As you are now the owner of this document which should have come to you for free, please consider making a donation of £1 or more for the upkeep of the (Radar) website which holds this document. I give my time for free, but it costs me in excess of £300 a year to bring these documents to you. You can donate here https://blunham.com/Radar

Please do not upload this copyright pdf document to any other website. Breach of copyright may result in a criminal conviction.

This document was generated by me Colin Hinson from a document held at Henlow Signals Museum. It is presented here (for free) and this version of the document is my copyright (along with the Signals Museum) in much the same way as a photograph would be. Be aware that breach of copyright can result in a criminal record.

The document should have been downloaded from my website https://blunham.com/Radar, if you downloaded it from elsewhere, please let me know (particularly if you were charged for it). You can contact me via my Genuki email page: https://www.genuki.org.uk/big/eng/YKS/various?recipient=colin

You may not copy the file for onward transmission of the data nor attempt to make monetary gain by the use of these files. If you want someone else to have a copy of the file, please point them at the website (https://blunham.com/Radar).

Please do not point them at the file itself as the file may move or be updated.

I put a lot of time into producing these files which is why you are met with this page when you open the file.

In order to generate this file, I need to scan the pages, split the double pages and remove any edge marks such as punch holes, clean up the pages, set the relevant pages to be all the same size and alignment. I then run Omnipage (OCR) to generate the searchable text and then generate the pdf file.

Hopefully after all that, I end up with a presentable file. If you find missing pages, pages in the wrong order, anything else wrong with the file or simply want to make a comment, please drop me a line (see above).

It is my hope that you find the file of use to you personally – I know that I would have liked to have found some of these files years ago – they would have saved me a lot of time !

Colin Hinson
In the village of Blunham, Bedfordshire.

RRE JOURNAL

APRIL 1959

R.R.E. JOURNAL APRIL 1959

R. R. E. JOURNAL

CONTENTS

Issued by the Royal Radar Establishment, Ministry of Supply

No. 43 April 1959

THE TRAVELLING-WAVE MASER

by P. N. Butcher

1. Introduction

One fundamental discovery led to the development of the solid-state maser (microwave amplification by stimulated emission of radiation). It was found that certain solids can be maintained in an active condition with respect to a microwave frequency band, i.e. they amplify radiation falling on them in this band instead of attenuating it. The solids which have this remarkable property are certain crystals doped with paramagnetic ions belonging to the transition groups of the Periodic Table (Fe^{3+}, Cr^{3+}, Ni^{2+}, Gd^{3+} etc.) which have a permanent magnetic moment. The paramagnetic ions are the active constituents of the crystal. Hereafter we usually call them simply 'the ions'; leaving understood the fact that they are paramagnetic. The surrounding crystal serves to hold the ions in place and modifies their behaviour in various useful ways. The relevant properties of the ions are described in Section 2. The discussion is confined to ions from the iron transition group (Sc, Ti, V, Cr, Mn, Fe, Co, Ni) which are the most important in practical applications.

The ions can be activated in a variety of ways.[1] The most successful method is the continuous microwave pumping technique suggested by Bloembergen[2]. The technique is described in Section 3 where the quantum theory of maser action is outlined qualitatively. We do not go into very much detail about the quantum theory because the classical treatment developed in Section 4 is far more illuminating. It gives one an intuitive grasp of the non-reciprocal behaviour of the paramagnetic ions which is exploited in the travelling-wave maser.

Once it was appreciated that a paramagnetic crystal could be maintained in an active conditions, making a practical amplifier out of it was a relatively easy step. One simply has to put the active crystal into some sort of structure which carries the signal to the crystal and takes it away again after it has been amplified. The obvious structure to choose is a resonant cavity with waveguide feeds. All but one of the solid-state masers built to date have been of this type. However, the cavity maser has several disadvantages: bi-directional gain, instability, high gain-sensitivity, narrow bandwidth

and tuning difficulties. It has been recognised for some time that the solution to these problems lies in distributing the active crystal along a waveguide instead of putting it in a cavity. The maser group at Bell Telephone Laboratories has recently built a travelling-wave maser with a performance which completely fulfils the theoretical predictions[3].

The waveguide in a travelling-wave maser must have two particular properties: extremely low group velocity (less than 1% of the velocity of light) and distinct regions of contra-rotating circularly polarised magnetic field. In Section 5 we show how these characteristics can be achieved by using an array of parallel wires to guide the wave, taking as a specific example the comb structure used at Bell Telephone Laboratories. Section 6 is concerned with the engineering characteristics of the travelling-wave maser: forward gain, reverse loss, gain sensitivity, bandwidth and noise figure. It is probably unnecessary to remind the reader that the current interest in masers has arisen because they have noise figures of the order of 0.1 db.

2. Spin, Magnetic Moment and Energy Levels

The electron is not just a point having mass $m = 0.91 \times 10^{-27}$ gm. and charge of magnitude $e = 4.8 \times 10^{-10}$ e.s.u. It is a particle of finite size which is always spinning about an internal axis. The spin can be described by drawing a vector $\underline{s}$ pointing along the axis of rotation and having a magnitude equal to $\sqrt{3}/2$, which is the angular momentum of the electron in units of $\hbar$, Planck's constant divided by 2π (see Fig. 1). Now $\hbar = 10^{-27}$ erg-sec. which is an extremely small angular momentum by macroscopic standards. For example, an electron moving in a circular orbit of radius 1 cm. with velocity 10^8 cm/sec. (kinetic energy 2.8 eV.) has an 'orbital' angular momentum of $mvr = 0.91 \times 10^{-19}$ erg-sec. We may therefore neglect spin when discussing the macroscopic motion of electrons. The situation is quite different when several electrons come together around a nucleus to form a stable ion. The orbit radii are then about 10^{-8} cm. and the kinetic energies are of the order of 2.8 eV. Hence $mvr = 0.91 \times 10^{-27}$ erg-sec. which has the same order of magnitude as the spin.

For an ion it is no longer legitimate to neglect the spin in comparison to the orbital angular momentum. In fact, when the ion belongs to the iron transition group and is embedded in a crystal lattice, precisely the reverse is true: we may neglect the orbital angular momentum in comparison to the spin. This situation occurs in the two most successful maser crystals: aluminium oxide doped with Cr^{3+} ions (ruby) and potassium cobalti-cyanide doped with Cr^{3+} ions.

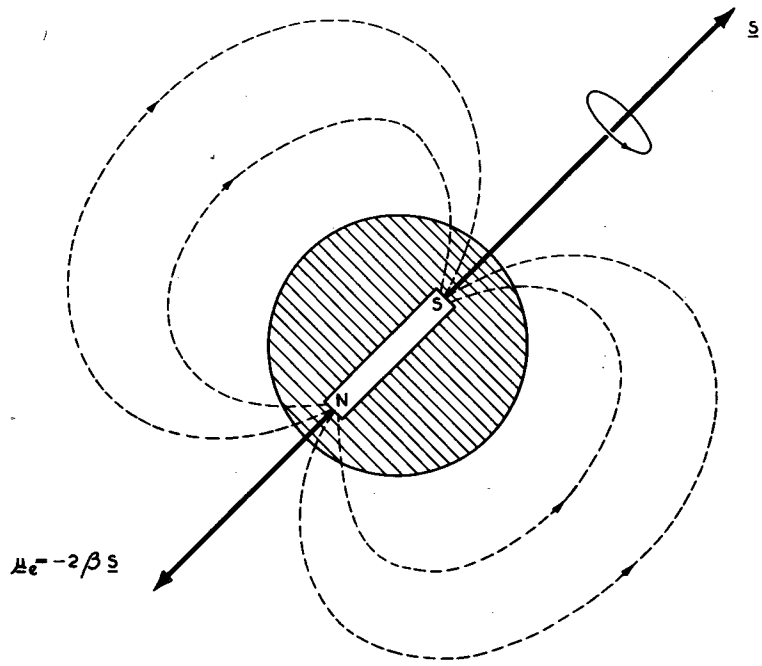


Figure 1 The spinning electron

It comes about because the electrons re-orient their orbits in such a way that their total orbital angular momentum vanishes. They do this in an effort to reduce their interaction with the surrounding crystal. As one might expect, the spins tend to associate in anti-parallel pairs in an attempt to make the total spin vanish as well. The characteristic property of the ions of the iron transition group is that complete pairing off of the spins is prevented by the combined effects of the Pauli exclusion principle and the Coulomb interaction between the electrons. (Parallel spins cannot approach one another and so have a low Coulomb interaction energy in comparison to anti-parallel spins which can approach one another). These ions have a residual total spin which can be described by a (total) spin vector $\underline{S}$ in the same way as we described the spin of an individual electron. The magnitude of $\underline{S}$ is equal to $\sqrt{S(S+1)}$ where $S = 1/2, 1, 3/2$ etc. is one half of the number of unpaired electronic spins in the ion. (For an isolated electron $S = 1/2$ and $|\underline{S}| = \sqrt{3}/2$. For Cr^{3+} $S = 3/2$ and $|\underline{S}| = \sqrt{15}/2$.) It would take us too far afield to investigate the reason for this curious formula for the magnitude of the spin vector which has its origin in the Uncertainty Principle. The number S is characteristic of the ion and is simply called its spin, since no confusion with the spin vector $\underline{S}$ is likely to arise.

It might perhaps seem more appropriate to call $\sqrt{S(S+1)}$ the spin, but S is easier to remember and it is more useful in calculations.

Now that we have given up the idea that the electron is a point particle, we might try to picture it as a tiny spinning ball of charge. The spinning charge creates a magnetic field which can be calculated from Maxwell's equations. One finds that the magnetic field is just the field which would be produced by a little bar magnet lying along the spin direction and having a magnetic moment (pole strength times length) equal to $(e/2mc)\hbar|\underline{s}|$ erg/gauss in magnitude where $c = 3 \times 10^{10}$ cm./sec. is the velocity of light. (See Fig. 1). The quantity $(e\hbar/2mc)$ is called the Bohr magneton and is denoted by $\beta = 0.93 \times 10^{-20}$ erg/gauss. The magnet has its North pole pointing in the opposite direction to $\underline{s}$ because the electronic charge is negative (e denotes the magnitude of the electronic charge). We may describe the magnetic properties of this model of the electron by drawing a 'magnetic moment vector' $\underline{\mu}_e = -\beta\underline{s}$, pointing in the opposite direction to $\underline{s}$ and having a magnitude equal to $\beta|\underline{s}|$. In so doing we would be wrong in only one respect: our simple model gives a magnitude to the magnetic moment which is too small by a factor of 2, the famous Landé g-factor for electron spin.[4] In fact the magnetic moment (vector) of an electron is

$$\underline{\mu}_e = -2\beta\underline{s} \tag{1}$$

as shown in Fig. 1.

We are concerned with whole ions rather than isolated electrons. By adding together the magnetic moments of all the electrons, we obtain for the magnetic moment of an ion with spin vector $\underline{S}$

$$\underline{\mu}_i = -2\beta\underline{S} \tag{2}$$

Ions of the transition group elements have a permanent magnetic moment because they have a non-zero spin vector. The ions of all other elements have zero spin and no permanent magnetic moment.

The significance of the permanent magnetic moment becomes apparent when we consider what happens when the ions are placed in a d.c. magnetic field $\underline{H}_o$. Suppose, for simplicity, that $\underline{H}_o$ is directed along the z-axis of a Cartesian coordinate system and has magnitude H_o. Then the energy which the ion has in this field is, as for a bar magnet,[4]

$$\begin{aligned} E &= -H_o\,\mu_{iz} \\ &= 2\beta H_o S_z \end{aligned} \tag{3}$$

from equation (2). The primary distinction between a macroscopic

bar magnet and an ion is that the bar magnet can take up any orientation with respect to the field while the orientation of the ion is quantised. The quantisation rule is: S_z must be integral when the spin S is integral and S_z must be half-integral when S is half-integral. By taking into account the obvious geometrical restriction that $|S_z|$ must be less than or equal to $|\underline{S}| = \sqrt{S(S+1)}$, we find that the permissible values of S_z are $-1/2$, $+1/2$ when $S = 1/2$; $-1, 0, +1$ when $S = 1$; $-3/2, -1/2, +1/2, +3/2$ when $S = 3/2$ and $-S$, $1-S$, $2-S$ ..... $S-1$, S in the general case. There are $2S + 1$ possible values of S_z going up from $-S$ in steps of one. Hence the ion has a series of $2S + 1$ equally spaced energy levels going up from $-2\beta H_o S$ in steps of $2\beta H_o$. Figure 2 is a sketch of the possible energy levels in the cases $S = 1/2$ (Ti^{3+}), 1(Ni^{2+}) and $3/2$ (Cr^{3+}). The energy levels are labelled by the values of S_z to which they correspond. With one trivial exception, this convention will be adhered to throughout the paper.

So far we have only taken account of the applied magnetic field. There is also an internal magnetic field produced by the electrons themselves as they move in their orbits.[5] The magnetic field produced by the orbital motion of <u>all</u> the electrons vanishes outside the ion because their total orbital angular momentum and magnetic moment both vanish. Inside the ion, however, the details of the orbital motion come into play and each electron experiences a finite magnetic field produced by all the others. The effect which the internal magnetic field has on the energy of the ion depends upon the angle which it makes with the total spin. Different energy levels in the external field correspond to different orientations of the total spin and so the internal field will affect them differently. Hence, one actually finds an unequally spaced series of energy levels as indicated in Fig. 2 for the case $S = 3/2$. The diagram applies to some particular direction of the applied field relative to the crystal. For some other direction of the applied field the energy levels will be different because there will be a change of the angles between the directions of the total spins of the various levels (which are, roughly speaking, determined by the applied field) and the direction of the internal magnetic field (which is determined by the electronic orbits and so, in the last analysis, by the surrounding crystal). Thus, because of the internal magnetic field, the energy levels are unequally spaced and depend upon the direction of the applied field as well as its magnitude. Unequally spaced energy levels are essential for maser action. The anisotropy of the energy levels is a help when tuning the maser.

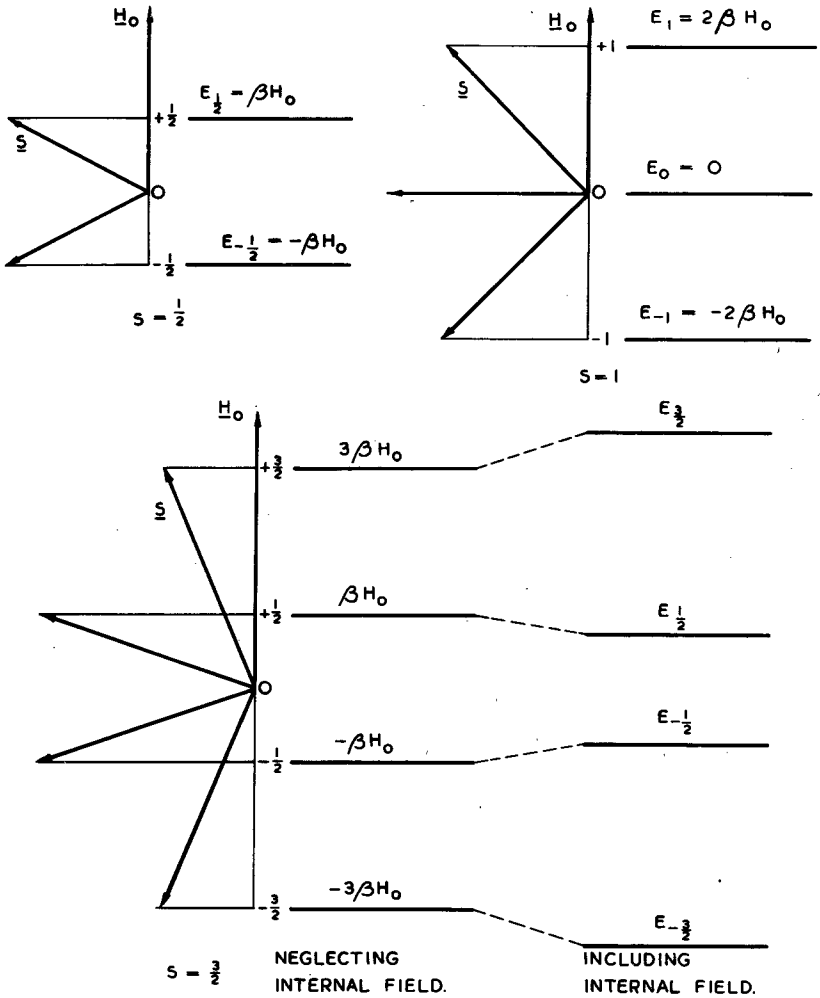


Figure 2 Energy levels of an ion in a magnetic field

3. Populations, Transitions, Pumping and Maser Action

At absolute zero the ions achieve the lowest possible total energy by all residing in the lowest energy level. At a temperature T greater than $0^{\circ}K$ some of the ions are shaken out of the lowest energy level by the thermal vibrations of the crystal. The fraction of ions in the level with $S_z = M$ is proportional to $\exp(-E_M/kT)$ where $k = 1.4 \times 10^{-16}$ erg/degree is Boltzmann's constant. Now E_M/kT is approximately equal to $M2\beta H_o/kT = 0.16$ when $M = \frac{1}{2}$, $H_o = 3000$ gauss, and $T = 1.25^{\circ}K$. As a rough approximation we may therefore expand

the exponential as a power series and keep just the first two terms. Since there are 2S + 1 levels in all, the fraction of ions in the level with $S_z = M$ is

$$\rho_M^o = \frac{1 - E_M/kT}{\sum_M (1 - E_M/kT)}$$

$$\simeq \frac{1 - E_M/kT}{2S + 1} \qquad (4)$$

It is interesting to digress for a moment in order to see why ions of the transition group elements are called paramagnetic ions. In thermal equilibrium the ions prefer to occupy the lower energy levels, i.e. there are more ions with positive values of μ_{iz} than there are with negative values of μ_{iz}. Consequently the total magnetic moment per unit volume of the crystal has a positive z-component. It is parallel to the applied field and the crystal is paramagnetic in the usual sense of the term. The paramagnetism is entirely due to the permanent magnetic moment of the transition group ions. Ions from other groups of the Periodic Table have no permanent magnetic moment. Lenz's law governs their behaviour in an applied magnetic field. The electronic orbits precess so as to set up a magnetic moment opposing the applied field, i.e. the ions are diamagnetic.

Paramagnetic ions can be induced to jump from one level to another by subjecting them to a magnetic field oscillating at an angular frequency equal to the energy difference divided by $\hbar$. For adjacent levels the transition frequency is (neglecting the internal magnetic field) $2\beta H_o/\hbar = 2\pi \times 2.8 H_o$ megaradians/sec., which is a microwave frequency when H_o is a few thousand gauss. It would appear from what has been said that the ions only respond to a microwave magnetic field at precisely the transition frequency. This is not true because there is yet another internal magnetic field acting on the ions; the field which the ions produce outside themselves by virtue of their spin magnetic moments.[6] Each ion experiences the magnetic field produced by all the other ions and any nuclear magnetic moments which may be in the crystal. This field is different at the sites of different ions and so the energy levels are spread out over a band. The ions therefore have a bell-shaped frequency response curve, the precise shape of which is determined by the density of levels in the various parts of the energy bands. It is called the line shape and is about 50 Mc/sec. wide in good maser crystals.

When a particular transition is excited by microwave radiation with angular frequency ω, ions in the lower level jump up to the upper level by absorbing a quantum $\hbar\omega$ of radiation energy and ions in the upper level drop down to the lower level by emitting a quantum of radiation energy. The transition rate per ion, V, is the same in both directions.[1] Consequently the net power absorbed from the radiation per unit volume of crystal is

$$P_m = \hbar\omega . N(\rho_\ell - \rho_u) . V \qquad (5)$$

where N is the number of ions per unit volume, ρ_ℓ is the fraction of ions in the lower level and ρ_u is the fraction of ions in the upper level. The important thing to notice is that the absorbed power is proportional to the 'population difference' across the transition $N(\rho_\ell - \rho_u)$.

When the ions are in thermal equilibrium, $\rho_\ell > \rho_u$ and there is a net positive absorption of radiation, i.e. the crystal is passive. In a maser the thermal equilibrium populations are upset in such a way that $\rho_\ell < \rho_u$. The incident radiation then stimulates a net negative absorption, i.e. a net emission of extra radiation and the crystal is active. There are several ways of achieving this condition.[1] The most satisfactory method is that suggested by Bloembergen[2] which utilises three energy levels of an ion with a spin of one or more. Let us number these levels 1, 2 and 3 in order of increasing energy (see Fig. 3). The signal to be amplified induces

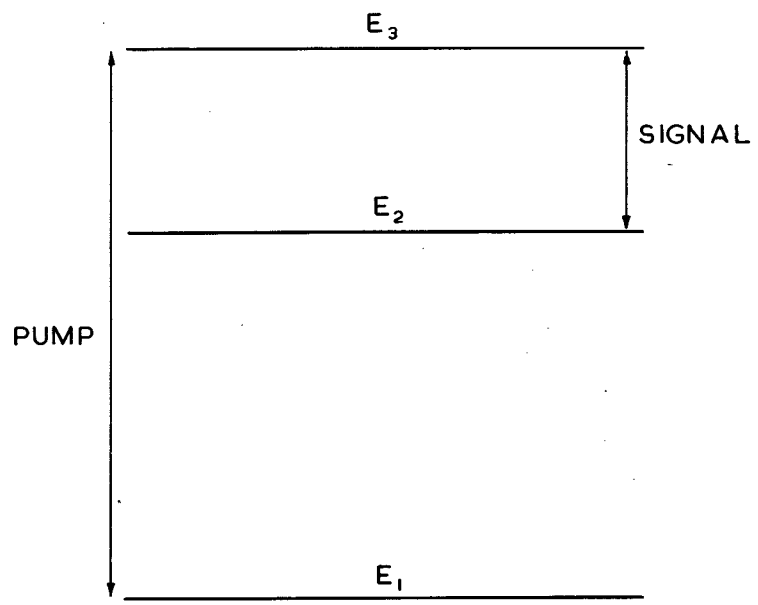


Figure 3 The three-level maser

transitions between levels 2 and 3. The population of level 3 is made greater than that of level 2 by 'pumping' ions up from level 1 with an intense microwave field at the resonant frequency of the 1 to 3 transition.

The pumping action comes about as follows. When the pump field has a low intensity the populations are not much disturbed from their thermal equilibrium values. The ions therefore absorb (positive) power from the pump because more ions jump up from level 1 to level 3 than drop down from level 3 to level 1. Obviously the population of level 3 is increased by the absorption process. It is prevented from increasing indefinitely by the thermal vibrations of the crystal which do their best to restore thermal equilibrium. The net result of the competition between the pump and the thermal vibrations is that the populations reach steady values with rather more ions in level 3 and rather less ions in level 1 than is the case in thermal equilibrium. As the pump intensity is raised the population of level 3 increases and the population of level 1 decreases until, when the pump completely overcomes the thermal vibrations, both levels have the same populations.

Efficient pumping of suitable crystals makes the population level 3 greater than that of level 2 and the ions become active at the signal frequency. Indeed, when levels 1 and 3 are equally populated, the population differences across the 2 to 3 transition and the 1 to 2 transition are equal in magnitude and opposite in sign. Hence the crystal is active either at the frequency of the 2 to 3 transition (as is assumed in Fig.3) or at the frequency of the 1 to 2 transition. Efficient pumping activates the crystal at one frequency or the other (disregarding the unlikely event of equal populations in all three levels). The three-level maser scheme fails, in principle, only when the energy levels are equally spaced. In that case the signal stimulates both the active transition and the passive transition. The net effect is a small absorption or emission depending on which transition has the largest transition rate. Thus, the unequally spaced energy levels produced by the internal magnetic field are essential for maser action.

We can easily estimate the fractional population difference across the 2 to 3 transition when the pump field is intense. Let ρ_1^o, ρ_2^o and ρ_3^o be the fractional populations of the three levels in thermal equilibrium. If the pump only affects the populations of levels 1 and 3, then the populations after pumping are

$$\rho_3 = \rho_1 = \tfrac{1}{2}(\rho_1^o + \rho_3^o)$$

$$\rho_2 = \rho^o{}_2$$

Hence, by using equation (4) with a slight change of notation, we obtain

$$\rho_2 - \rho_3 = \rho_2^o - \tfrac{1}{2}\left(\rho_1^o + \rho_3^o\right)$$

$$= \left[\tfrac{1}{2}(E_1 + E_3) - E_2\right] \Big/ \left[(2S + 1)kT\right] \qquad (6)$$

which is negative provided level 2 is nearer to level 3 than it is to level 1. The numerator in this expression is typically of the order of $-\frac{1}{2}\beta H_o$ so that $\rho_2 - \rho_3 \sim -\frac{1}{2}\beta H_o / \left[(2S + 1)kT\right] = -2.0\%$ when $H_o = 3000$ gauss, $S = 3/2$ and $T = 1.25^oK$. Clearly the population difference after pumping has the same order of magnitude as the thermal equilibrium population difference, but of course its sign has been changed by the pump.

The power necessary for efficient pumping is determined by: the strength of the coupling between the thermal vibrations and the ions, the amplitude of the thermal vibrations and the density of ions. Some ions are more weakly coupled to the thermal vibrations than others. Cr^{3+} in ruby or potassium cobalti-cyanide is particularly good in this respect. Many crystals, which might otherwise be suitable maser materials, have to be discarded because the ions are coupled too tightly to the thermal vibrations. The amplitude of the thermal vibrations can be reduced by cooling the crystal. As has been indicated above, it is usual to cool to liquid helium temperatures. Maser action has been observed in ruby at 60^oK, but the effect is less pronounced because the temperature appears in the denominator of expression (6) for the fractional population difference across the active transition.[7] The density of ions also plays a significant role in determining the pump power. As the density is raised, the ions come closer together and couple more strongly through their magnetic fields. It then becomes more difficult to disturb the thermal equilibrium of the ions. Taking ruby as an example: light ruby containing of the order of 0.05% Cr^{3+} can be effectively pumped with a few hundred microwatts of pump power per unit volume of crystal; dark ruby containing 1% Cr^{3+} or more is unaffected at this power level. A similar situation exists in other paramagnetic crystals. We see that pumped paramagnetic crystal comes in two grades: active crystal (low density) and passive crystal (high density). This fact is exploited in the travelling-wave maser.

4. Classical Theory of Maser Action

The quantum-theory of maser action has been outlined in Section 3. To complete the analysis we have to calculate the transition

rate induced by the signal field. While it is not difficult to do this quantum-mechanically, [5,8] it is more illuminating to carry out a classical calculation of the response of the ions to the signal field. The classical approach has one defect: we cannot easily take into account the internal magnetic field which is, in fact, essential for maser action. The difficulty can be overcome by a trick which gives precisely the right answer in some instances and gives the right physical picture in the general case. The author feels that the classical theory is more interesting than the standard quantum-mechanical calculation; it gives one an intuitive grasp of the non-reciprocal behaviour of the paramagnetic ions which is exploited in the travelling-wave maser.

First of all we give a classical description of the behaviour of an ion with spin vector $\underline{S}$ when it is subjected to a d.c. magnetic field $\underline{H}_o$. This topic has already been covered from a quantum-mechanical point of view, but a classical treatment is necessary before we attempt to handle the response to the signal field by classical methods. The angular momentum of the ion in c.g.s. units is $\hbar\underline{S}$, and the magnetic moment is $\underline{\mu}_i = -2\beta\underline{S}$. The d.c. magnetic field exerts a couple on the magnetic moment which is at right angles to both $\underline{H}_o$ and $\underline{\mu}_i$ and has magnitude $H_o|\underline{\mu}_i|\sin\theta$ where θ is the angle between $\underline{H}_o$ and $\underline{\mu}_i$. In vector notation, the couple is $\underline{\mu}_i \times \underline{H}_o$, the vector product of $\underline{\mu}_i$ and $\underline{H}_o$. Newton's law for rotational motion states that the rate of change of the angular momentum is equal to the couple. Hence

$$\hbar \frac{d}{dt}\underline{S} = \underline{\mu}_i \times \underline{H}_o \qquad (7)$$

i.e., since $\underline{S} = -\underline{\mu}_i/2\beta$ and $\underline{\mu}_i \times \underline{H}_o = -\underline{H}_o \times \underline{\mu}_i$,

$$\frac{d}{dt}\underline{\mu}_i = \gamma \underline{H}_o \times \underline{\mu}_i \qquad (8)$$

where $\gamma = 2\beta/\hbar = 2\pi \times 2.8$ megaradians/gauss is the 'gyro-magnetic ratio' of the ion.

The solution of equation (8) is well known; it is illustrated in Fig.4. The magnetic moment precesses about the d.c. magnetic field with an angular velocity $\omega_o = \gamma H_o$. The direction of rotation is related to the direction of $\underline{H}_o$ in the same way as the direction of rotation and the direction of motion when driving a right-handed screw. For brevity, we call this a positive rotation about $\underline{H}_o$, a rotation in the opposite sense is called a negative rotation. The angle θ between $\underline{\mu}_i$ and $\underline{H}_o$ can have any value. The validity of this

solution can be verified by substitution in equation (8).

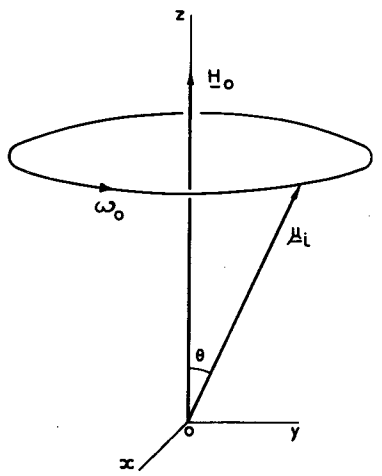


Figure 4 The precessing magnetic moment

We have just seen that the angle between $\underline{\mu}_i$ and $\underline{H}_o$ is arbitrary. This conclusion appears to contradict completely the quantisation rule for the orientation of the magnetic moment. Nevertheless, the classical approach is rigorously correct in this respect! We saw in the quantum-mechanical treatment that the ions are not all in the same state. Some have one allowed orientation, some have another. The uncertainty is removed when we consider (as we must do) a large number of ions. Their mean magnetic moment has a well-defined direction; but, since it is a mean, it need not have any of the allowed orientations, it can make any angle with $\underline{H}_o$. The mean magnetic moment is the quantum-mechanical analogue of the classical magnetic moment which appears in equation (8).

Equation (8) is not the whole story, obviously there is something missing because it predicts that the magnetic moment will go on precessing about $\underline{H}_o$ for ever, whereas we know that, in fact, it will lower its energy as far as possible by coming into line with $\underline{H}_o$. The coupling of the ion to the other ions and to the surrounding crystal causes this 'relaxation' of the precessing moment towards an equilibrium value. A description of the relaxation process which is adequate for our purpose can be obtained by adding on to the right-hand side of equation (8) the extra term $-(\underline{\mu}_i - \underline{\mu}_o)/T_2$ where $\underline{\mu}_o$ is the equilibrium moment (parallel to $\underline{H}_o$) and T_2 is the relaxation time (strictly: the spin-spin relaxation time[9]). The equation then becomes

$$\frac{d}{dt}\underline{\mu}_i = \gamma \underline{H}_o \times \underline{\mu}_i - (\underline{\mu}_i - \underline{\mu}_o)/T_2 \qquad (9)$$

Since $\underline{\mu}_o$ is parallel to $\underline{H}_o$, equation (9) obviously has the steady-state solution $\underline{\mu}_i = \underline{\mu}_o$. If the magnetic moment is moved away from the equilibrium position it spirals back towards it, rotating at the precession frequency ω_o and decreasing the angle of precession exponentially with a time constant equal to T_2. To prove that this

is so, we have only to verify by substitution in equation (9) that it has the general solution

$$\underline{\mu}_i = \underline{\mu}_o + \exp(-t/T_2)\,\underline{\mu}'$$

where $\underline{\mu}'$ is a solution of equation (8).

We are now ready to calculate the response of the ion to the magnetic field of the microwave signal

$$\underline{H}_1 = \mathrm{Re}\left[\underline{H}\exp(j\omega t)\right] \qquad (10)$$

where $\underline{H} = (H_x, H_y, H_z)$ is the complex vector amplitude of the field and $j = \sqrt{-1}$. The equation of motion of $\underline{\mu}_i$ in the presence of $\underline{H}_1$ is obtained from Equation (9) by substituting the total field $\underline{H}_o + \underline{H}_1$ for $\underline{H}_o$ alone. When $\underline{H}_1$ is small (as we shall assume) $\underline{\mu}_i$ is perturbed only slightly from $\underline{\mu}_o$ and becomes $\underline{\mu}_o + \underline{\mu}_1$ say. By multiplying out the vector product and neglecting the second order term $\underline{H}_1 \times \underline{\mu}_1$ we obtain the equation of motion of the perturbation $\underline{\mu}_1$:

$$\begin{aligned} d\underline{\mu}_1/dt &= \gamma(\underline{H}_o \times \underline{\mu}_1 + \underline{H}_1 \times \underline{\mu}_o) - \underline{\mu}_1/T_2 \\ &= \gamma\underline{H}_o \times (\underline{\mu}_1 - K\underline{H}_1) - \underline{\mu}_1/T_2 \end{aligned} \qquad (11)$$

since $\underline{\mu}_o = K\underline{H}_o$, where K is some proportionality constant, and $\underline{H}_1 \times K\underline{H}_o = -\underline{H}_o \times K\underline{H}_1$. The steady-state solution of equation (11) may be written in the form

$$\underline{\mu}_1 = \mathrm{Re}\left[\underline{\mu}\exp(j\omega t)\right] \qquad (12)$$

where $\underline{\mu} = (\mu_x, \mu_y, \mu_z)$ is the complex vector amplitude of $\underline{\mu}_1$. By taking the z-axis along $\underline{H}_o$ (see Figure 4) and writing out the Cartesian components* of equation (11) with $\underline{\mu}_1$ expressed as in equation (12) we obtain

$$\begin{aligned} (j\omega + 1/T_2)\,\mu_x &= -\omega_o(\mu_y - KH_y) \\ (j\omega + 1/T_2)\,\mu_y &= +\omega_o(\mu_x - KH_x) \\ (j\omega + 1/T_2)\,\mu_z &= 0 \end{aligned} \qquad (13)$$

---

* The Cartesian components of the vector product $\underline{A} \times \underline{B}$ are

$$\begin{aligned} &A_yB_z - A_zB_y \\ &A_zB_x - A_xB_z \\ &A_xB_y - A_yB_x \end{aligned}$$

Hence $\mu_z = o$ while μ_x and μ_y are determined by the first pair of equations in the set (13). Instead of solving them directly, it is more instructive to work with the quantities

$$\begin{aligned} \mu^+ &= \mu_x + j\mu_y \\ \mu^- &= \mu_x - j\mu_y \\ H^+ &= H_x + jH_y \\ H^- &= H_x - jH_y \end{aligned} \tag{14}$$

These quantities have a simple interpretation to be explained in a moment. By multiplying the second equation in the set (13) by j and then adding and subtracting it from the first equation we obtain immediately explicit formulae for μ^+ and μ^-:

$$\mu^+ = \frac{-jK\omega_o H^+}{j(\omega - \omega_o) + 1/T_2} \tag{15}$$

$$\mu^- = \frac{+jK\omega_o H^-}{j(\omega + \omega_o) + 1/T_2} \tag{16}$$

Before going on we must see what the quantities introduced by equation (14) actually represent. The first thing to notice is that μ_z and H_z do not appear in equations (15) and (16); μ_z is zero and H_z does not couple to the magnetic moment. In what follows we shall ignore H_z so that $\underline{\mu}_1$ and $\underline{H}_1$ both lie in the xy-plane. It is sufficient to consider $\underline{H}_1$ alone. By using the definitions (10) and (14) and ignoring H_z we have

$$\begin{aligned} \underline{H}_1 &= (H_{1x}, H_{1y}) \\ &= \mathrm{Re}\left[(H_x, H_y) \exp(j\omega t) \right] \\ &= \mathrm{Re}\left[\left((H^+ + H^-)/2, -j(H^+ - H^-)/2 \right) \exp(j\omega t) \right] \\ &= \mathrm{Re}\left[\tfrac{1}{2}H^+(1, -j) \exp(j\omega t) \right] + \mathrm{Re}\left[\tfrac{1}{2}H^-(1, j) \exp(j\omega t) \right] \end{aligned} \tag{17}$$

In the last line $\underline{H}_1$ has been broken up into the superposition of two fields. The first field is

$$\begin{aligned} \underline{H}_1^+ &= \mathrm{Re}\left[\tfrac{1}{2}H^+(1, -j) \exp(j\omega t) \right] \\ &= \tfrac{1}{2}|H^+| (\cos\phi, \sin\phi) \end{aligned} \tag{18}$$

where $\phi = \omega t +$ the phase angle of H^+. The phase angle ϕ increases as time progresses. When $\phi = 0$, $\underline{H}_1^+$ is directed along the x-axis, a quarter-cycle later, when $\phi = \frac{1}{2}\pi$, $\underline{H}_1^+$ is directed along the y-axis (see Figure 5). Moreover, the field $\underline{H}_1^+$ obviously has the constant magnitude $|\underline{H}_1^+| = \frac{1}{2}|H^+|$. We can conclude that $\underline{H}_1^+$ is a circularly polarized field rotating in the positive sense about Oz ($\underline{H}_o$). The second part of $\underline{H}_1$:

$$\underline{H}_1^- = \mathrm{Re}\left[\tfrac{1}{2}H^-(1,j)\ \exp(j\omega t)\right] \qquad (19)$$

is also circularly polarized, but rotates in the negative sense because of the different sign in front of j (see Figure 5).

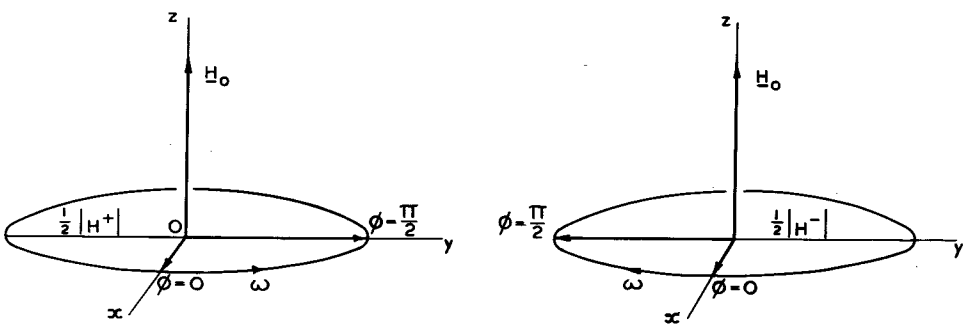


Figure 5 The circularly polarized components of the microwave magnetic field

We see that H^+ and H^- specify the amplitude and phase of the two contra-rotating circularly polarized fields into which the microwave magnetic field can be analysed. Similarly, μ^+ and μ^- specify the amplitude and phase of the two contra-rotating circularly polarized components ($\underline{\mu}_1^+$ and $\underline{\mu}_1^-$) which together make up the perturbation of the magnetic moment. The resolution of $\underline{H}_1$ and $\underline{\mu}_1$ into circularly polarized components enters naturally into the solution of the equation of motion because the natural motion of the magnetic moment is (apart from relaxation) circularly polarized. Moreover, we know that the natural motion is a rotation in the positive sense at the precession frequency ω_o. Consequently we expect to find that $\underline{\mu}_1^+$ is strongly excited by $\underline{H}_1^+$ when $\omega \approx \omega_o$ while $\underline{\mu}_1^-$ is only weakly excited by $\underline{H}_1^-$ at all frequencies. These expectations are realized in equations (15) and (16): μ^+ is excited by H^+ and exhibits a resonance when $\omega = \omega_o$ while μ^- is excited by H^- and does not exhibit a resonance.

To summarize the results of the above discussion we can say that, when $\omega \simeq \omega_o$, the ions only couple to the circularly polarized component of the magnetic field which rotates in sympathy with the natural precessional motion and the coupling is a maximum when the frequency of the field is equal to the frequency of precession. It follows from this fact that a paramagnetic crystal is a non-reciprocal transmission medium for circularly polarized electromagnetic waves. Consider what happens when a wave rotating in the positive sense about the direction of propagation passes through the crystal, first in the direction of $\underline{H}_o$ and then in the opposite direction. In the first case the microwave magnetic field rotates in sympathy with the precessional motion, couples strongly to the ions and exchanges energy with them. In the second case the microwave magnetic field rotates in the opposite sense to the precessional motion, does not couple to the ions and does not exchange energy with them. That is what we mean by a non-reciprocal transmission medium. Contrast the behaviour of a paramagnetic crystal with that of a lossy dielectric, which is a typical reciprocal medium. The attenuation of the wave by the lossy dielectric is the same for both directions of propagation because it depends only on the magnitude of the electric field and is independent of the direction of rotation. The non-reciprocal behaviour of the paramagnetic ions is utilised in the travelling-wave maser to obtain gain in one direction and attenuation in the opposite direction.

We are principally interested in the power absorbed from the microwave field by the magnetic moment. The absorbed power is given by

$$P_m = \mathrm{Av}(\underline{H}_1 \cdot d\underline{\mu}_1/dt) \qquad (20)$$

where 'Av' means 'time-average' and the dot means scalar product, i.e. the product of the vectors' magnitudes and the cosine of the angle between them. (Equation (20) is the magnetic analogue of the mean potential difference times current formula for electrical circuits.) To evaluate p_m we write $\underline{H}_1 = \underline{H}_1^+ + \underline{H}_1^-$ as in equation (17) and use the corresponding resolution of $\underline{\mu}_1$: $\underline{\mu}_1^+ + \underline{\mu}_1^-$. Thus

$$p_m = \mathrm{Av}\left[(\underline{H}_1^+ + \underline{H}_1^-) \cdot d(\underline{\mu}_1^+ + \underline{\mu}_1^-)/dt\right]$$

We can neglect both $\underline{\mu}_1^-$ and $\underline{H}_1^-$ in this expression because $\underline{\mu}_1^-$ is negligible compared to $\underline{\mu}_1^+$ near resonance and $\underline{H}_1^-$ rotates in the opposite sense to $d\underline{\mu}_1^+/dt$ so that their scalar product averages to zero. Hence

$$p_m = \mathrm{Av}(\underline{H}_1^+ \cdot d\underline{\mu}_1^+/dt) \qquad (21)$$

Now $\underline{H}_1^+$ and $d\underline{\mu}_1^+/dt$ both rotate in the same direction with the same angular velocity, keeping the same lengths and the same angle between them. Consequently $\underline{H}_1^+ \cdot d\underline{\mu}_1^+/dt$ is in fact constant. The lengths of the two vectors are $\frac{1}{2}|H^+|$ and $\frac{1}{2}\omega|\mu^+|$ and the angle between them is $\psi = \frac{1}{2}\pi +$ phase angle of μ^+ - phase angle of H^+; the ω and $\frac{1}{2}\pi$ appear because of the time derivitive. Hence, by using the rules for complex multiplication and substituting for μ^+ from equation (15), we obtain

$$p_m = \frac{\omega}{4}|H^+||\mu^+|\cos\psi$$

$$= \mathrm{Re}\,(j\omega\overline{H^+}\,\mu^+/4)$$

$$= \frac{K\omega\omega_o}{4}|H^+|^2 \frac{1/T_2}{(1/T_2)^2 + (\omega-\omega_o)^2} \qquad (22)$$

In the second line $\overline{H^+}$ denotes the complex conjugate of H^+. Equation (22) is essentially the desired formula but it is convenient to make a few changes at this point so as to put the formula into a more useful form. First, we multiply by N, the number of ions per unit volume, so as to obtain the power absorbed per unit volume. Then we multiply and divide by $2T_2^2$ so as to put the last factor into its standard form. Finally, we write $\omega_o = \gamma H_o = 2\beta H_o/h$ and $KH_o = \mu_o$, the z-component of the equilibrium moment. The final result for the power absorbed by the paramagnetic ions per unit volume of the crystal is

$$P_m = Np_m = \hbar\omega\,(N\mu_o\beta/4\hbar^2)|H^+|^2 \frac{2T_2}{1 + T_2^2(\omega-\omega_o)^2} \qquad (23)$$

The last factor in the expression for P_m specifies the frequency-dependence of the absorption, i.e. the line-shape. The width of the line between the half-power points is $2/T_2$ radians/sec $= 1/(\pi T_2)$ cycles/sec., which is about 50 Mc/sec in active maser crystals, while $\omega/2\pi \sim \omega_o/2\pi$ is a microwave frequency. It is therefore legitimate to neglect the frequency dependence of the ω in the first factor of P_m when considering the line-shape near resonance.

At the beginning of this section it was remarked that it would be necessary to use a trick to get the right answer because we are neglecting the internal magnetic field. There has been no trickery so far; it becomes necessary when we evaluate $N\mu_o$, the z-component of the equilibrium magnetic moment per unit volume - which is the

only unknown quantity in the formula for P_m. The difficulty arises because the energy levels are equally spaced in the absence of the internal magnetic field. Consequently the signal field stimulates transitions between every pair of adjacent levels, whereas in real maser crystals the levels are unequally spaced and only one transition is excited. This unrealistic feature of our model can be obviated by ensuring that only the transition we are interested in absorbs power, in spite of the fact that the levels are equally spaced.

To see how this can be achieved we consider the case of the transition between the $-\frac{1}{2}$ and $+\frac{1}{2}$ levels of Cr^{3+} ($S = 3/2$, see Figure 6). The quantum-mechanical theory of Section 3 showed

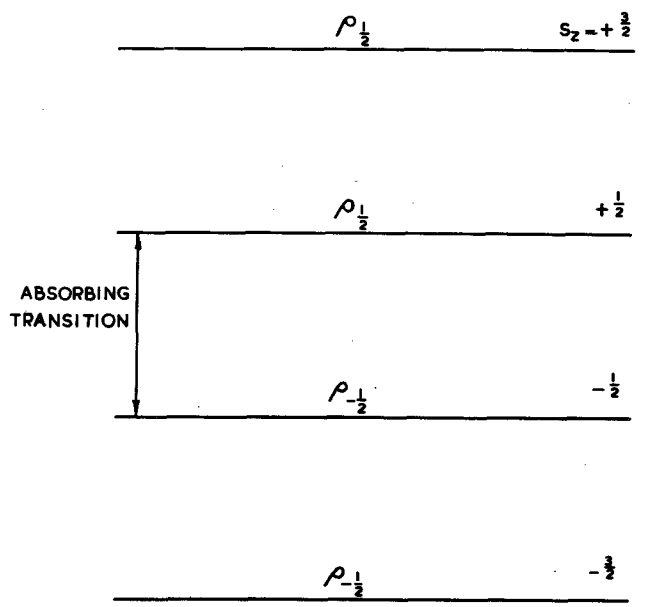


Figure 6 Artificial populations which supress the -3/2 to -1/2 and +1/2 to +3/2 transitions of Cr^{3+}

that the net power absorbed by a transition is proportional to the population difference across it (see equation (5)). The absorption is positive when there are more ions in the lower level than there are in the upper level, it is negative when the reverse is true (net emission) and it vanishes when the two populations are equal. With this fact in mind, we consider an artificial 'equilibrium' situation in which the fractional populations of the various levels

have the values shown in Figure 6. The fractional populations $\rho_{-1/2}$ and $\rho_{1/2}$ are determined by requiring that $\rho_{-1/2} - \rho_{1/2}$ have the value actually established by the pump and that $2(\rho_{-1/2} + \rho_{1/2}) = 1$, i.e. the sum of the fractional populations is unity, as it should be. In this situation we have the true population difference across the -1/2 to +1/2 transition and zero population difference across the other two transitions stimulated by the signal field. Consequently, all the power absorbed by the ions goes into the -1/2 to +1/2 transition, and the power absorbed by this transition alone is obtained by substituting in equation (23) the value of $N\mu_o$ appropriate to the artificial equilibrium situation

Referring to Figure 6 we see that, of the N ions per unit volume, a fraction $\rho_{1/2}$ have $S_z = 3/2$, a fraction $\rho_{1/2}$ have $S_z = 1/2$, a fraction $\rho_{-1/2}$ have $S_z = -1/2$ and a fraction $\rho_{-1/2}$ have $S_z = -3/2$. Hence, using equation (2), the z-component of the 'equilibrium' moment per unit volume is

$$N\mu_o = -2\beta N \left(\frac{3}{2}\rho_{1/2} + \frac{1}{2}\rho_{1/2} - \frac{1}{2}\rho_{-1/2} - \frac{3}{2}\rho_{-1/2}\right)$$
$$= 4\beta N (\rho_{-1/2} - \rho_{1/2}) \qquad (24)$$

By substituting this value of $N\mu_o$ in equation (23) we obtain the final result for the power absorbed by the -1/2 to +1/2 transition

$$P_m = \hbar\omega \cdot N(\rho_{-1/2} - \rho_{1/2}) \cdot \left[(\beta/\hbar)^2 |H^+|^2 \frac{2T_2}{1+T_2^2(\omega-\omega_o)^2} \right] \qquad (25)$$

The expression has been broken up into the product of the three factors which occur in equation (5) of Section3: the microwave quantum, the population difference per unit volume and the transition rate per ion. A quantum-mechanical calculation yields precisely the same expression provided it is correct to label the energy levels with values of S_z as we have done. Actually this is only true in special cases (e.g. $\underline{H}_o$ along the c-axis in ruby) because the internal magnetic field usually mixes up with the S_z values, but that is a technical complication of secondary importance from our point of view.

The same technique can be applied in the general case to calculate the power absorbed by the transition from $S_z = M$ to $S_z = M + 1$ when the ions have spin S. The result is

$$P_m = \hbar\omega \cdot N(\rho_M - \rho_{M+1}) \cdot \left[(\beta/2\hbar)^2 (S-M)(S+M+1) |H^+|^2 \frac{2T_2}{1+T_2^2(\omega-\omega_o)^2} \right] \qquad (26)$$

which reduces to equation (25) when $M = -1/2$ and $S = 3/2$. The author hopes that the experts will be as delighted as he was to find the matrix element of the magnetic moment appearing correctly in this expression.

5. The Comb Structure

We have seen how to obtain active crystal. To make an amplifier out of it, we have simply to put the crystal in some sort of structure to carry the signal power (and the pump power) to the active crystal and take it away again after it has been amplified. The obvious structure to choose is a resonant cavity with waveguide feeds (see Fig. 7). It has the advantage

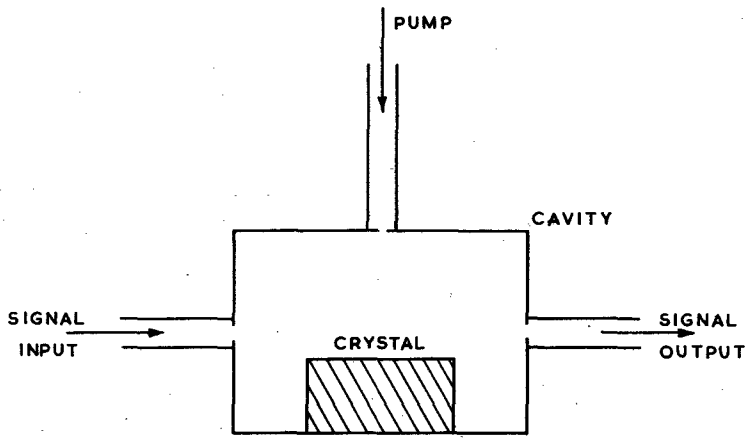


Figure 7 The transmission cavity maser

of producing large fields in the cavity for small input powers. But it has a number of disadvantages:

(i) Small gains are easily obtained, but to achieve the 20-30 db gain required of a practical device, the coupling to the waveguide feeds must be reduced so that less power escapes from the cavity and the field strength in it builds up.[10] If the coupling is reduced too far, the maser breaks into oscillation because the emission from the active crystal stimulated by noise in the cavity is more than enough to supply the ohmic losses and the

waveguide feeds. High gain is obtained by operating the maser on the verge of oscillation, which is obviously a most sensitive and unstable condition.

(ii) As the gain increases the amplification bandwidth narrows, falling off as the inverse square root of the gain.[10] At 20 db gain the maximum possible bandwidth of a cavity maser (reflection type) is 20% of the linewidth, i.e. typically 10 Mc/sec.

(iii) It is a simple matter to tune the transition frequencies of the paramagnetic ions by altering the magnitude and direction of the d.c. magnetic field. But the cavity must be tuned at the same time which is not so simple.

(iv) The amplification is reciprocal, i.e. power sent through the cavity in either direction is amplified by the same factor.[10] Consequently small reflections in the input and output waveguides of a high-gain maser can easily send it into oscillation. The reciprocity of the gain may be surprising in view of the fact that the paramagnetic crystal is non-reciprocal. It comes about because the magnetic field in a cavity is linearly polarised, i.e. it consists of two contra-rotating circularly polarised components of equal magnitude. Whatever direction the power is coming from, one of these circularly polarised components couples strongly to the paramagnetic ions. Non-reciprocal gain can only be achieved in cavity masers by adding non-reciprocal ferrite elements outside the cavity. These have negligible loss in one direction and high loss in the opposite direction.

All the above disadvantages of putting the active crystal in a cavity are alleviated to a considerable extent when the active crystal is distributed along a waveguide instead (see Fig. 8).

(i) High gain can be achieved by increasing the length of waveguide and by decreasing the velocity at which energy propagates through it, so as to keep the signal in the active crystal for a long time. Neither procedure leads to excessive gain sensitivity or instability.

(ii) High gain results from an exponential growth along the waveguide instead of from positive feedback. Consequently, as we shall see later, the bandwidth falls off as the inverse square root of the gain in dbs instead of falling off as the inverse square root of the actual gain.

(iii) Tuning the paramagnetic ions alone will tune the entire

device over the pass-band of the waveguide.

(iv) The amplification can be made non-reciprocal by placing the active crystal in regions of circularly polarised magnetic field. In fact, as we shall see later, by putting active crystal (low density) and passive crystal (high density) in distinct regions of contra-rotating circular polarisation we can obtain gain in one direction and an equal or larger loss in the opposite direction. The device cannot oscillate under these conditions, however large the reflections in the input and output waveguides may be.

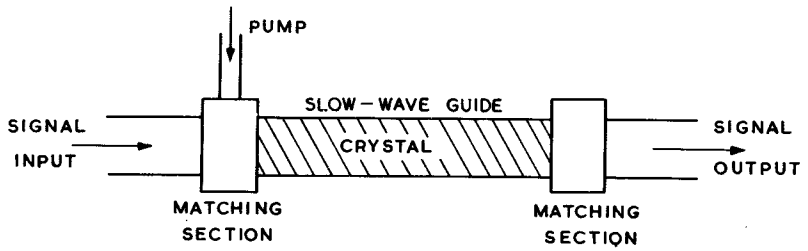


Figure 8 The travelling-wave maser

The essential characteristics desired of the waveguide structure for a travelling-wave maser are: low velocity of energy propagation (group velocity) and distinct regions of contra-rotating circular polarisation. Both these characteristics are possessed by structures in which the wave is guided along an array of parallel wires set transverse to the direction of propagation. Structures of this type have already found many applications to electronic travelling-wave tubes in which the active element is an electron beam. Examples are: helices, ladder lines, meander lines, interdigital lines and combs.[11] We shall consider the comb structure because it has been used successfully in a travelling-wave maser at Bell Telephone Laboratories.[3]

The 'pure' comb consists of an array of identical, equally spaced, parallel wire fingers, short-circuited at one end and open-circuited at the other (see Fig. 9). However, the pure comb will not transmit electromagnetic energy along its length. If we excite a particular finger, the wave bounces backwards and forwards between the short-circuit at one end and the open-circuit at the other; it is not passed on to adjacent fingers. In other words,

the group velocity is zero for a pure comb. For our purpose this is an advantage, because, by introducing a small perturbation, such as the capacitive loading plate indicated in Fig. 9, we can provide a connecting link between the fingers and propagate a signal along the comb (in the y-direction) with as small a group velocity as we desire. We shall see that group velocities of about 1% of the velocity of light are required. This represents a very high degree of wave-slowing which the comb structure is ideally suited to perform.

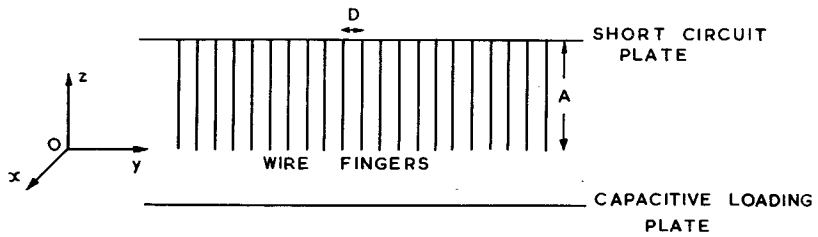


Figure 9 The comb structure

The finger-length A appropriate to the signal frequency ω is determined by the following considerations. When the pure comb is excited at the signal frequency a standing-wave is set up on each finger consisting of two waves travelling along it in opposite directions with the velocity of light. Since there is an open-circuit at the finger tip, the current along it varies like $\sin(2\pi z/\lambda)$ where z is measured from the finger tip and λ is the free-space wavelength at the signal frequency. At the base of the finger there is a short circuit, i.e. a current maximum. Hence $2\pi A/\lambda = \pi/2$ and $A = \lambda/4$. The fingers must be a quarter-wavelength long if the pure comb is to be capable of supporting a field at the signal frequency.

We see that the pure comb has an infinitesimal pass-band, limited to the frequency for which the fingers are precisely a quarter-wavelength long. The capacitive loading plate produces a finite pass-band. We may estimate its width very roughly by remembering that the group velocity is defined by the equation[12]

$$v_g = d\omega/d\beta_g \qquad (27)$$

where β_g is the phase-change coefficient in the direction of propagation. Now the comb is a periodic structure; the pass-band

limits are therefore $\beta_g = 0$ (zero-mode) and $\beta_g = \pi/D$ (π-mode) where D is the period of the comb. Very roughly, then, the width of the pass-band is

$$\Delta\omega/2\pi = (d\omega/d\beta_g)\Delta\beta_g/2\pi \qquad (28)$$

$$= (v_g\pi/D)/2\pi$$

which has the value 750 Mc/sec. when $v_g = c/100$ and D = 2 mm.. The pass-band is very much wider than the line-width of the paramagnetic crystal. It provides a large tuning band for the travelling-wave maser which can be increased by reducing the period of the comb.

The magnetic field distribution is hardly disturbed by the capacitive loading plate. Even so, the calculation of the details of the field distribution is obviously a formidable task.[11] Before deriving an approximate solution of the problem, it is useful to observe one or two simple properties of the field. The magnetic field encircles the fingers and the electric field points radially from them. Hence both fields are in the xy-plane (see Figure 9). Moreover, the x-component of the magnetic field is symmetrical about the plane of the comb while the y-component is anti-symmetrical. The last two statements are obviously true close to the fingers, but they apply quite generally because, apart from a small end-effect due to the capacitive loading plate, there is no geometrical feature to render them invalid as we move away from the comb. The symmetry properties allow us to confine our attention to the field below the comb ($x < 0$).

The essential properties of the magnetic field distribution below the comb can be obtained by recognising that the primary function of the fingers is to prevent conduction in the y-direction while still providing perfect conduction in the z-direction. Consider a wave whose wavelength λ_g in the y-direction is several times the period of the comb. The structure of the comb is averaged out over the wavelength and it behaves like a smooth sheet which conducts perfectly in the z-direction and does not conduct at all in the y-direction. To this degree of approximation the comb is uniform in the y-direction and the field below it depends on y through a factor $\exp(-j2\pi y/\lambda_g)$. Moreover, the magnetic field depends on z through the same factor as the current along the fingers: $\sin(2\pi z/\lambda)$. Finally, since the field must fall to zero away from the comb, it must depend on x through a factor $\exp(+\Gamma x)$ where Γ is the attenuation coefficient in the negative x-direction. Hence, apart from an amplitude factor, we have

$$H_x = \sin(2\pi z/\lambda)\exp\left[\Gamma x - j2\pi y/\lambda_g\right]$$

$$H_y = B\sin(2\pi z/\lambda)\exp\left[\Gamma x - j2\pi y/\lambda_g\right] \qquad (29)$$

where B is a constant determined by the ratio H_y/H_x.

The attenuation coefficient Γ is determined by the fact that the field components satisfy the wave equation

$$\frac{d^2U}{dx^2} + \frac{d^2U}{dy^2} + \frac{d^2U}{dz^2} + \left(\frac{2\pi}{\lambda}\right)^2 U = 0 \qquad (30)$$

By substituting H_x or H_y for U in equation (30) we find that

$$\Gamma = 2\pi/\lambda_g \qquad (31)$$

i.e. the attenuation coefficient away from the guiding structure is equal to the phase-change coefficient in the direction of propagation. This is a general property of slow-waves.

To evaluate the constant B we invoke Ampères Law: the line-integral of the magnetic field round any closed path is equal to 4π times the current threaded through it. The rectangular path in the xy-plane shown in Figure 10 (which is closed at infinity) has no current threaded through it; no conduction current because it is in free space and no displacement current because $E_z = 0$.

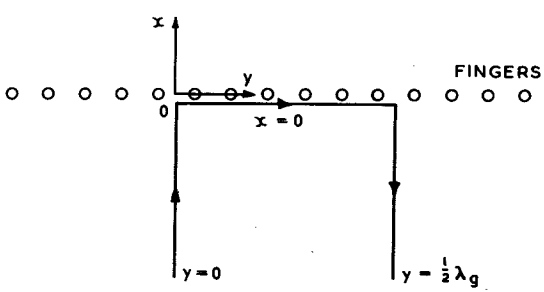


Figure 10 Path used in applying Ampères Law

Hence

$$\int_{-\infty}^{0} dx\, H_x(x,0,z) + \int_{0}^{\frac{1}{2}\lambda_g} dy\, H_y(0,y,z) + \int_{0}^{-\infty} dx\, H_x(x,\tfrac{1}{2}\lambda_g,z) = 0$$

i.e., from equations (29) and (31),

$$\sin(2\pi z/\lambda).\left[\lambda_g/2\pi + B\,\lambda_g/j\pi + \lambda_g/2\pi\right] = 0$$

whence

$$B = -j \qquad (32)$$

To summarise: the magnetic field distribution below the comb is given by

$$H_x = \sin(2\pi z/\lambda)\,\exp\left[2\pi(x-jy)/\lambda_g\right]$$

$$H_y = -jH_x \qquad (33)$$

In our discussion of the interaction of the paramagnetic ions and the signal field we saw that it is convenient to analyse the magnetic field into two contra-rotating circularly polarised components. The component which rotates in the positive sense about Oz has magnitude

$$\tfrac{1}{2}|H^+| = \tfrac{1}{2}|H_x + jH_y| = \tfrac{1}{2}|H_x + H_x| = |H_x| \qquad (34)$$

while the component which rotates in the negative sense about Oz has magnitude

$$\tfrac{1}{2}|H^-| = \tfrac{1}{2}|H_x - jH_y| = \tfrac{1}{2}|H_x - H_x| = 0 \qquad (35)$$

Thus the magnetic field below the comb has pure circular polarisation with a positive sense of rotation about Oz. On the other hand, the field above the comb ($x > 0$) has pure circular polarisation with a negative sense of rotation about Oz because, when we pass through the comb, H_y changes sign while H_x remains unaltered.

It was tacitly assumed above that λ_g is positive. By substituting equation (33) into equation (10) we see that the wave is therefore travelling in the positive y-direction. We call it the forward wave. Consider now a wave travelling in the negative y-direction (the reverse wave). To do this we have merely to change the sign of the coefficient of y in equation (29). Following through the above calculation of Γ and B we find that H_x and H_y are both replaced by their complex conjugates. Thus the magnetic field of the reverse wave is the complex conjugate of the magnetic field of the forward

wave. This is a general result. For the forward wave we found that $H_y = -jH_x$ below the comb and $H_y = jH_x$ above the comb. Taking the complex conjugates of these equations we find, for the reverse wave, $H_y = jH_x$ below the comb and $H_y = -jH_x$ above the comb. Thus the direction of rotation is reversed on both sides of the comb when the direction of propagation is reversed. This is the ideal situation for obtaining non-reciprocal gain.

Suppose we fill the space below the comb with active crystal and subject it to a d.c. magnetic field in the positive z-direction. Then, within the active crystal, the magnetic field of the forward wave is rotating in sympathy with the precessional motion, it couples tightly to the paramagnetic ions and the wave is amplified by their stimulated emission. On the other hand, within the active crystal, the magnetic field of the reverse wave is rotating in the opposite sense to the precessional motion, it does not couple to the paramagnetic ions and is not amplified.

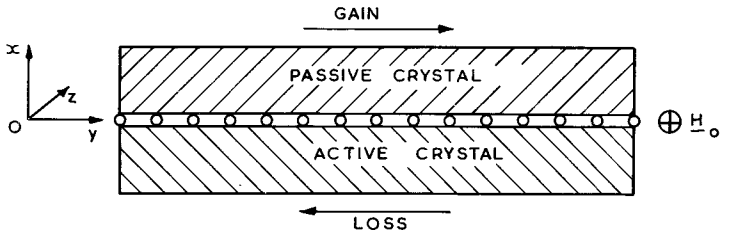


Figure 11 The crystal-loaded comb structure

We can do even better than this by filling the space above the comb with passive crystal (see Figure 11). The forward wave couples only to the active crystal and is amplified. The reverse wave couples only to the passive crystal and is attenuated. If the comb is symmetrically loaded with active and passive crystal having the same response to a microwave magnetic field rotating in sympathy with the precessional motion, then the forward gain is equal to the reverse loss. Reflections in the input and output waveguides can never lead to oscillation.

The reader may be in some doubt as to which side of the comb is which. We defined the regions below ($x < 0$) and above ($x > 0$) the comb with reference to the coordinate system shown in Figure 9. By inspecting the figure, and from the above discussion, we arrive at the following rule. Hold the comb horizontally and look along the fingers in either direction. Then waves travelling to the

right rotate in the positive sense (about the direction of viewing) below the comb and rotate in the negative sense above the comb. The directions of rotation are reversed for waves travelling to the left.

6. Engineering Characteristics

The physical principles of the travelling-wave maser have been explained in some detail. In this section we derive explicit expressions for the important engineering characteristics: forward gain, reverse loss, gain sensitivity, bandwidth and noise figure. In so doing we can safely ignore the ohmic loss of the comb structure because the device will be operating at liquid helium temperatures where the conductivity of normal metals becomes very high.

First of all we evaluate the forward gain. A forward wave is amplified because it couples only to the active crystal below the comb. The amplification can be allowed for by multiplying the magnetic field (33) by the factor $\exp(\alpha y)$ where α is the amplification coefficient in nepers/cm. The power flow P in the positive y-direction is proportional to the square of the field, i.e. it is proportional to $\exp(2\alpha y)$. Hence, in a small length Δy, the power flow increases by

$$\Delta P = (dP/dy)\,\Delta y = (2\alpha P)\Delta y \qquad (36)$$

The increase of P is due to the power emitted by the active ions in the length Δy. Hence

$$\Delta P = P_e \Delta y = \Delta y \int\int dx dz(-P_m) \qquad (37)$$

where P_e is the power emitted per unit length and P_m is the power absorbed per unit volume. The integration is taken over the cross-section of the active crystal. If the active crystal fills the region below the comb in which the field strength is appreciable then the integration limits are $x = -\infty$ to $x = 0$ and $z = 0$ to $z = A$. Equating the two values of ΔP we have

$$\alpha = P_e/(2P) \qquad (38)$$

Now the energy stored on the comb travels with the group velocity v_g. Hence

$$P = Wv_g \qquad (39)$$

where W is the energy stored per unit length.[12] By substituting equation (39) into equation (38) and using the conventional definition of the Q-factor of the wave, viz

$$1/Q_m = - P_e/(\omega W), \tag{40}$$

we obtain

$$\alpha = - \omega/(2Q_m v_g) \tag{41}$$

This is a standard result in the theory of lossy guides.[12] In our case however the absorbed power P_m is negative, the 'magnetic Q' Q_m is negative and we have amplification instead of attenuation.

Suppose that the active crystal fills a length ℓ of the comb structure. Then the power flow increases by the factor $\exp(2\alpha\ell)$ from one end to the other. Hence the forward gain is

$$G = \exp(2\alpha\ell) \tag{42}$$

and the forward gain in dbs is

$$\begin{aligned} G' &= 10 \log_{10} \exp(2\alpha\ell) \\ &= - 27.3 \frac{\ell}{\lambda} \cdot \frac{c}{v_g} \cdot \frac{1}{Q_m} \end{aligned} \tag{43}$$

In the last line we have substituted for α and written $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the free-space wavelength at the signal frequency.

To complete the calculation of the forward gain we have to evaluate Q_m. The power absorbed per unit volume by the paramagnetic ions is given by equation (26) in Section 4. By integrating that equation over the cross section of the active crystal and assuming that $N(\rho_M - \rho_{M+1})$ and T_2 are constant throughout we obtain

$$P_e = - \hbar\omega \cdot N(\rho_M - \rho_{M+1}) \times$$

$$\left[(\beta/2\hbar)^2 (S - M)(S + M + 1) \cdot \frac{2T_2}{1+T_2^2(\omega-\omega_o)^2} \int\int dx dz |H^+|^2\right] \tag{44}$$

The store energy per unit length is equal to 2 x 2 times the mean magnetic energy stored below the comb. One factor of 2 comes from the symmetry of the magnetic field and the other appears because the mean stored electric energy per unit length is equal to the mean stored magnetic energy per unit length. (This fact can be verified by calculating, from Maxwell's equations, the electric field going with the magnetic field (33)). Hence, by making use of equations (33) and (34) we obtain

$$W = \frac{4}{8\pi} \mathrm{Av}\left[\int\int dx\, dz\, \underline{H}_1^2\right]$$

$$= \frac{4}{16\pi} \int\int dx\, dz(|H_x|^2 + |H_y|^2)$$

$$= \frac{1}{8\pi} \int\int dx\, dz|H^+|^2 \qquad (45)$$

where the limits of integration are the same as before.

When the expressions (44) and (45) are substituted into equation (40) the field integrals cancel out leaving

$$\frac{1}{Q_m} = (2\pi\beta^2/\hbar).N(\rho_M - \rho_{M+1}).(S - M)(S + M + 1). \frac{2T_2}{1+T_2^2(\omega-\omega_o)^2} \qquad (46)$$

Putting in some typical numbers: $S = 3/2$, $M = -1/2$, $N = 2 \times 10^{19}$ ions/cc., $\rho_{-\frac{1}{2}} - \rho_{+\frac{1}{2}} = -2\%$ and $1/\pi T_2 = 50$ Mc./sec. gives $Q_m = -95$ when $\omega = \omega_o$. Referring back to equation (43) we see that 20 to 30 db forward gain can therefore be obtained from a crystal-loaded comb one wavelength long (3 cm. at X-band) provided $c/v_g \sim 100$. We are helped in this respect by the high refractive index of most paramagnetic crystals. Ruby, for example, has a refractive index of 3.3 so that it automatically reduces the velocity of propagation in the ratio 1:3.3. The required extra factor of 1:30 is easily obtained with a slightly perturbed comb structure. (We neglected the effect of the dielectric surroundings when discussing the comb structure in Section 5. The wavelength λ which appears there is in fact the dielectric wavelength and not the free-space wavelength. The only significant result of this correction, apart from its influence on the group velocity, is that the finger-length must be one-quarter of the dielectric wavelength at the centre of the pass-band).

It is hardly necessary to calculate the reverse loss of the travelling-wave maser. A reverse wave is attenuated because it couples only to the passive crystal underneath the comb. The magnetic Q of the reverse wave is given by Equation (46) when N, T_2 and $(\rho_M - \rho_{M+1})$ are given the values appropriate to the passive crystal. The density of ions in the passive crystal is high enough to ensure that their thermal equilibrium is not disturbed by the pump. Now, increasing the density of ions brings them closer together, they couple to one another more strongly through their magnetic fields and the paramagnetic resonance line width increases (which is why the pump is ineffective). Hence,

roughly speaking, T_2 is proportional to $1/N$ and NT_2 is the same in both active and passive crystals. If we also assume that $\rho_M - \rho_{M+1}$ has the same magnitude in both crystals, but opposite sign, then, from equation (46), Q_m has the same magnitude for both forward and reverse waves when $\omega = \omega_o$, but has opposite signs in the two cases. Consequently the reverse loss is equal to the forward gain. There are of course many factors which can alter this conclusion. Equally well, there are many ways in which we can adjust the ratio of reverse loss to forward gain.

We turn our attention now to the question of gain sensitivity. It is clear that the travelling-wave maser is a less sensitive device than the cavity maser. It is possible to make a numerical comparison by calculating the sensitivity of the gain to variations of the magnetic Q. The gain sensitivity is defined as the fractional gain change per unit fractional change of Q_m, i.e.

$$s = \left| \frac{dG/G}{dQ_m/Q_m} \right| = \left| \frac{Q_m}{G} \frac{dG}{dQ_m} \right| \tag{48}$$

Since $G = \exp(-\omega\ell/Q_m v_g)$, we have

$$s = \left| \frac{Q_m}{G} \cdot \frac{\omega\ell}{Q_m^2 v_g} \cdot G \right|$$

$$= \log_e G$$

$$= 0.230\, G' \tag{49}$$

The gain sensitivity of the cavity maser (reflection type) is $\sqrt{G}$. Hence at 20 db gain the sensitivities in the two cases are 4.6 and 10 and at 30 dbs they are 6.9 and 31.6.

The amplification bandwidth is easily calculated. It is limited because $-1/Q_m$ falls off as the frequency departs from the centre of the paramagnetic resonance line. The bandwidth between the 3 db points is obtained by doubling the frequency shift necessary to reduce G' by 3 dbs from its peak value. The magnetic Q can be eliminated from the answer by expressing it in terms of the peak gain. We find

$$B = \frac{1}{\pi T_2} \sqrt{\frac{3}{G'_p - 1}} \tag{50}$$

where G'_p is the peak gain in dbs. The first factor on the right-hand

side of equation (50) is the linewidth of the paramagnetic resonance ($\sim$ 50 Mc/sec). The second factor is equal to 0.42 at 20 db gain. Thus, at 20 db gain, the bandwidth of the travelling-wave maser is 42% of the linewidth ($\sim$ 20 Mc/sec). In the case of a cavity maser having 20 db gain the bandwidth cannot exceed 20% of the linewidth and is usually much less than this limiting value.[10]

Finally, there remains the all important question of noise performance. Very little noise arises from the ohmic losses of the comb structure because it is cooled with liquid helium. This not only gives the ohmic losses a very low temperature, it also gives them a very low resistance. Nyquist's theorem shows that both these factors reduce the noise output. The passive ions contribute very little to the noise because they do not couple to waves travelling towards the output. The bulk of the noise output from the cooled parts of the maser comes from the active ions.

We shall calculate the noise output in a plausible but roundabout way. We neglect noise from the ohmic losses and the passive ions entirely and consider first of all the case in which the 'active' ions are in thermal equilibrium at the standard noise temperature $T_o = 290^oK$. The 'active' ions are then passive, and the forward gain G is less than one. If the impedance of the signal generator is also at temperature T_o, then the entire device is in thermal equilibrium at temperature T_o and the total noise power output per unit bandwidth must be

$$kT_o = GkT_o + (1 - G)kT_o \tag{51}$$

On the right-hand side of this equation we have broken up kT_o into the part GkT_o coming from the source impedance and the part which is left: $(1 - G)kT_o$, which must come from the 'active' ions.

Next we consider the case in which the source impedance is at temperature T_o while the 'active' ions are in thermal equilibrium at some other temperature T. The temperature of the 'active' ions can be expressed in terms of the fractional population difference across the M to M + 1 transition which is excited by the microwave signal. By using equation (4) we obtain

$$\rho^o_M - \rho^o_{M+1} = (E_{M+1} - E_M)\Big/\left[(2S + 1)kT\right]$$

i.e.

$$T = \hbar\omega\Big/\left[(2S + 1)k(\rho^o_M - \rho^o_{M+1})\right] \tag{52}$$

since $E_{M+1} - E_M = \hbar\omega$. Now the noise output from the 'active' ions is proportional to their temperature. Hence the total noise power

output per unit bandwidth in this case is (cf. equation (51))

$$n = GkT_o + (1 - G)kT \tag{53}$$

Finally, we consider the practical case in which the 'active' ions are in fact activated by the pump. The population of levels M and M + 1 change to ρ_M and ρ_{M+1}, the 'effective temperature' of the ions becomes (cf. equation (52))

$$T_m = \hbar\omega \Big/ \left[(2S + 1)k(\rho_M - \rho_{M+1})\right] \tag{54}$$

which is negative because $\rho_{M+1} > \rho_M$, and the forward gain becomes larger than one for the same reason. The total noise power output per unit bandwidth in this, the case of interest, is (cf. equation (53))

$$n = GkT_o + (1 - G)kT_m \tag{55}$$

Notice that the active ions make a positive contribution to n because $G > 1$ and $T_m < 0$.

The noise figure of the travelling-wave maser is equal to the total noise power output per unit bandwidth divided by the fraction thereof coming from the source impedance:

$$\begin{aligned} F &= \left[GkT_o + (1 - G)kT_m\right] \Big/ (GkT_o) \\ &= 1 + |T_m|/T_o \end{aligned} \tag{56}$$

since $G >> 1$ and $T_m < 0$. Putting in some typical numbers: $\omega = 2\pi \times 10^4$ megaradians/sec (X-band), $S = 3/2$ and $\rho_M - \rho_{M+1} = -2.0\%$ gives $T_m = -6^\circ K$ and $F = 1.02$, i.e. 0.089 dbs. The travelling-wave maser proper has an almost ideal noise performance. The overall noise figure of the amplifier is therefore determined largely by the noise emitted from the uncooled losses in the waveguide input which are amplified by the maser.[13,14]

7. Conclusion

In the next few years the cavity maser will be completely eclipsed by the travelling-wave maser, which is a superior device in every respect. To convince the reader that this is not just another rash prediction by a theoretician, it will be as well to close with a brief description of the travelling-wave maser recently constructed at Bell Telephone Laboratories.[3] The comb is located in a rectangular waveguide with the fingers parallel to the broad face (see Figure 12). The active crystal is ruby containing 0.05% Cr^{3+}. The passive crystal is ruby containing 1% Cr^{3+}. The pump power at

19 kMc/sec is transmitted in the usual TE_{01} waveguide mode which is not disturbed by the comb. The signal at 6 kMc/sec is fed onto the

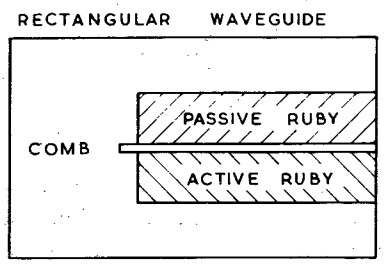


Figure 12 The B.T.L. travelling-wave maser

comb and taken off again by making two fingers extensions of the inner conductors of coaxial line feeds. With 100 mW pump power and 5" of loaded comb the device gives 23 db forward gain 29 db reverse loss, 25 Mc/sec bandwidth and 350 Mc/sec tuning band. The noise figure is 0.051 db excluding the input losses and 0.16 db including the input losses.[15]

Many problems remain to be solved. For the engineer there is the problem of designing suitable waveguide structures. For the physicist there is the problem of understanding the behaviour of paramagnetic crystals so that, ultimately, the crystal can be designed as well as the waveguide. The physical principles are clear and the future is bright.

References

1. J. P. Wittke, 'Molecular amplification and generation of microwaves', Proc. I.R.E. 45, p.291 (March 1957).

2. N. Bloembergen, 'Proposal for a new type solid-state maser', Phys. Rev. 104, p.324 (Oct 15, 1956).

3. R. W. DeGrasse, E. O. Schulz-DuBois and H. E. D. Scovil, 'A unilateral three-level maser employing a ruby-loaded comb structure', Trans. I.R.E (PGCT), to be published (Feb. 1959).

4. F. K. Richtmyer, E. H. Kennard and T. Lauritsen, 'Introduction to Modern Physics', Chap. 7 (McGraw-Hill Book Co. Inc., 1955).

5. D. J. Howarth, 'The physics of the solid state maser', R.R.E. Journal, April 1958, p. 7.

6. B. Bleaney and K. W. H. Stevens, 'Paramagnetic Resonance', Rep. Progr. Phys. 16, p. 108 (1953).

7. C. R. Ditchfield and P. A. Forrester, 'Maser action in the region of 60°K', Phys. Rev. Letters 1, p.448 (Dec. 15th, 1958).

8. P. N. Butcher, 'Theory of three-level paramagnetic masers. Part 1. - Quantum theory', Paper No. 2641, International Convention on Microwave Valves, London, May, 1958 (to be published in Proc. I.E.E. 105, Part B Supplement No. 11).

9. G. E. Pake, 'Nuclear magnetic resonance', Solid State Physics, Vol. 2, p.15 (Academic Press, 1956).

10. P. N. Butcher, 'Theory of three-level paramagnetic masers. Part 2. - Amplification and oscillation', Paper No. 2642, International Convention on Microwave Valves, London, May, 1958 (to be published in Proc. I.E.E., 105, Part B Supplement No. 11).

11. P. N. Butcher, 'The coupling impedance of tape structures', Proc. I.E.E., 104, Part B, p.177 (March, 1957).

12. J.C. Slater, 'Microwave Electronics', Chap. 3, (D. Van Nostrand, Co. Inc., 1950).

13. P. N. Butcher, 'Theory of three-level paramagnetic masers. Part 3 - Output noise power spectrum',,Paper No. 2643, International Convention on Microwave Valves, London, May, 1958 (to be published in Proc. I.E.E., 105, Part B Supplement No. 11).

14. P. N. Butcher, 'Theory of three-level paramagnetic masers. Part 4 - Noise figure', Paper No. 2644, International Convention on Microwave Valves, London, May, 1958 (to be published in Proc. I.E.E., 105, Part B Supplement No. 11).

15. R. W. DeGrasse, private communication.

"ARITHMO had been bred up to Accounts all his Life, and thought himself a compleat Master of NUMBERS. But when he was pushed hard to give the square root of the Number 2, he tried at it, and laboured long in millesimal Fractions, until he confessed there was no End of the Enquiry"

Isaac Watts
1674-1748

NUMERICAL INTEGRATION OF FUNCTIONS GIVEN IN DECIBELS

by R. Hensman and D. P. Jenkins

1. Introduction

Many practical situations require the numerical evaluation of the logarithm of the integral of a function tabulated in logarithmic form. This problem arises, for example, in estimating the power transmitted in part of the beam of an aerial whose polar diagram is given in decibels. There are also cases where the integration of an approximately exponential function may be most easily carried out by taking its logarithm, the more complex integration formula being compensated by the larger integration step which can be used.

A convenient method of performing such calculations has been given by Jacchia (1955) and this note describes his technique in a modified form.

2. Numerical formulae

When the function to be integrated is given in decibels at equal intervals of the independent variable, the usual difference table can be formed:

Independent variable	Function value in db	Differences			
$x_o - 2h$	y_{-2}				
		$\delta_{-1\frac{1}{2}}$			
$x_o - h$	y_{-1}		δ''_{-1}		
		$\delta_{-\frac{1}{2}}$		$\delta'''_{-\frac{1}{2}}$	
x_o	y_o		δ''_o		δ^{iv}_o
		$\delta_{\frac{1}{2}}$		$\delta'''_{\frac{1}{2}}$	
$x_o + h$	y_1		δ''_1		
		$\delta_{1\frac{1}{2}}$			
$x_o + 2h$	y_2				

In terms of these central differences Jacchia's basic result can be written

$$10 \log_{10}\left[\frac{1}{h}\int_{x_o}^{x_o+h} 10^{y/10}dx\right] =$$

$$y_o + \tfrac{1}{2}\delta_{\frac{1}{2}} + \frac{\log_e 10}{240}(\delta_{\frac{1}{2}})^2 - \frac{1}{12}\delta''_o + \frac{1}{24}\delta'''_{\frac{1}{2}} \cdots$$

The quantity $10^{y/10}$ is, of course, the 'natural' value corresponding to y db. This formula can be derived using natural logarithms by expanding $y(x_o + mh)$ in powers of m, δ, δ'', etc. followed by integration with respect to m. A more convenient form analogous to Simpson's rule is

$$10 \log_{10}\left[\int_{x_o-h}^{x_o+h} 10^{y/10}dx\right] =$$

$$10 \log_{10}(2h) + y_o + \frac{\log_e 10}{240}\left[\delta_{\frac{1}{2}} + \delta_{-\frac{1}{2}}\right]^2 + \frac{1}{6}\delta''_o \cdots$$

Both these formulae neglect fourth order terms and it must be remembered that they are not valid for $\delta > 10$ even if all other differences vanish. However, while they are useful for precise numerical integration, a formula sufficiently accurate for most purposes is obtained by dropping all the difference terms in the second form, and it is recommended that this be used in practice. The error of this approximation will be less than 0.1 db provided δ is less than about 1 db; note that this δ is for an interval of h while the ordinate is only taken at interval 2h where a difference of about 2 db could be tolerated.

When evaluating an integral over a long range this formula gives the contribution to each sub-range as the db value at the centre plus the sub-range length in db. Although these contributions cannot be added directly because they are in logarithmic form, the db value of the sum can readily be found using a table of addition logarithms. These are defined by

$$\log(a + b) = \text{greater of } (\log a, \log b) + \text{add log } |\log a - \log b|$$

where greater means algebraically greater. Tables of addition logarithms are given, for example, in Milne-Thomas and Comrie (1931)

but for limited accuracy the critical table given in the Appendix to this note will be adequate. This table is for use with decibels, so that if D_a and D_b are two powers expressed in decibels the db value of the sum of the powers D_{a+b} is given by

$$D_{a+b} = \text{greater of } (D_a, D_b) + \text{add db}|D_a - D_b|$$

The 'add db' is a function of the power ratio and the table is arranged so that for any power ratio between two tabular values of this quantity the appropriate value of 'add db' is also on the level between them. The 'add db' for a tabular value of power ratio is the higher of the alternatives - in critical cases ascend. Because of the rapid decrease of 'add db' as the power ratio increases, the integrals over sub-ranges need not be calculated so accurately when they are small compared with the total integral. For example an error of 1 db in an increment which is 10 db down on the total will produce an error of only 0.1 db; this relaxes the limit on δ to 5 db.

3. Numerical Example

To illustrate the use of the method without differences, the following table shows the evaluation of

$$10 \log_{10}\left[\int_0^{\pi} \frac{\sin^4 x}{x^4}\, dx\right]$$

given the integrand in dbs for $x = \pi/10, 3\pi/10 \ldots 9\pi/10$ so that $2h = \pi/5$ and $10 \log_{10}(2h) = -2.02$. This correction for the interval length has here been made in the sub-range contributions, but is more conveniently added on at the end after 'adding' the function values.

x	$(\sin^4 x)/x^4$	$\int_{x-\pi/10}^{x+\pi/10}$	working	$\int_0^{x+\pi/10}$
	(db)	(db)		(db)
0.1π	-0.28	-2.30		-2.30
0.3π	-2.64	-4.66	-2.30 + add db|4.66 - 2.30| =	-0.30
0.5π	-7.84	-9.86	-0.30 + add db|9.86 - 0.30| =	0.20
0.7π	-17.36	-19.38	0.20 + add db|19.38 + 0.20| =	0.20
0.9π	-38.45	-40.47		0.20

The correct value of the integral is 0.19 and this result may be obtained at the same interval by using more accurate values of the add dbs.

4. Logarithmic independent variable

When the function is to be integrated over a wide range of the independent variable it may be more convenient to express the latter also in logarithmic form. Alternatively the function may be given numerically at equal intervals in the logarithm of the independent variable - possibly at octave intervals. To find

$$10 \log_{10} \left[\int_a^b 10^{y/10} dx \right]$$

given y at equal intervals of

$$w = 10 \log_{10} x$$

the integral is transformed by this substitution giving

$$10 \log_{10} \left[\frac{\log_e 10}{10} \int_{10 \log_{10} a}^{10 \log_{10} b} 10^{(y+w)/10} dw \right] =$$

$$-6.37784 + 10 \log_{10} \left[\int_{D_a}^{D_b} 10^{(y+w)/10} dw \right]$$

This integral can be evaluated as described earlier treating (y+w) as the dependent variable given at equal intervals of the independent variable w.

5. References

Jacchia, L., Math. Tables Aids Comput. 9, 63 (1955).

Milne-Thomson, L.M. and Comrie, L.J., Standard Four-Figure Mathematical Tables (Macmillan, 1931).

6. Appendix

CRITICAL TABLE OF ADDITION DECIBELS

Power ratio in dbs	Add dbs	Power ratio in dbs	Add dbs
0		5.63	
	3.0		1.0
0.12		6.11	
	2.9		0.9
0.32		6.65	
	2.8		0.8
0.53		7.24	
	2.7		0.7
0.75		7.92	
	2.6		0.6
0.97		8.69	
	2.5		0.5
1.20		9.6	
	2.4		0.4
1.43		10.7	
	2.3		0.3
1.68		12.2	
	2.2		0.2
1.93		14.5	
	2.1		0.1
2.19		19.3	
	2.0		0
2.46			
	1.9		
2.74			
	1.8		
3.04			
	1.7		
3.35			
	1.6		
3.67		15.9	
	1.5		0.10
4.01		16.7	
	1.4		0.08
4.38		17.9	
	1.3		0.06
4.76		19.3	
	1.2		0.04
5.18		21	
	1.1		0.02
5.63		26	
	1.0		0
6.11			

"..... they look upon these things as Tales and Fancies, and will tell you that the Glasses do but delude your Eyes with vain Images; and even when they themselves consult their own Eyesight in the Use of these Tubes, the NARROWNESS OF THEIR MIND is such, that they will scarce believe their Senses when they dictate Ideas so new and strange."

Isaac Watts
1674-1748

FREQUENCY RESPONSE AND RESOLUTION OF DISPLAYS

by F. D. Boardman

1. Introduction

It has long been known that the information available in any display system is directly related to the resolving power of its constituent components, e.g. lenses, cathode ray tubes, etc. Some of the earliest work on the resolving power of optical systems was reported in 1834 by Airy who discussed the fundamental limit in resolution due to the finite aperture and this was later studied in greater detail by Rayleigh from the point of view of the ability of an astronomical telescope to resolve two adjacent stars. This led to a theoretical study of the nature of the image of an isolated star and the way in which the images of two stars (regarded as self-luminous points) combine. Any practical work, however, underlined the difficulty of determining accurately at what separation two stars just resolved. As Rayleigh remarks "a definite limit to an operation such as visual resolution (involving in some degree a mental judgement) is, of course, not to be expected". Furthermore the limiting resolution of each component of a display system is not sufficient to define the resolution of the whole system consisting of several such components. For example in a television system the camera, lens and cathode ray tube all degrade the sharpness of the final picture and it is not usually possible or even desirable to ensure that the resolving power of all the components except one is markedly greater than the resolving power of that one. Some means is required of defining the resolution of the individual components of a display in such a way that they can be combined to give the resolution of the whole system.

In a paper on the resolving power of microscopes Rayleigh[1] extends the study of the image of two points to that of a self-luminous grating. He showed that as the spacing was decreased the contrast in the image also decreased and the relation between contrast and frequency (the reciprocal of the spacing) was an accurate and quantitative measure of resolution. The use of frequency response to define the resolution of a system is quite familiar in electronic circuits, as, for example, the bandwidth of a radar determines its range resolution. A similar concept can be applied to display components. A frequency response can be defined analogous to that of a linear filter and with certain limitations can be used in a similar way, e.g. the frequency response of the

whole system is the product of the frequency responses of individual components. The method is applicable to television and film scanning systems, airborne surveillance and reconnaissance radars, line-scan reconnaissance, radio astronomy and to any system which attempts to produce as faithfully as possible a representation of analogue data of high information content.

The purpose of this article is to draw attention to the work which has already been done in this field notably by Schade in the U.S. and to describe briefly some recent contributions in this country. All that follows can be found in greater detail in the literature and it is hoped that the references listed will provide a convenient if not comprehensive introduction to the subject.

2. Frequency Response

Consider the factors which determine the sharpness of a television picture. The television system attempts to produce on the tube face the light distribution in the object scene as faithfully as possible. In practice, the imperfect resolving power of the various components of the system, camera, transmitting circuits, receiving circuits, cathode ray tube, destroys some of the fine detail, edges become less sharp and adjacent features merge. In particular an isolated light source of infinitesimal dimensions in the object scene produces on the tube face a spot of light of finite extent. The distribution of light intensity in this spot characterises the resolution of the system; the smaller the spot the better the definition and the sharper the picture.

To develop the theory further we must make the assumption that the system is linear in the sense that if there are two adjacent isolated light sources in the object scene the light distribution on the tube is the sum of the light distributions each would produce separately. Since this will only apply to small changes in light intensity the theory is limited to an object scene of low relative contrast. This is clearly a limitation in its application.

Making this assumption, let $G(y)$ be the light distribution on the tube face produced by an isolated small light source at $x = 0$ in the object scene, and $A(x)$ be the light distribution along one line of the object scene. Then the resulting light distribution on the tube face due to the object scene is

$$B(y) = \int_{-\infty}^{+\infty} A(x)\, G(y - x)\, dx \qquad (1)$$

Now any waveform in time can also be described by its complex

frequency spectrum. In a similar way we can describe the variation of light intensity with distance along the line by a spectrum of 'spatial' frequency-components. Thus, if we denote spatial frequency by s, the spectrum of the object scene is

$$a(s) = \int_{-\infty}^{+\infty} A(x) \exp(j2\pi sx)\, dx \qquad (2)$$

(The interpretation of the term 'spatial frequency' should present no difficulty. For example, $I = \sin 2\pi sx$ represents a sinusoidal distribution of intensity having s periods per unit length). The output light distribution also has a frequency spectrum defined by

$$b(s) = \int_{-\infty}^{\infty} B(y) \exp(j2\pi sy)\, dy \qquad (3)$$

The frequency response g(s) of the system can now be defined by analogy with linear filter theory by

$$b(s) = g(s) \cdot a(s) \qquad (4)$$

It must be emphasised that for g(s) to have any useful properties the system must be linear in the sense discussed above. Then, again by analogy with filter theory, it will be evident that the frequency response of the whole system is the product of the frequency responses of the various components of the system. It will thus be possible to estimate the effect on the resolution of the system of improving or degrading the resolving power of any one component.

In general the frequency response g(s) of a display component shows a low-pass characteristic (Fig. 1) having the value unity at low spatial frequencies and falling to zero at high frequencies.

The introduction of such a component into a display will have two prime effects on picture quality. Firstly the picture is less sharp, since the high frequency components in an edge are eliminated. Secondly it will improve picture quality by increasing the signal/noise ratio if the spatial frequency components in the noise are of wider bandwidth than the signal. Noise is introduced into displays by components having a grainy structure, such as film emulsion or the phosphor of a cathode ray tube. In most cathode ray tubes the spot size is the limiting factor in resolution and the finite width of the electron beam integrates the noise introduced by the phosphor;

consequently a reduction in spot size results in higher resolution but a wider bandwidth of noise due to phosphor grain. In a series

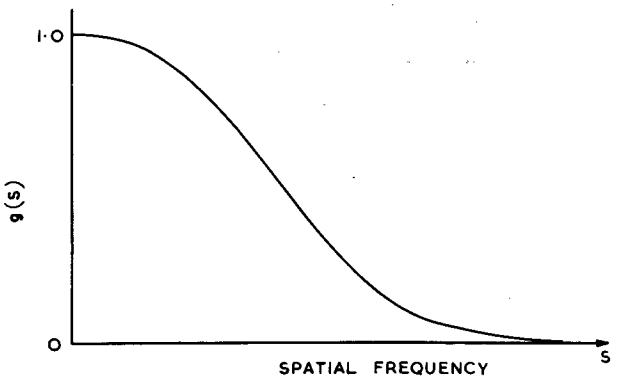


Figure 1 Spatial frequency response of a cathode ray tube

of papers Schade[2,3] studies in detail the frequency response of various forms of limiting aperture including the electron beam in a cathode ray tube, making some assumptions about the light distribution across a stationary spot. He estimates the effects of these apertures on signal/noise, graininess and picture quality in television and film scanning systems. Cameras, optical lenses, cathode ray tubes, films and so on are examined and techniques for aperture correction are discussed.

To simplify the assessment of the overall resolution of a system Schade describes the characteristics of a component by an 'equivalent pass-band' analogous to the noise bandwidth of a linear filter. This has obvious applications to the study of graininess. The equivalent pass-band is

$$N_e = \int_0^\infty |g(s)|^2 ds \qquad (5)$$

If the frequency response of a component is half-Gaussian, i.e.

$$g(s) = \exp(-s^2/2\sigma^2) \qquad (s > 0)$$

the equivalent pass-band is $\frac{1}{2}\sigma\sqrt{\pi}$. If the frequency responses of the several components of a system can be represented by half-Gaussian shapes with, in general, different values for σ, the frequency response of the whole system is also half-Gaussian with total σ given by

$$1/\sigma^2 = (1/\sigma_1^2) + (1/\sigma_2^2) + \ldots$$

It is evident, therefore, that the equivalent pass-band N_e of the system is given by

$$1/N_e^2 = (1/N_1^2) + (1/N_2^2) + \ldots + (1/N_r^2) \quad (6)$$

where N_1, N_2 ... N_r are the equivalent pass-bands of the individual components. In practice the shape of the frequency response of most components does approximate to half-Gaussian, particularly at low frequencies, and this is a useful method of assessing quantitatively the effect on the system of a change in resolution of a component.

3. Frequency Response of Components

We shall now turn to the problem of obtaining a measure of the frequency response of some of the components we have been considering. In recent years some work has been done on the measurement of optical systems and cathode ray tubes and this is outlined briefly below.

Two principal methods are used. The first is similar to that used in measuring the frequency response of linear filters, and is to measure the amplitude and phase of the output of the component for a sinusoidal input of constant amplitude and variable frequency. The difficulty with this method usually lies in providing the sinusoidal input in a convenient and practical form. The second method depends on the property that the frequency response is the Fourier Transform of the response to a unit impulse or delta function. For example in equation (2), if A(x) is a delta function at the origin, a(s) is unity. Substituting then in equation (4) gives g(s) = b(s). With this method the input generally presents few problems, but the measurement of the response to a sufficiently high accuracy is often difficult and calls for high stability in the operation of the equipment and the measuring gear.

(i) Optical Systems

The quality of an optical system is usually defined in terms of its aberration function. In Fig. 2 a point in the object plane is focussed by the optical system at a point P in the image plane. The wavefront emerging from the lens will in general not coincide with the sphere centred on P owing to the imperfections of the lens; the distance between the practical wavefront and the ideal one is the aberration. This is usually expressed as a function W(x,y) of the pupil co-ordinates (x,y). The skill in lens design consists of

adjusting the various parameters, radii of curvature, refractive indices and element spacing, to minimise W(x,y).

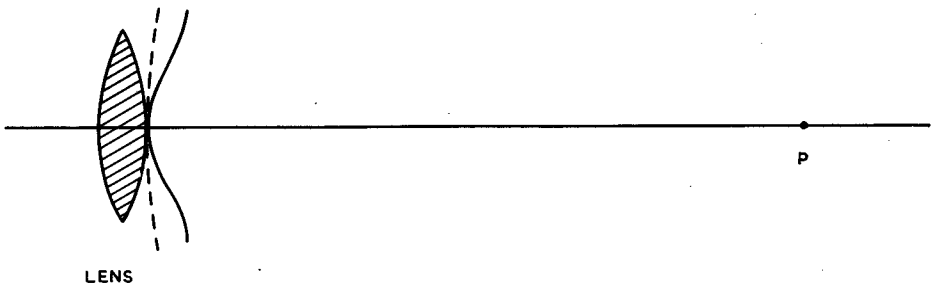


Figure 2 Aberrated wavefront from a lens

Hopkins[4] has evolved an elegant method of calculating the frequency response of a lens from the aberration function W(x,y). The response is expressed in terms of a normalised spatial frequency s defined by

$$s = \lambda R / \sin a \qquad (7)$$

where R = spatial frequency of object in cycles per unit length

λ = wavelength of the light

a = angular semi aperture subtended by the optical system at the object

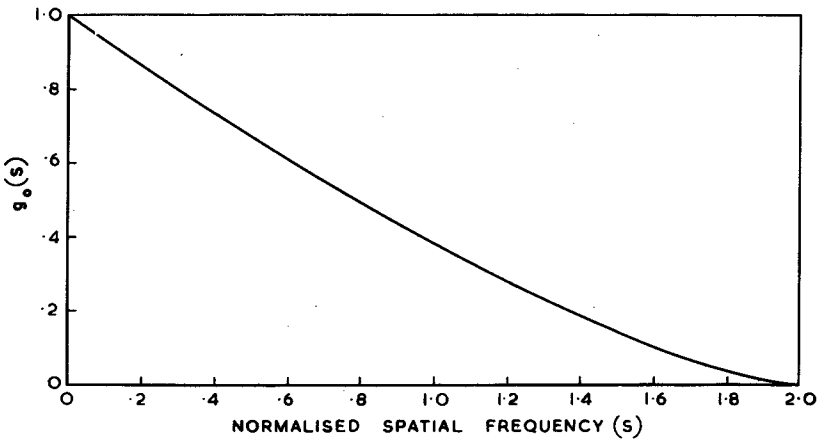
Assuming that there is no loss of light in passing through the lens and that it has a circular pupil of unit radius, Hopkins shows that the frequency response g(s) is proportional to

$$\int\int_{-\infty}^{\infty} \exp\left(2\pi j \Delta/\lambda\right) dx\, dy \qquad (8)$$

where $\Delta = W(x + \frac{1}{2}s, y) - W(x - \frac{1}{2}s, y)$

and x,y are the coordinates in the pupil. The proportionality factor is chosen to make g(0) = 1. Since the transmission of the lens is zero outside the pupil $|z| = 1$ $(z = x + jy)$, the area of integration reduces to that common to the two circles $|z - \frac{1}{2}s| = 1$ and $|z + \frac{1}{2}s| = 1$. This decreases as s increases and is zero for s greater than 2, which sets a theoretical upper limit to the bandwidth of the system.

The frequency response $g_o(s)$ of an aberration-free lens of finite aperture is obtained from equation (8) by writing $W(x,y) = 0$ over the area of integration. Fig. 3 is a plot of $g_o(s)$. In a practical lens the frequency response is further reduced by



Frequency 3 Frequency response of aberration-free circular lens

aberrations and Hopkins[5] shows how to calculate the further reduction in response due to defocussing, spherical aberration and astigmatism. To the equipment designer, who has little or no control over the resolution of his lens beyond careful selection from a shelf, defocussing is his main source of worry. Hopkins shows that for an otherwise aberration-free system the maximum axial displacement (δz) of the focal plane such that the response is not less than 0.8 of its value $g_o(s)$ when properly focussed is given to a close approximation by

$$\delta z = \pm 0.2/(R \sin a) \qquad (9)$$

The theoretical work of Hopkins on the effect of defocussing an optical system of high quality has largely been confirmed by measurements by Goodbody (as yet unpublished). The apparatus used is described in outline by Hopkins[6] and is basically a shear interferometer (Fig. 4). L is the lens under test illuminated by an object line source; D is a half-silvered mirror; M and P are corner-cube mirrors capable of lateral motion. Interference takes place at EE' between the aberrated wavefront via M and that via P. Hopkins shows that the total light intensity falling on EE' measures the response of the system to a spatial frequency determined by the lateral movements of the mirrors M and P from a mean position. Since the operation of the device depends upon the relative path differences in terms of optical wavelengths between rays travelling

via M and via P, very precise movement of the mirrors is required together with a complete absence of vibration in the system. A further difficulty is that the object line source must, of course, be monochromatic in the optical sense. Although an invaluable research tool the apparatus is not likely to find application in the practical measurement of lens performance.

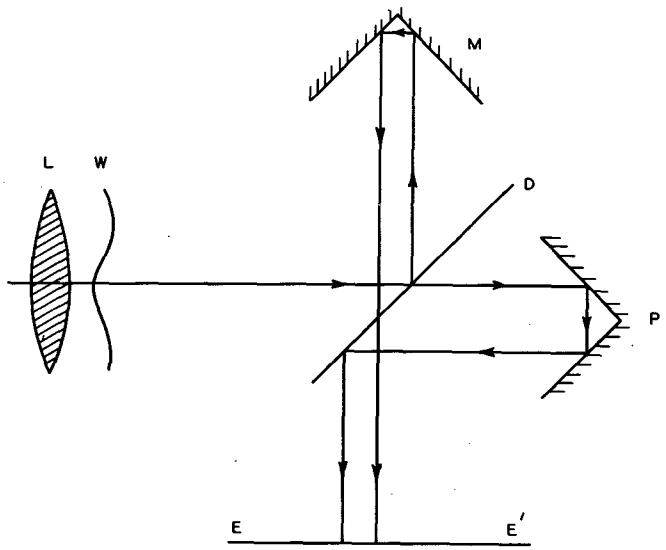


Figure 4 Shear interferometer for measuring spatial frequency response of a lens

A more straightforward method of obtaining the frequency response of a lens is described by Sproson[7] of B.B.C. Research Labs. This uses the delta function input method. A brightly illuminated narrow slit of light is the object of the lens under test and the light intensity in the image is scanned by a second slit of effective width much less than that of the image. The light output from the scanning slit is measured by a photocell and recorded. As we have seen, the frequency response is obtained by calculating the Fourier Transform of this record. The apparatus is bulky and demands a high degree of mechanical rigidity and accurate movement of the scanning slit. It has the advantage over the interferometer method that if the lens under test is well colour-corrected the source is not required to be monochromatic. Naish at R.A.E. has pointed out that the computation of the frequency response can be carried out directly by multiplying the output from the photocell by sin $2\pi sx$ and cos $2\pi sx$ in sin and cos potentiometers whose shafts are geared to the movement of the scanning slit. The image must be rescanned, of course, for each spatial frequency, s.

There seems little doubt that the frequency response of an optical system can be measured more simply by an object formed by a filter whose transmission varies sinusoidally in space when illuminated by a parallel light beam. To the author's knowledge such a light filter has not yet been realised practically. A cathode ray tube with a stationary sinusoidal line trace may provide a suitable object if a tube of sufficiently high resolution can be obtained, although difficulty will be experienced in obtaining the phase characteristic of the frequency response.

(ii) Cathode Ray Tubes

In the past the resolution of cathode ray tubes has usually been defined in terms of spot width or line width or by a 'shrinking raster' technique to obtain the limiting frequency. These methods are often very subjective and inaccurate. Recently Gill[8] has developed at R.R.E. an electronic method of measuring the frequency response of a cathode ray tube which is both objective and accurate. Fig. 5 is a block diagram of the apparatus. A sine-wave modulated

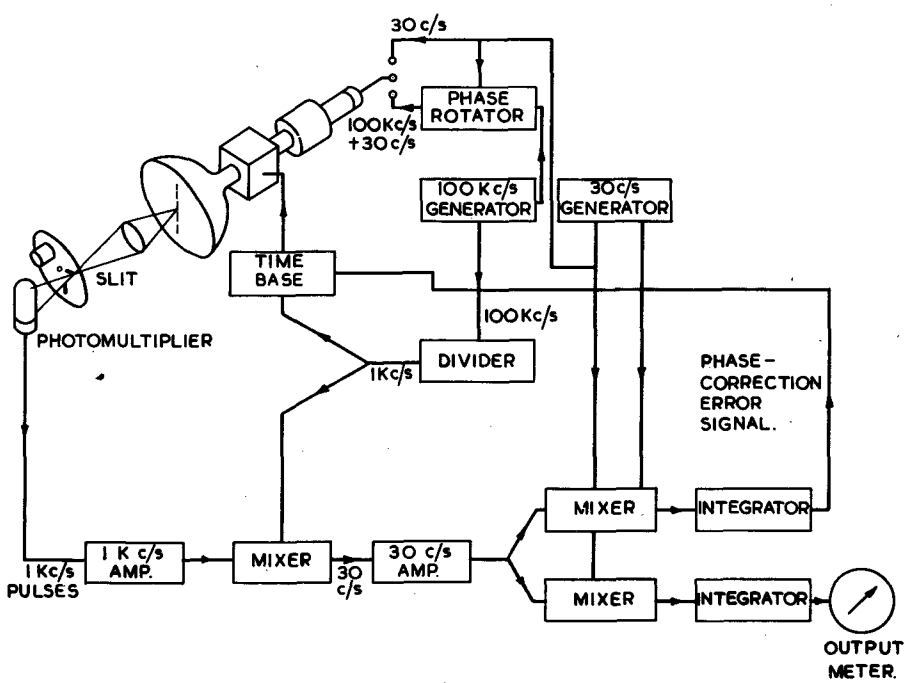


Figure 5 Apparatus for measuring spatial frequency-response of a cathode ray tube

line pattern is made to drift across the face of the tube, a narrow

slice of which is examined by a slit and a photocell. The pulses received by the photocell are modulated at the drift rate (30 c/s) of the sine wave pattern and the modulation depth is a measure of the response of thetube to the spatial frequency of that sine wave pattern. The spatial frequency in this equipment is selected by adjusting the scan speed of the trace. It will be seen (Fig. 5) that the low frequency (30 c/s) modulation is phase detected with a reference to give improved signal/noise ratio for work at low brightness levels. In a high resolution tube very high stability of the power supplies to the tube are required to maintain the sine wave pattern in its correct position to a small fraction of the spatial wavelength being measured. (In practice this implies a small fraction of the spot size.) To overcome this difficulty a servo is used on the position of the trace using information derived from the output modulation after phase detection with a quadrature reference. This, of course, discards any information on the phase of the frequency response but in most cases this is justified since the spot when properly focussed is symmetrical and its spectrum is thus entirely real.

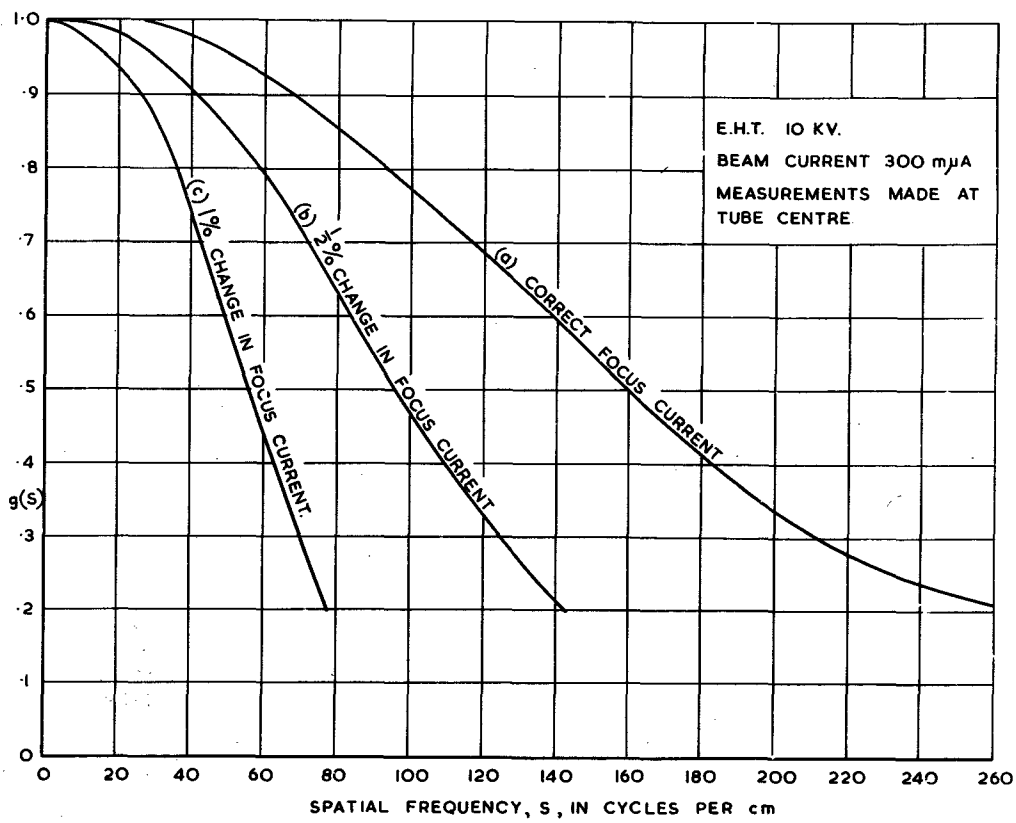


Figure 6 Effect of focus current variation on resolution

The frequency response of several tubes under different operating conditions has been measured with this apparatus. In particular the effect of variation of focus field and of deflection defocussing are shown in Figs. 6 and 7. In Fig. 6 the bandwidth

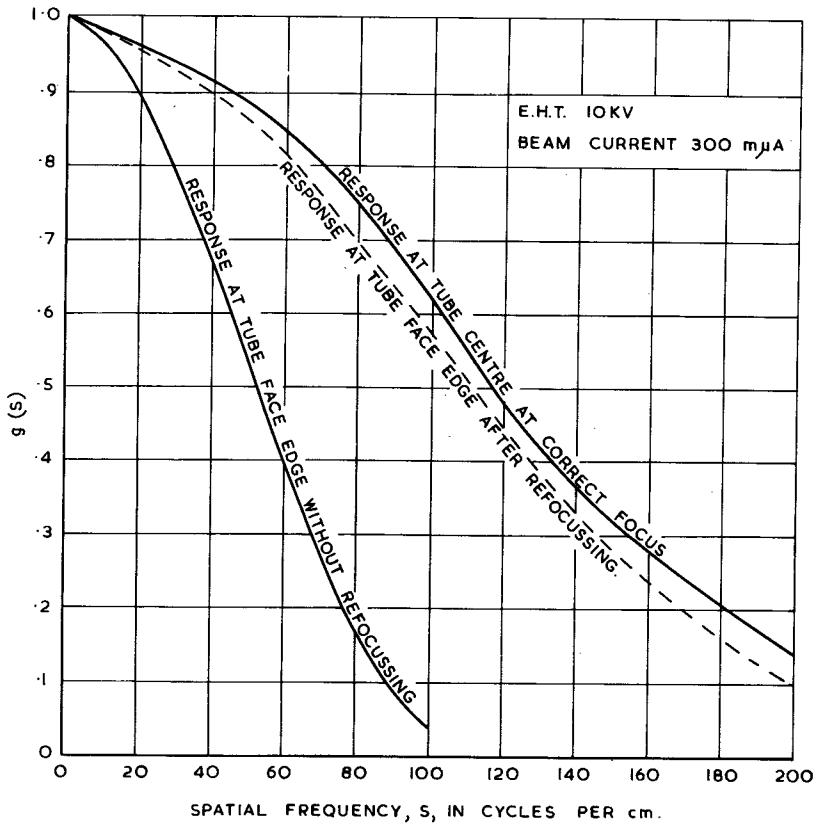


Figure 7 Effect of deflexion on resolution

for correct focus current (curve a) will be determined by aberrations in the electron lens and also possibly by the finite size of the grain of the emulsion. Curves b anc c show the reduction in bandwidth due to a combination of the above aberrations and defocussing. The focal length of an electron lens is known to be inversely proportional to the square of the focus current[9] and simple lens theory shows that, in a system in which the aberration is entirely due to defocussing, the spot size of a defocussed image is proportional to the change in focal length. Thus the bandwidth of an aberration-free system should be inversely proportional to the fractional change in focus current. Now the defocussed spot can be regarded as the convolution of the focussed spot with that obtained by defocussing a perfect point spot. The frequency responses are therefore multiplicative and assuming that the curves are approximately half-Gaussian, the principles of equation (6) will hold.

Using an obvious notation, the bandwidth σ_b' of a perfect system defocussed by an amount corresponding to a half per cent change in focus current will be given by

$$(1/\sigma_b')^2 = (1/\sigma_b)^2 - (1/\sigma_a)^2$$

From curves a and b we find

$$\sigma_a = 138 \text{ cycles/cm}$$

$$\sigma_b = 84 \text{ cycles/cm}$$

and hence

$$\sigma_b' = 106 \text{ cycles/cm}$$

Similarly

$$\sigma_c' = 54 \text{ cycles/cm}$$

and it is evident that the bandwidth is inversley proportional to the fractional change in focus current, as predicted.

In its present form Gill's apparatus is not suitable for the measurement of long after-glow tubes and a more direct method appears necessary. Ciuciura[10] has described an apparatus which measures the light distribution across a stationary line, a method analogous to that of Sproson for optical lenses. A very high order of electrical and mechanical stability is required but the designer claims that an accuracy of plus and minus three per cent has been achieved on relatively low resolution tubes. Its performance on the higher resolution tubes measured on Gill's apparatus is not known.

To the author's knowledge no attempt has yet been made to relate the frequency response to the aberrations of an electron lens in the way applied to optical systems by Hopkins and some work requires to be done here.

(iii) Aerial Polar Diagram

In surveillance radars and in radio-astronomy a limiting factor in the resolution of the system can be the polar diagram of the aerial. For example, as the aerial beam of a radar system scans a point reflector on the ground, the echoes are modulated by the square of the polar diagram and this in turn is converted by the display to a function $G(y)$ of the display co-ordinates y in the

direction perpendicular to that representing the beam. The Fourier Transform g(s) of this function will define the frequency response of the whole system. If the display resolution is sufficiently high g(s) will define the frequency response of the aerial polar diagram in the relevant co-ordinates.

Assuming a linear amplitude distribution across the aperture of the aerial, the polar diagram is approximately

$$\text{sinc}\ (L\theta/\lambda) \quad (\text{sinc}\ x = \frac{\sin \pi x}{\pi x})$$

where L is the width of the aperture. At a slant range R, the display co-ordinate is related to the angle θ by $y = R\theta$. Hence, since the polar diagram enters twice into the equation,

$$G(y) = \text{sinc}^2\ (Ly/R\lambda) \qquad (10)$$

In practice it is often more convenient to express the polar diagram as an approximation to a Gaussian function

$$\exp\ (-\theta^2/2\beta^2)$$

Then

$$G(y) = \exp(-y^2/2Y^2) \qquad (11)$$

where

$$R^2\beta^2 = 2Y^2.$$

The normalised frequency response is then

$$g(s) = \exp(-s^2/2\sigma^2) \quad (s > o)$$

where

$$2\pi Y\sigma = 1.$$

Referring to (10) it is easy to show that the shape of the beam near the centre approximates to (11) where

$$Y = \sqrt{\frac{3}{2}}\, R\lambda/\pi L.$$

Thus the frequency response is

$$g(s) = \exp(-s^2/2\sigma^2)$$

where

$$\sigma = (L/\sqrt{6})R\lambda$$

4. Sharpening Techniques

The possibility of defining the resolution of a component in a linear system by means of its frequency response simplifies, in theory, the problem of deducing the input distribution from the observed output. We have seen (equations 1 - 4) that if A(x) is the input distribution with complex spectrum a(s) and if g(s) is the frequency response of the system, then the Fourier Transform b(s) of the observed output B(y) is

$$b(s) = g(s).\ a(s)$$

Thus

$$a(s) = b(s)/g(s) \qquad (12)$$

In applications to radio-astronomy, where the input distribution A(x) is relatively simple and time can be spent analysing the output distribution B(y) this method has some uses. A comprehensive discussion is given by Burr[11] who also considers the uniqueness of the solution. However, in most applications, e.g. television, surveillance radar or reconnaissance, the output distribution is very complicated and the task of computing its Fourier Transform prohibitive.

In an attempt to overcome these difficulties Gouriet[12] proposes a method based on operations on the output waveform. If we can assume that at low frequencies the system response can be expressed approximately by

$$g(s) = (1 + g_1 p + g_2 p^2 + \ldots + g_r p^r)^{-1}$$

where $p = j2\pi s$, then the input A(x) is given by

$$A(x) = \left(1 + g_1 \frac{d}{dx} + g_2 \frac{d^2}{dx^2} + \ldots + g_r \frac{d^r}{dx^r}\right) B(x).$$

In a television or radar system the co-ordinate x in the direction of the timebase scan is proportional to time so that derivatives with respect to x can be carried out with respect to time in an electronic circuit. In practice Gouriet found that for television,

derivatives higher than the fourth are not usually required and useful sharpening of the picture can be obtained by use of the 1st, 2nd and 3rd order derivatives.

Ratcliffe[13] points out that the amount of sharpening which can be achieved is limited fundamentally by the fact that frequencies which have been removed by the limiting component cannot be reintroduced uniquely into the spectrum. In equation (12) if g(s) is zero at some frequency, a(s) is indeterminate. A further practical limit is determined by the signal/noise ratio in the output. For example, Ratcliffe shows that if the frequency response of the system is half-Gaussian,

$$g(s) = \exp(-s^2/2\sigma^2),$$

a gain in resolution of 1.6 times can in theory be achieved but at the expense of a loss of 23 db is signal/noise ratio. This clearly puts a severe limitation on the applications of these sharpening techniques.

In conclusion it is perhaps worth mentioning a subjective experiment carried out by Sproson[14] who invited a team of observers to estimate the relative sharpness of television pictures which had been degraded by certain low pass filters. The filters were of two basic types, one having a purely real frequency response to simulate a well-focussed cathode ray tube with a circular spot, the other having a non-zero phase component in the frequency response, e.g. an electrical lag. Sproson's conclusions were that the subjective sharpness correlated well with the bandwidth of the filter and that, for components and circuits having the same bandwidth, the distortion produced by non-phase shifting components (e.g. a cathode ray tube) is more acceptable to the eye than that of an electronic circuit which produces phase shift.

5. Conclusion

The resolution of a linear display system can be defined by its response to varying spatial frequencies. This will be determined by the frequency response of its individual components.

Methods have been evolved of defining and measuring the frequency response of an optical system and relating it to the aberration function. A method of measuring the frequency response of a cathode ray tube has also been developed.

Provided the signal/noise ratio in the output is adequate some picture sharpening can be achieved using the frequency response of

the system. This is probably more applicable to radio-astronomy than to television and radar applications.

6. Acknowledgments

The author is indebted to S. Ratcliffe and other colleagues for helpful discussion and to W. A. Woodley who carried out the measurements on cathode ray tubes.

7. References

(1) Baron Rayleigh, "Theory of optical images with special reference to the microscope," Phil. Mag. 42, pp 167-195 (1896).

(2) O. H. Schade, "Image gradation, graininess and sharpness in television and motion-picture systems," Jour. S.M.P.T.E. (February 1951, March 1952, August 1953).

(3) O. H. Schade, "Electro-optical characteristics of television systems" R.C.A. Rev. 9, pp 5-13 (March 1948), pp 13-37 (March 1948), pp 245-286 (June 1948), pp 490-530 (September 1948), pp 653-686 (December 1948).

(4) H. H. Hopkins, "The frequency response of optical systems," Proc. Phys. Soc. B 69, p 562 (May 1956).

(5) H. H. Hopkins, "The aberration permissible in optical systems," Proc. Phys. Soc. B 70, p 459 (May 1957).

(6) H. H. Hopkins, "Interferometric methods for the study of diffraction images," Opt. Act. 2, 1, pp 23-29 (April 1955).

(7) W. N. Sproson, "A new photo-electric optical bench," BBC Report T049.

(8) J. C. Gill, "A method of measuring the resolution of cathode ray tubes," RRE Tech. Note 638.

(9) L. Jacob, "An introduction to electron optics" p. 96 (Methuen).

(10) A. Ciuciura, "A cathode ray tube photo micrometer," Mullard Rep. VMA 529A.

(11) E. J. Burr, "Sharpening of observational data," Aust. Jour. Phys. 8, 1 pp 30-53 (March 1955)

(12) G. G. Gouriet, "Spectrum equalisation", Wireless Engineer, 30, p 112 (May 1953).

(13) S. Ratcliffe, "Correction for aperture distribution," RRE Memo. 1221.

(14) W. N. Sproson, "Subjective sharpness of television pictures," Electronic and Radio Engineer, 35, 4, p 124 (April 1958).

"There are many other things relating to MECHANICAL EXPERIMENTS, and to the Properties of the Air, Water, Fire, Iron, the Loadstone, and other Minerals and Metals, as well as the Doctrine of the sensible Qualities, viz. Colours, Sounds, Tastes, etc. which this Rank of Men cannot believe for want of a greater AMPLITUDE OF MIND.

The best way to convince them, is by giving them some Acquaintance with the VARIOUS EXPERIMENTS in Philosophy, and proving by OCCULAR DEMONSTRATION the multiform and amazing Operations of the AIR PUMP, the LOADSTONE, the CHEMICAL FURNACE, OPTICAL GLASSES, and MECHANICAL ENGINES. By this means the Understanding will stretch itself by Degrees, and when they have found there are so many new and strange Things that are most evidently true, they will not be so forward to condemn every new Proposition in any of the other Sciences, or in the Affairs of Religion or civil Life."

Isaac Watts
1674-1743

THIN FERROMAGNETIC FILMS

by A. C. Moore

Summary This article gives an account of the work at RRE on the preparation and properties of thin ferromagnetic films. The work provides the prospect of a computer store in which a million bits would occupy less than an eighteen inch cube, and which would possess a switching time faster than twenty milli-microseconds. The store would be cheap and easily constructed.

- List of Symbols -

B	magnetic induction.
B_{sat}	maximum value of B at saturation.
B_{rem}	value of B when H is reduced to zero.
H	field.
H_c	coercive field.
H_{app}	applied field.
t	time.
t_s	time for magnetization to change. (Usually time for dB/dt to pass from 10% of its max. value through maximum and then back to its 10% value. For fast switching, timing should start at the commencement of the field pulse. Some American workers use time for dB/dt to pass between $\frac{1}{2}$ max. value points, which gives faster times.)
M	intensity of magnetization.
M_s	saturation intensity of magnetization.
d	a critical dimension of the specimen.
oe	oersteds.
μsec	10^{-6} seconds.

mμsec	10^{-9} seconds.
μ	micron 10^{-4} cm.
A^{o}	angstrom unit 10^{-8} cm.
γ	gyromagnetic ratio.
σ	electrical conductivity.
A	exchange energy (spins).
K	anisotropy energy.
K_n	anisotropy constant.
E	free energy.
F_k	magneto crystalline energy.
λ	magnetostrictive constant.
p	uniaxial tension.
Ni	Nickel.
Fe	Iron.
α	damping factor.

Part I - General Survey

1. Introduction

Amongst the many devices which are needed in the construction of a modern digital computer, one which plays a very large part in determining the design is the high speed storage element. The specification of this element calls for (1) two well defined states which do not need power to maintain them and are therefore retained during a power failure, (2) a means of detecting rapidly in which state the element is, (3) a means of switching rapidly from one state to the other and (4) a well defined switching threshold to enable some form of X-Y coordinate system to be used. From the commercial point of view the question of cost is also important when a store consisting of a million bits or more may be under consideration; and in a store this size the cost includes not only that of the fundamental element, but also the cost of knitting it into a complete store.

One very obvious class that fulfils all the requirements of the storage element defined above is that of those ferromagnetic or ferrimagnetic materials which possess a rectangular hysteresis loop (see Fig. 1). Elements which possess the characteristics shown in

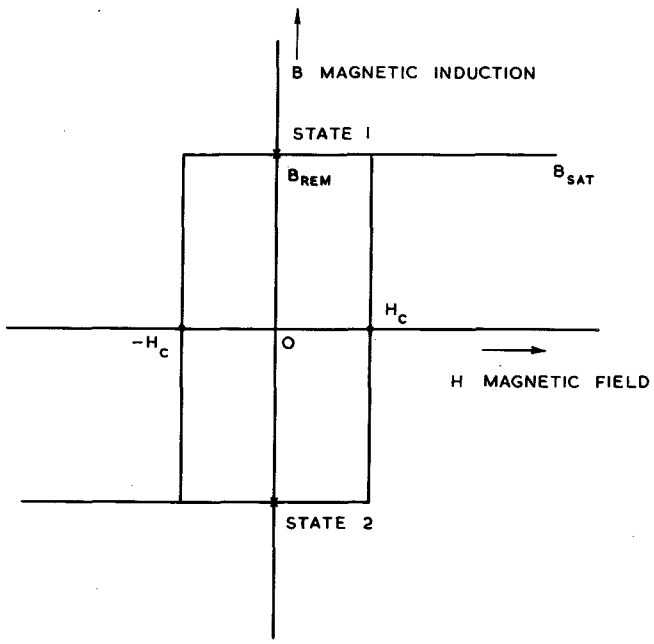


Figure 1 Ideal hysteresis loop

the figure are well on their way to satisfying the requirements of the storage element. In the absence of an applied field, states 1 and 2 are well defined and are maintained indefinitely. Which state the element is in can be detected by applying (say) a positive field for a sufficient length of time, and observing whether there is any change in the magnetic flux in the element. If it was in state 1 no change would ideally be observed; if in state 2, a change equal to twice the product of the cross sectional area of the element and the remanent value of the magnetic induction would be obtained. The element can be returned to state 2 by applying a field in the reverse direction of magnitude greater than the coercive field, H_c, for a suitable time. It is also possible to detect the state of magnetization by applying a large pulse of magnetic field for a time too short to change the magnetization of the specimen permanently. If the element is being driven further towards saturation, a smaller output is obtained than if it is being driven towards the knee of the B-H curve. This for obvious reasons is called non-destructive readout.

The existence of a well defined coercive force with a sharp transition from one state to the other also allows 'coincident current' working to be used. Strictly this should be called 'coincident field'. A field H_{app} less than H_c will not change the magnetization of the element. One greater, say $2H_{app}$, will do so if $2H_{app}$ exceeds H_c. Therefore two field-producing controls can be applied to each element and action occurs only if both are energised, each producing a field H_{app} in the same direction such that $\frac{1}{2}H_c < H_{app} < H_c$.

The one physical factor which has not so far been mentioned is the speed of the transition from one state to the other. The order of speed of change-over in a field of $2H_c$ for materials in current use is about 1 μsec but a fuller discussion of switching times will be reserved until later. This speed is too slow for proposed future computers, and the object of most of the thin film work is to overcome this limitation.

Finally there is the question of cost of the element and of the fabrication of a store. Small ferrite rings are the most popular high speed storage elements in present use. The cost per ring plus the cost of threading at least four wires through each core in a pre-arranged pattern works out at about 10d/bit. A megabit store will therefore cost over £40,000 without any of its associated circuits. Larger stores will cost more in proportion, and there is obviously room for economy here. It must be pointed out that the computer programmer is never satisfied with the amount of storage provided. He always wants more, and prefers to have it in a simple high speed form in which he can reach any bit of it in the same short time. For this reason large stores of the magnetic drum or tape variety are not adequate for high speed working. The time taken to abstract information from a serial store of these latter kinds can be quite long, although they may serve a useful purpose as a backing-up store, especially if their use can be foreseen and good arrangements made to transfer the relevant information into a high speed store.

2. Switching speed of magnetic elements

The change of magnetization in a conducting ferromagnet is in general accompanied by the induction of eddy currents following a well known law of Nature - resistance to change. In bulk materials these effects can be calculated from the classical equations of electromagnetism for a given shape and size of specimen, making various assumptions as to the continuous distribution of magnetization throughout the specimen, a constant value of permeability, and so on. In general the equations give the time for switching t_s as

an expression of the form

$$t_s = C\sigma M_s d^2/(H-H_c)$$

where σ is the electrical conductivity, C depends on the shape, d is a critical dimension of the specimen, M_s the intensity of magnetization, $H-H_c$ the applied excess field over a critical value H_c (below which no change takes place). It follows for example, that in thin sheets of material when the change in magnetization takes place from one surface to the other through the thin direction of the sheet that the time for complete change of magnetization with a given excess field is much reduced by thinning the sheet. This is a well known fact in the design of transformers. Unfortunately at thicknesses below 10μ in many metallic ferromagnetic materials the equation breaks down, and for a given excess field a constant time of changeover is obtained, independent of thickness. The limit of $(H-H_c)t_s$ is approximately 1/4 μsec. oe.

The reason for this failure must lie in one of the assumptions used to derive the equations; the assumption of a homogenous permeability is at fault. The Weiss theory of ferro-magnetism postulates quite large domains in which the direction of the elementary 'atomic' magnets all interact and point in the same preferred direction. This direction is one of a few chosen by the anisotropy of the material. Each domain is separated from its neighbour, whose magnetization points in a different direction, by a 'domain wall' which is of finite thickness. This thickness is determined by the balance between the energy of the exchange interaction between adjacent atomic magnets and the anisotropy energy of the material i.e. that energy which tends to keep the magnetization in the preferred directions in the material. The thickness is of the order of a few thousand A^o in permalloy (80% Ni 20% Fe). The material may be pictured as in Fig. 2. We can see that change of

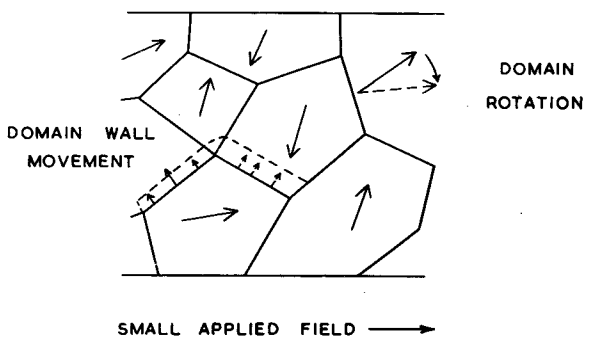


Figure 2 Magnetic domains

magnetization may take place in two distinct ways. In one, each domain rotates as a whole into line with the field against the anisotropy of the material and in the second, a domain pointing roughly in the direction of the applied field grows at the expense of its neighbour by domain wall movement.

The first case, that of domain rotation (but without anisotropy), was implicitly assumed in taking a uniform permeability throughout the material. As for the second case, it can be shown that in a thick slab of conducting material the motion of a domain wall is again limited by eddy current losses and the equation for the time of switching for a thick sheet, using the model of one or two domain walls sweeping through it, is of exactly the same form as the classical equation quoted above. In metallic films less than a few microns thick, however, and in ferrites, macroscopic eddy currents do not limit domain wall velocities and the product of t_s and $H-H_c$ is found to be either a constant, or proportional to d as in bulk ferrites. The limitation in domain wall velocity arises from some other factor such as the interaction between the elementary magnets and the crystal lattice.

Similar effects in the study of ferromagnetic resonance give rise to line broadening and a popular pastime at present is to relate the two effects by a simple equation of motion of the magnetic vector with an arbitrary damping term. The first and most popular equation is that of Landau and Lifshitz

$$d\underline{M}/dt = \gamma\underline{T} - \alpha\underline{M} \times \underline{T}/|M|^2$$

where

$\underline{M}$ = magnetic vector

$\underline{T}$ = torque, typically $\underline{M} \times \underline{H}$ in a simple case

γ = gyromagnetic ratio

α = damping factor

One is often left with some arbitrary constant; the number of domain walls in the switching problem. This can be adjusted to give the right answers.

In all cases so far mentioned change of magnetization has taken place by a domain wall motion, or its equivalent, through a thin direction of the specimen, the velocity being limited by eddy currents in thick metals and some intrinsic mechanism in thin ones and ferrites. It was pointed out earlier that a domain wall has a

finite thickness proportional to $\sqrt{(A/K)}$ where A is exchange energy and K the anisotropy energy. This is of the order of a few thousand A^{o} in nickel iron. If now we make a film too thin to allow domain walls to exist in its thin direction, what happens?

At first it was hoped that the thin film would change its magnetization by rotation i.e. all the elementary magnets would go round together in the plane of the film. This is obviously a much faster mode of operation than the successive changeover which occurs in domain wall motion. It was in fact found at first that domain wall motion occurred giving similar speeds of changeover to much thicker layers and to ferrites. The domain walls nucleated at the edges of the specimen and at imperfections in the film and spread across. Switch constants of $t_s(H-H_c) = 1/3$ μsec. oe. were obtained. Under suitable conditions, however, rotation could occur and fast changeover be obtained.

If it is assumed that rotation of magnetization takes place uniformly and simultaneously throughout the film, the application of one of the equations of motion of the magnetic vector gives some idea of the speeds that ought to be achieved. A number of points arise. If the reversing field H is exactly opposite to the magnetization direction no reversing takes place as the system is in unstable equilibrium (as far as rotational processes are concerned). Some form of misalignment or cross field or lack of uniformity in the direction of magnetization of the specimen is necessary to start the motion. If the misalignment is small a long induction period may occur before switching apparently starts. If the torque to change the magnetization $\underline{M} \times (\underline{H} - \underline{H}_c)$ is perpendicular to the plane of the film it is only after a small movement of $\underline{M}$ in that direction (which is opposed by the demagnetizing field of the thin film in that direction) that switching round takes place. If this movement is neglected, switch times in a small field of the order of one millimicrosecond are expected - using values for the damping factor derived from magnetic resonance experiments. If the initial motion is taken into account, changeover times of 5-10 m μsec. are forecast. It is perhaps relevant to point out that the damping factor apparently varies somewhat with frequency and the switch times are very sensitive to its value. Furthermore some recent experimental evidence suggests that $\underline{M}$ is not conserved and that the magnetic film separates into distinct regions which rotate independently of each other. The significance of this model is not apparent.

3. Thin film properties and switching by rotation

Obviously the film must be sufficiently thin to prevent domain walls moving in the thin direction, as this gives the easiest mode

of changeover and would have the lowest coercive force. Next, the coercive force for domain wall motion across the film must be as high as possible, and finally the coercive force for magnetic rotation must be made as low as possible, preferably less than that for domain wall motion if that can be achieved.

The first condition can be realized by using material with a low anisotropy and making the film thin enough - for 80-20 nickel iron the films must be less than 2000 A^{o} thin to be safe. The second condition is not yet easily controllable, for the theory of coercive force in magnetic materials has still to be fully worked out. Kersten has shown the importance of non-magnetic inclusions in modifying the domain wall energy, particularly at low fields, whilst Neel has shown the effect of stress centres particularly at high fields. Both theories require imperfections in the material of one kind or another, to oppose the motion of the domain walls. It is indeed suggested that in a perfect crystal (of sufficient size) no coercivity would be observed as there would be no impediment to domain wall motion, provided a domain wall could be formed, or existed for example at the surface. This suggests, fortunately, that absolute perfection is unwanted. Only experiment can tell us whether the coercive force for domain wall motion is sufficiently large. The demagnetizing energy of a domain wall of this kind increases as the film thickness decreases down to about 500 A^{o}. Below that thickness a new form of lower energy wall is formed. We should therefore make the films as thin as is reasonable to obtain a detectable output (but greater than 500 A^{o}), as this increases the domain wall coercivity. Finally there is the question of rotational coercivity. If domain wall motion is excluded and the magnetic film exists as a single domain and switches as a whole, in order to obtain two well defined magnetic states with a sharp transition threshold between them, the film must have some form of uniaxial anisotropy. In this preferred direction of magnetization a rectangular B-H loop will be obtained with coercivity equal to the rotational coercivity. At right angles to it there will be no remanent magnetization and the slope of B-H will also give a measure of the rotational coercivity (see Fig. 3).

To show this, consider the static case. M represents the saturation magnetization, which rotates as a whole and always lies in the plane of the film, making an angle θ with the preferred axis of magnetization. If K is the anisotropy constant, then for a field H_s applied along the preferred axis of magnetization the energy of the system is given by

$$E = K \sin^2\theta - H_s M \cos\theta$$

The hysteresis loop in the preferred direction is given by $M(\cos\theta)/H_S$ and it can be shown by equating to zero the first two partial derivatives of E with respect to θ that a reversible flux-change of 2M takes place when $H_S = 2K/M$. Similarly when a transverse field H_T is applied, we have

$$E = K \sin^2\theta - H_T M \sin\theta$$

and the hysteresis loop is given by $M(\sin\theta)/H_T$. It can be shown again by equating partial dE/dB to zero that saturation occurs when $H_T = 2K/M$. The results are given in Fig. 3. Incidently from the combined equation

$$E = K \sin^2\theta - H_S \cos\theta - H_T M \sin\theta$$

the effect of a transverse field on the hysteresis loop in the preferred direction can be found (for further details see D.O. Smith).

The immediate importance of the constant of anisotropy K can be seen. Although we require the rotational coercivity to be fairly small, too small a value could lead to trouble from stray fields (such as the earth's magnetic field) and from demagnetizing effects. Too high a value - even if the domain wall coercivity is higher still - makes it difficult for circuit designers to produce the required fields economically. Values of the orders of 1 oersted are a reasonable compromise.

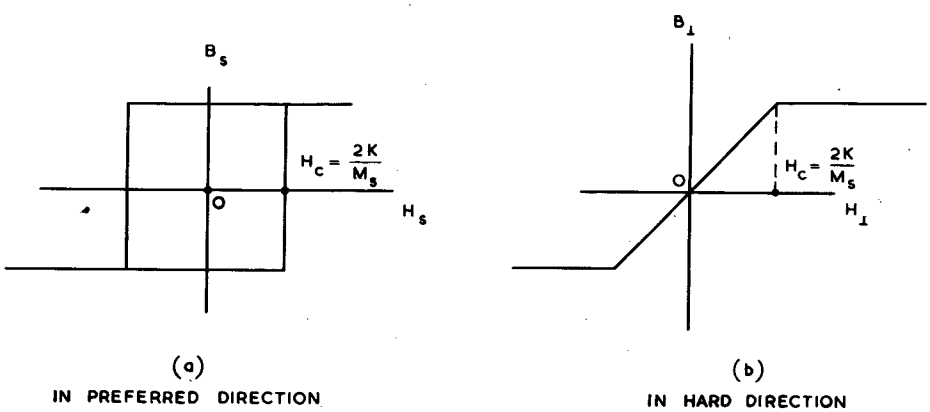


Figure 3 Rotational B/H loops of a material with uniaxial anisotropy

With films of this coercivity up to 2000 A° thick and more than 1 cm in diameter, demagnetizing effects in the plane of the film are negligible, although with thicker films an apparent shearing of the B-H loop will be obtained (see below).

4. Magnetic Anisotropy

We must next consider the various forms of anisotropy which can arise in a thin film. As their effects are additive, the greatest in magnitude will preponderate under a given set of conditions. From a practical point of view they can be classified as (1) Shape (2) Crystalline (3) Stress (4) Magnetic Field. There may be others.

(1) Shape Anisotropy If a magnetic body is magnetized in a given direction then magnetic poles will exist on the surfaces of the body. The field produced by these poles will tend to neutralize the applied magnetic field. The demagnetizing field ΔH is nearly proportional to the intensity of magnetization M in the desired direction. The constant proportionality N is called the demagnetizing factor. N can in general be calculated only for solids of revolution - infinite planes and similar convenient shapes. It is tabulated in various places, e.g. Stoner and Osborn.

In general a thin circular disc of diameter d and thickness t would show an isotropic demagnetization factor N approximately equal to $\pi^2 t/d$ in the plane of the disc and a factor $N = 4\pi$ normal to the disc. If the film was not symmetrical in its plane this would introduce an anisotropy into the energy expression. In general for a thin film of thickness 1000 A^o and other dimensions of 1 cm, if M_{sat} is 800 e.m.u. max. then ΔH is approximately 0.1 oe. max. and is just about negligible. If however the thin film is not continuous and comprises small islands of magnetic material with a preferred long direction, then this would cause a troublesome anisotropy.

(2) Magneto-crystalline anisotropy In a single crystal of magnetic material the energy of magnetization is a minimum along certain crystal axes, (100) in iron, (111) in nickel. For a cubic crystal (such as iron and nickel) the magneto-crystalline energy is

$$K_o + K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2 + \ldots$$

where α_s are the direction cosines referred to the crystal axes.

In poly crystalline materials the effect of crystalline anisotropy shows up in the approach to saturation of the B-H curve. Fields of the order of K_1/M_s are needed to reach 90% saturation.

(3) Magneto-strictive effects The effect of a magnetic field in changing the dimensions of a magnetic material is well known, although the details are somewhat complex. The inverse is also true. A uniaxial mechanical tension applied to a magnetic material with a positive coefficient of magnetostriction tends to

make that direction a preferred axis of magnetization and introduces a term of the form $(3/2)\lambda p \sin^2\theta$ into the expression for the energy, λ being the magnetostriction constant, p the tension, θ the angle between M and p. For a negative magnetostriction coefficient, the preferred direction of magnetization is at right angles to the tension. The converse is true of compression. Both the magneto-crystalline anisotropy and magnetostrictive anisotropy are explained at present in terms of spin-orbit interactions and are related to each other.

(4) Magnetic field anisotropy Magnetic materials of certain compositions, when annealed in a magnetic field, retain the direction of that field on cooling as a preferred direction of magnetization. The reason for this phenomenon has not been fully explained in the case of homogeneous alloys. Suggestions have been made that long range order, short range order or preferred arrangement of small amounts of impurities might be responsible for the effect. It does however offer a prospect of inducing a preferred direction of magnetization by an external control, provided the other anisotropies can be subdued.

It would appear that the best hope for producing a thin film with a controlled preferred direction of magnetization, low anisotropy and a low coercive force for rotation, is to choose a material with zero crystalline anisotropy and zero magnetostriction, ensure that no shape anisotropy occurs and induce a magnetic

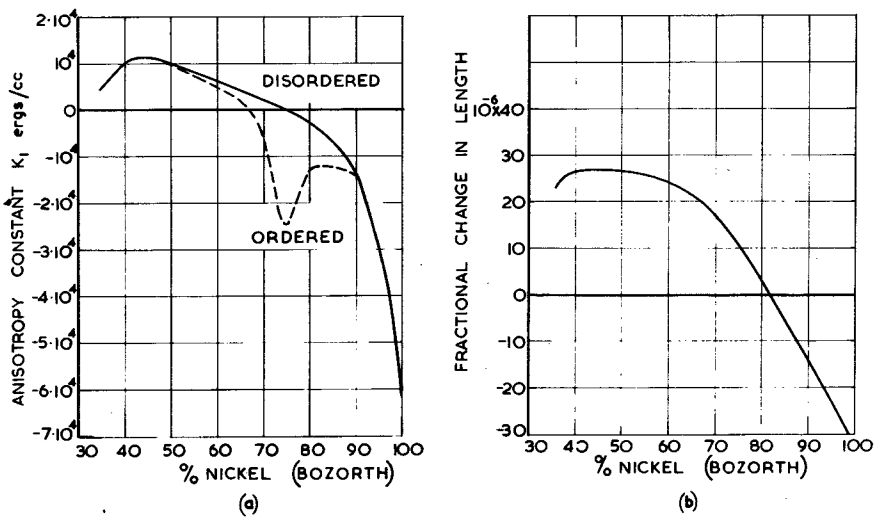


Figure 4

(a) Magnetocrystalline anisotropy v. composition of NiFe alloys

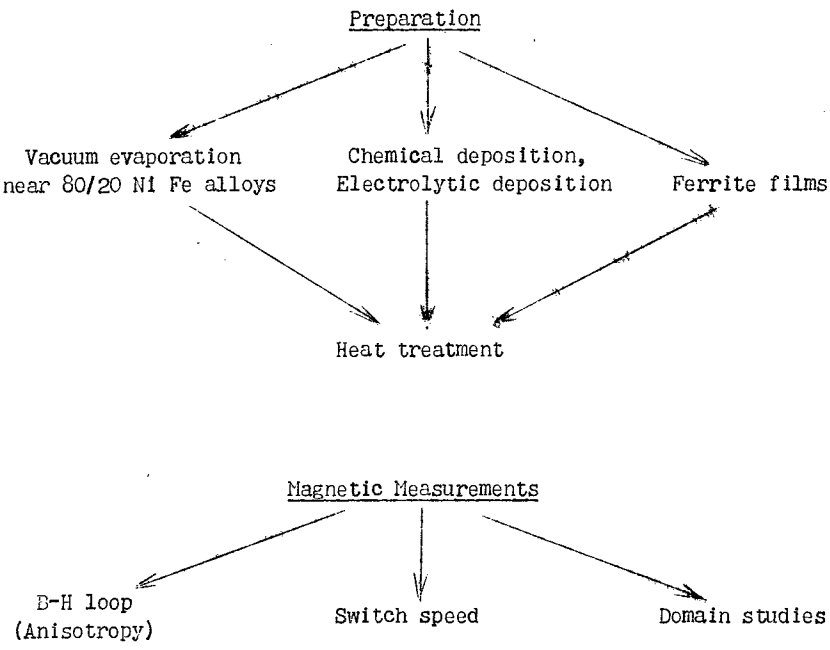
(b) Magnetostriction (saturated) v. composition of NiFe alloys

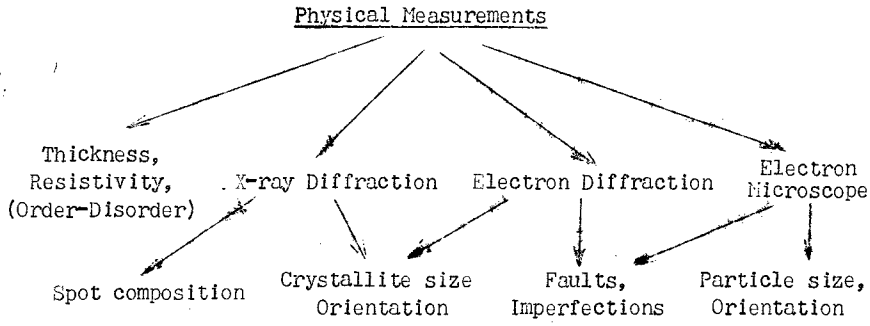
anisotropy in the chosen direction. Films of the magnetic elements are ruled out. The only two component alloy approximating to this requirement is nickel-iron which gives zero magnetostriction in polycrystalline material at about 83% Ni 17% Fe and zero magneto-crystalline anisotropy at about 72% Ni 18% Fe (see Figs. 4, a, b). To obtain a material in which both the anisotropies are zero at the same composition, a three-component alloy is needed, but this introduces more difficulties in controlling its composition in thin films. Most of the alloys containing nickel show Magnetic field anisotropy.

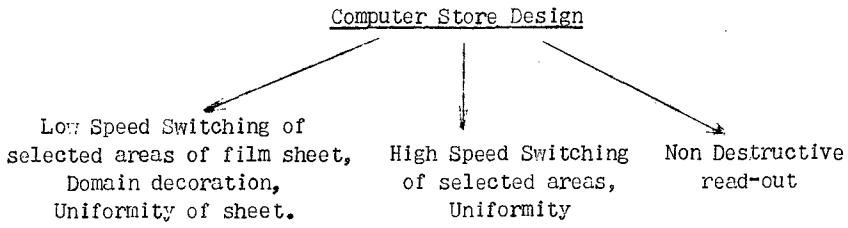
Part II - Programme of work at RRE

5. Introduction to Part II

There now follows a survey of the work at RRE on the preparation and properties of thin films. This programme must be viewed in the light of Part I of this article, and it must be borne in mind that the work itself is not by any means at a conclusive stage. It is summarized in the following diagram.







6. Preparation of films

Some initial experiments by A. S. Young on the direct rolling of magnetic tape to produce thin layers ceased at about one micron thinness. After that a lace effect was produced. There is a prospect of using a sandwich technique similar to that used for Wollaston wires but this has yet to be proved.

The vacuum evaporation of the magnetic alloy and its deposition on to a suitable substrate is a far quicker and easier method and offers the possibility of being easily adaptable to many different starting materials with a minimum of trouble. This method has been used by earlier workers in the field, notably Blois, and has also been developed at RRE. The basic conception is simple. The chosen material is placed in an enclosure which is evacuated to a pressure of about 10^{-5} mm of mercury. It is heated until it starts evaporating and the material then condenses on any (and all) convenient surfaces in the vacuum enclosure in a direct line from the starting material. The mean free path of N_2 at 10^{-5} mm is 6.5 metres, so diffusion out of the beam does not occur appreciably. As is only to be expected the process is not so simple as it appears. Blois pointed out that if equilibrium obtained between the starting alloy and its vapour the chemical composition of the vapour would be

changed. In the case of 80/20 Ni Fe, assuming Raoult's Law to hold (that the lowering of the vapour pressure of a solvent is proportional to the concentration of solute), an iron enrichment of about 16% should occur. Work by Goldfinger of Brussels University using a mass spectrometer to analyse the vapour, suggests that an enrichment of this order can occur under conditions which presumably approach equilibrium. On deposition on a cooler substrate it is assumed that instant freezing takes place with no subsequent re-evaporation and that the deposited film is the same composition as the vapour. The fact that our deposited films are near the starting composition shows that equilibrium is certainly not obtained. The evidence suggests that the faster evaporation takes place the nearer the composition is to that of the starting material. With evaporation rates of a few thousand A°/minute the iron enrichment does not exceed two per cent and is fairly reproducible. It is better to bring the starting material up to the desired temperature while the surface to be covered is shuttered off from the beam. The shutter is then removed and evaporation proceeds at a fairly constant rate until the desired thickness is obtained. This can be estimated by time alone or, better, by measuring the electrical resistance of the film between two control electrodes. The shutter is then replaced.

Heating of the starting material presents certain problems. Temperatures near 1400°C are needed. But molten nickel iron affects most metals commonly used for heaters (e.g. tungsten, molybdenum). The starting material should be placed in a crucible, but the crucible material is again critical. Silica has so far been ruled out because of the effects of traces of silicon on the magnetic properties, particularly the initial permeability, of common iron nickel alloys. Whether it affects the switching properties is not known. Alumina is quite suitable, apart from the scum which it forms on the surface of the melt and which reduces the rate of evaporation. Zirconia appears to be the best material so far. A $3\frac{1}{2}$ kilowatt induction heater gives adequate power to heat the starting ingot, whose mass is of the order of twenty grams at most. Considerable skill is required in the arrangement of heater, crucible and substrate to obtain uniform evaporation on to a surface which is, at present, about three inches square. As the surface is flat, a suitably large distance between crucible and surface is necessary to obtain negligible falling off in thickness of film at the edges.

The substrate itself is of very great importance. It must be inert, to prevent chemical changes in the deposited films. It should be amorphous to discourage the growth of large oriented crystals in the film. It should be rigid to prevent strains being developed. It should have a similar thermal coefficient of expansion to the nickel iron and should be capable of being heated to temperatures of near 500°C without softening, to enable annealing to

take place. Its surface must be smooth to prevent shape anisotropies developing. It happens that soft glass microscope slides and cover glasses fit these requirements remarkably well and they are extensively used.

The preferred direction of magnetization is, in practice, artificially set up during evaporation by applying a uniform magnetic field of about 50 oersteds parallel to the surface to be deposited upon. The substrate is heated to about 300°C. This has been found to work best in practice. Temperatures of this order are sufficiently high to enable the magnetic anneal to be effective. It is also believed that they encourage the formation of a uniform film with no isolated islands. The coercive force of films deposited on substrates at room temperature can be as high as 50-60 oersteds or more, but this drops rapidly to 1 or 2 oe when the substrate has been heated during evaporation. The magnetic preferred direction can be changed by subsequently annealing in a different magnetic field at temperatures above 300°C.

There is a limit to the size of uniform evaporated sheet which can be produced in present evaporators. Physical limitations are introduced by the size of the evaporation chamber. The vacuum pumps must be capable of keeping the chamber evacuated to about 10^{-5}mm even when the starting material and its crucible are heated red hot. Even so, the prospects of producing thin films for storage purposes by evaporation is quite feasible and looks fairly cheap compared with ferrite rings. Some form of building up in blocks a few inches square could be used, although if sufficiently large numbers were needed a suitable evaporator of much larger blocks could be built. An even cheaper method of producing as large a sheet as required is that of chemical or electrolytic deposition. Once again, however, similar difficulties arise in the control of composition of the deposited layer and the control of its mechanical condition. The properties of the substrate are again of major importance. In fact the whole problem is on a similar scale to that of preparation by evaporation. Enough preliminary work has been carried out at RRE to show that the method is promising, but no films have yet been obtained with as low a coercivity for rotation as the evaporated films. Most of them have had no preferred orientation and switching has been relatively slow, probably by domain wall movement.

7. Magnetic Measurements

There are two fundamental measurements which must be made on all magnetic films, namely the B-H loop in the preferred direction and at right angles to it, and also the changeover time.

B-H loops can be obtained in various ways, mostly rather

tedious. As high accuracy is not required and speed is essential if any number of specimens are to be examined, and as it is desirable to have a continuous display of the loop to enable the preferred direction to be found, a cathode ray tube presentation is indicated. A film 1 cm wide, 1000 A^{o} thick, and of saturation induction 10,000 gauss, produces a flux change of 0.2 line, on changing magnetization. If the substrate is 1 mm thick and of the same permeability as air, the flux change in the substrate alone, due to a field change of 2 oersteds is the same. In fact the 'air' linkage of any pick-up coil round the specimen is greater than this, by at least a factor of ten. This air flux can be easily eliminated by using a pair of balanced pick-up coils in opposition, and measuring the departure from balance obtained by inserting a specimen in one of them.

The basic apparatus is shown in Fig. 5, which is self explanatory. A few points must be noted. To portray a rapid

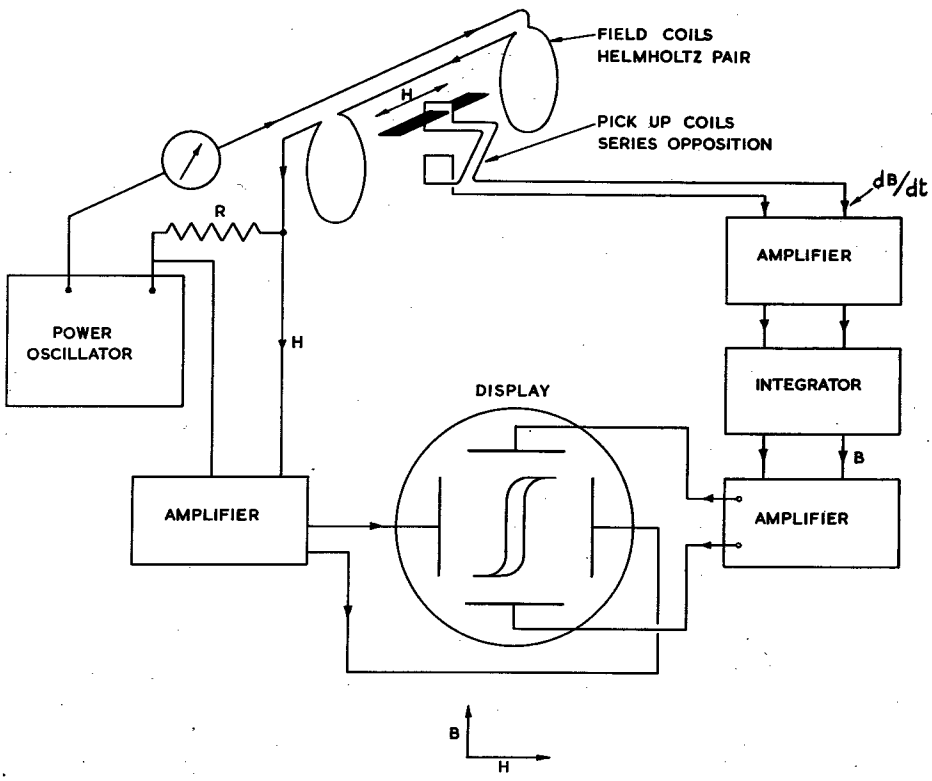


Figure 5 B-H loop apparatus

change in magnetization, the resonant frequency of the pick-up coils and the upper frequency limit of the amplifiers must be kept high in relation to the recurrence frequency of the whole cycle. If a rapid change takes place the output voltage dB/dt of the pick-up coils is large and the preamplifier must have a large signal handling capacity. The preamplifier gain must be sufficient to make any noise contributions from the post-integrator circuits negligible. The recurrence frequency should be as high as possible to minimise the amount of flicker noise from the preamplifier but the low frequency cut off of the amplifiers and integrators must not introduce any phase shifts at the recurrence frequency. Finally correction circuits, apart from the initial balance adjustment between the two pick-up coils should be eliminated. It is possible to obtain any shape of B-H loop required if an adequate number of correction adjustments are provided. Typical results for variation in coercivity with thickness in the preferred direction are shown in Fig. 6. The reasons for the sudden increases near 3000 A° are not known.

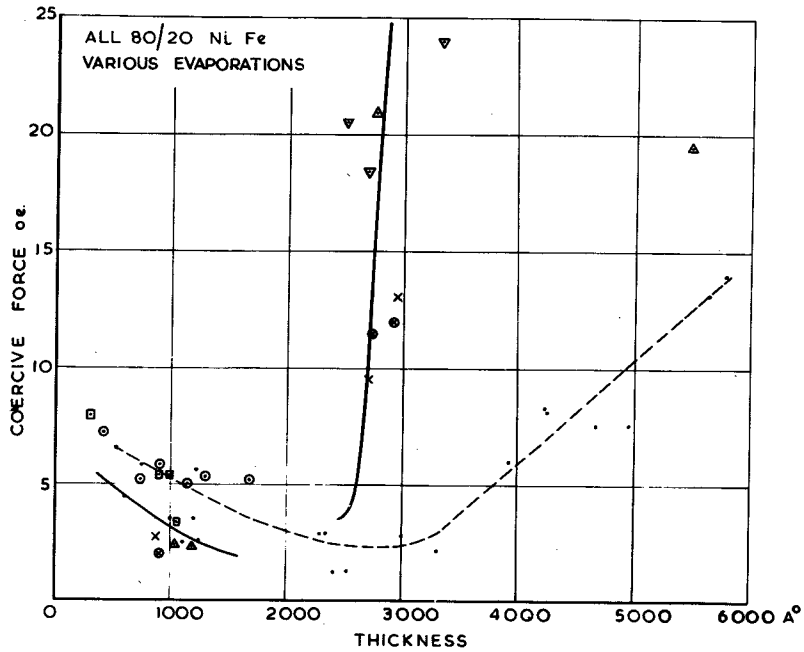
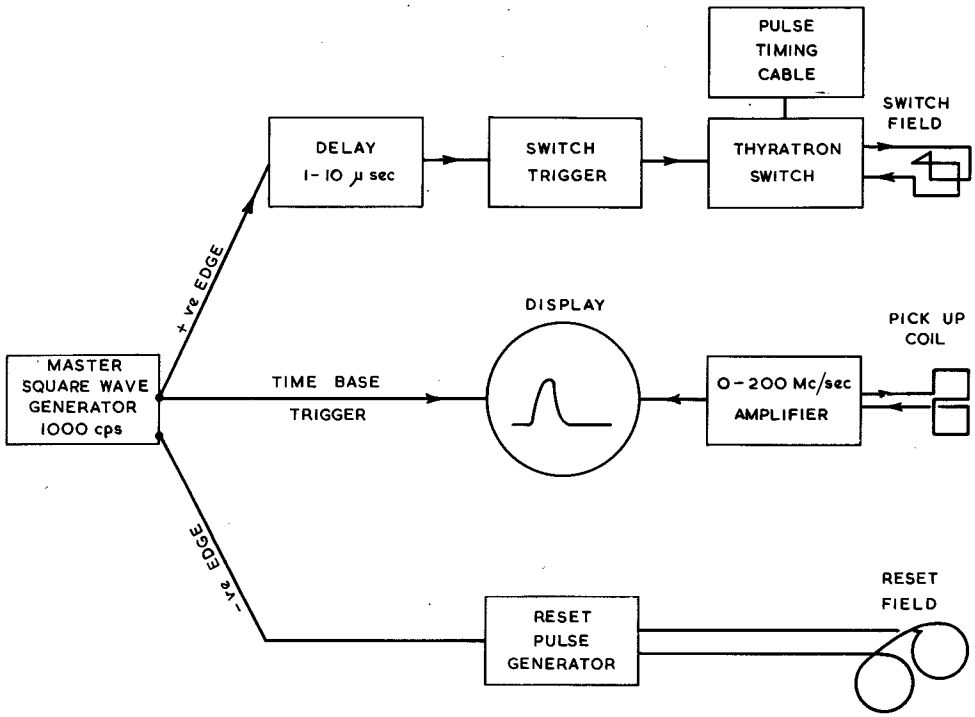


Figure 6 Coercivity v. thickness of evaporated NiFe films

8. Measurements of rate of change of magnetization

These measurements are carried out in an apparatus similar

in principal to the B-H loop apparatus, but scaled up in frequency. The integrator is left out and dB/dt is given as a function of time. The apparatus is shown in Fig. 7. A uniform field over a reasonable area is produced by an Helmholtz pair of coils of one turn each,



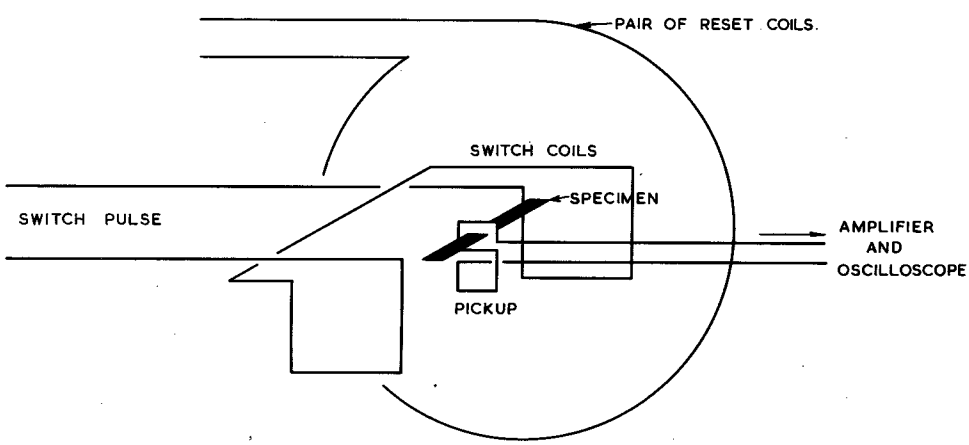


Figure 7 High speed switch apparatus

driven from a length of charged coaxial cable using a CV372 hydrogen thyratron as a switch. Rectangular pulses of current up to 15 amps can be obtained with a rise time of about 30 mμsec. and a duration depending on the length of the cable (up to 1 μsec). Switching times faster than 30 mμsec. can be measured, as the 30 mμ seconds include the slow start to the rising edge of the pulse. This is ineffective as switching takes place only when the coercive force has been exceeded. In fact the limit of the apparatus is about 10 mμsec. from this point of view. No oscilloscope amplifier is adequate at these speeds, so a distributed line amplifier of gain 2000 and frequency up to 160 Mc/sec. is used to drive directly the deflection plates of a high speed oscilloscope. Typical results are shown in Figures 8 and 9 in which the reciprocal of switching time is plotted against field for the preferred direction of a film

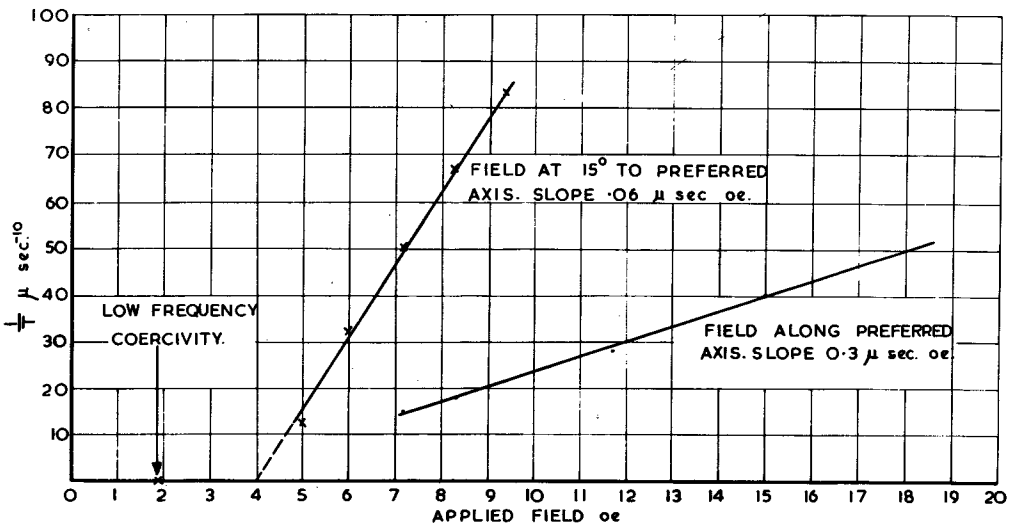
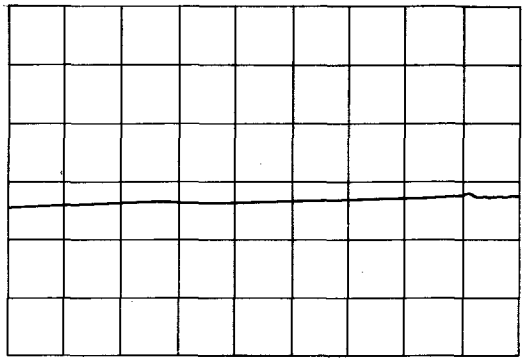


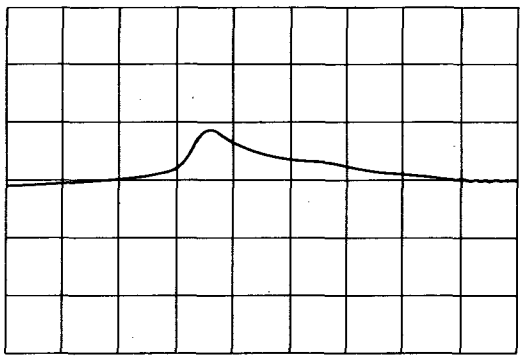
Figure 8 Switching time v. field for film in Fig. 9

and at 15° to that direction. The speed-up is very apparent and it is obvious that very fast switching of the order of 20 mμsec. can be obtained in a field of twice the intercept of the curve with the axis. This is perhaps unfair, as the DC coercive force is only about two oersteds. In practice, however, pulsed fields would be used, and it is probable that no switching would occur with fields below four oersteds provided their duration was less than 20 mμsec. This remains to be investigated.

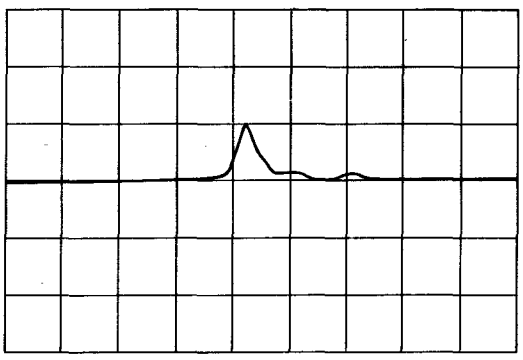
It is unfortunate that only a few specimens show quite such an improvement in switching speed on misalignment with the switching field although all with a preferred direction show it to some extent. One of the objects of the RRE research is to be able to produce elements with these fast characteristics every time.



a. 10 oe DRIVE PULSE NO RESET.



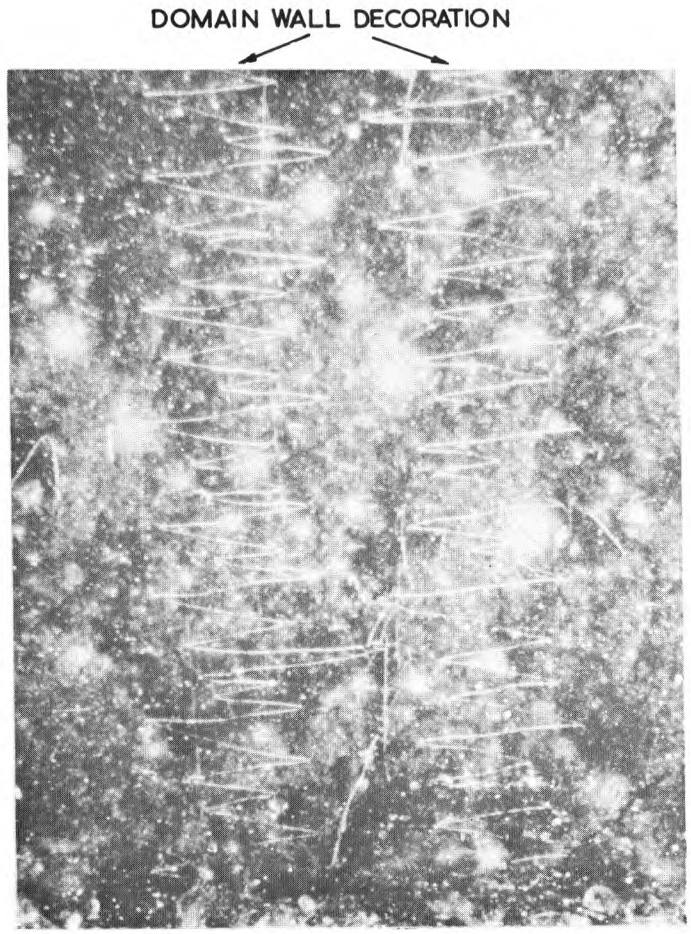
b. 5 oe DRIVE PULSE WITH RESET.



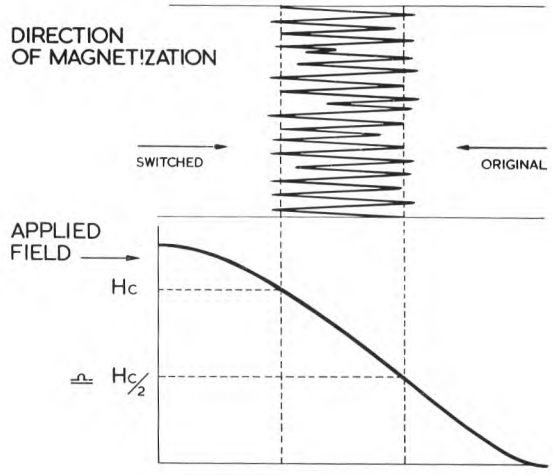
c 6 oe DRIVE PULSE WITH RESET

Figure 9

Output pulses from high speed switch apparatus 80/20 NiFe film 2,000 A^0 thick 15^0 to preferred direction. Scale: horizontal 20 μsec./div, vertical 200 mv./div.



(a)



(b)

FIG. 10 DOMAIN WALLS ON Ni Fe FILM

9. Domain Wall Decoration

In a thin film of NiFe, the direction of magnetization is usually in the plane of the film, yet when a domain boundary occurs, a component of magnetization normal to the film can occur as the direction of magnetization rotates from one domain direction to the next. It is then possible to show up this boundary by using the modern version of the 'iron filings technique' for showing up the poles of a magnet; colloidal magnetite is used in solution and concentrates on the boundaries. This method - first suggested by Bitter - usually relies upon a variation on a recipe given by Elmore for producing and stabilizing the colloid. However these recipes have all proved too corrosive for long time observation of domain patterns on thin films, and the one used at present was developed by C. Fuller. In this, equimolecular proportions of FeO and Fe_2O_3 are coprecipitated, filtered, washed and boiled up with soap solution; no hydrochlorine acid is used (as in the other recipes) to peptize the Fe_3O_4. Domain walls can be seen easily on NiFe films whose thickness lies between 700-2000 A^o. Above 2000 A^o the domain walls become broad and diffuse, below 700 A^o they become too narrow and a new form of domain wall develops in which the magnetization rotates in the plane of the film. This does not attract the colloid.

This technique is of great use in determining the extent of the change in the magnetization of a thin film when it is subjected to a non-uniform field. Some illustrations of this effect are shown in Fig. 10. Here the magnetic film has been magnetized to saturation in one direction and exists as a single domain. A wire carrying a current is laid on the surface to produce a reversing field underneath it. The horizontal component of the reversing field in the plane of the film is calculated, and it can be seen that portions of the film are changed over until the field has fallen to approximately a half of H_c. The boundary zig-zags to minimise the energy, but exact details are again somewhat obscure. Considerable fine detail can exist in the structure of these domain walls, all to minimise the wall energy, and it is apparent from the figure that two different types coexist, one showing as a faint fine line, the other a heavy double line looking slightly out of focus and apparently dotted.

This method enables the uniformity of the film to be tested from spot to spot and the preferred axis can be derived from a study of zig-zags.

10. Chemical Tests

The only chemical tests carried out on the films have been to

determine their average composition. A film 5 cms by 2 cms by 1000 A° contains about 8 x 10^{-4} gms of material and if it is desired to carry out an analysis to better than 1% then an accuracy of a few micrograms must be achieved. Nickel and iron are not very convenient substances to estimate in the presence of each other and a preliminary separation must first be carried out. The quantities present are then estimated by a polarograph, and some results are shown in Fig. 11.

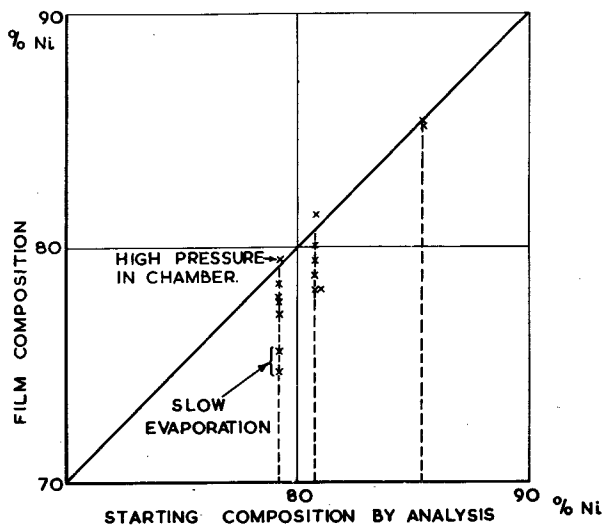


Figure 11 Chemical analysis of deposited films v. starting material

11. Physical Tests

The chemical tests give only a mean composition over the whole film; it would be advantageous to obtain spot compositions to check for varying composition (and hence anisotropy) over one film. Now X-ray defraction techniques will determine to high accuracy the lattice parameter of the material. As the nickel and iron form a continuous solid solution in the range of compositions we are interested in, the lattice parameter varies linearly throughout the range. There is evidence of a slight departure from linearity near 75 Ni 25 Fe, occasioned by the formation of Ni_3Fe, and this is shown in Fig. 12 where lattice parameter is plotted against composition. By using back-reflection Debye Scherrer patterns, a small change in lattice parameter shows up as a large change in ring diameter and it is possible to plot composition against ring diameter assuming there are no other causes of variation in lattice parameter (Fig. 13).

Possible sources of variation other than composition are temperature and strain. A strain corresponding to the elastic limit

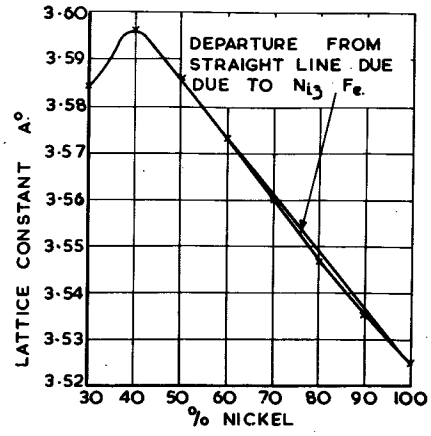


Figure 12 Lattice constant v. composition of NiFe films at 15°C

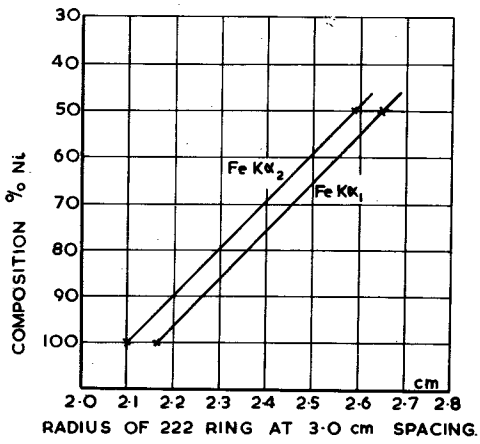


Figure 13 Composition v. diameter of diffracted rings at 3 cm. spacing

gives a 0.001% change in lattice parameter corresponding to a composition change of 3%. This is about the limit of accuracy of the measurements. This limit arises from the diffuseness of the diffracted rings caused by finite crystallite size in the direction normal to the thin film. (The diffuseness is too much to be attributed solely to strain variations). A very approximate estimate of crystallite size can be made - see Fig. 14. Annealing the films improves the sharpness of the rings a little and is essential for film thicknesses below 1500 A° to see the rings at all

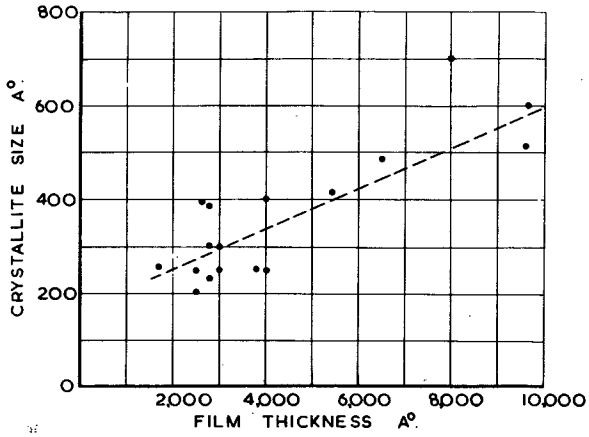


Figure 14 X-ray crystallite size v. thickness of NiFe film

No change in mean diameter of the rings is seen at different spots across the film surface, so no detectable variation in composition occurs. No perceptible change occurs on annealing. All the rings characteristic of a face centred cubic lattice are present with expected intensities and they are complete; therefore no orientation of the crystallites exists. In one particular case, after annealing, the (222) ring was missing, suggesting orientation of the (111) direction normal to the plane of the thin film. The other ring was complete, showing no orientation in the plane of the film.

Although iron and nickel have similar X-ray scattering coefficients, it is possible on bulk materials under extreme conditions to detect ordering of the atoms. This is not very possible on thin films, as the diffracted lines characteristic of order are too faint and are below the background level.

12. Electron Microscopy

Carbon replicas of the nickel iron films reveal a certain amount of structure reminiscent of boiling porridge (Fig. 15). What structure there is is mostly circular, about 500 A° or so across, and is randomly distributed. It is unlikely to cause anisotropy in the film.

Direct electron transmission of unannealed films does not allow many electrons through except near a few holes, and little detail can be seen. It is hoped to continue this work using annealed films. Electron diffraction of thin areas near holes

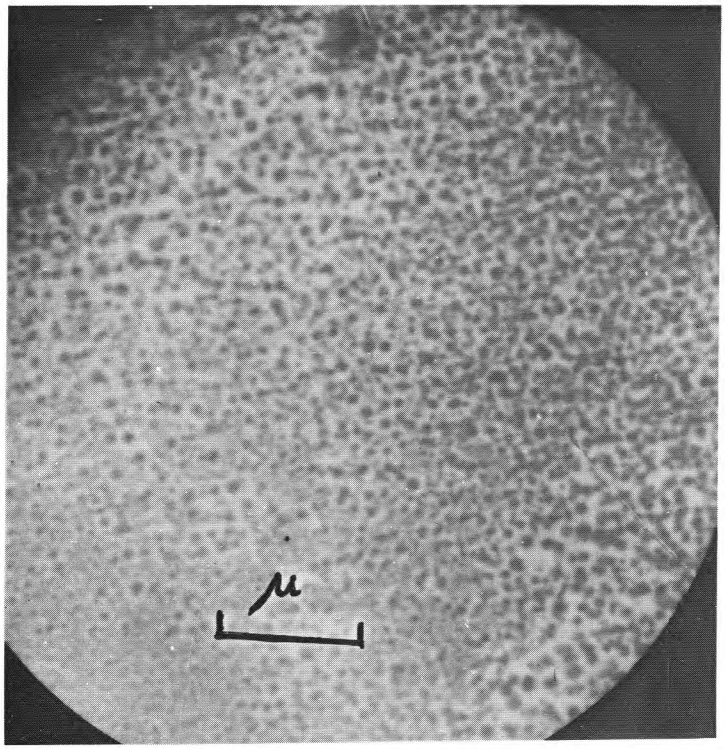


Figure 15 Electron micrograph of carbon replica of surface of NiFe film

showed complete rings characteristic of a face centred cubic system. There were no signs of orientation of crystallites. There was present, however, an additional diffracted ring probably caused by a further component in the film. It is thought that this is significant and warrants further investigation.

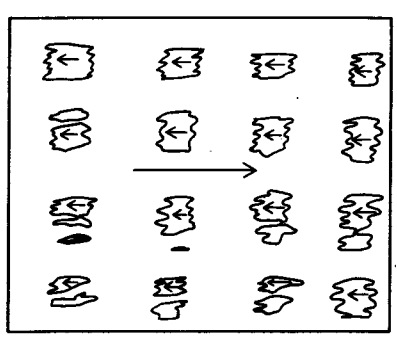
13. Thickness Measurements

The thickness of the films can be measured most easily by leaving one end of the substrate free of film and measuring the height of the step with a 'Talysurf'. In this machine a fine diamond stylus is drawn across the surface and its vertical motion is amplified electrically and fed onto a pen recorder. An accuracy of about 200 A° can be obtained. More accurate results can be obtained by silvering over the top of the step and forming multiple interference fringes between it and a half silvered optical flat with monochromatic light. The method is due to Tolansky and accuracy of about 100 A° can be achieved without taking any extreme precautions or using optically flat substrates.

Electrical resistance can be measured simply, and resistivity calculated if the thickness is known. Films below 1000 A^{o} thick have a higher resistivity than normal and this may be caused by the film consisting of islands not completely joined together. Above 1000 A^{o} the bulk resistivity is obtained, although some variation in resistivity occurs which is not entirely explained by variation in mean composition. This may be due to different degrees of ordering.

14. Magnetic Store

The final object of the investigation is to produce elements in a form suitable for use in a high speed computer store, and to this end a certain amount of work has been started in order to see what form the elements should take and how they would behave in non-uniform fields. The problem of interaction of the elements is also important when more than one is to be considered. Switching time measurements indicated that some adverse effects were arising with small specimens which did not occur with long specimens, whose centres only were switched. It was therefore thought desirable to investigate the possibility of using complete sheets of nickel iron and switching selected areas of them by conductors in planes parallel to the magnetic film plane, so arranged as to produce a field exceeding the coercive force at the particular points desired. To this end, the first line of attack was to determine by domain wall decoration methods how far the change in magnetization spread. Two insulated conducting strips were placed over a sheet of nickel iron film magnetized to saturation, and the strips were arranged to approach each other and the film at selected points. A DC current which would not produce any reversal of magnetization if passed singly through either of the conductors changed the magnetization when passed through both at the same time as shown in Fig. 16. From this evidence it would seem possible to space



DIRECTION OF MAGNETIZATION
INDICATED BY ARROWS.

Figure 16(a) Selected area switching of NiFe film

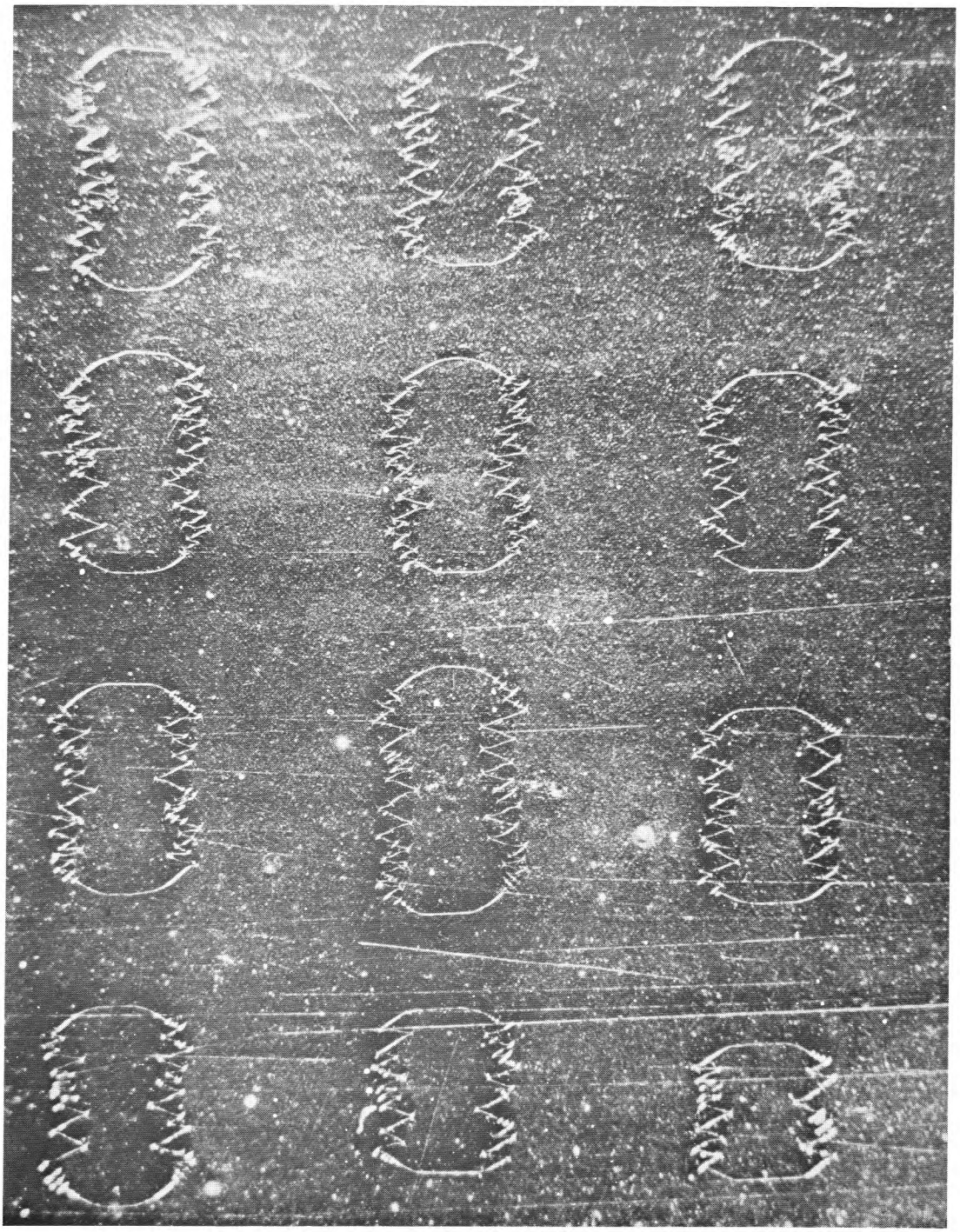


FIG. 16(b) SELECTED AREA SWITCHING OF Ni Fe FILM

SCALE:
1mm

individual elements in a store at a maximum of 0.5 cm apart and a complete 100 x 100 sheet in 50 cms square (maximum). The matrix of conductors would be arranged on a series of parallel sheets, and the packing in the direction at right angles to the sheets should be similar. This indicates that a megabit store could be compacted into an eighteen inch cube. There are troubles, however, when it comes to driving such a store at high speeds, and further experiments are being carried out using symmetrical strip transmission lines.

As a final item, mention should be made of digital storage with non-destructive read-out. This can be achieved by using two films A and B, one with a uniaxial anisotropy and a higher coercive force than the other. The higher coercivity film A, acts as the memory and if the lower coercivity film B is switched by a read pulse, then on removal of the pulse, the film A resets film B. It is therefore possible to determine the magnetic state of film A without switching it. On the other hand this state can be changed by applying a sufficiently large field to film A. Other techniques have been reported by other workers. In particular by applying a short enough pulse in the plane of the film at right angles to the preferred direction the magnetization can be changed elastically by enough to be detectable without any permanent change being caused. As the change is all rotation under maximum torque, it can take place very rapidly.

15. Acknowledgments

The work described in Part II has been conducted in collaboration with A. S. Young, who has been responsible for the evaporation of films and the investigation of their low frequency characteristics. The domain wall photographs also are the work of A. S. Young. To him and to R. J. Heritage, who has worked on the chemistry, acknowledgment is gladly made.

16. References

The following references are those which the author has had occasion to refer to more than once. A full list would be unnecessarily lengthy.

Books

R. M. Bozorth "Ferromagnetism" (Bell Lab. Series, Van Nostrand 1951) This is the current reference book on most ferromagnetic matters, though it contains little on the subject of ferrites or thin films. The bibliography includes nearly 2,000 references.

R. M. Stewart "Ferromagnetic Domains" (Cambridge, 1954) This contains more than its title implies and is concise.

Review Articles

C. Kittel and J. K. Galt "Ferromagnetic Domain Theory" (Solid State Physics Vol. 3, Academic Press, New York, 1956).

E. C. Stoner Rep. Prog. Phys. 13 p.83 (1950) and 11, p.43 (1948).

Journals

J. A. Osborn "Demagnetization factors of ellipsoids" Phys. Rev. 67 p.351 (1945).

M. S. Blois "Preparation and properties of thin films of NiFe J. App. Phys., 26 p.975 (1955).

J. App. Phys., 29 p.237 et seq. (March 1958) reports the proceedings of an A.I.E.E. conference on Magnetic Materials. An article by D. O. Smith "Static and Dynamic behaviour of Permalloy Films", p.264 is of particular relevance.