge \ldots \wedge x_k = 0$.

4. The symmetric algebra ΣV [5,6])

$\Sigma_2 V$ is the linear space spanned by the *symmetric* 2-tensors,

$$x \vee y := \tfrac{1}{2}(x \otimes y + y \otimes x); \quad x, y \in V.$$

Clearly, $x \vee y = y \vee x$ and $\dim \Sigma_2 V = \binom{d+1}{2}$.

$$(x \vee y, u \vee v) = \tfrac{1}{2}((x,u)(y,v) + (x,v)(y,u)) = \tfrac{1}{2} \operatorname{per} \begin{pmatrix} (x,u) & (x,v) \\ (y,u) & (y,v) \end{pmatrix}.$$

IPR2018-1557
HTC EX1023, Page 123

Likewise, for $0 \leqslant k < \infty$ we define

$$\Sigma_k V := \langle x_1 \vee \ldots \vee x_k | x_i \in V \rangle_R ,$$

where

$$x_1 \vee \ldots \vee x_k = \frac{1}{k!} \sum_{\sigma \in \mathcal{S}_k} x_{\sigma 1} \otimes \ldots \otimes x_{\sigma k} .$$

Then $\Sigma_k V$ has basis $\{e_{i_1} \vee \ldots \vee e_{i_k} |\ 1 \leqslant i_1 \leqslant \ldots \leqslant i_k \leqslant d\}$, and

$$\dim \Sigma_k V = \binom{d+k-1}{k} = h_k(j).$$

The inner product derives from

$$(x_1 \vee \ldots \vee x_k, y_1 \vee \ldots \vee y_k) = \frac{1}{k!} \operatorname{per}((x_i, y_j)).$$

The *symmetric algebra* $\Sigma V = \sum_{k \geqslant 0}^{\oplus} (\Sigma_k V)$ has Poincaré polynomial

$$p_\Sigma(t) = (1-t)^{-d} = H(t, j).$$

Remark

The basis for $\Sigma_k V$ may be normalized using the factors $\left(\frac{k!}{k!}\right)^{\frac{1}{2}}$, where

$k! = k_1! \ldots k_d!$ and $|k| = k_1 + \ldots + k_d = k$, $e_{i_1} \vee \ldots \vee e_{i_k} = e_1^{k_1} \vee \ldots \vee e_d^{k_d}$.

5. Relating Λ to Σ [7,8])

We shall consider power series whose coefficients are vector spaces, with addition defined as direct sum, and multiplication as tensor product.

$$\Lambda(t; V) := \sum_{k \geqslant 0} (\Lambda_k V) t^k; \quad \Sigma(t; V) := \sum_{k \geqslant 0} (\Sigma_k V) t^k.$$

Theorem 5.1

$$\Lambda(t; V \oplus W) = \Lambda(t; V) \otimes \Lambda(t; W); \ \Sigma(t; V \oplus W) = \Sigma(t; V) \otimes \Sigma(t; W).$$

Proof

By considering bases we prove $\Lambda_k(V \otimes W) = \sum_{i+j=k}^{\oplus} \Lambda_i(V) \otimes \Lambda_j(W)$, and the analogous formula for $\Sigma_k(V \oplus W)$.
These imply the relations of the theorem. □

Theorem 5.2

$$\Lambda(-t; V) \otimes \Sigma(t; V) = 1.$$

Proof

First observe that for $\dim V = 1$ we have

$$\Lambda(t; V) = 1 + tV; \quad \Sigma(t; V) = 1 + tV + t^2 V + \ldots = (1 - tV)^{-1}.$$

IPR2018-1557
HTC EX1023, Page 124

For an arbitrary V (which is the direct sum of one-dimensional subspaces), the theorem follows by application of theorem 5.1. □

6. Relating Λ and Σ to S [4,8])

The semigroup $(N, +)$ can be extended to the ring $(Z, +, .)$ by a well-known construction. Similarly, the isomorphism classes $[V]$ of vector spaces V, over a fixed field form a semigroup with respect to direct sums, which can be extended to the Grothendieck ring K, with operations direct sum and tensor product. The map $[V] \to \dim V$ identifies the rings K and Z. For V of dimension d we have

$$[\Lambda_k V] \to \binom{d}{k} \quad \text{and} \quad [\Sigma_k V] \to \binom{d+k-1}{k}.$$

More interesting is the isomorphism of the Grothendieck ring K and the ring S of symmetric functions (in infinitely many variables). Here we have the following correspondences: (V of dimension d, $x = (x_1, \ldots, x_d, 0, 0, \ldots)$)

$$\Lambda_k V \to e_k(x); \quad \Lambda(t; V) \to E(t; x),$$
$$\Sigma_k V \to h_k(x); \quad \Sigma(t; V) \to H(t; x).$$

This shows the connection between theorems 5.2 and 1.1.

Remark

The ring S of the symmetric functions is the underlying ring for the representations of the symmetric group $\mathcal{S}$ in the following sense: Let $\pi: \mathcal{S}_k \to \mathrm{Aut}(V_\pi)$ denote an irreducible representation of $\mathcal{S}_k$ on the $\mathcal{S}$-module V_π. Let $R(\mathcal{S})$ denote the ring of all such $\mathcal{S}$-modules. The fundamental theorem [5] says that there is an isomorphism (of λ-rings)

$$\Theta : R(\mathcal{S}) \simeq S.$$

We indicate how to obtain the symmetric function $\Theta(V_\pi)$: Let X be any vector space, consider $V_\pi \otimes X^{\otimes k}$ and its $\mathcal{S}_k$-invariant subspace $\pi(X)$. For instance, if $\pi = 1$ then $\pi(X) = \Sigma_k(X)$, if $\pi = +/-$ then $\pi(X) = \Lambda_k(X)$. Any linear $T: X \to X$ induces $T_k(T)$: $T_k X \to T_k X$ (see sec. 9) and $\pi(T)$: $\pi(X) \to \pi(X)$. Now the function $\mathrm{spec}(T) \to \mathrm{trace}\,\pi(T)$ is the desired symmetric function in S which is attached to π.

7. Duality [6])

The dual V^* of the vector space V consists of the linear functionals on V. The action of V^* on V may conveniently be described by a pairing $[V, V^*]$: $[v, w^*] := w^*(v)$. If we identify $(T_k V)^*$ with $T_k(V^*)$ we may write

IPR2018-1557
HTC EX1023, Page 125

$$[v_1 \otimes \ldots \otimes v_k, w_1^* \otimes \ldots \otimes w_k^*] = \prod_{i=1}^{k} [v_i, w_i^*] = \prod_{i=1}^{k} w_i^*(v_i).$$

Likewise we have $(\Lambda_k V)^* \simeq \Lambda_k(V^*)$ and $(\Sigma_k V)^* \simeq \Sigma_k(V^*)$ with the induced pairings

$$[v_1 \wedge \ldots \wedge v_k, w_1^* \wedge \ldots \wedge w_k^*] = \frac{1}{k!} \det (w_j^*(v_i)),$$

$$[v_1 \vee \ldots \vee v_k, w_1^* \vee \ldots \vee w_k^*] = \frac{1}{k!} \operatorname{per} (w_j^*(v_i)).$$

Remark

ΣV^* is isomorphic with the ring of polynomial functions on V, with ordinary multiplication. In a certain sense ΛV^* corresponds to the set of square-free polynomials in d variables, cf. sec. 9.

8. Metrics[6])

Let V have basis $e_1, \ldots, e_d$ and let $e_1^*, \ldots, e_d^*$ be the dual basis for V^*, i.e., $[e_i, e_j^*] = \delta_{ij}$.

In V a non-degenerate inner product is defined by a non-singular map $G: V \to V^*$ and $(v, w) := [v, Gw]$, $v, w \in V$. For the orthogonal geometry $O(r, s)$ the standard matrix for G, with respect to the bases e_i and e_j^* is

$$G = \begin{pmatrix} I_r & 0 \\ 0 & -I_s \end{pmatrix}, \quad \text{and} \quad G = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}$$

for the symplectic geometry. The map $G: V \to V^*$ induces $T_k G: T_k V \to T_k V^*$: $x_1 \otimes \ldots \otimes x_k \to Gx_1 \otimes \ldots Gx_k$, which in $T_k V$ leads to the inner product

$$(t, u) := [t, T_k Gu]; \quad t, u \in T_k V,$$

or

$$(x_1 \otimes \ldots x_k, y_1 \otimes \ldots y_k) = \prod_{i=1}^{k} [x_i, Gy_i] = \prod_{i=1}^{k} (x_i, y_i).$$

Likewise the maps $\Lambda_k G: \Lambda_k V \to \Lambda_k V^*$ and $\Sigma_k G: \Sigma_k V \to \Sigma_k V^*$ are defined and

$$(x_1 \wedge \ldots \wedge x_k, y_1 \wedge \ldots \wedge y_k) = \frac{1}{k!} \det ((x_i, y_j)),$$

$$(x_1 \vee \ldots \vee x_k, y_1 \vee \ldots \vee y_k) = \frac{1}{k!} \operatorname{per} ((x_i, y_j)).$$

9. Exterior and symmetric powers of a matrix[5,6,9,10])

Let $A: V \to V$ denote a linear map of V. The k-th tensor power $T_k A$ is defined by $T_k A: x_1 \otimes \ldots \otimes x_k \to Ax_1 \otimes \ldots \otimes Ax_k$. Likewise the maps $\Lambda_k A$ and $\Sigma_k A$ are defined.

IPR2018-1557
HTC EX1023, Page 126

Notation:

For k and l, with $|k| = |l| = k$ the matrix $A(k, l)$ is the $k \times k$ matrix which is obtained from (the $d \times d$ matrix) A by repeating k_i times the i-th row, and l_j times the j-th column, for $i, j = 1, \ldots, d$.

Lemma 9.1

In $(Ax)^k$ the coefficient of x^l equals $\frac{1}{l!}$ per $A(k, l)$.

Proof

$(Ax)^k = \prod_{i=1}^{d} \left(\sum_{j=1}^{d} a_{ij} x_j \right)^{k_i}$. Arrange the terms $a_{ij} x_j$ in a $k \times d$ matrix. Change this into a $k \times k$ matrix with columns corresponding to l. Suppress x^l, then we are left with $A(k, l)$ and the result follows. □

In the final remark of sec. 7 we mentioned that $\Lambda_k V$ corresponds in a certain sense to the square-free polynomials, the sense being that

$$x_{i_1} x_{i_2} \ldots x_{i_k} = (-1)^\sigma x_{i_1} \ldots x_{i_k}$$

for $\sigma \in \mathcal{S}_k$, i.e., the polynomials are *skew*. In this sense one has the following skew version of lemma 9.1:

Lemma 9.1′

In $(Ax)^k$ the coefficient of x^l equals $\det A(k, l)$. Note that $\det A(k, l) \neq 0$ implies $l! = 1$. By use of the orthonormal bases for $\Lambda_k V$ and $\Sigma_k V$ introduced in secs 3 and 4, we can now calculate the entries of the power matrices. □

Theorem 9.2

$$\Lambda_k(A)_{k,l} = \det A(k, l); \quad \Sigma_k(A)_{k,l} = \frac{\operatorname{per} A(k, l)}{\sqrt{k!}\sqrt{l!}}. \qquad \square$$

From the definition we infer:

Theorem 9.3

Let A have eigenvalues $\alpha_1, \ldots, \alpha_d$. The eigenvalues of $\Lambda_k(A)$ are the $\binom{d}{k}$ squarefree, those of $\Sigma_k(A)$ the $\binom{d+k-1}{k}$ homogeneous monomials of degree k in $\alpha_1, \ldots, \alpha_d$. □

The following expansion of the determinant is well-known:

$$\det(I + tA) = 1 + t \sum_{i=1}^{d} a_{ii} + t^2 \sum_{i<j} \begin{vmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{vmatrix} + \ldots + t^d \det(A).$$

Its counterpart for permanents is used less frequently:

IPR2018-1557
HTC EX1023, Page 127

Theorem 9.4

$$\det(I + tA) = \sum_{k=0}^{d} t^k \cdot \operatorname{trace} \Lambda_k(A) = \sum_{k} t^{|k|} \det A(k, k);$$

$$\det{}^{-1}(I - tA) = \sum_{k \geqslant 0} t^k \cdot \operatorname{trace} \Sigma_k(A) = \sum_{k} t^{|k|} \frac{\operatorname{per} A(k, k)}{k!}.$$

Proof

The equalities on the right hand sides follow from theorem 9.2. As for the left hand sides, theorem 9.3 implies that trace $\Sigma_k(A)$ is the k-th complete symmetric polynomial in the eigenvalues α of A, i.e., we have

$$\operatorname{trace} \Sigma_k(A) = h_k(\alpha), \quad \text{similarly trace } \Lambda_k(A) = e_k(\alpha).$$

Hence the sums of the traces are the generating functions $E(t; \alpha)$ and $H(t; \alpha)$ of sec. 1. The theorem now follows from theorem 1.1. □

Corollary 9.5

$$\det(I + tA) = \operatorname{trace} \Lambda(t; A) = \operatorname{trace}^{-1} \Sigma(t; A),$$

for the generating functions

$$\Lambda(t; A) = \sum_{k \geqslant 0} t^k \Lambda_k(A) \text{ and } \Sigma(t; A) = \sum_{k \geqslant 0} t^k \Sigma_k(A).$$

□

10. Applications

I. G. Macdonald told us the following proof of MacMahon's master theorem.

Theorem 10.1

The coefficient of x^k in the symmetric product $(Ax)^k$ equals the coefficient of x^k in $1/\det(I - A\Delta(x))$, where $\Delta(x) = \operatorname{diag}(x_1, \ldots, x_d)$.

Proof

By 9.1 and 9.2 the coefficient of x^k in $(Ax)^k$ equals the (k, k) entry of $\Sigma_k(A)$ where $k = |k|$. Hence it is the coefficient of x^k in trace $\Sigma_k(A\Delta(x))$. However, $\sum_{k \geqslant 0} \operatorname{trace} \Sigma_k(A\Delta(x)) = 1/\det(I - A\Delta(x))$, according to theorem 9.4. □

Theorem 10.2 [1])

$$\frac{(x, Ay)^k}{k!} = \sum_{|k| = |l| = k} \frac{x^k y^l}{k!\, l!} \operatorname{per} A(k, l).$$

Proof

Write $x = x_1 e_1 + \ldots + x_d e_d$, and take the inner products of the symmetric tensors $\frac{1}{k!} x \vee \ldots \vee x$ and $\frac{1}{k!} Ay \vee \ldots \vee Ay$, using

IPR2018-1557
HTC EX1023, Page 128

$$\frac{1}{k!}\, x \vee \ldots \vee x = \sum_{|k|=k} \frac{x^k}{k!}\, e_1^{k_1} \vee \ldots \vee e_d^{k_d}. \qquad \square$$

Corrolary 10.3 (N. G. de Bruijn)

For any rectangular matrix $A = (a_{ij})$ we have

$$\frac{1}{k!} \left(\sum_{i,j} a_{ij}\right)^k = \sum_{|k|=|l|=k} \frac{\operatorname{per} A(k, l)}{k!\, l!}.$$

Proof

Make A into a square matrix by supplying zeros, and apply theorem 10.2. $\square$

11. Fredholm's integral equation [2,3,11])

From the theory of integral equations it follows that

$$u(x) = f(x) + \lambda \int_0^1 K(x, t)\, u(t)\, \mathrm{d}t$$

has the solution

$$u(x) = f(x) + \int_0^1 \frac{D(x, y; \lambda)}{D(\lambda)} f(y)\, \mathrm{d}y,$$

provided $D(\lambda) \neq 0$. To explain the notation, we divide the interval $[0, 1]$ into d equal parts by $0 < \frac{1}{d} < \ldots < \frac{d-1}{d} < 1$. Let f be the d-vector with components $f_i = f\left(\frac{i}{d}\right)$, and let M be the $d \times d$ matrix with entries $M_{ij} = \frac{1}{d} K\left(\frac{i}{d}, \frac{j}{\mathrm{d}}\right)$. Then the integral equation is approximated by the following set of matrix equations of increasing size $d = 1, 2, 3, \ldots$

$$(I - \lambda M)\boldsymbol{u} = f.$$

These equations may be solved using Cramer's rule. Fredholm's determinant $D(\lambda)$ is defined by

$$1 - \lambda \int_0^1 K(t, t)\, \mathrm{d}t + \frac{\lambda^2}{2!} \int_0^1 \int_0^1 \begin{vmatrix} K(t_1, t_1) & K(t_1, t_2) \\ K(t_2, t_1) & K(t_2, t_2) \end{vmatrix} \mathrm{d}t_1\, \mathrm{d}t_2 - \frac{\lambda^3}{3!} \iiint + \ldots .$$

It is the limit, for $d \to \infty$, of

$$\Delta(\lambda) := 1 - \lambda \sum_i M_{ii} + \frac{\lambda^2}{2!} \sum_{i,j} \begin{vmatrix} M_{ii} & M_{ij} \\ M_{ji} & M_{jj} \end{vmatrix} + \ldots + (-1)^d \lambda^d \det M.$$

Fredholm's first minor $D(x, y; \lambda)$ is defined by

IPR2018-1557
HTC EX1023, Page 129

$$\lambda K(x,y) - \lambda^2 \int_0^1 \begin{vmatrix} K(x,y) & K(x,t) \\ K(t,y) & K(t,t) \end{vmatrix} \mathrm{d}t + \frac{\lambda^3}{2!} \int_0^1 \int_0^1 \begin{vmatrix} K(x,y) & K(x,s) & K(x,t) \\ K(s,y) & K(s,s) & K(s,t) \\ K(t,y) & K(t,s) & K(t,t) \end{vmatrix} \mathrm{d}s\,\mathrm{d}t + \ldots .$$

It is the limit for $d \to \infty$, of

$$\Delta(x,y;\lambda) := \lambda M_{xy} - \lambda^2 \sum_i \begin{vmatrix} M_{xy} & M_{xi} \\ M_{iy} & M_{ii} \end{vmatrix} + \frac{\lambda^3}{2!} \sum_{i,j} \begin{vmatrix} M_{xy} & M_{xi} & M_{xj} \\ M_{iy} & M_{ii} & M_{ij} \\ M_{jy} & M_{ji} & M_{jj} \end{vmatrix} + \cdots .$$

By replacing determinants by permanents we get expressions which we denote by $P(-\lambda)$, $\Pi(-\lambda)$, $-P(x,y;-\lambda)$, $-\Pi(x,y;-\lambda)$, respectively. Now in ref. 2 it is proved that

$$D(\lambda)\,P(\lambda) = 1, \quad D(x,y;\lambda)\,P(\lambda) = P(x,y;\lambda)\,D(\lambda). \qquad (*)$$

As a consequence the solution of Fredholm's equation may also be written in terms of permanents:

$$u(x) = f(x) + \int_0^1 \frac{P(x,y;\lambda)}{P(\lambda)} f(y)\,\mathrm{d}y.$$

Now the first equality in (*) is implied by

$$\Delta(\lambda) = \det(I - \lambda M) = (\Pi(\lambda))^{-1}.$$

One can verify that $\Delta(x,y;\lambda)\Pi(\lambda) = \Pi(x,y;\lambda)\Delta(\lambda)$ also holds, but we lack an elegant proof of this fact.

REFERENCES

1) N. Bebiano, Pacific J. Math. **101**, 1 (1982).
2) D. Kershaw, J. Integral Equ. **1**, 281 (1979).
3) N. G. de Bruijn, J. Math. Anal. Appl. **92**, 397 (1983).
4) I. G. MacDonald, Symmetric functions and Hall polynomials, Clarendon Press, Oxford 1979.
5) W. H. Greub, Multilinear algebra, 2nd edition, Springer, New York 1978.
6) R. Shaw, Linear algebra and group representations II, Academic Press, New York 1982.
7) S. Lang, Algebra, Addison-Wesley, Reading Mass. 1965.
8) D. Knutson, λ-rings and the representation theory of the symmetric group, Lecture Notes **308**, Springer, New York 1973.
9) C. C. MacDuffee, The theory of matrices, Chelsea, London 1946.
10) H. Minc, Permanents, Addison-Wesley, Reading Mass. 1978.
11) W. V. Lovitt, Linear integral equations, Dover, New York 1950.
12) H. Minc, Lin. Multilin. Alg. **12**, 227 (1983).

IPR2018-1557
HTC EX1023, Page 130

Philips J. Res. **39**, *121-142, 1984* *R 1087*

SOME ELECTROSTATIC PROBLEMS FOR A SPINDLE

by J. BOERSMA and R. NIJBORG

Department of Mathematics and Computing Science,
Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

Abstract

The classical treatment of the Laplace equation by separation of variables in bispherical coordinates is used to obtain analytical expressions and numerical results for the solution of various electrostatic problems for a perfectly conducting spindle: charged spindle in free space, spindle in a uniform axial or transverse electric field. Special attention is paid to the limit case of a spindle passing into a tight torus, that is a torus without central hole. It is shown that in this limit case the solutions of the potential problems for a spindle tend continuously to the corresponding solutions for a tight torus, as obtained by Belevitch and Boersma. In addition, approximations to the capacity and to the induced dipole moments are derived for the case of a spindle that is almost a tight torus.

Math. Rev.: 78A30, 33A45, 33A40

1. Introduction

In a recent paper [1]) Belevitch and Boersma treated some electrostatic problems for a perfectly conducting torus: charged torus in free space, torus in a uniform axial or transverse electric field. The solutions of these potential problems were derived by separation of variables applied to the Laplace equation in toroidal coordinates. For the limit case of a tight torus, that is a torus without central hole, the solutions of the mentioned problems were obtained by use of Kelvin's inversion. Next it was shown that the solutions of the potential problems for the tight torus coincide with the limits of the solutions for the general torus when the radius of the central hole tends to zero.

In this paper we consider the same electrostatic problems for a perfectly conducting spindle. The spindle is a surface of revolution that is generated by rotating a circular arc around its supporting chord. Both the spindle and the torus can be described by two parameters a, b, where a is the radius of the circular cross-section with a plane through the revolution axis, and b is the signed distance of the cross-section center to the revolution axis of the spindle or torus. The parameters a, b are restricted by $-a < b < a$ for the spindle, and by $b > a$ for the torus. The tight torus is described by equal radii $a = b$, and is therefore limit case of both the torus and the spindle.

IPR2018-1557
HTC EX1023, Page 131

The spindle is a coordinate surface for the orthogonal system of bispherical coordinates. In these coordinates the Laplace equation is known to be R–separable [2]). As detailed in sec. 2 of this paper, the separated solution modes involve conical functions, which are a special type of associated Legendre functions. The electrostatic problems for a spindle can now be solved by the usual separation of variables approach and the solutions are represented by integrals of conical functions. The potential problem of a charged spindle in free space is treated in sec. 3, while secs 4 and 5 deal with the spindle in a uniform axial or transverse electrostatic field. Analytical expressions are presented for the potential, the surface charge density on the spindle, the electrostatic capacity, and the induced dipole moments. Part of our results is known; for example, the capacity of the spindle has been determined by Szegö [3]), and the results for the dipole moments go back at least to Schiffer and Szegö [4]).

For the tight torus, considered as the limit case $b \to a$ of the spindle, the transformation to bispherical coordinates breaks down and the solutions of the potential problems become nugatory. This difficulty is overcome by a suitable rescaling of the bispherical coordinates whereupon, in the limit for $b \to a$, these coordinates tend to tangent-sphere coordinates corresponding to Kelvin's inversion. The limiting process for the tight torus is discussed in detail in sec. 6. Our main purpose is to show that, in the limit for $b \to a$, the solutions of the potential problems for the spindle tend to the corresponding solutions for the tight torus, as determined by Belevitch and Boersma [1]) by use of Kelvin's inversion. Consequently the electrostatic problems for a spindle and for a torus match continuously in the common limit case of a tight torus.

In sec. 7 numerical results are presented for the capacity and the induced dipole moments of the spindle, as functions of the aspect ratio $v = b/a$. Analytical approximations to the capacity and to the dipole moments are derived in sec. 8 for the case of a spindle that is almost a tight torus; this case corresponds to $v \lesssim 1$ or $b \lesssim a$. These approximations are based on two-term Taylor expansions at $v = 1$. It is found that the results of sec. 8 are identical to the approximations [1]) for the case of an almost tight torus, corresponding to $v \gtrsim 1$ or $b \gtrsim a$. Thus we conclude that the capacity and dipole moments considered as functions of $v = b/a$, are continuously differentiable at $v = 1$.

In this paper extensive use is made of various formulas for Legendre and Bessel functions. For convenience these are collected in an appendix, where some new formulas are established as well; a reference such as (A1) is to the corresponding equation of the appendix. The mathematical notations for all special functions are in accordance with our main references [5,6,7]). Rationalized electromagnetic units are used throughout.

IPR2018-1557
HTC EX1023, Page 132

2. The Laplace equation in bispherical coordinates

Let ϱ, θ, z denote the usual cylindrical coordinates as shown in fig. 1, where θ is the angle of rotation in the counterclockwise direction around 0z. Then we define bispherical coordinates β, α, θ (right-handed in this order) by retaining θ and introducing the conformal representation

$$z + i\varrho = c \coth \frac{\alpha - i\beta}{2} \quad (-\infty < \alpha < \infty,\ 0 \leqslant \beta \leqslant \pi), \tag{2.1}$$

where c is a constant. From (2.1) one deduces

$$\varrho = \frac{c \sin \beta}{\cosh \alpha - \cos \beta}, \quad z = \frac{c \sinh \alpha}{\cosh \alpha - \cos \beta}, \tag{2.2}$$

and, through elimination of α,

$$z^2 + \varrho^2 - 2\varrho c \cot \beta = c^2.$$

Hence, the equation $\beta = \beta_1$ defines the spindle

$$z^2 + (\varrho - b)^2 = a^2 \tag{2.3}$$

of fig. 1 with

$$a = \frac{c}{\sin \beta_1}, \quad b = c \cot \beta_1, \quad c^2 = a^2 - b^2, \quad \frac{b}{a} = \cos \beta_1. \tag{2.4}$$

The coordinate surface $\beta = \beta_1$ is apple-shaped if $0 < \beta_1 < \pi/2$ or $0 < b < a$, and spindle-shaped if $\pi/2 < \beta_1 < \pi$ or $-a < b < 0$; see fig. 1. In both cases the surface of revolution is called a spindle. The surface $\beta = \pi/2$ is the sphere of radius $c = a$ around the origin. In the limit case $b = a$ the spindle passes into a tight torus without central hole.

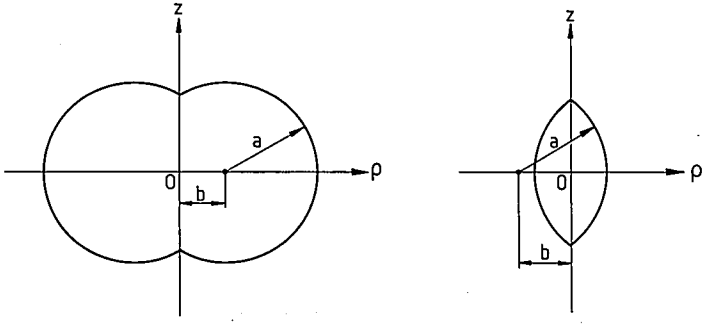


Fig. 1. Cross-section of the spindle $\beta = \beta_1$ with $0 < \beta_1 < \pi/2$ (left) and $\pi/2 < \beta_1 < \pi$ (right).

IPR2018-1557
HTC EX1023, Page 133

The transformation (2.1) sets up a one-to-one correspondence between the half-plane $\varrho \geqslant 0$ of the $(z + i\varrho)$-plane and the strip $-\infty < \alpha < \infty$, $0 \leqslant \beta \leqslant \pi$. Inside the spindle of eq. (2.3) one has $\beta_1 \leqslant \beta \leqslant \pi$, whereas outside the spindle one has $0 \leqslant \beta \leqslant \beta_1$. Both coordinates α and β tend to zero at large distance, i.e. for large

$$r = (\varrho^2 + z^2)^{\frac{1}{2}} = c\left(\frac{\cosh\alpha + \cos\beta}{\cosh\alpha - \cos\beta}\right)^{\frac{1}{2}}. \tag{2.5}$$

The squared line-element is

$$(\mathrm{d}\varrho)^2 + (\mathrm{d}z)^2 = \frac{(\mathrm{d}\alpha)^2 + (\mathrm{d}\beta)^2}{g^2}, \tag{2.6}$$

where

$$g = \frac{\cosh\alpha - \cos\beta}{c}. \tag{2.7}$$

In cylindrical coordinates and for a dependence on θ by a factor $\cos(m\theta)$ or $\sin(m\theta)$, the Laplace equation reduces to

$$\frac{\partial^2 V}{\partial\varrho^2} + \frac{1}{\varrho}\frac{\partial V}{\partial\varrho} + \frac{\partial^2 V}{\partial z^2} - \frac{m^2}{\varrho^2}V = 0. \tag{2.8}$$

Transformation into bispherical coordinates yields the equation

$$\frac{\partial}{\partial\alpha}\left(\frac{\sin\beta}{\cosh\alpha - \cos\beta}\frac{\partial V}{\partial\alpha}\right) + \frac{\partial}{\partial\beta}\left(\frac{\sin\beta}{\cosh\alpha - \cos\beta}\frac{\partial V}{\partial\beta}\right) - \frac{m^2 V}{\sin\beta(\cosh\alpha - \cos\beta)} = 0,$$

which is known to be R–separable. Then, through the substitution [2])

$$V = U(\cosh\alpha - \cos\beta)^{\frac{1}{2}}, \tag{2.9}$$

one gets the simply separable equation

$$\frac{\partial^2 U}{\partial\alpha^2} + \frac{\partial^2 U}{\partial\beta^2} + \cot\beta\frac{\partial U}{\partial\beta} - \left(\frac{1}{4} + \frac{m^2}{\sin^2\beta}\right)U = 0. \tag{2.10}$$

Since the variable α ranges from $-\infty$ to $+\infty$, the solution of (2.10) can be represented by the Fourier integral

$$U(\alpha, \beta) = \int_0^\infty u(t, \beta)\begin{matrix}\sin\\\cos\end{matrix}(\alpha t)\,\mathrm{d}t. \tag{2.11}$$

Substitution into (2.10) yields the ordinary differential equation

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\beta^2} + \cot\beta\frac{\mathrm{d}u}{\mathrm{d}\beta} - \left(\frac{1}{4} + t^2 + \frac{m^2}{\sin^2\beta}\right)u = 0,$$

IPR2018-1557
HTC EX1023, Page 134

which, by the substitution $\xi = \cos\beta$, becomes

$$(1 - \xi^2)\frac{d^2u}{d\xi^2} - 2\xi\frac{du}{d\xi} - \left(\frac{1}{4} + t^2 + \frac{m^2}{1 - \xi^2}\right)u = 0. \qquad (2.12)$$

The latter equation is recognized as the Legendre differential equation with parameters $\nu = -\frac{1}{2} \pm it$, $\mu = \pm m$, where the double signs do not yield distinct solutions. The solutions of (2.12) are thus the associated Legendre functions $P^m_{-\frac{1}{2}+it}$ and $Q^m_{-\frac{1}{2}+it}$ of argument $\xi = \cos\beta$. These special Legendre functions also go by the name of conical functions. Since only potentials vanishing at large distance r (i.e. for small α and β) are needed, they will be systematically presented as integrals in t of modes of the form

$$(\cosh\alpha - \cos\beta)^{\frac{1}{2}} P^m_{-\frac{1}{2}+it}(\cos\beta) \begin{matrix}\sin\\ \cos\end{matrix}(\alpha t) \begin{matrix}\sin\\ \cos\end{matrix}(m\theta), \qquad (2.13)$$

where the initial factor originates from (2.9).

3. The charged spindle in free space

We consider the spindle at constant potential V_0 and isolated in free space. The potential V is represented by an integral of modes (2.13) with $m = 0$ (since V is independent of θ) and even in α. By imposing the boundary condition $V = V_0$ at $\beta = \beta_1$ (on the spindle) and by use of (A2), the solution is found to be

$$V = 2^{\frac{1}{2}} V_0(\cosh\alpha - \cos\beta)^{\frac{1}{2}} \int_0^\infty \frac{P_{-\frac{1}{2}+it}(-\cos\beta_1)}{\cosh(\pi t) P_{-\frac{1}{2}+it}(\cos\beta_1)} \times P_{-\frac{1}{2}+it}(\cos\beta)\cos(\alpha t)\,dt. \qquad (3.1)$$

At large distance r, given by (2.5), the potential V tends to zero as

$$V = \frac{Q}{4\pi\varepsilon r}, \qquad (3.2)$$

where Q is the total charge and ε is the dielectric constant of free space. Thus the value of Q obtains from the limit of rV as $r \to \infty$, or equivalently, as $\alpha, \beta \to 0$. In the product rV the factor $(\cosh\alpha - \cos\beta)^{\frac{1}{2}}$ cancels whereupon the required limit is found by simply setting $\alpha = 0$, $\beta = 0$; notice also that for $\beta = 0$ one has[8]) $P_{-\frac{1}{2}+it}(1) = 1$. From the resulting value of Q we deduce the capacity

$$C = \frac{Q}{V_0} = 4\pi^2 \varepsilon a C_0, \qquad (3.3)$$

IPR2018-1557
HTC EX1023, Page 135

where the normalized capacity is

$$C_0 = \frac{2}{\pi} \sin\beta_1 \int_0^\infty \frac{P_{-\frac{1}{2}+it}(-\cos\beta_1)}{P_{-\frac{1}{2}+it}(\cos\beta_1)} \frac{\mathrm{d}t}{\cosh(\pi t)}. \tag{3.4}$$

The results (3.1) and (3.3) agree with those of Szegö [3]), who points out that (3.1) goes back to C. Neumann in 1881. The result (3.1) is also mentioned by Lebedev et al.[9]).

The charge density on the spindle is

$$\sigma = -\varepsilon \frac{\partial V}{\partial n}\bigg|_{\beta=\beta_1} = \varepsilon g \frac{\partial V}{\partial \beta}\bigg|_{\beta=\beta_1}, \tag{3.5}$$

where g is given by (2.7) and the change of sign is due to the fact that β decreases in the outward normal direction. The derivative in (3.5) is computed by using

$$X \frac{\partial V}{\partial \beta} = \frac{\partial}{\partial \beta}(XV) - V \frac{\partial X}{\partial \beta}, \tag{3.6}$$

where $X = (\cosh\alpha - \cos\beta)^{-\frac{1}{2}}$. In the right-hand side of (3.6) the derivative of XV is determined from (3.1), while $\partial X/\partial\beta$ is replaced by the derivative of (A2); furthermore, $V = V_0$ at $\beta = \beta_1$. Then by use of the Wronskian identity (A7) with $m = 0$, one obtains

$$\sigma = \frac{2^{\frac{3}{2}} \varepsilon V_0}{\pi c} \frac{(\cosh\alpha - \cos\beta_1)^{\frac{3}{2}}}{\sin\beta_1} \int_0^\infty \frac{\cos(\alpha t)}{P_{-\frac{1}{2}+it}(\cos\beta_1)} \mathrm{d}t. \tag{3.7}$$

The surface area element of the spindle is $(\varrho/g)\,\mathrm{d}\alpha\,\mathrm{d}\theta$ and the total charge is

$$Q = \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} \frac{\sigma\varrho}{g} \mathrm{d}\alpha\,\mathrm{d}\theta = 2\pi \int_{-\infty}^{\infty} \frac{\sigma\varrho}{g} \mathrm{d}\alpha. \tag{3.8}$$

Insert the values (2.2) and (2.7) for ϱ and g, and use (3.7) for σ, then the integrand $\sigma\varrho/g$ contains the factor $(\cosh\alpha - \cos\beta_1)^{-\frac{1}{2}}$ which is replaced by the Fourier integral (A2). As a result one obtains an integral of a product of two Fourier transforms, which is evaluated by Parseval's theorem [10]). The final result for Q checks with (3.3).

4. The spindle in a uniform axial electric field

A uniform electric field E_0 is applied along $0z$ corresponding to the potential

$$V_a = -E_0 z = \frac{-E_0 c \sinh\alpha}{\cosh\alpha - \cos\beta} \tag{4.1}$$

IPR2018-1557
HTC EX1023, Page 136

by (2.2). The spindle is understood to be at zero potential, i.e. $V = 0$ at $\beta = \beta_1$. The solution of the potential problem is

$$V = 2^{\frac{3}{2}} c E_0 (\cosh\alpha - \cos\beta)^{\frac{1}{2}} \int_0^\infty \left[\frac{P_{-\frac{1}{2}+it}(-\cos\beta_1)}{P_{-\frac{1}{2}+it}(\cos\beta_1)} P_{-\frac{1}{2}+it}(\cos\beta) - P_{-\frac{1}{2}+it}(-\cos\beta) \right] \frac{t \sin(\alpha t)}{\cosh(\pi t)} \, dt, \qquad (4.2)$$

which is justified as follows. The contribution of the term $P_{-\frac{1}{2}+it}(-\cos\beta)$ is precisely the applied potential (4.1) by (A4). The remainder of (4.2) is an integral of modes (2.13) with $m = 0$ (since V is independent of θ) and odd in α as the potential (4.1). Finally, (4.2) vanishes for $\beta = \beta_1$.

The solution (4.2) consists of the applied potential V_a augmented by a reaction potential. At large distance the reaction potential must reduce to the one of a vertical dipole of moment M at the origin, hence we must have

$$V - V_a = \frac{Mz}{4\pi\varepsilon r^3}, \qquad (4.3)$$

where r is given by (2.5). Thus the moment M can be determined by identifying (4.3) with the value resulting from (4.2) for small α and β. By a computation similar to the one following (3.2) one obtains

$$M = 8\pi^2 \varepsilon a^3 E_0 M_0, \qquad (4.4)$$

where the normalized moment is

$$M_0 = \frac{4}{\pi} \sin^3\beta_1 \int_0^\infty \frac{P_{-\frac{1}{2}+it}(-\cos\beta_1)}{P_{-\frac{1}{2}+it}(\cos\beta_1)} \frac{t^2 \, dt}{\cosh(\pi t)}. \qquad (4.5)$$

The results (4.2) and (4.4) are identical to those of Schiffer and Szegö [4]). The result (4.2) is also mentioned by Lebedev et al. [9]).

The induced charge density on the spindle is given by (3.5), in which $\partial V/\partial\beta$ is computed as in (3.6) and simplified by use of the Wronskian identity (A7) with $m = 0$. Thus we find

$$\sigma = \frac{2^{\frac{5}{2}} \varepsilon E_0}{\pi} \frac{(\cosh\alpha - \cos\beta_1)^{\frac{3}{2}}}{\sin\beta_1} \int_0^\infty \frac{t \sin(\alpha t)}{P_{-\frac{1}{2}+it}(\cos\beta_1)} \, dt. \qquad (4.6)$$

The densities $\sigma(\varrho, z)$ and $\sigma(\varrho, -z) = -\sigma(\varrho, z)$ at opposite points on the same vertical form a moment $2\sigma z$. By integrating this over the upper half of the

IPR2018-1557
HTC EX1023, Page 137

spindle one obtains the total moment

$$M = 2\pi \int_{-\infty}^{\infty} \frac{\sigma z \varrho}{g} \, d\alpha. \tag{4.7}$$

From (2.2), (2.7) and (4.6) it is found that the integrand $\sigma z \varrho / g$ contains the factor $\sinh\alpha(\cosh\alpha - \cos\beta_1)^{-\frac{3}{2}}$ which is replaced by the Fourier integral (A4). Then (4.7) reduces to an integral of a product of two Fourier transforms, which is evaluated by Parseval's theorem. The result for M checks with (4.4).

5. The spindle in a uniform transverse electric field

In the plane perpendicular to $0z$ of fig. 1, we introduce rectangular coordinates x, y related to the polar coordinates ϱ, θ by $x = \varrho\cos\theta$, $y = \varrho\sin\theta$. A uniform electric field E_0 is applied along $0x$ corresponding to the potential

$$V_a = -E_0 x = -E_0 \varrho \cos\theta = \frac{-E_0 c \sin\beta}{\cosh\alpha - \cos\beta} \cos\theta \tag{5.1}$$

by (2.2). Again the spindle is at zero potential, i.e. $V = 0$ at $\beta = \beta_1$. The solution of the potential problem, vanishing for $\beta = \beta_1$, is

$$V = 2^{\frac{3}{2}} c E_0 \cos\theta(\cosh\alpha - \cos\beta)^{\frac{1}{2}} \int_0^{\infty} \left[\frac{P^1_{-\frac{1}{2}+it}(-\cos\beta_1)}{P^1_{-\frac{1}{2}+it}(\cos\beta_1)} P^1_{-\frac{1}{2}+it}(\cos\beta) - P^1_{-\frac{1}{2}+it}(-\cos\beta) \right] \frac{\cos(\alpha t)}{\cosh(\pi t)} \, dt. \tag{5.2}$$

Here the contribution of the term $P^1_{-\frac{1}{2}+it}(-\cos\beta)$ is precisely the applied potential (5.1) by (A3). The remainder of (5.2) is an integral of modes (2.13), even in α and with $m = 1$ as implied by the factor $\cos\theta$ in (5.1).

At large distance the reaction potential $V - V_a$ must be equivalent to the one of a dipole of moment N in the x–direction at the origin; hence we must have

$$V - V_a = \frac{Nx}{4\pi\varepsilon r^3} = \frac{N\varrho\cos\theta}{4\pi\varepsilon r^3}, \tag{5.3}$$

where r is given by (2.5). The moment N is determined by identifying (5.3) with the value resulting from (5.2) for small α and β. Here we need the approximation [8])

$$P^1_{-\frac{1}{2}+it}(\cos\beta) \cong (t^2 + \tfrac{1}{4}) \sin\left(\frac{\beta}{2}\right) \tag{5.4}$$

IPR2018-1557
HTC EX1023, Page 138

valid for small β. The further computation is similar to that following (3.2) and we find

$$N = 2\pi^2 \varepsilon a^3 E_0 N_0, \tag{5.5}$$

where the normalized moment is

$$N_0 = \frac{8}{\pi} \sin^3 \beta_1 \int_0^\infty \frac{P^1_{-\frac{1}{2}+it}(-\cos\beta_1)}{P^1_{-\frac{1}{2}+it}(\cos\beta_1)} \frac{t^2 + \frac{1}{4}}{\cosh(\pi t)} \, dt. \tag{5.6}$$

The results (5.2) and (5.5) are identical to those of Schiffer and Szegö [4]).

The induced charge density on the spindle is given by (3.5), in which $\partial V/\partial\beta$ is computed as in (3.6) and simplified by use of the Wronskian identity (A7) with $m = 1$. Thus we find

$$\sigma = \frac{2^{\frac{5}{2}} \varepsilon E_0}{\pi} \cos\theta \frac{(\cosh\alpha - \cos\beta_1)^{\frac{3}{2}}}{\sin\beta_1} \int_0^\infty \frac{(t^2 + \frac{1}{4}) \cos(\alpha t)}{P^1_{-\frac{1}{2}+it}(\cos\beta_1)} \, dt. \tag{5.7}$$

The densities $\sigma(x, y, z)$ and $\sigma(-x, y, z) = -\sigma(x, y, z)$ at opposite points on the same line parallel to $0x$ form a moment $2\sigma x$. By integrating this over the appropriate half of the spindle one obtains the total moment

$$N = 2 \int_{-\infty}^{\infty} \int_{-\pi/2}^{\pi/2} \frac{\sigma x \varrho}{g} \, d\alpha \, d\theta = \pi \int_{-\infty}^{\infty} \frac{\sigma_1 \varrho^2}{g} \, d\alpha, \tag{5.8}$$

in which $\sigma_1 = \sigma/\cos\theta$. From (2.2), (2.7) and (5.7) it is found that the integrand $\sigma_1 \varrho^2/g$ contains the factor $(\cosh\alpha - \cos\beta_1)^{-\frac{3}{2}}$ which is replaced by the Fourier integral (A3). Then (5.8) reduces to an integral of a product of two Fourier transforms, which is evaluated by Parseval's theorem. The final result for N checks with (5.5).

6. The tight torus as a limit of the spindle for $b \to a$

Consider the spindle of fig. 1 with radii a, b, where $b < a$. When b tends to a by increasing values the spindle passes into a tight torus without central hole. In the limit one has $b = a$, hence $c = 0$ in (2.4), and the conformal representation (2.1) breaks down. To overcome this difficulty we change the bispherical coordinates α and β into $c\alpha/a$ and $c\beta/a$, respectively, before letting c tend to zero. Then the rescaled transformation (2.1) tends to

$$z + i\varrho = \frac{2a}{\alpha - i\beta}, \tag{6.1}$$

IPR2018-1557
HTC EX1023, Page 139

which is Kelvin's inversion. From (6.1) one deduces

$$\varrho = \frac{2a\beta}{\alpha^2 + \beta^2}, \quad z = \frac{2a\alpha}{\alpha^2 + \beta^2}, \tag{6.2}$$

and, through elimination of α,

$$z^2 + \left(\varrho - \frac{a}{\beta}\right)^2 = \left(\frac{a}{\beta}\right)^2. \tag{6.3}$$

Hence the limiting tight torus is described by the equation $\beta = 1$. The coordinates β, α, θ, defined by (6.1), are called tangent-sphere coordinates [11]). The range of α, β is $-\infty < \alpha < \infty$, $\beta \geqslant 0$. Inside the torus one has $1 \leqslant \beta \leqslant \infty$, with $\beta = \infty$ at $\varrho = 0$, $z = 0$. Outside the torus one has $0 \leqslant \beta \leqslant 1$, and both α and β tend to zero at large distance $r = (\varrho^2 + z^2)^{\frac{1}{2}}$.

The electrostatic problems of a charged tight torus in free space, and of a tight torus in a uniform axial or transverse electric field, have been solved by Belevitch and Boersma [12]) using separation of variables in tangent-sphere coordinates. The solutions obtained involve integrals of modified Bessel functions. It is the aim of this section to verify that the solutions for the tight torus ($b = a$) also obtain as limits of the solutions of the corresponding potential problems for the spindle ($b < a$), when $b \to a$. To that end we start from the results for the spindle as derived in secs 3 to 5. In these results we change α, β into $c\alpha/a$, $c\beta/a$, then the latter α, β tend to tangent-sphere coordinates as c tends to zero. To be consistent, one must also change β_1 into $c\beta_1/a$; then the first eq. (2.4) reduces to

$$\frac{c\beta_1}{a} = \arcsin\frac{c}{a}. \tag{6.4}$$

When $c \to 0$, this yields $\beta_1 = 1$ as the equation of the tight torus in tangent-sphere coordinates in agreement with the observation below (6.3). Obviously, $c\beta_1/a$ may be further changed into $c/a = h$, for short. Thus we arrive at the following tentative conclusion: In order to obtain the tight torus as the limit of the spindle, one must make the substitutions

$$\beta_1 \to h, \quad \alpha \to \alpha h, \quad \beta \to \beta h, \tag{6.5}$$

in the results for the spindle established in bispherical coordinates, and then let h tend to zero by positive values; this limit transition should yield the results for the tight torus in tangent-sphere coordinates.

As an example of the limiting process we consider the potential of the charged spindle in free space, given by (3.1). By making the substitutions (6.5) and by changing the integration variable t into t/h, the potential becomes

IPR2018-1557
HTC EX1023, Page 140

$$V = 2^{\frac{1}{2}} V_0 \frac{[\cosh(\alpha h) - \cos(\beta h)]^{\frac{1}{2}}}{h} \int_0^\infty \frac{P_{-\frac{1}{2}+it/h}(-\cos h)}{\cosh(\pi t/h)\, P_{-\frac{1}{2}+it/h}(\cos h)}$$

$$\times P_{-\frac{1}{2}+it/h}(\cos \beta h) \cos(\alpha t)\, dt. \qquad (6.6)$$

Clearly, the factor in front of the integral tends to $V_0(\alpha^2 + \beta^2)^{\frac{1}{2}}$ as $h \to 0$. Let the integrand (6.6) be shortly denoted by $F(t, h)$, then one has, by means of (A13) and (A19),

$$\lim_{h\to 0} F(t, h) = \frac{2}{\pi} \frac{K_0(t)}{I_0(t)} I_0(\beta t) \cos(\alpha t). \qquad (6.7)$$

Furthermore, by use of (A27) and (A29) we establish the estimate

$$|F(t, h)| \leqslant \left(\frac{6\beta h}{\sin \beta h}\right)^{\frac{1}{2}} \frac{K_0(t)}{I_0(t)} I_0(\beta t) = G(t), \qquad (6.8)$$

where it is pointed out that $\int_0^\infty G(t)\, dt$ is convergent. Then by Lebesgue's dominated convergence theorem [13]) the limit transition ($h \to 0$) and the integration may be interchanged in (6.6), and we find

$$\lim_{h\to 0} V = \frac{2V_0}{\pi} (\alpha^2 + \beta^2)^{\frac{1}{2}} \int_0^\infty \frac{K_0(t)}{I_0(t)} I_0(\beta t) \cos(\alpha t)\, dt. \qquad (6.9)$$

The latter expression is recognized as the potential [14]) of the charged tight torus in free space, expressed in the tangent-sphere coordinates α, β. In the same manner it can be shown that all results for the spindle, as derived in secs 3 to 5, tend to the corresponding results [12]) for the tight torus, when h tends to zero. For all results the limit values as $h \to 0$ are determined by means of (A13-17) and (A19-20), and the dominated convergence can be established by use of the estimates (A27-30) and (A34). The latter estimate is needed to demonstrate dominated convergence in the cases of the potential (5.2) and of the dipole moment (5.6). Indeed, consider for example the reaction potential

$$V - V_a = 2^{\frac{3}{2}} a E_0 \cos\theta \frac{[\cosh(\alpha h) - \cos(\beta h)]^{\frac{1}{2}}}{h}$$

$$\times \int_0^\infty \frac{h P^1_{-\frac{1}{2}+it/h}(-\cos h)}{\cosh(\pi t/h)\, P^1_{-\frac{1}{2}+it/h}(\cos h)} P^1_{-\frac{1}{2}+it/h}(\cos \beta h) \cos(\alpha t)\, dt, \qquad (6.10)$$

obtained from (5.2) with the substitutions (6.5) and with t changed into t/h. Let the integrand (6.10) be shortly denoted by $F(t, h)$, then by use of (A28)

IPR2018-1557
HTC EX1023, Page 141

and (A34) we establish the estimate

$$|F(t,h)| \leqslant \frac{\beta h}{\sin(\beta h)} \left(\frac{h}{\sin h}\right)^{\frac{1}{2}} \frac{I_1(\beta t)}{I_1(t)} \left[\frac{2}{\pi} + 6^{\frac{1}{2}}(t^2 + \tfrac{1}{4}h^2)\left(\frac{1}{t^2} - \frac{K_1(t)}{t}\right)\right]. \tag{6.11}$$

Introducing the short notation $G(t)$ for the right-hand side of (6.11), we observe that $\int_0^\infty G(t)\,\mathrm{d}t$ is convergent if $\beta < 1$, hence, Lebesgue's dominated convergence theorem applies. Thus we find, by means of (A15) and (A20),

$$\lim_{h\to 0}(V - V_a) = \frac{4a E_0}{\pi} \cos\theta(\alpha^2 + \beta^2)^{\frac{1}{2}} \int_0^\infty \frac{K_1(t)}{I_1(t)} I_1(\beta t)\, t \cos(\alpha t)\,\mathrm{d}t, \tag{6.12}$$

which is identical to the reaction potential [15]) of the tight torus in a uniform transverse electric field, expressed in the tangent-sphere coordinates α, β.

Summarizing, we have shown that, in the limit for $b \to a$, the solutions of the potential problems for the spindle tend to the corresponding solutions for the tight torus. The latter solutions also coincide with the limits of the solutions of the potential problems for the general torus [1]) when the radius of the central hole tends to zero. Thus we conclude that the electrostatic problems for the spindle ($b < a$) and for the torus ($b > a$) match continuously in the common limit case of the tight torus ($b = a$).

7. Numerical results

In this section we present numerical results for the capacity of the spindle and for the dipole moments induced by a uniform axial or transverse electric field. The normalized quantities C_0, M_0, N_0, are given by (3.4), (4.5), (5.6), represented by integrals of ratios of Legendre functions. The spindle is described by the equation $\beta = \beta_1$ in bispherical coordinates, where β_1 varies from $\beta_1 = \pi$ for the needle-shaped spindle to $\beta_1 = 0$ for the tight torus. Instead of β_1, it is more convenient to take as independent variable in the tabulations the aspect ratio $v = b/a$ of the spindle; by the last eq. (2.4) this ratio is related to β_1 by

$$v = \frac{b}{a} = \cos\beta_1 \tag{7.1}$$

and v now varies from -1 (needle) to 1 (tight torus).

The quantities C_0, M_0, N_0, are tabulated as functions of v in table I for $v = -0.95(0.05)0.95(0.01)1$. For $v < 1$, the numerical results are based on the representations by integrals of Legendre functions, which have been evaluated

IPR2018-1557
HTC EX1023, Page 142

TABLE I

v	$C_0(v)$	$M_0(v)$	$N_0(v)$	v	$C_0(v)$	$M_0(v)$	$N_0(v)$
−0.95	0.036037	0.000795	0.000280	0.15	0.355388	0.203564	0.922637
−0.90	0.058016	0.002650	0.001627	0.20	0.367576	0.219320	1.033255
−0.85	0.077186	0.005406	0.004580	0.25	0.379686	0.235529	1.151919
−0.80	0.094839	0.008996	0.009580	0.30	0.391724	0.252179	1.278871
−0.75	0.111497	0.013378	0.017016	0.35	0.403692	0.269257	1.414352
−0.70	0.127432	0.018521	0.027251	0.40	0.415595	0.286749	1.558599
−0.65	0.142811	0.024398	0.040626	0.45	0.427435	0.304644	1.711850
−0.60	0.157744	0.030986	0.057467	0.50	0.439217	0.322928	1.874340
−0.55	0.172308	0.038265	0.078088	0.55	0.450942	0.341588	2.046302
−0.50	0.186560	0.046217	0.102793	0.60	0.462613	0.360611	2.227969
−0.45	0.200544	0.054822	0.131877	0.65	0.474233	0.379985	2.419572
−0.40	0.214293	0.064066	0.165629	0.70	0.485803	0.399696	2.621342
−0.35	0.227835	0.073932	0.204332	0.75	0.497327	0.419733	2.833506
−0.30	0.241192	0.084404	0.248263	0.80	0.508805	0.440082	3.056294
−0.25	0.254382	0.095468	0.297694	0.85	0.520240	0.460731	3.289931
−0.20	0.267422	0.107108	0.352893	0.90	0.531632	0.481667	3.534644
−0.15	0.280324	0.119311	0.414126	0.95	0.542985	0.502877	3.790656
−0.10	0.293099	0.132062	0.481653	0.96	0.545251	0.507151	3.843234
−0.05	0.305759	0.145348	0.555733	0.97	0.547515	0.511435	3.896275
0.00	0.318310	0.159155	0.636620	0.98	0.549778	0.515729	3.949781
0.05	0.330761	0.173469	0.724567	0.99	0.552039	0.520034	4.003752
0.10	0.343118	0.188276	0.819823	1.00	0.554299	0.52349	4.058192

by numerical integration. For $v = 0$, corresponding to $\beta_1 = \pi/2$, the Legendre functions drop out and the resulting integrals can be evaluated analytically, yielding

$$C_0(0) = \frac{1}{\pi}, \quad M_0(0) = \frac{1}{2\pi}, \quad N_0(0) = \frac{2}{\pi}. \tag{7.2}$$

It is recalled that for $\beta_1 = \pi/2$ the spindle reduces to a sphere. From (7.2) we then deduce

$$C = 4\pi\,\varepsilon\,a, \quad \frac{M}{E_0} = \frac{N}{E_0} = 4\pi\,\varepsilon\,a^3, \tag{7.3}$$

in accordance with the known results for the capacity and the dipole moment of a sphere of radius a. For $v = 1$, the quoted numerical results are those for

IPR2018-1557
HTC EX1023, Page 143

the tight torus, based on the analytical expressions [16])

$$C_0(1) = \frac{4}{\pi^2} \int_0^\infty \frac{dt}{I_0^2(t)} = 0.554299, \tag{7.4}$$

$$M_0(1) = \frac{8}{3\pi^2} \int_0^\infty \frac{t^2}{I_0^2(t)}\, dt = 0.524349, \tag{7.5}$$

$$N_0(1) = \frac{16}{3\pi^2} \int_0^\infty \frac{t^2}{I_1^2(t)}\, dt = 4.058192. \tag{7.6}$$

The finer tabulation step used in the range $0.95 \leqslant v \leqslant 0.99$ clearly shows that the computed values tend continuously to the limit values for the tight torus. Moreover, from a graphical representation it appears that the approach to the limit values for $v = 1$ is close to linear versus v. This point is further elucidated in sec. 8 where analytical expressions for the limit slopes are derived.

8. The spindle that is almost a tight torus

Consider a spindle with radii a, b, such that $b \lesssim a$ or equivalently $v = b/a \lesssim 1$, i.e. the spindle is almost a tight torus. In sec. 7 it was noted that the quantities C_0, M_0, N_0, considered as functions of the parameter v approach almost linearly their limit values for $v = 1$. Therefore it is expected that for $v \lesssim 1$ all quantities can be accurately represented by two-term Taylor expansions at $v = 1$. Thus, for example, the capacity C_0 given by (3.4), is represented by

$$C_0(v) \cong C_0(1) + C_0'(1)\,(v - 1), \tag{8.1}$$

where $C_0(1)$ is the limit value (7.4) for the tight torus and $C_0'(1)$ is the limit slope. Similar two-term expansions hold for the dipole moments $M_0(v)$ and $N_0(v)$. In this section we shall derive analytical expressions for the limit slopes $C_0'(1)$, $M_0'(1)$, $N_0'(1)$. Our analysis resembles that of Belevitch and Boersma [17]) who derived similar analytical approximations to the capacity and to the dipole moments of an almost tight torus, that is a torus with radii a, b, such that $b \gtrsim a$ or $v \gtrsim 1$.

We first state the results which are all expressed in terms of infinite integrals involving Bessel functions. Some of these integrals already occur in (7.4)–(7.6); additional integrals are [18])

IPR2018-1557
HTC EX1023, Page 144

$$\Omega_0 = \frac{4}{\pi^2} \int_0^\infty \left[\frac{4}{t^2} - \frac{1}{I_1^2(t)} \right] \mathrm{d}t = 1.409252, \tag{8.2}$$

$$\Omega_1 = \frac{1}{3\pi^2} \int_0^\infty \left[1 - \frac{1}{I_0^2(t)} \right] \frac{\mathrm{d}t}{t^2} = 0.0411277. \tag{8.3}$$

The results are

$$C_0'(1) = \tfrac{1}{3} C_0(1) + \Omega_1 \quad = 0.225893, \tag{8.4}$$

$$M_0'(1) = M_0(1) - \tfrac{1}{6} C_0(1) = 0.431966, \tag{8.5}$$

$$N_0'(1) = N_0(1) + \Omega_0 \quad = 5.467444. \tag{8.6}$$

It has been checked that the numerical results for the limit slopes agree very well with the values obtained by numerical differentiation in table I.

The derivation of the above results for the limit slopes is explained in some detail for the example of M_0, which for $v < 1$, is given by the integral (4.5). Introduce the small parameter

$$h = 2 \sin\left(\frac{\beta_1}{2}\right), \tag{8.7}$$

and replace the integration variable t by t/h in (4.5), then one has

$$M_0(v) = \frac{4}{\pi} \left(\frac{\sin\beta_1}{2 \sin(\beta_1/2)} \right)^3 \int_0^\infty \frac{P_{-\frac{1}{2}+it/h}(-\cos\beta_1)}{P_{-\frac{1}{2}+it/h}(\cos\beta_1)} \frac{t^2\,\mathrm{d}t}{\cosh(\pi t/h)}. \tag{8.8}$$

The factor in front of the integral is replaced by the two-term approximation

$$\left(\frac{\sin\beta_1}{2 \sin(\beta_1/2)} \right)^3 = \left(1 - \frac{h^2}{4} \right)^{\frac{3}{2}} \cong 1 - \frac{3h^2}{8}. \tag{8.9}$$

Likewise, the ratio of Legendre functions in the integrand (8.8) is replaced by the two-term asymptotic expansion (A19) with δ, ξ changed into h, t. As a result we find for M_0, to order h^2,

$$M_0(v) \cong \frac{8}{\pi^2} \int_0^\infty \frac{t^2 K_0(t)}{I_0(t)} \mathrm{d}t - \frac{h^2}{\pi^2} \int_0^\infty \left[\frac{3t^2 K_0(t)}{I_0(t)} + \frac{t^2 - 1}{3I_0^2(t)} \right] \mathrm{d}t. \tag{8.10}$$

IPR2018-1557
HTC EX1023, Page 145

From (8.7) and (7.1) we have

$$h^2 = 2(1 - \cos\beta_1) = 2(1 - v), \tag{8.11}$$

hence, the two-term expansion (8.10) is of the form (8.1). Its first term yields, by use of (A35), the known value (7.5) for $M_0(1)$. The second term of (8.10) yields

$$M_0'(1) = \frac{2}{\pi^2}\int_0^\infty \left[\frac{3t^2 K_0(t)}{I_0(t)} + \frac{t^2 - 1}{3I_0^2(t)}\right] dt = \frac{2}{3\pi^2}\int_0^\infty \frac{4t^2 - 1}{I_0^2(t)}\, dt \tag{8.12}$$

by means of (A35). The latter result is identical to (8.5) by (7.4) and (7.5). This derivation of (8.5) is admittedly highly heuristic and could only be made rigorous by working along the lines of sec. 6, i.e. by extending the inequalities (A27) and (A29) to the next order. Such an effort is however out of proportion with the interest of the result, which is sufficiently substantiated by the numerical values of table I. Thus we continue with the same heuristic approach for the other results.

The dipole moment N_0 is given by the integral (5.6). Proceeding as before, we insert (8.9) and the two-term asymptotic expansion (A20). As a result we obtain the approximation, to order h^2,

$$N_0(v) \cong \frac{16}{\pi^2}\int_0^\infty \frac{t^2 K_1(t)}{I_1(t)}\, dt - \frac{2h^2}{\pi^2}\int_0^\infty \left[\frac{3t^2 K_1(t)}{I_1(t)} - \frac{2K_1(t)}{I_1(t)} + \frac{t^2 + 3}{3I_1^2(t)}\right] dt. \tag{8.13}$$

The first term yields, by use of (A35), the known value (7.6) for $N_0(1)$. The second integral in (8.13) can be simplified by means of (A35) and (A36), thus leading to the result (8.6) for $N_0'(1)$.

Consider next the capacity C_0, given by the integral (3.4). Replace again t by t/h, and insert the two-term asymptotic expansion (A19) for the ratio of Legendre functions. Then one finds for C_0, to order h^2,

$$C_0(v) \cong \frac{4}{\pi^2}\int_0^\infty \frac{K_0(t)}{I_0(t)}\, dt - \frac{h^2}{2\pi^2}\int_0^\infty \left[\frac{K_0(t)}{I_0(t)} + \frac{1}{3I_0^2(t)} - \frac{1}{3t^2 I_0^2(t)}\right] dt. \tag{8.14}$$

The first term yields, by use of (A35), the known value (7.4) for $C_0(1)$. The second integral in (8.14) is divergent because the final term of the integrand has a singularity $-t^{-2}/3$ at $t = 0$. The same difficulty was encountered in the approximate calculation of the capacity of an almost tight torus [17]). There it was shown that the divergent integral may be replaced by its finite part, obtained by subtracting the singularity of the integrand at $t = 0$. It is expected

IPR2018-1557
HTC EX1023, Page 146

that the same device applies to the second integral in (8.14). Thus by adding the term $t^{-2}/3$ to the integrand and by using (A35) we find

$$C_0'(1) = \frac{1}{3\pi^2}\int_0^\infty \left[\frac{4}{I_0^2(t)} + \frac{1}{t^2}\left(1 - \frac{1}{I_0^2(t)}\right)\right] \mathrm{d}t. \tag{8.15}$$

The latter result is identical to (8.4) by (7.4) and (8.3).

Finally, we compare our results for the spindle (where $b \leq a$) with those of Belevitch and Boersma [17]) for the almost tight torus (where $b \geq a$). Notice the difference in notation: For the almost tight torus the capacity C_0 and the dipole moments M_0, N_0, have been normalized by the radius b, instead of a; furthermore, C_0, M_0, N_0 are considered as functions of the ratio $u = a/b$, which is the inverse of our ratio v. Therefore, to enable a proper comparison, we introduce the renormalized capacity and dipole moments of the spindle, indicated by an upper tilde and defined by

$$\tilde{C}_0(u) = u\, C_0\left(\frac{1}{u}\right), \quad \tilde{M}_0(u) = u^3 M_0\left(\frac{1}{u}\right), \quad \tilde{N}_0(u) = u^3 N_0\left(\frac{1}{u}\right). \tag{8.16}$$

Then the limit slopes at $u = 1$ are readily found to be

$$\tilde{C}_0'(1) = C_0(1) - C_0'(1) = \tfrac{2}{3} C_0(1) - \Omega_1 \qquad = 0.328405, \tag{8.17}$$

$$\tilde{M}_0'(1) = 3M_0(1) - M_0'(1) = 2M_0(1) + \tfrac{1}{6} C_0(1) = 1.141080, \tag{8.18}$$

$$\tilde{N}_0'(1) = 3N_0(1) - N_0'(1) = 2N_0(1) - \Omega_0 \qquad = 6.707132. \tag{8.19}$$

Rather surprisingly, the present results are in exact agreement with the limit slopes [19]) calculated for the almost tight torus. Thus we conclude that the properly normalized capacity and dipole moments, considered as functions of u or v, are continuously differentiable at $u = v = 1$.

APPENDIX. FORMULAS FOR LEGENDRE FUNCTIONS AND BESSEL FUNCTIONS

We start from the integral representation [20])

$$P^m_{-\frac{1}{2}+it}(-\cos\beta) = \frac{2^{\frac{1}{2}}}{\pi^{\frac{3}{2}}}\Gamma(m+\tfrac{1}{2})\cosh(\pi t)(\sin\beta)^m \int_0^\infty \frac{\cos(\alpha t)}{(\cosh\alpha - \cos\beta)^{m+\frac{1}{2}}}\,\mathrm{d}\alpha,$$

valid for non-negative integral m and $0 < \beta \leqslant \pi$. The integral may be considered as a Fourier transform and the inverse transform is

$$(\sin\beta)^m (\cosh\alpha - \cos\beta)^{-m-\frac{1}{2}} = \frac{(2\pi)^{\frac{1}{2}}}{\Gamma(m+\frac{1}{2})}\int_0^\infty \frac{P^m_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)}\cos(\alpha t)\,\mathrm{d}t. \tag{A1}$$

IPR2018-1557
HTC EX1023, Page 147

For $m = 0$ and $m = 1$, (A1) becomes

$$(\cosh\alpha - \cos\beta)^{-\frac{1}{2}} = 2^{\frac{1}{2}} \int_0^\infty \frac{P_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \cos(\alpha t)\, dt, \tag{A2}$$

$$\sin\beta\,(\cosh\alpha - \cos\beta)^{-\frac{3}{2}} = 2^{\frac{3}{2}} \int_0^\infty \frac{P^1_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \cos(\alpha t)\, dt. \tag{A3}$$

Notice that (A3) also results from (A2) by differentiation with respect to β. The derivative of (A2) with respect to α is

$$\sinh\alpha\,(\cosh\alpha - \cos\beta)^{-\frac{3}{2}} = 2^{\frac{3}{2}} \int_0^\infty \frac{P_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} t \sin(\alpha t)\, dt. \tag{A4}$$

The Wronskian[21]) of the Legendre functions $P^m_{-\frac{1}{2}+it}(\cos\beta)$ and $Q^m_{-\frac{1}{2}+it}(\cos\beta)$ is, after simplification by the duplication formula for the Γ-function,

$$P^m_{-\frac{1}{2}+it}(\cos\beta)\frac{d}{d\beta}Q^m_{-\frac{1}{2}+it}(\cos\beta) - Q^m_{-\frac{1}{2}+it}(\cos\beta)\frac{d}{d\beta}P^m_{-\frac{1}{2}+it}(\cos\beta) = \frac{-1}{\sin\beta}\frac{\Gamma(\frac{1}{2}+m+it)}{\Gamma(\frac{1}{2}-m+it)}. \tag{A5}$$

Alternatively, in view of the relation[22])

$$P^m_{-\frac{1}{2}+it}(-\cos\beta) = (-1)^m i\sinh(\pi t) P^m_{-\frac{1}{2}+it}(\cos\beta) + \frac{2}{\pi}(-1)^m \cosh(\pi t) Q^m_{-\frac{1}{2}+it}(\cos\beta), \tag{A6}$$

one may also use $P^m_{-\frac{1}{2}+it}(\cos\beta)$ and $P^m_{-\frac{1}{2}+it}(-\cos\beta)$ as a pair of independent solutions of the Legendre differential equation. The corresponding Wronskian is

$$P^m_{-\frac{1}{2}+it}(\cos\beta)\frac{d}{d\beta}P^m_{-\frac{1}{2}+it}(-\cos\beta) - P^m_{-\frac{1}{2}+it}(-\cos\beta)\frac{d}{d\beta}P^m_{-\frac{1}{2}+it}(\cos\beta) = \frac{2}{\pi}\frac{(-1)^{m+1}}{\sin\beta}\cosh(\pi t)\frac{\Gamma(\frac{1}{2}+m+it)}{\Gamma(\frac{1}{2}-m+it)}. \tag{A7}$$

For small positive β, large $|\nu|$ and integral m, the asymptotic expansions of the Legendre functions of argument $\cos\beta$ are[23])

$$P^{-m}_{\nu}(\cos\beta) = (\nu+\tfrac{1}{2})^{-m}\left[J_m(x) + \frac{\delta^2}{4}\left(\frac{x}{6}J_{m+3}(x) - J_{m+2}(x) + \frac{1}{2x}J_{m+1}(x) + \frac{m}{2}J_m(x)\right)\right], \tag{A8}$$

$$Q^{-m}_{\nu}(\cos\beta) = -\frac{\pi}{2}(\nu+\tfrac{1}{2})^{-m}\left[Y_m(x) + \frac{\delta^2}{4}\left(\frac{x}{6}Y_{m+3}(x) - Y_{m+2}(x) + \frac{1}{2x}Y_{m+1}(x) + \frac{m}{2}Y_m(x)\right)\right] \tag{A9}$$

IPR2018-1557
HTC EX1023, Page 148

to the second order in δ, where

$$\delta = 2\sin\left(\frac{\beta}{2}\right), \quad x = (2\nu + 1)\sin\left(\frac{\beta}{2}\right); \tag{A10}$$

J_m and Y_m are Bessel functions, and the variable x is understood to remain finite. Changing ν into $-\frac{1}{2} + it$, we replace x by

$$x = i\xi, \quad \xi = 2t\sin\left(\frac{\beta}{2}\right). \tag{A11}$$

Then, $J_m(i\xi)$ and $Y_m(i\xi)$ can be expressed in terms of modified Bessel functions I_m and K_m of argument ξ, viz.

$$J_m(i\xi) = i^m I_m(\xi), \quad Y_m(i\xi) = -\frac{2}{\pi}(-i)^m K_m(\xi) + i^{m+1} I_m(\xi). \tag{A12}$$

Furthermore we set $m = 0$ or -1 in (A8) and (A9), retaining only the real part of the expansion (A9). Then all modified Bessel functions can be expressed in terms of functions of orders 0 and 1 by means of their recurrence formulas[24]) and the relations $I_{-m} = I_m$, $K_{-m} = K_m$. As a result the expansions (A8-9) reduce to

$$P_{-\frac{1}{2}+it}(\cos\beta) = I_0(\xi) + \frac{\delta^2}{24}\left[\left(\xi - \frac{1}{\xi}\right) I_1(\xi) + 2I_0(\xi)\right], \tag{A13}$$

$$\mathrm{Re}\, Q_{-\frac{1}{2}+it}(\cos\beta) = K_0(\xi) - \frac{\delta^2}{24}\left[\left(\xi - \frac{1}{\xi}\right) K_1(\xi) - 2K_0(\xi)\right], \tag{A14}$$

$$P^1_{-\frac{1}{2}+it}(\cos\beta) = t\left\{I_1(\xi) + \frac{\delta^2}{24}\left[I_1(\xi) + \left(\xi + \frac{3}{\xi}\right) I_0(\xi)\right]\right\}, \tag{A15}$$

$$\mathrm{Re}\, Q^1_{-\frac{1}{2}+it}(\cos\beta) = -t\left\{K_1(\xi) + \frac{\delta^2}{24}\left[K_1(\xi) - \left(\xi + \frac{3}{\xi}\right) K_0(\xi)\right]\right\}. \tag{A16}$$

Since $P^m_{-\frac{1}{2}+it}(\cos\beta)$ is real-valued, it follows from (A6) that

$$\mathrm{Re}\, Q^m_{-\frac{1}{2}+it}(\cos\beta) = (-1)^m \frac{\pi}{2} \frac{P^m_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)}. \tag{A17}$$

Using the Wronskian identity[25])

$$I_n(\xi) K_{n+1}(\xi) + I_{n+1}(\xi) K_n(\xi) = \frac{1}{\xi} \tag{A18}$$

and working to order δ^2, we then deduce the following expansions for quotients of Legendre functions:

$$\frac{P_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t) P_{-\frac{1}{2}+it}(\cos\beta)} = \frac{2}{\pi}\left\{\frac{K_0(\xi)}{I_0(\xi)} - \frac{\delta^2}{24}\left(1 - \frac{1}{\xi^2}\right)\frac{1}{I_0^2(\xi)}\right\}, \tag{A19}$$

$$\frac{P^1_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t) P^1_{-\frac{1}{2}+it}(\cos\beta)} = \frac{2}{\pi}\left\{\frac{K_1(\xi)}{I_1(\xi)} - \frac{\delta^2}{24}\left(1 + \frac{3}{\xi^2}\right)\frac{1}{I_1^2(\xi)}\right\}. \tag{A20}$$

IPR2018-1557
HTC EX1023, Page 149

By expressing [26]) P_ν^m in terms of P_ν^{-m}, for integral $m \geqslant 0$, and by using the Mehler-Dirichlet integral representation [27]) of $P_\nu^{-m}(\cos\beta)$, one obtains

$$P^m_{-\frac{1}{2}+it}(\cos\beta) = (-1)^m \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}+m+it)}{\Gamma(\frac{1}{2}-m+it)\,\Gamma(m+\frac{1}{2})\,(\sin\beta)^m} \times \int_0^\beta \cosh(ts)\,(\cos s - \cos\beta)^{m-\frac{1}{2}}\,ds. \tag{A21}$$

Notice that the factor in front of the integral is positive. Next, replace β by $\pi - \beta$ and s by $\pi - s$ in (A21), and divide by $\cosh(\pi t)$. Then by use of

$$\frac{\cosh t(\pi - s)}{\cosh(\pi t)} \leqslant 2\exp(-ts), \quad (s \leqslant \pi,\ t \geqslant 0)$$

one has the inequality

$$0 \leqslant \frac{P^m_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \leqslant (-1)^m \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{2\Gamma(\frac{1}{2}+m+it)}{\Gamma(\frac{1}{2}-m+it)\,\Gamma(m+\frac{1}{2})\,(\sin\beta)^m} \times \int_\beta^\pi \exp(-ts)\,(\cos\beta - \cos s)^{m-\frac{1}{2}}\,ds. \tag{A22}$$

The integrals (A21) and (A22) are now compared with integral representations of I_m and K_m, given by [28])

$$I_m(\beta t) = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma(m+\frac{1}{2})} \left(\frac{t}{2\beta}\right)^m \int_0^\beta \cosh(ts)\,(\beta^2 - s^2)^{m-\frac{1}{2}}\,ds, \tag{A23}$$

$$K_m(\beta t) = \frac{\pi^{\frac{1}{2}}}{\Gamma(m+\frac{1}{2})} \left(\frac{t}{2\beta}\right)^m \int_\beta^\infty \exp(-ts)\,(s^2 - \beta^2)^{m-\frac{1}{2}}\,ds. \tag{A24}$$

In the comparison we need the inequalities

$$\frac{\sin\beta}{\beta} \leqslant 2\,\frac{\cos s - \cos\beta}{\beta^2 - s^2} \leqslant 1, \qquad (0 \leqslant s \leqslant \beta) \tag{A25}$$

$$\frac{8}{3\pi^2} \leqslant 2\,\frac{\cos\beta - \cos s}{s^2 - \beta^2} \leqslant \frac{\sin\beta}{\beta}, \qquad \left(\beta \leqslant s \leqslant \pi,\ 0 \leqslant \beta \leqslant \frac{\pi}{2}\right) \tag{A26}$$

which readily follow from $\cos s$ being a decreasing convex function of s^2 for $0 \leqslant s \leqslant \pi$. The restriction $0 \leqslant \beta \leqslant \pi/2$ applies only to the left inequality in (A26), whereby the equality sign holds if $\beta = \pi/2$ and $s = \pi$. Then by using (A25) in the comparison of (A21) with (A23) for $m = 0$ or $m = 1$, one establishes

$$I_0(\beta t) \leqslant P_{-\frac{1}{2}+it}(\cos\beta) \leqslant \left(\frac{\beta}{\sin\beta}\right)^{\frac{1}{2}} I_0(\beta t), \tag{A27}$$

$$\left(\frac{\beta}{\sin\beta}\right)^{\frac{1}{2}} \frac{I_1(\beta t)}{t} \leqslant \frac{P^1_{-\frac{1}{2}+it}(\cos\beta)}{t^2 + \frac{1}{4}} \leqslant \frac{\beta}{\sin\beta}\,\frac{I_1(\beta t)}{t}. \tag{A28}$$

IPR2018-1557
HTC EX1023, Page 150

Similarly, from (A22), (A24) and (A26) we deduce

$$0 \leqslant \frac{P_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \leqslant 6^{\frac{1}{2}} K_0(\beta t), \qquad (0 \leqslant \beta \leqslant \frac{\pi}{2},\ t \geqslant 0) \tag{A29}$$

$$0 \leqslant \frac{P^1_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \leqslant \frac{4}{\pi}\left(\frac{\beta}{\sin\beta}\right)^{\frac{1}{2}} (t^2 + \tfrac{1}{4}) \frac{K_1(\beta t)}{t}, \quad (t \geqslant 0). \tag{A30}$$

Consider next the relation [29])

$$\frac{\mathrm{d}}{\mathrm{d}\beta}[\sin\beta\, P^1_{-\frac{1}{2}+it}(-\cos\beta)] = -(t^2 + \tfrac{1}{4}) \sin\beta\, P_{-\frac{1}{2}+it}(-\cos\beta). \tag{A31}$$

Since, for small β, one has [30])

$$P^{\mu}_{-\frac{1}{2}+it}(-\cos\beta) \cong \frac{\cosh(\pi t)\,\Gamma(\mu)}{\pi\,(\sin\beta/2)^{\mu}}, \quad (\mu > 0) \tag{A32}$$

one obtains, by integrating (A31) and using (A32) with $\mu = 1$,

$$(t^2 + \tfrac{1}{4}) \int_0^{\beta} \sin s \frac{P_{-\frac{1}{2}+it}(-\cos s)}{\cosh(\pi t)}\,\mathrm{d}s = \frac{2}{\pi} - \sin\beta \frac{P^1_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)}. \tag{A33}$$

In the latter integrand we use the estimate (A29) and we set $\sin s \leqslant s$, whereupon the resulting integral is evaluated by

$$\int_0^{\beta} s K_0(s t)\,\mathrm{d}s = -\frac{s K_1(s t)}{t}\Big|_0^{\beta} = \frac{1}{t^2} - \frac{\beta K_1(\beta t)}{t}.$$

Thus we obtain the inequality

$$0 \leqslant \frac{2}{\pi} - \sin\beta \frac{P^1_{-\frac{1}{2}+it}(-\cos\beta)}{\cosh(\pi t)} \leqslant$$

$$\leqslant 6^{\frac{1}{2}}(t^2 + \tfrac{1}{4})\left(\frac{1}{t^2} - \frac{\beta K_1(\beta t)}{t}\right), \quad (0 \leqslant \beta \leqslant \frac{\pi}{2}, t \geqslant 0). \tag{A34}$$

Finally, we present the Bessel-function integrals [31])

$$\int_0^{\infty} \frac{t^{\mu-1} K_{\nu}(t)}{I_{\nu}(t)}\,\mathrm{d}t = \frac{1}{\mu}\int_0^{\infty} \frac{t^{\mu-1}}{I^2_{\nu}(t)}\,\mathrm{d}t, \quad (\mu > 2\nu \geqslant 0) \tag{A35}$$

$$\int_0^{\infty} \left(\frac{K_1(t)}{I_1(t)} - \frac{2}{t^2}\right)\mathrm{d}t = \int_0^{\infty} \left(\frac{1}{I^2_1(t)} - \frac{4}{t^2}\right)\mathrm{d}t. \tag{A36}$$

IPR2018-1557
HTC EX1023, Page 151

Acknowledgement

The authors are indebted to Mr. A. P. M. Baaijens, Eindhoven University of Technology, for his help in carrying out the numerical calculations which underlie table I.

REFERENCES

[1]) V. Belevitch and J. Boersma, Philips J. Res. **38**, 79 (1983).
[2]) P. Moon and D. E. Spencer, Field theory handbook, Springer, Berlin, 1961; cf. p. 110-112.
[3]) G. Szegö, Bull. Amer. Math. Soc. **51**, 325 (1945); cf. p. 345.
[4]) M. Schiffer and G. Szegö, Trans. Amer. Math. Soc. **67**, 130 (1949); cf. p. 149.
[5]) A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, Higher transcendental functions, Vol. 1, McGraw-Hill, New York, 1953.
[6]) L. Robin, Fonctions sphériques de Legendre et fonctions sphéroïdales, Tomes II, III, Gauthier-Villars, Paris, 1958, 1959.
[7]) G. N. Watson, A treatise on the theory of Bessel functions, 2nd Ed., Cambridge University Press, Cambridge, 1958.
[8]) Ref. 5, p. 163, eq. (9).
[9]) N. N. Lebedev, I. P. Skalskaya and Y. S. Uflyand, Problems of mathematical physics, Prentice-Hall, Englewood Cliffs, 1965; cf. p. 246, probl. 510 and 511.
[10]) I. N. Sneddon, Fourier transforms, McGraw-Hill, New York, 1951; cf. p. 25.
[11]) Ref. 2, p. 104-106.
[12]) Ref. 1, secs 8 to 10.
[13]) T. M. Apostol, Mathematical analysis, Addison-Wesley, Reading, 1963; cf. p. 459.
[14]) Ref. 1, p. 96, eq. (8.1).
[15]) Ibid., p. 100, eq. (10.2).
[16]) Ibid., p. 96, eq. (8.4); p. 99, eq. (9.4); p. 101, eq. (10.4).
[17]) Ibid., sec. 14.
[18]) Ibid., p. 102, eq. (10.10), and p. 112; Ω_0 replaces the notation J_0.
[19]) Ibid., p. 112, eqs (14.2), (14.4), (14.6).
[20]) Ref. 6, Tome III, p. 145, eq. (192) with $\theta = \pi - \beta$, $p = t$; a missing multiplicative factor $\sin^m \theta$ has been corrected.
[21]) Ref. 5, p. 146, eq. (25) with $x = \cos\beta$, $\nu = -\frac{1}{2} + it$, $\mu = m$.
[22]) Ibid., p. 144, eq. (14) with $x = \cos\beta$, $\nu = -\frac{1}{2} + it$, $\mu = m$.
[23]) Ref. 6, Tome II, p. 241-250, eqs (43), (45) and (55).
[24]) Ref. 7, p. 79, eq. (1).
[25]) Ibid., p. 80, eq. (20).
[26]) Ref. 5, p. 144, eq. (17).
[27]) Ibid., p. 159, eq. (27).
[28]) Ref. 7, p. 79, last eq. (9); p. 172, eq. (4).
[29]) Ref. 5, p. 161, eqs (19) and (16).
[30]) Ibid., p. 164, eq. (13).
[31]) Ref. 1, p. 126-127, eqs (B11) and (B12).

IPR2018-1557
HTC EX1023, Page 152

Philips J. Res. 39, 143-152, 1984 R 1088

THREE-WEIGHT CODES AND ASSOCIATION SCHEMES

by A. R. CALDERBANK*) and J.-M. GOETHALS**)

*) *AT&T Bell Laboratories, Murray Hill, New Jersey, USA*
**) *Philips Research Laboratory, B-1170 Brussels, Belgium*

Abstract

Three-weight projective codes C are considered for which the restriction to C of the Hamming association scheme $H_n(q)$ is an association scheme with three classes. Sufficient conditions are established and restrictions on the three weights of C are obtained.
It is shown in the binary case that the three-weight subcodes of the shortened second-order Reed-Muller codes provide a large class of examples. Previously known examples were the duals of perfect 3-error-correcting or uniformly packed 2-error-correcting codes.

Math. Rev.: 94B25, 05B30

1. Introduction

Codes C with few distances having the property that the restriction to C of the Hamming association scheme is itself an association scheme were studied by Delsarte[1]), who for linear codes obtained a necessary and sufficient condition: if C is a code with s weights, then the restriction to C of the Hamming scheme is an association scheme with s classes if and only if, among the cosets of its dual code $C^{\perp}$, exactly $s+1$ distinct weight distributions occur. Previously known examples were obtained from the observation made by Delsarte[1]) that if the minimum distance d of $C^{\perp}$ satisfies $d \geqslant 2s-1$, then the weight distribution of any coset of $C^{\perp}$ is uniquely determined by its minimum weight. It was shown by Goethals and van Tilborg[2]) that this situation occurs if and only if $C^{\perp}$ is a uniformly packed quasi perfect code.

In this paper, we investigate further the implications of Delsarte's condition in the case of three-weight codes. We obtain, in particular, a new sufficient condition which applies also when $d<5$. Along the way we also obtain restrictions on the weights of C by a method similar to the one used by Calderbank and Goethals[3]) for the case when $d \geqslant 2s-1$. We also find a large class of examples satisfying our new condition in the binary case.

The paper is organized as follows. In sec. 2, we examine the conditions for a three-weight projective code to yield a 3–class association scheme and obtain

IPR2018-1557
HTC EX1023, Page 153

in theorem 2.2 a new sufficient condition. In sec. 3, we obtain a large class of examples by showing that all three-weight cyclic subcodes of the shortened second-order Reed-Muller codes satisfy our new condition. Finally, in sec. 4, we analyze in more detail the parameters of the association schemes thus obtained.

Our study of three-weight cyclic codes was motivated by the fact that these codes provide periodic sequences with good crosscorrelation properties, as it was shown by Sarwate and Pursley[4]).

2. Three-weight projective codes and association schemes

Let C be a three-weight projective code of length n and dimension k over the finite field GF(q). Thus we assume:
(i) only three distinct nonzero distances occur among codewords of C;
(ii) the dual code $C^{\perp}$ has minimum distance at least equal to 3.
The *distribution matrix* of $C^{\perp}$ has as its set of rows the weight distributions of all of the cosets of $C^{\perp}$. Delsarte[1]) (cf. theorem 6.10, p. 91) proved that the distance relations in an s–weight linear code C define an association scheme A with s classes on C if and only if the distribution matrix of the dual code $C^{\perp}$ contains $s+1$ distinct rows. In this case the set of cosets of $C^{\perp}$ can be partitioned into $s+1$ disjoint subsets, each characterized by a given weight distribution, which we denote by $S_0 = \{C^{\perp}\}, S_1, S_2, \ldots, S_s$. The $s+1$ relations $R_0, R_1, \ldots, R_s$, defined by

$$(C^{\perp} + x,\ C^{\perp} + y) \in R_i \text{ iff } C^{\perp} + (x+y) \in S_i,$$

then yield an s-class association scheme B which is dual to A (cf. Delsarte[1]), Goethals[5]), and Calderbank and Goethals[3])).

It is our purpose to study the classes of three-weight projective codes which yield a pair of dual association schemes as above. From the abovementioned result of Delsarte, it is sufficient to examine which of these codes have the property that among the cosets of their duals only four distinct weight distributions occur.

Since only three weights occur in C, the covering radius of $C^{\perp}$ is at most equal to 3 (cf. Delsarte[6])). Thus the minimum weight of any coset of $C^{\perp}$ can only be one of the following values: 0, 1, 2 or 3. If we assume that all four values, and only four distinct weight distributions, occur among the cosets of $C^{\perp}$, we have to conclude that the weight distribution of any coset is uniquely determined by its minimum weight. Although it is in principle possible to think of other possibilities, we shall restrict our attention to this case.

We know already from a result of Delsarte[6]) (cf. MacWilliams and Sloane[7]), theorem 20, p. 169) that the weight distribution $A_0(x), A_1(x), \ldots, A_n(x)$ of

IPR2018-1557
HTC EX1023, Page 154

any coset $C^{\perp} + x$ is uniquely determined by the first three values $A_0(x)$, $A_1(x)$ and $A_2(x)$, where $A_i(x)$ denotes the number of codewords of $C^{\perp}$ at distance i from x. To be more precise we restate here Delsarte's result in the form of a lemma adapted to our case.

Lemma 2.1 (Delsarte)

Let the Krawtchouk expansion *) of the polynomial

$$F(z) = |C| \prod_{i=1}^{3} \left(1 - \frac{z}{w_i}\right) \tag{1}$$

be given by

$$F(z) = \sum_{i=0}^{3} \alpha_i K_i(z),$$

where w_1, w_2, w_3 are the three weights occurring in C. Moreover, for $k = 1, 2, \ldots, n-3$, let the Krawtchouk expansion of $z^k F(z)$ be given by

$$z^k F(z) = \sum_{i=0}^{k+3} \beta_i K_i(z).$$

Then, for any coset $C^{\perp} + x$, the coefficients $A_i(x)$ of its weight enumerator are related by

$$\sum_{i=0}^{3} \alpha_i A_i(x) = 1, \tag{2}$$

$$\sum_{i=0}^{k+3} \beta_i A_i(x) = 0. \tag{3}$$

The relations (2) and (3), clearly show how all coefficients of the weight enumerator can be obtained from the first three. Let us now examine what the possibilities are for $A_0(x)$, $A_1(x)$ and $A_2(x)$. We first observe that:

(i) for $C^{\perp}$ itself, we have $A_0 = 1$, $A_1 = A_2 = 0$, since the minimum distance is at least 3.

(ii) for any coset of minimum weight equal to 3, we have $A_0 = A_1 = A_2 = 0$. Note that, in this case, we must have, from (2), $\alpha_3 A_3 = 1$.

In both cases the weight distribution is uniquely determined. For each of the remaining cosets $C^{\perp} + x$, let us assume that there exist integers λ_1, λ_2 such that the number of codewords from $C^{\perp}$ at distance 2 from x is given by λ_1, respectively λ_2, if the minimum weight of $C^{\perp} + x$ is 1, respectively 2. This

*) For a definition of the Krawtchouk expansion, we refer, for example, to MacWilliams and Sloane[7]), p. 168.

IPR2018-1557
HTC EX1023, Page 155

means that for the weight distribution of $C^{\perp} + x$, we have either

$$A_0(x) = 0, \quad A_1(x) = 1, \quad A_2(x) = \lambda_1,$$

or

$$A_0(x) = A_1(x) = 0, \quad A_2(x) = \lambda_2,$$

and the conditions are fulfilled to have a pair of dual association schemes.

We now restate the above result in the form of a theorem, which also provides restrictions on the weights occurring in C.

Theorem 2.2

Let C be a three-weight projective code with weights w_1, w_2, w_3, and let its dual $C^{\perp}$ have covering radius equal to 3. If there exist integers λ_1, λ_2 such that for any vector x, the number of codewords from $C^{\perp}$ at distance 2 from x is given by

$$\lambda_1, \text{ if } d(x, C^{\perp}) = 1,$$
$$\lambda_2, \text{ if } d(x, C^{\perp}) = 2,$$

and is zero otherwise, the restriction of the Hamming scheme to C is an association scheme with 3 classes and

$$\frac{q^3}{2\lambda_2} \prod_{i>j} (w_i - w_j)$$

is an integer dividing $|C|^3$.

Proof

From the assumptions, it follows that the weight distribution of any coset of $C^{\perp}$ depends only on its minimum weight. For the first three coefficients we have one of the following four possibilities, depending on the minimum weight:

- if the minimum weight is 0: $A_0 = 1, A_1 = A_2 = 0$;
- if the minimum weight is 1: $A_0 = 0, A_1 = 1, A_2 = \lambda_1$;
- if the minimum weight is 2: $A_0 = A_1 = 0, A_2 = \lambda_2$;
- if the minimum weight is 3: $A_0 = A_1 = A_2 = 0$.

The other coefficients A_i are obtained by the relations (2), (3) of lemma 2.1. Hence, by theorem 6.10 of Delsarte [1]), the distance relations define an association scheme A on C. As indicated at the beginning of this section, we have also a 3–class association scheme B on the set X of cosets of $C^{\perp}$, and A and B are dual schemes. Let S_i be the set of cosets of weight i, for $i = 1, 2, 3$, and let D_i the adjacency matrix of the relation R_i on X, with R_i defined as above, i.e.

$$(C^{\perp} + x, C^{\perp} + y) \in R_i \text{ iff } C^{\perp} + (x + y) \in S_i.$$

IPR2018-1557
HTC EX1023, Page 156

The eigenvalues of D_1 are $n(q-1)$ with multiplicity 1 and $n(q-1)-q\,w_i$ with multiplicity B_i, where B_i is the number of codewords of C with weight w_i, for $i = 1, 2, 3$.
Thus

$$\prod_{i=1}^{3} [D_1 - (n(q-1) - q\,w_i)\,I] = \frac{q^3\,w_1\,w_2\,w_3\,J}{|C|}, \tag{4}$$

where J denotes the all-one matrix of size $|C| = q^k$, and $(D_1 - n(q-1)I)J = 0$.

From our assumptions, it follows that

$$D_1^2 = n(q-1)\,I + (q-2)\,D_1 + 2(\lambda_1 D_1 + \lambda_2 D_2).$$

Hence the eigenvalues of D_2 and of $D_3 = J - I - D_1 - D_2$ can be obtained from those of D_1. Defining the polynomials $F(z)$ and $P_i(z)$ for $i = 1, 2, 3$, as follows:

$$F(z) = |C| \prod_{i=1}^{3} \left(1 - \frac{z}{w_i}\right) = \frac{|C|}{q^3\,w_1\,w_2\,w_3} \prod_{i=1}^{3} (K_1(z) - K_1(w_i)),$$

$$P_1(z) = K_1(z) = n(q-1) - q\,z,$$

$$P_2(z) = \frac{1}{2\lambda_2}\,(K_1^2(z) - 2\lambda_1 K_1(z) - (q-2)\,K_1(z) - n(q-1)),$$

$$P_3(z) = F(z) - (1 + P_1(z) + P_2(z)),$$

we can give the eigenmatrix of the coset scheme the form:

$$P = \begin{bmatrix} 1 & P_1(0) & P_2(0) & P_3(0) \\ 1 & P_1(w_1) & P_2(w_1) & P_3(w_1) \\ 1 & P_1(w_2) & P_2(w_2) & P_3(w_2) \\ 1 & P_1(w_3) & P_2(w_3) & P_3(w_3) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & K_1(0) & K_1^2(0) & K_1^3(0) \\ 1 & K_1(w_1) & K_1^2(w_1) & K_1^3(w_1) \\ 1 & K_1(w_2) & K_1^2(w_2) & K_1^3(w_2) \\ 1 & K_1(w_3) & K_1^2(w_3) & K_1^3(w_3) \end{bmatrix} T, \tag{5}$$

where T is an upper triangular matrix with diagonal entries

$$T_{00} = T_{11} = 1,\ T_{22} = \frac{1}{2\lambda_2},\ T_{33} = \frac{|C|}{q^3\,w_1\,w_2\,w_3}.$$

From (5) we easily compute the determinant of P

$$\det P = \frac{q^3}{2\lambda_2} \prod_{i>j} (w_i - w_j)\,|C|\,.$$

If Q is the dual eigenmatrix then $PQ = |C|\,I$ and the divisibility result is obtained by taking determinants. □

IPR2018-1557
HTC EX1023, Page 157

A special case of theorem 2.2 is when $C^{\perp}$ is a 2–error–correcting code. In this case $\lambda_1 = 0$, $\lambda_2 = 1$, and Calderbank and Goethals [3]) proved as a corollary that there exists an integer t such that either

(i) $w_3 - w_2 = w_2 - w_1 = p^t$ (with $q = p^m$),

(ii) $p = 3$, $w_3 - w_2 = 2.3^t$, and $w_2 - w_1 = 3^t$, or

(iii) $p = 3$, $w_3 - w_2 = 3^t$, and $w_2 - w_1 = 2.3^t$.

Note that the minimum distance d of $C^{\perp}$ is at most equal to 7, cf. Delsarte [6]). If $d = 7$, then $C^{\perp}$ is a *perfect 3–error–correcting code*, as MacWilliams [8]) proved that an e–error–correcting linear code is perfect if and only if there are exactly e weights in the dual code. The only example of a perfect 3–error–correcting code is provided by the [23, 12, 7] binary Golay code whose dual has weights 8, 12 and 16. If $d = 5$ or 6, then $C^{\perp}$ is a *uniformly packed 2–error–correcting code*. Uniformly packed codes are a generalization of perfect codes and were introduced by Semakov, Zinovjev and Zaitzev [9]). Goethals and van Tilborg [2]) proved that an e–error–correcting linear code is uniformly packed if and only if there are exactly $e + 1$ weights in the dual code. For example, the [21, 12, 5] and [22, 12, 6] binary codes obtained by puncturing the [23, 12, 7] Golay code are uniformly packed. Another family of examples is provided by the 2–error–correcting binary BCH codes of length $n = 2^{2m+1} - 1$ ($m \geqslant 2$) and dimension $2^{m+1} - 4m - 3$, for which Kasami [10]) proved that the non-zero codewords in the dual code have weight $2^{2m} - 2^m$, 2^{2m} or $2^{2m} + 2^m$. Further examples and/or possibilities are mentioned by the authors in a previous paper [3]). We note that in all of the above cases, we have $\lambda_1 = 0$, and $\lambda_2 = 1$. In the next section we shall describe some families of examples for which $\lambda_1 \neq 0$ and $\lambda_2 > 1$.

3. Subcodes of the second-order Reed-Muller codes

In this section we shall show that all three-weight cyclic subcodes of RM(2,m)*, the shortened second-order binary Reed-Muller code, satisfy the conditions of theorem 2.2. Kasami [10]) first obtained the weight distributions of these codes. Goethals [11]) described a method for analyzing these codes, based on the theory of alternating bilinear forms. From these results it follows that the only three-weight cyclic codes which are subcodes of RM(2,m)* are those described in the following lemma. For more details the reader may consult MacWilliams and Sloane [7]), chap. 15.

Lemma 3.1

Let C be the cyclic code of length $n = 2^m - 1$ ($m \geqslant 4$) and parity-check polynomial $h(x) = m_1(x)\, m_s(x)$, where $s = 2^i + 1$, for some $i \leqslant m/2$. Then C is a three-weight code if and only if either s is relatively prime to n or m is even and $s = 2^{m/2} + 1$.

IPR2018-1557
HTC EX1023, Page 158

The weight distribution of these codes can be described as follows.

When n and s are relatively prime, we have $(m, 2i) = (m, i) = d$, say. Then, with r defined by $r = \frac{1}{2}(m - d)$, the weight distribution of C is as given in the table below (table I).

TABLE I

w = weight	B_w = number of codewords
$2^{m-1} - 2^{m-1-r}$	$(2^m - 1)(2^r + 1)2^{r-1}$
2^{m-1}	$(2^m - 1)(2^m - 2^{2r} + 1)$
$2^{m-1} + 2^{m-1-r}$	$(2^m - 1)(2^r - 1)2^{r-1}$

For the special case $m = 2t$, $n = 2^{2t} - 1$, $s = 2^t + 1$, the weight distribution is as given in the table below (table II).

TABLE II

w = weight	B_w = number of codewords
$2^{2t-1} - 2^{t-1}$	$(2^t - 1)(2^{2t-1} + 2^{t-1})$
2^{2t-1}	$(2^{2t} - 1)$
$2^{2t-1} + 2^{t-1}$	$(2^t - 1)(2^{2t-1} - 2^{t-1})$

Delsarte and Goethals [12]) showed the relation between the latter codes and multiplicative groups consisting of 2^t Hadamard matrices of order 2^{2t}. Sarwate and Pursley [4]) called the periodic sequences derived from the codewords of these codes "Kasami sequences". Many of the sequences derived from the codes described above have interesting crosscorrelation properties.

Lemma 3.2

Let $C^\perp$ be the dual of the cyclic code described in lemma 3.1 and let C' be the code of length $n + 1$ obtained from $C^\perp$ by adding an overall parity-check. Then C' is left invariant by the doubly transitive group of affine permutations

$$x \rightarrow a x + b$$

in the field GF(2^m).

Proof

The codewords of C' can be described by the subsets X of GF(2^m) satisfying

$$\sum_{x \in X} x = 0, \quad \sum_{x \in X} x^{\tau+1} = 0, \quad |X| = 0 \bmod 2, \tag{6}$$

where we have written τ for 2^i. Under the affine transformation $x \rightarrow ax + b$ the set X is transformed into the set $Y = \{ax + b \mid x \in X\}$.

IPR2018-1557
HTC EX1023, Page 159

Clearly we have $|Y| = |X|$, and

$$\sum_{y \in Y} y = a \sum_{x \in X} x + b|X|,$$

$$\sum_{y \in Y} y^{\tau+1} = a^{\tau+1} \sum_{x \in X} x^{\tau+1} + a^{\tau} b \sum_{x \in X} x^{\tau} + ab^{\tau} \sum_{x \in X} x + b^{\tau+1}|X|.$$

Since moreover $\sum x^{\tau} = (\sum x)^{\tau}$, it follows that Y satisfies the conditions (6) whenever X does. □

Corollary 3.3

The dual $C^{\perp}$ of the codes described in lemma 3.1 have the property that, for any $t \geqslant 2$ and for any pair of coordinate places, the number of codewords of weight $2t - 1$ or $2t$ having a "1" in both places is a constant $\lambda(t)$ independent of the pair of coordinate places.

Proof

This follows from the fact that the extended code C' is left invariant by a doubly transitive group of permutations. Hence, for any $t \geqslant 2$, the codewords of weight $2t$ in C' form a 2–design $2\text{–}(2^m, 2t, \lambda(t))$. □

Corollary 3.4

The codes C described in lemma 3.1 satisfy the conditions of theorem 2.2.

Proof

The dual code $C^{\perp}$ is generated by $m_1(x)m_s(x)$, hence has minimum weight at least equal to 3. From corollary 3.3 it follows that the number of codewords at distance 2 from any vector x is

$$\lambda_1 = \lambda(2), \text{ if } d(x, C^{\perp}) = 1,$$
$$\lambda_2 = 1 + \lambda(2), \text{ if } d(x, C^{\perp}) = 2,$$

with $\lambda(2)$ defined as in corollary 3.3. □

4. Association schemes defined by three-weight subcodes of RM(2, m)*

In this section, we shall study in more detail the association schemes defined by the codes of sec. 3.

4.1. *The special case* $m = 2t, s = 2^t + 1$

In this case the code C, with weight distribution as given in table II, has dimension $3t$. Its dual $C^{\perp}$ has minimum weight equal to 3 and the number of codewords of weight 3 is given by

$$A_3 = \frac{(2^{t-1} - 1)(2^{2t} - 1)}{3}.$$

IPR2018-1557
HTC EX1023, Page 160

The constant $\lambda(2)$ is then given by

$$\lambda(2) = \frac{3A_3}{(2^{2t}-1)} = 2^{t-1} - 1,$$

and $\lambda_1 = \lambda(2)$, $\lambda_2 = 1 + \lambda(2)$, cf. corollary 3.3.

The Krawtchouk expansion of the polynomial $F(z)$, cf (1), is easily computed from the weights in table II. It is given by

$$F(z) = \frac{1}{2^{t-1}}\left[K_0(z) + K_1(z) + \frac{1}{(2^{2t}-1)/3}(K_2(z) + K_3(z))\right].$$

From these data, one can obtain the eigenmatrix (5) of the coset scheme, which is given by

$$P = \begin{array}{c} \begin{matrix} I & D_1 & D_2 & D_3 \end{matrix} \\ \begin{bmatrix} 1 & (2^{2t}-1) & (2^t-1)(2^{2t}-1) & (2^t-1) \\ 1 & (2^t-1) & -(2^t-1) & -1 \\ 1 & -1 & -(2^t-1) & (2^t-1) \\ 1 & -(2^t+1) & (2^t+1) & -1 \end{bmatrix} \end{array}.$$

We observe that D_2 and D_3 are the adjacency matrices of strongly regular graphs, since apart from their valencies, they have only two distinct eigenvalues. Moreover one of these eigenvalues is actually equal to the valency for D_3. This means that the graph defined by the cosets of weight 3 is a disjoint union of cliques of size 2^t. In fact it is easily verified that the association scheme is *imprimitive* in the sense defined by Cameron, Goethals and Seidel [13]). The clique property means that any two cosets of weight 3 are mutually at distance 3. Hence adjoining this set of cosets to the code $C^{\perp}$ gives a code with minimum distance 3 which is 2^t times as large. It is not difficult to verify that the latter is the Hamming code.

The dual scheme which is defined by the distance relations in C is also imprimitive. It is easily verified, for example, that the graph defined by the distance $w_2 = 2^{2t-1}$ is a disjoint union of cliques of size 2^{2t}. The 2^{2t} codewords in any clique are in a one-to-one correspondence with the rows of one of Hadamard matrices described by Delsarte and Goethals [12]).

4.2. *The general case* $s = 2^i + 1$, $(2^i + 1, 2^m - 1) = 1$

In this case the code C, with weight distribution as given in table I, has dimension $2m$. The number of codewords of weight 3 in its dual is given by

$$A_3 = \frac{(2^m-1)(2^{m-1-2r}-1)}{3},$$

where $r = \frac{1}{2}(m-d)$, $d = (m, 2i) = (m, i)$.

IPR2018-1557
HTC EX1023, Page 161

We observe that $A_3 = 0$ if $d = 1$, which occurs if and only if m is odd and i is relatively prime to m. This is the case when $i = 1$, i.e. $s = 3$. We then have the 2-error-correcting BCH codes of length $2^m - 1$, m odd, which are known to be uniformly packed, cf. Goethals and van Tilborg [2]). The above result shows that uniformly packed codes with the same parameters can be obtained for all i that are relatively prime to m. For these codes, we have indeed $A_3 = 0$, hence $\lambda(2) = 0$, $\lambda_1 = 0$, $\lambda_2 = 1$. In the remaining cases, $\lambda(2)$ is given by

$$\lambda(2) = \frac{3A_3}{(2^m - 1)} = 2^{m-1-2r} - 1.$$

The Krawtchouk expansion of $F(z)$, cf. (1), is given by

$$F(z) = \frac{1}{2^m - 2^{m-2r}} [(3 \cdot 2^m - 2^{2m-2r} - 2)(K_0(z) + K_1(z)) + 6(K_2(z) + K_3(z))]$$

and the eigenmatrix of the coset scheme, by

$$P = \begin{array}{c} \begin{matrix} I & D_1 & D_2 & D_3 \end{matrix} \\ \begin{bmatrix} 1 & (2^m - 1) & (2^m - 1)(2^{2r} - 1) & (2^m - 1)(2^m + 1 - 2^{2r}) \\ 1 & (2^{m-r} - 1) & (2^r - 1)(2^{m-r} - 2^r - 1) & -(2^m + 1 - 2^{2r}) \\ 1 & -1 & -(2^{2r} - 1) & (2^{2r} - 1) \\ 1 & -(2^{m-r} + 1) & (2^r + 1)(2^{m-r} - 2^r + 1) & -(2^m + 1 - 2^{2r}) \end{bmatrix} \end{array}.$$

Here also we observe that the graph defined by the cosets of weight 3 is strongly regular. A similar property holds for the graph defined by the distance w_2 in the distance scheme.

REFERENCES

1) P. Delsarte, Philips Res. Repts. **28** Suppl. 10 (1973).
2) J.-M. Goethals and H. C. A. van Tilborg, Philips Res. Repts. **30**, 9 (1975).
3) A. R. Calderbank and J.-M. Goethals, Europ. J. Combinatorics, to appear.
4) D. V. Sarwate and M. B. Pursley, Proc. IEEE **68**, 593 (1980).
5) J.-M. Goethals, Association schemes, in: Algebraic coding theory and applications (G. Longo, ed.), CISM Courses and Lectures No. 258, Springer-Verlag, Wien-New York, 1979, cf. p. 243-283.
6) P. Delsarte, Info. and Control **23**, 407 (1973).
7) F. J. MacWilliams and N. J. A. Sloane, The theory of error-correcting codes, North-Holland Publishing Company, Amsterdam, 1977.
8) F. J. MacWilliams, Combinatorial problems of elementary group theory, Thesis, Dept. of Math., Harvard University, 1962.
9) N. V. Semakov, V. A. Zinovjev and G. V. Zaitzev, Problemy Peredači Informacii **6**, 38 (1971).
10) T. Kasami, Proc. conference on Combinatorial mathematics and its applications, in: R. C. Bose and T. A. Dowling (eds), The Univ. of North Carolina Press, Chapel Hill, N.C., 1969, cf. p. 335-357.
11) J.-M. Goethals, Info. and Control **31**, 43 (1976).
12) P. Delsarte and J.-M. Goethals, Info. and Control **15**, 196 (1969).
13) P. J. Cameron, J.-M. Goethals and J. J. Seidel, Proc. Kon. Nederl. Akad. Wetenschappen, Amsterdam, **A 81**, 196 (1978).

IPR2018-1557
HTC EX1023, Page 162

Philips J. Res. **39**, *153-177, 1984* *R 1089*

A METHOD OF CODE GENERATION FOR ALGORITHMIC LANGUAGES

by P. BRANQUART

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

A method of code generation using a high level intermediate code, the H-code, is described. First the H-code is characterized and general principles of storage allocation are explained, while the notion of static access is defined. Then a simple scheme of H-code generation is sketched and a few examples of application to high level language constructions are given. It is then shown how machine code can be obtained from H-code in a modular way, solving the problems of register allocation, transformation of static accesses into instruction operand addresses, and decomposition of the source functions into instruction *op*-codes. Finally, the machine dependencies of the code generator are summarized and some conclusions based on practical results are drawn.

Math. Rev.: 68B99

1. Introduction

During the past years two important compiler projects have been developed at PRLB:

— an ALGOL 68 [1]) compiler for the X8-ELECTROLOGICA (1968-1973) which has been a pure research project [2]),

— a code generator for CHILL [3,4]) and the ZILOG Z8000 computer (1982-1983), which has been made in collaboration with the development (PLS group, then with MBLE-PTI, now with APT) [5]).

The method used for generating code for ALGOL 68 appeared to be very reliable, while it did not require sophisticated tools. It has been considered appropriate and thus also used for the second project.

The present paper attempts to explain the main lines of that method which is sensibly different from the classical strategies as those explained in the literature, e.g. [6]): instead of introducing the machine dependencies after a relatively low level decomposition of the source language functions (into triples or qua-

IPR2018-1557
HTC EX1023, Page 163

druples), the method keeps the source function level unchanged as long as possible while machine independent storage allocation conventions are introduced at an early stage. This leads to the definition of a high-level intermediate code, the H-code.

In addition to being portable, the H-code provides for a highly modular decomposition of the problem of code generation (hence its reliability), and it allows a good quality for the machine code ultimately generated. Finally, thanks to its high level, it makes it easy to take profit of hardware features implementing possibly intricate source functions.

2. Generalities

A compiler is a program transforming source programs written in a high-level language into target programs which are generally written in a particular machine code. The transformation is made in two main steps: the analysis and the code generation. In the present scheme, code generation is in turn made in two steps: first the H-code is produced and then the machine code. Notice that the separation in steps is purely conceptual: it does not necessarily involve several sequential passes.

In this section we try to situate the intermediate forms of the programs during the compilation (result of the analysis and H-code) with respect to source language and machine code characteristics.

(1) Source languages allow to express actions on data to be taken in some order. Thus source language definitions are based on three things:
   — the source data,
   — the source actions and
   — the order of execution of the actions.

   From a semantic point of view, source data are sets in the mathematical sense, source actions are functions applying to the sets. The order is determined either by the composition rules of the functions or by a more explicit control structure.

(2) Machine languages also express actions on data to be executed in a given order. Machine data and functions are generally much more elementary than source data and source functions. Roughly,
   — data correspond to machine memory units (words or bytes), and
   — functions to machine instructions.

   However, we have no composition rule as such: these are replaced by storage allocation conventions and sequential execution.

   To generate machine code we must thus, on the one hand, define a storage allocation scheme, on the other hand, decompose the source data and functions into target ones. We deal with:

IPR2018-1557
HTC EX1023, Page 164

— data decomposition, by defining a mapping between data and memory units, thus defining the rules of memory representation of the data structures,
— with function decomposition, by defining a mapping between the source functions and sequences of instruction *op*-codes.

Instruction *op*-codes are finally provided with operand addresses on the basis of the storage allocation scheme.

(3) The analysis of the programs gives them a form more manageable for the code generation than the program text; it performs no deep transformation and has two purposes:
— to factor out the source composition rules,
— to detect static (i.e. independent from execution) program errors.

The analysis deals not only with the lexical (i.e. atomic) and context-free (i.e. tree-like) structure of the programs, but also it makes explicit the visibility of the declarations and performs the static mode (type) checking. The result of analysis can be recorded by so called 'decorated trees' as specified by attributed context-free grammars [7]); actually the trees are then represented by linked lists, the elements of which are the sets of properties (e.g. the mode) of the corresponding tree nodes.

Such a representation is very powerful and allows very elaborate optimizations. It is clear however that there is a price to pay for this generality: decorated trees consume memory space, and non-sequential tree traversal algorithms for attribute management are time consuming.

We have used a less general solution:
— the tree is stored in a linear prefix form (PRETREE),
— the declaration structure is implemented by means of a symbol table (SYMBTAB) with an entry per declaration: each tree terminal node corresponding to an application of a declared notation (identifier or operator) is connected to the symbol table entry of its declaration,
— each tree node is connected to a mode (i.e. linked to an entry in a mode table MODETAB): a terminal node corresponding to an object is connected to the mode of the object, and a non-terminal node corresponding to a function is connected to the mode of the result of the function. In addition, we assume that mode analysis leaves explicit tracks of the detection of implicit mode transformations under form of explicit unary operators.

(4) The H-code only makes two assumptions on the target machine:
— the memory of the machine consists of a number of addressable units,
— there is a mechanism of dynamic addressing.

Thus the highly machine dependent particularities are ignored i.e.:

IPR2018-1557
HTC EX1023, Page 165

— the *op*-codes,
— the register structure,
— the addressing modes.

The H-code does not decompose source functions and data, but the composition rules are replaced by storage allocation conventions and sequential execution. In addition the following optimizations are performed:

— the data manipulations and the dynamic memory management instructions are minimized,
— the dynamic checks are replaced by static ones wherever possible.

3. Storage allocation and access mechanisms

3.1. *Language features*

A storage allocation scheme depends on the properties of the source language and moreover it is not always easy to isolate the relevant characteristics from the language definitions. Both ALGOL 68 (A68 for short) and CHILL definitions explicitly make the difference between static and dynamic semantics, which simplifies the task of specifying the storage allocation rules.

We sketch very briefly the main characteristics of the two languages as far as the storage allocation is concerned; later, we shall make some choices when having to be more concrete to illustrate the code generation scheme. Except when explicitly stated otherwise, the described features are available in both languages.

(1) There are values which are set elements, they exist as such; to each set corresponds a specific mode characteristic of the set properties. All values can be manipulated as entities, e.g. they can be passed as parameter or obtained as result of functions.

Values may be simple or composed in which case they consist of other values. The main classical simple values are the integers, and the real and Boolean values, in addition CHILL defines the enumerated set values and the static ranges. The main classical composed values are the arrays and the records. Both A68 and CHILL admit dynamic arrays but CHILL limits their use which allows a more static storage allocation (more precisely in CHILL dynamic arrays can only be allocated by allocators, not by declared variables, and they cannot be elements of other composed values). We shall ignore the A68 flexible arrays in this paper.

In CHILL the memory representation of the composed data structures can be influenced by program information (e.g. using the 'pack' specification).

A68 defines union modes, the use of which is completely secure, CHILL

IPR2018-1557
HTC EX1023, Page.166

defines variant records, the use of which is more flexible but also less secure.

(2) There are locations which are objects with a life-time in the sense that to exist they have to be explicitly created (allocated), and that they may disappear after some time either automatically or because explicitly deallocated. A location has a mode; the mode of the values that can be assigned to it. A68 considers locations as special values called references, and thus a location inherits all the general properties of values, in particular it can be a component of a composed value (practically, an implementation will use a pointer to the location in the representation of the composed value). In CHILL, the locations are a special class of objects, to be a component of a value, a location must first be explicitly *described* by a reference value. The difference between the two philosophies amounts to distinguish explicitly (CHILL) or not (A68) between the locations and their address at the language level.

(3) There are procedures, possibly recursive, that can have parameters (values or locations) and that can provide a result (value or location); the procedures themselves can be manipulated dynamically e.g. assigned to locations, passed as parameters and obtained as result. Both languages consider procedures (text plus environment) as values.

(4) There is a block structure, influencing not only the visibility structure of the programs but also the life-time of some locations: the temporary locations.

(5) There are synonym object declarations associating notations (identifiers) to objects possibly specified dynamically. CHILL, allows only static synonym declarations for values but dynamic ones for locations.

(6) There are allocators, creating locations dynamically each time they are executed. Both languages allow stack and heap allocators. Locations created by stack allocators are deallocated automatically at block exit. Locations created by heap allocators can be explicitly deallocated in CHILL, but in A68, an automatic garbage collection can be implemented.

(7) There are variable declarations which have two effects: the creation of a location, and the association of an identifier to that location. A variable declaration is executed once per activation of the block where it is directly included. It combines the effects of two more primitive concepts: the allocator and the synonym object declaration. (Remark that the parameter passing mechanisms, value and location, can also be expressed in terms of these primitives.) Both languages allow variable declarations but CHILL disallows variables giving rise to locations with dynamic size, thus locations involving dynamic arrays, and variables creating locations on the heap.

IPR2018-1557
HTC EX1023, Page 167

(8) There are expressions producing a result of the same static mode each time they are executed. CHILL never gives rise to partial results without a static limitation in size.
(9) There are control structures consisting of goto, conditional, case and loop statements. A68 being an expression language, each construction (clause) is liable to yield a value. In CHILL, control structures have no result.

Both languages have thus features involving dynamic memory allocation, namely:

— data structures with dynamic size, like dynamic arrays,
— recursive procedures, and
— allocators appearing outside stack variable declaration.

However, CHILL is more restrictive, the essential restrictions being that dynamic arrays can only be created by allocators.

3.2. *Memory representation of values*

The scheme we adopt here is mostly inspired from A68: the use of dynamic arrays is not restricted, but all values are considered to be stored in addressable memory units (no packing information is taken into account).

Values are represented by bit patterns meant to be contained into locations. The bit patterns of simple values have a static size, the same for all values of a given mode. Dynamic arrays involve values with dynamic size.

The composition of values implies a representation where the components can be accessed; for this purpose values with a dynamic size are split into two parts, a static and a dynamic part. The static part occupies contiguous memory units and has a size and a structure known at compile-time; it allows to access to the components of the value. Notice that, since CHILL has no values of dynamic size as components of other values, the splitting is not needed.

For a particular language, the memory representation rules of the values depend on the target computer, and so does the H-code which is based on these rules. However in practice, the machine dependency can be characterized by a few static parameters, namely the static sizes of the simple value modes (i.e. the number of memory units necessary to store simple values of such modes). These parameters are the only machine dependencies of the H-code.

3.3. *Memory organization*

Notational convention: in what follows, memory zones present at run-time are given a notation with the postfix %; this allows to make more explicit the differences between compile-time and run-time devices and avoids confusions.

IPR2018-1557
HTC EX1023, Page 168

In both languages, the execution of a program requires three memory zones, STATIC %, STACK % and HEAP %.

(1) STATIC % is a memory zone with size and structure completely controlled at compile-time. It contains in particular the constant table CONSTAB % (containing the constant values referred to by the generated code) and the locations with a static size and a permanent life-time (i.e. whose life-time is the complete duration of the program). Memory for STATIC % is allocated once and for all at the beginning of the program execution, and never recovered before the end of that execution.

(2) STACK % contains locations that can be reserved and recovered in a last-in first-out manner, i.e. the language and compiler temporary locations which are bound to the nested procedure and block structure of the programs: mainly, the locations implementing the procedure parameters and the temporary variables. Partial results of calculations are also allocated on STACK %. In the sequel, the part of the stack devoted to partial results will be called working stack (WOST %). In general STACK % has a dynamic size.

STACK % is organized in memory frames, one per procedure entered but not yet left. As for composed values, a frame is split up into a static and a dynamic part. In this way, in a frame, data can always be characterized by a static displacement 'p'.

At a given moment of the program execution, only some frames are accessible: there is one accessible frame for the last activation of each procedure surrounding the part of the program currently executed. The frames possibly have a dynamic address, they must in general be accessed via a dynamic addressing mechanism. At each run-time moment, the information needed to access the locations in the accessible procedure frames is a vector containing these frame addresses, thus with one element per nesting level. This vector is called DISPLAY %, it has a static size which corresponds to the maximum nesting level of the procedures in the program; it can thus be in STATIC %. If the DISPLAY % is continuously kept up-to-date dynamically, we can statically characterize data stored on STACK % by pairs consisting of a nesting level n and a frame offset p, say [n.p]. The main difference between CHILL and ALGOL 68 is that in CHILL, if we disallow stack allocators (what most implementations do), the organization of each procedure frame is entirely static, dynamic size objects only appear on the HEAP %.

(3) HEAP % contains the locations not satisfying the above criteria e.g. locations created by allocators and the life-time of which is not bound to the nested program structure. The HEAP % is of such a dynamic

IPR2018-1557
HTC EX1023, Page 169

nature that to access it, an additional indirection level has to be used. Any accessible location on the HEAP % is reached by a chain of pointers (initially a single one) whose first element is necessarily on the STACK %, and can thus be accessed there via a pair [n.p] as explained above.

3.4. *Accesses*

At H-code level we ignore the particularities of the addressing modes and of the register structure of the computer. We replace them by the notion of *access*. An access is an abstraction specifying static informations that will be transformed later into instruction operand addresses, but which is machine independent. The main access classes required by the three memory zones defined in the previous section are as follows:

(1) STATIC % consists of a number of tables of static size, the relative position of which should not be defined now. An access into STATIC % is thus specified by the name of the table and an offset d into it: e.g. {*stat*,CONSTAB %,d} will be a CONSTAB % access. In general, *stat* accesses are very efficient and should be used wherever possible instead of more dynamic ones. Another example is an access to a DISPLAY % element that can be written: {*stat*,DISPLAY %,n}, where n is a procedure nesting level.

(2) A location or a value stored in the static part of a STACK % frame at an address characterized by [n.p], will be given an access {*dirst*,[n.p]}, *dirst* meaning direct stack access. To deal with more dynamic accesses, as HEAP % accesses or those obtained as results of computations (indexings for example), we are led to store dynamic addresses of objects in the static parts of the frames. An object the address of which is obtained by adding a static offset d to the address stored at the access {*dirst*,[n.p]} will be characterized by the indirect access {*indst*,[n.p],d}. It will become clear later why the static displacement d is part of the static access and is not directly added to the dynamic address.

(3) Accesses to the HEAP % must always be indirect, as explained in sec. 3.3.

By analogy, accesses to the code are defined. These are supposed to represent entry points in the programs each being characterized by a sequential numbering L. A code access will be written {*lab*,L}.

Finally, it is useful to be able to characterize static values by special accesses specifying the value itself, e.g. by its relative address d in a compile-time table CONSTAB. Such an access will be written {*const*,d}.

We thus have the following access classes:

IPR2018-1557
HTC EX1023, Page 170

{*stat*,TAB %,offset},
{*dirst*,[n.p]},
{*indst*,[n.p],d},
{*lab*,L}, and
{*const*,d}.

As it is now, the access system is oversimplified, in particular the problems involved by the packing of data structures (non-addressability) as they appear in CHILL are ignored. It is however sufficient to illustrate the H-code generation (for more details see refs 2 and 5).

4. A simple scheme of H-code generation

4.1. *Principle*

Terminology: in what follows we distinguish the notions of *translation* and of *code generation* proper: translation includes code generation, and in addition involves the management of static information such as compile-time table updatings. In some cases, translation involves no code generation at all (several examples are given in sec. 4.3).

At H-code level, source functions and data are not decomposed (there is thus roughly one H-code instruction per source function), but source composition rules are replaced by storage allocation conventions. More precisely, when a source function deals with an object, the corresponding H-code instruction will have as parameters the 'static properties' characterizing that object: a MODE and an ACCESS, written [MODE,ACCESS].

— The mode tells whether the object is a value or a location, and specifies the statically known structure of the value itself, or of the values that are to be put in the locations respectively. Thus the mode is characteristic of the memory structure of the object, according to the memory representation rules. A location mode will be written {*loc*,BM}, and a value mode {*val*,BM}, where BM is called 'base mode' and corresponds to the more classical notion of mode as used in the previous sections.
— The access tells how to find the value, given the memory allocation scheme as explained above.

We generate H-code from the result of the analysis, using a very simple algorithm scanning the prefix form of the tree (PRETREE) sequentially from left to right, and from the symbol table SYMBTAB and the mode table MODETAB, to which the nodes of the tree are connected. This algorithm, due to its sequentiality, does not implement all possible expression optimizations; this must be considered a separate issue without interference with the principle of the H-code.

IPR2018-1557
HTC EX1023, Page 171

We explain the algorithm for conventional functions, those having objects as parameters and possibly delivering an object as result. We shall show in the examples how the same mechanism is easily applicable to other language constructions like declarations and control structures.

The algorithm of translation uses two stacks the bottom stack (BOST) and the top stack (TOPST) to save and recover information:

— TOPST contains the functions which, at each moment of the translation are only partially translated (i.e. all the parameters of which have not yet been dealt with). TOPST makes it possible to recover the function between the translation of two of its parameters and after the translation of all of them. In fact TOPST transmits information from top to bottom in the tree, it implements a particular case of inherited attributes calculated in an ad-hoc way when the tree is scanned from left to right. In the sequel, we shall assume that the attribute 'function result mode', as resulting from mode analysis, is put on TOPST with each function.

— BOST contains the static properties [MODE-ACCESS] of the already translated parameters of the still partially translated functions. In fact, BOST transmits information from bottom to top during the left to right scanning of the tree. It implements a particular case of synthesized attributes.

The algorithm works as follows:

— PRETREE is scanned sequentially,

— when a function is met, it is put on TOPST with the mode of its result (i.e. a pointer to MODETAB), and prefix translation is possibly performed,

— when an elementary parameter is met (i.e. a parameter which is not another function, thus necessarily a constant or an identifier), its static properties are pushed on BOST. The static properties of a constant are deduced from the constant itself, and those of an identifier (denoting a declared object) are found in SYMBTAB, supposing that the declaration has already been treated. Then, if the parameter is not the last one of the current function (as represented at the top of TOPST), infix translation is generated if needed, otherwise postfix translation is generated. For most of functions, postfix translation is in fact the H-code generation proper. It is based on
  — the static properties of the function parameters at the top of BOST,
  — the storage allocation scheme.

Note that the mechanism allows generalizations:

(1) more inherited attributes can be put on TOPST,

(2) TOPST can be scanned from top to bottom to get additional information allowing to generate better code. For example, this process allows to know that the result of the current function will be assigned, in which case the

IPR2018-1557
HTC EX1023, Page 172

result can be directly constructed in the destination and not on WOST %. This mechanism is called the prevision mechanism.

When a function has been translated, it is popped from TOPST and the static properties of its parameters are replaced by those of its result on BOST; the access to the latter result is determined on the basis of the code that has been generated, itself based on the storage allocation scheme.

Finally, the result of the translated function becoming the next parameter of the function at the top of TOPST after the former has been popped, the mechanism goes on recursively.

In what follows, the frame of the algorithm will be assumed and no longer made explicit:

— the current function will be the one at the top of TOPST where the mode of its result can also be found,
— the static properties of its parameters will be those appearing at the top of BOST, and when the static properties of the result of a function are specified, they will be supposed to replace those of the parameters at the top of BOST.

In addition, we suppose that mode constraints are satisfied, and we adopt a scheme where as in CHILL an explicit splitting is made between values and locations, and where the partial results always have a static size.

4.2. *Application to declarations*

Let us take the case of a synonym value declaration associating a value, possibly dynamic, with an identifier. This construction is executed once per activation of the directly surrounding block. The relation identifier-value remains valid up to the moment the current activation of the block is left. We may consider that this function has two parameters specifying respectively

— the identifier which is defined and which, after the visibility analysis, is connected to a SYMBTAB entry,
— the value to be associated with the identifier; at source level this object is specified by an expression.

To translate the declaration, we translate the expression first, we get its mode and the access to its value at the top of BOST. Then, we must ensure that the value can be reaccessed from the identifier up to the moment the current block is left. The strategy is as follows:

— if according to its static access, the value is a CONSTAB value (access of the form {*const*,d}) no code has to be generated and the static properties of the constant are put at the SYMBTAB entry corresponding to the identifier. We thus recover the efficiency of constant definitions,

IPR2018-1557
HTC EX1023, Page 173

— otherwise, code has to be generated to save the value on the STACK %, where it is given an access of the form {*dirst*, – } which is put in SYMBTAB as above. (In the notations, a hyphen replaces information irrelevant in the current context.)

In this way, at each use of the identifier, the static properties of the associated value can be found in SYMBTAB, an assumption which we have made above.

For the sake of simplicity, we have ignored the fact that a declaration can be preceded, in the program text, by uses of the declared notation (for details on the subject see ref. 5).

If the declaration is a synonym location declaration, it is the address of the location which is put on STACK % and the location is characterized by an indirect access {*indst*, – }.

If the declaration is a variable declaration, a location is reserved on STACK % and is characterized by an access {*dirst*,[n.p]}.

4.3. *Applications to conventional functions*

4.3.1. Binary operator

Let OP be any binary operator, $[M_1,A_1]$ be the static properties of its first operand and $[M_2,A_2]$ those of its second operand. In general, the result does not preexist in memory and code has to be generated to perform the calculation and to store the result on the WOST % (registers are ignored). There are however particular cases worth to be treated in a more ad hoc way. The strategy of translation is determined from the static properties of the operands; follow some examples among many other possibilities:

— if the accesses of both operands show that they are CONSTAB values, constant folding is possible, the static result is stored in CONSTAB at an offset d and the result is given an access A_r of the form {*const*,d}. No code is generated,

— suppose OP is the integer adding operator: if one of the operands is the constant 0, the result is characterized by the static properties of the other operand, no code is generated,

— if the result has to be calculated dynamically, we know its static size on the basis of its mode M_r. Space for it is systematically allocated on WOST %, at an access A_r of the form {*dirst*, – }. The H-code to be generated is thus:

$$OP([M_1,A_1],[M_2,A_2],[M_r,A_r]).$$

4.3.2. Assignation

Assignation is an operation by which a value (the source) is stored in a location (the destination). In the program, both source and destination are specified by expressions. The assignation is supposed to have no resulting value. Let

IPR2018-1557
HTC EX1023, Page 174

$[M_d,A_d]$ and $[M_s,A_s]$ be the static properties of the destination and source respectively.

In principle, we just have to generate an H-code instruction of the form ':=([M_d,A_d],[M_s,A_s])', thus delaying any decomposition up to the level of machine code generation. However, we can also take the dynamic checks explicitly into account and replace them by static ones when possible. For example, in case of an assignment to a range location, it is possible, on the basis of the modes M_s and M_d, to detect the cases where the set of values defined by M_s is necessarily included in the one defined by M_d; in this case no dynamic check is needed and we generate a special H-code instruction :='(−), characteristic of this fact.

4.3.3. Selection

The selection is an operation by which a component (field) of a record (value or location) is provided as result. At source level, a selection is specified by an expression delivering the record and a field selector having the form of an identifier. The memory representation of a record is such that in its static part each field has a static offset, say d.

Let $[M_s,A_s]$ be the static properties of the structure, and let M_f be the mode of the field. We define the strategy of H-code generation according to the different access classes of A_s,

- if A_s is of the form {*const*,X} meaning that the record is in the compile-time table CONSTAB, we just characterize the result by $[M_f,A_f]$, where A_f is {*const*,X+d}, assuming here that the displacement d in CONSTAB is the same as in the target computer. No code is generated;
- if A_s is of the form {*dirst*,[n.p]}, the result is similarly characterized by A_f = {*dirst*,[n.p+d]};
- if A_s is of the form {*indst*,[n.p],X}, remember that X is a static displacement with respect to the dynamic object address stored at {*dirst*,[n.p]}. In this case the access to the field is {*indst*,[n.p],X+d}, and no code has to be generated.

If we deal with variant records, and if we select into the variant part, a check of the tag field is generally required: the value of the tag field in the record has to belong to the set 'S' of tag values associated to the variant alternative to which the selector belongs. Such a dynamic check requires the generation of an H-code instruction with as parameters the set S and the static properties of the tag field $[M_t,A_t]$. The latter can be calculated statically in the same manner as $[M_f,A_f]$ above. Thus for tag checking, an H-code instruction of the form

TAGCHECK(S,[M_t,A_t])

has to be generated.

IPR2018-1557
HTC EX1023, Page 175

4.3.4. Indexing

Indexing is an operation by which an array element (value or location) corresponding to a generally dynamic index, is obtained as result. For the sake of simplicity, we make no special case for static indexes, and we only consider one dimensional arrays. Indexing is a typical dynamic action and we have to generate code. In the source program, the array and the index are specified by expressions. The dynamic actions are the following ones:

— execution of the two expressions,
— bound checking, and
— calculation of the address of the resulting array element, i.e. the indexing proper.

(1) Bound checking: if we suppose that the array bounds are static and available in the mode of the array, bound checking is quite similar to range checks in assignation. The same optimization can be done by inspecting the range of the index expression mode with respect to the range of the array index, otherwise code has to be generated.

(2) Indexing proper: to calculate the address of the array element we need the static properties $[M_a,A_a]$ of the array itself and those $[M_i,A_i]$ of the index: A_a characterizes the array address C %, M_a the static memory size S of an array element and $[M_i,A_i]$ the value I % of the index. The element address EA % is thus C % + I %*S, assuming that the array lower bound is 0. The obvious solution to represent the result of the indexing is to store that address EA % on the WOST % at [n1.p1], and to characterize the result by the access A_r = {*indst*,[n1.p1],0}. Thus, if M_r is the mode of the array element, we generate the following H-code instruction:

$$\text{INDEX}([M_a,A_a],[M_i,A_i],[M_r,A_r]),$$

supposed to involve bound checking, and a similar instruction specially flagged if no bound checking is needed.

This strategy can be refined: when the array element is a simple value that takes less or the same space as its address, it is generally more efficient to store that element directly on the working stack at an access A_r of the form {*dirst*,[ni.pi]}. Appropriate H-code instructions must be generated if this strategy is adopted.

4.3.5. Contents taking

Contents taking is an operation (most of the time implicit in the programs), consisting in passing from a location to its associated value or, in terms of implementation, from a location to its contents. Contents taking applies thus to an object with a mode of the form {*loc*,Y} and with some access A.

IPR2018-1557
HTC EX1023, Page 176

To translate the contents taking we just give to the result the mode {*val*,Y} and the same access A. These properties characterize the contents of the original location, no code has to be generated.

4.3.6. Allocators

An allocator is a function which given a mode M creates a location of that mode, with some life-time, and provides as its result a reference value to the created location. We shall treat the case where the location is allocated on the HEAP % and has a static size.

The parameters to be provided to the H-code instruction corresponding to the function are:

— the mode M of the reserved location from which the static size of the location can be deduced, and
— the access A_r of the form {*dirst*, − } of the working space where to store the reference value.

Thus, we generate an instruction of the form:

ALLOC(M,A_r).

4.3.7. Dereferencing

Dereferencing is an operation by which a location is obtained from a reference value. The parameters of this function are the static properties [M_r,A_r] of the reference. We shall chose our strategy of H-code generation according to the access class of A_r, assumed here to be either *dirst* or *indst*.

— If A_r is {*dirst*,[n.p]}, the location can be characterized by [M_l,A_l], where A_l = {*indst*,[n.p],0}. No code has to be generated.
— If however A_r is {*indst*,[n.p],d}, we have no way, with the set of accesses specified so far, to operate on the access only. We have to generate code by which the reference value (i.e. the address of the location) is stored on WOST % e.g. at [nl.pl], and we characterize the location by the access A = {*indst*,[nl.pl],0}. We generate:

DEREF([M_r,A_r],[M,A]),

where M is the mode of the location deduced from M_r.

4.4. *Application to control structures*

4.4.1. Label definition and GOTO

For simplicity, we suppose that GOTO statements are restricted in such a way that they can be executed without precaution about dynamic storage allocation (among other things, we suppose that it is not possible to jump to a label outside of a procedure). For circumventing the problems of forward jumps, we suppose that a previous phase has associated an access A of the form {*lab*,L} to each label definition in SYMBTAB.

IPR2018-1557
HTC EX1023, Page 177

Under the above assumptions, we can consider the GOTO as a function with one parameter A = {*lab*,L}. This function has no result. For a GOTO we thus generate: GOTO(A), or simply GOTO(L).

4.4.2. Conditional statement

Let us consider a source statement of the form:

IF ⟨b⟩ THEN ⟨s1⟩ ELSE ⟨s2⟩ FI,

where ⟨b⟩ is a Boolean expression and, ⟨s1⟩ and ⟨s2⟩ two statements. To simplify we do not treat the case of conditional expressions. The machine code to be generated is of the form:

```
    code for ⟨b⟩
    if ⟨b⟩ is false goto L1
    code for ⟨s1⟩
    goto L2
L1: code for ⟨s2⟩
L2:.
```

We can consider four functions:

— IF, without any action, that could be ignored, but which is kept explicit for the sake of readability.
— THEN, which given the Boolean value of ⟨b⟩ characterized by $[M_b,A_b]$ and L1, causes a jump to L1 if the value of ⟨b⟩ is false,
— ELSE, which causes a jump to L2 and defines L1, and
— FI, which defines L2.

At H-code level we generate:

```
IF
H-code for ⟨b⟩
THEN ([Mb,Ab],L1)
H-code for ⟨s1⟩
ELSE(L1,L2)
H-code for ⟨s2⟩
FI(L2).
```

5. Machine code generation

There are three main problems for generating machine code from H-code:

— register allocation,
— access transformation into instruction addresses, and
— source function decomposition into *op*-codes.

From now on, we have to be machine dependent, but we shall see that we can modularize the problem and isolate the peculiarities of the target computer.

IPR2018-1557
HTC EX1023, Page 178

5.1. O-*code and pseudo-addresses*

Actually, we do not generate machine code as such, but a slightly different form called O-code that is adapted to obtain easily either assembly code or loadable code or executable bit patterns. Each O-code instruction corresponds to one machine instruction; as the latter, it has an *op*-code and operand address specifications. The *op*-code of an O-code instruction represents the same *op*-code as in the machine instruction, but the operand address specifications differ from those in the machine instructions essentially because they are kept relocatable; they are called 'pseudo-addresses'.

A pseudo-address (ps for short) has a category corresponding to an addressing mode in the computer, and a specification allowing to construct the corresponding instruction operand address in an unambiguous way.

Let us assume we have a computer with 4 addressing modes:

— register, where the operand is in a specific register R,
— absolute, where the operand is at a specific address A,
— immediate, where the operand is a specific value V,
— indexed, where the operand is at an address obtained by adding a specific static displacement d to the contents of a specific index register R_x.

The ps corresponding to each addressing mode are as follows:

— to the register mode corresponds a ps of the form {*reg*,R}, where R is the register specification,
— to the absolute mode corresponds a ps of the form {*abs*,SYMB,d}, where SYMB is the symbolic representation of a memory address and d an offset from that address (Note that we can easily pass from that ps to assembler conventions: assuming the assembler directive 'SYMB = −', the address corresponding to the above ps is written 'SYMB + d'.),
— to the immediate addressing mode corresponds a ps {*im*,V}, where V is the integer representation of the operand value,
— to the indexed mode corresponds a ps of the form {*idx*,SYMB,d,R_x}, where SYMB and d have the same interpretation as above, and where R_x specifies the index register.

5.2. *Register management*

5.2.1. Register allocation

Classically, register allocation is based on the following strategies:

— the program is separated in basic blocks, i.e. into parts that are executed sequentially and that have no entry points except the one at the beginning of the block,
— within a basic block track is kept statically in a register table REGTAB, of the dynamic register contents according to the instructions which are

IPR2018-1557
HTC EX1023, Page 179

generated; this contents is used to avoid loading a register with information already contained in a register,

— at each basic block entry, the register table is reset, taking into account the contents of this table at all the program points possibly leading to this basic block,

— the order in which an expression tree is scanned to generate code, is sometimes guided by Sethi-Ullman numbers attached to the nodes; this allows to minimize the number of temporaries and thus the number of register savings and restorings.

These strategies are well known[6]), we briefly describe their adaptation to the H-code interface:

— first of all, the order of translation of the functions is settled at H-code level, it will not be changed here. Thus if we want to take profit of Sethi-Ullman numbers, this must be done at H-code generation level. This obviously requires that we abandon the idea of pure sequential translation, but it does by no means interfere with the principles of the H-code,

— the second adaptation is due to the fact that working space on WOST % has been allocated to all partial results, ignoring the register structure. We shall simply leave that working space unused in the generated code when we can deal with registers only for those partial results. This means waste of space but in small quantity (Notice that it is possible to recover that space by renumbering the STACK % accesses if the memory is very scarce.),

— finally, we must explain the means used to keep a static track of register contents:

(1) If in an H-code instruction, such as OP([M_1,A_1],[M_2,A_2],[M_r,A_r]), a partial result has been given space on WOST %, A_r is a direct access to this space. If during machine code generation, it appears that the value characterized by A_r is in a register R, no instruction is generated to store that value at A_r, but A_r is associated with R in REGTAB. Later, when A_r characterizes an operand in another H-code instruction, REGTAB is first consulted and if there is still a register associated with A_r the value has not to be loaded.

(2) This system can easily be extended to the comparison register (or equivalent hardware flags).

(3) If a partial result is characterized by an indirect access A_i, and if we have generated instructions to load the dynamic address of the operand in an index register R_x, we associate that access to R_x in REGTAB. This allows to avoid reloading an(other) index register with that address.

(4) If an instruction is generated to load the n-th DISPLAY % element into a register R_d, the access {*stat*,DISPLAY %,n} is associated with R_d in REGTAB.

IPR2018-1557
HTC EX1023, Page 180

Many extensions are possible, e.g. variables can be implemented in registers; but we shall limit ourselves here to the above cases.

5.2.2. Tools for register management

To have a systematic register management we have to define a number of tools; the most important are sketched below. These tools are intended to hide the machine dependencies of the register allocation and to increase the portability. We need tools for:

- choosing a register on the basis of required properties and of the contents of REGTAB,
- recording in REGTAB the access A_r to the result of an operation when such a result is known to be in a register,
- erasing the contents of a register table element when the recorded information is no longer valid. For example, at the end of a procedure call with nesting level n, all accesses with pairs of the form [n. −] as recorded in the register table become invalid,
- updating the register table at a program entry point.

In addition, given register allocation is hidden within the above tools, we need means to lock and unlock the use of registers during code generation. An example of application of this is an H-code instruction which has two operands; if the first one is accessible through a register, we must be sure that the transformation of the access (see next section) for the second operand will not destroy the register allowing to access the first one: the registers involved in the first operand are temporarily locked. They can be unlocked as soon as that operand is no longer used. At a given moment a number of registers can be locked, it is the designer responsibility to ensure that there is always a sufficient number of registers available. Note that at the interface between two H-code instructions all registers are unlocked (except possibly those globally locked for special purposes such as Rd0 as explained in the next section). Obviously tools for locking and unlocking registers have to be defined.

The different tools can be implemented as procedure or macro calls provided with appropriate parameters; combined with the data structures defining REGTAB, such a set of tools can be considered as abstract data structure.

5.3. *Access transformation*

The problem is the following: given an access, define the corresponding pseudo-address. The transformation will be made by a procedure (ACCTRANS) which given one of the accesses

IPR2018-1557
HTC EX1023, Page 181

{*stat*,TAB %,d},
{*dirst*,[n.p]},
{*indst*,[n.p],d},
{*lab*,L}, or
{*const*,d},

will deliver one of the following ps:

{*reg*,R},
{*im*,V},
{*abs*,SYMB,d}, or
{*idx*,SYMB,d,R}.

(1) An access of the form {*stat*,TAB %,d} may simply be transformed into {*abs*,TAB %,d}.

(2) An access of the form {*dirst*,[n.p]} may be transformed as follows:
- if there is in REGTAB a register R associated to that access, the ps {*reg*,R} is delivered, otherwise,
- if n = 0: we can access the procedure frame using absolute addressing, and we get the ps {*abs*,STACK %,p} where STACK % is the symbolic representation of the address of the beginning of the STACK %, supposed to be static,
- if n corresponds to the procedure currently translated, we reserve an index register, say Rd0, to contain the address of the current procedure memory frame. To access that frame, we can systematically use the ps{*idx*,ZERO,p,Rd0}, where ZERO is the symbolic representation of the address zero,
- otherwise, we must choose an index register, say R_x, according to the register allocation strategy: if there is a register associated with {*stat*,DISPLAY %,n} in REGTAB, it is taken and has not to be loaded; if not, R_x has to be chosen and a load instruction has to be generated:

LOAD {*abs*,DISPLAY %,n},{*reg*,R_x}.

Then the ps{*idx*,ZERO,p,R_x} can be delivered. Obviously, R_x must be locked as long as required in the context of its use.

The above scheme ensures a good efficiency for computers without hardware facility for display addressing; if such a facility exists, the access transformation becomes straightforward. This shows that keeping the H-code at a sufficient high level allows one to take easily profit of rather sophisticated hardware features.

(3) If the access is of the form {*indst*,[n.p],d}, we need an index register, say R_x, which would contain the indirect address stored at the access {*dirst*,[n.p]}. If according to REGTAB such a register already exists it is

IPR2018-1557
HTC EX1023, Page 182

taken; otherwise we have to load one. For this purpose, we apply the access transformation to {*dirst*,[n.p]} as explained under (2), and get a ps of the form {*idx*,ZERO,p,R_y} corresponding to the access to the dynamic address itself. We then generate

LOAD {*idx*,ZERO,p,R_y},{*reg*,R_x}

and we deliver the ps

{*idx*,ZERO,d,R_x}.

Remark that in most computers, the presence in an indexed addressing mode of a static offset which possibly is not zero does not affect the efficiency; it is to take profit of this property that we have kept such an offset in the indirect accesses.

(4) If the access is of the form {*lab*,L}, it indicates an entry point in the program, and it is to be used in instructions like jumps, subroutine calls or possibly load address instructions. We could deliver the ps {*abs*,CODE %,a}, where CODE % specifies the address of the beginning of the code and 'a' the offset in the code corresponding to L. However it is not advisable to deal with the offsets as such now, and we prefer an intermediate form of ps similar to the original access, say {*lab*,L}. This access will be easily transformed into an instruction address, e.g. if we generate assembly code of the form:

```
L1: —
    —
    JP   L1
```

where L1 is a symbolic address corresponding to L.

(5) The access {*const*,V} is transformed into the ps {*im*,V} when the mode of V shows that it can be used in an immediate addressing mode, otherwise, a ps {*abs*,CONSTAB %,d} is yielded, while the necessary directives are produced to store the constant V in CONSTAB % at the relative address d, at load time.

This simple mechanism supposes that the computer has an orthogonal addressing system, i.e. that most instructions admit most of logically possible addressing modes. Under this condition, the result of an access transformation can always be used as such. If this condition is not fulfilled, we have to parameterize the access transformation in order to forbid some addressing modes according to the context in which the resulting ps will be used. As an example, we could have to forbid the immediate addressing mode in some contexts, in which case it would have to be systematically replaced by a ps of the form {*abs*,CONSTAB %,−}.

IPR2018-1557
HTC EX1023, Page 183

5.4. *Function decomposition*

Although function decomposition might seem to be an ad-hoc task highly dependent on the target computer, the H-code allows a high level of modularity: each H-code instruction can be seen as a separate module, with pre and post conditions completely specified independently from the context. The only interaction between H-code instructions is due to the storage allocation as settled at H-code generation and to the register allocation settled here through the tools defined above.

Moreover, it will be shown that additional tools allow one to take very easily profit of machine peculiarities. This set of tools can even be seen as forming the instructions of a low level code (I-code) which, as compared to the O-code would be more easily adaptable to a class of hardwares by a mechanism of parameterization.

The information we have to generate O-code for an H-code instruction is the following:

— the source function and its semantics as defined at source level,
— the static properties [MODE,ACCESS] of its parameters and result,
— the register allocation and access transformation mechanisms, and
— the definition of the *op*-codes of the computer.

We shall now give two small examples of I-code tools, and thereafter briefly explain the main lines of translation of H-code instructions into O-code. In a last section we shall summarize the problems posed by the retargeting of a code generator based on the principles explained here, and we shall factor out its machine dependencies.

5.4.1. Examples of I-code tools

(1) Suppose we have a computer with a general integer addition instruction (ADD) having two operands, by which the first operand specified by any kind of addressing mode is added to a second operand necessarily contained in a register. In addition, assume the computer has a more efficient instruction increase (INCR), allowing to perform an increment (1st operand) to the contents of a location (2d operand) specified by any addressing mode, but under the condition that the increment is a static (immediate) value in the range (0..63). We define a tool of the form INCREMENT(psi,psd), performing the following actions, where GEN causes the generation of the O-code instruction specified as its parameter, and where psi is the ps of the increment and psd, the ps of the location to be incremented:

— in case psi = {*im*,V} and V is in the range (0..63);

GEN(INCR,psi,psd),

IPR2018-1557
HTC EX1023, Page 184

— in case psd = {*reg*,R}:

GEN(ADD,psi,psd),

— in case psi = {*reg*,R} and psd does not correspond to a register:

GEN(ADD,psd,psi),
GEN(STO,psi,psd),

— otherwise getting a free register R from the register allocation mechanism:

GEN(LOAD,psi,{*reg*,R}),
GEN(ADD,psd,{*reg*,*R*}),
GEN(STO,{*reg*,*R*},psd).

Obviously, we shall use this tool each time we have to generate code for performing an increment (under the condition that if the increment is in a register, it may be destroyed). Only the tool definition is machine dependent, its implementation on a given machine may take profit of the machine peculiarities, or even be, in a first step, implemented in a straightforward way and improved thereafter.

(2) Suppose we have a computer with classical integer multiplication and arithmetic shift instructions. As it is generally the case, the multiplication by a power of 2 can advantageously be replaced by a shift whenever possible. We proceed as above defining a tool MULTIPLY(ps1,ps2), that will choose the best instruction according to ps1 and ps2.

5.4.2. H-code instruction translation

In general, similar things are done for translating all H-code instructions:

— accesses involved in the parameters are transformed into ps, while the necessary register lockings are performed (ACCTRANS),
— then the decomposition proper takes place, and
— finally, registers are unlocked.

Remarks:

(1) When we have a hardware instruction equivalent to the H-one, there is no decomposition, and code generation is straightforward. We can see here another example of the advantage of having kept the H-code at a high level: if we retarget the code generator for a computer with instructions modeling elaborated source functions, we can take profit of them without difficulty. The classical example is an instruction performing an indexing including bound checking. It would be much more difficult to use that instruction if the H-code had decomposed the indexing operation into more elementary ones.

IPR2018-1557
HTC EX1023, Page 185

(2) If there exists a hardware instruction combining several source functions, taking profit of it requires the combination of several H-code instructions. In such a case, it is probably better to make the combination at H-code generation level making use of the prevision mechanism (sec. 4.1). As an example, instead of generating two H-code instructions for 'x := x + a', we could generate a single one which would exactly correspond to a hardware increment instruction.

(3) For two cases, we have to deal with genuine decomposition:
   — The first case corresponds to a source function applying to objects of any mode: we generally have to split the decomposition into several subcases on the basis of that mode. E.g., for the assignation, there are trivial cases where hardware move instructions can be directly used, but there are cases where this is not sufficient and where a dynamic traversal of the data to be assigned and of the destination is needed. The classical case is the one where the data consist of a nesting of records and dynamic arrays for which bound checking at any level of nesting is needed. It is to be noted that the frame of the algorithm of data structure traversal is not machine dependent and that it can be kept nearly unmodified when the compiler is retargeted. (For further details see ref. 2.)
   — The second case is bound to the administration of the procedure memory frames, the structure of the latter being the basis on which the H-code has defined the accesses to the objects. If there is no special hardware facility for such an administration there must be a genuine decomposition. Otherwise, either the conventions adopted at H-code level can be mapped on the hardware facilities and no decomposition is needed, or we have the choice between two strategies: (a) we ignore the hardware facilities and we decompose as in the first case or, (b) we adapt the procedure frame administration to the hardware facilities, which implies some modifications of the static storage allocation tools at the H-code level.

6. Summary of the machine dependencies

The machine dependencies of the code generation mechanism described above are localized in the following parts:

— H-code: the algorithm associating memory sizes with the data modes, but this amounts in general to the redefinition of a few constants,
— register allocation: some tools may need adaptations when the register structure changes significantly,
— access transformation: must be adapted to the available addressing modes. As explained in sec. 5.3, the availability of elaborated addressing modes to similate display addressing can easily be used,

IPR2018-1557
HTC EX1023, Page 186

— I-code tools: must be redefined, but as stated above this can be done in several steps, the first ones not necessarily taking all peculiarities of the new hardware into account,
— function decomposition: may be kept unchanged, if it only uses I-code tools for generating O-code. Moreover, as explained in sec. 5.4.2, it is easy to take profit of new hardware instructions modeling the source functions.

To deal with new hardware instruction implementing several source functions at the same time, it is advisable to define new H-code instructions. Similarly, for being able to use new hardware conventions dealing with the administration of the procedure memory frames, some modifications at H-code level are required. These last two modifications of the H-code are not compulsory, they can be ignored in a first step and introduced later to improve the efficiency.

6. Conclusion

The method described in this paper is essentially based on a logical division of the problems leading to a highly modular design. It results in a final product which is easily maintainable, and which seems to be a good compromise between the efficiency of the produced code and the portability of the code generator.

The ALGOL 68 compiler has proved to be very reliable and its efficiency can be advantageously compared with the one of an existing ALGOL 60 compiler for the X8.

The CHILL code generator, due to an early unavailability of the hardware, has not been much used. It has however served as a basis for other CHILL code generators, where it has proved to be easily adaptable to changes both in source language and operating system environment, and easily portable even to computers sensibly different from the Z8000.

REFERENCES

[1]) A. van Wijngaarden, B. J. Mailloux, J. E. L. Peck and C. H. A. Koster, Report on the algorithmic language ALGOL 68, Num. Math. 14, Springer-Verlag, Berlin, 1969.
[2]) P. Branquart, J.-P. Cardinael, J. Lewi, J.-P. Delescaille and M. Vanbegin, An optimized translation process and its application to ALGOL68, Lecture Notes in Computer Science 38, Springer-Verlag, Berlin, 1976.
[3]) CCITT Recommendation Z200, CHILL language definition, 1983.
[4]) P. Branquart, G. Louis and P. Wodon, An analytical description of CHILL, the CCITT high level language, Lecture Notes in Computer Science 128, Springer-Verlag, Berlin, 1983.
[5]) P. Branquart, Code generation for CHILL, PRLB manuscript M82, June 1984.
[6]) A. Aho and J. Ullman, Principles of compiler design, Addision-Wesley, Reading, 1977.
[7]) W. Waite and G. Goos, Compiler construction, Springer-Verlag, Berlin, 1983.

IPR2018-1557
HTC EX1023, Page 187

Philips J. Res. **39**, *178-194, 1984* *R 1090*

BLOCK DECOMPOSITION AND ITERATION IN STOCHASTIC MATRICES

by P. J. COURTOIS and P. SEMAL

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

The theory of regular splittings for singular M–matrices is used to derive the necessary and sufficient conditions for the convergence of iterative decomposition and aggregation techniques in the computation of the Perron-Frobenius eigenvector of a stochastic matrix. These conditions appear to be somewhat less restrictive than those which must be satisfied by the block Jacobi or the block Gauss-Seidel method. A method is also given to compute lower and upper bounds for the subvector corresponding to a block when that block only is accessible or is of interest. Results previously obtained by the authors prove that these bounds are the best ones which can be determined in this case.

CCCS 41.10, 41.40, 11.40, 11.40c

1. Introduction

The problem which is addressed here is that of evaluating the stationary probability distribution vector $\boldsymbol{\nu}$ of a finite homogeneous Markov chain with transition probability matrix $\mathbf{Q}$. This vector $\boldsymbol{\nu}$ is the left positive eigenvector of $\mathbf{Q}$ which, by the Perron-Frobenius theory, is unique if the stochastic matrix $\mathbf{Q}$ is irreducible, that is if the chain is ergodic.

In many applications where this stationary distribution plays a significant role, and in particular in models of computer system behaviour, the number of states in the Markov chain can be quite large. It is typical for example in the study of queueing models for the performance or the reliability analysis of computer systems, of data communication networks, or of telephone exchange systems, to find models with 10.000 states or more*).

Fortunately, such large matrices often present features which can be advantageously exploited for the computation of their eigensystem. First, they

*) The large state space of these models is basically due to the fact that the number of states of a queueing network grows combinatorially as fast as $\binom{L+N-1}{L-1}$ with the number L of queues and the number N of customers in the network. See ref. 1 for an example of a two node computer network model with 4×10^4 states which grows to 10^9 states for three nodes only.

IPR2018-1557
HTC EX1023, Page 188

are in general very sparse, typically a dozen or so of non-null elements per row[1]). This sparsity makes the use of iterative methods computationally attractive compared to direct methods.

Besides, the states of the Markov chain can be usually rearranged and assembled into aggregates in such a way that the matrix of transition probabilities has a block structure with certain regularities which reflect the structural properties of the system being modeled. In queueing network model matrices, for example, many diagonal blocks are, except for the main diagonal elements, identical to each other. In the case of systems which consist of weakly coupled subsystems, the matrix will also enjoy the property of *near-complete decomposability*[2-5]). The elements of the off-diagonal blocks are then small compared to those of the diagonal blocks.

During the last decade, different techniques[6-8]) which combine block decomposition and iteration have been proposed to take advantage of such properties in the evaluation of the stationary eigenvector ν. These techniques are closely related to methods of aggregation as they were first proposed by Simon and Ando[3]) to analyze input-output econometric models. The basic principle is to calculate for each aggregate, represented by a diagonal block, an approximation of the corresponding subvector of ν. These approximated subvectors are then used to aggregate the states belonging to a same block into one single state and to obtain an estimation of the matrix of transition probabilities between these aggregated states. The stationary probability vector of this aggregated transition matrix yields, with the approximated subvectors, an approximation of the whole stationary vector ν. Since the stationary vector of the aggregated transition matrix provides new scalings for the subvectors, the process may be iterated until a sufficient accuracy has been reached.

In this paper we start by showing that these decomposition techniques are essentially particular implementations of the block Gauss-Seidel or Jacobi iterative methods. We then show that the regular splitting theorem for the singular consistent case of M–matrices[9]), provides a unique framework to derive the necessary and sufficient conditions for the convergence of these techniques. This derivation is neater and more general than previous developments[10,11]). Of particular interest is the fact that the conditions obtained are somewhat less restrictive for these decomposition techniques than for the Jacobi or for the Gauss-Seidel methods.

In the second part, we give a simple method to compute lower and upper bounds for the subvector corresponding to a submatrix when that submatrix only is accessible or is of interest. These bounds are especially useful in the case of very large matrices. They are derived from results which have been recently obtained in refs 12 and 13, and which have a direct probabilistic inter-

IPR2018-1557
HTC EX1023, Page 189

pretation in terms of state visit rates. This interpretation is useful for the determination of low and upper bounds for the steady-state probabilities when a subset of the states of a Markovian model is studied in isolation. An example is given. The bounds have also potential useful applications for the analysis of queueing network models by decomposition and aggregation.

It is remarkable that, by opposition to previous work [7,10,11]), none of our developments requires that the matrix $\mathbf{Q}$ has the property of near-complete decomposability. In particular, the convergence conditions turn out to be independent of the degree of coupling between the blocks. However, a small coupling improves the rate of convergence of the iterative decomposition techniques as well as the accuracy of the subvector bounds.

2. State aggregation

Let us consider a Markov process, the states of which have been partitioned into two sets, denoted $\mathbf{S}_1$ and $\mathbf{S}_2$, which we will subsequently call aggregates. Let also the matrix

$$\mathbf{Q} = \begin{bmatrix} [\mathbf{Q}_{11}]_{n\times n} & [\mathbf{Q}_{12}]_{n\times m} \\ [\mathbf{Q}_{21}]_{m\times n} & [\mathbf{Q}_{22}]_{m\times m} \end{bmatrix} \tag{1}$$

be the associated matrix of transition probabilities, decomposed according to the same partition.

The matrix $\mathbf{Q}_{11}$ is the transition matrix for the states of aggregate $\mathbf{S}_1$, and matrix $\mathbf{Q}_{22}$ for the states of aggregate $\mathbf{S}_2$. We will consider two aggregates only for simplicity reasons. Our results can be easily generalized for an arbitrary number of aggregates.

There are several reasons why a decomposition like (1) can be useful. For example, the interesting states from the user's point of view can be restricted to those of one aggregate only. Or the entire matrix can be so large that one aggregate at a time only can be manipulated. Another reason may be that the matrix is nearly completely decomposable [3,14]) into several aggregates. In this latter case, the elements of the off-diagonal blocks ($\mathbf{Q}_{12}$, $\mathbf{Q}_{21}$) are small compared to those of the diagonal blocks ($\mathbf{Q}_{11}$, $\mathbf{Q}_{22}$), and it is interesting, both from a modelling and from a numerical point of view, to analyze separately the short-term transient of the system determined by the dominant diagonal blocks, and the long-term transient which results from the weak couplings between these blocks.

One of the most important characteristic of a Markov process is its equilibrium probability vector, denoted ν, which satisfies

$$\nu = \nu \mathbf{Q}. \tag{2}$$

IPR2018-1557
HTC EX1023, Page 190

If $\mathbf{Q}$ is irreducible, this eigenvector ν is positive and unique up to a normalization constant. It is generally normalized so that its components sum up to unity

$$\nu \mathbf{1} = 1. \tag{3}$$

Each component will then give the steady-state probability of being in the corresponding state.

This eigenvector ν can be partitioned in the same way as matrix $\mathbf{Q}$

$$\nu \triangleq (\alpha_1 \nu_1 \alpha_2 \nu_2), \tag{4}$$

where the positive scalars α_1 and α_2 give the steady-state probability of being in each of the aggregates and the positive normalized subvectors ν_1 and ν_2 give the relative equilibrium probabilities within a same aggregate. Thus, these variables satisfy

$$\begin{cases} \nu_1 \in \mathbf{R}_n^+, \nu_1 \mathbf{1} = 1, \\ \nu_2 \in \mathbf{R}_m^+, \nu_2 \mathbf{1} = 1, \\ \alpha_1, \alpha_2 \in \mathbf{R}^+, \alpha_1 + \alpha_2 = 1. \end{cases} \tag{5}$$

Two different objectives can now be addressed. Either we are interested in the whole vector ν and, for some of the reasons already mentioned, an analysis by block would be advantageous. Or we are interested in one particular vector only, say ν_1, and the corresponding submatrix only is available.

3. Block iterative methods

With the first objective in mind, let us decompose eq. (2)

$$(\alpha_1 \nu_1 \alpha_2 \nu_2) = (\alpha_1 \nu_1 \alpha_2 \nu_2)\, \mathbf{Q}, \tag{6}$$

which can be rewritten as

$$\begin{cases} \alpha_1 \nu_1 = \alpha_2 \nu_2 \mathbf{Q}_{21} (\mathbf{I} - \mathbf{Q}_{11})^{-1}, \\ \alpha_2 \nu_2 = \alpha_1 \nu_1 \mathbf{Q}_{12} (\mathbf{I} - \mathbf{Q}_{22})^{-1}, \end{cases} \tag{7}$$

or, by substituting $\alpha_2 \nu_2$ in the first equation, we obtain for ν_1 the classical Gauss-Aitken-Bodewig formula [15])

$$\nu_1 = \nu_1 [\mathbf{Q}_{11} + \mathbf{Q}_{12} (\mathbf{I} - \mathbf{Q}_{22})^{-1} \mathbf{Q}_{21}], \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix.

Let us now recall the convergence conditions of the most classical iterative algorithms. System (7) immediately suggests the following iteration scheme

$$\begin{cases} \alpha_1^{(m+1)} \nu_1^{(m+1)} = \alpha_2^{(m)} \nu_2^{(m)} \mathbf{Q}_{21} (\mathbf{I} - \mathbf{Q}_{11})^{-1}, \\ \alpha_2^{(m+1)} \nu_2^{(m+1)} = \alpha_1^{(m)} \nu_1^{(m)} \mathbf{Q}_{12} (\mathbf{I} - \mathbf{Q}_{22})^{-1}, \end{cases} \tag{9}$$

IPR2018-1557
HTC EX1023, Page 191

which can be recognized as the block Jacobi algorithm, or

$$\begin{cases} \alpha_1^{(m+1)}\, \nu_1^{(m+1)} = \alpha_2^{(m)}\, \nu_2^{(m)}\, \mathbf{Q}_{21}(\mathbf{I} - \mathbf{Q}_{11})^{-1}, \\ \alpha_2^{(m+1)}\, \nu_2^{(m+1)} = \alpha_1^{(m+1)}\, \nu_1^{(m+1)}\, \mathbf{Q}_{12}(\mathbf{I} - \mathbf{Q}_{22})^{-1}, \end{cases} \tag{10}$$

which is the block Gauss-Seidel algorithm. System (10) differs from (9) by the fact that the iterated approximations $(\alpha_1^{(m+1)}\, \nu_1^{(m+1)})$ are introduced as soon as they are available, here, in the computation of $\alpha_2^{(m+1)}\, \nu_2^{(m+1)}$. Both algorithms can be described in a matrix form. Let us decompose the matrix $(\mathbf{I} - \mathbf{Q})$ as

$$(\mathbf{I} - \mathbf{Q}) \triangleq \mathbf{L} + \mathbf{D} + \mathbf{U},$$

where matrix $\mathbf{D}$ is block diagonal and, matrices $\mathbf{L}$ and $\mathbf{U}$ are strictly lower and strictly upper block triangular respectively. Systems (9) and (10) can now be rewritten as

$$\nu^{(m+1)} = \nu^{(m)}\, (-\mathbf{L} - \mathbf{U})\, \mathbf{D}^{-1} \triangleq \nu^{(m)}\, \mathbf{J}, \tag{9'}$$

$$\nu^{(m+1)} = \nu^{(m)}\, (-\mathbf{L})\, (\mathbf{D} + \mathbf{U})^{-1} \triangleq \nu^{(m)}\, \mathbf{H}. \tag{10'}$$

Iterative schemes of this type have been generalized by Plemmons and Neumann [9]) who introduced the concept of *regular splitting*. A splitting of the matrix $(\mathbf{I} - \mathbf{Q})$ is regular if it satisfies the relations

$$(\mathbf{I} - \mathbf{Q}) = \mathbf{M} - \mathbf{N}, \quad \mathbf{M}^{-1} \geqslant \mathbf{0}, \quad \mathbf{N} \geqslant \mathbf{0}. \tag{11}$$

To every regular splitting of $(\mathbf{I} - \mathbf{Q})$ corresponds a non-negative iteration matrix

$$\mathbf{A} = \mathbf{N}\, \mathbf{M}^{-1}$$

and an iterative scheme

$$\nu^{(m+1)} = \nu^{(m)}\, \mathbf{A} \tag{12}$$

Equations (9') and (10') give the splittings which correspond to the block Jacobi and block Gauss-Seidel algorithms. Since $\mathbf{Q}$ is irreducible, $\mathbf{D}$ and $\mathbf{D} + \mathbf{U}$ are non-singular M-matrices and the splittings are regular. If $\mathbf{M}$ is chosen equal to the unity matrix, $\mathbf{A}$ reduces to $\mathbf{Q}$ and the iterative scheme (12) is equivalent to the power method [16]).

The conditions for the convergence of an iterative scheme (12) are the following (see ref. 9, or ref. 17 p. 197):

$$\begin{cases} (c_1) \quad \varrho(\mathbf{A}) \leqslant 1, \\ (c_2) \quad \text{if } \varrho(\mathbf{A}) = 1 \text{ then index } (\mathbf{I} - \mathbf{A}) \leqslant 1, \\ (c_3) \quad \text{if } \varrho(\mathbf{A}) = 1 \text{ and if } \lambda \in \sigma(\mathbf{A}) \text{ with } |\lambda| = 1, \text{ then } \lambda = 1, \end{cases}$$

where $\sigma(\mathbf{A})$ and $\varrho(\mathbf{A})$ denote the spectrum and the spectral radius of the iteration matrix $\mathbf{A}$. Condition c_1 requires that the spectral radius of the iteration

IPR2018-1557
HTC EX1023, Page 192

matrix **A** is not greater than one. Otherwise, if the spectral radius is 1, conditions c_2 and c_3 require that the unit eigenvalue be simple and that no other eigenvalue of module one exist. These latter conditions are satisfied if the iteration matrix **A** is irreducible (c_2) and acyclic (c_3).

Neumann and Plemmons[9]), (see also ref. 17), proved that the iteration matrix derived from any regular splitting of the matrix $(\mathbf{I} - \mathbf{Q})$ satisfies conditions c_1 and c_2 if the matrix **Q** is stochastic and irreducible. However, the acyclicity of **Q** is not sufficient to guarantee the acyclicity of the iteration matrix which is required by the last condition c_3. As a simple example, the stochastic, irreducible and acyclic matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q}_{1N} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{32} & \mathbf{Q}_{33} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q}_{NN} \end{bmatrix},$$

leads, for the block Jacobi and for the block Gauss-Seidel splittings, to an iteration matrix which is cyclic, so that the iterations (9′) or (10′) would not converge in this case. In general, the partition into cyclic classes of **A** will not be the same as the partition into blocks of **Q**.

But the iteration matrix, when it is cyclic, is cogredient, by the Perron-Frobenius theorem, to the matrix

$$\begin{bmatrix} \mathbf{0} & \mathbf{A}_{12} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{23} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_{(h-1)h} \\ \mathbf{A}_{h1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}.$$

In theorem 1 of the appendix, we prove that the h eigenvectors of a cyclic matrix which are associated with the eigenvalues of unit module have all the form

$$(\lambda^r \xi_1 \lambda^{2r} \xi_2 \dots \lambda^{hr} \xi_h), \quad r = 0, \dots, h-1, \tag{13}$$

where the scalar λ is complex and where the set of the subvectors ξ_i, $i = 1, \dots, h$ corresponding to the cyclic classes of **A** is unique up to a multiplicative constant. The subvectors ν_i, $i = 1, \dots, N$, which correspond to the diagonal blocks $\mathbf{Q}_{ii}$ and to the aggregates $\mathbf{S}_i$ will consequently converge in the iteration scheme

$$\nu^{(m+1)} = \nu^{(m)} \mathbf{A},$$

even when **A** is cyclic, if and only if all states of a same aggregate $\mathbf{S}_i$ of **Q**

IPR2018-1557
HTC EX1023, Page 193

belong to a same cyclic class of the corresponding iteration matrix **A**. In theorem II of the appendix, we prove that this property will be verified if the aggregates are not completely decomposable, i.e. if the submatrices $\mathbf{Q}_{ii}$ are not block diagonal.

With this simple construction rule for the blocks $\mathbf{Q}_{ii}$, and regardless of whether the iteration matrix is cyclic or not, the iteration scheme (12) can therefore be used to obtain the subvectors v_i. When **A** is cyclic, however, once the subvectors v_i have converged, a separate calculation is needed to obtain the exact scalings.

The algorithms due to Koury et al.[8]), Takahashi[6]) and Vantilborgh[7]), perform this separate calculation of the scaling factors at each iteration, the iterative matrix being cyclic or not. These algorithms use the schemes (9) or (10) to obtain the iterated subvectors v_i and, after each iteration, they perform a rescaling of the factors α_i as follows. The post-multiplication of system (7) by **1**, a column vector of all ones, yields, after some simplifications

$$\begin{cases} \alpha_1 = \alpha_1(v_1 \mathbf{Q}_{11} \mathbf{1}) + \alpha_2(v_2 \mathbf{Q}_{21} \mathbf{1}), \\ \alpha_2 = \alpha_1(v_1 \mathbf{Q}_{12} \mathbf{1}) + \alpha_2(v_2 \mathbf{Q}_{22} \mathbf{1}). \end{cases} \tag{14}$$

The introduction into this system of equations of the iterated approximations $v_i^{(m+1)}$ yields the iterated factors $\alpha_i^{(m+1)}$ univocally since the matrix **Q** is irreducible and acyclic.

This separate computation of the scaling factors presents two types of advantages. First, it has been shown numerically[8,11]) that it could speed up the convergence of the block Jacobi or the block Gauss-Seidel method when the matrix **Q** is nearly completely decomposable. But the reasons and the exact conditions for this improvement to take place are not yet clearly understood[11]).

Secondly, and this results clearly from the arguments hereabove, this separate computation becomes necessary for the convergence of the block Jacobi or Gauss-Seidel methods, when the iteration matrix (**J** or **H**) is cyclic. This property leads to the conjecture that such a separate computation performed regularly after a few block Jacobi or Gauss-Seidel iterations will help also when the matrix, although not cyclic, has nevertheless a similar structure which causes the iteration scheme (12) to converge predominantly on subvectors.

The parallel algorithm due to Mitra[16]) can be seen as an asynchronous implementation of the power method. In this algorithm, the i^{th} parallel processor computes independently the iterated approximation $v_i^{(m+1)}$ for block i, by using for each block j ($j \neq i$), the most recent iterated approximation $v_j^{(k)}$ currently available. As we have seen, the power method corresponds to a splitting

IPR2018-1557
HTC EX1023, Page 194

where $\mathbf{M} = \mathbf{I}$, and $\mathbf{A} = \mathbf{N} = \mathbf{Q}$. The irreducibility and the acyclicity of $\mathbf{Q}$ are therefore necessary and sufficient to guarantee the convergence of the entire vector $\boldsymbol{\nu}$, in this case.

Note that all the above arguments remain valid whether the matrix $\mathbf{Q}$ is nearly completely decomposable or not. They are remarkably independent of ε, the degree of coupling between the aggregates. Contrary to previous work[7,10,11]), these arguments show that necessary and sufficient conditions for the convergence do not require ε being small, a small degree of coupling ε increasing simply the convergence rates if the matrix $\mathbf{Q}$ is nearly completely decomposable.

4. Polyhedron of Perron-Frobenius eigenvectors

Suppose now that the irreducible block $\mathbf{Q}_{11}$ only is known or accessible, and that, on the basis of this information only, we are interested in an approximation to $\boldsymbol{\nu}_1$. Any iterative algorithm based on the recurrence (9) or (10) is helpless in this case since it requires the knowledge of the matrices $\mathbf{Q}_{12}$, $\mathbf{Q}_{21}$ and $\mathbf{Q}_{22}$.

But, if we start again from system (7), the following development, made here for $\boldsymbol{\nu}_1$, leads to an interesting approximation of this vector
Let us define the diagonal normalization matrix $\boldsymbol{\Sigma}$ by

$$diag(\boldsymbol{\Sigma}) \triangleq (\mathbf{I} - \mathbf{Q}_{11})^{-1}\,\mathbf{1}; \tag{15}$$

and rewrite $\boldsymbol{\nu}_1$ as

$$\boldsymbol{\nu}_1 = \frac{\alpha_2}{\alpha_1}\,\boldsymbol{\nu}_2\,\mathbf{Q}_{21}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{I} - \mathbf{Q}_{11})^{-1};$$

or as

$$\boldsymbol{\nu}_1 = \boldsymbol{\beta}\,\mathbf{Z}, \tag{16}$$

with

$$\boldsymbol{\beta} = \frac{\alpha_2}{\alpha_1}\,\boldsymbol{\nu}_2\,\mathbf{Q}_{21}\,\boldsymbol{\Sigma}, \tag{17}$$

and

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-1}\,(\mathbf{I} - \mathbf{Q}_{11})^{-1}. \tag{18}$$

As $(\mathbf{I} - \mathbf{Q}_{11})$ is an M–matrix, it results that $\boldsymbol{\Sigma} \geqslant \mathbf{0}$ and thus that $\boldsymbol{\beta} \geqslant \mathbf{0}$. It results also from (7) and from the definition of $\boldsymbol{\Sigma}$ that

$$\boldsymbol{\beta}\,\mathbf{1} = \frac{\alpha_1\,\boldsymbol{\nu}_1\,\mathbf{1}}{\alpha_1} = 1. \tag{19}$$

Eqs (16) to (19) simply prove that the vector $\boldsymbol{\nu}_1$ can be expressed as a convex combination of the normalized rows of the matrix $(\mathbf{I} - \mathbf{Q}_{11})^{-1}$, or, in other words, that the vector $\boldsymbol{\nu}_1$ belongs to the polyhedron, the vertices of which are

IPR2018-1557
HTC EX1023, Page 195

given by the normalized rows of the matrix $(\mathbf{I} - \mathbf{Q}_{11})^{-1}$. A more general and stronger result is established in refs 12 and 13, where it is proven that this polyhedron is the smallest convex set which is guaranteed to contain the vector ν_1. Consequently, the polyhedron vertices are the best bounds one can obtain when matrix $\mathbf{Q}_{11}$ only is available:

$$\min_i \mathbf{Z}_{ij} \leqslant (\nu_1)_j \leqslant \max_i \mathbf{Z}_{ij} . \tag{20}$$

This result gives a particular meaning to the iterative algorithms discussed in the previous section: the iteration schemes (9) and (10) aim at finding the most accurate convex combination of the normalized rows of $(\mathbf{I} - \mathbf{Q}_{11})^{-1}$.

It has also an interesting probabilistic interpretation. Consider the absorbing Markov chain associated with matrix $\mathbf{Q}_{11}$

$$\begin{bmatrix} [\mathbf{Q}_{11}]_{n\times n} & [\mathbf{Q}_{12}\mathbf{1}]_{n\times 1} \\ [\mathbf{0}]_{1\times n} & [\mathbf{1}]_{1\times 1} \end{bmatrix}; \tag{21}$$

and for any pair (i,j) of states of $\mathbf{S}_1$, let $\mathbf{M}_{ij}$ denote the average number of times that the transient process defined by $\mathbf{Q}_{11}$, started in state i, is in state j, $(i,j \in \mathbf{S}_1)$, before being absorbed. By definition, these quantities satisfy the following recurrence equations (see e.g. ref. 18)

$$\mathbf{M}_{ij} = \mathbf{I}_{ij} + \sum_{k\in \mathbf{S}_1} [\mathbf{Q}_{11}]_{ik} \mathbf{M}_{kj}, \tag{22}$$

which can be rewritten in matrix form as

$$\mathbf{M} = \mathbf{I} + \mathbf{Q}_{11}\mathbf{M} = (\mathbf{I} - \mathbf{Q}_{11})^{-1}. \tag{23}$$

Let us now call

$$m_{ij} \triangleq \frac{\mathbf{M}_{ij}}{\sum_{k\in \mathbf{S}_1} \mathbf{M}_{ik}} \triangleq \frac{\mathbf{M}_{ij}}{t_i} \tag{24}$$

the rate of visit to state j when the process defined by $\mathbf{Q}_{11}$ is started in state i. The sum t_i is the average total number of transitions before absorption when $\mathbf{Q}_{11}$ is started in state i.

As a result of eqs (18) and (23), we have the following equalities

$$m_{ij} = \mathbf{Z}_{ij} \quad i,j = 1, \ldots, n. \tag{25}$$

And, eq. (17) can now be rewritten as:

$$\exists\, \boldsymbol{\beta} \in \mathbf{R}_n^+,\ \boldsymbol{\beta}\,\mathbf{1} = 1, \text{ such that: } \nu_1 = \sum_{i=1}^{n} \beta_i m_i , \tag{26}$$

where:

$$m_i \triangleq (m_{i1}, m_{i2}, \ldots, m_{in}); \tag{27}$$

IPR2018-1557
HTC EX1023, Page 196

is the vector of visit rates to the different states of $\mathbf{S}_1$, when the process is started in state i.

This result has several interesting consequences. Assume that in the decomposition (1), the matrix $\mathbf{Q}_{21}$ has only one non-null column, say the i^{th} one. Then, the only state by which subset $\mathbf{S}_1$ can be entered from outside is state i. In this case, the vector $\boldsymbol{m}_i$ is precisely equal to the vector $\boldsymbol{v}_1$. More generally: if matrix $\mathbf{Q}_{21}$ has two non-zero columns, say columns i and j, then the solution $\boldsymbol{v}_1$ will be a convex combination of $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$; and if all the columns of matrix $\mathbf{Q}_{21}$ are non-null then $\boldsymbol{v}_1$ will be a convex combination of all the $\boldsymbol{m}_i, i = 1, \ldots, n$. An algebraic proof of this result can be easily obtained from eq. (17).

We have also proven in ref. 13 that the visit rate to a state j is maximum when the process is started from that state j. Eq. (20) can therefore be rewritten as

$$\min_i (m_{ij}) \leqslant (\boldsymbol{v}_1)_j \leqslant \max_i (m_{ij}) = m_{jj} . \tag{28}$$

Note that the state i from which the relative visit rate to a state j is minimum cannot receive a more precise general characterization; in each case, it will depend on the relative values of the visit numbers $\mathbf{M}_{ij}$ and on the absorption times t_i.

The inequalities (20) and (28) are new results. These bounds are useful to approximate the steady-state probabilities for a subset of the states of a Markov chain when that subset alone is of interest or when that part only of the Markov transition process is accessible.

This is specially true when that embedded subset of the states corresponds to a particular submodel such as a random walk or a queue. The matrix $\mathbf{Z}$ and the visit rates corresponding to such subsets can often be efficiently computed. They may even sometimes receive an analytical form. As an example to illustrate this last point, consider the Markovian matrix

$$\begin{array}{cccccccc|c}
0 & p & 0 & 0 & \ldots & 0 & 0 & 0 & \\
q & 0 & p & 0 & \ldots & 0 & 0 & 0 & \\
0 & q & 0 & p & \ldots & 0 & 0 & 0 & \\
\cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \mathbf{Q}_{12} \\
0 & 0 & 0 & 0 & \ldots & 0 & p & 0 & \\
0 & 0 & 0 & 0 & \ldots & q & 0 & p & \\
0 & 0 & 0 & 0 & \ldots & 0 & q & 0 & \\
\hline
\multicolumn{7}{c|}{\mathbf{Q}_{21}} & \mathbf{Q}_{22}
\end{array}$$

where $p + q = 1$, and where

IPR2018-1557
HTC EX1023, Page 197

$$\mathbf{Q}_{12}\mathbf{1} = \begin{bmatrix} q \\ 0 \\ \vdots \\ 0 \\ p \end{bmatrix}.$$

The $(n-1)\times(n-1)$ submatrix $\mathbf{Q}_{11}$, when considered in isolation, behaves as a random walk with two absorbing states, 1 and $n-1$. In this case, we have (see e.g. ref. 18, p. 151) for $i,j = 1,\ldots,n-1$

$$\mathbf{M}_{ij} = \frac{1}{(p-q)(r^n-1)} \times \begin{cases} (r^j-1)\times(r^{n-i}-1), & \text{if } j \leqslant i \\ (r^i-1)\times(r^{n-i}-r^{j-i}), & \text{if } j \geqslant i, \end{cases}$$

where $r = (p/q) \neq 1$. For $p = q = \frac{1}{2}$, (i.e. $r = 1$), this expression simplifies to

$$\mathbf{M}_{ij} = \frac{2}{n} \times \begin{cases} j(n-i) & \text{if } j \leqslant i, \\ i(n-j) & \text{if } j \geqslant i, \end{cases}$$

while

$$t_i = \sum_{j=1}^{n-1} \mathbf{M}_{ij} = i(n-i).$$

Consequently, the bounds

$$\min_i (m_{ij}) \leqslant (\nu_1)_j \leqslant m_{jj},$$

become

$$\frac{2}{n} \times \min_i \left(\frac{n-j}{n-i}, \frac{j}{i}\right) \leqslant (\nu_1)_j \leqslant \frac{2}{n},$$

or

$$\frac{2}{n(n-1)} \times \min(n-j, j) \leqslant (\nu_1)_j \leqslant \frac{2}{n},$$

which, according to the theory developed in refs 12 and 13, are the tightest bounds one can obtain when $\mathbf{Q}_{11}$ is only known.

In a similar way, we have recently obtained bounds for a subset of states which corresponds to a single server queue. Such bounds are useful for the analysis by decomposition and aggregation of large Markovian models of queueing networks.

5. Acknowledgement

We are grateful to our colleague Paul Vandooren for his many useful advices during the preparation of this paper.

IPR2018-1557
HTC EX1023, Page 198

6. Appendix

Let **Q** be a stochastic, irreducible and acyclic matrix, and let **A** be its corresponding block Jacobi or block Gauss-Seidel iteration matrix. We are interested in the behaviour of the iteration process

$$\xi^{(m+1)} = \xi^{(m)} \mathbf{A}, \tag{A.1}$$

when the matrix **A**, which is also non-negative, irreducible and has a unit spectral radius, is cyclic of index h.

Theorem 1

The h left eigenvectors of **A** which are associated with the h eigenvalues of module one all have the form

$$(\lambda^r \xi_1^{2r} \xi_2 \ldots \lambda^{hr} \xi_h), \quad r = 0, \ldots, h-1, \tag{A.2}$$

where the scalar λ is complex and the set of subvectors $\xi_1, \ldots, \xi_h$ is unique up to a multiplicative constant.

Proof

By the Perron-Frobenius theorem, if the non-negative, irreducible matrix **A** is cyclic of index h, then

— the h eigenvalues of module one are: λ^r, $r = 0, 1, \ldots, h-1$ with

$$\lambda = e^{i\frac{2\pi}{h}}, \tag{A.3}$$

— and, there exists a permutation matrix **P** such that

$$\mathbf{P}\,\mathbf{A}\,\mathbf{P}^T = \begin{bmatrix} \mathbf{0} & \mathbf{A}_{12} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{23} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_{(h-1)h} \\ \mathbf{A}_{h1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \end{bmatrix}, \tag{A.4}$$

where the zero blocks along the diagonal are square.

Without loss of generality, we may assume that the states of **Q** have been reordered so that $\mathbf{P} = \mathbf{I}$. Let us now consider the left eigenvector of **A**, associated to the unit eigenvalue ($r = 0$). If we apply to this eigenvector the same partitioning as that of the cyclic classes, we can write

$$\xi \triangleq (\xi_1 \xi_2 \ldots \xi_h) = (\xi_1 \xi_2 \ldots \xi_h)\,\mathbf{A}. \tag{A.5}$$

This last equation can be rewritten as

IPR2018-1557
HTC EX1023, Page 199

$$\xi_i = \xi_i \mathbf{A}_{i(i+1)} \ldots \mathbf{A}_{(i-1)i} = \xi_i \prod_{k=i}^{i-1} \mathbf{A}_{k\bar{k}^+}, \tag{A.6}$$

where the index $\bar{k}^+$ is defined by

$$\bar{k}^+ \triangleq (k \bmod h) + 1.$$

We can prove now, by verification, that the vector

$$\boldsymbol{\beta}^{(r)} = (\beta_1^{(r)} \ldots \beta_h^{(r)}) \triangleq (\lambda^r \xi_1 \lambda^{2r} \xi_2 \ldots \lambda^{hr} \xi_h), \quad r = 1, \ldots, h-1, \tag{A.7}$$

is the left eigenvector of $\mathbf{A}$ corresponding to the r^{th} eigenvalue λ^r

$$\boldsymbol{\beta}^{(r)} \mathbf{A} = \lambda^r \boldsymbol{\beta}^{(r)}. \tag{A.8}$$

Developing the matrix $\mathbf{A}$ in this last equation gives

$$\beta_i^{(r)} \mathbf{A}_{i\bar{i}^+} = \lambda^r \beta_{\bar{i}^+}^{(r)}.$$

If we solve this system of equations, we obtain

$$\beta_i^{(r)} \prod_{k=i}^{i-1} \mathbf{A}_{k\bar{k}^+} = \lambda^{hr} \beta_i^{(r)},$$

or, using eqs (A.3) and (A.7)

$$\lambda^{ir} \xi_i \prod_{k=i}^{i-1} \mathbf{A}_{k\bar{k}^+} = \lambda^{ir} \xi_i,$$

which proves, with (A.6), the desired result. □

Therefore, when the iteration matrix $\mathbf{A}$ is cyclic, the iteration scheme (A.1) will ultimately produce vectors $\xi^{(m)}$ which are linear combinations of the h eigenvectors (A.2) only. If the state partition of $\mathbf{A}$ into cyclic classes is known, one can then extract, from the iterated vectors, normalized subvectors which are parallel to the ξ_i.

And if all states of an aggregate $\mathbf{S}_r$ of $\mathbf{Q}$ belong to a single cyclic class of $\mathbf{A}$, say the i^{th} one, then the vector $\boldsymbol{v}_r$ will also be parallel to the corresponding subvector of ξ_i.

The following theorem gives a sufficient condition for this last property to hold. This condition is proven for the block Jacobi and the block Gauss-Seidel iteration matrices.

Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N$, be a partition of the set of states of the Markov chain such that the matrix of transition probabilities $\mathbf{Q}_{rr}$ of any aggregate $\mathbf{S}_r$ cannot, by a reordering of the states, have a block diagonal form. Such an aggregate is said *connected* because any pair of states (i_a, i_b) of that aggregate is connected by an undirect path, i.e.

IPR2018-1557
HTC EX1023, Page 200

$$\forall\, i_a, i_b \in \mathbf{S}_r: \begin{cases} \mathbf{Q}_{i_a i_b} > 0, \\ \text{or} \\ \mathbf{Q}_{i_b i_a} > 0, \\ \text{or} \\ \exists\, i_0, i_1, \ldots, i_n \in \mathbf{S}_r: \begin{cases} \mathbf{Q}_{i_a i_0} > 0 \text{ or } \mathbf{Q}_{i_0 i_a} > 0, \text{ and} \\ \mathbf{Q}_{i_k i_{k+1}} > 0 \text{ or } \mathbf{Q}_{i_{k+1} i_k} > 0, \text{ and} \\ \mathbf{Q}_{i_n i_b} > 0 \text{ or } \mathbf{Q}_{i_b i_n} > 0. \end{cases} \end{cases}$$

Theorem 2

When the block Jacobi or the block Gauss-Seidel iteration matrix **A** is cyclic, the states of a same aggregate of **Q** will belong to a same cyclic class of **A**, if this aggregate is connected.

Proof

Because of the definition of a connected aggregate, it is sufficient to prove that any two directly connected states of a same connected aggregate,

$$\forall\, i_0, i_a \in \mathbf{S}_r: \mathbf{Q}_{i_0 i_a} > 0, \tag{B.1}$$

belong to the same cyclic class. The proof is given for the block Jacobi iteration matrix; its accommodation to the block Gauss-Seidel iteration matrix is explained hereinafter.

Let us remember the triangular block decomposition of the matrix $(\mathbf{I} - \mathbf{Q})$,

$$(\mathbf{I} - \mathbf{Q}) \triangleq \mathbf{L} + \mathbf{D} + \mathbf{U},$$

and the associated block Jacobi iteration matrix,

$$\mathbf{A} = (-\mathbf{L} - \mathbf{U})\,\mathbf{D}^{-1}.$$

The following lemma is a direct consequence of the fact that $\mathbf{D}^{-1}$ can be expressed as an infinite sum of matrix powers of the diagonal blocks:

Lemma 1

$$\begin{gathered} \forall\, r, \forall\, i, j \in \mathbf{S}_r: \mathbf{D}_{ij}^{-1} > 0 \\ \text{iff} \\ \exists\, n \geqslant 0, [\mathbf{Q}_{rr}{}^n]_{ij} > 0. \end{gathered} \tag{B.2}$$

This lemma simply states that the (i,j) element of inverse matrix $\mathbf{D}^{-1}$, $i, j \in \mathbf{S}_r$, is non-null, if state j is reachable from state i, without leaving the aggregate $\mathbf{S}_r$. Because of our assumption (B.1), we have thus

$$\begin{cases} \mathbf{D}_{i_0 i_0}^{-1} > 0, \\ \mathbf{D}_{i_0 i_a}^{-1} > 0. \end{cases} \tag{B.3}$$

IPR2018-1557
HTC EX1023, Page 201

If the state i_0 can be reached from a state i_1 which does not belong to the aggregate,

$$\exists\, i_1 \notin \mathbf{S}_r\colon \mathbf{Q}_{i_1 i_0} > 0, \tag{B.4}$$

then, we have

$$\begin{cases} \mathbf{A}_{i_1 i_0} \geqslant (-\mathbf{L} - \mathbf{U})_{i_1 i_0}\, \mathbf{D}^{-1}_{i_0 i_0} > 0 \\ \mathbf{A}_{i_1 i_a} \geqslant (-\mathbf{L} - \mathbf{U})_{i_1 i_0}\, \mathbf{D}^{-1}_{i_0 i_a} > 0 \end{cases} \tag{B.5}$$

which proves that the states i_0 and i_a belong to the same cyclic class.

Otherwise, if such a state cannot be found,

$$\forall\, i_1 \notin \mathbf{S}_r\colon \mathbf{Q}_{i_1 i_0} = 0, \tag{B.6}$$

the irreducibility of the matrix $\mathbf{Q}$ guarantees that the state i_0 must be reachable from a state i_1, which belongs to the aggregate

$$\exists\, i_1 \in \mathbf{S}_r,\ i_1 \neq i_0\colon \mathbf{Q}_{i_1 i_0} > 0. \tag{B.7}$$

One can assume, without loss of generality, that there exists such an index i_1 which is different from i_a. Indeed, if the only index i_1 which satisfies eq. (B.7) is equal to i_a, then, the states i_0 and i_a form a loop which, because of the irreducibility of $\mathbf{Q}$, should be accessible from another third state i_1. We may therefore assume that the labels i_0 and i_a have been chosen so that the loop can be accessed through i_0. We have thus

$$\exists\, i_1 \in \mathbf{S}_r,\ i_1 \neq i_0, i_1 \neq i_a\colon \mathbf{Q}_{i_1 i_0} > 0, \tag{B.8}$$

which, by lemma 1, ensures that

$$\begin{aligned} &\mathbf{D}^{-1}_{i_1 i_1} > 0, \\ &\mathbf{D}^{-1}_{i_1 i_0} > 0, \\ &\mathbf{D}^{-1}_{i_1 i_a} > 0. \end{aligned} \tag{B.9}$$

The reasoning developed for state i_0 in eq. (B.3), can now be applied again to state i_1 in eq. (B.9) and iterated. At the k^{th} iteration, the current state i_{k-1} of the aggregate $\mathbf{S}_r$, is either, reachable from outside the aggregate

$$\exists\, i_k \notin \mathbf{S}_r\colon \mathbf{Q}_{i_k i_{k-1}} > 0, \tag{B.10}$$

which, by the lemma, ensures that we have

$$\begin{cases} \mathbf{A}_{i_k i_{k-1}} \geqslant (-\mathbf{L} - \mathbf{U})_{i_k i_{k-1}}\, \mathbf{D}^{-1}_{i_{k-1} i_{k-1}} > 0 \\ \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ \mathbf{A}_{i_k i_0} \geqslant (-\mathbf{L} - \mathbf{U})_{i_k i_{k-1}}\, \mathbf{D}^{-1}_{i_{k-1} i_0} > 0 \\ \mathbf{A}_{i_k i_a} \geqslant (-\mathbf{L} - \mathbf{U})_{i_k i_{k-1}}\, \mathbf{D}^{-1}_{i_{k-1} i_a} > 0\,, \end{cases} \tag{B.11}$$

IPR2018-1557
HTC EX1023, Page 202

which guarantee that the states

$$i_{k-1}, \ldots, i_0, i_a, \tag{B.12}$$

belong to the same cyclic class. Otherwise, the current state i_{k-1} can be shown to be reachable from a state i_k which belongs to the aggregate $\mathbf{S}_r$ and, which does not belong to the sequence (B.12). Indeed, if the only state from which the state i_{k-1} is accessible, is a state of the sequence (B.12), then we have a loop of states, which again must be accessible from outside. The irreducibility of the matrix $\mathbf{Q}$, which ensures that the aggregate is reachable from outside, and the finite size of the aggregate guarantee that an index k satisfying (B.10) will be found.

A similar proof can be established for the Gauss-Seidel iteration matrix

$$\mathbf{A} = -\mathbf{L}(\mathbf{D} + \mathbf{U})^{-1}.$$

An equivalent lemma states that the (i,j) element of inverse matrix $(\mathbf{D} + \mathbf{U})^{-1}$ is non-null,

$$\forall\, i \in \mathbf{S}_r,\ \forall\, j \in \mathbf{S}_s,\ r \leqslant s\colon (\mathbf{D} + \mathbf{U})^{-1}_{ij} > 0,$$

if one can construct a directed path from state i to state j, which goes only through states belonging to aggregates of non-decreasing labels. The proof of the theorem is also based on the construction of a sequence of states which, no more belong to the same aggregate, but to aggregates of non-increasing labels. The sequence terminates at the k^{th} iteration if the current state i_{k-1} which belongs to aggregate $\mathbf{S}_s$ can be accessed from a state i_k which belong to an aggregate with a larger label

$$\exists\, t > s,\ \exists\, i_k \in \mathbf{S}_t\colon \mathbf{Q}_{i_k i_{k-1}} > 0.$$

This ensures that

$$-\mathbf{L}_{i_k i_{k-1}} > 0.$$

which again, yields

$$\begin{cases} \mathbf{A}_{i_k i_{k-1}} \geqslant (-\mathbf{L})_{i_k i_{k-1}} (\mathbf{D} + \mathbf{U})^{-1}_{i_{k-1} i_{k-1}} > 0 \\ \quad \ldots \quad \ldots \qquad \ldots \qquad \ldots \\ \mathbf{A}_{i_k i_0} \geqslant (-\mathbf{L})_{i_k i_{k-1}} (\mathbf{D} + \mathbf{U})^{-1}_{i_{k-1} i_0} > 0 \\ \mathbf{A}_{i_k i_a} \geqslant (-\mathbf{L})_{i k i_{k-1}} (\mathbf{D} + \mathbf{U})^{-1}_{i_{k-1} i_a} > 0\,. \end{cases}$$

Again, the existence of such an index is guaranteed by the irreducibility of $\mathbf{Q}$ and its finite size. □

IPR2018-1557
HTC EX1023, Page 203

REFERENCES

[1]) E. F. Wunderlich, L. Kaufman and B. Gopinath, Proc. 5th Int. Conference on Computer Communication, Atlanta, 1980.
[2]) P. J. Courtois, Decomposability: Queueing and computer system application, Academic Press, New York, 1977.
[3]) H. A. Simon and A. Ando, Econometrica **29**, 111 (1961).
[4]) P. J. Courtois, Comm. Assoc. Comput. Mach. **18**, 371 (1975).
[5]) P. J. Courtois and H. Vantilborgh, Acta Informatica **6**, 251 (1976).
[6]) Y. Takahashi, Research Report B-18, Dept. of Information Sciences, Tokyo Institute of Technology, June 1975.
[7]) H. Vantilborgh, The error of aggregation, A contribution to the theory of decomposable systems and applications, Thèse de Docteur en Sciences Appliquées, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1983.
[8]) R. Koury, D. F. McAllister and W. J. Stewart, SIAM J. Alg. Disc. Meth. **5**, 164 (1984).
[9]) M. Neumann and R. J. Plemmons, Numerische Mathematik **31**, 265 (1978).
[10]) G. W. Stewart, J. Assoc. Comput. Mach. **30**, 271 (1983).
[11]) G. W. Stewart, W. J. Stewart and D. F. McAllister, A two stage iteration for solving nearly uncoupled Markov chains, to be published.
[12]) P. J. Courtois and P. Semal, J. Assoc. Comput. Mach. **31**, no. 4 (1984).
[13]) P. J. Courtois and P. Semal, to appear in Linear Algebra and its Applications.
[14]) P. J. Courtois, Econometrica **43**, 691 (1975).
[15]) E. Bodewig, Matrix Calculus, North-Holland Publishing Company, Amsterdam, 1956; cf. p. 117.
[16]) B. Lubachevski and D. Mitra, A chaotic, asynchronous algorithm for computing the fixed point of a non-negative matrix of unit spectral radius, to be published.
[17]) A. Berman and R. J. Plemmons, Nonnegative matrices in the mathematical sciences, Academic Press, New York (1979), cf. p. 195-202.
[18]) J. G. Kemeny and J. L. Snell, Finite Markov chains, D. Van Nostrand Co., Inc., New York 1960, cf. p. 150-152.

IPR2018-1557
HTC EX1023, Page 204

Philips J. Res. **39**, 195-205, 1984 R 1091

CODES BETWEEN BCH AND RS CODES

by CH. COUVREUR and PH. PIRET

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

The codes obtained as the shortened binary image of linear codes over GF(2^m) are studied in detail. Their dimension is computed and it is shown that they may be used as unequal error protection codes.

Math. Rev.: 94A05

1. Introduction

The study of Reed-Solomon (RS) codes is a major part of algebraic coding theory [1]). From a theoretical point of view these codes have many interesting algebraic properties and from a practical point of view, they have optimum burst error correction capabilities. Moreover, binary BCH codes which are powerful binary codes can be defined as being subfield subcodes of RS codes. In this paper we also define binary codes by means of codes over larger alphabets. Our goal will be to analyse the structure and the error correcting capabilities of a class of binary linear codes that are obtained when one shortens [2]) the binary image of some cyclic codes over GF(2^m) [1,2]). The content of the paper is as follows. The codes are introduced in sec. 2 and their dimension over GF(2) is computed in sec. 3. Some of their error correcting capabilities are described in sec. 4 where it is noted that they are well adapted for unequal error protection (U.E.P.) [3,4]). Some examples are discussed in sec. 5.

2. Definition of the shortened binary image codes

Let N be an odd positive integer and let m be the smallest positive integer such that N is a divisor of $2^m - 1$. The Galois field GF(2^m) is thus the smallest extension field of GF(2) that contains all N^{th} roots of unity. Let β be a primitive N^{th} root of 1 and let U be any subset of $S = \{0, 1, \ldots, N-1\}$. To U we associate the polynomial

$$g(X) = \prod_{u \in U} (X - \beta^u). \tag{1}$$

It is a divisor of $X^N - 1$ and it generates a linear cyclic block code of length N

IPR2018-1557
HTC EX1023, Page 205

and dimension $K = N - \text{card}\, U$ over GF(2^m). This code will be denoted by $C(m)$. Any codeword of $C(m)$ can be represented by a polynomial

$$v(X) = \sum_{i=0}^{N-1} v_i X^i$$

over GF(2^m) or by the corresponding N–tuple $v = (v_0, \ldots, v_i, \ldots, v_{N-1})$. We shall now express any $v_i \in \text{GF}(2^m)$ as a binary m–tuple. Let α be an element of GF(2^m) that is not in any proper subfield of GF(2^m). Any $v_i \in \text{GF}(2^m)$ has a unique expression

$$v_i = \sum_{j=0}^{m-1} v_i^j \alpha^j \qquad v_i^j \in \text{GF}(2),$$

which leads to the representation of the word v of $C(m)$ by an $m \times N$ matrix over GF(2):

$$v = \begin{bmatrix} v_0^0 & \ldots & v_i^0 & \ldots & v_{N-1}^0 \\ \vdots & & \vdots & & \vdots \\ v_0^j & \ldots & v_i^j & \ldots & v_{N-1}^j \\ \vdots & & & & \\ v_0^{m-1} & \ldots & v_i^{m-1} & \ldots & v_{N-1}^{m-1} \end{bmatrix}. \tag{2}$$

Defining the polynomial $v^j(X)$, $(0 \leqslant j \leqslant m-1)$ over GF(2) as being

$$\sum_{i=0}^{N-1} v_i^j X^i,$$

we can also represent v by

$$v(X) = \sum_{j=0}^{m-1} v^j(X)\, \alpha^j.$$

From $C(m)$ we define $C(t)$ $(1 \leqslant t \leqslant m)$ as follows

$$C(t) \triangleq \{v(X) \in C(m) : v^j(X) = 0 \quad \text{for} \quad j \geqslant t\}.$$

Any polynomial $v(X) \in C(t)$ can thus be written as

$$v(X) = \sum_{j=0}^{t-1} v^j(X)\, \alpha^j. \tag{3}$$

It can also be described by a matrix v (as in (2)) but containing only t rows. This matrix v is called the binary image [1]) of $v(X)$ (with respect to α). We may thus consider a code $C(t)$ as being a binary code of length Nt. This will be our point of view but however we shall often represent a binary Nt–tuple by an expression like (3).

IPR2018-1557
HTC EX1023, Page 206

Since any polynomial of $C(m)$ is a multiple of $g(X)$, (1), any $v(X)$ as given in (3) will be in $C(t)$ if and only if it satisfies

$$v(\beta^u) = 0, \qquad \text{all } u \in U. \tag{4}$$

The code $C(t)$ will be called a shortened binary image (SBI or t-SBI) of the code $C(m)$. For $t \leqslant m-1$, $C(t)$ is a linear code over GF(2) but not over any larger field.

3. The dimension of the SBI codes

In this section we determine the dimension of the codes $C(t)$ defined in sec. 2 when considered as being vector spaces over GF(2). The elements α and β of GF(2^m) will have the same meaning as in sec. 2.

Let us associate to any u in $S = \{0, 1, \ldots, N-1\}$ its cyclotomic coset $S(u) = \{u\,2^r \bmod N$: all integers $r \in [0, m-1]\}$, and in any such $S(u)$ let us choose a distinguished element called the coset representative of $S(u)$. The set of all these coset representatives will be denoted by S^*. The polynomial $X^N - 1$ is then factorized as

$$X^N - 1 = \prod_{u \in S^*} h_u(X), \quad \deg h_u(X) = \operatorname{card} S(u),$$

where $h_u(X)$ is the minimal polynomial of β^u over GF(2). To any subset U of S we associate the subset U^* of S^* defined by

$$U^* = \{u \in S^* : S(u) \cap U \text{ is not empty}\}.$$

The set U^* thus contains the representatives of cosets some elements of which are in U. We use it to define the polynomial

$$g^*(X) = \prod_{u \in U^*} h_u(X),$$

which is the smallest degree polynomial over GF(2) that is a multiple of $g(X)$. Thus $g^*(X)$ generates the binary subfield subcode of $C(m)$.

Let us now denote by A the algebra of polynomials in X over GF(2) taken modulo $X^N - 1$. The set of polynomials in A that are multiples of $g^*(X)$ is an ideal denoted by I^*. Similarly any polynomial $(X^N - 1)/h_u(X)$ generates also an ideal (denoted by $I(u)$) in A. These last ideals are minimal and in each of them we may choose an idempotent polynomial $e_u(X)$ as generator. The algebra A can be represented as the *direct* sum

$$A = I^* + \sum_{u \in U^*} I(u),$$

to which corresponds the (unique) expression [1])

$$v^j(X) = a^j(X)\, g^*(X) + \sum_{u \in U^*} b_u^j(X)\, e_u(X) \tag{5}$$

of an arbitrary polynomial $v^j(X)$ of A. In this expression $e_u(\beta^s)$ is 1 for $s \in S(u)$

IPR2018-1557
HTC EX1023, Page 207

and 0 for $s \notin S(u)$, while the formal degree of the coefficient $b_u^j(X)$ is equal to the cardinality of $S(u)$ minus one (card $S(u) - 1$). From (5) and the condition (4) we may conclude that the polynomial $v(X)$ of (3) is in $C(t)$ if and only if for any u in U^* we have

$$\sum_{j=0}^{t-1} b_u^j(\beta^s)\,\alpha^j = 0, \quad \text{all } s \in S(u) \cap U. \tag{6}$$

To compute the dimension of $C(t)$ is to compute how many degrees of freedom (in bits) we have in the choice of the polynomials $a^j(X)$ and $b_u^j(X)$ occurring in the representation (5) of the polynomials $v^j(X)$, $(0 \leqslant j \leqslant t-1)$, and constrained by the eq. (6). Let us first consider the polynomials $a^j(X)$. They are not constrained by (6) and may thus be freely chosen. Their formal degree is $m(I^*) - 1$ with $m(I^*) = N - \deg g^*(X)$. Since there are t such $a^j(X)$, they introduce $t\,m(I^*)$ degrees of freedom (in bits) in the code word $v(X)$. Let us now consider the polynomials $b_u^j(X)$. Since any $s \in S(u)$ can be written as $u2^{f(s)}$, the system (6) becomes

$$\sum_{j=0}^{t-1} b_u^j(\beta^{u2^{f(s)}})\,\alpha^j = 0, \quad \text{all } u \in U^*, \quad \text{all } s \in S(u) \cap U. \tag{7}$$

We denote card $S(u)$ by $m(u)$. These numbers $m(u)$ are divisors of m:

$$m(u) \mid m, \quad 1 \leqslant m(u) \leqslant m$$

and β^u is in GF($2^{m(u)}$) which is a proper subfield of GF(2^m) when $m(u)$ is a proper divisor of m. We solve now (7) for a fixed $u \in U^*$. For each $s \in S(u) \cap U$, (7) specifies one equation to be satisfied by the polynomials $b_u^j(X)$, $(0 \leqslant j \leqslant t-1)$. By squaring this equation $r\,m(u) - f(s)$ times for $r = 1, \ldots, m/m(u)$, we get $m/m(u)$ new equations, and by doing this for all s in $S(u) \cap U$ we get

$$\sum_{j=0}^{t-1} b_u^j(\beta^u)\,\alpha^{j2^{rm(u)-f(s)}} = 0, \quad 1 \leqslant r \leqslant m/m(u), \quad \text{all } s \in S(u) \cap U. \tag{8}$$

Denoting card $[S(u) \cap U]$ by $\lambda(u)$, we see that the t elements $b_u^j(\beta^u)$, $(0 \leqslant j \leqslant t-1)$, of GF($2^{m(u)}$) have to satisfy $m\lambda(u)/m(u)$ equations. We write these equations in matrix form as

$$b_u H_u = 0, \tag{9}$$

where b_u denotes the row vector $[b_u^0(\beta^u), \ldots, b_u^{t-1}(\beta^u)]$ and H_u is a $t \times m\lambda(u)/m(u)$ Vandermonde-like matrix. In other words, its (i,j) entry has the form $(\eta_j)^{i-1}$ where the elements η_j run through the set of the $m\lambda(u)/m(u)$ different elements $\alpha^{2^{rm(u)-f(s)}}$ appearing in (8). As a consequence the rank of H_u is equal to its smallest dimension. If $m\lambda(u)/m(u) \geqslant t$ the only solution of (9) is $b_u = 0$. If $m\lambda(u)/m(u) < t$, the $t - m\lambda(u)/m(u)$ last components of b_u may be chosen

IPR2018-1557
HTC EX1023, Page 208

arbitrarily in $GF(2^{m(u)})$ and the $m\lambda(u)/m(u)$ first ones are then obtainable from (9) since the $m\lambda(u)/m(u)$ first rows of H_u form a square nonsingular matrix. Moreover the elements of b_u to be computed from the freely chosen ones are also in $GF(2^{m(u)})$. Indeed, when $m(u)$ is a proper divisor of m and the system (9) contains the equation $\sum b_u^j(\beta^u)\gamma^j = 0$, then it also contains the $m/m(u) - 1$ equations $\sum b_u^j(\beta^u)\gamma^{j2^{rm(u)}} = 0$, $r = 1, \ldots, m/m(u) - 1$. Such a system is said to be self-conjugate with respect to the subfield $GF(2^{m(u)})$ of $GF(2^m)$. As it is well known by the theory of BCH codes, this self-conjugation of the parity-check equations guarantees that the solutions have all their components in $GF(2^{m(u)})$ [1].

Up to now we have only considered the value of the polynomials $b_u^j(X)$ for $X = \beta^u$. However since the formal degree of $b_u^j(X)$ is equal to $m(u) - 1$, this polynomial is completely determined as soon as its value is known for a value of X that is a proper element of $GF(2^{m(u)})$. Since β^u is such a proper element, the polynomials $b_u^j(X)$ are completely determined as soon as $b_u^j(\beta^u)$ is known. As a consequence, for any $u \in U^*$ there are $m(u)[t - m\lambda(u)/m(u)] = tm(u) - m\lambda(u)$ degrees of freedom (in bits) in the choice of polynomials $b_u^j(X)$. We are now in a position to compute the dimension of the code $C(t)$ viewed as a vector space over GF(2). We state this result as a theorem.

Theorem 3.1.

The dimension $K_2(t)$ of the binary linear code $C(t)$ is given by:

$$K_2(t) = t\left[N - \sum_{u \in U^*} m(u)\right] + \sum_{u \in U^*} [t\,m(u) - m\,\lambda(u)]^+ \qquad (10)$$

with $m(u) = \operatorname{card} S(u)$, $\lambda(u) = \operatorname{card} S(u) \cap U$, where $a^+ = a$ for $a \geqslant 0$ and $a^+ = 0$ for $a \leqslant 0$.

Proof

In the right member of (10), the first term represents the dimension afforded by the polynomials $a^j(X)$ and the other terms represent the dimension afforded by the polynomials $b_u^j(X)$. □

As an example we choose $m = 4$, $N = 15$, $t = 3$ and $U = \{0, 1, 2, 3, 4, 5, 6, 9\}$. From these data we obtain

$$U^* = \{0, 1, 3, 5\},$$
$$m(0) = 1,\ m(5) = 2,\ m(1) = m(3) = 4,$$
$$\lambda(0) = \lambda(5) = 1,\ \lambda(1) = \lambda(3) = 3.$$

To compute $K_2(3)$ we evaluate

$$N - \sum_{u \in U^*} m(u) = 4,$$
$$[t\,m(u) - m\,\lambda(u)]^+ = 0, \quad u = 0, 1, 3,$$
$$[t\,m(5) - m\,\lambda(5)]^+ = 2.$$

IPR2018-1557
HTC EX1023, Page 209

The SBI code corresponding to these parameters has length equal to 45 and dimension equal to $(3 \times 4) + 2 = 14$. Further properties of this code are given in sec. 5.

From the discussion above it is natural to associate a matrix $B(v)$ to any polynomial $v(X) = \sum_{j=0}^{t-1} v^j(X)\,\alpha^j$ where the $v^j(X)$ are polynomials over GF(2) that are expressed as in (5). This matrix $B(v)$ has $M = 1 + \text{card}\, U^*$ rows and t columns. Its first $M-1$ rows are indexed by the elements u of U^* and the last one is indexed by $\bar{U}$, the complementary set of U in S. Its columns are indexed by the integers from 0 up to $t-1$. Writing b_u^j for $b_u^j(\beta^u)$ we construct $B(v)$ as follows. For $u \in U^*$, $B(v)$ contains in position (u,j) the element b_u^j of GF($2^{m(u)}$). In position $(\bar{U},j)$, $B(v)$ contains the polynomials $a^j(X)$ of formal degree $N - \deg g^*(X) - 1$.

$$B(v) = \begin{bmatrix} b_{u_1}^0 & \dots & b_{u_1}^j & \dots & b_{u_1}^{t-1} \\ \vdots & & \vdots & & \vdots \\ b_{u_s}^0 & \dots & b_{u_s}^j & \dots & b_{u_s}^{t-1} \\ \vdots & & \vdots & & \vdots \\ b_{u_{M-1}}^0 & \dots & b_{u_{M-1}}^j & \dots & b_{u_{M-1}}^{t-1} \\ a^0(X) & \dots & a^j(X) & \dots & a^{t-1}(X) \end{bmatrix}. \tag{11}$$

When $v(X)$ is in $C(t)$, the row $b_u = (b_u^0, \dots, b_u^j, \dots, b_u^{t-1})$, $(u \in U^*)$, of $B(v)$ has to satisfy (9). This row vector b_u is thus a vector of a linear code $E(u)$ of length t over GF($2^{m(u)}$) specified by the parity check matrix H_u. We shall denote by $K(u)$ the dimension of $E(u)$ over GF($2^{m(u)}$). When $K(u) \geqslant 1$ holds, the minimum distance of $E(u)$ will be denoted by $D(u)$. The last row of $B(v)$ is always unrestricted. The corresponding code $E(\bar{U})$ when viewed as a code of length t over the set of polynomials of formal degree $m(I^*) - 1$ thus satisfies $D(\bar{U}) = 1$.

These notations are used in the next section where the U.E.P. properties of $C(t)$ are considered.

4. Application to unequal error protection

Let G be an arbitrary $k \times n$ generator matrix for an (n,k) linear code C over GF(q). Let $a = (a_1, \dots, a_k)$ be an information k–tuple and let $v = (v_1, \dots, v_n)$ be the corresponding encoded n–tuple: $v = a\,G$. For any integer i in $[1,k]$ we define $d(i)$ to be the minimum Hamming weight of all codewords v of C that are obtained from the encoding of an information k–tuple a with $a_i \neq 0$. It turns out that at the decoding a_i can be correctly estimated from a noisy version w of v as long as the Hamming weight of $v - w$ *is* $\leqslant (d(i) - 1)/2$. This

IPR2018-1557
HTC EX1023, Page 210

$d(i)$ is called the minimum distance relative to the information symbol a_i. The point is that the k minimum distances $d(i)$ are not necessarily equal. The levels of protection of the information symbols a_i may thus be different. This phenomenon is called *unequal error protection* (U.E.P.)[3,4].

In this section we show that the codes $C(t)$ described above can be used for U.E.P. when a specified encoding rule is applied. This rule is as follows. For all $u \in U^*$ we choose one vector b_u in the code $E(u)$ described at the end of sec. 3. This corresponds to, let us say, the preencoding of $m(u)K(u)$ binary information digits. As explained in sec. 3, each t–tuple b_u specifies the t polynomials $b_u^j(X)$ that have to be used in (5), together with an arbitrary t–tuple $a = (a^0(X); \ldots, a^{t-1}(X))$ of polynomials $a^j(X)$ of formal degree $m(I^*) - 1$, to construct a word $v(X) = \sum_{j=0}^{t-1} \alpha^j v^j(X)$ of $C(t)$. From the t–tuples b_u and a we construct a matrix as $B(v)$, (11). Of course some of the codes $E(u)$ may be of dimension zero, $(K(u) = 0)$. The corresponding rows of $B(v)$ are then identically zero and the 'choice' of the corresponding b_u $(= 0)$ does not preencode any information. We shall order the rows of $B(v)$ in such a way that these identically zero rows are the first ones of the matrix $B(v)$.

Suppose now that for a given code $C(t)$, the s^{th} row of these matrices $B(v)$ is the first one that is not identically zero. It follows that the minimum weight of any nonzero component $v^j(X)$ of $v(X)$ is equal to the minimum distance of the binary cyclic code of length N generated by $g_s(X) = \prod_{j<s} h_{uj}(X)$. This minimum distance will be denoted by $\Delta(u_s)$. On the other hand, a word $v(X)$ to which is associated a matrix $B(v)$ with nonzero entries in its s^{th} row will contain at least $D(u_s)$ nonzero components $v^j(X)$ (see sec. 3). Consequently the minimum bit weight of this $v(X)$ will be at least equal to $d(u_s) = D(u_s)\Delta(u_s)$. This $d(u_s)$ is the minimum distance of $C(t)$ relative to the $K(u_s)m(u_s)$ information bits that are preencoded into the row of $B(v)$ indexed by u_s. This means that if $v(X) = \sum_{j=0}^{t-1} \alpha^j v^j(X) \in C(t)$ is transmitted and received as $v_0^*(X) = \sum_{j=0}^{t-1} \alpha^j v_0^j(X)$ with at most $t(u_s) = \lfloor (d(u_s) - 1)/2 \rfloor$ bit errors, then the information contained in the row u_s of $B(v)$ will be correctly decoded. Having estimated the polynomials $b_{u_s}^j(X)$ we can reset $v_0^*(X)$ as follows

$$v_0^*(X) \to v_1^*(X) = v_0^*(X) - \sum_{j=0}^{t-1} \alpha^j b_{u_s}^j(X)\, e_{u_s}(X).$$

This $v_1^*(X)$ is a noisy version of $v_1(X) = v(X) - \sum_{j=0}^{t-1} \alpha^j b_{u_s}^j(X)\, e_{u_s}(X)$ which is a word of a subcode of $C(t)$, namely the subcode of all words the B matrix (11) of which has zeros not only in its $(s-1)$ first rows but also in the s^{th} one.

IPR2018-1557
HTC EX1023, Page 211

We can now estimate the $(s+1)^{\text{th}}$ row of $B(v)$ just as we have estimated the s^{th} one. The minimum distance $d(u_{s+1})$ relative to the $K(u_{s+1})$ $m(u_{s+1})$ information bits preencoded in the $(s+1)^{\text{th}}$ row of B is obtained as above

$$d(u_{s+1}) = D(u_{s+1})\,\Delta(u_{s+1}),$$

where $D(u_{s+1})$ is the minimum distance of $E(u_{s+1})$ and $\Delta(u_{s+1})$ is the minimum distance of the binary cyclic code of length N generated by

$$g_{s+1}(X) = \prod_{j<s+1} h_{uj}(X).$$

As a consequence the correct decoding of the row $b_{u_{s+1}}$ of B is possible any time b_{u_s} was correctly estimated and $v_0^*(X)$ contains at most

$$t(u_{s+1}) = \left[(d(u_{s+1}) - 1)/2\right]$$

bit errors. We can continue on this way to estimate the successive rows of $B(v)$ up to the last one. This last one will be correctly estimated provided the preceding rows were correctly estimated and the received word $v_0^*(X)$ contains at most $(D(I^*) - 1)/2$ bit errors where $D(I^*)$ is the minimum distance of the binary cyclic code generated by $g^*(X)$. Let us now consider the transmission of a codeword $v(X) = \sum_{j=0}^{t-1} v^j(X)\,\alpha^j$ that is received as $v(X) + e(X)$ where $e(X) = \sum_{j=0}^{t-1} e^j(X)\,\alpha^j$ denotes the error polynomial. On the basis of the discussion above we may first state the following property.

Property 4.1.

Suppose that at the decoding of $v(X) + e(X)$ the s first rows of $B(v)$ (for some s in $[1, M-1]$) were correctly determined. Then the $(s+1)^{\text{st}}$ row of $B(v)$ will also be correctly determined if $e(X)$ contain less than $D(u_{s+1})\,\Delta(u_{s+1})/2$ nonzero binary digits.

However this property does not give the complete error correcting capabilities of the SBI codes. Therefore it may be useful to consider SBI codes as being generalized concatenated codes in the sense of Zinov'ev. We refer the reader to the original paper [5]) or to [1]) for a definition of Zinov'ev codes. We also leave to the reader to make explicit how our codes can be seen as generalized concatenated codes. In his paper, Zinov'ev presents a very general decoding algorithm to achieve the performances promised by property 4.1. Moreover it can also (and at the same time) correct the errors described in the two following statements.

Property 4.2.

Let us assume that for some $s \in [1, M-1]$ the s first rows of $B(v)$ were correctly determined and that less than $\Delta(u_{s+1})/2$ components $e_j(X)$ of $e(X)$ are nonzero. In this case the $(s+1)^{\text{st}}$ row of $B(v)$ will be correctly estimated by the decoding algorithm of Zinov'ev.

IPR2018-1557
HTC EX1023, Page 212

Property 4.3.

Let us represent the error sequence $e(X)$ as $\sum_{i=0}^{N-1} e_i X^i$ with e_i in GF(2^m). Let us assume that the s first rows b_{uj} of $B(v)$ were correctly determined and that the number of nonzero e_i in $e(X)$ is less than $D(u_{s+1})/2$. In this case the $(s+1)^{st}$ row of $B(v)$ will be correctly obtained by the algorithm of Zinov'ev.

5. Examples

Let us now present two examples.

— First let α be a primitive element of GF(2^4) : $\alpha^4 = \alpha + 1$. For

$$U = \{0, 1, 2, 3, 4, 5, 6, 9\}$$

and

$$g(X) = \prod_{u \in U} (X - \alpha^u)$$

we define $C(3)$ to be the set of polynomials

$$v(X) = \sum_{j=0}^{2} v^j(X)\, \alpha^j,\ v^j(X) \in \mathrm{GF}(2)\,[X]/(X^{15} - 1),$$

that are multiples of $g(X)$. The matrix B associated to a nonzero $v(X)$ can be given the form:

$$B = \begin{bmatrix} b_0^0 & b_0^1 & b_0^2 \\ b_1^0 & b_1^1 & b_1^2 \\ b_3^0 & b_3^1 & b_3^2 \\ b_5^0 & b_5^1 & b_5^2 \\ a^0 & a^1 & a^2 \end{bmatrix}$$

with, for all j,

$b_0^j \in \mathrm{GF}(2)$,
$b_1^j \in \mathrm{GF}(2)\,[X]/(X^4 + X + 1) \simeq \mathrm{GF}(2^4)$,
$b_3^j \in \mathrm{GF}(2)\,[X]/(X^4 + X^3 + X^2 + X + 1) \simeq \mathrm{GF}(2^4)$,
$b_5^j \in \mathrm{GF}(2)\,[X]/(X^2 + X + 1) \simeq \mathrm{GF}(2^2)$,

The elements a^j of the last row of B are polynomials of formal degree 3 over GF(2). The entries of a matrix B associated to a codeword $v(X)$ of $C(3)$ have to satisfy

$$\sum_{j=0}^{2} b_0^j \alpha^j = 0,$$

$$\sum_{j=0}^{2} b_1^j \alpha^{j2^r} = 0, \qquad r = 0, 2, 3$$

$$\sum_{j=0}^{2} b_3^j \alpha^{j2^r} = 0, \qquad r = 0, 1, 3$$

$$\sum_{j=0}^{2} b_5^j \alpha^{j4^r} = 0, \qquad r = 0, 1.$$

IPR2018-1557
HTC EX1023, Page 213

The solution of this system is given by

$$b_0^j = b_1^j = b_3^j = 0 \qquad \text{all } j$$
$$b_5^0 = b,\ b_5^1 = b_5^2 = \alpha^{10} b \qquad b \subset \mathrm{GF}(2^2). \tag{12}$$

The first nonzero row of B is the fourth one corresponding to $u = 5$, and the minimum weight of a nonzero element in this row is 3 : $D(5) = 3$. The minimum weight $\Delta(5)$ of a nonzero $v^j(X)$ is the minimum weight of the (15,6) binary cyclic code generated by $(X-1)(X^4+X+1)(X^4+X^3+X+1)$, and thus $\Delta(5) = 6$ holds. As a consequence $d(5) = D(5)\Delta(5)$ is equal to 18. On the other hand, $D(7) = 1$ and $\Delta(7) = 8$ hold, so that $d(7)$ is equal to 8. We have thus constructed a binary code of length 45 and dimension 14. Two information bits are encoded by the choice of the element b of $\mathrm{GF}(2^2)$ in (12) and they are protected with minimum distance 18. The 12 last information bits are encoded by the choice of the 3 polynomials $a^j(X)$, $(0 \leqslant j \leqslant 2)$, and they are protected with minimum distance 8. We note that the best known (45,14) code for *equal* error protection has minimum distance 14 [1]).

— For the second example, let α be primitive in $\mathrm{GF}(2^3)$: $\alpha^3 = \alpha + 1$ and let us consider the cyclic (7,2) RS code over $\mathrm{GF}(2^3)$, generated by

$$g(X) = \prod_{i=1}^{5} (X - \alpha^i).$$

We choose $t = m = 3$. The matrix B associated to a nonzero $v(X)$ can be given the form

$$B = \begin{bmatrix} b_1^0 & b_1^1 & b_1^2 \\ b_3^0 & b_3^1 & b_3^2 \\ b_0^0 & b_0^0 & b_0^2 \end{bmatrix},$$

with

$$b_1^j \in \mathrm{GF}(2)[X]/(X^3+X+1) \simeq \mathrm{GF}(2^3),$$
$$b_3^j \in \mathrm{GF}(2)[X]/(X^3+X^2+1) \simeq \mathrm{GF}(2^3),$$
$$b_0^j \in \mathrm{GF}(2).$$

The entries of B have to satisfy

$$\sum_{j=0}^{2} b_1^j \alpha^{j2^r} = 0, \qquad r = 0, 1, 2$$
$$\sum_{j=0}^{2} b_3^j \alpha^{j2^r} = 0, \qquad r = 0, 1 \ .$$

The solutions of this system are

$$b_1^j = 0, \quad \text{all } j;$$
$$b_3^0 = b,\ b_3^1 = \alpha\, b,\ b_3^2 = \alpha^4 b, \quad \text{for any } b \in \mathrm{GF}(2^3);$$
$$b_0^j \text{ arbitrary.}$$

IPR2018-1557
HTC EX1023, Page 214

As a consequence we get

$$D(3) = 3,\ \Delta(3) = 3;$$
$$D(0) = 1,\ \Delta(0) = 7.$$

The three first information bits are encoded by the choice of the element b of GF(2^3) that specifies the elements b_3^j of B. These 3 bits are protected by a minimum distance $d(3) = 3 \times 3 = 9$. The three last information bits are the three bits b_0^j ($0 \leqslant j \leqslant 2$); these are protected by a minimum distance $d(0) = 1 \times 7 = 7$. We obtain on this way a binary code of length 21 and dimension 6, that has the same performances as a (21,6) binary cyclic code recently considered by Van Gils [4]).

6. Acknowledgment

The authors acknowledge the useful comments of the referees. These comments have lead to several improvements of the paper.

REFERENCES

[1]) F. J. MacWilliams and N. J. A. Sloane, The theory of error correcting codes, North-Holland Publishing Company, Amsterdam 1977.
[2]) E. Berlekamp, Algebraic coding theory, McGraw-Hill, New York 1968.
[3]) I. M. Boyarinov and G. L. Katsman, IEEE Trans. on Information Theory, **IT-27**, 168 (1981).
[4]) W. J. van Gils, IEEE Trans. on Information Theory, **IT-29**, 866 (1983).
[5]) V. A. Zinon'ev, Problems of Information Transmission, **17**, 254 (1981).

Ned. Philips Bedrijven B.V.
PHILIPS NATUURKUNDIG LAB
BIBL... -KAMP Y - 1
P.O. ... 000
5600 JA EINDHOVEN
NEDERLAND

IPR2018-1557
HTC EX1023, Page 215

Philips J. Res. **39**, *206-225, 1984* *R 1092*

ALGORITHMIC ASPECTS OF DIGITAL SYSTEM DESIGN

by M. DAVIO

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

The complexity of very large scale integrated circuits has deeply modified the design techniques; the new methods have two essential characteristics: they are both computerized and hierarchical in nature. As a consequence, it becomes necessary to develop new algorithmic and algebraic tools to achieve acceptable tradeoffs between conflicting optimization criteria. This will be illustrated by the design of data paths and it will be shown how graph theory and complexity theory can help the designer to achieve his objectives. The paper first presents a structural interpretation algorithm for high level language programs, i.e. a formal method for translating such a program into a hardware structure. It also covers the selection of an appropriate set of computation resources and the detection and modification of the intrinsic parallelism.

Math. Rev.: 94C10, 94C15

1. Introduction

A simple way of modelling a digital system is to say that it performs a *computation*. The word computation is taken here with its general mathematical meaning [1]). The validity of this model is easily perceived in the case of special purpose circuits, such as digital filters or pocket calculators; it also holds true in the case of general purpose systems: for example, a microprocessor executes the interpretation algorithm of its instruction set.

The design of a digital system will accordingly start with a description of the computation to be performed; the complexity of modern digital systems, the wish to achieve correctness proofs makes it a must to start from a *formal specification*. In general, the designer will thus initially select a suitable algorithm and express it in some high level language.

In the context of Very Large Scale Integrated circuits, the design process identifies with a compilation as it has to translate the initial high level specification into a target language, which is the description of the integrated circuit layout. That compilation is generally known as *silicon compilation* [2]). Silicon compilation has two important characteristics, both due to the complexity of the target systems measured say by the number of constituent transistors: typical target systems may involve up to one million transistors; it has to be

IPR2018-1557
HTC EX1023, Page 216

both *computer aided* and *hierarchical* in nature, i.e. it should proceed in a top down manner along a number of levels of progressively refined descriptions.

The algorithmic aspect of this design process is at once perceived. At each level of the hierarchy, the designer will receive the formal specification of a behaviour, a set of primitives characteristic of the hierarchical level and a set of acceptable *design rules* selected to insure the design correctness. For example, at the familiar logical level, the specified behaviour could be a synchronous counter, the primitives could be NAND gates and the design rules would describe the acceptable ways of interconnecting gates.

If we scan the literature devoted to silicon compilation[3,4,5,6]), we shall observe that the first tools developed for coping with the complexity of large chips mainly addressed layout problems (design rule checking, routing, placement), simulation and timing problems (electrical, logical, mixed level simulation). So far, very little has been done at the functional level, in spite of its fundamental importance: the choice of the algorithm to be executed to perform any suitable function has indeed a crucial importance on the performances of the computation. We face here classical software problems such as correctness proof and program transformation but in the renewed perspective of hardware implementation.

This is roughly the frame of the present discussion: in sec. 2, we shortly describe the general model of algorithmic state machine and show how it is possible to map an algorithm, described in some high level language on a corresponding hardware structure; this is the *structural interpretation algorithm*. We next focus on the design of processing units and present various tools allowing one to select appropriate computation resources: the optimization problems are numerous here and we attempt to show how algebraic methods can be used to solve these problems; sec. 3 studies the problem of choosing a minimum cost set of resources preserving the intrinsic parallelism of the given algorithm; this is the *resource selection algorithm*. Finally, sec. 4 tackles the sequentialization problem and shows how it is possible to minimize the number of resources for performing a computation within a given target time.

These are only sample problems selected to show how mathematical techniques can help one to model various situations encountered in digital design. There is however no doubt that important results can be reached in this new emerging field.

2. Algorithmic state machines

2.1. *The design objectives*

The starting point of a digital design is an algorithm expressed for example as a program in some high level language. This will be the data we have to

IPR2018-1557
HTC EX1023, Page 217

translate or *compile* into a hardware device. Basically, such a program consists of five type of entities:

1. *variable names* such as $x, y, \ldots$, representing scalars of predefined types (integers, characters, ...) or such as $\mathbf{v}, \mathbf{w}, \ldots$, representing vectors;
2. *function names* i.e. transformations on the value space of the variables; function names will not only include the standard operators $+, -, \cdot, /$ but other operators such as SHIFT, ROTATE, COMPARE, when they appear as "natural" in the expression of the algorithm;
3. *assignments* such as $x := f(\mathbf{v})$ or $\mathbf{u} := \mathbf{g}(\mathbf{v}, \mathbf{w})$; the introduction of vector assignments will take for us the particular meaning of *simultaneous* or *parallel* assignment;
4. *predicates* such as $(a \geqslant b)$ or $(\mathbf{x} > \mathbf{0})$: predicates are most often binary valued functions of the variables to be used for test purposes;
5. *constructs* such as *if ... then ... else, while ... do, repeat ... until, case ... of, for.*

The choice of the algorithm is in general a critical one as it may deeply influence the performance (cost and computation time) of the resulting circuit.

At the other end of the design process, we find the basic digital components. For sake of simplicity, we shall restrict ourselves to two types of components, *combinational logic circuits* and *synchronous registers*. A *combinational logic circuit* is viewed as a black box with n data inputs, p control inputs and m data outputs. For each value of its control inputs, the circuit realizes a mapping from its inputs to its outputs.

A combinational logic circuit CC is characterized by its cost $\gamma(CC)$ and its delay $\delta(CC)$; the cost is to be taken in the wide sense, i.e. it may be a chip area, a power consumption, etc. For example, if we take as cost unit the bit of equivalent ROM, we immediately obtain:

$$\gamma(CC) \leqslant m \cdot 2^{n+p}.$$

The delay is the maximum time an input modification requires to propagate its effects to all the outputs; this propagation is viewed as *asynchronous* as there is at this level no clock signal to control the propagation.

The second type of component we require is the *synchronous register*. We adopt here a very simple register model: the register has n input and n output data bits. It has two control lines: the CK or clock line and the LD or load line: the register is loaded with the input data at, say, the clock signal risetime if the load signal is high. For the simplicity of this introductory presentation, we shall assume that the registers and flip-flops can be read and loaded with a new value during a single clock period; from the technological point of view, this amounts to assume that the flip-flops are of the master-slave type or,

IPR2018-1557
HTC EX1023, Page 218

equivalently, that the system has been provided with a two phase clock scheme.

Our problem can now be stated as follows: *given an algorithm represented as a program in some high level language, how do we "map" that algorithm onto a hardware structure that implements that algorithm*, i.e. that, from any set of data, generates the right results while following the sequencing indications given by the program.

As such, the objective may look rather vague; in the early days of digital system design, it looked more like an art than like a science or a technique and, to some extent, this is still the case today. Still, we immediately feel that to improve over the present state of the art, and, in particular, if we wish to apply to the problem some of the computing power available today, we have to formalize the relationship between the given program and the target implementation. This translation from a given program to a target hardware architecture bears a close analogy with the general problem of compilation and it is thus natural to use in this context the term "*silicon compiler*": the problem is indeed the translation of a high level language to another language, expressing the same behaviour in physical terms of floorplan, layout, masks, ...

We shall however aim at a more limited objective and we shall call it *structural interpretation*: it is *structural*, because we shall limit ourselves to the translation of the high level language up to the structural (or functional) level, without going down to the geometrical or physical level; it is an *interpretation* in the sense that it will follow step by step the original program, without attempting any global optimization.

2.2. *Algorithmic state machines*

The structural interpretation, i.e. the link between the initial program and the basic circuit elements described above will be established thanks to the model of *algorithmic state machine*[7,8,9,10]); the system appears as decomposed in two interconnected parts the *control unit* and the *processing unit* or *data path*.

The two units cooperate by exchanging various signals: the control unit provides the processing unit with *command signals*, to inform the latter of the next operation to be carried out. Typically, command lines correspond to control variables of programmable computation resources or to register control lines. On the other hand, the processing unit provides the control unit with binary signals called *condition variables*. These condition variables provide the control unit with the relevant information about the past history of the computation to allow decisions about the next step of the computation.

It should be clear at this point that, in the model of algorithmic state machine, both the processing and the control units are viewed as *finite state*

IPR2018-1557
HTC EX1023, Page 219

machines[11,12]). A finite state machine is implemented as a *synchronous sequential circuit*, i.e. as the loop interconnection of a combinational circuit *CC* and of a register *R*. The contents of the register, available at its output, is the present state. From the present state and from the applied input signal the combinational circuit *CC* computes the output signal and the next state, to be stored in the register *R* at the next clock pulse.

We shall show in the forthcoming section how it is possible to achieve a *structural interpretation* of the initial program, i.e. a formal mapping from that initial program onto a corresponding hardware structure; in the present section, we limit ourselves to some general guidelines:

1. the *processing unit* will take care of the variable names, of the functions and of the assignments; rather intuitively, the variable names will be assigned registers and the function names will be assigned combinational logic circuits; observe that, by function names, we not only mean the functions appearing in the right hand members of the assignments but also the functions required to evaluate the truth value of the conditions appearing in the program; finally, the assignments will become *functional register transfers* of the type:

$$\mathbf{R} := \mathbf{F}(\mathbf{R}); \tag{1}$$

   the latter equation means that the contents of the set of registers **R** is to be loaded with a function **F** of the contents of these registers.
2. the *control unit* will take care of the program itself i.e. of the constructs

   while ... do, if ... then ... else,

   etc., of their sequencing and of the *condition variables*, i.e. of the binary variables providing the truth value of the conditions to be evaluated.

2.3. *The processing unit*

The processing unit basically consists of three types of entities:

1. *combinational switching circuits*, viewed as computation primitives and performing two types of functions, viz. evaluation of the results required by the right hand side of the assignments and evaluation of the truth value of the predicates appearing in the algorithm;
2. *registers and flip-flops* allowing the appropriate storage of data, of intermediate and final results and of condition variables;
3. *controlled connection networks* allowing one to connect, at each step of the computation process, the appropriate data to the computation resources and to store in appropriate registers the results of the computations carried out by the resources.

IPR2018-1557
HTC EX1023, Page 220

With this model in mind, it is not difficult to imagine the main steps of the design of the processing unit; to reach this goal, we should bear in mind the fact that the processing unit is to be implemented as a synchronous sequential machine, i.e. that all the operations pertaining to a specific computation step have to be carried out during a single clock period.

The structural interpretation algorithm

- STEP 1: for each step of the algorithm, prepare a list of the required *behaviours* or function names;
- STEP 2: define a set of (possibly programmable) computation resources sufficient to realize, at each step, the required behaviours, taking into account the intrinsic possibilities of parallelism;
- STEP 3: *resource allocation*: for each step of the computation, allocate an appropriate resource to each of the required behaviours;
- STEP 4: for each step of the algorithm, prepare a list of the variables to be stored at the end of that step; determine accordingly the number and the size of the required registers;
- STEP 5: *memory allocation*: for each step of the computation, allocate an appropriate register to all the variables to be stored at the end of that step.

It should be clear that the above five steps completely determine the processing unit, including the connection network and its position at each step. The memory allocation appears as a critical step in the above method as it critically influences the complexity of the connection network.

2.4. *The control unit*

We now turn to the second constituent part of the algorithmic state machine, viz. its *control unit*. Our primary goal is to implement the control unit as a finite state machine. The first thing to do, once that decision has been taken, is to ask oneself what are the *natural* constructs associated with a finite state machine. It is actually easy to translate in software terms the concepts of transition and of output functions. For a control automaton having as input signals the minterms $m_i(\mathbf{x})$ in the condition variables $\mathbf{x}$, and originating commands σ_j to the processing unit, the transition and output functions can be described by a set of **case** statements.

To illustrate such a **case** statement, we first introduce some notations. Consider a finite automaton A with the present state q and receiving as input literal the minterm $m_i(\mathbf{x})$; then, its next state q^+ and its output signal, the command σ, will be denoted $q^+ = qM_i$ and $\sigma = qN_i$, respectively.

A typical case statement describing the behaviour of a finite automaton is then

IPR2018-1557
HTC EX1023, Page 221

$$
\begin{array}{llll}
\langle \text{label } q \rangle & \text{case} & \mathbf{x} & \text{of} \\
 & & m_0(\mathbf{x}) & \sigma_0 = qN_0 \quad \text{goto } \langle \text{label } qM_0 \rangle \\
 & & m_1(\mathbf{x}) & \sigma_1 = qN_1 \quad \text{goto } \langle \text{label } qM_1 \rangle \\
 & & \cdots & \cdots \quad \cdots \\
 & & m_{2^n-1}(\mathbf{x}) & \sigma_{2^n-1} = qN_{2^n-1} \quad \text{goto } \langle \text{label } qM_{2^n-1} \rangle \\
 & \text{end.} & &
\end{array}
$$

Clearly, in the above **case** statement, the internal state q of the corresponding automaton is associated with the statement label; once that correspondence has been grasped, the analogy becomes straightforward.

The next step of the process would be to match the control structures of the given program onto **case** statements of the above type. As an example, consider the two basic structures of *structured* or *gotoless* programming

(i) the statement

$$\text{if } x \text{ then } \sigma \text{ else } T$$

is translated as

$$
\begin{array}{llll}
\langle q \rangle & \text{case} & x & \text{of} \\
 & & 0 & T \quad \text{goto } \langle q+1 \rangle \\
 & & 1 & \sigma \quad \text{goto } \langle q+1 \rangle \\
 & \text{end} & &
\end{array}
$$

(ii) similarly, the statement

$$\text{while } x \text{ do } \sigma$$

is translated as

$$
\begin{array}{llll}
\langle q \rangle & \text{case} & x & \text{of} \\
 & & 0 & \lambda \quad \text{goto } \langle q+1 \rangle \\
 & & 1 & \sigma \quad \text{goto } \langle q \rangle \\
 & \text{end} & &
\end{array}
$$

where λ denotes the absence of operation, commonly known as a no-operation or *NOOP* statement.

These two examples, together with the universality of structured programs, immediately show the universality of our model, i.e. show that the control structures of an arbitrary program may always be modelled by a finite state machine. The central point in that mapping is the correspondence between the used labels and the internal states of the finite state machine.

A *canonical program* P is first defined as a sequence of *instructions*

$$N_0 N_1 \ldots N_T ,$$

containing an *initial instruction* N_0 and a *final instruction* N_T. If we assume

IPR2018-1557
HTC EX1023, Page 222

that there are exactly n condition variables, a typical instruction N_i will be of the form

$$\begin{array}{llll} N_i & m_0(\mathbf{x}) & \sigma_0 & N_{i,0} \\ & m_1(\mathbf{x}) & \sigma_1 & N_{i,1} \\ & \cdots & \cdots & \cdots \\ & m_j(\mathbf{x}) & \sigma_j & N_{i,j} \\ & \cdots & \cdots & \cdots \\ & m_{2^n-1}(\mathbf{x}) & \sigma_{2^n-1} & N_{i,2^n-1}\ , \end{array}$$

where $m_j(\mathbf{x})$ stands for the j-th minterm in the condition variables $\mathbf{x}$, σ_j stands for a functional register transfer of the processing unit (it will correspond to appropriate values of the control lines) and N_i, $N_{i,j}$ stand for instruction labels. The interpretation of the instruction is the following one: during the execution of instruction N_i, if $\mathbf{x} = \mathbf{j}$, i.e. if $m_j(\mathbf{x}) = 1$ and $k \neq j \Rightarrow m_k(\mathbf{x}) = 0$ then, execute the command σ_j and goto the instruction wearing the label $N_{i,j}$.

With that interpretation, the above instruction immediately appears as another notation of a **case** statement and accordingly defines the transition and output functions of the corresponding automaton. We are immediately in position of designing the control unit as a synchronous sequential circuit. The input alphabet of that automaton is the set $\{m_i(\mathbf{x})\}$ of minterms in the condition variables $\mathbf{x}$; the output alphabet is formed by the binary coded commands σ_j and the set of internal states is given by the set of instruction labels $\{N_k\}$. The above general instruction then corresponds to an internal state of the system and defines, for each input signal, a next state and an output signal.

2.5. *Conclusion*

In the preceding sections, we laid the principles of a structural interpretation of programs written in a high level language thanks to the model of algorithmic state machines. The main steps of the design may be summarized as follows:

- — choice of the algorithm; possible algorithm transformation; detection of the intrinsic parallelism;
- — design of the processing unit, itself decomposed into five steps: determination of the functions required at each step; selection of an appropriate set of resources; resource allocation; selection of an appropriate set of memory elements; memory allocation;
- — design of the control unit; transcription of the program in the canonical form; if required, state minimization; encoding of the commands and of the internal states; optimization of the combinational part of the control unit.

IPR2018-1557
HTC EX1023, Page 223

As a matter of fact, these are only the main lines of a possible strategy and this discussion should be completed by a study of the techniques available to describe various implementations of a given program and to provide help in the selection of acceptable tradeoffs. Some of these techniques are discussed in the following sections.

3. Data path design

3.1. *Introduction*

The design of processing units has been paid little attention in the literature [13,14]: in general, the hardware architecture of the *processing unit* or *data path* is considered as given and the algorithm to be implemented is first transformed so as to become executable on the given data path. If, in spite of its lack of flexibility, that position is acceptable in the case of general purpose processors, it does not help much the designer in case of custom design (or special purpose) devices. We shall now investigate various techniques that can be put at work to formalize the data path design.

As starting point of the design process, we shall assume available

- a list of computation steps to be performed. Thus, if the program is expressed in a PASCAL-like form, the steps roughly correspond to assignment statements,
- a list of data items available before the initial step of the computation,
- for each step of the computation, a list of the required functions or behaviours. We shall assume that these functions are members of the set

$$\mathcal{F} = \{f_0, f_1, \ldots, f_{c-1}\},$$

- for each step of the computation, a list of data items required for the following steps of the computation.

Our general objective is to describe design tools that will generate a data path able to support the intrinsic parallelism of the given algorithm: this has a double aspect: on the one hand, it should be possible to realize simultaneously, at each step of the algorithm all the functions required at that step. On the other hand, it should be possible to perform simultaneously all the functional transfers required at that step. Thus the preservation of the intrinsic parallelism imposes requirements both on the number and type of available resources and on the connecting capabilities of the processing unit. In the following subsection, we study the first of these two requirements.

3.2. *Resource requirements*

According to our definitions, the functions required for the algorithm execution are members of the set

IPR2018-1557
HTC EX1023, Page 224

$$\mathscr{F} = \{f_0, f_1, \ldots, f_{c-1}\}.$$

Let us call $w_i(t)$ the number of functions of type f_i required at time t and define

$$w_i = \max_t \{w_i(t)\}. \tag{2}$$

Thus, if we had at our disposal w_i resources realizing the behaviour f_i, we would have enough of these resources at any step to satisfy the needs expressed by the algorithm. However, a behaviour of type f_i can not only be realized by a resource having this single behaviour but also on programmable resources having, among others, that particular behaviour. We shall accordingly assume that the candidate resources are in the set

$$\mathscr{R} = \{R_S | S \subseteq \mathbf{Z}_c = \{0, 1, \ldots, c-1\}\},$$

where R_S is a programmable resource having the behaviours $f_i \forall i \in S$. Thus, a resource $R_{0,1}$ is a programmable resource having the behaviours f_0 and f_1, etc. A resource R_S will be characterized by is cost γ_S. As usual, we shall assume that all the resources R_S have identical delays, chosen as unit of time.

Then, for any $S \subseteq \mathbf{Z}_c$ and such that $|S| > 1$, we define

$$w_S(t) = \sum_{i \in S} w_i(t), \tag{3}$$

and

$$w_S = \max_t \{w_S(t)\}. \tag{4}$$

Observe that one has

$$w_S \leqslant \sum_{i \in S} w_i, \tag{5}$$

as the needs in functions f_i are in general not simultaneously maximal. Finally, we denote by x_S the number of resources of type R_S used in the design.

The unknowns of our problem are the quantities x_S, $S \subseteq \mathbf{Z}_c$. A solution will be said optimal if it minimizes the total cost:

$$\sum_{S \subseteq Z_c} x_S \gamma_S. \tag{6}$$

This optimization problem is rather complex and we shall only propose an heuristic solution. We first describe a necessary and sufficient set of conditions[13]) to be satisfied by any acceptable set of values x_S. We first assume $c = 2$. This situation will serve as initial point to the following proof by induction.

Theorem 1.

A necessary and sufficient condition to be satisfied by the integers x_0, x_1 and $x_{0,1}$ to allow a parallelism preserving implementation of an algorithm using two types of functions is to satisfy the system of inequations:

IPR2018-1557
HTC EX1023, Page 225

$$\begin{aligned} x_0 + x_{0,1} &\geqslant w_0; \\ x_1 + x_{0,1} &\geqslant w_1; \\ x_0 + x_1 + x_{0,1} &\geqslant w_{0,1} \ . \end{aligned} \tag{7}$$

Proof

The condition is obviously necessary; indeed, the only way to realize the required behaviours f_0 is to use resources of types R_0 or $R_{0,1}$, etc.

To prove the sufficiency, we distinguish two cases:

(a) $x_0 \geqslant w_0$. In this case, at time t, we select $w_0(t) \leqslant w_0 \leqslant x_0$ of the resources R_0 to realize the required functions f_0. According to the second eq. (7), the $x_1 + x_{0,1}$ remaining resources are sufficient to realize the required behaviours f_1.

(b) $x_0 < w_0$. Let us consider a time t such that $x_0 < w_0(t)$. We use the x_0 resources R_0 to realize as many behaviours f_0 and use $(w_0(t) - x_0)$ resources $R_{0,1}$ to realize the still required behaviours f_0. We are thus left with $(x_{0,1} - w_0(t) + x_0)$ resources of type $R_{0,1}$ and the first eq. (7) shows that this number is actually positive. Now, there are $(x_1 + x_{0,1} - w_0(t) + x_0)$ resources available to realize the requires functions f_1. This number is sufficient, as the third eq. (7) yields

$$x_0 + x_1 + x_{0,1} \geqslant w_{0,1} \geqslant w_{0,1}(t) = w_0(t) + w_1(t) \geqslant w_0(t).$$

□

It is now easy to extend theorem 1 to the general situation involving c distinct functions f_i. This is the purpose of the following theorem.

Theorem 2.

A necessary and sufficient condition to be satisfied by the integers x_S to allow a parallelism preserving implementation of an algorithm using c types of functions is to satisfy the system of inequations:

$$\sum_{T:T\cap S\neq\phi} x_T \geqslant w_S; \forall S \subseteq \mathbf{Z}_c. \tag{8}$$

Proof

The condition is clearly necessary. Indeed, the resources appearing in the left hand member are the only ones able to realize the behaviours required by the right hand member. The proof of the sufficiency is achieved by induction on the number c of types of functions and theorem 1 is the initial step of that recurrence. We assume that the proposition holds true for c types of functions and show that it also holds true for $c + 1$ types of functions. We furthermore assume that we dispose of x_S resources of type R_S, $c \notin S$ and of $x_{S,c}$ resources of type $R_{S,c}$ and we have thus to show that, if the numbers x_S and $x_{S,c}$ satisfy

IPR2018-1557
HTC EX1023, Page 226

the hypothesis, then, there exists a strategy allowing one to realize the required functions at each step of the algorithm.

The strategy can be outlined as follows: the behaviours $f_0, f_1, \ldots, f_{c-1}$ will be realized in the following way:

— x_0 resources R_0 and $y_{0,c}$ resources $R_{0,c}$;
— x_1 resources R_1 and $y_{1,c}$ resources $R_{1,c}$;
— $x_{0,1}$ resources $R_{0,1}$ and $y_{0,1,c}$ resources $R_{0,1,c}$;
— etc. ...
— $\forall T \subseteq \mathbf{Z}_c$, x_T resources R_T and $y_{T,c}$ resources $R_{T,c}$.

We shall next try to reduce as far as possible the quantities $y_{T,c}$ and show that there remain enough resources to realize the required behaviours f_c. To satisfy the recurrence hypothesis, the numbers x_T and $y_{T,c}$ should verify the inequations

$$\forall S \subseteq \mathbf{Z}_c, \sum_{T: T \cap S \neq \phi} (x_T + y_{T,c}) \geqslant w_S. \tag{9}$$

We next try to reduce the quantities $y_{T,c}$, without violating the conditions (9) and attempting to minimize the sum

$$\sum_{T \subseteq Z_c} y_{T,c}.$$

We now show that, if that sum is minimal, there exists some set S such that the two following conditions hold

$$\sum_{T: T \cap S \neq \phi} (x_T + y_{T,c}) = w_S; \tag{10}$$

$$\sum_{T: T \cap S \neq \phi} y_{T,c} = 0. \tag{11}$$

Clearly, if none of the eqs (9) is an equality, one of the $y_{T,c}$ can be reduced in such a way that one at least of these equations becomes an equality. Assume that the inequation with index S has become an equality. Now, if the obtained solution is minimal, then

$$\tilde{T} \cap S = \phi \Rightarrow y_{\tilde{T},c} = 0.$$

Indeed, if this is not the case, either it is possible to reduce $y_{\tilde{T},c}$ by one unit in all the inequations of index $\tilde{S} \neq S$, or one at least of these inequations reduces to an equality. Assume that that equality corresponds to the index $\tilde{S}$. It is then simple to show that the inequation (9) corresponding to the index $S \cup \tilde{S}$ also reduces to an equality. Keeping increasing the set S in a similar way, we finally reach the situation described by eqs (10, 11). Combining these equations, we obtain

IPR2018-1557
HTC EX1023, Page 227

$$\sum_{T:T\cap S\neq\phi} x_T + \sum_{T\subseteq Z_c} y_{T,c} = w_S . \tag{12}$$

Now, the number of resources available to realize the behaviours f_c is given by

$$\sum_{T\subseteq Z_c} (x_{T,c} - y_{T,c}),$$

and according to eq. (12), this number is equal to

$$\sum_{T\subseteq Z_c} x_{T,c} + \sum_{T:T\cap S\neq\phi} x_T - w_S . \tag{13}$$

Now, the eq. (8) with index S,c yields

$$\sum_{T:T\cap S\neq\phi} x_T + \sum_{T\subseteq Z_c} x_{T,c} \geqslant w_{S,c} ;$$

at time t, the right hand member of this expression is equal to $w_S(t) + w_c(t)$ and this shows that the number of available resources, as given by eq. (13) is sufficient to realize the required behaviours.

Let us now go back to the main problem, i.e. to determine the number of resources x_S so as to minimize the cost function (6) while satisfying conditions (8). The obvious starting point is to choose:

$$x_i = w_i; \quad x_S = 0, \quad \forall S : |S| > 1. \tag{14}$$

This solution indeed uses exactly as many non-programmable or elementary resources as required if no programmable resources are available.

The algorithm will then run along a series of steps during each of which one shall replace elementary resources by programmable resources. More precisely, one shall replace δ sets consisting of a single copy of the resources R_i, $i\in U$ by δ programmable resources R_U. The quantity δ will be selected so as to keep satisfying the conditions (8).

Note that, for each of the above δ substitutions, the cost return is given by

$$\lambda U = \sum_{i\in U} \gamma_i - \gamma_U . \tag{15}$$

Now, to compute the maximal value of δ, we observe that the modification of x_i, $i\in U$ and of x_U modifies all the inequations (8) such that $U\cap S\neq\phi$. Such an inequation contains in its left hand member $|U\cap S|$ of the variables x_i, $i\in U$ as well as the variable x_U itself. If we denote by ε_S the excess of the left hand member on the right hand member

$$\varepsilon_S = \sum_{T:T\cap S\neq\phi} x_S - w_S ,$$

we immediately see that δ should satisfy the inequality

$$\varepsilon_S \geqslant \delta(|(S\cap U)| - 1),$$

IPR2018-1557
HTC EX1023, Page 228

so that we finally obtain

$$\delta = \min_{U: U \cap S \neq \phi} \left\{ \left\lfloor \frac{\varepsilon_S}{(|S \cap U| - 1)} \right\rfloor \right\}. \tag{16}$$

The above discussion justifies the following algorithm.

The resource selection algorithm

— STEP 1. $x_i := w_i$; $|S| > 1 \Rightarrow x_S := 0$.

— STEP 2. For each subset $U \subseteq \mathbf{Z}_c$ not yet considered in previous executions of step 2, compute the maximum value of δ by (16) and the corresponding return $\delta\lambda_U$. If the maximum possible return is less than or equal to zero, then end, else, select the subset U corresponding to the maximum return and carry out the δ substitutions $\{R_i | i \in U\} \rightarrow R_U$. Repeat step 2.

4. Sequentialization

4.1. *Computation schemes*

In many applications, it is not required to preserve the intrinsic parallelism of a given algorithm as there is no stringent requirement on the speed of the process. In that case, it becomes interesting to reduce as far as possible the cost of the implementation and a typical technique for reaching this goal is that of *sequentialization*: this is the process by which one realizes in sequence a number of independent tasks that could otherwise have been executed simultaneously. We shall study the sequentialization process and the related optimization problems on a special type of algorithms called *computation schemes*.

As any algorithm, a *computation scheme* is expressed in terms of variable names, of *computation primitives* and of *admissible constructs*. The computation primitives are m-tuples of binary functions

$$\mathbf{f}: \{0, 1\}^n \rightarrow \{0, 1\}^m. \tag{17}$$

assumed to be available as combinational switching circuits and displayed as black boxes.

Among the computation primitives, we shall always assume the presence of *projection functions*. The i–th projection $\mathrm{pr}_i(\mathbf{x})$ of the vector

$$\mathbf{x} = [x_{n-1}, \ldots, x_1, x_0],$$

is simply the i–th component x_i of $\mathbf{x}$. The projection function in fact allows us to isolate specific connections within bundles of connections. In software terms, the projection function correspond to less frequently encountered bit extraction.

IPR2018-1557
HTC EX1023, Page 229

We can now give the formal definition of a *computation scheme*. Consider the set of computation primitives

$$\Pi = \{\Pi_i | \, i \in I \text{ and } \Pi_i : \{0, 1\}^{n_i} \rightarrow \{0, 1\}^{m_i}\}. \tag{18}$$

Thus Π_i is a black box having n_i binary inputs and m_i binary outputs. A computation scheme is a set of equalities giving the value of a particular variable V_i. Thus

$$C = \{V_0, V_1, \ldots, V_R\}, \tag{19}$$

where, for every k, one of the following situation holds:

— V_k is a constant;

— V_k is an input variable;

— V_k is a projection function, i.e., for example,

$$V_k(\mathbf{x}) = \mathrm{pr}_j(V_t(\mathbf{x})); \tag{20}$$

— V_k is obtained by function composition:

$$V_k(\mathbf{x}) = \Pi_i(V_{k_1}, V_{k_2}, \ldots, V_{k_{n_i}}) \tag{21}$$

such that:

$$\forall \, i = 1, 2, \ldots, n_i : k_i < k. \tag{22}$$

Condition (22) guarantees the loopfree nature of the algorithm.

It is important to note, in the definition of a computation scheme, that the order of the definitions of the V_i's is irrelevant as the order of the partial computations is fixed by the indexes or by the names of these variables. It should also be clear that the replacement of argument variables by their expressions actually correspond to function composition. *Function composition* actually appears as the single admissible construct of computation schemes.

Many algorithms of practical interest are computation schemes. This is in particular the case of fixed length iterations. As a matter of fact, any algorithm may be replaced by some equivalent computation scheme and, whenever possible, the designer should start his design process by considering implementations of computation schemes.

A computation scheme has an obvious structural interpretation as a loopfree interconnection of combinational circuits computing the algorithm primitives. It obviously suffices to interconnect these operators as indicated in the computation scheme description.

Let us conclude this introductory section by presenting a rough classification of computation schemes based on the type of computation primitives they put at work. Let us first observe that the computation primitives may be

IPR2018-1557
HTC EX1023, Page 230

proper or *improper*: by improper computation primitive, we mean primitives having a zero delay or a negligible delay, such as the projection function, the left and right shifts, etc. Improper computation primitives thus appear as fixed connection patterns requiring no logic elements. At the opposite, proper computation primitives carry out an actual computation such as addition or amplitude comparison.

A computation scheme is a *unit execution time* scheme iff all its proper computation primitives have identical delays: this particular delay will be chosen as unit of time. Similarly, a computation scheme is *homogeneous* iff all its proper computation resources are identical. An homogeneous computation scheme is thus also a unit execution time scheme. The serial addition provides us with an example of homogeneous computation scheme. As a matter of fact, all the finite iterations are of this type.

In what follows, we discuss the possibilities of implementing computation schemes as synchronous sequential circuits. To reach this design goal, we essentially have to discover a technique for decomposing the computation into a number of consecutive steps. Once this has been done, we find ourselves in a position similar to the one discussed about algorithmic state machines: we can indeed evaluate the behaviour and memory requirements, decide on the set of required resources and carry out the resource and memory allocations. This will yield the processing unit. The control unit is particularly simple in the case of a computation scheme: indeed, as the number of computation steps is fixed, the control unit reduces to a fixed sequence generator and can for example be designed as the combination of a counter and of a decoder.

The interest of computation schemes lies in the fact that it is possible here to develop formal techniques of parallelism detection, of parallelism increase and of sequentialization. These techniques will be studied in the following sections.

4.2. *The connection graph*

Consider a computation scheme

$$C = \{V_1, V_2, \ldots, V_N\}. \tag{23}$$

To that computation scheme, we associate a labelled graph $\langle \mathbf{V}, \mathbf{E}, \mathbf{W} \rangle$ as follows:

— the vertices of the graph are in one to one correspondence with the variables V_i; the constants and the input variables will account for a single vertex called *data* or *begin* vertex; similarly, the outputs will also be accounted for by a single node called *result* or *end* vertex.

IPR2018-1557
HTC EX1023, Page 231

— the graph contains an edge from V_i to V_j iff the result of the computation of V_i is used as such in the computation of V_j, i.e. iff, in the combinational implementation, there is a physical connection from an output of operator V_i to an input of operator V_j. Formally

$$(V_i, V_j) \in \mathbf{E} < \Leftrightarrow V_j = \Pi_s(\ldots, V_i, \ldots)$$

or

$$V_j = Pr_t(V_i).$$

— To every edge, we associate an integer label denoting the number of bits carried by the corresponding connection.

The obtained graph is called *connection graph.*

Two binary relations are associated with the connection graph: the *precedence relation* and the *immediate precedence relation.* Intuitively, the precedence relation is defined as follows: V_i precedes V_j if the results obtained in the computation of V_i are used, directly or not, during the computation V_j, i.e. if there is a connection from V_i to V_j in the corresponding combinational implementation. In algebraic terms, the precedence relation is the transitive closure of the connection relation. If V_i precedes V_j, we write $V_i < V_j$. We say furthermore that V_i is an *immediate predecessor* of V_j if $V_i < V_j$ and there exists no vertex V_k such that $V_i < V_k < V_j$. The immediate precedence relation, which is the smallest relation having the precedence relation as transitive closure is useful when discussing the resource requirements.

The precedence graph has two important parameters: the *cost* is the number of nodes corresponding to proper computation primitives; the *computation time* is the length of the longest path from the data node to the result node. The cost and the computation time of the precedence graph are global parameters of the graph that can be viewed as measures of the cost and the computation time of the corresponding combinational implementation, at least for homogeneous computation schemes.

4.3. *Acceptable labellings*

We consider a computation scheme:

$$C = \{V_1, V_2, \ldots, V_N\}$$

together with its connection graph $\langle \mathbf{V}, \mathbf{E}, \mathbf{W} \rangle$ and its precedence graph $\langle \mathbf{V}, < \rangle$. An *acceptable labelling* of the precedence graph is a mapping

$$\phi : \mathbf{V} \rightarrow \mathbf{N} \tag{24}$$

of the set $\mathbf{V}$ of vertices into the set $\mathbf{N}$ of natural numbers such that

$$V_i < V_j \Rightarrow \phi(V_i) < \phi(V_j). \tag{25}$$

IPR2018-1557
HTC EX1023, Page 232

There are two ideas associated with the choice of an acceptable labelling:
- the label of a particular node will represent the execution step during which one plans to execute the task corresponding to that particular node;
- once the labelling has been chosen, it becomes easy to determine the number of behaviours of any relevant type required at any step of the algorithm. In particular, it becomes elementary to compute the parameters w_S used in the previous section. For homogeneous computation schemes, there is a single of these parameters, w_c, called *computation width* and representing the maximum number of vertices carrying the same label.

4.4. *Optimization problems*

Consider again the immediate precedence graph of a computation scheme. The length L of this graph is the number of edges on the longest path relating the BEGIN vertex to the END vertex. If n is the number of vertices in the graph and if the BEGIN vertex is always labelled with 0, the target time (i.e. the label carried by the END vertex) can range between L, its minimum value and $n + 1$ (including the END vertex). There is obviously no interest in considering target times larger than $n + 1$.

That observation suffices to show that there are numerous acceptable labellings, each with associated values of W_S. In the present section, we shall discuss the following problem, related to homogeneous computation schemes: *given a target time T, find an acceptable labelling ϕ minimizing the computation width w_c*. Intuitively, this means that we shall try to realize the computation within the given target time and using a minimum number of resources. Clearly enough, this problem is very restricted in scope but the discussion of its solutions will throw some light on the type of questions raised by these optimization problems.

Our optimization problem is closely related to a classical optimization problem known as the *scheduling problem*[15,16]). For unit execution time systems, the scheduling problem can be stated as follows: *given a number of resources, obtain an acceptable labelling minimizing the execution time.*

The scheduling problem is known to be a NP-complete problem, except when the given number of resources is two or when the immediate precedence graph is a tree. In this section, we show that our optimization problem and the scheduling problem are polynomially equivalent, i.e. that their computational complexities at most differ by a polynomial factor.

Assume first given a procedure *optschedule* that, given a precendence graph and an appropriate *number of resources* will provide an acceptable labelling resulting in a *minimal length* execution. Then, the program

IPR2018-1557
HTC EX1023, Page 233

```
program optlabels;
    begin
        number of resources := 0;
        repeat
            number of resources := number of resources + 1;
            optschedule
        until minimal length ⟨= target time
    end.
```

will provide a minimal width acceptable labelling. As the number of resources will remain smaller than n, we obtain, between the execution times $T(optlabels)$ and $T(optschedule)$, the relationship

$$T(optlabels) = O(nT(optschedule)). \tag{26}$$

Similarly, assume given a procedure *optlabels* that, given a precedence graph and an appropriate *target time* will provide an acceptable labelling resulting in a *minimal width* implementation. Then, the program

```
program optschedule;
    begin
        comp time := 0;
        repeat
            comp time := comp time + 1;
            optlabels
        until minimal width ⟨= target width
    end.
```

will result in an optimal schedule. As the target time will never exceed n, we obtain:

$$T(optschedule) = O(nT(optlabels)). \tag{27}$$

The eqs (26) and (27) show that the computational complexities of the two optimization problem differ at most by a factor n^2.

In practice, our optimization problem is solved by the above program optlabels together with one of the available scheduling algorithms. It should be pointed out that, as the scheduling problem is NP-complete, most of these algorithms ultimately rest upon some form of (possibly restricted) enumeration.

5. Conclusions

This paper first discussed the concept of structural interpretation of a high level language program and proposed an algorithm for this purpose. On the one hand, such an attempt is a must if we wish to translate automatically programs

IPR2018-1557
HTC EX1023, Page 234

into hardware structures. As a byproduct, optimization problems can be seen in the now familiar frame of formal program transformations. On the other hand, there certainly exist other choices of the initial language and of the structural interpretation rules: there is no doubt that this discussion is far from being completed. We feel however that the general rules proposed here are sufficiently well supported by practical experience to assert their actual value.

The paper next focused on two optimization problems: the choice of a minimum cost set of resources able to support a given degree of parallelism and the minimization of the number of computation resources within a given target time. These two sample problems showed the role played by formal program transformations, graph theory and complexity theory in digital design. Many other examples could be given to support that assertion.

Clearly, the incentive problem of creating appropriate design tools at the algorithmic or functional level is still in its infancy. We can conjecture that adequate methods will result both of a better "physical" or "intuitive" understanding of the decomposition problems at the system or at the algorithmic level but also of an in depth knowledge and application of adequate algebraic techniques.

Acknowledgement

The author is indebted to his colleague J.-M. Barthelemy and to Mr F. Messemaekers for useful discussions and suggestions.

REFERENCES

[1]) M. Davis, Computability and unsolvability, McGraw Hill, New York, 1958.
[2]) R. Ayres, VLSI, Silicon compilation and the art of automatic microchip design, Prentice Hall, Englewoods Cliffs, 1983.
[3]) C. A. Mead and L. S. Conway, Introduction to VLSI systems, Addison-Wesley, Reading, 1980.
[4]) D. L. Johannsen, Silicon compilation, California Institute of Technology, Dept of Computer Science, Technical report #4530, 1981.
[5]) A. R. Deas, The unit silicon compiler, University of Edinburgh, Dept of Computer Science, Internal Report CSR-145-83, 1983.
[6]) N. J. Elias and A. W. Wetzel, Proc. ACM-IEEE 20-th Design Automation Conference, 46, (1983).
[7]) V. M. Glushkov, Kibernetica **1**, 1 (1965).
[8]) C. C. Clare, Designing logic systems using state machines, McGraw Hill, New York, 1973.
[9]) D. Mange, E. Sanchez and A. Stauffer, Systèmes logiques programmés, Presses Polytechniques Romandes, Lausanne, 1982.
[10]) M. Davio, J. P. Deschamps and A. Thayse, Digital systems with algorithm implementation, J. Wiley, Chichester, 1983.
[11]) F. C. Hennie, Finite state models for logical machines, Wiley, New York, 1968.
[12]) A. Ginzburg, Algebraic theory of automata, Academic Press, New York, 1968.
[13]) J. P. Deschamps, Matérialisation de machines algorithmiques, Ph. D. Dissertation, Ecole Polytechnique Fédérale, Lausanne, 1983.
[14]) C. Y. Hitcock III and D. E. Thomas, Proc. ACM-IEEE 20-th Design Automation Conference, 484 (1983).
[15]) E. Coffman, *Computer and Job-Shop Scheduling*, Wiley Interscience, New York, 1976.
[16]) M. Davio and A. Thayse, Philips J. Res. **34**, 26 (1979).

IPR2018-1557
HTC EX1023, Page 235

Philips J. Res. 39, 226-241, 1984 *R 1093*

APPLICATION OF THE INDEX THEORY OF PSEUDO-LOSSLESS FUNCTIONS TO THE BISTRITZ STABILITY TEST

by P. DELSARTE, Y. GENIN and Y. KAMP

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

This paper contains an investigation of the Bistritz stability test, a new powerful method to count the zeros of a polynomial in the unit disk, with the help of the index theory of pseudo-lossless rational functions. It is shown how an appropriate iterative procedure to compute the index of an arbitrary pseudo-lossless function provides a framework in which the Bistritz test can be derived in a direct way, extended to the case of complex polynomials, and generalized so as to cope with the exceptional situations where the standard version fails to be applicable.

Math. Rev.: 12D10, 30C15

1. Introduction

The problem of counting the zeros of a polynomial $x(z)$ in a given region of the z–plane is standard in mathematics. It arises naturally in numerous applications, due to its obvious significance in stability analysis of physical systems which can be described by rational transfer functions. The two regions of most interest in that respect are the right half-plane ($\mathrm{Re}\, z > 0$) for continuous-time systems and the unit disk ($|z| < 1$) for discrete-time systems. The classical algorithms known respectively as the Routh-Hurwitz test and the Schur-Cohn test provide simple and efficient methods to treat these two cases. The reader is especially referred to Gantmacher [1]), Marden [2]) and Barnett [3]) for a comprehensive discussion of these tests and their various implementations, as well as for the mathematical environment of the general question of locating the zeros of a polynomial.

Due to its long history and to the many results available in the context of the Routh-Hurwitz and Schur-Cohn tests, this subject was considered up to quite recently as somewhat exhausted from a theoretical viewpoint. However, two significant new results in the field have been obtained in the last two years.

The first result is concerned with a new test, due to Bistritz, for counting the zeros of a polynomial $x(z)$ in the unit disk [4,5]). The following comment explains an important specific property of the Bistritz test. Both the Routh-

IPR2018-1557
HTC EX1023, Page 236

Hurwitz test and the Schur-Cohn test are based upon well-defined polynomial recurrence relations. In the first case, the polynomials involved have some amount of structure so that half of their coefficients have actually to be computed, while, in the second case, they have no special structure. This dissymmetry between both situations led Bistritz to a new approach to the unit disk problem. The resulting algorithm[4,5]) involves polynomials with the same amount of structure as in the Routh-Hurwitz test; its complexity is about one half of the complexity of the Schur-Cohn test.

The second result is of a very different nature. It consists in a new general approach to the problem of locating the zeros of a polynomial, based on the index theory of pseudo-lossless rational functions[6]). These functions constitute a subclass of the class of pseudo-positive functions introduced by Oono in classical network theory[7,8]). They can be defined with respect to various regions of the complex plane. In particular, a rational function $F(z)$ is said to be pseudo-lossless in the unit disk if its real part vanishes almost everywhere on the boundary $|z| = 1$, i.e., if one has $F(z) + \bar{F}(1/\bar{z}) = 0$. The index of a pseudo-lossless function $F(z)$ is of direct relevance to the problem considered, since it is precisely defined as the number of zeros in the unit disk of the sum of the numerator and denominator polynomials of $F(z)$. It turns out that the index enjoys a remarkable additivity property with respect to pseudo-lossless decompositions. The interest of approaching the general question of locating the zeros of a polynomial via the method of pseudo-lossless functions is threefold. First, this approach is quite flexible and allows one, for example, to derive the Routh-Hurwitz, Schur-Cohn and Bistritz tests within the same mathematical framework (in the right half-plane formulation as well as in the unit disk formulation). Secondly, the cases of real and complex polynomials can be treated without essential distinction. Thirdly, the method provides the material required to cope with the singular cases of the stability tests. (Informally speaking, these are the cases for which the standard versions fail to apply. Each test has its own singular cases.)

This paper contains an approach to the Bistritz stability test based on the index theory for rational functions that are pseudo-lossless in the unit disk. It turns out that, by such a method, the Bistritz test can be derived in a simple manner, extended to the case of complex polynomials, and generalized so as to accommodate the singular cases.

Section 2 contains some basic definitions and properties of the class of pseudo-lossless rational functions, which are required in the sequel. In particular, it is reminded how a pseudo-lossless component can be associated to any pole in the closed unit disk $|z| \leqslant 1$ of a pseudo-lossless function $F(z)$. Moreover, this component can be extracted from $F(z)$ to produce a residual

IPR2018-1557
HTC EX1023, Page 237

pseudo-lossless function whose index equals the difference between the indices of $F(z)$ and the component considered. Finally, a pseudo-lossless rational function is shown to be the ratio of the reciprocal part over the antireciprocal part of a complex polynomial. The number of zeros of this polynomial in the unit disk appears to be equal to the index of the function thus defined, provided the numerator and the denominator are coprime.

Section 3 describes a novel iterative decomposition procedure for pseudo-lossless rational functions. In its most general form, an iteration loop consists of four successive steps involving complete and partial extractions of pseudo-lossless components relative to two given points z_0 and z_1 of the complex plane, which belong respectively to the unit disk ($|z_0| < 1$) and to the unit circle ($|z_1| = 1$). The pseudo-lossless function obtained at the end of an iteration loop is proved to have a degree which is reduced by two units at least, and an index which is not increased and simply related to the index of the initial function. The algorithm stops when the degree of the pseudo-lossless function considered has been driven to zero; the index of the function is then obtained by concatenating the index relations corresponding to all iteration loops.

In sec. 4, this algorithm is applied to count the zeros in the unit disk $|z| < 1$ of a complex polynomial $x(z)$, or, equivalently, to compute the index of the pseudo-lossless function defined as the ratio of the reciprocal part over the antireciprocal part of $x(z)$. It turns out that in most cases (which are the regular cases) two steps only among the four steps implied in an iteration loop are not trivial. Moreover, if the points z_0 and z_1 are selected as $z_0 = 0$ and $z_1 = 1$, these two steps exhibit exactly the same algebraic form. As a result, a three-term polynomial recurrence relation is brought out, which produces a sequence of reciprocal polynomials of decreasing degree. As a remarkable fact, the index of the pseudo-lossless function considered can be deduced in a simple way from the values assumed by these polynomials at the point $z = 1$. This yields the formulation of the Bistritz stability test in the regular cases.

In sec. 5, those exceptional situations where the iterative procedure of sec. 3 does not reduce to the three-term recurrence relation of sec. 4 are investigated. By relying on the general algorithm to determine the index of an arbitrary pseudo-lossless rational function, one shows that polynomial relations of a more complex form can be substituted for those of the regular case. As a result, a completely general stability test is obtained, which can accommodate any complex polynomial and which spontaneously simplifies into the standard version in the regular case.

Let us finally mention that, at least in the regular case, the Bistritz test can be implemented in a simple tabular form[5]) similar to that of the Routh-Hurwitz test.

IPR2018-1557
HTC EX1023, Page 238

2. Pseudo-lossless rational functions

According to the terminology introduced by Oono [8]), a rational function $F(z)$ with complex coefficients is said to be *pseudo-positive* if it satisfies the inequality

$$\operatorname{Re} F(e^{i\theta}) \geqslant 0, \qquad 0 \leqslant \theta < 2\pi, \tag{1}$$

except when $e^{i\theta}$ is a pole of $F(z)$. A *pseudo-lossless* rational function is a pseudo-positive function for which the constraint (1) is satisfied with equality. Equivalently, a defining property of a pseudo-lossless rational function is the identity

$$F(z) + \bar{F}\left(\frac{1}{\bar{z}}\right) = 0. \tag{2}$$

Of course, a pseudo-lossless function is in general not analytic in the unit disk $|z| < 1$. The class of pseudo-positive functions is partitioned by means of the value of an integer parameter, called the index, which can be interpreted as a certain measure of analyticity in the unit disk. By definition [8,6]), the *index* $k(F)$ of a pseudo-positive rational function $F(z)$ equals the number of poles in $|z| < 1$ of the Cayley transform $S(z)$ of $F(z)$, given by

$$S(z) = \frac{1 - F(z)}{1 + F(z)}. \tag{3}$$

The pseudo-positive and pseudo-lossless functions of index zero can be verified to be the positive and lossless functions in the usual sense of circuit theory [9]).

The index theory of pseudo-lossless rational functions is of direct interest in the problem of counting the zeros of a polynomial in the unit disk $|z| < 1$. Indeed, let $x(z)$ be a complex polynomial of degree n and let $\hat{x}(z) = z^n \bar{x}(1/\bar{z})$ be its reciprocal. The rational function

$$F(z) = \frac{x(z) + \hat{x}(z)}{x(z) - \hat{x}(z)}$$

is pseudo-lossless by construction. Moreover, its index $k(F)$ appears to be precisely defined as the number $N(x)$ counting the zeros of the polynomial $x(z)$ in $|z| < 1$, under the assumption that $x(z)$ and $\hat{x}(z)$ are coprime. (Throughout this paper, the zeros of a polynomial are counted with their multiplicities.)

The index enjoys some remarkable properties of invariance and additive decomposition. A subset of these properties, required for our approach to the Bistritz test, is summarized in the following three theorems.

IPR2018-1557
HTC EX1023, Page 239

Theorem 1

Let $F(z)$ be a pseudo-lossless rational function. The functions $F(z)^{-1}$ and $F(z) + iX$, with X a real constant, are pseudo-lossless of the same degree and index as $F(z)$.

Theorem 2

Let $F_1(z)$ and $F_2(z)$ be pseudo-lossless rational functions without common poles. Then their sum $F(z) = F_1(z) + F_2(z)$ is pseudo-lossless of degree $\deg F = \deg F_1 + \deg F_2$ and of index $k(F) = k(F_1) + k(F_2)$.

Theorem 3

Let $z_1, z_2, \ldots, z_s$ be the distinct poles of a pseudo-lossless rational function $F(z)$ in the closed unit disk $|z| \leqslant 1$, with multiplicities $\mu_1, \mu_2, \ldots, \mu_s$. Then $F(z)$ can be decomposed as a sum of s pseudo-lossless rational functions in the form $F(z) = F_1(z) + F_2(z) + \ldots + F_s(z) + iX$, with X a real constant, where the pseudo-lossless components $F_j(z)$ are given by

$$F_j(z) = \sum_{u=1}^{\mu_j} \left[g_u \left(\frac{1 - \bar{z}_j z}{z - z_j} \right)^u - \bar{g}_u \left(\frac{z - z_j}{1 - \bar{z}_j z} \right)^u \right], \quad \text{for } |z_j| < 1, \quad (4')$$

$$F_j(z) = \sum_{u=1}^{\mu_j} h_u i^{u-1} \left(\frac{z_j + z}{z_j - z} \right)^u, \qquad \text{for } |z_j| = 1, \quad (4'')$$

for appropriate complex constants g_u and real constants h_u. Moreover, the index contributions of these pseudo-lossless components add up to give $k(F) = k(F_1) + k(F_2) + \ldots + k(F_s)$ with

$$k(F_j) = \mu_j, \qquad \text{for } |z_j| < 1,$$

$$k(F_j) = \frac{\mu_j}{2}, \qquad \text{for } |z_j| = 1 \text{ and } \mu_j \text{ even},$$

$$k(F_j) = \frac{\mu_j - \operatorname{sgn} h_{\mu_j}}{2} \quad \text{for } |z_j| = 1 \text{ and } \mu_j \text{ odd}. \qquad (5)$$

The reader is referred to a paper by the authors [6]) and the references therein for a proof of the theorems above and for a general discussion of the index of pseudo-positive rational functions.

3. Theoretical basis of the Bistritz test

The index of a pseudo-lossless rational function can be computed in different ways. The Bistritz test relies on a new iterative procedure to determine

IPR2018-1557
HTC EX1023, Page 240

the index, consisting of extractions of appropriate pseudo-lossless components relative to a given point z_1 on the unit circle ($|z_1| = 1$) and a given point z_0 in the unit disk ($|z_0| < 1$). By contrast with other methods[6]), the extractions involved are not all complete. In spite of this fact, it turns out that the resulting iterative procedure reduces the degree and does not increase the index. In principle, each iteration loop includes four successive steps; but, in most cases, two of them only are not trivial.

Without loss of generality, let us assume the pseudo-lossless rational function $F(z)$ to have a pole at the point $z = z_1$ of the unit circle. The basic idea is to perform extractions of pseudo-lossless components alternatively in z_1 and z_0 as follows. (Recall that z_1 and z_0 are *given* points, with $|z_1| = 1$ and $|z_0| < 1$.) First, a *partial* extraction is performed on $F(z)$ at $z = z_1$ while using the available degree of freedom so as to create a zero of the remaining function at $z = z_0$ (step 2 below). Next, the pseudo-lossless component at $z = z_0$ is extracted from the inverse of this remainder (step 4). The only purpose of the auxiliary steps 1 and 3 is to get around some special cases. A full iteration loop can in fact be described as follows.

Step 1

In case $F(z)$ has a pole at $z = z_0$, the corresponding pseudo-lossless component $F_0(z)$ is first extracted from $F(z)$ to yield

$$F(z) = F_0(z) + G_1(z), \qquad k(F) = k(F_0) + k(G_1), \tag{6}$$

with $F_0(z)$ of the form (4′) for $z_j = z_0$. If z_0 is not a pole of $F(z)$, one simply sets $F_0(z) = 0$, i.e., $F(z) = G_1(z)$. In any case one has the inequalities

$$k(G_1) \leqslant k(F), \qquad \deg G_1 \leqslant \deg F. \tag{7}$$

Step 2

By hypothesis, $F(z)$ has a pole at the point z_1 of the unit circle. Hence $G_1(z)$ has a pole at z_1. Should we extract the pseudo-lossless component $F_1(z)$ of $G_1(z)$ at this point, we would obtain

$$G_1(z) = F_1(z) + H(z), \qquad k(G_1) = k(F_1) + k(H), \tag{8}$$

with $F_1(z)$ of the form (4″) for $z_j = z_1$. Let us however not make this extraction and substitute for it the partial extraction implied by the decomposition

$$G_1(z) = \phi_1(z) + G_2(z)^{-1}, \qquad \text{with} \quad \phi_1(z) = \frac{\alpha z_1 + \bar{\alpha} z}{z_1 - z}, \tag{9}$$

where α is the unique complex number yielding $G_1(z_0) = \phi_1(z_0)$. The function $\phi_1(z)$, although being pseudo-lossless, is in general not equal to $F_1(z)$. Note in particular that the index $k(G_1)$ generally differs from $k(\phi_1) + k(G_2)$ since

IPR2018-1557
HTC EX1023, Page 241

$\phi_1(z)$ and $G_2(z)^{-1}$ have normally a common pole in z_1 so that theorem 2 cannot be applied. Eliminating $G_1(z)$ between (8) and (9), one derives the identity $G_2(z)^{-1} = F_1(z) - \phi_1(z) + H(z)$, whence $k(G_2) = k(F_1 - \phi_1) + k(H)$, in view of theorems 1 and 2, since the pseudo-lossless functions $F_1(z) - \phi_1(z)$ and $H(z)$ have no common poles (by construction). Substituting in the last index relation the value of $k(H)$ deduced from (8) one finally obtains the result

$$k(G_2) = k(G_1) + k(F_1 - \phi_1) - k(F_1). \tag{10}$$

Due to theorem 3, the modulus of the difference $k(F_1 - \phi_1) - k(F_1)$ cannot exceed unity. Hence one has the inequalities

$$k(G_2) \leqslant k(G_1) + 1, \qquad \deg G_2 \leqslant \deg G_1. \tag{11}$$

Step 3

In the case $F_1(z) = (\mathrm{Re}\,\alpha)(z_1 + z)/(z_1 - z)$, the pseudo-lossless function $G_2(z)$ does not vanish for $z = z_1$ and may even have a pole (possibly multiple) at this point. In such an exceptional situation, the pseudo-lossless component $F_1^*(z)$ at z_1 is extracted from $G_2(z)$, together with a suitable imaginary constant iX_1, to yield

$$G_2(z) = F_1^*(z) + iX_1 + G_3(z), \qquad k(G_2) = k(F_1^*) + k(G_3), \tag{12}$$

with $F_1^*(z)$ of the form (4″) for $z_j = z_1$. The constant term iX_1 is chosen so that the pseudo-lossless function $G_3(z)$ has a zero at the point z_1, i.e.,

$$iX_1 = \lim_{z \to z_1} [G_2(z) - F_1^*(z)]. \tag{13}$$

In any other situation we simply set $G_3(z) = G_2(z)$. The following inequalities are clearly satisfied in all cases

$$k(G_3) \leqslant k(G_2), \qquad \deg G_3 \leqslant \deg G_2. \tag{14}$$

Step 4

By construction, the pseudo-lossless function $G_2(z)$ has a pole at the point z_0. Hence so has the function $G_3(z)$. Let us denote by $F_0^*(z)$ the pseudo-lossless component of $G_3(z)$ at z_0. Since the value $F_0^*(z_1)$ is imaginary, theorems 1 and 2 yield the decomposition

$$G_3(z) = F_0^*(z) - F_0^*(z_1) + \mathscr{F}(z)^{-1}, \qquad k(G_3) = k(F_0^*) + k(\mathscr{F}), \tag{15}$$

with $F_0^*(z)$ of the form (4′) for $z_j = z_0$. Due to theorem 3, the following inequalities are then obvious

$$k(\mathscr{F}) \leqslant k(G_3) - 1, \qquad \deg \mathscr{F} \leqslant \deg G_3 - 1. \tag{16}$$

It turns out that the pseudo-lossless rational function $\mathscr{F}(z)$ obtained at the end of step 4 has a pole at the point $z = z_1$. The imaginary constant $F_0^*(z_1)$ has

IPR2018-1557
HTC EX1023, Page 242

indeed been introduced in the construction (15) of $\mathscr{F}(z)$ so as to produce $\mathscr{F}(z_1)^{-1} = G_3(z_1) = 0$. Therefore, the whole procedure above can be iterated by substituting $\mathscr{F}(z)$ for $F(z)$ in step 1. The resulting algorithm reduces the degree and does not increase the index, for each iteration loop, as it is shown by concatenating the relations (7), (11), (14) and (16); indeed, these imply

$$k(\mathscr{F}) \leqslant k(F), \qquad \deg \mathscr{F} \leqslant \deg F - 2. \tag{17}$$

Thus, for a given function $F(z)$ of degree n, the algorithm stops when the degree has been driven to zero after a maximum of $n/2$ or $(n+1)/2$ iteration loops, depending on whether n is even or odd. Note that the last loop is not necessarily complete. In any case, the value of the index $k(F)$ is obtained at the end of the algorithm by adding together the index identities (6), (10), (12), (15) relative to each step. As for the index ingredients $k(F_0)$, $k(F_1)$, $k(F_1 - \phi_1)$, $k(F_1^*)$ and $k(F_0^*)$, they are computed by repeated use of theorem 3.

Although the method just described actually yields the index of any pseudo-lossless rational function, its implementation into an efficient numerical algorithm is certainly not obvious at this point. Its sophistication stems from three different origins. First, provisions have been made in the general procedure to cope with exceptional situations, such as the function $F(z)$ having a pole at z_0 (step 1) or the function $G_2(z)$ not vanishing at z_1 (step 3). Secondly, the actual computation of the extracted pseudo-lossless components can be rather cumbersome (especially for multiple poles). Thirdly, the points z_0 and z_1 can be arbitrarily selected (with $|z_0| < 1$ and $|z_1| = 1$) and could even be modified after each iteration loop.

In the next section it will be shown that the choice $z_0 = 0$ and $z_1 = 1$ leads to a remarkably simple algorithm provided each iteration loop is *regular*, in the sense that the following four conditions are satisfied:

(i) $F_0(z) = 0$. Then $F(z)$ has no pole at $z = 0$ and step 1 is skipped.
(ii) $\deg F_1 = 1$. Then $F(z)$ has a simple pole at $z = 1$ (step 2).
(iii) $F_1(z) \neq (\mathrm{Re}\,\alpha)(1+z)/(1-z)$. Then $G_2(z)$ has a zero at $z = 1$ and step 3 is skipped.
(iv) $\deg F_0^* = 1$. Then $G_3(z)$ has a simple pole at $z = 0$ (step 4).

4. The Bistritz test in the regular case

Let $x(z)$ be any complex polynomial of degree n. As explained in sec. 2, the index $k(F)$ of the pseudo-lossless rational function

$$F(z) = \frac{x(z) + \hat{x}(z)}{x(z) - \hat{x}(z)}$$

equals precisely the number $N(x)$ counting the zeros of $x(z)$ in the unit disk

IPR2018-1557
HTC EX1023, Page 243

$|z| < 1$, under the assumption that $x(z)$ and $\hat{x}(z)$ are coprime. Let us further assume $x(z)$ to be normalized in such a way that the value $x(1)$ is real.

Introducing the reciprocal polynomials

$$y_n(z) = x(z) + \hat{x}(z), \qquad y_{n-1}(z) = \frac{x(z) - \hat{x}(z)}{1 - z}, \tag{18}$$

of degree n and $n-1$, let us use the notation $f_n(z) = y_n(z)/(1-z)y_{n-1}(z)$ instead of $F(z)$ to emphasize the degree of the pseudo-lossless function under consideration. With the simple choice

$$z_0 = 0 \quad \text{and} \quad z_1 = 1, \tag{19}$$

let us then apply the iterative procedure of the preceding section to determine the value $k(f_n) = N(x)$ under the additional assumption that all iteration loops are regular. It turns out that, in such a situation, the successive steps involved in each loop take the following form.

Step 1

As the regularity condition (i) implies $y_{n-1}(0) \neq 0$, this step reduces trivially to $G_1(z) = f_n(z)$.

Step 2

By the regularity condition (ii), the pole of $G_1(z)$ at $z = 1$ is simple, which means $y_{n-1}(1) \neq 0$. Hence the expressions (8) and (9) of $G_1(z)$ have the form

$$f_n(z) = \frac{h_1(1+z)}{1-z} + H(z), \qquad f_n(z) = \frac{\alpha + \bar{\alpha} z}{1 - z} + G_2(z)^{-1}, \tag{20}$$

where $h_1 = y_n(1)/2y_{n-1}(1)$ is nonzero due to the coprimeness assumption. The parameter α is determined by $\alpha = y_n(0)/y_{n-1}(0)$ in order to force a pole at $z = 0$ for $G_2(z)$. In view of (20), the pseudo-lossless function $G_2(z)$ can be written as

$$G_2(z) = -\frac{(1-z)y_{n-1}(z)}{z\, y_{n-2}(z)}, \tag{21}$$

where $y_{n-2}(z)$ denotes the reciprocal polynomial defined via

$$z\, y_{n-2}(z) = (\alpha + \bar{\alpha} z)y_{n-1}(z) - y_n(z). \tag{22}$$

From (10) it follows that the index of $G_2(z)$ is determined by

$$2k(G_2) = 2k(f_n) + \operatorname{sgn} h_1 + \operatorname{sgn}(\operatorname{Re}\alpha - h_1).$$

Finally, in view of the relation $2(\operatorname{Re}\alpha - h_1) = y_{n-2}(1)/y_{n-1}(1)$, the index $k(G_2)$ can be expressed as

$$k(G_2) = k(f_n) + \tfrac{1}{2}\operatorname{sgn}\frac{y_n(1)}{y_{n-1}(1)} + \tfrac{1}{2}\operatorname{sgn}\frac{y_{n-1}(1)}{y_{n-2}(1)}. \tag{23}$$

IPR2018-1557
HTC EX1023, Page 244

Step 3

By regularity condition (iii), the function $G_2(z)$ has a zero at $z = 1$, which means $y_{n-2}(1) \neq 0$ (so that (23) makes sense). This step trivially reduces to $G_3(z) = G_2(z)$.

Step 4

The pseudo-lossless function $G_2(z) = G_3(z)$ has a simple pole at $z = 0$, by regularity condition (iv). Thus $y_{n-2}(0)$ does not vanish. The pseudo-lossless component $F_0^*(z)$ has the form (4') with $z_j = 0$ and $\mu_j = 1$, i.e.,

$$F_0^*(z) = -g_1 z^{-1} + \bar{g}_1 z, \qquad \text{with} \quad g_1 = \frac{y_{n-1}(0)}{y_{n-2}(0)}. \tag{24}$$

In view of (15) and (24), the pseudo-lossless function $\mathscr{F}(z)$ is determined via

$$G_2(z) = -(g_1 + \bar{g}_1 z)(1 - z) z^{-1} + \mathscr{F}(z)^{-1}. \tag{25}$$

Defining the reciprocal polynomial $y_{n-3}(z)$ from the relation

$$z\, y_{n-3}(z) = (g_1 + \bar{g}_1 z)\, y_{n-2}(z) - y_{n-1}(z), \tag{26}$$

and setting $f_{n-2}(z) = \mathscr{F}(z)$, one is led at the end of the loop to the pseudo-lossless function of degree $n - 2$ given by

$$f_{n-2}(z) = \frac{y_{n-2}(z)}{(1 - z)\, y_{n-3}(z)}, \tag{27}$$

in view of (25) and (26). Moreover, the index of $f_{n-2}(z)$ is found via (23) and (24) to have the form

$$k(f_{n-2}) = k(f_n) - 1 + \tfrac{1}{2}\left(\operatorname{sgn} \frac{y_n(1)}{y_{n-1}(1)} + \operatorname{sgn} \frac{y_{n-1}(1)}{y_{n-2}(1)}\right). \tag{28}$$

From the key observation that the relations (22) and (26) determining the polynomials $y_{n-2}(z)$ and $y_{n-3}(z)$ have the same algebraic structure it follows that the iterative procedure to compute $k(f_n)$, whence $N(x)$, reduces to a surprisingly simple form. Let us indeed define the sequence of reciprocal polynomials $y_t(z)$, with $\deg y_t = t$, by the descending recurrence formula

$$z\, y_{t-2}(z) = (\alpha_t + \bar{\alpha}_t z)\, y_{t-1}(z) - y_t(z), \tag{29}$$

for $t = n, n - 1, \ldots, 1$, with the initialization (18) and the convention $y_{-1}(z) = 0$, i.e., $0 = (\alpha_1 + \bar{\alpha}_1 z) y_0(z) - y_1(z)$. Of course, the numbers α_t are given by

$$\alpha_t = \frac{y_t(0)}{y_{t-1}(0)}. \tag{30}$$

IPR2018-1557
HTC EX1023, Page 245

Introduce then the *stability parameters* $\omega_1, \omega_2, \ldots, \omega_n$ by defining

$$\omega_t = \frac{y_t(1)}{y_{t-1}(1)}. \tag{31}$$

One is then led to the following simple test to count the zeros of the polynomial $x(z)$ in the unit disk $|z| < 1$.

Theorem 4 [Bistritz test [4,5])]

Let $x(z)$ be a complex polynomial of degree n, with $x(1)$ real, written in the form $x(z) = [y_n(z) + (1-z)y_{n-1}(z)]/2$ with $y_n(z) = x(z) + \hat{x}(z)$ and $(1-z)y_{n-1}(z) = x(z) - \hat{x}(z)$. If the $n+1$ reciprocal polynomials $y_n(z), y_{n-1}(z), \ldots, y_0(z)$ generated by the recurrence relation (29) all exist and do not vanish at the point $z = 1$, then the number $N(x)$ counting the zeros of $x(z)$ in $|z| < 1$ is equal to the number of negative elements in the associated sequence of stability parameters $(\omega_n, \omega_{n-1}, \ldots, \omega_1)$, given by (31).

Proof

It follows from (29) that the greatest common divisor of $y_n(z)$ and $y_{n-1}(z)$ must be a factor of $y_t(z)$ for all $t \geqslant 0$. Hence, as $y_0(z) = y_0(1) \neq 0$, the polynomials $y_n(z)$ and $y_{n-1}(z)$ are coprime. Due to the assumption $y_n(1) \neq 0$, this implies that $x(z)$ and $\hat{x}(z)$ are coprime. As a result, one has $k(f_n) = N(x)$ by definition of the index.

On the other hand, the assumptions $y_t(0) \neq 0$ and $y_t(1) \neq 0$ for all t are easily verified to be equivalent to the regularity conditions (i)–(iv) for all iteration loops. An iterative use of the index formula (28) then yields the desired expression for $k(f_n)$, namely

$$k(f_n) = \tfrac{1}{2}\left(n - \sum_{t=1}^{n} \operatorname{sgn} \omega_t\right). \tag{32}$$

This is immediate when n is even, and makes use of the obvious result $2k(f_1) = 1 - \operatorname{sgn}\omega_1$, with $f_1(z) = y_1(z)/(1-z)y_0(z)$, when n is odd. □

It turns out that the Bistritz test enjoys a definite advantage, in terms of computational complexity, with respect to the classical Schur-Cohn test for obtaining the same number $N(x)$. As all polynomials $y_t(z)$ produced by the recurrence relation (29) are spontaneously reciprocal, only half of their coefficients have actually to be computed. Thus the Bistritz test requires, roughly speaking, half the number of arithmetic multiplications involved in the Schur-Cohn test.

Let us further observe from (29), (30) and (31) that the stability parameters satisfy the recurrene relation

$$\omega_t + \omega_{t-1}^{-1} = 2 \operatorname{Re} \alpha_t, \tag{33}$$

IPR2018-1557
HTC EX1023, Page 246

for $t = 2, 3, \ldots, n$, with the initialization $\omega_1 = 2 \operatorname{Re} \alpha_1$ (i.e., $\omega_0 = \infty$). Therefore, they can alternatively be computed in terms of the parameters α_t instead of the values $y_t(1)$. In the same line, it is also interesting to note the following algebraic property. From (29) one deduces that the numbers $y_t(1)$ satisfy the system of linear equations

$$\begin{bmatrix} 2 \operatorname{Re} \alpha_1 & -1 & & & \\ -1 & 2 \operatorname{Re} \alpha_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & \\ & & -1 & 2 \operatorname{Re} \alpha_n & \end{bmatrix} \begin{bmatrix} y_0(1) \\ y_1(1) \\ \vdots \\ y_{n-1}(1) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ y_n(1) \end{bmatrix}. \quad (34)$$

Let A denote the tridiagonal matrix of the system (34) and let $\Delta_1, \Delta_2, \ldots, \Delta_n$ be the nested principal minors of A (from the upper corner). These minors are known to obey the recurrence relation

$$\Delta_t = (2 \operatorname{Re} \alpha_t) \Delta_{t-1} - \Delta_{t-2}, \quad (35)$$

with the initialization $\Delta_{-1} = 0, \Delta_0 = 1$. A comparison of (33) and (35) shows that the stability parameter ω_t can be expressed as the ratios

$$\omega_t = \frac{\Delta_t}{\Delta_{t-1}}, \qquad 1 \leqslant t \leqslant n. \quad (36)$$

In view of Sylvester's law of inertia, theorem 4 can therefore be recast as follows in terms of the tridiagonal matrix A of the system (34). The number $N(x)$ counting the zeros of $x(z)$ in the unit disk equals the number of negative eigenvalues of A.

A standard problem in system theory is to check whether a given polynomial $x(z)$ is *stable*, in the sense that it is devoid of zeros in the unit disk $|z| < 1$. Since any nonregular loop in the iterative computation of the index $k(F)$ of a pseudo-lossless rational function $F(z)$ produces a strictly positive contribution to this index, theorem 4 directly yields the following criterion. A polynomial $x(z)$ of degree n, assumed to be coprime with $\hat{x}(z)$, is stable if and only if its stability parameters $\omega_1, \omega_2, \ldots, \omega_n$ are well defined and are all positive. In view of (33) and the obvious identity $\omega_t \omega_{t-2} = \dfrac{y_t(1)}{y_{t-2}(1)}$, it is seen that this criterion can be given the following simplified form.

Corollary 5 [Bistritz stability criterion [4,5]]

A complex polynomial $x(z)$ of degree n, coprime with $\hat{x}(z)$ and taking a real value at the point $z = 1$, is stable if and only if the polynomials $y_t(z)$ produced by the recurrence relation (29) fulfil both conditions

IPR2018-1557
HTC EX1023, Page 247

$$\operatorname{Re}\alpha_t > 0, \qquad \text{for } t = n, n-1, \ldots, 1,$$

$$y_{n-2j}(1)\, y_{n-2j-2}(1) > 0, \qquad \text{for } j = 0, 1, \ldots, \left[\frac{n-2}{2}\right].$$

5. The general Bistritz test, including the singular cases

The simple version of the Bistritz test given by theorem 4 fails to apply whenever the general method involves a *singular* loop, i.e., an iteration loop violating the regularity conditions (i)–(iv). Such a case is characterized by $y_t(0) = 0$ or $y_t(1) = 0$ for a certain value of t, which may cause an interruption in the sequence of polynomials $y_s(z)$. To deal with such singularities, one has to rely on the general form of the iterative procedure to determine the index of an arbitrary pseudo-lossless rational function, as established in sec. 3.

Note first that the case $y_n(0) = 0$, whence $\alpha_n = 0$, is not singular, and that the possibility $y_n(1) = 0$ has to be ruled out due to the assumption of coprimeness between $x(z)$ and $\hat{x}(z)$. In view of the definition of regularity at the end of sec. 3, an iteration loop is singular when at least one of its four successive steps is singular in the sense that it satisfies

$$F_0(z) \neq 0, \ \deg F_1 > 1, F_1(z) = (\operatorname{Re}\alpha)\,\frac{1+z}{1-z}, \text{ or } \deg F_0^* > 1.$$

It turns out that these cases are respectively equivalent to $y_{n-1}(0) = 0$, $y_{n-1}(1) = 0$, $y_{n-2}(1) = 0$ and $y_{n-2}(0) = 0$. In any of them, appropriate substitutes for the recurrence formula (29) will be produced by working out the successive steps of the general loop in a polynomial form.

With $F(z) = f_n(z) = y_n(z)/(1-z)\,y_{n-1}(z)$, let us consider the general situation where $y_{n-1}(0)$ may vanish. Thus we write $y_{n-1}(z) = z^u y_{n-2u-1}(z)$, with $u \geqslant 0$, where $y_{n-2u-1}(z)$ is a reciprocal polynomial, of degree $n-2u-1$, not vanishing at $z = 0$. In view of theorem 3, the pseudo-lossless component $F_0(z)$ extracted in step 1 has the form $F_0(z) = z^{-u}\pi_{2u}(z)$, with $\pi_{2u}(z)$ an antireciprocal polynomial of degree $2u$, i.e., $\pi_{2u}(z) = -\hat{\pi}_{2u}(z)$. Note that the coefficients of $\pi_{2u}(z)$ are obtained from the data by solving the equation

$$y_n(z) \equiv (1-z)\,\pi_{2u}(z)\,y_{n-2u-1}(z) \mod z^u.$$

Setting $p = n - 2u$ and introducing the reciprocal polynomial $y_p(z)$ defined by the relation

$$z^u y_p(z) = y_n(z) - (1-z)\,\pi_{2u}(z)\,y_{p-1}(z), \tag{37}$$

one is led at the end of step 1 to the results

$$G_1(z) = f_p(z) = \frac{y_p(z)}{(1-z)\,y_{p-1}(z)}, \qquad k(f_n) = u + k(f_p). \tag{38}$$

IPR2018-1557
HTC EX1023, Page 248

As one has $y_{p-1}(0) \neq 0$, by construction, step 2 is based on the same relation as in (22). Thus the reciprocal polynomial $y_{p-2}(z)$ is defined via

$$z\, y_{p-2}(z) = (\alpha_p + \bar{\alpha}_p z)\, y_{p-1}(z) - y_p(z), \tag{39}$$

with $\alpha_p = y_p(0)/y_{p-1}(0)$. The function $G_2(z)$ and the index difference $k(G_1) - k(G_2)$ are then determined from y_{p-1} and y_{p-2} as follows. One has

$$G_2(z) = -\frac{(1-z)\, y_{p-1}(z)}{z\, y_{p-2}(z)}, \qquad k(f_p) = k(G_2) + a, \tag{40}$$

where $a = k(F_1) - k(F_1 - \phi_1)$. It turns out that a can only take the values -1, 0 or 1. Three different situations can actually be met.

1°) If $\deg F_1 > 1$, one has $y_{p-1}(1) = 0$, $y_{p-2}(1) \neq 0$ and $a = 0$ (trivially).

2°) If $\deg F_1 = 1$ and $F_1(z) \neq \dfrac{(\operatorname{Re}\alpha_p)(1+z)}{1-z}$, one has $y_{p-1}(1) \neq 0$, $y_{p-2}(1) \neq 0$, whence $2a = -(\operatorname{sgn}\omega_p + \operatorname{sgn}\omega_{p-1})$ as explained in sec. 4.

3°) If $F_1(z) = \dfrac{(\operatorname{Re}\alpha_p)(1+z)}{1-z}$, one has $y_{p-1}(1) \neq 0, y_{p-2}(1) = 0$, whence $2a = 1 - \operatorname{sgn}\omega_p$ in view of theorem 3 and the identity $\omega_p = 2\operatorname{Re}\alpha_p$ resulting from (39).

In the third situation, step 3 of the iteration loop is not trivial. The polynomial $y_{p-2}(z)$ has a zero at $z = 1$ and hence has the form

$$y_{p-2}(z) = (i - i\,z)^v\, y_{p-v-2}(z)$$

with $v \geqslant 1$ and $y_{p-v-2}(1) \neq 0$. The constant i^v has been inserted in the factor $(1-z)^v$ of $y_{p-2}(z)$ so that the polynomial $y_{p-v-2}(z)$ be reciprocal. In view of (40) and $y_{p-1}(1) \neq 0$, the pseudo-lossless component $F_1^*(z) + iX$ extracted from $G_2(z)$ can be written as $F_1^*(z) + iX_1 = \sigma_{v-1}(z)/i^v(1-z)^{v-1}$, where $\sigma_{v-1}(z)$ is the reciprocal polynomial of degree $v-1$ uniquely determined from the equation $y_{p-1}(z) + z\,\sigma_{v-1}(z)\, y_{p-v-2}(z) \equiv 0 \bmod (1-z)^v$. Setting then $q = p - v$ and defining the reciprocal polynomial $y_{q-1}(z)$ of degree $q-1$ via the relation

$$(i - i\,z)^v\, y_{q-1}(z) = z\,\sigma_{v-1}(z)\, y_{q-2}(z) + y_{p-1}(z), \tag{41}$$

one obtains after completion of step 3 the expressions

$$G_3(z) = -\frac{(1-z)\, y_{q-1}(z)}{z\, y_{q-2}(z)}, \qquad k(G_2) = k(G_3) + b, \tag{42}$$

with $b = k(F_1^*)$. In view of (4″), (5) and the equality $\sigma_{v-1}(1) = -y_{p-1}(1)/y_{q-2}(1)$ resulting from (41), the value of b is found to be

IPR2018-1557
HTC EX1023, Page 249

$$b = \tfrac{1}{2}(v - 1) \qquad \text{if } v \text{ is odd,}$$

$$b = \tfrac{1}{2}\left(v - 1 - \operatorname{sgn}\frac{y_{p-1}(1)}{y_{q-2}(1)}\right), \quad \text{if } v \text{ is even.} \tag{43}$$

By construction, $G_3(z)$ has a pole at $z = 0$, of multiplicity larger than one in case $y_{q-2}(0) = 0$. In the general situation let us write $y_{q-2}(z) = z^w y_{q-2w-2}(z)$, with $w \geqslant 0$, where $y_{q-2w-2}(z)$ is a reciprocal polynomial, of degree $q - 2w - 2$, not vanishing at $z = 0$. In view of (42) and theorem 3, the pseudo-lossless component $F_0^*(z) - F_0^*(1)$ extracted from $G_3(z)$ in step 4 takes the form $F_0^*(z) - F_0^*(1) = -(1 - z)\tau_{2w+1}(z)/z^{w+1}$, where $\tau_{2w+1}(z)$ is the unique reciprocal polynomial of degree $2w + 1$ solution of the equation

$$y_{q-1}(z) \equiv \tau_{2w+1}(z)\, y_{q-2w-2}(z) \mod z^{w+1}.$$

With $r = q - 2w$, define the reciprocal polynomial $y_{r-3}(z)$, of degree $r - 3$, via the relation

$$z^{w+1} y_{r-3}(z) = \tau_{2w+1}(z)\, y_{r-2}(z) - y_{q-1}(z). \tag{44}$$

The final outcome of the iteration loop is then found to be

$$\mathscr{F}(z) = f_{r-2}(z) = \frac{y_{r-2}(z)}{(1 - z)\, y_{r-3}(z)}, \qquad k(G_3) = k(f_{r-2}) + w + 1. \tag{45}$$

To sum up, the singular loops of the Bistritz test are characterized by the occurrence of one or several of the situations

$$y_{n-1}(0) = 0,\ y_{n-1}(1) = 0,\ y_{n-2}(1) = 0, \text{ and } y_{n-2}(0) = 0.$$

In such cases, one has to subsitute for the pair of polynomials $y_{n-2}(z), y_{n-3}(z)$ resulting from the standard relation (29) the pair of polynomials $y_{r-2}(z), y_{r-3}(z)$ obtained with the help of the general relations (37), (39), (41) and (44). The index of the updated pseudo-lossless function $f_{r-2}(z) = y_{r-2}(z)/(1 - z)y_{r-3}(z)$ is then related to the index of $f_n(z)$ through the concatenation of the index identities (38), (40), (42) and (45), which are used instead of the standard version (28).

The general polynomial relations above are easily seen to simplify into the standard from (29) both in the regular case and in the singular case

$$y_{n-1}(0) \neq 0,\ y_{n-1}(1) = 0,\ y_{n-2}(1) \neq 0,\ y_{n-2}(0) \neq 0.$$

Note indeed that in the latter situation one has $\omega_n = \infty$, $\omega_{n-1} = 0$ but

$$\omega_n \omega_{n-1} = \frac{y_n(1)}{y_{n-2}(1)} < 0$$

due to (29), which yields $k(f_{n-2}) = k(f_n) - 1$, in agreement with (28); as established from (38), (40) and (45) with $u = w = a = 0$. Let us further observe

IPR2018-1557
HTC EX1023, Page 250

from the identity $n = r + 2u + v + 2w$ that in all other singular cases one has both inequalities $k(f_{r-2}) < k(f_n)$ and $\deg f_{r-2} < \deg f_n - 2$. On the other hand, let us emphasize that the four steps of a given loop cannot be all singular. In particular, the cases $y_{n-1}(1) = 0$ and $y_{n-2}(1) = 0$ are mutually exclusive. Finally, let us note that repeated singular loops are quite possible.

REFERENCES

[1]) F. R. Gantmacher, The theory of matrices, Vol. II, Chelsea, New York, 1959.
[2]) M. Marden, Geometry of polynomials, Amer. Math. Soc., Providence, 1966.
[3]) S. Barnett, Polynomials and linear control systems, Dekker, New York, 1983.
[4]) Y. Bistritz, Proc. Int. Symp. Mathematical Theory of Networks and Systems, Beer-Sheva, Israel, 69 (1983).
[5]) Y. Bistritz, IEEE Trans. Circuits and Systems, **CAS-30**, 917 (1983).
[6]) P. Delsarte, Y. Genin and Y. Kamp, Pseudo-lossless functions with application to the problem of locating the zeros of a polynomial, IEEE Trans. Circuits and Systems (to be published).
[7]) Y. Oono, IECE Japan **57**, no. 10, 12 (1974) and **58**, no. 1, 2, 3 (1975).
[8]) Y. Oono, Proc. Int. Symp. Circuits and Systems, Chicago, 469 (1981).
[9]) V. Belevitch, Classical network theory, Holden Day, San Francisco, 1968.

IPR2018-1557
HTC EX1023, Page 251

Philips J. Res. **39**, 242-253, 1984 R 1094

HELMHOLTZ DECOMPOSITION ON MULTIPLY CONNECTED DOMAINS

by C. DIERIECK and F. CROWET

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

This paper describes the complete characterization of potentials and of stream functions, associated to general two-dimensional vector fields, as encountered in the Helmholtz decomposition problem. We especially focus attention on multiply connected domains. All the characterizations are formulated in a variational form and are thus well suited to finite element approximations.

Math. Rev.: 35A15, 35J20, 35J67, 35Q99

1. Introduction

Recently, the authors obtained results on the *stream function* characterization of *divergence free vector fields*[1]). This includes a new variational characterization and an explicit determination of the constants attained by the stream function on the connected components of the boundary. These results are summarized in the sec. 2. The ideas that led to these results can easily be applied to characterize *potentials* which represent *irrotational vector fields*, in the most general cases. This is considered in sec. 3. Main contribution of this paper concerns the decomposition of arbitrary vector fields on $\mathcal{R}^2$, which have least squares components, and are defined on multiply connected domains. Such *Helmholtz decomposition* of a vector field as a sum of a divergence free vector field and an irrotational vector field, is indeed known since long (see ref. 2). However the applicability of such decompositions as well as precise characterizations in a wide variety of cases, and in particular on multiply connected domains, have been given little attention. In particular we fill this gap and analyse such Helmholtz decompositions of vector fields defined on *multiply connected domains*. We especially focus attention on the variational formulations and on the complete characterization of all possible realizations (sec. 4). Moreover we try to better understand the underlying mechanism of decomposition, by discovering the origins of its multiplicity (sec. 5).

IPR2018-1557
HTC EX1023, Page 252

Throughout the paper we consider a domain Ω which is supposed to be an open bounded subset of $\mathscr{R}^2$, having a Lipschitz-continuous boundary. This domain is also supposed to be multiply connected. The boundary has thus several connected components Γ_i for $i = 0, \ldots, n$, Γ_0 being the outer boundary. Most of further notations are standard (see refs 3 and 4).

2. Stream function representation of divergence free vector fields

We consider here the characterization problem of *divergence free vector fields* in $L^2(\Omega)^2$ (also called solenoidal vector fields). Further, some particularizations are obtained for more regular vector fields. In practical cases, most of these fields are related to problems of hydraulics and rheology[5,6].

For every vector field $\boldsymbol{v}$ in the space

$$W = \{w \in L^2(\Omega)^2;\ \operatorname{div} w = 0,\ \langle \gamma_\nu w, 1\rangle_{\Gamma_i} = 0,\ i = 0, \ldots, n\}, \tag{1}$$

there exists a unique stream function class $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$, such that

$$\boldsymbol{v} = \mathbf{curl}\,\psi = (\partial_2 \psi, -\partial_1 \psi),\ \psi \in \dot{\psi},$$

and this correspondence

$$S : \boldsymbol{v} \in W \to \dot{\psi} \in H^1(\Omega)/\mathscr{R}. \tag{2}$$

is a *Hilbert-space isomorphism*[3]. The main characterization for such associated stream function class is given in the following theorem.

Theorem 1

For every $\boldsymbol{v} \in W$, the stream function class $\dot{\psi} = S\boldsymbol{v}$ is unique solution of the variational problem: $\dot{\psi} \in H^1(\Omega)/\ $:

$$(\mathbf{grad}\,\psi, \mathbf{grad}\,\chi)_{0,\Omega} = (\boldsymbol{v}, \mathbf{curl}\,\chi)_{0,\Omega} \quad \forall \chi \in H^1(\Omega).$$

Proof

We first prove that the variational equation characterizes a unique solution. Obviously the form $(\mathbf{grad}\,.\,, \mathbf{grad}\,.\,)_{0,\Omega}$ belongs to $L(H^1(\Omega)/\mathscr{R}, H^1(\Omega)\mathscr{R}; \mathscr{R})$ and is $H^1(\Omega)/\mathscr{R}$-elliptic, since $|.|_{1,\Omega}$ in $H^1(\Omega)/\mathscr{R}$ is a norm equivalent to $\|.\|_{1,\Omega}$. On the other hand, the form $(\boldsymbol{v}, \mathbf{curl}\,.\,)_{0,\Omega} \in L(H^1(\Omega)/\mathscr{R}, \mathscr{R})$, since

$$|(\boldsymbol{v}, \mathbf{curl}\,\chi)_{0,\Omega}| \leqslant 2\|\boldsymbol{v}\|_{0,\Omega}\|\dot{\chi}\|_{1,\Omega}.$$

Then by the Lax-Milgram theorem we obtain the existence and unicity of the solution. To show that the variational equation is equivalent to state that $\dot{\psi} = S\boldsymbol{v}$ in W, we proceed as follows. For any solution $\psi \in \dot{\psi}$ of the equation, we have $(\mathbf{curl}\,\psi, \mathbf{curl}\,\chi)_{0,\Omega} = (\boldsymbol{v}, \mathbf{curl}\,\chi)_{0,\Omega}$, for $\forall \dot{\chi} \in H^1(\Omega)/\mathscr{R}$. Obviously

IPR2018-1557
HTC EX1023, Page 253

$v - \mathbf{curl}\,\psi \in W^{\perp}$ in $L^2(\Omega)^2$. On the other hand v and $\mathbf{curl}\,\psi$ are in W, and consequently we obtain the desired result. □

In some particular cases, this characterization can be proved, to be equivalent with a *single* boundary value problem. This mainly concerns vector fields in $W \cap H(\mathrm{curl};\Omega)$ and $W \cap H_0(\mathrm{div};\Omega)$.

For vector fields in $W \cap H(\mathrm{curl};\Omega)$, we can formulate a more appropriate characterization. It is important to point out, that only for such vector fields $\mathrm{curl}\,v \in L^2(\Omega)$. The characterization of the stream function class can thus equivalently be stated as a *Neumann problem*. The compatibility condition being always satisfied by a Green formula. We obtain:

Corollary 2

For every $v \in W \cap H(\mathrm{curl};\Omega)$, the following statements are equivalent

(1) $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$ and $\dot{\psi} = Sv$.

(2) $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$:
$(\mathbf{grad}\,\psi, \mathbf{grad}\,\chi)_{0,\Omega} = (\mathrm{curl}\,v, \chi)_{0,\Omega} - \langle \gamma_\tau v, \gamma_0 \chi\rangle_\Gamma \quad \forall \chi \in H^1(\Omega)$.

(3) $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$: $-\Delta\psi = \mathrm{curl}\,v$ in Ω,
$\partial_\nu \psi = -\gamma_\tau v$ on Γ.

For vector fields in $W \cap H_0(\mathrm{div};\Omega) = \{w \in L^2(\Omega)^2; \mathrm{div}\,w = 0, \gamma_\nu w = 0\}$, there exists a unique stream function that vanishes on the outer boundary. Indeed stream functions related to vector fields in $W \cap H_0(\mathrm{div};\Omega)$ are constant on every connected component of the boundary. It turns out to be useful to introduce the following H^1 *geometrical basis functions*, which only depend on geometrical data, and are defined for every $i = 1, \ldots, n$ as:

$$\begin{aligned} \varphi_i \in H^1(\Omega)\colon\ -\Delta\varphi_i &= 0 \quad \text{in } \Omega, \\ \varphi_i &= \delta_{ij} \quad \text{on } \Gamma_j;\ \ j = 0, \ldots, n. \end{aligned} \tag{3}$$

We deduce then the following result, where we obtain an explicit characterization of the constants c_i, which the stream function attains on the connected components of the boundary. This is a new result and replaces known indirect methods as described in ref. 7.

Corollary 3

For every $v \in W \cap H_0(\mathrm{div};\Omega)$ there exists a unique stream function, which vanishes on the outer boundary, and is in

$$\Phi = H_0^1(\Omega) \oplus \mathrm{span}\,\{\varphi_i;\ i = 1, \ldots, n\}. \tag{4}$$

IPR2018-1557
HTC EX1023, Page 254

The following statements are equivalent:

(1) $\psi \in S\boldsymbol{v} \cap \Phi$.

(2) $\psi = \psi_0 + \sum_{i=1}^{n} c_i \varphi_i$ where;

(i) $\psi_0 \in H_0^1(\Omega)$: $-\Delta\psi_0 = \operatorname{curl} \boldsymbol{v}$ in Ω,

(ii) φ_i; $i = 1, \ldots, n$ are defined by (3),

(iii) $(c_1, \ldots, c_n) \in \mathscr{R}^n$:

$$\sum_{i=1}^{n} c_i (\mathbf{grad}\, \varphi_i, \mathbf{grad}\, \varphi_j)_{0,\Omega} = (\boldsymbol{v}, \mathbf{curl}\, \varphi_j)_{0,\Omega} \quad j = 1, \ldots, n.$$

Proof

For every $\boldsymbol{v} \in W \cap H_0(\mathrm{div}; \Omega)$ we can construct $\psi = \psi_0 + \sum_{i=1}^{n} c_i \varphi_i$ according to (2.i–iii). Indeed for such problem there exists a unique solution, and in particular for (iii) since the linear operator is positive definite. By construction we have thus that $\psi = \psi_0 + \sum_{i=1}^{n} c_i \varphi_i$ satisfies $(\boldsymbol{v} - \mathbf{curl}\, \psi, \mathbf{curl}\, \chi)_{0,\Omega} = 0$ for all $\chi \in \Phi$, which is equivalent to state that $\boldsymbol{v} = \mathbf{curl}\, \psi$ in $W \cap H_0(\mathrm{div}; \Omega)$. The converse is deduced from theorem 1 and eq. (4). □

It is interesting to remark that span $\{\varphi_i;\ i = 1, \ldots, n\}$ is the orthogonal complement of $H_0^1(\Omega)$ in Φ, endowed with the norm $\| \mathbf{grad}\, . \|_{0,\Omega}$. We obtain thus a *representation formula* for elements in Φ which relies on its orthogonal direct sum decomposition. This result is an immediate consequence of the variational characterization of the φ_i, defined in eq. (3):

$$(\mathbf{grad}\, \varphi_i, \mathbf{grad}\, \chi)_{0,\Omega} = 0 \qquad \forall \chi \in H_0^1(\Omega). \tag{5}$$

In the light of the two preceding corollaries 2 and 3, we can now obtain a better insight into the mechanism that governs the stream function characterization. It becomes then clear, why the general characterization of theorem 1 is *not* directly linked up with a *single* operator equation. Indeed, for some arbitrary vector fields $\boldsymbol{v} \in W$, the function $\operatorname{curl} \boldsymbol{v}$, is not in general contained in $L^2(\Omega)$. Consequently the stream function cannot be defined by a single Neumann problem. Nevertheless and in order to associate an operator formulation to the characterization of theorem 1, we can first determine an element $\boldsymbol{w}_0 \in W \cap H_0(\mathrm{div}; \Omega)$ such that $\operatorname{curl} \boldsymbol{w}_0 = \operatorname{curl} \boldsymbol{v}$ in $H^{-1}(\Omega)$. At this stage, what is important to remark, is that by this construction $[\boldsymbol{v} - \boldsymbol{w}_0] \in W \cap H(\mathrm{curl}; \Omega)$. Consequently we can then apply the characterization of corollary 2 to the vector $[\boldsymbol{v} - \boldsymbol{w}_0]$ and obtain:

Corollary 4

For every $\boldsymbol{v} \in W$ we have $\psi_0 + \psi_1 \in S\boldsymbol{v}$, where the stream functions are uni-

IPR2018-1557
HTC EX1023, Page 255

quely characterized, up to a constant as follows[8]):

$$\begin{array}{lll} (1) & \psi_0 \in H_0^1(\Omega); & -\Delta\psi_0 = \operatorname{curl} \boldsymbol{v} \quad \text{in } \Omega, \\ (2) & \dot{\psi}_1 \in H^1(\Omega)/\mathscr{R}: & -\Delta\psi_1 = 0 \quad \text{in } \Omega, \\ & & \partial_\nu \psi_1 = -\gamma_\tau(\boldsymbol{v} - \mathbf{curl}\,\psi_0) \quad \text{on } \Gamma. \end{array}$$

The latter result is equivalent to an *orthogonal direct sum decomposition* of W. Indeed denoting

$$W_0 = \{\mathbf{curl}\,\psi;\ \psi \in H_0^1(\Omega)\}, \tag{6}$$

$$W_1 = \{\boldsymbol{v} \in W;\ \operatorname{curl} \boldsymbol{v} = 0 \quad \text{in } \Omega\}, \tag{7}$$

we have that

$$W = W_0 \oplus W_1, \tag{8}$$

where W_1 is the orthogonal complement of W_0 in W, as subspace of $L^2(\Omega)^2$.

3. Characterization of potentials for irrotational vector fields

In this section, we turn now to the characterization of the *potentials* related to *irrotational vector fields*, as are encountered e.g. in the so called potential flows[5]) and also in electric field calculations. It is easy to show that all characterizations for potentials of irrotational vector fields are immediately obtained from the results on the stream function characterizations, reported in the preceding paragraph.

From the preceding paragraph we deduce, by considering a *linear isometry* $l \in O^+(L^2(\Omega)^2)$ defined as: $l\boldsymbol{v} = (-v_2, v_1)$ for $\boldsymbol{v} \in L^2(\Omega)^2$ that for every vector field $\boldsymbol{v}$ in the space:

$$X = \{\boldsymbol{w} \in L^2(\Omega)^2;\ \operatorname{curl} \boldsymbol{w} = 0,\ \langle \gamma_\tau \boldsymbol{w}, 1\rangle_{\Gamma_i} = 0,\ i = 0, \ldots, n\}, \tag{9}$$

there exists a unique class $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$ such that $\boldsymbol{v} = \mathbf{grad}\,\varphi$, for $\varphi \in \dot{\varphi}$. Moreover this correspondence, defined as:

$$P : \boldsymbol{v} \in X \to \dot{\varphi} \in H^1(\Omega)/\mathscr{R} \tag{10}$$

is a *Hilbert space isomorphism*. We deduce from theorem 1, the following characterization:

Theorem 5

For every $\boldsymbol{v} \in X$, there exists a unique class of potentials $\dot{\varphi} = P\boldsymbol{v}$, characterized by the variational problem: $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$:

$$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\chi)_{0,\Omega} = (\boldsymbol{v}, \mathbf{grad}\,\chi)_{0,\Omega} \qquad \forall \chi \in H^1(\Omega).$$

In some particular situations, this characterization can here also be transformed as is done in the preceding section.

IPR2018-1557
HTC EX1023, Page 256

For vector fields in $X \cap H(\mathrm{div};\Omega) = \{w \in X; \mathrm{div}\, w \in L^2(\Omega)\}$, we deduce from corollary 2 an appropriate characterization. We remark that only irrotational vector fields in $X \cap H(\mathrm{div};\Omega)$ can be represented by potentials that are solutions of a *Neumann problem*, the compatibility condition being trivially satisfied.

Corollary 6

For every $v \in X \cap H(\mathrm{div};\Omega)$, the following statements are equivalent:

(1) $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$ and $v = \mathbf{grad}\,\varphi$, $\varphi \in \dot{\varphi}$.

(2) $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$:
$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\chi)_{0,\Omega} = (-\mathrm{div}\, v, \chi)_{0,\Omega} + \langle \gamma_\nu v, \gamma_0 \chi\rangle_\Gamma \quad \forall \chi\, H^1(\Omega)$.

(3) $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$: $-\Delta\varphi = -\mathrm{div}\, v$ in Ω,
$\partial_\nu \varphi = \gamma_\nu v$ on Γ.

For vector fields in $X \cap H_0(\mathrm{curl};\Omega) = \{w \in L^2(\Omega)^2; \mathrm{curl}\, w = 0, \gamma_\tau w = 0\}$, there exists a unique potential which vanishes on the outer boundary. Indeed potentials related to vector fields in $X \cap H_0(\mathrm{curl};\Omega)$ are constant on every connected component of the boundary, and are elements of Φ, defined in eq. (4). From corollary 3, we immediately deduce:

Corollary 7

For every $v \in X \cap H_0(\mathrm{curl};\Omega)$ there exists a unique potential

$$\varphi \in \Phi = H_0^1(\Omega) \oplus \mathrm{span}\,\{\varphi_i;\ i = 1, \ldots, n\}.$$

The following statements are equivalent:

(1) $\varphi \in P v \cap \Phi$.

(2) $\varphi = \varphi_0 + \sum_{i=1}^{n} d_i \varphi_i$, where

(i) $\varphi_0 \in H_0^1(\Omega)$: $-\Delta\varphi_0 = -\mathrm{div}\, v$ in Ω,

(ii) φ_i; $i = 1, \ldots, n$ are defined by (3),

(iii) $(d_1, \ldots, d_n) \in \mathscr{R}^n$:
$\sum_{i=1}^{n} d_i (\mathbf{grad}\,\varphi_i, \mathbf{grad}\,\varphi_j)_{0,\Omega} = (v, \mathbf{grad}\,\varphi_j)_{0,\Omega}$, $j = 1, \ldots, n$.

Note that the constants $\{d_i;\ i = 1, \ldots, n\}$ *exist* and are *uniquely defined*, since the matrix of the linear system (2–ii) is *positive definite*.

By extension of the previous results, we readily deduce the following equivalent characterization of theorem 5, for the class of potentials related to general vector fields in X.

Corollary 8

For every $v \in X$, we have $(\varphi_0 + \varphi_1) \in P v$, where the potentials are uniquely characterized, up to a constant as follows:

IPR2018-1557
HTC EX1023, Page 257

$$
\begin{aligned}
&(1) \qquad \varphi_0 \in H_0^1(\Omega): \quad -\Delta\varphi_0 = -\operatorname{div} \boldsymbol{v} \quad \text{in } \Omega,\\
&(2) \qquad \dot{\varphi}_1 \in H^1(\Omega)/\mathcal{R}: \quad -\Delta\varphi_1 = 0 \quad \text{in } \Omega,\\
&\qquad\qquad\qquad\qquad\quad \partial_\nu\varphi_1 = \gamma_\nu(\boldsymbol{v} - \mathbf{grad}\,\varphi_0) \quad \text{on } \Gamma.
\end{aligned}
$$

Here again it is important to stress an *orthogonal direct sum decomposition* of X, which is implicitly contained in the preceding corollary 8. Indeed if we define:

$$X_0 = \{\mathbf{grad}\,\varphi;\ \varphi \in H_0^1(\Omega)\}, \tag{11}$$

$$X_1 = \{\boldsymbol{v} \in X;\ \operatorname{div} \boldsymbol{v} = 0 \quad \text{in } \Omega\}, \tag{12}$$

we readily obtain that

$$X = X_0 \oplus X_1, \tag{13}$$

since X_1 is the orthogonal complement of X_0 in X, as subspace of $L^2(\Omega)^2$.

4. Representation of vector fields from $L^2(\Omega)^2$

We are now ready to turn to our main topic, namely the *Helmholtz decomposition problem* for general vector fields in $L^2(\Omega)^2$. For some given $\boldsymbol{v} \in L^2(\Omega)^2$ we have to find all $\boldsymbol{g} \in X$ and $\boldsymbol{c} \in W$ such that $\boldsymbol{v} = \boldsymbol{g} + \boldsymbol{c}$. We immediately deduce that such equality, implies $[\boldsymbol{v} - \boldsymbol{g}] \in W$ and $[\boldsymbol{v} - \boldsymbol{c}] \in X$. Let us first recall the well-known fact that any element in $L^2(\Omega)^2$ can be split up into two parts: one in X and the other in W. We have: $L^2(\Omega)^2 = \{\boldsymbol{g} + \boldsymbol{c};\ \boldsymbol{g} \in X, \boldsymbol{c} \in W\}$. Taking into account eqs (8) and (13) we deduce from the preceding theorem a first decomposition of $L^2(\Omega)^2$, which however, is not orthogonal:

$$L^2(\Omega)^2 = X_0 + X_1 + W_0 + W_1.$$

Indeed we only have that: $X_0 \perp (W_0 \oplus W_1)$, and $W_0 \perp (X_0 \oplus X_1)$. In the sequel we are now particularly interested in obtaining all possible decompositions of vector fields in $L^2(\Omega)^2$. What is important to remark first, is that for a general $\boldsymbol{v}$ in $L^2(\Omega)^2$ neither $\gamma_\nu \boldsymbol{v}$ nor $\gamma_\tau \boldsymbol{v}$ are defined. Moreover for a general $\boldsymbol{g} \in X$ and $\boldsymbol{c} \in W$ neither $\gamma_\nu \boldsymbol{g}$ nor $\gamma_\tau \boldsymbol{c}$ exist in $H^{-\frac{1}{2}}(\Gamma)$, only $\gamma_\tau \boldsymbol{g}$ and $\gamma_\nu \boldsymbol{c}$ are well defined in $H^{-\frac{1}{2}}(\Gamma)$. Therefore it is not possible, in general, to prescribe neither $\gamma_\nu \boldsymbol{g}$ nor $\gamma_\tau \boldsymbol{c}$, in order to do the decomposition. However the trace $\gamma_\nu[\boldsymbol{v} - \boldsymbol{g}]$ (resp. $\gamma_\tau[\boldsymbol{v} - \boldsymbol{c}]$) will be well defined in $H^{-\frac{1}{2}}(\Gamma)$ since we require $[\boldsymbol{v} - \boldsymbol{g}] \in H(\operatorname{div}; \Omega)$ (resp. $[\boldsymbol{v} - \boldsymbol{c}] \in H(\operatorname{curl}; \Omega)$). A first type of decomposition can then be obtained by constructing $[\boldsymbol{v} - \boldsymbol{g}]$ or $\boldsymbol{c}$ with a prescribed trace $\gamma_\nu[\boldsymbol{v} - \boldsymbol{g}] = \gamma_\nu \boldsymbol{c} = h$, h being any element of $H^{-\frac{1}{2}}(\Gamma)$ such that $\langle h, 1\rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$. Secondly, another type of realization of the decomposition relies on the following considerations: $[\boldsymbol{v} - \boldsymbol{c}]$ or $\boldsymbol{g}$ has a prescribed trace;

IPR2018-1557
HTC EX1023, Page 258

$\gamma_\tau[v - c] = \gamma_\tau g = p$, where p is any element of $H^{-\frac{1}{2}}(\Gamma)$ which satisfies $\langle p, 1\rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$.

In the sequel we turn to the detailed construction of the separate components of some given vector field in $L^2(\Omega)^2$.

4.1. *General decompositions*

We consider first the decomposition of an arbitrary vector field v in $L^2(\Omega)^2$, where the solenoidal component has a prescribed trace γ_ν.

Theorem 9

For every $v \in L^2(\Omega)^2$, and for every $h \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies

$$\langle h, 1\rangle_{\Gamma_i} = 0, \quad i = 0, \ldots, n,$$

there exists a unique decomposition of v as $g + c$ where $c \in W$, $\gamma_\nu c = h$ and $g \in X$. The characterization of $\dot{\varphi}$ and $\dot{\psi}$ such that: $v = \mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$ in $L^2(\Omega)^2$, is as follows:

(1) $\dot{\varphi} \in H^1(\Omega)/\mathcal{R}$:

$$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\chi)_{0,\Omega} = (v, \mathbf{grad}\,\chi)_{0,\Omega} - \langle h, \gamma_0\chi\rangle_\Gamma, \qquad \forall \chi \in H^1(\Omega), \quad (14)$$

(2) $\dot{\psi} \in H^1(\Omega)/\mathcal{R}$:

$$(\mathbf{grad}\,\psi, \mathbf{grad}\,\chi)_{0,\Omega} = (v, \mathbf{curl}\,\chi)_{0,\Omega} + \langle \gamma_\tau \mathbf{grad}\,\varphi, \gamma_0\chi\rangle_\Gamma, \quad \forall \chi \in H^1(\Omega). \quad (15)$$

Proof

We first determine some irrotational vector field $g \in X$ such that $[v - g] \in W$, with $\gamma_\nu[v - g] = h$, in $H^{-\frac{1}{2}}(\Gamma)$. We can always construct such $g \in X$ uniquely, as $g = \mathbf{grad}(\varphi_0 + \varphi_1)$, by using corollary 8, where: $\varphi_0 \in H_0^1(\Omega)$ and $\dot{\varphi}_1 \in H^1(\Omega)/\mathcal{R}$. The characterization of $\varphi = \varphi_0 + \varphi_1$ can equivalently be replaced by (14), owing to theorem 5. The stream function class associated to $[v - g]$ follows then immediately from theorem 1, such that: $\mathbf{curl}\,\psi = v - \mathbf{grad}(\varphi_0 + \varphi_1)$ in W, where $\dot{\psi}$ is characterized by eq. (15), taking into account that X_0 and W are orthogonal. □

We consider now a second type of decomposition of an arbitrary vector field $v \in L^2(\Omega)^2$, where the irrotational component has a prescribed trace γ_τ.

Theorem 10

For every $v \in L^2(\Omega)^2$, and for every $p \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies

$$\langle p, I\rangle_{\Gamma_i} = 0, \quad i = 0, \ldots, n,$$

there exists a unique decomposition of v as $g + c$ where $c \in W$, $g \in X$ and $\gamma_\tau g = p$. The characterization of $\dot{\varphi}$ and $\dot{\psi}$ such that: $v = \mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$ in $L^2(\Omega)^2$, is

IPR2018-1557
HTC EX1023, Page 259

as follows:

(1) $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$:

$$(\mathbf{grad}\,\psi, \mathbf{grad}\,\chi)_{0,\Omega} = (\boldsymbol{v}, \mathbf{curl}\,\chi)_{0,\Omega} + \langle p, \gamma_0 \chi \rangle_\Gamma, \qquad \forall \chi \in H^1(\Omega), \quad (16)$$

(2) $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$:

$$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\chi)_{0,\Omega} = (\boldsymbol{v}, \mathbf{grad}\,\chi)_{0,\Omega} - \langle \gamma_\nu \mathbf{curl}\,\psi, \gamma_0 \chi \rangle_\Gamma, \quad \forall \chi \in H^1(\Omega). \quad (17)$$

Proof

This decomposition amounts to the construction of some irrotational field $\boldsymbol{c} \in W$ such that $[\boldsymbol{v} - \boldsymbol{c}] \in X$, and $\gamma_\tau[\boldsymbol{v} - \boldsymbol{c}] = p$, in $H^{-\frac{1}{2}}(\Gamma)$. The construction of such element $\boldsymbol{c} \in W$ can be realized using corollary 4. There exists a uniquely defined $\psi_0 \in H_0^1(\Omega)$ and $\dot{\psi}_1 \in H^1(\Omega)/\mathscr{R}$ such that $\boldsymbol{c} = \mathbf{curl}(\psi_0 + \psi_1)$. The stream function $\psi = \psi_0 + \psi_1$ can equivalently be characterized, owing to theorem 1 by eq. (16). The potential class, associated to $[\boldsymbol{v} - \boldsymbol{c}]$ follows then immediately from theorem 5. We have: $\mathbf{grad}\,\varphi = \boldsymbol{v} - \mathbf{curl}(\psi_0 + \psi_1)$ in X, where $\dot{\varphi}$ is characterized by eq. (17), taking into account that W_0 and X are orthogonal. □

4.2. *Decompositions of vector fields of* $H(\mathrm{div}; \Omega)$ *and* $H(\mathrm{curl}; \Omega)$

If $\boldsymbol{v} \in H(\mathrm{div}; \Omega)$ (resp. $\boldsymbol{v} \in H(\mathrm{curl}; \Omega)$) then the trace $\gamma_\nu \boldsymbol{v}$ (resp. $\gamma_\tau \boldsymbol{v}$) is defined in $H^{-\frac{1}{2}}(\Gamma)$, and the potential class $\dot{\varphi}$ (resp. the stream function class $\dot{\psi}$) obtained in the two preceding decompositions, can be equivalently characterized by a *Neumann problem*. In this case, the two previous decompositions (theorems 9 and 10) can be done in a special way. If $\boldsymbol{v} \in H(\mathrm{div}; \Omega)$, the existence of the trace $\gamma_\nu \boldsymbol{v}$ in $H^{-\frac{1}{2}}(\Gamma)$ allows to define $\gamma_\nu \boldsymbol{c}$. Clearly, $\gamma_\nu \boldsymbol{c}$ can also be determined by prescribing the trace of g in $H^{-\frac{1}{2}}(\Gamma)$: $\gamma_\nu g = k$. Thus, for any k, arbitrarily chosen in $H^{-\frac{1}{2}}(\Gamma)$ such that $\langle \gamma_\nu \boldsymbol{v} - k, 1 \rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$, it is possible to realize the decomposition of theorem 9 with h chosen as $\gamma_\nu \boldsymbol{v} - k$. The following theorem states this result.

Theorem 11

For every $\boldsymbol{v} \in H(\mathrm{div}; \Omega)$, and for every $k \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies

$$\langle k, 1 \rangle_{\Gamma_i} = \langle \gamma_\nu \boldsymbol{v}, 1 \rangle_{\Gamma_i}, \quad i = 0, 1, \ldots, n,$$

there exists a unique decomposition of $\boldsymbol{v}$ as $g + \boldsymbol{c}$ where $\boldsymbol{c} \in W$, $g \in X$ and $\gamma_\nu g = k$. The characterization of $\dot{\varphi}$ and $\dot{\psi}$ such that: $\boldsymbol{v} = \mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$ in $L^2(\Omega)^2$, is as follows:

(1) $\dot{\varphi} \in H^1(\Omega)/\mathscr{R}$:

$$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\chi)_{0,\Omega} = (-\mathrm{div}\,\boldsymbol{v}, \chi)_{0,\Omega} + \langle k, \gamma_0 \chi \rangle_\Gamma, \quad \forall \chi \in H^1(\Omega).$$

(2) $\dot{\psi} \in H^1(\Omega)/\mathscr{R}$: defined by eq. (15).

Similarly if $\boldsymbol{v} \in H(\mathrm{curl}; \Omega)$, then $\gamma_\tau \boldsymbol{v}$ exists in $H^{-\frac{1}{2}}(\Gamma)$ and we can consider $\gamma_\tau g$. The knowledge of the trace $\gamma_\tau \boldsymbol{c}$ in $H^{-\frac{1}{2}}(\Gamma)$ can determine $\gamma_\tau g$. Thus for any q,

IPR2018-1557
HTC EX1023, Page 260

arbitrarily chosen in $H^{-\frac{1}{2}}(\Gamma)$ such that $\langle \gamma_\tau v - q, 1\rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$, it is possible to perform the decomposition of theorem 10 with p chosen as $\gamma_\tau v - q$. We deduce the following result.

Theorem 12

For every $v \in H(\text{curl}; \Omega)$, and for every $q \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies

$$\langle q, 1\rangle_{\Gamma_i} = \langle \gamma_\tau v, 1\rangle_{\Gamma_i}, \quad i = 0, \ldots, n,$$

there exists a unique decomposition of v as $g + c$ where $g \in X$, $c \in W$ and $\gamma_\tau c = q$. The characterization of $\dot{\varphi}$ and $\dot{\psi}$ such that: $v = \mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$ in $L^2(\Omega)^2$, is as follows:

(1) $\dot{\psi} \in H^1(\Omega)/\mathcal{R}$:
$(\mathbf{grad}\,\psi, \mathbf{grad}\,\chi)_{0,\Omega} = (\text{curl}\, v, \chi)_{0,\Omega} - \langle q, \gamma_0 \chi\rangle_\Gamma, \quad \forall \chi \in H^1(\Omega)$.

(2) $\dot{\varphi} \in H^1(\Omega)/\mathcal{R}$: defined by eq. (17).

It becomes now clear, that the only arbitrariness in the decomposition of an element $v \in L^2(\Omega)^2$ as $v = g + c$, $g \in X$, $c \in W$, is the *choice of the traces* $\gamma_\nu c$ and $\gamma_\tau g$. Consequently, the theorems 9–12 provide *all* possible ways in realizing the decomposition of v.

5. Interpretation of the multiplicity of decompositions in $L^2(\Omega)^2$

In the preceding sections, we gained insight into the variety of decompositions which are available for general vector fields in $L^2(\Omega)^2$. To better understand the underlying mechanism of decomposition, we introduce a subspace of $L^2(\Omega)^2$ and review some of the basic results in decomposition.

We consider:

$$R = \{v \in L^2(\Omega)^2;\ \text{div}\, v = 0,\ \text{curl}\, v = 0\}, \tag{18}$$

which in general, includes the space $X \cap W$. In particular for a simply connected domain, we have that $R = X \cap W$. The space R is a subspace of $H(\text{div}; \Omega)$ and of $H(\text{curl}; \Omega)$. Consequently for every $v \in R$, the traces $\gamma_\nu v$ and $\gamma_\tau v$ always exist in $H^{-\frac{1}{2}}(\Gamma)$. From theorem 9 we deduce that for every $v \in R$, and for every $h \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies $\langle h, 1\rangle_{\Gamma_i} = 0$ for $i = 0, \ldots, n$, there exists a unique decomposition of v as $\mathbf{grad}\,\varphi_h + \mathbf{curl}\,\psi_h$ where $\gamma_\nu \mathbf{curl}\,\psi_h = h$. Similarly from theorem 10 we deduce that for every $v \in R$, and for every $p \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies $\langle p, 1\rangle_{\Gamma_i} = 0$ for $i = 0, \ldots, n$, there exists a unique decomposition of v as $\mathbf{grad}\,\varphi_p + \mathbf{curl}\,\psi_p$ where $\gamma_\tau \mathbf{grad}\,\varphi_p = p$. Consequently we can also note the following equivalent characterization for R defined in eq. (18):

$$R = \{g + c;\ g \in X_1,\ c \in W_1\}. \tag{19}$$

Moreover, it is important to stress, that whenever the domain Ω is simply con-

IPR2018-1557
HTC EX1023, Page 261

nected, then $R = X_1 \cap W_1$. For a given vector field $\boldsymbol{v} \in L^2(\Omega)^2$ we can now, interpret the decomposition. We first determine two vectors one in X_0 and the other in W_0. This amounts to the construction of $\boldsymbol{g}_0 = \mathbf{grad}\,\varphi_0$ in X_0, where $\varphi_0 \in H_0^1(\Omega)$: $-\Delta\varphi_0 = -\operatorname{div}\boldsymbol{v}$ in Ω, and also of $\boldsymbol{c}_0 = \mathbf{curl}\,\psi_0$ in W_0 where $\psi_0 \in H_0^1(\Omega)$: $-\Delta\psi_0 = \operatorname{curl}\boldsymbol{v}$ in Ω. These two vectors $\boldsymbol{g}_0$ and $\boldsymbol{c}_0$ are uniquely determined for a given $\boldsymbol{v}$. Main fact, which is to be stressed now, is that the vector $[\boldsymbol{v} - \boldsymbol{g}_0 - \boldsymbol{c}_0]$ is in R, and is obviously *uniquely* defined. Moreover it is easy to verify that X_0, W_0 and R are *two by two orthogonal.* We obtain thus the *following canonical decomposition* of $\boldsymbol{v}$ in $L^2(\Omega)^2$: $\boldsymbol{v} = \boldsymbol{g}_0 + \boldsymbol{c}_0 + \boldsymbol{r}$, with $\boldsymbol{g}_0 \in X_0$, $\boldsymbol{c}_0 \in W_0$, $\boldsymbol{r} \in R$. The following theorem is a new and basic result.

Theorem 13

The space $L^2(\Omega)^2$ admits an orthogonal direct sum decomposition:

$$L^2(\Omega)^2 = X_0 \oplus W_0 \oplus R,$$

where X_0, W_0, R are two by two orthogonal subspaces, defined resp. by eqs (11), (6), (18) or (19). If Ω is *simply connected*, then moreover

$$R = X_1 \cap W_1 = X \cap W.$$

As results from the secs 2 and 3, there exists a unique class of potentials representing $\boldsymbol{g}_0$ and also a unique stream function class representing $\boldsymbol{c}_0$. On the contrary, the vector $\boldsymbol{r}$ *uniquely defined in R, has no unique representation as* $\mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$. Indeed as was stressed in the preceding, *we can represent* $\boldsymbol{r} \in R$ *in several ways.* As is shown before, for every $h \in H^{-\frac{1}{2}}(\Gamma)$ satisfying $\langle h, 1\rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$ we can represent $\boldsymbol{r}$ as: $\boldsymbol{r} = \mathbf{grad}\,\varphi_h + \mathbf{curl}\,\psi_h$, where $\gamma_\nu\,\mathbf{curl}\,\psi_h = h$. Consequently, for every h, there exists a unique representation for $\boldsymbol{v} \in L^2(\Omega)^2$ as: $\boldsymbol{v} = \mathbf{grad}\,(\varphi_0 + \varphi_h) + \mathbf{curl}\,(\psi_0 + \psi_h)$, where

$$\gamma_\nu\,\mathbf{curl}\,(\psi_0 + \psi_h) = h.$$

Similarly, for every $p \in H^{-\frac{1}{2}}(\Gamma)$ which satisfies $\langle p, 1\rangle_{\Gamma_i} = 0$, $i = 0, 1, \ldots, n$, we can also represent $\boldsymbol{r}$ uniquely as: $\boldsymbol{r} = \mathbf{grad}\,\varphi_p + \mathbf{curl}\,\psi_p$, $\gamma_\tau\,\mathbf{grad}\,\varphi_p = p$. Thus, for every p, there exists a unique representation for $\boldsymbol{v}$ as

$$\boldsymbol{v} = \mathbf{grad}\,(\varphi_0 + \varphi_p) + \mathbf{curl}\,(\psi_0 + \psi_p),$$

where $\gamma_\tau\,\mathbf{grad}\,(\varphi_0 + \varphi_p) = p$, as was pointed out above.

6. Conclusion

Throughout the paper, we aimed to obtain precise and complete characterizations for the potentials and the stream functions that represent two-dimensional vector fields of $L^2(\Omega)^2$ in most practical situations. About the

IPR2018-1557
HTC EX1023, Page 262

stream function representation of *divergence free vector fields*, elements of eq. (1), the main characterization is formulated in theorem 1. Some variants are obtained in general (corollary 4) and in two particular cases: for vector fields in $W \cap H(\text{curl}; \Omega)$ and $W \cap H_0(\text{div}; \Omega)$, resp. in corollary 2 and 3. The characterization of the potentials associated to *irrotational vector fields*, elements of (9) is obtained in theorem 5, and consists of the basic result, governing such correspondence. In two particular cases, equivalent formulations are obtained, i.e. for vector fields in $X \cap H(\text{div}; \Omega)$ (corollary 6) and $X \cap H_0(\text{curl}; \Omega)$ (corrollary 7).

These results and considerations enabled us to solve completely the *Helmholtz decomposition problem*. Indeed in sec. 4, we considered two basic types of realizations, for a decomposition of arbitrary vector fields in $L^2(\Omega)^2$. They are considered in the theorems 9 and 10. For any vector field $\boldsymbol{v} \in L^2(\Omega)^2$ it is thus possible to realize the decomposition of $\boldsymbol{v}$ as $\mathbf{grad}\,\varphi + \mathbf{curl}\,\psi$, where one of the following conditions is realized: $\gamma_\nu \mathbf{curl}\,\psi = h$ or $\gamma_\tau \mathbf{grad}\,\varphi = p$, with h and p being any function in $H^{-\frac{1}{2}}(\Gamma)$ which satisfies $\langle h, 1\rangle_{\Gamma_i} = 0$, $i = 0, \ldots, n$, and $\langle p, 1\rangle_{\Gamma_i} = 0$, i $= 0, \ldots, n$. In two particular cases, we could realize these two fundamental decompositions in a special way. This concerns vector fields in $H(\text{div}; \Omega)$ (theorem 11) or in $H(\text{curl}; \Omega)$ (theorem 12). A final result is given in theorem 13, where we obtained a canonical decomposition of $L^2(\Omega)^2$. This result governs the multiplicity of decompositions, and moreover allows an interpretation of the mechanism of decomposition (sec. 5).

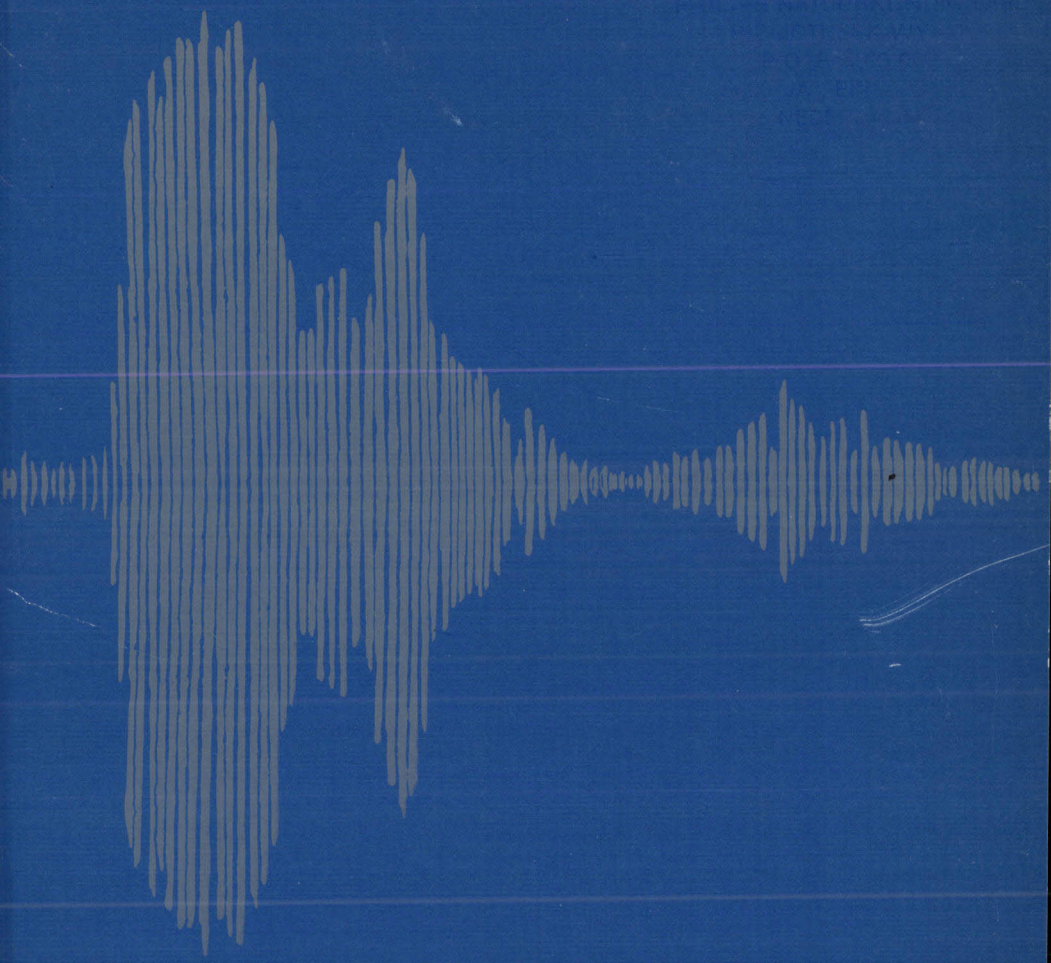
REFERENCES

1) F. Crowet and C. Dierieck, Jl. Mécanique Théorique & Appliquée **2**, 67 (1983).
2) P. M. Morse and H. Feshbach, Methods of theoretical physics, Part I. McGraw Hill, New York, 1953, cf. p. 52-53.
3) V. Girault and P. A. Raviart, Finite element approximation of the Navier-Stokes equations, Springer Verlag, Berlin, 1979.
4) J. L. Lions and E. Magenes, Problèmes aux limites non homogènes et applications (Tome I), Dunod, Paris, 1968.
5) A. Chorin and J. Marsden, A mathematical introduction to fluid mechanics, Springer Verlag, Berlin, 1979.
6) L. D. Landau and E. H. Lifshitz, Fluid mechanics, Pergamon Press, London, 1963.
7) R. Glowinski and O. Pironneau, Siam Rev., **21**, 167 (1979).
8) R. Temam, Navier-Stokes equations. Theory and numerical analysis, North-Holland Publishing Company, Amsterdam, 1977, cf. Ch. I.

IPR2018-1557
HTC EX1023, Page 263

hilips Journal of Research

Vol. 39 No. 6 1984



HILIPS

IPR2018-1557
HTC EX1023, Page 264

Philips Journal of Research

Philips Journal of Research, published by the 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 39 is Dfl. 75, 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.

Editorial Board

A. van Oostrom (General Editor) Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Y. Genin, Brussels, Belgium
R. Kersten, Aachen, Germany
R. Memming, Hamburg, Germany
R. F. Milsom, Redhill, Surrey, U.K.
W. A. Smith, Briarcliff Manor, N.Y., U.S.A.
W. T. Stacy, Sunnyvale, CA., U.S.A.
J. B. Theeten, Limeil-Brévannes, France

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

© Philips International B.V., Eindhoven, The Netherlands, 1984. Articles or illustrations reproduced in whole or in part must be accompanied by a full acknowledgement of the source: Philips Journal of Research.

IPR2018-1557
HTC EX1023, Page 265

Philips J. Res. 39, 255-274, 1984 *R 1095*

FORMATION, THERMAL STABILITY AND PHYSICAL PROPERTIES OF AMORPHOUS 3d-BASED ALLOYS

by K. H. J. BUSCHOW

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

A few technological applications of amorphous alloys are briefly discussed, stressing some unique properties of amorphous alloys. A comparative study is made of the glass forming tendency and the thermal stability in several Ni–base amorphous alloys. New experimental data for crystallization temperatures are given for amorphous alloys in which Ni is combined with Nb, Gd and Th. The occurrence of compositional short range ordering and its effect on the crystallization temperature, activation energy for crystallization, heat of formation and magnetic properties is discussed. New experimental data of the magnetic properties of amorphous $Zr_{1-x}Fe_x$, $La_{1-x}Ni_x$ and $W_{1-x}Co_x$ alloys are given. The interpretation of the magnetic properties of these alloys and of alloys reported in the literature, is made in the light of experimental results obtained by means of various techniques used to study the electronic properties of amorphous alloys.

PACS numbers: 81.50-e, 75.50kj, 71.20.+c.

1. Introduction

Many amorphous alloys have superior properties compared to their crystalline counterparts which make them suitable for various technological applications. These comprise applications as superconducting materials, starting materials for permanent magnets, materials for high frequency transformer cores or recording heads. Thin films of amorphous alloys are applied in various types of high density recording but also in corrosion resistant coatings. By way of introduction a few examples of such applications will be briefly discussed.

The application of amorphous alloys in high frequency devices and transformer cores requires low hysteresis losses. Here one may benefit from the absence of grain boundaries in amorphous alloys which usually can act as pinning centres for Bloch walls. The metastable nature of amorphous alloys offers a further advantage that can be used to improve their performance. It is well known by now that subtle atomic rearrangements can take place in amorphous alloys. These so-called structural relaxations lead to slightly enhanced stabilities and the associated activation energies are low compared to the corresponding activation energies for the amorphous to crystalline transforma-

IPR2018-1557
HTC EX1023, Page 266

tion. Structural rearrangements can be initiated by annealing below the crystallization temperature T_x. The distribution of directions associated with the atomic rearrangements is isotropic in character. But the presence of domain walls and a preferred magnetization direction within the domains induces an anisotropy in the atomic rearrangements which leads to domain wall fixing. Much improved results can be obtained by annealing in a magnetic field which reduces the hysteresis energy losses. Oblique field annealing can be shown to be useful, in particular, since it increases the number of 180° domain walls and hence reduces the losses due to eddy currents [1]).

Thin amorphous films based on rare earths (R) and either Co or Fe (T) are applied in high density magneto-optical recording [2]). The suitability of these materials stems from a high positive uniaxial anisotropy which fixes the magnetization in a direction perpendicular to the film plane. This anisotropy, in combination with the strong ferromagnetic coupling between the T moments and the weaker antiferromagnetic coupling between the R and T moments, is the reason that alloys of an appropriate R/3d ratio exhibit the favourable feature of being magnetically hard at room temperature but magnetically soft at slightly higher temperatures. This feature is used for thermomagnetic writing by means of a laser beam which locally reverses the magnetization direction (see fig. 1). The presence of the positive uniaxial anisotropy is known to depend on the preparatory conditions. In Fe–based materials prepared by co-evaporation of the constituent metals it originates from column shaped

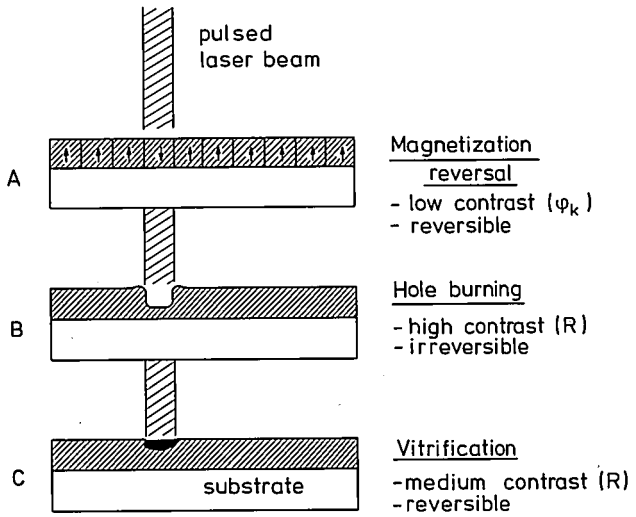


Fig. 1. Schematic representation of different types of high density recording devices based on amorphous alloys (φ_K stands for Kerr effect read-out, R for reflectance difference read-out).

IPR2018-1557
HTC EX1023, Page 267

microstructure. In Co–based materials made by carefully controlled sputtering it is thought to arise from a directional anisotropy in the arrangement of pairs of Co atoms [3]). A more detailed discussion of the origin of the uniaxial anisotropy in rare earth 3d–transition metal films is given in ref. 4. High density recording can be achieved also with other types of amorphous alloys where one uses the difference in optical reflectance or transmittance between the amorphous and crystalline state (see fig. 1).

It will be clear that in order to provide a broad experimental basis for these applications and in order to have a good understanding of the physical processes involved, some fundamental research regarding the stability of amorphous alloys and their electronic and magnetic properties is desirable. In the first place one has to keep in mind that the metastable character of amorphous alloys, apart from the advantages mentioned above, has also a great disadvantage since it makes amorphous alloys go over to the stable crystalline state as a function of temperature and time. A large part of this report will therefore be devoted to thermal stability. A simple model described earlier will be discussed and applied to the many experimental data now available for Ni alloys. This will be followed by a discussion of the occurrence of Compositional Short Range Order (CSRO). Although the occurrence of CSRO strongly affects the stability of the amorphous alloy (making it less metastable) its influence on the thermal stability is only of moderate importance. Finally, it will be shown that the occurrence of CSRO is also of influence when dealing with magnetic properties.

2. Model description

In several previous publications [4,5]) a simple model has been developed in which the crystallization temperature T_x was described in terms of a semi-empirical relationship of the form

$$T_x = 7.5\,\Delta H_h, \tag{1}$$

where T_x is given in K and ΔH_h in kJ per mol of the amorphous alloy $A_{1-x}B_x$. The quantity ΔH_h is the hole formation enthalpy and pertains to a hole of the same size as the smaller type of atom in the amorphous alloy $A_{1-x}B_x$. For a given alloy the hole enthalpy can easily be derived by means of the expression

$$\Delta H_h = \bar{x}\,\Delta H^B_{1V}(1-\bar{x})\left(\frac{V_B}{V_A}\right)^{\frac{5}{6}}\Delta H^A_{1V} \tag{2}$$

using the values of the monovacancy energies ($\Delta H^A_{1V}, \Delta H^B_{1V}$) and the values of the molar volumes (V_A, V_B) listed for almost all metals by Miedema [6]). The

IPR2018-1557
HTC EX1023, Page 268

effective concentrations $\bar{x}_A$ and $\bar{x}_B$ are obtained from the actual concentrations by weighing these latter with the corresponding cross-sectional areas of the A and B atoms. Using the fact that the cross-sectional areas are proportional to $V^{\frac{2}{3}}$ (V is the atomic volume) one has

$$\bar{x}_B = \bar{x} = x V_B^{\frac{2}{3}} [(1 - x) V_A^{\frac{2}{3}} + x V_B^{\frac{2}{3}}]^{-1}; \quad \bar{x}_A = 1 - \bar{x}. \tag{3}$$

Examples in which the experimental T_x values are compared with model calculations are given in the bottom part of fig. 2. The results shown in the top part of the figure stress the fact that the thermal stability of amorphous alloys is not correlated with the heat of compound formation in the corresponding binary system. Experimental results obtained for a relatively large number of different binary Ni–based systems are compared with model calculations in fig. 3.

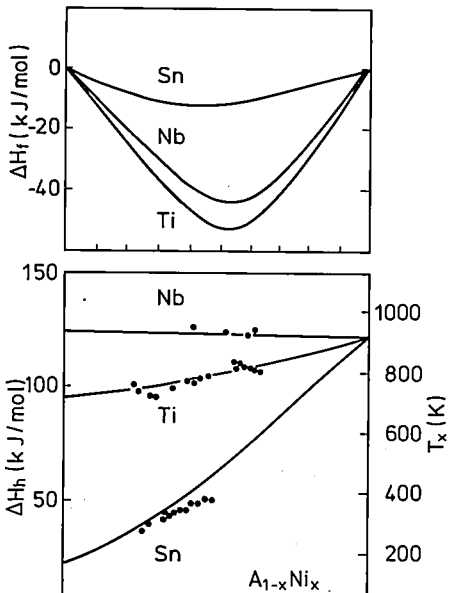


Fig. 2. Bottom part: Comparison of the concentration dependence of the crystallization temperatures (T_x) in various amorphous $A_{1-x}Ni_x$ alloys. The data for A = Ti and Nb were obtained with a heating rate s = 50 K/min on melt spun ribbons. The data for A = Sn were obtained on vapour-deposited alloys with a heating rate s = 2 K/min, kindly provided by Chr. Janot. The full curves represent model calculations made on the basis of eqs (1–3) given in the main text. The scales on the left and right vertical axes correspond to the relation $T_x = 7.5\,\Delta H_h$ where T_x is given in K and ΔH_h is given in kJ per mol alloy $A_{1-x}Ni_x$. Top part: Concentration dependence of the heat of formation in various $A_{1-x}Ni_x$ alloys A = Sn, Nb and Ti derived from the data published by Niessen et al. [12]).

IPR2018-1557
HTC EX1023, Page 269

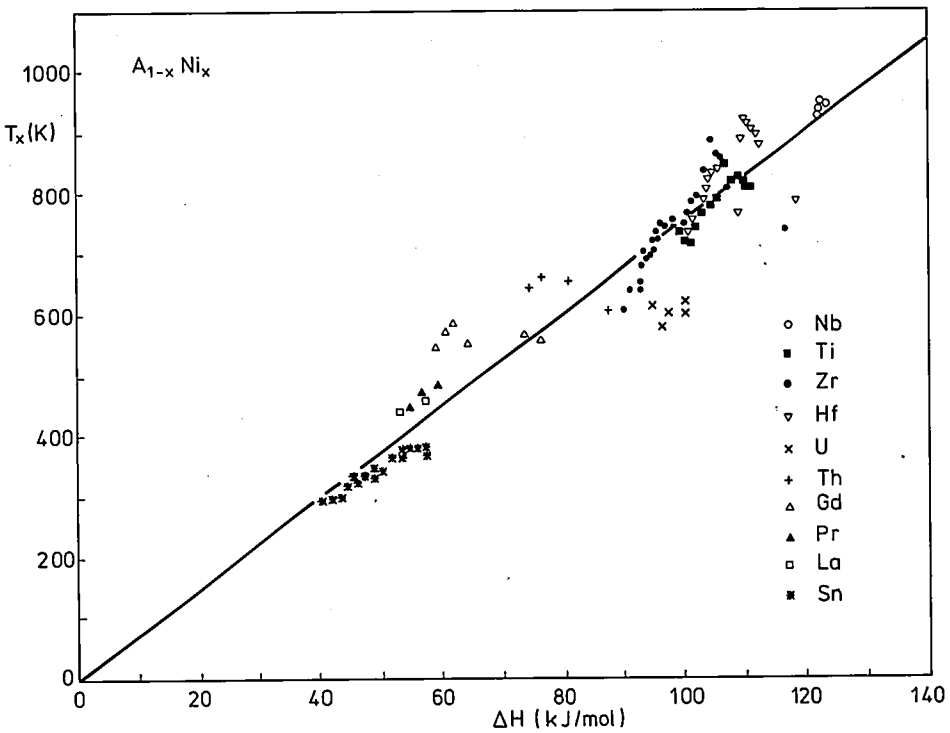


Fig. 3. Comparison of experimental T_x values with calculated hole enthalpies ΔH_h in various $A_{1-x}Ni_x$ alloys. The data were taken from the following sources: A = Nb (this investigation, see table I) A = Ti (ref. 13), A = Zr (ref. 14), A = Hf (ref. 15), A = Th,Gd (present investigation, see table I), A = La,Pr (ref. 5). The data for A = Sn were kindly provided by Chr. Janot.

TABLE I

Crystallization temperatures, determined at a heating rate of 50 K/min, for several Ni–based amorphous alloys considered in fig. 3.

composition	T_x (K)	composition	T_x (K)
$Gd_{1-x}Ni_x$		$Th_{1-x}Ni_x$	
x = 0.28	547	x = 0.25	645
0.29	545	0.30	662
0.30	549	0.40	653
0.32	559	0.55	607
0.34	572	0.70	642
0.36	585	$Nb_{1-x}Ni_x$	
0.40	551	x = 0.42	943
0.52	565	0.52	933
0.56	568	0.60	925
0.58	557	0.61	943

IPR2018-1557
HTC EX1023, Page 270

In general, the agreement between model predictions and experiment is quite satisfactory. Yet there are two unsatisfactory aspects associated with the model description.

(i) The above description of the thermal stability of amorphous alloys is based on a kinetic approach which assumes proportionality between the temperature of incipient crystallization (T_x) and the activation energy ΔE for viscous flow [7,8]). The latter activation energy was taken to be proportional to the hole formation enthalpy ($\Delta E \propto \Delta H_h$). Calorimetric measurements make it possible to obtain experimental values for ΔE so that one can compare the experimental T_x values directly with the corresponding experimental activation energies. When this is done one is faced with the rather peculiar situation that the experimental T_x values are not proportional to the experimental activation energies ΔE. Remarkably enough, they are proportional to the calculated values ΔH_h.

(ii) One might ask the question why the predictions of the model are not affected by the occurrence of CSRO. When the heat of alloying (ΔH_m) or heat of compounding (ΔH_f) is strongly negative one expects the occurrence of CSRO. In such cases the average number of dissimilar neighbours is larger than that expected on the basis of a random distribution of atoms. Moreover, the degree to which CSRO occurs will be the larger, the more negative the corresponding value of ΔH_f [5]). For instance, one expects CSRO in $Ti_{1-x}Ni_x$ to be substantially higher than in $Sn_{1-x}Ni_x$ (see top part of fig. 3), although the T_x values in both systems behave in satisfactory agreement with the predictions of the model.

3. Heat of formation and CSRO

Experimental evidence for the presence of compositional short range ordering in different types of amorphous alloys was obtained on several occasions [9,10]). In this report we will discuss some results of calorimetric measurements. The evidence for CSRO is an indirect one but the discussion of the calorimetric results offers the advantage that a connection can be made with the relative stability of crystalline and amorphous alloys.

The values of the formation enthalpy ΔH_f for a number of Zr–Ni compounds have been plotted as a function of composition (filled circles) in fig. 4. These values were obtained by Henaff et al. [11]) by means of solution calorimetry. The open squares in fig. 4 represent the formation enthalpies of amorphous $Zr_{1-x}Ni_x$ alloys of various concentrations. These values were derived from the heat of crystallization ΔH_{cr} and from the heat of formation of the two-phase mixtures ($\overline{\Delta H_f}$) in which the amorphous alloys had crystallized. Since in a given binary system the heat of formation in the two-phase region

IPR2018-1557
HTC EX1023, Page 271

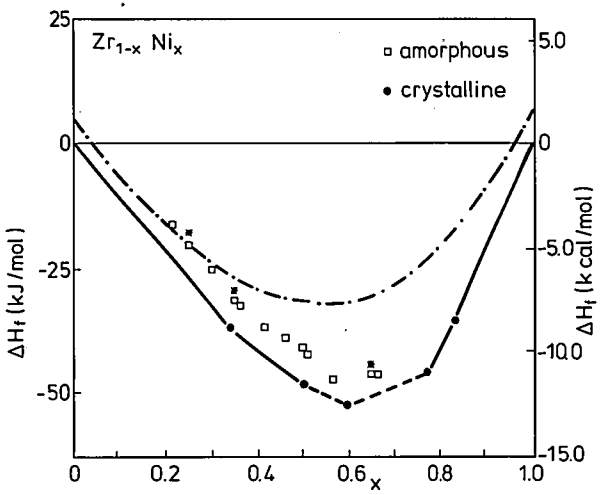


Fig. 4. Heat of formation (ΔH_f) of several Zr–Ni compound obtained by solution calorimetry (filled circles, ref. 11). The lines connecting the data points represent the enthalpies of the corresponding two-phase mixtures ($\overline{\Delta H_f}$). The open squares and stars represent values for the heat of formation obtained for various amorphous alloys (see text).

varies linearly between the ΔH_f values of the two adjacent compounds one may obtain these $\overline{\Delta H_f}$ values relatively easily by interpolation (see fig. 4). For three amorphous alloys Henaff et al. determined the ΔH_f values of the amorphous alloys also by direct calorimetric measurements. These values are presented by stars in fig. 4. There is satisfactory agreement between the two sets of ΔH_f values for the amorphous alloys.

The dependence of ΔH_f on concentration for binary alloys $A_{1-x}B_x$ can be represented by means of the expression [12])

$$\Delta H_f = 2\,F(x)\,[(1-x)V_A^{\frac{2}{3}} + xV_B^{\frac{2}{3}}]\,C, \qquad (4)$$

where C is a constant for a given choice of A and B components. The function $F(x)$ is equal to $F_C(x) = \bar{x}_A\bar{x}_B[1 + 8(\bar{x}_A\bar{x}_B)^2]$ in the case of compound formation. For regular solid solutions and amorphous alloys it reduces to $F_A(x) = \bar{x}_A\bar{x}_B$. For the crystalline compounds, eq. (4) in combination with $F_C(x)$ gives a smooth line passing through the data points in fig. 4 (filled circles). For the sake of clarity this line was not included in the figure since it does not differ much from the lines connecting the data points. For the amorphous alloys eq. (4) in combination with $F_A(x)$ gives the broken curve in fig. 4 after this curve has been shifted in the positive direction by 5 kJ/mol. This

IPR2018-1557
HTC EX1023, Page 272

shift may be somewhat arbitrary. It takes account of the fact that the concentration dependence of ΔH_f for the amorphous alloys should be given with respect to amorphous Zr and Ni rather than with respect to the crystalline metals Zr and Ni. This enthalpy difference was estimated to be of the order of 5 kJ/mol [11]). Inspection of the results in fig. 4 makes it clear that the enthalpy values found for the amorphous alloys tend to be close to the values expected for crystalline materials. This is the case in particular for alloys of a concentration near $x = 0.6$, where the difference from the shifted solid solution curve (broken line) is quite distinct and indicates the occurrence of appreciable short range ordering. This can also be seen by direct comparison of the experimental values of the ratios $\Delta H_f(\text{cryst}) - \Delta H_f(\text{amorph})/\Delta H_f(\text{cryst}) = \Delta H_{cr}/\overline{\Delta H_f}$. It follows that the enthalpies of the amorphous state are about 30% lower than those of the crystalline state near $x = 0.2$. But they are only less than 10% lower near $x = 0.6$.

4. Glass forming ability, thermal stability and their relation to binary phase diagrams

Three different Ni–based binary systems are compared in fig. 5. The experimental data for the crystallization temperatures are given at the top. Experimental values of the activation energies, derived from rate-dependent measurements, are given in the middle. More details regarding these experimental data were presented elsewhere [13–15]). General features of the phase diagrams (in arbitrary units) are shown at the bottom [16]). Note that in order to assess possible relationships between T_x, ΔE and features of the phase diagram fully we have used complete scales rather than truncated scales to represent the concentration dependences of T_x and ΔE.

It has been generally accepted that regions of so-called easy glass formation are situated around the concentrations corresponding to deep eutectics, where the depression of the melting point from the ideal solid solutions liquidus is relatively high (see, for instance, ref. 17). The three $A_{1-x}Ni_x$ systems considered in fig. 5 seem to be no exception in this respect. By contrast, the results given at the top of fig. 5 show that the crystallization temperatures (T_x) are rather insensitive to the details of the phase diagram. The temperatures T_x generally tend to increase with Ni content, which is in accordance with the predictions of the model (broken line). In all three systems one finds *no* enhancement of T_x in the region of the deepest eutectic temperatures. This means that one cannot speak of a correlation between the thermal stability (as measured by T_x) and the glass-forming ability. This result can also be expressed by saying that alloys lending themselves to easy vitrification do not have better thermal stabilities than alloys that are less prone to vitrification.

IPR2018-1557
HTC EX1023, Page 273

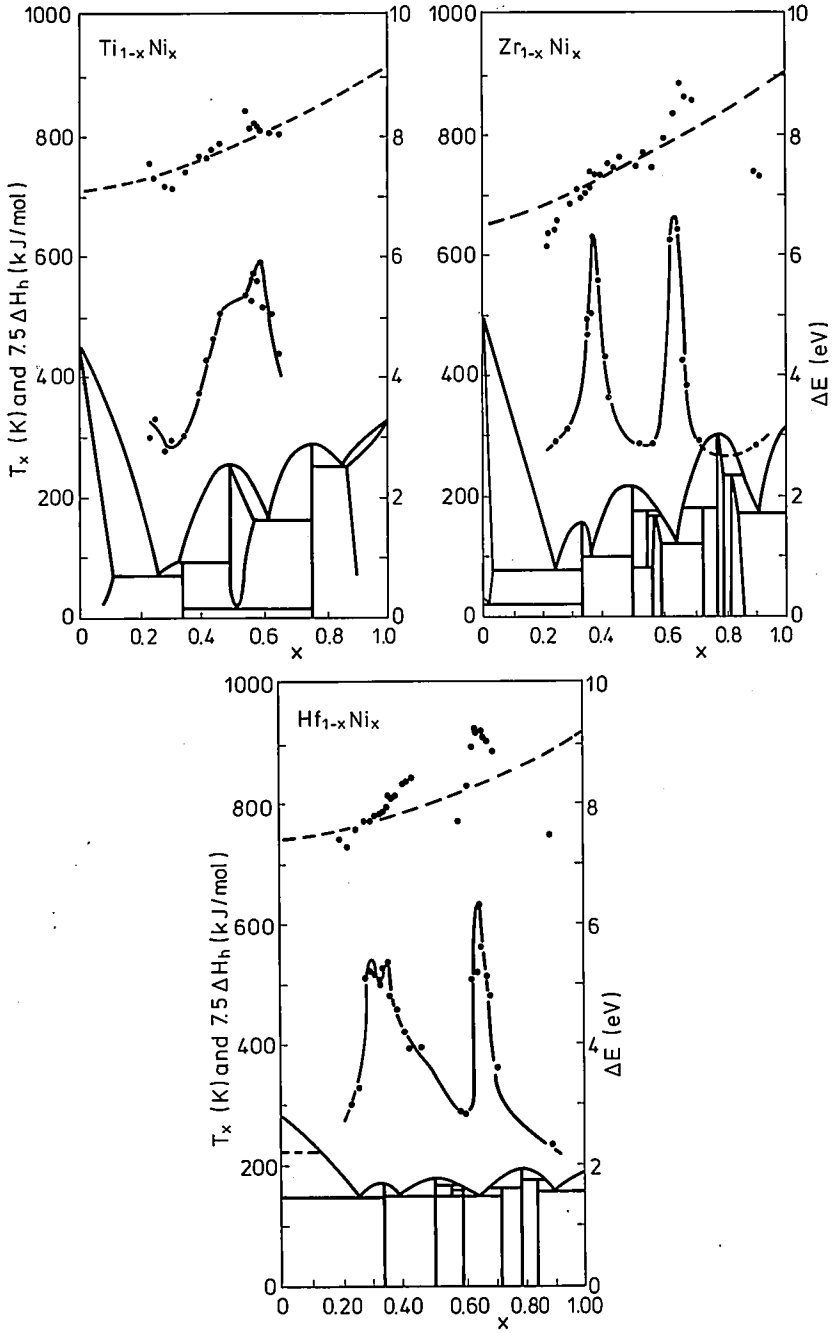


Fig. 5. Comparison of the crystallization temperature T_x (top), activation energies ΔE (middle) and phase relationships (bottom) in three $A_{1-x}Ni_x$ systems. For more details see refs 13-15.

IPR2018-1557
HTC EX1023, Page 274

The quantities T_x and ΔE both reflect the thermal stability of amorphous alloys. It is noteworthy that the trends shown by the concentration dependences of ΔE are mimicked only to a very small extent by the concentration dependences of the corresponding T_x values. The latter values give rise to only moderate excursions (about 10%) from the model predictions while the activation energies sometimes vary by more than 100%. It has already been mentioned above that according to the kinetic approach T_x and ΔE are proportional to each other [7-8]). This is a consequence of the diffusion constant being proportional to the reciprocal of the viscosity ($\eta = \eta_0 \exp \Delta E/S_c T$). Viscous flow and subsequent crystallization become possible when a critical value ($\eta_{cr} \approx 10^{13}$ P) is reached, leading to the relation

$$T_x \propto \frac{\Delta E}{S_c}. \tag{5}$$

In deriving this expression it is usually assumed that the configurational entropy S_c is a temperature-independent constant since crystallization takes place in temperature ranges well below melting temperatures. In previous analyses [4-5]) the possibility of the occurrence of compositional short-range ordering was not taken into consideration. The large negative values of the enthalpy of alloying for all three systems favours the occurrence of CSRO, and some of the experimental evidence for it has been discussed in the previous section. When CSRO does occur S_c may no longer be regarded as temperature-independent. In the same manner as found in atomic order-disorder transformations in crystalline alloys, one would expect the configurational entropy to give rise to reversible temperature variations in particular in alloys with CSRO. Irreversible changes in amorphous alloys were studied by Chen [7]), who showed that the associated temperature variation of S_c leads to apparent values of the activation energy (ΔE) that are larger than the actual values (ΔE_0) by a factor $[1 + (\mathrm{d}(\ln S_c))/(\mathrm{d}(\ln T))]$. The same applies to reversible changes of S_c. This means that even in the absence of irreversible changes, an enhancement of the ΔE values may be expected in alloys with CSRO. There is ample experimental evidence for the occurrence of reversible as well as irreversible changes in a number of different amorphous alloys [18-19]). It can be concluded from the results discussed above that the linear relationship between T_x and the experimental values of ΔE may break down in alloy systems exhibiting CSRO.

From the concentration dependence of ΔE shown for the three systems in fig. 5 one may infer that the degree of CSRO is comparatively high in the middle concentration regions. When amorphous alloys are no longer considered as having atomic arrangements of basically random distributions, alloys

IPR2018-1557
HTC EX1023, Page 275

of a different degree of CSRO will also have different values for S_c. From eq. (5) it then follows that in the middle concentration ranges mentioned above one would expect S_c to be comparatively low. As a result, the corresponding T_x values would be comparatively high. The effect of a lower S_c value may be counteracted, however, by a low activation energy ΔE_o. Here one has to consider that CSRO leads to a larger number of dissimilar neighbour atoms than would have been present in a completely random alloy and that the motion of Ni atoms in a surrounding rich in Ti, Zr or Hf is easier than in a Ni-rich surrounding. The effect of CSRO is therefore to lower ΔE_o. Owing to the mutually compensating influence of both effects one may expect that the occurrence of CSRO will not much affect the predictions of the hole model. If one disregards possible differences in the crystallization temperature associated with differences in the crystallization models of the various alloys, and ascribes the variations around the calculated concentration dependences (broken lines) exclusively to CSRO, one finds that these variations are of the order of 10% in the cases of Zr–Ni and Hf–Ni and only 5% in the case of Ti–Ni. These relatively small variations are to be compared with variations of more than 100% in the case of ΔE (see the middle parts of fig. 5). Of course, one has to keep in mind that the extreme sensitivity of ΔE towards CSRO is the result of the fact that ΔE does not probe S_c itself but rather its logarithmic temperature derivative.

5. Electronic properties

Information regarding the electronic properties of amorphous alloys can be obtained by means of the same techniques currently used for crystalline materials. These techniques comprise photoemission and Mössbauer spectroscopy, NMR, ESR and measurements of the magnetic susceptibility and the low-temperature specific heat. The random nature of the atomic arrangements in amorphous alloys leads to a certain degree line of broadening which is particularly pronounced in NMR experiments and the data analysis requires special care in cases where local density of states are probed by means of Knight shift measurements. In ESR experiments on Gd-doped amorphous alloys the linewidths (ΔH) range from about 150 G in amorphous $Pd_{1-x}Si_x$ alloys to more than 1000 G in amorphous Zr–based alloys [20–21]). However, in all these cases, experiments are sufficiently accurate to separate the residual linewidths (ΔH_o) from the Korringa rates ($d(\Delta H)/dT$). Based on the latter data values of the density of d–band states were obtained. Results obtained for various Zr–based amorphous alloys have been listed in table II [20]). There is satisfactory agreement of these data with those obtained by means of specific heat measurements made on several of similar Zr–based amorphous alloys [22–23]). All these

IPR2018-1557
HTC EX1023, Page 276

TABLE II

Experimental values of the Korringa rate (in GK^{-1}) in various amorphous alloys and the corresponding value of the d-electron density of states (in eV^{-1} per atom per spin).

$Zr_{1-x}T_x$	$d(\Delta H)/dT$	$N_d(E_F)$
$Zr_{0.72}Cu_{0.28}$	36.2	0.90
$Zr_{0.50}Cu_{0.50}$	25.5	0.73
$Zr_{0.44}Cu_{0.56}$	27.0	0.76
$Zr_{0.38}Cu_{0.62}$	22.8	0.70
$Zr_{0.67}Ni_{0.33}$	39.1	0.94
$Zr_{0.67}Pd_{0.33}$	30.1	0.82
$Zr_{0.78}Pt_{0.22}$	37.2	0.92
$Zr_{0.67}Co_{0.33}$	30.0	0.82
$Zr_{0.74}Rh_{0.26}$	35.4	0.90

experimental data have in common that they show that the d–electron density of states in amorphous Zr–rich alloys is larger than in crystalline Zr metal itself. It is interesting to compare these results with bandstructure calculations and results obtained by means of photoemission spectroscopy [24]). Upon alloying the d–bands of Zr and the late transition metal T do not coalesce into a common band. Both d–bands tend to become narrower and the repulsion between the d–bands leads to a shift to higher and lower energies for the d–states of Zr and T, respectively. Furthermore it follows from the bandstructure calculations that the partial d–density of states of Zr has a minimum slightly above E_F and increases below E_F. Owing to the d–band repulsion one would expect therefore that as a consequence of the Zr d–band shift towards higher energies the Fermi level would move into a region of a higher d–density of states. This is exactly what the data in table 2 and the specific heat data suggest, since they indicate an increase in $Nd(E_F)$ in the amorphous alloys relative to pure Zr metal.

Useful information regarding the bonding in amorphous alloys can often be obtained from results of Mössbauer spectroscopy. The Mössbauer isomer shift is a direct measure of the electronic charge density at the nuclear site of a given Mössbauer isotope and its changes observed upon alloying can provide valuable information regarding the concomitant changes in electronic properties. A systematic analysis of the Isomer Shift (IS) was first given by Miedema and van der Woude and applied to Au–alloys and compounds [25]). Recently their model was also applied to amorphous Fe–alloys [26]). If $\delta(IS)_{max}$ is the IS

IPR2018-1557
HTC EX1023, Page 277

(relative to α-Fe) in a dilute system where each Fe atom is surrounded by A atoms only, the formula used in the data analysis can be represented by:

$$\delta(\mathrm{IS})_{\max} = 0.75\,\Delta\phi^* - 1.65\,\frac{\Delta n_{\mathrm{WS}}}{n_{\mathrm{WS}}^{\mathrm{Fe}}}. \qquad (6)$$

Values for the electronegativity differences $\Delta\phi^* = \phi_A^* - \phi_{\mathrm{Fe}}^*$ and the electron density difference $\Delta n_{\mathrm{WS}} = n_{\mathrm{WS}}^A - n_{\mathrm{WS}}^{\mathrm{Fe}}$ can be obtained by means of the data tabulated for all metals by Niessen et al.[12]).

In practice the strainfree dilute limit $\delta(\mathrm{IS})_{\max}$ is reached when $\bar{x} \to 0$, i.e. it can be determined by means of measurements on various metals A containing Fe impurities. In dilute crystalline alloys these Fe impurities occupy crystallographic position of the A atoms in the host lattice which entails a size mismatch. This size mismatch can introduce a considerable error in the determination of $\delta(\mathrm{IS})_{\max}$. In this respect amorphous alloys represent suitable materials since there is virtually no difference between A and Fe sites. Although amorphous alloys with very low Fe concentrations cannot be prepared, the values of $\delta(\mathrm{IS})_{\max}$ can easily be obtained via measurements on less dilute alloys by extrapolation to $\bar{x} \to 0$ where $\bar{x}$ is an effective concentration (see eq. (3)) which measures the fractional area of contact of Fe atoms with other Fe atom neighbours. In the extrapolations mentioned use is made of the linear dependence of the IS on the effective concentration of the A component in $A_{1-x}\mathrm{Fe}_x$

$$\delta(\mathrm{IS})_{\max} = (1 - \bar{x})\,\delta(\mathrm{IS})_{\max}. \qquad (7)$$

The advantages of this type of analysis can be summarized as follows:

(i) It applies to a large variety of different amorphous alloys. This is illustrated by means of fig. 6 where the experimental values[26]) of $\delta(\mathrm{IS})_{\max}$ and values tabulated for ϕ^* and n_{WS} have been used to plot $\delta(\mathrm{IS})_{\max}/\Delta\phi^*$ versus $\Delta n_{\mathrm{WS}}\,[n_{\mathrm{WS}}^A \Delta\phi^*]^{-1}$. The slope and the intercept on the vertical axis correspond to the coefficients of the two terms on the right hand side of eq. (6).

(ii) Experimental values of IS can be broken down in two contributions comprising the interatomic charge transfer term $\mathrm{C.T} = 0.75\,\Delta\phi^*$ and the intraatomic d–s electron conversion term $\mathrm{E.C} = 1.65\,\Delta n_{\mathrm{WS}}/n_{\mathrm{WS}}^{\mathrm{Fe}}$. Using simple scaling laws these contributions can easily be translated into electron numbers[26]).

As an example, let us consider the situation in amorphous $Zr_{1-x}Fe_x$. It can be inferred from the results shown in fig. 7 that the critical concentration for the occurrence of a moment on the Fe atoms in $Zr_{1-x}Fe_x$ is just below $x = 0.4$. This corresponds to $\bar{x} = 0.3$. It can be derived from eqs (6) and (7) that at this concentration the charge transfer from Zr to Fe equals 0.5 electrons, the s–d conversion being even smaller (less than 0.2 electrons per Fe atom). It follows

IPR2018-1557
HTC EX1023, Page 278

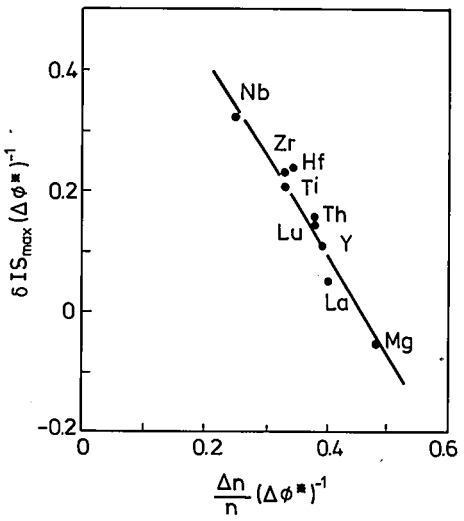


Fig. 6. Analysis of the experimental values of the strain-free dilute limit of the ^{57}Fe isomer shift in various amorphous alloys in terms of eq. (6) (see text). Data were taken from ref. 26.

from these results that the reduction in Fe moment from 2.2 μ_B in Fe metal to zero in $Zr_{0.6}Fe_{0.4}$ is not the result of a filling up of the 3d band due to charge transfer from Zr to Fe.

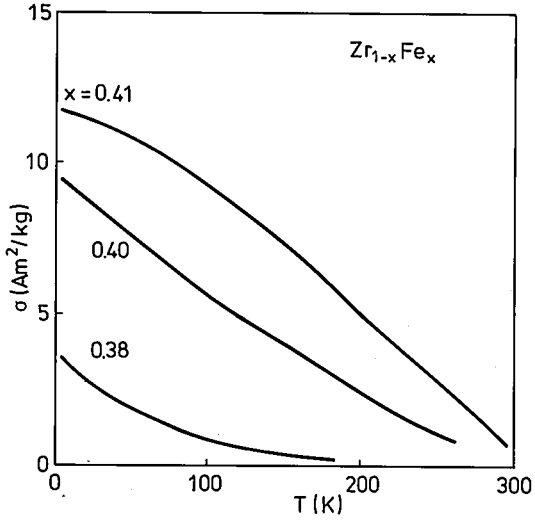


Fig. 7. Temperature dependence of several amorphous $Zr_{1-x}Fe_x$ alloys made by melt spinning. The measurements were made in an applied field of 240 kA/m, using an adaption of the Faraday method.

IPR2018-1557
HTC EX1023, Page 279

It is interesting to compare the results of the ^{57}Fe isomer shift analysis with those obtained from photoemission experiments [24]). This comparison can be made by means of the data listed in table III for the compound Zr_3Fe. Like in amorphous alloys the fractional areas of contact of an Fe atom with Zr and Fe neighbours differ from the actual atomic fractions x in $Zr_{1-x}Fe_x$. In inter-

TABLE III

Comparison of the experimental values determined from XPS core level shifts and ^{57}Fe Mössbauer spectroscopy for the changes of the number of s,p-electrons ($\Delta n_{s,p}$) and d-electrons (Δn_d) when Fe is compounded with Zr in Zr_3Fe.

Method	$\Delta n_{s,p}$(el./Fe)	Δn_d(el./Fe)
XPS	+0.4	+0.5
^{57}Fe(IS)	+0.4	+0.3

metallics like Zr_3Fe, which have a strong heat of formation, there is a tendency of the Fe atoms to have exclusively Fe nearest neighbours. It can be shown that owing to this feature and owing to the relatively large size of the Zr atoms one has a situation in Zr_3Fe which corresponds virtually to the dilute limit ($\bar{x} = 0$). From the analysis of the Mössbauer data in terms of eqs (6) and (7) one derives an increase in s,p–electrons due to charge transfer equal to $C.T \approx 0.7$ electrons per Fe, and a decrease in s,p–electrons due to s–d conversion equal to $E.C \approx 0.3$ electrons per Fe. This means that the effective increase in s,p–electrons is equal to $\Delta n_{s,p} \approx 0.4$ el./Fe, while $\Delta n_d \approx 0.3$. These numbers are in good agreement with those derived from an analysis of the XPS core level data. The results shown in table III make one point quite clear. There is only a charge transfer of a few tenths of an electron from Zr to Fe even in the case of high Zr concentrations. Although the experimental results of XPS and IS pertain to Fe–based alloys, there is evidence that the situation is much the same in other alloys of 3d–metals with more electropositive metals such as the rare earths and actinides [4,27–29]).

6. Magnetic properties

The large atomic disorder and the concomitant absence of lattice periodicity in amorphous solids cause these latter materials to have magnetic properties which are different from those found in their crystalline counterparts. Results for several amorphous $La_{1-x}Ni_x$ alloys are shown in fig. 8. It is seen that the transition from the paramagnetic to the ferromagnetic state is rather sluggish.

IPR2018-1557
HTC EX1023, Page 280

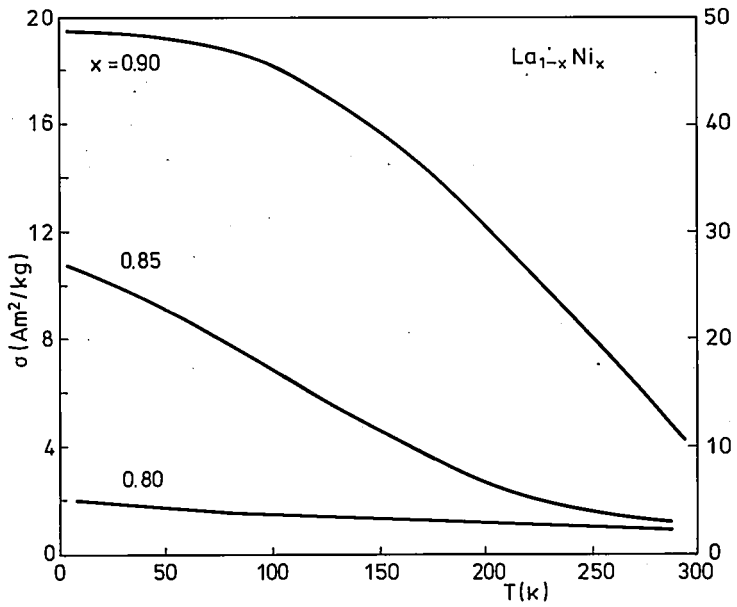


Fig. 8. Temperature dependence of the magnetization (measured at $H = 240$ kA/m) of three amorphous $La_{1-x}Ni_x$ alloys made by co-evaporation of La and Ni.

The flattening of the temperature dependence of the magnetization in random systems was discussed extensively elsewhere [30]).

The formation of amorphous alloys in which 3d–metals are combined with more electropositive metals leads generally to a reduction in the 3d–atom moment. Examples of experimental results shown for Ni, Co and Fe alloys are discussed in refs 4, 26 and 29–32. Whereas the Ni– and Co–based amorphous alloys have a concentration dependence of the 3d–moment that extrapolates to the moment of the pure crystalline 3d–metals ($\bar{x} \rightarrow 1$) one often finds a different situation in amorphous Fe–based alloys. This is also the case in amorphous $Hf_{1-x}Fe_x$ where the extrapolation would lead to a value significantly below the value $2.2\,\mu_B$/Fe found in α–Fe [31]). Most likely this result is explained in terms of a distribution of exchange interactions between the 3d–moments [33]). In the case of Fe–based alloys one may expect a substantial portion of negative exchange interactions to be present which causes deviations from colinear ferromagnetic ordering and entails a relatively too small net saturation magnetization when determined from bulk magnetization measurements. In comparing the moment reduction in amorphous Fe–based alloys it is therefore desirable not to rely too heavily on the results of bulk magnetic measurements but to use the results derived from ^{57}Fe hyperfine field measurements instead.

IPR2018-1557
HTC EX1023, Page 281

It has already been mentioned briefly in sec. 5 that a description of the moment reduction in various amorphous 3d–based alloys is often given in terms of a charge transfer model. Upon alloying, electrons are transferred from the less electronegative A component to the 3d–component where they fill up the 3d–band and thus lower the 3d–moment. It is misleading that the charge transfer model, when applied to a certain class of alloys, is able to account reasonably well for the experimental data. The model is, however, less satisfactory when applied to alloys differing widely in components and composition. One of the reasons for this is probably the presence of still another effect which overrules the effect of charge transfer on the magnetic properties, which will be discussed below. Another reason is that the amount of the charge transferred from the nonmagnetic component A to the transition metal T in $A_{1-x}T_x$ is actually rather small, i.e. it is of the order of a few tenths of an electron (see sec. 5). In descriptions based on charge transfer the electron numbers needed are much higher. For instance Heiman and Kazama had to assume a charge transfer equal to 2.21 electrons per Zr atom in $Zr_{1-x}Co_x$ in order to give an adequate description of the corresponding concentration dependence of the 3d–moment [34]). In view of the experimental evidence presented at the end of sec. 5 one may conclude therefore that charge transfer does occur to some degree but that its magnitude is too small to explain the moment reduction. For this reason we will focus our attention on another mechanism that may be employed for describing moment variations in 3d–metal systems.

Based on the ideas that the degree of atomic short range order is not the same in all amorphous alloys a model has been proposed for describing the differences in magnetic properties of various amorphous 3d–based alloys [26,31]): In order to take account of the size differences that may arise between A and T in $A_{1-x}T_x$, only alloys of similar effective concentration $\bar{x}$ were compared. For a fixed 3d–element these alloys are expected to have the same average T moment when CSRO is absent. Depending on the nature and the degree of CSRO substantially different values for the average T moment will arise. Qualitatively, the nature and the degree of CSRO was estimated by the sign and magnitude of the heat of alloying or heat of compounding (ΔH) [4]). Negative values of ΔH in amorphous alloys $A_{1-x}T_x$ will lead to average atomic arrangements in which the number of dissimilar nearest neighbours is larger than would be the case in a statistical atomic distribution. In these cases the average T moment is relatively low. The reverse arguments apply if ΔH is positive. By using the sign and magnitude of ΔH as a measure of, respectively, the nature and degree of CSRO it has been possible to analyse successfully the magnetic properties of various ferromagnetic amorphous alloys in which a

IPR2018-1557
HTC EX1023, Page 282

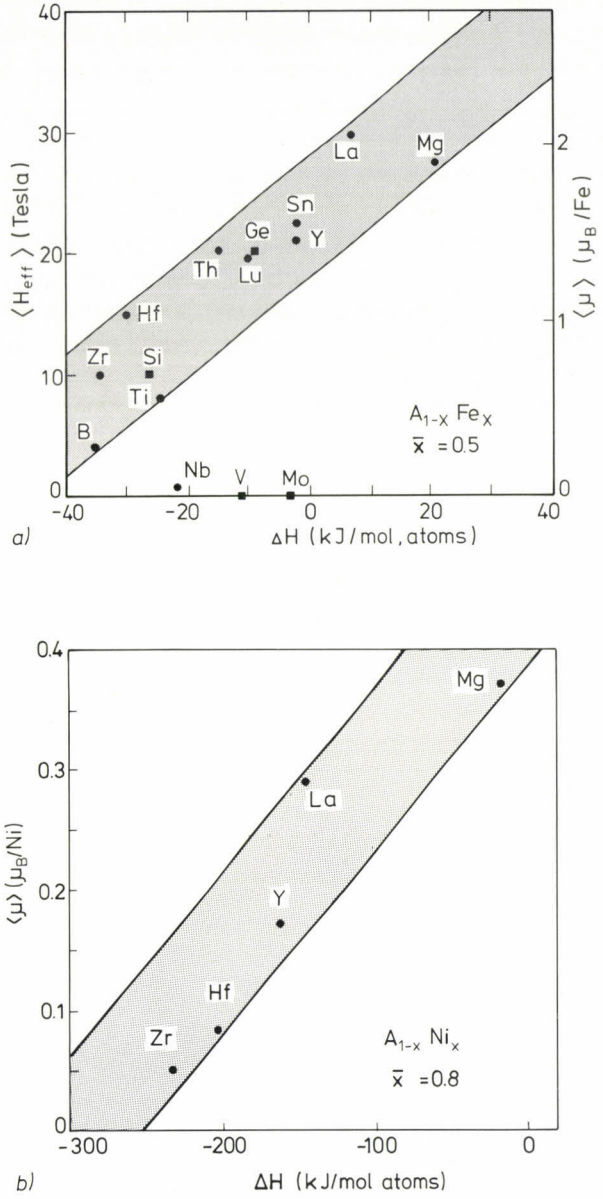


Fig. 9. *a*) Average effective hyperfine field $|\Delta H_{\text{eff}}|$ (full circles, left scale) and average Fe moment (full squares, right scale) in amorphous *A*–Fe alloys with effective Fe concentration $\bar{x} = 0.5$ versus heat of formation. *b*) Average Ni moment in various amorphous Ni alloys ($\bar{x} = 0.80$) versus heat of alloying. Data were reproduced from refs 26 and 31.

IPR2018-1557
HTC EX1023, Page 283

3d–metal is combined with a nonmagnetic partner element. Results for Fe and Ni alloys are shown in fig. 9, where we used the data of Niessen et al. [12]) for obtaining ΔH. All the alloys in the same figure have the same effective concentration (i.e. the 3d–atoms would have a similar fractional area of contact with other 3d–neighbours in all these alloys if CSRO were absent). The degree of CSRO, in which the 3d–atoms are surrounded by more than an average number of A neighbours, is expected to increase in going from Mg to Zr. As a consequence the average 3d–moment of the alloys decreases.

In this model the effect of CSRO on the magnetic properties is overemphasized and no account has been taken of the fact that there may be differences in the moment reducing power between the various A components in alloys of the same effective concentrations $\bar{x}$. In fact, the descriptions based on CSRO and the Friedel model discussed elsewhere [31]) may be regarded as being supplementary to each other. The alloys with T = Nb, V and Mo in fig. 9*a* represent examples where the moment reducing power is expected to be particularly high. This is probably the reason why they fall outside the broad band containing the other alloys.

7. Concluding remarks

In this report the effect of CSRO on various physical properties of amorphous 3d–based alloys has been discussed. Calorimetric measurements have shown that the absolute value of the formation enthalpy of amorphous alloys may be smaller than that of the crystalline state by about 30% when CSRO is small or absent. In the presence of CSRO this enthalpy difference can become much smaller (less than 10%).

The transformation temperature of the amorphous-to-crystalline transition defining the resistance against crystallization of amorphous alloys is not much affected by CSRO. By contrast, the (apparent) activation energies for crystallization obtained from rate-dependent measurements may change appreciably when CSRO is present. It was also found that amorphous alloys of a composition corresponding to the lowest eutectic composition in a binary system, where the depression of the melting point from the ideal solution liquidus and the glass-forming ability are highest, are not necessarily characterized by a high thermal stability.

The effect of CSRO in 3d–based alloys is to change the fractional number of nearest 3d–atom neighbours. When the degree of CSRO is taken to increase with increasing heat of alloying one finds a correlation between the average 3d–moment and the correspondent heat of alloying.

IPR2018-1557
HTC EX1023, Page 284

REFERENCES

[1]) H. Fujimori, H. Yoshimoto, T. Masumoto and T. Mitea, J. Appl. Phys. **52**, 1893 (1981).
[2]) M. Hartmann, J. Braat and B. J. Jacobs, IEEE Trans. Magn. **MAG-12**, 1013 (1984).
[3]) H. K. Lachowicz, IEEE Trans. Magn. **MAG-12**, 1417 (1984).
[4]) K. H. J. Buschow, 'The physics and chemistry of rare earths', in: K. A. Gschneidner and L. Eyring (eds), Vol. 7, North-Holland Physics Publishing, Amsterdam, 1984.
[5]) K. H. J. Buschow, Solid State Comm. **43**, 171 (1982).
[6]) A. R. Miedema, Z. Metallk. **70**, 345 (1979).
[7]) H. S. Chen, Appl. Phys. Lett. **29**, 12 (1976).
[8]) H. A. Davies, Phys. Chem. Glasses **17**, 159 (1976).
[9]) M. Sakata, N. Cowlam and H. Davies, Rapidly quenched metals IV, in: T. Masumoto and K. Suzuki (eds), Vol. 1, p. 327, The Japan Inst. of Metals 1982.
[10]) T. Fukunaga, K. Kai, M. Naka, N. Watanabe and K. Suzuki, Rapidly quenched metals IV, in: T. Masumoto and K. Suzuki (eds), Vol. 1, p. 347, The Japan Inst. of Metals 1982.
[11]) M. P. Henaff, C. Colinet, A. Pasturel and K. H. J. Buschow, J. Appl. Phys. **56**, 307 (1984).
[12]) A. K. Niessen, F. R. de Boer, R. Boom, P. F. de Châtel, W. C. M. Mattens and A. R. Miedema, Calphad **7**, 51 (1983).
[13]) K. H. J. Buschow, J. Appl. Phys. **56**, 304 (1984).
[14]) K. H. J. Buschow, J. Phys. F **14**, 593 (1984).
[15]) K. H. J. Buschow, J. Non-Cryst. Solids (in press).
[16]) C. J. Smithells, Metals reference book, Fifth edition, Butterworths, London 1976.
[17]) R. W. Cahn, Contemp. Phys. **21**, 43 (1980).
[18]) E. Balanzat and J. Hillairet, J. Phys. F **21**, 2907 (1982).
[19]) M. G. Scott, Scripta Met. **15**, 1073 (1981).
[20]) H. J. Eifert, B. Elschner and K. H. J. Buschow, Phys. Rev. B **25**, 7441 (1982); B **29**, 2905 (1984).
[21]) K. H. J. Buschow, H. J. Eifert and B. Elschner, Phys. Stat. Sol. (b) **115**, 455 (1983).
[22]) K. Samwer, PhD thesis University of Göttingen, 1981 (unpublished).
[23]) J. E. Gräbner, B. Golding, R. J. Schutz, F. S. L. Hsu and H. S. Chen, Phys. Rev. Lett. **39**, 1480 (1977).
[24]) J. Kübler, K. H. Bennemann, R. Lapka, F. Rösel, P. Oelhafen and H. J. Güntherodt, Phys. Rev. B **23**, 5176 (1981).
[25]) A. R. Miedema and F. van der Woude, Physica B **100**, 145 (1980).
[26]) A. M. van der Kraan and K. H. J. Buschow, IEEE Trans. Magn. **MAG-12**, 1284 (1984).
[27]) J. A. Zoulay and L. Ley, Solid State Comm. **31**, 131 (1979).
[28]) M. Cyrot and M. Lavagna, J. Appl. Phys. **50**, 2333 (1979).
[29]) J. M. D. Coey, D. Givord, A. Lienard and J. P. Rebouillat, J. Phys. F **11**, 2707 (1981).
[30]) K. Handrich, Phys. Stat. Sol. **32**, K55 (1969).
[31]) K. H. J. Buschow, J. Physique (1984, Proc. MRS Conf. Strasbourg 1984).
[32]) K. H. J. Buschow, J. Magn. Magn. Mater. **28**, 20 (1982).
[33]) J. M. D. Coey, J. Appl. Phys. **49**, 1646 (1978).
[34]) N. Heiman and N. Kazama, Phys. Rev. B **17**, 2215 (1978).

IPR2018-1557
HTC EX1023, Page 285

Philips J. Res. 39, 275-292, 1984 R 1096

CONTRIBUTION TO THE SYSTEM Mg–Au–Hg

by J. L. C. DAAMS and J. H. N. VAN VUCHT

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Reported are supercells for the so-called Na_3As–type compounds Mg_3Hg and Mg_3Au, the unit cell dimensions of a rhombohedral compound of supposedly the composition Mg_5Hg_2, isomorphous to Na_5Hg_2, the atomic structures of Mg_2Hg, isostructural to Co_2Si, and Mg_2Au which is a superstructure of the Co_2Si cell. Also reported are some crystallographic dimension data on ternary alloys of Mg, Hg and Au of 50 at.% Mg and more.

PACS numbers: 61.55. Hg.

1. Introduction

In the course of a search for metallic materials suitable for mercury-vapour dispensers we obtained data that may contribute to the knowledge of the ternary system Mg–Au–Hg and its binary boundary systems Mg–Hg and Au–Mg. Only the magnesium-rich part of the system was studied.

2. Experimental

Samples of the alloys were made by weighing calculated amounts of the constituent elements in thick-walled molybdenum crucibles. These were covered with a molybdenum lid in an atmosphere of about 0.1 atm of pure argon and then closed by welding in an argon-arc furnace. Subsequently, the crucibles were encapsulated in silica envelopes for the heat treatments.

We used gold of Drijfhout (50 ppm Ag, 10 ppm Fe, 7 ppm Cu, 6 ppm Pt, 0.7 ppm Mg), magnesium of New Metals (70 ppm Fe, 40 ppm Si, 20 ppm Al, 3 ppm B) and double-distilled mercury.

IPR2018-1557
HTC EX1023, Page 286

After ample time of reacting at a temperature of mostly 800 °C for a start, and a subsequent annealing procedure, first the encapsulated samples were cooled down in water and then the molybdenum crucibles were opened in a glove box filled with pure argon. This proved to be necessary since most of the alloys reacted with air rather quickly. The contents were then pulverized and the powder was used to fill diffraction-sample holders specially designed to permit transport in air and diffraction in vacuum. After removing the holders from the glove box and evacuating them to get rid of the argon gas their X-ray diffractogram was made by using CuK-α radiation on a Philips diffractometer system equipped with a graphite monochromator. Chemical analyses showed no trace of molybdenum in the reacted samples. Indexing, refining of lattice constants, filling of unit cells and calculations of reliability factors and interatomic distances were done by using the appropriate self-written algol computing programs running on an IBM VM370.

3. Results

A review of our results is given in the tables 1*a*, 1*b*, 1*c* and fig. 1.

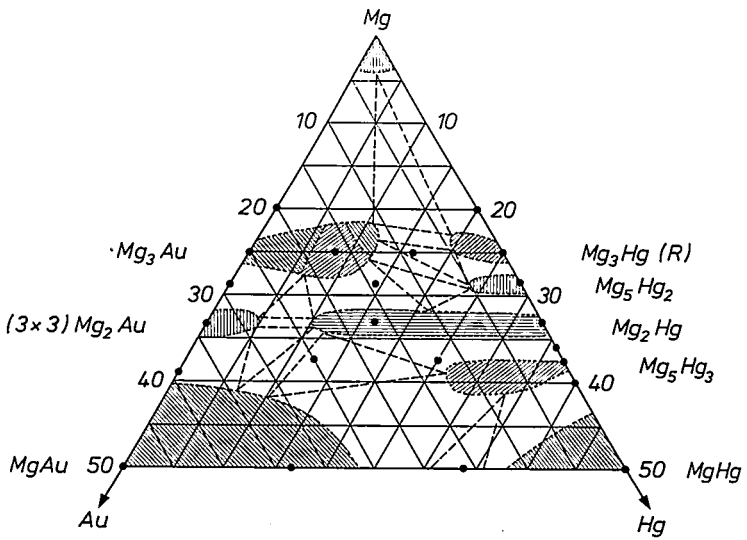


Fig. 1. Schematic phase diagram for the Mg-rich part of the system Mg-Au-Hg ($T \sim 400$ °C).

IPR2018-1557
HTC EX1023, Page 287

TABLE I*a*

Mg–Au samples

composition Mg	Au	treatment time	temp. °C	phases	lattice dimensions (Ångström)	remarks
80	20	4 d	600	h.c.p. Mg	$a = 3.2084(2)$ $c = 5.2055(5)$	
				Cu_3P–type P −3 c 1 Mg_3Au	$a = 8.072(1)$ $c = 8.488(2)$	substructure Na_3As–type P 6_3/m m c $a = 4.660(1)$ Å $c = 8.488(2)$ Å
71.4	28.6	3 d	750	Cu_3P–type P −3 c 1 Mg_3Au	$a = 8.072(1)$ $c = 8.488(2)$	substructure Na_3As–type P 6_3/m m c as above
				P n m a Mg_2Au	$a = 18.315(6)$ $b = 13.548(2)$ $c = 8.251(4)$	substructure Co_2Si–type P n m a $a = 6.105(2)$ Å $b = 4.516(1)$ Å $c = 8.251(2)$ Å
66.7	33.3	3 h	400	P n m a Mg_2Au	$a = 18.345(6)$ $b = 13.566(2)$ $c = 8.258(5)$	substructure Co_2Si–type P n m a $a = 6.115(2)$ Å $b = 4.522(1)$ Å $c = 8.258(2)$ Å
61	39	3 d	750	CsCl–type MgAu	$a = 3.268(1)$	
				P n m a Mg_2Au	$a = 18.345(6)$ $b = 13.563(2)$ $c = 8.258(3)$	substructure Co_2Si–type P n m a as above
50	50	3 d	750	CsCl–type MgAu	$a = 3.270(1)$	

IPR2018-1557
HTC EX1023, Page 288

TABLE I*b*

Mg–Hg samples

composition Mg	composition Hg	treatment time	treatment temp. °C	phases	lattice dimensions (Ångström)	remarks
80	20	5 d	600	h.c.p. Mg	$a = 3.208(1)$ $c = 5.202(2)$	
		17 d	220	R 3 2 rhombohedral Mg_3Hg	$a = 8.35(1)$ $c = 25.84(1)$ or $a = 9.87$ $\alpha = 50.0$ degr.	substructure Na_3As–type P $6_3/m$ m c with $a = 4.821(1)$ Å $c = 8.614(2)$ Å
75	25	5 d 17 d	600 220	R 3 2 rhombohedral Mg_3Hg	$a = 8.351(2)$ $c = 25.845(2)$ or $a = 9.872$ $\alpha = 50.04$ degr.	substructure Na_3As–type P $6_3/m$ m c with $a = 4.823(1)$ Å $c = 8.618(1)$ Å
71.4	28.6	21 d	500	R 3 2 rhombohedral Mg_3Hg	$a = 8.335(4)$ $c = 25.746(9)$ $a = 9.839$ $\alpha = 50.12$ degr.	Na_3As–type substructure as above
				R (−) 3 c Na_5Hg_2–type Mg_5Hg_2	$a = 8.254(1)$ $c = 47.660(1)$ or $a = 16.586$ $\alpha = 28.81$ degr.	substructure P $6_3/m$ m c $a = 4.765(1)$ Å $c = 15.881(1)$ Å
				P n m a Co_2Si–type Mg_2Hg	$a = 6.214(3)$ $b = 4.620(6)$ $c = 8.794(5)$	see below

IPR2018-1557
HTC EX1023, Page 289

TABLE I*b* (continued)

composition Mg	composition Hg	treatment time	treatment temp. °C	phases	lattice dimensions (Ångström)	remarks
66.7	33.3	4 h	600	P n m a Co_2Si–type Mg_2Hg	a = 6.220(1) b = 4.618(1) c = 8.800(2)	see structure determination below
64.3	35.7	2 d	600	P n m a Co_2Si–type Mg_2Hg	a = 6.224(1) b = 4.619(1) c = 8.794(1)	as above
				P 6_3/m c m Mn_5Si_3–type Mg_5Hg_3	a = 8.219(1) c = 5.911(1)	(about 25%)
62.5	37.5	1 d	600	P n m a Co_2Si–type Mg_2Hg	a = 6.222(3) b = 4.629(4) c = 8.805(3)	see above concentration (about 15%)
				P 6_3/m c m Mn_5Si_3–type Mg_5Hg_3	a = 8.220(1) c = 5.913(1)	
				remaining peaks of unindexable phase		
60	40	1 d	600	P 6_3/m c m Mn_5Si_3–type Mg_5Hg_3	a = 8.216(1) c = 5.912(1)	
				CsCl–type MgHg	a = 3.4468(1)	(about 30%)
50	50	1 d	600	CsCl–type MgHg	a = 3.4480(1)	

IPR2018-1557
HTC EX1023, Page 290

TABLE Ic

Mg–Au–Hg samples

composition Mg	Au	Hg	treatment time	temp. °C	phases	lattice dimensions (Ångström)	remarks
50	33.3	16.7	22 d	800	CsCl–type	$a = 3.3161(1)$	
50	16.7	33.3	22 d	500	P $6_3/m$ c m Mn_5Si_3–type Mg_5Hg_3	$a = 8.078(2)$ $c = 5.857(2)$	
					probably a second Mn_5Si_3–type phase but deformed (or a third phase).		
62.5	25	12.5	21 d	600	CsCl–type	$a = 3.345(3)$	
					P $6_3/m$ c m Mn_5Si_3–type	not measured	(about 5%)
					P n m a Co_2Si–type	$a = 6.074(2)$ $b = 4.497(1)$ $c = 8.560(2)$	(about 50%)
62.5	12.5	25	4 d	600	P $6_3/m$ c m Mn_5Si_3–type	$a = 8.127(2)$ $c = 5.880(2)$	
					P n m a Co_2Si–type	$a = 6.081(2)$ $b = 4.521(2)$ $c = 8.681(4)$	(about 10%)
66.7	16.7	16.7	1 d quenched	800	P n m a Co_2Si–type	$a = 6.081(1)$ $b = 4.528(1)$ $c = 8.654(3)$	
71.4	14.3	14.3	21 d	600	P $6_3/m$ m c Na_3As–type substructure	$a = 4.713(1)$ $c = 8.503(2)$	(about 70%) rhombohedral superstructure as in Mg_3Hg
					P n m a Co_2Si	$a = 6.137(2)$ $b = 4.573(1)$ $c = 8.732(1)$	

IPR2018-1557
HTC EX1023, Page 291

TABLE Ic (continued)

composition Mg	Au	Hg	treatment time	temp. °C	phases	lattice dimensions (Ångström)	remarks
75	16.7	8.3	1 d	800 quenched	P 6_3/m m c Na_3As–type substructure	$a = 4.690(1)$ $c = 8.489(1)$	rhombohedral superstructure as in Mg_3Hg
75	8.3	16.7	4 h	600	P 6_3/m m c Na_3As–type substructure	$a = 4.728(1)$ $c = 8.511(2)$	(about 50%) with superstructure as in Mg_3Au
					R (−) 3 c Na_5Hg_2–type	$a = 8.224(2)$ $c = 47.58(1)$ or $a = 16.56$ $\alpha = 28.76$ degr.	(about 25%)
					P 6_3/m m c Na_3As–type substructure	$a = 4.808(2)$ $c = 8.592(2)$	(about 25%) rhombohedral superstructure as in Mg_3Hg

3.1. *The system Mg–MgAu*

In the range 0–25 at.% Au the compound Mg_3Au, as found in agreement with ref. 1, is in equilibrium with the h.c.p. magnesium lattice. In literature of before 1965 Mg_3Au was reported to have a so-called Na_3As–type P 6_3/m m c *) unit cell [2]), but Mansmann [3]) showed convincingly that its structure is isomorphous with that of Cu_3P which has a three times larger hexagonal cell. Mansmann stated that the latter compound does not have the structure as given by Pearson in his Handbook [2]), but one in which a copper atom is situated at 0,0,1/4 instead of 0,0,0, the so-called anti-LaF_3 structure [3]). In 1972 the structure of Cu_3P was redetermined by Olofsson [4]) using a single crystal. Our observation of a superstructure of the 'Na_3As' cell with $a = 8.072(1)$ Å and $c = 8.488(2)$ Å [2]) is in full agreement with Mansmann's work. Although our powder diagram was not free from preferred orientation, R–factor calculations (for 36 reflexions) on all three structure models, i.e. Pearson's Cu_3P type

*) Space-group notations according to international conventions, see ref. 10.

IPR2018-1557
HTC EX1023, Page 292

(R = 22%), Olofsson's model for Cu_3P (R = 20%) and the Mansmann anti-LaF_3 type (R = 18.8%) were in favour of the latter. See the description in sec. 4.

Between 25 and 33.3 at.% Au Mg_3Au is in equilibrium with a compound Mg_2Au, with what at first seemed to be an orthorhombic P n m a unit cell of structure type Co_2Si. However, the diffraction pattern of this compound was always accompanied by a number of additional weak peaks which eventually could be indexed as superstructure peaks, indicating that the real unit cell is three times that of a 'Co_2Si-type Mg_2Au' in two dimensions (with a = 18.34(1) Å, b = 13.56(1) Å and c = 8.26(1) Å. This appeared in agreement with work of van Look and Schubert [5]).

At 33.3 at.% Au the sample was practically single-phase Mg_2Au, containing only a small amount of CsCl-type MgAu, so we used its diffractogram for a provisional structure determination. For the substructure of Mg_2Au (the Co_2Si-type approximation) the refinement procedure consisted of first a refinement on intensities of the parameters of the heavy atoms followed by a refinement on minimum total overlap, keeping the gold atoms fixed and taking for the radii values given for 12 coordination. Starting from this Co_2Si-type subcell we constructed a supercell in space group P n m a (nr. 62) and again refined alternatingly the gold parameters on intensities and subsequently the magnesium positions on minimum overlap. This time we used 15 position parameters for the gold atoms and 30 position parameters for the magnesium atoms. The diagram measured was cleaned beforehand of peaks belonging to the CsCl-type MgAu compound. The superstructure peaks calculated in these regions were also excluded. However, the possibility remains that the composition is no longer $(Mg_2Au)_9$ but that one or more atoms are missing. In that case we should expect to observe some extra peaks of the neighbouring phases in our Mg_2Au diagram. Regretfully these peaks would coincide with some peaks of the superlattice, and therefore we omitted the latter in the R-factor calculation. The structure that we arrived at combines as such a reasonable fit of the calculated intensities and a geometrically acceptable positioning of the atoms, as is shown in sec. 5 where the final structure proposal is described. Table V shows the comparison of observed and calculated diagram. The R-factor is 16.9%, which is indeed not too bad in view of the overall quality of the diagram and the unavoidable presence of preferred orientation.

At 50 at.% Au the sample was found to be single-phase CsCl-type, in substantial agreement with ref. 6.

3.2. *The system Mg–MgHg*

Analogous to the above system the first compound found, starting from Mg, is Na_3As-type Mg_3Hg. Here, however, every sample showed a number

IPR2018-1557
HTC EX1023, Page 293

of small additional peaks, and this induced us to describe the real unit cell as a large rhombohedron. The dimensions of the (hexagonal) cell are clearly related to a hypothetical 'Cu_3P–type' Mg_3Hg, such that the c axis is multiplied by three. Obviously the Mg_3Hg structure seems a stacking variant of the Cu_3P structure (see fig. 2). With a cell filling program the trial structure given in sec. 4 was found most promising. The R–factor was 12.65% for the first 58 calculated peaks including the non-observed reflexions (see table III). Refinement did not lead to a substantially better value. The structure shows the same kind of mixed heavy atom–magnesium layers as are present in the anti-LaF_3 structure, which seems to increase the probability that both structures are right.

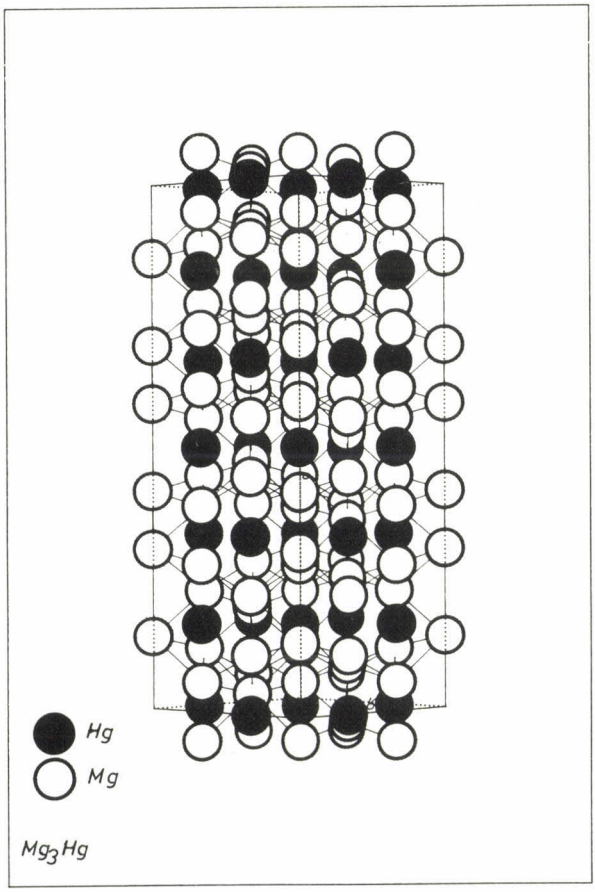


Fig. 2*a*). The structure of Mg_3Hg as proposed by a cell-filling program, being obviously a stacking variant of the anti-LaF_3 type.

IPR2018-1557
HTC EX1023, Page 294

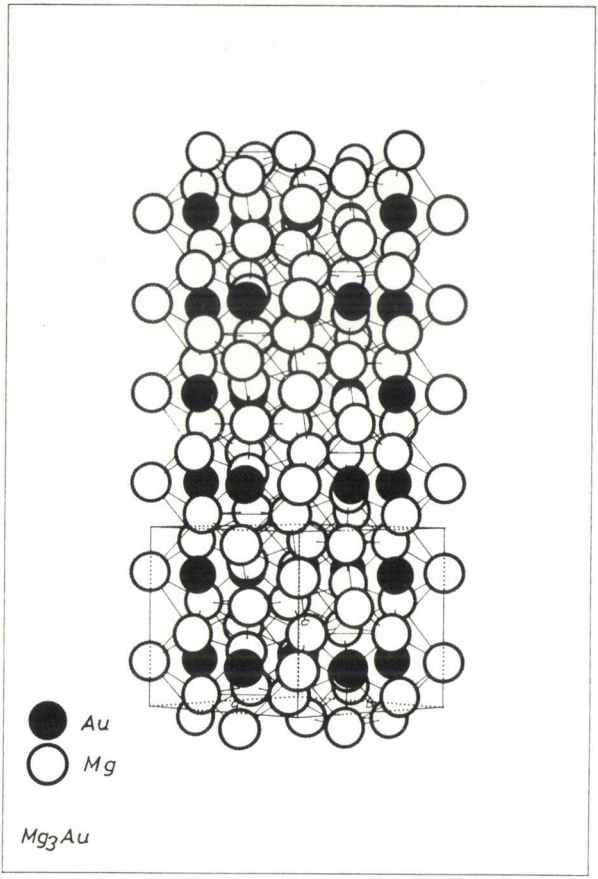


Fig. 2*b*. Three unit cells of Mg_3Au with the anti-LaF_3 type of structure. Note that the heavy atoms in both types are situated in layers which also contain some magnesium atoms, in contrast to the Cu_3P type structure as given by Pearson.

The 72 at.% Mg sample contained three phases, probably due to the peritectic melting of the majority phase denoted as Mg_5Hg_2 by ref. 7. Besides a small amount of Mg_3Hg and Co_2Si–type Mg_2Hg we found a rhombohedral (R 3 c or R − 3 c) unit cell with $a = 8.254(1)$ Å and $c = 47.660(1)$ Å ($a = 16.586$ Å, $\alpha = 28.81$ degr.). The latter pattern could also be described as having a P 6_3/m m c substructure with unit-cell dimensions $a = 4.765(1)$ Å and $c = 15.881(1)$ Å. Although the type is unknown and even the composition is not quite certain, we still think to have found a compound that is isostructural with the so-called 'Na_5Hg_2' with $a = 9.39$ and $c = 53.1$ Å ($a = 18.52$ Å and $\alpha = 29.38$ degr.), reported by Nielsen and Baenzinger [8]).

IPR2018-1557
HTC EX1023, Page 295

A sample with 33.3 at.% Hg yielded a diffractogram that was indexed single-phase on the basis of an orthorhombic P n m a unit cell with $a = 6.220(1)$ Å, $b = 4.618(1)$ Å and $c = 8.800(2)$ Å and could be used for a structure determination.

The structure is of the Co_2Si–type and consequently the composition of the compound is probably Mg_2Hg, in agreement with Hansen [1]). The refinement of Mg_2Hg was performed in the same way as described for Mg_2Au. The structural data are given in sec. 5. Table VI shows the comparison of the observed and the calculated diagram.

In the 35.7 at.% Hg sample the majority phase is hexagonal Mn_5Si_3–type Mg_5Hg_3. At the composition Mg_5Hg_3 this should have been the only phase present, but we also found some Mg_2Hg and moreover a few small unexplained peaks.

The adjacent two-phase region, represented by a sample of composition 40 at.% Hg, showed the presence of Mg_5Hg_3 and CsCl–type MgHg, in agreement with Hansen [1]).

3.3. *The cross-section MgAu–MgHg*

The single-phase regions of the CsCl–type compounds MgAu and MgHg do not join. In between a region exists where at least one Mn_5Si_3–type lattice is stable.

3.4. *The cross-section Mg_5Au_3–Mg_5Hg_3*

This section comes from the two-phase (Mg_2Au–MgAu) region, ends in the single-phase Mg_5Hg_3 region and does not reveal any new ternary phase in between. The single-phase Mg_5Hg_3 region does not even extend to the point at which one third of the mercury atoms is replaced by gold.

3.5. *The cross-section Mg_2Au–Mg_2Hg*

Here we studied only the composition halfway, which was purely single-phase Co_2Si–type, indicating that at least in this lattice the mercury atoms may easily be replaced by gold atoms. As no trace of a superstructure was observed, we must conclude that the region of stability of the superstructure is probably much smaller than that of the substructure.

3.6. *The cross-section Mg_5Au_2–Mg_5Hg_2*

The halfway composition we studied here was two-phase Mg_3Hg–type + Co_2Si–type. The lattice constants of both phases reveal a solid solution to a certain extent at both ends of the pseudo binary system.

IPR2018-1557
HTC EX1023, Page 296

3.7. *The cross-section Mg_3Au–Mg_3Hg*

This section appears to be composed of a single-phase region extending from pure Mg_3Au to compounds in which more than one third of gold is replaced by mercury. At two thirds replacement, however, a three-phase region exists where at 600 °C Mg_3Au, Mg_3Hg and 'Mg_5Hg_2' seem to be in equilibrium.

4. Structural information on the compounds Mg_3Hg and Mg_3Au

Mg_3Hg		Mg_3Au
hR72		hP24
R 3 2 (155)		P −3 c 1 (165)
hexagonal cell	rhombohedral cell	(anti-LaF_3 type)
a = 8.351(2) Å	a = 9.872 Å	a = 8.072(1) Å
c = 25.845(2) Å	α = 50.04 degr.	c = 8.488(2) Å
V = 1560.9 Å^3	V = 520.3 Å^3	V = 478.5 Å^3

Atomic position parameters are

9 Hg at (d)	1/3	1/3	0	6 Au at (f)	.659	0	1/4
9 Hg at (e)	2/3	2/3	1/6	12 Mg at (g)	.365	.056	.080
6 Mg at (c)	2/3	1/3	7/36	4 Mg at (d)	1/3	2/3	.193
6 Mg at (c)	1/3	2/3	3/36	2 Mg at (a)	0	0	1/4
6 Mg at (c)	2/3	1/3	1/36				
18 Mg at (f)	2/3	2/3	1/18				
18 Mg at (f)	1/3	1/3	1/9				

The interatomic distances are listed in table II.

TABLE II

Interatomic distances

Mg_3Hg			Mg_3Au		
atoms	number	distance (Å)	atoms	number	distance (Å)
Mg–Mg	2	2.8713	Mg–Mg	1	2.9962
	1	2.8722		4	3.0353
	1	2.8739		2	3.0582
	4	2.8743		1	3.0677
	12	2.8749		7	3.1013
	6	3.1315		4	3.1754
Hg–Mg	2	2.8705		3	3.2921
	2	2.8713		1	3.2922
	2	2.8747	Au–Mg	3	2.7038
	2	2.8749		3	2.7513
	6	3.1323		1	2.8003
	6	3.1327		1	2.8238
				1	3.0041

IPR2018-1557
HTC EX1023, Page 297

TABLE III

Observed and calculated diffraction angles θ (in degrees) and intensities I (in arbitrary units) of the intermetallic compound Mg_3Hg

observed					calculated	
θ	I	h	k	l	θ	I
10.29	46.0	0	0	6	10.30	37.6
		0	1	5	10.55	.1
10.60	43.0	1	1	0	10.63	36.0
11.81	125.0	1	1	3	11.83	123.2
14.87	25.0	1	1	6	14.89	23.6
15.52	2.0	0	0	9	15.56	.0
16.72	3.5	1	2	2	16.74	.0
17.32	1.0	0	2	7	17.35	.6
17.81	1.0	2	1	4	17.82	.3
		1	2	5	18.59	.0
		3	0	0	18.63	71.9
18.61	67.0	2	0	8	18.65	.0
18.97	85.0	1	1	9	19.00	83.2
19.40	4.0	3	0	3	19.38	.0
20.48	1.5	2	1	7	20.52	.7
20.95	6.0	0	0	12	20.96	6.4
		3	0	6	21.48	36.1
21.47	41.0	0	2	10	21.49	.0
22.30	24.0	2	2	3	22.31	24.4
		1	0	13	23.69	.0
		1	3	4	23.71	.1
23.71	10.0	1	1	12	23.73	6.8
		2	2	6	24.20	6.5
24.20	4.0	2	1	10	24.21	.0
24.61	1.0	3	0	9	24.66	.0
27.13	12.0	2	2	9	27.14	30.3

5. Structural information of the compounds Mg_2Hg and Mg_2Au

Mg_2Hg	Mg_2Au	
oP12 (C23)	substructure oP12 (C23)	superstructure oP108
Co_2Si–type	Co_2Si–type	
P n m a (62)	P n m a (62)	P n m a (62)
$a = 6.219(1)$ Å	$a = 6.114(1)$ Å	$a = 18.328(6)$ Å
$b = 4.617(1)$ Å	$b = 4.520(1)$ Å	$b = 13.561(7)$ Å
$c = 8.799(2)$ Å	$c = 8.256(1)$ Å	$c = 8.253(2)$ Å
$V = 252.6$ Å^3	$V = 228.4$ Å^3	$V = 2051.1$ Å^3

IPR2018-1557
HTC EX1023, Page 298

Atomic position parameters are:

Mg	.184	3/4	.945	Mg	.189	3/4	.929	Mg	.063	1/4	.925
Mg	.558	3/4	.212	Mg	.532	3/4	.225	Mg	.372	1/4	.928
Hg	.281	1/4	.123	Au	.284	1/4	.101	Mg	.763	1/4	.921
B(Hg) = 3.12 Å^2				B(Au) = 2.58 Å^2				Mg	.180	1/4	.251
R = 15.7%				R = 14.9%				Mg	.517	1/4	.218
				(on substructure peaks)				Mg	.849	1/4	.209
								Mg	.052	.580	.929
								Mg	.406	.571	.916
								Mg	.731	.588	.937
								Mg	.170	.571	.213
								Mg	.505	.591	.205
								Mg	.842	.566	.247
								Au	.081	3/4	.121
								Au	.381	3/4	.093
								Au	.789	3/4	.102
								Au	.105	.087	.073
								Au	.424	.108	.113
								Au	.756	.077	.112
								R = 16.9% over 197 peaks including all non observed reflections			

The interatomic distances are listed in table IV.

For both structures the coordination numbers are 13, 13 and 10 for the Mg, Mg and the Hg/Au atoms respectively[9]).

TABLE IV

Interatomic distances

Mg_2Hg			Mg_2Au *)		
atoms	number	distance (Å)	atoms	number	distance (Å)
Mg–Mg	2	3.1102	Mg–Mg	2	3.0158
	4	3.1247		2	3.0859
	2	3.1722		4	3.1037
	2	3.2962		4	3.1272
	2	3.3798		2	3.2180
	4	3.4241		2	3.4349
Hg–Hg	2	3.8220	Au–Au	2	3.8514
Hg–Mg	1	2.8368	Au–Mg	1	2.7146
	2	2.8449		2	2.7316
	1	2.9447		1	2.9028
	2	2.9748		2	2.9051
	2	3.0537		1	2.9119
	1	3.1014		2	3.0928
	1	3.3707		1	3.2314

*) Substructure. The superstructure has a shortest distance of 2.582 Å for Mg–Au and 2.855 Å for Mg–Mg.

IPR2018-1557

HTC EX1023, Page 299

TABLE V

Observed and calculated diffraction angles θ (in degrees) and intensities I (in arbitrary units) of the intermetallic compound Mg_2Au

observed					calculated	
θ	I	h	k	l	θ	I
		2	0	1	5.82	7.0
5.84	9.1	1	1	0	5.87	4.1
9.01	11.0	3	1	0	9.02	7.6
10.22	36.6	4	0	1	10.21	27.3
10.74	24.5	0	2	0	10.75	16.9
		3	1	2	11.16	5.0
11.20	129.5	0	1	3	11.20	123.9
		1	2	1	11.51	7.2
11.53	7.6	4	1	1	11.56	2.4
11.72	7.0	4	0	2	11.70	6.3
		2	1	3	12.22	1.5
12.26	9.9	2	2	1	12.26	8.3
13.00	193.0	3	2	0	13.01	201.4
		3	1	3	13.39	244.1
13.39	245.0	3	2	1	13.43	2.4
13.84	4.3	4	0	3	13.84	4.1
14.02	4.5	2	0	4	14.02	4.0
		6	0	0	14.59	62.0
14.61	73.5	3	2	2	14.61	2.3
		1	2	3	14.84	1.0
		5	1	2	14.85	9.6
14.84	8.9	4	1	3	14.88	3.4
		4	2	1	14.92	.5
15.58	4.6	6	1	0	15.59	3.4
		3	2	3	16.40	1.5
		4	0	4	16.41	1.3
16.40	1.7	1	3	0	16.44	1.5
16.68	2.3	5	2	1	16.65	2.8
		1	2	4	17.28	2.2
		4	1	4	17.31	14.2
17.32	11.4	2	3	1	17.32	.8
		4	2	3	17.66	3.8
17.70	28.5	6	0	3	17.71	25.0
		1	3	2	17.76	4.3
17.89	3.3	3	3	0	17.87	1.4
		2	3	2	18.27	6.8
		6	2	0	18.27	17.4
18.26	26.9	7	1	1	18.28	3.3
		6	1	3	18.55	99.1
18.55	103.5	6	2	1	18.58	2.2
		3	3	2	19.10	.2

IPR2018-1557
HTC EX1023, Page 300

TABLE V (continued)

observed					calculated	
θ	I	h	k	l	θ	I
		0	3	3	19.13	166.8
		5	2	3	19.18	1.6
19.14	173.0	7	1	2	19.19	13.4
		8	0	1	19.92	.7
19.93	67.0	0	0	6	19.93	77.2
20.24	17.8	4	3	2	20.22	1.9
		2	0	6	20.55	4.1
		3	3	3	20.55	21.6
20.55	33.8	6	1	4	20.60	2.2
		7	1	3	20.63	11.2
		7	2	1	20.66	33.7
20.66	58.6	8	1	1	20.69	.7
20.90	7.5	6	2	3	20.91	2.4
		3	2	5	21.25	1.2
21.24	14.5	1	3	4	21.28	7.1
		7	2	2	21.48	5.9
21.51	6.6	8	1	2	21.51	.4
21.91	30.5	0	4	0	21.91	20.7
		3	1	6	22.04	1.9
22.03	2.6	1	4	0	22.06	.0
		6	3	0	22.12	11.9
22.11	10.0	8	0	3	22.13	.2
		2	4	0	22.49	.0
22.51	22.7	7	1	4	22.52	24.2
		8	1	3	22.84	.3
		8	2	1	22.86	1.4
		0	2	6	22.87	4.1
		5	3	3	22.90	.7
22.87	9.6	9	1	0	22.91	2.4
		6	3	2	23.16	4.0
		9	1	1	23.17	5.6
23.19	27.5	3	4	0	23.20	10.1
		2	2	6	23.43	.4
		4	3	4	23.44	7.2
23.43	4.8	3	4	1	23.45	.4
		8	2	2	23.62	.0
23.64	4.8	1	3	5	23.64	1.2
		0	1	7	24.11	.7
24.10	71.0	3	2	6	24.11	58.6
		4	4	0	24.16	.3
		4	4	1	24.40	5.9
24.41	49.5	6	3	3	24.42	45.4
		9	2	0	24.93	12.8
24.93	22.5	7	3	2	24.93	1.5

IPR2018-1557
HTC EX1023, Page 301

TABLE VI

Observed and calculated diffraction angles θ (in degrees) and intensities I (in arbitrary units) of the intermetallic compound Mg_2Hg

observed					calculated	
θ	I	h	k	l	θ	I
8.79	6.0	1	0	1	8.75	5.9
10.92	103.0	0	1	1	10.89	93.2
12.45	133.0	1	0	2	12.42	138.6
13.11	108.0	1	1	1	13.07	110.2
14.42	90.0	2	0	0	14.38	51.5
15.28	8.5	2	0	1	15.27	4.4
15.85	1.0	1	1	2	15.81	1.5
16.95	4.5	1	0	3	16.93	5.2
17.45	28.5	2	1	0	17.43	25.7
17.75	3.0	2	0	2	17.71	3.1
		0	1	3	18.18	60.8
18.21	129.0	2	1	1	18.19	77.9
		0	2	0	19.55	42.7
19.60	90.0	1	1	3	19.62	58.1
20.57	26.5	0	0	4	20.55	15.8
21.21	8.5	2	0	3	21.22	8.2
21.58	2.0	1	2	1	21.57	.9
21.88	2.0	1	0	4	21.86	.1
22.21	1.5	0	2	2	22.20	.0
22.52	5.5	3	0	1	22.51	1.6
		1	2	2	23.44	38.2
23.48	63.5	2	1	3	23.50	24.2
24.32	18.0	3	0	2	24.32	17.2
		2	2	0	24.63	18.1
24.67	30.5	3	1	1	24.69	11.2
25.22	4.0	2	2	1	25.21	1.6
25.46	13.5	2	0	4	25.47	18.3
		1	2	3	26.33	2.6
26.38	17.5	3	1	2	26.40	8.0
26.86	3.5	2	2	2	26.89	1.8
		1	0	5	27.13	.0
27.13	3.5	3	0	3	27.14	2.5
27.48	8.0	2	1	4	27.48	2.4
27.99	3.5	0	1	5	28.01	3.9
		0	2	4	29.01	8.5
		1	1	5	29.07	9.9
29.04	28.0	3	1	3	29.07	5.4
29.49	5.0	2	2	3	29.54	5.2
29.75	7.5	4	0	0	29.79	2.6
		2	0	5	30.28	.2
30.25	11.5	4	0	1	30.30	5.3
		3	2	1	30.56	.8
30.60	7.5	0	3	1	30.64	3.1
30.78	5.0	3	0	4	30.79	.2

IPR2018-1557
HTC EX1023, Page 302

6. Acknowledgement

Thanks are due to H. C. Donkersloot, who assisted at the major part of the experimental work, and P. Hokkeling who was responsible for the preparation of the samples. Especially we thank J. Hornstra for his valuable remarks.

REFERENCES

[1]) M. Hansen and K. Anderko, Constitution of binary alloys, McGraw-Hill, New York, 1958.
[2]) W. B. Pearson, A handbook of lattice spacings and structures of metals and alloys, Pergamon Press, New York, 1967, cf. Vol. 2, p. 206.
[3]) M. Mansmann, Z. für Kristallog. **122**, 399 (1965).
[4]) O. Olofsson, Acta Chem. Scand. **26**, 2777 (1972).
[5]) N. van Look and K. Schubert, Metall. **23**, 4 (1969).
[6]) G. Brauer and W. Haucke, Z. physik. Chem. **B33**, 304 (1963).
[7]) G. Brauer and R. Rudolph, Z. anorg. Chem. **248**, 405 (1941).
[8]) J. W. Nielsen and N. C. Baenzinger, Acta Cryst. **7**, 277 (1954).
[9]) W. B. Pearson, The crystal chemistry and physics of metals and alloys, John Wiley and Sons Inc., New York, 1972, cf. p. 532.
[10]) N. F. M. Henry and K. Lonsdale (eds), International tables for X-ray crystallography, Kynoch Press, Birmingham, England, vol. 1, 1962.

IPR2018-1557
HTC EX1023, Page 303

Philips J. Res. 39, 293-304, 1984 *R 1097*

SOME CONSTRUCTIONS OF OPTIMAL BINARY LINEAR UNEQUAL ERROR PROTECTION CODES

by W. J. VAN GILS

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

This paper describes a number of constructions of binary Linear Unequal Error Protection (LUEP) codes. The separation vectors of the constructed codes include those of all optimal binary LUEP codes of length less than or equal to 15.

AMS: 94B05, 94B60.

1. Introduction

Consider a binary linear code C of length n and dimension k with generator matrix G to be used on a binary symmetric channel. In many applications it is necessary to provide different protection levels for different components m_i of the input message word $\boldsymbol{m}$. For example in transmitting numerical (binary) data, errors in the more significant bits are more serious than are errors in the less significant bits, and therefore more significant bits should have more protection than less significant bits.

A suitable measure for these protection levels for separate positions in input message words is the separation vector [1]).

Definition

For a binary linear $[n,k]$ code C the separation vector $\boldsymbol{s}(G) = (s(G)_1, s(G)_2, \ldots, s(G)_k)$ with respect to a generator matrix G of C is defined by

$$s(G)_i := \min\{\mathrm{wt}(\boldsymbol{m}\,G) \mid \boldsymbol{m} \in \{0,1\}^k,\ m_i = 1\},$$

where wt(.) denotes the Hamming weight function.

This separation vector $\boldsymbol{s}(G)$ guarantees the correct interpretation of the i^{th} message bit whenever Nearest Neighbour Decoding (ref. 2 p. 11) is applied and no more than $(s(G)_i - 1)/2$ errors have occurred in the transmitted codeword [1]).

A linear code that has a generator matrix G such that the components of the corresponding separation vector $\boldsymbol{s}(G)$ are not mutually equal is called a Linear

IPR2018-1557
HTC EX1023, Page 304

Unequal Error Protection (LUEP) code[1]). By permuting the rows of a gennerator matrix G we may obtain a generator matrix G' for the code such that $s(G')$ is nonincreasing, i.e. $s(G')_i \geqslant s(G')_{i+1}$ for $i = 1, 2, \ldots, k-1$. In this paper we always assume that the rows in generator matrices are so ordered that the corresponding separation vectors are nonincreasing.

Any LUEP code C has a so-called optimal generator matrix G^*. This means that the separation vector $s(G^*)$ is componentwise larger than or equal to the separation vector $s(G)$ of any generator matrix G of C[1]), denoted by $s(G^*) \geqslant s(G)$ ($x \geqslant y$ means $x_i \geqslant y_i$ for all i). The vector $s = s(G^*)$ is called the separation vector of the linear code C. We use the notation $[n,k,s]$ for C.

For any $k \in \mathbb{N}$ and $s \in \mathbb{N}^k$ we define $n(s)$ to be the length of the shortest binary linear code of dimension k with a separation vector of at least s, and $n^{ex}(s)$ to be the length of the shortest binary linear code of dimension k with separation vector (exactly) s[3,4]). An $[n(s),k,s]$ code is called length-optimal[3]). It is called optimal if an $[n(s),k,t]$ code with $t \geqslant s$, $t \neq s$ does not exist[3,4]). In refs 3 and 4 a number of bounds for the functions $n(s)$ and $n^{ex}(s)$ are derived. In ref. 5 methods for constructing LUEP codes from shorther codes are described.

In refs 3 and 4 an incomplete list of the separation vectors of the optimal binary LUEP codes of length less than or equal to 15 is given. In this paper we provide the complete list of the separation vectors of all optimal binary LUEP codes of length less than or equal to 15, together with examples of generator matrices having these separation vectors. Furthermore, we give a number of constructions of infinite series optimal binary LUEP codes.

2. Constructions

Table I provides the separation vectors of all optimal binary LUEP codes of length less than or equal to 15. In this table, n denotes the length of the code, k denotes the dimension, and $d(n,k)$ denotes the maximal minimum distance of a binary code of length n and dimension k. The brackets and commas commonly appearing in separation vectors have been deleted. Only in the cases where a component of a separation vector is larger than 9, it is followed by a point (.). Examples of codes having the parameters given in table I are constructed below. The bounds in ref. 4 can be used to show that certain LUEP codes are optimal. They are also useful in showing that table I is complete. In cases where these bounds did not work, methods of exhaustive search were used to show that codes with certain parameters do not exist. Table I is the same table as table I in ref. 4, extended by the parameters [14,10,(4333322222)], [15,3,(994)], [15,8,(73333333)], [15,8,(55554443)], [15,8,(55544444)] and [15,11,(43333222222)]. In (ref. 4 table I) no references to constructions were given, which has been done in this paper.

IPR2018-1557
HTC EX1023, Page 305

TABLE I

The separation vector of all binary optimal LUEP codes of length less than or equal to 15.

n	k	$d(n,k)$	separation vector
4	2	2	A 32
5	2	3	A 42
5	3	2	A 322
6	2	4	A 52
6	3	3	A 422
6	4	2	A 3222
7	2	4	A 62, I 54
7	3	4	A 522
7	4	3	A 4222
7	5	2	A 32222
8	2	5	A 72, I 64
8	3	4	A 622, C 544
8	4	4	A 5222
8	5	2	A 42222, J 33332
8	6	2	A 322222
9	2	6	A 82, I 74
9	3	4	A 722, C 644, G 554
9	4	4	A 6222, C 5444
9	5	3	A 52222, J 44442, B 43333
9	6	2	A 422222, J 333322
9	7	2	A 3222222
10	2	6	A 92, I 84, I 76
10	3	5	A 822, C 744, L 664
10	4	4	A 7222, C 6444, G 5544
10	5	4	A 62222, C 54444
10	6	3	A 522222, J 444422, J 433332
10	7	2	A 4222222, J 3333222
10	8	2	A 32222222
11	2	7	A 10.2, I 94, I 86
11	3	6	A 922, C 844, K_1 764
11	4	5	A 8222, C 7444, E 6644
11	5	4	A 72222, C 64444, G 55444
11	6	4	A 622222, J 544442, B 533333
11	7	3	A 5222222, J 4444222, J 4333322
11	8	2	A 42222222, J 33332222
11	9	2	A 322222222
12	2	8	A 11.2, I 10.4, I 96
12	3	6	A 10.22, C 944, E 864, K_2 774, K_1 766
12	4	6	A 9222, C 8444, K_1 7644
12	5	4	A 82222, C 74444, E 66444, M 55554
12	6	4	A 722222, C 644444, G 554444
12	7	4	A 6222222, J 5444422, J 5333332

IPR2018-1557
HTC EX1023, Page 306

TABLE I (cont.)

n	k	$d(n,k)$	separation vector
12	8	3	*A* 52222222, *J* 44442222, *J* 43333222
12	9	2	*A* 422222222, *J* 333322222
12	10	2	*A* 3222222222
13	2	8	*A* 12.2, *I* 11.4, *I* 10.6, *I* 98
13	3	7	*A* 11.22, *C* 10.44, K_1 964, *E* 884, *L* 866
13	4	6	*A* 10.222, *C* 9444, *L* 8644, *F* 7744, K_1 7666
13	5	5	*A* 92222, *C* 84444, K_1 76444, *L* 66664, *H* 66555
13	6	4	*A* 822222, *C* 744444, *D* 664444, *M* 555544
13	7	4	*A* 7222222, *J* 6444442, *B* 6333333, *J* 5544442, K_1 5444444
13	8	4	*A* 62222222, *J* 54444222, *J* 53333322
13	9	3	*A* 522222222, *J* 444422222, *J* 433332222
13	10	2	*A* 4222222222, *J* 3333222222,
13	11	2	*A* 32222222222
14	2	9	*A* 13.2, *I* 12.4, *I* 11.6, *I* 10.8
14	3	8	*A* 12.22, *C* 11.44, *L* 10.64, K_1 984, K_1 966
14	4	7	*A* 11.222, C 10.444, K_1 9644, *L* 8844, *L* 8666
14	5	6	*A* 10.2222, *C* 94444, *L* 86444, *F* 77444, *N* 76666
14	6	5	*A* 922222, *C* 844444, *E* 764444, *L* 666644, *J* 665552
14	7	4	*A* 8222222, *C* 7444444, *J* 6644442, *Q* 6544444, *M* 5555444
14	8	4	*A* 72222222, *J* 64444422, *J* 63333332, *J* 55444422, K_1 54444444
14	9	4	*A* 622222222, *J* 544442222, *J* 533333222
14	10	3	*A* 5222222222, *J* 4444222222, *J* 4333322222
14	11	2	*A* 42222222222, *J* 33332222222
14	12	2	*A* 322222222222
15	2	10	*A* 14.2, *I* 13.4, *I* 12.6, *I* 11.8
15	3	8	*A* 13.22, *C* 12.44, K_1 11.64, K_1 10.84, *L* 10.66, K_2 994, K_1 988
15	4	8	*A* 12.222, *C* 11.444, *L* 10.644, K_1 9844, K_1 9666
15	5	7	*A* 11.2222, *C* 10.4444, K_1 96444, *L* 88444, *L* 86666
15	6	6	*A* 10.22222, *C* 944444, *L* 864444, K_2 774444, *J* 766662, K_1 766644, *O* 765554
15	7	5	*A* 9222222, *C* 8444444, *P* 7644444, *L* 6666444, *J* 6655522
15	8	4	*A* 82222222, *J* 74444442, *B* 73333333, *J* 66444422, *J* 65444442, *L* 64444444, *R* 55554443, *S* 55544444
15	9	4	*A* 722222222, J 644444222, *J* 633333322, *J* 554444222, K_1 544444444
15	10	4	*A* 6222222222, *J* 5444222222, *J* 5333332222
15	11	3	*A* 52222222222, *J* 44442222222, *J* 43333222222
15	12	2	*A* 422222222222, *J* 333322222222
15	13	2	*A* 3222222222222

IPR2018-1557
HTC EX1023, Page 307

In this paper we frequently use two results of ref. 4. Hence we repeat these results in the following two theorems.

Theorem 1 (ref. 4, theorem 12)

For any $k \in \mathbb{N}$ and nonincreasing $s \in \mathbb{N}^k$ we have that

$$n^{ex}(s_1, \ldots, s_k) \geqslant s_i + n(\hat{s}_1, \ldots, \hat{s}_{i-1}, \hat{s}_{i+1}, \ldots, \hat{s}_k)$$

holds for any $i \in \{1, \ldots, k\}$, where

$$\hat{s}_j := \begin{cases} s_j - \lfloor s_j/2 \rfloor & \text{for } j < i \\ \lceil s_j/2 \rceil & \text{for } j > i, \end{cases}$$

(where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x, and $\lceil x \rceil$ denotes the smallest integer larger than or equal to x).

Theorem 2 (ref. 4, corollary 14)

For any $k \in \mathbb{N}$ and any nonincreasing $s \in \mathbb{N}^k$, the function $n(s)$ satisfies the following inequalities,

a. $n(s_1, s_2, \ldots, s_k) \geqslant s_1 + n(\lceil s_2/2 \rceil, \ldots, \lceil s_k/2 \rceil)$,

b. $n(s_1, s_2, \ldots, s_k) \geqslant \sum_{i=1}^{k} \lceil s_i/2^{i-1} \rceil$.

Construction A

For $n, k \in \mathbb{N}$, $n \geqslant k+1$, the k by n matrix

$$\left[\begin{array}{c|c} & 11111\ldots\ldots111 \\ I_k & 1 \\ & \vdots \quad 0_{k-1,n-k-1} \\ & 1 \end{array}\right] \qquad (1)$$

is a generator matrix of an optimal binary $[n, k, (n-k+1, 2, 2, \ldots, 2)]$ code (I_k denotes the identity matrix of order k, $0_{k-1,n-k-1}$ denotes the all-zero $k-1$ by $n-k-1$ matrix).

Proof

It is easy to check that the parameters of the code are correct. Furthermore by theorem 2*b* the length of a k–dimensional binary code with separation vector $(n-k+1, 2, 2, \ldots, 2)$ is at least n, and with separation vector larger than $(n-k+1, 2, 2, \ldots, 2)$ is at least $n+1$ (by $s > t$ (s larger than t) we mean $s \geqslant t, s \neq t$).

IPR2018-1557
HTC EX1023, Page 308

Construction B

For $k \in \mathbb{N}$, $k \geqslant 4$, the k by $2k-1$ matrix

$$\left[\begin{array}{c|c|c} 00\ldots\ldots0 & 11\ldots\ldots1 & 0 \\ \hline & & 1 \\ I_{k-1} & I_{k-1} & 1 \\ & & \vdots \\ & & 1 \end{array}\right] \tag{2}$$

is a generator matrix of an optimal binary $[2k-1, k, (k-1, 3, 3, \ldots, 3)]$ code.

Proof

It is easy to verify that the parameters of the code are correct. By theorem 2*b*, we have that the length of a k–dimensional binary code with separation vector $(k-1, 3, 3, \ldots, 3)$ is at least $2k-1$. Application of theorem 2*b* to a k–vector s with $s_1 \geqslant k$ and $s_i \geqslant 3$ for $i = 2, \ldots, k$ shows that $n(s) \geqslant 2k$. Application of the theorems 1 and 2 to a k–vector s such that $s_1 = k-1$, $s_2 \geqslant 4$, $s_i \geqslant 3$ for $i = 3, \ldots, k-1$, and $s_k = 3$ shows that

$$\begin{aligned} n^{ex}(s) &\geqslant 3 + n(s_1 - 1, \ldots, s_{k-1} - 1) \\ &\geqslant 3 + s_1 - 1 + n(\lceil (s_2 - 1)/2 \rceil, \ldots, \lceil (s_{k-1} - 1)/2 \rceil) \\ &\geqslant 3 + k - 2 + \underbrace{n(2, 1, 1, \ldots, 1)}_{k-2} \\ &\geqslant 3 + k - 2 + k - 1 = 2k. \end{aligned}$$

Furthermore it is not difficult to check that a binary $[2k-1, k, (k-1, 4, 4, \ldots, 4)]$ code does not exist. Finally, by theorem 2*b* the length of a k–dimensional binary code with a separation vector of at least $(k-1, 5, 4, 4, \ldots, 4)$ is at least $2k$. These observations show that the code in construction B is optimal.

Construction C

For $n, k \in \mathbb{N}$, $n \geqslant \max\{2k, k+4\}$, the k by n matrix

$$\left[\begin{array}{c|c|c|c} 00\ldots\ldots0 & 11\ldots\ldots1 & 11\ldots1 & 10 \\ \hline & & & 11 \\ I_{k-1} & I_{k-1} & 0 & 11 \\ & & & :: \\ & & & 11 \end{array}\right] \tag{3}$$

is a generator matrix of an optimal binary $[n, k, (n-k, 4, 4, \ldots, 4)]$ code.

IPR2018-1557
HTC EX1023, Page 309

Proof

Similar to the proof of construction A.

Construction D

For $p, q \in \mathbb{N}$, $p \geqslant q \geqslant 2$, the $p+q+2$ by $2p+3q+3$ matrix

$$\left[\begin{array}{c|c|c|c|c} 00\ldots\ldots\ldots\ldots0 & 1\,1\,1\,0 & 11\ldots\ldots1 & 00\ldots0 & \overset{q-1}{\longleftrightarrow}\ 11..1 \\ 00\ldots\ldots\ldots\ldots0 & 1\,1\,0\,1 & 00\ldots\ldots0 & 11\ldots1 & 11..1 \\ \hline & \begin{array}{c} 0\,1\,0\,1 \\ 0\,1\,0\,1 \\ :\ :\ :\ : \\ :\ :\ :\ : \\ 0\,1\,0\,1 \end{array} & & & \\ I_{p+q} & & I_p & 0 & 0 \\ & \begin{array}{c} 1\,0\,1\,0 \\ 1\,0\,1\,0 \\ :\ :\ :\ : \\ 1\,0\,1\,0 \end{array} & & & \\ & & 0 & I_q & 0 \end{array}\right] \tag{4}$$

is a generator matrix of an optimal binary

$$[2p+3q+3, p+q+2, (p+q+2, 2q+2, 4, 4, \ldots, 4)]$$

code.

Proof

Similar to the proof of construction A.

Construction E

For $p, q, r \in \mathbb{N}$, $p \geqslant 3$, $r \geqslant 2$, $q \geqslant r-p+2$, the p by $(2p+q+2r-4)$ matrix

$$\left[\begin{array}{c|c|c|c|c} 11\ldots\ldots1 & 00\ldots\ldots0 & 11\ldots1 & 00\ldots0 & 11..1 \\ 00\ldots\ldots0 & 00\ldots\ldots0 & 11\ldots1 & 11\ldots1 & 00..0 \\ \hline & & \begin{array}{c} 1 \\ 1 \\ :\ 0 \\ : \\ 1 \end{array} & \begin{array}{c} 1 \\ 1 \\ :\ 0 \\ : \\ 1 \end{array} & \\ I_{p-2} & I_{p-2} & & & 0 \\ & & \overset{r}{\longleftrightarrow} & \overset{r}{\longleftrightarrow} & \overset{q}{\longleftrightarrow} \end{array}\right] \tag{5}$$

is a generator matrix of an optimal binary

$$[2p+q+2r-4, p, (p+q+r-2, 2r, 4, 4, \ldots, 4)]$$

code.

IPR2018-1557
HTC EX1023, Page 310

Proof

Similar to the proof of construction A.

Construction F

For $p,q \in \mathbb{N}$, $p \geqslant 3$, $q \geqslant 2$, $q \geqslant p-2$, the p by $p+3q$ matrix

$$\left[\begin{array}{cc|c|c|c} 00\ldots.0 & 11\ldots.1 & 11\ldots1 & 00\ldots0 & 11\ldots.1 \\ 00\ldots.0 & 11\ldots.1 & 00\ldots0 & 11\ldots1 & 11\ldots.1 \\ \hline I_{p-2} & I_{p-2} & \begin{array}{c}1\\1\ 0\\ \vdots\\1\end{array} & \begin{array}{c}1\\1\ 0\\ \vdots\\1\end{array} & 0 \\ & & \overleftrightarrow{q+1} & \overleftrightarrow{q+1} & \overleftrightarrow{q-(p-2)} \end{array}\right] \tag{6}$$

is a generator matrix of an optimal binary

$$[p+3q, p, (2q+1, 2q+1, 4, 4, \ldots, 4)]$$

code.

Proof

Similar to the proof of construction A.

Construction G

For $p \in \mathbb{N}$, the $2p$ by $4p$ matrix

$$\left[\begin{array}{c|c|c|c} 00\ldots\ldots0 & 1\,1\,1\,0 & 11..1 & 00..0 \\ 00\ldots\ldots0 & 1\,1\,0\,1 & 00..0 & 11..1 \\ \hline & \begin{array}{c}1\,0\,1\,0\\ \vdots\ \vdots\ \vdots\ \vdots\\1\,0\,1\,0\end{array} & & \\ I_{2p-2} & & 0 & I_{p-1} \\ & \begin{array}{c}0\,1\,0\,1\\ \vdots\ \vdots\ \vdots\ \vdots\\0\,1\,0\,1\end{array} & & \\ & & I_{p-1} & 0 \end{array}\right] \tag{7}$$

is a generator matrix of a binary $[4p, 2p, (p+2, p+2, 4, 4, \ldots, 4)]$ code. For $p = 2, 3$ the codes are optimal, but in general they are not.

In ref. 6 the codes from construction G are treated extensively, the weight enumerators and automorphism groups are determined completely and a majority logic decoding method for these codes is given. For $p = 3$ we obtain a [12,6,(5,5,4,4,4,4)] optimal LUEP code. By deleting the row and column pairs (6,4), (5,3) and (4,2) successively we obtain [11,5,(5,5,4,4,4)], [10,4,(5,5,4,4)] and [9,3,(5,5,4)] optimal LUEP codes respectively.

IPR2018-1557
HTC EX1023, Page 311

Construction H

For $p \in \mathbb{N}$, $p \geqslant 3$, the $(p+2)$ by $(4p+1)$ matrix

$$\left[\begin{array}{c|c|c|c|c} 00\ldots0 & 11\ldots1 & 11\ldots1 & 00\ldots0 & 0 \\ 00\ldots0 & 11\ldots1 & 00\ldots0 & 11\ldots1 & 0 \\ \hline & & & & 1 \\ I_p & I_p & I_p & I_p & 1 \\ & & & & \vdots \\ & & & & 1 \end{array}\right] \tag{8}$$

is a generator matrix of a length-optimal binary

$$[4p+1, p+2, (2p, 2p, 5, 5, \ldots, 5)]$$

LUEP code.

Proof

It is easy to check that the code has the given parameters. By theorem 2*b* the length of a $(p+2)$-dimensional binary code with separation vector $(2p, 2p, 5, 5, \ldots, 5)$ is at least $4p+1$.

For $p = 3$ this construction gives a $[13, 5, (6, 6, 5, 5, 5)]$ optimal LUEP code. Furthermore table I refers to the following trivial constructions.

Construction I

For $p, q \in \mathbb{N}$, $p > q$, the 2 by $(p+2q)$ matrix

$$\left[\begin{array}{c|c|c} 11\ldots\ldots1 & 00\ldots.0 & 11\ldots.1 \\ 00\ldots\ldots0 & 11\ldots.1 & 11\ldots.1 \\ \longleftrightarrow & \longleftrightarrow & \longleftrightarrow \\ p & q & q \end{array}\right] \tag{9}$$

is a generator matrix of an optimal binary $[p+2q, 2, (p+q, 2q)]$ LUEP code.

Construction J

If the matrix G_1 has separation vector $s(G_1)$ such that $s(G_1)_k \geqslant 2$, then the matrix

$$G_2 = \left[\begin{array}{c|c} 0 & \\ 0 & \\ \vdots & G_1 \\ \vdots & \\ 0 & \\ \hline 1 & 100000\ldots\ldots0 \end{array}\right] \tag{10}$$

has separation vector $s(G_2) = (s(G_1), 2)$.

IPR2018-1557
HTC EX1023, Page 312

Construction K_i

If the matrix G_1 has separation vector $s(G_1)$ then the matrix

$$G_2 := [\, G_1 \mid e_i \,], \tag{11}$$

where e_i is the vector with a 1 on the i^{th} position and zeros elsewhere, has separation vector $s(G_2) = s(G_1) + e_i$.

The following theorem can be used to determine whether construction K_i gives an optimal code.

Theorem 3

If s is such that for all $t \geqslant s$, $t \neq s$, it holds that $n(t) > n(s)$ and if G is a generator matrix of a binary optimal $[r + n(s),k,(r,2s)]$ code, then the code generated by $[G \mid e_1 \mid e_1 \mid \ldots \mid e_1]$ (t times) is a binary optimal $[r + t + n(s),k,(r + t,2s)]$ code for t in IN arbitrary.

Proof

Let s and G fulfill the conditions mentioned above. By theorem 2*a* we have that

a) $n(r + t,2s) \geqslant r + t + n(s)$.
b) $n(r + t + 1,2s) \geqslant r + t + 1 + n(s) > r + t + n(s)$.
c) $n(r + t,2s + u) \geqslant r + t + n(\lceil s_1 + u_1/2 \rceil, \ldots, \lceil s_{k-1} + u_{k-1}/2 \rceil) \geqslant r + t + 1 + n(s)$ for $u \geqslant 0$, $u \neq 0$.

Combination of a), b) and c) shows that the code generated by $[G \mid e_1 \mid e_1 \mid \ldots \mid e_1]$ (t times) is optimal.

Construction L

Adding an overall parity-check bit to a binary $[n,k,s = (s_1, \ldots, s_k)]$ code gives a binary $[n + 1,k,s' = (2\lfloor (s_1 + 1)/2 \rfloor, \ldots, 2\lfloor (s_k + 1)/2 \rfloor)]$ code.

Sporadic constructions referred to in table I are the following.

Construction M

The 7 by 14 matrix

$$\begin{bmatrix} 00011111000000 \\ 00011000111000 \\ 00010100100101 \\ 00001010010011 \\ 00110001100000 \\ 01001010001000 \\ 10000000000111 \end{bmatrix} \tag{12}$$

IPR2018-1557
HTC EX1023, Page 313

is a generator matrix of an optimal binary [14,7,(5,5,5,5,4,4,4)] LUEP code. Deleting the first column and the last row from the matrix in (12) gives an optimal binary [13,6,(5,5,5,5,4,4)] code. Deleting the first two columns and the last two rows from the matrix in (12) gives an optimal binary [12,5,(5,5,5,5,4)] LUEP code.

Construction N

Application of [5,construction 1] with $m = 1$, $q = 2$ and G_1 a generator matrix of the [7,4,(3,3,3,3)] Hamming code gives an optimal binary [14,5,(7,6,6,6,6)] LUEP code.

Construction O

The 6 by 15 matrix

$$\begin{bmatrix} 000011111000111 \\ 000000000111111 \\ 100011000100100 \\ 010010100010010 \\ 001010010001001 \\ 000100001001001 \end{bmatrix} \tag{13}$$

is a generator matrix of an optimal binary [15,6,(7,6,5,5,5,4)] LUEP code.

Construction P

The 7 by 15 matrix

$$\begin{bmatrix} 000001111111000 \\ 000001110000111 \\ 100001001000100 \\ 010001001000010 \\ 001000100100001 \\ 000100100010001 \\ 000010100001001 \end{bmatrix} \tag{14}$$

is a generator matrix of an optimal binary [15,7,(7,6,4,4,4,4,4)] LUEP code.

Construction Q

By deleting the 8^{th} column from the matrix in (14) we obtain a generator matrix of an optimal binary [14,7,(6,5,4,4,4,4,4)] code.

Construction R

The 8 by 15 matrix

IPR2018-1557
HTC EX1023, Page 314

$$\left[\begin{array}{c|c} 0 & \\ 0 & \\ 0 & \\ 0 & G \\ 0 & \\ 0 & \\ 0 & \\ \hline 1 & 00000100010000 \end{array}\right] \tag{15}$$

where G is the matrix in (12), is a generator matrix of an optimal binary [15,8,(5,5,5,5,4,4,4,3)] LUEP code.

Construction S

The 8 by 15 matrix

$$\left[\begin{array}{l} 100000001100110 \\ 010000001010101 \\ 001000001001111 \\ 000100000110100 \\ 000010001110000 \\ 000001001101000 \\ 000000101011000 \\ 000000010111000 \end{array}\right] \tag{16}$$

is a generator matrix of an optimal binary [15,8,(5,5,5,4,4,4,4,4)] LUEP code.

REFERENCES

[1]) L. A. Dunning and W. E. Robbins, Inform. Contr. **37**, 150 (1978).
[2]) F. J. MacWilliams and N. J. A. Sloane, The theory of error-correcting codes, North-Holland Mathematical Library, Amsterdam, Vol. 16, 1978.
[3]) W. J. van Gils, EUT-Rep. 82-WSK-02, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, The Netherlands, June 1982.
[4]) W. J. van Gils, IEEE Trans. Inform. Theory **IT-29**, 866 (1983).
[5]) W. J. van Gils, IEEE Trans. Inform. Theory **IT-30**, 544 (1984).
[6]) L. M. H. E. Driessen, IEEE Trans. Inform. Theory **IT-30**, 392 (1984).

IPR2018-1557
HTC EX1023, Page 315

Philips J. Res. **39**, *305-316, 1984* *R 1098*

NETWORKS FOR SORTING WITH FUSION

by C. RONSE

Philips Research Laboratory, B-1170 Brussels, Belgium

Abstract

We consider a particular type of sorting where objects having the same weight go to the same destination. We call it *regular fusion-sorting*. We show that if in a sorting network built with sorting cells, we replace these cells by regular fusion-sorting cells, then the network becomes a regular fusion-sorting network. We give some applications, as the extraction of sets from multisets, occurrence counting in lists of objects, multiconnection networks.

CCCS: 42.30, 51.20 and 52.10.

1. Introduction and statement of the result

Sorting consists in reordering n objects $X_0, \ldots, X_{n-1}$ with respective weights $w_0, \ldots, w_{n-1}$, in such a way that the weights are in increasing order.

Sorting networks, in particular those built with binary cells, are widely used in computer science and communications [1,2]).

The purpose of this paper is to study a generalization of sorting where two objects having the same weight get the same place. Consider for example five objects X_0, X_1, X_2, X_3, X_4 having respective weights 6, 1, 3, 1, 6. With this type of sorting, the objects will be recorded as follows, with the ordering from top to bottom:

$$\begin{matrix} \{X_1, X_3\} \\ X_2 \\ \{X_0, X_4\}. \end{matrix} \qquad (1)$$

We call such a type of sorting *fusion-sorting*. We will study the design of switching networks which realize the operation of fusion-sorting; we call them *fusion-sorting networks*. In particular, we will consider *cellular* fusion-sorting networks; we mean by it that these networks are built by interconnecting small standard fusion-sorting networks called cells.

We should explain what the superposition of several object (e.g., $\{X_1, X_3\}$ and $\{X_0, X_4\}$ in (1)) means in practice. From a general hardware point of view, if each object is represented with k bits, the superposition of m objects will

IPR2018-1557
HTC EX1023, Page 316

form a km bit vector. Taking into account the fact that one should either reserve a value of each k-bit vector to the possible absence of object at a given output, or generate a validation binary signal indicating whether there is effectively an object at that output, the wiring capacity necessary to accommodate fusion-sorting would be very large. In this case it would be better to perform a sorting first, and then a fusion; we describe indeed a network for fusion in sec. 2.

However this general hardware point of view will be used only for the sake of the exposition and illustration of the possibilities of fusion-sorting. There are several particular cases where fusion-sorting does not lead to a significant increase in wiring. We list them briefly here, they will be described with more details in sec. 3:

(a) Suppose that two objects X_i and X_j have the same weight $w_i = w_j$ if and only if they are identical. When they are fused, it is not necessary to store at the output the superposition of the two objects, but only the single object $X_i = X_j$. It is thus only necessary to have some code for the absence of object (by reserving a special value of the k bit vector representing objects or by a validation signal).

(b) Although the connecting of several inputs to a single output is not a current practice in hardware design; the reverse is quite feasible. In multiconnection networks a single input can be connected to any number of outputs. We get thus a multiconnection network by turning a fusion-sorting network backwards. Here fusion corresponds to the connection of an input to several outputs. It is easily seen that no real wiring expansion is involved in this operation.

(c) One can find applications towards parallel list processing. Objects correspond to words in lists. The superposition of objects corresponds to lists. Then the operation of fusion is the concatenation of two or more lists into a larger one. Implementations would be possible with languages supporting some form of parallelism.

Let us now describe in more detail the working of fusion-sorting. As explained above, we use the hardware point of view, but only for exposition purposes.

Let n be the number of objects to sort. A sorting network has n *double* inputs (and outputs), each one being subdivided into one for the object and one for the weight. Then the sorting network will realize a sorting on both the objects and the weights. This is shown in fig. 1 for $n = 4$.

For a fusion-sorting network, things are more complicated. If at least two objects have the same weight, then the destination of the objects will not be uniquely determined, since we will require as set of outputs only a subset of the n available ones.

IPR2018-1557
HTC EX1023, Page 317

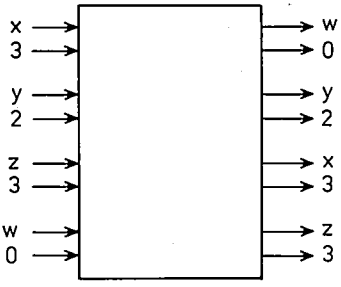


Fig. 1. A sorting 4-cell in action.

However, for technical reasons, the choice of the set of outputs will be determined by the number of repetitions of each distinct weight among the n objects.

Let us number the inputs and outputs $0, \ldots, n-1$. Write $X_0, \ldots, X_{n-1}$ for the objects, and $w_0, \ldots, w_{n-1}$ for their corresponding weights. We assume that these n weights form m distinct values

$$v_0, \ldots, v_{m-1}, \tag{2}$$

where $m \leqslant n$ and

$$v_0 < \ldots < v_{m-1}. \tag{3}$$

We define for $i = 0, \ldots, m-1$:

$$F_i = \text{the set of all } j\,(j = 0, \ldots, n-1) \text{ such that } w_j = v_i; \tag{4}$$

$$f_i = \text{the size of } F_i; \tag{5}$$

$$g_i = 0 \text{ if } i = 0;$$

$$= \sum_{k=0}^{i-1} f_k \text{ otherwise.} \tag{6}$$

For $j = 0, \ldots, n-1$, we must determine the destination of object x_j, in other words the output corresponding to input j. We write it j^δ.

One would like to have $j^\delta = i$ for $j \in F_i$ (so that only the first m outputs would be used), but this cannot be achieved with fusion-sorting networks derived from ordinary sorting networks, as we will see in the next section.

We have the smallest difference with an ordinary sorting network if we assume that the fusion-sorting network realizes an ordinary sorting on the

IPR2018-1557
HTC EX1023, Page 318

weights. Thus the weight v_i will be found at outputs $g_i, \ldots, g_i + f_i - 1$. But then for $j \in F_i$, j^δ will be one of these values. We can choose the lowest one

$$j^\delta = g_i \text{ for } j \in F_i. \tag{7}$$

With such a choice of δ, the fusion-sorting will be called *regular*. A network realizing a regular fusion-sorting on the objects and an ordinary sorting on the weights will be called a *regular fusion-sorting network*. We illustrate this in fig. 2 for $n = 8$.

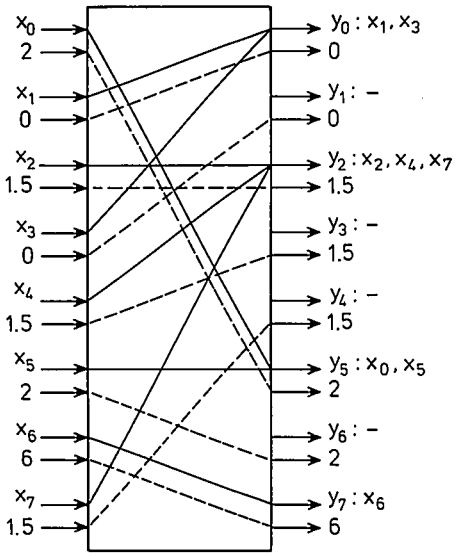


Fig. 2. A regular fusion-sorting 8-cell in action.

Now the interest in choosing this particular type of sorting resides in the following result:

Theorem

Let **N** be a sorting network built by an interconnection of smaller sorting networks $\mathbf{S}_0, \ldots, \mathbf{S}_c$. If we replace these smaller networks by regular fusion-sorting neworks $\mathbf{S}'_0, \ldots, \mathbf{S}'_c$, then the resulting network $\mathbf{N}'$ is a regular fusion-sorting network.

As a consequence, the numerous known designs for cellular sorting networks (triangular, diamond, etc.) built with binary cells can be used as regular fusion-sorting networks, provided that the binary cells are replaced by the 3-state cells shown in fig. 3.

We will prove our theorem in sec. 2, where we will also examine other forms of fusion-sorting and other designs of regular fusion-sorting networks. Applications will be discussed in sec. 3.

IPR2018-1557
HTC EX1023, Page 319

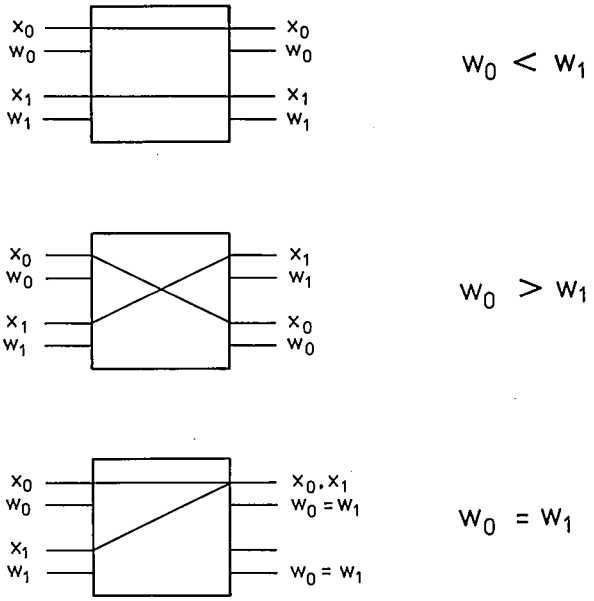


Fig. 3. A regular fusion-sorting 2-cell with its 3 states.

2. Discussion

Let us first demonstrate our theorem.

Let $i^* \in \{0, \ldots, m-1\}$ and $j^* \in F_i$. We will show that the destination of X_{j^*} through $\mathbf{N}'$ is g_{i^*}.

Let $\varepsilon > 0$ such that $v_{i^*} - \varepsilon > v_{i^*-1}$ if $i^* > 0$. We will compare the path followed by the objects and weights in the following two situations:

(i) with the network $\mathbf{N}'$;

(ii) with the network $\mathbf{N}$, where we change the weight $w_{j^*} = v_{i^*}$ of X_{j^*} to $w'_{j^*} = v_{i^*} - \varepsilon$.

We can subdivide $\mathbf{N}$ and $\mathbf{N}'$ into s successive stages, in each one of them every object passes through at most one subnetwork. After t stages ($0 \leqslant t \leqslant s$), the objects and weights have certain positions between 0 and $n-1$. We show the following by induction on t.

After t stages, there is an integer p_t with $0 \leqslant p_t \leqslant n-1$, such that:

(a) X_{j^*} has position p_t in both (i) and (ii), with weight w_{j^*} in (i) and w'_{j^*} in (ii).

(b) For any integer p with $0 \leqslant p \leqslant n-1$ and $p \neq p_t$, we have the same weight at position p in (i) and (ii).

This is obviously true for $t = 0$. Let us now show that if it is true for some $t \geqslant 0$, then it is true for $t+1$.

IPR2018-1557
HTC EX1023, Page 320

Consider a subnetwork $\mathbf{S}_d$ of $\mathbf{N}$ lying in stage $t+1$, and the corresponding subnetwork $\mathbf{S}'_d$ of $\mathbf{N}'$. Let q be its size, and write $x_0, \ldots, x_{q-1}$ and $y_0, \ldots, y_{q-1}$ for the levels corresponding to its q successive inputs and outputs respectively.

By induction hypothesis, for $u = 0, \ldots, q-1$ we have either $x_u = p_t$ and X_{j^*} has position x_u in both (i) and (ii), with respective weights w_{j^*} and w'_{j^*} or x_u has the same weight in both (i) and (ii). We can adapt (eqs 4, 5 and 6) to the present situation; we define for $i = 0, \ldots, m-1$:

Φ_i = the set of all x_u ($u = 0, \ldots, q-1$) such that in (i) the weight corresponding to the position x_u is v_i.

φ_i = the size of Φ_i.

γ_i = 0 if $i = 0$;

$$= \sum_{k=0}^{i-1} \varphi_k \text{ otherwise.}$$

In (ii), $x_u \in \Phi_i$ has weight v_i, except for $i = i^*$ and $x_u = p_t$, where it has weight $v_{i^*} - \varepsilon$.

Let us compare the outputs of $\mathbf{S}_d$ and $\mathbf{S}'_d$. In (i), $\mathbf{S}'_d$ operates an ordinary sorting on the weights. Thus for $i = 0, \ldots, m-1$ and $u = \gamma_i, \ldots, \gamma_i + \varphi_i - 1$, y_u has weight v_i. If p_t is one of $x_0, \ldots, x_{q-1}$ (i.e. if X_{j^*} is in one of the inputs of $\mathbf{S}'_d$), the output $y_{\gamma_{i^*}}$ will contain the object X_{j^*}.

In (ii), y_u will have the same weight as in (i), except for $u = \gamma_{i^*}$ when p_t is one of $x_0, \ldots, x_{q-1}$. In this case $y_{\gamma_{i^*}}$ has weight $v_{i^*} - \varepsilon$ and the object X_{j^*} will go to the output $y_{\gamma_{i^*}}$.

Thus in both cases we find the same value for p_{t+1}, and for every position $p \neq p_{t+1}$, one has the same weight in both (i) and (ii).

Thus our assertion is true for $t+1$.

Therefore our assertion is true for every t, and it is in particular true for $t = s$. This means that in (i) X_{j^*} goes to the output g_{i^*} and that for $i = 0, \ldots, m-1$ and $j = g_i, \ldots, g_i + f_i - 1$, the output j has weight v_i. As j^* was arbitrarily chosen, this means that $\mathbf{N}'$ operates a regular fusion-sorting on the object and an ordinary sorting on the weights.

This achieves the proof of our theorem.

Now let us examine alternative methods for fusion-sorting. In the preceding section, we said that it would be interesting to have a fusion-sorting network such that the destination of an object X_j would be i for $j \in F_i$. As we said there, we cannot do this with a network derived from an ordinary sorting network.

IPR2018-1557
HTC EX1023, Page 321

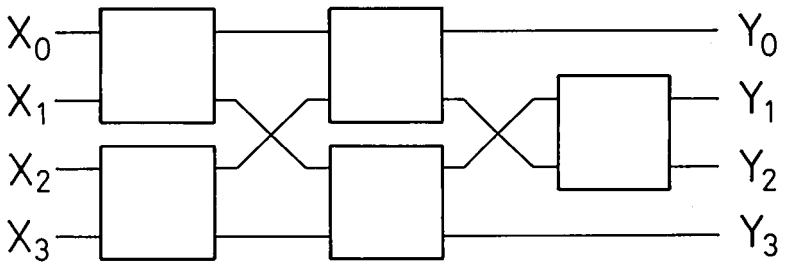


Fig. 4.

We give a counterexample here. Take $n = 4$ and consider the cellular sorting network shown in fig. 4. Suppose that we take the weights,

$$w_0 = 0,$$
$$w_1 = 1,$$
$$w_2 = 0,$$
and $$w_3 = 2. \tag{8}$$

The destination j^δ of object of x_j would then be given by,

$$0^\delta = 0,$$
$$1^\delta = 1,$$
$$2^\delta = 0,$$
and $$3^\delta = 2. \tag{9}$$

It is not hard to see that it is impossible to realize (9) with that network after replacing the 2-state cells by 3-state fusion-sorting cells. This is also impossible with 4-state cells (with the 4^{th} state being the 3^{rd} one turned upside down), even with a centralized (external) control of the cells.

It is thus justified to restrict the design of a cellular fusion-sorting network from an ordinary one to the regular fusion-sorting network described in the theorem.

Now it is possible to realize the building of a regular fusion-sorting network by combining two networks in succession: the first one is an ordinary sorting network, while the second one will realize a *regular fusion*, which is an input to output mapping φ such that for $j = 0, \ldots, n-1$ we have:

$$j^\varphi \leqslant j$$
and $$(j^\varphi)^\varphi = j^\varphi. \tag{10}$$

In fact, given eqs (4), (5) and (6), one defines $j^\varphi = g_i$ for $j = g_i, \ldots, g_i + f_i - 1$.

IPR2018-1557
HTC EX1023, Page 322

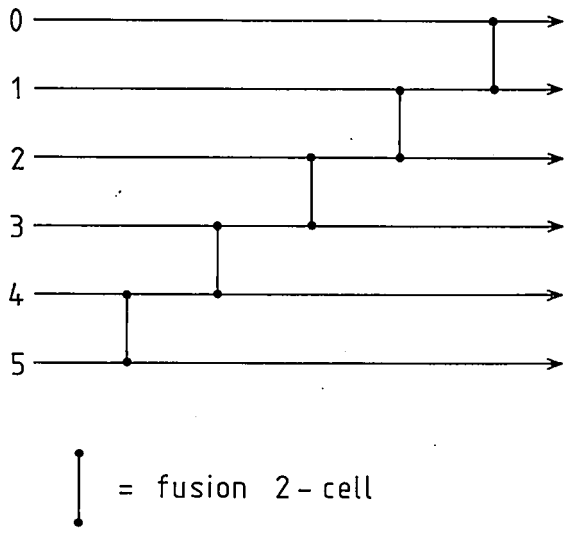


Fig. 5. A regular fusion network, $n = 5$.

This second network will be called a *regular fusion network*. There are several possible cellular designs for it. Let us give here two examples using binary cells.

A fusion 2-cell has 2 states, the first and third ones from fig. 3. Now the first design of a regular fusion network using 2-cells is derived from the 'branching network' described in ref. 3; we show it in fig. 5. The second design is based on the fact that the fusion of the numbers $j+1, \ldots, j+t$ with j can be realized as follows: write

$$t = \sum_{r=0}^{s} \alpha_r 2^r,$$

where $\alpha_r \in \{0, 1\}$ for reach r; then this fusion can be realized by fusing $j+t$ with $j+t-\alpha_s 2^s$, then $j+t-\alpha_s 2^s$ with $j+t-\alpha_s 2^s - \alpha_{s-1} 2^{s-1}$, etc. If $2^k < n \leqslant 2^{k+1}$, then the network can be realized by a succession of $2k+1$ stages numbered $1, \ldots, 2k+1$:

- In stage 1, a fusion cell is placed between j and $j+2^k$ for $0 \leqslant j \leqslant n - 2^k$.
- In stage $2v$ ($v = 1, \ldots, k$), a fusion cell is placed between j and $j + 2^{k-v}$ for $0 \leqslant j \leqslant n - 2^{k-v}$ and $j \equiv 0, \ldots, 2^{k-v} - 1$ modulo 2^{k-v+1},
- In stage $2v+1$ ($v = 1, \ldots, k$), a fusion cell is placed between j and $j + 2^{k-v}$ for $0 \leqslant j \leqslant n - 2^{k-v}$ and $j \equiv 2^{k-v}, \ldots, 2^{k-v+1} - 1$ modulo 2^{k-v+1}.

We illustrate this design in fig. 6 for $n = 10$. When n is a power of 2, this network can be given a structure similar to Lawrie's omega network [4])

IPR2018-1557
HTC EX1023, Page 323

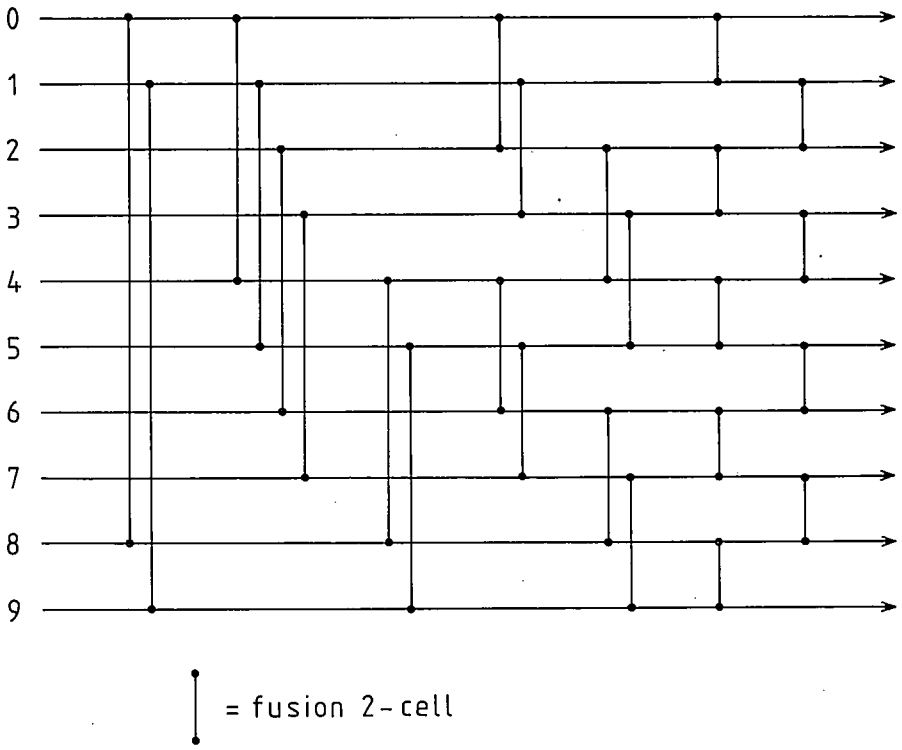


Fig. 6. A regular fusion network, $n = 10$.

if one precedes the stages 1, 2, 4, 6, etc. by a perfect shuffle. We show this in fig. 7.

The first design has cost and delay (in terms of 2-cells) linear in n, while the second one has cost and delay linear in $n \cdot \log(n)$ and $\log(n)$ respectively.

The question arises: when is the design of a regular fusion-sorting network by the modification of cells in an ordinary sorting network (as explained in the theorem) more appropriate than the juxtaposition of an ordinary sorting network with a regular fusion network described here? The answer depends upon the circumstances. If the flow of several objects at the same position in a cell leads to an increase of wiring capacity or of processing time, then the combination of a sorting network and a fusion network is preferable, since the complication is restricted to the end of the design. On the other hand, we will describe in the next section several applications (multiconnection networks, extraction of sets from multisets) where this complication does not arise; then it will be better to use the design based on our theorem.

IPR2018-1557
HTC EX1023, Page 324

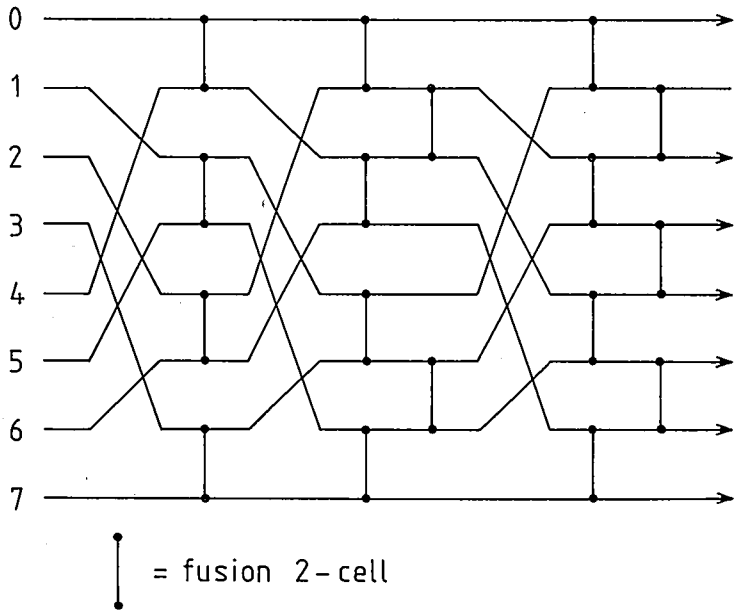


Fig. 7. A regular fusion network, $n = 8$.

3. Applications

Fusion-sorting networks can be used for the extraction of a 'dictionnary' (i.e., the list of all words in alphabetical order) from a text. This operation is equivalent to the sorting of a multiset into a set. It can be achieved by operating a fusion-sorting of the word occurrences, with their respective weights being in alphabetical order.

It is also possible to count the number of occurrences of every word by associating to each word an occurrence count (initialized to 1 at the input) and adding a counter to each fusion-sorting 2-cell. Suppose that for the 2-cell shown in fig. 3 the occurrence counts of its two input words are c_0 and c_1. Then the occurrence counts of its two outputs will be

c_0 and c_1 in the first state.
c_1 and c_0 in the second state.
$c_0 + c_1$ and 0 in the third state.

It should be noted that in the above example the fusion operation does not cause an increase in wiring capacity or processing time, because when two words are 'fused', the corresponding output does not contain the repetition of two identical words, but one word.

IPR2018-1557
HTC EX1023, Page 325

Fusion-sorting networks turned backwards can be used as multiconnection networks (a multiconnection is an input-output connection where every output is connected to one input, but an input can be connected to any number of outputs). A multiconnection realizes a map from the outputs to the inputs, while a fusion-sorting realizes a map from the inputs to the outputs. Thus if we turn a regular fusion-sorting network backwards, then it can realize multiconnections in the following way: we associate to every output a weight corresponding to the input to which it must be connected. Then by performing a regular fusion-sorting from the outputs to the inputs, we will obtain a multiconnection which is similar to the desired one, up to a reordering of the inputs. This reordering can be achieved by inserting a permutation network on the side of the inputs. We illustrate this design in fig. 8*a*.

This type of multiconnection networks, combining a permutation network followed by a regular fusion-sorting network turned backwards, is to be compared to the basic design proposed in ref. 3, which consists in the combination of three networks, namely a 'branching network' enclosed between two permutation networks. This is shown in fig. 8*b*. A 'branching network' is a multiconnection network realizing only a particular subset of multiconnections; it corresponds in fact to a regular fusion network turned backwards. It is thus easy to see that this branching network and the permutation network on the side of the outputs correspond to our regular fusion-sorting network turned backwards.

Other applications of fusion-sorting can be imagined. For example in a parallel list processor, a fusion-sorting network could be used for sorting or

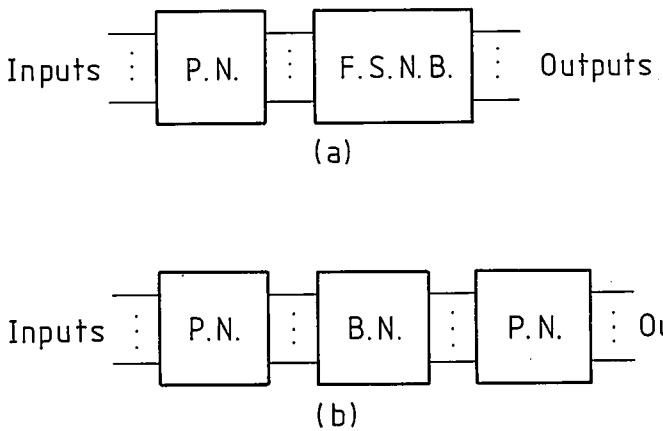


Fig. 8. Multiconnection networks: P.N. = permutation network, B.N. = branching network, F.S.N.B. = fusion-sorting network turned backwards.

IPR2018-1557
HTC EX1023, Page 326

concatenating lists. Given two lists (list 0) and (list 1), the fusion-sorting 2-cell of fig. 3 will apply to them the following operations:

— No change in the first state.
— The inversion between (list 0) and (list 1) in the second state.
— The concatenation of the two lists in the third state; (list 0) becomes (list 0 list 1) and (list 1) becomes the empty list ().

This can be done efficiently if these list are coded dynamically. We take n position records $P_0, \ldots, P_{n-1}$, containing each two address pointers *begin* and *end*, n object records $Z_0, \ldots, Z_{n-1}$, containing each the description of an object X_i, its weight w_i and an address pointer *next*, and finally an empty record *NIL*. For each P_i, the pointers *begin* and *end* give the address of the records Z_j corresponding to the first and last elements of the corresponding list. For each Z_i, the pointer *next* gives the address of the record Z_j such that X_j is the successor of X_i in its corresponding list. When the list associated to P_i is empty or when Z_j corresponds to the last element of a list, the corresponding pointers *begin*, *end* and *next* are set or *NIL*. At the start, we set in each P_i the pointers *begin* and *end* to Z_i, and in each Z_i the pointer *next* to *NIL*. Then the operation of fusion-sorting on two lists corresponding to positions P_i and P_j can be performed by manipulations of the pointers *begin*, *end* and *next*. At the end of the process, the lists can be accessed thanks to the chains of *begin* and *next* pointers.

4. Conclusion

By our theorem, we know that every cellular design for sorting networks (triangular, diamond, Batcher's, Base and Nelson's, etc.) can be used for the design of regular fusion-sorting networks.

It is also possible to build regular fusion-sorting networks by combining an ordinary sorting network with a regular fusion network. We give two different designs for a regular fusion network built from binary cells.

There are several applications of regular fusion-sorting which do not lead to an increase of wiring complexity due to the fusion operation, for example the design of multiconnection networks or the extraction of a set from a multiset.

Acknowledgement

The author is indebted to M. J. Fischer for suggesting a shortening of the proof of the theorem.

REFERENCES

[1]) D. E. Knuth, The art of computer programming, Vol. 3, Sorting and Searching, Addison-Wesley, Reading Mass., 1973.
[2]) Computer 12, no. 6 (1979).
[3]) S. Sakata, H. Hirayama and Y. Shibata, Electronics and Communications in Japan 58-A, 51 (1975).
[4]) D. H. Lawrie, IEEE Trans. Computers C-24, 1145 (1975).

IPR2018-1557
HTC EX1023, Page 327

AUTHOR INDEX TO VOLUME 39

IPR2018-1557
HTC EX1023, Page 328

IPR2018-1557

HTC EX1023, Page 329

IPR2018-1557
HTC EX1023, Page 330