THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIV SEPTEMBER 1955 NUMBER 5

Alloyed Junction Avalanche Transistors

By S. L. MILLER and J. J. EBERS

(Manuscript received March 25, 1955)

A new device, the avalanche transistor, is described. Its properties derive from the utilization of the multiplication inherent in the breakdown process of reverse-biased semiconductor junctions. These junction transistors have regions of designable alpha greater than unity and are similar in many respects to point-contact and hook-collector transistors. They should, however, have advantages in speed and designability. Large regions of negative resistance of moderate magnitude can also be obtained in appropriate circuits. The device should find wide application in both switching and transmission. Design information for avalanche transistors is given.

1. INTRODUCTION

It has been shown recently that reverse biased silicon and germanium junctions break down as a result of a multiplicative process which is analogous to multiplicative breakdown in a gas.[1, 2, 3] Minority carriers which are thermally or otherwise generated diffuse to the high field region of the reverse biased junction, where they are accelerated, producing hole-electron pairs by collision with the atoms of the crystal lattice. As in the case of ionization of a gas, the rate of pair production is dependent upon the electric field distribution; however, it has been found that in the case of semiconductors holes and electrons are comparably effective in producing additional current carriers. The holes and

electrons which are produced by collision may themselves produce additional pairs, and so on, resulting in an avalanche. As the junction reverse bias is increased, the space charge layer widens, the maximum electric field increases, and the total rate of pair production increases. At a particular value of reverse voltage, called the breakdown voltage, the multiplication of the minority current carriers becomes essentially infinite. The current then increases very rapidly, being limited only by the external circuit resistance.

In semiconductors the number of minority carriers which are multiplied may be augmented and controlled by means of an emitter junction which is placed in close proximity to the reverse-biased, multiplying collector junction. The result, of course, is a simple n-p-n or p-n-p junction transistor which with proper design may exhibit considerable multiplication of the emitter current at voltages well below the breakdown voltage. The existence of this multiplication mechanism, which has been called avalanche multiplication, in reverse biased transistor collector junctions is of very great interest both from a transmission as well as from a switching point of view. So far as switching is concerned, it means that a simple junction device may have a current gain, or alpha, greater than unity and in appropriate circuits, be capable of exhibiting a negative resistance just as point-contact or hook-collector transistors do. Such a device may be used as the active element in pulse generator, regenerative pulse amplifier, or counter circuits. The effect of this multiplicative process on transmission applications is also important. Large regions of designable negative resistance of moderate magnitude may be attained in two terminal circuits and used to reduce losses in transmission systems. In general, the alteration in collector characteristics due to avalanche multiplication means that single stage amplifiers may be unstable and/or may exhibit a high degree of distortion depending on the operating conditions.

Further discussion of the negative resistance characteristic will be presented in a later section. Additional objectives of this paper are to indicate how an avalanche transistor can be characterized relative to its terminal behavior and to present design information which will enable the design of avalanche transistors for specific applications. The discussion will be in terms of alloyed transistors; however, the ideas presented are applicable to other structures.

It should be pointed out here that many junctions break down at or near the surface at a voltage considerably below that expected from the bulk properties of the junction. In these cases the multiplication of the bulk junction never rises very much above unity before "surface break-

down" occurs. Even though the surface breakdown also is frequently multiplying, only a small percentage of the emitter current is multiplied. The discussion in this paper is primarily applicable to transistors which exhibit "body breakdown."

2. THEORY

The multiplication of reverse biased step junctions, such as those of transistors made by the alloy-diffusion method, closely follows the empirical expression[3]

$$M = \frac{1}{1 - (V/V_B)^n} \tag{1}$$

where V_B is the junction body breakdown voltage and n is a parameter which varies with the resistivity and resistivity type of the material on the high resistivity side of the junction. In alloyed junction transistors this side is the base. Presumably the value of the parameter n varies also from semiconductor to semiconductor. For alloyed step junctions on p-type germanium, measured n values for different resistivities have ranged from 4.5 to 6.5. On n-type germanium n is approximately 3 throughout the investigated range. The breakdown voltage, V_B, rises monotonically with the resistivity of the base layer.

Equation (1) says that the alpha of a transistor has the form (neglecting space charge layer widening effects)

$$\alpha(V) = \alpha_0 M(V) = \frac{\alpha_0}{1 - (V/V_B)^n} \tag{2}$$

where α_0 is the value of the current gain at very low voltage, or, more accurately, the fraction of the emitter current which is collected, neglecting multiplication. This equation implies a designable alpha greater than unity for junction transistors. From equation (2) the voltage, V_S, at which alpha becomes unity is given by

$$V_S = V_B \sqrt[n]{1 - \alpha_0} \tag{3}$$

Thus V_S is completely determined by the body breakdown voltage of the collector junction, the low-voltage alpha and the value of n for the particular junction. With proper design V_S can be made only a small fraction of V_B. Obviously a device designed to take advantage of the multiplication effect should have as high an α_0 as possible and as low a value of the parameter n as is available. The advantage clearly lies with the p-n-p as opposed to the n-p-n transistor in germanium. Another

requirement is that the transistor technology be sufficiently advanced that the body breakdown of the collector junction can be observed. At present this means roughly that the breakdown voltage must be kept low, that is, below about 50 volts.

In linear applications which impose that α must remain below unity or that distortion be minimized, different requirements are necessary. For such applications high V_s is desired, which in turn calls for high base resistivity. Fulfilling this requirement, however, may necessitate a compromise, since high base resistivity results in limited frequency cut-off at low voltages because of space-charge layer punch-through.[4] Base resistance considerations may likewise put an upper limit on permissable base resistivity. In such a case the use of n-p-n transistors seems advisable,[5] and the operating voltage should remain always comfortably below V_s.

In the circuit configuration of Fig. 1 the avalanche transistor behaves like a gas discharge device in several respects. Its characteristic includes

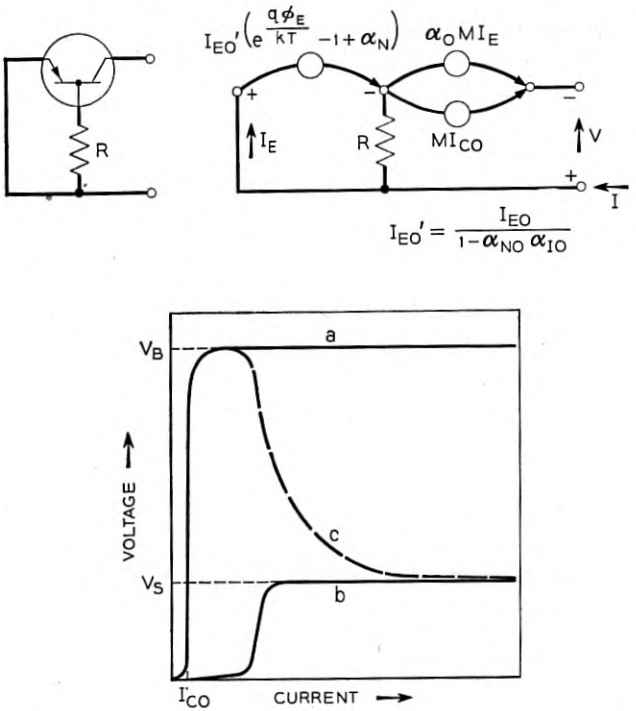


Fig. 1 — Avalanche transistor negative resistance circuit.

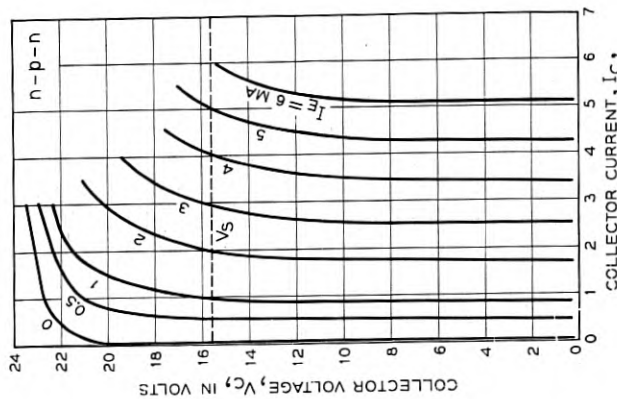
a high resistance region and a low resistance region separated by a region of negative resistance. The origins of the negative resistance in the two cases, however, exhibit distinctions which should be pointed out. In the gas discharge, negative resistance results after breakdown primarily from the increased cathode efficiency due to changes in the field configuration because of the space charge effects of the carriers. In the solid state analogue, the charges already resident in the discharge region in the form of the ionized chemical impurities in the lattice outnumber the carriers traversing the high field region by many orders of magnitude at reasonable current densities. Hence the field configuration is not significantly current-dependent. Furthermore, there is no cathodic regenerative mechanism in the solid discharge (gamma process), and photoregeneration in germanium is only $\sim 10^{-4}$ efficient. Townsends' β mechanism has been demonstrated to be the dominant agency in maintaining the discharge.[2, 3] Therefore, the avalanche breakdown in the solid generally does not inherently lead to negative resistances. There would, however, be the possibility of negative resistance at very high current densities.

The negative resistances observable are the result of the fact that with avalanche transistors it is possible to look across the discharge in two different ways. One is directly across the avalanche multiplication region alone (from the base to the collector of a transistor) and the other is across a source of minority carriers and the multiplication region in series (from emitter to collector with base floating). In the former case base-to-collector breakdown leading to near zero incremental resistance occurs at that voltage at which the avalanche multiplication becomes infinite. In the latter case, because of continuity of current requirements, emitter-to-collector breakdown occurs at V_S, that voltage at which the multiplication times the α_0 of the transistor becomes unity. Any circuit configuration which in effect goes from the former to the latter condition with increasing current will exhibit negative resistance.

Curve (a) in Fig. 1 is the normal reverse characteristic of the collector junction, as measured collector to base with emitter open, indicating the breakdown voltage, V_B. For a theoretical transistor this curve is determined by the equation

$$I = MI_{CO} = \frac{I_{CO}}{1 - (V/V_B)^n} \tag{4}$$

Since the emitter is open circuited it does not contribute any minority carriers to the discharge. Curve (b) in Fig. 1 is the reverse characteristic of the collector junction (as measured collector to emitter with base



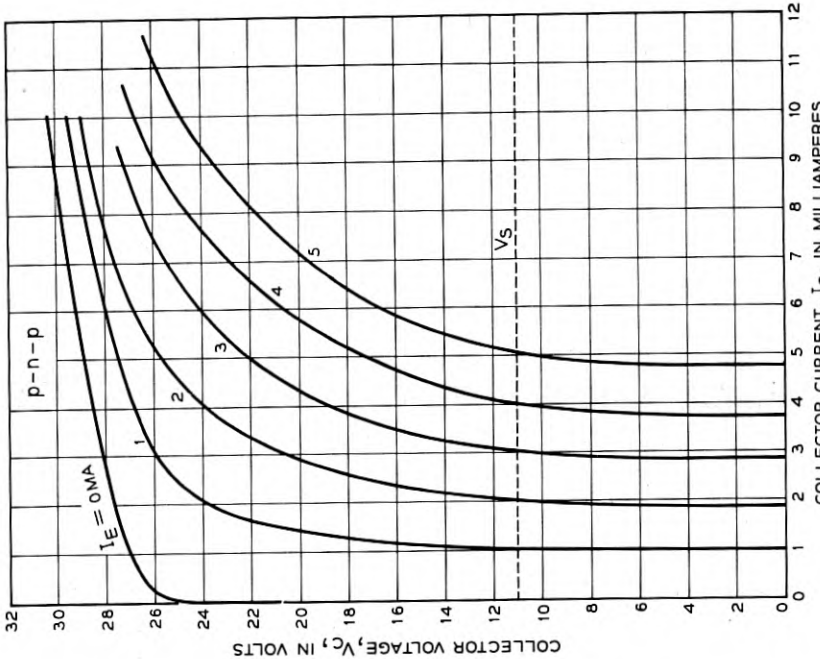


Fig. 2 — Collector characteristics of alloyed junction transistors exhibiting avalanche multiplication. (a) n-p-n and (b) p-n-p.

open). According to the equivalent circuit of the transistor the current is given by the equation

$$I_C = \frac{MI_{CO}}{1 - \alpha_0 M} \tag{5}$$

and it is seen that current increases, subject only to limitation by circuit resistance, when the collector voltage reaches V_S, the voltage corresponding to unity total alpha or $\alpha_0 M = 1$. Since the emitter current of alloyed junction transistors is primarily made up of minority carriers emitted into the base region, the emitter efficiency in this case can be said to be high. It is apparent that if the emitter efficiency could be made to vary with current, then it would be possible to obtain a negative emitter-to-collector resistance. This is the purpose of the base resistor shown in Fig. 1. For low currents the impedance of the emitter junction is high, since the voltage-current characteristic is exponential in character, and most of the current flows through the base resistor. As the voltage is increased essentially curve (a) is traced out. Near the breakdown voltage the current increases, the emitter-to-base voltage increases in the forward direction, and the impedance of the emitter junction becomes smaller in comparison with the base resistance. This process results in increased emitter current. Thus the characteristic, curve (c), begins to depart from curve (a), and as a larger and larger fraction of the total current is transferred from the base circuit to the emitter circuit, curve (c) approaches curve (b) asymptotically.

3. CHARACTERIZATION

In view of the above description of the multiplication properties of transistor junctions, it is clear that the collector characteristics of all transistors, in which a surface breakdown does not intervene, will look like those shown in Figs. 2(a) and 2(b). In Fig. 2(a), the measured collector characteristics for a representative n-p-n transistor are shown. Fig. 2(b) gives the same information for a representative p-n-p unit. The horizontal dashed line in each case indicates the voltage V_S at which α becomes unity. This voltage is of course a function of the particular value of α_0 for each transistor. Either of these transistor types would have somewhat more distortion as a common base amplifier than has hitherto been expected from transistor theory neglecting avalanche effects.

As has already been emphasized, the p-n-p has a decided advantage for monostable, astable, or bistable switching applications. A suitable characterization for an avalanche transistor designed for switching use

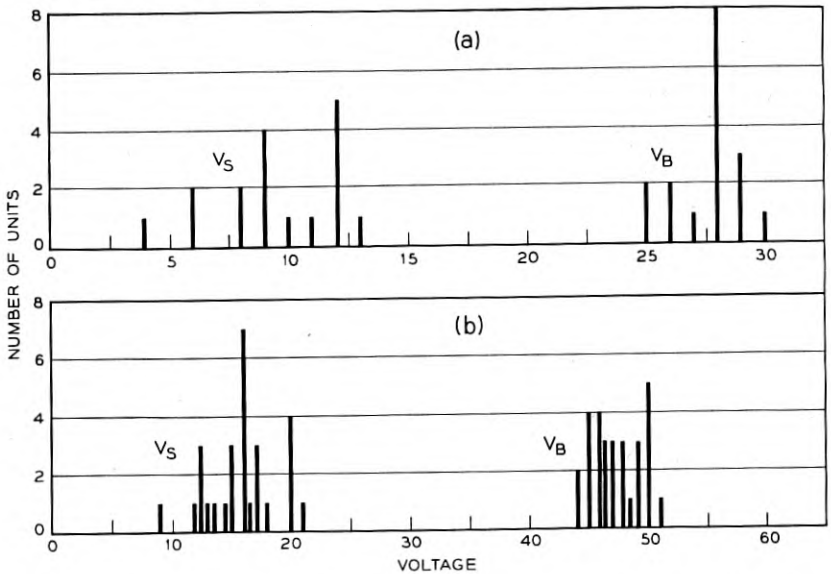


Fig. 3 — Distributions of V_B and V_S for avalanche transistors. (a) Base resistivity = 0.2 Ω-cm; (b) Base resistivity = 0.5 Ω-cm.

would certainly involve the body breakdown voltage of the collector, V_B, and the voltage at which the total α of the transistor is unity, V_S, which can be thought of as a sustain voltage. The breakdown voltage is a function of the base layer resistivity, while the ratio of the sustain voltage to the breakdown voltage is a function of the low-voltage alpha of the transistor. Therefore, the ability to control these parameters involves only the ability to control base layer resistivity and α_0. These are problems which have already claimed considerable attention with good results in other transistor designs.

Groups of such avalanche transistors have been made at Bell Telephone Laboratories. Figs. 3(a) and 3(b) give distributions of V_B and V_S for groups of about 25 transistors made on 0.2 Ω-cm and 0.5 Ω-cm n-type material. In Figs. 4(a) and 4(b) are given the relations between V_S and α_0 for the two groups.

The spreads in V_B for both groups of transistors largely reflect slight deviations from the nominal resistivity. Although the germanium in both cases was zone levelled, it is estimated that there is a ± 5 per cent variation in resistivity for individual wafers. Undoubtedly some variation is due to surface breakdowns very close to but below the body breakdown. The spread in V_S is, of course, partially the result of the

spread in V_B insofar as the variation in V_B is the result of variation in the bulk material. The remainder stems from the distribution of α_0 values. No vigorous attempt was made to control α_0 carefully or to hold rigid limits on base resistivity for these sample groups.

From (3) it can be seen that log V_S/V_B when plotted versus log $(1 - \alpha_0)$ for transistors of different α_0 should yield a straight line of slope $1/n$. Solid lines of slope $\frac{1}{3}$ have been drawn in Fig. 4 along with dashed lines of slope $\frac{1}{6}$. The adherence to the $n = 3$ law is confirmed.

Equation (2) gives the total α of a transistor when multiplication is taken into account. Values calculated from this equation are plotted in Fig. 5 in comparison with values determined by ac α measurements and with α values obtained from the measured static collector characteristics for a representative unit from the group having base layer resistivity

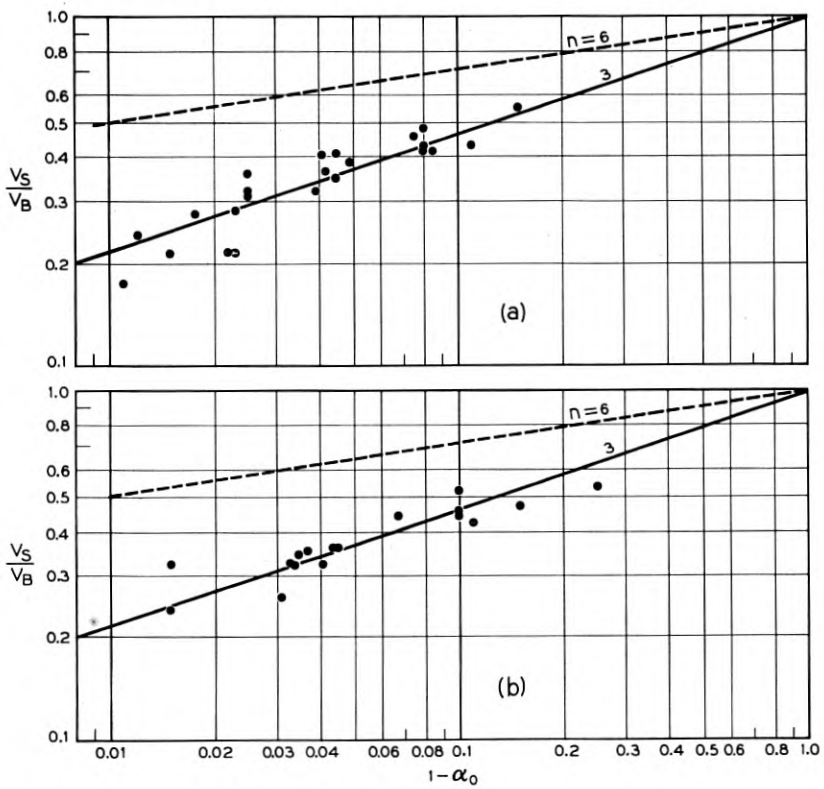


Fig. 4 — Plot of V_S/V_B vs $1 - \alpha_0$. (a) Base resistivity = 0.2 Ω-cm. (b) Base resistivity = 0.5 Ω-cm.

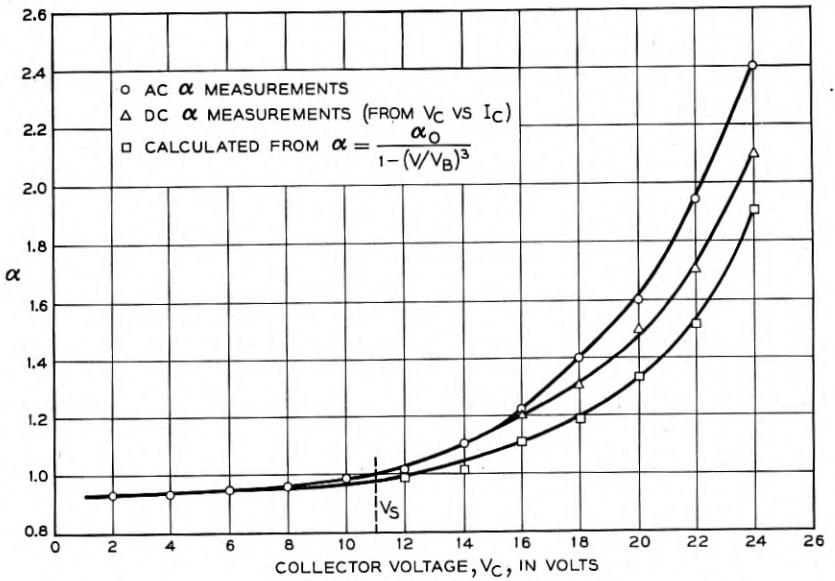


Fig. 5 — Alpha versus collector voltage for an avalanche transistor.

of 0.2 Ω-cm. The agreement is good. It is quite general for the formula to give a slightly lower value of α than is observed in this type of experiment. The functional form of the multiplication and especially the value of the exponent were determined from an experiment which minimized effects on α_0 of changes in the base current flow pattern.[3] The experimental α as determined above would be expected to be higher than theoretical because when α goes above unity, the majority carrier flow from the collector to the base biases the emitter so that α_0 increases with increasing total α. This effect is the opposite of the effect in which α_0 decreases with increasing large emitter current in ordinary junction transistor operation with α less than unity.

The frequency characteristic of the multiplicative increase in α is of great interest. Theoretically the multiplication process should be extremely fast. The response time of the multiplication should be comparable with the transit time for carriers across the high field region in the reverse biased junction. This time would be a small fraction of a millimicrosecond in all practical junctions. Therefore, the frequency cutoff of the total alpha should not be measurably smaller than the frequency cutoff of the α_0. Actually the effect discussed above, which increases α_0 with increasing total α (when above unity), and therefore, with

increasing collector voltage, could have an even greater effect on the frequency cutoff of α_0. Again it would be expected that the change would be in the direction of increasing f_N. Fig. 6 gives measured frequency cutoffs of total α vs. the collector voltage for the same representative unit which was used for Fig. 5. The theoretical curve was calculated[5] from

$$f_N = \frac{f_{NO}}{\left[1 - \left(\frac{V_c f_{NO}}{3.65 \times 10^{-13} N_I A}\right)^{1/2}\right]^2} \tag{6}$$

This equation merely takes into account the variation of base width due to widening of the collector space-charge layer with increasing voltage.[4, 6] There is a sharp departure upwards from the theoretical curve for the points corresponding to total α greater than one. This behavior is in line with the above explanation. Again, the reverse effect is observed with increasing large emitter currents in junction transistors operated in the ordinary manner. Thus the frequency characteristics of the multiplied alpha are as good as, and can be better than, the frequency char-

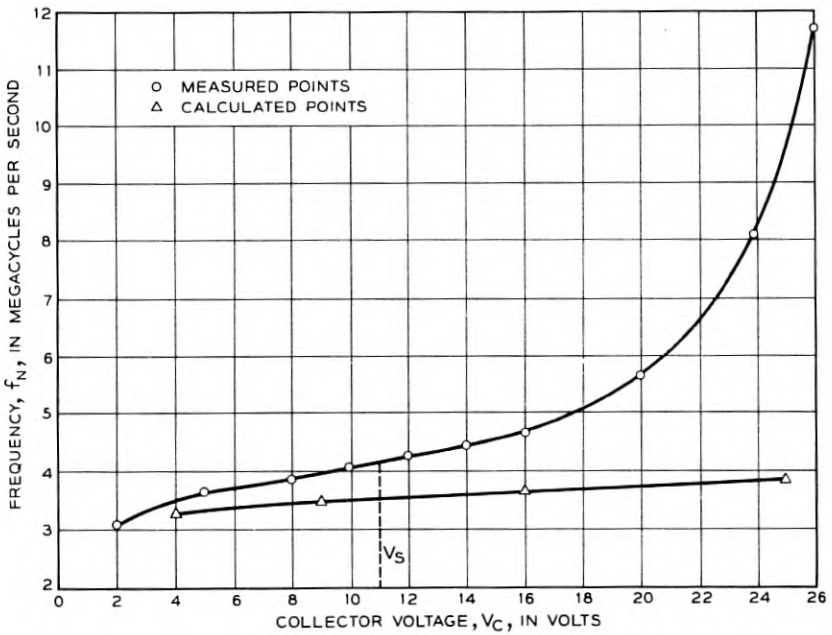


Fig. 6 — Frequency cut-off of alpha versus collector voltage for an avalanche transistor.

acteristics of α_0. This means that any device which utilizes the negative resistance inherent in the α greater than unity will be limited in frequency only by the dispersion in transit time across the base.

The temperature variation of both V_B and V_S is comparatively small. The breakdown voltage of np junctions increases slowly with increasing temperature at the rate of about 0.1 per cent per degree centigrade.[7] Aside from the change in V_B, V_S will be affected also by changes in $1 - \alpha_0$ with temperature. In the case of p-n-p germanium transistors a drastic change of $1 - \alpha_0$ by a factor of two with temperature would give only a 25 per cent change in V_S. Data given by Ebers and Miller[5] show that such changes would require very large temperature changes. However, it should not be supposed that devices based on the avalanche principle would always be very temperature insensitive. For example, in the circuit shown in Fig. 1, the mechanism which switches from the high breakdown voltage to the low breakdown voltage path may be highly temperature sensitive, unless remedial steps are taken, because the saturation current is intimately involved.

4. DESIGN DATA

The principal designable features of an avalanche transistor are the V_B and V_S values. From the discussion in the next section, it will be

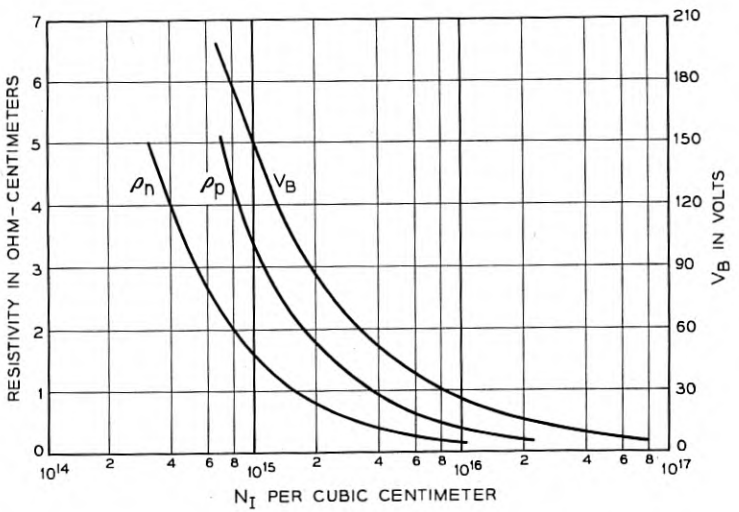


Fig. 7 — Resistivity and breakdown voltage versus net density of impurity centers.

TABLE I

n	High Resistivity Side of Step Junction
4.7	0.15 Ω-cm, p type
5.5	0.25 Ω-cm, p type
6.6	0.5 Ω-cm, p type
~6	2.0 Ω-cm, p type
3	0.1 Ω-cm, n type
3.4	0.6 Ω-cm, n type
3	2.0 Ω-cm, n type

seen that some of the considerations in ordinary transistor design, like keeping the base resistance low, sometimes are no longer important. The control of V_B, as pointed out above, is simply a matter of controlling the base layer resistivity. Miller[3] has given the breakdown voltage as a function of the net density of impurity centers on the high resistivity side of a germanium step junction like those made by the alloy process. These data are given in Fig. 7. This illustration also contains plots of the resistivity of the high resistivity side vs. the net density of impurity centers for n and p germanium material. These latter curves were redrawn from data given by M. B. Prince.[8] They represent the best available information on this subject at this time.

V_S is a function of V_B, $1 - \alpha_0$, and the pertinent parameter n. n is not completely independent of V_B since for a given type of transistor, n-p-n or p-n-p, both n and V_B are functions of the base layer resistivity. The experimental values of n for various base layer resistivities and resistivity types in germanium have been determined by Miller[3] and are given in Table I.

It is important that the space charge region of the collector not punch through the base layer below the breakdown voltage if the full interval between V_B and V_S is to be utilized. This requirement effectively puts an upper limit on the frequency cutoff for each base resistivity value, since the base width must be at least as wide as the space-charge region at the breakdown voltage. Fig. 8 gives the maximum frequency cutoff vs. the resistivity of the base region for p-n-p germanium avalanche transistors.

It is conceivable that if base layer width could be controlled very accurately, this punch-through phenomenon could be used to determine the peak of the negative in the resistance curve shown in Fig. 1 instead of relying on uniformly well etched junctions to show exactly the same breakdown voltage.

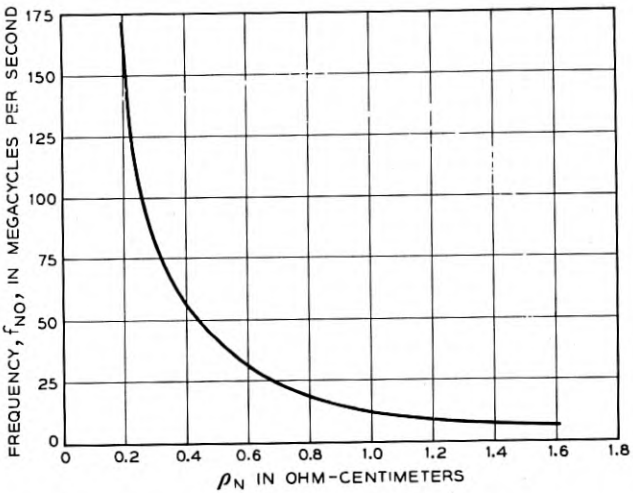


Fig. 8 — Maximum low-voltage frequency cut-off of alpha versus resistivity of base region for alloyed junction pnp transistors.

5. APPLICATIONS

In the region of collector voltage for which the alpha of an avalanche transistor is greater than unity, its terminal characteristics are quite similar to those of a point-contact transistor or a hook-collector transistor.[9, 10, 11] Hence most of the switching circuits designed for point-contact transistors could also use avalanche transistors. To provide a large voltage swing in such circuits the breakdown voltage of the collector junction should correspond to body breakdown and α_0 should be as large as possible in order to give a low value of the sustaining voltage, V_S. This latter requirement also leads to a low power dissipation when the transistor switch is in the ON (closed) condition.

One of the most serious limitations on the usefulness of hook-collector transistors is switching speed. Since the multiplication phenomenon involves the emission of minority carriers by the hook junction and the transport of these carriers across the hook region, a dispersion in transit time exists which results in a low alpha cutoff frequency for the device. In fact, the effective cutoff frequency may be an order of magnitude less than the cutoff frequency which would be calculated on the basis of the transport of the minority carriers across the base layer. As has already been pointed out, avalanche transistors exhibit alpha cutoff frequencies indistinguishable from those of non-multiplying devices. Frequency cutoffs of alpha in the range of 5 to 10 mc/sec are easily obtainable, even in pnp avalanche transistors, since the resistivity of the base layer can

be quite low and hence punch through is inhibited for the necessarily thin base layers. Since the initial rate of change of current or voltage in pulse circuits is proportional to both alpha and alpha cutoff frequency, the rise times obtainable with avalanche transistors easily fall in the tenth microsecond range.

Another limitation on the switching speed of point-contact or hook-collector transistors results from the storage of minority carriers. This phenomenon manifests itself when one attempts to open the transistor switch after it has been driven into the current saturation region. In bistable type circuits using avalanche transistors the minimum collector voltage is automatically limited to the sustaining voltage, V_S, and the transistor never goes into the current saturation region. The collector junction never becomes forward biased, and there is none of the usual storage of minority carriers.[12]

One of the most interesting applications of avalanche transistors is as a two terminal, negative impedance element. Such a circuit has many applications in both switching and transmission. The circuit analysis of this application will be presented since it exemplifies the usefulness of the large-signal junction transistor theory and since the negative resistance characteristic that is obtained is typical of that for avalanche transistors.

The circuit to be discussed is shown in Fig. 1. An equivalent circuit, based on the work of Ebers and Moll,[13] is also shown in Fig. 1. The quantities α_{NO} and α_{IO} are the low voltage normal and inverse alphas. The voltage Φ_E is the voltage across the emitter junction. It is assumed that the base resistance is included in R. It is apparent that

$$I = \alpha_{NO} M I_E + M I_{CO} \tag{7}$$

If (1) and (7) are combined the following equation can be obtained

$$V = V_B \sqrt[n]{1 - \frac{I_{CO} + \alpha_{NO} I_E}{I}} \tag{8}$$

if it is assumed that the applied voltage, V, is large compared to Φ_E. In order to plot a voltage-current characteristic it is necessary to have a relation between I_E and I. From the equivalent circuit

$$I = \frac{\Phi_E}{R} + I_E \tag{9}$$

Since

$$\Phi_E = \frac{kT}{q} \ln \frac{I_E + (1 - \alpha_{NO}) I'_{EO}}{I'_{EO}} \tag{10}$$

where the quantity I'_{EO} is the reverse current of the emitter junction with the collector shorted to the base, it follows that

$$I = I_E + \frac{kT}{qR} \ln \frac{I_E + (1 - \alpha_{NO})I'_{EO}}{I'_{EO}} \tag{11}$$

(8) and (11) completely determine the negative resistance characteristic of the circuit of Fig. 1. A measured curve for a particular transistor in this circuit is shown in Fig. 9.

In some applications the slope of the negative resistance characteristic is of primary importance. It can be shown that

$$\frac{dV}{dI} = -\frac{V}{nI} \frac{1 - \alpha_{NO} M \dfrac{dI_E}{dI}}{1 - M} \tag{12}$$

or

$$\frac{dV}{dI} = -\frac{V}{nI}\left[1 - \left(1 - \alpha_{NO} \frac{dI_E}{dI}\right)\left(\frac{V_B}{V}\right)^n\right] \tag{13}$$

For values of R in the vicinity of 5,000 ohms and I_E greater than a milliampere,

$$\frac{dI_E}{dI} = \frac{1}{1 + \dfrac{kT}{qR}\dfrac{1}{I_E}} \simeq 1 \tag{14}$$

and

$$\frac{dV}{dI} = -\frac{V}{nI}\left[1 - (1 - \alpha_{NO})\left(\frac{V_B}{V}\right)^n\right] = -\frac{V}{nI}\left[1 - \left(\frac{V_S}{V}\right)^n\right] \tag{15}$$

This relation, in itself, is of very little utility since V and I are related in a complicated manner as can be seen by examining (8) and (11). However, if a plot of the negative resistance characteristic is available, this equation provides a theoretical check on the negative resistance at a given operating point. For example, consider the negative resistance characteristic shown in Fig. 9. The transistor used in obtaining these data had the following characteristics: $V_B = 36$ and $(1 - \alpha_0) = 0.047$. From the curve, for an operating point of 16 volts and 6 milliamperes, the measured negative resistance is 450 ohms. Substitution of appropriate values in equation (15) yields a negative resistance of 400 ohms. In every case that has been investigated a similar discrepancy has been found to exist. The higher value of negative resistance is attributed to an increase in α_0 due to the focusing effect of the voltage drop in the base

layer. Actually what happens is that the center of the emitter, under these conditions of operation, becomes more forward biased than the outer edge, thus improving the transport efficiency since there is proportionately less surface recombination.

In some applications it would be very convenient if the type of negative resistance characteristic described above could be obtained without resorting to external circuit elements; in other words, if a true two terminal negative resistance could be obtained. As was explained above, the purpose of the external resistance in Fig. 1 is effectively to switch the current from the base collector loop to the emitter collector loop. Another way of interpreting the effect of the resistor is that it changes the effective alpha of the transistor by causing a higher percentage of the total current to be injected minority carrier current. A p-n-p (or n-p-n) structure designed in such a way that the alpha increases with emitter current would yield the same result. Such a device is shown in Fig. 10 along with the negative resistance characteristic obtained between emitter and collector. It is observed that the structure is not significantly different from a nonsymmetrical alloyed junction transistor in which the roles of the emitter and collector have been interchanged. For low values of current the alpha of the transistor is low and the peak voltage attained may be as much as 95 per cent of the breakdown voltage. As the current increases, a voltage drop occurs in the base layer. This drop biases

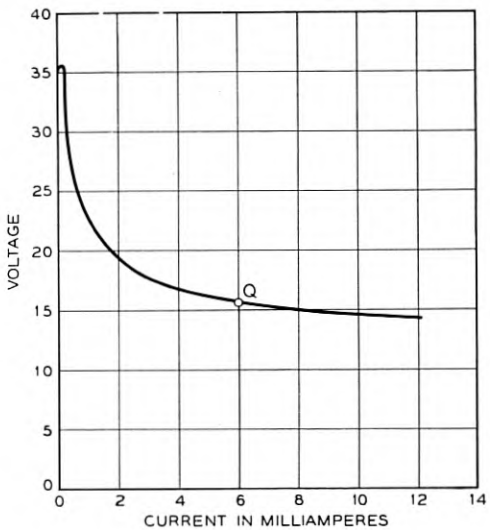


Fig. 9 — Measured avalanche transistor negative resistance characteristic.

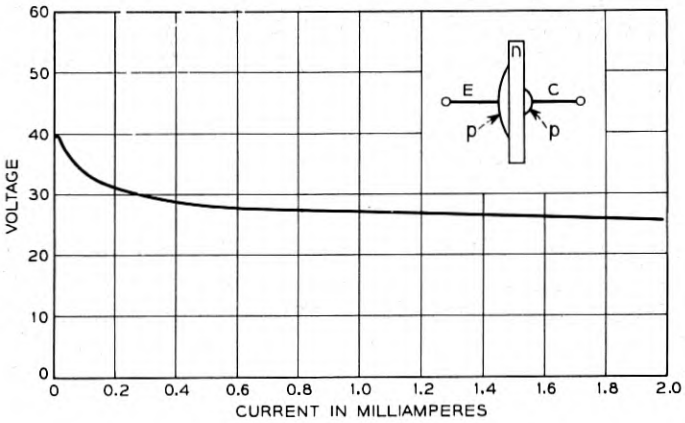


Fig. 10. — Two terminal avalanche device characteristic.

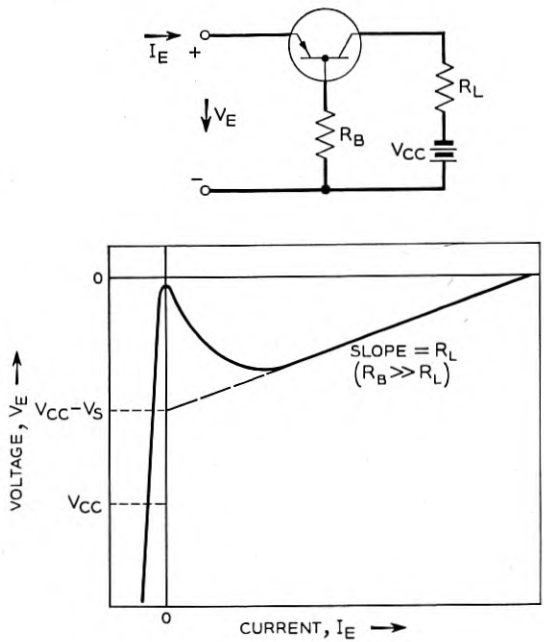


Fig. 11 — Emitter negative resistance characteristic of the avalanche transistor.

the outside of the emitter less forward than the center, and the alpha increases, resulting in a negative resistance.

It is worthwhile, in passing, to make a few statements concerning the type of negative resistance obtained in the circuit of Fig. 11. Such circuits, using point-contact transistors, have been amply discussed by Anderson.[9] Therefore this circuit will be discussed with the objective of pointing out the differences in the behavior when avalanche transistors are used. The general features of the negative resistance characteristic are also shown in Fig. 11. First of all the peak point is depressed below the origin an amount equal to the emitter floating potential plus the $MI_{co}R_B$ drop in the external base resistance (neglecting internal base resistance). The initial slope of the negative resistance characteristic is governed by the collector supply voltage and how nearly it approaches the breakdown voltage, V_B. Since the magnitude of the negative resistance characteristic is given approximately by $(1 - \alpha)R_B$, it is apparent that the negative resistance approaches zero as the collector voltage approaches the sustaining voltage, V_S.

6. CONCLUSIONS

A new device has been described which is similar in some respects to point-contact transistors and hook-collector transistors. Avalanche transistors should prove to be very useful in both switching and transmission applications.

As has been shown, the behavior of avalanche transistors is quite well understood. In addition the terminal characteristics are sufficiently well related to the structure of the device to enable the design of devices to meet the needs of specific applications. It is believed that the presence of avalanche multiplication in transistor junctions may open up a whole new class of devices which can perform circuit functions not previously feasible with any single device.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the benefit of many discussions with J. L. Moll. J. J. Kleimack fabricated the first devices on which body breakdown could consistently be observed. Many of the data were gathered by W. C. Meyer.

Note: Approximately coincident with the submission of this article a paper covering related material appeared in the March, 1955, R. C. A. Review. It is entitled "Delayed Collector Conduction, A New Effect in Junction Transistors" by M. C. Kidd, W. Hasenberg, and W. M. Webster. As the divergence in titles indicates, the subject is treated from a somewhat different viewpoint in the two papers.

REFERENCES

1. McKay, K. G. and McAfee, K. B., Electron Multiplication in Silicon and Germanium, Phys. Rev., **91,** pp. 1079–1084, Sept. 1, 1953.
2. McKay, K. G., Avalanche Breakdown in Silicon, Phys. Rev., **94,** pp. 877–884, May 15, 1954.
3. Miller, S. L., Avalanche Breakdown in Germanium, Phys. Rev. (in publication).
4. Shockley, W., Transistor Electronics: Imperfections, Unipolar and Analogue Transistors, Proc. I.R.E., **40,** pp. 1289–1313, Nov., 1952.
5. Ebers, J. J. and Miller, S. L., Design of Alloyed Junction Germanium Transistors for High Speed Switching, B.S.T.J. **34,** pp. 761–782, July, 1955.
6. Early, J. M., Effects of Space-Charge Layer Widening in Junction Transistors, Proc. I.R.E., **40,** pp. 1401–1406, Nov., 1952.
7. Pearson, G. L. and Sawyer, B., Silicon p-n Junction Alloy Diodes, Proc. I.R.E., **40,** pp. 1348–1351, Nov., 1952.
8. Prince, M. B., Drift Mobilities in Semiconductors, I. Germanium, Phys. Rev., **92,** pp. 681–687, Nov. 1, 1953.
9. Anderson, A. E., Transistors in Switching Circuits, Proc. I.R.E., **40,** pp. 1541–1558, Nov., 1952.
10. Shockley, W., Sparks, M., and Teal, G. K., p-n Junction Transistors, Phys. Rev., **83,** pp. 151–162, July 1, 1951.
11. Ebers, J. J., Four-Terminal p-n-p-n Transistors, Proc. I.R.E., **40,** pp. 1361–1364, Nov., 1952.
12. Moll, J. L., Large-Signal Transient Response of Junction Transistors, Proc. I.R.E., **42,** pp. 1773–1784, Dec., 1954.
13. Ebers, J. J. and Moll, J. L., Large-Signal Behavior of Junction Transistors, Proc. I.R.E., **42,** pp. 1761–1772; Dec., 1954.

Effect of Dislocations on Ultrasonic Wave Attenuation in Metals*

By W. P. MASON

(Manuscript received April 6, 1955)

The causes of energy dissipation and mechanical instabilities of the elastic constants in metals can usually be traced to the presence of an imperfection in the crystal lattice called a dislocation. Edge dislocations are regions in the lattice where an extra plane of atoms has been added or subtracted from an otherwise perfect crystal. Such dislocations can move through the crystal under the application of shearing stresses or because of thermal agitation. It is shown that the primary causes of energy dissipation in a metal are dislocation loops pinned at irregular intervals by impurity atoms. At very low temperatures these dislocations lie along minimum energy positions but at higher temperatures they can be displaced to the next minimum energy position. In going to the next position, the dislocation meets an energy barrier determined by the energy required to overcome the limiting shearing stress T_{13_0} and the energy to stretch the dislocation. This barrier causes a relaxation effect for which the dislocations lag behind the applied stress and abstract energy from the mechanical vibrations. By measuring the position and height of the relaxation peak as a function of frequency and temperature, evidence is obtained for the value of the limiting shearing stress, the number of dislocations per square cm., and the average loop length. The values obtained agree with other methods for measuring these quantities.

At higher temperatures thermal agitation causes the loops to break away from their pinning impurity atoms. In the process, it is shown that a loss occurs which is independent of frequency and amplitude but which varies exponentially with the temperature. The activation energy found agrees with the calculated value for the binding energy of an impurity atom. Dislocations also occur at the boundaries between grains in the metal and produce a peak in the measured attenuation of a polycrystal which reaches a maximum at high temperatures and low frequencies. The activation energy for this process is determined by the energy required for a vacancy to diffuse

* *Editor's Note:* The present paper is a chapter of a new book entitled "Ultrasonics in Solids", which is scheduled for publication by D. Van Nostrand in 1956.

from one position to another. This energy is somewhat less than the bulk diffusion energy on account of the strains in the grain boundary.

Highly worked metals show two other effects due to dislocations, called the Köster effect and the viscosity effect. It appears that these are due to zig zag dislocations which do not lie in minimum energy positions. In the course of time free dislocations become pinned and the Köster effect disappears.

INTRODUCTION

Metals are used very considerably in conducting sound waves in such applications as mechanical filters, reed relays, low frequency delay lines, and ultrasonic processing devices. In all of these devices, the energy loss and the stability of the elastic properties of the metals are of prime importance. The causes of energy losses and instabilities in all solid materials are associated with the molecular motions that can take place under thermal agitation and under the stresses that are applied to the materials. For metals, most of the irreversible processes have been interpreted in terms of a type of imperfection known as a dislocation. While it is not within the scope of this book to discuss dislocation theory in detail,[1] a short introduction is given in order to provide a background for the ultrasonic measurements discussed in this chapter.

Dislocations were first introduced into the theory of metals to explain the fact that the limiting shear stress required to cause plastic flow in very pure single crystals was such a small fraction of the shear elastic constant of the crystal. Various measurements for very pure metals have shown that the limiting shear stress may be only 1/60,000 times the elastic shear modulus μ at room temperature and not more than three times this value at absolute zero. Without invoking imperfections of some type it is very difficult[2] to explain why the limiting shearing stress

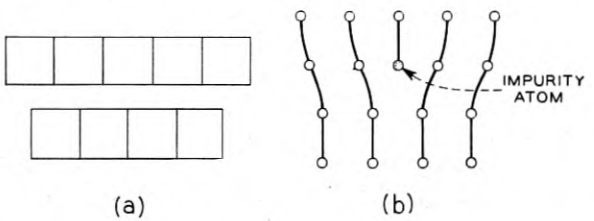


Fig. 1 — Edge dislocation showing position for small impurity atom.

[1] Complete discussions of dislocation theory have been given in two recent books, A. H. Cottrell, Dislocations and Plastic Flow in Crystals, Oxford University Press, 1953, and W. T. Read, Dislocations in Crystals, McGraw-Hill Company, 1953.

[2] See A. H. Cottrell, loc. cit., pp. 8–12.

should be less than $1/30$ of μ. A type of imperfection known as an edge dislocation was first introduced by G. I. Taylor and E. Orowan to explain this effect. An edge dislocation, as shown by Fig. 1, is a region in a crystal lattice where an extra row of atoms is either introduced or abstracted from an otherwise perfect crystal. This can be accommodated only if the crystal is severely strained in the joining region, with a high compression in the region of too many planes and a high tension in the region of too few planes.

Experiment has shown that these dislocations are mobile and move in close packed atomic planes since the energy they have to overcome in these planes is less than in other planes. For face centered metals such as aluminum, lead, copper, and silver this plane is the (111) plane which is the plane perpendicular to the cube diagonal as shown by Fig. 2. The direction of motion is in the $10\bar{1}$ direction which is the direction for which successive atoms are closest together. As can be seen from Fig. 2, the distance b that they are separated is $1/\sqrt{2}$ times the cube edge for the unit cell. Body-centered crystals glide along the (110) plane in the [111] direction. Table I shows[3] the number of glide planes, glide directions and atomic spacings for several types of crystal structures.

Returning to the problem of the limiting shearing stress, it is obvious

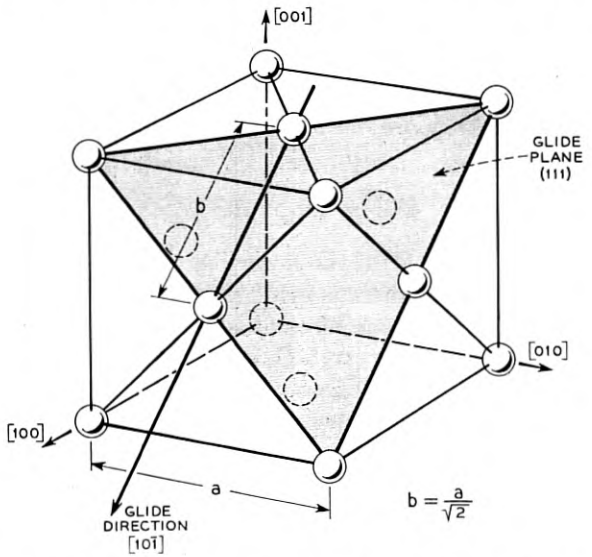


Fig. 2 — Glide plane, glide direction and glide distance b for a face centered metal.

[3] See W. T. Read, loc. cit., p. 22.

TABLE I

Structure	Glide Plane	Glide Direction	Magnitude in terms of cube edge	Number of glide directions
Simple cubic	010	100	1	3
Face-centered cubic	111	110	$\frac{1}{\sqrt{2}}$	6
Body-centered cubic	110	111	$\frac{1}{2}\sqrt{3}$	4

that with this type of imperfection it will require less force to move a plane of atoms to the next minimum energy position one molecular distance b from the first minimum energy position. This follows from the fact that the large force required to push one atom by another one is partly cancelled by the attraction of the next atom for its opposite number on the other side of the dislocation. This force depends on the width of the dislocation, i.e., the number of atom planes over which the dislocation is spread. Various assumptions are considered by Cottrell[4] who finds that the limiting shearing stress should be in the range

$$T_{13_0} = 4.0 \times 10^{-6}\,\mu \text{ to } 3.6 \times 10^{-4}\,\mu \tag{1}$$

As discussed in the next section, an ultrasonic relaxation at low temperatures has been found which correlates with the energy required to move a dislocation from one minimum energy position to the next one at a distance b from the first against the limiting shear stress T_{13_0} and it is found that

$$T_{13_0} \doteq 6.0 \times 10^{-6}\,\mu, \tag{2}$$

in satisfactory agreement with the lower limit of (1).

In a pure single crystal there is evidence that the dislocations form a network which outlines mosaic blocks having slightly different orientations from each other in the crystal. Evidence for such blocks is obtained from the width of X-ray reflections from a metal[5] which can be interpreted to indicate that the number of dislocation lines is of the order of

$$N_0\ell = 10^8 \text{ dislocation per sq cm} \tag{3}$$

Here N_0 is the total number of dislocation loops per cubic centimeter and ℓ is the average loop length as shown in the model of Fig. 4. If these dis-

[4] A. H. Cottrell, loc. cit., p. 62–64.
[5] A. H. Cottrell, loc. cit., Chapter IV, pp. 99–102.

locations form the edges of a network we must have

$$N_0 \ell^3 \doteq 1 \tag{4}$$

and hence equations (3) and (4) would indicate

$$N_0 \doteq 10^{12}; \qquad \ell = 10^{-4} \text{ cm} \tag{5}$$

Data from the etching of crystals[6] for which pits are delineated at dislocation ends, however, indicate that for aluminum the number of dislocations per square centimeter is in the order of

$$N_0 \ell \doteq 10^5 \text{ to } 10^6 \tag{6}$$

For germanium[7] the number of dislocations per square cm. is about 10^4 to 10^5 per sq cm.

According to Mott[8] the most likely form for a network of dislocations is one for which dislocations from three intersecting glide planes meet in a point as shown in Fig. 3. The Burger's vectors then add up to zero and

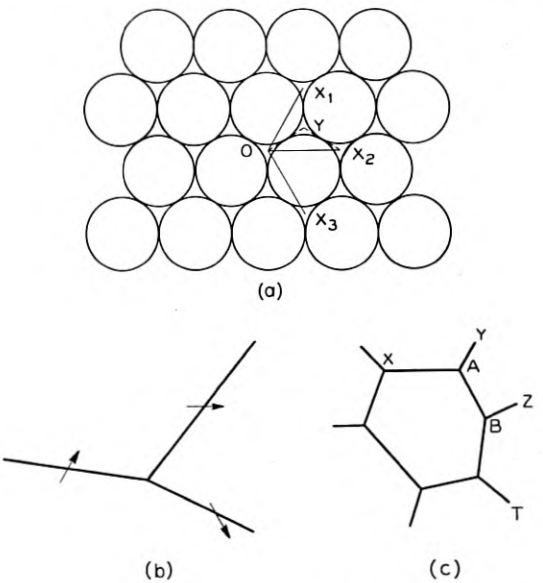


Fig. 3 — Dislocation network forming crystal mosaic (after Mott) (a) A close packed plane, showing the Burgers vector for complete and half dislocations (b) Junction of three dislocations (c) Network of dislocations.

[6] I. and H. Suzuki, Dislocation Networks in Crystals, Rep. Res. Inst. Tohoku Univ. **A6**, No. 6, Dec., 1954.
[7] F. Vogel, private communication.
[8] N. F. Mott, Phil. Mag., **43**, 1151, 1952.

the nodes are stable. A network of such dislocations can outline the mosaic blocks as already discussed. However, the ultrasonic data presented shortly indicate that the effective loops are delineated by impurity atoms rather than by dislocation nodes. This is a possibility even with the model of Fig. 3 if impurity atoms settle along dislocation lines and clamp them at definite points. Impurity atoms having different radii than the solvent atoms are attracted to dislocations since they relieve energy by their presence. If the radius of the impurity atom is smaller than that of the solute atom, the former will take the position shown by the dotted line of Fig. 1 in the compressed region since this position corresponds to lower energy. If the radius is larger than that of the solvent atoms, the impurities will settle in the region of crystal extension.

Using elastic theory Cottrell[9] has shown that the binding energy of an impurity atom is given by the equation

$$U_B = \frac{4}{3}\left(\frac{1+\sigma}{1-\sigma}\right)\frac{\mu b r^3 \epsilon \sin\theta}{R} \tag{7}$$

where r is the radius of solvent atom, $\epsilon = (r' - r)/r$ where r' is the radius of the impurity atom, θ is the angle between the glide plane and the line joining the dislocation center and the impurity atom and R is the distance of the dislocation center from the impurity atom. At low temperatures the dislocation atoms will settle in the position of minimum energy which is $\sin\theta = -1$; $R = r$. Equation (7) was calculated on the basis of elastic theory which is not strictly valid for such large strains. Comparison with experimental values[10] shows that the measured energy is about $\frac{1}{3}$ to $\frac{1}{2}$ that calculated from (7). Hence we take

$$U_B = \frac{1}{3}\left(\frac{1+\sigma}{1-\sigma}\right)\mu b^3 \epsilon \tag{8}$$

for a face-centered metal for which $r = b/\sqrt{2}$. Consequently the model considered for a pure single crystal is the one shown in Fig. 4. It consists of the basic network of dislocations shown in Fig. 3 with a distribution of impurity atoms along the dislocations determining the average loop length between pinning points. With this model

$$N_0 \ell^3 \lesssim 1 \tag{9}$$

The ultrasonic data presented in the next sections indicate that, for aluminum, $N_0\ell^3 \doteq 0.08$, while for lead it is about 0.5.

The fundamental unit considered for ultrasonic attenuation is then the

[9] A. H. Cottrell, loc. cit., p. 57.
[10] A. H. Cottrell, loc. cit., p. 134.

pinned dislocation of average length ℓ_0. Such a loop acts like a stretched string and it can be shown[11] to have a tension T equal to

$$T = \mu b^2 \tag{10}$$

At absolute zero such loops remain stationary in their minimum energy positions, but as the temperature rises, thermal agitation occurs and a dislocation loop may become displaced to the next minimum energy position. When a dislocation moves from one minimum energy position it has to overcome the shearing stress tending to return it to the mini-

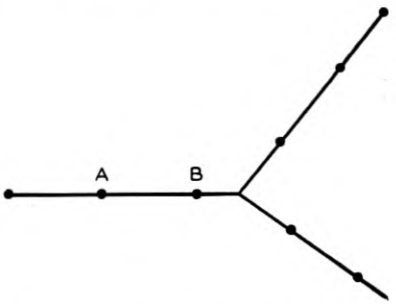


Fig. 4. — Dislocation model with impurity atoms.

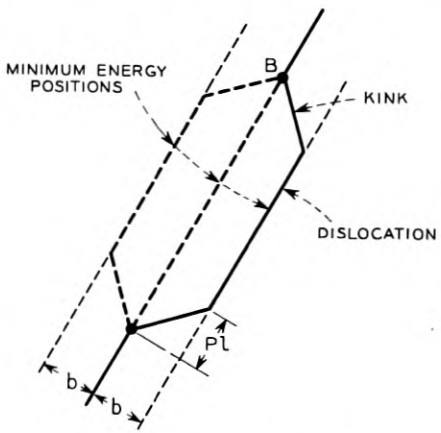


Fig. 5 — Dislocation loop with pinning atoms and form of dislocation loop.

[11] A. H. Cottrell, loc. cit., p. 53.

mum energy position; this is usually taken as a sinusoidal stress

$$T_{13} = T_{13_0} b \sin \frac{2\pi x}{b} \tag{11}$$

where T_{13_0} is the limiting Peierls shearing stress required to surmount the barrier and x is the loop displacement. In addition to this, energy is required to stretch the dislocation against its tension T. The model that has usually been considered[12] for this type of motion is the one shown by Fig. 5. This consists of a straight section of dislocation connected to the pinning points by two "kinks" or approximately straight sections of dislocations cutting across between two minimum energy positions. The energy associated with the dislocation loop at any displacement d from a minimum energy position can be calculated as follows. If $p\ell$ is the percentage of the total length of the loop covered by a "kink", the increase in length for a displacement d is

$$\Delta\ell = 2(\sqrt{(p\ell)^2 + d^2} - p\ell) \doteq \frac{d^2}{p\ell} \tag{12}$$

The work done against the tension T is

$$W_1 = T\Delta\ell = \frac{\mu b^2 d^2}{p\ell} \tag{13}$$

Work is also done against the restoring force of (11) in an amount

$$W_2 = T_{13_0} b \int_0^y dy \int_0^{d(y)} \sin \frac{2\pi x}{b}\, dx \tag{14}$$

Performing the integration, we find that

$$W_2 = \frac{T_{13_0} b^2 \ell}{2\pi}\left[1 - \cos\frac{2\pi d}{b} + 2p\left(\cos\frac{2\pi d}{b} - \frac{b}{2\pi d}\sin\frac{2\pi d}{b}\right)\right] \tag{15}$$

Hence the total energy W is the sum of (13) and (15).

When the loop reaches the next energy minimum, $\cos 2\pi d/b = 1$, and

$$A = \frac{T_{13_0} b^2 p\ell}{\pi} + \frac{\mu b^2 d^2}{p\ell} \tag{16}$$

To obtain the minimum energy, the length $p\ell$ is determined by making $W_{\min}$ a minimum with respect to $p\ell$. Differentiating $W_{\min}$ by $p\ell$ and set-

[12] This type of loop was first considered by N. F. Mott and F. R. N. Nabarro, Dislocation Theory and Transient Creep, Report of a Conference on Strength of Solids, published by The Physical Society, 1948, and has been elaborated by Read, loc. cit., p. 47. Read investigated the conditions for kink stability.

ting the result equal to zero, we find

$$p\ell = \sqrt{\frac{\pi\mu}{T_{13_0}}}\, d = \sqrt{\frac{\pi\mu}{T_{13_0}}}\, b \tag{17}$$

since $d = b$ at this position. Hence this length is independent of the total loop length. For the ratio of T_{13_0}/μ found experimentally, i.e., 6.0×10^{-6}, the length of each kink is about

$$p\ell = 720b = 2.5 \times 10^{-5} \text{ cm for lead} \tag{18}$$

The energy for both kinks is from equation (13)

$$W_1 = b^3 \sqrt{\frac{T_{13_0}\mu}{\pi}} \tag{19}$$

The total energy at the first minimum A is then

$$A = 2b^3 \sqrt{\frac{T_{13_0}\mu}{\pi}} \tag{20}$$

The energy H, at the maximum, which occurs when $\cos 2\pi d/b = -1$ becomes

$$H = \frac{T_{13_0} b^2 \ell}{\pi} \tag{21}$$

There are other minima at distances $\pm 2b$, $\pm 3b$, etc., on each side of the central position and it can be shown that the successive minima have values of $2A$, $3A$, etc., while the height of the energy barrier remains at H. The question arises as to whether these other positions should be included in the calculation. This model is an ideal situation in which no strains of any kind are permitted in the medium. Actually, impurity atoms above and below the glide plane introduce stresses which distort the dislocation from its straight line position and have the effect of increasing the energies at the bottom of the potential wells. This effect will be much larger for position of $\pm 2b$, $\pm 3b$, etc., and will probably wipe out these minimum energy positions. For the central line and for the positions $\pm b$, the effect is smaller and hence it appears that the model should have only two alternate positions in addition to the central position. For an absolutely pure crystal these other energy positions would probably exist. Calculations including them show that the form of the Q^{-1} curve given by equation (37) is essentially unchanged but that the multiplying constant $\Delta\mu/\mu$ is increased.

The potential well model for a dislocation displaced from its minimum energy position to the next minimum on either side will then be that

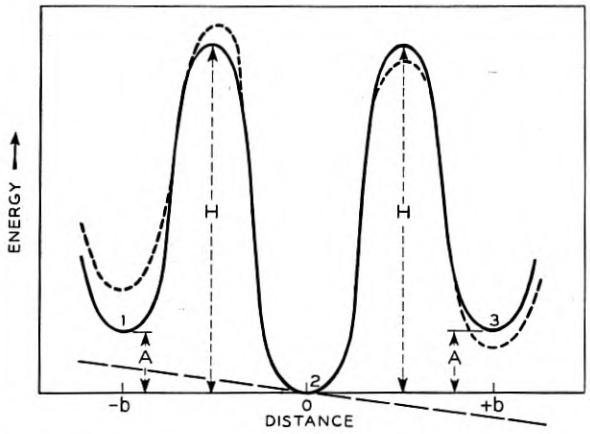


Fig. 6 — Potential well model corresponding to dislocation loop.

shown by Fig. 6. The two energy wells on either side are higher than the center one by the energy A of equation (20). In between these minimum energy positions there are potential barriers of height H given by equation (21). The physical picture given for the process is then that the dislocation vibrates in its lowest potential well (2) until it acquires enough thermal energy to overcome the barrier H and land in wells 1 or 3. According to reaction rate theory, the number of times per second that this process is likely to occur is given by the equation

$$\alpha = \gamma e^{-U/RT} \tag{22}$$

where γ is the number of times that the dislocation attacks the barrier per second, U is the height of the energy barrier and RT the thermal energy. If U is expressed in calories per mole, then R the gas constant per mole is 2 calories per degree increase in temperature. The number of times that the dislocation attacks the barrier should be approximately equal to the resonant frequency of the dislocation in its potential well. Experimentally, it is found that γ is close to $2\pi f = \omega$, the value for the natural vibration.

The resonant frequency of a dislocation in its potential well can be calculated from the restoring force and the mass per unit length which[13]

[13] J. S. Koehler, Chapter VII, Imperfections in Nearly Perfect Crystals, Wiley, 1952. Koehler considers dissipation to be due to dissipation of freely vibrating dislocations. For the potential well model, freely vibrating dislocations do not exist.

is $\pi\rho b^2$. The restoring force can be determined from the derivative of $(W_1 + W_2)$ with respect to d as d becomes small. The result of differentiation, using equations (13) and (15) is

$$\left[2\pi T_{13_0}\ell + b\sqrt{\pi T_{13_0}\mu}\left(\frac{2}{\pi} - \frac{8}{3}\pi\right)\right]d \tag{23}$$

Since the last term is much smaller than the first, we neglect it, and have the equation

$$\pi\rho b^2\ell\frac{d^2d}{dt^2} + 2\pi T_{13_0}\ell d = 0 \tag{24}$$

For simple harmonic motion $d^2d/dt^2 = -\omega^2 d$ so that

$$\pi\rho b^2\ell\omega^2 = 2\pi T_{13_0}\ell \tag{25}$$

Hence the resonant frequency for a dislocation in its potential well is independent of the loop length and equal to

$$f = \frac{1}{2\pi b}\sqrt{\frac{2T_{13_0}}{\rho}} \tag{26}$$

For lead, copper, aluminum, and silver, using the value of $T_{13_0} = 6.0 \times 10^{-6}\ \mu$, the relaxation values of ω for the metals are

	Pb	Cu	Al	Ag	
$\omega = 2\pi f \doteq$	7.8×10^9	2.8×10^{10}	3.6×10^{10}	1.9×10^{10}	(27)

All of these values are considerably higher than any angular frequencies ω which have so far been used. For frequencies approaching these values resonance effects may be expected.

The potential well model of Fig. 6 results in a relaxation type loss and a change in the elastic constant. As with all relaxation effects, energy for low frequency vibrations has time to equilibrate among the various potential wells and is returned at a later part of the cycle without appreciable loss. Since the dislocations can be displaced from their equilibrium positions, a plastic component of strain results and the metal has a greater elastic compliance than would occur without dislocation displacement. As the period of force application becomes comparable with the equilibrating or "relaxation" time an appreciable fraction of the energy is not returned to the vibration but is converted into heat. In this frequency

range, the attenuation becomes a maximum and the elastic constant is intermediate between its high frequency and low frequency values. Finally if the frequency of vibration is sufficiently high there is not time for the dislocation to move from its potential well and the material becomes stiffer and less lossy. At room temperature sufficiently high [illegible] sonic frequencies have not been used to observe this complete process. However, as one lowers the temperature, the relaxation frequency decreases due to the activation energy term $e^{-U/RT}$ of (22); this process was first observed by Bordoni.[14] An explanation in terms of dislocation loops was first published by the writer.[15]

The loss and change in the elastic constant can be obtained by applying reaction rate theory to the model of Fig. 6. The effect of applying a shearing stress along the glide plane is to lower one potential well and raise the other one as shown by the dashed line of Fig. 6. The effect of any other stress is to provide a component of shearing stress in the glide plane and the magnitude of the relaxation effect is then related to the relaxation in the glide plane as discussed in the next section. In general a relaxation measured in one stress system has to be multiplied by a factor F to equal that in the glide plane.

The lowering of the potential well 3 by the shearing stress T_{13} is equal to

$$\Delta = T_{13}b^2\ell(1 - p) \tag{28}$$

Potential well 1 is raised by a similar amount. This results in a redistribution of the number of dislocations in the three types of wells and hence a plastic strain equal to

$$S_{13}{}^P = (N_3 - N_1)(1 - p)b^2\ell, \tag{29}$$

where $(1 - p)b\ell$ is the area swept out by a single loop. The rate at which the shearing stress changes with time is determined by combining the four transition probabilities for the three wells and since the details of the calculation have been given previously,[16] only the final result is given.

---

[14] P. C. Bordoni, Elastic and Anelastic Behavior of Some Metals at Very Low Temperatures, J. Acous. Soc. Am., **26**, July, 1954.

[15] W. P. Mason, Dislocation Relaxations at Low Temperatures and the Determination of the Limiting Shearing Stress of a Metal, Phys. Rev., **98**, pp. 1136–1138, May 15, 1955.

[16] W. P. Mason, Relaxations in the Attenuation of Single Crystal Lead at Low Temperatures and Their Relation to Dislocation Theory, J. Acous. Soc. Am., **27**, July, 1955. For other measurements see also paper K4 by B. Welber, Program of the 49th meeting of the Acoutica Society of America, July 2, 1955. The relaxation in lead at 50 kc was shown as a slide.
July, 1955.

For finite stresses the formula for the plastic strain takes the form

$$S_{13}{}^{P} = \frac{2e^{-A/RT}N_0(1-p)b^2\ell \sinh \frac{\Delta}{kT} \cdot \left[1 - \frac{j\omega}{2\gamma}\left(\frac{1+e^{A/RT}}{\cosh \Delta/2kT}\right)e^{(H-A)/RT}\right]}{1 + 2e^{-A/RT}\left[\cosh \frac{\Delta}{kT} - \frac{j\omega}{\gamma}\right] \cdot (4 + 2e^{-A/RT}) \cosh \frac{\Delta}{2kT}\, e^{(H-A)/RT} + \frac{3\omega^2}{\gamma^2}\, e^{2(H-A)/RT}} \tag{30}$$

if H and A are expressed in calories per mole. For stresses such that

$$\frac{\Delta}{kT} = \frac{T_{13}b^2\ell(1-p)}{kT} \ll 1$$

we can replace the sinh by the argument and the cosh by 1. Hence equation (30) reduces to

$$\frac{S_{13}{}^{P}}{T_{13}} = \frac{\frac{2e^{-A/RT}}{1+2e^{-A/RT}}\left(\frac{N_0\ell^2(1-p)^2b^4}{kT}\right) \cdot \left[1 - \frac{j\omega}{\gamma}\left(\frac{1+e^{A/RT}}{2}\right)e^{(H-A)/RT}\right]}{\left(1 - \frac{j\omega}{\gamma}\left(\frac{3}{1+2e^{-A/RT}}\right)e^{(H-A)/RT}\right)\left(1 - \frac{j\omega}{\gamma}e^{(H-A)/RT}\right)} \tag{31}$$

If we expand the right hand side of (31) into real and imaginary parts, we find

$$\frac{S_{13}{}^{P}}{T_{13}} = \frac{2e^{-A/RT}}{1+2^{-A/RT}}\left[\frac{N_0\ell^2(1-p)^2b^4}{kT}\right] \left\{\frac{1 + \left(\frac{\omega}{\omega_0}\right)^2\left(\frac{2e^{A/RT}+e^{-A/RT}}{1+2e^{-A/RT}}\right) + j\frac{\omega}{\omega_0}\left[\frac{5+2e^{-A/RT}-e^{A/RT}}{2(1+2e^{-A/RT})} + \left(\frac{\omega}{\omega_0}\right)^2\left(\frac{3(1+e^{A/RT})}{2(1+2e^{-A/RT})}\right)\right]}{1 + \frac{\omega^2}{\omega_0{}^2}\left(\frac{10+4e^{-A/RT}+4e^{-2A/RT}}{1+4e^{-A/RT}+4e^{-2A/RT}}\right) + \left(\frac{\omega}{\omega_0}\right)^4\left(\frac{3}{1+2e^{-A/RT}}\right)^2}\right\} \tag{32}$$

and

$$\omega_0 = \gamma e^{-(H-A)/RT}$$

If A is small compared to RT this equation reduces to the usual expression

$$\frac{S_{13}{}^{P}}{T_{13}} = \frac{2e^{-A/RT}}{1 + 2e^{-A/RT}} \left[\frac{N_0\ell^2(1 - p)^2b^4}{kT}\right] \frac{1 + j\dfrac{\omega}{\omega_0}}{1 + \left(\dfrac{\omega}{\omega_0}\right)^2}$$

An intermediate equation is obtained by expanding the three terms of (31) into a series for the two factors in the denominator divided by the factor in the numerator. The first two terms yield in the equation,

$$\frac{S_{13}{}^{P}}{T_{13}} = \frac{2e^{-A/RT}}{1 + 2e^{-A/RT}} \left(\frac{N_0\ell^2(1 - p)^2b^4}{kT}\right) \left(\frac{1 + j\dfrac{\omega}{\omega_0}}{1 + \left(\dfrac{\omega}{\omega_0}\right)^2}\right) \tag{33}$$

$$\text{where} \qquad \omega_0 = \frac{\gamma e^{-(H-A)/RT}}{F}$$

and

$$F = \frac{2 + 5e^{A/RT} - e^{2A/RT}}{2(2 + e^{A/RT})}$$

This equation will be used in evaluating the measurements to be discussed in the next section.

If we add to this the purely elastic strains $S_{13}{}^{P} = T_{13}\,\mu^{E}$, the ratio of the total strain to the applied stress is

$$\frac{S_{13}{}^{P} + S_{13}{}^{E}}{T_{13}} = \frac{1}{\mu} = \frac{1}{\mu^{E}} + \frac{2e^{-A/RT}}{1 + 2e^{-A/RT}} \frac{N_0(1 - p)^2b^4\ell^2}{kT} \left(\frac{1 + j\omega/\omega_0}{1 + (\omega/\omega_0)^2}\right) \tag{34}$$

This equation shows that the elastic shear modulus μ varies with frequency and the presence of an imaginary term indicates that there is a dissipation associated with this relaxation. For the real part we find

$$\frac{\mu^{E} - \mu}{\mu} = \frac{\Delta\mu}{\mu} = \left(\frac{2e^{-A/RT}}{1 + 2e^{-A/RT}}\right) \frac{N_0\ell^2(1 - p)^2b^4}{kT}\,\mu^{E} \left(\frac{1}{1 + \omega^2/\omega_0{}^2}\right) \tag{35}$$

The total change in elastic constant $\Delta\mu^{0}$ which occurs when the frequency goes from zero to infinity is then

$$\frac{\Delta\mu^{0}}{\mu^{0}} = \left(\frac{2e^{-A/RT}}{1 + 2e^{-A/RT}}\right) \frac{N_0\ell^2(1 - p)^2b^4\mu^{E}}{kT} \tag{36}$$

Since the imaginary part of (34) represents the dissipation of energy it

can be shown that

$$\frac{1}{Q} = \frac{\Delta\mu^0}{\mu^0} \frac{\omega/\omega_0}{1 + (\omega/\omega_0)^2}. \tag{37}$$

These are the equations connected with a single relaxation which, in this case, would be associated with the displacement of a single length dislocation. Since there is a distribution of loop lengths about the most probable value, the activation energy, H, will vary and hence each length will have a different relaxation frequency. Equations (35) and (37) can be generalized to the forms

$$Q^{-1} = \sum_{i=1}^{n} \frac{\Delta_{\mu i}}{\mu_0} \left(\frac{\omega/\omega_i}{1 + (\omega/\omega_i)^2} \right); \qquad \frac{\mu}{\mu_0} = 1 + \sum_{i=1}^{n} \frac{\Delta_{\mu i}}{\mu_0} \left(1 \frac{(\omega/\omega_i)^2}{1 + (\omega/\omega_i)^2} \right) \tag{38}$$

As will be discussed in the next section, these equations can be used to evaluate the characteristics of a relaxation at low temperatures which is connected with the displacement of dislocations from their minimum energy positions to adjacent minimum energy positions. In the process they give experimental evidence on the limiting shearing stress in metals, on the number of dislocations per square centimeter in metals and on the average loop lengths. The values found are in good agreement with other methods for measuring these properties.

EXPERIMENTAL EVIDENCE FOR DISLOCATION RELAXATIONS AT LOW TEMPERATURES

The first experimental evidence for a relaxation at low temperatures was provided by the work of Bordoni.[14] Using an electrostatic drive method, Bordoni measured the inverse Q values for a number of metals down to liquid helium temperatures. All these measurements were made with very small strains, i.e., less than 10^{-8}, which are in themselves too small to displace dislocations from their minimum energy positions. They can, however, bias the potential wells of the model of Fig. 6. Then, temperature induced motions arise so that, for slowly applied motions, equilibrium of dislocation distributions can occur between the potential wells, with the applied stress thereby superimposing a plastic strain on the elastic strain. For very high frequencies there is not time for the dislocations to redistribute themselves and only the elastic strain occurs. This process results in a relaxation which extends over a frequency range. As discussed in the previous section, the dislocations have too high a natural frequency for this process to be observed at room temperature.

All of Bordoni's measurements were made for polycrystals or single crystals of a definite length, i.e., 6.4 cms. The natural resonance fre-

angular relaxation frequencies $\omega_0 = 2\pi f_0$ can be represented by the equation

$$\omega_0 = 5.3 \times 10^9 \, e^{-975/RT} \tag{40}$$

so that we are dealing with a process having an activation energy of about 975 calories per mole. This is a very low activation energy and the only known process which has as low an energy as this is the one discussed in the previous section, namely, the displacement of a dislocation, by one atomic spacing, against the limiting shearing stress of the crystal.

To show that this is of the right order of magnitude, we may evaluate the activation energy of the process from (20) and (21):

$$H - A = \left[\frac{T_{13_0} b^2 \ell}{\pi} - 2b^3 \sqrt{\frac{T_{13_0}\mu}{\pi}}\right] \frac{6.025 \times 10^{23}}{4.182 \times 10^7} \text{ (cal./mole)} \tag{41}$$

As will be shown presently, the energy average loop length is of the order of 4×10^{-4} cm and since it is known that

$$b = 3.5 \times 10^{-8} \text{ cm}; \quad \text{and} \quad \mu = 7.0 \times 10^{10} \text{ dynes/cm}^2 \tag{42}$$

we find for the limiting shearing stress a value of 4.8×10^5 dynes/cm^2.

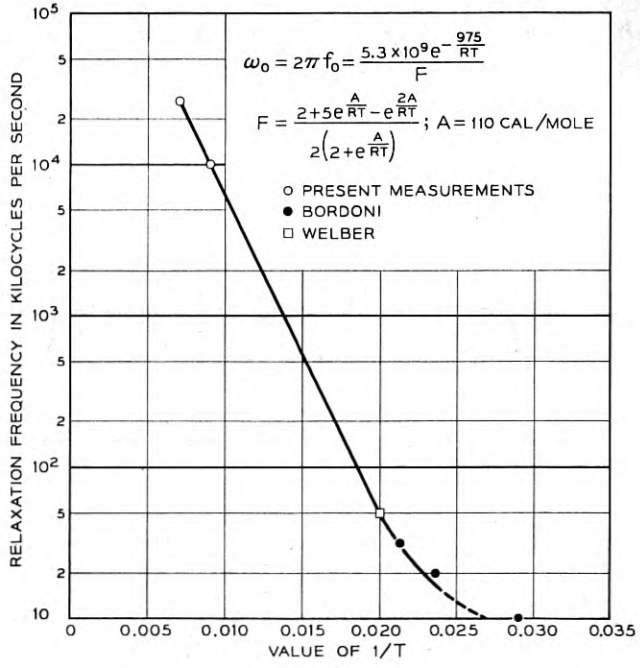


Fig. 9 — Plot of log of relaxation frequency against $1/T$.

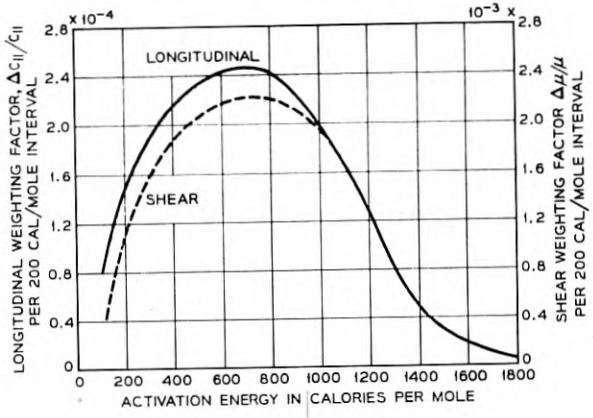


Fig. 10 — Weighting factors for longitudinal and shear waves plotted against activation energies.

Due to the factor, F, a slightly nonlinear relation is obtained between the relaxation frequencies and the reciprocal of the temperature as shown in Fig. 9. Hence we find

$$T_{13_0} = 4.8 \times 10^5 \text{ dynes}; \qquad T_{13_0}/\mu = 6.8 \times 10^{-6} \tag{43}$$

values in good agreement with Peierls' lower theoretical value in (1). The value of $\gamma = 5.3 \times 10^9$ is within a factor of 1.5 of the calculated angular resonance frequency of a dislocation in its potential well as given by (26).

In calculating T_{13_0} an assumed value was used for the average length of a dislocation loop. Some evidence on the size of the dislocation loops and their distribution around the average size can be obtained from the data of Fig. 8 for the attenuation measured in a lead single crystal. The form of the $1/Q$ curve for a single loop length is given by (32). The best fit to the measured values is shown by the dashed curve of Fig. 8(a), and, as can be seen therein, a single relaxation does not agree fully with the measured values. If we assume a distribution of activation energies, the measured curves can be fitted by using the weighting function given in Fig. 10. This verifies the existence of loops shorter and longer than the mean value, in agreement with the model shown in Fig. 4. Since the measured activation energies, $H - A$, are nearly proportional to the loop lengths, the distribution in activation energies also corresponds to the distribution of loop lengths. The distribution function for the shear measurements is also shown in Fig. 10. The height of the shear curve is about 9 times as high as that for the longitudinal measurement. It will be

evident from subsequent discussion of equation (45) that this is to be expected. The number of dislocations for the shear crystal is about 40 per cent higher than for the longitudinal crystal.

Some data on the average loop length can be obtained from the data of Fig. 8(a). If we consider a single loop length as shown by the dashed curve and increase the maximum until the area under the loop is equal to that under the measured curve, the value of the weighting factor $\Delta c_{11}/c_{11}$ becomes

$$\frac{\Delta c_{11}}{c_{11}} = 8.4 \times 10^{-4} \tag{44}$$

Since all the formulas used in dislocation theory refer to isotropic materials rather than cubic crystals, we take

$$c_{11} = \lambda + 2\mu$$

To determine how a measured change in $\Delta c_{11}/c_{11}$, determines a weighting factor $\Delta\mu/\mu$, we note that the bulk modulus, B, does not have any relaxation effect, i.e., a pure compression will cause no shearing motion in the glide plane. Since

$$B = \lambda + \tfrac{2}{3}\mu; \qquad \Delta B = \Delta\lambda + \tfrac{2}{3}\Delta\mu = 0; \qquad \text{Hence} \quad \Delta\lambda = -\tfrac{2}{3}\Delta\mu$$

Therefore

$$\frac{\Delta c_{11}}{c_{11}} = \frac{\Delta\lambda + 2\Delta\mu}{\lambda + 2\mu} = \left(\frac{\frac{4}{3}\Delta\mu}{\mu}\right)\frac{\mu}{\lambda + 2\mu},$$

$$\text{and} \qquad \frac{\Delta\mu}{\mu} = \frac{3}{4}\left(\frac{\lambda + 2\mu}{\mu}\right)\frac{\Delta c_{11}}{c_{11}} \tag{45}$$

It is readily shown that the relation between the change in Young's modulus and the shearing modulus is

$$\frac{\Delta\mu}{\mu} = \frac{3\mu}{Y_0}\left(\frac{\Delta Y_0}{Y_0}\right), \tag{46}$$

a relation needed later.

From (35), (44), and (45), we have

$$4.3 \times 10^{-3} = \frac{2e^{-AR/T_0}}{1 + 2e^{-A/RT_0}}\left[\frac{\bar{N}\ell(1-p)^2b^4}{kT_0}\right]\mu^E \tag{47}$$

where $\bar{N} = N_0\ell$ is the number of dislocations per square centimeter, which is a measurable quantity if one uses the etch technique.

Since $A = 110$ calories per mole, $T_0 = 140°\text{K}$, $(1-p)^2 = 0.90$; $b = 3.5$

$\times 10^{-8}$; $k = 1.38 \times 10^{-16}$ and $\mu^E = 7.0 \times 10^{10}$ dynes per sq cm we find

$$\bar{N}\ell \doteq 1.3 \times 10^3 \tag{48}$$

Data from the mechanical hysteresis effect discussed in the next section indicate that

$$\bar{N} \doteq 3.1 \times 10^6 \tag{49}$$

and hence the average loop length is about

$$\ell \doteq 4 \times 10^{-4} \text{ cm} \tag{50}$$

Hence we have

$$\bar{\ell} \doteq 4 \times 10^{-4} \text{ cm}, \quad \text{and} \quad N_0\bar{\ell}^3 \doteq 0.5 \tag{51}$$

One can use the measurements of Bordoni shown in Fig. 7 to show that the present interpretation, attributing the relaxation to the displacement of a dislocation by one atomic spacing against the limiting shear stress of the crystal, is consistent for several face-centered metals. Since the temperature of maximum attenuation results when the measuring frequency equals the relaxation frequency we have

$$\frac{\omega}{\gamma e^{-(H-A)/RT}} = 1$$

Since the ratio of each measuring frequencies listed in (39) to the corresponding resonant frequency for dislocations (27) is nearly a constant, the activation energy, $H - A$, will be proportional to the temperature of maximum loss. The activation energy of lead was evaluated as 975 calories per mole at a temperature of 35°K. Hence, we can evaluate the activation energy of each metal; as shown in Table II, the average value of (T_{13_0}/μ) is about 6×10^{-6}.

In view of the approximate nature of this calculation (including the tacit assumption that all the loop lengths are the same for the different metals), the results in Table II constitute satisfactory confirmation of the

TABLE II

Metal	Temp. of Max. Atten.	H-A in cals mole	b in cm $\times 10^8$	μ dynes/cm^2	T_{13_0} dynes/cm^2	T_{13_0}/μ
Pb....	35°K	975	3.5	7.0×10^{10}	4.8×10^5	6.8×10^{-6}
Cu....	85°K	2360	2.55	4.6×10^{11}	2.4×10^6	5.2×10^{-6}
Al.....	100°K	2760	2.86	2.5×10^{11}	1.9×10^6	7.6×10^{-6}
Ag....	60°K	1710	2.88	2.7×10^{11}	1.3×10^6	4.8×10^{-6}

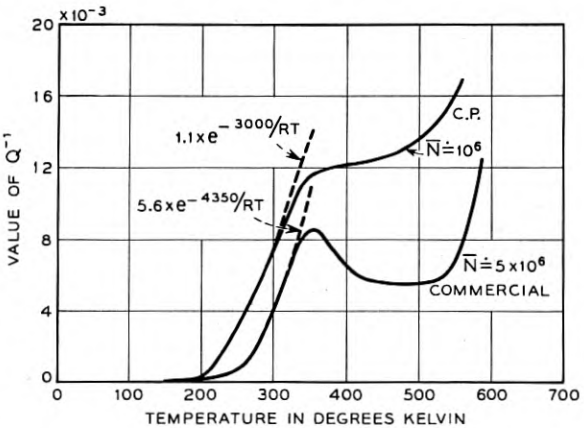


Fig. 11 — Higher temperature values of mechanical hysteresis effect (Bordoni and Nuovo).

proposed origin of the relaxation frequency. The values found are in good agreement with the most probable values given by Cottrell.[4]

TEMPERATURE ACTIVATED MECHANICAL HYSTERESIS MECHANISMS

In addition to the relaxation mechanism discussed in the last section, the measured values in Figs. 8 show that there is another source of loss which is independent of frequency for a given specimen. Fig. 11 shows an extension of the measurements of Bordoni, by Bordoni and Nuovo,[17] to a higher temperature for pure and commercial lead. Both materials show an increase in attenuation of the form

$$Q^{-1} = Ce^{-U/RT}$$

with the values

$$\underset{\text{pure lead}}{C = 1.1; \quad U = 3000 \text{ cal/mole}; \quad \text{and}} \qquad (52)$$

$$\underset{\text{commercial lead}}{C = 5.6; \quad U = 4350 \text{ cal/mole}.}$$

Above 350°K, this increase reaches a maximum and then drops off. A similar rise was found by Ké[18] for aluminum single crystals at 0.8 cycles

[17] P. G. Bordoni and M. Nuovo, Sulla Dissipozione Delle Onde Elastiche Nel Piombo Ad Alta Tempetura, Il Nuovo Cimento, **11,** pp. 127–140, Feb. 1, 1954.

[18] T. S. Ké, Experimental Evidence of the Viscous Behavior of Grain Boundaries in Metals, Phys. Rev., **71,** p. 533, 1947. See also C. Zener, Elasticity and Anelasticity of Metals, p. 151, Fig. 48, Chicago Univ. Press, 1948.

as shown in Fig. 12. This rise follows the same formula with

$$C = 0.37 \qquad \text{and} \qquad U = 5200 \text{ cals/mole} \tag{53}$$

Since this effect is independent of frequency it is called a temperature activated mechanical hysteresis.

The model of Fig. 4 admits only two possibilities. One is that the pinning points can momentarily be torn from the dislocation, thereby allowing energy to be transmitted from one loop to the other. The other possibility is that thermal energy will be sufficient to generate an unstable Frank-Read loop which will carry off energy. It is readily shown, however, that it would require 10,000 times the activation energy to cause the loop breakdown and hence the Frank-Read loop cannot be the mechanism. As discussed later, the energy required to remove an impurity atom is of the right order of magnitude and hence this is considered the cause of the hysteresis loss.

The question is how a momentary breakaway of a pinning point can abstract energy from the vibration. All the loops undergo thermal vibration with a velocity determined by the equation

$$\tfrac{1}{2}(\dot{x}_T)^2 m = kT \tag{54}$$

where m is the mass of the loop which is $\pi\rho b^2 \ell$, where ρ is the density of the medium, and $\dot{x}_T$ is the thermal velocity as the loop crosses the equilibrium position. For lead, with loops of 4×10^{-4} cm in length, the mass of the dislocation is about 1.7×10^{-17} grams so that the thermal vibration velocity is about 70 cm/sec at room temperature. This is large

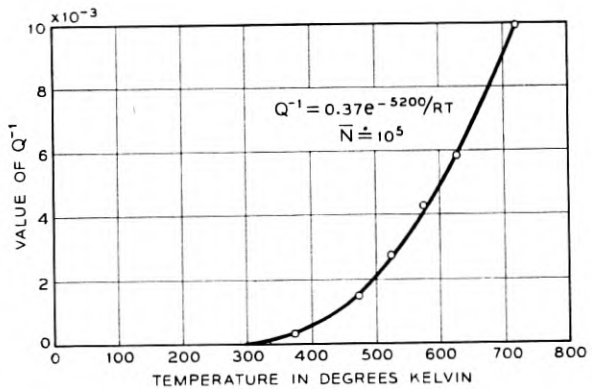


Fig. 12 — Temperature activated mechanical hysteresis effect in aluminum single crystal (after Ké).

compared to the velocity of any particle in the crystal. However, the particle velocity $\dot{u}$ can add to or substract from the thermal motion when $\dot{u}$ and $\dot{x}_T$ are in or 180° out of phase; hence the energy of vibration may be written

$$\tfrac{1}{2}(\dot{x}_T + \dot{u})^2 m + \tfrac{1}{2}(\dot{x}_T - \dot{u})^2 m = (\dot{x}_T + \dot{u})^2 m$$

since it is equally probable for the thermal velocity to add or to subtract from the particle velocity. As long as the loop is pinned at the ends, this energy is returned to the crystal vibration without loss since the vibrations are coherent. When a pinning point is broken, on the average, energy equal to

$$\frac{(\dot{u})^2 m}{2} \tag{55}$$

is abstracted from each loop and since two loops are involved for each impurity atom, an energy of $\dot{u}^2 m$ is abstracted from the crystal vibration.

To calculate the acoustic loss to be expected from this source, consider a shear strain of the form shown in Fig. 13 to be propagated in the direction perpendicular to the glide plane. At $t = 0$, the particle displacement is zero at $x = 0$ and u at $x = \ell$. After an interval of time dt, the strain will be displaced a distance $V_s dt$, where V_s is the shear velocity which is 8.0×10^4 cm/sec for lead. The particle velocity $\dot{u}$ is a constant over the time of the wave. The energy lost in going a distance $\ell_t = V_s dt$ can be calculated as follows. The energy lost per dislocation vibration is $m(\dot{u})^2/2$. For an interval of time dt this has to be multiplied by fdt where f is the frequency of vibration of a dislocation. Hence the loss from a

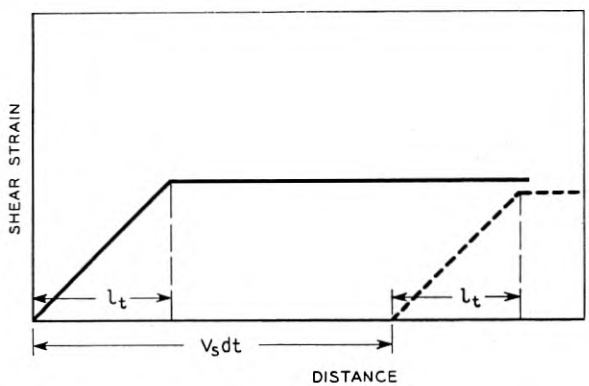


Fig. 13 — Form of shear wave for calculating mechanical hysteresis losses.

single unpinned vibration, in time dt is

$$\dot{u}^2 mf\,dt = \frac{\dot{u}^2(\pi\rho b^2\ell)\sqrt{\frac{2T_{13_0}}{\rho}}}{4\pi b} = \frac{\dot{u}^2 b\ell\sqrt{\frac{T_{13_0}\rho}{2}}\,dt}{2} \tag{56}$$

To get the total loss we have to multiply this by the number of unpinned dislocation loops. From Boltzmann's principle, the number of unpinned loops per cc will be

$$2N_0 e^{-U/RT} \tag{57}$$

where N_0 is the number pinned loops per cc and U the activation energy for tearing an impurity atom away. We have to multiply this by the volume of the disturbance which may be written $V_s dt$ if a unit cross-sectional area is considered. Hence the total loss in time dt is

$$\left(\dot{u}b^2\sqrt{\mu\rho}\sqrt{\frac{T_{13_0}}{2\mu}}\,V_s N_0\ell e^{-U/RT}\right)(dt)^2 \tag{58}$$

The total input energy for the wave, in time dt, is

$$W_0 = \dot{u}^2\sqrt{\mu\rho}\,dt, \tag{59}$$

where $\sqrt{\mu\rho}$ is the characteristic impedance of the medium. Hence the energy transmitted in the first interval of time dt takes the form

$$W = W_0\left[1 - \frac{\left(\sqrt{\frac{2T_{13_0}}{\mu}}\,bV_s N_0\ell e^{-UR/T}\right)}{2}dt\right]. \tag{60}$$

This is the first term of the expansion of the equation

$$W = W_0 e^{-\frac{\left(\sqrt{\frac{2T_{13_0}}{\mu}}\,bV_s N_0\ell e^{-U/RT}\right)t}{2}} = W_0 e^{-\delta t}$$

where δ is the decrement. Since the decrement $\delta = \pi/Q$, we have finally

$$\frac{1}{Q} = \frac{\sqrt{\frac{2T_{13_0}}{\mu}}\,bV_s N_0\ell e^{-U/RT}}{2\pi} \tag{62}$$

This loss is independent of frequency since it does not depend on ℓ_t which determines the steepness of the displacement-time curve. It does not depend on the amplitude up to a strain which can cause breakaways due to the applied strain alone.

To see if (62) yields a reasonable value for a single crystal, assume a temperature of 250°K for which measurements in Fig. 8 along the [100]

direction show that $Q^{-1} = 1.55 \times 10^{-4}$. Multiplying this by 5.2, [see equation (45)], to give the value for a shear wave propagated along the [111] direction we have

$$\frac{1}{Q} = 8.0 \times 10^{-4} = \frac{3.48 \times 10^{-3} \times 3.5 \times 10^{-8} \times 8.0 \times 10^{4} \times 1.7 \times 10^{-4}(N_0\ell)}{2\pi} \tag{63}$$

Solving for $N_0\ell$, the number of dislocations per sq cm, we have

$$\bar{N} = N_0\ell \doteq 3.1 \times 10^{6} \tag{64}$$

If we perform the same calculations with the data for polycrystalline lead, multiplying the resultant values of Q^{-1} by

$$\frac{\Delta\mu}{\mu} = \frac{3\mu}{Y_0}\left(\frac{\Delta Y_0}{Y_0}\right) = 1.2\,\frac{\Delta Y_0}{Y_0}, \tag{65}$$

we find

C. P. lead	Commercial lead
$\bar{N} = 1.2 \times 10^{6}$	$\bar{N} \doteq 6 \times 10^{6}$

(66)

Hence the number of dislocations for an unstrained single crystal or polycrystal appears to be in the neighborhood of 10^6 to 6×10^6.

A somewhat lower value is found for aluminum from the data of Fig. 12. This curve shows an increase in Q^{-1} for a single crystal which can be represented by the equation

$$Q^{-1} = 0.37e^{-5200/RT} \tag{67}$$

These measurements were made for a single crystal wire using a torsional oscillation of 0.8 cycles per second. Since this is for a shear vibration, no correction will be required in the isotropic approximation. To make this agree with equation (62) for $b = 2.86 \times 10^{-8}$ and $V_s = 3 \times 10^5$ cm/sec, we must have

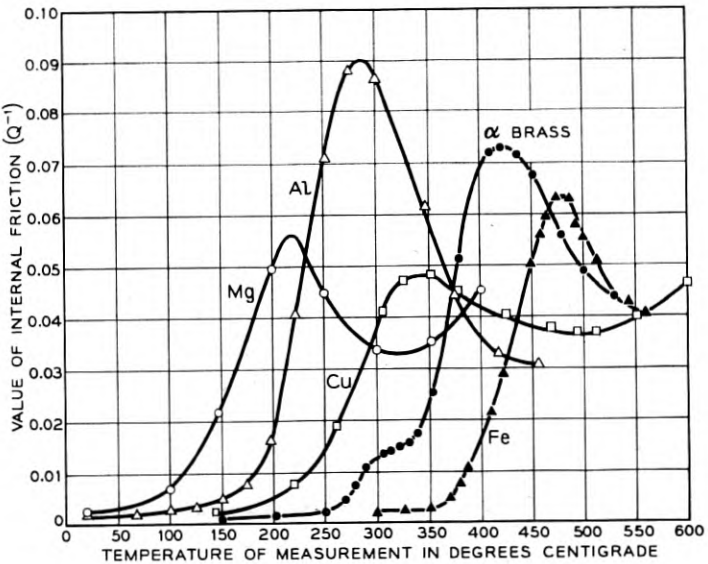
$$\bar{N} = N_0\ell = 10^{5} \tag{68}$$

which is in fair agreement with the values of 10^5 to 10^6 determined from etch pit data.

This temperature actuated mechanical hysteresis is a property of all the metals measured at high temperatures as is shown by the data of Ké[19] given in Fig. 14(a). Here a peak associated with grain boundary

[19] J. S. Ké, J. Appl. Phys., **21**, p. 414, 1950.

(a)



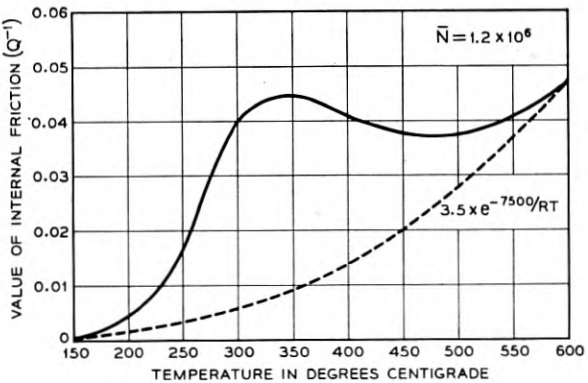


Fig. 14 — (a) Mechanical hysteresis and grain boundary losses in several polycrystalline metals, (b) Separation of losses for copper (after Ké).

motion (discussed in the next section) is followed by an exponentially rising value of internal dissipation as measured by Q^{-1}. The curve for copper is shown in detail in Fig. 14(b). We can separate out the grain boundary loss from the hysteresis loss as shown by the dashed line. This curve can be represented by the equation of the figure with a constant $C = 3.5$ and an activation energy of 7500 calories per mole. Analysis

similar to that outlined for lead and aluminum yields about 1.2×10^6 for the number of dislocations per sq cm.

It remains to be shown that the theoretical activation energy for unpinning a pinning point is of the right order of magnitude to agree with experiment. The binding energy of an impurity atom is given by (8). To obtain $U = 3{,}000$ to 4,350 calories per mole for lead, 5,200 for aluminum, and 7,500 for copper, we must have the following ratios of impurity atom radius to metal atom radius;

$$\text{Pb};\ \epsilon = 0.108 \text{ to } 0.156: \qquad \text{Al};\ \epsilon = 0.102: \qquad \text{Cu};\ \epsilon = 0.11 \tag{69}$$

These values are close to what would be expected for the principal impurities which are bismuth and antimony for lead, iron and copper for aluminum, and silicon for copper.

This process should be present for all metals but the activation energy is usually too high for this effect to be observed at room temperature. For lead, as seen from Fig. 11, the effect reaches a maximum at a temperature of about 350°K. This temperature is not far from the critical temperature T_0 of a Cottrell atmosphere, for which the impurity atoms cease to be condensed around a dislocation but, due to thermal agitation, form a "Maxwell" atmosphere around the dislocation. The energy to pull the dislocation away from the atmosphere increases and the number of free dislocation loops decreases. This temperature[20] is

$$T_0 = U/R \log (1/c_0) \tag{70}$$

where c_0 is the concentration of impurities in the metal.

For room temperature measurements for most single crystals and unstrained polycrystals, it appears that most of the loss is due to the relaxation mechanism previously discussed. This loss becomes nonlinear with amplitude for large amplitudes as shown by the measurements of Nowick[21] for copper single crystals. These results are shown in Fig. 15. For strain amplitudes above 1×10^{-7}, the value of Q^{-1} increases as a function of increasing amplitude. Similar results have been found by Read[21] and others. This can be expected from the model of Fig. 6. When the value of

$$\frac{\Delta}{kT} = \frac{T_{13} b^2 \ell (1 - p)}{kT} = 1, \tag{71}$$

the argument can no longer replace the sinh. For copper, this corresponds

[20] A. H. Cottrell, Dislocations and Plastic Flow in Crystals, p. 141. Oxford University Press, 1953.

[21] A. S. Nowick, Phys. Rev., **80**, p. 249, 1950. T. A. Read, Trans. A.I.M.E., **143**, p. 30, 1941.

to

$$T_{13} = \frac{1.38 \times 10^{-16} \times 300}{(2.55 \times 10^{-8})^2 \times 4 \times 10^{-4}} = 1.6 \times 10^5 \text{ dynes/cm}^2 \quad (72)$$

Since the shear modulus is 4.6×10^{11} dynes cm^2, this corresponds to a strain of 3.5×10^{-7}. Furthermore, the potential well model is no longer adequate since the dislocations can be displaced by several atomic spacings and their ultimate displacement will be determined by the increase in length of the dislocation. Hence the nonlinear effect is consistent with consideration of the dislocation loop displacement as the cause of dissipation in a metal crystal.

It has been found that the larger the number of impurities, and hence the shorter the loop length ℓ, the larger is the stress required to make Q^{-1} vary non-linearly. This is in agreement with (71). Also since the attenuation is proportional to the number of dislocations $\bar{N}$ times the average loop length ℓ the effect of increased impurities is to lower ℓ and also the attenuation. On the other hand the hysteresis loss depends directly on the number of dislocations only, and should be independent of the impurity content. This appears to be borne out by the data for the pure single crystal, 99.99 pure, in Fig. 8 and for the commercial ma-

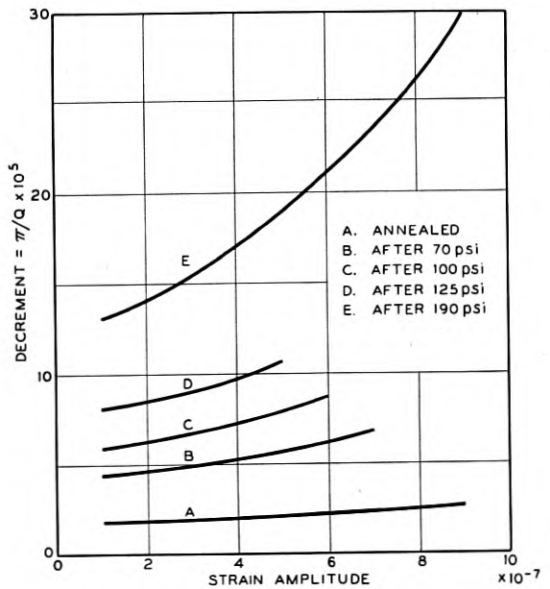


Fig. 15 — Internal friction in copper at room temperature as a function of cold work and strain amplitude (after Nowick).

terial in Fig. 11. An extension of the measurements of Fig. 8 to higher temperatures shows that the hysteresis attenuation reaches its maximum at 260°K and decreases for higher temperatures. This lowering of the peak from 350°K to 260°K, by the increased purity of the lead, is consistent with (70) for the Cottrell atmosphere temperature. Below the peak, however, when one multiplies the crystal value $C = 0.92$ by 5.2 and the polycrystal value C by 1.2, to reduce them to the values for the glide plane, the two values are within a few per cent of each other. Hence the effect of impurities appears to be small for the hysteresis term.

VISCOUS GRAIN BOUNDARY LOSSES

For polycrystals, another dislocation relaxation effect occurring in grain boundaries, contributes to acoustic loss at high temperatures and low frequencies. The boundaries[22] between grains of different orientations are regions of vacancies and considerable strain in the lattice. One of the simplest types of grain boundaries, called the Burgers boundary, is shown by Fig. 16. Here two slightly misoriented crystals are made to join by a

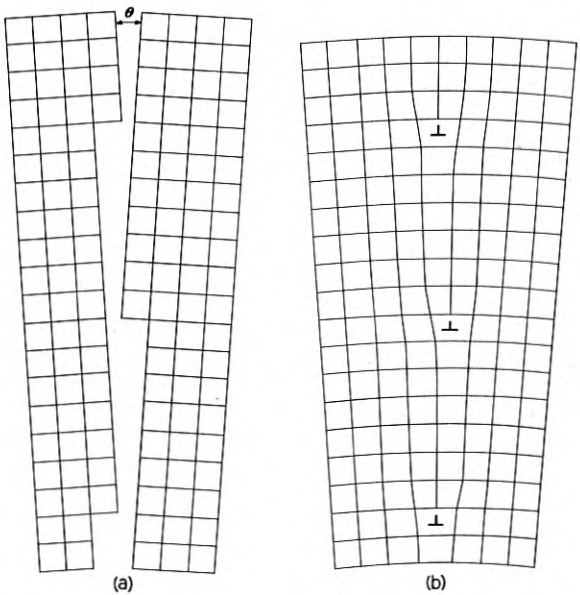


Fig. 16 — Burgers dislocation grain boundary with dislocations along cube edges (after Read).

[22] A complete discussion of dislocation grain boundary models is given by W. T. Read, Dislocations in Crystals, Chapters 11 to 14, McGraw-Hill Co.

series of edge dislocations whose spacings depend on the angle of misfit. Such a grain boundary can be made to glide on the application of a shearing stress, as has been shown by Parker and Washburn.[23] More complex boundaries occur when the plane of joining is not a cube edge. For such boundaries as shown in Fig. 17, two sets of dislocations are required to join the two differently oriented crystals. This type of boundary is called a tilt boundary. If the two grains are given a relative rota-

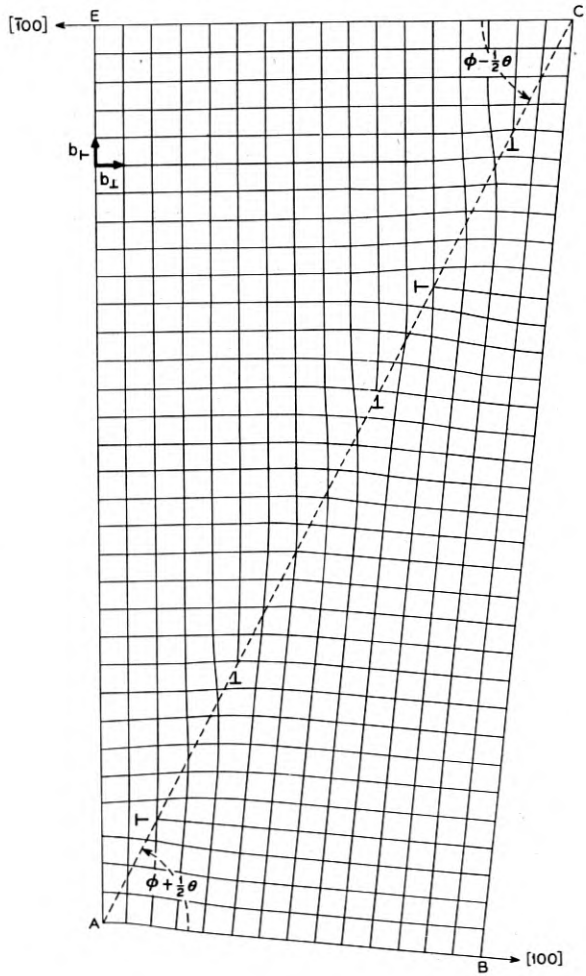


Fig. 17 — Read-Shockley dislocation tilt boundary with two sets of dislocations.

[23] Trans. A.I.M.E., **194,** pp. 1076–1078, 1952.

tion with respect to each other, the dislocations connecting them are of the screw type and such boundaries are called twist boundaries.

In the more complicated types of boundaries, the motion cannot be simple glide as in the boundary of Fig. 16 because, due to the various dislocation systems, glide requires sending one set of dislocations through the other and such a motion is strongly resisted. Such a boundary can move only if atoms diffuse into adjacent vacancies and this takes place only at high temperatures where thermal energies can overcome the high activation energies of diffusion. Hence when a cyclic stress is applied to a polycrystal, there is not enough time for a grain boundary to move appreciably so that the elastic constant of a polycrystal does not differ much from a single crystal at room temperature.

Fig. 18 shows the measurements of Ké[24] on the elastic and dissipation properties of polycrystals and single crystals of aluminum for a torsional vibration at 0.8 cycles, between 100°C and 450°C. The measurements

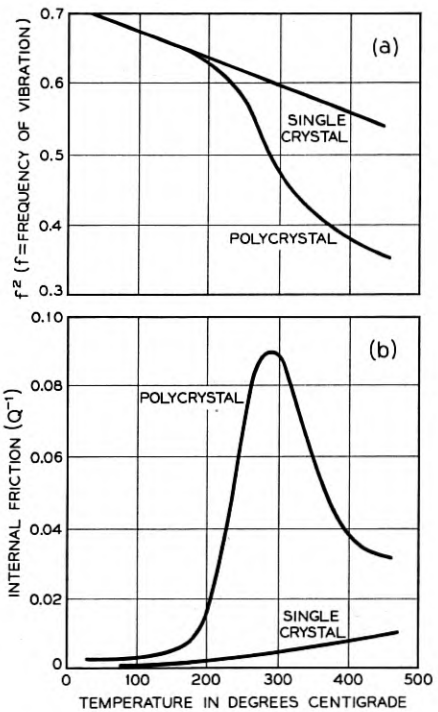


Fig. 18 — Elastic constants and internal dissipation in polycrystal and single crystal aluminum showing grain boundary effect (after Ké).

[24] T. S. Ké, Phys. Rev., **70**, p. 105A, 1946, and **71**, p. 533, 1947.

show a relaxation for the polycrystal but none for the single crystal. Both crystal types, however, show the hysteresis effect. By measuring the temperature shift of the maximum attenuation as the driving frequency is changed, Ké found an average activation energy of about 31,000 calories per mole, which is approximately 3.5 kilocalories under the self-diffusion constant of about 34.5 kilocalories per mole for aluminum in aluminum. This decrease of the average value is probably due to the average shearing stress exerted along the boundary. Shearing stress can lower the activation energy for diffusion, as has been shown by the measurement of diffusion bonding in solderless wrapped connections.[25] The measurements of Ké confirm the fact that grain boundaries can only move by diffusion of their components under a stress bias. With an activation energy of 31,000 calories per mole, the equation for the angular relaxation frequency becomes

$$\omega_0 = 4 \times 10^{12} e^{-31{,}000/RT} \tag{73}$$

The value of 4×10^{12} is close to what one would expect for the vibration of a molecule in its own potential well and agrees with the idea that the fundamental process is the motion of the domain wall by diffusion of its separate parts.

The breadth of the relaxation is greater than would be expected for a single activation energy. In fact it requires an activation energy distribution from 27.5 kilocalories to 34.5 kilocalories to account for the width. If we compare this with the measured[25] activation energy curve as a function of shear stress, shown in Fig. 19, this range can be explained by shear stresses ranging from 0 to 2000 pounds per square inch with an

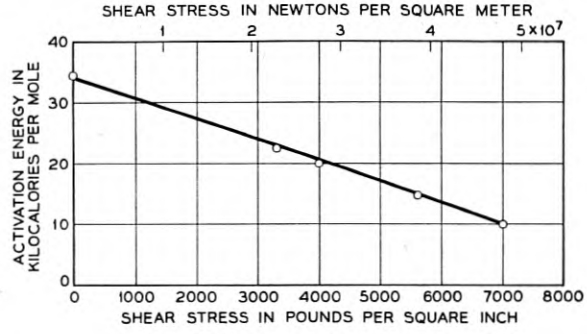


Fig. 19 — Lowering of activation energy of diffusion in aluminum by the application of a shearing stress.

[25] W. P. Mason and O. L. Anderson, Stress Systems in the Solderless Wrapped Connection and Their Permanence, B.S.T.J., **33**, pp. 1093–1111, Sept., 1954.

average value of 1,000 pounds per square inch. The strength of the relaxation shows that when the motion of the domain walls can adjust in a time much less than one period, about $\frac{1}{3}$ of the strain occurs by grain boundary motion and $\frac{2}{3}$ by elastic straining.

EFFECTS OF SMALL AND LARGE AMOUNTS OF COLD WORK ON ACOUSTIC ATTENUATION IN METALS

The curves of Bordoni, shown by Fig. 7 and Nowick in Fig. 15 show that small amounts of cold work can increase the attenuation due to the relaxation effect previously discussed. Since the activation energy does not change, the indications are that this increase is due to the production of more dislocations lying along minimum energy positions. An increase in dislocations by factors of 10 or more can occur by this process.

When large amounts of cold work are applied to metals, two other effects have been discovered by Köster[26] and his collaborators. The first effect, as shown in Fig. 20, consists of a temporary decrease of the

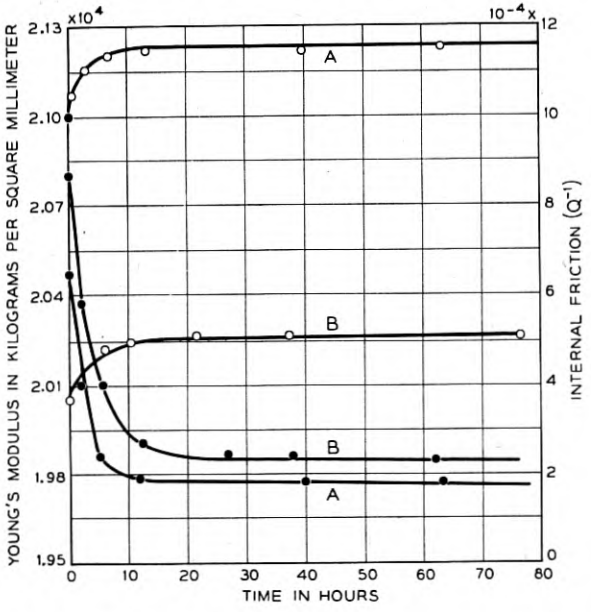


Fig. 20 — Internal friction (solid circles) and dynamic Young's modulus (open circles) of Armco iron, measured at 20°C as a function of time after deformation. Degree of cold drawing: A — 25 per cent, B — 80 per cent. (After Köster.)

[26] W. Köster and K. Rosenthal, Z. Metallkunde, **30,** p. 345, 1938; W. Köster, Arch. Eisenhüttenw, **14,** p. 271, 1940–41; and F. Förster and W. Köster, Naturwiss, **25,** p. 436, 1937.

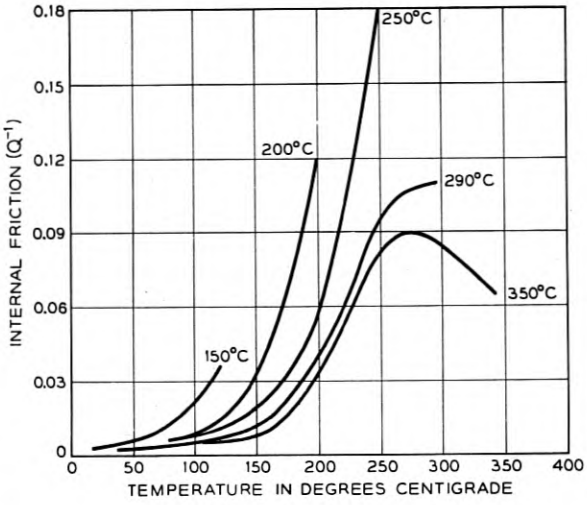


Fig. 21 — Dependence of internal friction of heavily cold worked aluminum on the temperature of measurement, after a series of anneals. The corresponding annealing temperature is marked at the top of each curve (after Ké).

Young's modulus and an increase in the internal friction which relax out in several hours at room temperature, i.e., far below a recrystallization temperature. For very high values of cold drawing, i.e., 80 per cent, a permanent change occurs in the value of Young's modulus until the metal is recrystallized. Although complete measurements[27] have not been made it appears that this effect is independent of amplitude for strains up to 10^{-5}, and is not strongly dependent on frequency or temperature from -100°C to 0°C. This effect has been called the Köster effect.[27]

Heavily deformed metals, particularly aluminum, show another effect called the "viscosity" effect after the Köster effect has disappeared. A residual internal friction is observed provided that the measurements are made at sufficiently high temperatures or low frequencies. Fig. 21 shows the internal friction of heavily worked aluminum after a series of anneals at the temperatures indicated in the figure. The most complete study of this effect is that of Ké[28] and Zener and Ké[28] in which low frequency (torsional pendulum) and static measurements (relaxation) were made. These curves show that the internal friction keeps on rising with temperature and shows no relaxation effect. At the recrystallization

[27] A. S. Nowick, Internal Friction and Dynamic Modulus of Cold Worked Metals, J. Appl. Phys., **25**, pp. 1129–1134, Sept., 1954.

[28] T. S. Ké, Trans. A.I.M.E., **188**, p. 575, 1950 and T. S. Ké and C. Zener, Symposium on the Plastic Deformation of Crystalline Solids, U. S. Office of Naval Research, Pittsburgh, 1950.

temperature, 300°C, this effect disappears and the normal grain boundary relaxation returns. This effect is not observable in brass or iron at these temperatures since the lower melting temperature of aluminum gives it a much lower activation energy.

While sufficient data for the Köster effect has not been obtained to make certain any interpretation in terms of dislocation theory, two possibilities may be mentioned. When the material is highly worked, large numbers of new dislocations are produced principally along slip bands in the interior of the crystal grains. From the absence of a temperature dependence of the Köster effect they must be of the zigzag types[29] which run across minimum energy positions, such as those illustrated in Fig. 5, in a zigzag manner as shown in Fig. 22. Such dislocations

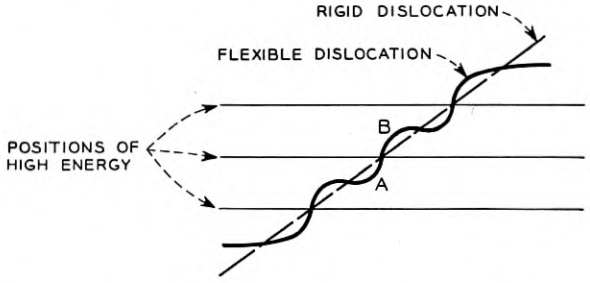


Fig. 22 — Form of zigzag dislocation that crosses lattice rows in a slip plane (after Cottrell).

are not bound by potential wells and their frequency of vibration is determined by their tension rather than by the limiting shearing stress for a potential well which was discussed in the first section. The equation for a stretched dislocation, as discussed by Koehler,[13] can be written in the form

$$m\frac{\partial^2 x}{\partial t^2} + B\frac{\partial x}{\partial t} - \frac{T}{\pi}\frac{\partial^2 x}{\partial y^2} = T_{13}b \tag{74}$$

where m, the mass per unit length, is $\pi\rho b^2$; B is a dissipation constant, T is the tension of the dislocation equal to μb^2, x is the displacement of the dislocation at a distance y from one end, and T_{13} the applied shearing stress. The natural frequency for a vibration loop is obtained by setting B and T_{13} equal to zero. The preferred shape of a vibration is then

$$x = A(y\ell - y^2) \tag{75}$$

[29] For a discussion of zigzag dislocations, see A. H. Cottrell, Dislocations and Plastic Flow in Crystals, p. 65, Oxford University Press, 1953.

which represents a zero displacement at $y = 0$ and $y = \ell$ with a parabolic shape in between. The average displacement for the dislocation is

$$\bar{x} = \frac{A}{\ell}\int_0^{\ell} (y\ell - y^2)\, dy = \frac{A\ell^2}{6} \tag{76}$$

Inserting this value in (74) and introducing the value $\partial^2/\partial t^2 = -\omega^2$, we have

$$\frac{\omega^2(\pi\rho b^2\ell^2)}{6} = \frac{2\mu b^2}{\pi} \tag{77}$$

Solving for the frequency we find

$$f = \frac{1}{2\pi\ell}\sqrt{\frac{12}{\pi^2}\frac{\mu}{\rho}} \tag{78}$$

When the zigzag dislocations are originally formed, some of them are probably relatively free from entanglement with other dislocations or with impurity atoms. In the course of time they gravitate towards other dislocations or atoms and form pinning points with them, thereby restricting their possible motion and the associated dissipation.

While they are free to vibrate, one possible cause of loss is the damped vibrations of the dislocations discussed by Koehler.[13] Koehler considers that the dislocations can be made to follow the applied stress and due to the damping constant B of (74) abstract energy from the vibration. To a first approximation, the loss caused by this mechanism is proportional to the first power of the frequency and the fourth power of the loop length.[13] Since no activation energy is involved, this loss should be independent of the temperature. As the loop length becomes smaller due to an increased number of pinning points, this loss rapidly decreases.

If the Köster loss is independent of the frequency, another possibility may be the thermal type of hysteresis loss, i.e., loss independent of frequency. For the case of free dislocations it is supposed that they can move in loops but the pinning points are not fixed points. Hence, energy abstracted from the mechanical vibration by the motion of the loop will not be coherent and therefore will not be returnable to the vibration. The loss calculated from such a mechanism will be

$$Q^{-1} = \frac{\pi b^2\rho\ell f N_0 V_s}{\pi\sqrt{\mu\rho}} \tag{79}$$

Inserting f from (78) we find

$$Q^{-1} = \frac{\sqrt{3}b^2 N_0 V_s}{\pi^2} \tag{80}$$

where N_0 is the total number of loops present in the unpinned dislocations. This should be independent of the temperature at the time of forming. As these free dislocations gradually become pinned by other dislocations or atoms, the number, N_0, decreases and the attenuation decreases while the value of Young's modulus increases, since the degree of displacement diminishes as the loops become pinned. Since $b \doteq 2.5 \times 10^{-8}$; $V_s \doteq 3.2 \times 10^5$ cm/sec for iron, Q^{-1} should be multiplied by 1.2 to transform from a Young's modulus strain along the axis to a shearing strain in the glide plane. Hence N_0 should have a value of

$$N_0 \doteq 1.3 \times 10^7 \tag{81}$$

Some idea of the effect of such a model on the elastic constants can be obtained by calculating the ratio of the plastic to elastic shear for a static force. From equations (74) and (76) the average displacement $\bar{x}$ of a dislocation loop is

$$\bar{x} = \frac{\pi T_{13} \ell^2}{12 \mu b} \tag{82}$$

The total plastic strain, due to the displacement of N_0 loops, is then

$$S_{13}{}^P = N_0 S b = N_0 \bar{x} \ell b = N_0 \frac{\pi T_{13} \ell^3}{12 \mu} \tag{83}$$

where S is the average area of a loop. To this we add the elastic strain so that

$$\frac{S_{13}{}^P + S_{13}{}^E}{T_{13}} = \frac{1}{\mu^E} + \frac{\pi N_0 \ell^3}{12 \mu} = \frac{1}{\mu} \tag{84}$$

Hence the difference between two constants is

$$\frac{\mu^E - \mu}{\mu \mu^E} \doteq \frac{\Delta \mu}{\mu \mu^E} = \frac{\pi N_0 \ell^3}{12 \mu} \quad \text{or} \quad \frac{\Delta \mu}{\mu} = \frac{\pi N_0 \ell^3 \mu^E}{12 \mu} \tag{85}$$

Using (46) to relate the change in elastic constant along the glide plane to the change in the elastic constant Y_0 for an isotropic material, we multiply the change in elastic constant of Fig. 20 by a factor of 1.2. Hence

$$\frac{\Delta \mu}{\mu} = 1.2 \times 0.01 \doteq \frac{\pi N_0 \ell^3}{12} \tag{86}$$

Since

$$b = 2.5 \times 10^{-8} \text{ cm} \tag{87}$$

we find

$$N_0 \ell^3 = 0.046; \qquad \ell \doteq 1.5 \times 10^{-3} \text{ cm} \tag{88}$$

where we have used the value of $N_0 = 1.3 \times 10^7$. This appears to be a possible value for a free dislocation loop not pinned down by other dislocations. According to Nowick[27] the permanent change in the elastic modulus for heavily worked material, which does not recover below the recrystallization temperature, stems from the lowering of the true Young's modulus due to the decrease in the average interatomic force constants resulting from the deformation.

As soon as the free dislocations become tied down by other dislocations, this source of dissipation and modulus change disappears. As shown in Fig. 21, however, there remains another source of dissipation which appears to be related to the temperature actuated hysteresis effect previously discussed. If we take the difference between the actual attenuation curves and the grain boundary relaxation effect, which should always be present, the added attenuation can be represented by the equation

$$Q^{-1} = Ce^{-(10,500/RT)} \tag{89}$$

where the constant C varies from

$$C = 2.0 \times 10^4 \text{ to zero} \tag{90}$$

as the annealing temperature goes from 125°C to the recrystallization temperature. Since this effect does not reach a peak value in the manner of the grain boundary relaxation effect, this cannot be a relaxation effect.

The mechanism considered here is that it is a temperature actuated hysteresis effect similar to that discussed previously except that the dislocations are not in a potential well and are bound by other dislocations rather than by impurity atoms. It is known that under conditions of severe deformations, each crystallite is broken into sub-grains by the production of slip bands. Hence the dislocations are present in these bands and in the course of time join into a network of dislocation loops. The slip bands are under a shearing stress equal to the limiting shearing stress of the crystal and under these conditions the data of Fig. 19 show that the activation energy for diffusion can be lowered to 10.5 kilocalories per mole. Hence the loss considered here is that due to thermal energy abstracted from the mechanical vibration by the incoherent energy of unpinned dislocation loops.

From (80) it is readily seen that the form of the loss equation should

be

$$Q^{-1} = \frac{\sqrt{3}b^2 V_s N_0 e^{-(10,500/RT)}}{\pi^2} \tag{91}$$

From the value of C in (90), the number of loops per cubic centimeter should be

$$N_0 \doteq 10^{15} \tag{92}$$

If we consider that the dislocations outline the form of an approximately regular mosaic structure, with $N_0 \ell^3 \doteq 1$; $\ell \doteq 10^{-5}$ cm; we find

$$N_0 \ell \doteq 10^{10} \text{ dislocations per sq cm} \tag{93}$$

which is a reasonable value for a hard worked material. In all probability, the dislocations concentrate in the slip bands and have a considerably higher density and smaller average loop length than given by (93). This type of loss does not show up in the same temperature range for brass and iron, as found by Köster, since the activation energy for these materials cannot be reduced to such a low value as for aluminum.

Magnetic Pulse Modulators

By K. J. BUSCH, A. D. HASLEY and CARL NEITZERT

(Manuscript received March 11, 1955)

The impetus for the development of magnetic pulse modulators for radar gear stems from the extreme reliability possible for magnetic devices. Descriptions of magnetic modulators developed for this purpose are given. Mathematical treatments of both the ac and dc charged series type magnetic modulators add to the understanding of core resetting in the ac case and reveal new areas of operation in the dc case, such as a possibility of voltage amplification and automatic core resetting. Means are described for obtaining very short pulses and for absorbing unwanted stored energy in parasitics following the output pulse. These means may also be applied to other pulse modulators. Also, a way is suggested for operating the cathode of the thyratron at ground potential in a dc charged magnetic pulse modulator.

INTRODUCTION

Increasing emphasis on the reliability of the components used in electronic equipment is leading component and system designers to greater use of magnetic devices to take advantage of their almost unlimited life. These devices have been used to replace limited life items, such as hydrogen thyratrons, electron tubes, etc., or to relegate such items to a part of the circuit where long life can be obtained by operation well below rating. Examples of the application of magnetic devices to improve the reliability of radar system modulators are the ac and dc magnetic pulse modulators to be described herein.

These radar modulators provide short-duration high-voltage high-current negative pulses to a magnetron which generates microwaves for the system. An essential function of the modulator, accordingly, is one of intermittently switching a high-voltage high-current source across a load. In a typical system, the duration of the switching action might be one microsecond or less, and a thousand such switches a second would be required. A typical voltage might be 30,000 volts and the peak current

quency or a sub-multiple of it. If the supply is dc and S_1 is omitted, the repetition rate is equal to twice the frequency at which the charging inductor, L, and the capacitance of the pulse network resonate. If the switch is present, repetition rates which are less than this may be used. During the charging period of the network, S_2 is open and the network acts simply as a capacitor, C. When the network is fully charged and the charging current has fallen to zero, S_2 is closed. The network discharges through the pulse transformer into the magnetron load. Switch, S_2, is then opened and the cycle is repeated.

The switch, S_2, must carry a large peak current during the discharge of the network. This switch may be a spark-gap, a thyratron or a thyractor. In order that a thyractor may be used, its core must be unsaturated during the charging of the network and must saturate at the instant the network is due to discharge. Also, after the thyractor functions as a switch, its core must be reset to be ready for the next pulse.

This circuit, with a thyractor, would be a practical one except for the fact that the network must discharge through the saturated reactance of the thyractor. Since this reactance becomes part of the network during discharge, it limits the shortness of the pulse obtainable. As a practical matter, this reactance cannot be made arbitrarily small since the cross-sectional area of the core, the flux density and the number of turns of wire on the winding must satisfy the relation

$$NAB = \int e\,dt \tag{3}$$

during the charging period without prematurely saturating the core.

To avoid this difficulty, a step-by-step method of charging the network is used. The saturation flux linkages of the thyractors may be made smaller with each step providing a large enough ratio of saturated to un-

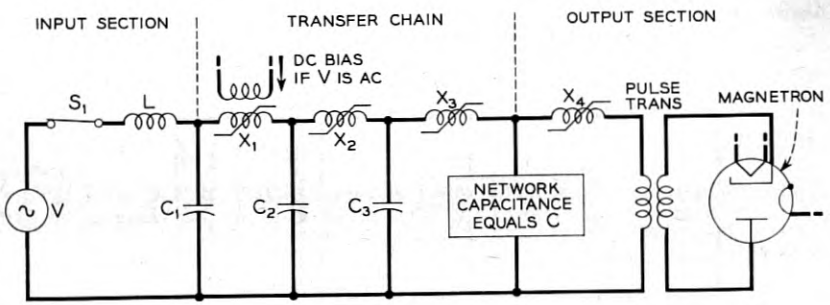


Fig. 3 — Basic series type magnetic pulse modulator.

saturated inductance is provided in the thyractor. Fortunately, the permalloys are ideal for this use and three or four thyractors are all that are needed to cut the charging time of the network from the order of thousands of microseconds to the order of a microsecond. Referring to Fig. 3, the inductor, L, has the same function as it had in Fig. 2, and, neglecting losses, the capacitors, C_1, C_2 and C_3 each may have a value equal to the capacitance of the network. The thyractors X_1, X_2, X_3 and X_4 have descending values of saturation flux linkages. With this circuit, when the first capacitor is charged to its peak voltage, X_1 saturates and connects C_2 across C_1 through the saturated inductance of X_1. It will be shown later that all the energy stored on C_1 will be transferred to C_2 in a time equal to one-half cycle of the resonant frequency of the saturated inductance of X_1 and the capacitance of C_1 and C_2 in series. Once C_2 is charged, X_2 saturates and the charge on C_2 is transferred to C_3 in a shorter time than the previous transfer since the saturated inductance of X_2 is less than that of X_1. This process continues until the network is charged. The time of charge of the network is so fast that thyractor, X_4, needs only a few turns and its saturated inductance usually ends up being about that of a section of the network.

It is seen from the above that the thyractors are designed to have descending values of saturated inductance as the energy is stepped from one capacitor to the next, until a value is reached in the last thyractor, X_4, that is small enough to be incorporated into the pulse network. Under this circumstance, the switching action of the last thyractor may be considered as ideal. Referring again to Fig. 3, it is the section of the circuit labeled "Transfer Chain" that provides this desired decrease in charging time of the network or, in other words, pulse shortening action. This action and the way the thyractor cores are reset after each pulse will be given detailed attention in the Parts that follow.

Part II — AC-Charged, Series-Type, Magnetic Pulse Modulator

The circuit diagram of a typical ac charged modulator is shown in Fig. 4. Power transformer, $TR1$, transforms the voltage available from the source to the value best suited to the transfer chain. The linear reactor L_i may be on either the primary or the secondary side of the power transformer but placing it on the primary side generally results in a more economical design. The dc bias shown on the first thyractor is necessary if pulses of one polarity are to be obtained. Bias current would ordinarily be obtained from the ac source by means of a dry-disk rectifier. The linear inductance in the bias circuit decouples the bias supply from the first

thyractor and will be assumed large enough to keep the bias current constant.

For purposes of analysis, the modulator will be divided into an input section, a transfer chain and an output section. The input section, up to and including the first thyractor, will be analyzed first. It will be shown how this portion of the circuit produces one negative current pulse through thyractor X_1 for each cycle of the ac source. The action of a section of the transfer chain in converting this current pulse into a shorter current pulse will then be explained. A more detailed explanation of the transfer chain is included in Part III. With this as background, the special requirements imposed on the thyractor core material will be discussed. The output section will then be analyzed in detail and a method given for producing shorter pulses than are possible with the circuit of Fig. 4. Finally, the core resetting action, which automatically takes place between main pulses, will be analyzed.

In the actual design of a modulator, the effect of dissipation must be considered. This may be done by an approximate method, as given in Part III, or exact numerical computations may be made. However, in order to make the main action of the modulator more apparent, dissipation will be neglected in this Part.

Thyractors will be assumed to have zero reluctance cores when unsaturated. The core will be assumed to saturate suddenly when the core flux reaches its saturation value Φ_s, after which the thyractor will have a constant saturated inductance L. In order to satisfy these assumptions, the idealized hysteresis loop must have the form shown in Fig. 5.

INPUT SECTION

If the power source and the charging inductance are referred to the secondary of the power transformer, the circuit of the first section will

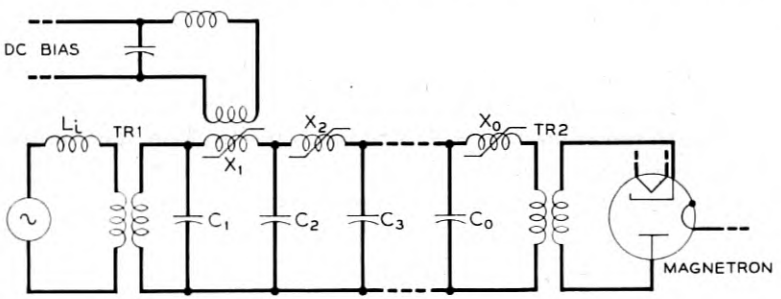


Fig. 4 — Circuit diagram of a typical ac charged series-type modulator.

be that shown in Fig. 6. The apparent source, v_i, will be assumed to have no internal impedance and to provide a sinusoidal voltage

$$v_i = V_i \sin \omega t_i \tag{4}$$

where $\omega/2\pi$ is the pulse repetition rate required at the magnetron. A set of initial conditions will be assumed for time $t_i = 0$ and will be justified by showing that the circuit returns to these conditions at the end of one cycle. These conditions are $i_i(0) = i_1(0) = I_b$, $v_1(0) = v_2(0) = 0$ and $\varphi_1(0) = -\Phi_{1s}$. In these relations, i_i, i_1, v_1 and v_2 are the instantaneous currents and voltages labeled on the circuit of Fig. 6, I_b is a constant component of i_i and i_1 which is to be evaluated, φ_1 is the core flux of thyractor X_1 and Φ_{1s} is the saturation value of φ_1. It is further assumed that all other capacitors in the chain are initially discharged and that all other thyractors are saturated in the direction of positive flux. These latter assumptions are approximate as will be seen when the resetting action is considered.

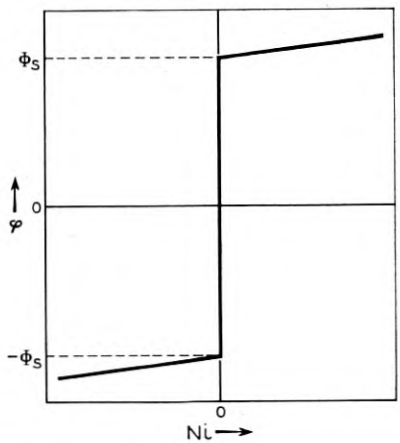


Fig. 5 — Idealized hysteresis loop upon which the theoretical analysis is based.

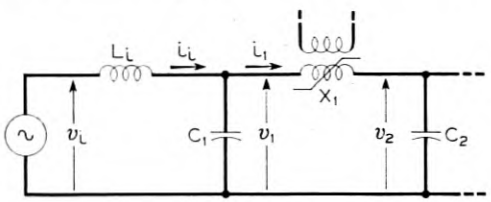


Fig. 6 — Input section of an ac charged modulator. This circuit produces one negative current pulse through X_1 for each cycle of the ac source.

The differential equations for the circuit are

$$V_i \sin \omega t_i = L_i \frac{di_i}{dt_i} + v_1 \tag{5}$$

and

$$i_1 = i_i - C_1 \frac{dv_1}{dt_i} \tag{6}$$

With the initial conditions assumed, and with L_i chosen to make $\omega^2 L_i C_1 = 1$, the solutions to these equations are

$$i_i = I_b + \frac{V_i}{2\omega L_i} \omega t_i \sin \omega t_i \tag{7}$$

and

$$v_1 = \frac{V_i}{2} (\sin \omega t_i - \omega t_i \cos \omega t_i) \tag{8}$$

In solving for these equations, it is assumed that i_1 remains constant at I_b. This is justified by the fact that v_1 becomes positive immediately after time $t_i = 0$ causing the flux φ_1 of the first thyractor to rise into the unsaturated region. The high unsaturated inducatance of the thyractor then keeps the current constant.

In the unsaturated condition the net mmf acting on the thyractor core must be zero, therefore the mmf of the main thyractor winding, due to the current I_b, must be balanced by the mmf of the bias winding. Since the bias current can be arbitrarily adjusted, the value selected should give optimum conditions in the remainder of the circuit. The proper value of I_b is

$$I_b = \frac{V_i}{2\omega L_i} \tag{9}$$

This value of I_b makes the average value of i_i equal to zero and thus avoids dc saturation in the core of the power transformer. It has the additional advantage of making the rms source current a minimum. This gives maximum source power factor and minimum copper loss in the charging choke and the power transformer. The value of the source power factor corresponding to this value of I_b is slightly over 0.96 and the rms value of i_i is 1.15 $V_i/\omega L_i$.

During the time that (7) and (8) apply, the current I_b flows through all the thyractors and the primary winding of the pulse transformer. Since none of these, except the first thyractor, have any core bias, their

cores are all held at positive saturation which is their reset condition. If the windings are dissipationless, all capacitors, except C_1, are short circuited.

With no voltage across C_2, the voltage v_1 is applied across the first thyractor. This causes its core flux φ_1 to vary in accordance with the equation

$$\begin{aligned}\varphi_1 &= \frac{1}{N_1}\int v_1\, dt_i \\ &= \frac{V_i}{2\omega N_1}(2 - 2\cos\omega t_i - \omega t_i \sin\omega t_i) - \Phi_{1s}\end{aligned} \tag{10}$$

in which N_1 is the number of turns on the main winding of X_1.

Equations (7), (8) and (10) are plotted in Fig. 7 for two cycles of the source. The source voltage v_i is included for reference. The charging current of capacitor C_1 equals $i_i - I_b$. This current is a sinusoid whose amplitude increases in direct proportion to time. During the first half cycle, C_1 charges to a positive maximum voltage of $\pi V_i/2$. The capacitor current then reverses and C_1 discharges, becoming completely discharged when ωt_i equals 4.49 radians. During the interval $\omega t_i < 4.49$, v_1, the voltage across X_1, is positive, causing φ_1 to increase. At $\omega t_i = 4.49$, φ_1 passes through a maximum value

$$\varphi_{1\max} = \frac{3.41\ V_i}{\omega N_1} - \Phi_{1s} \tag{11}$$

To avoid saturation of the core at this time, the inequality

$$2\omega N_1\Phi_{1s} > 3.41\ V_i \tag{12}$$

must be satisfied. Fig. 7 shows $\varphi_{1\max} = \Phi_{1s}$. This calls for the least amount of core material but is not an essential condition. Ordinarily a small margin of safety should be allowed.

For values of $\omega t_i > 4.49$, C_1 charges in the negative direction. The capacitor voltage reaches a negative maximum value of $-\pi V_i$ at the end of the cycle. During this time, flux φ_1 decreases and, since the integral of v_1 over a complete cycle is zero, φ_1 returns to its initial value, $-\Phi_{1s}$, at the end of the cycle. The core of thyractor X_1 then saturates and C_1 discharges through the saturated inductance into capacitor C_2. Under ideal conditions, C_1 discharges completely in a very small fraction of one cycle of the source, during which time i_i is held constant by the charging inductor. At the end of the discharge, $i_i = i_1 = I_b$, $v_1 = 0$ and $\varphi_1 = -\Phi_{1s}$ which are the same as the assumed initial conditions. Capacitor C_1 then starts to recharge and the cycle is repeated.

If (12) were made an equality instead of an inequality, the thyractor core would saturate and C_1 would discharge at the end of each half cycle. Alternate pulses applied to the chain, would be of reverse polarity and a rectifier action would be necessary at some point. In this case no bias would be required on the first thyractor and current I_b could be zero. However, the core resetting action, to be described for the chain thyractors, would be upset and those thyractors beyond the rectifier would require some other means of resetting.

It is not possible to increase the charging period beyond one full cycle

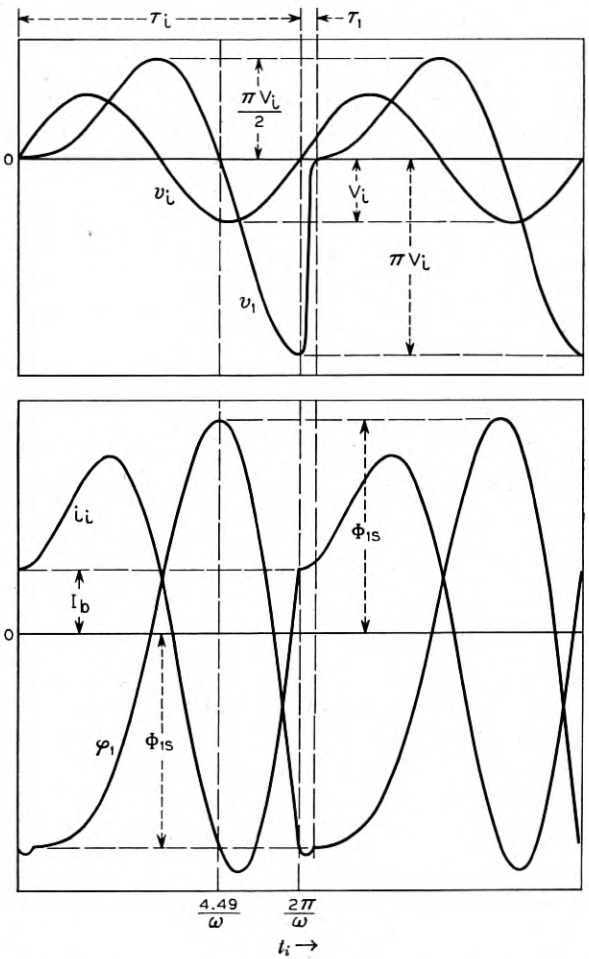


Fig. 7 — Waveforms of source voltage v_i, first-capacitor voltage v_1, source current i_i and first thyractor core flux ϕ_1 for two complete cycles of the ac source.

if the initial thyractor core flux is to be $-\Phi_{1s}$. This can be seen from (10) which gives a negative maximum value in φ_1 of $-\Phi_{1s} - 2.96\ V_i/\omega N_1$ at $\omega t_i = 7.72$ radians. This is not possible since the thyractor saturates and C_1 discharges when $\varphi_1 = -\Phi_{1s}$. The fact that φ_1 is changing at its maximum rate at $\omega t_i = 2\pi$ is advantageous though in that it makes the design of the thyractor core less critical.

Since the voltage v_1 is also across the secondary of the power transformer, the core flux of this transformer will be given by an equation similar to (10). If the average value of i_i is zero, the average value of this flux will also be zero. The transformer core must therefore be designed for a peak core flux of

$$\varphi_{T1\max} = \frac{3.41 V_i}{\omega N_s} \tag{13}$$

in which N_s is the number of secondary turns. The root-mean-square value of i_i, which is also the transformer secondary current, is $1.15\ V_i/\omega L_i$.

The linear charging reactor must be designed for a maximum energy storage of

$$\tfrac{1}{2} L_i i_{i\max}^2 = 1.82\ V_i^2 C_1 \text{ joules} \tag{14}$$

TRANSFER CHAIN

The need for and the basic action of the transfer chain have been discussed in Part I. A general analysis of a single section of the chain and of the chain as a unit is given in Part III. It will suffice here to set down the special conditions which apply in the case of ac charging.

Figure 8 shows the 1st and 2nd sections of the chain. As previously stated thyractor X_1 saturates once each cycle allowing C_1 to discharge into C_2. This occurs when v_1 has its maximum negative value. In the case of the ac charged modulator, it is desirable that C_1 be completely discharged when X_1 becomes unsaturated. In the absence of dissipation,

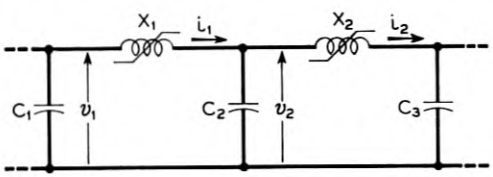


Fig. 8 — Circuit diagram of the first two sections of the transfer chain. The chain serves to produce successively shorter current pulses through each thyractor.

this requires that $C_1 = C_2$. In this case, all of the energy stored in C_1 will be transferred to C_2 and the peak value of v_1 will be the same as the peak value of v_2, that is $-\pi V_i$.

During the discharge of C_1 into C_2, with a negligibly small error,

$$i_1 = I_b - \frac{\pi V_i}{\omega_1 L_1} \sin \omega_1 t_1 \tag{15}$$

$$v_1 = -\frac{\pi V_i}{2} (1 + \cos \omega_1 t_1) \tag{16}$$

$$v_2 = -\frac{\pi V_i}{2} (1 - \cos \omega_1 t_1) \tag{17}$$

and since v_2 also appears across thyractor X_2

$$\varphi_2 = \frac{-\pi V_i}{2\omega_1 N_2} (\omega_1 t_1 - \sin \omega_1 t_1) + \Phi_{2s} \tag{18}$$

In these equations i_1, v_1 and v_2 are the current and voltages designated in Fig. 8, φ_2 is the core flux of X_2, N_2 is the number of turns on X_2, L_1 is the saturated inductance of X_1, t_1 is the time measured from the instant X_1 saturates and ω_1 is resonant angular frequency of the discharge circuit.

Figure 9 shows curves of v_1, v_2, i_1 and φ_2 plotted versus t_1. The current pulse in i_1 is one half of a cycle of a sine wave superimposed on the constant value I_b. The duration of this pulse is $\tau_1 = \pi/\omega_1$. This circuit transfers the stored energy from C_1 to C_2. During this interval φ_2 varies from positive to negative saturation. If thyractor X_1 is designed to satisfy the equation

$$4N_2\Phi_{2s} = \pi V_i \tau_1 \tag{19}$$

φ_2 will reach negative saturation at the same time that $v_1 = 0$ and $v_2 = -\pi V_i$. Since at this time $i_1 = I_b$, X_1 will become unsaturated after which i_1 will be held at I_b. When X_2 saturates, C_2 starts to discharge into the third capacitor. The current pulse through X_2 is shown as i_2 in Fig. 9. The duration of this pulse is

$$\tau_2 = \frac{\pi}{\omega_2} = \pi \sqrt{\frac{L_2 C_2 C_3}{C_2 + C_3}} \tag{20}$$

in which $C_3 = C_2$ for optimum operation of the dissipationless ac charged modulator.

The current pulse thus travels down the chain, becoming greater in magnitude and smaller in duration in each successive section. The ratio

of τ_1 to τ_2 can be shown to be approximately proportional to the saturation flux density and the square root of the core volume of thyractor X_2. It is also inversely proportional to the square root of the pulse energy. Theoretically this ratio can be made large at pleasure but it is generally more economical to use several sections to obtain the desired reduction in pulse duration.

When dissipation is considered the capacitors should be graded in size and the peak capacitor voltage should increase toward the input of the chain. A simple approximate method of including dissipation consists of raising the peak voltage of say capacitor C_1 so that it will have an excess of stored energy equal to the estimated copper loss in X_1 and core loss in X_2. An alternate approximate method is given in Part III.

Neither of the above methods give a satisfactory account of the core loss in that portion of the chain where the pulse becomes very short. Unless the product of C_2 and the shunt coreless-resistance of X_2 is large, the analysis of Fig. 8 should be made on an exact basis. This involves 3rd

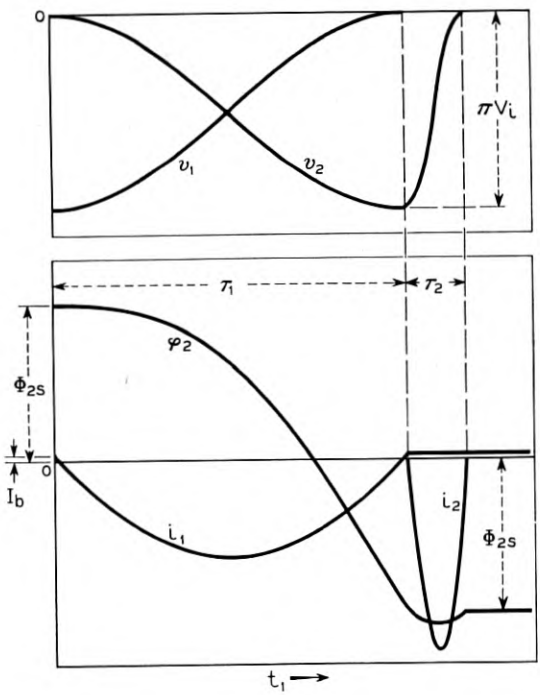


Fig. 9 — Curves illustrating the reduction in the current-pulse width from the first to the second thyractor. Capacitor voltages and thyractor core flux are also shown. These curves are typical of all of the transfer sections.

order differential equations and therefore should be done numerically for individual cases.

CORE CONSIDERATIONS

The cores of thyractors require special consideration. It has been previously assumed that the thyractor cores have zero reluctance when unsaturated and become suddenly and completely saturated at a certain value of flux. The practical approximation of this is a core that reaches complete saturation at a small value of magnetizing force. This inherently requires that the unsaturated permeability be high. Examples of core materials that meet this requirement are the nickel-iron alloys such as deltamax, molybdenum permalloy and supermalloy. Others of equal importance exist but they will not be discussed here. The chief advantage of deltamax over the other two is its high saturation flux density. Its use is therefore indicated when the saturating time is long, as in the case of the input thyractor. The fact that its hysteresis loop has a striking rectangular appearance is, in itself, no great virtue for this application.

For thyractors following the first, the swing in flux linkage is much smaller and the high saturation flux of deltamax is not required. In all thyractors, the flux swings from saturation in one direction to saturation in the other and back again for each pulse applied to the magnetron. With respect to core loss, supermalloy is superior to both deltamax and molybdenum permalloy in that it has a higher resistivity and a much lower hysteresis loss. Supermalloy is also suitable for use in the first thyractor in certain cases where the repetition rate is high.

In order to utilize the advantages of these high-permeability materials, the thyractor cores must be of the gapless wound-tape type. In most cases, the tape thickness will be made smaller as the output of the modulator is approached and may reach values below one mil.

OUTPUT SECTION

When a rectangular voltage pulse of relatively long duration is required for the magnetron, the output capacitor, C_0 of Fig. 4, is replaced by a line-type pulse-forming network having a total capacitance equal to the normal value of C_0. During the relatively long charging period, the network acts like a capacitance C_0, whereas during the relatively short discharge period it acts as a pulse-forming network. The saturated inductance of the output thyractor is in series with the network during discharge. This adds to the design complication. Either the thyractor saturated inductance must be made small enough to cause negligible

distortion in the pulse shape or the network must be designed to include the inductance as part of the network. In the former case, the reduction in pulse width by the output section of the modulator is appreciably less than could otherwise be obtained.

As the required magnetron pulse width decreases, the output thyractor, the pulse transformer and the shunt capacitance cause deterioration of the pulse shape to the extent that a pulse-forming network can hardly be justified. If a capacitor is used instead of a network, the output section can be designed as a unit in accordance with the following analysis.

Fig. 10 shows the output section with all elements referred to the secondary of the pulse transformer. In this circuit L_0 represents the sum of the transformer leakage inductance and the saturated inductance of the output thyractor, capacitance C_M is the sum of the transformer capacitance and the magnetron capacitance. Current i_M is thus the total

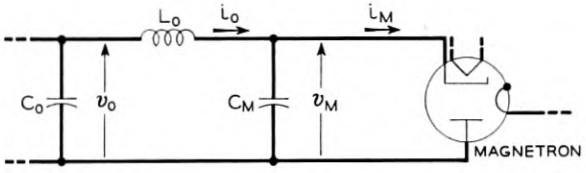


Fig. 10 — Output section with a magnetron load. All elements are referred to the secondary of the pulse transformer.

magnetron current less the current required to charge the magnetron capacitance.

For the period under consideration, v_0 has an initial value $-V_0$ and the initial values of i_0, v_M and i_M are all zero. The thyractor saturates at the beginning of the period. Capacitor C_0 discharges through L_0 into C_M until v_M reaches the magnetron firing voltage, $-V_M$. If the magnetron is approximated by a dissipationless diode in series with a biasing emf, v_M then remains constant while C_0 continues to discharge through the magnetron.

During the charging of C_M

$$i_0 = -\frac{V_0}{\omega_0 L_0} \sin \omega_0 t_0 \tag{21}$$

where

$$\omega_0^2 L_0 \frac{C_0 C_M}{C_0 + C_M} = 1 \tag{22}$$

$$v_0 = -\frac{V_0 C_0}{C_0 + C_M} - \frac{V_0 C_M}{C_0 + C_M} \cos \omega_0 t_0 \tag{23}$$

and

$$v_M = -\frac{V_0 C_0}{C_0 + C_M} (1 - \cos \omega_0 t_0) \tag{24}$$

in which t_0 is measured from the instant that X_0 saturates.

When $v_M = -V_M$ at $t_0 = \tau_0$, the magnetron fires. At this time

$$v_M(\tau_0) = -V_M = -\frac{V_0 C_0}{C_0 + C_M} (1 - \cos \omega_0 \tau_0) \tag{25}$$

from which

$$\cos \omega_0 t_0 = 1 - \frac{V_M}{V_0} \frac{C_0 + C_M}{C_0} \tag{26}$$

In addition,

$$v_0(\tau_0) = -V_0 + V_M \frac{C_M}{C_0} \tag{27}$$

and

$$i_0(\tau_0) = \frac{-V_0}{\omega_0 L_0} \left[\frac{2V_M(C_0 + C_M)}{V_0 C_0} - \frac{V_M^2 (C_0 + C_M)^2}{V_0^2 C_0^2} \right]^{1/2} \tag{28}$$

After the magnetron fires, v_M remains constant at $-V_M$ and

$$i_0 = i_M = \frac{v_0(\tau_0) + V_M}{\sqrt{L_0/C_0}} \sin \frac{t_0 - \tau_0}{\sqrt{L_0 C_0}} + i_0(\tau_0) \cos \frac{t_0 - \tau_0}{\sqrt{L_0 C_0}} \tag{29}$$

If the magnetron current pulse has a width τ_M at its base, and it is assumed that C_0 is completely discharged and that $i_0 = 0$ at the end of the magnetron current pulse, then from (29)

$$\frac{v_0(\tau_0) + V_M}{\sqrt{L_0/C_0}} \sin \frac{\tau_M}{\sqrt{L_0 C_0}} + i_0(\tau_0) \cos \frac{\tau_M}{\sqrt{L_0 C_0}} = 0 \tag{30}$$

Also, since $v_0 = 0$ the voltage across L_0 must equal V_M, or

$$L_0 \left. \frac{di_0}{dt_0} \right|_{t_0 = \tau_0 + \tau_M} = V_M$$

that is

$$[v_0(\tau_0) + V_M] \cos \frac{\tau_M}{\sqrt{L_0 C_0}} - \sqrt{\frac{L_0}{C_0}}\, i_0(\tau_0) \sin \frac{\tau_M}{\sqrt{L_0 C_0}} = V_M \tag{31}$$

Combining (30) and (31) and substituting the values of $v_0(\tau_0)$ and $i_0(\tau_0)$ from (27) and (28) respectively and simplifying gives

$$\frac{C_0 V_0^2}{2} + \frac{C_M V_M^2}{2} = C_0 V_0 V_M \tag{32}$$

The first term on the left is the energy initially stored in C_0 and the second term is the energy remaining in C_M at the end of the magnetron current pulse. The difference between these energies is the magnetron pulse energy W_M, that is

$$\frac{C_0 V_0^2}{2} - \frac{C_M V_M^2}{2} = W_M \tag{33}$$

When C_M is negligibly small these equations are identical since for this case $V_0 = 2V_M$. In the case being considered here, however, they may be solved for V_0 and C_0 giving

$$V_0 = V_M \frac{2W_M + C_M V_M^2}{W_M + C_M V_M^2} \tag{34}$$

and

$$C_0 = \frac{1}{V_M^2} \frac{(W_M + C_M V_M^2)^2}{2W_M + C_M V_M^2} \tag{35}$$

If the chain capacitance is given, (35) fixes the pulse transformer turns ratio after which (34) fixes the chain voltage. On the other hand, if the chain voltage is given, these equations determine the turns ratio and the chain capacitance.

It can also be shown that

$$\cos \frac{\tau_M}{\sqrt{L_0 C_0}} = 1 - \frac{V_0}{V_M} + \frac{C_M}{C_0} \tag{36}$$

and

$$\cos \sqrt{\frac{C_0 + C_M}{L_0 C_0 C_M}}\, \tau_0 = 1 - \frac{V_M}{V_0} \frac{C_0 + C_M}{C_0} \tag{37}$$

from which L_0 and τ_0 are found.

Equations (34) through (37) assume that the magnetron pulse energy and pulse width are specified and that for a given magnetron, its shunt capacitance and firing voltage are known. If these quantities are not known, some other form of the equations may be more useful.

The shape of the magnetron current pulse depends upon the fraction

of the total energy left in C_M. Fig. 11 shows curves of i_0, i_M and v_M versus t_0 measured from the instant that the output thyractor saturates. During time τ_0, capacitor C_M charges to the firing voltage of magnetron. During time τ_M the magnetron conducts and $i_0 = i_M$. Two curves are shown for i_M during time τ_M. It is seen that if the energy stored in C_M is less than 30.9 per cent of the magnetron pulse energy, the magnetron current pulse is a better approximation of a rectangular pulse. In the terminal case where $C_M = 0$, i_M is a complete half cycle of a sine curve.

For magnetron pulse widths less than 0.1 microsecond it becomes very difficult to make $C_M V_M^2 < .618 W_M$ and as a result the magnetron current pulse deteriorates. In addition a large fraction of the energy is left in C_M at the end of the pulse and must be dissipated in the circuit. This reduces the overall efficiency of the modulator. A definite improvement can be obtained by separating the transformer capacitance from the magnetron capacitance by placing the output thyractor on the secondary side of the pulse transformer. Figs. 10 and 11 still apply, but now C_M is the total shunt capacitance across the magnetron and C_0 is the transformer capacitance plus any added capacitance that may be required. The values of V_0, C_0, L_0 and τ_0 are again given by (34), (35), (36) and (37) respectively. The pulse transformer can be designed so that its capacitance is equal to the required value of C_0 or capacitance can be added across the secondary to make the total equal to C_0. The leakage inductance of the transformer is considered as a part of the saturated inductance of the previous thyractor.

An alternate point of view is to consider the transformer capacitance

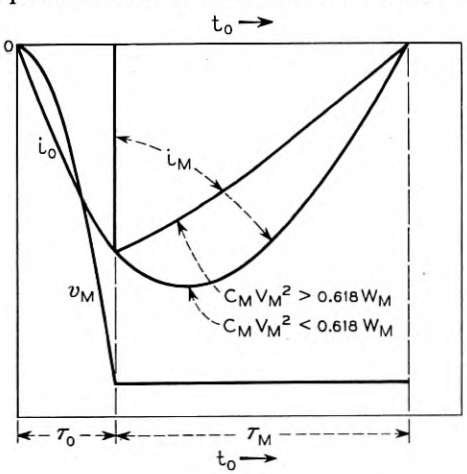


Fig. 11 — Waveforms of the magnetron current and voltage showing the effect of shunt capacitance upon the shape of the magnetron current pulse.

and leakage inductance, capacitance C_0 and the saturated inductance L_0 of the output thyractor to constitute a pulse forming network.

Placing the output thyractor on the secondary side of the pulse transformer results in a more reasonable value of L_0 unless a very high chain voltage is used. For chain voltages of the order of 10,000 volts and with very short magnetron pulses, the value of L_0 required on the primary of the pulse transformer may be a fraction of a microhenry. A much larger value is required on the secondary side so that stray inductance is far less important.

A disadvantage of placing the last thyractor on the secondary side of the pulse transformer is that the thyractor and any added capacitance across the transformer secondary must be insulated for a very high voltage, sometimes as high as 70,000 volts.

The core of the output thyractor presents special problems for very short pulses. In order to keep eddy current loss from being excessive core tape as thin as 0.25 mil may have to be used. Because of the fragility of such thin core material and the high voltage encountered, a toroidal core mounted in a core box of a high dielectric strength material such as teflon is generally desirable.

As previously mentioned, a considerable amount of energy is left in the magnetron capacitance. This energy must be dissipated in the modulator between main pulses. The result may be a considerable amount of ringing which may or may not have undesirable effects on the magnetron. In Part III the use of a damping thyractor is described which discharges C_M at the end of the magnetron pulse.

CORE RESETTING

Consideration of the first section has shown that the core flux of the first thyractor returns to its initial value at the end of each cycle. No further consideration of the resetting of this core is required. In the case of other thyractors, however, the cores must be at positive saturation at the beginning of the pulse and are left at negative saturation at the end of the pulse. Between pulses the cores must be reset to positive saturation. This resetting action is provided by the component I_b of i_i which flows even though the first thyractor is unsaturated.

The exact analysis involves long, complicated equations which add very little to a working knowledge of the modulator. To avoid including so much detail, the present analysis will be largely descriptive and only the important terms of the equations will be given.

Resetting of the second thyractor core starts at the end of the main-pulse discharge of C_2. At this time C_3 is fully charged and C_2 is dis-

charged. The voltage of C_3 is therefore applied across the second thyractor and its polarity is in the resetting direction. The resetting action of this voltage is small, if C_3 is allowed to discharge immediately, and will be neglected.

The main resetting occurs as a result of current $i_i = I_b$, which is given by (9), during the time that the first thyractor core is unsaturated. If time is measured from the instant C_2 completes its discharge of the main pulse, the voltage v_2 rises linearly according to the equation

$$v_2 = \frac{I_b t}{C_1} = \frac{\omega V_i}{2} t = \pi V_i \frac{t}{\tau_i} \tag{38}$$

Since, except for the very short time during which C_3 is discharging, $v_3 = 0$, v_2 is applied across the second thyractor and causes its core flux to rise in the positive direction in accordance with the equation

$$\varphi_2 = \frac{1}{N_2} \int v_2 \, dt = \frac{\omega V_i t^2}{4N_2} - \Phi_{2s} \tag{39}$$

Using (19), this becomes

$$\varphi_2 = \Phi_{2s} \left[\frac{2t^2}{\tau_i \tau_1} - 1 \right] \tag{40}$$

in which τ_i is the charging time and τ_1 is the discharging time of C_1.

When $t = \sqrt{\tau_i \tau_1}$, $\varphi_2 = +\Phi_{2s}$ and $v_2 = \pi V_i \sqrt{\tau_1/\tau_i}$. Assuming $\tau_i/\tau_1 = 30$ to be a typical value, this time is 0.183 τ_i and the voltage is 0.183 πV_i. When $\varphi_2 = \Phi_{2s}$, the second thyractor saturates and C_2 discharges into C_3 raising the voltage v_3 to $+\pi V_i \sqrt{\tau_1/\tau_i}$.

Fig. 12 shows curves of v_2, φ_2, v_3 and φ_3 plotted versus time measured from the end of the main discharge of C_2. Starting with the main discharge of C_1, from a to b C_1 discharges into C_2, v_2 goes negative to $-\pi V_i$ and φ_2 varies from $+\Phi_{2s}$ to $-\Phi_{2s}$. During this interval v_3 is zero and φ_3 is constant at Φ_{3s}. At point b the core of X_2 saturates. From b to c, C_2 discharges into C_3, φ_2 makes a small excursion into the saturated region, and v_2 returns to zero completing the main voltage pulse on C_2. At the same time v_3 goes negative to $-\pi V_i$ and φ_3 varies from $+\Phi_{3s}$ to $-\Phi_{3s}$. At point c the core of X_3 saturates. Immediately following c Fig. 12 shows the excursion of φ_3 into the saturated region and the return of v_3 to zero, thus completing the main voltage pulse on C_3. The effect of v_3 upon φ_2 during this discharge period τ_3 is neglected in Fig. 12. The resetting of the core of the second thyractor starts at c. The voltage and flux of the second thyractor rise according to (38) and (39) respectively. At d, the core of X_2 saturates in the positive direction. Between d and e

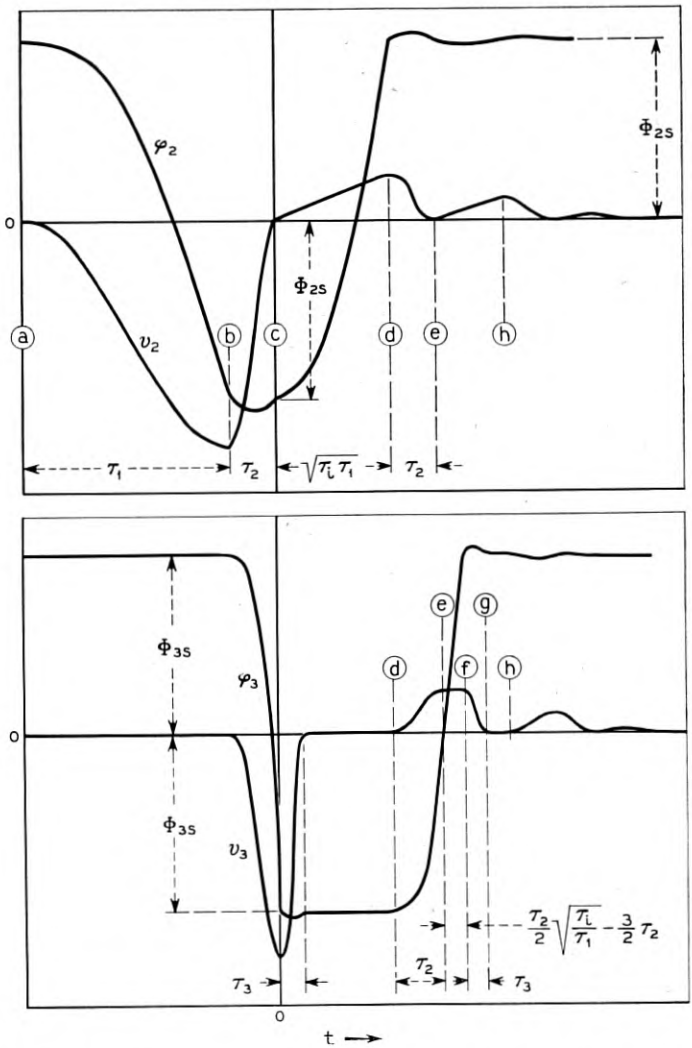


Fig. 12 — Curves of capacitor voltages and core fluxes illustrating the resetting action, of the cores of X_2 and X_3, between main pulses. This action is made possible by the core bias of X_1.

C_2 discharges into C_3. This discharge is of the same nature as the main pulse discharge in the interval b — c except that the variation is reversed and the magnitude of the voltage is much smaller. The time required, however, is the same, that is τ_2. The third thyractor does not saturate immediately at the end of this discharge because of the low

voltage. A charge is thus trapped on C_3 for an additional time equal to $\tau_2/2\sqrt{\tau_i/\tau_1} - \frac{3}{2}\tau_2$. At f the core of X_3 saturates and C_3 discharges into C_4.

A positive-going pulse is thus formed on C_2 in the interval c – d. This pulse is transferred to C_3 in the interval d – e, to C_4 in the interval f – g and so on along the chain. In the dissipationless case this pulse should reset all of the thyractor cores.

However the current I_b given by (9) continues to flow through the first thyractor. At point e, therefore, C_2 starts to recharge and v_2 again rises linearly. In the interval e – g the voltage across the second thyractor is the difference between v_2 and v_3. Since at first $v_3 > v_2$, φ_2 makes a small excursion into the unsaturated region. At g, v_3 becomes zero and φ_2 returns to positive saturation at h. A second positive-going pulse is thus formed on C_2 in the interval e – h. This pulse also moves down the chain.

This action continues, each pulse being smaller and shorter than the preceding one. Beyond h Fig. 12 is not intended to be accurate but shows in a qualitative manner that all capacitor voltages approach zero and all core fluxes approach positive saturation as required for the initial condition for the next main pulse.

In Fig. 12 the time intervals are not shown in their true relative magnitudes but are, in each case, made large enough to show the curve shape. Actual oscillograms will also show considerable deviations from Fig. 12 because of dissipation and because of small charges left on the capacitors by the main pulse.

Part III. DC-Charged Series-Type Magnetic Pulse Modulator

The dc-charged series-type magnetic pulse modulator to be discussed in this part draws power from a dc source and provides unidirectional high-voltage, short-duration pulses to a load. The basic differences between the ac- and dc-charged magnetic modulators are the type input sections required, the resulting limitations imposed upon the first section of the transfer chain, and the manner in which the saturable elements of the chain are reset.

Four input or charging circuits that periodically charge a capacitor from a dc source at a rate determined by an external trigger supply are shown in Figure 13. All these arrangements may be used for either resonant or linear charging. The first or input stage of the charging circuits shown in parts a and b of this figure are familiar arrangements employed in other type modulators. The diode indicated in Figure 13(b) permits resonant operation when the charging period is much smaller

than the pulse repetition period, or when a variable inter-pulse period is required. After C_1 in either of these two circuits has attained its desired voltage either by resonant or linear charging, the switching element is activated by an external trigger voltage and the stored energy is transferred to C_2 in a time interval short compared to the charging time of C_1. The switch is then re-opened by the completion of this transfer action and by the removal of the trigger voltage, and C_1 again begins to charge while C_2 is discharged by the action of the transfer chain (not shown in the figure).

The diode of Figure 13(b) can be replaced by the active switching device since it, too, normally has rectifying properties. This arrangement shown in Figure 13(c) operates the switching device at a lower peak current since the charging time of C_1 is long compared to that of C_2. However, it does introduce an additional large delay between the start of the trigger pulse and the generation of the modulator output pulse.

Figure 13(d) is a special charging arrangement suggested by Professor C. Neitzert that provides energy during part of the charging period of C_1 to reset the saturable elements of the transfer chain, in a manner similar to the resetting action obtained in the pulse transformer of a line type modulator. In addition, it permits operation of one element of the active switch at ground potential, which is advantageous if filamentary power is required in this device.

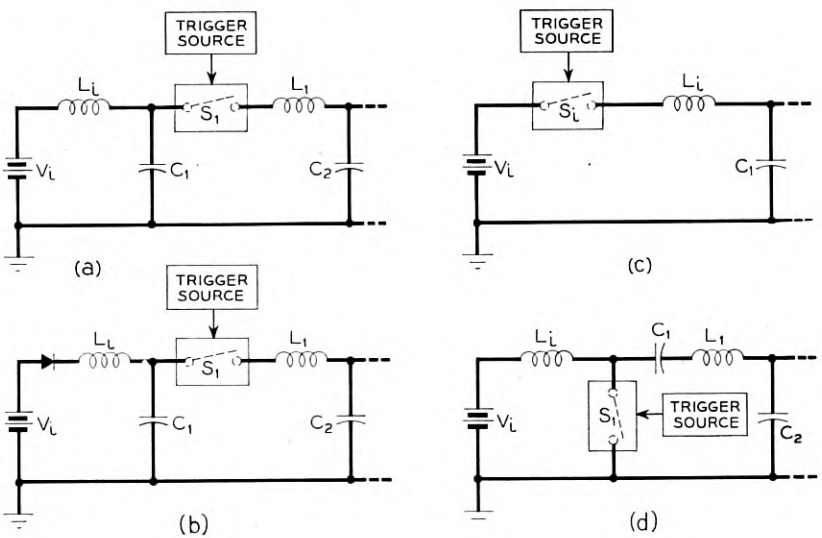


Fig. 13 — DC-charging circuits.

A typical dc-charged magnetic modulator, consisting of an input section, n transfer sections, an output section, a damping section, and load is shown in Figure 14.[5] The remainder of this part will be restricted to the analysis and discussion of the input section, a typical transfer section, and the damping section, since the response of the output section and the characteristics of a magnetron load have been treated in detail in Part II. After this is done, a composite analysis of the input section and transfer chain will be made by drawing upon the results of the first two analyses.

Before this is undertaken, certain assumptions that involve all sections and the modulator operation in general will be made. These assumptions embody an attempt to include the effect of losses which have pronounced influence on observed modulator performance. In order to facilitate the mathematical treatment of the problem, only series dissipative elements will be used. Generally, a good approximation is obtained by the following assumptions:

1. The dc source will be considered ideal.
2. The constant voltage drop generally associated with the active switching device during the charging period will be subtracted from the ideal source potential yielding a net source potential of V_i.
3. No current will flow through either an open active switching device or an unsaturated thyractor.
4. The core losses associated with any thyractor and transformer

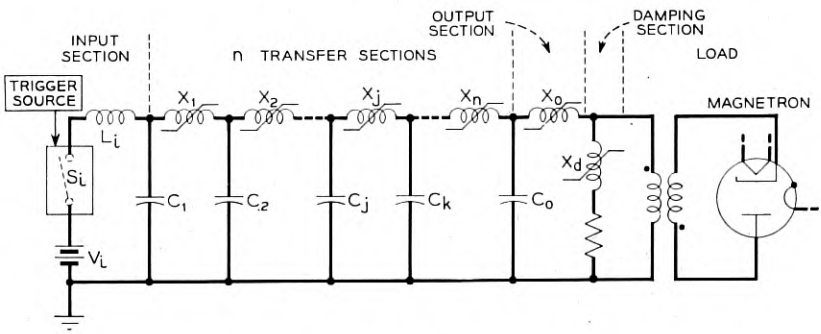


Fig. 14 — Typical dc-charged magnetic modulator.

[5] In this figure, the typical modulator operates from a dc source of positive potential and employs the active switch in the first section. This potential is used, rather than a negative one which would appear more practical, in order to facilitate the analysis that follows. The active switch could have been employed in the second section but since the only difference between such an arrangement and the one indicated is merely a matter of triggering no generality has been sacrificed.

will be represented by a series resistance in the preceding section, effective only during the time the switch associated with this section is closed.

5. The winding losses associated with any thyractor and transformer will be represented by a series resistance in the section associated with the element, effective only during the time the switch of this section is closed.

6. All capacitors will be considered lossless.

An equivalent circuit of the typical modulator based on the above assumptions is shown in Figure 15. In this figure the inductances L_1, L_2, L_j, L_n, L_0 and L_d represent the saturated inductances of the thyractors. Inductance L_t and capacitance C_d are the leakage inductance of the transformer and the distributed capacitance of the transformer and magnetron, respectively. The losses associated with the practical transformer have been included in resistances R_n and R_0 as explained in the assumptions.

Assume that a pulse of energy has been initiated down the chain and this energy is now stored in a capacitor between C_2 and C_j, while the preceding pulse has been completely dissipated in the damping section and load. Hence, the sections shown explicitly in Figure 15 are in an inactive or quiescent state. However, voltages may still exist on the capacitors of these sections, and these quiescent voltages are indicated in the figure. The energy storage associated with these voltages makes possible, as will be demonstrated later, a second method of stably operating the saturable elements without the necessity of bias windings. Now,

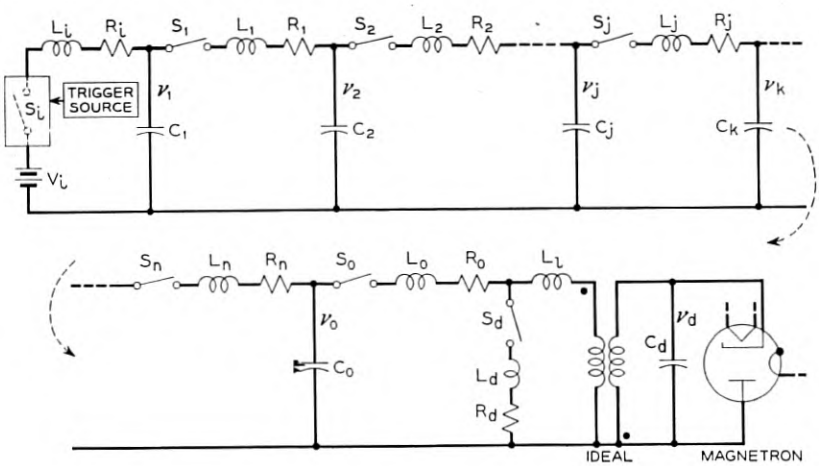


Fig. 15 — Equivalent circuit of typical dc-charged magnetic modulator.

if the assumption is made that all thyractor switch closures are initiated only when the capacitor preceding the switch stores all the available pulsed energy and the switch is re-opened when this energy, less losses, is transferred to the following capacitor, the circuit operation will be completely defined. Such an assumption means that for each pulse of energy initiated down the chain by the switch S_i, the switches close and subsequently re-open only once; that is, secondary saturations of any element are precluded.[6]

INPUT CHARGING SECTION

The input charging section of Figure 15 is shown in detail in Figure 16. Observe that the rectifying property usually associated with the active switching device is indicated explicitly by the presence of an ideal diode

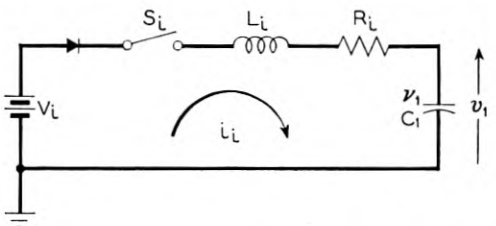


Fig. 16 — Equivalent circuit of input charging section.

in this figure. Although this circuit has already been thoroughly analyzed in past work on charging circuits,[7] the resonant charging case will be briefly treated here in order to provide the necessary background material to support the later discussion on the over-all modulator performance.

At the time switch S_i closes, time t_i equals zero, the current through the charging inductor L_i is zero, and the capacitor C_1 is charged to ν_1. For these quiescent conditions, the current i_i through the switch is of the form

$$i_i = \frac{V_i - \nu_1}{L_i \omega_i} \epsilon^{-\alpha_i t_i} \sin \omega_i t_i \tag{41}$$

[6] Possibly it would be more desirable to store energy in either or both of the first and last capacitors of the chain that generate, during the period that the modulator is inactive, secondary voltage pulses of the proper polarity traveling in a direction which would tend to reset all magnetic elements. Such operation, however, will not be considered here.

[7] G. W. Glasoe and J. V. Lebacqz, Pulse Generators, McGraw-Hill Book Company, Inc., 1948.

where

$$V_i > \nu_1 \tag{42}$$

$$\alpha_i = \frac{R_i}{2L_i} \tag{43}$$

$$\omega_i = \sqrt{\beta_i^2 - \alpha_i^2} \tag{44}$$

and

$$\beta_i = \sqrt{\frac{1}{L_i C_1}} \tag{45}$$

for the oscillatory case. This is the only case of interest, since under this condition the charging current eventually tends to reverse, consequently initiating deactivation of the switch. This deactivation is completed by removal of the trigger voltage if it is still present. It is seen from equation (41) that the current is unidirectional over the period τ_i, the charging time of the section, where τ_i is

$$\tau_i = \frac{\pi}{\omega_i} \tag{46}$$

The voltage across C_1 as a function of time t_i is v_1, where

$$v_1 = V_i - [V_i - \nu_1]\left[\frac{\alpha_i}{\omega_i} \sin \omega_i t_i + \cos \omega_i t_i\right] \epsilon^{-\alpha_i t_i} \tag{47}$$

Since, as previously noted the charging current through C_1 has been unidirectional during the entire charging period, capacitor C_1 will be charged to its maximum voltage V_1 at the end of this period, that is, at time t_i equals τ_i. From (47) this peak voltage is

$$V_1 = V_i + (V_i - \nu_1)\delta_i \tag{48}$$

where δ_i, the loss factor of the input stage, is

$$\delta_i = \epsilon^{-\alpha_i \tau_i} \tag{49}$$

Equation (48) may also be written in the following form:

$$V_i = \left[\frac{1}{1 + \delta_i}\right] V_1 + \left[\frac{\delta_i}{1 + \delta_i}\right] \nu_1 \tag{50}$$

Upon normalizing the voltages of equations (48) and (50) and rearranging the following simple linear functions result:

$$\underline{V_1} = (1 - \underline{\nu_1})\delta_i + 1 \tag{51}$$

$$\underline{V_1} = (-\delta_i)\underline{\nu_1} + (1 + \delta_i) \tag{52}$$

and

$$\underline{V_i} = [\delta_i/(1 + \delta_i)]\underline{\nu_1} + 1/(1 + \delta_i) \tag{53}$$

In these three equations the underscored voltages indicate normalized values; the voltages of the first two equations have been normalized to V_i while those of the last equation have been normalized to V_1. Typical curves for equations (51), (52) and (53) are shown in Figs. 17, 18 and 19, respectively. Note that the ranges of the variables in the figures are restricted to represent only the practical cases where the energy stored in capacitor C_1 is increased by the charging current ($V_1 > \nu_1$), and the loss factor is restricted to the realizable values ($0 < \delta_i < 1$). These curves present an easy graphical means of determining any one of the following parameters when the other two are known: input voltage V_i, output voltages ν_1 and V_1, and input loss factor δ_i. Once all the above parameters are determined, it becomes a simple matter to calculate losses, efficiency, etc.

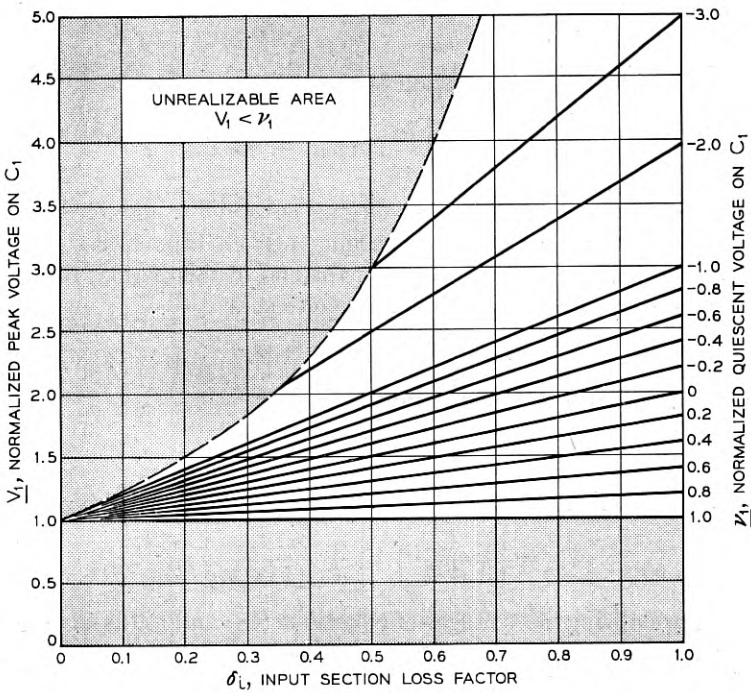


Fig. 17 — Straight line plot for the input section for V_i equals unity.

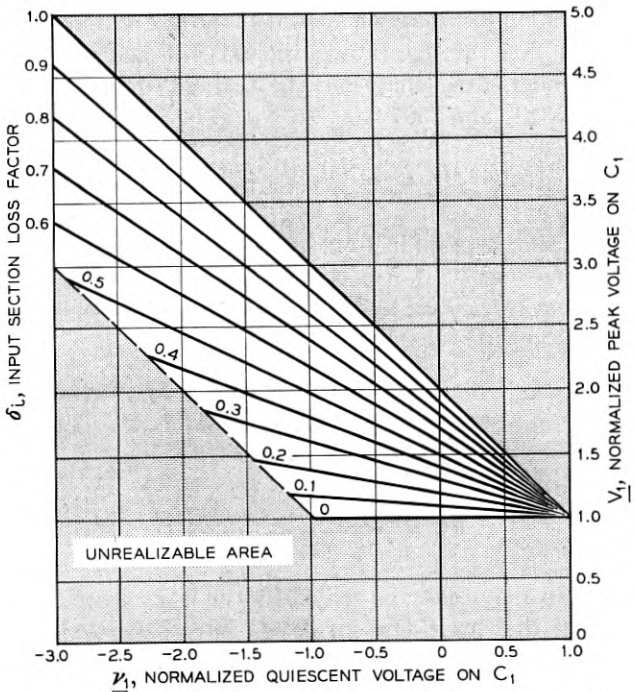


Fig. 18 — Straight line plot for the input section for V_i equals unity.

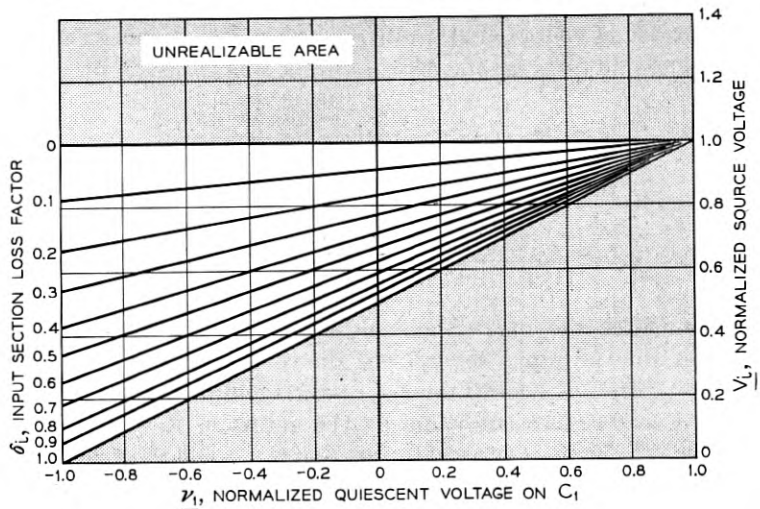


Fig. 19 — Straight line plot for the input section for V_1 equals unity.

Since the switch, S_i, has been opened at the end of the charging period, no current will flow through it until the switch is again triggered at time t_i equals T. Under these conditions the average current, $\bar{i}_i$, through the switch is from (41) and (50)

$$\bar{i}_i = \frac{C_1}{T}(V_1 - \nu_1) \tag{54}$$

Although a simple graphical method for determining the functions of energy has been suggested, it is still noteworthy to write these relations explicitly. From (41), (50), and (54), the energy per pulse, W_i, that is dissipated in the series resistance, R_i, is

$$W_i = \frac{\bar{i}_i^2 T^2(1 - \delta_i)}{2C_1(1 + \delta_i)} \tag{55}$$

or

$$W_i = \frac{C_1(1 - \delta_i)(V_1 - \nu_1)^2}{2(1 + \delta_i)} \tag{56}$$

These equations may easily be solved for the input loss factor, δ_1, yielding expressions that are useful for design purposes.

Since explicit relations for the energy loss per pulse have been derived, an expression for the efficiency, η_i, of the energy transfer from the source, V_i, to the first capacitor, C_1, may be written. From equations (55) and (56), respectively, two expressions for this efficiency, which is the ratio of the increase of energy stored in C_1 at the end of the charging period to the energy supplied by the source during this period, are

$$\eta_i = \frac{2C_1(1 + \delta_i)V_i - T(1 - \delta_i)\bar{i}_i}{2C_1(1 + \delta_i)V_i} \tag{57}$$

or

$$\eta_i = \frac{(1 + \delta_i)(V_1 + \nu_1)}{(1 + \delta_i)(V_1 + \nu_1) + (1 - \delta_i)(V_1 - \nu_1)} \tag{58}$$

To summarize the preceding analysis, typical voltage and current waveforms for the input section are shown in Fig. 20. Observe that capacitor C_1 after it has attained its peak voltage is not immediately discharged to its quiescent value by the action of the transfer chain. Instead this discharge is delayed by a period G_1, called the guard interval of the first transfer section. It was noted that the active switch at the end of the charging period, τ_i, was opened by a combination of circuit actions. Now some devices, such as a thyratron, require a definite time

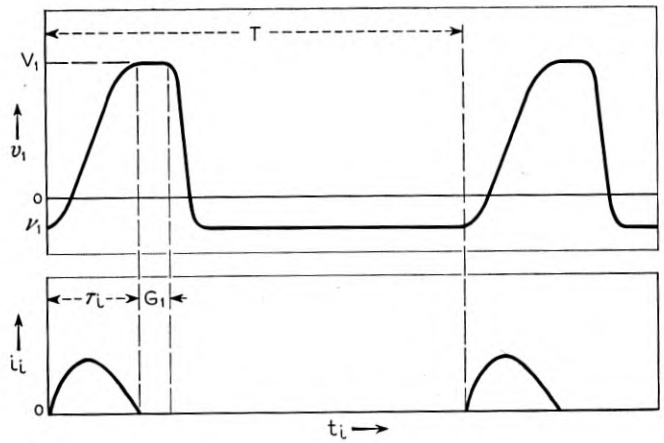


Fig. 20 — Typical waveforms for the input section.

period for deactivation to insure that the switch remains open. The guard interval, G_1, provides this period by forestalling the discharge action.

An expression that completely defines the operation of the input charging section would greatly facilitate the following endeavor to synthesize the over-all modulator performance. Such an expression is easily constructed by writing equations (50) and (54) in the following matrix form:

$$\begin{bmatrix} V_i \\ \overline{i_i} \end{bmatrix} = \begin{bmatrix} 1/(1+\delta_i) & \delta_i/(1+\delta_i) \\ C_1/T & -C_1/T \end{bmatrix} \begin{bmatrix} V_1 \\ \nu_1 \end{bmatrix} \tag{59}$$

Equation (59) relates the input and output conditions of the charging section in terms of the circuit constants independent of the time variable, t_i. Note that there is a similarity between this equation and the equation that relates the input and output voltages and currents of a four-terminal linear passive network by its $ABCD$ constants. The similarity is only in form; hence, the lower case letters *abcd* will be used to denote the constants of the above equation in the following fashion:

$$\begin{bmatrix} V_i \\ \overline{i_i} \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} V_1 \\ \nu_1 \end{bmatrix} \tag{60}$$

where

$$a_i = 1/(1+\delta_i) \tag{61}$$

$$b_i = \delta_i/(1+\delta_i) \tag{62}$$

$$c_i = C_1/T \tag{63}$$

and

$$d_i = -C_1/T \tag{64}$$

These relations, as mentioned above, will be employed later.

TYPICAL TRANSFER SECTION

The j*th* or typical section of the transfer chain of Fig. 15, is shown in detail in Figure 21. According to the assumptions concerning switch operation, switch S_j will remain open as shown until the total available pulsed energy previously initiated by the charging section is stored in capacitor C_j. Consequently, at the time switch S_j closes, starting at time t_j equals zero, the current through the saturated j*th* thyractor is zero, and capacitor C_j is charged to its peak voltage V_j while C_k is still at its quiescent value ν_k. For these conditions, the current, i_j, through the

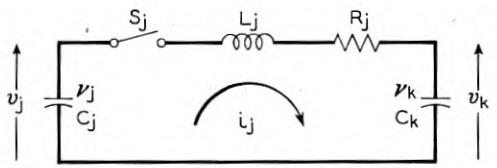


Fig. 21 — Equivalent circuit of a typical transfer section.

switch is of the form

$$i_j = \frac{V_j - \nu_k}{L_j\omega_j}\,\epsilon^{-\alpha_j t_j}\sin\omega_j t_j \tag{65}$$

where

$$V_j > \nu_k \tag{66}$$

$$\alpha_j = \frac{R_j}{2L_j} \tag{67}$$

$$\omega_j = \sqrt{\beta_j^2 - \alpha_j^2} \tag{68}$$

$$\beta_j = \sqrt{\frac{1+\lambda_j}{L_jC_j}} \tag{69}$$

and

$$\lambda_j = \frac{C_j}{C_k} \tag{70}$$

for the oscillatory case. This again is the only case of interest since the transfer current will eventually tend to reverse, driving the thyractor out of saturation and hence extinguishing the current. Under these conditions, it is seen from (65) that the period of current flow is τ_j, called the transfer period of the j*th* section, where

$$\tau_j = \frac{\pi}{\omega_j}. \tag{71}$$

The voltages across C_j and C_k as functions of time, t_j, v_j and v_k respectively, may be shown to be

$$v_j = V_j - \frac{V_j - \nu_k}{1 + \lambda_j}\left[1 - \left(\frac{\alpha_j}{\omega_j}\sin \omega_j t_j + \cos \omega_j t_j\right)\epsilon^{-\alpha_j t_j}\right] \tag{72}$$

and

$$v_k = \nu_k + \frac{\lambda_j(V_j - \nu_k)}{1 + \lambda_j}\left[1 - \left(\frac{\alpha_j}{\omega_j}\sin \omega_j t_j + \cos \omega_j t_j\right)\epsilon^{-\alpha_j t_j}\right] \tag{73}$$

Since as noted above the current has been unidirectional during the entire transfer period, the capacitor C_j will be charged to its minimum or quiescent voltage ν_j and capacitor C_k will be charged to its maximum voltage V_k at the end of this period. Hence, at time t_j equals τ_j (72) and (73) become

$$\nu_j = \left[1 - \frac{1 + \delta_j}{1 + \lambda_j}\right] V_j + \left[\frac{1 + \delta_j}{1 + \lambda_j}\right] \nu_k \tag{74}$$

and

$$V_k = \left[\lambda_j \frac{1 + \delta_j}{1 + \lambda_j}\right] V_j + \left[1 - \lambda_j \frac{1 + \delta_j}{1 + \lambda_j}\right] \nu_k \tag{75}$$

where, δ_j, the loss factor of the j*th* transfer section, is

$$\delta_j = \epsilon^{-\alpha_j \tau_j} \tag{76}$$

Rearrangement of equations (74) and (75) and normalization of the voltages to V_j (underscored voltages again represent normalized values) yield equations (77) and (78):

$$\underline{V_k} = \left[\frac{\delta_j}{1 + \delta_j}(1 - \underline{\nu_j})\right] \lambda_j + \left[\frac{\delta_j}{1 + \delta_j}\left(1 + \frac{\underline{\nu_j}}{\delta_j}\right)\right] \tag{77}$$

$$\underline{\nu_k} = \left[\frac{1}{1 + \delta_j}(\underline{\nu_j} - 1)\right] \lambda_j + \left[\frac{\delta_j}{1 + \delta_j}\left(1 + \frac{\underline{\nu_j}}{\delta_j}\right)\right] \tag{78}$$

which are both families of straight line functions of the variables λ_j and

$\underline{V_k}$ or $\underline{\nu_k}$, respectively. These families result when a particular value of loss factor, δ_j (any value between 0 and 1), is assigned and $\underline{\nu_j}$ is given a set of values within its permissible range (its maximum range is from -1 to 1; however, it is further restricted as will be demonstrated below).

Observe that both families of straight lines of equations (77) and (78) have the same zero intercept ($\lambda_j = 0$), namely

$$\underline{V_k}(\lambda_j = 0) = \underline{\nu_k}(\lambda_j = 0) = \frac{\delta_j}{1 + \delta_j}\left(1 + \frac{\underline{\nu_j}}{\delta_j}\right) \tag{79}$$

Also, it may be shown that for any value of $\underline{\nu_j}$ the family of lines of equation (77) will go through the point

$$(\lambda_j, \underline{V_k}) = \left(\frac{1}{\delta_j}, 1\right) \tag{80}$$

And similarly for any value of both $\underline{\nu_j}$ and δ_j the family of lines of equations (78) will go through the point

$$(\lambda_j, \underline{\nu_k}) = (-1, 1) \tag{81}$$

Hence, equations (77) and (78) present an easily constructed graphical method of displaying all possible quiescent and peak voltages on the two capacitors, C_j and C_k, for a particular value of δ_j and any value of capacitance ratio λ_j. Two such plots are shown in Figs. 22 and 23, one for the lossless case, δ_j equals unity, and the other for a practical case, δ_j equals $\frac{2}{3}$, respectively. With the latter type plot that accounts for circuit dissipation, the loss and efficiency involved in the transfer can quickly be computed since the energy storage in the capacitors before and after the transfer action are readily determined. Both these figures suggest that peak voltage amplification can be realized if λ_j is made greater than $1/\delta_j$. This interesting possibility will be discussed later.

In Figure 23, the shaded area is unrealizable and is defined by the following considerations. At the start of the transfer action energy is stored in capacitor C_j and C_k. Although, depending on the capacitance ratio, the energy storage in C_k may be greater, $\underline{\nu_k}$ must always be less than unity as seen from (66). Now, at the end of the transfer period the only condition of interest is the case where the energy storage, despite circuit dissipation, has been increased in C_k. This means that $\underline{V_k}$ must always be greater than $|\underline{\nu_k}|$. A line that includes all the points for V_k equal to $|\underline{\nu_k}|$ can readily be constructed graphically. Such a line is shown in Fig. 23, and all points below it represent conditions where $\underline{V_k}$ is less than $|\underline{\nu_k}|$, and consequently are unrealizable.

A figure that represents the general case with circuit dissipation is

shown in Figure 24. The four regions which are indicated in this figure are defined by the voltage conditions of Table I. A typical voltage waveform for each region is shown in Fig. 25. In this figure, capacitor C_j charges to its peak voltage over a time interval τ_h, the transfer period of the h*th* section which precedes the j*th* section in the transfer chain. Capacitor C_k is discharged in a time interval τ_k, the transfer period of the following or k*th* section. Again it should be noted that both capacitors C_j and C_k are not immediately discharged once they attain their peak

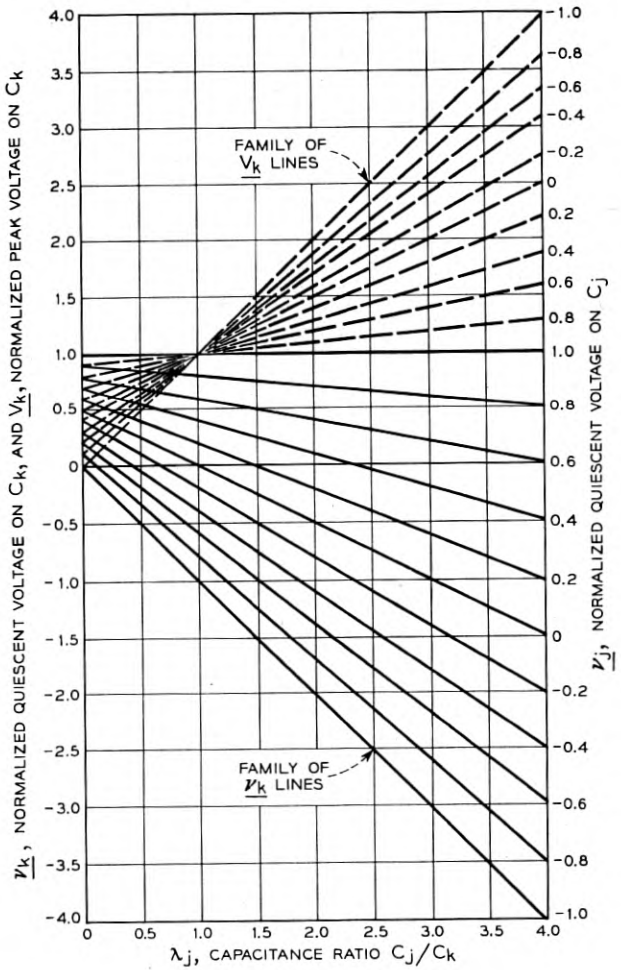


Fig. 22 — Straight line plot for a lossless transfer section, δ_j equals 1.

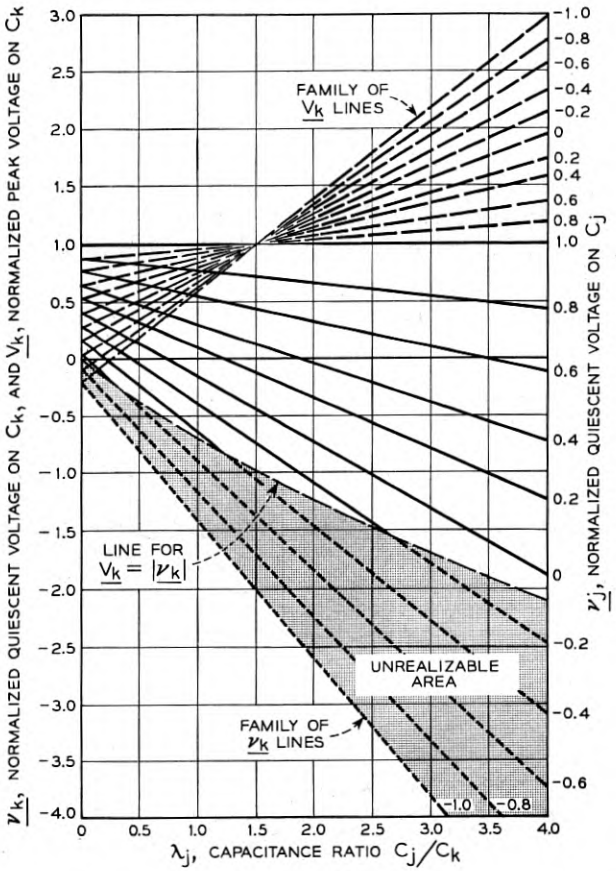


Fig. 23 — Straight line plot for a practical transfer section, δ_j equals $\frac{2}{3}$.

voltages but remain charged for a period of G_j and G_k, the guard intervals of the j*th* and k*th* transfer sections, respectively. This time delay, or guard interval, plays an important part both in permitting circuit operation without the necessity of bias windings on the thyractors and in designing these elements to operate properly even when wide production tolerances are allowed. This former aspect will be treated later. Capacitor C_1 that was discussed in the charging stage and C_k are discharged in the same manner as capacitor C_j.

Since the switch, S_j, has been opened at the end of the transfer period, no current will flow through it until the switch is again closed at time t_j equals T. Under these conditions the average current, $\bar{i}_j$, through the

TABLE I

Region	ν_j	ν_k
I	>0	>0
II	>0	<0
III	<0	>0
IV	<0	<0

switch from (65), (74) and (75) is

$$\bar{i}_j = \frac{C_j}{T}(V_j - \nu_j) \tag{82}$$

or

$$\bar{i}_j = \frac{C_k}{T}(V_k - \nu_k) \tag{83}$$

Although again a simple graphical method for determining the functions of energy has been suggested, explicit forms of these relations will be written. From equations (65), (74) and (75), the energy per pulse

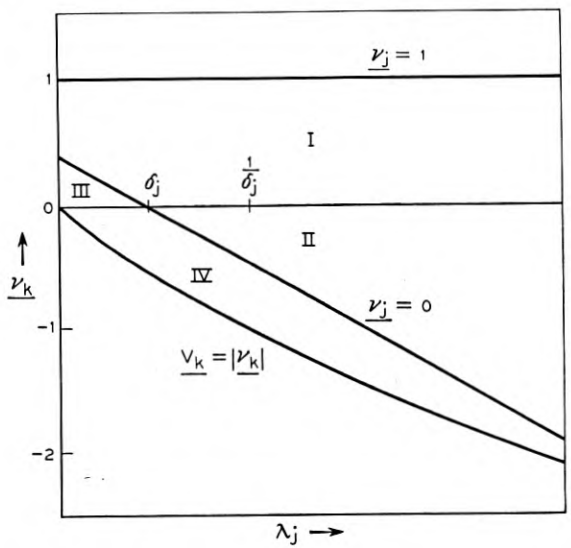


Fig. 24 — General straight line plot for a typical transfer section with losses, indicating the four regions of operation.

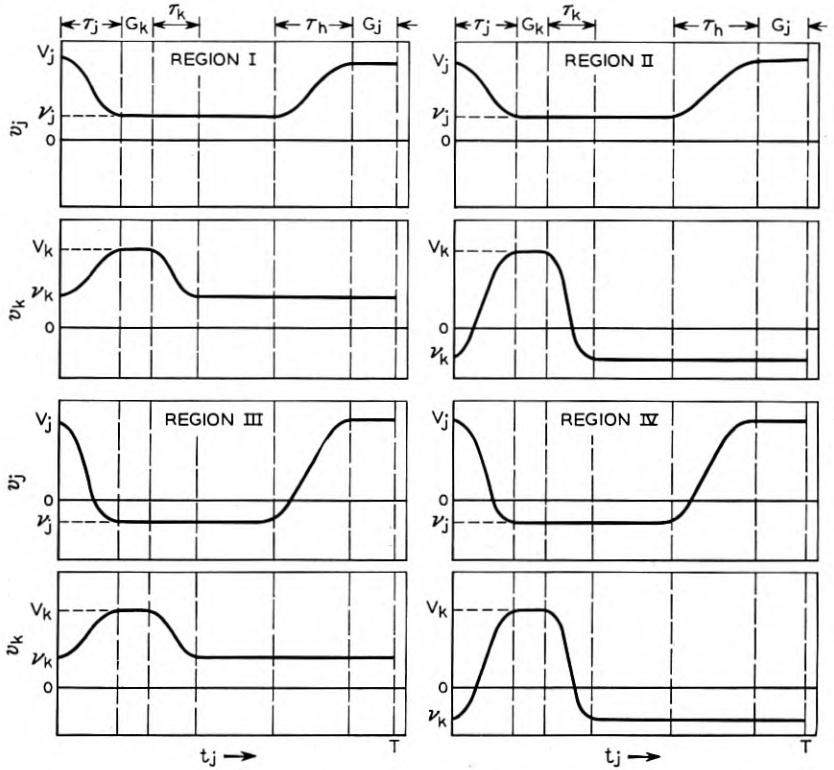


Fig. 25 — Typical voltage waveforms of a transfer section experienced in the four possible operating regions.

W_j, that is dissipated in the series resistance, R_j, is

$$W_j = \frac{C_j(V_j - \nu_j)^2(1 + \lambda_j)(1 - \delta_j)}{2(1 + \delta_j)} \tag{84}$$

or

$$W_j = \frac{C_k(V_k - \nu_k)^2(1 + \lambda_k)(1 - \delta_j)}{2\lambda_j(1 + \delta_j)} \tag{85}$$

These equations may be solved for the transfer loss factor, δ_j, which once again yield expressions that are useful in design computations.

From equations (84) and (85), expressions for the efficiency, η_j, of the energy transfer from capacitor C_j to C_k may be written. The resulting expressions for this efficiency, which is the ratio of the increase of energy

in C_k to the decrease of energy in C_j due to the transfer action, are

$$\eta_j = \frac{(1 + \delta_j)(V_j + \nu_j) - (1 + \lambda_j)(1 - \delta_j)(V_j - \nu_j)}{(1 + \delta_j)(V_j + \nu_j)} \tag{86}$$

and

$$\eta_j = \frac{\lambda_j(1 + \delta_j)(V_k + \nu_k)}{\lambda_j(1 + \delta_j)(V_k + \nu_k) + (1 + \lambda_j)(1 - \delta_j)(V_k - \nu_k)} \tag{87}$$

Up to this point, little has been said about how the thyractors of the transfer chain switch in the manner as outlined in the assumptions. Such operation is now worth consideration. In order to realize the equivalent circuit of Fig. 21, the j*th* thyractor must have the hysteresis loop shown in Figure 26. It was assumed that S_j closes at time t_j equals zero when capacitor C_j is charged to its maximum voltage, V_j. The switch closure is associated with the saturation of the j*th* thyractor. Assume the core at this time is at the positive saturation point designated by $+B_s$ in Fig. 26. During the current discharge it moves out to some value of maximum field intensity represented by point e. For any practical core material this value of H, which can be derived from (65), is normally hundreds of times its coercive force. At time equal to τ_j the current is again zero, and hence the core has returned to $+B_s$. It was noted before, in the discussion following (65), that the transfer current by its tendency to reverse drives the core into an unsaturated region. This region, lying between $+B_s$ and $-B_s$, has extremely high permeability; consequently, the high thyractor impedance cuts off the current. In order

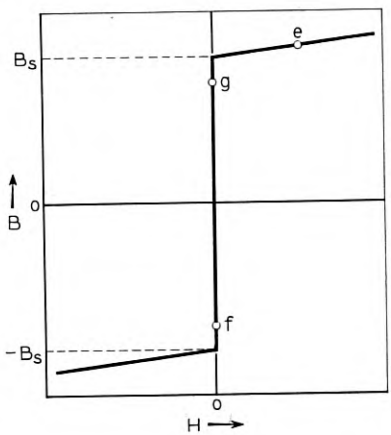


Fig. 26 — Hysteresis loop of the j*th* thyractor.

to best utilize the core material, the core must be driven to some point near or equal to $-B_S$ such as f before the next pulse that is initiated by the charging stage starts to recharge C_j to its peak voltage.

Rather than use a bias winding to establish a polarizing field as suggested by Melville, the resetting action can be accomplished by insuring that the proper voltage waveform exists across the thyractor. The flux density swing in the core is proportional to the time integral of voltage across the main winding. This voltage for the j*th* thyractor inductor, which is merely the difference of the voltages on capacitors C_j and C_k, over the pulse repetition period T can have any of the three forms shown in Fig. 27 depending on the quiescent values, ν_j and ν_k. To avoid confusion, the voltage across this element during the period τ_j is not shown, since it was demonstrated that the core over this period will experience no net flux swing due to such voltages. As can be verified from Figs. 22 and 23, these waveforms are associated with the values λ_j, indicated in the figure.

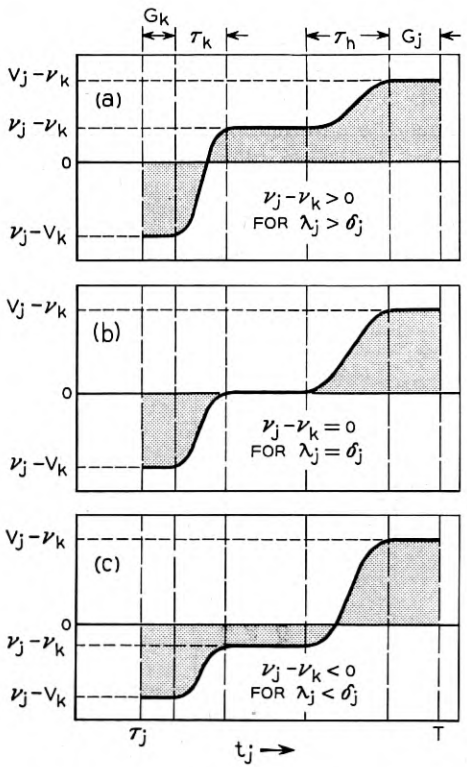


Fig. 27 — Possible voltage waveforms across the j*th* thyractor.

If the net area under the voltage curve from time t_j equals τ_j to T is zero, the core magnetization will have traversed a path similar to $B_s g f g B_s$ of Fig. 26. The subsequent saturation of the core and resulting transfer action will complete the cycle in the manner previously described by driving the magnetization from $+B_s$ to e and returning to $+B_s$. The negative area in Fig. 27 swings the core on the first part of this path from $+B_s$ to its smallest value of flux density represented by point f. This value of flux density must be greater than $-B_s$ in order to avoid secondary saturations. Furthermore, for the most economical utilization of the core material, the core magnetization should reach this value approximately at the start of the charging period of capacitor C_j as in Figure 27(b) or (c). The flux swing resulting from the positive area up to the time capacitor C_j is charged to its peak voltage should bring the core to some point less than $+B_s$ such as g on the loop. Capacitor C_j will then remain charged for a length of time G_j, the guard interval, such that the additional area causes the core to saturate completing the path in the unsaturated region.

Now it can be demonstrated from (74) and (75) that the peak positive voltage across the thyractor is always greater than the magnitude of the peak negative voltage for any practical transfer section ($\delta_j < 1$). Since pulse shortening is required in successive transfer sections, that is, the transfer periods of successive sections must be smaller and smaller, it is obvious from Fig. 27(b), that for λ_j equals δ_j, the only manner in which to make the net area under the voltage curve exactly zero, as is required, is to make the guard interval of the following section, G_k, greater than the guard interval of the j*th* section, G_j.

From this line of reasoning, it is seen from Figure 27(a) that for λ_j greater than δ_j the guard interval G_k must be made even larger. However, for λ_j less than δ_j as in Fig. 27(c), guard interval shortening in addition to pulse shortening may be realized. The advantages of such operation will be discussed in the composite analysis of the input section and transfer chain that follows.

To complete the analysis of the typical transfer section, it is possible to write an expression, similar to the one derived for the input section, that relates the peak and quiescent voltages on the capacitors C_j and C_k. Rearrangement of equations (74) and (75) admits of the following:

$$\begin{bmatrix} V_j \\ \\ \nu_j \end{bmatrix} = \left[\begin{array}{c|c} \dfrac{1+\lambda_j}{\lambda_j(1+\delta_j)} & 1 - \dfrac{1+\lambda_j}{\lambda_j(1+\delta_j)} \\ \hline \dfrac{1+\lambda_j}{\lambda_j(1+\delta_j)} - \dfrac{1}{\lambda_j} & 1 + \dfrac{1}{\lambda_j} - \dfrac{1+\lambda_j}{\lambda_j(1+\delta_j)} \end{array}\right] \begin{bmatrix} V_k \\ \\ \nu_k \end{bmatrix} \tag{88}$$

or

$$\begin{bmatrix} V_j \\ v_j \end{bmatrix} = \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \begin{bmatrix} V_k \\ v_k \end{bmatrix} \tag{89}$$

where

$$a_j = \frac{1 + \lambda_j}{\lambda_j(1 + \delta_j)} \tag{90}$$

$$b_j = 1 - \frac{1 + \lambda_j}{\lambda_j(1 + \delta_j)} \tag{91}$$

$$c_j = \frac{1 + \lambda_j}{\lambda_j(1 + \delta_j)} - \frac{1}{\lambda_j} \tag{92}$$

and

$$d_j = 1 + \frac{1}{\lambda_j} - \frac{1 + \lambda_j}{\lambda_j(1 + \delta_j)} \tag{93}$$

More will be said later concerning the application of these equations in the evaluation of the modulator performance.

DAMPING SECTION

Since the magnetron current at the end of the RF pulse is extinguished when considerable energy is still stored in circuit inductances and capacitances, the magnetron voltage pulse decays in an oscillatory fashion.[8] Normally, this oscillation will be reinforced by energy remaining in the energy storage device of the modulator. The resultant negative voltage peaks generally give rise to low-power secondary RF pulses which are most undesirable since they occur during the listening period of the radar, masking even the strongest echos.

The undesirable voltage transient may be avoided by minimizing the inductively stored energy and by dissipating the capacitively stored energy with the damping section shown in Fig. 14. In this section the thyractor, called a damping thyractor, is designed to saturate at the end of the main RF pulse. The resistance in series with this element is chosen to effectively terminate the energy storage device, discharging it before oscillations can arise. The saturated inductance of the damping thyractor is proportioned such that the distributed capacitance across the magnetron is terminated in a slightly underdamped circuit in order to complete its discharge in a relatively short period. Hence, the magnitude of

[8] *Pulse Generators*, op. cit.

the positive voltage backswing across the magnetron is controlled by the degree of damping and can be reduced to such a value that no objectionable voltage reversal will follow.

The damping circuit that exists across the distributed capacitance when the damping thyractor saturates could be analyzed in the same fashion as the two preceding type sections. However, since this occurs after the generation of the RF pulse, it would contribute little to the synthesis of the over-all performance of the modulator and hence only the results will be briefly noted.

When the damping thyractor saturates, the distributed capacitance across the magnetron is charged to the magnetron firing voltage, $-V_M$. For an underdamped circuit, the current through the thyractor, similar to that of (41), discharges this capacitance over a period τ_d, the damping period. At the end of this period, the current is extinguished by its tendency to reverse through the thyractor, and the magnetron capacitance remains charged to some positive potential, v_d, where

$$v_d = \delta_d V_M \tag{94}$$

The loss factor, δ_d, which is also the backswing ratio in this case, is defined in terms of the circuit parameters of the damping section similar to the form evolved for the input loss factor, δ_i, of (49). It may be demonstrated that the shortest damping period, $\pi R_T C_d$ seconds, is realized when the total series inductance of the damping path is made equal to $R_T^2 C_d/2$ henries, where R_T is the total resistance referred to the same winding as the distributed capacitance, C_d. The backswing ratio for this condition is $\epsilon^{-\pi}$, about four per cent. Hence, the damping circuit is completely defined by the above relations since R_T is made to be the characteristic impedance of the energy storage device and C_d is fixed by the pulse transformer and magnetron combination.

Finally the damping efficiency, η_d, that is the ratio of the energy dissipated over the damping period to the energy originally stored in C_d is:

$$\eta_d = 1 - \delta_d^2 \tag{95}$$

For the condition of minimum damping time, the efficiency is 99.8 per cent. This figure was derived on the basis that no energy was left in the *PFN* at the beginning of the damping period.

COMPOSITE ANALYSIS OF THE INPUT SECTION AND TRANSFER CHAIN

In Part I it was explained that in order to obtain pulse shortening successive thyractors in the transfer chain must have descending values of saturating flux linkages. Part of these flux linkages in any thyractor

are required for the charging of the input capacitor of the particular transfer section associated with the thyractor. The remainder provide the thyractor guard interval. In the first section the amount of flux linkages required for the guard interval may be quite small compared to those needed for the input charging time. However, if the successive capacitance ratios are such that increasing guard intervals must be employed in order to operate without bias, the proportion of flux linkages needed for successive guard intervals becomes larger and larger. Eventually, it will become necessary to increase the saturating flux linkages of successive thyractors, and hence not pulse shortening but pulse lengthening will result.

$$\overline{i_h} = \frac{C_h}{T}(V_h - \nu_h),$$

$$\overline{i_h} = \frac{C_j}{T}(V_j - \nu_j) = \overline{i_j} = \frac{C_j}{T}(V_j - \nu_j),$$

$$\overline{i_j} = \frac{C_k}{T}(V_k - \nu_k) = \overline{i_k} = \frac{C_k}{T}(V_k - \nu_k),$$

$$\overline{i_k} = \frac{C_l}{T}(V_l - \nu_l)$$

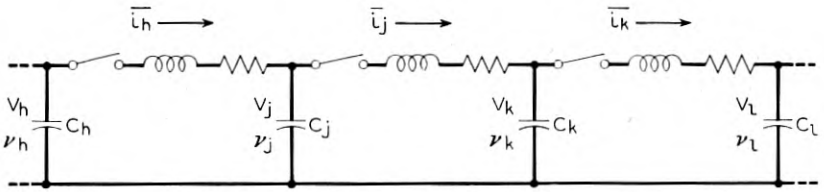


Fig. 28 — Illustration of current continuity by extension of the average current relations of equations (82) and (83).

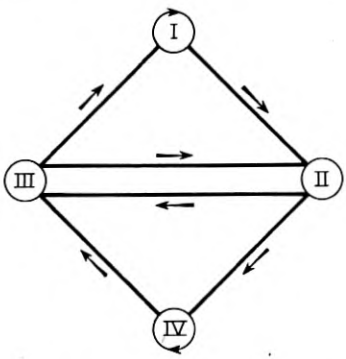


Fig. 29 — Representation of the possible operating regions of successive transfer sections of the transfer chain.

These conditions occur when the sections have capacitance ratios that are greater than their loss factors. The only advantage in such operation would be to obtain the peak voltage amplification which results when this ratio is increased to a value greater than the reciprocal of the loss factor. This further increase results in restricting the pulse shortening to a smaller number of sections if bias free operation is still to be maintained. On the other hand, if the capacitance ratios are smaller than the loss factors, there is no theoretical limit on the number of sections in which pulse shortening can be realized since guard interval shortening is possible. However, the required voltage amplification would then have to be provided by a pulse transformer.

From the average current relations of (51), (82) and (83), it is observed that current continuity through the successive sections does not depend upon the values of δ and λ. Fig. 28 illustrates this point. However, since the peak and quiescent voltages out of a section must equal the peak and quiescent voltages into the next section, there are restrictions on the regions in which two adjacent sections may operate. These restrictions can be seen by inspecting either Table I or Fig. 25. Fig. 29 indicates all the possible transfer chains in the following manner: any number of successive sections up to an entire chain that all operate in either region I or IV are possible. This is indicated by the arrowhead on the circle inscribing the Roman numerals I and IV in the figure. Beyond these possibilities, successive sections must be operating in the regions indicated by the arrows. For example, a section operating in region II must be followed by one that operates either in region III or IV.

When these voltage restrictions are applied to a modulator consisting of an input section and a transfer chain of n sections, the following relations result:

$$\begin{bmatrix} V_i \\ \bar{i}_i \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} V_1 \\ \nu_1 \end{bmatrix} \tag{96}$$

$$\begin{bmatrix} V_1 \\ \nu_1 \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \begin{bmatrix} V_2 \\ \nu_2 \end{bmatrix} \tag{97}$$

. . . . . .

$$\begin{bmatrix} V_j \\ \nu_j \end{bmatrix} = \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \begin{bmatrix} V_k \\ \nu_k \end{bmatrix} \tag{98}$$

. . . . . .

$$\begin{bmatrix} V_n \\ \nu_n \end{bmatrix} = \begin{bmatrix} {}_n & b_n \\ c_n & d_n \end{bmatrix} \begin{bmatrix} V_0 \\ \nu_0 \end{bmatrix} \tag{99}$$

from which follows:

$$\begin{bmatrix} V_1 \\ \nu_1 \end{bmatrix} = \begin{bmatrix} a_{1n} & b_{1n} \\ c_{1n} & d_{1n} \end{bmatrix} \begin{bmatrix} V_0 \\ \nu_0 \end{bmatrix} \tag{100}$$

and

$$\begin{bmatrix} V_i \\ \bar{i}_i \end{bmatrix} = \begin{bmatrix} a_{in} & b_{in} \\ c_{in} & d_{in} \end{bmatrix} \begin{bmatrix} V_0 \\ \nu_0 \end{bmatrix} \tag{101}$$

where

$$\begin{bmatrix} a_{1n} & b_{1n} \\ c_{1n} & d_{1n} \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \cdots \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \cdots \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \tag{102}$$

and

$$\begin{bmatrix} a_{in} & b_{in} \\ c_{in} & d_{in} \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} a_{1n} & b_{1n} \\ c_{1n} & d_{1n} \end{bmatrix} \tag{103}$$

Equation (100), which relates the input and output voltages of the transfer chain, can also be used to determine the voltages on any capacitor if the conditions on any one capacitor are known. This is readily accomplished by letting n equal the number of sections that separate the two capacitors in question. The relation of (101) is of particular interest since the over-all operation is completely defined by the network parameters alone. If either the input or output conditions are given, the other is determined.

Equation (101) was derived on the basis that energy is stored in the capacitors during inactive periods and that the thyractors do not experience secondary saturations. Experimental evidence has verified that a modulator whose capacitors are initially discharged can build up to these stable conditions. But this cannot easily be predicted, hence empirical methods must be used to ascertain that reasonable guard intervals exist during this initial build-up period.

Obviously, the special case that was originally discussed by Melville and was treated in Part II in which all quiescent voltages are zero will operate stably. As can readily be seen from Fig. 24, this requires that the capacitance ratio of each section be equal to the section loss factor. However, since such conditions do not facilitate pulse shortening when bias-free operation is desired, smaller capacitance ratios with the resulting quiescent voltages should be employed.

Since a number of transformers with any reasonable turns ratio can

be used, the voltage level in any portion of the modulator is not rigidly fixed by the input and output requirements. In choosing these levels, paramount consideration should be given to the operational requirements imposed upon the active switching device, such as the peak inverse voltage, peak current, duty cycle, etc., so that long life may be realized. Since the size of the capacitors is primarily determined by the energy per pulse and the size of the thyractors by this, the over-all pulse shortening and the number of transfer sections, the voltage level affects these elements very little unless it is increased to a point where the end margins must be larger than those required for purely mechanical reasons. This may become important when the required output pulse is greater than 10,000 volts. It should be noted that although a larger number of sections may radically reduce the total amount of core material, and hence the losses, the size of the modulator will increase considerably since additional capacitors are required. These factors make any general theoretical attempt to ascertain the optimum design of little value. Present design experience on high-power modulators has indicated that usually no more than three or four transfer sections are required.

COMPOSITE ANALYSIS OF THE HOMOGENEOUS GEOMETRIC CASE

A homogeneous geometric modulator, that is a modulator composed of sections that have identical loss factors and capacitance ratios, can be analyzed with much less difficulty than a completely general modulator wherein each section is different and distinct. Such an analysis may be useful as a first approximation to the response of any geometric modulator provided that the loss factor of the homogeneous modulator is judiciously chosen. Furthermore, it will indicate the best performance attainable when the loss factor is made unity, and the methods outlined will serve as a guide for analyzing any modulator.

Consider a homogeneous geometric modulator composed of an input section and n transfer sections. For such a modulator[9]

$$\delta_i = \delta_1 = \delta_2 = \cdots \delta_n = \delta \tag{104}$$

and

$$\lambda_1 = \lambda_2 = \cdots \lambda_n = \lambda \tag{105}$$

Since all transfer sections will now have the same network constants,

[9] The value of any capacitor and the sum of all the capacitance in the chain can be explicitly written in terms of either the input or output capacitor by employing the well-known relations derived for a geometric progression.

abcd, (102) can be written in the following form:

$$\begin{bmatrix} a_{1n} & b_{1n} \\ c_{1n} & d_{1n} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^n \tag{106}$$

From equations (90) through (93) inclusive, it may be demonstrated that for this particular modulator:

$$a_{1n} = \frac{(1+\lambda)(\lambda^n - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} - \frac{\lambda^{n-1} - 1}{\lambda^n - \lambda^{n-1}} \tag{107}$$

$$b_{1n} = \frac{\lambda^n - 1}{\lambda^n - \lambda^{n-1}} - \frac{(1+\lambda)(\lambda^n - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} \tag{108}$$

$$c_{1n} = \frac{(1+\lambda)(\lambda^n - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} - \frac{\lambda^n - 1}{\lambda^{n+1} - \lambda^n} \tag{109}$$

$$d_{1n} = \frac{\lambda^{n+1} - 1}{\lambda^{n+1} - \lambda^n} - \frac{(1+\lambda)(\lambda^n - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} \tag{110}$$

It is interesting to note that the determinant of the square matrix composed of the above elements equals $1/\lambda^n$. In addition note that if λ is made equal to δ, c_{1n} and d_{1n} become zero and unity respectively. This means from equation (100) that the quiescent voltages on all capacitors are identical and will all be made zero if capacitor C_0 is completely discharged by the output and damping sections. This substantiates what has previously been deduced in the general discussion of the composite input section and transfer chain arrangement.

The final matrix multiplication indicated in equation (103) can be performed with the aid of (61) through (64) yielding:

$$a_{in} = \frac{\delta}{1+\delta} + \frac{(1-\delta)(\lambda^{n+1} - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} \tag{111}$$

$$b_{in} = \frac{\delta}{1+\delta} - \frac{(1-\delta)(\lambda^n - 1)}{(1+\delta)(\lambda^{n+1} - \lambda^n)} \tag{112}$$

$$c_{in} = C_0/T \tag{113}$$

$$d_{in} = -C_0/T \tag{114}$$

The determinant of the square matrix composed of the above elements is $-C_0/T$.

From (101) and the preceding results the efficiency, η, of the energy

transfer from the dc source to the output capacitor is

$$\eta = \frac{\frac{1}{2}(1+\delta)(\lambda^{n+1}-\lambda^{n})\left[1+\frac{\nu_0}{V_0}\right]}{(\lambda^{n+1}-1)\left[1+\delta\frac{\nu_0}{V_0}\right]-(\lambda^{n}-1)\left[\delta+\frac{\nu_0}{V_0}\right]} \tag{115}$$

For the special case previously discussed in which all quiescent voltages were zero, that is, the case which results when ν_0 equals zero and λ equals δ, the efficiency becomes

$$\eta = \tfrac{1}{2}\delta^{n}(\delta+1) \tag{116}$$

Part IV — Conclusions

The analyses of the ac- and dc-charged modulators in the two preceding parts, although approximate to some degree, do provide a reasonable understanding of observed performance. This work has indicated the manner in which automatic core resetting in both devices can be achieved in all thyractors but the first of the ac-charged arrangement. All the regions of operation of a section of the transfer chain have been explored and have yielded the possibility of obtaining voltage amplification. However, such operation limits the pulse shortening that can be attained in the modulator; consequently, a transformer is still employed to provide the necessary voltage step-up.

Three practical innovations, in addition to the automatic core resetting, have been suggested. First, in order to provide short duration pulses without undue complication of the pulse transformer, the output thyractor section has been placed on the load side of this transformer. Second, a damping thyractor is employed such that the residual energy stored in the modulator after the generation of the main RF pulse can be safely dissipated without causing RF after pulsing. Both of these innovations can also be applied to other pulse modulators. And last, a dc-charged arrangement that provides still another means of automatic core resetting is presented in which the thyratron that may be required can be operated with its cathode at ground potential.

Several experimental ac- and dc-charged magnetic modulators that embody the automatic resetting feature without pulse amplification in the transfer chain have been developed. Fig. 30 illustrates an ac-charged modulator that provides an output pulse length of less than 0.1 microsecond. This very short pulse length has been achieved by placing the output thyractor, which has a ¼ mil molybdenum permalloy tape core, on the load side of the pulse transformer. The damping thyractor prin-

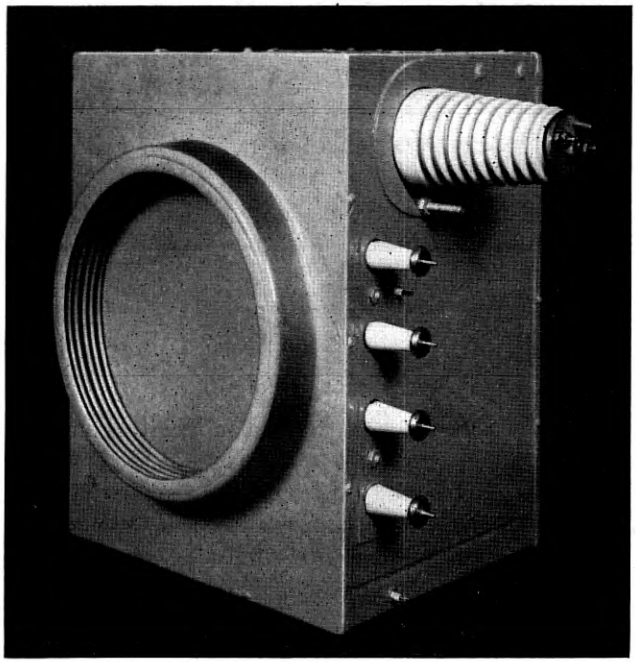


Fig. 30 — Experimental ac-charged magnetic modulator for a pulse duration of less than 0.1 microsecond.

ciple has been successfully employed in line type modulators as well as in these magnetic modulators. However, the grounded thyratron arrangement possible in the dc-charged case has not yet been incorporated in any design.

LIST OF SYMBOLS

	Quantity*
a	Section, chain or modulator constant
A	Area
b	Section, chain or modulator constant
B	Flux density
c	Section, chain or modulator constant
C	Capacitance
d	Section, chain or modulator constant
G	Guard interval
i	Instantaneous current

$\bar{i}$	Average current
I	Peak or constant current
L	Inductance
n	Number of sections in the transfer chain
N	Number of turns of wire
R	Resistance
t	Time
T	Pulse repetition period
v	Instantaneous voltage
V	Peak or constant voltage
$\underline{V}$	Normalized peak or constant voltage
W	Energy
X	Thyractor reference designation
α	Reciprocal of time constant
β	Constant, $\sqrt{\omega^2 + \alpha^2}$
δ	Loss factor
η	Efficiency
λ	Capacitance ratio
ν	Quiescent voltage
$\underline{\nu}$	Normalized quiescent voltage
τ	Time interval
ϕ	Magnetic flux
ω	Angular frequency

* All quantities are in rationalized MKS units.

CORRECTION

A. Uhlir, Jr., author of the paper *The Potentials of Infinite Systems of Sources and Numerical Problems in Semiconductor Engineering*, which appeared in the January, 1955, issue of the B.S.T.J., pages 105 to 128, has brought the following corrections to the attention of the editors.

In the last text sentence on page 107, for Q read P.

On page 124, equation (37) contains the term

$$-2 \ln \frac{\sinh \pi\lambda/k}{\pi\lambda/k}$$

This term should be

$$-2 \ln \frac{\sinh \pi\lambda/k}{2\pi/k}$$

CORRECTION

A. Uhlir, Jr., author of the paper *The Potentials of Infinite Systems of Sources and Numerical Problems in Semiconductor Engineering*, which appeared in the January, 1955, issue of the B.S.T.J., pages 105 to 128, has brought the following corrections to the attention of the editors.

In the last text sentence on page 107, for Q read P.

On page 124, equation (37) contains the term

$$-2 \ln \frac{\sinh \pi\lambda/k}{\pi\lambda/k}$$

This term should be

$$-2 \ln \frac{\sinh \pi\lambda/k}{2\pi/k}$$

Conversion of Maxwell's Equations into Generalized Telegraphist's Equations*

By S. A. SCHELKUNOFF

(Manuscript received May 3, 1955)

In this paper it is explained how Maxwell's field equations together with the appropriate boundary conditions may be converted into equations analogous to those for coupled transmission lines. This makes it possible to use the well-known techniques of dealing with transmission lines to solve certain field problems in those cases in which either the method of separating the variables fails or the boundary conditions are too complicated for the conventional method. For example, this method may be applied to studying waveguide to horn junctions, bending of waveguides, propagation of waves over an imperfect earth in the vicinity of the source, etc. Other applications are suggested in the course of the paper.

On the theoretical side, this conversion of field equations into transmission line equations brings together two heretofore independent theories of wave propagation on wires, namely, Lord Kelvin's theory based on circuit concepts and Kirchhoff's laws and Mie's theory based on field concepts and Maxwell's equations.

The "Generalized Telegraphist's Equations" derived in this paper differ from Kelvin's classical Telegraphist's Equations in two respects. Firstly, for a pair of conductors Kelvin obtained one pair of differential equations implying the existence of only one mode of propagation. For the same pair of conductors we obtain an infinite set of equations implying an infinite number of modes, from which Kelvin's equations are obtained by neglecting the

* The substance of this paper was presented at a joint meeting of the International Scientific Radio Union (U. S. A. National Committee), the Institute of Radio Engineers (Professional Groups on Antennas and Propagation, and Microwave Theory and Techniques), and American Geophysical Union (Section of Terrestrial Magnetism and Electricity) on May 4, 1954, Washington, D.C., under the title "Solution of Field Problems with the Aid of Distributed Circuit Parameter Concepts." Subsequently, it was presented at a Conference on Fields held on January 31, 1955 during the Winter General Meeting of the American Institute of Electrical Engineers in New York. This paper was also presented at the University of Bristol, England, on May 25, 1955, and at the Technische Hochschule in Zurich, Switzerland, on June 13, 1955.

coupling between the principal mode and the higher order modes. Secondly, our equations for some transmission structures contain additional "circuit parameters" which do not appear in the classical equations. These parameters are of the same nature as those in the corresponding equations for waveguides of uniform cross-section with perfectly conducting walls and filled with heterogeneous dielectric medium. In the present case they arise from the boundaries of conductors rather from lack of homogeneity.

The mathematics of converting Maxwell's Equations into Generalized Telegraphist's Equations is straightforward, although in the most general cases rather lengthy. The essential point is that a function, which for practical purposes is sufficiently arbitrary, may be represented in numerous ways by a series of orthogonal functions; and that when some such series are non-differentiable, the required relations between the coefficients of series representing the various field components may be obtained from Maxwell's equations by integration rather than by conventional substitution followed by differentiation.

CONTENTS

1. INTRODUCTION

For certain structures Maxwell's equations together with boundary conditions can be converted into exact or nearly exact equations similar to telegraphist's equations for coupled transmission lines. These structures include conventional dissipative wire transmission lines, dissipative

coaxial conductors, dissipative waveguides of either constant or variable cross-section, bent waveguides, plane and curved earth, etc. The coefficients in these equations play the role of "distributed circuit parameters;" but they are obtained from Maxwell's equations and boundary conditions rather than from consideration of static electric and magnetic fields. The distributed circuit parameters of some structures may be interpreted as distributed self and mutual series impedances and shunt admittances. But, in general, there are other distributed parameters which may be called "voltage and current transfer coefficients." The general equations are thus of the same form as the equations previously obtained by the author for waveguides of constant cross-section with perfectly conducting walls and filled with nonhomogeneous dielectric and magnetic media.

The possibility of converting Maxwell's equations into generalized telegraphist's is important from theoretical and practical points of view. This possibility removes a nagging feeling that the classical telegraphist's equations, useful as they are in practice, are fundamentally inconsistent with Maxwell's field theory. We shall find that they *are* consistent although approximate. We shall find that for conventional transmission lines, such as coaxial pairs, the generalized telegraphist's equations reduce to classical telegraphist's equations when the distributed coupling of the principal mode to the higher order modes is neglected. We also find that the classical equations can be used at much higher frequencies than one would expect from their conventional derivation based on the assumption of quasi-stationary fields. On the practical side, the generalized telegraphist's equations represent a method for solving boundary value problems using the well-known transmission line concepts and techniques. In a gentle waveguide to horn junction, for instance, we can obtain in the first approximation the transmission equations for the dominant mode and then calculate the higher order modes, generated by the expanding boundaries, as "crosstalk" between the dominant and higher order modes in the same way we calculate the crosstalk between adjacent conventional transmission lines in a cable. Thus we can look at three dimensional wave propagation from another angle, from the point of view of one dimensional propagation. We can also treat problems in curvilinear coordinates when the variables are not separable.

The conversion of Maxwell's equations into generalized telegraphist's equations brings together two independent theories of wave propagation, based on quite different concepts, which have merely "coexisted" for more than three quarters of a century. Lord Kelvin obtained his telegraphist's equations for cables[1] (transmission lines) ten years before

Maxwell formulated his field equations. In modern notation the telegraphist's or transmission line equations for two conductors, Fig. 1, are

$$\frac{\partial V^i}{\partial z} = -RI^i - L\frac{\partial I^i}{\partial t}, \qquad \frac{\partial I^i}{\partial z} = -GV^i - C\frac{\partial V^i}{\partial t} \tag{1}$$

where R, L, G, C are respectively the resistance, inductance, conductance and capacitance per unit length along the line. The dependent variables I^i, V^i are the instantaneous values of the current in one conductor and the transverse voltage from it to the other conductor. The distributed circuit parameters R, L, G, C are computed from static considerations. In computing R it is assumed that direct current is flowing in one conductor and returning via the other. The same assumption is made in computing the magnetic flux linkage per unit length and hence in computing L. In computing G a constant voltage is assumed to exist be-

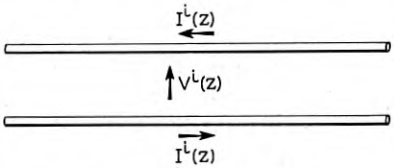


Fig. 1 — Parallel wires.

tween two conductors. The ratio of the resulting transverse direct current per unit length to this voltage is G. Finally, in computing the capacitance per unit length, C, it is assumed that $G = 0$ and that there is a constant voltage between the conductors. The ratio of the charge to this voltage is C.

On solving these equations for a sinusoidally varying applied voltage we find that the current and voltage are propagated with a finite velocity and that their amplitudes diminish exponentially with the distance from the generator. But in deriving these equations it has been assumed that these amplitudes are independent of the distance from the generator. Hence one would expect the equations to deteriorate steadily as the frequency increases. One derives the same impression from the point of view of Maxwell's theory. And yet experiments have shown that in many practical situations the errors are too small for detection even at very high frequencies. Since the "engineering theory," based on Kelvin's equations of Kirchhoff's type, is much simpler in practical applications than Maxwell's theory, it had continued to play the dominant role in electrical communication until the coming of radio and waveguides. To appreciate the difference in the "orders of complexity" of these two

theories one should glance at the forty-nine pages of Mie's paper on wave propagation along parallel wires[2] and compare them with the engineering solution of the same problem. In Mie's paper the reader is confronted with an elaborate and difficult mathematical analysis while the engineering solution is just a simple problem of elementary calculus. Mie's analysis is good only for an infinitely long pair of parallel metal cylinders imbedded in a homogeneous medium. On the other hand, the engineering solution applies to wires of variable cross-section, to twisted pairs of wires, to wires which are not straight and parallel, to wires insulated with layers of different media, to wires supported by insulators on poles — that is, to a wide range of cases in which an analysis based on Maxwell's field equations seems hopeless. On the other hand, there are problems of radiation whose solutions can readily be deduced from field equations and which apparently are not amenable to treatment with the aid of classical concepts of distributed circuit parameters.

Thus, the two theories have coexisted side by side but not on "speaking terms with each other." This situation has been one of continued challenge to students of electromagnetic theory. John R. Carson,[3] for instance, derived the classical telegraphist's equations from the Lorentz solution of Maxwell's equations in terms of retarded potentials and stated clearly the approximations he had to make. He then concluded that the accuracy of telegraphist's equations decreases with increasing frequency. Recently the author had an occasion to discuss the subject of this paper with A. Clavier. He informed me that many years ago when he taught electromagnetic theory at Ecolè Superieure d'Electricité, he became interested in the relation between Kirchhoff's type of theory of long lines and Maxwell's field theory. At that time he found that, in the case of simple geometry and no loss, the Lorentz solution of Maxwell's equations in terms of retarded potentials yielded a set of equations, identical in form with (1) but with a different meaning ascribed to V^i. The same result may be obtained directly from Maxwell's equations[4, 5] if in the case $R = 0$ we restrict ourselves to TEM waves, in which case V^i has the meaning identical with that ascribed to it by Lord Kelvin.

For continuously coupled transmission lines, for several parallel wires for instance, telegraphist's equations are

$$\frac{\partial V_m{}^i}{\partial z} = -\sum_n \left(R_{mn} I_n{}^i + L_{mn} \frac{\partial I_n{}^i}{\partial t} \right), \qquad \frac{\partial I_m{}^i}{\partial z} = -\sum_n \left(G V_{mn}{}^i + C_{mn} \frac{\partial V_n{}^i}{\partial t} \right) \tag{2}$$

where $I_m{}^i$ is the instantaneous current in the m-th line and $V_m{}^i$ the in-

stantaneous transverse voltage associated with the m-th line. The coefficients corresponding to $n = m$ are the distributed circuit parameters for the m-th line, and those corresponding to unequal values of m and n are the distributed coupling parameters for the m-th and n-th lines. For steady state these equations reduce to a system of ordinary differential equations. This reduction is accomplished by regarding the instantaneous voltages and currents, $V_m{}^i$ and $I_m{}^i$, as the real parts of complex voltages and currents, $V_m \exp(j\omega t)$ and $I_m \exp(j\omega t)$. Thus (2) is transformed into

$$\frac{dV_m}{dz} = -\sum_n Z_{mn} I_n , \qquad \frac{dI_m}{dz} = -\sum_n Y_{mn} V_n \tag{3}$$

where the distributed complex impedances per unit length, Z_{mn}, and complex admittances per unit length, Y_{mn}, are

$$Z_{mn} = R_{mn} + j\omega L_{mn} , \qquad Y_{mn} = G_{mn} + j\omega C_{mn} \tag{4}$$

The usefulness of (1) and (2) is severely restricted because even for relatively slowly varying currents the resistance R of a conductor is not independent of time. The voltage drop across a section of a conductor depends not only on the current but on the second and higher time derivatives of the current. It is for this reason that (1) is properly named "telegraphist's" rather than "telephonist's" equations. However, (3) may be used even at quite high frequencies provided we use ac resistances R_{mn}, which include the skin effect, in place of dc resistances. A similar allowance should be made for the internal inductances of the conductors.

It has been shown[6] that for each mode of propagation in a perfectly conducting waveguide of uniform cross-section it is possible to obtain equations analogous to telegraphist's equations. Thus for TM waves the steady state equations of propagation are

$$\frac{dV}{dz} = -\left(j\omega\mu + \frac{\chi^2}{g + j\omega\epsilon}\right) I, \qquad \frac{dI}{dz} = -(g + j\omega\epsilon) V \tag{5}$$

and for TE waves

$$\frac{dV}{dz} = -j\omega\mu I, \qquad \frac{dI}{dz} = -\left(g + j\omega\epsilon + \frac{\chi^2}{j\omega\mu}\right) V \tag{6}$$

where the constant χ depends on the shape and size of each conductor and on the field distribution in a typical transverse plane. The "voltage" V and the "current" I are related to the magnitudes of the transverse components of electric and magnetic intensities.

In this paper we shall be concerned primarily with the steady-state

equations. This entails no loss in generality because the Laplace transform method would enable us to find the more general solutions from the steady state solutions. It is possible, however, to convert such equations as above into forms applicable to non-periodic time variations in the dependent variables. Thus (6) would become

$$\frac{\partial V^i}{\partial z} = -\mu \frac{\partial I^i}{\partial t}, \qquad \frac{\partial I^i}{\partial z} = -gV^i + \epsilon \frac{\partial V^i}{\partial t} + \mu^{-1}\chi^2 \int_{-\infty}^{t} V^i(\tau)\, d\tau \tag{7}$$

The first equation of the set (5) can be transformed either into

$$g \frac{\partial V^i}{\partial z} + \epsilon \frac{\partial^2 V^i}{\partial t \partial z} = -g\mu \frac{\partial I^i}{\partial t} + \mu\epsilon \frac{\partial^2 I^i}{\partial t^2} - \chi^2 I^i \tag{8}$$

or into

$$\frac{\partial V^i}{\partial z} = -\mu \frac{\partial I^i}{\partial t} - \chi^2 \epsilon^{-1} \int_{-\infty}^{t} I^i(\tau) e^{-(g/\epsilon)(t-\tau)}\, d\tau \tag{9}$$

However, the steady state equations combined with Laplace transforms are, as a rule, more convenient for dealing with general time varying phenomena than the nonsteady state equations.

2. HEURISTIC DISCUSSION OF THE PROBLEM OF CONVERTING FIELD EQUATIONS INTO GENERALIZED TELEGRAPHIST'S EQUATIONS

Consider two coaxial conductors. If they are perfectly conducting, the field between them may be expressed in terms of TEM, TE, and TM modes. Each of these modes can exist independently of the others. Suppose now that we have excited a pure TEM mode. Let us then introduce a small resistive spot on one of the cylinders. Some current will flow across the spot and will give rise to a non-vanishing electric intensity tangential to the spot. This intensity will act as an impressed longitudinal intensity and will thus generate a large number of modes traveling in opposite directions from the spot. Let us introduce another spot, and then another and another until both cylinders are covered by resistive films. At each step various modes will be generated and regenerated. The argument suggests that we should be able to express the field *between* the imperfectly conducting cylinders in terms of modes appropriate to perfectly conducting cylinders. However, none of the latter modes can now exist independently of the others. The surface impedance of the cylinders provides continuous coupling between various modes.

In the case of imperfectly conducting cylinders the longitudinal electric intensity does not vanish at the surface of either conductor and

yet the above physical argument leads us to believe that we can express it in terms of functions which vanish there. Are we facing a contradiction? The answer is, no. The functions representing the longitudinal intensity over a given cross-section when the cylinders are perfectly conducting form a complete orthogonal set. This set is sufficient for representing arbitrary continuous functions, which do not vanish on the boundary of the cross-section, at all points *except on the boundary itself*. The situation is analogous to that existing in Fourier analysis. A function which is bounded and continuous in the closed interval $(0, \ell)$ may be represented by a sine series in an open interval even when the function does not vanish at the ends of the interval. However, the series will be non-uniformly convergent and non-differentiable. For this reason such series cannot be substituted in Maxwell's equations when differentiation is required. However, there is a way of overcoming this difficulty which can best be illustrated by an example. As far as the representation of the longitudinal electric intensity is concerned, we shall have one series for points in the interior of the waveguide and another on its boundary. The latter is obtained from the boundary condition, that is from the product of the surface impedance and the tangential magnetic intensity.

A waveguide with continuously varying cross-section may be regarded as the limit of a waveguide made up of a large number of very short waveguides with constant but different cross-sections. Consider only one sudden change in the cross-section. The effect of this discontinuity on a wave in one mode is to produce waves in many other modes traveling in opposite directions from the discontinuity. Hence, the discontinuity couples various modes and an expanding boundary represents continuous coupling. Bending also represents continuous coupling.

In some structures the modes of propagation will be spherical or systems of spherical and plane modes. Take for instance a perfectly conducting cone. There will be two systems of spherical modes of propagation, internal and external, completely independent of each other. If the perfectly conducting cone is replaced by a sheet of finite thickness and conductivity, there will exist a linear relation between electric and magnetic intensities tangential to the internal and external surfaces of the sheet; thus

$$\begin{aligned} E_{\tan}{}^{e} &= Z_{ee}H_{\tan}{}^{e} + Z_{ei}H_{\tan}{}^{i} \\ E_{\tan}{}^{i} &= Z_{ie}H_{\tan}{}^{e} + Z_{ii}H_{\tan}{}^{i} \end{aligned} \tag{10}$$

where the Z's are the surface and transfer impedances of the sheet. This equation expresses the coupling between external and internal waves. If

the conical conductor is deformed, further coupling arises from the deformation. The cone may be deformed into a cylinder, in which case the external waves will still be spherical while the internal waves will become plane.

We can make our calculations of fields step by step as suggested by the heuristic argument. For example, we can calculate the scattering from a typical resistive spot and integrate the scattered field over a continuous distribution of spots. Since we would be neglecting the second order scattering, our result would be approximately true only for a sufficiently small perturbation of the original field. This is the method used by S. P. Morgan[8] to obtain mode conversion losses in transmission of circular electric waves through slightly non-cylindrical guides. If the first order perturbation is not good enough, one presumably could calculate higher order perturbations. However, this direct method, although very useful in some situations, has its limitations. For instance, no matter how small is the dissipation, the amplitude of the wave will be attenuated with the increasing distance from the source while the amplitude of the wave "unperturbed" by the resistance would have remained constant.

In the next section we shall state the generalized telegraphist's equations. The remainder of the paper will be devoted to the mathematical technique of obtaining them from Maxwell's equations. This technique is simple in principle but in general cases requires rather lengthy mathematical manipulation which might obscur the main ideas. For this reason the technique will be illustrated by a series of simple examples.

3. THE FORM OF GENERALIZED TELEGRAPHIST'S EQUATIONS

In a previous paper[7] we obtained from Maxwell's equations the following equations for waveguides of uniform cross-section, bounded by perfectly conducting walls, and filled with nonhomogeneous dielectric and magnetic media

$$\begin{aligned} \frac{dV_m}{dz} &= -\sum_n Z_{mn} I_n - \sum_n {}^{V}T_{mn} V_n \\ \frac{dI_m}{dz} &= -\sum_n Y_{mn} V_n - \sum_n {}^{I}T_{mn} I_n \end{aligned} \tag{11}$$

The equations which we shall obtain in this paper are of the same form. They are more general than the classical telegraphist's equations, given in (3), for conventional transmission lines since, in addition to distributed series impedances Z_{mn} and shunt admittances Y_{mn}, (11) contains "voltage transfer coefficients" ${}^{V}T_{mn}$ and "current transfer coefficients" ${}^{I}T_{mn}$.

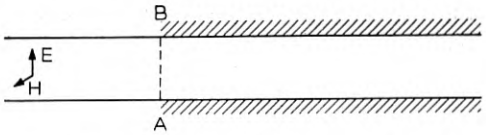


Fig. 2 — Two parallel planes, perfectly conducting to the left of the junction AB and imperfectly conducting to the right of it.

The voltages, V_m, and currents, I_m, are related to the amplitudes of electric and magnetic intensities associated with each particular mode. To each pair, V_m and I_m, there corresponds a certain field pattern in the transverse plane. The choice of these field patterns is essentially arbitrary.* Often the most convenient choice is the one for which the mutual coefficients are as small as possible so that the corresponding modes are as independent as possible. But this is not always the case. For example, take a junction between a pair of perfectly conducting parallel planes and a pair of imperfectly conducting planes, Fig. 2. Consider what happens near the junction AB when the TEM mode is traveling from the left toward the junction. In this mode E is constant in the vertical direction. Between the imperfectly conducting planes we can represent the entire field in terms of certain independent modes by solving the appropriate boundary value problem.[9] If the distance between the planes is sufficiently large in comparison with wavelength, there is no mode in which E is either constant in the vertical direction or nearly constant. No matter how small is the surface resistance of the planes (as long as it is different from zero), by spreading the planes we can reach a condition in which the vertical electric intensity is distributed almost sinusoidally with height, the maximum occurs half way between the planes, and the minima near the planes. It is quite evident that these modes are not the best for representing the field near the junction. From physical considerations we expect that after the TEM wave enters the space between the imperfectly conducting planes, it is still the same wave for considerable distance except near the planes. If we expand the field to the right of AB in terms of modes appropriate to perfectly conducting planes, we *will* have a mode with constant vertical electric intensity. This mode will be feebly coupled to higher order modes. On account of this feeble coupling the field near the junction is not much different from that which would exist between perfectly conducting planes. However, under the postulated conditions there are many higher order modes which travel with almost the same velocity as the principal mode. For this reason the conversion from the principal mode to these higher order

* Just as arbitrary as the choice of "meshes" in writing Kirchhoff's equations.

modes will be cumulative and eventually the transverse field pattern will become totally unlike that near the junction. This final pattern is best obtained by solving the conventional boundary value problem; but there is a large region near the junction where the representation in terms of modes appropriate to perfectly conducting planes is much more practical.

In some instances, particularly when the mutual impedances and admittances are small and transfer coefficients vanish, it is possible to calculate the mutual coefficients from the power flow along the guide and the power absorbed by its wall. In the middle thirties this author obtained in this way the coupling between TM and TE waves in dissipative cylindrical waveguides. The result agreed with that obtained from the appropriate characteristic equation for hybrid waves (unpublished work). Much more important was the application of this idea by W. J. Albersheim[10] to the propagation of circular electric waves round a bend. It is equally possible to obtain small coupling coefficients due to small irregularities in the dielectric medium from the unperturbed field which would exist if these irregularities were removed by calculating the response to the relative polarization currents.

In general, (11) is simpler to work with than the original Maxwell's equations. In particular, when the mutual coefficients (those corresponding to the unequal subscripts) are small, we can solve the equations by successive approximations as in problems of cross-talk between conventional transmission lines in a cable. That is, we first neglect the coupling between various modes and obtain the first approximate solution. Then we calculate the voltages and currents induced from each mode into every other mode in a typical element of length dz (that is, $Z_{mn}I_n^{(0)}\,dz$, $Y_{mn}V_n^{(0)}\,dz$, etc. where $I_n^{(0)}$ and $V_n^{(0)}$ represent the first approximations). These induced voltages and currents we regard as impressed voltages and currents exciting waves in the corresponding modes. The effects of these impressed voltages and currents are then obtained by integration. This process can be repeated indefinitely. But usually the second approximation is sufficient for practical purposes.

However, if two or more modes have the same propagation constant, we have a situation analogous to that existing in directional couplers. No matter how small is the coupling, all power may pass from one mode to the other. In this case, on account of the coupling the common propagation constant will be split into several nearly equal propagation constants.

In concluding this section we would like to call the reader's attention to a rather curious situation which existed before the present derivation

of generalized telegraphist's equations from Maxwell's field equations. The conventional derivation of classical telegraphist's equations (3) led one to expect that in general the mutual distributed parameters will differ from zero. The *independent* modes of propagation are obtained *after* these equations have been solved. On the other hand, in each case in which telegraphist's equations, such as (5) and (6), were obtained from Maxwell's equations, the modes were invariably independent. Obviously, this independence of modes was due to the fact that selected situations were rather trivial: Maxwell's equations were separable in the chosen coordinates and the boundary conditions were particularly simple. The independence was purely accidental, inherent in the popular method of solving Maxwell's equations, and limited to the problems which could be handled by that method.

4. UNIFORM STRIP TRANSMISSION LINES — THE PRINCIPAL MODE

The simplest mode of propagation between perfectly conducting parallel plane sheets is the TEM mode in which the electric lines of force are normal to the planes and the magnetic lines are parallel to them. Let us assume that the x and y axes are parallel respectively to electric and magnetic lines. The field of this mode will then be independent of the y coordinate, and Maxwell's equations reduce to

$$\frac{\partial E_x}{\partial z} = -j\omega\mu H_y + \frac{\partial E_z}{\partial x}, \qquad \frac{\partial H_y}{\partial z} = -(g + j\omega\epsilon)E_x \tag{12}$$

$$E_z = \frac{1}{g + j\omega\epsilon}\frac{\partial H_y}{\partial x} \tag{13}$$

For the mode under consideration E_z vanishes identically and therefore H_y and E_x are independent of the x coordinate as well. Essentially the same situation will exist if we cut the planes as shown in Fig. 3 to form a strip transmission line with "guards" to keep the field from spreading into the outer space.

If the sheets are not perfectly conducting, E_z does not vanish on their surface but is proportional to the linear current densities, that is, to the tangential magnetic intensities

$$E_z(0, z) = Z_1 H_y(0, z), \qquad E_z(a, z) = -Z_2 H_y(a, z) \tag{14}$$

The coefficients Z_1 and Z_2 are the surface impedances of the sheets. Hence, E_z will not vanish between the sheets. From the heuristic argument expounded in Section 2 we attribute this effect of finite conductivity to the production of higher modes of propagation. For good

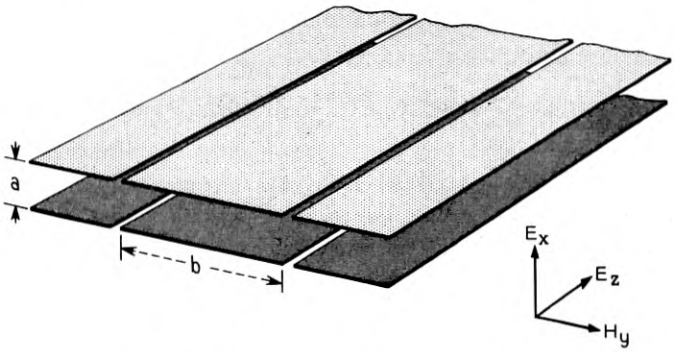


Fig. 3 — Uniform strip transmission line.

conductors Z_1 and Z_2 are extremely small so that while E_z cannot vanish, it can be very small. This "almost TEM" mode is often called the principal mode.

The transverse components of the field between and on the impedance strips will now be expressed in the following form

$$\begin{aligned} E_x &= \frac{V_0(z)}{a} + \sum N_n^{-1} V_n(z) \cos \frac{n\pi x}{a} \\ H_y &= \frac{I_0(z)}{b} + \sum N_n^{-1} I_n(z) \cos \frac{n\pi x}{a} \end{aligned} \tag{15}$$

where the summation index assumes all integral values from 1 to ∞. The normalization factors

$$N_n^{\,2} = \int_0^a \int_0^b \cos^2 \frac{n\pi x}{a}\, dx\, dy = \frac{1}{2}\, ab \tag{16}$$

are chosen to make the expression for the power flow identical with that for a multiple-conductor conventional transmission line, that is,

$$P = \frac{1}{2} \int_0^a \int_0^b E_x H_y^*\, dx\, dy = \frac{1}{2}\, V_0(z) I_0^*(z) + \frac{1}{2} \sum V_m(z) I_m^*(z) \tag{17}$$

When the strips are perfectly conducting the voltages and currents, $V_n(z)$ and $I_n(z)$, are independent of each other; otherwise, they are not. From the purely mathematical point of view we can regard expressions (15) as representations of the solutions between the impedance strips by cosine series. Such representations exist because E_x and H_y are continuous functions of x in the *closed* interval $(0, a)$.

Similarly, we represent the longitudinal electric intensity by a sine series

$$E_z = \sum e_n(z) \sin \frac{n\pi x}{a}, \qquad n = 1, 2, 3, \cdots \tag{18}$$

When the boundaries are not perfectly conducting, such a representation is possible only in the *open* interval $0 < x < a$ since the sine terms vanish at the ends of the interval while E_z does not. *On* the boundaries we use (14) and (15),

$$\begin{aligned} E_z(0, z) &= \frac{Z_1}{b} I_0(z) + \sum Z_1 N_n^{-1} I_n(z) \\ E_z(a, z) &= \frac{Z_2}{b} I_0(z) + \sum (-)^{n+1} Z_2 N_n^{-1} I_n(z) \end{aligned} \tag{19}$$

In the closed interval the series (18) represents a discontinuous function and therefore does not converge uniformly. Moreover, its coefficients diminish so slowly that after term by term differentiation, the series will diverge. Hence, we may not substitute this series in the first equation of the set (12) in order to obtain the relations between $V_n(z)$, $I_n(z)$, and $e_n(z)$ in the usual way. There is another way, however.

To obtain the equations for the principal mode we merely integrate (12) with respect to x from 0 to a and note that

$$\int_0^a E_x \, dx = V_0(z), \qquad \int_0^a H_y \, dz = \frac{a}{b} I_0(z) \tag{20}$$

Thus we find

$$\begin{aligned} \frac{dV_0(z)}{dz} &= -\frac{j\omega\mu a}{b} I_0(z) + E_z(a, z) - E_z(0, z) \\ \frac{a}{b} \frac{dI_0(z)}{dz} &= -(g + j\omega\epsilon) V_0(z) \end{aligned} \tag{21}$$

We now substitute from (19) into (21),

$$\begin{aligned} \frac{dV_0(z)}{dz} &= -\left(\frac{j\omega\mu a}{b} + \frac{Z_1 + Z_2}{b}\right) I_0(z) - \sum \frac{Z_1 + (-)^n Z_2}{N_n} I_n(z) \\ \frac{dI_0(z)}{dz} &= -\frac{(g + j\omega\epsilon)b}{a} V_0(z), \end{aligned} \tag{22}$$

where the summation extends over the sequence $n = 1,2,3, \cdots$. The classical form of telegraphist's equations is obtained if we neglect the

summation, that is, the coupling of the principal mode to the higher order modes.

It is worth noting that (18) for the longitudinal electric intensity has not been used.

5. UNIFORM STRIP TRANSMISSION LINES — HIGHER ORDER MODES

Telegraphist's equations for the typical higher order mode will be obtained if we multiply equations (12) by $N_m^{-1} \cos (m\pi x/a)\, dxdy$ and integrate over the cross-section of the strip line. Thus, we have

$$\int_0^a \int_0^b \frac{\partial E_x}{\partial z} N_m^{-1} \cos \frac{m\pi x}{a}\, dx\, dy = -j\omega\mu \int_0^a \int_0^b H_y N_m^{-1} \cos \frac{m\pi x}{a}\, dx\, dy + \int_0^a \int_0^b N_m^{-1} \cos \frac{m\pi x}{a} \frac{\partial E_z}{\partial x}\, dx\, dy \tag{23}$$

In the first and second terms of this equation we substitute from (15). The last term we integrate by parts. Thus we find

$$\frac{dV_m(z)}{dz} = -j\omega\mu I_m(z) - bN_m^{-1}[E_z(0, z) + (-)^{m+1} E_z(a, z)] + \int_0^a \int_0^b \frac{m\pi}{aN_m} E_z \sin \frac{m\pi x}{a}\, dx\, dy \tag{24}$$

To evaluate the last term we substitute from (13), integrate once more by parts, and substitute from (15),

$$\int_0^a \int_0^b \frac{m\pi}{aN_m} E_z \sin \frac{m\pi x}{a}\, dx\, dy = \frac{m\pi}{(g + j\omega\epsilon)aN_m} \int_0^a \int_0^b \frac{\partial H_y}{\partial x} \sin \frac{m\pi x}{a}\, dx\, dy = \frac{m\pi}{(g + j\omega\epsilon)aN_m} \left[bH_y(x, z) \sin \frac{m\pi x}{a} \Big|_0^a - \frac{m\pi}{a} \int_0^a \int_0^b H_y \cos \frac{m\pi x}{a}\, dx\, dy \right] = -\frac{m^2\pi^2}{(g + j\omega\epsilon)a^2} I_m(z) \tag{25}$$

In view of this and (19), (24) becomes

$$\frac{dV_m(z)}{dz} = -\left[j\omega\mu + \frac{m^2\pi^2}{(g + j\omega\epsilon)a^2} + \frac{(Z_1 + Z_2)b}{N_m^2} \right] I_m(z) + \frac{Z_1 + (-)^m Z_2}{N_m} I_0(z) + \sum{}' \frac{[Z_1 + (-)^{m+n} Z_2]b}{N_m N_n} I_n(z), \tag{26}$$

$$m = 1, 2, 3, \cdots$$

where the prime after the summation sign indicates that the summation is to be extended over the sequence $n = 1,2,3, \cdots$ *except* $n = m$.

Similarly, we obtain from the second equation of the set (12)

$$\frac{dI_m(z)}{dz} = -(g + j\omega\epsilon)V_m(z), \quad m = 1, 2, 3, \cdots \tag{27}$$

Again it should be noted that in the above derivation *we have not used* the non-uniformly convergent series (18) for the longitudinal electric intensity. We could have used it. In that case, however, we would have been faced with the necessity of justifying certain steps. There is a theorem[12] to the effect that a uniformly convergent series may be integrated term by term. But the series (18) is not uniformly convergent. Hence, if we substitute from (18) in the last term of (24), we would have to prove that in this special instance the term by term integration is permissible. Actually the non-uniform convergence is only *sufficient* condition for term by term integration and *not a necessary* condition. Even the examples given in Reference 12 to show that some non-uniformly convergent series may not be integrated term by term in certain closed intervals are somewhat misleading without an explicit qualification; for it so happens that these series may be integrated term by term in slightly smaller intervals and correct results then obtained by passing to the limit. Nevertheless in the present case there is no reason why we should have complicated our derivation by using steps requiring special justification.

To obtain the longitudinal electric intensity we substitute from (15) and (18) in (13) and differentiate term by term. This differentiation is permissible if the series of derivatives is uniformly convergent. In the present case this means that the differentiation should be restricted to an open interval $0 < x < a$. Thus

$$e_n(z) = -\frac{n\pi}{(g + j\omega\epsilon)aN_n} \tag{28}$$

and

$$E_z = -\sum' \frac{n\pi}{(g + j\omega\epsilon)aN_n} I_n(z) \sin\frac{n\pi x}{a} \tag{29}$$

For $x = 0$, a E_z may be obtained from (19). Very near the boundaries the series (29) converges very slowly. However, we know that E_z is very small there and normally we would not be interested in it. If we are, the best way to find it is by interpolation from the boundary values (19) and the interior values sufficiently far from the boundaries where the con-

vergence of (29) is more satisfactory. The slow convergence of (29) near the boundaries does not affect, of course, the validity of our telegraphist's equations.

6. STRIP TRANSMISSION LINE WITH VARIABLE CROSS-SECTION — THE PRINCIPAL MODE

In this section we shall consider strip transmission lines with variable cross-section, Fig. 4, which exemplify horns and waveguide to horn junctions. Here we can use either cartesian coordinates or curvilinear while in the parallel plane case the former seemed obviously the most

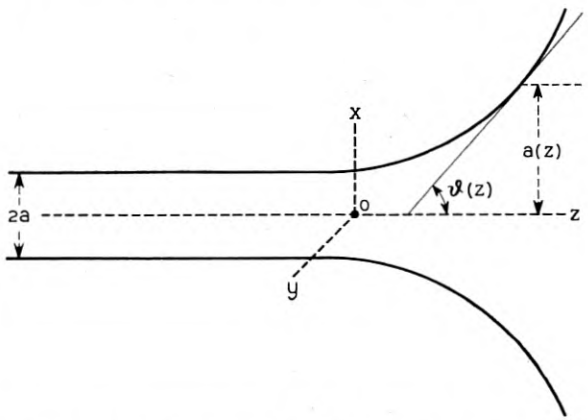


Fig. 4 — Strip transmission line with variable cross-section.

appropriate. It appears that cartesian coordinates are still the most convenient when the shape of the boundaries is arbitrary; in a subsequent section, however, we shall consider an example of curvilinear coordinates.

For the sake of simplicity we shall confine ourselves to the symmetric transmission line, both in geometry and in the impedance of the strips (that is, we shall assume $Z_2 = Z_1$). In the case of symmetric mode we can then insert a perfectly conducting plane $y0z$ in the middle of the strip transmission line without disturbing the field.* Hence the boundary conditions will be

$$E_z(0, z) = 0, \qquad E_t(a, z) = -ZH_y(a, z) \tag{30}$$

where Z is the surface impedance of the upper strip and E_t is the com-

* In the case of antisymmetric modes we can introduce an infinite impedance sheet.

ponent of electric intensity tangential to the strip. Since

$$E_t = E_z \cos \vartheta + E_x \sin \vartheta \tag{31}$$

where $\vartheta(z)$ is the angle between the axial plane and the plane tangent to the strip, the second boundary condition becomes

$$E_z(a, z) = -E_x \tan \vartheta - Z \sec \vartheta H_y(a, z) \tag{32}$$

In line with the heuristic argument propounded in Section 2, we shall express the field in the present strip line with variable distance $a(z)$ between the strips in terms of modes appropriate to the case of constant a and perfectly conducting boundaries. From the mathematical point of view this amounts to expanding the field intensities in a typical transverse plane in Fourier series in x. The coefficients of the series are to be determined from Maxwell's equations and boundary conditions. Thus we shall express E_x and H_y by series (15) and E_z by (18). The latter expression will hold only for $x < a$. When $x = a$, we find from (15) and (32)

$$\begin{aligned} E_z(a, z) = &- \frac{\tan \vartheta}{a} V_0(z) - \sum (-)^n N_n^{-1} \tan \vartheta \, V_n(z) \\ &- \frac{Z \sec \vartheta}{b} I_0(z) - \sum (-)^n N_n^{-1} Z \sec \vartheta \, I_n(z) \end{aligned} \tag{33}$$

To obtain the equations for principal waves we proceed as in Section 4 and integrate (12) with respect to x, taking into consideration (15). Thus we obtain (21). Then we substitute from (33) into (21),

$$\begin{aligned} \frac{dV_0(z)}{dz} = &- \left(\frac{j\omega\mu a}{b} + \frac{Z \sec \vartheta}{b} \right) I_0(z) - \frac{\tan \vartheta}{a} V_0(z) \\ &- \sum (-)^n N_n^{-1} Z \sec \vartheta \, I_n(z) - \sum (-)^n N_n^{-1} \tan \vartheta \, V_n(z) \end{aligned} \tag{34}$$

$$\frac{dI_0(z)}{dz} = - \frac{(g + j\omega\epsilon)b}{a} V_0(z)$$

Note the appearance of voltage transfer coefficients in addition to the mutual series impedances.

7. STRIP TRANSMISSION LINE WITH VARIABLE CROSS-SECTION — HIGHER ORDER MODES

The telegraphist's equations for higher order modes are obtained in much the same way as in Section 5. We must only remember that a is a function of z. The variation of a with z will introduce extra terms in our

final equations. Thus if we multiply the first equation of the set (12) by $N_n^{-1} \cos n\pi x/a$ and integrate over the cross-section of the line, we find first

$$\int_0^a \int_0^b \frac{\partial E_x}{dz} N_m^{-1} \cos \frac{m\pi x}{a}\, dx\, dy = -\left[j\omega\mu + \frac{m^2}{(g + j\omega\epsilon)a^2}\right] I_m(z) + b(-)^m N_m^{-1} E_z(a, z) \tag{35}$$

Differentiating the series for E_x as given by (15), we obtain

$$\frac{\partial E_x}{\partial z} = \frac{1}{a}\frac{dV_0(z)}{dz} - \frac{a'(z)}{a^2} V_0(z) + \sum \left[\frac{dV_n}{dz} N_n^{-1} \cos \frac{n\pi x}{a} - \frac{N_n'(z)}{N_n^2} V_n \cos \frac{n\pi x}{a} + \frac{n\pi x a'(z)}{N_n a^2} V_n \sin \frac{n\pi x}{a}\right] \tag{36}$$

Substituting from (36) in (35) and using (33), we have

$$\frac{dV_m}{dz} = -Z_{mm} I_n - \sum{}' Z_{mn} I_n - \sum {}^{v}T_{mn} V_n \tag{37}$$

where the prime denotes that the summation is extended over the sequence $n = 1,2,3, \cdots$ except $n = m$, and

$$\begin{aligned}
Z_{mn} &= j\omega\mu + \frac{m^2\pi^2}{(g + j\omega\epsilon)a^2} + bN_m^{-2} Z \sec\vartheta \\
Z_{mn} &= (-)^{m+n} N_m^{-1} N_n^{-1} bZ \sec\vartheta \quad \text{if} \quad m \neq n,\ m \neq 0,\ n \neq 0 \\
{}^{v}T_{mn} &= (-)^{m+n} N_m^{-1} N_n^{-1} b \tan\vartheta \\
&\quad - \frac{n\pi a'(z)}{N_m N_n a^2} \int_0^a \int_0^b x \cos \frac{m\pi x}{a} \sin \frac{n\pi x}{a}\, dx\, dy \\
&\qquad\qquad m \neq n, \quad m \neq 0, \quad n \neq 0,
\end{aligned} \tag{38}$$

$$\begin{aligned}
{}^{v}T_{mm} &= N_m^{-2} b \tan\vartheta + N_m' N_m^{-1} - \frac{m\pi a'(z) b}{2N_m^2 a^2} \int_0^a x \sin \frac{2m\pi x}{a}\, dx \\
Z_{m0} &= (-)^m Z N_m^{-1} \sec\vartheta, \quad m \neq 0 \\
{}^{v}T_{m0} &= \frac{(-)^m b \tan\vartheta}{N_m a}
\end{aligned}$$

Similarly we find

$$\frac{dI_m}{dz} = -(g + j\omega\epsilon)V_m - \sum {}^{I}T_{mn}I_n \tag{39}$$

where

$$\begin{aligned} {}^{I}T_{mn} &= -\frac{n\pi a'(z)}{a^2 N_m N_n}\int_0^a\int_0 x\cos\frac{m\pi x}{a}\sin\frac{n\pi x}{a}\,dx\,dy, \\ &\qquad m \neq n, \quad m \neq 0, \quad n \neq 0 \\ {}^{I}T_{mm} &= N_m' N_m^{-1} - \frac{m\pi a'(z)}{2a^2 N_m^2}\int_0^a\int_0^b x\sin\frac{2m\pi x}{a}\,dx\,dy \\ {}^{I}T_{m0} &= 0 \end{aligned} \tag{40}$$

8. BENT STRIP TRANSMISSION LINES — THE PRINCIPAL MODE

Let us now suppose that the strip transmission line (with the guard strips) shown in Fig. 3 is bent uniformly in the xz plane. After bending, the x lines will be radii emerging from the axis of bending, the y lines will be straight and parallel to the axis, and the z lines will be circular arcs coaxial with the axis. The section of this structure by the plane $y = 0$ is shown in Fig. 5. The curved z axis of the bent coordinate system

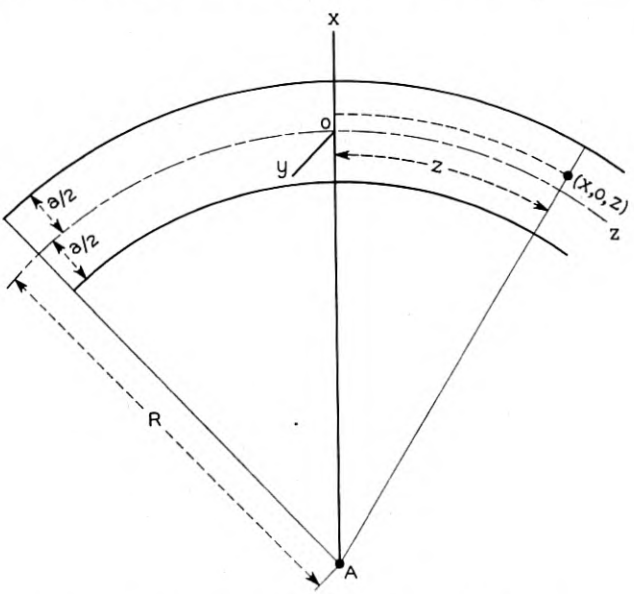


Fig. 5 — Uniformly bent strip transmission line.

will be chosen half way between the strips. The "distance" z between the radial xz planes will be measured along this curved z axis. The coordinate x is the shortest distance between the given point and the $y0z$ coordinate surface. The differential distance between two points will then be

$$ds^2 = dx^2 + dy^2 + \left(1 + \frac{x}{R}\right)^2 dz^2 \tag{41}$$

where R is the radius of curvature of the z axis and is, in general, a function of z. The last term is obtained from the fact that the distances along the z lines between radial planes are proportional to the radii of curvature. Hence, if ds_z is the differential distance along a typical z line, the ratio ds_z/dz should equal the ratio $(R + x)/R$, or $ds_z = (R + x)\, dz/R$.

In this bent cartesian coordinate system Maxwell's equations take the following form

$$\begin{aligned}
\frac{\partial E_x}{\partial z} &= -j\omega\mu\left(1 + \frac{x}{R}\right) H_y + \frac{\partial}{\partial x}\left[\left(1 + \frac{x}{R}\right) E_z\right] \\
\frac{\partial H_y}{\partial z} &= -j\omega\epsilon\left(1 + \frac{x}{R}\right) E_x + \frac{\partial}{\partial y}\left[\left(1 + \frac{x}{R}\right) H_z\right] \\
\frac{\partial E_y}{\partial z} &= j\omega\mu\left(1 + \frac{x}{R}\right) H_x + \frac{\partial}{\partial y}\left[\left(1 + \frac{x}{R}\right) H_z\right] \\
\frac{\partial H_x}{\partial z} &= j\omega\epsilon\left(1 + \frac{x}{R}\right) E_y + \frac{\partial}{\partial x}\left[\left(1 + \frac{x}{R}\right) H_z\right] \\
E_z &= \frac{1}{j\omega\epsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right), \qquad H_z = -\frac{1}{j\omega\mu}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)
\end{aligned} \tag{42}$$

There is no loss of generality in the apparent assumption that $g = 0$ since the general results may be obtained if we replace ϵ by $\epsilon + (g/j\omega)$. When the field is independent of the y coordinate, the equations become

$$\begin{aligned}
\frac{\partial E_x}{\partial z} &= -j\omega\mu\left(1 + \frac{x}{R}\right) H_y + \frac{\partial}{\partial x}\left[\left(1 + \frac{x}{R}\right) E_z\right] \\
\frac{\partial H_y}{\partial z} &= -j\omega\epsilon\left(1 + \frac{x}{R}\right) E_x, \qquad E_z = \frac{1}{j\omega\epsilon}\frac{\partial H_y}{\partial x}
\end{aligned} \tag{43}$$

We shall now express the field in terms of modes appropriate to perfectly conducting plane strips. To do this we can use (15) provided we replace x by $x + (a/2)$, the transformation being needed because x is

now measured from a different reference surface. Thus we obtain

$$E_x = \frac{V_0(z)}{a} + \sum N_n^{-1} V_n(z) \cos \frac{n\pi}{a}\left(x + \frac{a}{2}\right), \quad n = 1, 2, 3, \cdots$$
$$H_y = \frac{I_0(z)}{b} + \sum N_n^{-1} I_n(z) \cos \frac{n\pi}{a}\left(x + \frac{a}{2}\right) \tag{44}$$

where the normalization factors are still given by (16). The boundary conditions are

$$E_z\left(-\frac{a}{2}, z\right) = \frac{Z_1}{b} I_0(z) + \sum Z_1 N_n^{-1} I_n(z)$$
$$E_z\left(\frac{a}{2}, z\right) = -\frac{Z_2}{b} I_0(z) - \sum Z_2 (-)^n N_n^{-1} I_n(z) \tag{45}$$

Telegraphist's equations for the principal mode are obtained once more merely by integrating the first two equations of the set (43) with respect to x and using (44) and the boundary conditions (45). Thus we find

$$\frac{dV_0(z)}{dz} = -\left[\frac{j\omega\mu a}{b} + \frac{Z_1 + Z_2}{b} + \frac{(Z_2 - Z_1)a}{2Rb}\right] I_0(z) - \sum Z_{0n} I_n(z)$$
$$\frac{dI_0(z)}{dz} = -\frac{j\omega\epsilon b}{a} V_0(z) - \sum Y_{0n} V_n(z) \tag{46}$$

where

$$Z_{0n} = \frac{(-)^n Z_2 + Z_1}{N_n} + \frac{[(-)^n Z_2 - Z_1]a}{2RN_n} + \frac{j\omega\mu}{RN_n} \int_{-a/2}^{a/2} x \cos \frac{n\pi}{a}\left(x + \frac{a}{2}\right) dx \tag{47}$$
$$Y_{0n} = \frac{j\omega\epsilon b}{RaN_n} \int_{-a/2}^{a/2} x \cos \frac{n\pi}{a}\left(x + \frac{a}{2}\right) dx$$

These equations are valid even if the strips are bent non-uniformly so that R is a function of z.

9. BENT STRIP TRANSMISSION LINES — HIGHER ORDER MODES

To obtain telegraphist's equations for the higher order modes we multiply the first two equations of the set (43) by

$$N_m^{-1} \cos \frac{m\pi}{a}\left(x + \frac{a}{2}\right)$$

and integrate over the cross-section of the strip line. As in previous examples the last term in the first equation should be integrated by parts. Then we should substitute for E_z under the integral sign the third equation of the set (43) and once more integrate by parts. Finally, we should substitute for H_y its series representation. Thus we shall find

$$\begin{aligned} \frac{dV_m}{dz} &= -\left[j\omega\mu + \frac{m^2\pi^2}{j\omega\epsilon a^2} + \frac{2(Z_1 + Z_2)}{a} + \frac{Z_2 - Z_1}{R}\right] I_m - \sum{}' Z_{mn} I_n \\ \frac{dI_m}{dz} &= -j\omega\epsilon V_m - \sum{}' Y_{mn} V_n \end{aligned} \tag{48}$$

where the summations are extended over the sequence $n = 0,1,2, \cdots$ excepting $n = m$ and

$$\begin{aligned} Z_{mn} = {} & \frac{2[Z_1 + (-)^{m+n} Z_2]}{a} + \frac{(-)^{m+n} Z_2 - Z_1}{R} \\ & + \frac{1}{Ra}\left(j\omega\mu + \frac{mn\pi^2}{j\omega\epsilon a^2}\right) \int_{-a/2}^{a/2} x \cos\left[(m - n)\pi\left(\frac{x}{2} + \frac{1}{2}\right)\right] dx \\ & + \frac{1}{Ra}\left(j\omega\mu - \frac{mn\pi^2}{j\omega\epsilon a^2}\right) \int_{-a/2}^{a/2} x \cos\left[(m + n)\pi\left(\frac{x}{a} + \frac{1}{2}\right)\right] dx \\ Y_{mn} = {} & \frac{2j\omega\epsilon}{Ra} \int_{-a/2}^{a/2} x \cos\frac{m\pi}{a}\left(x + \frac{a}{2}\right) \cos\frac{n\pi}{a}\left(x + \frac{a}{2}\right) dx \end{aligned} \tag{49}$$

for $n \neq 0, m$. For $n = 0$ the mutual parameters are given by (47).

In the open interval $-a/2 < x < a/2$ the series for H_y may be differentiated term by term. Hence, the longitudinal electric intensity may be obtained from the last equation of the set (43). Thus, *between* the boundaries we have

$$E_z(x, z) = -\sum (n\pi/a) N_n^{-1} I_n(z) \sin\frac{n\pi}{a}\left(x + \frac{a}{2}\right), \qquad -a/2 < x < a/2 \tag{50}$$

On the boundaries we have (45).

The above equations are still valid when R is a function of z; but a and b must be constants.

10. EXPANDING STRIP TRANSMISSION LINES IN CURVILINEAR COORDINATES — A CASE IN WHICH MAXWELL'S EQUATIONS ARE NOT SEPARABLE

The separate sets of terms in the series representing various field components in all preceding problems satisfied Maxwell's equations. The

entire series were required to satisfy the boundary conditions. In the present section we shall express the field in terms of sets of functions which individually *do not* satisfy Maxwell's equations and which *may or may not* satisfy the boundary conditions. The example we are about to consider will illustrate a method for solving Maxwell's equations, when the variables are not separable, by reducing them to generalized telegraphist's equations.

Let us assume that the boundaries of the expanding portion of the strip transmission line in Fig. 4 are circular cylinders tangential to the plane boundaries to the left of the $x0y$ plane. This is, of course, a special case of the problem treated in Sections 6 and 7. In the present section, however, we shall use curvilinear coordinates. In such coordinates Maxwell's equations are

$$\begin{aligned}
\frac{\partial(e_1E_u)}{\partial w} &= -j\omega\mu(e_3e_1/e_2)(e_2H_v) + \frac{\partial(e_3E_w)}{\partial u} \\
\frac{\partial(e_2H_v)}{\partial w} &= -j\omega\epsilon(e_2e_3/e_1)(e_1E_u) + \frac{\partial}{\partial v}(e_3H_w) \\
\frac{\partial(e_2E_v)}{\partial w} &= j\omega\mu(e_2e_3/e_1)(e_1H_u) + \frac{\partial}{\partial v}(e_3E_w) \\
\frac{\partial(e_1H_u)}{\partial w} &= j\omega\epsilon(e_3e_1/e_2)(e_2E_v) + \frac{\partial}{\partial u}(e_3H_w) \\
E_w &= \frac{1}{j\omega\epsilon e_1e_2}\left[\frac{\partial(e_2H_v)}{\partial u} - \frac{\partial(e_1H_u)}{\partial v}\right] \\
H_w &= \frac{1}{j\omega\mu e_1e_2}\left[\frac{\partial(e_1E_u)}{\partial v} - \frac{\partial(e_2E_v)}{\partial u}\right]
\end{aligned} \tag{51}$$

These equations have been arranged in a form convenient for problems in which wave propagation takes place along the w lines. In some cases it is convenient to treat the products e_1E_u , e_2E_v , e_1H_u , e_2H_v rather than the field components themselves as dependent variables and the parentheses around these products in the preceding equations are intended to call attention to this fact.

The choice of a particular coordinate system for solving a physical problem depends on various factors. The cartesian system chosen in Sections 6 and 7 is good for several reasons: Maxwell's equations have a particularly simple form, boundary conditions are easy to express for almost arbitrary boundaries, the basic transverse field patterns con-

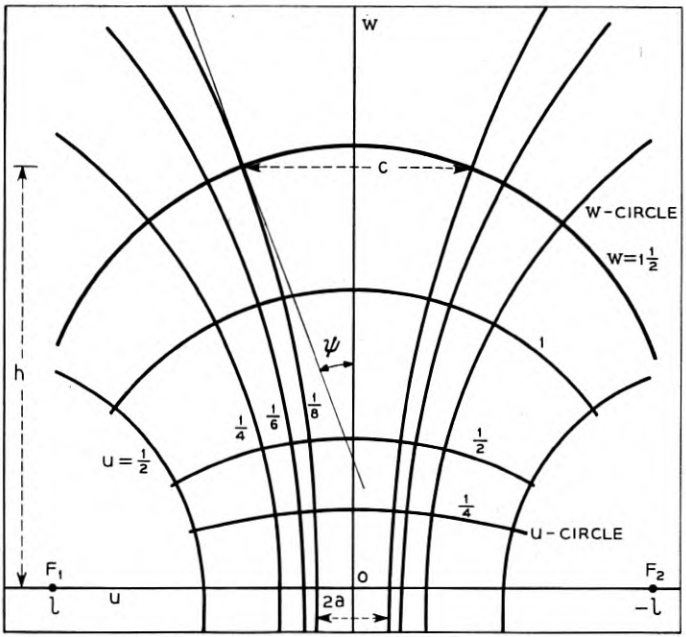


Fig. 6 — Biaxial coordinates.

form to those in waveguides of uniform cross-section. For this last reason, the cartesian system is particularly convenient for the analysis of junction sections between two waveguides of rectangular cross-section. Biaxial coordinates[13] are more convenient in some respects for the analysis of junctions between two-dimensional waveguides and two-dimensional horns, Fig. 6, although we are not prepared to say that they are more convenient than cartesian coordinates when all factors are taken into consideration. Here we shall use biaxial coordinates solely to illustrate the conversion of Maxwell's equations in curvilinear coordinates into generalized telegraphist's equations.

Biaxial coordinate system consists of two orthogonal systems of circular cylinders perpendicular to a system of parallel planes. A section by one of these planes is shown in Fig. 6. Circles of one system are non-intersecting and their centers lie on the horizontal axis. Circles of the other system intersect at the foci F_1 and F_2; their centers lie on the vertical axis. The non-intersecting circles will be called the w-lines (lines of constant u and varying w), and the intersecting circles the u-lines. The coordinate u is the shortest distance between the w-circle and the vertical axis; w is the intercept of the u-circular arc on the vertical axis (each

u-circle will be split for our purposes into two arcs, one above and the other below the focal line F_1F_2).

The radius of a typical w-circle is

$$\frac{1}{2}\ell\left(\frac{\ell}{u}-\frac{u}{\ell}\right)$$

and the distance between the center and the vertical axis

$$\frac{1}{2}\ell\left(\frac{\ell}{u}+\frac{u}{\ell}\right)$$

The radius of a u-circle is

$$\frac{1}{2}\ell\left(\frac{w}{\ell}+\frac{\ell}{w}\right)$$

and the distance of its center from the horizontal axis

$$\frac{1}{2}\left(\frac{w}{\ell}-\frac{\ell}{w}\right)$$

Depending on whether this distance is positive or negative, the center is above or below the horizontal axis. In this coordinate system

$$e_1=\frac{1+(w/\ell)^2}{1+(uw/\ell^2)^2},\qquad e_2=1,\qquad e_3=\frac{1-(u/\ell)^2}{1+(uw/\ell^2)} \tag{52}$$

A section of a waveguide to horn junction is characterized by the following parameters: the length h, the width of the narrow aperture $2a$, the width of the wide aperture $2c$, and the horn angle 2ψ at the wide aperture. If h/a and ψ are given, then

$$c/a=1+(h/a)\tan\tfrac{1}{2}\psi,\qquad \ell/a=[1+(2h/a\sin\psi)]^{1/2} \tag{53}$$

From the last equation we determine the semi-focal distance ℓ for the coordinate system. The coordinate w_0 of the "wave-front" at the wide aperture may be obtained from

$$aw_0/\ell^2=\tan\tfrac{1}{2}\psi \tag{54}$$

As in Sections 6 and 7 we shall consider those modes for which the field is independent of v and for which (51) becomes

$$\begin{aligned}\frac{\partial(e_1E_u)}{\partial w}&=-j\omega\mu e_1e_3H_v+\frac{\partial}{\partial u}(e_3E_w)\\ \frac{\partial H_v}{\partial w}&=-j\omega\epsilon(e_3/e_1)(e_1E_u),\qquad E_w=\frac{1}{j\omega\epsilon e_1}\frac{\partial H_v}{\partial u}\end{aligned} \tag{55}$$

We have already examined several cases with imperfectly conducting boundaries and in the present example we shall assume that the boundaries are perfectly conducting. We shall confine ourselves to symmetric modes for which E_u is perpendicular to the vw plane. We shall express our field in the form analogous to (15); thus

$$e_1 E_w = \frac{V_0(w)}{a} + \sum N_n^{-1} V_n(w) \cos \frac{n\pi u}{a}$$

$$H_v = \frac{I_0(w)}{b} + \sum N_n^{-1} I_n(w) \cos \frac{n\pi u}{a} \tag{56}$$

$$N_n^2 = ab/2, \qquad n = 1, 2, 3, \cdots$$

where b is the width of the strips. Substituting in the last equation of the set (55) we find

$$j\omega\epsilon e_1 E_w = -\sum \frac{n\pi}{aN_n} I_n(w) \sin \frac{n\pi u}{a} \tag{57}$$

The boundary conditions are thus satisfied automatically. If we integrate the first series in the set (56) along a typical u-line, we obtain

$$\int_0^a e_1 E_u \, du = \int_0^a E_u \, ds_u = V_0(w) \tag{58}$$

Hence, $V_0(w)$ is the transverse voltage from the middle plane of the strip line to the upper strip, measured along a u-line. It should be noted that in the narrow aperture $w = 0$, $u = x$, $e_1 = 1$ and series (56) are identical with (15).

To obtain telegraphist's equations for the principal mode we integrate the first two equations of the set (55) along a u-line and substitute from (56); thus we have

$$\frac{dV_0}{dw} = -Z_{00} I_0 - Z_{01} I_1 - Z_{02} I_2 - \cdots$$

$$\frac{dI_0}{dw} = -Y_{00} V_0 - Y_{01} V_1 - Y_{02} V_2 - \cdots \tag{59}$$

where

$$Z_{00} = \frac{j\omega\mu}{b} \int_0^a e_1 e_3 \, du, \qquad Y_{00} = \frac{j\omega\epsilon b}{a} \int_0^a \frac{e_3}{e_1} \, du$$

$$Z_{0n} = \frac{j\omega\mu}{N_n} \int_0^a e_1 e_3 \cos \frac{n\pi u}{a} \, du, \qquad Y_{0n} = \frac{j\omega\epsilon b}{N_n} \int_0^a \frac{e_3}{e_1} \cos \frac{n\pi u}{a} \, du, \tag{60}$$

$$n > 0$$

The integrals for Z_{00}, Y_{00}, and Y_{0n} may be expressed in terms of elementary functions, and the integrals for Z_{0n} by power series since

$$\begin{aligned} e_1e_3 &= [1 + (w/\ell)^2][1 - (u/\ell)^2][1 + (uw/\ell^2)^2]^{-2} \\ &= [1 + (w/\ell)^2][1 - (u/\ell)^2][1 - 2(uw/\ell^2)^2 + 3(uw/\ell^2)^4 - \cdots] \end{aligned} \tag{61}$$

In practical cases u/ℓ and uw/ℓ^2 are relatively small and a few terms of the series will suffice.

To obtain the corresponding equations for the higher order modes we should multiply the first two equations of the set (55) by $N_m^{-1}\cos(m\pi u/a)$ and integrate over the cross-section of the strip line by the uv surface. Thus, we find

$$\begin{aligned} Z_{mn} &= \frac{j\omega\mu b}{N_mN_n}\int_0^a e_1e_3\cos\frac{m\pi u}{a}\cos\frac{n\pi u}{a}\,du \\ &\quad + \frac{mn\pi^2 b}{j\omega\epsilon a^2N_mN_n}\int_0^a \frac{e_3}{e_1}\sin\frac{m\pi u}{a}\sin\frac{n\pi u}{a}\,du \\ Y_{mn} &= \frac{j\omega\epsilon b}{N_mN_n}\int_0^a \frac{e_3}{e_1}\cos\frac{m\pi u}{a}\cos\frac{n\pi u}{a}\,du \end{aligned} \tag{62}$$

for all m, n not equal to zero.

11. TRANSVERSE ELECTRIC WAVES BETWEEN PARALLEL PLANES

Let us now see what happens in the case of TE modes. Again we shall consider the simplest case, the case of parallel planes, Fig. 3, and assume that the field is independent of the y coordinates. The only non-vanishing field components are E_y, H_x and H_z, and Maxwell's equations become

$$\begin{aligned} &\frac{\partial E_y}{\partial z} = j\omega\mu H_x, \qquad \frac{\partial H_x}{\partial z} = (g + j\omega\epsilon)E_y + \frac{\partial H_z}{\partial x} \\ &H_z = -\frac{1}{j\omega\mu}\frac{\partial E_y}{\partial x} \end{aligned} \tag{63}$$

For perfectly conducting planes the general solution is of the following form

$$E_y = \sum N_n^{-1}V_n(z)\sin(n\pi x/a), \qquad H_x = -\sum N_n^{-1}I_n(z)\sin(n\pi x/a) \tag{64}$$

where the normalizing factors are given by (16). We now assume that *between* $(0 < x < a)$, the imperfectly conducting planes the general solution has still the same form. Putting it differently we expand the new solution in a sine series. Since the sine terms vanish on the boundaries,

the series will represent the new solution only between the boundaries. Since these series represent discontinuous functions, their coefficients will ultimately vary as $1/n$; therefore the derivative series will diverge. Hence, we cannot obtain H_z by substituting from (64) into the third equation of the set (63). For this reason, we assume an independent series for H_z,

$$H_z = \sum N_n^{-1} i_n(z) \cos(n\pi x/a) \tag{65}$$

On the boundaries the ratios of the tangential electric and magnetic intensities equal surface impedances of the boundaries, with appropriate signs,

$$\begin{aligned} E_y(0, z) &= -Z_1 H_z(0, z) = -\sum Z_1 N_n^{-1} i_n(z) \\ E_y(a, z) &= Z_2 H_z(a, z) = \sum (-)^n Z_2 N_n^{-1} i_n(z) \end{aligned} \tag{66}$$

The cosine series (65) represents a continuous function and its coefficients will decrease fast enough to make the derivative series convergent. So we substitute from (64) and (65) in (63), combine the terms containing similar sine terms, and equate the coefficients of the resulting sine series to zero. Thus we obtain

$$\frac{dV_n(z)}{dz} = -j\omega\mu I_n(z), \qquad \frac{dI_n(z)}{dz} = -(g + j\omega\epsilon)V_n(z) + \frac{n\pi}{a} i_n(z) \tag{67}$$

We now multiply the third equation in the set (63) by $N_m^{-1} \cos(m\pi x/a)$ and integrate,

$$\begin{aligned} \int_0^b \int_0^a H_z N_m^{-1} \cos\frac{m\pi x}{a}\, dx\, dy \\ = -\frac{1}{j\omega\mu} \int_0^b \int_0^a \frac{\partial E_y}{\partial x} N_m^{-1} \cos\frac{m\pi x}{a}\, dx\, dy \end{aligned} \tag{68}$$

On the left we substitute from (65), and on the right we integrate by parts,

$$\begin{aligned} i_m(z) = -\frac{b}{j\omega\mu N_m} \cos\frac{m\pi x}{a} E_y(x, z) \Big|_0^a \\ - \frac{1}{j\omega\mu} \int_0^b \int_0^a E_y \frac{m\pi}{aN_m} \sin\frac{m\pi x}{a}\, dx\, dy \end{aligned} \tag{69}$$

In the first term on the right we substitute from (66). In the second term we substitute the series for E_y. Since this series is not uniformly convergent in the closed interval $0 \leqq x \leqq a$, we cannot be sure that we shall

get the right answer by integrating the series term by term. What we can do is to integrate in a slightly smaller interval in which the series converges uniformly, and then pass to the limit. In the present case the answer turns out to be the same as that obtained when we integrate term by term in the closed interval. Thus we find

$$i_m(z) = -b \sum_n \frac{Z_1 + (-)^{m+n} Z_2}{j\omega\mu N_m N_n} i_n(z) - \frac{m\pi}{j\omega\mu a} V_m(z) \tag{70}$$

If we solve this set of equations for $i_n(z)$ and substitute in (67), we shall obtain telegraphist's equations.

Rearranging the terms in (70), we have

$$\left[1 + \frac{2(Z_1 + Z_2)}{j\omega\mu a}\right] i_m(z) + \sum{}' \frac{2[Z_1 + (-)^{m+n} Z_2]}{j\omega\mu a} in(z) = -\frac{m\pi}{j\omega\mu a} V_m(z) \tag{71}$$

Neglecting the summation, we obtain an approximate solution

$$i_n(z) = -\frac{m\pi}{j\omega\mu a}\left[1 + \frac{2(Z_1 + Z_2)}{j\omega\mu a}\right]^{-1} V_m(z) \tag{72}$$

and approximate telegraphist's equations,

$$\frac{dV_n}{dz} = -j\omega\mu I_n,$$
$$\frac{dI_n}{dz} = -\left[g + j\omega\epsilon + \frac{n^2\pi^2}{j\omega\mu a^2 + 2(Z_1 + Z_2)a}\right] V_n \tag{73}$$

Instead of solving (70) for $i_n(z)$ we can obtain $V_m(z)$ from (70) and substitute it in (67), after replacing n in (73) by m.

12. WAVES ON INFINITE CONDUCTORS

In this section we shall consider waves on two semi-infinite conductors tapering to a point, Fig. 7(a), and waves outside a certain sphere (S), Fig. 7(b), which encloses the terminals of conductors which are not tapered to a point. In the latter case the sphere (S) will enclose a source of power; in the former case we assume an idealized point source of power at the origin O. For simplicity we shall assume that the structure possesses circular symmetry about OA and plane symmetry about the plane perpendicular to OA at O. In this case there will be waves in which the magnetic lines are circles coaxial with the axis of the structure. In

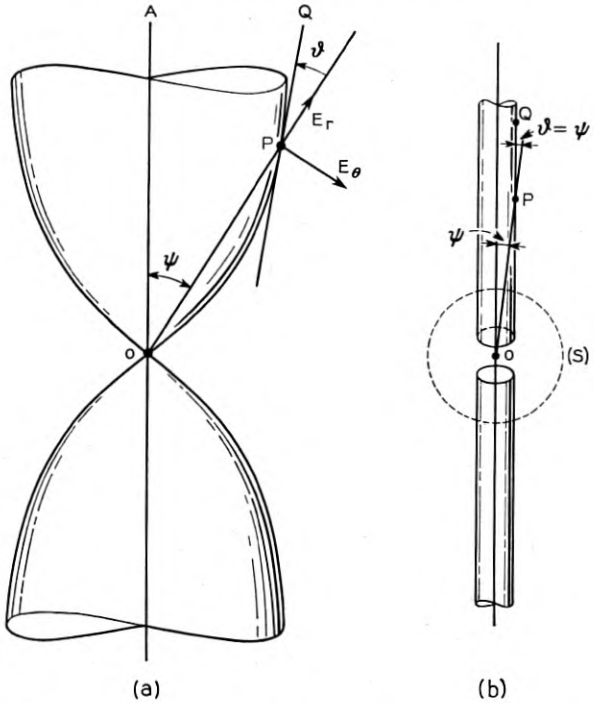


Fig. 7 — Infinite conductors excited by a point source (a) and by a source of finite size (b).

spherical coordinates the appropriate field equations are

$$\frac{\partial}{\partial r}(rE_\theta) = -j\omega\mu(rH_\varphi) + \frac{\partial E_r}{\partial \theta}, \qquad \frac{\partial}{\partial r}(rH_\varphi) = -(g + j\omega\epsilon)(rE_\theta)$$
$$E_r = \frac{1}{(g + j\omega\epsilon)r^2}\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, rH_\varphi) \tag{74}$$

Let P be a typical point on the upper half of the structure and ψ be the angle between the radius $0P$ and the axis $0A$. Let ϑ be the angle from the radius to the tangent plane PQ. Then the boundary conditions are

$$E_r(r,\psi)\cos\vartheta - E_\theta(r,\psi)\sin\vartheta = ZH_\varphi(r,\psi)$$
$$E_r(r,\pi-\psi)\cos\vartheta + E_\theta(r,\pi-\psi)\sin\vartheta = -ZH_\varphi(r,\pi-\psi) \tag{75}$$

where Z is the surface impedance. Thus, *at the surface* of each conductor the radial electric intensity may be expressed in terms of the meridian

electric intensity and the magnetic intensity

$$\begin{aligned} E_r(r, \psi) &= E_\theta(r, \psi) \tan \vartheta + Z \sec \vartheta H_\varphi(r, \psi) \\ E_r(r, \pi - \psi) &= -E_r(r, \psi) \end{aligned} \tag{76}$$

Let us further confine ourselves to the symmetric modes in which the currents passing through the cross-sections of the upper and the lower conductors equidistant from 0 are equal and similarly directed. Then we shall represent the field by the following series

$$\begin{aligned} rE_\theta &= \frac{V_0(r)}{N_0(r) \sin \theta} + \sum V_n(r) N_n^{-1}(r) \frac{\partial M_n(\cos \theta)}{\partial \theta}, \\ & \qquad \psi \leqq \theta \leqq \pi - \psi \\ rH_\varphi &= \frac{I_0(r)}{2\pi \sin \theta} + \sum I_n(r) N_n^{-1}(r) \frac{\partial M_n(\cos \theta)}{\partial \theta}, \\ & \qquad \psi \leqq \theta \leqq \pi - \psi \\ E_r &= -\sum \frac{n(n+1)}{(g + j\omega\epsilon) r^2 N_m(r)} I_n(r) M_n(\cos \theta), \qquad \psi < \theta < \pi - \psi \end{aligned} \tag{77}$$

where

$$M_n(\cos \theta) = \tfrac{1}{2}[P_n(\cos \theta) - P_n(-\cos \theta)] \tag{78}$$

and the P-functions are Legendre functions. The summations are extended over the roots n of the following equation:

$$M_n(\cos \psi) = 0 \tag{79}$$

The normalizing factors are

$$\begin{aligned} N_0 &= \int_\psi^{\pi-\psi} \frac{d\theta\, d\varphi}{\sin \theta} = 2 \log \cot \frac{\psi}{2} \\ N_n^2 &= 2\pi \int_\psi^{\pi-\psi} \left[\frac{\partial}{\partial \theta} M_n(\cos \theta)\right]^2 \sin \theta\, d\theta \\ &= 2\pi n(n+1) \int_\psi^{\pi-\psi} M_n(\cos \theta)]^2 \sin \theta\, d\theta \end{aligned} \tag{80}$$

Each individual term in (77) will satisfy Maxwell's equations and the boundary conditions if the conductors happen to be perfectly conducting cones. Otherwise we need the entire series. The function $V_0(r)$ is the transverse voltage between the conductors along a typical meridian; $I_0(r)$ is the current in the upper conductor associated with the principal wave. The remaining functions, $V_n(r)$ and $I_n(r)$, are proportional to the

electric and magnetic intensities of various modes. In view of (77) the boundary conditions become

$$E_r(r, \psi) = \frac{\tan \vartheta}{rN_0 \sin \psi} V_0(r) + \sum \frac{\tan \vartheta}{rN_n} \frac{\partial M_n(\cos \psi)}{\partial \psi} V_n(r) + \frac{Z \sec \vartheta}{2\pi r \sin \psi} I_0(r) + \sum \frac{Z \sec \vartheta}{rN_n} \frac{\partial M_n(\cos \psi)}{\partial \psi} I_n(r) \tag{81}$$

Equations for the principal mode are obtained as in previous examples by integrating the first two field equations (74). Thus, we find

$$\begin{aligned} \frac{dV_0(r)}{dr} &= -\frac{j\omega\mu}{\pi} \log \cot \frac{\psi}{2} I_0(r) + E_r(r, \pi - \psi) - E_r(r, \psi) \\ \frac{dI_0(r)}{dr} &= -\frac{\pi(g + j\omega\epsilon)}{\log \cot (\psi/2)} V_0(r) \end{aligned} \tag{82}$$

Substituting from (81), we obtain the final result.

To obtain the equations for the higher mode we shall multiply the field equations by the normalized characteristic functions and integrate. It is important to remember that ψ, ϑ and therefore n and N_n are functions of r. On one occasion we shall have to integrate by parts as follows

$$\begin{aligned} &\int\int N_m^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \sin \theta \frac{\partial E_r}{\partial \theta} d\theta \, d\varphi \\ &\quad = 2\pi N_m^{-1} E_r(r, \theta) \frac{\partial M_m(\cos \theta)}{\partial \theta} \sin \theta \Big|_{\psi}^{\pi-\psi} + \int\int N_m^{-1} E_r(r, \theta) m \\ &\qquad (m + 1) M_m(\cos \theta) \sin \theta \, d\theta \, d\varphi \\ &= -\frac{m(m + 1)}{(g + j\omega\epsilon) r^2} I_m(r) + 2\pi N_m^{-1} E_r(r, \theta) \frac{\partial M_m(\cos \theta)}{\partial \theta} \sin \theta \Big|_{\psi}^{\pi-\psi} \end{aligned} \tag{83}$$

On another occasion we have to take into consideration the above-mentioned dependence of n and N_n on r,

$$\begin{aligned} &\int\int N_m^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \sin \theta \frac{\partial}{\partial r} \left[V_n(r) N_n^{-1}(r) \frac{\partial M_n(\cos \theta)}{\partial \theta} \right] d\theta \, d\varphi \\ &= V_n(r) \int\int N_m^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \frac{\partial}{\partial r} \left[N_n^{-1}(r) \frac{\partial M_n(\cos \theta)}{\partial \theta} \right] \\ &\qquad \sin \theta \, d\theta \, d\varphi, \qquad \text{if } n \neq m, \\ &= \frac{\partial V_m(r)}{\partial r} + V_m(r) \int\int N_m^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \frac{\partial}{\partial r} \left[N_m^{-1}(r) \frac{\partial M_m(\cos \theta)}{\partial \theta} \right] \\ &\qquad \sin \theta \, d\theta \, d\varphi \end{aligned} \tag{84}$$

if $n = m$.

In the end we shall obtain (11) with the following values of the transmission line parameters,

$$Z_{00} = \frac{j\omega\mu}{\pi} \log \cot (\psi/2) + \frac{Z \sec \vartheta}{\pi 2/\sin \psi}$$

$$Y_{00} = \frac{\pi(g + j\omega)}{\log \cot (\psi/2)}, \qquad I_{T_{00}} = 0$$

$${}^{V}T_{00} = \frac{\tan \vartheta}{r \sin \psi \log \cot (\psi/2)}$$

$${}^{V}T_{0n} = \frac{2 \tan \vartheta}{rN_n} \frac{\partial M_n(\cos \psi)}{\partial \psi}, \qquad n \neq 0$$

$$Z_{0n} = \frac{2Z \sec \vartheta}{rN_n} \frac{\partial M_n(\cos \psi)}{\partial \psi}$$

$$Y_{0n} = I_{T_{0n}} = 0, \qquad n \neq 0$$

$$Z_{mm} = j\omega\mu + \frac{m(m+1)}{(g + j\omega\epsilon)r^2} + \frac{4\pi Z \sec \vartheta}{rN_m{}^2} \left[\frac{\partial M_m(\cos \psi)}{\partial \psi}\right]^2 \sin \psi, \qquad m \neq 0$$

$$Z_{mn} = \frac{4\pi Z \sec \vartheta}{rN_m N_n} \frac{\partial M_m(\cos \psi)}{\partial \psi} \frac{\partial M_n(\cos \psi)}{\partial \psi} \sin \psi, \qquad n \neq m, \quad m \neq 0 \tag{85}$$

$${}^{V}T_{mn} = \frac{4\pi \tan \vartheta}{rN_m N_n} \frac{\partial M_m(\cos \psi)}{\partial \psi} \frac{\partial M_n(\cos \psi)}{\partial \psi} \sin \psi + \int\int N_m{}^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \frac{\partial}{\partial r}\left[N_n{}^{-1}(r) \frac{\partial M_n(\cos \psi)}{\partial r}\right] \sin \theta \, d\theta \, d\varphi \qquad m \neq 0, \quad n \neq 0;$$

$$Z_{m0} = Z_{0m}, \qquad Z_{mn} = Z_{nm},$$

$$V_{T_{m0}} = \frac{2\pi \tan \vartheta}{rN_m \log \cot (\psi/2)} \frac{\partial M_n(\cos \psi)}{\partial \psi}, \qquad m \neq 0$$

$$Y_{mm} = g + j\omega\epsilon, \qquad Y_{mn} = Y_{nm} = 0, \qquad n \neq m$$

$${}^{I}T_{mn} = \int\int N_m{}^{-1} \frac{\partial M_m(\cos \theta)}{\partial \theta} \frac{\partial}{\partial r}\left[N_n{}^{-1}(r) \frac{\partial M_n(\cos \psi)}{\partial r}\right] \sin \theta \, d\theta \, d\varphi. \qquad m \neq 0, \quad n \neq 0$$

$${}^{I}T_{m0} = 0$$

13. WAVES ON SEMI-INFINITE CONDUCTORS

The telegraphist's equations for a single conductor, the upper conductor let us say, may be obtained by a few modifications of the equations in the preceding section. There will be no terms outside the summation signs in (77). The function M_n (cos θ) should be replaced by P_n ($-$cos θ). The integrals with respect to θ should be evaluated from $\theta = \psi$ to $\theta = \pi$ rather than from $\theta = \psi$ to $\pi - \psi$.

14. WAVES OVER A PLANE IMPEDANCE SHEET

Under some conditions a plane earth may be approximated by an impedance sheet. Such a sheet is a cone of angle $\psi = \pi/2$ and the telegraphist's equations for it will be obtained if we replace M_n (cos θ) by P_{2n+1} (cos θ) where $n = 0,1,2, \cdots$. The integrals should be calculated over the upper hemisphere. Of course, $\vartheta = 0$, and hence all the voltage and current transfer coefficients vanish.

The normalization factor becomes

$$N_n = \frac{\sqrt{2\pi}\,\sqrt{n(n+1)}}{\sqrt{2n+1}} \tag{86}$$

If the distributed self-impedance of a typical mode is expressed as

$$Z_{mm} = Z_{mm}^{\,0} + Z_{mm}' \tag{87}$$

where $Z_{mm}^{\,0}$ is the distributed self-impedance for a perfectly conducting sheet and Z_{mm}' is due to the finite surface impedance, then

$$\begin{aligned} Z_{mm}^{\,0} &= j\omega\mu + \frac{m(m+1)}{(g + j\omega\epsilon)r^2} \\ Z_{mm}' &= (-)^{m+1}\,\frac{(2m+1)(1.3.5\cdots m)^2 Z}{m(m+1)[2.4.6\cdots(m-1)]^2 r} \end{aligned} \tag{88}$$

The distributed mutual impedances are given by

$$Z_{mn} = \sqrt{Z_{mm}' Z_{nn}'} \tag{89}$$

The distributed admittances are independent of the surface impedance of the sheet; hence

$$Y_{mm} = g + j\omega\epsilon, \qquad Y_{mn} = 0 \quad \text{if} \quad m \neq n \tag{90}$$

In all these equations m and n are odd integers.

15. DERIVATION OF APPROXIMATE TELEGRAPHIST'S EQUATIONS FOR THE TE_{11} MODE IN A CIRCULAR WAVEGUIDE-TO-HORN JUNCTION

It is probably safer not to make approximations sooner than absolutely necessary provided we are willing to tolerate a mass of detail which later turns out to be unnecessary. Still, if the technique of conversion of Maxwell's equations into telegraphist's equations is thoroughly understood, it may be possible to make *ab initio* approximations without undue risk of omitting something more important than we are willing to neglect. To illustrate such *ab initio* approximations we shall obtain telegraphist's equations for the dominant mode in a gentle waveguide-to-horn circular junction. At the start we shall neglect all the coupling coefficients except those between TE_{11} and TM_{11} modes. Even these will be retained only part of the way in order to explain what we should do if we neglect them from the beginning. In the next section we shall discuss cases in which we *should not neglect* all the coupling coefficients.

First of all we shall exhibit azymuthal variation of the field.

$$\begin{aligned} E_\rho &= \pi^{-1/2}\hat{E} \sin\varphi \\ H_\varphi &= \pi^{-1/2}\hat{H}_\varphi \sin\varphi \\ E_\varphi &= \pi^{-1/2}\hat{E}_\varphi \cos\varphi \\ H_\rho &= \pi^{-1/2}\hat{H}_\rho \cos\varphi \\ E_z &= \pi^{-1/2}\hat{E}_z \sin\varphi \\ H_z &= \pi^{-1/2}\hat{H}_z \cos\varphi \end{aligned} \tag{91}$$

The factor $\pi^{-1/2}$ has been introduced to normalize the sine and cosine. If we retain only the first radial TE and TM modes, we have

$$\begin{aligned} \hat{E}_\rho &\cong N^{-1}V(z)(\chi\rho)^{-1}J_1(\chi\rho) + \bar{N}^{-1}\bar{V}(z)J_1'(\bar{\chi}\rho) \\ \hat{H}_\varphi &\cong N^{-1}I(z)(\chi\rho)^{-1}J_1(\chi\rho) + \bar{N}^{-1}\bar{I}(z)J_1'(\bar{\chi}\rho) \\ \hat{E}_\varphi &\cong N^{-1}V(z)J_1'(\chi\rho) + \bar{N}^{-1}\bar{V}(z)(\bar{\chi}\rho)^{-1}J_1(\bar{\chi}\rho) \\ \hat{H}_\rho &\cong N^{-1}I(z)J_1'(\chi\rho) - \bar{N}^{-1}\bar{I}(z)(\bar{\chi}\rho)^{-1}J_1(\bar{\chi}\rho) \end{aligned} \tag{92}$$

where

$$\chi a = 1.841\cdots \qquad \bar{\chi}a = 3.83\cdots \tag{93}$$

and

$$
\begin{aligned}
N^2 &= \int_0^a ([J_1'(\chi\rho)]^2 + (\chi\rho)^{-2}[J_1(\chi\rho)]^2)\rho\, d\rho \\
&= \int_0^a [J_1(\chi\rho)]^2\rho\, d\rho = \frac{1}{2}a^2\left[1 - \frac{1}{(1.84)^2}\right][J_1(1.84)]^2 = (0.345a)^2 \\
\bar{N}^2 &= \int_0^a ([J_1'(\bar{\chi}\rho)]^2 + (\bar{\chi}\rho)^{-2}[J_1(\bar{\chi}\rho)]^2)\rho\, d\rho \\
&= \int_0^a [J_1(\bar{\chi}\rho)]^2\rho\, d\rho = \tfrac{1}{2}a^2[J_0(3.83)]^2 = (0.285a)^2.
\end{aligned} \tag{94}
$$

Maxwell's transmission equations for transverse field components in cylindrical coordinates are

$$
\begin{aligned}
\frac{\partial E_\rho}{\partial z} &= -j\omega\mu H_\varphi + \frac{\partial E_z}{\partial \rho} \\
\frac{\partial H_\varphi}{\partial z} &= -j\omega\epsilon E_\rho + \frac{\partial H_z}{\rho\partial\varphi} \\
\frac{\partial E_\varphi}{\partial z} &= j\omega\mu H_\rho + \frac{\partial E_z}{\rho\partial\varphi} \\
\frac{\partial H_\rho}{\partial z} &= j\omega\epsilon E_\varphi + \frac{\partial H_z}{\partial \rho}
\end{aligned} \tag{95}
$$

In addition we have the following equations for the longitudinal field components

$$
\frac{\partial}{\partial\rho}(\rho E_\varphi) - \frac{\partial E_\rho}{\partial\varphi} = -j\omega\mu\rho H_z, \qquad \frac{\partial}{\partial\varphi}(\rho H_\varphi) - \frac{\partial H_\rho}{\partial\varphi} = j\omega\epsilon\rho E_z \tag{96}
$$

In view of (91), (95) becomes

$$
\begin{aligned}
\frac{\partial \hat{E}_\rho}{\partial z} &= -j\omega\mu\hat{H}_\varphi + \frac{\partial \hat{E}_z}{\partial \rho} \\
\frac{\partial \hat{H}_\varphi}{\partial z} &= -j\omega\epsilon\hat{E}_\rho - \rho^{-1}\hat{H}_z \\
\frac{\partial \hat{E}_\varphi}{\partial z} &= j\omega\mu\hat{H}_\rho + \rho^{-1}\hat{E}_z \\
\frac{\partial \hat{H}_\rho}{\partial z} &= j\omega\epsilon\hat{E}_\varphi + \frac{\partial \hat{H}_z}{\partial \rho}
\end{aligned} \tag{97}
$$

while from (92) and (96) we obtain

$$\begin{aligned} \hat{H}_z &= V(z)(j\omega\mu N)^{-1}\chi J_1(\chi\rho) \\ \hat{E}_z &= -\bar{I}(z)(j\omega\epsilon\bar{N})^{-1}\bar{\chi}J_1(\bar{\chi}\rho) \end{aligned} \tag{98}$$

The expression for $\hat{E}_z$, even when completed by inclusion of the higher order radial modes, is valid only in the interior ($\rho < a$) of the junction. On the boundary we have

$$\hat{E}_z(a, z) = -\hat{E}_\rho(a,z) \tan \vartheta(z) \tag{99}$$

To obtain the telegraphist's equations we multiply the first column of (97) by

$$N^{-1}(\chi\rho)^{-1}J_1(\chi\rho)\rho \, d\rho$$

and

$$N^{-1}J_1(\chi\rho)\rho \, d\rho$$

respectively, add and integrate from $\rho = 0$ to $\rho = a$. The second column is similarly treated. The following are auxiliary calculations. In view of (92)

$$\int_0^a [\hat{H}_\varphi N^{-1}(\chi\rho)^{-1}J_1(\chi\rho) + \hat{H}_\rho N^{-1}J_1'(\chi\rho)]\rho \, d\rho = -I(z) \tag{100}$$

The terms involving $\bar{I}(z)$ have disappeared after integration. To obtain

$$\int_0^a \frac{\partial \hat{E}_z}{\partial \rho} N^{-1}(\chi\rho)^{-1}J_1(\chi\rho)\rho \, d\rho + \int_0^a \hat{E}_z N^{-1}J_1'(\chi\rho) \, d\rho \tag{101}$$

we integrate the first term by parts

$$\hat{E}_z N^{-1}\chi^{-1}J_1(\chi\rho)\Big|_0^a - \int_0^a \hat{E}_z N^{-1}J_1'(\chi\rho) \, d\rho \tag{102}$$

The last term of this expression cancels the last term in (101); thus the total is

$$\begin{aligned} \hat{E}_z(a,z)\chi^{-1}N^{-1}J_1(\chi a) &= -\hat{E}_\rho(a,z) \tan \vartheta(z)\chi^{-1}N^{-1}J_1(\chi a) \\ &= (\chi N)^{-2}a^{-1}[J_1(\chi a)]^2 \tan \vartheta V(z) \\ &\quad - (\chi N\bar{N})^{-1}J_1(\chi a)J_1'(\bar{\chi}a) \tan \vartheta \bar{V}(z) \end{aligned} \tag{103}$$

At this point let us note that if we had decided to neglect the TM_{11} mode at the beginning, we would have set $\hat{E}_z = 0$ in equations (98). But in obtaining the telegraphist's equations from (97) it would still have been necessary to retain $\hat{E}_z$ until after the integration has been per-

formed and the boundary condition utilized. To obtain

$$\int_0^a \left[-\rho^{-1}\hat{H}_z N^{-1}(\chi\rho)^{-1}J_1(\chi\rho) - \frac{\partial \hat{H}_z}{\partial \rho} N^{-1}J_1'(\chi\rho) \right] \rho \, d\rho \tag{104}$$

we also integrate by parts. This time expression (104) is found to equal

$$-\int_0^a \hat{H}_z N^{-1}\chi\rho J_1(\chi\rho)\, d\rho = -\frac{\chi^2}{j\omega\mu} V(z) \tag{105}$$

In the calculation of

$$\int_0^a \left[\frac{\partial \hat{E}_\rho}{\partial z} N^{-1}(\chi\rho)^{-1}J_1(\chi\rho) + \frac{\partial \hat{E}_\varphi}{\partial z} N^{-1}J_1(\chi\rho) \right] \rho \, d\rho \tag{106}$$

and a similar integral involving $\hat{H}_\varphi$ and $\hat{H}_\rho$ we must remember that a, χ, and N are functions of z. Thus, this integral will be equal to

$$\begin{aligned} &\frac{dV}{dz} + V(z) \int_0^a \Big((\chi N\rho)^{-1}J_1(\chi\rho) \frac{\partial}{\partial z} [(\chi N\rho)^{-1}J_1(\chi\rho)] \\ &\qquad + N^{-1}J_1'(\chi\rho) \frac{\partial}{\partial z} [N^{-1}J_1'(\chi\rho)] \Big) \rho \, d\rho \\ &+ \bar{V}(z) \int_0^a \Big((\chi N\rho)^{-1}J_1(\chi\rho) \frac{\partial}{\partial z} [\bar{N}^{-1}J_1'(\bar{\chi}\rho)] \\ &\qquad + N^{-1}J_1'(\chi\rho) \frac{\partial}{\partial z} [(\bar{\chi}\bar{N}\rho)^{-1}J_1(\bar{\chi}\rho)] \Big) \rho \, d\rho \end{aligned} \tag{107}$$

At this point we should point out another reason why we temporarily retained $\bar{V}(z)$. Each equation in a complete set of telegraphist's equations contains only one derivative of either a voltage function or a current function. To derive such a set of equations we must perform a weighted integration of Maxwell's equations with appropriate weighting factors as in (106). When $\bar{V}(z)$ is retained and wrong weighting factors are used, the derivative of $\bar{V}(z)$ with respect to z will not be eliminated and, hence, we shall be warned of our error. But when we neglect $\bar{V}(z)$ *ab initio*, we lose this self-checking feature. However, after we acquire some experience with this technique, we should not need the self-checking inherent in the retention of other modes.

The final equations for the dominant mode in a wave-guide-to-horn junction are

$$\frac{dV}{dz} = -ZI - {}^{V}TV, \qquad \frac{dI}{dz} = -YV - {}^{I}TI \tag{108}$$

where

$$Z = j\omega\mu, \qquad Y = j\omega\epsilon + \frac{(1.841)^2}{j\omega\mu[a(z)]^2}$$

$$^{I}T = T, \qquad ^{V}T = T + \frac{[J_1(\chi a)]^2 \tan \vartheta(z)}{\chi^2 N^2 a} \tag{109}$$

$$= T + 0.837a^{-1} \tan \vartheta(z),$$

with T given by the integral assoicated with $V(z)$ in (107).

16. EFFECT OF COUPLING ON DEGENERATE OR NEARLY DEGENERATE MODES

Degenerate modes are the modes which have the same velocity of propagation when the coupling is absent. With such modes the coupling may be very important even when its magnitude is small. The reason is: the transfer of wave motion from one such mode to the other will be cumulative in the direction of propagation. This effect is illustrated by directional couplers or by beats in two coupled pendulums having the same resonant frequencies. In such cases the resistance of the waveguide wall should not be neglected for it may have an important effect aside from introducing attenuation. Thus, no matter how small is the coupling, the degenerate modes should be considered as a group even though their coupling to other modes may be neglected.

The same is true of nearly degenerate modes as in the case of waves over a plane impedance sheet at large distances from the source, such as the current element in Fig. 8.

17. COAXIAL CONDUCTORS-CIRCULARLY SYMMETRIC MODES

Heretofore, we have considered waves in waveguides completely shielded from the external space. A complete shielding implies a coating of that surface of a waveguide which is exposed to the external space with a substance which is either a perfect electric conductor or a perfect magnetic conductor. In practice such a perfect shielding is impossible. The foregoing equations are thus approximate, even through the effect of approximations on waves in the guide may be negligible for all practical purposes. On the other hand, the effect of imperfect shielding on the "cross-talk" or interference between two waveguides may be im-

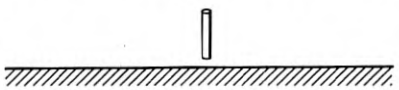


Fig. 8 — A vertical current element above an impedance sheet.

portant, especially at relatively low frequencies, even though the magnitude of cross-talk is small. The most practical way to calculate this cross-talk between two parallel waveguides, let us say, is to solve the above approximate telegraphist's equations for one waveguide. Then, we can obtain the tangential electric intensity on the outer surface of this waveguide from that on the inner surface. This electric intensity will be impressed on the "two-wire line" formed by the two waveguides. Resulting currents can be calculated, and from them one can obtain the tangential electric intensity on the inner surface of the second waveguide. Finally, we can obtain the waves in the second waveguide which are stirred up by the tangential electric intensity. This method is illustrated elsewhere.[14]

The same method can be used for a single waveguide in empty space — an impractical situation — if we wish to calculate the first approximation to the feeble external field. The rest of this section is of theoretical interest only. Our object is to show that it is possible to obtain a set of telegraphist's equations for a waveguide which includes external waves as well as internal.

As a concrete example we shall take a pair of coaxial cylinders and consider circularly symmetric modes. First, we shall derive the equations for the internal waves only — as we did in the preceding sections — and then point out the modifications which must be introduced in order to include the external waves. As usual we start with Maxwell's equations

$$\frac{\partial E_\rho}{\partial z} = -j\omega\mu H_\varphi + \frac{\partial E_z}{\partial \rho}, \qquad \frac{\partial H_\varphi}{\partial z} = -(g + j\omega\epsilon)E_\rho$$
$$E_z = \frac{1}{(g + j\omega\epsilon)\rho}\frac{\partial(\rho H_\varphi)}{\partial \rho} \tag{110}$$

and the boundary conditions

$$E_z(a, z) = Z_1 H_\varphi(a, z), \qquad E_z(b, z) = -Z_2 H_\varphi(b, z) \tag{111}$$

It is the boundary conditions that we shall have to modify when we wish to include the external modes. The rest of the derivation follows along the lines already discussed. We have the expansions for the transverse field components in terms of modes appropriate to perfectly conducting coaxial cylinders

$$E_\rho = \frac{V_0(z)}{\rho \log (b/a)} + \sum N_m^{-1} V_m(z) R_m'(\rho)$$
$$H_\varphi = \frac{I_0(z)}{2\pi\rho} + \sum N_m^{-1} I_m(z) R_m'(\rho) \tag{112}$$

The radial functions are defined by

$$\rho \frac{d^2R_m}{d\rho^2} + \frac{dR_m}{d\rho} + \chi_m{}^2\rho R_m = 0, \qquad R_m(a) = R_m(b) = 0 \tag{113}$$

Hence,

$$R_m(\rho) = J_0(\chi_m\rho)N_0(\chi_m b) - N_0(\chi_m\rho)J_0(\chi_m b) \tag{114}$$

where χ_m is a root of

$$J_0(\chi_m a)N_0(\chi_m b) - N_0(\chi_m a)J_0(\chi_m b) = 0 \tag{115}$$

The normalizing factors are obtained from

$$N_m{}^2 = 2\pi \int_a^b [R_m{}'(\rho)]^2\rho \, d\rho = 2\pi\chi_m{}^2 \int_a^b [R_m(\rho)]^2\rho \, d\rho \tag{116}$$

The longitudinal electric intensity may be obtained (in the present instance) from the third equation of the set (110) and from (112),

$$E_z = -\sum \frac{\chi_m{}^2}{(g + j\omega\epsilon)N_m} I_m(z)R_m(\rho), \qquad 0 \leqq \rho < a \tag{117}$$

However, we should remind the reader that the telegraphist's equations may be obtained without this equation since we can eliminate E_z from (110) before embarking on their derivation.

Multiplying the first equation of the set (110) by $N_p{}^{-1}R_p{}'(\rho)\rho \, d\rho \, d\varphi$ and integrating, we find

$$\begin{aligned}
\int_0^{2\pi}\int_a^b \frac{\partial E_z}{\partial\rho} N_p{}^{-1}R_p{}'(\rho)\rho \, d\rho \, d\varphi &= 2\pi E_z(\rho, z)N_p{}^{-1}R_p{}'(\rho)\rho \Big|_a^b \\
&\quad - \int_0^{2\pi}\int_a^b N_p{}^{-1}E_z \frac{d}{d\rho}[\rho R_p{}'(\rho)] \, d\rho \, d\varphi \\
&= 2\pi E_z(b,z)N_p{}^{-1}bR_p{}'(b) - 2\pi E_z(a,z)N_p{}^{-1}aR_p{}'(a) \\
&\quad + \chi_p{}^2 N_p{}^{-1} \int_0^{2\pi}\int_a^b E_z R_p(\rho)\rho \, d\rho \, d\varphi \\
&= 2\pi E_z(b,z)N_p{}^{-1}bR_p{}'(b) - 2\pi E_z(a,z)N_p{}^{-1}aR_p{}'(a) - \frac{\chi_p{}^2}{g + j\omega\epsilon} I_p(z)
\end{aligned} \tag{118}$$

Using the boundary conditions (111) and treating the second equation of the set (110) in the already familiar manner we obtain the distributed

parameters for the internal modes

$$
\begin{aligned}
Z_{pp} &= j\omega\mu + \frac{\chi_p^2}{g + j\omega\epsilon} + 2\pi N_p^{-2}(Z_1 a[R_p'(a)]^2 + Z_2 b[R_p'(b)]^2), \quad p \neq 0 \\
Z_{p0} &= N_p^{-1}[Z_1 R_p'(z) + Z_2 R_p'(b)], \quad p \neq 0 \\
Z_{00} &= \frac{j\omega\mu}{2\pi}\log(b/a) + \frac{Z_1}{2\pi a} + \frac{Z_2}{2\pi b}, \qquad Y_{00} = \frac{2\pi(g + j\omega\epsilon)}{\log(b/a)} \\
Y_{pm} &= g + j\omega\epsilon.
\end{aligned}
\tag{119}
$$

To include the external modes we shall pick a center for their origin. For a semi-infinite coaxial pair this center may be chosen on the axis near the end of the pair. For coaxial cylinders extending to infinity in both directions the center may be chosen arbitrarily on the axis but preferably near the source of internal waves. The external modes are then defined as in Sections 12 and 13 and the coupling between the external and internal modes is given by

$$
\begin{aligned}
E_z^i &= -(\eta_c \coth \sigma_c h)H_\varphi^i + (\eta_c \operatorname{csch} \sigma_c h)H_\varphi^e \\
E_z^e &= -(\eta_c \operatorname{csch} \sigma_c h)H_\varphi^i + (\eta_c \coth \sigma_c h)H_\varphi^e
\end{aligned}
\tag{120}
$$

where E_z^i and H_φ^i are taken at the inner surface of the outer cylinder and E_z^e and H_φ^e at the outer surface. In these equations η_c and σ_c are respectively the intrinsic impedance and propagation constant of the substance from which the outer cylinder is made. The thickness h of this outer cylinder is assumed to be small compared with its radius. Otherwise, the self and mutual impedances in (120) should be expressed in terms of the modified Bessel functions. Another assumption is that σ_c is very large compared with the propagation constants of various modes under consideration. For metal walls this assumption is highly satisfactory for all modes except those of exceedingly high order. In (118) we must substitute E_z^i for $E_z(b,z)$. In a corresponding equation for external mode we should use E_z^e as given by (120).

18. VANE ATTENUATORS

Our last example will be the "vane attenuator" in a rectangular waveguide, Fig. 9. The dotted line passing through AB represents a thin resistive sheet, so thin that the vertical current under the influence of a vertical field is distributed uniformly through the thickness of the sheet. Hence, the vertical electric intensity is continuous across the sheet. It is

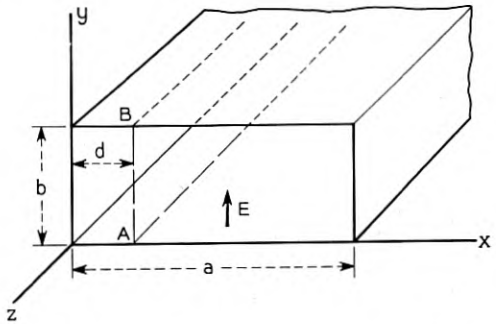


Fig. 9 — A rectangular waveguide with a thin resistive sheet.

not difficult to solve an appropriate boundary value problem. If we assume for simplicity that the guide walls are perfectly conducting and confine our attention to waves whose intensities are independent of the y-coordinate, Maxwell's equations separate in two sets, one involving E_y, H_x, and H_z and the other H_y, E_x, and E_z. We shall consider the first set [see (63)]. The resistive sheet implies a discontinuity in H_z,

$$H_z(d - 0, z) - H_z(d + 0, z) = YE_y(d, z) \tag{121}$$

where Y is the admittance of a unit area of the sheet. This will include not only the conductance of a thin metallic film but also the capacitance of a thin plastic film on which the metal may be deposited. The usual solution of the boundary problem will be obtained by assuming two separate fields, one for the region to the left of the sheet and one for the region to the right of it. Taking into consideration the continuity of E_y and the discontinuity in H_z, we shall find a transcendental equation for the propagation constants of the various modes appropriate to the waveguide with a thin resistive sheet.

Here, however, we shall express the field in the waveguide with the sheet in terms of modes appropriate to the same waveguide without the sheet. Thus we assume expansions (64) for E_y and H_x and (65) for H_z. Since E_y and H_x are continuous functions, their sine series are uniformly convergent as well as differentiable. On the other hand, H_z is discontinuous and neither differentiable nor uniformly convergent. This non-differentiability affects the calculation of the following integral

$$P_m = \int_0^b \int_0^a \frac{\partial H_z}{\partial x} N_m^{-1} \sin \frac{m\pi x}{a} \, dx \, dy \tag{122}$$

needed in the conversion of Maxwell's equations into telegraphist's

equations. First we have to split it into two integrals

$$P_m = \int_0^b \int_0^{d-0} + \int_0^b \int_{d+0}^a \tag{123}$$

Then we have to integrate it by parts,

$$P_m = bH_z N_m^{-1} \sin \frac{m\pi x}{a} \Big|_0^{d-0} + bH_z N_m^{-1} \sin \frac{m\pi x}{a} \Big|_{d+0}^a - \int_0^b \int_0^a \frac{m\pi}{a} H_z \cos \frac{m\pi x}{a} \, dx \, dy \tag{124}$$

And finally we have to substitute H_z from the third equation of the set (63) into the integrand of (124) and integrate by parts once more before substituting the series for E_y from (64). In this way we find

$$P_m = bN_m^{-1}[H_z(d-0, z) - H_z(d+0, z)] \sin \frac{m\pi d}{a} + \frac{m^2\pi^2}{j\omega\mu a^2} V_m(z) \tag{125}$$

The bracketed term may be expressed in terms of V_n's if we use (64) and (121). In this way, we obtain the following telegraphist's equations

$$\frac{dV_m}{dz} = -j\omega\mu I_m$$

$$\frac{dI_m}{dz} = -\left(g + j\omega\epsilon + \frac{m^2\pi^2}{j\omega\mu a^2} + \frac{2Y}{a} \sin^2 \frac{m\pi d}{a}\right) V_m - \sum_n{}' \frac{2Y}{a} \sin \frac{m\pi d}{a} \sin \frac{n\pi d}{a} V_n \tag{126}$$

where the prime after the summation signs signifies the omission of the term corresponding to $n = m$.

19. ARBITRARINESS OF MODAL TRANSVERSE FIELD PATTERNS

In almost all examples considered by us the variations of transverse field components in transverse planes were expressed in terms of functions associated with orthogonal modes in waveguides of uniform cross-section and with perfectly conducting walls. An exception was made in Section 10 where we used curvilinear coordinates. The guiding principle in selecting the basic set of transverse field patterns for general field representation should be in most cases, but not in all cases, the minimization of coupling coefficients. That there are exceptions was made clear in

connection with the junction between two waveguides, one with perfectly conducting and the other imperfectly conducting walls, Fig. 2 (see remarks toward the end of Section 3). Aside from convenience the choice of transverse modal patterns is rather arbitrary. A set must be complete, that is, adequate for representing any field which can exist inside the guide. It should be an orthogonal set; this will enable us to obtain a set of telegraphist's equations in which each equation contains only one derivative with respect to the direction of propagation. But, as we have already seen, the sets of terms representing the individual "modes" do not have to satisfy either Maxwell's equations or the boundary conditions. The situation is similar to that which confronts us when we choose a set of meshes in a network in order to write Kirchhoff's equations in terms of mesh currents.

In the case of circular waveguides, for instance, we can express E_ρ and H_φ in terms of the "sawtooth" functions in which case E_z will be expressed in terms of "square sine" functions. It is not a convenient set; but, certainly, it is a permissible set.

20. CONCLUDING REMARKS

In the preceding sections we have illustrated the technique of conversion of Maxwell's equations into generalized telegraphist's equations by several typical examples. In many instances this technique is a practical method for solving field problems. This method may be valuable even when the more conventional methods can be used. Consider a slightly deformed rectangular waveguide in which two faces are arcs of coaxial cylinders and the other two faces are radial planes, Fig. 10. If we use cylindrical coordinates, we can separate the variables and obtain a set of orthogonal modes in which fields are expressed in terms of Bessel functions. As the curvature decreases these modes become more and more like the corresponding modes in a strictly rectangular waveguide. Nevertheless the mathematical machinery remains different. No matter how

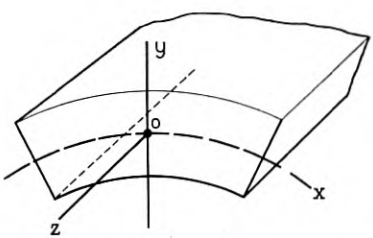


Fig. 10 — A deformed rectangular waveguide.

small is the curvature we still have to deal with Bessel functions rather than with sines and cosines. We can and will, of course, replace Bessel functions by their asymptotic expansions. This will simplify mathematics. But we still would be left with a "discontinuity in thinking" about the zero and non-zero curvature cases. At any rate it seems that we can gain something in understanding the effect of the gradual deformation on the field, particularly if this deformation is varying along the guide, by formulating the problem in terms of "deformed cartesian" coordinates. Then the effect of deformation will be thought of as coupling between various modes in a strictly rectangular waveguide. The coupling coefficients can be evaluated and numerical results thus obtained for more general conditions than is possible by the conventional method.

In other cases, numerical calculations, although possible in theory, would perhaps be prohibitive in practice. Even then this technique may contribute toward the qualitative understanding of physical phenomena. Consider two wires diverging from the terminals A, B of a generator, Fig. 11. Let us imagine a family of spheres concentric with the midpoint of the segment AB. Let us consider the sections of wires intercepted by a typical sphere as sections of two cones with their apices at the center of the spheres. For such cones we can obtain a set of orthogonal modes. The transverse field distributions associated with these modes we now take for representing the field distribution in the actual case, just as we did in previous examples. One of the infinite system of such modes will be the principal mode which at sufficiently large distances from A,B will be the usual "transmission line" mode for two parallel wires (that is, when the wires actually do become parallel). It would not be difficult as a matter of fact to obtain telegraphist's equations for this mode together with coupling coefficients to the higher order modes. For perfectly conducting wires these coupling coefficients become progressively smaller

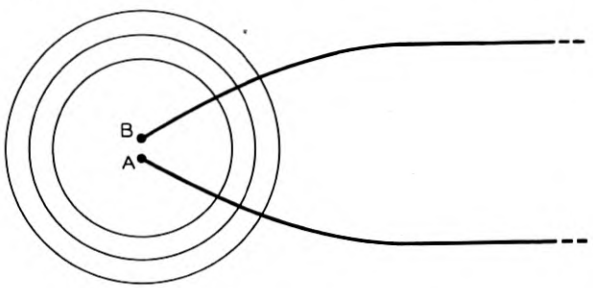


Fig. 11 — Wires diverging from the terminals of a generator.

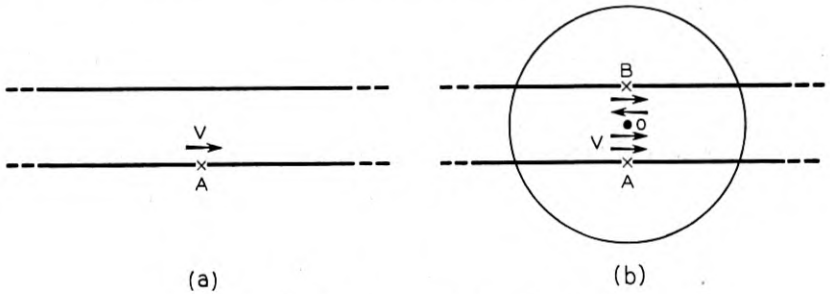


Fig. 12 — Parallel wires with a generator at A.

as the distance from the generator increases. Thus the higher order modes will be generated largely in the vicinity of the generator if we define the "vicinity" as the interior of a sphere whose radius is a reasonably large multiple of the final distance between the wires. These are the modes which will carry off to infinity what we usually call the "radiated energy." There will be very little radiation if the distance between the wires never exceeds a small fraction of the quarter-wavelength. This is because the higher order modes are substantially non-propagating at distances close to the center of their origin.

For thin wires the calculation of transverse patterns needed for telegraphist's equations requires the solution of a transcendental equation.[15] To use this equation in the present case we should replace the oval traces of the wires on a typical "wavefront" sphere by equivalent circles, that is, circles giving the same shunt capacitance in the principal mode. A more accurate analysis would be possible but hardly worth the effort.

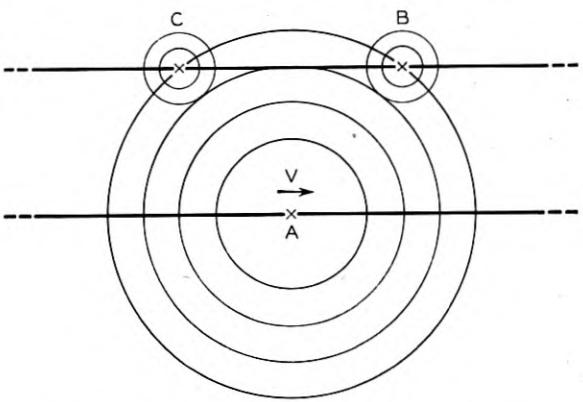


Fig. 13 — Parallel wires and a succession of primary and secondary waves.

An analysis of waves on two parallel wires, such as Mie's,[2] is not realistic since he assumed that his generator is at infinity. Sometimes such an assumption is not objectionable; but at other times one is better off without it. If two parallel wires are infinite in both directions and a generator is connected to one wire, Fig. 12(a), and if the distance between the wires is small, one can conveniently replace the impressed voltage by the sum of push-push and push-pull voltages, Fig. 12(b), to take advantage of the symmetry. Then outside some sphere concentric with the mid-point 0, we have four wires "diverging" from 0 and the analysis may proceed along the lines suggested for two wires. If, however, the distance between the wires is large, we shall find it more expedient to consider waves on a single wire generated at point A, Fig. 13, which in their turn generate waves on the second wire at points where the spherical wavefronts intersect it. Those waves impinge on the first wire and generate tertiary waves.

Many other examples will occur to the reader in which the telegraphist's equations will be useful to a greater or lesser extent.

REFERENCES

1. Sir William Thomson (Lord Kelvin), Mathematical and Physical Papers, **2**, p. 79, Cambridge University Press (1884).
2. G. Mie, Electrische Wellen an Zwei Parallelen Drahten, Ann. der Phys. Ser. 4, **2**, pp. 201–249, 1900.
3. John R. Carson, Electromagnetic Theory and the Foundations of Electric Circuit Theory, B.S.T.J. **6**, pp. 1–17, Jan., 1927.
4. S. A. Schelkunoff, Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields, B.S.T.J., Oct., 1934.
5. S. A. Schelkunoff, Electromagnetic Waves, p. 282, D. Van Nostrand Co., 1943.
6. S. A. Schelkunoff, Transmission Theory of Plane Electromagnetic Waves, Proc. I.R.E., **25**, pp. 1457–1492, Nov., 1937.
7. S. A. Schelkunoff, Generalized Telegraphist's Equations for Waveguides, B.S.T.J., **31**, pp. 784–801, July, 1952.
8. S. P. Morgan, Jr., Mode Conversion Losses in Transmission of Circular Electric Waves through Slightly Non-cylindrical Guides, J. Appl. Phys., **21**, pp. 329–338, April, 1950.
9. S. A. Schelkunoff, Electromagnetic Waves, D. Van Nostrand Co., 1943, pp. 485–488.
10. W. J. Albersheim, Propagation of TE_{01} Waves in Curved Waveguides, B.S.T.J., pp. 1–32, Jan., 1949.
11. Reference 5, pp. 92–94.
12. E. T. Whitaker and G. N. Watson, A Course of Modern Analysis, Cambridge University Press, pp. 78–80, 1927.
13. S. A. Schelkunoff, Applied Mathematics for Engineers and Scientists, D. Van Nostrand Co., pp. 154–155; 1948.
14. S. A. Schelkunoff and T. M. Odarenko, Crosstalk Between Coaxial Transmission Lines, B.S.T.J., **16**, pp. 144–164, April, 1937.
15. S. A. Schelkunoff, Advanced Antenna Theory, John Wiley & Sons, 1952, equation (312), p. 98.

A Method for Synthesizing Sequential Circuits

By GEORGE H. MEALY

(Manuscript received May 6, 1955)

The theoretical basis of sequential circuit synthesis is developed, with particular reference to the work of D. A. Huffman and E. F. Moore. A new method of synthesis is developed which emphasizes formal procedures rather than the more familiar intuitive ones. Familiarity is assumed with the use of switching algebra in the synthesis of combinational circuits.

CONTENTS

1. INTRODUCTION

1.1 *Foreword*

The designer of a sequential switching circuit — a circuit with storage or "memory" — faces a far more difficult problem than is faced by the

designer of, say, a simple translating circuit. In the latter case, comparatively simple and straightforward methods of synthesis are known.[1] In the former case, the designer frequently does not even know how to begin to solve the problem. Only recently did D. A. Huffman develop a method which, at an early point in the design, gives rather explicit procedures for carrying the design through to completion.[2] The method relies for its success on a tabular method of presenting the circuit requirements. This table, called a *flow chart*, may be subjected to simple manipulations which remove redundancies in the verbal statement of the circuit requirements. When supplemented by somewhat more complicated procedures, the flow chart is reduced to a form which leads directly to a circuit having a minimal number of storage elements. This process will be called *reduction* in this paper, and direct manipulation of the flow chart will be called *merging*.

Independently, E. F. Moore investigated the abstract properties of sequential circuits.[3] In particular, Moore asked what can be said about a circuit when one knows nothing about it except what may be inferred by performing experiments involving only the input and output terminals of the circuit. A by-product of Moore's theory was a general method for reducing (if necessary) a circuit whose description is completely known.* This method is essentially the same as Huffman's methods, *sans* flow chart manipulation.

The situation, then, is the following: Once a flow chart, or some equivalent statement of circuit requirements, has been obtained, one may use Moore's procedure for reducing the circuit. Once the circuit has been completely reduced, the remainder of the synthesis procedure is fairly uncomplicated. On the other hand, one may use the merging process of Huffman on the flow table. Very often this will result in complete reduction; less often it will be necessary to use additional procedures equivalent to the Moore process. Merging, when it is possible, is easier to use than is the Moore procedure, hence one would like to find a method which is as simple as merging and at the same time results in complete reduction more often than does merging.

Huffman's method was originally developed in connection with relay circuits, although it is applicable in other instances. It does not, however, always work in its unmodified form when applied to switching circuitry of the type that is commonly used in the design of digital computers.[4, 5] One then asks, how can Huffman's method be extended to cover such instances?

* We shall use the word "circuit" to refer both to physical circuits and to abstract representations of circuit requirements (such as flow charts). The latter of course, may correspond to many physical circuits.

This paper offers one possible solution to both questions. After describing an abstract model for sequential circuits, we develop Moore's method for reduction, as it applies to our model. We then develop a new method applicable to synchronous circuitry, which is commonly used in computer design. Finally, the method is extended to relay circuitry as an example of asynchronous circuitry. The relationship between our method and Huffman's method, as they are applied to this class of circuits, is then explained.

1.2 *Introductory Remarks*

It is very tempting at the outset to make the flat statement: There is no such thing as a synchronous circuit. This would be strictly true if we defined a synchronous circuit as one with the properties:

(*S1*) *Any lead or device within the circuit may assume, at any instant of time, only one of two conditions, such as high or low voltage, pulse or no pulse.*

(*S2*). *The behavior of the circuit may be completely described by the consideration of conditions in the circuit at equally-spaced instants in time.** Because it is quite clear that no *physical* circuit satisfies (S1) and (S2), such a blanket statement would be a quibble, for the engineer does recognize a certain class of circuits which he calls synchronous. The unfortunate fact is that the distinction between a synchronous and an asynchronous circuit is very hazy in many cases of actual engineering interest. Roughly, we may say that the more nearly a circuit satisfies (S1) and (S2), the more likely will an engineer be to identify it as a synchronous circuit.

As intuitive guides to the usual properties of a synchronous circuit, these characteristics are offered:

(1). There is a so-called clock which supplies timing pulses to the circuit.

(2). Inputs and outputs are in the form of voltage or current pulses which occur synchronously with pulses from the clock.

(3). The repetition rate of the clock pulses may be varied, within limits, without affecting the correct operation of the circuit, so long as input pulses remain synchronized with the clock.

Another assumption that is commonly made, although it does not bear on the distinction between synchronous and asynchronous circuits, should nevertheless be mentioned. If this assumption is made, then we may distinguish between combinational and sequential circuits.

* Actually, these need not be equally-spaced. However, the instants considered must not depend on any property of any sequence of inputs presented to the circuit, such as the duration of a pulse.

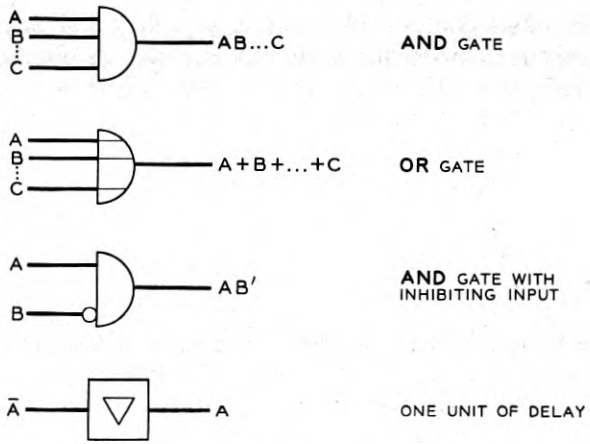


Fig. 1

(*D*). *Certain circuits contain no time delay — their input combinations in every case completely determine their output combinations.*

We will be concerned mainly with a technology in which these assumptions are nearly satisfied — that of the type employed in Leiner *et al.*[4, 5] In this technology, one uses AND gates (with or without inhibiting inputs), OR gates, delay lines, and amplifiers. For our purposes, we may ignore the need for amplifiers. The other basic circuits are as shown in Fig. 1. The properties of these circuit blocks are defined by the algebraic expressions in the illustration.*

The familiar switching (or Boolean) algebra is used, where 0 stands for no pulse, 1 for pulse, + for OR, · for AND and ( )′ for NOT. It is assumed that the reader is familiar with switching algebra and its use in practical design problems. We recall from switching algebra:

(1) A *switching function* is any (finite) expression in switching algebra.

(2) A *minimal polynomial* of n variables is any product of the form:

$$x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$$

where

$$x_i^{a_i} = \begin{cases} x_i' & a_i = 0 \\ x_i & a_i = 1 \end{cases}$$

(3) We define

$$P_j = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$$

* The unit of delay is the interval between the start of two successive clock pulses. The notation "$\bar{A}$", used in Fig. 1, will be explained in Section 2.1.

where j is the decimal form of $a_1a_2 \cdots a_n$, considered as a binary number. For example, if $n = 3$, $P_0 = x_1'x_2'x_3'$, $P_1 = x_1'x_2'x_3$, etc.

(4) Every switching function of n variables may be brought into a unique *canonical form*:

$$f(x_1, \ldots, x_n) = \sum_{i=0}^{2^n-1} f_iP_i$$

where

$$f_j = f(a_1, a_2, \cdots, a_n)$$

(5) Corresponding to each function is a *truth-table* which displays the value of the function for each set of arguments. For $n = 2$, the truth-table corresponding to the canonical form is found in Table I. The correspondence between the truth-table and canonical form is one-to-one.

For further information about switching algebra see, for instance, Reference 9.

As an example, consider the function

$$f(x, y) = x' + y'$$

Its truth-table is Table II, and, therefore, $f_0 = f_1 = f_2 = 1$ and $f_3 = 0$. The canonical form is

$$f(x, y) = x'y' + x'y + xy'$$

2. A MODEL FOR SEQUENTIAL CIRCUITS

2.1 *The Model*

We begin by giving an abstract definition of a *switching circuit*:

A switching circuit is a circuit with a finite number of inputs, outputs.

TABLE I

x_1	x_2	$f(x_1, x_2)$
0	0	f_0
0	1	f_1
1	0	f_2
1	1	f_3

TABLE II

x	y	$f(x, y)$
0	0	1
0	1	1
1	0	1
1	1	0

and (internal) states. Its present output combination and next state are determined uniquely by the present input combination and the present state. If the circuit has one internal state, we call it a combinational circuit; *otherwise, we call it a* sequential circuit.

We have now to explain what we mean by this definition when we apply it to the technology introduced in Section 1. First, we assume a circuit has n binary-valued *input variables*, $x_1, x_2, \cdots, x_n$; m binary-valued *output variables*, $y_1, y_2, \cdots, y_m$; s binary-valued *excitation variables*, $\bar{q}_1, \bar{q}_2, \cdots, \bar{q}_s$; and s binary-valued *state variables*, $q_1, q_2, \cdots, q_s$, corresponding one-to-one with the excitation variables. In order to facilitate discussion, we note that a set of minimal polynomials may be associated with each set of variables. Specifically, corresponding to the input variables, we have the *input combinations*, X_j; associated with the output variables are the *output combinations*, Y_l; corresponding to the excitation variables are the *next states*, $\bar{Q}_k$; and with the state variables, we associate the *present states*, Q_i. For example, if $n = m = s = 3$, we have:

$$X_4 = x_1x_2'x_3'$$

$$Y_2 = y_1'y_2y_3'$$

$$\bar{Q}_1 = \bar{q}_1'\bar{q}_2'\bar{q}_3$$

$$Q_7 = q_1q_2q_3$$

We will use this notation and terminology for convenience. Rather than stating that, at some time, $x_1 = 1$, $x_2 = 0$, and $x_3 = 0$, we will say that input combination X_4 (or its equivalent — input combination 100) is present. That is, $X_4 = 1$ and thus the inputs are, respectively, 1, 0, and 0.

Now, according to the definition given above, to each circuit we must be able to assign some set of equations relating the $\bar{q}_i$ and y_i to the x_i and q_i. These equations will have the general form:

$$\bar{Q}_k = \bar{Q}(Q_i, X_j)$$

$$Y_\ell = Y(Q_i, X_j)$$

That is, k and ℓ must be uniquely determined by i and j. Each circuit is associated with a truth-table with its columns headed (in order):

$$q_1, \cdots, q_s, \quad x_1, \cdots, x_n; \quad \bar{q}_1, \cdots, \bar{q}_s, \quad y_1, \cdots, y_m.$$

The number of circuits having n input, m output, and s internal vari-

ables is equal to

$$2^{(m+s)2^{(n+s)}}$$

since the truth table has $2^{(n+s)}$ rows and $m + s$ columns which must be filled in with 0's and 1's.

The interpretation of this model is now fairly straightforward. We have assumed (S1), (S2), and (D) and know that, physically, the delay unit provides storage. We assign the $\bar{q}_i$, the excitation variables, to the inputs of delay lines, and we assign the q_j, the state variables, to delay line outputs. The present state of the circuit is the combination of conditions on the delay line outputs. The next state is the combination of conditions on the delay line inputs, since one time unit later this combination will be present on the outputs.

To make the discussion concrete, consider Fig. 2. The circuit equations are:

$$\bar{q}_1 = q_1'q_2' + x'q_2'$$
$$\bar{q}_2 = q_1q_2' + xq_1$$
$$y = q_1'q_2'$$

From these equations, we write Table III.

2.2 *State Diagrams*

It is usually not clear from an examination of the circuit diagram or circuit equations just what a sequential circuit does. The truth-table

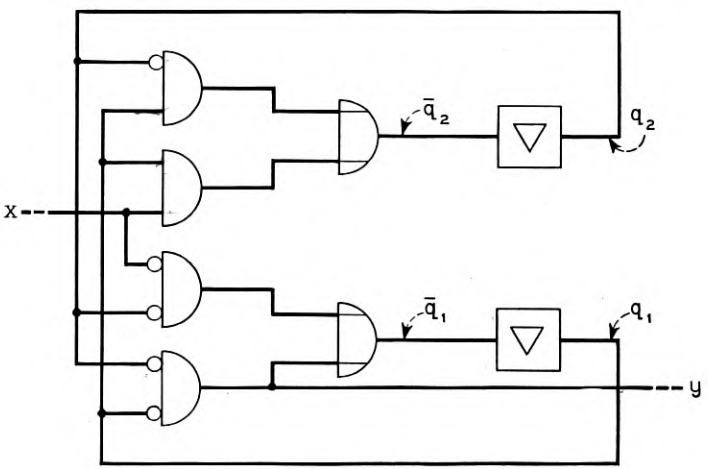


Fig. 2

TABLE III

q_1	q_2	x	$\bar{q}_1$	$\bar{q}_2$	y
0	0	0	1	0	1
0	0	1	1	0	1
0	1	0	0	0	0
0	1	1	0	0	0
1	0	0	1	1	0
1	0	1	0	1	0
1	1	0	0	0	0
1	1	1	0	1	0

is more helpful and tells the whole story if we put it in a different form, called a *state diagram*. In this diagram, circles will represent states. Each line of the truth-table will be represented by an arrow going from the present to the next state. A label on the arrow will give the corresponding input and output combination. The state diagram for the circuit discussed in Section 2.1 is given in Fig. 3.

The arrows in the state diagram correspond to changes of state of the associated circuit, and both the arrows and the changes of state are called *transitions*. A transition begins at a present state and ends at the next state. The transition is labeled X/Y. X is an input combination and Y is the corresponding output combination.

As an example, consider Table IV, which gives the sequences of states and outputs which correspond to each initial state of the circuit and the input sequence 100. Depending upon what state the circuit is started in, the input sequence 100 produces three different output sequences. It is

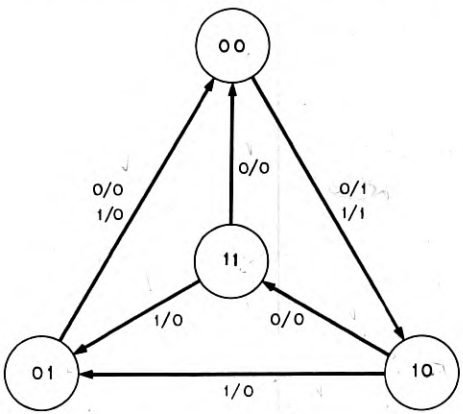


Fig. 3

TABLE IV

x	0 0 1	1 0 0	1 0 0	1 0 0
q_1	0 0 1	1 0 0	1 0 0	0 1 1
q_2	1 0 0	1 1 0	0 1 0	0 0 1
$\bar{q}_1$	0 1 1	0 0 1	0 0 1	1 1 0
$\bar{q}_2$	0 0 1	1 0 0	1 0 0	0 1 0
y	0 1 0	0 0 1	0 0 1	1 0 0

difficult and probably of little value to put into words exactly what this particular circuit does. However, given any initial state and any sequence of inputs, we can immediately tell what happens from the state diagram. (The truth table may be used for the same purpose, but less easily. It is far more difficult to determine circuit behavior by chasing signals around the circuit diagram.) The problem of circuit analysis is now completely solved. Given any circuit, we may immediately write its circuit equations. A truth-table is easily obtained from the equations. Given the truth-table or given the associated state diagram, we may determine exactly how the circuit behaves for any initial state and input sequence.

Conversely, once a state diagram or truth-table is found for a proposed circuit, the above steps may be traced backwards in order to arrive at a circuit diagram. The only problem here is designing combinational circuits economically. The really significant problem in sequential circuit synthesis is finding a suitable state diagram or truth-table. This problem, in turn, may be subdivided into two problems:

(1) finding *any* state diagram or its equivalent which fulfills the circuit requirements and

(2) reducing this to the state diagram which is to be used for the final part of the design process.

The next section of this paper develops Moore's method of reduction and is basic in justifying the methods developed in the succeeding sections.

3. CIRCUIT EQUIVALENCE

3.1 *Moore's Theory*

The key to the synthesis of sequential circuits is the concept of circuit equivalence which was discovered independently by Huffman[2] and Moore.[3] We are concerned mainly with the portions of Moore's theory which have direct application to synthesis; certain differences in treatment are necessary since Moore's model for sequential machines is differ-

ent from ours. All of Moore's arguments carry over with only slight changes.

Roughly speaking, we call two circuits equivalent if we cannot tell them apart by performing experiments involving only their inputs and outputs. Once we have solved the first problem of synthesis by finding any state diagram which fulfills the circuit requirements it will usually be found that the state diagram has more states than are necessary to perform the assigned task. In such a case, we usually wish to simplify the circuit by removing redundant states in such a way that the final circuit is equivalent to the original one.

We must now make the concept of equivalence more precise. We define:

(1) *Two* states, Q_i *in circuit* S *and* Q_j *in circuit* T, *are called* equivalent *if, given* S *initially in state* Q_i *and* T *initially in state* Q_j, *there is no sequence of input combinations which, when presented to both* S *and* T, *will cause* S *and* T *to produce different sequences of output combinations.*

(2) *Two* circuits, S *and* T, *are called* equivalent *if, corresponding to each state* Q_i *of* S, *there is at least one state* Q_j *of* T *such that* Q_i *is equivalent to* Q_j; *and corresponding to each state* Q_j *of* T *there is at least one state* Q_k *of* S *such that* Q_j *is equivalent to* Q_k.

In (1), it should be noted that T may be a copy of S. Hence (1) is also a definition for equivalence between states in the same machine. Moore has shown that even if no two states in a given machine are equivalent, it is not always possible to find out what state the machine started in by some experiment. That is, there is not always a sequence of input combinations which will result in a different sequence of output combinations for each possible initial state of the circuit. The state diagram of Fig. 3 is the example used by Moore to prove this; state 11 may not be distinguished from state 10 by any experiment which begins with a 1, and state 01 may not be distinguished from state 11 by any experiment which begins with a 0.

If there are two states in a circuit which are equivalent, it should be possible to eliminate one of them. This will result in a circuit equivalent to the original circuit. This is indeed possible, and the process of reduction may be carried out in an essentially unique manner, as is stated by

Theorem 1 (Moore). Corresponding to each circuit, S, *is a circuit* T *which has the properties: (1)* T *is equivalent to* S, *(2)* T *has a minimal number of states, (3) no two states in* T *are equivalent, and (4)* T *is unique, except for circuits that result from* T *by relabeling its states.* T *is called the* reduced form *of* S.

We shall state the procedure to be followed in deriving T from S,

referring the reader to Reference 3 for a complete proof of Theorem 1. First, divide the states of S into sets such that (1) all states in a given set are equivalent, (2) if a state is in a given set then all states equivalent to that state are also in the same set, and (3) no state is in two different sets. These sets are called *equivalence sets* or *classes.* Now, assign a state of T to each equivalence set of states. If there is a transition, bearing the symbol X/Y, from a state in one equivalence set of S to a state in a different equivalence set of S, insert a transition bearing the same symbol X/Y between the corresponding states in T. If there is a transition between two states in the same equivalence set of S, insert a transition in T which begins and ends at the corresponding state of T. Do this for all transitions in S.

We have not given as yet an effective procedure for determining the equivalence sets. This procedure will be provided by the method of proof of the next theorem. Before stating the theorem, we state a precise definition of what we mean by "experiment." By an *experiment of length* k, we mean the process of presenting a circuit which is in some specified initial state with a sequence of k successive input combinations. By the *result* of an experiment, we mean the sequence of output combinations produced by the experiment. We say that two states are *indistinguishable by any experiment of length* k if for all experiments of length k the result does not depend on which was the initial state. We may now state

Theorem 2 (Moore). Given a circuit S whose reduced form has a total of p states, then for any two states, Q_i *and* Q_j *, in S,* Q_i *is equivalent to* Q_j *if and only if* Q_i *is not distinguishable from* Q_j *by any experiment of length* $(p - 1)$.*

Proof: Consider all experiments of length k. All states may be divided into equivalence sets by the rule: put two states in the same equivalence set if and only if they are indistinguishable by any experiment of length k. For each k, there is now defined a set of equivalence sets which we will call P_k .

Consider two states, a and b, that are not equivalent but are indistinguishable by any experiment of length k. Since a and b are not equivalent, there is an experiment of some minimum length, say n, that will distinguish a from b. Consider the two states, $\bar{a}$ and $\bar{b}$, that a and b are taken into by the first $(n - k - 1)$ input combinations of the experiment. $\bar{a}$ and $\bar{b}$ are then distinguishable by an experiment of length $(k + 1)$ but by no shorter experiment.

We have now proved that P_k is not already the set of equivalence sets

* This theorem is a trivial extension of Moore's result.

TABLE V

q_1	q_2	x	$\bar{q}_1$	$\bar{q}_2$	y
0	0	0	0	0	0
0	0	1	1	0	0
0	1	0	0	0	1
0	1	1	0	0	1
1	0	0	0	1	0
1	0	1	0	1	0
1	1	0			
1	1	1			

As a more complete example, including the construction of a reduced machine, consider Fig. 4(a). Applying Rule I, we get:

$$\begin{array}{ll} \bar{P}_1 : (0)\ (1, 2, 3, 4) & \\ \quad\ \ \uparrow\!\_\_\_\_\!\urcorner\ \ \uparrow & 1 \\ \bar{P}_2 : (0)\ (2, 4)\ (1, 3) & \end{array}$$

To construct the reduced circuit, assign state A in the new circuit to (0), B to (1, 3), and C to (2, 4). The resulting circuit is shown as Fig. 4(b).

In order to develop a physical circuit, it is necessary to assign a binary code to the states. The assignment is more or less arbitrary for synchronous circuits, but will in general affect the number of circuit elements used. In this instance, we choose to make the assignment:

$$\text{A} \rightarrow 01$$

$$\text{B} \rightarrow 00$$

$$\text{C} \rightarrow 10$$

Rewriting the state diagram as a truth-table, we get Table V. Two rows in the right half of the truth-table are blank, since state 11 does not appear in the state diagram. It is legitimate to fill these rows in in any way, and it is preferable to fill them in in a manner that results in simplification of the final circuit. Taking advantage of this fact, we may set:

$$\bar{q}_1 = q_1' q_2' x$$

$$\bar{q}_2 = q_1$$

$$y = q_2$$

The final circuit is shown in Fig. 5. Fig. 6 shows the state diagram for the completed circuit. As it happens, state 11 is not equivalent to any other state.

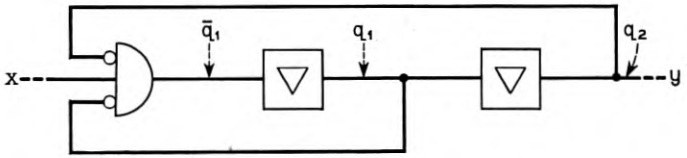


Fig. 5

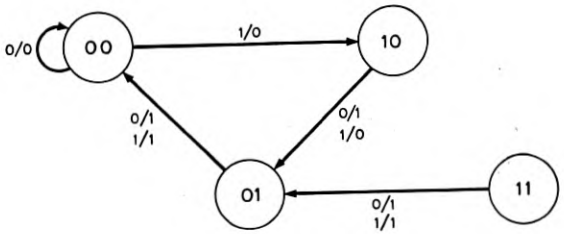


Fig. 6

This concludes the material on circuit equivalence. In the following section, we develop the method for synchronous circuits. As will be seen, an essential feature of the method is the use of truth-tables rather than state diagrams (which become unmanageable for circuits with more than a few variables) and a very much simplified form of Rule I which may be applied directly to truth-tables. Our program will be (1) to describe the kind of argument used in going from verbal circuit requirements to a truth-table; (2) to restate Rule I in a form (Rule II) which is adapted to synthesis and applies to truth-tables; (3) to develop Rule III, a generalized form of Huffman's merging process; (4) to discuss "don't care" situations, familiar to the reader from the study of combinational circuits; and (5) to give a summary of the method. A complete design example will be given in Section 5.5, following application of the method to asynchronous circuits.

4. DEVELOPMENT OF THE METHOD FOR SYNCHRONOUS CIRCUITS

4.1 *Introductory Remarks*

As seen in the last section, the first problem in synthesis is finding some state diagram that will behave according to the circuit requirements. The state diagram need not be very efficient in the sense that it may have far more states than are actually needed, for the procedures developed in the last section give a straightforward procedure for removing redundant states. Unfortunately, the initial step in the process relies heavily on

the designer's ingenuity. However, we can outline procedures that are of some assistance in finding an initial state diagram.

The simplest case, and indeed the only wholly straightforward case, is that in which the circuit must always return to its initial state after it has received some fixed number of input combinations. Essentially, this case is simple because we may consider all possible input sequences. We assign a new state any time anything happens, up to the last input. The last input then takes us back to the initial state. For instance, suppose that we want a circuit which receives sets of three binary digits in serial form and puts a pulse out on one of eight leads during the third digit to indicate the number that was received. The state diagram may immediately be written down, as shown in Fig. 7. Rather than write sets of 8 binary digits for the output symbols, we have designated the lead that should be energized, if any, and otherwise have written "0".

It is immediately clear that this is even a reduced machine — no two states are equivalent. This is an extreme case; usually there will be certain sequences of inputs which will never occur and/or certain sequences of inputs for which (in Huffman's words) we do not care to specify the circuit action. More often, however, there will be patterns of successive input combinations that will produce the same circuit action. For instance, suppose that in a sequence of 4 inputs we wish to have a final output only if the input sequence is 1010 or 0101. Then we can draw a state diagram showing all sequences which is shown as Fig. 8(a). How-

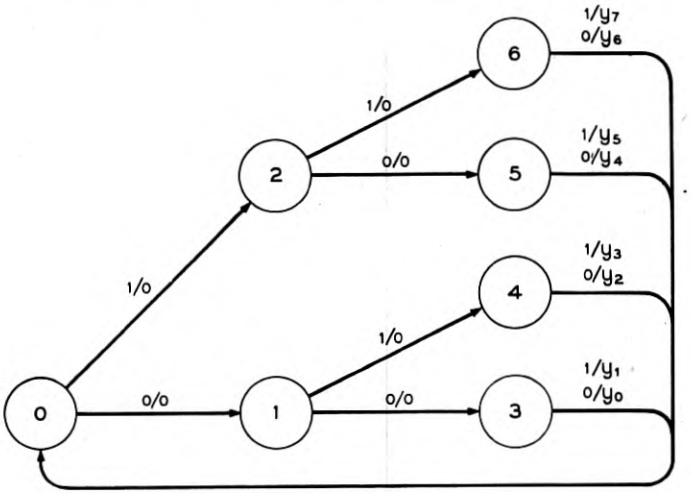


Fig. 7

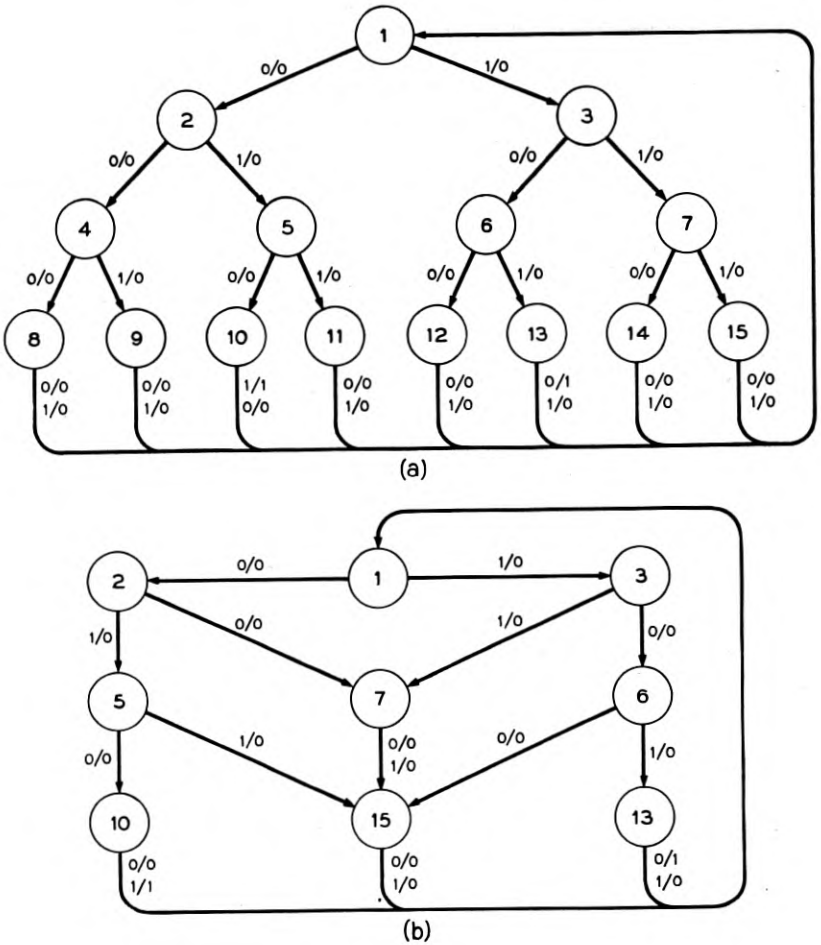


Fig. 8

ever, with a modest amount of ingenuity, we might have drawn Fig. 8(b) as our first attempt. In fact, it is clear that Fig. 8(b) shows the reduced form of the diagram in Fig. 8(a).

On the other hand, if there is no state which is entered cyclically, as above, no really explicit directions may be given for drawing an initial state diagram. In practice, one starts to draw a branching diagram such as the above. To terminate each branch, it is necessary to recognize that each transition from the state at the end of the branch may terminate in

TABLE VI

Input combination....	0	1	0	1
Present State	Next State		Output Combination	
1	2	3	0	0
2	4	5	0	0
3	6	7	0	0
4	8	9	0	0
5	10	11	0	0
6	12	13	0	0
7	14	15	0	0
8	1	1	0	0
9	1	1	0	0
10	1	1	0	1
11	1	1	0	0
12	1	1	0	0
13	1	1	1	0
14	1	1	0	0
15	1	1	0	0

some state which is already in the diagram. To the author's knowledge, no more specific directions for this are possible.

In practice, large state diagrams become very messy to draw. Where this is the case, it is better to revert to the truth-table, recast in a matrix form with states corresponding to rows and input combinations corresponding to columns. One of the most valuable features of this mode of presentation is that the truth table may be used directly to perform a large part of the reduction process. To illustrate the truth-table in a simple case consider Table VI which is the truth-table corresponding to the state diagram of Fig. 8(a). Of the two portions of the table, the left hand one represents the next states and the right hand one gives the output combinations.

4.2 *Modification of First Reduction Process*

At this point, we give an extension of Rule I which applies to truth-tables. It was noted above that Moore's theory assumes that each machine is completely specified, although the specification is not known to the experimenter. In our restatement of Rule I, we must allow for the possibility of blank entries in the truth-table. This provision amounts to calling two circuits equivalent if there is no evidence for believing that they are not equivalent.*

* This procedure is essentially that stated in Reference 2, pp. 183–185.

Rule II: Separate the rows of the truth-table into sets such that two rows are in the same set if and only if no corresponding entries in the right-hand portion of the rows are contradictory. (A blank entry is not considered to contradict any entry.) Call these sets "$\overline{P}_1$." Given the set of sets $\overline{P}_k$, find if possible two rows in the same set of $\overline{P}_k$ such that for some input combination the two rows have row designations (next states) which are not blank and correspond to rows in different sets of $\overline{P}_k$. Put one of these rows into a new set in $\overline{P}_{k+1}$ together with all rows in the original set of $\overline{P}_k$ which go into the same set in $\overline{P}_k$ for the given row and input combination. Leave the other sets in $\overline{P}_k$ fixed in $\overline{P}_{k+1}$. If this is not possible, the process terminates. Now apply the truth-table analog of the process described following Theorem 1.

Except for the stipulations concerning blank entries, Rule II is merely a reworded form of Rule I.

4.3 *Second Reduction Process*

Rule II, given above, seems rather complicated. Although this complication is more apparent than real, one still wishes to find a reduction rule that has both the effect and the appearance of simplicity. Presumably, one must pay for this in one way or another — the surprising thing is that one is not required to pay too heavily. In point of fact the reduction rule given below, when applied to asynchronous circuits, is somewhat more powerful than Huffman's rule for merging.

We ask, then, what are the simplest circumstances in which a state may be eliminated by using Rule II? Is it possible to consider only pairs of states instead of considering larger sets of states? To answer these questions, consider any pair of rows that are in the same set of $\overline{P}_1$. That is, no corresponding entries in the right-hand portion of the rows may be contradictory. Now if in addition no corresponding entries in the left-hand portion of the rows are contradictory, then the two states have the same output combination for a given input combination and the next state is the same, or may be made to be the same by filling in a "don't care" entry, for any given input combination.* Therefore, the two states are equivalent. This means that we may eliminate one and keep the other. If we eliminate state A in favor of state B, then any appearance in the table of "A" must be changed to read "B".

We restate the above more formally as Rule III. This process is called *merging*, after Huffman, since we will see that it is a general form of his merging process.

* Or the present state is also the next state in both cases.

TABLE VII

Input combination....	0	1	0	1
Present State	Next State		Output Combination	
1	2	3	0	0
2	7	5	0	0
3	6	7	0	0
5	10	15	0	0
6	15	13	0	0
7	15	15	0	0
10	1	1	0	1
13	1	1	1	0
15	1	1	0	0

TABLE VIII

Input combination....	0	1	0	1
Present State	Next State		Output Combination	
1	1	2	0	0
2	3	1	0	1
3	2	4	0	1
4	4	3	0	0

Rule III: To merge state A with stage B, change all appearances of "A" in the table to read "B" and copy the entries of row A into row B. Eliminate row A.

Rule III may be used whenever, after the "A's" have been changed to "B's", to each entry in row A corresponds either the same entry in row B or a blank in row B.

As an example, consider Table VI. We see that states 8, 9, 11, 12, and 14 may be merged with state 15. Then state 4 may be merged with state 7. The resulting table, Table VII, now corresponds to the state diagram of Fig. 8(b).*

Note that Rule III may not always give complete reduction. An example is Table VIII, to which Rule III may not be applied. However, Rule II leads to the conclusion that states 2 and 3 are equivalent, as are states 1 and 4.

* The reader is urged to write out the intermediate truth-tables derived by carrying out the mergers step by step.

4.4 *Blank Entries; Uniqueness of Reduction*

The provision for blank entries in Rules II and III corresponds to "don't care" situations, which usually result from restrictions on the input sequences. The result of merging rows in different orders is not always unique. The reason for this is simple — when truth-tables have blanks, they may usually be filled in in different ways so as to result in circuits which are not equivalent. Since merging usually results in filling in blanks, different orders of merging may result in blanks being filled in differently. This situation is not in contradiction with Moore's theory; there it is assumed that the state diagram is completely specified at the outset.

As an example, consider Table IX(a). Here, there are four output leads. The designation of which lead is to be energized is given in the right portion of the table — a dash indicates that no lead is to be energized. Clearly, we may merge 8 and 9 with 7; 4 and 5 with 3; and 6 with 1. The result is shown in Table IX(b). A final merging of 7 with 3 and 2 with 1 leaves the table of Table IX(c). On the other hand, if we merge 2 with 1; 4 with 3; 6 with 5; and 8 and 9 with 7 we get Table IX(d) instead, and Rule II tells us that this is a completely reduced circuit.

We have, incidentally, demonstrated that reduction is not necessarily unique even if only Rule II is used, since Rule III is a restricted form of Rule II. Therefore, Theorem 1 is not necessarily valid unless the initial truth-table has no blank entries. Again, this does not mean that the theorem as originally stated is false — it means only that we are applying it under conditions which are somewhat more general than those obtaining in Moore's theory. Actually, we are really considering sets of circuits in synthesis. Each circuit is described only partly by the initial truth table and the truth table is, in a mathematical sense, a kind of domain of definition for the circuits in the set considered. Within this domain all circuits in the set are identical while outside this domain the circuits are specified only by "don't cares" and therefore may differ. Moore's theory applies to each individual circuit. We, on the other hand, are applying it to sets of circuits and must therefore be prepared to find some differences in detail.

4.5 *Summary of Method*

In general we start synthesis by writing either (1) a state diagram or (2) a truth-table, as outlined in Section 4.1. Following this step, we use Rule I supplemented by stipulations concerning "don't cares" or Rule III followed by Rule II to achieve reduction. In case (1), the state dia-

TABLE IX(a)

Input Combination....	00	01	11	10	00	01	11	10
Present State	Next State				Output Combination			
1	1	6		2	—	AA		AB
2			3	2			—	AB
3		4	3	5		DB	—	DA
4	1	4			—	DB		
5	1			5	—			DA
6		6	7			AA	—	
7		9	7	8		DB	—	DA
8	1			8	—			DA
9	1	9			—	DB		

TABLE IX(b)

Input Combination....	00	01	11	10	00	01	11	10
Present State	Next State				Output Combination			
1	1	1	7	2	—	AA	—	AB
2			3	2			—	AB
3	1	3	3	3	—	DB	—	DA
7	1	7	7	7	—	DB	—	DA

TABLE IX(c)

Input Combination....	00	01	11	10	00	01	11	10
Present State	Next State				Output Combination			
1	1	1	3	1	—	AA	—	AB
3	1	3	3	3	—	DB	—	DA

TABLE IX(d)

Input Combination....	00	01	11	10	00	01	11	10
Present State	Next State				Output Combination			
1	1	5	3	1	—	AA	—	AB
3	1	3	3	5	—	DB	—	DA
5	1	5	7	5	—	AA	—	DA
7	1	7	7	7	—	DB	—	DA

gram must now be translated into a truth-table. At this point in the process binary coding must be assigned to the states in order to complete synthesis with two-valued storage elements. Two remarks are in order here:

(1) The simplicity of the final circuit will be affected by the exact coding assigned as well as by the truth-table finally chosen, if reduction is not unique.

(2) Using a minimum number of storage elements is not always wise. In practical situations, the choice of components dictates one's criterion for minimality, and this criterion must ultimately be based on considerations of economy and reliability. For instance, the present writer has seen an example in which it was much more economical to use seven, rather than three, storage elements in order to achieve eight states. In fact one has doubts that complete reduction, itself, is always desirable.

5. THE METHOD APPLIED TO ASYNCHRONOUS CIRCUITS

5.1 *Introductory Remarks*

In this section we carry out the transition from synchronous to asynchronous circuitry. A more exhaustive treatment of the subject of asynchronous circuitry is contained in Huffman.[2]

We agree (1) that no clock will be used and (2) that "1" in switching algebra will correspond to a high voltage or current, an energized relay coil, or operated relay contacts. We must now pay careful attention to circuit conditions at *every* instant of time. One very real difficulty arises since time delays inherent in the "combinational" circuit elements may frequently be of the same order of magnitude as the time required to change the state of a storage element. This may mean that spurious inputs to flip-flops may be produced by changes of input combination solely because of nonuniform delays in portions of the "combinational" circuitry. These difficulties will not be considered further since little can be said about them over and above noting their existence. Another problem — that of *race conditions* (a definition of this term will be given below) — can be resolved by logical methods; we shall treat this problem in moderate detail.

5.2 *Interpretation of the Model*

For the purpose of illustrating the pertinent facts and methods which relate to asynchronous circuits, we use relay circuitry as being typical of asynchronous circuitry. Fig. 9 illustrates our conventions and notations.

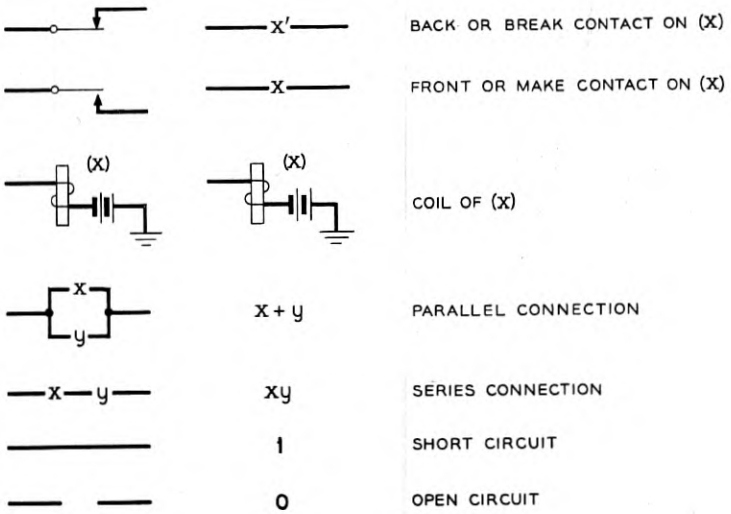


Fig. 9

Our interpretation of the abstract model for sequential circuits given in Section 2 must be changed somewhat. To be concrete, consider the circuit of Fig. 10. We think of this circuit as having two types of relays — to *primary* relays correspond input variables and to *secondary* relays correspond excitation and state variables. The general situation is shown in Fig. 11. The primary relays are controlled directly by the inputs; we shall use "x_i" to denote both the i^{th} input and the contacts on relay (x_i). The secondary relays are controlled by contacts on any or all relays; they furnish the storage in the circuit. Considering relay (q_i), we shall say that $\bar{q}_i = 1$ whenever the coil of (q_i) is energized and that $q_i = 1$

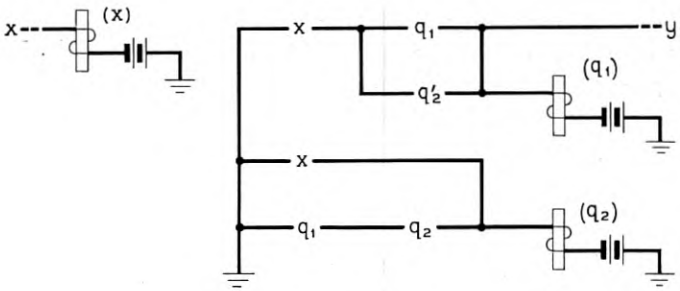


Fig. 10

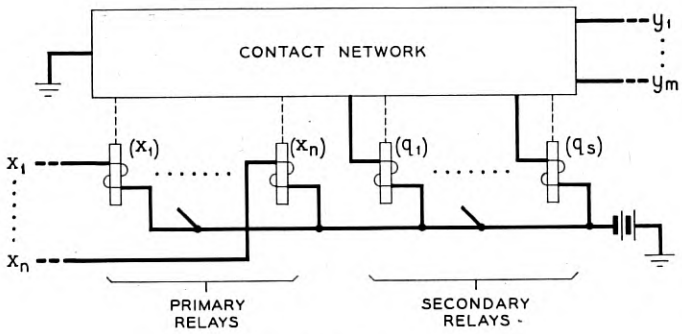


Fig. 11

whenever (q_i) is fully operated. Note carefully the distinction between these two statements!

5.3 *Race Conditions; Coding of States*

The meaning of "present state" is clear enough — it is determined by which secondary relays are *operated*. We shall say that the "next" state is determined by which secondary relays are *energized*. However, the "next" state may never be realized as a *present* state! We shall now reconsider the circuit of Fig. 10. On the basis of our previous agreement, we may draw a truth-table and state diagram. The truth-table is that given by Table X, and the state diagram is shown in Fig. 12.

In order to study the action of asynchronous circuits, it is often convenient to make use of *sequence diagrams*.[6] These are essentially pictures of what happens in a circuit as a function of time;* a line opposite a relay or lead designation represents an operated relay or a grounded lead. For instance, assume that both relays in Fig. 10 are released, a ground is applied and then released later on, and moreover that (q_1) is faster in operating than (q_2). The corresponding sequence diagram is shown in Fig. 13(a). Clearly, in this case, the circuit does almost what one would expect from consideration of the state diagram, except that the circuit goes from state 00 to state 11 by way of state 10! The situation is quite different if (q_2) is faster in operating than (q_1), as shown by Fig. 13(b). In this case, although state 11 is the "next state," it is never reached, since (q_2) in operating breaks the operating path of (q_1). A situation such as this is called a race condition. Whether it is harmful or not depends on the circuit requirements.

* The time scale is usually distorted, sequence of events being more important than their duration.

TABLE X

q_1	q_2	x	$\bar{q}_1$	$\bar{q}_2$	y
0	0	0	0	0	0
0	0	1	1	1	1
0	1	0	0	0	0
0	1	1	0	1	0
1	0	0	0	0	0
1	0	1	1	1	1
1	1	0	0	1	0
1	1	1	1	1	1

We say that a *race condition* exists in a circuit for input combination X_i and present state Q_j if the next state Q_k is such that the binary forms of j and k disagree by more than one binary digit. For, if they do, more than one relay is attempting to change its state of operation, and differences in operate and/or release times may lead to differences in circuit behavior.

In order to avoid races, it is necessary and sufficient that any distinct states directly connected by a transition disagree in exactly one binary digit. We can always avoid races if we add enough extra states. On the other hand, if a race condition is not harmful, removing the race condition generally decreases circuit operating speed.

One further remark must be made: it is often very helpful to assume that only one input variable may change its value at any given instant and to arrange connecting circuits in a system so that this condition is satisfied. To appreciate why this might be the case, consider a system containing two interconnecting circuits. These circuits may be viewed together as a single, larger circuit. If the above condition on the interconnecting leads is not fulfilled, then race conditions may be present in the over-all circuit even though they are not present in either circuit considered by itself.

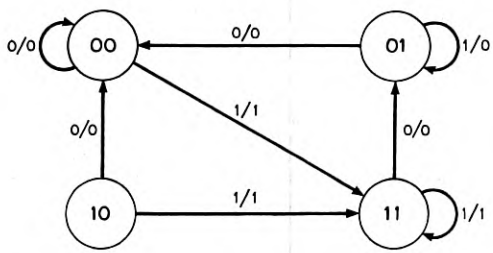


Fig. 12

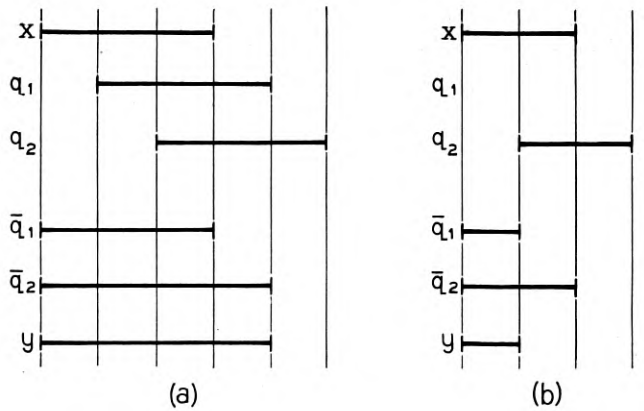


Fig. 13

The usual state of affairs in an asynchronous circuit is this: upon a change of input combination, if the "next" state of the circuit is different from its present state, the states of the individual storage elements will change until a final state of the circuit is reached in which no further change of state is possible. Two remarks are in order here. First, we have already seen that in the presence of race conditions the final state, if any, may depend on operate and release times as well as on the truth-table for the circuit. Second, there may be no final state — this is the case for certain pulse-generating circuits.* Usually however, if the new input combination is maintained for a sufficiently long interval, a final state will be reached. Since in most cases of practical interest the time required to reach the final state is much less than the interval during which any given input combination is held, design effort is fixed on the final states, rather than any possible intermediate states.

For the above reasons, the formal part of synthesis — that part of synthesis which ends with writing out circuit equations — is both different and more difficult in the case of asynchronous circuitry. Although it is true that we need not consider the possibility of race conditions until that point in synthesis in which we assign binary coding to the states, it is not true that the same truth table may always be used for both a synchronous and an asynchronous realization of a given circuit. (That this is possible for the circuit of Table IX is only accidental). The reason for this is tied in very closely with the fact that we speak of presence or absence of pulses in synchronous circuits but of

* See Reference 6, Chapter 18, for examples.

quasi-steady-state conditions on leads in asynchronous circuits. A pulse on lead x_2, for instance, might be represented by X_2 in a synchronous circuit but as X_0 followed by X_2 followed by X_0 in an asynchronous circuit.

5.4 *Huffman's Method*

The purpose of this section is not to outline Huffman's method of synthesis[2] but, rather, to support our claim made above that Rule III represents a slight generalization of Huffman's merging process. We shall assume familiarity with the contents of Reference 2.

The justification for this claim is immediate, if not already self-evident to the reader. Namely, suppose that an initial flow table is written down. By going immediately to the associated truth table, Huffman's rule for merging becomes the same as Rule III, except that Rule III allows somewhat more latitude for merging in that it is permissible to change the symbols corresponding to certain next states. In Huffman's method, it would be necessary to resort to equivalence arguments in such instances. We are considering here that the use of equivalence arguments is separate from the purely mechanical merging process, although there is evidence in Reference 2 that Huffman considers the use of such arguments to be a part of merging. Our point is that such arguments may be avoided in many cases if we work directly with the truth table and Rule III.

5.5 *Summary of Method*

We have now disposed of the basic principles of our method as applied to asynchronous circuits. The synthesis steps are:

(1) Write a truth-table which satisfies the circuit requirements.

(2) Use Rules II and III in reverse order, as applicable, to obtain a reduced truth-table.

(3) Code the states in a binary code. If possible, assign the code so that no harmful race conditions are present. Otherwise, add states in such a way as to make eliminate harmful races.[2, 6]

(4) Write the circuit equations.

(5) Synthesize the combinational networks.

As our final example, we consider the following problem, taken from Reference 6 (Problem 8–9):

A rotating shaft carries a single grounded brush which makes contact with three stationary commutator segments arranged symmetrically around the shaft. A relay circuit is required which will indicate the direction of shaft rotation by lighting a lamp when the shaft is rotating in the

clockwise direction. The shaft may reverse its direction at any time. Assume that the shaft is driven so that a brush contact closure is 0.25 second and that the open time between the brush leaving one segment and reaching another is 0.25 second. When the shaft changes direction, the output indication must change as quickly as possible, at most within 2 seconds.

Let the brush be grounded and the three segments be labelled "x_1," "x_2" and "x_3" respectively. For the output indication, let $y = 1$ when the shaft is rotating in the clockwise sense. Now, in order to write the initial truth-table, we may first consider what the circuit must do to keep track of the brush while it is rotating in only one direction. This situation is clearly taken care of by Table XI(a). All that remains is to enlarge the table to enable (say) the circuit to go from states 1–6 to states 7–12 when the direction is changed from clockwise to counterclockwise. A usable strategy is this: as one segment, say x_1, is passed the circuit expects x_2 to come up next. If x_3 comes up before x_2, we can cause the circuit to go to the counterclockwise state in which x_3 has occurred and x_2 is expected next. This has been done in Table XI(b).

With regard to the output note that it is sufficient to assign $y = 0$ to states $7 - 12$ and $y = 1$ to states $1 - 6$, regardless of input combination.

The possibilities for merging, (using Rule III), are obvious: merge 2 with 3, 4 with 5, 6 with 1, 8 with 9, 10 with 11, and 12 with 7. The result is Table XI(c). Now use Rule II to determine whether reduction is complete. Actually, literal use of Rule II is a waste of time, for we may use this argument:

$$\overline{P}_1 : (1, 3, 5)\ (7, 9, 11)$$

By examining input combination 100, we split off both 5 and 9 from the sets above, arriving at:

$$\overline{P}_2 : (1, 3)\ (5)\ (7, 11)\ (9)$$

By examining input combination 010, we see that 1 and 3 (7 and 11) are distinguishable. Hence, the circuit is completely reduced.

We now have to code the states. To assist in this process, we draw the state diagram shown in Fig. 14(a). Since there are two triangles in the diagram, we cannot assign codes to avoid races, and therefore extra states must be added. One way to do this is to insert new states between 5 and 1 and between 11 and 7 in such a way that the circuit will treat the new states as transient states. This has been done and coding has been assigned in Fig. 14(b). The corresponding truth-table is shown as Table XII.

TABLE XI(a)

Input Combination....	000	100	010	001	(All)
Present State	Next State				Output Combination
1	1	2			1
2	3	2			1
3	3		4		1
4	5		4		1
5	5			6	1
6	1			6	1
7	7			8	0
8	9			8	0
9	9		10		0
10	11		10		0
11	11	12			0
12	7	12			0

TABLE XI(b)

Input Combination....	000	100	010	001	(All)
Present State	Next State				Output Combination
1	1	2	10		1
2	3	2			1
3	3		4	8	1
4	5		4		1
5	5	12		6	1
6	1			6	1
7	7		4	8	0
8	9			8	0
9	9	2	10		0
10	11		10		0
11	11	12		6	0
12	7	12			0

TABLE XI(c)

Input Combination....	000	100	010	001	(All)
Present State	Next State				Output Combination
1	1	3	11	1	1
3	3	3	5	9	1
5	5	7	5	1	1
7	7	7	5	9	0
9	9	3	11	9	0
11	11	7	11	1	0

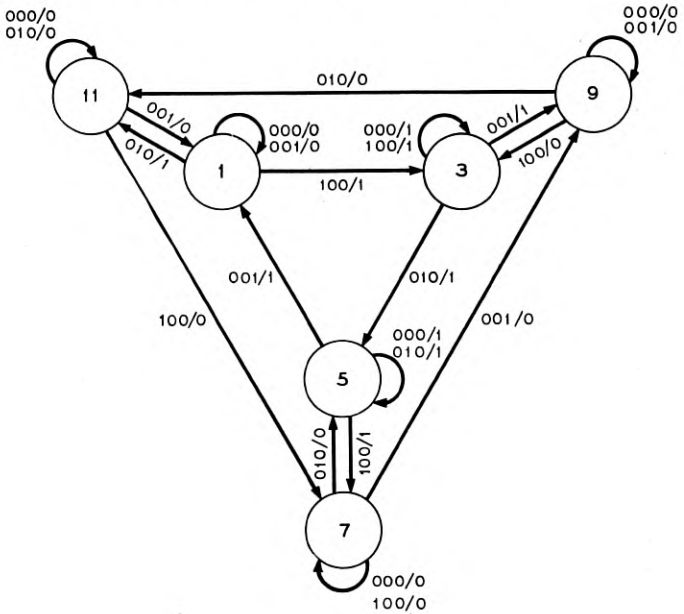


Fig. 14(a)

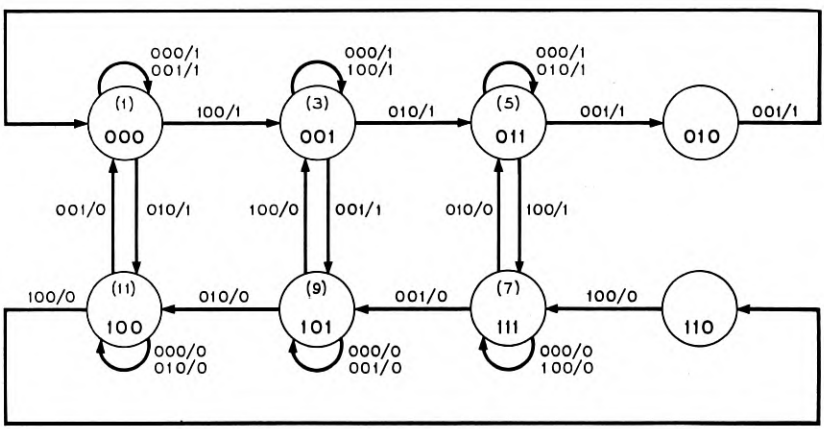


Fig. 14(b)

The circuit equations may be written as:

$$\bar{q}_1 = (q_2'q_3'x_2 + q_2q_3x_1 + q_2'q_3x_3 + q_1)\cdot(q_2'q_3'x_3 + q_2'q_3x_1 + q_2q_3x_2)'$$

$$\bar{q}_2 = (q_1'q_3x_2 + q_1q_3'x_1 + q_2)\cdot(q_3'x_3 + q_1'x_3)'$$

$$\bar{q}_3 = (q_1'q_2'x_1 + q_1q_2x_1 + q_3)\cdot(q_1'q_2x_3 + q_1q_2'x_2)'$$

In this particular case, if shunt-down operation[6] is not objectionable, it is even possible to dispense with primary relays.* A circuit that satisfies the stated conditions is given in Fig. 15. (The author does not guarantee that the circuit is minimal!)

6. DISCUSSION

Like any "systematic" method for synthesizing certain classes of switching circuits, our method leaves much to be desired. First, the problem of synthesizing really large circuits has not been touched — one wonders whether it is really possible to do this with any method that

TABLE XII

Input Combination....	000	100	010	001	(All)
Present State	Next State				Output Combination
000	000	001	100	000	1
001	001	001	011	101	1
011	011	111	011	010	1
010	d	d	d	000	1
100	100	110	100	000	0
101	101	001	100	101	0
111	111	111	011	101	0
110	d	111	d	d	0

relies on the use of a truth-table without making use of automatic design aids inasmuch as large truth-tables become unmanageable. Second, the first step of the process, as described in Section 4, has in no sense been eliminated — this is probably the step that asks the most of the designer's ingenuity and skill. Third, the process of coding the states may have a great effect on the final cost of the circuit — despite this, there are at present no rules for carrying out the coding in an optimal manner.

To compare our method with that of Huffman,[2] several pertinent comments may be made. First, our method applies equally well to synchron-

* This was pointed out to the writer by A. H. Budlong.

ous and asynchronous circuit synthesis whereas Huffman's method was formulated specifically for asynchronous circuit synthesis. We hasten to add, however, that the basic concepts of Huffman's paper are valid in both cases. Such changes in detail as are required to adapt his method to synchronous circuit synthesis would almost certainly result in a method identical with the method of this paper. Second, for asynchronous circuits, the initial truth table we write down is different only in appearance from the initial flow table that we might have written — neither method offers any advantage in this respect. Third, the ease of using Huffman's merging rule as opposed to the use of Rule III must be weighed against the necessity of translating the final flow table into a truth table in order to develop circuit equations. Finally, the present

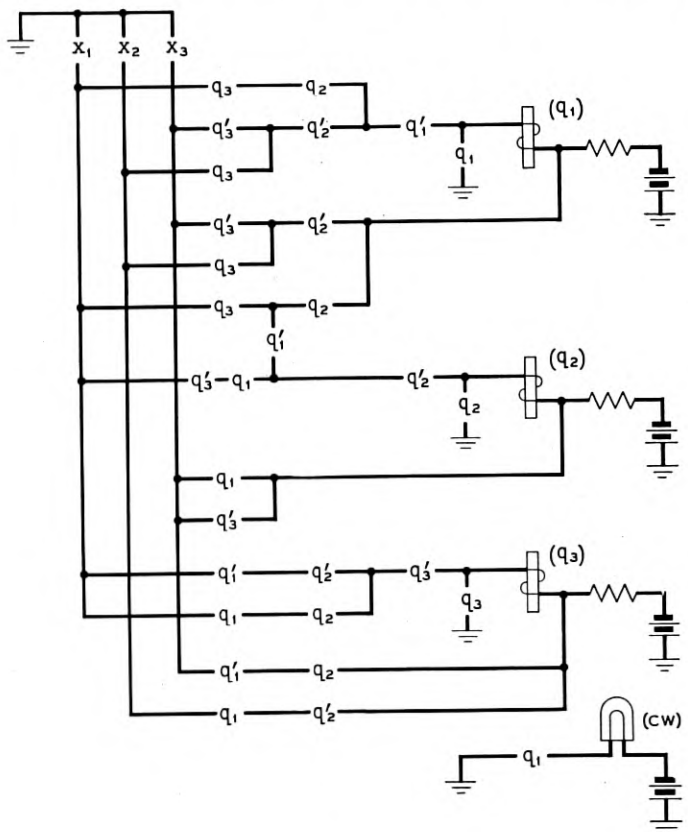


Fig. 15

method is more often successful (in principle, at least) in achieving complete reduction without the use of auxiliary equivalence arguments. Nevertheless, it is always advisable to use Rule II in order to test for complete reduction.

Finally, it should be pointed out that there are many cases where other, more intuitive, methods are more useful. Such methods for asynchronous circuit design are given in Reference 6.

In fact, the place of formal methods, such as that outlined in this paper, in the every day practice of synthesis is much smaller than might appear at first glance. It is probably fair to say that the theory furnishes, at present at least, only generalized methods of attack on synthesis together with a small handful of particularized tools for design. It is the author's belief that these methods are genuinely useful insofar as they aid in understanding the nature of sequential circuits and furnish a unified way of thinking about circuits during their design. It would be a mistake, however, to believe that they provide detailed design methods in the same sense in which such methods are available for electrical network synthesis. The engineer must make a judicious selection of his design tools and, most likely, must invent methods and diagrammatic devices which fit the particular problem at hand.

A few words should be said about the comparative originality of the author's treatment of this subject. The model proposed in Section 2 was suggested to the writer by the content of E. F. Moore, Reference 7, and, in the case of synchronous circuits, is almost identical with the discrete transducer of information theory.[8] Independently, S. H. Washburn proposed essentially the same model in an unpublished memorandum.

Our interpretation of the model for asynchronous circuits and consequences of that interpretation with relation to race conditions were independently treated by Huffman.[2] Our use of Rule III in the method owes much to Huffman's work.

7. ACKNOWLEDGEMENTS

The author gratefully acknowledges the constructive criticisms of many of his colleagues at Bell Telephone Laboratories, Inc. during the course of the work reported in this paper. He owes particular thanks to W. J. Cadden, E. F. Moore, and S. H. Washburn of Bell Telephone Laboratories, Inc. and D. A. Huffman of the Massachusetts Institute of Technology for their participation in many discussions, philosophical and otherwise, which have greatly aided the writer in clarifying his thoughts on this subject and which have resulted, it is hoped, in a far better presentation than would otherwise have been possible.

8. SELECTED BIBLIOGRAPHY

1. Karnaugh, M., The Map Method for Synthesis of Combinational Logic Circuits, Comm. and Electronics, No. 9, 1953.
2. Huffman, D. A., The Synthesis of Sequential Switching Circuits, J. Franklin Inst., **257**, pp. 161–190, 275–303, March and April, 1954.
3. Moore, E. F., Gedanken - Experiments on Sequential Machines, to be published in Automata Studies, Princeton University Press.
4. Leiner, A. L., Notz, W. A., Smith, J. L., and Weinberger, A., System Design of the SEAC and DYSEAC, Trans. I.R.E. Professional Group on Electronic Computers, Vol. EC-3, No. 2, June, 1954, pp. 8–22.
5. Felker, J. H., Typical Block Diagrams for a Transistor Digital Computer, Trans. A.I.E.E., **17-I**, pp. 175–182, 1952.
6. Keister, W., Ritchie, A. E., and Washburn, S. H., The Design of Switching Circuits, D. Van Nostrand, 1951.
7. Moore, E. F., A Simplified Universal Turing Machine, Proc. Assoc. for Computing Machinery, (Toronto meeting), 1952.
8. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., **27**, July and Oct., pp. 279–423, 623–656, 1948.
9. Nelson, E. C., An Algebraic Theory for Use in Digital Computer Design, Trans. I.R.E. Professional Group on Electronic Computers, Vol. EC-3, No. 3, pp. 12–21, Sept., 1954.
10. Burks, A. W. and Wright, J. B., Theory of Logical Nets, Proc. I.R.E., **41**, pp. 1357–1365, Oct., 1953.
11. Murray, F. J., Mechanisms and Robots, J. Assoc. for Computing Machinery, **2**, pp. 61–82, Apr., 1955.

Arcing of Electrical Contacts in Telephone Switching Circuits

Part V — Mechanisms of the Short Arc and Erosion of Contacts

By M. M. ATALLA

(Manuscript received March 4, 1955)

This is a presentation of a study of the mechanisms of the short arc between closely spaced contacts and its erosion effects. The study is based on optical measurements of the erosion obtained on contacts after repeated arcing on closure or opening. Most experiments reported here are essentially of the probing type designed to test specific postulates and assumptions. For short arcs initiated at 250 volts, clean palladium, iron and nickel contacts have shown a reversal, with arc duration, in the direction of net transfer. Net anode losses were obtained with short duration arcs and net cathode losses with longer duration arcs. This reversal, however, did not occur with silver, gold or copper. For longer arcs initiated as air breakdowns from 500 volts, all the above metals indicated a net loss from the cathode. For arcs initiated at 250 volts between fully activated contacts, shallow cathode losses were generally observed with little or no buildups on the anode.

The first section of this paper is a summary of the experimental work done and the results obtained. In the second section, the data are analyzed and a tentative working model is proposed for the short arc and its erosion effects.

INTRODUCTION

The problem of contact erosion due to arcing has been the subject of a large number of investigations. The literature includes a considerable accumulation of data on the erosion characteristics of many contact materials. Due, however, to the vast variations in testing conditions adopted, there are considerable disagreements and discrepancies among results from different investigations. Inconsistencies even within one investigation are not uncommon.

In general, the erosion behavior of contacts depends, to varying degrees, on the following main parameters: the physical properties of the contact material, surface conditions, contact geometry and separation, arc duration, arc current and the surrounding atmosphere. In our study, most of these parameters were considered separately, whenever possible, with the primary objective of clarifying the mechanisms involved. Most experiments reported here are, in effect, of the probing type designed to test specific postulates and assumptions. The first section of this paper is a summary of the experimental work done and the results obtained. In the second section, the data are analyzed and a tentative working model is proposed for the short arc and its erosion effects. Because of the rather extreme complexity of the phenomena and the lack of basic data on the conduction properties of metal vapors, this model is at best a simplified one and is probably incomplete in some respects.

NOTATION

F	Field strength
I	Total current
M	Mass of an atom
N	Gas concentration
T	Temperature
T_0	Ambient temperature
T_b	Boiling temperature
ΔT_b	$T_b - T_0$
V	Voltage
V_i	Minimum ionization potential of a metal atom
V_c	Voltage drop in cathode fall
Q_i	Ionization cross-section
Q_e	Excitation cross-section
W	Atomic weight
a	Radius of arc spot
d	Contact separation
e	Electron charge
j	Total current density at cathode
j_-	Electron current density at cathode
j_+	Ion current density at cathode
k	Boltzmann's constant
p	Gas pressure
t	Time

v Arc voltage
z Distance from anode surface
θ Angular location of a point between contacts with respect to the center of the anode arc spot
λ Thermal conductivity

MEASUREMENTS

Contacts tested were made of crossed cylinders 0.050 or 0.125 cm diameter. Their surfaces were prepared by fine polishing followed by washing with alcohol and distilled water. They were mounted on a sound-head* and operated at 60 cycles/sec. Care was taken to avoid additional arcing by eliminating chatter of the contacts on closure. In the low-voltage experiments this was done satisfactorily by mechanical adjustment of the contact separation and by choosing a proper charging resistor to avoid excessive recharging during chatter opening. For the high-voltage experiments, however, it was necessary to adopt a mechanical switching scheme which prevents recharging until the contacts were fully open. The behavior of the contacts was regularly observed on an oscilloscope.

In most experiments, the circuit consisted of a coaxial cable with a characteristic impedance of 75 ohms and a period of 3.5×10^{-9} sec per foot. The cable was charged, during contact opening, to any desired voltage through a proper resistor. All lines were matched with a 75-ohm resistor at the contact end, thus allowing only one discharge per closure without spurious reflections. In all cases, therefore, constant arc current pulses were obtained. Their amplitudes were controlled by varying the charging voltage. Their periods were controlled by varying the cable length.† The use of this constant current pulse scheme makes the interpretation of the data far simpler and more direct. In each experiment, the contacts were subjected to 20,000 to one million operations, depending on the arc energy.

Since the main interest was in the contribution of each electrode to the maintenance of the arc, conventional weight measurements would have been of little significance. An optical measurement scheme was therefore adopted. It allowed a discrimination between losses and gains

* To avoid contact activation by organic vapors, the construction of these units was free of organic materials except for varnish insulation on the winding. From observations of the eroded surfaces and oscilloscope traces, as discussed in a following section on activated contacts, these contacts were free of activation.

† The velocity of closure of the contacts is estimated at about 5 cms/sec. During the longest duration arc. of 10^{-6} sec used in these experiments, the contact motion is only 500 A compared to an initial separation of about 25,000 A.

as indicated by craters and build-ups. It also permitted examination of the geometries involved. This was particularly important in cases where each electrode indicated both loss and gain and the detection of matched patterns for one pair of contacts was quite significant in determining the most probable directions of transfer. A microscope was used, with magnifications as high as 740, and a quantitative measure of metal loss was made. The losses measured were of the order of 10^{-7} cc and the accuracy is better than ± 50 per cent.

1. EXPERIMENTS WITH VARYING ARC DURATION ON CLOSURE

Test contacts were operated in laboratory atmosphere, using matched cables in lengths ranging between 5 feet and 260 feet. In all cases they were charged to a fixed voltage of 250 volts and allowed to discharge on closure. The arc durations for these cables varied between 17.5×10^{-9} and 910×10^{-9} sec. For control of the current, separately matched multiple cables were used in parallel. In most cases, at least three runs were made for each cable length. The volume of metal loss, appearing as a

TABLE I — EROSION OF PALLADIUM CONTACTS ON CLOSURE BY SHORT ARCS INITIATED AT 250 VOLTS, 3.2 AMPS

	Arc Duration 10^{-9} sec	No. of operations 10^3	Erosion: (loss, gain) 10^{-7} cc		Loss/total loss		Rate of loss 10^{-14} cc/erg	
			Anode	Cathode	Anode	Cathode	Anode	Cathode
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	17.5	540	0.72, n*	n, build-up†	1.0	0.0	1.7	
(2)	35	430	2.01, n	0.08, buildup	0.96	0.04	2.9	
(3)	52.5	320	1.9, n	n, buildup	1.0	0.0	2.5	
(4)	70	430	2.02, buildup	0.85, buildup	0.7	0.3	1.8	0.75
(5)	87.5	108	0.21, buildup	0.21, buildup	0.5	0.5	0.5	0.5
(6)	105	108	0.43, buildup	0.57, buildup	0.43	0.57	0.83	1.1
(7)	140	108	n, buildup	1.79, n	0.0	1.0		2.7
(8)	280	108	0.6, buildup	2.67, n	0.18	0.82	0.69	3.2
(9)	385	18	n, buildup	0.54, n	0.0	1.0		2.7
(10)	912	36	n, buildup	loss, n	0.0	1.0		not measured

* "n" denotes no loss or no gain or those that are too small to measure.

† Volume of buildups were not measured. In general, they match the geometry of a hole on the opposite electrode. This includes observations on lines 4 to 6 where each electrode showed both gain and loss.

TABLE II — EROSION OF PALLADIUM CONTACTS ON CLOSURE BY SHORT ARCS INITIATED AT 250 VOLTS, 1.6 AMPS

	Arc Duration 10^{-9} sec	No. of operations 10^3	Erosion: (loss, gain) 10^{-7} cc		Loss/total loss		Rate of loss 10^{-14} cc/erg	
			Anode	Cathode	Anode	Cathode	Anode	Cathode
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	35	650	1.8, *n**	*n*, buildup*	1.0	0.0	3.5	
(2)	105	108	0.97, *n*	*n*, buildup	1.0	0.0	3.8	
(3)	140	360	1.84, buildup	0.51, buildup	0.78	0.22	1.6	0.45
(4)	280	540	1.86, buildup	4.17, buildup	0.31	0.69	0.55	1.23
(5)	385	51	*n*, buildup	1.6, *n*	0.0	1.0		3.7

* See footnotes below Table I.

depression or crater on an electrode surface, was measured and the geometry sketched.

Tables I and II show the results obtained for palladium contacts with currents of 3.2 and 1.6 amperes. In both cases, a characteristic change in the direction of transfer is observed. In Table I, for instance, for arc durations 52.5×10^{-9} sec and less, lines 1 to 3, the losses were predominantly from the anode. The geometries observed generally consisted of a rather irregular yet definite buildup on the cathode and a corresponding hole on the anode. The geometrical resemblance between the anode hole and the cathode buildup was in many cases rather striking. This and the absence of buildups surrounding the cathode hole, strongly suggest that the arc was mainly maintained through vapor from the anode. This, however, does not necessarily exclude the possibility of some evaporation from the cathode. These arcs are called *anode arcs*. For arcs of longer duration, 70×10^{-9} to 105×10^{-9} sec in the case of Table I, lines 4 to 6, the observed erosion was distinctly different. It was characterized by the appearance of both a hole and a buildup on each electrode. The geometrical resemblance between a hole on one electrode and a buildup on the opposite electrode is a strong indication that both electrodes were contributing more or less equally to the maintenance of the arc. This stage of the arc is called the *mixed arc* stage. Further increase in the arc duration, above 140×10^{-9} sec in the case of Table I, lines 7 to 10, the erosion character changed once more. Holes were obtained on the cathodes and matching buildups on the anode. These arcs are called *cathode arcs*. They probably still involve some evaporation from the anode. Table II shows similar data for palladium contacts at 1.6 amp where a reversal in transfer is also indicated. Fig. 1 is a plot of columns

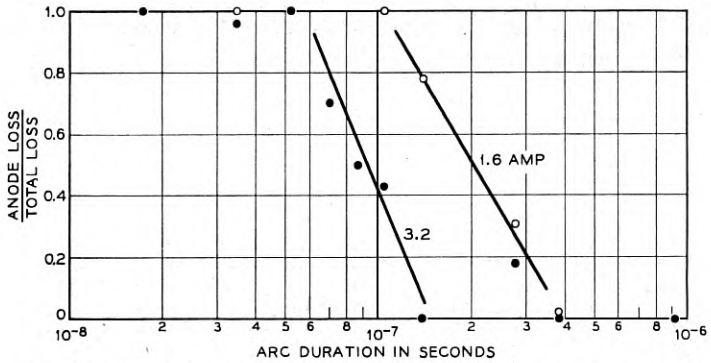


Fig. 1 — Reversal of transfer between Pd contacts on closure. Short arcs initiated at 250 volts.

5 and 6 from Tables I and II. It is shown that the reversal of transfer occurred later for the smaller current. It is believed, however, that due to the difficulty of premature closure, discussed below, not much emphasis should be given to the exact relative locations of the transition points for the different currents.

It should be pointed out that, while this observation of reversal in the arc transfer is unmistakable, its exact location is rather difficult to obtain with great consistency. This is because of the extreme proximity of the contacts when an arc strikes and the tendency of occurrence of premature closures. These are caused by the formation by the arc[1,2] of mounds which decrease the separation and the closure time. This difficulty was particularly noticeable with the longer cables. However, by proper adjustments such as the use of various retardation schemes for the moving contact, it was possible to satisfactorily decrease the frequency of premature closures. It is evident that the effect of premature closures is to allow only short duration arcs irrespective of the desired duration as set by the cable length. In extreme cases, where premature closures predominate, the phenomenon of reversal of transfer can be completely missed. The use of higher voltage presents additional means for decreasing the frequency of premature closures by initiating the arcs at wider separations. For experiments in air, however, one is limited by the minimum sparking potential of air.

Columns 7 and 8, Tables I and II, give the measured rate of metal loss from each electrode. This is defined as the volume of metal loss per unit arc energy. For instance at 3.2 amp, Table I, the rate of loss for

[1] L. H. Germer and F. E. Haworth, J. Appl. Phys., **20,** p. 1085, 1949.
[2] M. M. Atalla, B.S.T.J., **32,** p. 1503, 1953.

TABLE III — EROSION OF SILVER AND GOLD CONTACTS UNDER CONDITIONS BEYOND THE REVERSAL POINT OF PALLADIUM, INITIAL VOLTAGE 250

Current Amp	Arc Duration 10^{-9} sec	No. of operations 10^3	Erosion: (loss, gain) 10^{-7} cc		Loss/total loss		Rate of loss 10^{-14} cc/erg	
			Anode	Cathode	Anode	Cathode	Anode	Cathode
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ag, 1.6	385	430	4.3, *n**	*n*, buildup*	1.0	0.0	1.15	
Ag, 3.2	385	270	5.8, *n*	*n*, buildup	1.0	0.0	1.25	
Ag, 6.4	140	79	3.5, *n*	*n*, buildup	1.0	0.0	3.5	
Au, 3.2	140	90	1.7, *n*	*n*, buildup	1.0	0.0	3.0	

* See footnotes below Table I.

both the anode and cathode arc stages is between 1.7×10^{-14} and 3.0×10^{-14} cc/erg. For the mixed arc stage, lines 4 to 6 of Table I and lines 3 and 4 of Table II, the rate of loss is consistently lower. This is an indication of considerable exchange of metal between the two electrodes during this arc stage.

The aforementioned erosion behavior of palladium, as characterized by the reversal of transfer with arc duration, was also obtained for iron and nickel contacts. These tests were performed at 250 volts and 3.2 amp for two cable lengths of 10 and 110 feet. For silver, gold and copper, on the other hand, no reversal in transfer was observed at 250 volts for various currents. Table III shows some quantitative data for silver and gold obtained under conditions which would normally cause cathode erosion for palladium contacts. As indicated, the losses for both silver and gold were from the anode. By raising the arc current to 6.4 amps, silver still failed to indicate a reversal. A tentative explanation of this behavior is proposed in a later section.

2. EXPERIMENTS WITH LONGER AIR BREAKDOWN ARCS ON CLOSURE

To study the effects on erosion character of a gas present between the contacts in the arc channel, the following experiment was performed. Instead of the 250 volts used in the aforementioned experiment, corresponding to a separation of about 25,000 Å, a voltage of 500 was used. Arcs obtained were therefore initiated as air breakdowns. The corresponding separation at which an arc is initiated in air is about 3×10^{-3} cm which is of the order of 60 mean free paths of an electron in atmospheric air. In these experiments this large separation eliminated the previous difficulty of premature closure. Table IV presents erosion data

TABLE IV — EROSION DATA FOR PALLADIUM CONTACTS ON CLOSURE BY ARCS INITIATED AT 500 VOLTS AS AIR BREAKDOWNS*

Current Amp	Arc Duration 10^{-9} sec	No. of operations 10^3	Erosion: (loss, gain) 10^{-7} cc		Loss/total loss		Rate of loss 10^{-14} cc/erg	
			Anode	Cathode	Anode	Cathode	Anode	Cathode
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3.2	35	184	*n*, buildup†	0.72, *n*†	0.0	1.0		2.5
3.2	280	36	*n*, buildup	1.85, *n*	0.0	1.0		4.0
3.2	385	18	*n*, buildup	1.2, *n*	0.0	1.0		3.8
6.4	17.5	216	*n*, buildup	0.55, *n*	0.0	1.0		1.6
6.4	35	108	*n*, buildup	0.82, *n*	0.0	1.0		2.4
6.4	70	54	*n*, buildup	1.6, *n*	0.0	1.0		4.7

* In the course of these experiments, some metal loss from the anode was occasionally observed. This was believed to be due to the statistical time lags of air breakdown which would cause a decrease in the contact separation at which the arc was initiated. By illuminating the contacts with ultraviolet this difficulty was eliminated.

† See footnotes below Table I.

for palladium contacts obtained at 3.2 and 6.4 amp. The direction of transfer was independent of arc duration and consistently from cathode to anode. Each anode generally showed a well defined buildup closely matching a hole on the cathode. In contrast to the buildups obtained with short arcs, which were usually irregular and sometimes had more than one peak, these were more regular and usually had a single peak. This difference may be attributed to differences between the initiation mechanisms of short arcs and air breakdowns. Short arcs are initiated by field emission and a sharp point on the cathode surface determines the location of the arc. This point does not necessarily correspond to the smallest separation and on successive closures the arc channel is more or less randomly located. For air breakdowns, on the other hand, surface irregularities are not as significant and the location of the breakdown channel is mainly at the cathode point nearest to the anode.

The rate of cathode erosion for palladium contacts by 500-volt air breakdowns, Table IV, Column 8, is between 1.6×10^{-14} and 4.7×10^{-14} cc/erg depending on current and arc duration.

For silver and gold contacts, the same erosion behavior was obtained. For the 500-volt air breakdowns, holes were obtained on the cathode and buildups on the anode. Table V shows typical data obtained from two test runs with silver and gold contacts. It is of interest to note that their rate of erosion is 4 to 5 times less than for palladium at similar conditions.

TABLE V — EROSION OF SILVER AND GOLD CONTACTS ON CLOSURE BY 500-VOLT AIR BREAKDOWNS

Current Amp	Arc Duration 10^{-9} sec	No. of Operations 10^3	Erosion: (loss, gain) 10^{-7} cc		Loss/total loss		Rate of loss 10^{-14} cc/erg	
			Anode	Cathode	Anode	Cathode	Anode	Cathode
Ag, 3.2	385	230	*n**, build-up*	3.4, *n*	0.0	1.0		0.85
Au, 3.2	140	90	*n*, buildup	0.64, *n*	0.0	1.0		1.1

* See footnote below Table I.

3. EXPERIMENTS WITH SHORT ARCS BETWEEN ACTIVATED CONTACTS ON CLOSURE

Contacts activated by organic vapors[3] have been shown to arc more readily than clean contacts. They are initiated at appreciably lower fields[4] and maintained at appreciably lower currents.[3] The following experiments were carried out to study the erosion behavior of activated contacts. For such contacts an arc is initiated at fields as low as 10^5 volts/cm. For an initial voltage of 250 this corresponds to a separation of 2.5×10^{-3} or about 50 mean free paths of an electron in atmospheric air. This indicates that activation experiments performed in atmospheric air at such a voltage would give erosion results that may be influenced by the presence of air in the arc channel as discussed in the previous section. This difficulty was eliminated by operating the contacts in a vacuum of 10 microns. Organic materials left in the construction of the sound head used for operating the contacts provided sufficient organic vapors for rapid activation of the contacts. The voltage transient across the contacts during closure was observed on an oscilloscope. At the beginning of the test, when the contacts were clean, a certain frequency of premature or early closures was observed. As the contacts became more active the frequency of premature closures decreased and finally disappeared. This was an indication that gradual activation initiated the arcs at progressively increasing separation. The period of activation was usually between 2 and 5 minutes, at 60 operations/sec, with the test continued for about one hour thereafter. Further evidence of contact activation was the formation of considerable quantities of black sooty deposits which were not metallic as indicated by fuming solubility tests.

[3] L. H. Germer, J. Appl. Phys., **22,** p. 955, 1951.
[4] M. M. Atalla, B.S.T.J., **32,** p. 1493, 1953.

Tests were performed on the more or less noble metals palladium, silver and gold and on the base metals copper, nickel, tungsten, iron and aluminum. Not only did the noble metals become active but also the base metals copper, nickel and tungsten. The sooty deposit which is typical for activated contacts was observed on all these metals. Contacts of iron and aluminum, however, failed to show any sign of activation even after as many as 6×10^5 operations.

The metals that were activated have shown one common erosion behavior. Metal loss was almost entirely from the cathode in the form of a shallow depression spread over a considerably larger area than obtained with clean metals. The anode showed little or no metallic deposits in contrast to the sizable buildups obtained with clean metals. For activated palladium the rate of erosion was measured at about 1.0×10^{-14} cc/erg which is about one-half to one-fourth the rate of erosion for clean palladium.

Additional experiments were performed on activated palladium and silver contacts in the presence of air at 50 volts. The degree of activation of the contacts was controlled by varying the concentration of d-limonene vapor in air. Only one result of these experiments is reported here concerning a characteristic difference between the erosion of activated palladium and silver contacts. Palladium contacts showed loss from the cathode even for concentrations of d-limonene vapor as low as 4 per cent of the saturation concentration. Silver, on the other hand, did not show erosion from the cathode until appreciably higher concentrations, 10 to 20 times that for palladium, were introduced.

4 EXPERIMENTS ON BREAK

The objects of these experiments was to compare the erosion of contacts by arcs obtained on opening with the erosion of similar arcs obtained on closure. This was done by allowing a cable to discharge from approximately the same voltage of 250 through two pairs of contacts, one during closure and the other during opening.* Palladium contacts were used with arc durations of 35×10^{-9} and 380×10^{-9} sec at 3.2 amp. The erosion behavior was almost identical for both pairs of contacts. For the short arc duration both contacts exhibited anode loss whereas for the long arc duration cathode loss occurred in both cases.

Measurements on Pd were also made with air breakdown arcs initiated during contact opening at 500 volts. Cathode loss, observed in similar

* The discharge on opening was obtained during the charging of the cable following first separation of the contacts. By adjusting the charging resistor it was possible to control the breakdown voltage.

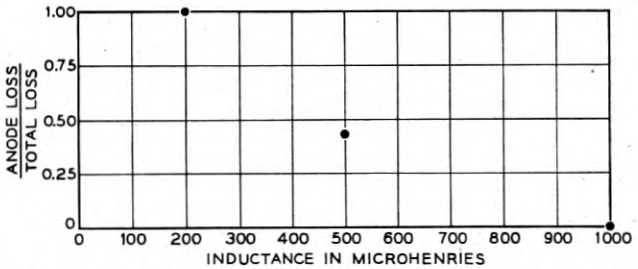


Fig. 2 — Erosion of Pd contacts on break of inductive circuit; I = 0.1 amp, V = 6 volts.

arcs on contact closure, was duplicated here. From these data one concludes that *there is no basic difference between the arcs and erosion effects occurring during the closure or opening of contacts, provided that the initiation conditions are the same.*

Another experiment was carried out on contact erosion due to arcing on break. While the results of this experiment did not yield additional basic information, beyond confirming the above findings, they are of some practical interest. A contact was made to open a 6-volt circuit containing a 60-ohm resistance and a variable inductance. Arcing on opening occurred in the form of a succession of short breakdowns whose duration varied with the circuit inductance. Three inductances, 200, 500, and 1000 microhenries, were tried with palladium contacts. The metal loss results are shown in Fig. 2. At 200 microhenries most of the arcing occurred at small contact separations thereby producing anode loss. At 500 microhenries, arcing was a mixture of short arc breakdowns and longer air breakdowns which caused loss from both electrodes. At 1,000 microhenries, arcing was predominantly due to air breakdowns at wider separations which gave loss mainly from the cathode. The results of this experiment should be useful in indicating the role of arcing in distorting results in low voltage experiments designed to study bridge transfer during contact opening.

In the following section an analysis of the data is presented, and a tentative mechanism of the short arc and contact erosion is proposed.

5. DISCUSSION — TENTATIVE MECHANISM OF THE SHORT ARC AND CONTACT EROSION

Germer and Smith[5] have attempted to record the voltage transient across a pair of contacts during the initiation of a short arc on a high

[5] L. H. Germer and J. L. Smith, J. Appl. Phys., **23**, p. 553, 1952.

speed oscilloscope. Their results have shown a rapid drop to the final arc voltage in a time less than the time resolution of the scope (about 2×10^{-9} sec). It was concluded that the arc initiation time was probably less than 10^{-9} sec. This indicates that in our experiments, all the discharges at small separations must have been maintained at the arc voltage for almost their entire duration, the shortest duration being 17.5×10^{-9} sec.

It has also been shown that short arcs in air[6] or in vacuum[7] are initiated by field emission electrons. Furthermore, from the size of arc pits obtained by Germer and Haworth, Kisliuk[8] has concluded that in the short arc, the electrons are emitted from the cathode primarily by field emission and the arc is maintained by ionization of the metal vapor from the electrodes by electron collision. For a metal with work function φ and minimum ionization potential V_i, the observed arc voltage usually exceeds the sum $V_i + \varphi$ by a volt or less. One may, therefore, postulate the existence of a cathode dark space, where electrons acquire enough energy to produce ionizing collisions, followed by an arc column where a plasma is maintained.

If V_c is the voltage drop through the cathode dark space, j_- the electron current density emitted from the cathode and j_+ the ion current density at the cathode edge of the plasma, the field strength F on an infinite plane cathode, is given by Mackeown's[9] equation:

$$F^2 = 7.57 \times 10^5 (V_c)^{1/2} j_- \left[\frac{j_+}{j_-} (1845W)^{1/2} - 1 \right] \qquad (1)$$

where W is the atomic weight of the ions, F is in volts per cm, V_c in volts and j_- is in amp/cm^2. For the short arc, where the separations are very small, the observed current densities indicate that the width of the arc is usually considerably larger than the contact separation or arc length. It is not too unreasonable, therefore, to neglect the edge effect and apply the above equation. Furthermore, this steady state equation should still be applicable to a changing arc, as will be shown to be the case for the short arc, provided that the changes occurring within an ion transit time are very small.

The cathode electron current density j_-, for an arc maintained by field emission, is further related to the field F at the cathode by the Fowler-

---

[6] M. M. Atalla, B.S.T.J., **34**, p. 203, 1955.
[7] W. S. Boyle, P. Kisliuk, and L. H. Germer, J. Appl. Phys., **26**, p. 571, 1955.
[8] P. Kisliuk, J. Appl. Phys., **25**, p. 897, 1954.
[9] S. S. Mackeown, Phys. Rev., **34**, p. 611, 1929.

TABLE VI — RELATION BETWEEN j_+ AND j_+/j_- AT AN INFINITE PLANE CATHODE; V_c = 10 VOLTS, W = 100 AND φ = 5, 4.5 AND 3 e. VOLTS.

j_+/j_-		0.05	0.1	0.2	0.4	0.6	1.0
j_+ 10^6 amp/cm²	$\varphi = 5$	10.5	8.40	7.07	6.04	5.53	5.00
	$\varphi = 4.5$	6.93	5.64	4.85	4.15	3.77	3.45
	$\varphi = 3$	1.43	1.22	1.04	0.905	0.844	0.780

Nordheim equation:[10]

$$j_- = 1.54 \times 10^{-6} \frac{F^2}{\varphi} \exp[-6.83 \times 10^7 \varphi^{3/2} f(y)/F] \qquad (2)$$

where $f(y)$ is the Nordheim elliptic function[11] of the variable $y = 3.79 \times 10^{-4} F^{1/2}/\varphi$ and φ is the work function of the cathode metal.

Physically, (1) and (2) must be satisfied simultaneously at the cathode. By combining the two equations one may eliminate the field term F and obtain a unique relation between j_+ and j_+/j_- for a fixed value of φ.* Table VI presents calculations made at φ = 5, 4.5 and 3 e. volts. This is essentially the same procedure previously followed by Wasserab.[12] One observes from Table VI that for a wide range of j_+/j_- (at constant ϕ) the change in the ion current density is relatively small. For instance, a 2-fold decrease in j_+ corresponds to a 20-fold increase in j_+/j_-. *Short arcs, therefore, and more generally all field emission arcs, are maintained at approximately a constant ion current density at the cathode.* For most contact metals this density is of the order of 10^6 amp/cm².

It has been shown[6,7] that the short arc is initiated when the power density of the field emission electrons bombarding an anode spot becomes sufficiently high to cause anode evaporation. From this, one may con-

[10] A. Sommerfeld and H. Bethe, Handbuch der Physik (Verlag. Julius Springer, Berlin) **24,** p. 441, 1933.

[11] L. W. Nordheim, Proc. Roy. Soc., **A121,** p. 626, 1928.

* A third equation may be introduced relating j_+/j_- to the collision cross-sections and the gas density distribution in the gap. For the one-dimensional case, neglecting recombination, the equation is given by:

$$\frac{j_+}{j_-} = \frac{Q_i}{Q_i + Q_e} \left[1 - \exp\left(-(Q_i + Q_e) \int_0^d N \, dx\right)\right]$$

where Q_i and Q_e are the ionization and excitation cross-sections of the metal vapor. Due to lack of data on atomic cross-sections and the physical complexity of the pressure distribution between the contacts, no attempt has been made to calculate ionization rates in the gap. Instead, the analysis was carried out by leaving j_+/j_- as an adjustable variable.

[12] T. Wasserab, Z. Physik, **130,** p. 311, 1951.

clude that in its earliest stages, an established arc runs primarily in *anode* metal vapor. The evaporating spot on the anode is then minimum in size. For a constant current arc, which also operates at constant power, the corresponding rate of anode evaporation must, therefore, be a maximum since the maximum rate of heat conduction into the anode is proportional to the size of the anode spot, its boiling temperature and thermal conductivity. Fig. 3 is a diagrammatic representation of the conditions between a pair of contacts at an early stage of the arc. For radius a of the boiling anode spot, an approximately equal area on the cathode must constitute the electron emitting area, since the cathode field is maintained by the approaching positive ions which have been formed by electron collisions with the anode vapor. For a plane at a dis-

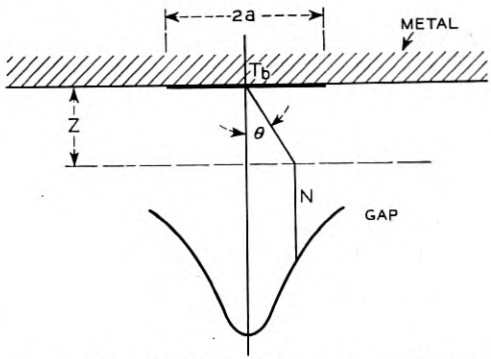


Fig. 3 — Vapor density distribution due to a small evaporating spot.

tance z from the anode, the density of vapor, originating at a small anode spot of radius a, is roughly inversely proportional to both z^2 and $(\cos\theta)^3$. From this, one should expect the electrons approaching the anode to have a strong tendency to scatter to the periphery. The resultant redistribution of the energy of the bombarding electrons over a larger area causes the boiling area to expand in size. On the average, therefore, the effective length of an electron path before reaching the anode will increase. The number of collisions, including ionizing collisions, will increase. Hence the ratio j_+/j_- increases with the size of the anode spot. This growth of the anode spot with time has been observed[13] by examination of the sizes of anode pits produced by single arcs of different durations.

According to Table VI, an increase in j_+/j_-, where both j_+ and j_- are measured at the cathode, corresponds to a proportionate decrease in

[13] W. S. Boyle and L. H. Germer, J. Appl. Phys., **26**, p. 571, 1955.

j_-, since the decrease in j_+ is relatively small. The rate of evaporation of the anode will decrease for three reasons: (1) the rate of energy dissipation by conduction increases with increase in anode spot size, (2) the electron current density is decreasing, and (3) the average energy of an electron reaching the anode is decreasing due to the increase in inelastic collisions. The cathode spot on the other hand, being approximately equal in size to the anode spot, is bombarded by ions of practically constant current density. Furthermore, each ion reaching the cathode will have the same high energy corresponding to the potential drop in the cathode fall. *All prevailing conditions, therefore, will tend to decrease the rate of energy dissipation at the anode while increasing it at the cathode. If the ion current becomes sufficiently high, cathode evaporation occurs.* This corresponds to a critical ratio j_+/j_- which is calculated in the following section.

For a plane cathode spot of radius a, the boiling temperature is reached at its center when

$$j_+(V_c + V_i - \varphi)a/\bar{\lambda}\Delta T_b = 1, \tag{3}$$

and evaporation takes place when it exceeds 1. The term in brackets is the energy of condensation of an ion on the cathode surface,[14] $\bar{\lambda}$ is an appropriate average thermal conductivity of the cathode metal for the temperature range between ambient and boiling, and ΔT_b is the temperature rise of the cathode to boiling. The total arc current I, in terms of j_+ and j_- at the cathode, is given by:

$$I = \pi a^2(j_+ + j_-) \tag{4}$$

Combining Equations (3) and (4) to eliminate a, the critical condition for maintenance of the cathode spot at boiling becomes:

$$j_+ \frac{(j_+)}{(j_-)} = \frac{\pi}{I} \frac{(\bar{\lambda} \cdot \Delta T_b)^2}{(V_c + V_i - \varphi)} \tag{5}$$

In a previous section it has been shown that a combination of the emission and space charge equations, (1) and (2), gives a relation of the form $j_+ = f(j_+/j_-, \varphi)$; See Table VI. This can be combined with (5) to eliminate j_+, thereby expressing j_+/j_- in terms of the cathode physical constants. *This is the critical ratio of j_+/j_- which must be exceeded to cause evaporation of the cathode spot.* Unfortunately, however, data on thermal conductivity above the melting point, are only available for the low melting point metals. For the majority of these, the change in the

[14] K. G. Compton, Phys. Rev., **37**, p. 1077, 1931.

thermal conductivity with temperature is rather small except at melting where a sudden substantial decrease in conductivity occurs. In Table VII values of λ_0, λ_b and λ_0/λ_b are given for various metals as obtained from the references indicated. For copper and silver, thermal conductivities were calculated from electric resistivity data using the Franz-Wiedemann[15] relation with the theoretical constant 2.45×10^{-8} (volt/°C)^2.

For metals whose thermal conductivities at high temperatures are not available, λ_b was taken as $0.5\lambda_0$ as suggested by the last column in Table VII. Table VIII is a summary of calculations of the critical ratio j_+/j_- for a number of metals. In these calculations, V_c was replaced by $V_i + \varphi$ and the term $V_c + V_i - \varphi$ in (5) by $2V_i$. The error involved is only a

TABLE VII — THERMAL CONDUCTIVITIES OF SOME METALS

Metal	λ_0[a]	λ_b	λ_b/λ_0
	watt/cm°C		
Cd	0.933	0.451[b]	0.45
Pb	0.352	0.209[b]	0.59
Sn	0.657	0.324[b]	0.49
Zn	1.13	0.602[b]	0.53
Al	2.03	0.84[b]	0.41
Ag	4.19	2.1[c]	0.50
Cu	3.88	1.9[c]	0.49

[a] Reference 16. [b] Reference 17. [c] Calculated from electric resistivity data for Ag (Reference 18) and Cu (Reference 19).

fraction of a volt[8] and, furthermore, one can carry out the calculations for metals for which the arc voltage is not certain. The thermal conductivity $\bar{\lambda}$ is taken as the arithmetic mean of λ_0 and λ_b.

Column 6 gives the minimum values of $j_+(j_+/j_-)$ which satisfy both the cathode emission and space charge equations, (1) and (2). This minimum value is a function of the work function and atomic weight of the cathode metal. Column 7 gives the values of $j_+(j_+/j_-)$ required for cathode evaporation as determined by the thermal conduction equation, (5). For a given current, these values are a function of the boiling temperature, the thermal conductivity and the minimum ionization potential of the metal. Column 8 gives values of j_+/j_- obtained from (1) and (2) at the given values of $j_+(j_+/j_-)$ of Column 7. These values of j_+/j_- must be exceeded in an arc discharge before cathode evaporation can occur.

[15] A. Sommerfeld and H. Bethe, Elektronentheorie der Metalle. Handbuch der Physik von Geiger und Scheel, Aufl. 24/2 (Berlin, Julius Springer, 1933).
[16] International Critical Tables.
[17] C. J. Smithells, Metals Reference Book, Interscience Publ. Inc., p. 576, 1949.
[18] Handbook of Metals.
[19] Handbook of Physics and Chemistry, p. 2247, 1953–1954.

TABLE VIII — CRITICAL RATIO j_+/j_- WHICH MUST BE EXCEEDED TO CAUSE CATHODE EVAPORATION. TOTAL CURRENT I = 1.0 AMP. TERM "INST" SIGNIFIES INSTANTANEOUS CATHODE EVAPORATION FROM BEGINNING OF ARC.

	Metal, ϕ[20] e. volts (1)	ΔT_b °C (2)	λ_b wat/cm °C (3)	$\bar{\lambda}$ watt/ cm °C (4)	$2V_i$ e. volts (5)	$j_+(j_+/j_-)$ min. in arc. Eq. (1), (2) amp/cm² (6)	$j_+(j_+/j_-)$ evap. Eq. (5) amp/cm² (7)	(j_+/j_-) evap. Eq. (1), (2), (5) (8)
(1)	Pd, 4.8	2200	0.34	0.51	16.66	1.9×10^5	1.42×10^4	INST
(2)	Ni, 4.84	2900	0.29	0.435	15.27	3.6×10^5	2.13×10^4	INST
(3)	Fe, 4.36	3000	0.31	0.465	15.79	2.5×10^5	2.45×10^4	INST
(4)	Pt, 5.29	4300	0.35	0.525	17.92	1.6×10^5	4.97×10^4	INST
(5)	Ag, 4.3	1950	2.1	3.15	15.19	1.2×10^5	5.12×10^5	0.11
(6)	Au, 4.58	2600	1.5	2.25	18.45	9×10^4	3.15×10^5	0.072
(7)	Cu, 4.47	2300	1.9	2.85	15.45	2.5×10^5	5.62×10^5	0.065

It is evident that *for any metal if the entry in Column 7 is less than that in Column 6, some cathode evaporation will take place even during the earliest stages of the arc.* As shown in Column 8, this is the case for Pd, Ni, Fe and Pt. For Ag, Au and Cu, on the other hand, *the arc may be initiated as a true anode arc and only when relatively high ratios j_+/j_- are obtained in the discharge will evaporation from the cathode take place.* This ratio is highest for silver, 0.11, followed by gold, 0.072 and then copper, 0.065. Unfortunately, the present analysis cannot be carried further to determine whether such ionization rates can or cannot be obtained in a discharge, due to the lack of data on collision cross-sections for vapors of these metals. The analysis as such, however, establishes some basic differences among metals in their erosion behavior, by showing some to have stronger tendencies than others for cathode evaporation.* Our observations are in accordance with this conclusion where with Pd,† Fe and Ni it was possible to have enough cathode evaporation

[20] H. B. Michaelson, J. Appl. Phys., **21**, p. 456, 1950.

* It is of interest to point out that Froome[21] has observed similar differences for arcs at low gas pressures on Hg and Cu cathodes. For 10^{-7} sec. arcs on Hg, multiple non-stationary cathode spots were observed while with Cu the spots were not visible and often non-existent. From heat conduction calculations, similar to the above, Froome concluded that while Hg could be easily vaporized, Cu would not even be heated to red heat. For 30×10^{-6} sec arcs, however, cathode spots on Cu were observed.

† For Pd, the observed time for the reversal of the transfer is of the order of 10^{-7} sec, Fig. 1. This time is appreciable in terms of the electron and ion transit times and is attributed to thermal relaxation of the contact metal. It is of the same order as the observed time lags preceding the initiation of the short arc which were shown[6] to correspond to the heating time of the anode spot.

[21] K. D. Froome, Proc. Phys. Soc., (London) **60**, p. 431, 1948.

to exceed that of the anode while for Ag, Au and Cu this reversal was not obtained.*

When evaporation from a cathode spot occurs, it modifies the gas density distribution between the contacts by introducing a high density region near the cathode. This causes additional scattering of the emitted electrons which enhances the spread of the bombarded anode spot and decreases its rate of evaporation. It is, therefore, conceivable that a condition could be reached where the anode spot, radius a_+, becomes large enough compared to the cathode spot, radius a_-, that the rate of evaporation of the cathode exceeds that of the anode. The conditions under which this may occur will now be derived. The power dissipated at the cathode is $j_+(\pi a_-^2)(2V_i)$, and the power dissipated at the anode is $(j_+ + j_-)\pi a_-^2\varphi$, where both j_+ and j_- are measured at the cathode. The specified anode power is actually a lower limit since all the electrons are assumed to reach the anode with zero kinetic energy. This rate of evaporation is assumed to correspond to the difference between the power delivered by electrons or ions and the power dissipated by conduction through the corresponding electrode. The heat dissipated by metal melting is neglected.† For two hemispherical spots, one on each electrode, maintained at boiling temperature, the rate of cathode evaporation exceeds that of the anode if:

$$j_+\pi a_-^2(2V_i - \varphi) - 2\pi a_-\lambda\cdot\Delta T_b > (j_+ + j_-)\pi a_-^2\varphi - 2\pi a_+\lambda\cdot\Delta T_b \qquad (6)$$

Combining with (4) and assuming that $j_-/j_+ \gg 1.0$, one gets

$$\frac{a_+}{a_-} > 1 + \frac{\varphi}{2\lambda\cdot\Delta T_b}\left[1 - \frac{j_+}{j_-}\left(\frac{2V_i}{\varphi} - 1\right)\right]\left[\frac{j_-}{j_+}\frac{Ij_+}{\pi}\right]^{1/2} \qquad (7)$$

If ionization is due mainly to ionizing collisions between electrons and metal atoms, it can be shown that (for electron energies slightly above V_i), the maximum ion to electron ratio obtainable is $Q_i/(Q_i + Q_e)$, where Q_i and Q_e are the ionization and excitation cross-sections. No data is available to permit a calculation of this ratio for any of the metals in this investigation. For mercury, however, this ratio is about 1/7, for electrons at 0.2 volt above the minimum ionization potential of mercury.

* This statement is not meant to exclude the possibility of some cathode evaporation for these metals since our testing method is not capable of detecting cathode evaporation if it is much less than that of the anode.

† This is metal leaving the cathode surface. The error involved is discussed later.

This was obtained from ionization data[22] and excitation data.[23] A tentative calculation was, therefore, carried out for Pd at the two values of 0.1 and 0.2 for j_+/j_-. The corresponding values of j_+ are 7×10^6 and 6×10^6 amp/cm^2 respectively. At $j_+/j_- = 0.1$, the calculated ratio a_+/a_- is between 6 and 12, the lower value based on λ_0 and the higher value at λ_b. At $j_+/j_- = 0.2$, a_+/a_- is between 3.5 and 6.*

For metals such as Pd which exhibit a reversal to the cathode arc by showing cathode loss, one can make another estimate of the probable ratio j_+/j_- for such arcs from the measurement of the rate of erosion. If a_- is the radius of the cathode spot, the power used in evaporation is taken as the difference between the power dissipated at the cathode and the power dissipated by conduction to maintain the cathode spot at the boiling temperature. This neglects the energy carried away by molten metal which may escape the cathode spot and deposit elsewhere. Observations on single arc anode pits, however, have shown[1] that each pit was surrounded by a rim which contains most of the metal from the pit. For Pt the volume of this metal was less by a factor of three than the amount which can be melted by the arc energy. To correct for this melting effect in calculating the rate of metal evaporation, one must not assume that the melting energy of the displaced metal is lost since this metal still remains on a rather narrow rim surrounding the pit.† From the photograph in reference 1, it appears that the average width of the rim is 10 to 15 per cent, the diameter of the pit. The effect of displacing this molten metal, therefore, is a redistribution of the initial arc energy where 70 per cent of the energy is dissipated in the pit area and 30 per cent dissipated on a surrounding rim 10 to 15 per cent the diameter of the pit. For the degree of accuracy desired in our calculations, it appears justifiable to neglect this effect.

---

[22] W. B. Nottingham, Phys. Rev., **55**, p. 203, 1939.

[23] H. Massey and E. Burhop, Electronic and Ionic Phenomena, p. 62. (Oxford, Clarendon Press, 1952).

* Single arc pit measurements were also made for Pd contacts at 3.2 amp arc of 0.39×10^{-6} sec duration initiated at 250 volts. The single anode pit observed had an average diameter of 16×10^{-4} cm corresponding to a current density of only 1.5×10^6 amp/cm^2. Comparing with Table VI, one finds that unless the cathode emitting spot is considerably smaller than the observed anode spot, this low density may be obtained only if high ratios of j_+/j_-, higher than 1.0, are attainable. This is unlikely for the low energy electrons in the short arc. Actually cathode observations, with 1,700 magnification, have shown a number of smaller individual pits, probably an indication of a non-stationary cathode spot in accordance with previous cathode observations,[24] of an average diameter of 2.4×10^{-4} cm. If only one of these pits carried the total current at any one time, the current density would be 70×10^6 amp/cm^2 corresponding to a j_+/j_- of about 0.1, Table VI. The measured ratio a_+/a_- is 6.7.

[24] J. D. Cobine, Gaseous Conduction, McGraw-Hill, 1941.

† It is evident that no correction is needed if the molten metal is not displaced.

No correction is needed for the heat of condensation of the metal deposited on the cathode spot in the form of neutral atoms, since these will be reflected, with or without loss of identity, from the cathode spot which is already at the boiling temperature.

The ratio of the evaporation power to the input power Iv is given by:

$$\frac{\text{Evaporation power}}{\text{Input power}} = \frac{1}{Iv}\left[j_+\pi a_-^2(2V_i) - 2\pi a_-\lambda\cdot\Delta T_b\right]$$

Eliminating a_-, through the introduction of the total current I, and setting $x = j_+/j_+ + j_-$, one gets:

$$\frac{\text{Evaporation power}}{\text{Input power}} = x\cdot\frac{2V_i}{v} - \frac{2\lambda\cdot\Delta T_b}{v}\left(\frac{\pi x}{Ij_+}\right)^{1/2} \tag{8}$$

For any value of j_+/j_-, or of x, j_+ is determined from Table VI for palladium and the power ratio in (8) may be obtained. From the physical properties of Pd* the volume evaporated per unit energy is about 1.8×10^{-12} cc/erg. If it is assumed that 50 per cent of the evaporated metal from the cathode is redeposited on the cathode, one can calculate the cathode loss per unit input energy from (8), for each value of j_+/j_-. Results of such calculations for Pd are given in Table IX. The erosion rate of the cathode of Pd contacts was measured at about 3.5×10^{-14} cc/erg, Tables I and II. These were obtained from measurements with arcs of durations sufficient to allow erosion reversal. During the first portion of each of these arcs, as much as 50 per cent of the total arc duration, metal was transferred from the anode to the cathode at an average rate of about 3×10^{-14} cc/erg. The rate of cathode loss† is probably as high as

$$(3.5 \times 10^{-14}) + (3 \times 10^{-14}) = 6.5 \times 10^{-14} \text{ cc/erg.}$$

From Table IX, one may therefore conclude that *for the latter stage or cathode stage of the short arc in Pd, an upper limit of 10 per cent of the total current is carried by positive ions.*

In the section on measurements, it was noted that for the longer arcs, initiated as air breakdowns, the erosion was consistently from the *cathode* for all the metals which were investigated. These experiments were performed in laboratory air at 500 volts and the corresponding contact separation was about 3×10^{-3} cm. At the high value of pd or Nd prevailing in the gap, the anode is, from the beginning, sufficiently

* Reference 17, p. 419.

† It is possible that this observed loss is not all due to surface evaporation but may be partly due to some metal leaving the surface in the molten stage. The calculated ratio j_+/j_- is only, therefore, an upper limit.

TABLE IX — RATIO OF EVAPORATION POWER TO INPUT POWER, AND RATE OF Pd METAL LOSS FOR A CATHODE ARC AT DIFFERENT VALUES OF j_+/j_-. CALCULATIONS ARE FOR $I = 1.0$ AMP. $\lambda = (\lambda_0 + \lambda_b)/2 = 0.51$ WATT/CM°C

j_+/j_-	0.04	0.06	0.08	0.10	0.20
$\frac{1}{x} = \frac{j_+ + j_-}{j_+}$	26	17.7	13.5	11	6
j_+ = amp/cm²	9.5×10^6	8.3×10^6	7.6×10^6	7×10^6	6×10^6
Evap. power / Input power	0.028	0.044	0.060	0.076	0.15
Cathode loss, rate, 10^{-14} cc/erg (based on 50 per cent of total evap. rate)	2.4	3.9	5.4	6.8	13

shielded while the cathode is being continuously bombarded over a relatively small area by high energy ions dropping through the cathode fall. If n_0 is a number of electrons leaving the cathode and Q is the total collision cross-section, then the number of electrons n_d reaching the anode without any collisions is given by:

$$n_d/n_0 = \exp(-NQ\,d) \tag{9}$$

At 10 volts, Q for both O_2 and N_2 is about 10^{-15} cm².[25] For atmospheric air at 300°K and $d = 3 \times 10^{-3}$ cm, (9) shows that practically no electron will reach the anode without an elastic or inelastic collision. Those having only elastic collisions will undergo little change in energy but will be scattered to an anode spot larger than the cathode emitting spot. If one assumes the inelastic collision cross-section to be 15–20 times less than the total collision cross-section, less than 1–3 per cent of the electrons will reach the anode with full energy. One concludes, therefore, that *for arcs initiated as gas breakdowns at small separations, erosion is generally confined to the cathode, provided that the product Nd is high enough to provide sufficient anode shielding.* These arcs initially run primarily in the gas between the contacts until cathode evaporation occurs when cathode vapor will contribute to the maintenance of the arc. The erosion data for palladium in air given in Table IV do substantiate this by showing an increase in the rate of cathode evaporation with increasing arc duration. At 3.2 amps, the rate of erosion increases from 2.5×10^{-14} cc/erg. for 35×10^{-9} sec arcs, to 4.0×10^{-14} cc/erg. for 280×10^{-9} sec. At 6.4 amp, it increases from 1.6×10^{-14} cc/erg. at 17.5×10^{-9} sec to 4.7×10^{-14} cc/erg. at 70×10^{-9} sec.

For *fully* activated contacts of Pd and Ag, erosion was also obtained

[25] R. B. Brode, Revs. Modern Phys., **5**, p. 257, 1933.

from the cathode. Since these experiments were performed at a vacuum of 10 microns, anode shielding must have been provided by means other than air between the contacts. Activated contacts are characterized by the sooty products deposited on the contact surfaces, the lower fields for arc initiation and the lower currents at which the arc can be maintained.[3,4] The activation deposits are organic and have poor conduction properties and probably low boiling temperatures. When the arc is initiated, the anode surface will present, at least temporarily, the physical properties of the deposit rather than those of the substrate. The anode arc stage, discussed above, will therefore be maintained in vapor from the anode deposit. Due to the low conductivity, boiling point and heat evaporation of the deposit, the evaporation rate and the rate of growth of the anode spot must be appreciably higher than for the clean metal.* Furthermore, according to (5), the cathode deposit will boil more readily and the transition to the cathode arc stage will occur sooner. *Shielding of the anode metal is provided, therefore, first by vapor from the anode deposit and then by vapor from the cathode deposit which may finally be mixed with cathode metal vapor.* Arc voltage transients across activated contacts, reported by Germer and Smith,[5] do substantiate this by showing a gradual transition, within one arc, from the higher arc voltage of the activating substance,[3] to the lower arc voltage of the contact metal. Finally, the lower cathode erosion rates observed for activated contacts are readily explainable as due to the arc energy expended in decomposing and evaporating the organic deposits.

* This is substantiated by the observation that arcs between active contacts can be maintained at much lower currents than for clean contacts. This indicates that at these lower currents, corresponding to lower dissipated power, enough vapor pressure was maintained between the contacts to provide the necessary ionization.

Bell System Technical Papers Not Published in This Journal

ACHENBACH, C. H.[1] (Retired)
Power Equipment for Telephone Central Offices, Telephony, **148,** Apr., 1955.

ANDERSON, P. W., see Weiss, M. T.

BAGNALL, V. B.[3]
Operation Dew Line, J. Franklin Inst., **259,** pp. 481–490, June, 1955.

BAKER, W. O., see Winslow, F. H.

BECK, A. C.[1]
Measurement Techniques for Multimode Waveguides, Proc. I.R.E., MTT-3, pp. 35–41, Apr., 1955.

BEMSKI, G.,[1] NIERENBERG, W. A.,[6] and SILSBEE, H. B.[6]
Cosine Interaction in CsF and RbF, Phys. Rev., **98,** pp. 470–473, Apr., 1955.

BENES, V. E.[1]
On the Consistency of an Axiom of Enumerability, J. Symbolic Logic, **20,** p. 29, Mar., 1955.

BEST, F. S., see Harrower, G. A.

BITTRICH, G., see Compton, K. G.

BOGERT, B. P.[1]
Sterophonic Sound Reproduction Enhancement Utilizing the Haas Effect, S.M.P.T.E., **64,** pp. 308–309, June, 1955.

BOORSE, H. A., see Smith, B.

BOYLE, W. S.[1]
Self-Propagating Intermittent Discharge, J. Appl. Phys., **26,** pp. 584–586, May, 1955.

[1] Bell Telephone Laboratories, Inc.
[3] Western Electric Company, Inc.
[6] University of California.

BOYLE, W. S.[1] and GERMER, L. H.[1]
Arcing at Electrical Contacts on Closure. Part VI — The Anode Mechanism of Extremely Short Arcs, J. Appl. Phys., **26,** pp. 571–574, May, 1955.

BOYLE, W. S.,[1] KISLIUK, P.,[1] and GERMER, L. H.[1]
Electrical Breakdown in High Vacuum, J. Appl. Phys., **26,** pp. 720–725, June, 1955.

BROWN, S. C., See Rose, D. J.

BUEHLER, E., see Tanenbaum, M.

BURNS, F. P.,[1] and QUIMBY, S. L.[1]
Ordering Processes in Cu_3Au, Phys. Rev., **97,** pp. 1567–1575, Mar. 15, 1955.

CETLIN, B. B., see Geller, S.

COMPTON, K. G.,[1] EHRHARDT, R. A.,[1] and BITTRICH, G.[1]
Brass Plating, J. Am. Electroplaters Soc., **41,** pp. 1431–1439, Dec., 1954.

DARLINGTON, S.[1]
An Introduction to Time-Variable Networks, Proc. of Symposium on Circuit Analysis, pp. 5-1 to 5-25, May, 1955.

EHRHARDT, R. A., see Compton, K. G.

FEHER, G.,[1] and KIP, A. F.[6]
Electron Spin Resonance Absorption in Metals. Part 1 — Experimental, Phys. Rev., **98,** pp. 337–348, Apr. 15, 1955.

FELCH, E. P.,[1] and ISRAEL, J. O.[1]
A Simple Circuit for Frequency Standards Employing Overtone Crystals, Proc. I.R.E., **43,** pp. 596–602, May, 1955.

FISHER, J. R.,[1] and POTTER, J. F.[1]
Significant Factors Affecting the Physical Structure of Dry Pressed Steatite, Am. Ceramic Soc. Bull., **34,** pp. 177–181, June, 1955.

FROST, H. B.[1]
Cathode Interference Impedance Desimplified, Trans. P.G.R.Q.C. 5, pp. 27–33, Apr., 1955.

[1] Bell Telephone Laboratories, Inc.
[6] University of California.

GEBALLE, T. H.,[1] and HULL, G. W.[1]
The Seebeck Effect in Silicon, Phys. Rev., **98,** pp. 940–947, May 15, 1955.

GELLER, S.,[1] and CETLIN, B. B.[1]
The Crystal Structure of $HhSe_2$, Acta Crys., **8,** pp. 272–274, May, 1955.

GELLER, S.,[1] and SCHAWLOW, A. L.[1]
Crystal Structure and Quadrupole Coupling of Cyanogen Bromide, BrCN, J. Chem. Phys., **23,** pp. 779–783, May, 1955.

GELLER, S.[1]
Crystal Structure of RhTe and $RhTe_2$, Am. Chem. Soc. J., **77,** pp. 2641–2644, May 5, 1955.

GERMER, L. H., see Boyle, W. C.

HANNAY, N. B., see Tanenbaum, M.

HARROWER, G. A.,[1] BEST, F. S.,[1] and MACHALETT, A. A.[1]
Coaxial Electrical Connection Into Vacuum, Rev. Sci. Instr., **26,** p. 404, Apr., 1955.

HEIDENREICH, R. D.[1]
Thermionic Emission Microscopy of Metals. Part I — General, J. Appl. Phys., **26,** pp. 757–765, June, 1955.

HOCHGRAF, LESTER,[1] and WATLING, R. G.[1]
Telephone Lines for Rural Subscriber, A.I.E.E. Commun. and Electronics, **18,** pp. 171–176, May, 1955.

HOHN, F.[1]
First Conference on Training Personnel for the Computing Machine Field, Am. Math. Monthly, **62,** pp. 8–15, Jan., 1955.

HOLDEN, A. N.,[1] MATTHIAS, B. T.,[1] MERZ, W. J.,[1] and REMEIKA, J. P.[1]
New Class of Ferroelectrics, Phys. Rev., Letter to the Editor, **98,** p. 546, Apr. 15, 1955.

HOOTON, J. A.,[1] and MERZ, W. J.[1]
Etch Patterns and Ferroelectric Domains in $BaTiO_3$ Single Crystals, Phys. Rev., **98,** pp. 409–413, Apr. 15, 1955.

[1] Bell Telephone Laboratories, Inc.

HULL, G. W., see Geballe, T. H.

HUTSON, A. R.[1]
Velocity Analysis of Thermionic Emission from Single-Crystal Tungsten, Phys. Rev., **98,** pp. 889–901, May 15, 1955.

ISRAEL, J. O., see Felch, E. P.

KARNAUGH, M.[1]
Pulse-Switching Circuits Using Magnetic Cores, Proc. I.R.E., **43,** 570–583, May, 1955.

KEYWELL, F.[1]
Measurements and Collision — Radiation Damage Theory of High-Vacuum Sputtering, Phys. Rev., **97,** pp. 1611–1619, Mar. 15, 1955.

KIP, A. F., see Feher, G.

KISLIUK, P., see Boyle, W. S.

KOHN, W.,[7] and LUTTINGER, J. M.[8]
Theory of Donor Levels in Silicon, Phys. Rev., Letter to the Editor, **97,** p. 1721, Mar. 15, 1955.

KOHN, W.,[7] and LUTTINGER, J. M.[8]
Theory of Donor States in Silicon, Phys. Rev., **98,** pp. 915–922, May 15, 1955.

LUTTINGER, J. M., see Kohn, W.

MACHALETT, A. A., see Harrower, G. A.

MALTHANER, W. A.,[1] and VAUGHAN, H. E.[1]
Control Features of Magnetic-Drum Telephone Office, I.R.E. Trans. P.G.E.C., **4,** 21–26, Mar., 1955.

MASON, W. P.[1]
Dislocation Relaxations at Low Temperatures and the Determination of the Limiting Shearing Stress of a Metal, Letter to the Editor, Phys. Rev., **98,** pp. 1136–1138, May 15, 1955.

---

[1] Bell Telephone Laboratories, Inc.
[7] Carnegie Institute of Technology.
[8] Univesity of Michigan.

MATREYEK, W., see Winslow, F. H.

PAPE, N. R., see Winslow, F. H.

MATTHIAS, B. T., see Holden, A. N.

McSKIMIN, H. J.[1]
Measurement of the Elastic Constants of Single Crystal Cobalt, J. Appl. Phys., **26,** pp. 406–409, Apr., 1955.

MERZ, W. J., see Holden, A. N.

MERZ, W. J., see Hooton, J. A.

MORRISON, J.,[1] and ZITTERSTROM, R. B.[1]
Barium Getters in Carbon Monoxide, J. Appl. Phys., **26,** pp. 437–442, Apr., 1955.

NEWHOUSE, R. C.[1]
Feedback Relations in Military Weapon Systems, I.R.E. Trans. P.G.A.N.E.-1, **3,** pp. 24–27, Sept., 1954.

NIERENBERG, W. A., see Bemski, G.

PFANN, W. G.[1]
Continuous Multistage Separation by Zone-Melting, J. Metals, **7,** pp. 297–303 (Part 2), Feb., 1955.

PIERCE, J. R.[1]
Interaction of Moving Charges with Wave Circuits, J. Appl. Phys., **26,** pp. 627–638, May, 1955.

POTTER, J. F., see Fisher, J. R.

PRINCE, M. B.[1]
Silicon Solar Energy Cells, J. Appl. Phys., **26,** pp. 534–540, May, 1955.

QUIMBY, S. L., see Burns, F. P.

REMEIKA, J. P., see Holden, A. N.

ROBERTSON, S. D.[1]
The Ultra-Bandwidth Finline Coupler, Proc. I.R.E., **43,** pp. 739–741, June, 1955.

[1] Bell Telephone Laboratories, Inc.

ROSE, D. J.,[1] and BROWN, S. C.[4]
High-Frequency Gas Discharge Plasma in Hydrogen, Phys. Rev., **98,** pp. 310–316, Apr. 15, 1955.

SCHAWLOW, A. L., see Geller, S.

SILSBEE, H. B., see Bemski, G.

SMITH, B.,[1] and BOORSE, H. A.[5]
Helium II Film Transport. Part I — The Role of Substrate, Phys. Rev., **98,** pp. 328–336, Apr. 15, 1955.

TANENBAUM, M.,[1] VALDES, L. B.,[1] BUEHLER, E.,[1] and HANNAY, N. B.[1]
Silicon n-p-n Grown Junction Transistors, J. Appl. Phys., **26,** pp. 686–692, June, 1955.

TUKEY, J. W.[1]
Unsolved Problems of Experimental Statistics, Am. Stat. Assoc. J., **49,** pp. 706–731, Dec., 1954.

TURNER, E. C.[9] (Retired)
Rural Telephone Growth Forecasting, Telephony, **148,** Apr., 1955.

VALDES, L. B., see Tanenbaum, M.

VAN ROOSBROECK, W.[1]
Injected Current Carrier Transport in a Semi-Infinite Semiconductor and the Determination of Lifetimes and Surface Recombination Velocities, J. Appl. Phys., **26,** pp. 380–391, Apr., 1955.

VAUGHAN, H. E., see Malthaner, W. A.

WATLING, R. G., see Hochgraf, Lester

WEISS, M. T.,[1] and ANDERSON, P. W.[1]
Ferromagnetic Resonance in Ferroxdure, Phys. Rev., **98,** pp. 925–926, May 15, 1955.

WINSLOW, F. H.,[1] BAKER, W. O.,[1] PAPE, N. R.,[1] and MATREYEK, W.[1]
Formation and Properties of Polymer Carbon, J. Poly. Sci., **16,** pp. 101–120, Apr., 1955.

ZETTERSTROM, R. B., see Morrison, J.

[1] Bell Telephone Laboratories, Inc.
[4] Massachusetts Institute of Technology.
[5] Barnard College, Columbia University.
[9] New York Telephone Company.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ANDERSON, R. E. D.
Magnetically Regulated Battery Charger, Monograph 2360.

BAKER, W. O., see Winslow, F. H.

BENNETT, W. R.
Transmission Through Periodically Operated Switches, Monograph 2402.

BRIGGS, H. B., and FLETCHER, R. C.
Absorption of Infrared Light by Free Carrier in Germanium, Monograph 2308.

BURTON, J. A.
Impurity Centers in Ge and Si, Monograph 2449.

CUTLER, C. C.
The Regenerative Pulse Generator, Monograph 2385.

DEWALD, J. F.
The Kinetics of Formation of Anode Films, Monograph 2361.

FINE, M. E., and KENNEY, N. T.
Low Temperature Acoustic Relaxation in Ni-Fe Ferrites, Monograph 2356.

FLETCHER, R. C., see Briggs, H. B.

GELLER, S., MATTHIAS, B. T., and GOLDSTEIN, R.
Some New Intermetallic Compounds with the "β-Wolfram" Structure, Monograph 2377.

GILMAN, G. W., see Kelly, M. J.

GOLDSTEIN, R., see Geller, S.

GREINER, E. S.
Plastic Deformation of Germanium and Silicon by Torsion, Monograph 2362.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

HALSEY, R. J., see Kelly, M. J.

HAMILTON, B. H.
Semiconductor Devices in Regulated Metallic Rectifiers, Monograph 2359.

HEIDENREICH, ROBERT
Transition Structure, Monograph 2374.

HERRING, C., see Pearson, G. L.

INSKIP, L. S., and WATSON, H. W.
Grounding of Portable Electric Equipment, Monograph 2396.

KARNAUGH, M.
Pulse-Switching Circuits Using Magnetic Cores, Monograph 2407.

KELLY, M. J., RADLEY, SIR GORDON, GILMAN, G. W., and HALSEY, R. J.
A Translantic Telephone Cable, Monograph 2434.

KENNEY, N. T., see Fine, M. E.

KOHN, W., and LUTTINGER, J. M.
Hyperfine Splitting of Donor States in Silicon, Monograph 2399.

KOHN, W., see Luttinger, J. M.

LOUISELL, W. H., and PIERCE, J. R.
Power Flow in Electron Beam Devices, Monograph 2422.

LUTTINGER, J. M., and KOHN, W.
Motion of Electrons and Holes in Perturbed Periodic Fields, Monograph 2400.

LUTTINGER, J. M., see Kohn, W.

MACNEE, A. B., see Talpey, T. E.

MALTHANER, W. A., and VAUGHAN, H. E.
Control Features of a Magnetic-Drum Telephone Office, Monograph 2423.

MASON, W. P.
Aging of Properties of Ferroelectric Ceramics, Monograph 2367.

MATREYEK, W., see Winslow, F. H.

MATTHIAS, B. T.
Superconductivity and Valence Electrons Per Atom, Monograph 2358.

MATTHIAS, B. T., see Geller, S.

McSKIMIN, H. J.
Transducer Design for Ultrasonic Delay Lines, Monograph 2397.

PAPE, N. R., see Winslow, F. H.

PEARSON, G. L., and HERRING, C.
Magneto-Resistance Effect and Band Structure of Silicon, Monograph 2415.

PFANN, W. G.
Continuous Multistage Separation by Zone-Melting, Monograph 2388.

PIERCE, J. R.
General Sources of Noise in Vacuum Tubes, Monograph 2380.

PIERCE, J. R., see Louisell, W. H.

RADLEY, SIR GORDON, see Kelly, M. J.

READ, W. T., JR.
Theory of Dislocations in Semiconductors, Monograph 2435.

SMITH, D. H.
Silicon Alloy Junction Diode as a Reference Standard, Monograph 2364.

TALPEY, T. E., and MACNEE, A. B.
Uncorrelated Component of Induced Grid Noise, Monograph 2436.

VAN ROOSBROECK, W.
Injected Current Carriers in a Semiconductor, Monograph 2411.

VAUGHAN, H. E., see Malthaner, W. A.

VON OHLSEN, L. H.
Small Signal Performance of the 416B Planar Triode, Monograph 2381.

WATSON, H. W., see Inskip, L. S.

WINSLOW, F. H., BAKER, W. O., PAPE, N. R., and MATREYEK, W.
Formation and Properties of Polymer Carbon, Monograph 2403.

Contributors to This Issue

M. M. Atalla, B.S., Cairo University, 1945; M.S., Purdue University, 1947; Ph.D., Purdue University, 1949; Studies at Purdue undertaken as the result of a scholarship from Cairo University for four years of graduate work. Bell Telephone Laboratories, 1950–. For the past three years he has been a member of the Switching Apparatus Development Department, in which he is supervising a group doing fundamental research work on contact physics and engineering. Current projects include fundamental studies of gas discharge phenomena between contacts, their mechanisms, and their physical effects on contact behavior; also fundamental studies of contact opens and resistance. He is a member of Sigma Xi, Sigma Pi Sigma, Pi Tau Sigma, the American Physical Society, and an associate member of the A.S.M.E.

Kenneth Joseph Busch, B.S. in electrical engineering, Lehigh University, 1951. He joined Bell Telephone Laboratories as a member of technical staff the same year. He has worked in the transmission transformer group, principally on high voltage pulse transformers and magnetic pulse modulators. Mr. Busch is a member of the Institute of Radio Engineers, Tau Beta Pi and Eta Kappa Nu.

J. James Ebers, B.S., Antioch College, 1946; M.S., Ph.D., Ohio State University, 1947, 1950. Bell Telephone Laboratories, 1951–. Prior to joining the Laboratories, Dr. Ebers served as an Instructor in Electrical Engineering at Ohio State University from 1947 to 1950, and as Assistant Professor from 1950 to 1951. He worked as a Research Foundation Assistant and Associate at Ohio State University from 1946 to 1951. His early work at the Laboratories was concerned with the development of transistors for switching applications, and for the past 2½ years he has been concerned with the development of the alloyed junction transistor. Member of the American Physical Society, Eta Kappa Nu, and Sigma Xi. Member of the I.R.E. and past chairman of its task force to standardize methods of tests for transistors in switching applications.

Andrew D. Hasley, B.S. in E.E., University of Michigan, 1930; Bell Telephone Laboratories, 1930–. His early work was concerned with the

design and development of transformers for use in telephone carrier systems. During World War II, he was engaged in the design and development of high-voltage pulse transformers for radar systems. Since January, 1945, he has been a supervisor of the pulse transformer group. Member of Tau Beta Pi and the Special Quality Transformer and Reactor Committee of the Radio Electronics and Television Manufacturers Association, 1950 to present. Senior member of the I.R.E.

WARREN P. MASON, B.S. in E.E., University of Kansas, 1921; M.A. and Ph.D. in Physics, Columbia University, 1925 and 1928. He joined the Engineering Department of Western Electric in 1921 before it was succeeded by Bell Telephone Laboratories. His first four years of work were spent in investigations of carrier-transmission systems. Since then he has been involved in investigation of wave transmission networks, both electrical and mechanical, in piezo-electric crystal research, and in the study of the mechanical properties of liquids and solids. At present he heads a subdepartment in the Mathematical Research Department concerned with Mechanics Research. President Acoustical Society of America. Fellow of the Institute of Radio Engineers and the American Physical Society.

GEORGE H. MEALY, A.B., Harvard College, 1951. Mr. Mealy joined Bell Telephone Laboratories in 1951. A member of the Switching Development Department, Mr. Mealy at present heads a group concerned with Fundamental Design Techniques. From 1946 to 1948 he was an electronics technician in the Navy. In 1954, Mr. Mealy taught courses in switching theory and digital computers at the College of the City of New York. Member of the Association for Computing Machinery and the Association for Symbolic Logic, and associate member of the Institute of Radio Engineers.

S. L. MILLER, B.S., Webb Institute of Naval Architecture, 1944; M.A. in Physics, Columbia University, 1949; Ph.D. in Physics, Columbia University, 1952; taught at City College of New York, 1948–1950; Bell Telephone Laboratories, 1952–. Since he has been at the Laboratories Dr. Miller has been engaged in exploratory development work on transistors. He is a member of the American Physical Society and Sigma Xi.

DR. CARL NEITZERT received his B.S. in E.E. in 1928 and his M.A. in Mathematics in 1929 from the University of Missouri. His Sc.D. in E.E. was obtained from the Massachusetts Institute of Technology in

1936 and an honorary degree of Master of Engineering from the Stevens Institute in the Fall of 1954. He was a teaching assistant, research assistant and instructor at M.I.T. during the period 1929 to 1940 and Assistant Professor, Associate Professor and Professor at Stevens from 1940 to 1955. While at Stevens he did electronic consulting work for the New Jersey State Police (1941 and 1942) and General Time Corporation Research Laboratory (1945 to 1955). Since June, 1955, he has been a full-time research engineer at the General Time Corporation Research Laboratory. He has published two papers, *The Measurement of Small Alternating Voltages at Audio Frequencies* (Co-author with E. A. Johnson) R.S.I., Vol. 5, May, 1934, and *Thermal Agitation Voltages in Resistors*, Physics, Vol. 5, October, 1934. His doctor's thesis was entitled, *The Synthesis of a Two-Terminal, Non-Dissipative Network for a Finite Band of Frequencies.* Dr. Neitzert is a member of the I.R.E. Symbols Committee (1946 to date), and was a member of the I.R.E. Circuits Committee (1945 to 1947). In addition to his membership in I.R.E., he is a member of A.I.E.E., Sigma Xi, Tau Beta Pi, Eta Kappa Nu and Pi Mu Epsilon.

S. A. SCHELKUNOFF, B.A. and M.A., State College of Washington, 1923; Ph.D., Columbia University, 1928. Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–26. The State College of Washington faculty, 1926–29. Bell Telephone Laboratories, 1929–. In the early thirties, Mr. Schelkunoff embarked on a continuing study of electromagnetic theory. This included early research in the transmission properties of coaxial lines. For many years he has worked on the theory of wave guides for transmitting microwaves and on theories of various antennas for radio communication. During World War II Mr. Schelkunoff served as a consultant on wave propagation to the National Defense Research Committee and the Navy. Other problems which have engaged Mr. Schelkunoff's attention have concerned resonators, attenuators, radiating horns, artificial grounds and similar aspects of electromagnetic theory. He holds a number of patents on wave guides, antennas and resonators. He is the author of three books, co-author of a fourth and the author of numerous technical papers and articles. Awarded the Morris Liebmann Memorial Prize by the Institute of Radio Engineers in 1942 and the Stuart Ballantine Medal by the Franklin Institute in 1949. Fellow of the Institute of Radio Engineers, Fellow of the American Institute of Electrical Engineers, Fellow of The Association for the Advancement of Science. Also member of the American Mathematical Society, and the Mathematical Association of America.