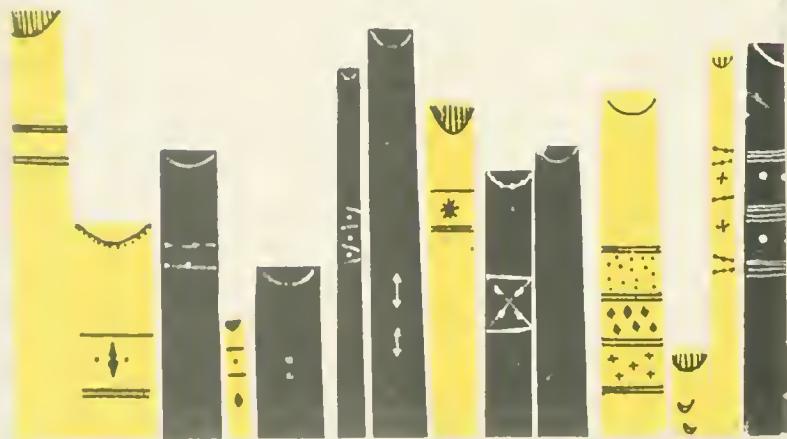


reference collection book



kansas city
public library
kansas city,
missouri



PUBLIC LIBRARY
KANSAS CITY
MO

From the collection of the

z n m
o Prelinger
v a
t p

San Francisco, California
2008

PUBLIC LIBRARY
KANSAS CITY
MO

WAAHALLI CHAMBERS
VITIS CAVOURIA
OM

H E B E L L I S Y S T E M
Technical Journal
VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXV

JANUARY 1956

KARTA NUMBER 1

FEB 1956

JAN 30 1956

Diffused Emitter and Base Silicon Transistors

M. TANENBAUM AND D. E. THOMAS 1

A High-Frequency Diffused Base Germanium Transistor C. A. LEE 23

Waveguide Investigations with Millimicrosecond Pulses

A. C. BECK 35

Experiments on the Regeneration of Binary Microwave Pulses

O. E. DELANGE 67

Crossbar Tandem as a Long Distance Switching System

A. O. ADAM 91

Growing Waves Due to Transverse Velocities

J. R. PIERCE AND L. R. WALKER 109

Coupled Helices

J. S. COOK, R. KOMPFLNER AND C. F. QUATE 127

Statistical Techniques for Reducing the Experiment Time in Reliability Studies MILTON SOBEL 179

A Class of Binary Signaling Alphabets

DAVID SLEPIAN 203

Bell System Technical Papers Not Published in This Journal

235

Recent Bell System Monographs

242

Contributors to This Issue

244

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Executive Vice President, American
Telephone and Telegraph Company*

EDITORIAL COMMITTEE

B. McMILLAN, <i>Chairman</i>	H. R. HUNTLEY
A. J. BUSCH	F. R. LACK
A. C. DICKIESON	J. R. PIERCE
R. L. DIETZOLD	H. V. SCHMIDT
K. E. GOULD	C. E. SCHOOLEY
E. I. GREEN	G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

JANUARY 1956

NUMBER 1

Copyright 1956, American Telephone and Telegraph Company

Diffused Emitter and Base Silicon Transistors*

By M. TANENBAUM and D. E. THOMAS

(Manuscript received October 21, 1955)

Silicon n-p-n transistors have been made in which the base and emitter regions were produced by diffusing impurities from the vapor phase. Transistors with base layers 3.8×10^{-4} -cm thick have been made. The diffusion techniques and the processes for making electrical contact to the structures are described.

The electrical characteristics of a transistor with a maximum alpha of 0.97 and an alpha-cutoff of 120 mc/sec are presented. The manner in which some of the electrical parameters are determined by the distribution of the doping impurities is discussed. Design data for the diffused emitter, diffused base structure is calculated and compared with the measured characteristics.

INTRODUCTION

The necessity of thin base layers for high-frequency operation of transistors has long been apparent. One of the most appealing techniques for controlling the distribution of impurities in a semiconductor is the diffusion of the impurity into the solid semiconductor. The diffusion coefficients of Group III acceptors and Group V donors into germanium and silicon are sufficiently low at judiciously selected temperatures so

* A portion of the material of this paper was presented at the Semiconductor Device Conference of the Institute of Radio Engineers, Philadelphia, Pa., June, 1955.

that it is possible to envision transistors with base layer thicknesses of a micron and frequency response of several thousand megacycles per second.

A major deterrent to the application of diffusion to silicon transistor fabrication in the past was the drastic decrease in lifetime which generally occurs when silicon is heated to the high temperatures required for diffusion. There was also insufficient knowledge of the diffusion parameters to permit the preparation of structures with controlled layer thicknesses and desired dopings. Recently the investigations of C. S. Fuller and co-workers have produced detailed information concerning the diffusion of Group III and Group V elements in silicon. This information has made possible the controlled fabrication of transistors with base layers sufficiently thin that high alphas are obtained even though the lifetime has been reduced to a fraction of a microsecond. In a cooperative program with Fuller, diffusion structures were produced which have permitted the fabrication of transistors whose electrical behavior closely approximates the behavior anticipated from the design. This paper describes these techniques which have resulted in high alpha silicon transistors with alpha-cutoff of over 100 mc/sec.

1.0 FABRICATION OF THE TRANSISTORS

Fuller's work¹ has shown that in silicon the diffusion coefficient of a Group III acceptor is usually 10 to 100 times larger than that of the Group V donor in the same row in the periodic table at the same temperatures. These experiments were performed in evacuated silica tubes using the Group III and Group V elements as the source of diffusant. Under these conditions a particular steady state surface concentration of the diffusant is produced and the depth of diffusion is sensitive to this concentration as well as to the diffusion coefficient. The experiments show that the effective steady state surface concentration of the donor impurities produced under these conditions is ten to one hundred times greater than that of the acceptor impurities. Thus, by the simultaneous diffusion of selected donor and acceptor impurities into n-type silicon an n-p-n structure will result. The first n-layer forms because the surface concentration of the donor is greater than that of the acceptor. The p-layer is produced because the acceptor diffuses faster than the donor and gets ahead of it. The final n-region is simply the original background doping of the n-type silicon sample. It has been possible to produce n-p-n structures by the simultaneous diffusion of several combinations of donors and acceptors. Often, however, the diffusion coefficients and surface concentrations of the donors and acceptors are such that opti-

¹ C. S. Fuller, private communication.

mum layer thicknesses (see Sections 3 and 4) are not produced by simultaneous diffusion. In this case, one of the impurities is started ahead of the other in a prior diffusion, and then the other impurity is diffused in a second operation.

With the proper choice of diffusion temperatures and times it has been possible to make n-p-n structures with base layer thicknesses of 2×10^{-4} cm. The uniformity of the layers in a given specimen is better than ten per cent of the layer thickness. Fig. 1 illustrates the uniformity of the layers. This figure is an enlarged photograph of a view perpendicular to the surface of the specimen. A bevel which makes an angle of five degrees with the original surface has been polished on the specimen. This angle magnifies the layer thickness by 11.5. The layer is defined by an etchant which preferentially stains p-type silicon¹ and the width of the layer is measured with a calibrated microscope.

After diffusion the entire surface of the silicon wafer is covered with the diffused n- and p-type layers, see Fig. 2(a). Electrical contact must now be made to the three regions of the device. The base contact can be made by polishing a bevel on the specimen to expose and magnify the base layer and then alloying a lead to this region by the same tech-

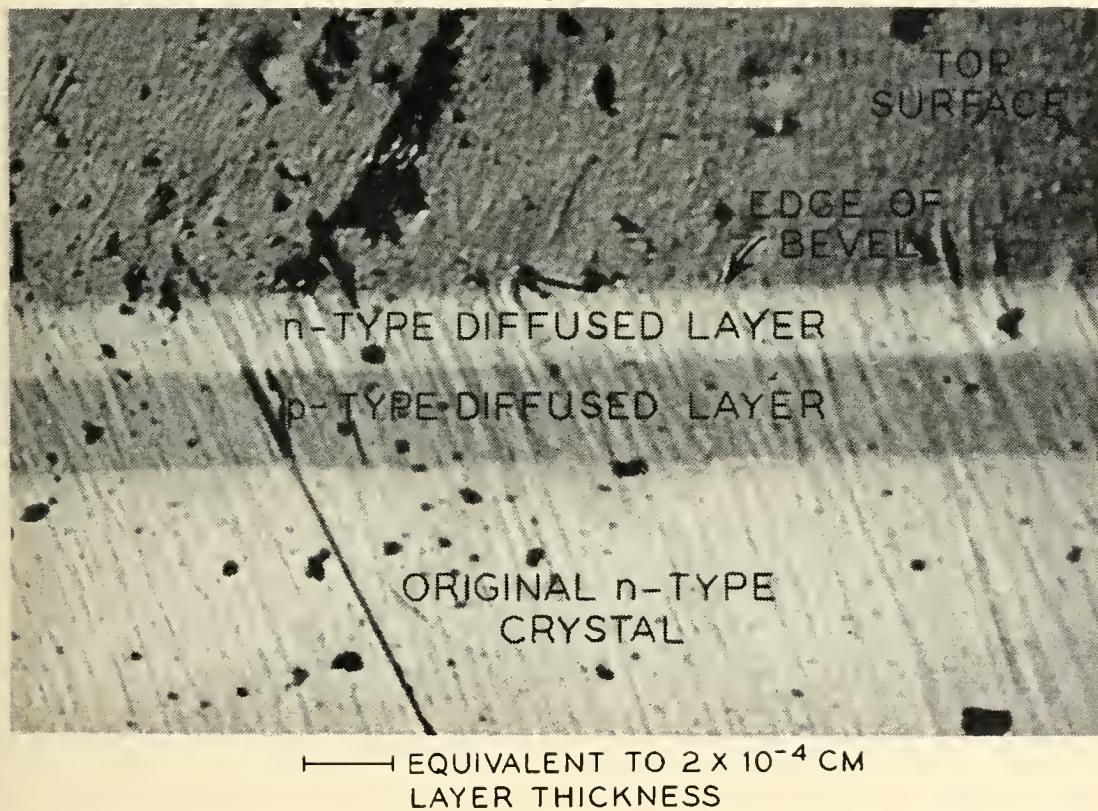


Fig. 1 — Angle section of a double diffused silicon wafer. The p-type center layer is approximately 2×10^{-4} cm thick.

niques employed in the fabrication of grown junction transistors, Fig. 2(b). However, a much simpler technique has been evolved. If the surface concentration of the donor diffusant is maintained below a certain critical value, it is possible to alloy an aluminum wire directly through the diffused n-type layer and thus make effective contact to the base layer, Fig. 2(e). Since the resistivity of the original silicon wafer is one to five ohm-cm, the aluminum will be rectifying to this region. It has been experimentally shown that if the surface concentration of the donor diffusant is less than the critical value mentioned above, the aluminum will also be rectifying to the diffused n-type region and the contact becomes merely an extension of the base layer. The n-layers produced by diffusing from elemental antimony are below the critical concentration and the direct aluminum alloying technique is feasible.

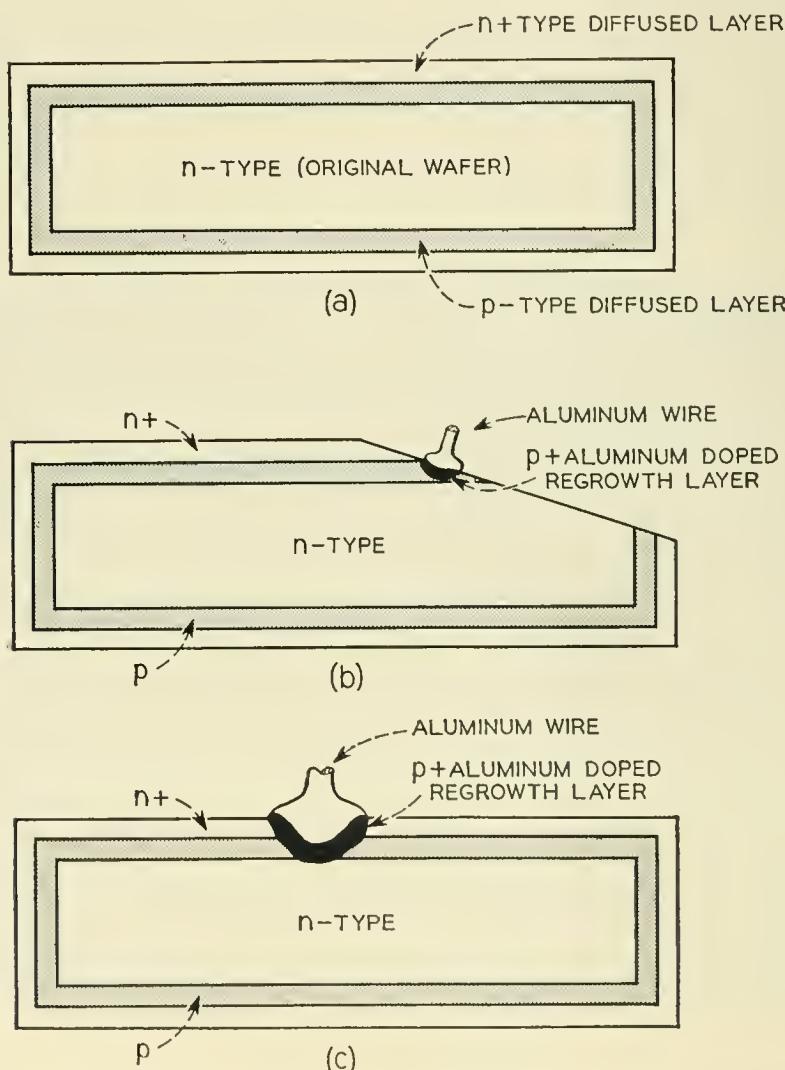


Fig. 2 — Schematic illustration of (a) double diffused n-p-n wafer, (b) angle section method of making base contact, and (c) direct alloying method of making base contact.

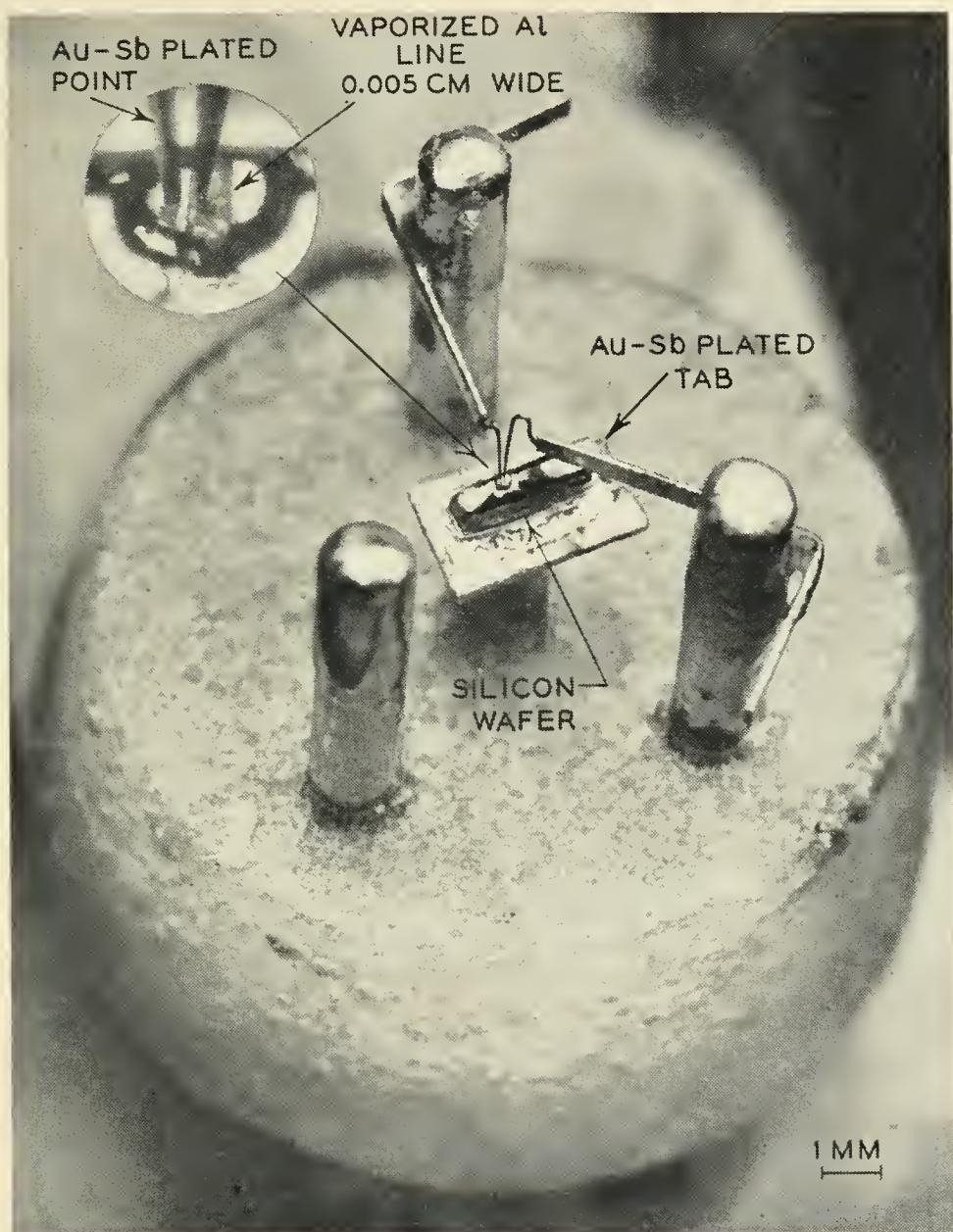


Fig. 3 — Mounted double diffused transistor.

Contact to the emitter layer is achieved by alloying a film of gold containing a small amount of antimony. Since this alloy will produce an n-type regrowth layer, it is only necessary to insure that the gold-antimony film does not alloy through the p-type base layer, thus shorting the emitter to the collector. This is controlled by limiting the amount of gold-antimony alloy which is available by using a thin evaporated film or by electroplating a thin film of gold-antimony alloy on an inert metal point and alloying this structure to the emitter layer.

Ohmic contact to the collector is produced by alloying the silicon wafer to an inert metal tab plated with a gold-antimony alloy.

The transistors whose characteristics are reported in this paper were prepared from 3 ohm-cm n-type silicon using antimony and aluminum as the diffusants. The base contact was produced by evaporating aluminum through a mask so that a line approximately 0.005×0.015 cm in lateral dimensions and 100,000 Å thick was formed on the surface. This aluminum line was alloyed through the emitter layer in a subsequent operation. The wafer was then alloyed onto the plated kovar tab. A small area approximately 0.015 cm in diameter was masked around the line and the wafer was etched to remove the unwanted layers. The unit was then mounted in a header. Electrical contact to the collector was made by soldering to the kovar tab. Contact to the base was made with a tungsten point pressure contact to the alloyed aluminum. Contact to the emitter was made by bringing a gold-antimony plated tungsten point into pressure contact with the emitter layer. The gold-antimony plate was then alloyed by passing a controlled electrical pulse between the plated point and the transistor collector lead. Fig. 3 is a photograph of a mounted unit.

2.0 ELECTRICAL CHARACTERISTICS

The frequency cutoffs of experimental double diffused silicon transistors fabricated as described above are an order of magnitude higher than the known cutoff frequencies of earlier silicon transistors. This is shown in Fig. 4 which gives the measured common base and common emitter current gains for one of these units as a function of frequency. The common base short-circuit current gain is seen to have a cutoff frequency of about 120 mc/sec.² The common emitter short-circuit current gain is shown on the same figure. The low-frequency current gain is better than thirty decibels and the cutoff frequency which is indicated by the frequency at which the gain is 3 db below its low-frequency value is 3 mc/sec. This is an exceptionally large common emitter bandwidth for a thirty db common emitter current gain and is of the same order of magnitude as that obtained with the highest frequency germanium transistors (e.g., p-n-i-p or tetrode) which had been made prior to the diffused base germanium transistor.³

² The increase in common base current gain above unity (indicated by current gain in decibels being positive) in the vicinity of 50 mc/sec is caused by a reactance gain error in the common base measurement. This error is caused by a combination of the emitter to ground parasitic capacitance and the positive reactance component of the transistor input impedance resulting from phase shift in the alpha current gain.

³ C. A. Lee, A High-Frequency Diffused Base Germanium Transistor, see page 23.

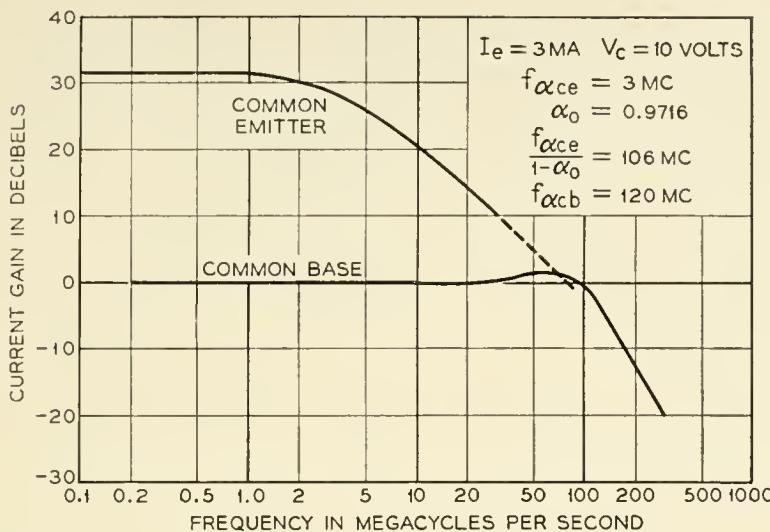


Fig. 4 — Short-circuit current gain of a double diffused silicon n-p-n transistor as a function of frequency in the common emitter and common base connections.

Fig. 5 shows a high-frequency lumped constant equivalent circuit for the double diffused silicon transistor whose current gain cutoff characteristic is shown in Fig. 4. External parasitic capacitances have been omitted from the circuit. The configuration is the conventional one for junction transistors with two exceptions. A series resistance r'_e has been added in the emitter circuit to account for contact resistance resulting from the fact that the present emitter point contacts are not perfectly ohmic. A second resistance r'_c has been added in the collector circuit to account for the ohmic resistance of the n-type silicon between the collector terminal and the effective collector junction. This resistance exists in all junction transistors but in larger area low frequency junction transistors its effect on alpha-cutoff is sufficiently small so that it has been ignored in equivalent circuits of these devices. The collector RC

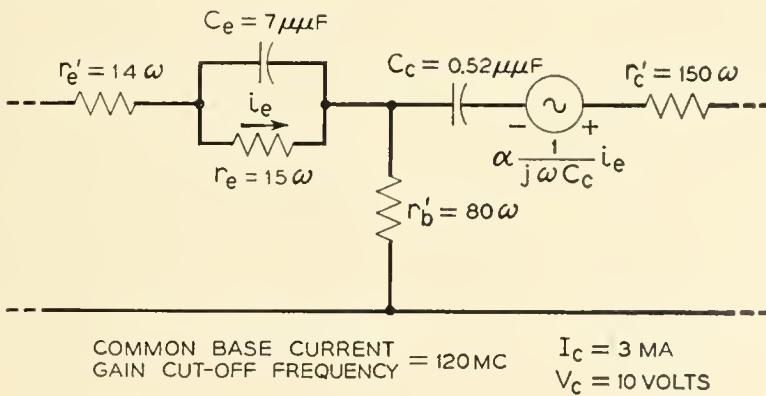


Fig. 5 — High-frequency lumped constant equivalent circuit for a double diffused silicon n-p-n transistor.

cutoff caused by the collector capacitance and the combined collector body resistance and base resistance is an order of magnitude higher than the measured alpha cutoff frequency and therefore is not too serious in impairing the very high-frequency performance of the transistor. This is due to the low capacitance of the collector junction which is seen to be approximately 0.5 mmf at 10 volts collector voltage. The base resistance of this transistor is less than 100 ohms which is quite low and compares very favorably with the best low frequency transistors reported previously.

The low-frequency characteristics of the double diffused silicon transistor are very similar to those of other junction transistors. This is illustrated in Fig. 6 where the static collector characteristics of one of these transistors are given. At zero emitter current the collector current is too small to be seen on the scale of this figure. The collector current

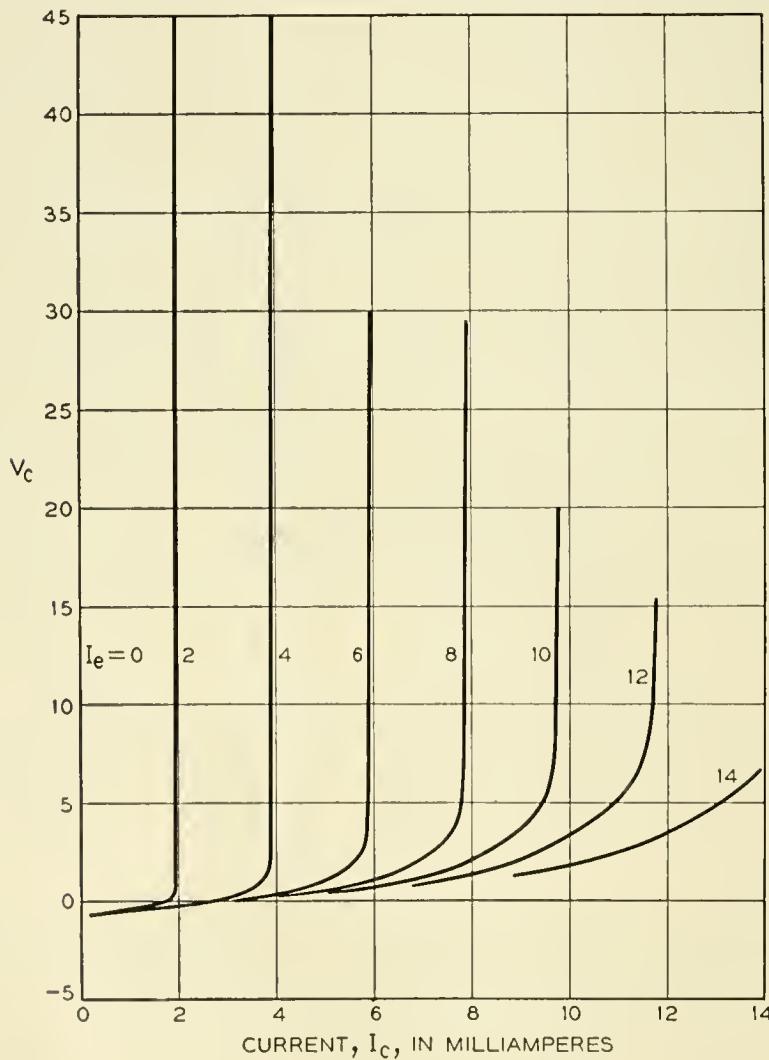


Fig. 6 — Collector characteristics of a double diffused silicon n-p-n transistor.

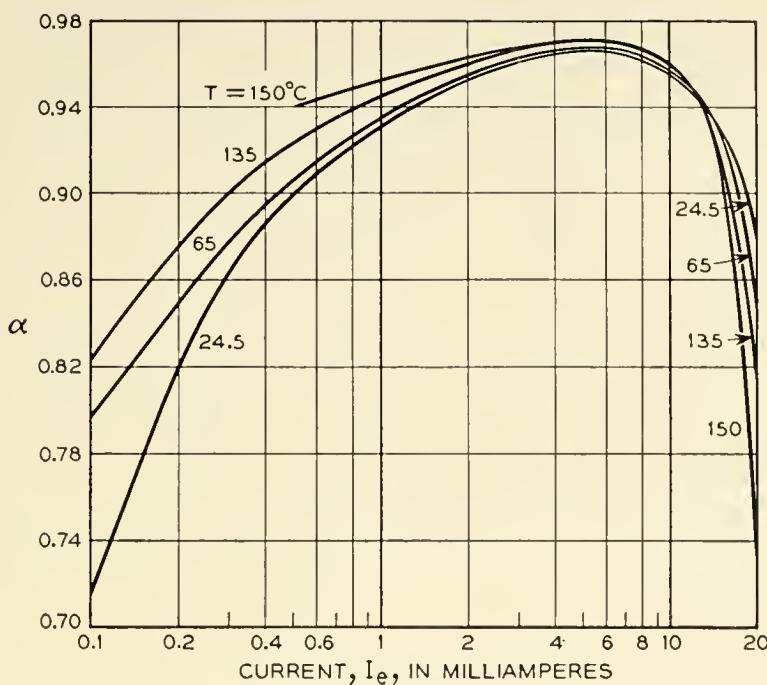


Fig. 7 — Alpha as a function of emitter current and temperature for a double diffused silicon n-p-n transistor.

under this condition does not truly saturate but collector junction resistance is very high. Collector junction resistances of 50 megohms at reverse biases of 50 volts are common.

The continuous power dissipation permissible with these units is also shown in Fig. 6. The figure shows dissipation of 200 milliwatts and the units have been operated at 400 milliwatts without damage. As illustrated in Fig. 3 no special provision has been made for power dissipation and it would appear from the performance obtained to date that powers of a few watts could be handled by these units with relatively minor provisions for heat dissipation. However, it can also be seen from Fig. 6 that at low collector voltages alpha decreases rapidly as the emitter current is increased. The transistor is, therefore, non-linear in this range of emitter currents and collector voltages. In many applications, this non-linearity may limit the operating range of the device to values below those which would be permissible from the point of view of continuous power dissipation.

Fig. 7 gives the magnitude of alpha as a function of emitter current for a fixed collector voltage of 10 volts and a number of ambient temperatures. These curves are presented to illustrate the stability of the parameters of the double diffused silicon transistor at increased ambient temperatures. Over the range from 1 to 15 milliamperes emitter current and 25°C to 150°C ambient temperature, alpha is seen to change only

by approximately 2 per cent. This amounts to only 150 parts per million change in alpha per degree centigrade change in ambient temperature.

The decrease in alpha at low emitter currents shown in Fig. 7 has been observed in every double diffused silicon transistor which has been made to date. Although this effect is not completely understood at present it could be caused by recombination centers in the base layer that can be saturated at high injection levels. Such saturation would result in an increase in effective lifetime and a corresponding increase in alpha. The large increase in alpha with temperature at low emitter currents is consistent with this proposal. It has also been observed that shining a strong light on the transistor will produce an appreciable increase in alpha at low emitter currents but has little effect at high emitter currents. A strong light would also be expected to saturate recombination centers which are active at low emitter currents and this behavior is also consistent with the above proposal.

3.0 DISCUSSION OF THE TRANSISTOR STRUCTURE

Although the low frequency electrical characteristics of the double diffused silicon transistor which are presented in Section 2 are quite similar to those usually obtained in junction transistors, the structure of the double diffused transistor is sufficiently different from that of the grown junction or alloy transistor that a discussion of some design principles is warranted. This section is devoted to a general discussion of the factors which determine the electrical characteristics of the transistors. In Section 4 the general ideas of Section 3 are applied in a more specialized fashion to the double diffused structure and a detailed calculation of electrical parameters is presented.

One essential difference between the double diffused transistor and grown junction or alloy transistors arises from the manner in which the impurities are distributed in the three active regions. In the ideal case of a double-doped grown junction transistor or an alloy transistor the concentration of impurities in a given region is essentially uniform and the transition from one conductivity type to another at the emitter and collector junctions is abrupt giving rise to step junctions. On the other hand in the double diffused structure the distribution of impurities is more closely described by the error function complement and the emitter and collector junctions are graded. These differences can have an appreciable influence on the electrical behavior of the transistors.

Fig. 8(a) shows the probable distribution of donor impurities, N_D , and acceptor impurities, N_A , in a double diffused n-p-n. Fig. 8(b) is a

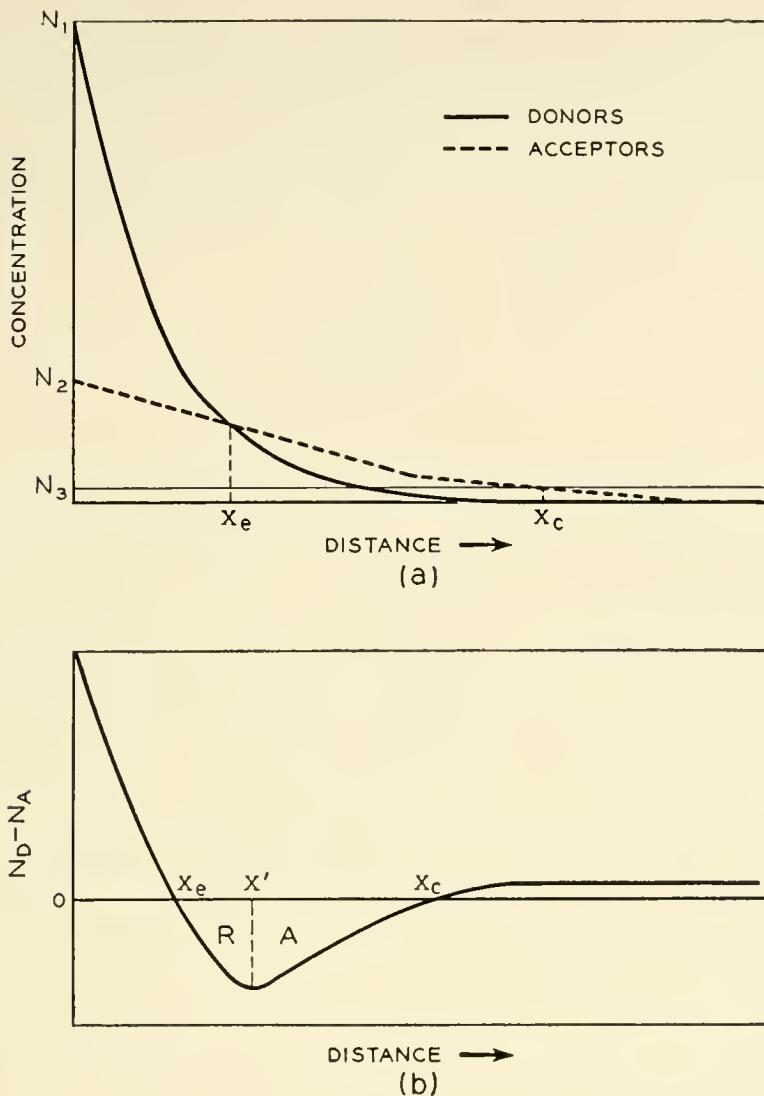


Fig. 8 — Diagrammatic representation of (a) donor and acceptor distributions and (b) uncompensated impurity distribution in a double diffused n-p-n transistor.

plot of $N_D - N_A$ which would result from the distribution in Fig. 8(a). Krömer⁴ has shown that a nonuniform distribution of impurities in a semiconductor will produce electric fields which can influence the flow of electrons and holes. For example, in the base region the fields between the emitter junction, x_e , and the minimum in the $N_D - N_A$ curve, x' , will retard the flow of electrons toward the collector while the fields between this minimum and the collector junction, x_c , will accelerate the flow of electrons toward the collector. These base layer fields will affect the transit time of minority carriers across the base and thus contribute

⁴ H. Krömer, On Diffusion and Drift Transistor Theory I, II, III, Archiv. der Electr. Übertragung, **8**, pp. 223-228, pp. 363-369, pp. 499-504, 1954.

to the frequency response of the transistor. In addition the base resistance will be dependent on the distribution of both diffusants. These three factors are discussed in detail below.

Moll and Ross⁵ have determined that the minority current, I_m , that will flow into the base region of a transistor if the base is doped in a non-uniform manner is given by

$$I_m = \frac{n_i^2 q D_m}{\int_b N(x) dx} e^{qV/kT} \quad (3.1)$$

where n_i is the carrier concentration in intrinsic material, q is the electronic charge, V is the applied voltage, D_m is the diffusion coefficient of the minority carriers, and the integral represents the total number of uncompensated impurities in the base. The primary assumptions in this derivation are (1) planar junctions, (2) no recombination in the base region, and (3) a boundary condition at the collector junction that the minority carrier density at this point equals zero. It is also assumed that the minority carrier concentration in the base region just adjacent to the emitter junction is equal to the equilibrium minority carrier density at this point multiplied by the Boltzman factor $\exp(qV/kT)$. It is of special interest to note that I_m depends only on the total number of uncompensated impurities in the base and not on the manner in which they are distributed.

In the double diffused transistor, it has been convenient from the point of ease of fabrication to make the emitter layer approximately the same thickness as the base layer. It has been observed that heating silicon to high temperatures degrades the lifetime of n- and p-type silicon in a similar manner.⁶ Both base and emitter layers have experienced the same heat treatment and to a first approximation it can be assumed that the lifetime in the two regions will be essentially the same. Thus assumptions (1) and (2) should also apply to current flow from base to emitter. If we assume that the surface recombination velocity at the free surface of the emitter is infinite, then this imposes a boundary condition at this side of the emitter which under conditions of forward bias on the emitter is equivalent to assumption (3). Thus an equation of the form of (3.1) should also give the minority current flow from base to emitter. Since the emitter efficiency, γ , is given by

⁵ J. L. Moll and I. M. Ross, The Dependence of Transistor Parameters on the Distribution of Base Layer Resistivity, Proc. I.R.E. in press.

⁶ G. Bemski, private communication.

$$\gamma = \frac{I_m(\text{emitter to base})}{I_m(\text{emitter to base}) + I_m(\text{base to emitter})}$$

proper substitution of (3.1) will give the emitter efficiency of the double diffused n-p-n transistor,

$$\gamma = \frac{1}{1 + \frac{D_p}{D_n} \frac{\int_b (N_A - N_D) dx}{\int_e (N_D - N_A) dx}} \quad (3.2)$$

In (3.2), D_p is the diffusion coefficient of holes in the emitter, D_n is the diffusion coefficient of electrons in the base and the ratio of integrals is the ratio of total uncompensated doping in the base to that in the emitter.

A calculation of transit time is more difficult. Krömer⁴ has studied the case of an aiding field which reduces transit time of minority carriers across the base region and thus increases frequency response. In the double diffused transistor the situation is more complex. Near the emitter side of the base region the field is retarding (Region R, see Fig. 8) and becomes aiding (Region A) only after the base region doping reaches a maximum. The case of retarding fields has been studied by Lee³ and by Moll.⁷ At present, the case for a base region containing both types of fields has not been solved. However, at the present state of knowledge some speculations about transit time can be made.

The two factors of primary importance are the magnitude of the built-in fields and the distance over which they extend. In the double diffused transistor, the widths of regions R and A are determined by the surface concentrations and diffusion coefficients of the diffusants. It can be shown by numerical computation⁷ that if region R constitutes no more than 30–40 per cent of the entire base layer width, then the overall effect of the built-in fields will be to aid the transport of minority carriers and to produce a reduction in transit time. In addition the absolute magnitude of region R is important. If the point x' should occur within an effective Debye length from the emitter junction, i.e., if x' is located in the space charge region associated with the emitter junction, then the retarding fields can be neglected.

The base resistance can also be calculated from surface concentrations and diffusion coefficients of the impurities. This is done by considering the base layer as a conducting sheet and determining the sheet con-

⁷ J. L. Moll, private communication.

ductivity from the total number of uncompensated impurities per square centimeter of sheet and the appropriate mobility weighted to account for impurity scattering.

4.0 CALCULATION OF DESIGN PARAMETERS

To calculate the parameters which determine emitter efficiency, transit time, and base resistance it is assumed that the distribution of uncompensated impurities is given by

$$N(x) = N_1 \operatorname{erfc} \frac{x}{L_1} - N_2 \operatorname{erfc} \frac{x}{L_2} + N_3 \quad (4.1)$$

where N_1 and N_2 are the surface concentrations of the emitter and base impurity diffusants respectively, L_1 and L_2 are their respective diffusion lengths, and N_3 is the original doping of the semiconductor into which the impurities are diffused. The impurity diffusion lengths are defined as

$$L_1 = 2 \sqrt{D_1 t_1} \quad \text{and} \quad L_2 = 2 \sqrt{D_2 t_2} \quad (4.2)$$

where the D 's are the respective diffusion coefficients and the t 's are the diffusion times.

Equation (4.1) can be reduced to

$$\Gamma(\xi) = \Gamma_1 \operatorname{erfc} \xi - \Gamma_2 \operatorname{erfc} \lambda \xi + 1 \quad (4.3)$$

where

$$\Gamma(\xi) = \frac{N(\xi)}{N_3}; \quad \Gamma_1 = \frac{N_1}{N_3}; \quad \Gamma_2 = \frac{N_2}{N_3}; \quad \xi = \frac{x}{L_1}; \quad \lambda = \frac{L_1}{L_2}$$

For cases of interest here, $\Gamma(\xi)$ will be zero at two points, α and β , and will have one minimum at ξ' . In the transistor structure the emitter junction occurs at $\xi = \alpha$ and the collector junction occurs at $\xi = \beta$. Thus the base width is determined by $\beta - \alpha$. The extent of aiding and retarding fields in the base is determined by ξ' . The integral of (4.3) from 0 to α , I_1 , and from α to β , I_2 , are the integrals of interest in (3.2) and thus determine emitter efficiency. In addition I_2 is the integral from which base resistance can be calculated.

The calculations which follow apply only for values of Γ_1/Γ_2 and Γ_2 greater than ten. Some of the simplifying assumptions which are made will not apply at lower values of these parameters where the distribution of both diffusants as well as the background doping affect the structure in all three regions of the device.

4.1 Base Width

From Fig. 8 and (4.3) it can be seen that for $\Gamma_2 \geq 10$, α is essentially independent of Γ_2 and is primarily a function of Γ_1/Γ_2 and λ . Fig. 9 is a plot of α versus Γ_1/Γ_2 with λ as the parameter. The particular plot is for $\Gamma_2 = 10^4$. Although as stated α is essentially independent of Γ_2 , at lower values of Γ_2 , α may not exist for the larger values of λ , i.e., the p-layer does not form.

In the same manner, it can be seen that β is essentially independent of Γ_1/Γ_2 and is a function only of Γ_2 and λ . Fig. 10 is a plot of β versus Γ_2 with λ as a parameter. This plot is for $\Gamma_1/\Gamma_2 = 10$ and at larger Γ_1/Γ_2 , β may not exist at large λ .

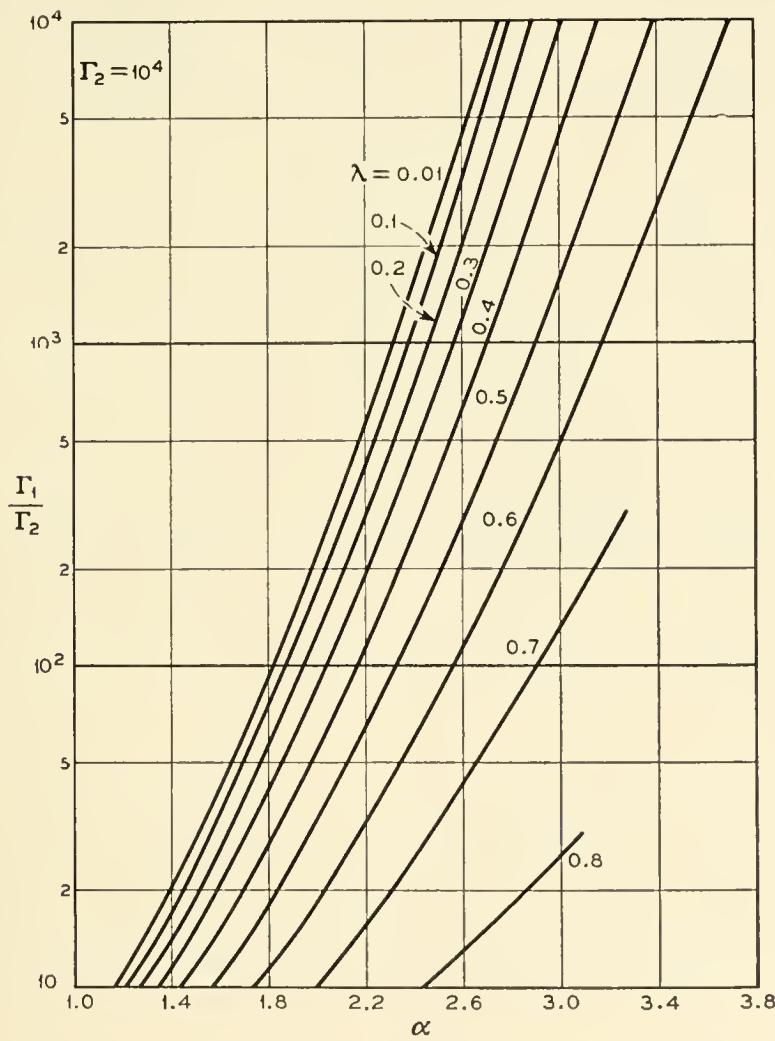


Fig. 9 — Emitter layer thickness (in reduced units) as a function of the ratio of the surface concentrations of the diffusing impurities (Γ_1/Γ_2) and the ratio of their diffusion lengths (λ).

The base width

$$w = \beta - \alpha$$

can be obtained from Figs. 9 and 10. α , β and w can be converted to centimeters by multiplying by the appropriate value of L_1 .

4.2 Emitter Efficiency

With the limits α and β determined above, the integrals I_1 and I_2 can be calculated. Examination of the integrals shows that I_1 is closely proportional to Γ_1/Γ_2 and also to Γ_2 . On the other hand I_2 is closely proportional to Γ_2 and essentially independent of Γ_1/Γ_2 . Thus, the ratio of I_2/I_1 which determines γ depends primarily on Γ_1/Γ_2 . Fig. 11 is a plot of the constant I_2/I_1 contours in the $\Gamma_1/\Gamma_2 - \lambda$ plane for I_2/I_1 in the range from -1.0 to -0.01 . The graph is for $\Gamma_2 = 10^4$. Since from (3.2)

$$\gamma = \frac{1}{1 - \frac{D_p}{D_n} \frac{I_2}{I_1}} \quad (4.4)$$

for an n-p-n transistor, and assuming $D_p/D_n = \frac{1}{3}$ for silicon, then

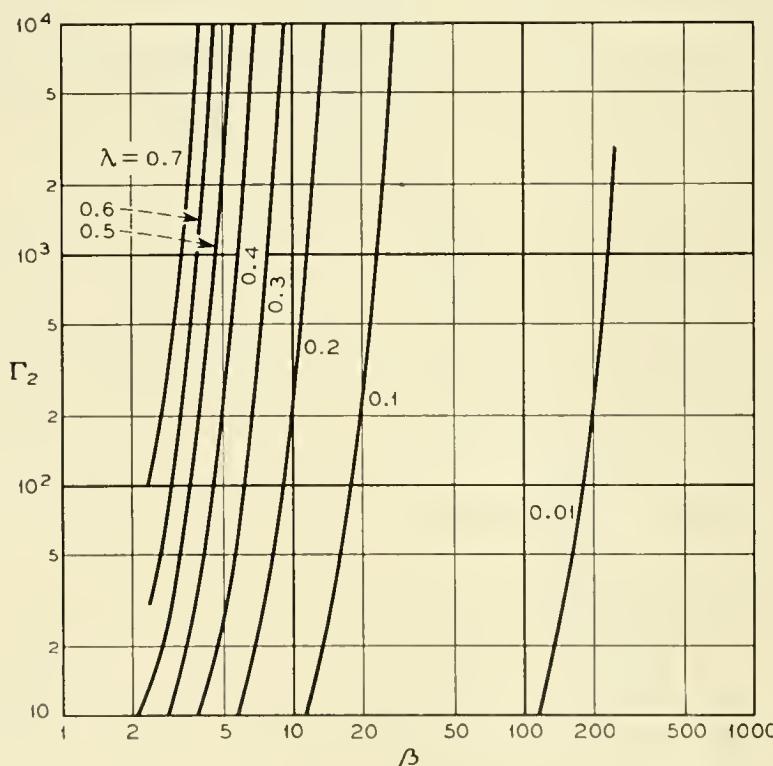


Fig. 10 — Collector junction depth (in reduced units) as a function of the surface concentration (in reduced units) of the diffusant which determines the conductivity type of the base layer (Γ_2) and the ratio of the diffusion lengths (λ) of the two diffusing impurities.

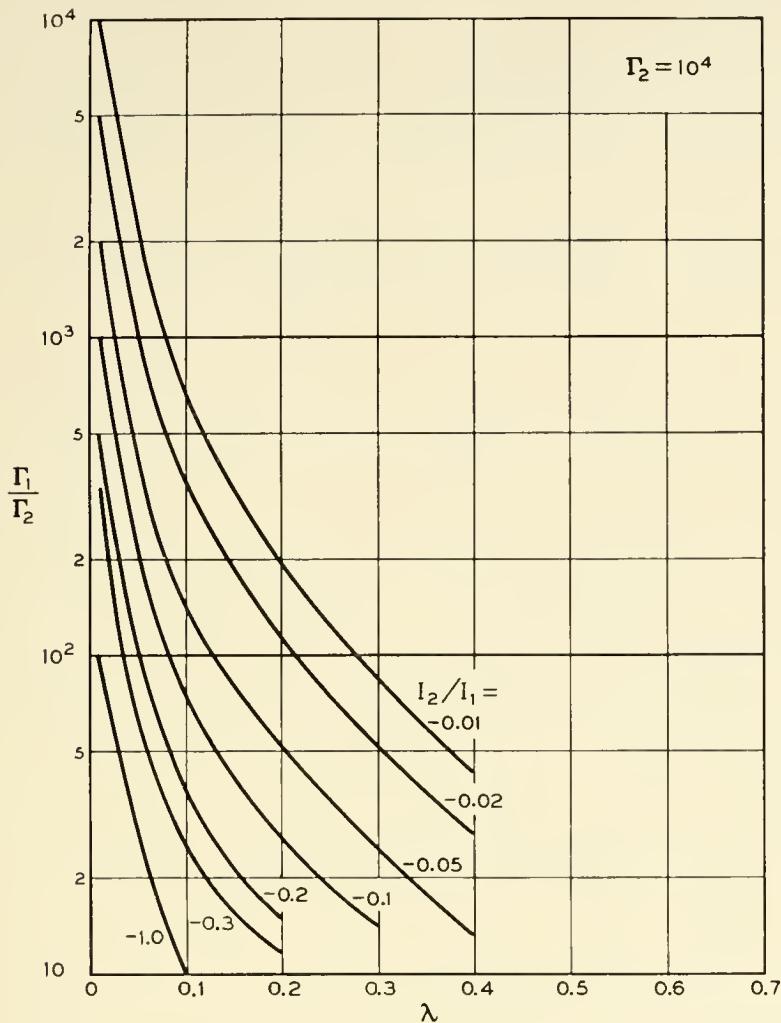


Fig. 11 — Dependence of emitter efficiency upon diffusant surface concentrations and diffusion lengths. The lines of constant I_2/I_1 are essentially lines of constant emitter efficiency. The ordinate is the ratio of surface concentrations of the two diffusants and the abscissa is the ratio of their diffusion lengths.

$I_2/I_1 = -1.0$ corresponds to a γ of 0.75 and $I_2/I_1 = -0.01$ corresponds to a γ of 0.997.

4.3 Base Resistance

It was indicated above that I_2 depends principally on Γ_2 and λ . Fig. 12 is a plot of the constant I_2 contours in the $\Gamma_2 - \lambda$ plane for I_2 in the range from -10^4 to -10 . The graph is for $\Gamma_1/\Gamma_2 = 10$. The base layer sheet conductivity, g_b , can be calculated from these data as

$$g_b = -q\bar{\mu}I_2L_1N_3 \quad (4.5)$$

where q , L_1 and N_3 are as defined above and $\bar{\mu}$ is a mobility properly weighted to account for impurity scattering in the non-uniformly doped base region. The units of g_b are mhos per square.

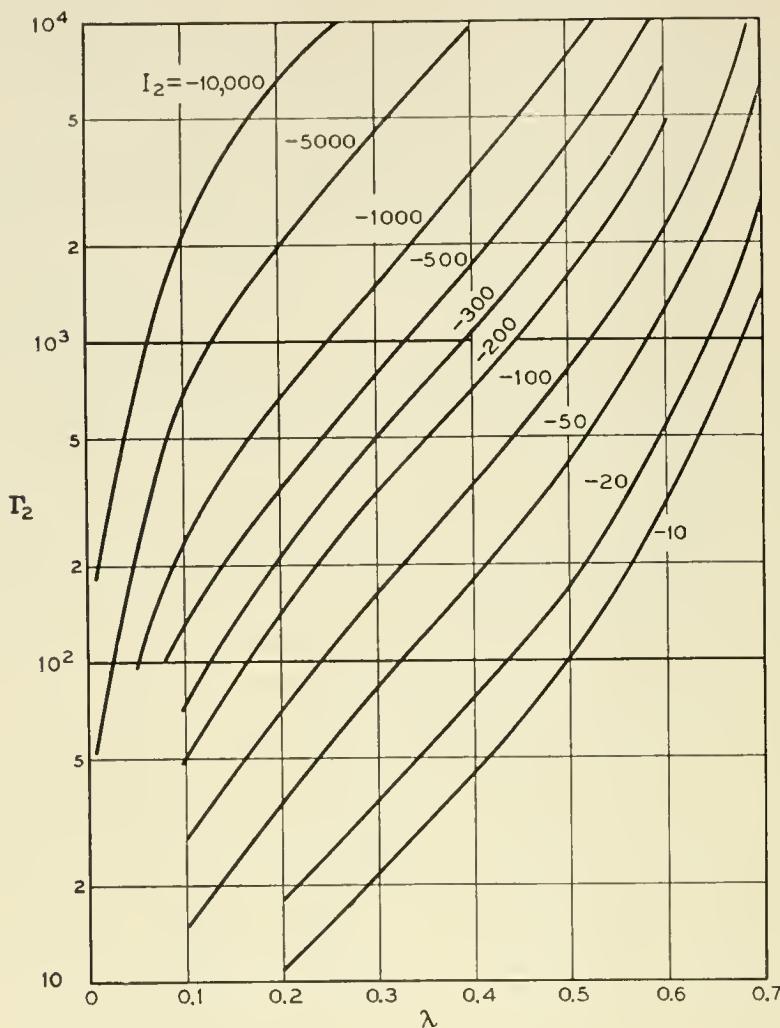


Fig. 12—Dependence of base layer sheet conductivity on diffusant surface concentrations and diffusion lengths. The lines of constant I_2 are essentially lines of constant base sheet conductivity. The ordinate is the surface concentration (in reduced units) of the diffusant which determines the conductivity type of the base layer and the abscissa is the ratio of the diffusion lengths of the two diffusing impurities.

4.4 Transit Time

With a knowledge of where the minimum value, ξ' , of (4.3) occurs, it is possible to calculate over what fraction of the base width the fields are retarding. The interesting quantity here is

$$\Delta R = \frac{\xi' - \alpha}{\beta - \alpha}$$

ξ' is a function of Γ_1/Γ_2 and λ and varies only very slowly with Γ_1/Γ_2 . α is also a function of Γ_1/Γ_2 and λ and varies only slowly with Γ_1/Γ_2 . The most rapidly changing part of ΔR is β which depends primarily on Γ_2 as noted above. Fig. 13 is a plot of the constant ΔR contours in the $\Gamma_2 - \lambda$ plane for values of ΔR in the range 0.1 to 0.3. This graph is

for data with $\Gamma_1/\Gamma_2 = 10$. As Γ_1/Γ_2 increases at constant Γ_2 and λ , ΔR decreases slightly. At $\Gamma_1/\Gamma_2 = 10^4$, the average change in ΔR is a decrease of about 25 per cent for constant Γ_2 and λ when $\Delta R \leq 0.3$. The error is larger for values of ΔR greater than 0.3. It was noted above that when ΔR becomes greater than 0.3, the retarding fields become dominant. Therefore, this region is of slight interest in the design of a high frequency transistor.

4.5 A Sample Design

By superimposing Figs. 11, 12 and 13 the ranges of Γ_2 , Γ_1/Γ_2 and λ which are consistent with desired values of γ , g_b and ΔR can be deter-

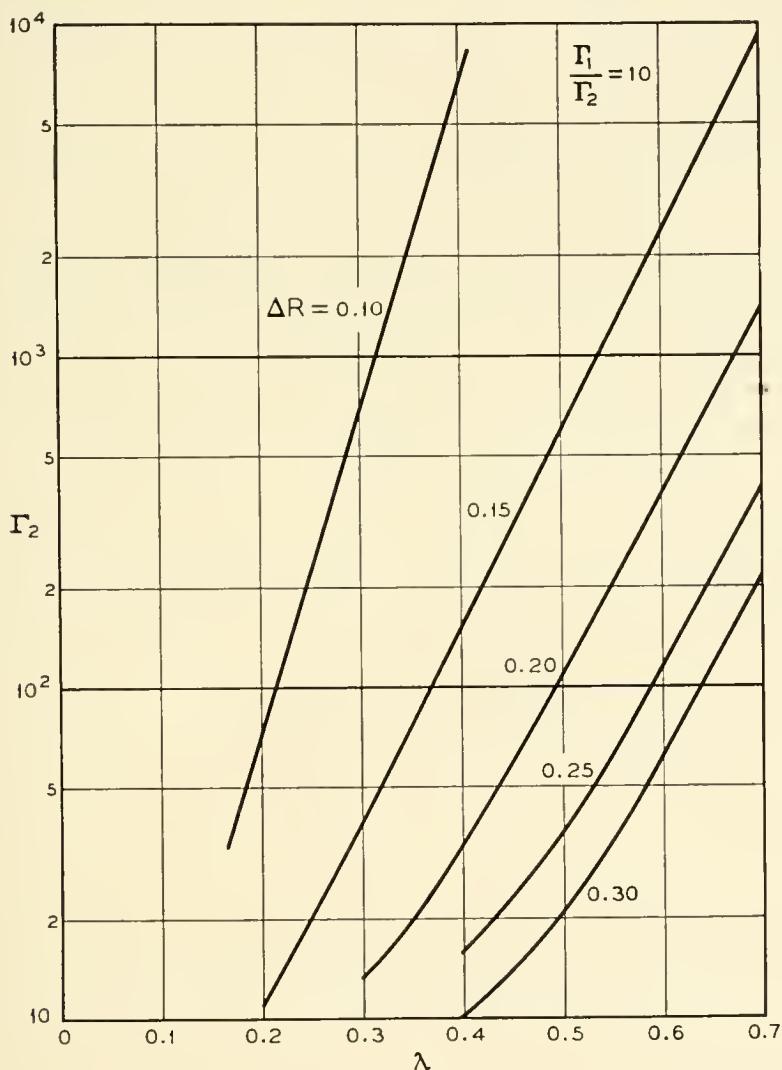


Fig. 13 — Dependence of the built-in field distribution on concentrations and diffusion lengths. The lines of constant ΔR indicate the fraction of the base layer thickness over which built-in fields are retarding. The ordinate is the surface concentration (in reduced units) of the diffusant which determines the conductivity type of the base layer and the abscissa is the ratio of the diffusion lengths of the two diffusing impurities.

mined by the area enclosed by the specified contour lines. It is also possible to compare the measured parameters of a specific device and observe how closely they agree with what is predicted from the estimated concentrations and diffusion coefficients. This is done below for the transistor described in Sections 1 and 2.

The comparison is complicated by the fact that the exact values of the surface concentrations and diffusion coefficients are not known precisely enough at present to permit an accurate evaluation of the design theory. However, the following values of concentrations and diffusion coefficients are thought to be realistic for this transistor.¹

$$N_1 = 5 \times 10^{18} \quad D_1 = 3 \times 10^{-12} \quad t_1 = 5.7 \times 10^3$$

$$N_2 = 4 \times 10^{17} \quad D_2 = 2.5 \times 10^{-11} \quad t_2 = 1.2 \times 10^3$$

$$N_3 = 10^{15}$$

From these values it is seen that

$$\Gamma_1/\Gamma_2 = 12.5; \quad \Gamma_2 = 400; \quad \lambda = 0.6$$

From Fig. 9, $\alpha = 1.9$ and from Fig. 10, $\beta = 3.6$ and therefore $w = 1.7$. Measurement of the emitter and base layer dimensions showed that these layers were approximately the same thickness which was 3.8×10^{-4} cm. Thus the measured ratio of emitter width to base width of unity is in good agreement with the value of 1.1 predicted from the assumed concentrations and diffusion coefficients.

From Fig. 11, $I_2/I_1 \approx -0.01$. If this value is substituted into (4.4), $\gamma = 0.997$. This compares with a measured maximum alpha of 0.972.

From Fig. 12, $I_2 = -15$. Assuming an average hole mobility of 350 $\text{cm}^2/\text{volt. sec.}$ and evaluating L_1 from the measured emitter thickness and the calculated α , (4.5) gives a value of $g_b = 1.7 \times 10^{-4}$ mhos per square. The geometry of the emitter and base contacts as shown in Fig. 3 makes it difficult to calculate the effective base resistance from the sheet conductivity even at very small emitter currents. In addition at the very high injection levels at which these transistors are operated the calculation of effective base resistance becomes very difficult. However, from the geometry it would be expected that the effective base resistance would be no greater than 0.1 of the sheet resistivity or 600 ohms. This is about seven times larger than the measured value of 80 ohms reported in Section 2.

From Fig. 13, ΔR is approximately 0.20. Thus there should be an overall aiding effect of the built-in fields. In addition the impurity gradient at the emitter junction is believed to be approximately $10^{21}/\text{cm}^4$ and the

space charge associated with this gradient will extend approximately 2×10^{-5} cm into the base region. The base thickness over which retarding fields extend is ΔR times the base width or 7.6×10^{-5} cm. Thus the first quarter of region R will be space charge and can be neglected.

The frequency cutoff from pure diffusion transit is given by

$$f_\alpha = \frac{2.43D}{2\pi W^2} \quad (4.6)$$

where W is the measured base layer thickness. Assuming $D = 25 \text{ cm}^2/\text{sec}$ for electrons in the base region, $f_\alpha = 67 \text{ mc/sec}$. Since the measured cutoff was 120 mc/sec, the predicted aiding effect of the built-in field is evidently present.

These computations illustrate how the measured electrical parameters can be used to check the values of the surface concentrations and diffusion coefficients. Conversely knowledge of the concentrations and diffusion coefficients aid in the design of devices which will have prescribed electrical parameters. The agreement in the case of the transistor described above is not perfect and indicates errors in the proposed values of the concentrations and diffusion coefficients. However, it is sufficiently close to be encouraging and indicate the value of the calculations.

The discussion of design has been limited to a very few of the important parameters. Junction capacitances, emitter and collector resistances are among the other important characteristics which have been omitted here. Presumably all of these quantities can be calculated if the detailed structure of the device is known and the structure should be susceptible to the type of analysis used above. Another fact, which has been ignored, is that these transistors were operated at high injection levels and a low level analysis of electrical parameters was used. All of these other factors must be considered for a detailed understanding of the device. The object of this last section has been to indicate one path which the more detailed analysis might take.

5.0 CONCLUSIONS

By means of multiple diffusion, it has been possible to produce silicon transistors with alpha-cutoff above 100 mc/sec. Refinements of the described techniques offer the possibility of even higher frequency performance. These transistors show the other advantages expected from silicon such as low saturation currents and satisfactory operation at high temperatures.

The structure of the double diffused transistor is susceptible to design

analysis in a fashion similar to that which has been applied to other junction transistors. The non-uniform distribution of impurities produces significant electrical effects which can be controlled to enhance appreciably the high-frequency behavior of the devices.

The extreme control inherent in the use of diffusion to distribute impurities in a semiconductor structure suggests that this technique will become one of the most valuable in the fabrication of semiconductor devices.

ACKNOWLEDGEMENT

The authors are indebted to several people who contributed to the work described in this paper. In particular, the double diffused silicon from which the transistors were prepared was supplied by C. S. Fuller and J. A. Ditzenberger. The data on diffusion coefficients and concentrations were also obtained by them.

P. W. Foy and G. Kaminsky assisted in the fabrication and mounting of the transistors and J. M. Klein aided in the electrical characterization. The computations of the various solutions of the diffusion equation, (4.3), were performed by Francis Maier. In addition many valuable discussions with C. A. Lee, G. Weinreich, J. L. Moll, and G. C. Dacey helped formulate many of the ideas presented herein.

A High-Frequency Diffused Base Germanium Transistor

By CHARLES A. LEE

(Manuscript received November 15, 1955)

Techniques of impurity diffusion and alloying have been developed which make possible the construction of p-n-p junction transistors utilizing a diffused surface layer as a base region. An important feature is the high degree of dimensional control obtainable. Diffusion has the advantages of being able to produce uniform large area junctions which may be utilized in high power devices, and very thin surface layers which may be utilized in high-frequency devices.

Transistors have been made in germanium which typically have alphas of 0.98 and alpha-cutoff frequencies of 500 mc/s. The fabrication, electrical characterization, and design considerations of these transistors are discussed.

INTRODUCTION

Recent work^{1, 2} concerning diffusion of impurities into germanium and silicon prompted the suggestion³ that the dimensional control inherent in these processes be utilized to make high-frequency transistors.

One of the critical dimensions of junction transistors, which in many cases seriously restricts their upper frequency limit of operation, is the thickness of the base region. A considerable advance in transistor properties can be accomplished if it is possible to reduce this dimension one or two orders of magnitude. The diffusion constants of ordinary donors and acceptors in germanium are such that, within realizable temperatures and times, the depth of diffused surface layers may be as small as 10^{-6} cm. Already in the present works layers slightly less than 1 micron (10^{-4} cm) thick have been made and utilized in transistors. Moreover, the times and temperatures required to produce 1 micron surface layers permit good control of the depth of penetration and the concentration of the diffusant in the surface layer with techniques described below.

If one considers making a transistor whose base region consists of such

a diffused surface layer, several problems become immediately apparent:

(1) Control of body resistivity and lifetime during the diffusion heating cycle.

(2) Control of the surface concentration of the diffusant.

(3) Making an emitter on the surface of a thin diffused layer and controlling the depth of penetration.

(4) Making an ohmic base contact to the diffused surface layer.

One approach to the solution of these problems in germanium which has enabled us to make transistors with alpha-cutoff frequencies in excess of 500 mc/sec is described in the main body of the paper.

An important characteristic feature of the diffusion technique is that it produces an impurity gradient in the base region of the transistor. This impurity gradient produces a "built-in" electric field in such a direction as to aid the transport of minority carriers from emitter to collector. Such a drift field may considerably enhance the frequency response of a transistor for given physical dimensions.⁴

The capabilities of these new techniques are only partially realized by their application to the making of high frequency transistors, and even in this field their potential has not been completely explored. For example, with these techniques applied to making a p-n-i-p structure the possibility of constructing transistor amplifiers with usable gain at frequencies in excess of 1,000 mc/sec now seems feasible.

DESCRIPTION OF TRANSISTOR FABRICATION AND PHYSICAL CHARACTERISTICS

As starting material for a p-n-p structure, p-type germanium of 0.8 ohm-cm resistivity was used. From the single crystal ingot rectangular bars were cut and then lapped and polished to the approximate dimensions: 200 × 60 × 15 mils. After a slight etch, the bars were washed in deionized water and placed in a vacuum oven for the diffusion of an n-type impurity into the surface. The vacuum oven consisted of a small molybdenum capsule heated by radiation from a tungsten coil and surrounded by suitable radiation shields made also of molybdenum. The capsule could be baked out at about 1,900°C in order that impurities detrimental to the electrical characteristics of the germanium be evaporated to sufficiently low levels.⁵

As a source of n-type impurity to be placed with the p-type bars in the molybdenum oven, arsenic doped germanium was used. The relatively high vapor pressure of the arsenic was reduced to a desirable range (about 10^{-4} mm of Hg) by diluting it in germanium. The use of germanium eliminated any additional problems of contamination by the

dilutant, and provided a convenient means of determining the degree of dilution by a measurement of the conductivity. The arsenic concentrations used in the source crystal were typically of the order of 10^{17} – 10^{19} /cc. These concentrations were rather high compared to the concentrations desired in the diffused surface layers since compensation had to be made for losses of arsenic due to the imperfect fit of the cover on the capsule and due to some chemical reaction and adsorption which occurred on the internal surfaces of the capsule.

The layers obtained after diffusion were then evaluated for sheet conductivity and thickness. To measure the sheet conductivity a four-point probe method⁶ was used. An island of the surface layer was formed by masking and etching to reveal the junction between the surface layer and the p-type body. The island was then biased in the reverse direction with respect to the body thus effectively isolating it electrically during the measurement of its sheet conductivity. The thickness of the surface layer was obtained by first lapping at a small angle to the original surface ($\frac{1}{2}^\circ$ – 1°) and locating the junction on the beveled surface with a thermal probe; then multiplying the tangent of the angle between the two surfaces by the distance from the edge of the bevel to the junction gives the desired thickness. Another particularly convenient method of measuring the thickness⁷ is to place a half silvered mirror parallel to the original surface and count fringes, of the sodium D-line for example, from the edge of the bevel to the junction. Typically the transistors described here were prepared from diffused layers with a sheet conductivity of about 200 ohms/square, and a layer thickness of $(1.5 \pm 0.3) \times 10^{-4}$ cm.

When the surface layer had been evaluated, the emitter and base contacts were made using techniques of vacuum evaporation and alloying. For the emitter, a film of aluminum approximately 1,000 Å thick was evaporated onto the surface through a mask which defined an emitter area of 1×2 mils. The bar with the evaporated aluminum was then placed on a strip heater in a hydrogen atmosphere and momentarily brought up to a temperature sufficient to alloy the aluminum. The emitter having been thus formed, the bar was again placed in the masking jig and a film of gold-antimony alloy from 3,000 to 4,000 Å thick was evaporated onto the surface. This film was identical in area to the emitter, and was placed parallel to and 0.5 to 1 mil away from the emitter. The bar was again placed on the heater strip and heated to the gold-germanium eutectic temperature, thus forming the ohmic base contact. The masking jig was constructed to permit the simultaneous evaporation of eight pairs of contacts on each bar. Thus, using a 3-mil diamond saw, a bar could be cut into eight units.

Each unit, with an alloyed emitter and base contact, was then soldered to a platinum tab with indium, a sufficient quantity of indium being used to alloy through the n-type surface layer on the back of the unit. One of the last steps was to mask the emitter and base contacts with a 6- to 8-mil diameter dot of wax and form a small area collector junction by etching the unit attached to the platinum tab, in CP4. After washing in solvents to remove the wax, the unit was mounted in a header designed to allow electrolytically pointed wire contacts to be made to the base and emitter areas of the transistor. These spring contacts were made of 1-mil phosphor bronze wire.

ELECTRICAL CHARACTERIZATION

Of the parameters that characterize the performance of a transistor, one of the most important is the short circuit current gain (α) versus frequency. The measured variation of α and $\alpha/(1 - \alpha)$ (short-circuit current gain in the grounded emitter circuit) as a function of frequency for a typical unit is shown in Fig. 1. For comparison the same parameters for an exceptionally good unit are shown in Fig. 2.

In order that the alpha-cutoff frequency be a measure of the transit time of minority carriers through the active regions of the transistor, any resistance-capacity cutoffs, of the emitter and collector circuits, must lie considerably higher than the measured f_α . In the emitter circuit, an external contact resistance to the aluminum emitter of the order of 10

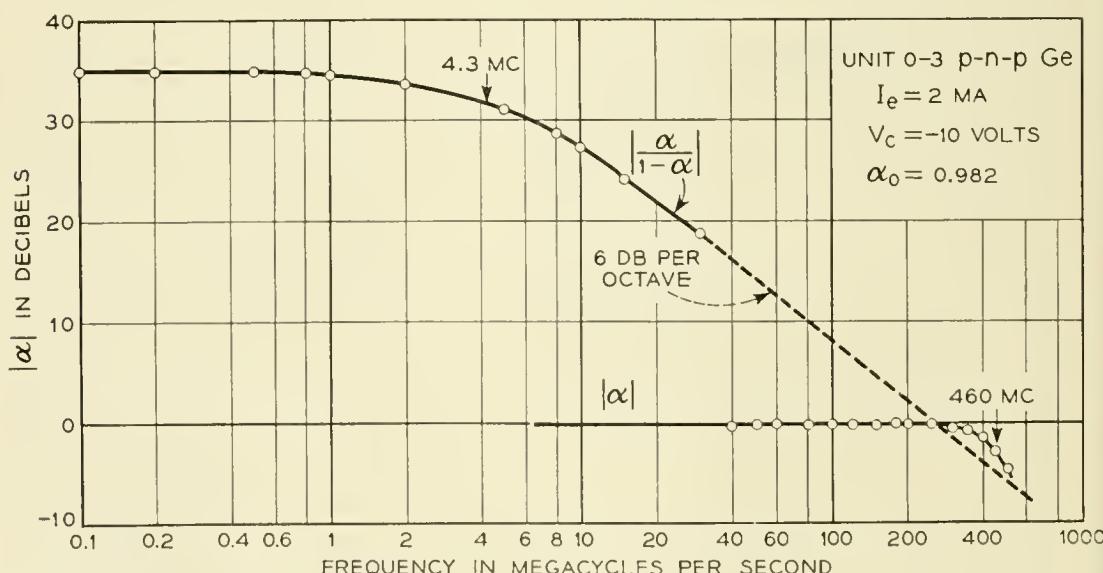


Fig. 1 — The grounded emitter and grounded base response versus frequency for a typical unit.

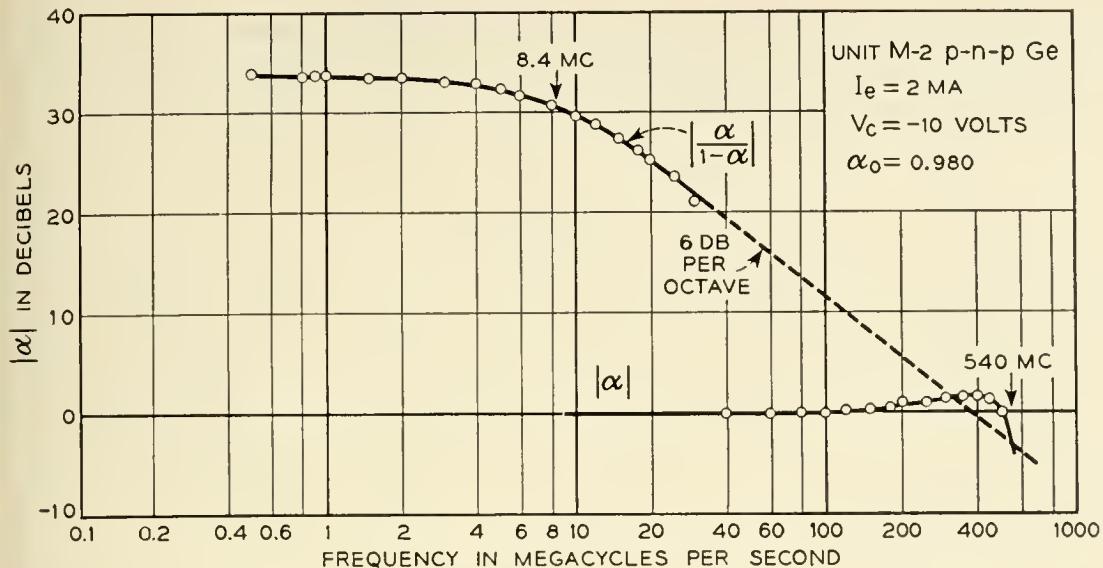


Fig. 2 — The grounded emitter and grounded base response versus frequency for an exceptionally good unit.

to 20 ohms and a junction transition capacity of $1 \mu\text{fd}$ were measured. The displacement current which flows through this transition capacity reduces the emitter efficiency and must be kept small relative to the injected hole current. With 1 milliampere of current flowing through the emitter junction, and consequently an emitter resistance of 26 ohms, the emitter cutoff for this transistor was above 6,000 mc/sec. One can now see that the emitter area must be small and the current density high to attain a high emitter cutoff frequency. The fact that a low base resistance requires a high level of doping in the base region, and thus a high emitter transition capacity, restricts one to small areas and high current densities.

In the collector circuit capacities of 0.5 to $0.8 \mu\text{fd}$ at a collector voltage of -10 volts were measured. There was a spreading resistance in the collector body of about 100 ohms which was the result of the small emitter area. The base resistance was approximately 100 ohms. If the phase shift and attenuation due to the transport of minority carriers through the base region were small at the collector cutoff frequency, the effective base resistance would be decreased by the factor $(1 - \alpha)$. The collector cutoff frequency is then given by

$$f_c = \frac{1}{2\pi C_c R_c}$$

where C_c = collector transition capacity

and R_c = collector body spreading resistance.

However, in the transistors described here the base region produces the major contribution to the observed alpha-cutoff frequency and it is more appropriate to use the expression

$$f_c = \frac{1}{2\pi C_c(r_b + R_c)}$$

where $r_b \equiv$ base resistance. This cutoff frequency could be raised by increasing the collector voltage, but the allowable power dissipation in the mounting determines an upper limit for this voltage. It should be noted that an increase in the doping of the collector material would raise the cutoff since the spreading resistance is inversely proportional to N_a , while the junction capacity for constant collector voltage is only proportional to $N_a^{1/2}$.

The low-frequency alpha of the transistor ranged from 0.95 to 0.99 with some exceptional units as high as 0.998. The factors to be considered here are the emitter efficiency γ and the transport factor β . The transport factor is dependent upon the lifetime in the base region, the recombination velocity at the surface immediately surrounding the emitter, and the geometry. The geometrical factor of the ratio of the emitter dimensions to the base layer thickness is > 10 , indicating that solutions for a planar geometry may be assumed.⁸ If a lifetime in the base region of 1 microsecond and a surface recombination velocity of 2,000 cm/sec is assumed a perturbation calculation⁹ gives

$$\beta = 0.995$$

The high value of β obtained with what is estimated to be a low base region lifetime and a high surface recombination velocity indicates that the observed low frequency alpha is most probably limited by the emitter injection efficiency. As for the emitter injection efficiency, within the accuracy to which the impurity concentrations in the emitter regrowth layer and the base region are known, together with the thicknesses of these two regions, the calculated efficiency is consistent with the experimentally observed values.

CONSIDERATIONS OF TRANSIT TIME

An examination of what agreement exists between the alpha-cutoff frequency and the physical measurements of the base region involves the mechanism of transport of minority carriers through the active regions of the transistor. The "active regions" include the space charge

region of the collector junction. The transit time through this region¹⁰ is no longer a negligible factor. A short calculation will show that with -10 volts on the collector junction, the space charge layer is about 4×10^{-4} em thick and that the frequency cutoff associated with transport through this region is approximately 3,000 mc/sec.

The remaining problem is the transport of minority carriers through the base region. Depending upon the boundary conditions existing at the surface of the germanium during the diffusion process, considerable gradients of the impurity density in the surface layer are possible. However, the problem of what boundary conditions existed during the diffusion process employed in the fabrication of these transistors will not be discussed here because of the many uncertainties involved. Some qualitative idea is necessary though of how electric fields arising from impurity gradients may affect the frequency behavior of a transistor in the limit of low injection.

If one assumes a constant electric field as would result from an exponential impurity gradient in the base region of a transistor, then the continuity equation may be solved for the distribution of minority carriers.⁴ From the hole distribution one can obtain an expression for the transport factor β and it has the form

$$\beta = e^\eta \frac{Z}{\eta \sinh Z + Z \cosh Z}$$

where

$$\eta \equiv \frac{1}{2} \ln \frac{N_e}{N_c} = \frac{1}{2} \frac{qE}{kT} w,$$

$$Z \equiv [i\varphi + \eta^2]^{1/2}$$

$$\varphi = \omega \frac{w^2}{D_p}$$

N_e ≡ donor density in base region at emitter junction

N_c ≡ donor density in base region at collector junction

E ≡ electric field strength

D_p ≡ diffusion constant for holes

w ≡ width of the base layer

A plot of this function for various values of η is shown in Fig. 3. For $\eta = 0$, the above expression reduces to the well known case of a uniformly doped base region. The important feature to be noted in Fig. 3 is that relatively small gradients of the impurity distribution in the base layer can produce a considerable enhancement of the frequency response.

It is instructive to calculate what the alpha-cutoff frequency would be for a base region with a uniform distribution of impurity. The effective thickness of the base layer may be estimated by decreasing the measured thickness of the surface layer by the penetration of the space charge region of the collector and the depth of the alloyed emitter structure. Using a value for the diffusion constant of holes in the base region appropriate to a donor density of about $10^{17}/\text{cc}$,

$$300 \text{ mc/s} \leq f_\alpha \leq 800 \text{ mc/s}$$

This result implies that the frequency enhancement due to "built-in" fields is at most a factor of two. In addition it was observed that the alpha-cutoff frequency was a function of the emitter current as shown in Fig. 4. This variation indicates that at least intermediate injection

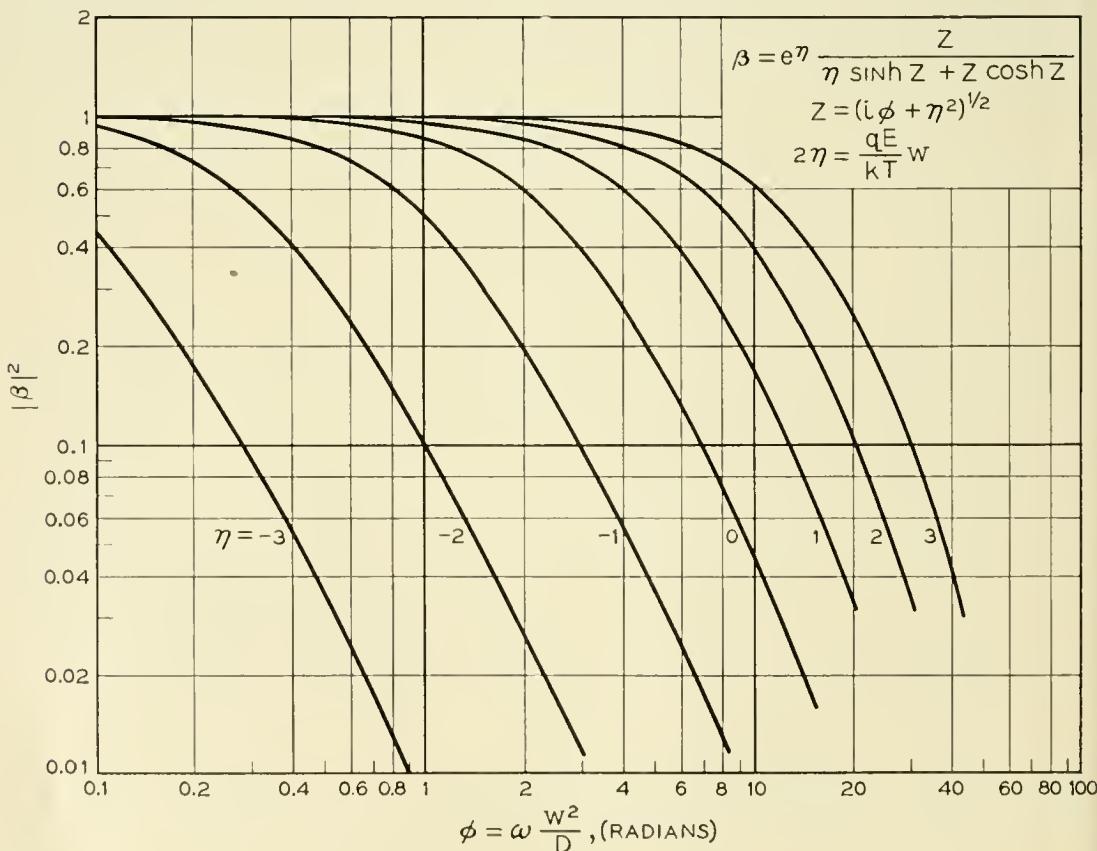


Fig. 3 — The variation of $|\beta|$ versus frequency for various values of a uniform drift field in the base region.

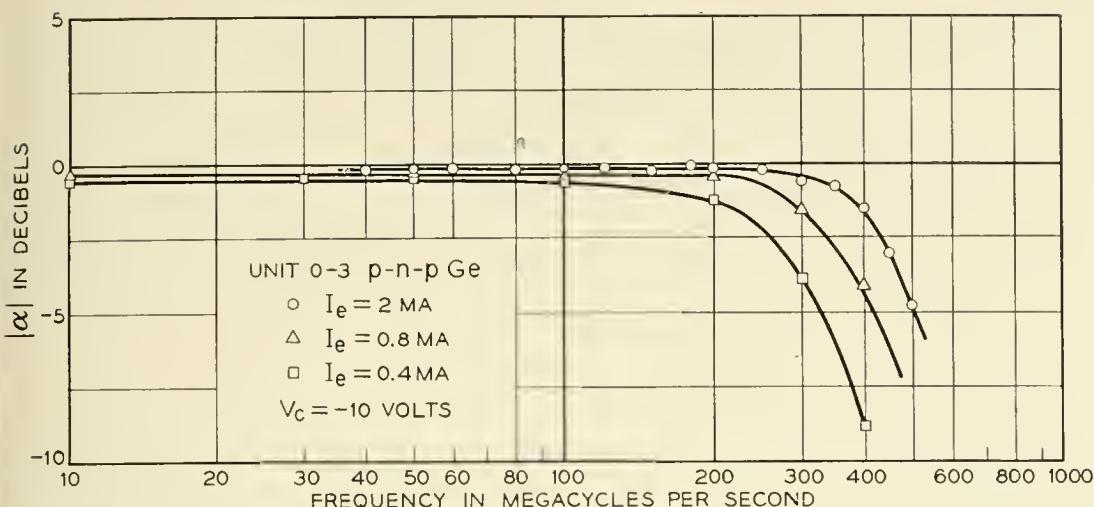


Fig. 4 — The variation of the alpha-cutoff frequency as a function of emitter current.

levels exist in the range of emitter current shown in Fig. 4. The conclusion to be drawn then is that electric fields produced by impurity gradients in the base region are not the dominant factor in the transport of minority carriers in these transistors.

The emitter current for a low level of injection could not be determined by measuring f_α versus I_e because the high input impedance at very low levels was shorted by the input capacity of the header and socket. Thus at very small emitter currents the measured cutoff frequency was due to an emitter cutoff and was roughly proportional to the emitter current. At $I_e \geq 1 \text{ ma}$ this effect is small, but here at least intermediate levels of injection already exist.

A further attempt to measure the effect of any "built-in" fields by turning the transistor around and measuring the inverse alpha proved fruitless for two reasons. The unfavorable geometrical factor of a large collector area and a small emitter area as well as a poor injection efficiency gave an alpha of only

$$\alpha = 0.1$$

Secondly, the injection efficiency turns out in this case to be proportional to $\omega^{-1/2}$ giving a cutoff frequency of less than 1 mc/sec. The square-root dependence of the injection efficiency on frequency may be readily seen. The electron current injected into the collector body may be expressed as

$$J_e = qD_nN \left[\frac{1 + i\omega\tau_e}{L_e^2} \right]^{1/2}$$

where $q \equiv$ electronic charge

D_n ≡ diffusion constant of electrons

$$N = \frac{q}{kT} v_1 n_c$$

v_1 ≡ voltage across collector junction

n_c ≡ density of electrons on the p-type side of the collector junction

τ_e ≡ lifetime of electrons in collector body

L_e ≡ diffusion length of electrons in the collector body

Since the inverse cutoff frequency is well below that associated with the base region, we may regard the injected hole current as independent of the frequency in this region. The injection efficiency is low so that

$$\gamma \approx \frac{J_p}{J_e} \ll 1$$

Thus at a frequency where

$$\omega \tau_c \gg 1$$

then

$$\gamma \propto \omega^{-1/2}$$

An interesting feature of these transistors was the very high current densities at which the emitter could be operated without appreciable loss of injection efficiency. Fig. 5 shows the transmission of a 50 millimicro-second pulse up to currents of 18 milliamperes which corresponds to a current density of 1800 amperes/cm². The injection efficiency should remain high as long as the electron density at the emitter edge of the base region remains small compared to the acceptor density in the emitter regrowth layer. When high injection levels are reached the injected hole density at the emitter greatly exceeds the donor density in the base region. In order to preserve charge neutrality then

$$p \approx n$$

where p ≡ hole density

n ≡ electron density

As the injected hole density is raised still further the electron density will eventually become comparable to the acceptor density in the emitter regrowth layer. The density of acceptors in the emitter regrowth

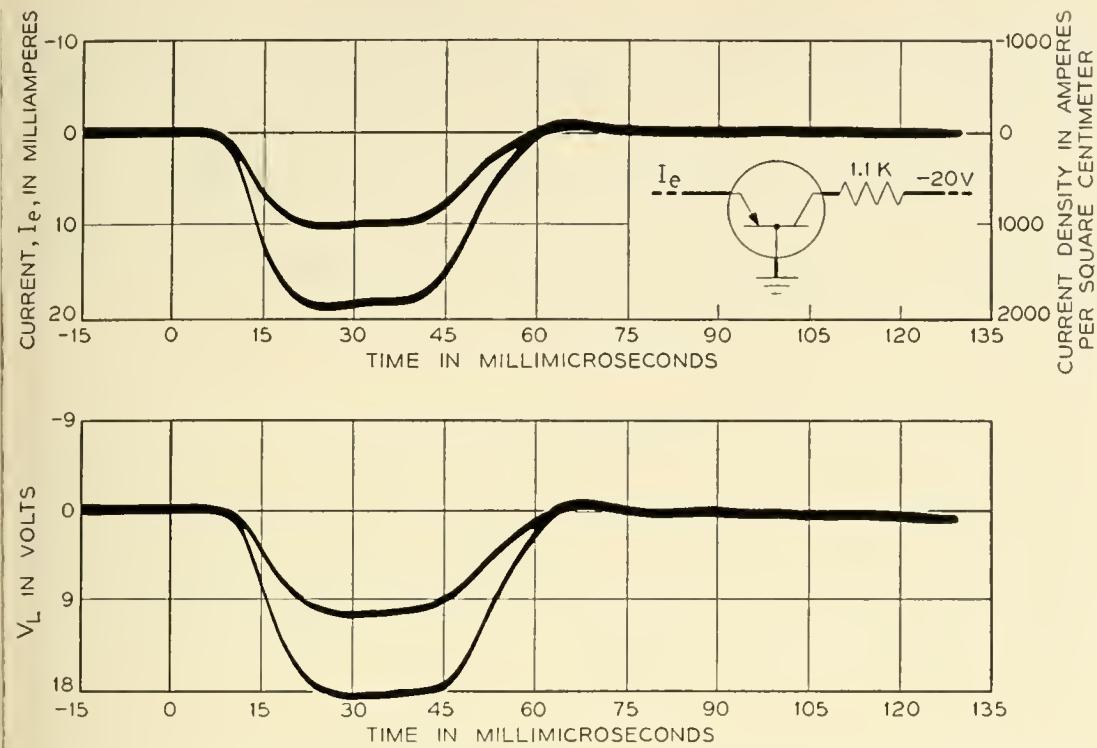


Fig. 5 — Transmission of a 50 millimicrosecond pulse at emitter currents up to 18 ma by a typical unit. (Courtesy of F. K. Bowers).

region is of the order of

$$N_A \approx 10^{20}/\text{cc}$$

and this is to be compared with injected hole density at the base region side of the emitter junction. The relation between the injected hole density and the current density may be approximated by⁸

$$J_p = \frac{2qD_p p_1}{w}$$

where p_1 = hole density at emitter side of base region

w = width of base region

A short calculation indicates that the emitter efficiency should remain high at a current density of an order of magnitude higher than 1,800 amp/cm². The measurements were not carried to higher current densities because the voltage drop across the spreading resistance in the collector was producing saturation of the collector junction.

CONCLUSIONS

Impurity diffusion is an extremely powerful tool for the fabrication of high frequency transistors. Moreover, of the 50-odd transistors which

were made in the laboratory, the characteristics were remarkably uniform considering the variations usually encountered at such a stage of development. It appears that diffusion process is sufficiently controllable that the thickness of the base region can be reduced to half that of the units described here. Therefore, with no change in the other design parameters, outside of perhaps a different mounting, units with a 1000 mc/s cutoff frequency should be possible.

ACKNOWLEDGMENT

The author wishes to acknowledge the help of P. W. Foy and W. Wiegmann who aided in the construction of the transistors, D. E. Thomas who designed the electrical equipment needed to characterize these units, and J. Klein who helped with the electrical measurements. The numerical evaluation of alpha for drift fields was done by Lillian Lee whose assistance is gratefully acknowledged.

REFERENCES

1. C. S. Fuller, Phys. Rev., **86**, pp. 136-137, 1952.
2. J. Saby and W. C. Dunlap, Jr., Phys. Rev., **90**, p. 630, 1953.
3. W. Shockley, private communication.
4. H. Krömer, Archiv. der Elek. Übertragung, **8**, No. 5, pp. 223-228, 1954.
5. R. A. Logan and M. Schwartz, Phys. Rev., **96**, p. 46, 1954.
6. L. B. Valdes, Proc. I.R.E., **42**, pp. 420-427, 1954.
7. W. L. Bond and F. M. Smits, to be published.
8. E. S. Rittner, Phys. Rev., **94**, p. 1161, 1954.
9. W. M. Webster, Proc. I.R.E., **42**, p. 914, 1954.
10. J. M. Early, B.S.T.J., **33**, pp. 517-533, 1954.

Waveguide Investigations with Millimicrosecond Pulses

By A. C. BECK

(Manuscript received October 11, 1955)

Pulse techniques have been used for many waveguide testing purposes. The importance of increased resolution by means of short pulses has led to the development of equipment to generate, receive and display pulses about 5 or 6 millimicroseconds long. The equipment is briefly described and its resolution and measuring range are discussed. Dominant mode waveguide and antenna tests are described and illustrated. Applications to multimode waveguides are then considered. Mode separation, delay distortion and its equalization, and mode conversion are discussed, and examples are given. The resolution obtained with this equipment provides information that is difficult to get by any other means, and its use has proved to be very helpful in waveguide investigations.

CONTENTS

1. Introduction	35
2. Pulse Generation	36
3. Receiver and Indicator	41
4. Resolution and Measuring Range	42
5. Dominant Mode Waveguide Tests	43
6. Testing Antenna Installations	45
7. Separation of Modes on a Time Basis	48
8. Delay Distortion	52
9. Delay Distortion Equalization	54
10. Measuring Mode Conversion from Isolated Sources	57
11. Measuring Distributed Mode Conversion in Long Waveguides	61
12. Concluding Remarks	65

1. INTRODUCTION

Pulse testing techniques have been employed to advantage in waveguide investigations in numerous ways. The importance of better resolution through the use of short pulses has always been apparent and, from the first, equipment was employed which used as short a pulse as possible. Radar-type apparatus using magnetrons and a pulse width of about one-tenth microsecond has seen considerable use in waveguide research, and many of the results have been published.^{1, 2}

To improve the resolution, work was initiated some time ago by S. E. Miller to obtain measuring equipment which would operate with much shorter pulses. As a result, pulses about 5 or 6 millimicroseconds long became available at a frequency of 9,000 mc. In a pulse of this length there are less than 100 cycles of radio frequency energy, and the signal occupies less than ten feet of path length in the transmission medium. The RF bandwidth required is about 500 mc. In order to obtain such bandwidths, traveling wave tubes were developed by J. R. Pierce and members of the Electronics Research Department of the Laboratories. The completed amplifiers were designed by W. W. Mumford, N. J. Pierce, R. W. Dawson and J. W. Bell assisted in the design and construction phases, and G. D. Mandeville has been closely associated in all of this work.

2. PULSE GENERATION

These millimicrosecond pulses have been produced by two different types of generators. In the first equipment, a regenerative pulse generator of the type suggested by C. C. Cutler of the Laboratories was used.³ This was a very useful device, although somewhat complicated and hard to keep in adjustment. A brief description of it will permit comparisons with a simpler generator which was developed a little later.

A block diagram of the regenerative pulse generator is shown in Fig. 1. The fundamental part of the system is the feedback loop drawn with heavy lines in the lower central part of the figure. This includes a traveling wave amplifier, a waveguide delay line about sixty feet long, a crystal expander, a band-pass filter, and an attenuator. This combination forms an oscillator which produces very short pulses of microwave energy. Between pulses, the expander makes the feedback loop loss too high for oscillation. Each time the pulse circulates around the loop it tends to shorten, due to the greater amplification of its narrower upper part caused by the expander action, until it uses the entire available bandwidth. A 500-mc gaussian band-pass filter is used in the feedback loop of this generator to determine the final bandwidth. An automatic gain control operates with the expander to limit the pulse amplitude, thus preventing amplifier compression from reducing the available expansion.

To get enough separation between outgoing pulses for reflected pulse measurements with waveguides, the repetition rate would need to be too low for a practical delay line length in the loop. Therefore a 12.8-mc fundamental rate was chosen, and a gated traveling wave tube amplifier was used to reduce it to a 100-ke rate at the output. This amplifier is kept in a cutoff condition for 127 pulses, and then a gate pulse restores

it to the normal amplifying condition for fifty millimicroseconds, during which time the 128th pulse is passed on to the output of the generator as shown on Fig. 1.

The synchronizing system is also shown on Fig. 1. A 100-ke quartz crystal controlled oscillator with three cathode follower outputs is the basis of the system. One output goes through a seven stage multiplier to get a 12.8-mc signal, which is used to control a pulser for synchronizing the circulating loop. Another output controls the gate pulser for the output traveling wave amplifier. Accurate timing of the gate pulse is obtained by adding the 12.8-mc pulses through a buffer amplifier to the gate pulser. The third output synchronizes the indicator oscilloscope sweep to give a steady pattern on the screen.

Although this equipment was fairly satisfactory and served for many

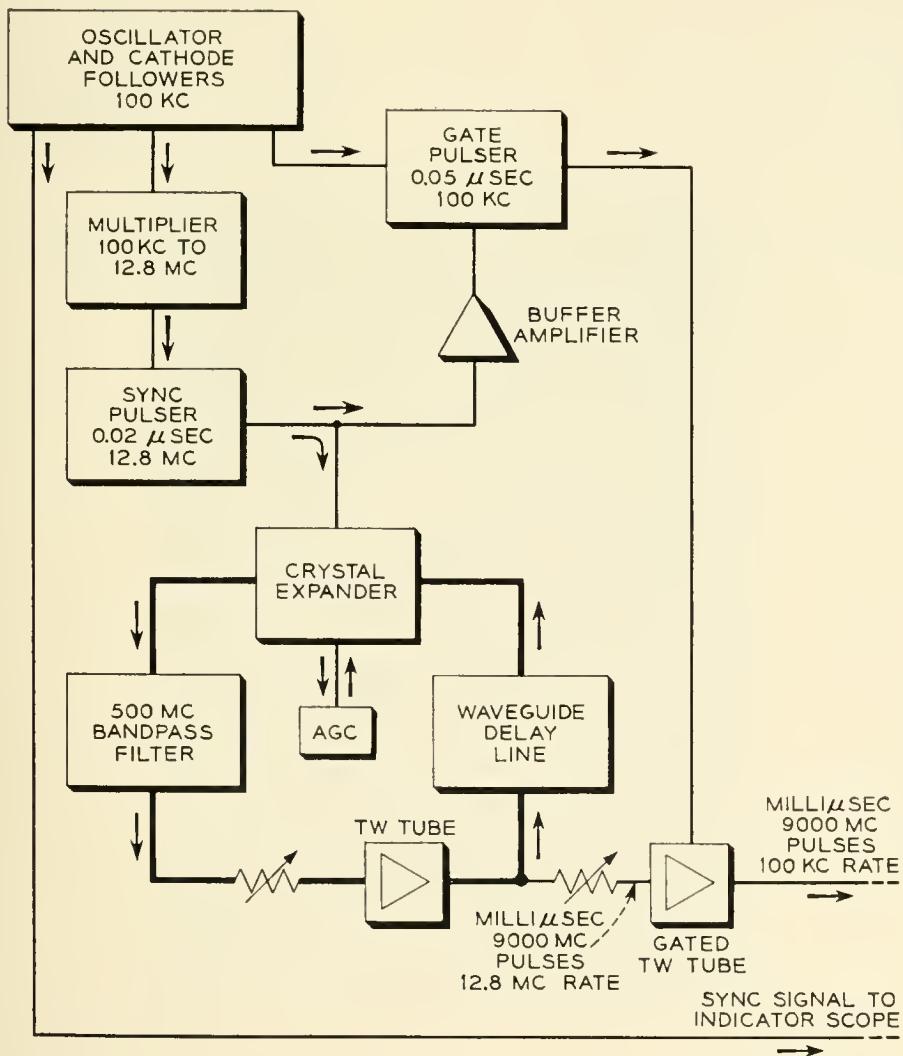


FIG. 1 — Block diagram of the regenerative pulse generator.

testing purposes, it was rather complex and there were some problems in its construction and use. It was difficult to obtain suitable microwave crystals to match the waveguide at low levels in the expander. This would make it even more difficult to build this type of pulse generator for higher frequency ranges. Stability also proved to be a problem. The frequency multiplier had to be very well constructed to avoid phase shift due to drifting. The gate pulser also required care in design and construction in order to get a stable and flat output pulse. It was somewhat troublesome to keep the gain adjusted for proper operation, and the gate pulse time adjustment required some attention. The pulse frequency could not be changed. For these reasons, and in order to get a smaller, lighter and less complicated pulse generator, work was carried out to produce pulses of about the same length by a simpler method.

If the gated output amplifier of Fig. 1 were to have a CW instead of a pulsed input, a pulse of microwave energy would nevertheless appear at the output because of the presence of the gating pulse. This gating pulse is applied to the beam forming electrode of the tube to obtain the gating action. If the beam forming electrode could be pulsed from cutoff to its normal operating potential for a very short time, very short pulses of output energy could be obtained from a continuous input signal. However, it is difficult to obtain millimicrosecond video gating pulses of sufficient amplitude for this purpose at a 100-kc repetition rate.

A traveling-wave tube amplifies normally only when the helix is within a small voltage range around its rated dc operating value. For voltages either above or below this range, the tube is cut off. When the helix voltage is raised through this range into the cutoff region beyond it, and then brought back again, two pulses are obtained, one during a small part of the rise time and the other during a small part of the return time. If the rise and fall times are steep, very short pulses can be obtained. Fig. 2 shows the pulse envelopes photographed from the indicator scope screen when this is done. For the top trace, the helix was biased 300 volts negatively from its normal operating potential, then pulsed to its correct operating range for about 80 millimicroseconds, during which time normal amplification of the CW input signal was obtained. The effect of further increasing the helix video pulse amplitude in the positive direction is shown by the succeeding lower traces. The envelope dips in the middle, then two separated pulses remain — one during a part of the rise time and one during a part of the fall time of helix voltage. The pulses shown on the bottom trace have shortened to about six millimicroseconds in length. The helix pulse had a positive amplitude of about 500 volts for this trace.

Since only one of these pulses can be used to get the desired repetition rate, it is necessary to eliminate the other pulse. This is done in a similar manner to that used for gating out the undesired pulses in the regenerative pulse generator. However, it is not necessary to use another amplifier, as was required there, since the same tube can be used for this purpose, as well as for producing the microwave pulses. Its beam forming electrode is biased negatively about 250 volts with respect to the cathode, and then is pulsed to the normal operating potential for about 50 millimicroseconds during the time of the first short pulse obtained by gating the helix. Thus, the beam forming electrode potential has been returned to the cutoff value during the second helix pulse, which is therefore eliminated.

A block diagram of the resulting double-gated pulse generator is shown in Fig. 3. Comparison with Fig. 1 shows that it is simpler than the regenerative pulse generator, and it has also proved more satisfactory in operation. It can be used at any frequency where a signal source and a traveling-wave amplifier are available, and the pulse

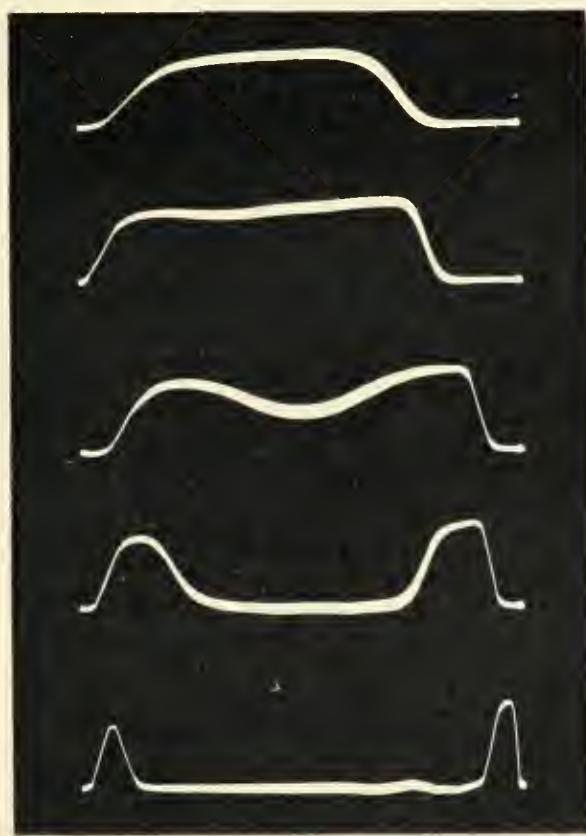


Fig. 2 — Envelopes of microwave pulses at the output of a traveling wave amplifier with continuous wave input and helix gating. The gating voltage is higher for the lower traces.

frequency can be set anywhere within the bandwidth of the traveling-wave amplifier by tuning the klystron oscillator.

The pulse center frequency is shifted from that of the klystron oscillator frequency by this helix gating process. An over-simplified but helpful explanation of this effect can be obtained by considering that the microwave signal voltage on the helix causes a bunching of the electron stream. This bunching has the same periodicity as the microwave signal voltage when the dc helix potential is held constant. However, since the helix voltage is continuously increased in the positive direction during the time of the first pulse, the average velocity of the last bunches of electrons becomes higher than that of the earlier bunches in the pulse, because the later electrons come along at the time of higher positive helix voltage. This tends to shorten the total length of the series of bunches, resulting in a shorter wavelength at the output end of the helix and therefore a higher output microwave frequency. On the second pulse, obtained when the helix voltage returns toward zero, the process is reversed, the bunching is stretched out, and the frequency is decreased. This second pulse is, however, gated out in this arrangement by the beam-forming electrode pulsing voltage. The result for this particular tube and pulse length is an effective output frequency approximately 150 mc higher than the oscillator frequency, but this figure is not constant over the range of pulse frequencies available within the amplifier bandwidth.

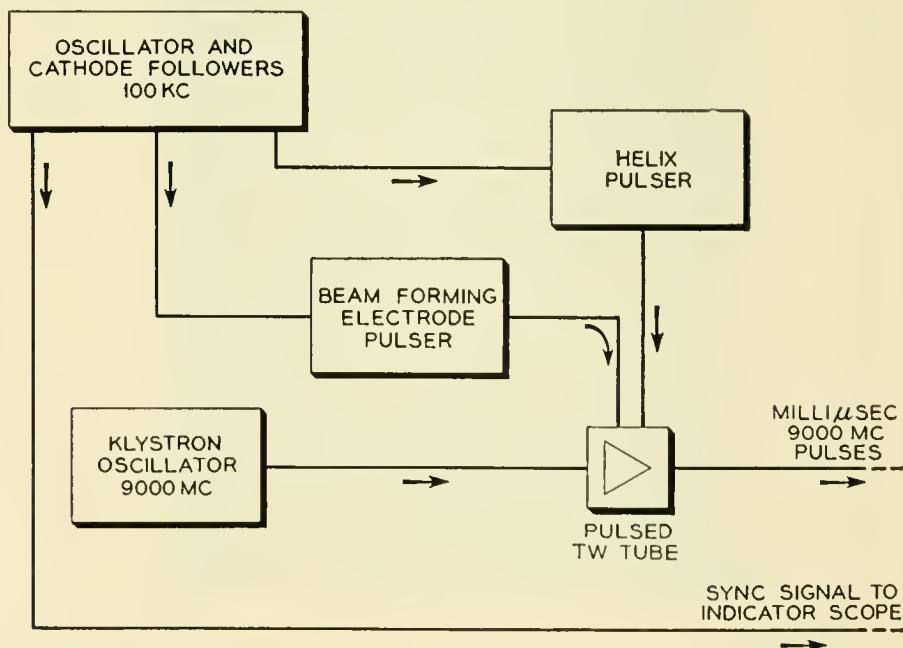


Fig. 3 — Block diagram of the double-gated traveling wave tube millimicrosecond pulse generator.

3. RECEIVER AND INDICATOR

The receiving equipment is shown in Fig. 4. It uses two traveling-wave amplifiers in cascade. A wide band detector and a video amplifier then follow, and the signal envelope is displayed by connecting it to the vertical deflecting plates of a 5 XP type oscilloscope tube. The video amplifier now consists of two Hewlett Packard wide band distributed amplifiers, having a baseband width of about 175 mc. The second one of these has been modified to give a higher output voltage. The sweep circuits for this oscilloscope have been built especially for this use, and produce a sweep speed in the order of 6 feet per microsecond. An intensity pulser is used to eliminate the return trace. These parts of the system are controlled by a synchronizing output from the pulse generator 100-kc oscillator. A precision phase shifter is used at the receiver for the same purpose that a range unit is employed in radar systems. This has a dial, calibrated in millimicroseconds, which moves the position of a pulse appearing on the scope and makes accurate measurement of pulse delay time possible.

Fig. 4 also shows the appearance of the pulses obtained with this equipment. The pulse on the left-hand side of this trace came from the

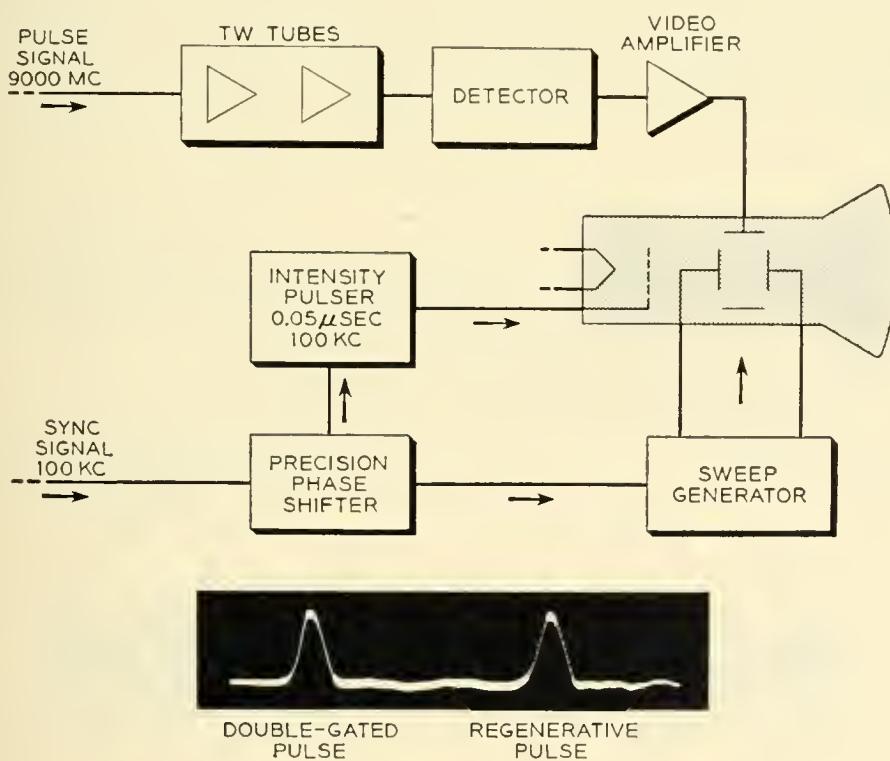


Fig. 4 — Block diagram of millimicrosecond pulse receiver and indicator. The indicator trace photograph shows pulses from each type of generator.

newer double-gated pulse generator, while the pulse on the right was produced by the regenerative pulse generator. It can be seen that they appear to have about the same pulse width and shape. This is partly due to the fact that the video amplifier bandwidth is not quite adequate to show the actual shape, since in both cases the pulses are slightly shorter than can be correctly reproduced through this amplifier. The ripples on the base line following the pulses are also due to the video amplifier characteristics when used with such short pulses.

4. RESOLUTION AND MEASURING RANGE

Fig. 5 shows a piece of equipment which was placed between the pulse generator and the receiver to show the resolution which can be obtained. This waveguide hybrid junction has its branch marked 1 connected to the pulse generator and branch 3 connected to the receiver. If the two side branches marked 2 and 4 were terminated, substantially no energy would be transmitted from the pulser straight through to the receiver. However, a short circuit placed on either side branch will send energy through the system to the receiver. Two short circuits were so placed that the one on branch 4 was 4 feet farther away from the hybrid junction than the one on branch 2. The pulse appearing first is produced by a signal traveling from the pulser to the short circuit on branch 2 and then through to the receiver, as shown by the path drawn with short dashes. A second pulse is produced by the signal which travels

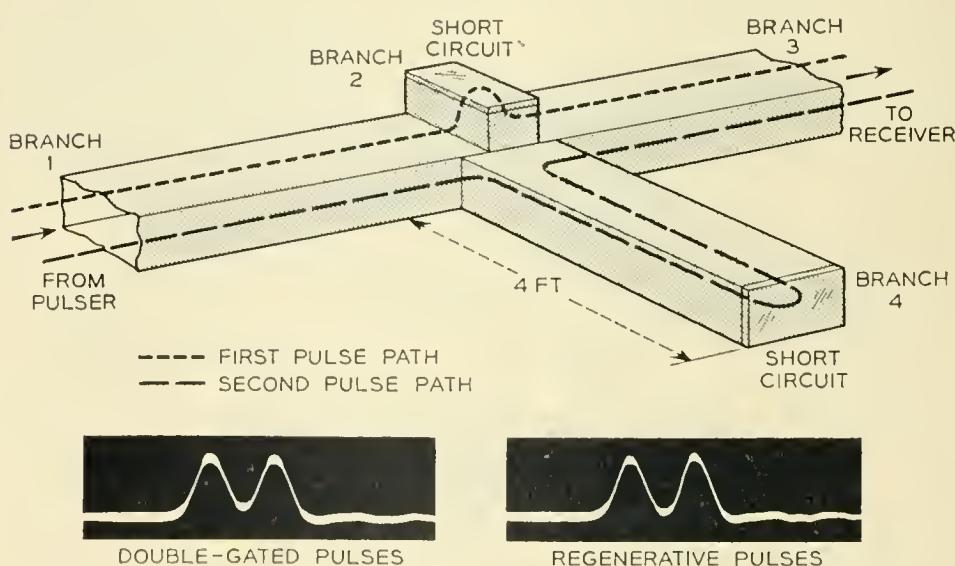


Fig. 5 — Waveguide hybrid circuit used to demonstrate resolution of millimicrosecond pulses. Trace photographs of pulses from each type of generator are shown.

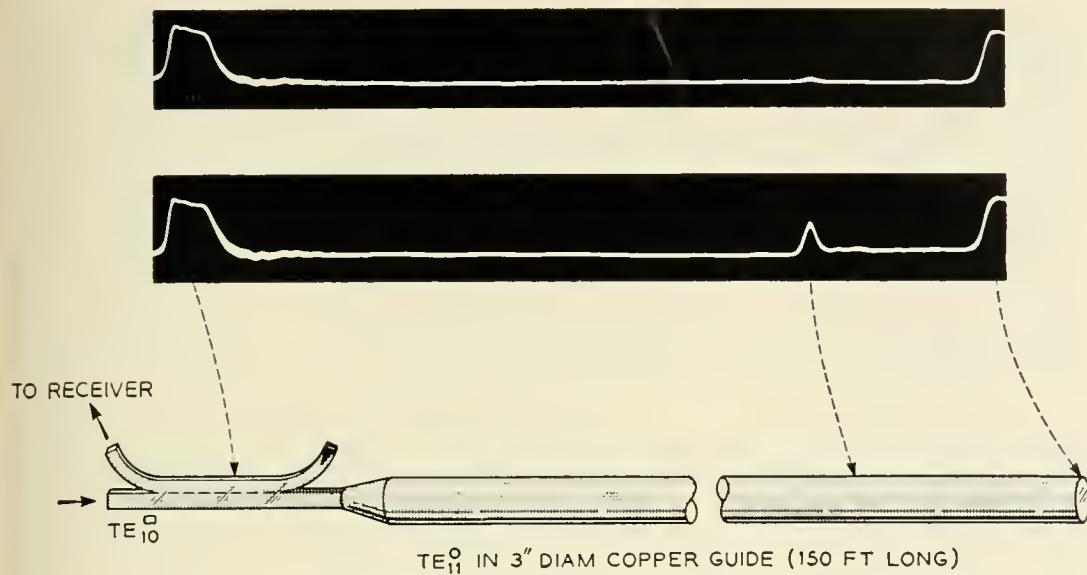


Fig. 6 — Waveguide arrangement and oscilloscope trace photos showing presence and location of defective joint. The dominant mode (TE_{11}^0) was used with its polarization changed 90 degrees for the two trace photos.

from the pulse generator through branch 4 to the short circuit and then to the receiver as shown by the long dashed line. This pulse has traveled 8 feet farther in the waveguide than the first pulse. This would be equivalent to seeing separate radar echoes from two targets about 4 feet apart. Resolution tests made in this way with the pulses from the regenerative pulse generator, and from the double-gated pulse generator, are shown on Fig. 5. With our video amplifier and viewing equipment, there is no appreciable difference in the resolution obtained using either type of pulse generator.

The measuring range is determined by the power output of the gated amplifier at saturation and by the noise figure of the first tube in the receiver. In this equipment the saturation level is about 1 watt, and the noise figure of the first receiver tube is rather poor. As a result, received pulses about 70 db below the outgoing pulse can be observed, which is enough range for many measurement purposes.

5. DOMINANT MODE WAVEGUIDE TESTS

Fig. 6 shows the use of this equipment to test 3" round waveguides such as those installed between radio repeater equipment and an antenna. This particular 150-foot line had very good soldered joints and was thought to be electrically very smooth. The signal is sent in through a transducer to produce the dominant TE_{11} mode. The receiver is connected through a directional coupler on the sending end to look for any

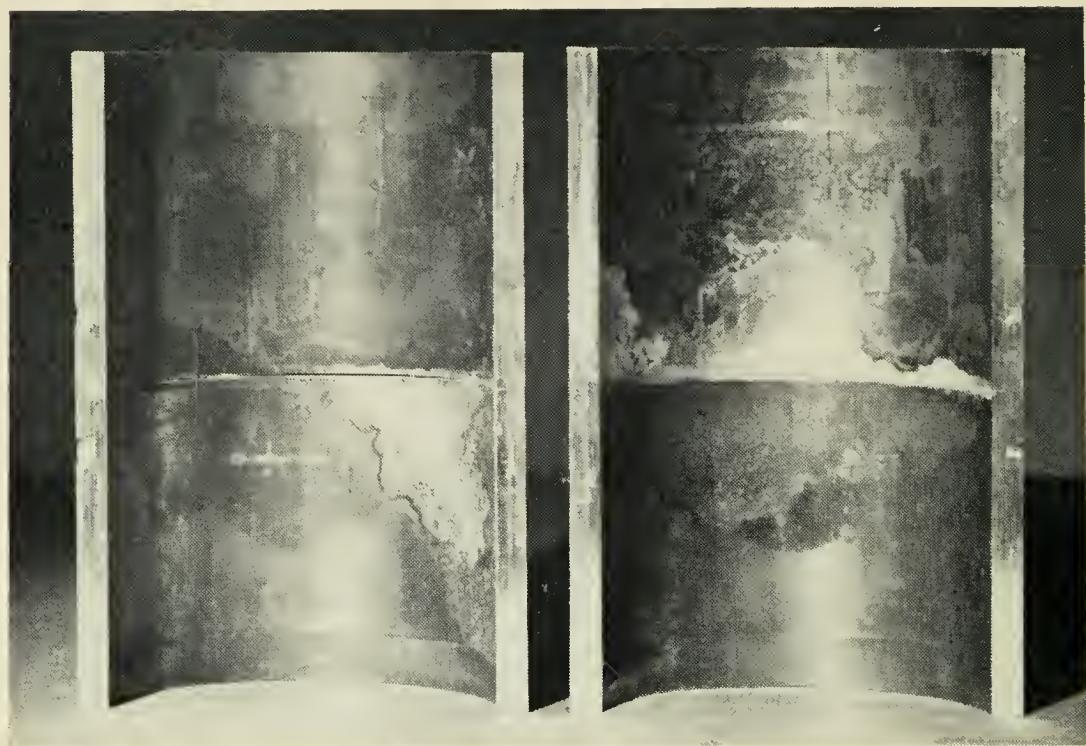


Fig. 7 --- Defective joint caused by imperfect soldering which gave the reflection shown on Fig. 6.

reflections from imperfections in the line. The overloaded signal at the left of the oscilloscope trace is produced by leakage directly through the directional coupler. The overloaded signal on the other end of this trace is produced by the reflection from the short circuit piston at the far end of the waveguide. The signal between these two, which is about 45 db down from the input signal, is produced by an imperfect joint in the waveguide. The signal polarization was oriented so that a maximum reflection was obtained in the case of the lower trace. In the other trace, the polarization was changed by 90°. It is seen that this particular joint produces a stronger reflection for one polarization than for the other. By use of the precision phase shifter in the receiver the exact location of this defect was found and the particular joint that was at fault was sawed out. Fig. 7 shows this joint after the pipe had been cut in half through the middle. The guide is quite smooth on the inside in spite of the discoloration of some solder that is shown here, but on the left-hand side of the illustration the open crack is seen where the solder did not run in properly. This causes the reflected pulse that shows on the trace. The fact that this crack is less than a semi-circumference in length causes the echo to be stronger for one polarization than for the other.

Fig. 8 shows the same test for a 3" diameter aluminum waveguide 250 feet long. This line was mounted horizontally in the test building with compression couplings used at the joints. The line expanded on warm days but the friction of the mounting supports was so great that it pulled open at some of the joints when the temperature returned to normal. These open joints produced reflected pulses from 40 to 50 db down, which are shown here. They come at intervals equal to the length of one section of pipe, about 12 feet. Some of these show polarization effects where the crack was more open on one side than on the other, but others are almost independent of polarization. These two photographs of the trace were taken with the polarization changed 90°.

Fig. 9 shows the same test for a 3" diameter galvanized iron waveguide. This line had shown fairly high loss using CW for measurements. The existence of a great many echoes from random distances indicates a rough interior finish in the waveguide. Fig. 10 shows the kind of imperfections in the zinc coating used for galvanizing which caused these reflections.

6. TESTING ANTENNA INSTALLATIONS

The use of this equipment in testing waveguide and antenna installations for microwave radio repeater systems is shown in Fig. 11. This particular work was done in cooperation with A. B. Crawford's antenna research group at Holmdel, who designed the antenna system. A directional coupler was used to observe energy reflections from the system under test. In this installation a 3" diameter round guide carrying the TE_{11}^0 mode was used to feed the antenna. Two different waveguide

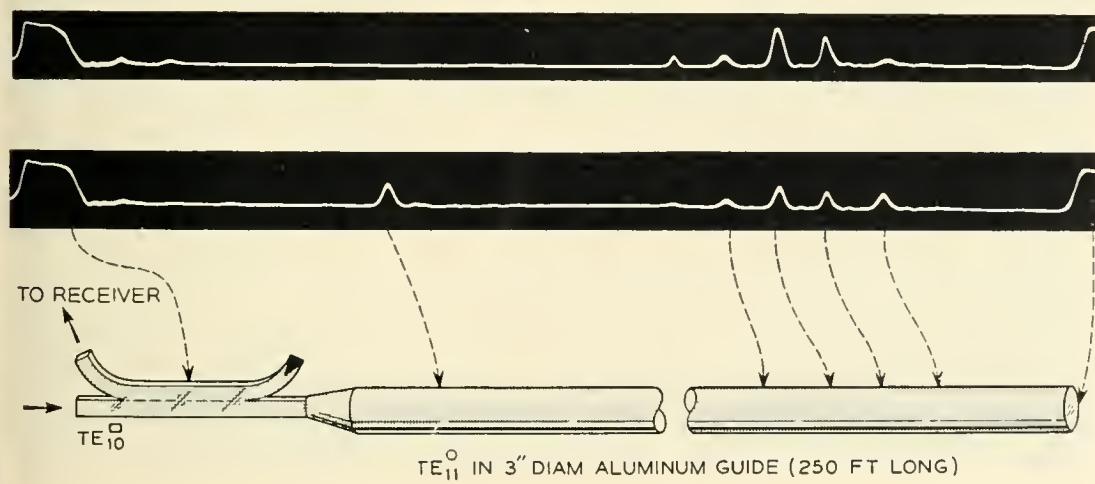


Fig. 8 — Reflections from several defective joints in a dominant (TE_{11}^0) mode waveguide. The two trace photos are for polarizations differing by 90 degrees.

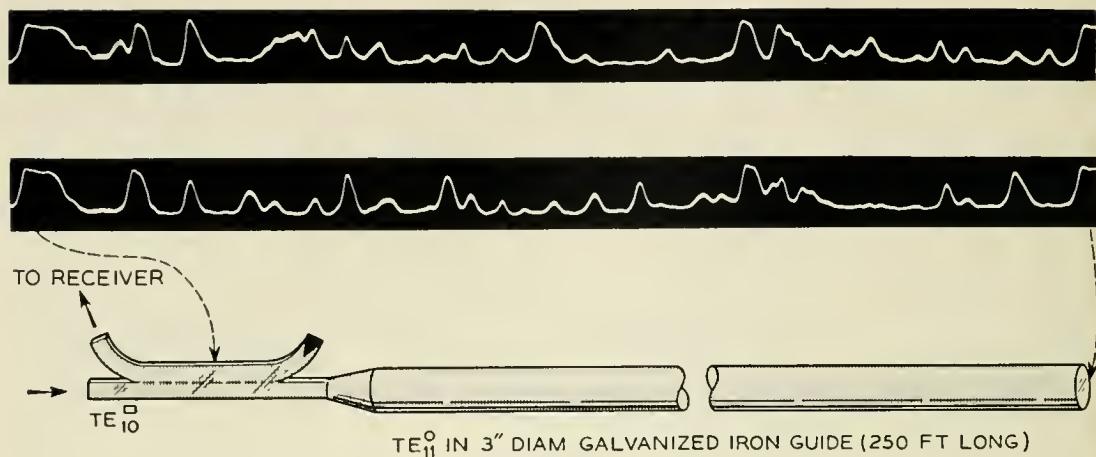


FIG. 9 — Multiple reflections from a dominant (TE₁₁) mode waveguide with a rough inside surface. The two trace photos are for polarizations differing by 90 degrees.

joints are shown here. In addition, a study was being made of the return loss of the transition piece at the throat of the antenna which connected the 3" waveguide to the square section of the horn. The waveguide sections are about 10 feet long. The overloaded pulse at the left on the traces is the leakage through the directional coupler. The

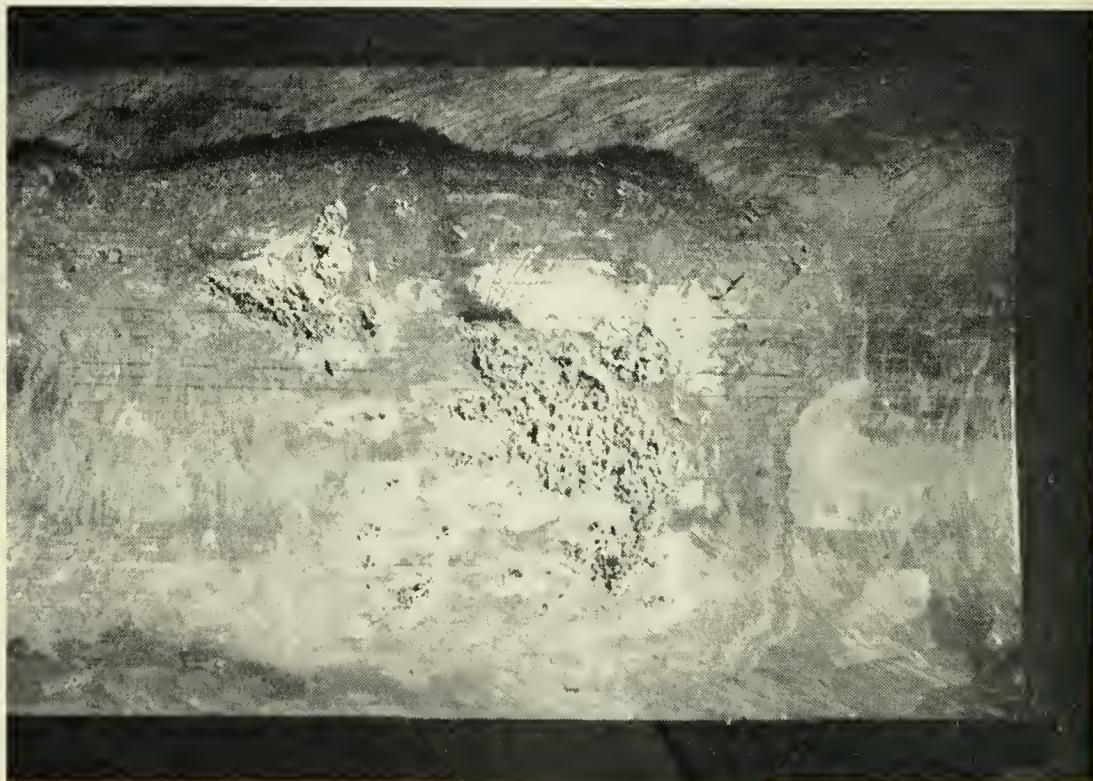


Fig. 10 — Rough inside surface of a galvanized iron waveguide which produced the reflections shown on Fig. 9.

other echoes are associated with the parts of the system from which they came by the dashed lines and arrows on the figure. A clamped joint in the line gave the reflection shown next following the initial overloaded pulse. A well made threaded coupling in which the ends of the pipe butted squarely is seen to have a very much lower reflection, scarcely observable on this trace. Since there is always reflection from the mouth and upper reflector parts of this kind of antenna, it is not possible to measure a throat transition piece alone by conventional CW methods, as the total reflected power from the system is measured. Here, use of the resolution of this short pulse equipment completely separated the reflection of the transition piece from all other reflections and made a measurement of its performance possible. In this particular case, the reflection from the transition is more than 50 db down from the incident signal which represents very good design. As can be seen,

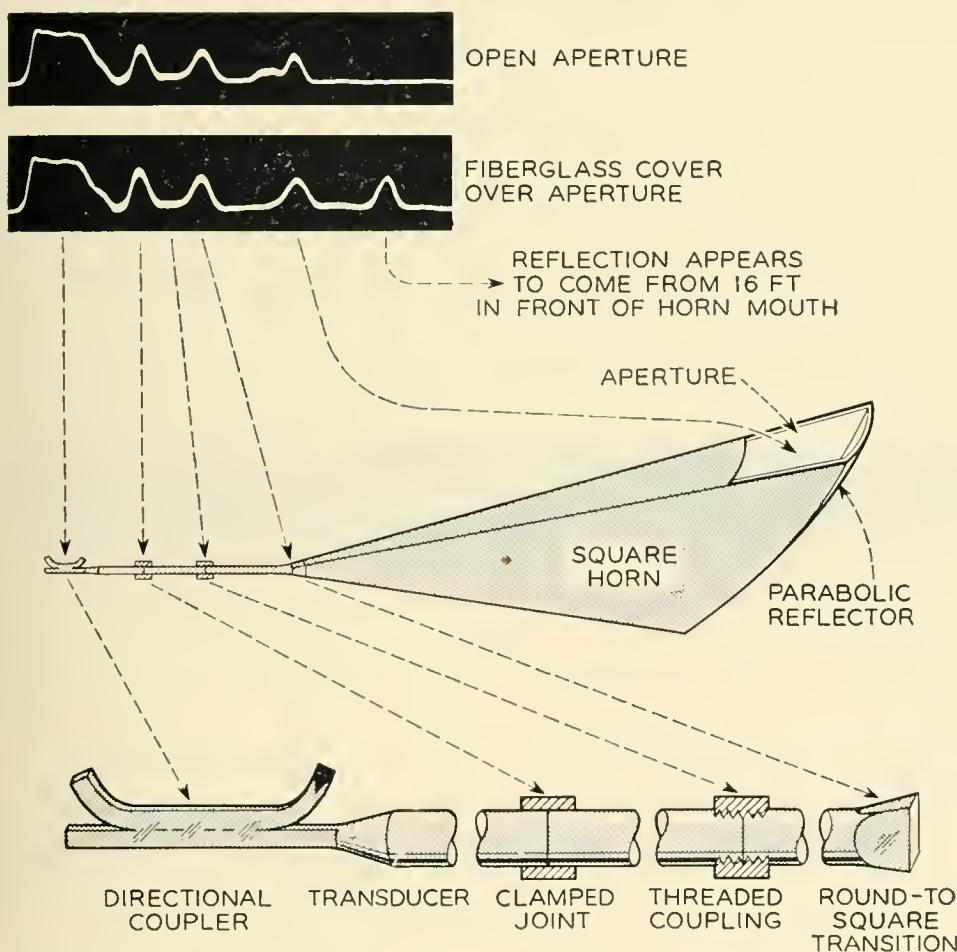


Fig. 11 — Waveguide and antenna arrangement with trace photos showing reflections from joints, transition section, and cover.

the reflection from the parabolic reflector and mouth is also quite low, and this characterizes a good antenna installation.

The extra reflected pulse on the right of the lower trace on Fig. 11 appeared when a fiberglass weatherproof cover was installed over the open mouth of the horn. This cover by itself would normally produce a troublesome reflection. However, in this antenna, it is a continuation of one of the side walls of the horn. Consequently, outgoing signals strike it at an oblique angle. Reflected energy from it is not focused by the parabolic section back at the waveguide, so the overall reflected power in the waveguide was found to be rather low. However, measuring it with this equipment, we found that an extra reflection appeared to come from a point 16 feet out in front of the mouth of the horn when the cover was in place. This is accounted for by the fact that energy reflected obliquely from this cover bounces back and forth inside the horn before getting back into the waveguide, thus traveling the extra distance that makes the measurement seem to show that it comes from 16 feet out in front.

7. SEPARATION OF MODES ON A TIME BASIS

If a pulse of energy is introduced into a moderate length of round waveguide to excite a number of modes which travel with different group velocities, and then observed farther along the line, or reflected from a piston at the end and observed at the beginning, separate pulses will be seen corresponding to each mode that is sent. This is illustrated

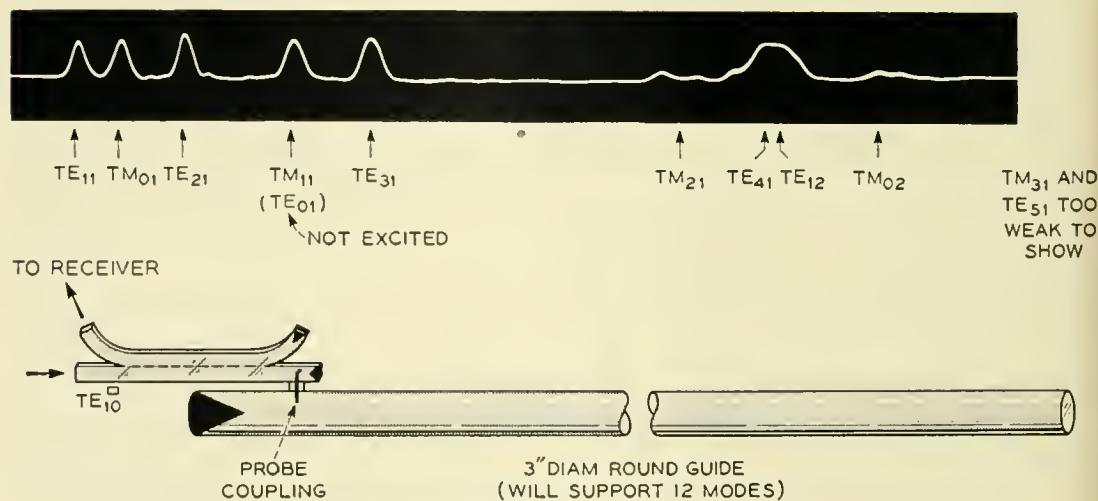


FIG. 12 — Arrangement for showing mode separation on a time basis in a multi-mode waveguide. The pulses in the trace photo have all traveled to the piston and back. The earlier outgoing pulse due to directional coupler unbalance is not shown.

in Fig. 12. In this arrangement energy was sent into the round line from a probe inserted in the side of the guide. This couples to all of the 12 modes which can be supported, with the exception of the TE_{01} circular electric mode. The sending end of the round guide was terminated. A directional coupler is connected to the sending probe so that the return from the piston at the far end can be observed on the receiver. Because of the different time that each mode takes to travel one round trip in this waveguide, which was 258 feet long, separate pulses are seen for each mode. The pulses in this figure have been marked to show which mode is being received.

The time of each pulse referred to the outgoing pulse was measured and found to check very well with the calculated time. The formula for the time of transit in the waveguide for any mode is:

$$T = \frac{L}{0.98322\sqrt{1 - v_{nm}^2}}$$

where T = time in millimicroseconds

L = length of pulse travel in feet

$v_{nm} = \lambda/\lambda_c$

λ = operating wavelength in air

λ_c = cutoff wavelength of guide for the mode involved.

TABLE I—CALCULATED AND MEASURED VALUE OF TIME FOR ONE ROUND TRIP

Mode Number	Mode Designation	Time in Millimicroseconds	
		Calculated	Measured
1	TE_{11}	545	545
2	TM_{01}	561	561
3	TE_{21}	587	587
4	TM_{11}	634	634
5	TE_{01}	634	—
6	TE_{31}	665	665
7	TM_{21}	795	793
8	TE_{41}	835	838
9	TE_{12}	838	—
10	TM_{02}	890	890
11	TM_{31}	1461	—
12	TE_{51}	1519	—

The calculated and measured value of time for one round trip is given in Table I.

In this experiment the operating wavelength was 3.35 centimeters. This was obtained by measurements based on group velocity in a number of guides as well as information about the pulse generator components. It represents an effective wavelength giving correct time of travel. The pulse occupies such a wide bandwidth that a measurement of its wavelength is difficult by the usual means.

The dashes in the measured column indicate that the mode was not excited by the probe or was too weak to measure. These modes do not appear on the oscilloscope trace photograph.

The relative pulse heights can be calculated from a knowledge of the probe coupling factors and the line loss. The probe coupling factors as given by M. Aronoff in unpublished work are expressed by the following:

For TE_{nm} modes:

$$P = 2.390 \frac{n^2}{K_{nm}^2 - n^2} \frac{\lambda_{g_{nm}}}{\lambda_{g_{11}}}$$

For TM_{nm} modes:

$$P = 1.195 \epsilon_n \frac{\lambda}{\lambda_{g_{11}}} \frac{\lambda}{\lambda_{g_{nm}}}$$

where

P = ratio of probe coupling power in mode nm to that in mode TE_{11}

n = first index of mode being calculated

K_{nm} = Bessel function zero value for mode being calculated = $\pi d/\lambda_c$

λ = wavelength in air

λ_g = wavelength in the guide for the mode involved

λ_c = cutoff wavelength of guide for the mode involved

ϵ_n = 1 for $n = 0$

ϵ_n = 2 for $n \neq 0$

d = waveguide diameter

Formulas for guide loss as given by S. A. Schelkunoff on page 390 of his book *Electromagnetic Waves* for this case where the resistivity of the aluminum guide is 4.14×10^{-6} ohms per cm cube are:

For TE_{nm} modes:

$$\alpha = 3.805 \left(\frac{n^2}{K_{nm}^2 - n^2} + v_{nm}^2 \right) (1 - v_{nm}^2)^{-1/2}$$

For TM_{nm} modes:

$$\alpha = 3.805(1 - v_{nm}^2)^{-1/2}$$

where:

α = attenuation of this aluminum guide in db

n = first index of mode being calculated

K_{nm} = Bessel function zero value for mode being calculated = $\pi d/\lambda_c$

v_{nm} = λ/λ_c

λ = operating wavelength in air

λ_c = cutoff wavelength of guide for the mode involved

d = waveguide diameter

Table II gives the calculated probe coupling factor, line loss, and relative pulse height for each mode. In the calculation of the latter, wave ellipticity and loss due to mode conversion were neglected, but the heat loss given by the preceding formulas has been increased 20 per cent for all modes, to take account of surface roughness. Relative pulse heights were obtained by subtracting the relative line loss from twice the relative probe coupling factor. The relative line loss is the number in the table minus 2.33 db, the loss for the TE_{11} mode.

The actual pulse heights on the photo of the trace on Fig. 12 are in fair agreement with these calculated values. Differences are probably due to polarization rotation in the guide (wave ellipticity) and conversion to other modes, effects which were neglected in the calculations, and which are different for different modes.

Calculated pulse heights with this guide length, except for modes near cutoff, vary less than the probe coupling factors, because line loss is high when tight probe coupling exists. This is to be expected, since both are the result of high fields near the guide walls.

The table of round trip travel time shows that the TE_{41} and TE_{12} modes are separated by only three millimicroseconds after the round trip in this waveguide. They would not be resolved as separate pulses by this equipment. However, the table of calculated pulse heights shows that the TE_{41} pulse should be about 22 db higher than the TE_{12} pulse.

TABLE II — CALCULATED PROBE COUPLING FACTOR, LINE LOSS AND PULSE HEIGHT FOR EACH MODE

Mode Number	Mode Designation	Relative Probe Coupling Factor, db	$1.2 \times$ Theoretical Line Loss, db	Calculated Relative Pulse Heights, db
1	TE ₁₁	0	2.33	0
2	TM ₀₁	+0.32	4.88	-1.91
3	TE ₂₁	+2.86	4.85	+3.20
4	TM ₁₁	+2.80	5.51	+2.42
5	TE ₀₁	-∞	1.73	-∞
6	TE ₃₁	+4.82	8.21	+3.76
7	TM ₂₁	+1.82	6.92	-0.95
8	TE ₄₁	+6.80	13.86	+2.07
9	TE ₁₂	-8.73	4.70	-19.83
10	TM ₀₂	-1.68	7.74	-8.77
11	TM ₃₁	-0.82	12.71	-12.02
12	TE ₅₁	+10.14	32.09	-9.48

Since the TE₁₂ pulse is so weak, it would not show on the trace even if it were resolved on a time basis. Coupling to the TM₀₂ mode is rather weak, and the gain was increased somewhat at its position on the trace to show its time location.

8. DELAY DISTORTION

Another effect of the wide bandwidth of the pulses used with this equipment can be observed in Fig. 12. The pulses that have traveled for a longer time in the guide are in the modes closer to cutoff, and are on the right-hand side of the oscilloscope trace. They are broadened and distorted compared with the ones on the left-hand side. This effect is due to delay distortion in the guide. This can be explained by reference to Fig. 13. On this figure the ratio of group velocity to the velocity in an unbounded medium is shown plotted as a function of frequency for each of the modes that can be propagated. The bandwidth of the transmitted pulse is indicated by the vertical shaded area. It will be noticed that the spacing of the pulses on the oscilloscope trace on Fig. 12 from left to right in time corresponds to the spacing of the group velocity curves in the bandwidth of the pulse from top to bottom. Delay distortion on these curves is shown by the slope of the line across the pulse bandwidth. If the line were horizontal, showing the same group velocity at all points in the band, there would be no delay distortion. The greater the difference in group velocity at the two edges of the band, the greater the delay distortion. The curves of Fig. 13 indicate

that there should be increasing amounts of delay distortion reading from top to bottom for the pulse bandwidth used in these experiments. The effect of this delay distortion is to cause a broadening of the pulse. Examination of the pulse pattern of Fig. 12 shows that the later pulses corresponding in mode designation to the lower curves of Fig. 13 do indeed show a broadening due to the increased delay distortion. One method of reducing the effect of delay distortion is to use frequency division multiplex so that each signal uses a smaller bandwidth. Another way, suggested by D. H. Ring, is to invert the band in a section of the waveguide between one pair of repeaters compared with that between an adjacent pair of repeaters so that the slope is, in effect, placed in the opposite direction, and delay distortion tends to cancel out, to a first order at least.

The quantitative magnitude of delay distortion has been expressed by S. Darlington in terms of the modulating base-band frequency needed to generate two side frequencies which suffer a relative phase error of 180° in traversing the line. This would cause cancellation of a single frequency AM signal, and severe distortion using any of the

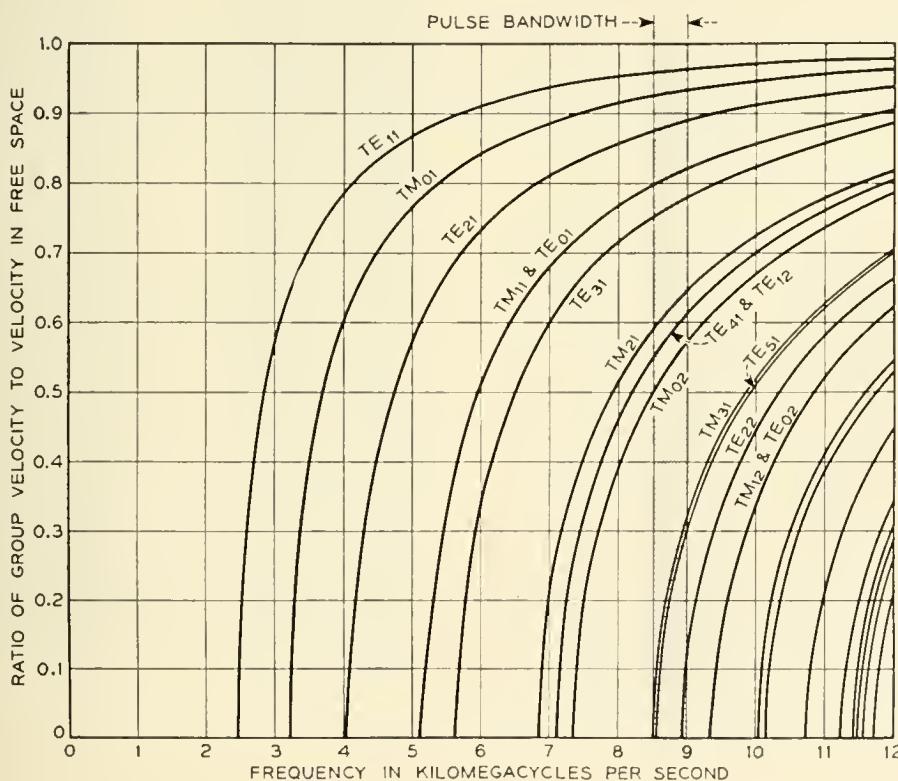


Fig. 13 — Theoretical group velocity vs. frequency curves for the 3" diameter waveguide used for the tests shown on Fig. 12. The vertical shaded area gives the bandwidth for the millimicrosecond pulses employed in that arrangement.

ordinary modulation methods. Darlington gives this formula:

$$fB = f \left(\frac{\lambda}{2L} \right)^{1/2} \frac{(1 - v_{nm}^2)^{3/4}}{v_{nm}}$$

where:

fB = base bandwidth for 180° out of phase sidebands

f = operating frequency (in same units as fB)

λ = wavelength in air

L = waveguide length (in same units as λ)

$v_{nm} = \lambda/\lambda_c$

λ_c = cutoff wavelength for the mode involved

With this equipment, the base bandwidth of the pulse is about 175 mc, and when fB from the formula above is about equal to or less than this, pulse distortion should be observed. The following Table III gives fB calculated from this formula for the arrangement shown on Fig. 12.

It is interesting to note that pulses in the TM_{11} and TE_{31} modes, for which fB is less than the 175-mc pulse bandwidth, are broadened, but not badly distorted. For the higher modes, where fB is much less than 175 mc, broadening and severe distortion are evident. Another example is given in the next section.

9. DELAY DISTORTION EQUALIZATION

If the distance which a pulse travels in a waveguide is increased, its delay distortion also increases. Since the group velocity at one edge of the band is different than at the other edge of the band, the amount by which the two edges get out of phase with each other increases with the total length of travel, causing increased distortion and pulse broadening. The Darlington formula in the previous section shows that fB varies inversely as the square root of the length of travel. This effect is shown on Fig. 14. In this arrangement the transmitter was connected to the end of a 3" diameter round waveguide 107 feet long through a small hole in the end plate. A mode filter was used so that only the TE_{01} mode would be transmitted in this waveguide. Through another small hole in the end plate polarized 90° from the first one, and rotated 90° around the plate, a directional coupler was connected as shown. The direct through guide of this directional coupler could be short circuited with a waveguide shorting switch. Energy reflected from this

TABLE III — CALCULATED VALUES OF fB FOR THE ARRANGEMENT SHOWN IN FIG. 12

Mode Number	Mode Designation	fB Megacycles	Remarks
1	TE_{11}	324.0	
2	TM_{01}	237.7	
3	TE_{21}	174.9	
4	TM_{11}	124.1	
5	TE_{01}	124.1	Not excited
6	TE_{31}	105.2	
7	TM_{21}	65.9	
8	TE_{41}	59.1	
9	TE_{12}	58.6	Very weakly excited
10	TM_{02}	51.8	
11	TM_{31}	21.3	Not observed
12	TE_{51}	20.0	Not observed

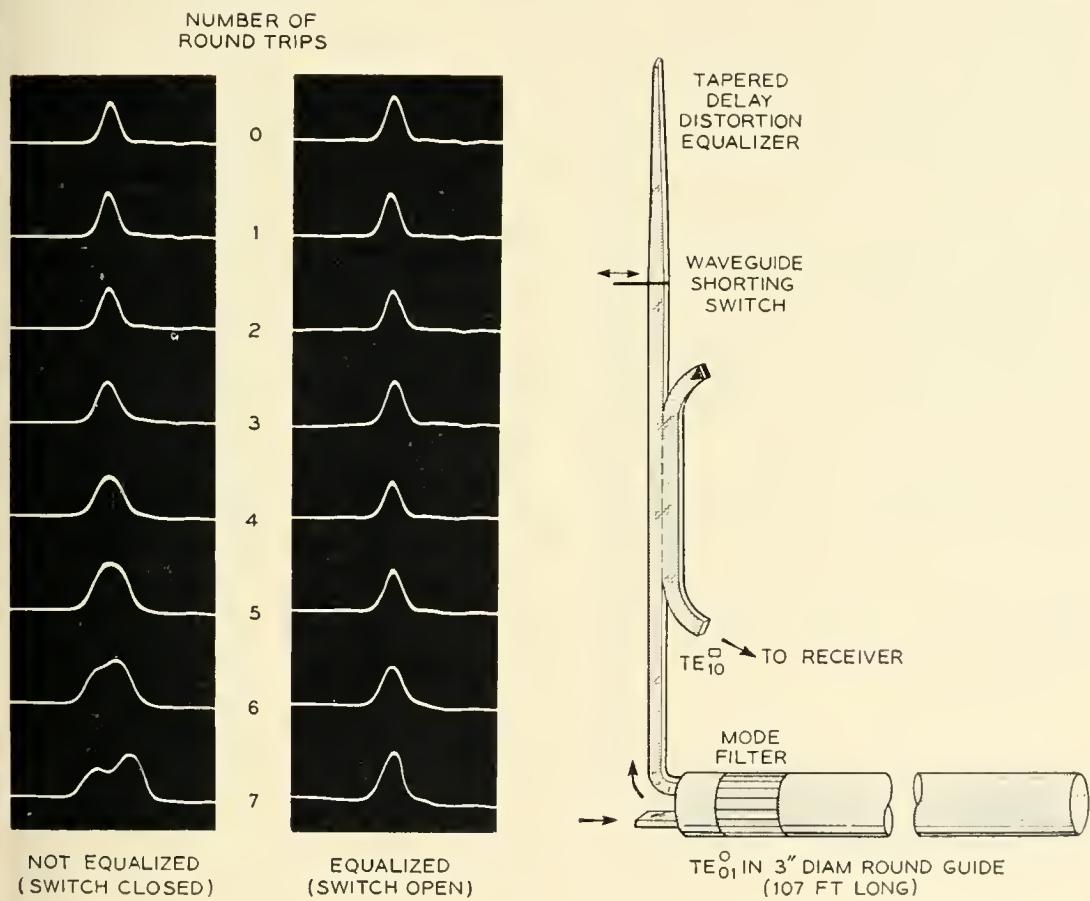


Fig. 14 — The left-hand series of pulses shows the build up of delay distortion with increasing number of round trips in a long waveguide. The right-hand series shows the improvement obtained with the tapered delay distortion equalizer shown at the right.

switch was then taken through the directional coupler to the receiver as shown by the output arrow. The series of pulses at the left-hand photograph of the oscilloscope traces was taken with this waveguide shorting switch closed. The top pulse shows the direct leakage across the inside of the end plate before it has traveled through the 3" round guide. The next pulse is marked one round trip, having gone therefore 214 feet in the TE_{01} mode in the round waveguide. The successive pulses have traveled more round trips as shown by the number in the center between the two photographs. The effect of increased delay distortion broadening and distorting the pulse can be seen as the numbers increase. The values of fB from the Darlington formula in the previous section for these lengths are given in Table IV.

It will be noticed that pulse broadening, and eventually severe distortion, occurs as fB decreases much below the 175-mc pulse bandwidth. The effect is gradual, and not too bad a pulse shape is seen until fB is about half the pulse bandwidth, although broadening is very evident earlier.

When the waveguide short-circuiting switch was opened so that the tapered delay distortion equalizer was used to reflect the energy, instead of the switch, the series of pulses at the right was observed on the indicator. It will be noted that there is much less distortion of these pulses, particularly toward the bottom of the series. The ones at the top have less distortion than would be expected, probably because of frequency modulation of the injected pulse. The equalizer consists of a long gradually tapered section of waveguide which has its size reduced to a point beyond cutoff for the frequencies involved. Reflection takes place at the point of cutoff in this tapered guide. For the high frequency part of the pulse bandwidth, this point is farther away from the shorting switch than for the low frequency part of the bandwidth. Consequently, the high frequency part of the pulse travels farther in one round trip into this tapered section and back than the low frequency part of

TABLE IV — VALUES OF fB FROM THE DARLINGTON FORMULA FOR THE ARRANGEMENT SHOWN IN FIG. 14

Round Trip Number	fB Megacycles	Round Trip Number	fB Megacycles
1	185.8	6	75.8
2	131.4	7	70.2
3	107.3	8	65.7
4	92.9	9	61.9
5	83.1	10	58.7

the pulse. This increased time of travel compensates for the shorter time of travel of the high frequency edge of the band in the 3" round waveguide, so equalization takes place. Since this waveguide close to cutoff introduces considerable delay distortion by itself, the taper effect must be made larger in order to secure the equalization. This can be done by making the taper sufficiently gradual. This type of equalizer introduces a rather high loss in the system. For this reason it might be used to predistort the signal at an early level in a repeater system. Equalization by this method was suggested by J. R. Pierce.

9. MEASURING MODE CONVERSION FROM ISOLATED SOURCES

One of the important uses of this equipment has been for the measurement of mode conversion. W. D. Warters has cooperated in developing techniques and carrying out such measurements. One of the problems in the design of mode filters used for suppressing all modes except the circular electric ones in round multimode guides is mode conversion. Since these mode filters have circular symmetry, conversion can take place only to circular electric modes of order higher than the TE_{01} mode. This conversion is, however, a troublesome one, since these higher order circular modes cannot be suppressed by the usual type of filter.

An arrangement for measuring mode conversion at such mode filters from the TE_{01} to the TE_{02} mode is being used with the short pulse equipment. This employs a 400-foot long section of the 5" diameter line. Because the coupled-line transducer available had too high a loss to TE_{02} , a combined $TE_{01} - TE_{02}$ transducer was assembled. It uses one-half of the round waveguide to couple to each mode. Fig. 15 shows this device.

The use of this transducer and line is illustrated in Fig. 16. Pulses in the TE_{01} mode are sent into the waveguide by the upper section of the transducer as shown. Some of the TE_{01} energy goes directly across to the TE_{02} transducer and appears as the outgoing pulse with a level down about 32 db. This is useful as a time reference in the system and is shown as the outgoing pulse in the photo of the oscilloscope trace above. The main energy in the TE_{01} mode propagates down the line as shown by dashed line 2, which is the path of this wave. Most of this energy goes all the way to the reflecting piston at the far end and then returns to the TE_{02} transducer where it gives a pulse which is marked TE_{01} round trip on the trace photograph above. Two thirds of the way from the sending end to the piston, the mode filter being measured is inserted in the line. When the TE_{01} mode energy comes to this mode filter, a small amount of it is converted to the TE_{02} mode. This

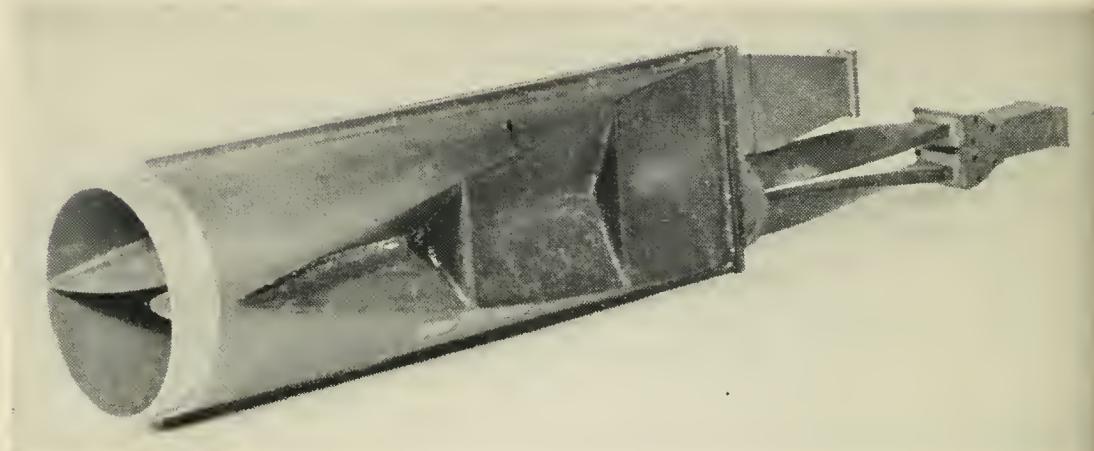


Fig. 15 — A special experimental transducer for injecting the TE_{01} mode and receiving the converted TE_{02} mode in a 5" diameter waveguide.

continues to the piston by path 4 (with dashed lines and crosses) and then returns and is received by the TE_{02} part of the transducer. This appears on the trace photo as the TE_{02} first conversion. When the main TE_{01} energy reflected by the piston comes back to the mode filter, conversion again takes place to TE_{02} . This is shown by path 3 having dashed lines and circles. This returns to the TE_{02} part of the transducer and appears on the trace photo as the TE_{02} second conversion. In addition, a small amount of energy in the TE_{02} mode is generated by the TE_{01} upper part of the transducer. It is shown by path 5, having

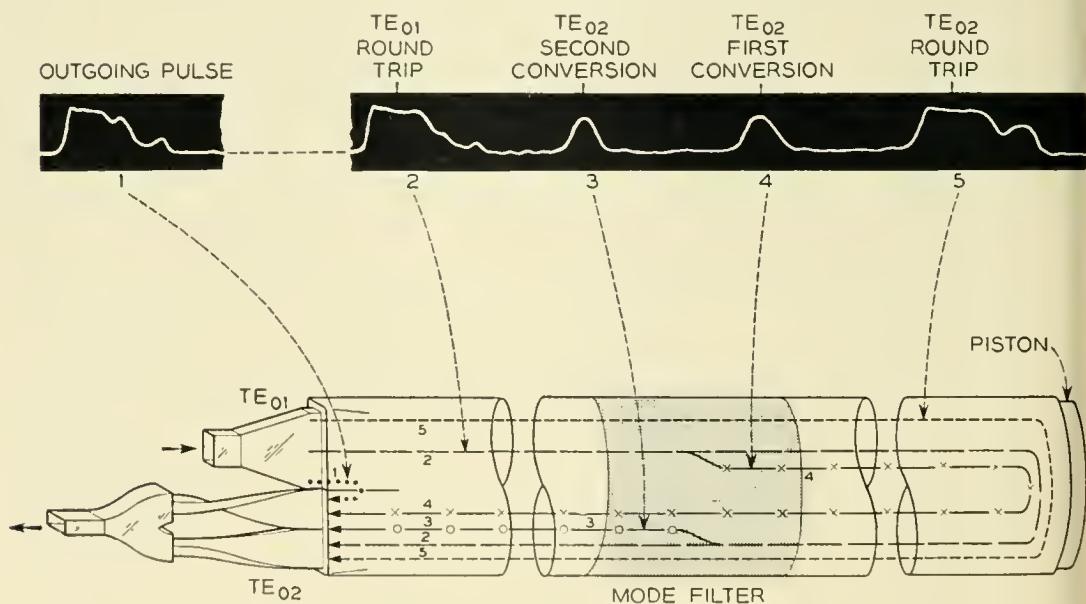


Fig. 16 — Trace photos and waveguide paths traveled when measuring TE_{01} to TE_{02} mode conversion at a mode filter with the transducer shown on Fig. 15.

short dashes. This goes down through the waveguide to the far end piston and back, and is received by the TE_{02} transducer and shown as the pulse marked TE_{02} round trip. The pulse marked TE_{01} round trip has a time separation from the outgoing pulse which is determined by the group velocity of TE_{01} waves going one round trip in the guide. The TE_{02} round trip pulse appears at a time corresponding to the group velocity of the TE_{02} mode going one round trip in the guide. Spacing the node filter two-thirds of the way down produces the two conversion pulses equally spaced between these two as shown in Fig. 16. The first conversion pulse appears at a time which is the sum of the time taken for the TE_{01} to go down to the filter and the TE_{02} to go from the filter to the piston and back to the receiver. Because of the slower velocity of the TE_{02} , this appears at the time shown, since it was in the TE_{02} mode for a longer time than it was in the TE_{01} mode. The second conversion, which happened when TE_{01} came back to the mode filter, comes earlier in time than the first conversion, since the path for this signal was in the TE_{01} mode longer than it was in the TE_{02} mode. This arrangement gives very good time separation, and makes possible a measurement of the amount of mode conversion taking place in the mode filters. Mode conversion from TE_{01} to TE_{02} as low as 50 to 55 db down, can be measured with this equipment.

Randomly spaced single discontinuities in long waveguides can be located by this technique if they are separated far enough to give individually resolved short pulses in the converted mode. Fig. 17 shows

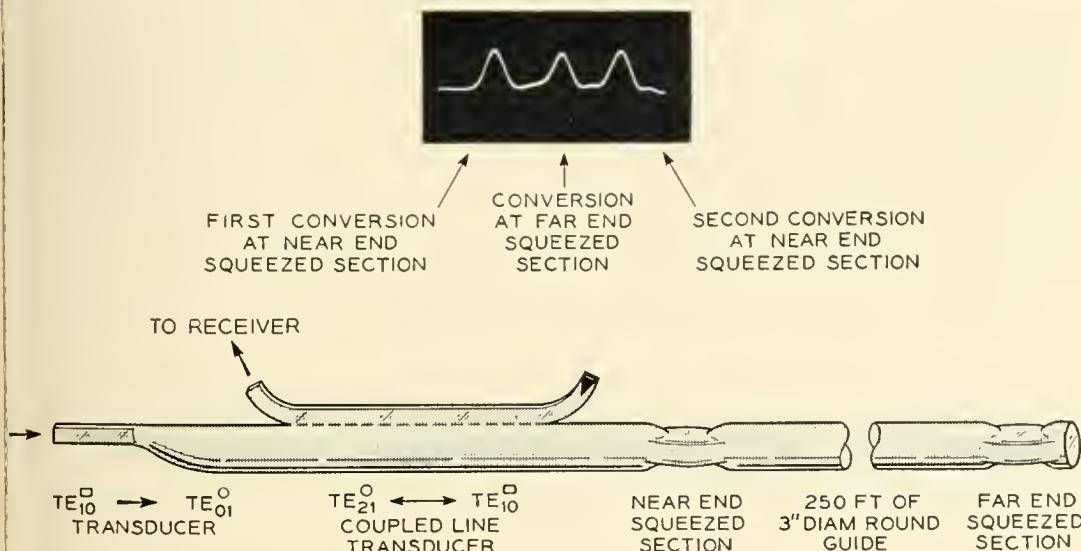


Fig. 17 — Arrangement used to explain the measurement and location of mode conversion from isolated sources. A deliberately squeezed section was placed at each end of the long waveguide, producing the pulses shown in the trace photo.

an arrangement having oval sections deliberately placed in the waveguide in order to explain the method. Pure TE_{01} excitation is used, and the converted TE_{21} mode observed with a coupled line transducer giving an output for that mode alone.

Let us consider first what would happen with the far-end squeezed section alone, omitting the near-end squeezed section from consideration. The injected TE_{01} mode signal would then travel down the 250 feet of 3" diameter round waveguide to the far end with substantially no mode conversion at the level being measured. At this point it goes through the squeezed section. Conversion now takes place from the TE_{01} mode to the TE_{21} mode. Both these modes after reflection from the piston travel back up the waveguide to the sending end. The group velocity of the TE_{21} mode is higher than the group velocity of the TE_{01} mode, so energy in these two modes separates, and if a coupling system were used to receive energy in both modes, two pulses would appear, with a time separation between them. In this case, since the receiver is connected to the line through the coupled line transducer which is responsive only to the TE_{21} mode, only one pulse is seen, that due to this mode alone. This is the center pulse in the trace photograph at the top of Fig. 17. If only one mode conversion point at the far end of the guide exists, only this one pulse is seen at the receiver. It would be spaced a distance away from the injected outgoing pulse that corresponds in time to one trip of the TE_{01} mode down to the far end and one trip of the TE_{21} mode from the far end back to the receiver.

Now let us consider what would happen if the near-end squeezed section alone were present. When the TE_{01} wave passes the oval section just beyond the coupled line transducer, conversion takes place, and the energy travels down the line in both the TE_{01} and the TE_{21} modes, at a higher group velocity in the TE_{21} mode. These two signals are reflected by the piston at the far end and return to the sending end. The TE_{21} signal comes through the coupled line transducer and appears as the pulse at the left of the photo shown on Fig. 17. Now the TE_{01} energy has lagged behind the TE_{21} energy, and when it gets back to the near-end squeezed section, a second mode conversion takes place, and TE_{21} mode energy is produced which comes through the coupled line transducer and appears at the receiver at the time of the right hand pulse. The spacing between these two pulses is equal to the difference in round trip times between the two modes.

In general, for a single conversion source occurring at any point in the line, two pulses will appear on the scope. The spacing between these pulses corresponds to the difference in group velocity between the modes.

from the point of the discontinuity down to the piston at the far end, and then back to the discontinuity. If the discontinuity is at the far end, this time difference becomes zero, and a single pulse is seen. By making a measurement of the pulse spacing, the location of a single conversion point can be determined.

In the arrangement illustrated in Fig. 17, two isolated sources of conversion existed. They were spaced far enough apart so that they were resolved by this equipment, and all three pulses were observed. The two outside pulses were due to the first conversion point. The center pulse was caused by the other squeeze, which was right at the reflecting piston. If this conversion point had been located back some distance from the piston, it would have produced two conversion pulses whose spacing could be used to determine the location of the conversion point.

The coupled-line transducers are calibrated for coupling loss by sending the pulse through a directional coupler into the branch normally used for the output to the receiver. This gives a return loss from the directional coupler equal to twice the transducer loss plus the round trip line loss.

1. MEASURING DISTRIBUTED MODE CONVERSION IN LONG WAVEGUIDES

Measurements of mode conversion from TE_{01} to a number of other modes have been made with 5" diameter guides using this equipment. The arrangement of Fig. 18 was set up for this purpose. This is the same as Fig. 17, except that a long taper was used at the input end of the 5" waveguide, and a movable piston installed at the remote end.

One of the converted modes studied with this apparatus arrangement was the TM_{11} mode, which is produced by bends in the guide. This mode has the same velocity in the waveguide as the TE_{01} mode. Therefore energy components converted at different points in the line stay in phase with the injected TE_{01} mode from which they are converted. There is never any time separation between these modes, and a single

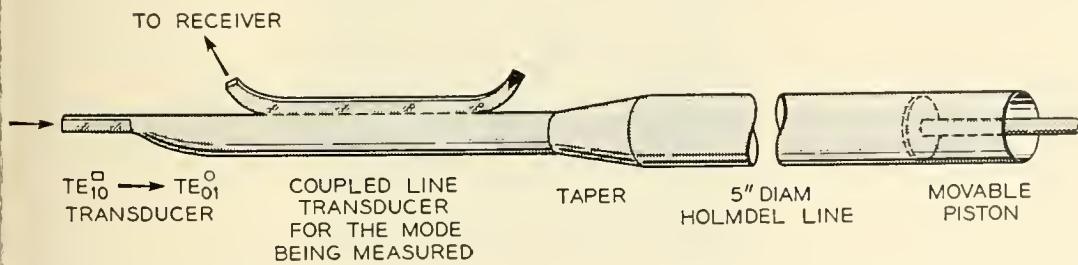


Fig. 18 — Arrangement used for measuring mode conversion in the 5" diameter waveguides at Holmdel.

narrow pulse like the transmitted one is all that appears on the indicator oscilloscope. It is not possible from this to get any information about the location or extent of the conversion points in the line. Moving the far end piston does not change the relative phases of the modes, so no changes are seen in indicator pattern or pulse level as the piston is moved. For the Holmdel waveguides, which are about 500 feet long, the total round trip TM_{11} mode converted level varies from 32 to 36 db below the input TE_{01} mode level over a frequency range from 8,800 to 9,600 me per second.

All the other modes have velocities that are different than that of the TE_{01} mode. When mode conversion takes place at many closely spaced points along the waveguide, the pulses from the various sources overlap, and phasing effects take place. In general, a filled-in pulse much longer than the injected one is observed. The maximum possible, but not necessary, pulse length is equal to the difference in time required for the TE_{01} mode and the converted mode to travel the total waveguide length being observed. The phasing effects within the broadened pulse change its height and shape as a function of frequency and line length.

Measurements of mode conversion from TE_{01} to TE_{31} in these waveguides illustrate distributed sources and piston phasing effects. The TE_{31} mode has a group velocity 1.4 per cent slower than the TE_{01} mode. For a full round trip in the 500-foot lines, assuming conversion at the input end, this causes a time separation of about two and one half pulse widths between these two modes. The received pulse is about two and a half times as long as the injected pulse, indicating rather closely spaced sources over the whole line length. For one far-end piston position, the received pattern is shown as the upper trace in Fig. 19. As the piston is moved, the center depressed part of the trace gradually

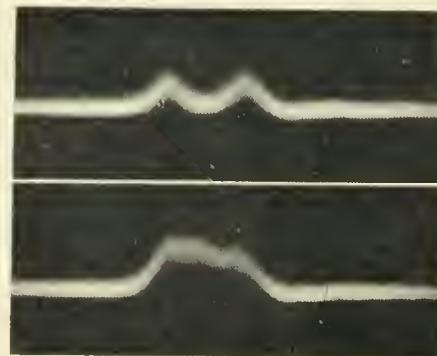


Fig. 19—Received pulse patterns with the arrangement of Fig. 18 used for studying conversion to the TE_{31} mode.

rises until the pattern shown in the lower trace is seen. As the piston is moved farther in the same direction the trace gradually changes to have the appearance of the upper photo again. Moving the far-end piston changes the phase of energy on the return trip, and thus it can be made to add to, or nearly cancel out, conversion components that originated ahead of the piston. When the time separation becomes great enough to prevent overlapping in the pulse width, phasing effects cannot take place, therefore, the beginning and end of the spread-out received pulse are not affected by moving the piston. Energy converted at the sending end of the guide travels the full round trip to the piston and back in the slower TE_{31} mode, and thus appears at the latest time, which is at the right-hand end of the received pulse. Conversion at the piston end returns at the center of the pulse, and conversion on the return trip comes at earlier times, at the left-hand part of the pulse. The TE_{01} mode has less loss in the guide than the TE_{31} mode. Since the energy in the earlier part of the received pulse spent a greater part of the trip in the lower loss TE_{01} mode before conversion, the output is higher here, and slopes off toward the right, where the later returning energy has gone for a longer distance in the higher loss mode. The pulse height at the maximum shows the converted energy from that part of the line to be between 30 and 35 db below the incident TE_{01} energy level over the measured bandwidth.

Measurements of mode conversion from TE_{01} to TE_{21} in these waveguides show these same effects, and also a phasing effect as a function of frequency. The TE_{21} mode has a group velocity 2.4 per cent faster than the TE_{01} mode. For a full round trip in the guides, this is a time separation of about four pulse widths between the modes. At one frequency and one far-end piston position, the TE_{21} response shown as the top trace of Fig. 20 was obtained. Moving the far-end piston gradually changed this to the second trace from the top, and further piston motion changed it back again. This is the same kind of piston phasing effect observed in the TE_{31} mode conversion studies. The irregular top of this broadened pulse indicates fewer conversion points than for the TE_{31} mode, or phasing effects along the guide length. Since the TE_{21} mode has a higher group velocity than the TE_{01} mode, energy converted at the beginning of the guide returns at the earlier or left-hand part of the pulse, and conversions on the return trip, having traveled longer in the slower TE_{01} mode, are on the right-hand side of the pulse. This is just the reverse of the situation for the TE_{31} mode. Since the loss in the TE_{21} mode is higher than in the TE_{01} mode, the right side of this broadened pulse is higher than the left side, as the energy in the left side has

gone further in the higher loss TE_{21} mode. Conversions from the piston end of the guide return in the center of the pulse, and only in this region do piston phasing effects appear. As the frequency is changed the pattern changes, until it reaches the extreme shape shown in the next-to-the-bottom trace, with this narrower pulse coming at a time corresponding to the center of the broadened pulse at the top. Further frequency change in the same direction returns the shape to that of the top traces. At the frequency giving the received pulse shown on the next-to-the-bottom trace, moving the far-end piston causes a gradual change to the shape shown on the lowest trace. This makes it appear as if the mode conversion were coming almost entirely from the part of the guide near the piston end at this frequency. The upper traces appear to show that more energy is converted at the transducer end of the waveguide at that frequency. It would seem that at certain frequencies some phase cancellation is taking place between conversion points spaced closely enough to overlap within the pulse width. At frequencies between the ones giving traces like this, the appearance is more like that shown for the TE_{31} mode on Fig. 19 except for the slope across the top of the pulse being reversed. The highest part of this TE_{21} pulse is

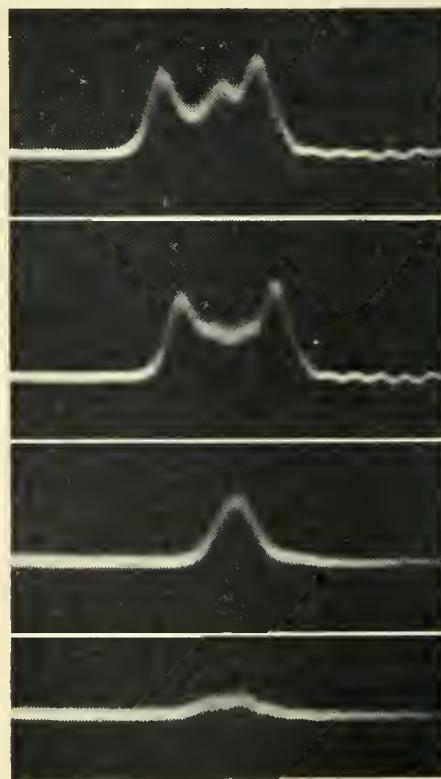


Fig. 20 — Received pulse patterns with the arrangement of Fig. 18 used for studying conversion to the TE_{21} mode.

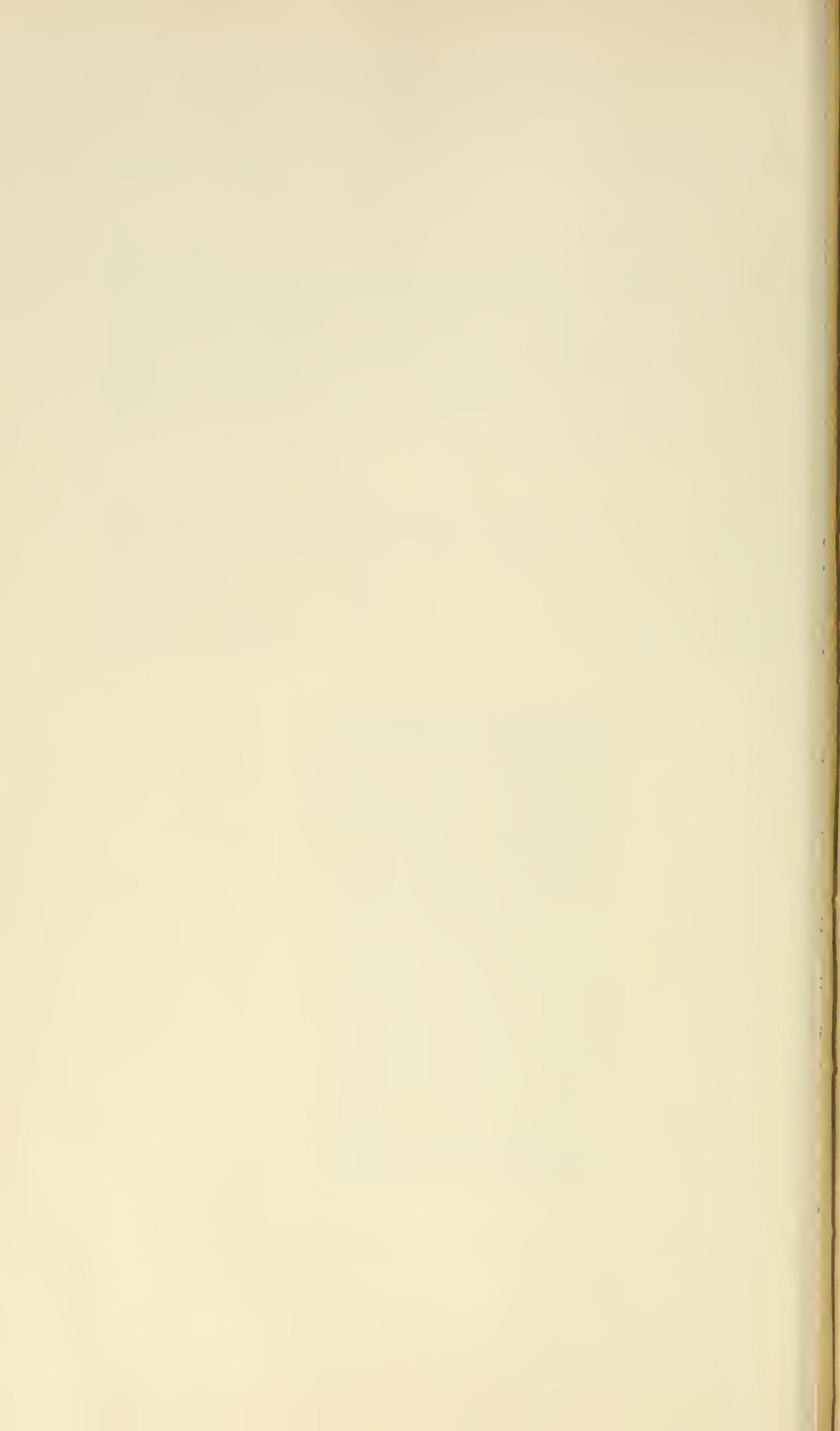
24 to 27 db below the injected TE_{01} pulse level for the 5" diameter Holmdel waveguides.

12. CONCLUDING REMARKS

The high resolution obtainable with this millimicrosecond pulse equipment provides information difficult to obtain by any other means. These examples of its use in waveguide investigations indicate the possibilities of the method in research, design and testing procedures. It is being used for many other similar purposes in addition to the illustrations given here, and no doubt many more uses will be found for such short pulses in the future.

REFERENCES

1. S. E. Miller and A. C. Beck, Low-loss Waveguide Transmission, Proc. I.R.E., **41**, pp. 348-358, March, 1953.
2. S. E. Miller, Waveguide As a Communication Medium, B. S. T. J., **33**, pp. 1209-1265, Nov., 1954.
3. C. C. Cutler, The Regenerative Pulse Generator, Proc. I.R.E., **43**, pp. 140-148, Feb., 1955.
4. S. E. Miller, Coupled Wave Theory and Waveguide Applications, B. S. T. J., **33**, pp. 661-719, May, 1954.



Experiments on the Regeneration of Binary Microwave Pulses

By O. E. DeLANGE

(Manuscript received September 7, 1955)

A simple device has been produced for regenerating binary pulses directly at microwave frequencies. To determine the capabilities of such devices one of them was included in a circulating test loop in which pulse groups were passed through the device a large number of times. Results indicate that even in the presence of serious noise and bandwidth limitations pulses can be regenerated many times and still show no noticeable deterioration. Pictures of circulated pulses are included which illustrate performance of the regenerator.

INTRODUCTION

The chief advantage of a transmission system employing binary pulses resides in the possibility of regenerating such pulses at intervals along the route of transmission to prevent the accumulation of distortion due to noise, bandwidth limitations and other effects. This makes it possible to take the total allowable deterioration of signal in each section of a long relay system rather than having to make each link sufficiently good to prevent total accumulated distortion from becoming excessive. This has been pointed out by a number of writers.¹⁻²

W. M. Goodall³ has shown the feasibility of transmitting television signals in binary form. Such transmission requires a considerable amount of bandwidth; a seven digit system, for example, would require transmission of seventy million pulses per second. This need for wide bands makes the microwave range an attractive one in which to work. S. E. Miller⁴ has pointed out that a binary system employing regeneration might prove to be especially advantageous in waveguide transmission.

¹ B. M. Oliver, J. R. Pierce and, C. E. Shannon, The Philosophy of PCM, Proc. I. R. E., Nov., 1948.

² L. A. Meacham and E. Peterson, An Experimental Multichannel Pulse Code Modulation System of Toll Quality, B. S. T. J., Jan. 1948.

³ W. M. Goodall, Television by Pulse Code Modulation, B. S. T. J., Jan., 1951.

⁴ S. E. Miller, Waveguide as a Communication Medium, B. S. T. J., Nov., 1954.

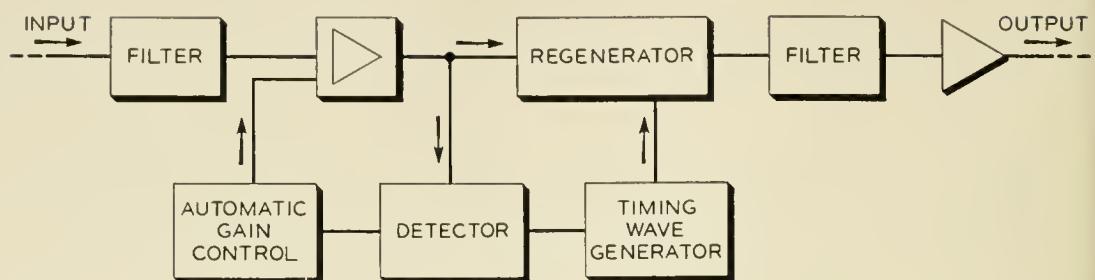


Fig. 1 — A typical regenerative repeater shown in block form.

That the Bell System is interested in the long-distance transmission of television and other broad-band signals is evident from the number of miles of such broad-band circuits, both coaxial cable and microwave radio,⁵ now in service. These circuits provide high-grade transmission because each repeater was designed to have a very flat frequency characteristic and linear phase over a considerable bandwidth. Furthermore, these characteristics are very carefully maintained. For a binary pulse system employing regeneration the requirements on flatness of band and linearity of phase can be relaxed to a considerable degree. The components for such a system should, therefore, be simpler and less expensive to build and maintain. Reduced maintenance costs might well prove to be the chief virtue of the binary system.

Since the chief advantage of a binary system lies in the possibility of regeneration it is obvious that a very important part of such a system is the regenerative repeater employed. Fig. 1 shows in block form a typical broad-band, microwave repeater. Here the input, which might come from either a radio antenna or from a waveguide, is first passed through a proper microwave filter then amplified, probably by a traveling-wave amplifier. The amplified pulses of energy are regenerated, filtered, amplified and sent on to the next repeater. The experiment to be described here deals primarily with the block labeled "Regenerator" on Fig. 1.

In these first experiments one of our main objectives was to keep the repeater as simple as possible. This suggests regeneration of pulses directly at microwave frequency, which for this experiment was chosen to be 4 kmc. It was suggested by J. R. Pierce and W. D. Lewis, both of Bell Telephone Laboratories, that further simplification might be made possible by accepting only partial instead of complete regeneration. This suggestion was adopted.

For the case of complete regeneration each incoming pulse inaugurates a new pulse, perfect in shape and correctly timed to be sent on to the

⁵ A. A. Roetken, K. D. Smith and R. W. Friis, The TD-2 System, B. S. T. J., Oct., 1951, Part II.

next repeater. Thus noise and other disturbing effects are completely eliminated and the output of each repeater is identical to the original signal which entered the system. For the case of partial regeneration incoming pulses are retimed and reshaped only as well as is possible with simple equipment. Obviously the difference between complete and partial regeneration is one of degree.

One object of the experiment was to determine how well such a partial regenerator would function and what price must be paid for employing partial instead of complete regeneration. The regenerator developed consists simply of a waveguide hybrid junction with a silicon crystal diode in each side arm. It appears to meet the requirement of simplicity in that it combines the functions of amplitude slicing and pulse retiming in one unit. A detailed description of this unit will be given later. Although the purpose of this experiment was to determine what could be accomplished in a very simple repeater we must keep in mind that superior performance would be obtained from a regenerator which approached more nearly the ideal. For some applications the better regenerator might result in a more economical system even though the regenerator itself might be more complicated and more expensive to produce.

METHOD OF TESTING

The regeneration of pulses consists of two functions. The first function is that of removing amplitude distortions, the second is that of restoring each pulse to its proper time. The retiming problem divides into two parts the first of which is the actual retiming process and the second that of obtaining the proper timing pulses with which to perform this function. In a practical commercial system timing information at a repeater would probably be derived from the incoming signal pulses. There are a number of problems involved in this recovery of timing pulses. These are being studied at the present time but were avoided in the experiment described here by deriving such information from the local synchronizing gear.

Since the device we are dealing with only partially regenerates pulses it is not enough to study the performance of a single unit — we should like to have a large number operating in tandem so that we can observe what happens to pulses as they pass through one after another of these regenerators. To avoid the necessity of building a large number of units the pulse circulating technique of simulating a chain of repeaters was employed. Fig. 2 shows this circulating loop in block form.

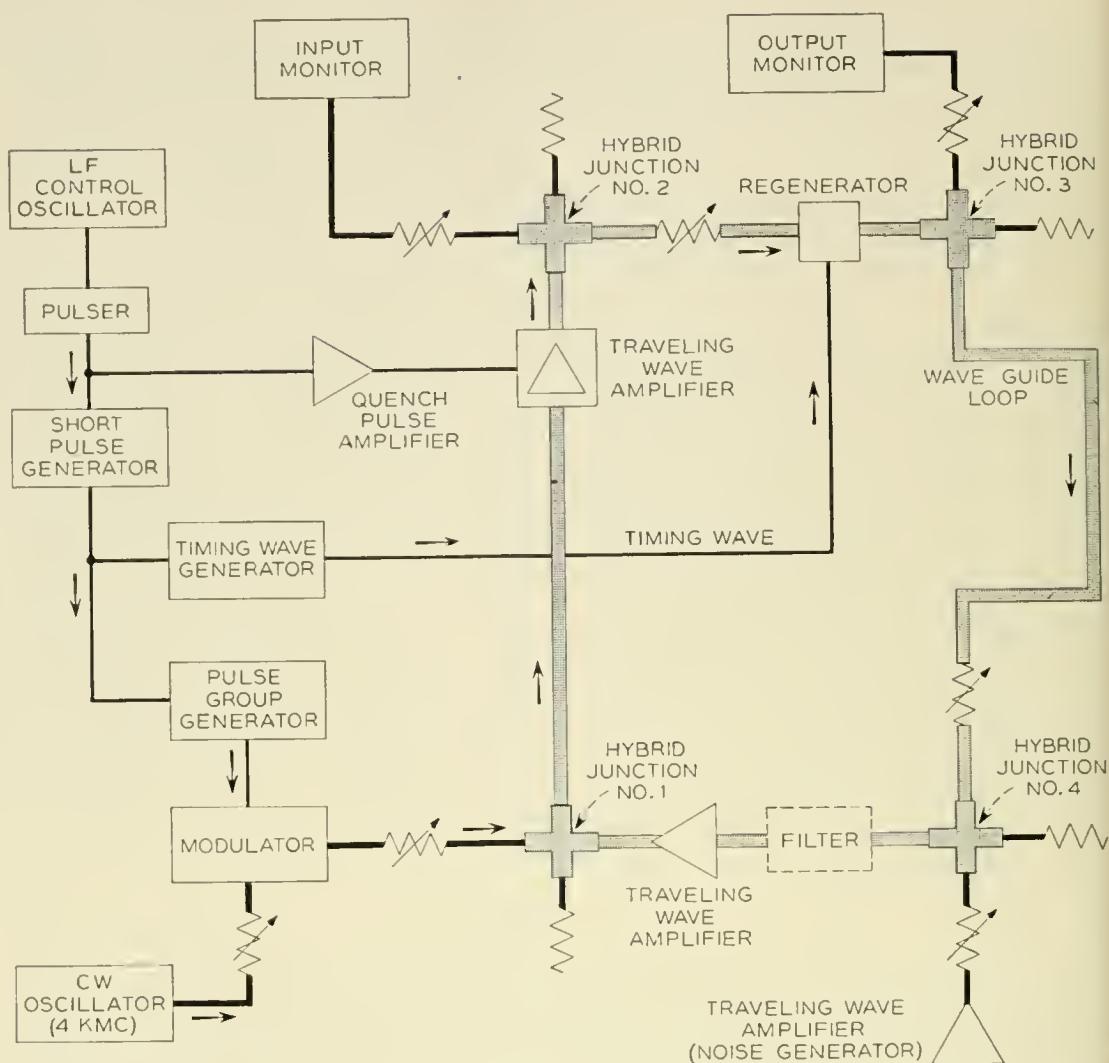


Fig. 2 — The circulating loop.

To provide RF test pulses for this loop the output of a 4 kmc, cw oscillator is gated by baseband pulse groups in a microwave gate or modulator. The resultant microwave pulses are fed into the loop (heavy line) through hybrid junction No. 1. They are then amplified by a traveling-wave amplifier the output of which is coupled to the pulse regenerator through another hybrid junction (No. 2). The purpose of this hybrid is to provide a position for monitoring the input to the regenerator. A monitoring position at the output of the regenerator is provided by a third hybrid, the main output of which feeds a considerable length of waveguide which provides the necessary loop delay. At the far end of the waveguide another hybrid (No. 4) makes it possible to feed noise, which is derived from a traveling-wave amplifier, into the loop. The combined output after passing through a band pass filter is ampli-

fied by another traveling-wave amplifier and fed back into the loop input thus completing the circuit.

The synchronizing equipment starts out with an oscillator going at approximately 78 kc. A pulse generator is locked in step with this oscillator. The output of the pulser is a negative 3 microsecond pulse as shown in Fig. 3A. After being amplified to a level of about 75 volts this pulse is applied to the helix of the first traveling-wave tube to reduce the gain of this tube during the 3-microsecond interval. Out of each 12.8 μ sec interval pulses are allowed to circulate for 9.8 μ sec but are blocked for the remaining 3 μ sec thus allowing the loop to return to the quiescent condition once during each period as shown on Figs. 3A and 3C.

The 3 μ sec pulse also synchronizes a short-pulse generator. This unit delivers pulses which are about 25 millimicroseconds long at the base and spaced by 12.8 μ sec, i.e., with a repetition frequency of 78 kc. See Fig. 3B.

In order to simulate a PCM system it was decided to circulate pulse

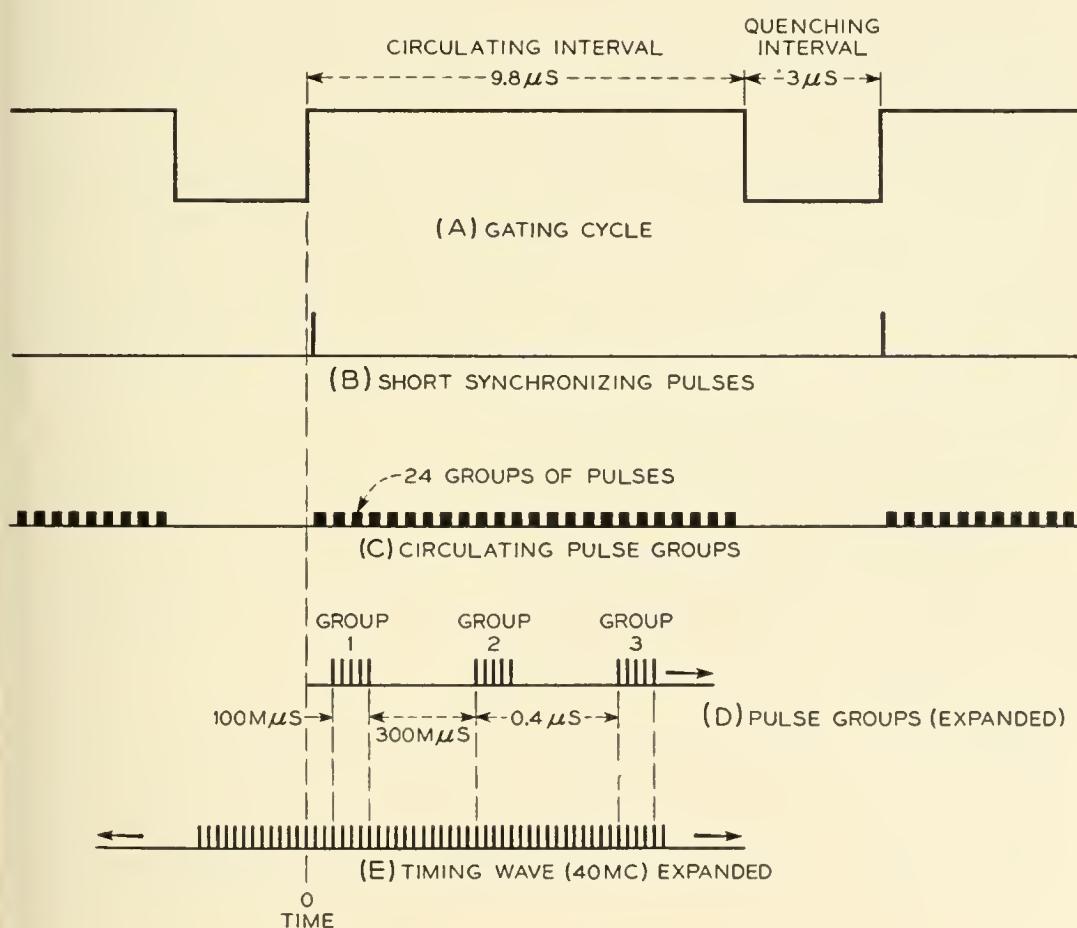


Fig. 3 — Timing events in the circulating loop.

groups rather than individual pulses through the system. These were derived from the pulse group generator which is capable of delivering any number up to 5 pulses for each short input pulse. These pulses are about 15 milli-microseconds long at the base and spaced 25 milli-microseconds apart. The amplitude of each of these pulses can be adjusted independently to any value from zero to full amplitude making it possible to set up any combination of the five pulses. These are the pulses which are used to gate, or modulate, the output of the 4-kmc oscillator.

The total delay around the waveguide loop including TW tubes, etc., was $0.4\mu\text{sec}$ or 400 milli-microseconds. This was sufficient to allow time between pulse groups and yet short enough that groups could circulate 24 times in the available $9.8\mu\text{sec}$ interval. This can be seen from Figs. 3C and 3D. The latter figure shows an expanded view of circulating pulse groups. The pulses in Group 1 are inserted into the loop at the beginning of each gating cycle, the remaining groups result from circulation around the loop.

When all five pulses are present in the pulse groups the pulse repetition frequency is 40 mc. (Pulse interval 25 milli-microseconds). For this condition timing pulses should be supplied to the regenerator at the rate of 40 million per second. These pulses are supplied continuously and not in groups as is the case with the circulating pulses. See Fig. 3E. In order to maintain time coincidence between the circulating pulses and the timing pulses the delay around the loop must be adjusted to be an exact multiple of the pulse spacing. In this experiment the loop delay is equal to 16-pulse intervals. Since timing pulses are obtained by harmonic generation from the quenching frequency as will be discussed later this frequency must be an exact submultiple of pulse repetition frequency. In this experiment the ratio is 512 to 1.

Although the above discussion is based on a five-pulse group and 40-mc repetition frequency it turned out that for most of the experiments described here it was preferable to drop out every other pulse, leaving three to a group and resulting in a 20-mc repetition frequency. The one exception to this is the limited-band-width experiment which will be described later.

For all of the experiments described here timing pulses were derived from the 78-kc quenching frequency by harmonic generation. A pulse with a width of 25 milli-microseconds and with a 78-kc repetition frequency as shown in Fig. 3B supplied the input to the timing wave generator. This generator consists of several stages of limiting amplifiers all tuned to 20 mc, followed by a locked-in 20-mc oscillator. The output of the amplifier consists of a train of 20-mc sine waves with constant ampli-

tude for most of the $12.8\mu\text{sec}$ period but falling off somewhat at the end of the period. This train locks in the oscillator which oscillates at a constant amplitude over the whole period and at a frequency of 20 mc. Timing pulses obtained from the cathode circuit of the oscillator tube provided the timing waves for most of the experiments. For the experiment where a 40-me timing wave was required it was obtained from the 20 mc train by means of a frequency doubler. For this case it is necessary for the output of the timing wave generator to remain constant in amplitude and fixed in phase for the 512-pulse interval between synchronizing pulses.

In spite of the stringent requirements placed upon the timing equipment it functioned well and maintained synchronism over adequately long periods of time without adjustment.

PERFORMANCE OF REGENERATOR

Performance of the regenerator under various conditions is recorded on the accompanying illustrations of recovered pulse envelopes. The first experiment was to determine the effects of disturbances which arise at only one point in a system. Such effects were simulated by adding disturbances along with the group of pulses as they were fed into the circulating loop from the modulator. This is equivalent to having them occur at only the first repeater of the chain.

Some of the first experiments also involved the use of extraneous pulses to represent noise or distortion since these pulses could be synchronized and thus studied more readily than could random effects. In Fig. 4A the first pulse at the left represents a desired digit pulse with its amplitude increased by a burst of noise, the second pulse represents a clean digit pulse, and the third pulse a burst of noise. This group is at the input to the regenerator. Fig. 4B shows the same group of pulses after traversing the regenerator once. The pulses are seen to be shortened due to the gating, or retiming, action. There is also seen to be some amplitude correction, i.e. the two desired pulses are of more nearly the same amplitude and the undesired pulse has been reduced in relative amplitude. After a few trips through the regenerator the pulse group was rendered practically perfect and remained so for the rest of the twenty-four trips around the loop. Fig. 4C shows the group after 24 trips. In another experiment pulses were circulated for 100 trips without deterioration. Nothing was found to indicate that regeneration could not be repeated indefinitely.

Figs. 5A and 5B represent the same conditions as those of 4A and 4B

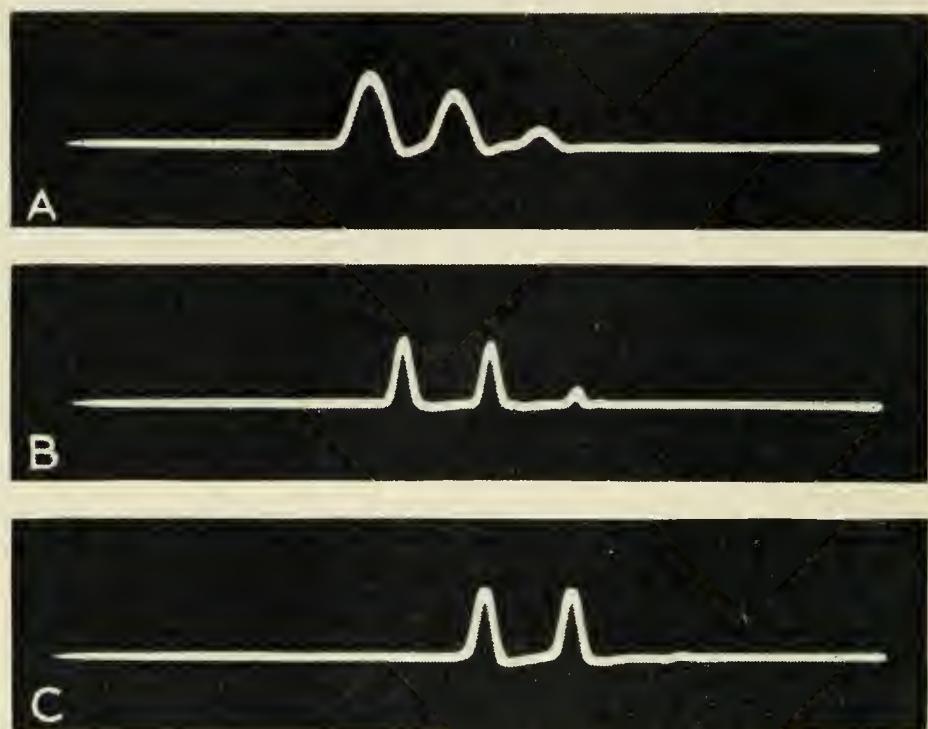


Fig. 4 — Effect of regeneration on disturbances which occur at only one repeater. A — Input to regenerator, original signal. B — Output of regenerator, first trip. C — Output of regenerator, 24th trip.

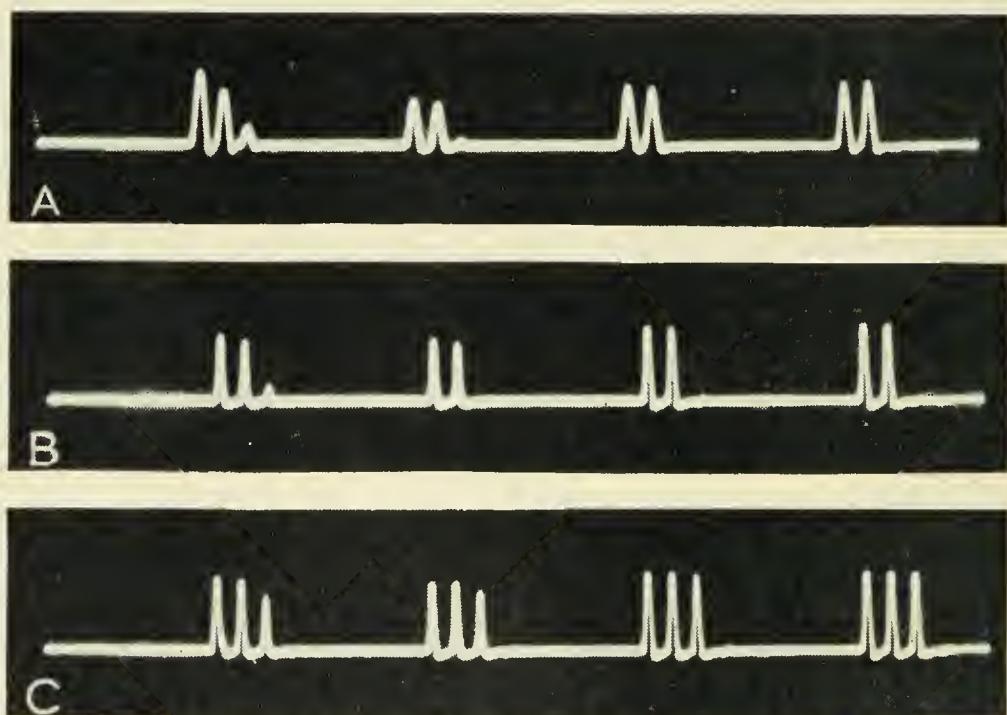


Fig. 5 — Effect of regeneration on disturbances which occur at only one repeater. A — Input to regenerator, first four groups. B — Output of regenerator, first four groups. C — Output of regenerator, increased input level.

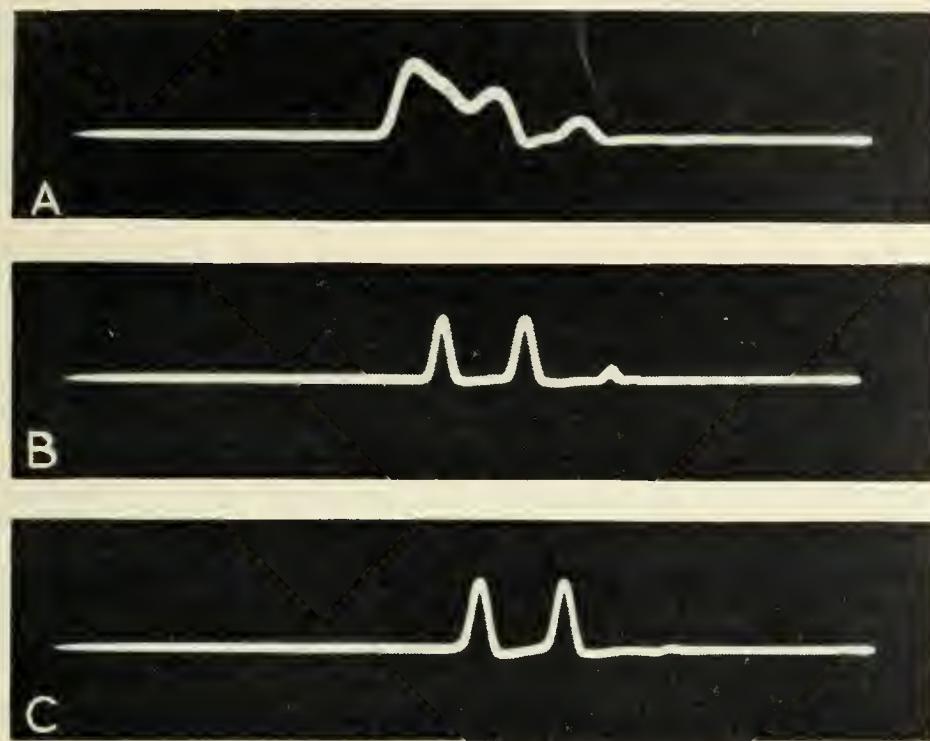


Fig. 6.—Effect of regeneration on disturbances which occur at only one repeater. A—Input to regenerator, original signal. B—Output of regenerator, first trip. C—Output of regenerator, 24th trip.

except that the oscilloscope sweep has been contracted in order to show the progressive effects produced by repeated passage of the signal through the regenerator. Fig. 5B shows that after the pulses have passed through the regenerator only twice all visible effects of the disturbances have been removed. Fig. 5C shows the effect of simply increasing the RF pulse input to the regenerator by approximately 4 db. The small "noise" pulse which in the previous case was quickly dropped out because of being below the slicing level has now come up above the slicing level and so builds up to full amplitude after only a few trips through the regenerator. Note that in the cases shown in Figs. 4 and 5 discrimination against unwanted pulses has been purely on an amplitude basis since the gate has been unblocked to pulses with amplitudes above the slicing level whenever one of these disturbing pulses was present.

For Fig. 6A conditions are the same as for Fig. 4A except that an additional pulse has been added to simulate intersymbol noise or interference. Fig. 6B indicates that after only one trip through the regenerator the effect of the added pulse is very small. After a few trips the effect is completely eliminated leaving a practically perfect group which continues on for 24 trips as shown by Fig. 6C. For the intersymbol pulse, discrimination is on a time basis since this interference occurs at a time

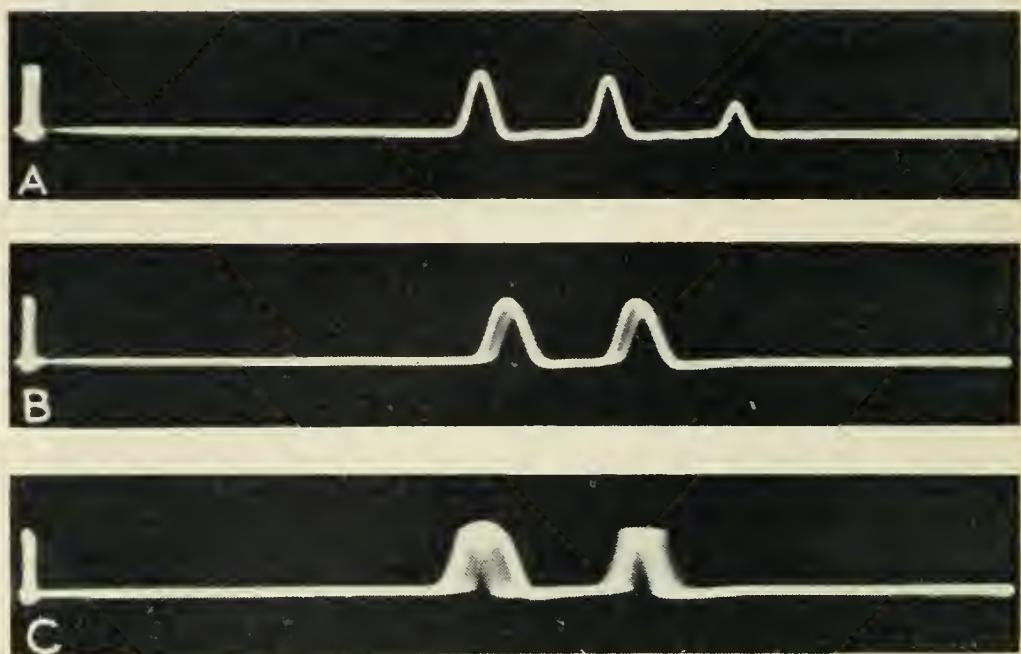


Fig. 7 — Effect of regenerating in amplitude without retiming. A — Output of regenerator, no timing, first trip. B — Output of regenerator, no timing, 10th trip. C — Output of regenerator, no timing, 23rd trip.

when no gating pulse is present and hence finds the gate blocked regardless of amplitude.

To show the need for retiming the pictures shown on Figs. 7 and 8 were taken. These were taken with the amplitude slicer in operation but with the pulses not being retimed. Figs. 7A, 7B and 7C, respectively, show the output of the slicer for the first, tenth and twenty-third trips. After ten trips, there is noticeable time jitter caused by residual noise in the system; after 23 trips this jitter has become severe though pulses are still recognizable. It should be pointed out that for this experiment no noise was purposely added to the system and hence the signal-to-noise ratio was much better than that which would probably be encountered in an operating system. For such a system we would expect time jitter effects to build up much more rapidly. For Fig. 8 conditions are the same as for Fig. 7 except that the pulse spacing is decreased by the addition of an extra pulse at the input. Now, after ten trips, time jitter is bad and after 23 trips the pulse group has become little more than a smear. This increased distortion is probably due to the fact that less jitter is now required to cause overlap of pulses. There may also be some effects due to change of duty cycle. For Fig. 9 there was neither slicing nor retiming of pulses. Here, pulse groups deteriorate very rapidly to nothing more than blobs of energy. Note that there is an increase of

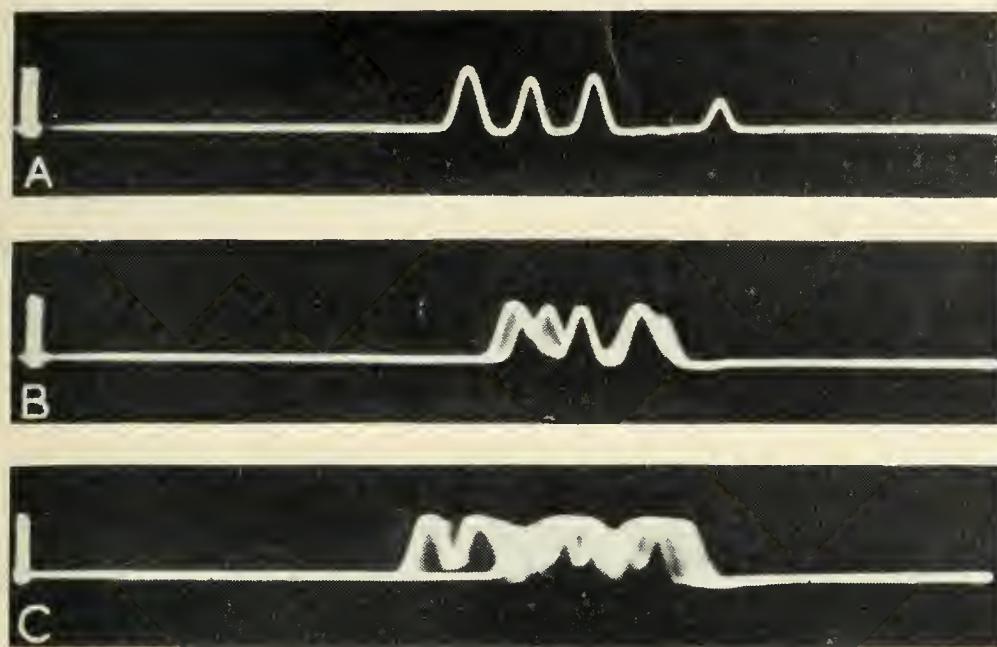


Fig. 8 — Effect of regenerating in amplitude without retiming. A — Output of regenerator, no timing, first trip. B — Output of regenerator, no timing, 10th trip. C — Output of regenerator, no timing, 23rd trip.

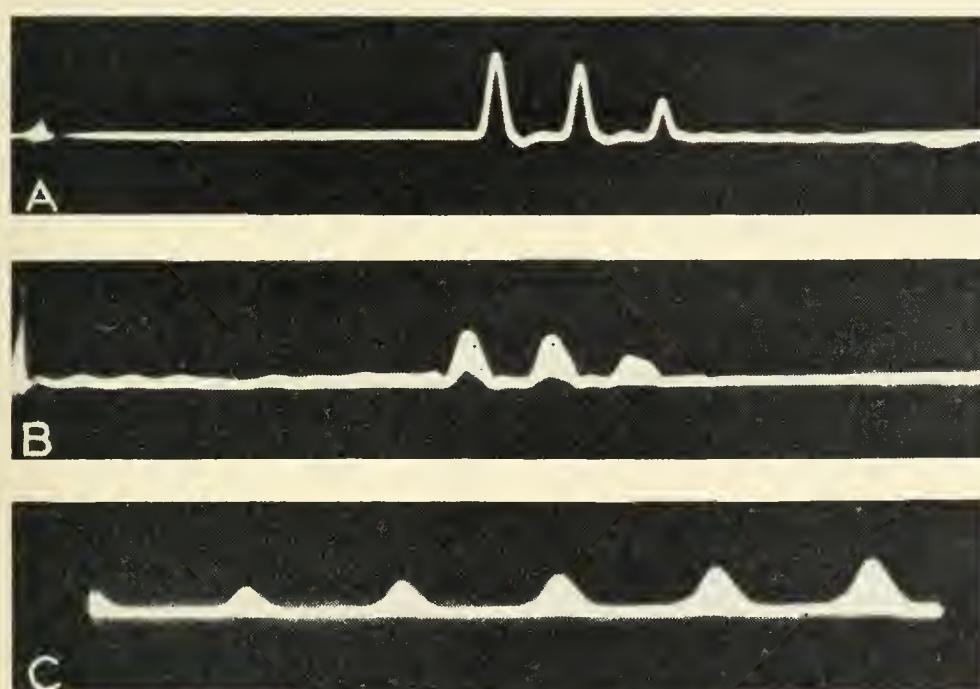


Fig. 9 — Pulses circulating through the loop without regeneration. A — Original input. B — 4th trip without regeneration. C — 20th to 24th trip without regeneration.

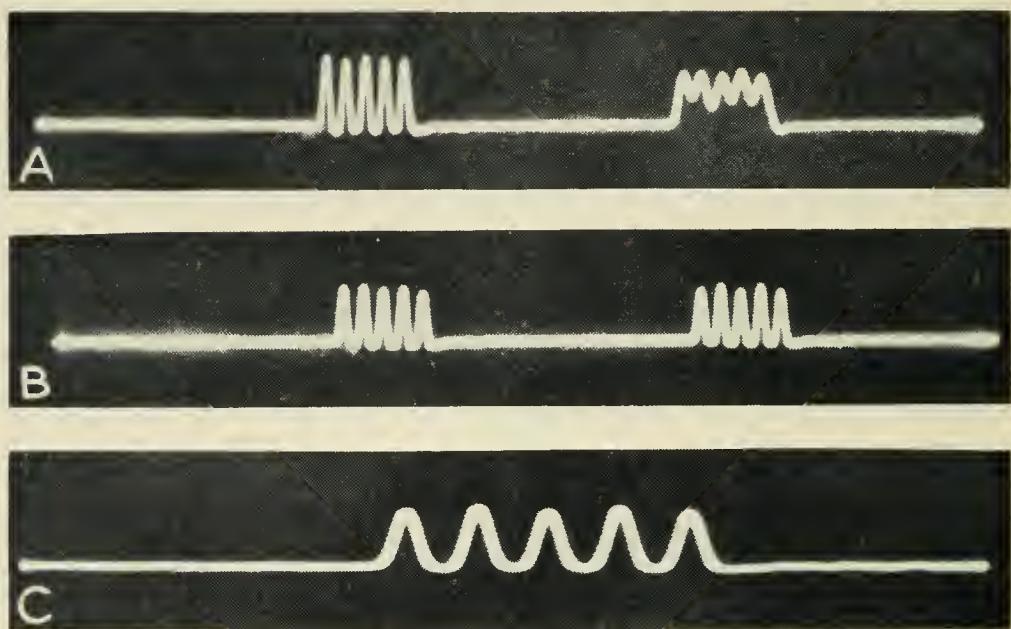


Fig. 10—The regeneration of band-limited pulses. A—Input to regenerator, first two groups. B—Output of regenerator, first two groups. C—Output of regenerator, 24th trip.

amplitude with each trip around the loop indicating that loop gain was slightly greater than unity. Without the slicer it is difficult to set the gain to exactly unity and the amplitude tends to either increase or decrease depending upon whether the gain is greater or less than unity. Results indicated by the pictures of Fig. 9 are possibly not typical of a properly functioning system but do show what happened in this particular system when regeneration was dispensed with.

Another important function of regeneration is that of overcoming band-limiting effects. Figs. 10 and 11 show what can be accomplished. For this experiment the pulse groups inserted into the loop were as shown at the left in Fig. 10A. These pulses were 15 milli-microseconds wide at the base and spaced by 25 milli-microseconds which corresponds to a repetition frequency of 40 mc. After passing through a band-pass filter these pulses were distorted to the extent shown at the right in Fig. 10A. From the characteristic of the filter, as shown on Fig. 12, it is seen that the bandwidth employed is not very different from the theoretical minimum required for double sideband transmission. This minimum characteristic is shown by the dashed lines on Fig. 12. Fig. 10B shows that at the output of the regenerator the effects of band limiting have been removed. This is borne out by Fig. 10C which shows that after 24 trips the code group was still practically perfect. It should be pointed out that the pulses traversed the filter once for each trip around the loop,

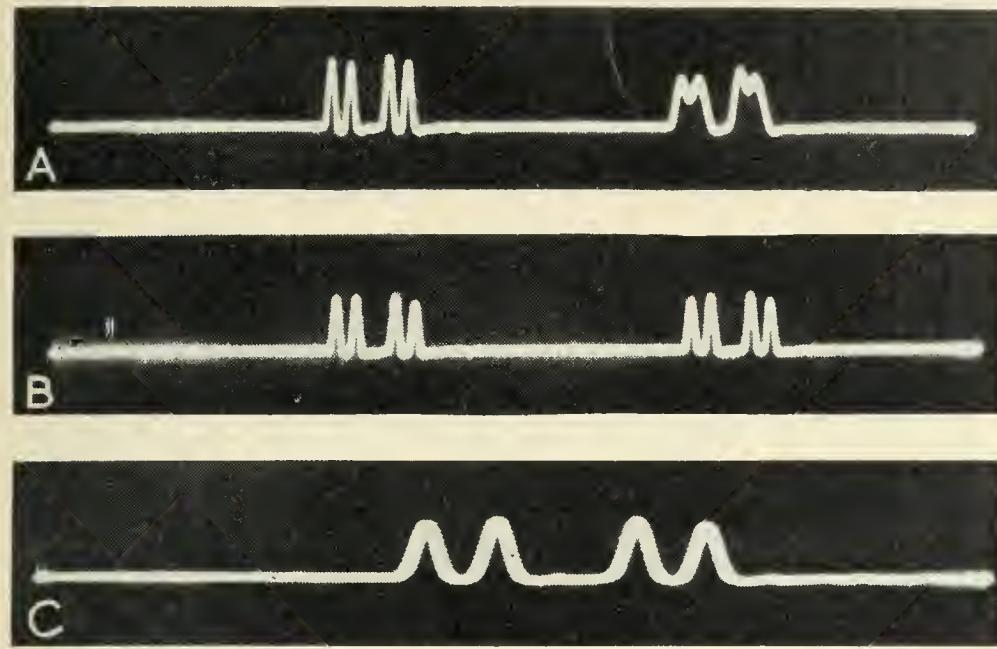


Fig. 11 — The regeneration of band-limited pulses. A — Input to regenerator, first two groups. B — Output of regenerator, first two groups. C — Output of regenerator, 24th trip.

that is for each trip the input to the regenerator was as shown at the right of Fig. 10A and the output as shown by Fig. 10B. It is important to note that Fig. 12 represents the frequency characteristic of a single link of the simulated system. The pictures of Fig. 11 show the same experiment but this time with a different code group. Any code group which we could set up with our five digit pulses was transmitted equally well.

In order to determine the breaking point of the experimental system, broad-band noise obtained from a traveling-wave amplifier was added into the system as shown on Fig. 2. The breaking point of the system is the noise level which is just sufficient to start producing errors at the output of the system.* The noise is seen to be band-limited in exactly the same way as the signal. With the system adjusted to operate properly the level of added noise was increased to the point where errors became barely discernible after 24 trips around the loop. Noise level was now reduced slightly (no errors discernible) and the ratio of rms signal to rms noise measured. Fig. 13A shows the input to the regenerator for the 23rd and 24th trips with this amount of noise added. Note that the noise has

* The type of noise employed has a Gaussian amplitude distribution and therefore there was actually no definite breaking point — the rate at which errors occurred increased continuously as noise amplitude was increased. The breaking point was taken as the noise level at which errors became barely discernible on the viewing oscilloscope. More accurate measurements made in other experiments indicate that this is a fairly satisfactory criterion.

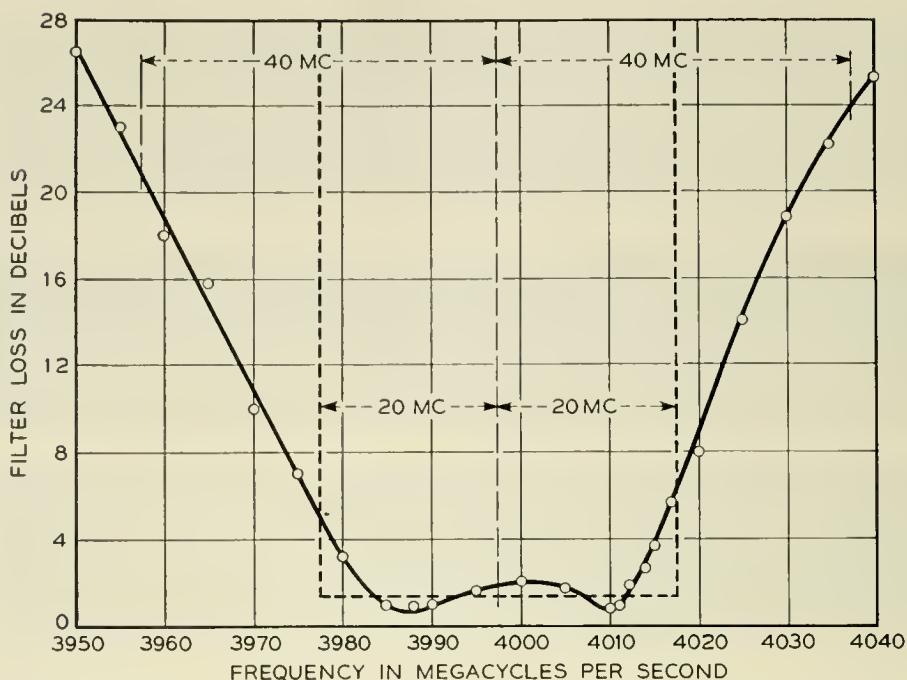


Fig. 12 — Characteristics of the band-pass microwave filter.

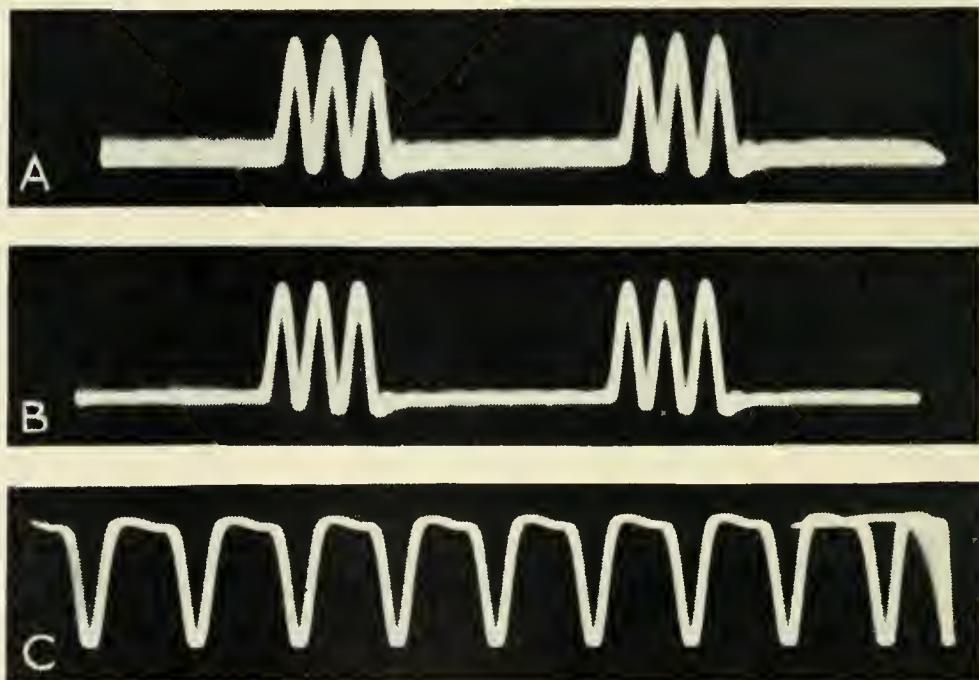


Fig. 13. — The regeneration of pulses in the presence of broad-band, random noise added at each repeater. A — Input to regenerator, 23rd and 24th trips, broad-band noise added. B — Input to regenerator, 23rd and 24th trips, no added noise. C — 20-mc timing wave.

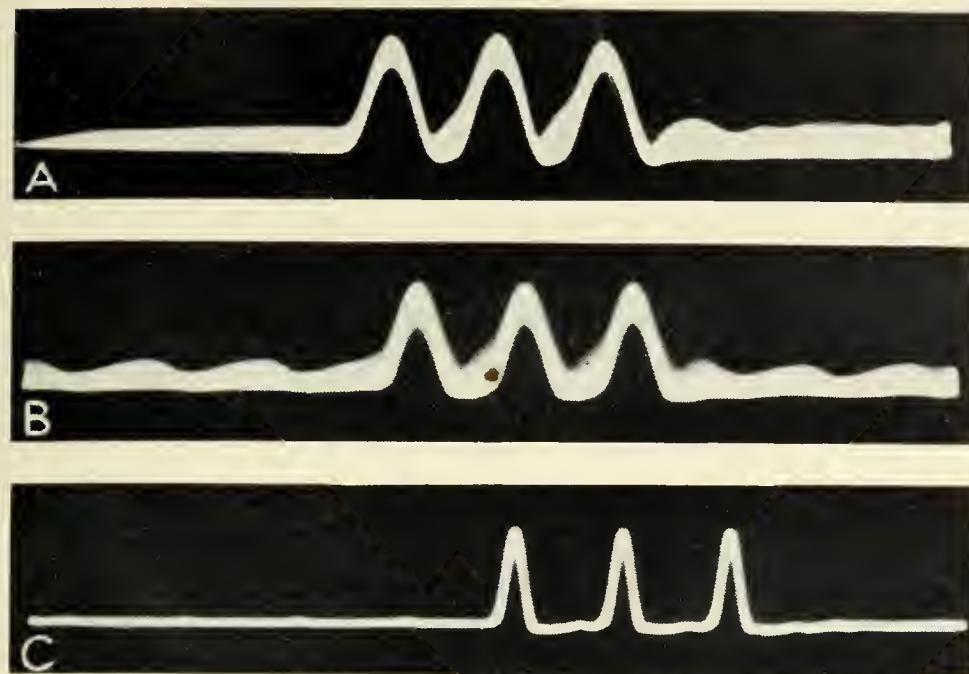


Fig. 14 — The regeneration of pulses in the presence of interference occurring at each repeater. A — Original signal with added modulated carrier interference. B — Input to regenerator, 24th trip, modulated carrier interference. C — Output of regenerator, 24th trip, modulated carrier interference.

produced a considerable broadening of the oscilloscope trace. Fig. 13B shows the same pulse groups with no added noise. These photographs are included to give some idea as to how bad the noise was at the breaking point of the system. Of course maximum noise peaks occur rather infrequently and do not show on the photograph. At the output of the regenerator effects due to noise were barely discernible. This output looked so much like that shown at Fig. 14C that no separate photograph is shown for it.

Figs. 14A, 14B and 14C show the effects of a different type of interference upon the system. This disturbance was produced by adding into the system a carrier of exactly the same frequency as the signal carrier (4 kmc) but modulated by a 14-mc wave, a frequency in the same order as the pulse rate. Here again the level of the interference was adjusted to be just below the breaking point of the system. A comparison between Figs. 14B and 14C gives convincing evidence that the regenerator has substantially restored the waveform.

For the case of the interfering signal a ratio of signal to interference of 10 db on a peak-to-peak basis was measured when the interference was just below the breaking point of the system. This, of course, is 4 db above the theoretical value for a perfect regenerator. For the case of

broad-band random noise an rms signal to noise ratio of 20 db was measured.* This compares with a ratio of 18 db as measured by Messrs. Meacham and Peterson for a system employing complete regeneration and a single repeater.†

Recently, A. F. Dietrich repeated the circulating loop experiment at a radio frequency of 11 kmc. His determinations of required signal-to-noise ratios are substantially the same as those reported here. From the various experiments we conclude that for a long chain of properly functioning regenerative repeaters of the type discussed here practically perfect transmission is obtained as long as the signal-to-noise ratio at the input to each repeater is 20 db or better on an rms basis. In an operating system it might be desirable to increase this ratio to 23 db to take care of deficiencies in automatic gain controls, power changes, etc.

From the experiments we also conclude that the price we pay for using partial instead of complete regeneration is about 3 to 4 db increase in the required signal-to-noise ratio. In a radio system which provides a fading margin this penalty would be less since the probability that two or more adjacent links will reach maximum fades simultaneously is very small. Under these conditions only one repeater at a time would be near the breaking point and the system would behave much as though the repeater provided complete regeneration.

TIMING

Although we have considered the problem of retiming of signal pulses up to now we have not discussed the problem of obtaining the necessary timing pulses to perform this function, but have simply assumed that a source of such pulses was available. As was mentioned earlier timing pulses would probably be derived from the signal pulses in a practical system. These pulses would be fed into some narrow band amplifier tuned to pulse repetition frequency. The output of this circuit could be made to be a sine wave at repetition frequency if gaps between the input pulses were not too great. Timing pulses could be derived from this sine wave. This timing equipment could be similar to that used in these experiments and described earlier. Further study of the problems of obtaining timing information is being made.

* For Gaussian noise it is not possible to specify a theoretical value of minimum S/N ratio without specifying the tolerable percentage of errors. For the number of errors detectable on the oscilloscope it seems reasonable to assume a 12 db peak factor for the noise. The peak factor for the signal is 3 db. The 6 db peak S/N which would be required for an ideal regenerator then becomes 15 db on an rms basis.

† L. A. Meacham and E. Peterson, B. S. T. J., p. 43, Jan., 1948.

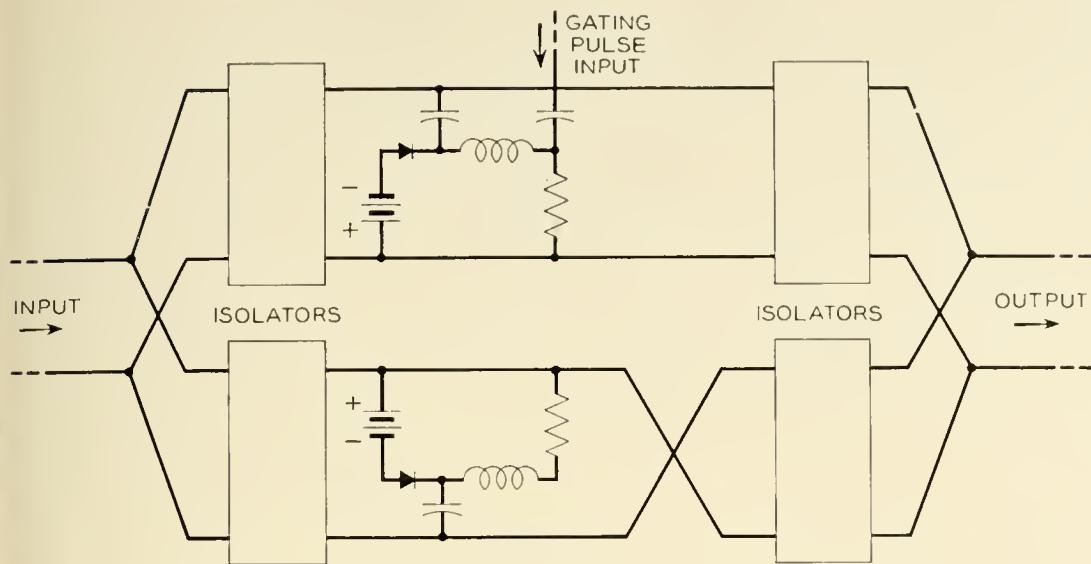


Fig. 15A — Low-frequency equivalent of the partial regenerator.

DESCRIPTION OF REGENERATOR

This device regenerates pulses by performing on them the operations of "slicing" and retiming.

An ideal slicer is a device with an input-output characteristics such as shown by the dashed lines of Fig. 15C. It is seen that for all input levels below the so-called slicing level transmission through the device is zero but that for all amplitudes greater than this value the output level is finite and constant. Thus, all input voltages which are less than the slicing level have no effect upon the output whereas all input voltages greater than the slicing level produce the same amplitude of output. Normally conditions are adjusted so that the slicing level is at one-half

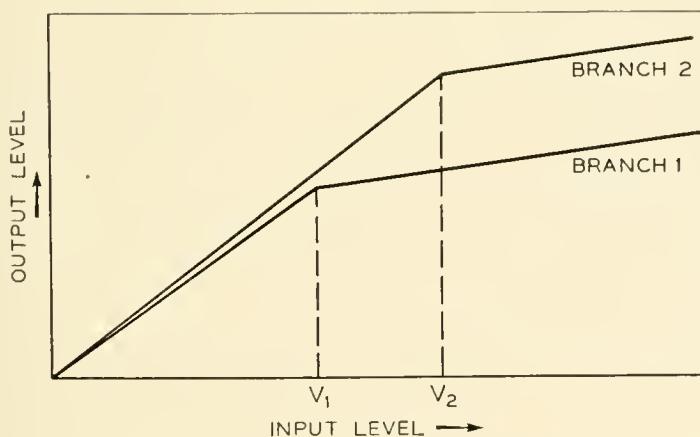


Fig. 15B — Characteristics of the separate branches with differential bias.

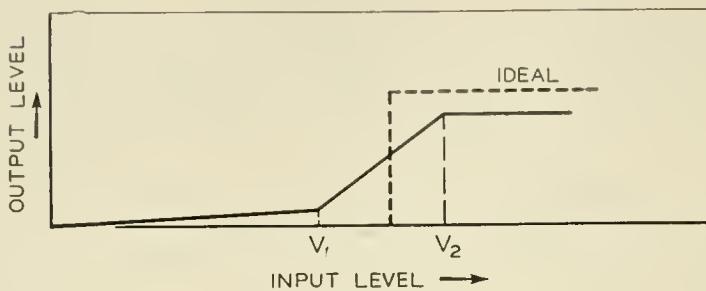


Fig. 15C — Resultant output with differential bias.

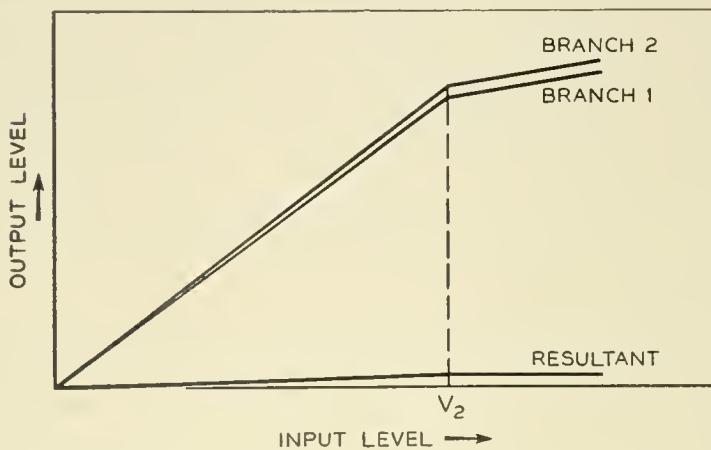


Fig. 15D — Characteristics of the separate branches and resultant output with equal biases.

of peak pulse amplitude — then at the output of the slicer there will be no effect whatsoever from disturbances unless these disturbances exceed half of the pulse amplitude. It is this slicing action which removes the amplitude effects of noise. Time jitter effects are removed by retiming, i.e., the device is made to have high loss regardless of input level except at those times when a gating pulse is present.

Fig. 15A shows schematically a low-frequency equivalent of the regenerator used in these experiments. Here an input line divides into two identical branches isolated from each other and each with a diode shunted across it. The outputs of the two branches are recombined through necessary isolators to form a single output. The phase of one branch is reversed before recombination, so that the final output is the difference between the two individual outputs.

Fig. 15B shows the input-output characteristics of the two branches when the diodes are biased back to be non-conducting by means of bias voltages V_1 and V_2 respectively. For low levels the input-output characteristic of both branches will be linear and have a 45° slope. As soon

as the input voltage in a branch reaches a value equal to that of the back bias the diode will start to conduct, thus absorbing power and decrease the slope of the characteristic. The output of Branch 1 starts to flatten off when the input reaches the value V_1 , while the output of Branch 2 does not flatten until the input reaches the value V_2 . The combined output, which is equal to the differences of the two branch outputs, is then that shown by the solid line of Fig. 15C and is seen to have a transition region between a low output and a high output level. If the two branches are accurately balanced and if the signal voltage is large compared to the differential bias $V_2 - V_1$ the transition becomes sharp and the device is a good slicer.

If the two diodes are equally biased as shown on Fig. 15D the outputs of the two branches should be nearly equal regardless of input and the total output, which is the difference between the two branch outputs, will always be small.

Fig. 16 shows a microwave equivalent of the circuit of Fig. 15A. In the microwave structure lengths of wave-guide replace the wire lines and branching, recombining and isolation are accomplished by means of hybrid junctions. The hybrid shown here is of the type known as the 1A junction.

Fig. 17 shows another equivalent microwave structure employing only one hybrid. This is the type used in the experiments described here. The output consists of the combined energies reflected from the two side arms of the junction. With the junction connected as shown phase relationships are such that the output is the difference between the reflec-

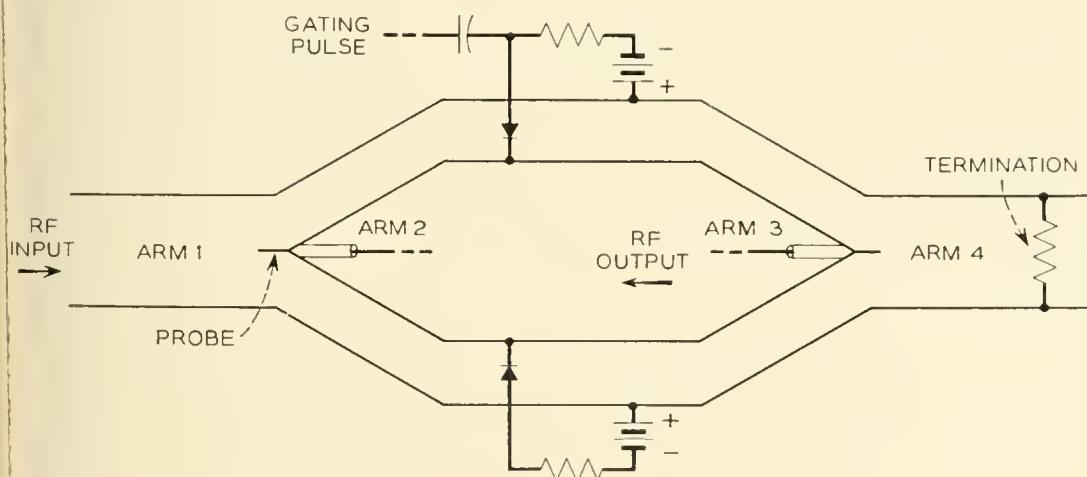


Fig. 16 — Microwave regenerator.

tions from the two side arms so that when conditions in the two arms are identical there is no output. The crystal diodes coupled to the side arms are equivalent to those shunted across the two lines of Fig. 15A.

Fig. 18, which is a plot of the measured input-output characteristic of the regenerator used in the loop test, shows how the device acts as a combined slicer and retimer. Curve A, obtained with equal biases on the two diodes, is the characteristic with no gating pulse applied i.e. the diodes are normally biased in this manner. It is seen that this condition produces the maximum of loss through the device. By shifting one diode bias so as to produce a differential of 0.5 volt the characteristic changes to that of Curve B. This differential bias can be supplied by the timing pulse in such a way that this pulse shifts the characteristic from that shown at A to that shown at B thus decreasing the loss through the device by some 12 to 15 db during the time the pulse is present. In this way the regenerator is made to act as a gate — though not an ideal one.

We see from curve B that with the differential bias the device has the characteristic of a slicer — though again not ideal. For lower levels of input there is a region over which the input-output characteristic is square law with a one db change of input producing a two db change of output. This region is followed by another in which limiting is fairly pronounced. At the 8-db input level, which is the point at which limiting sets in, the loss through the regenerator was measured to be approximately 12 db. The characteristic shown was found to be reproducible both in these experiments at 4 kmc and in those by A. F. Dietrich at 11 kmc.

For a perfect slicer only an infinitesimal change of input level is re-

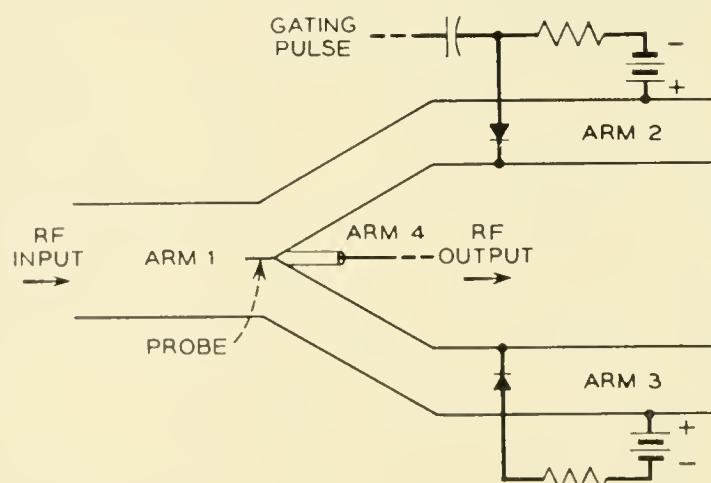


Fig. 17 — Microwave regenerator employing a single hybrid junction.

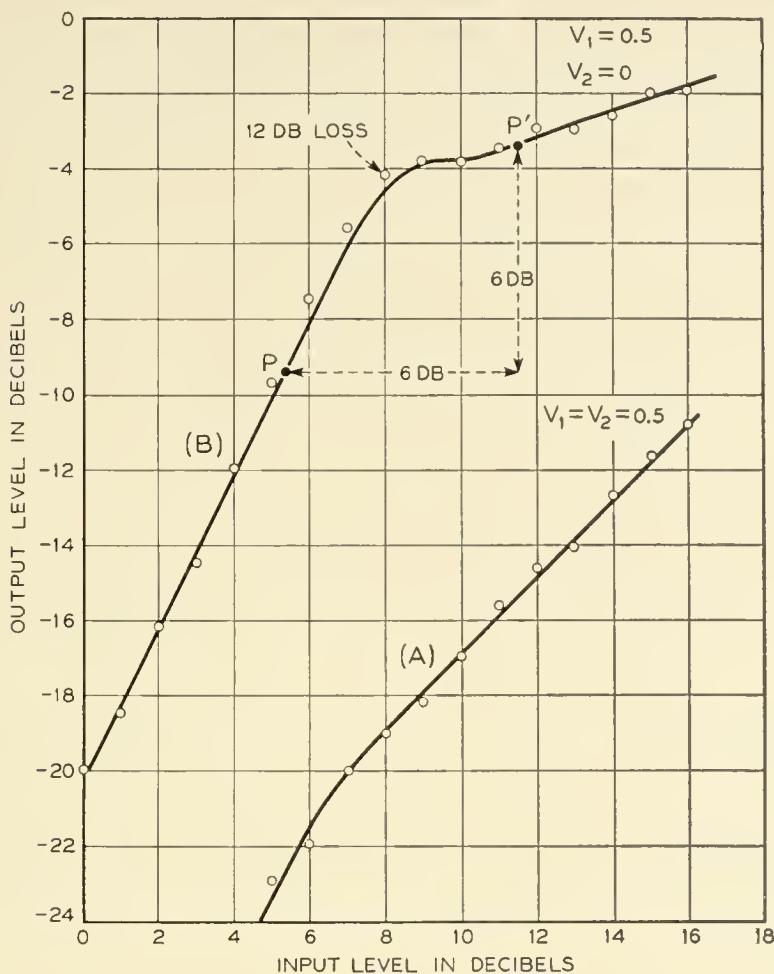


Fig. 18 — Static characteristics of the regenerator employed in these experiments.

quired to change the output from zero to maximum. The input level at which this transition takes place is the slicing level and has a very definite value. For a characteristic such as that shown on Fig. 18 this point is not at all definite and the question arises as to how one determines the slicing level for such a device. Obviously this point should be somewhere on the portion of the characteristic where expansion takes place. In the case of the circulating loop the slicing level is the level for which total gain around the loop is exactly equal to unity. Why this is so can be seen from Fig. 19 which is a plot of gain versus input level for a repeater containing a slicer with a characteristic as shown by curve B of Fig. 18. Amplifiers are necessary in the loop to make up for loss through the regenerator and other components. For Fig. 19 we assume that these amplifiers have been adjusted so that gain around the loop is exactly unity for an input pulse having a peak amplitude corresponding to the

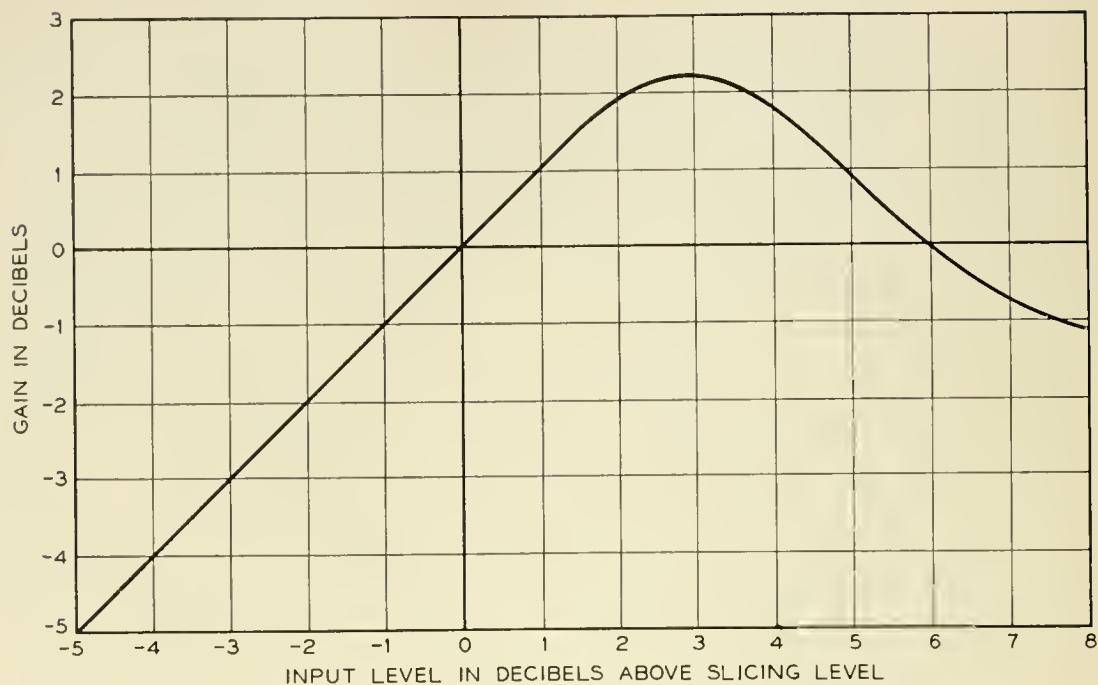


Fig 19 — Gain characteristics of a repeater providing partial regeneration.

point P' of Fig. 18. On Fig. 19 all other levels are shown in reference to this unity-gain value.

From Fig. 19 it is obvious that a pulse which starts out in the loop with a peak amplitude exactly equal to the reference, or slicing level, will continue to circulate without change of amplitude since for this level there is unity gain around the loop. A pulse with amplitude greater than the slicing level will have its amplitude increased by each passage through a regenerator until it eventually reaches a value of +6 db. It will continue to circulate at this amplitude, for here also the gain around the loop is unity.* Any pulse with peak amplitude less than the reference level will have its amplitude decreased by successive trips through the regenerator and eventually go to zero. We also see that the greater the departure of the amplitude of a pulse from the slicing level the more effect the regenerator has upon it. This means that the device acts much more powerfully on low level noise than on noise with pulse peaks near the slicing level. As examples consider first the case of noise peaks only 1 db below slicing level at the input ($\text{peak } S/N = 7 \text{ db}$). At this level there is a 1 db loss through the repeater so that at the output the noise peaks will be 2 db below reference to give a S/N ratio of 8 db. Next

* Note that the +6-db level is at a point of stable equilibrium whereas at the slicing level equilibrium is unstable.

consider noise with a peak level 5 db below slicing level ($S/N = 11$ db) at the input. The loss at this level is 5 db resulting in a noise level 10 db below reference to give a S/N ratio of 16 db. We see that a 4 db improvement in S/N ratio at the input results in an 8 db improvement in this ratio at the output.

Everything which was said above concerning the circulating loop applies equally to a chain of identical repeaters. To set the effective slicing level at half amplitude at each repeater in a chain one would first find two points on the slicer characteristics such as P and P' of Fig. 18. The point P should be in the region of expansion and P' in the limiting region. Also the points should be so chosen that a 6 db increase of input from that at point P results in a 6 db increase in output at the point P' . If now at each repeater we adjust pulse peak amplitude at the slicer input to a value corresponding to that at point P' we will have unity gain from one repeater to the next at levels corresponding to pulse peaks. We will also have unity gain at levels corresponding to one half of pulse amplitude. The effective slicing level is thus set at half amplitude. Obviously the procedure for setting the slicing level at some value other than half amplitude would be practically the same. It should be pointed out that although half amplitude is the preferred slicing level for baseband pulses this is not the case for carrier pulses. W. R. Bennett of Bell Telephone Laboratories has shown that for carrier pulses the probability that noise of a given power will reduce signal pulses below half amplitude is less than the probability that this same noise will exceed half amplitude. This comes about from the fact that for effective cancellation there must be a 180° phase relationship between noise and pulse carrier. For this reason the slicing level should be set slightly above half amplitude for a carrier pulse system.

The difference in performance between a perfect slicer and one with characteristics such as shown on Fig. 18 are as follows: For the perfect slicer no effects from noise or other disturbances are passed from one repeater to the next. For the case of the imperfect regenerator some effects are passed on and so tend to accumulate in a chain of repeaters. To prevent this accumulated noise from building up to the breaking point of the system it is necessary to make the signal-to-noise ratio at each repeater somewhat better than that which would be required with the ideal slicer. For the case of random noise the required S/N ratio seems to be about 5 or 6 db above the theoretical value. This is due in part to slicer deficiency and in part to other system imperfections.

CONCLUSIONS

It is possible to build a simple device for regenerating pulses directly at microwave frequencies. A long chain of repeaters employing this regenerator should perform satisfactorily as long as the rms signal-to-noise ratio at each repeater is maintained at a value of 20 db or greater. There are a number of remaining problems which must be solved before we have a complete regenerative repeater. Some of these problems are: (1) Recovery of information for retiming from the incoming pulse train; (2) Automatic gain or level control to set the slicing level at each repeater; (3) Simple, reliable, economical, broad-band microwave amplifiers. (4) Proper filters — both for transmitting and receiving. Traveling-wave tube development should eventually result in amplifiers which will meet all of the requirements set forth in (3) above. Any improvements which can be made in the regenerator without adding undue complications would also be advantageous.

ACKNOWLEDGMENTS

A. F. Dietrich assisted in setting up the equipment described here and in many other ways. The experiment would not have been possible without traveling-wave tubes and amplifiers which were obtained through the cooperation of M. E. Hines, C. C. Cutler and their associates. I wish to thank W. M. Goodall, and J. R. Pierce for many valuable suggestions.

Crossbar Tandem as a Long Distance Switching System

By A. O. ADAM

(Manuscript received March 4, 1955)

Major toll switching features are being added to the crossbar tandem switching system for use at many of the important long distance switching centers of the nationwide network. These include automatic selection of one of several alternate routes to a particular destination, storing and sending forward digits as required, highly flexible code conversion for transmitting digits different from those received, and a translating arrangement to select the most direct route to a destination. The system is designed to serve both operator and customer dialed long distance traffic.

INTRODUCTION

The crossbar tandem switching system,¹ originally designed for switching between local dial offices, will now play an important role in nationwide dialing. New features are now available or are being developed that will permit this system to switch all types of traffic. As a result, crossbar tandem offices will have widespread use at many of the important switching centers of the nationwide switching network.

This paper briefly reviews the crossbar tandem switching system and its application for local switching, followed by discussion of the general aspects of the nationwide switching plan and of the major new features required to adapt crossbar tandem to this plan.

CROSSBAR TANDEM OFFICES USED FOR LOCAL SWITCHING

Crossbar tandem offices are now used in many of the large metropolitan areas throughout the country for interconnecting all types of local dial offices. In these applications they perform three major functions. Basically, they permit economies in trunking by combining small amounts of

traffic to and from the local offices into larger amounts for routing over common trunk groups to gain increased efficiency resulting in fewer overall trunks.

A second important function is to permit handling calls economically between different types of local offices which are not compatible from the standpoint of intercommunication by direct pulsing. Crossbar tandem offices serve to connect these offices and to supply the conversion from one type of pulsing to another where such incompatibilities exist.

The third major function is that of centralization of equipment or services. For example, centralization of expensive charging equipment at a crossbar tandem office results in efficient use of such equipment and over-all lower cost as compared with furnishing this equipment at each local office requiring it. Examples of such equipment are remote control of zone registration and centralized automatic message accounting.² Centralization of other services such as weather bureau, time-of-day and similar services can be furnished.

The first crossbar tandem offices were installed in 1941 in New York, Detroit and San Francisco. These offices were equipped to interconnect local panel and No. 1 crossbar central offices in the metropolitan areas, and to complete calls to manual central offices in the same areas. The war years slowed both development and production and it was not until the late 40's that many features now in use were placed in service. These later features enable customers in step-by-step local central offices on the fringes of the metropolitan areas to interconnect on a direct dialing basis with metropolitan area customers in panel, crossbar, manual and step-by-step central offices. This same development also permitted central offices in strictly step-by-step areas to be interconnected by a crossbar tandem office where direct interconnecting was not economical. Facilities were also made available in the crossbar tandem system for completing calls from switchboards where operators use dials or multifrequency key pulsing sets.

Since a crossbar tandem office usually has access to all of the local offices in the area in which it is installed, it is attractive for handling short and long haul terminating traffic. The addition of toll terminal equipment at Gotham Tandem in New York City in 1947 permitted operators in New York State and northern New Jersey as well as distant operators to dial or key pulse directly into the tandem equipment for completion of calls to approximately 350 central offices in the New York metropolitan area. This method of completing these calls without the aid of the inward operators was a major advance in using tandem switching equipment for speeding completion of out-of-town calls.

CROSSBAR TANDEM SWITCHING ARRANGEMENT

The connections in a crossbar tandem office are established through crossbar switches mounted on incoming trunk link and outgoing office link frames shown on Fig. 1. The connections set up through these switches are controlled by equipment common to the crossbar tandem office which is held only long enough to set up each individual connection. Senders and markers are the major common control circuits.

The sender's function is to register the digits of the called number, transmit the called office code to the marker and then, as subsequently directed by the marker, control the outpulsing to the next office.

The marker's function is to receive the code digits from the sender for translation, return information to the sender concerning the details of the call, select an idle outgoing trunk to the called destination and close the transmission path through the crossbar switches from the incoming to the outgoing trunk.

GENERAL ASPECTS OF NATIONWIDE DIALING

Operator distance dialing, now used extensively throughout the country, as well as customer direct distance dialing are based on the division of the United States and Canada into numbering plan areas, interconnected by a national network through some 225 Control Switching Points (CSP's) equipped with automatic toll switching systems.

* An essential element of the nationwide dialing program is a universal numbering plan³ wherein each customer will have a distinctive number which does not conflict with the number of any other customer. The method employed is to divide the United States and Canada geographi-

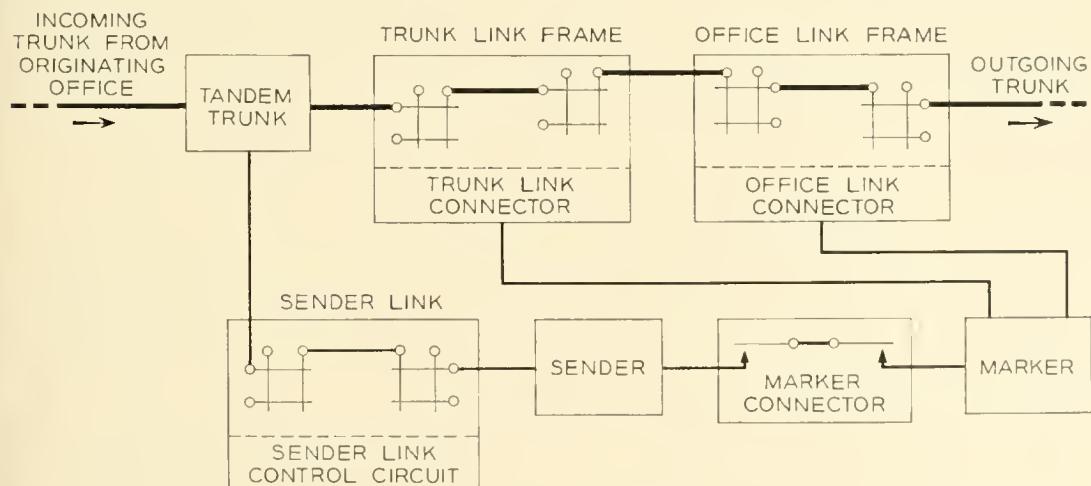


Fig. 1 — Crossbar tandem switching arrangement.

cially into more than 100 numbering plan areas and to give each of these a distinctive three digit code with either a 1 or 0 as the middle digit. Each numbering plan area will contain 500 or fewer local central offices each of which will be assigned a distinctive three-digit office code. Thus each of the telephones in the United States and Canada will have, for distance dialing purposes, a distinct identity consisting of a three digit area code, an office code of two letters and a numeral, and a station number of four digits. Under this plan, a customer will dial 7 digits to reach another customer in the same numbering area and 10 digits to reach a customer in a different numbering area.

A further requirement for nationwide dialing of long distance calls is a fundamental plan⁴ for automatic toll switching. The plan provides a systematic method of interconnecting all the local central offices and toll switching centers in the United States and Canada. As shown on Fig. 2, several local central offices or "end offices" are served by a single toll center or toll point that has trunks to a "home" primary center which serves a group of toll centers. Each primary center, has trunks to a "home" sectional center which serves a larger area of the country. Similarly, the entire toll dialing territory is divided into eleven very large areas called regions, each having a regional center to serve all the sectional centers in the region. One of the regional centers, probably St. Louis, Missouri, will be designated the national center. The homing arrangements are such that it is not necessary for end offices, toll centers, toll points and primary centers to home on the next higher ranking office since the complete final route chain is not necessary. For example, end offices may be served directly from any of the higher ranking switching centers also shown in Fig. 2.

Collectively, the national center, the regional centers, the sectional centers and the primary centers will constitute the control switching points for nationwide dialing. The basic switching centers and homing arrangements are illustrated in Fig. 3.

TANDEM CROSSBAR FEATURES FOR NATIONWIDE DIALING

The broad objective in developing new features for crossbar tandem is to provide a toll switching system that can be used in cities where the large capacity and the full versatility of the No. 4 toll crossbar switching system⁵ may not be economical.

The application of crossbar tandem two-wire switching systems at primary and sectional centers has been made possible by the extended use of high speed carrier systems. The echoes at the 2-wire crossbar tandem switching offices can be effectively reduced by providing a high

office balance and by the use of impedance compensators and fixed pads. A well balanced two-wire switching system, proper assignment of inter-toll trunk losses, and the use of carrier circuits with high speed of propagation will permit through switching with little or no impairment from an echo standpoint.

The new features for crossbar tandem will provide arrangements necessary for operation at control switching points (CSP's). These include automatic alternate routing, the ability to store and send forward

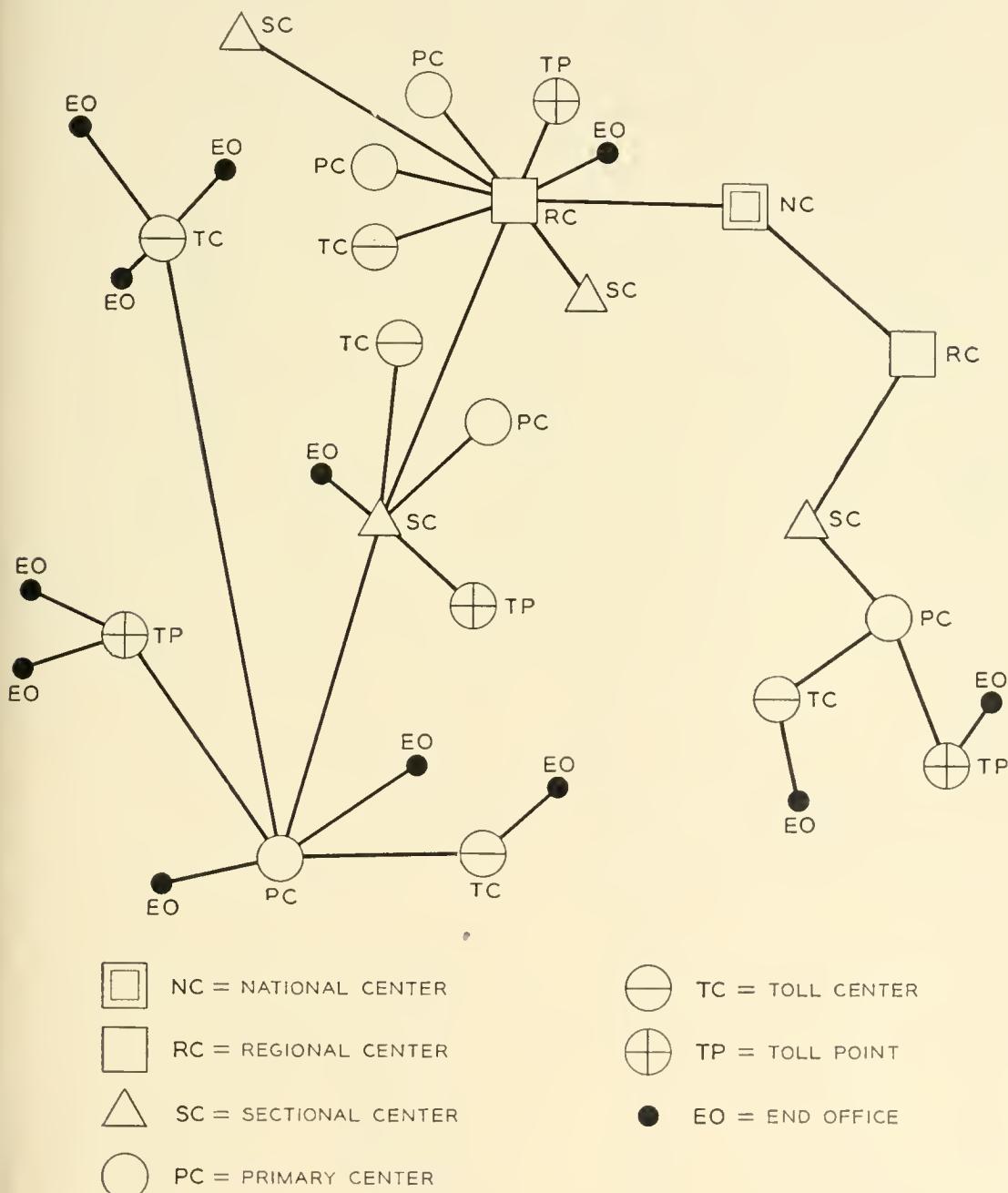
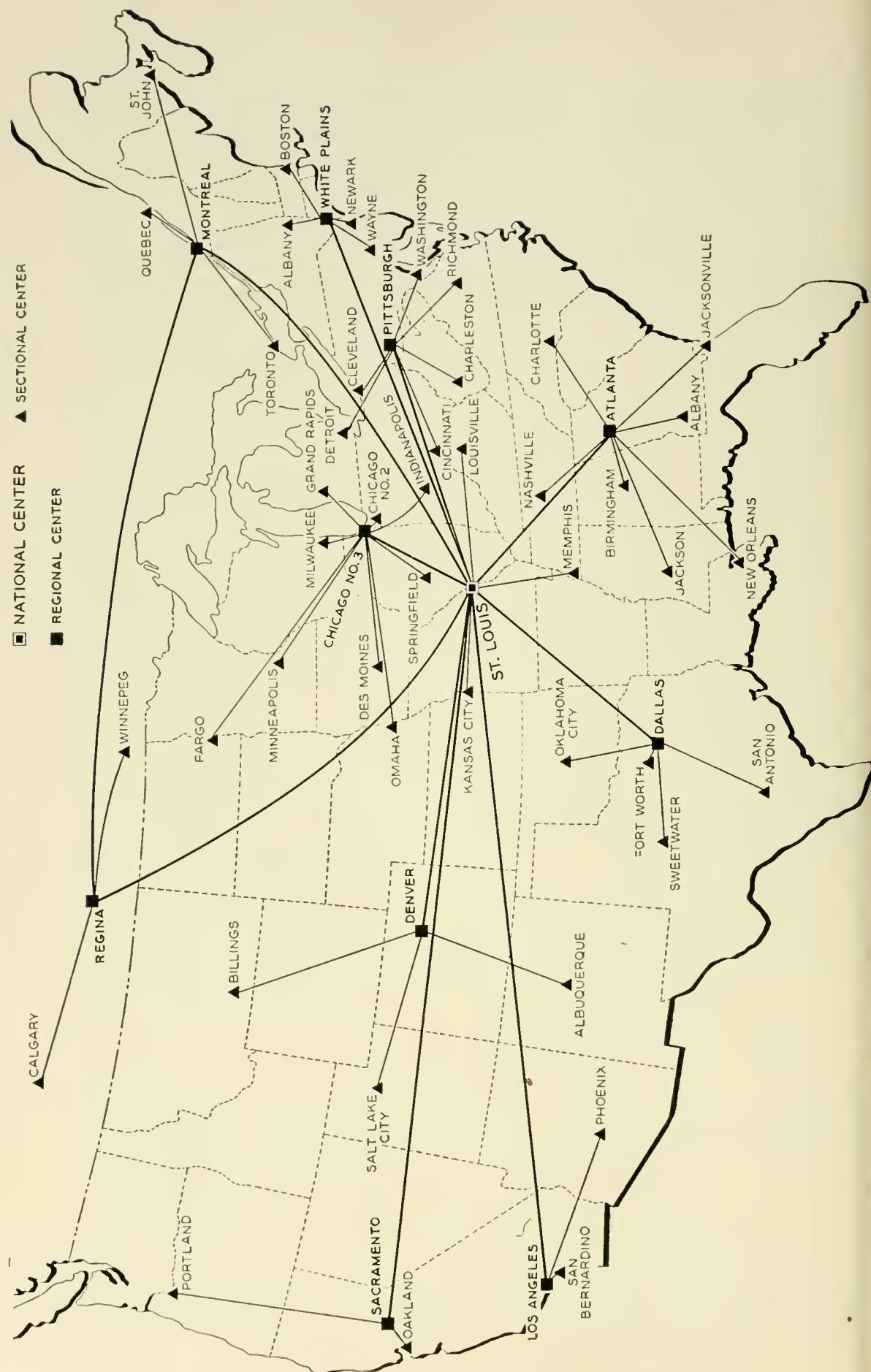


Fig. 2 — Homing arrangement for local central offices and toll centers.



Winter, — Toulon's view of the basic diatom switching network in 1965.

digits as required, highly flexible code conversion (transmitting forward different digits for the area or office code instead of the dialed digits), prefixing digits ahead of the called office code, and six-digit translation.

ALTERNATE ROUTING

The control switching points will be interconnected by a final or "backbone" network of intertoll trunks engineered so that very few calls will be delayed. In addition, direct circuits between individual switching offices of all classes will be provided as warranted by the traffic density. These are called "high-usage" groups and are not engineered to handle all the traffic offered to them during the busy hour. Traffic offered to a high-usage group which finds all trunks busy will be automatically rerouted to alternate routes^{6, 7} consisting of other high-usage groups or to the final trunk group. The ability of the crossbar tandem equipment at the control switching point to select one of several alternate routes automatically, when all choices in the first route are busy, contributes to the economy of the plant and provides additional protection against complete interruption of service when all circuits on a particular route are out of service.

Fig. 4 shows a hypothetical example of alternate routing when a crossbar tandem office at South Bend, Indiana, receives a call destined for Youngstown, Ohio. To select an idle path, using this plan, the switching equipment at South Bend first tests the direct trunks to Youngstown. If these are all busy, it tests the direct trunks to Cleveland where the call would be completed over the final group to Youngstown. If the group to Cleveland is also busy, South Bend would test the group

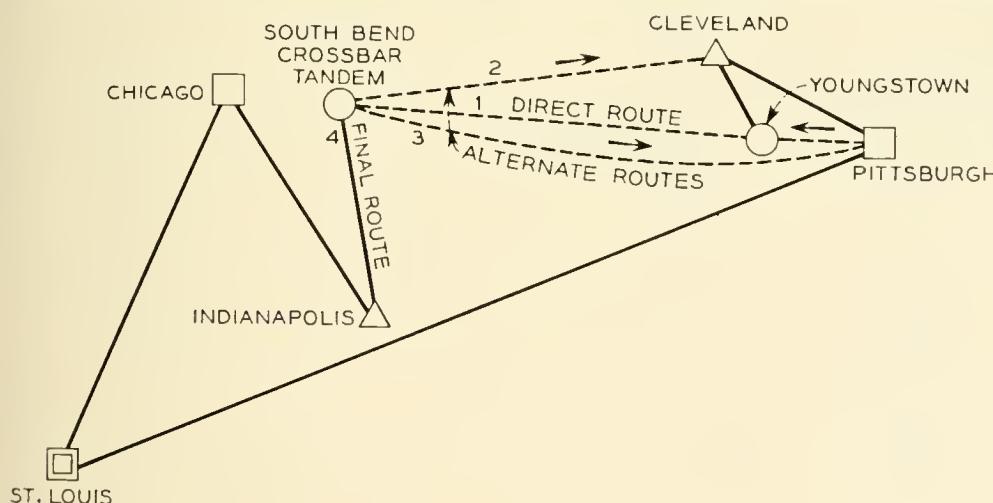


Fig. 4 — Toll network — alternate routing.

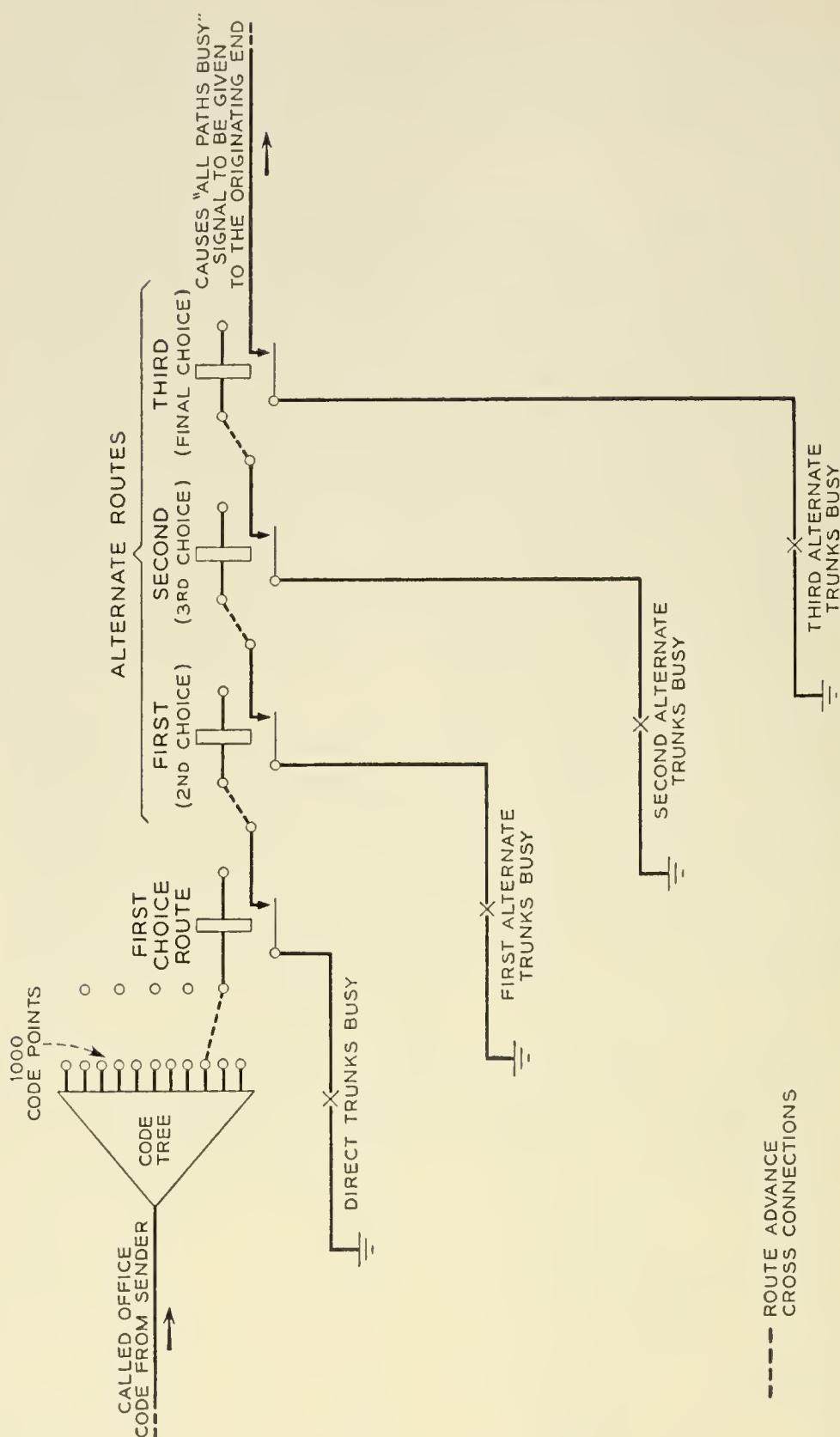


Fig. 5 — Method used for alternate routing.

to Pittsburgh and on its last attempt it would test the final group to Indianapolis. If the call were routed to Pittsburgh or Indianapolis, the switching equipment at these points would attempt by first choice and alternate routes to reach Youngstown. The final choice backbone route would be via Indianapolis, Chicago, St. Louis, Pittsburgh, Cleveland to Youngstown. Should all the trunks in any of the final groups tested be busy no further attempt to complete the call is made. It is unlikely that so many alternate routes would be provided in actual practice since crossbar tandem can test only a maximum of 240 trunks on each call and, in the case illustrated, the final trunk group to Indianapolis may be quite large.

The method employed by the crossbar tandem marker in selecting the direct route and subsequent alternate routes is shown in simplified form on Fig. 5. As a result of the translating operation, the marker selects the first choice route relay, corresponding to the called destination. Each route relay has a number of contacts which are connected to supply all the information required for proper routing of the call. Several of these contacts are used to indicate the equipment location of the trunks and the number of trunks to be tested. The marker tests all of the trunks in the direct route and if they are busy, the search for an idle trunk continues in the first alternate route which is brought into play from the "route advance" cross-connection shown on the sketch. As many as three alternate routes in addition to the first choice route can be tested in this manner.

STORING AND SENDING FORWARD DIGITS AS REQUIRED

The crossbar tandem equipment at control switching points must store all the digits received and send forward as many as are required to complete the call.

The called number recorded at a switching point is in the form of ABX-XXXX if the call is to be completed in the same numbering plan area. If the called destination is in another area, the area code XOX or XIX precedes the 7 digit number. The area codes XOX or XIX and the local office code ABX are the digits used for routing purposes and are sufficient to complete the call regardless of the number of switching points involved. Each control switching point is arranged to advance the call towards its destination when these codes are received. If the next switching point is not in the numbering area of the called telephone, the complete ten-digit number is needed to advance the call toward its destination. If the next switching point is in the num-

bering area of the called telephone the area code is not needed and seven digits will suffice for completing the call.

For example, suppose a call is originated by a customer in South Bend, Indiana, destined for customer NAtional 4-1234 in Washington, D.C. If it is assumed that the route to Washington is via a switching center in Pittsburgh, then the crossbar tandem equipment at South Bend pulses forward to Pittsburgh 202-NA4-1234, 202 being the area code for the District of Columbia. Pittsburgh in turn will delete the area code and send NA4-1234 to the District of Columbia terminating area.

As another example, suppose the crossbar tandem office at South Bend receives a call from some foreign area destined to a nearby step-by-step end office in Michigan. The crossbar tandem equipment receives and stores a ten-digit number comprising the area code and the seven digits for the office code and station number. Assuming that direct trunks to the step-by-step end office in Michigan are available, the area code and office code are deleted and the line number only is pulsed forward. To meet all conditions, the equipment is arranged to permit deletion of either the first three, four, five or six digits of a ten-digit number.

CODE CONVERSION

At the present time, some step-by-step primary centers reach other offices by the use of routing codes that are different from those assigned under the national numbering plan. This arrangement is used to obtain economies in switching equipment of the step-by-step plant and is acceptable with operator originated calls. However, with the introduction of customer direct distance dialing, it is essential that the codes used by customers be in accordance with the national numbering plan. The crossbar tandem control switching point must then automatically provide the routing codes needed by the intermediate step-by-step primary centers. This is accomplished by the code conversion feature which substitutes the arbitrary digits required to reach the called office through the step-by-step systems. Fig. 6 illustrates an application of this feature. It shows a crossbar tandem office arranged for completing calls through a step-by-step toll center to a local central office, GArden 8, in an adjacent area. A call reaching the crossbar tandem office for a customer in this office arrives with the national number, 218-GA8-1234. To complete this call, the crossbar tandem equipment deletes the area code 218 and pulses forward the local office code and number. If the

call is switched to an alternate route via the step-by-step primary center, it will be necessary for the crossbar tandem equipment to delete the area code 218 and substitute the arbitrary digits 062 to direct the call through the switches at the primary center, since the toll center requires the full seven digit number for completing the call.

PREFIXING DIGITS

It may be necessary to route a call from one area to another and back to the original area for completion. Such a situation arises on a call from Amarillo to Lubbock, Texas, both in area 915 when the crossbar tandem switching equipment finds all of the direct paths from Amarillo to Lubbock busy as illustrated on Fig. 7. The call could be routed to Lubbock via Oklahoma City which is in area 405. A seven-digit number for example, MAin 2-1234, is received in the crossbar tandem office at Amarillo. Assuming that the call is to be switched out of the 915 area through the 405 area and back to the 915 area for completion, it is necessary for the crossbar tandem office in Amarillo to prefix 915 to the MAin 2-1234 number so that the switching equipment in Oklahoma City will know that the call is for the 915 area and not for the 405 area.

Prefixing digits may also be needed at crossbar tandem offices to route calls through step-by-step primary centers. The crossbar tandem office in Fig. 8 receives the seven digit number MA2-1234 for a call to a

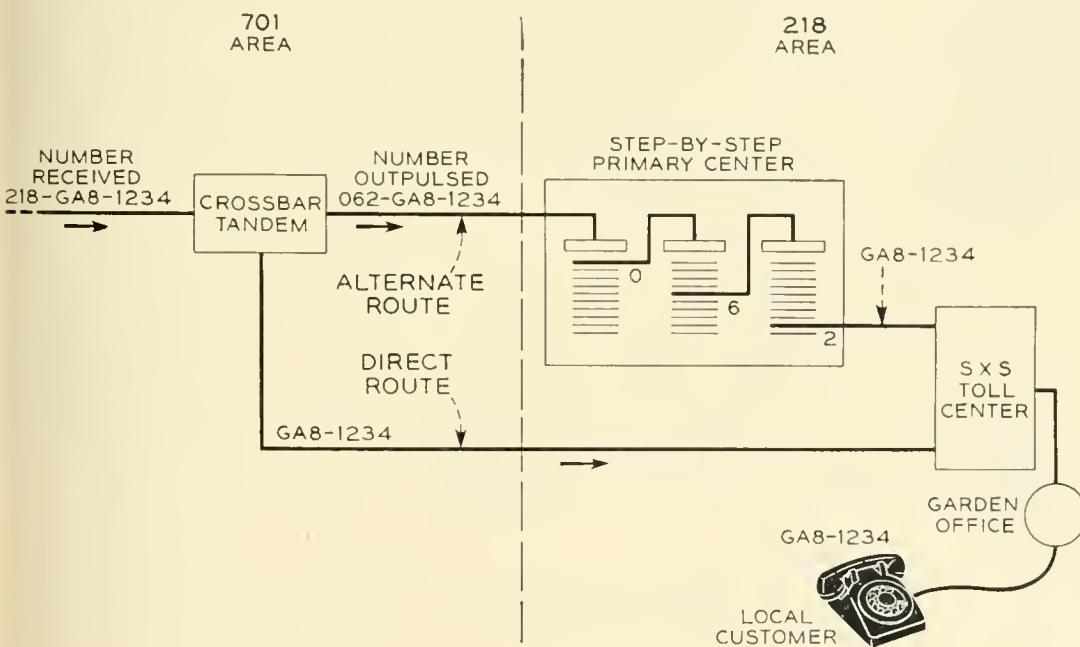


Fig. 6 — Code conversion.

customer in the Madison office in the same area. However, since the toll center needs the full seven digit number for completing the call and since the step-by-step switches at the primary center "use up" two digits (04) for its switching, the crossbar tandem equipment must prefix 04 to the seven digit number.

METHOD OF DETERMINING DIGITS TO BE TRANSMITTED

The circuitry involved for transmitting digits as received, prefixing, code conversion and for deletion involves both marker and sender functions. The senders have ten registers (1 to 10) for storing incoming digits and three registers (AA, AB, AC) for storing the arbitrary digits that are used for prefixing and code conversion.

On a ten-digit call into a crossbar tandem switching center the area code XOX, the office code ABX and the station number XXXX are stored in the impulsing or receiving registers of the sender. The code digits XOX-ABX are sent to the marker which translates them to determine which of the digits received by the sender should be outpulsed. It also determines whether arbitrary digits should be transmitted ahead of the digits received and, if so, the value of the arbitrary digits to be stored in the sender registers AA, AB and AC. Case 1 of Fig. 9 assumes that a ten-digit number has been stored in the sender registers 1 to 10

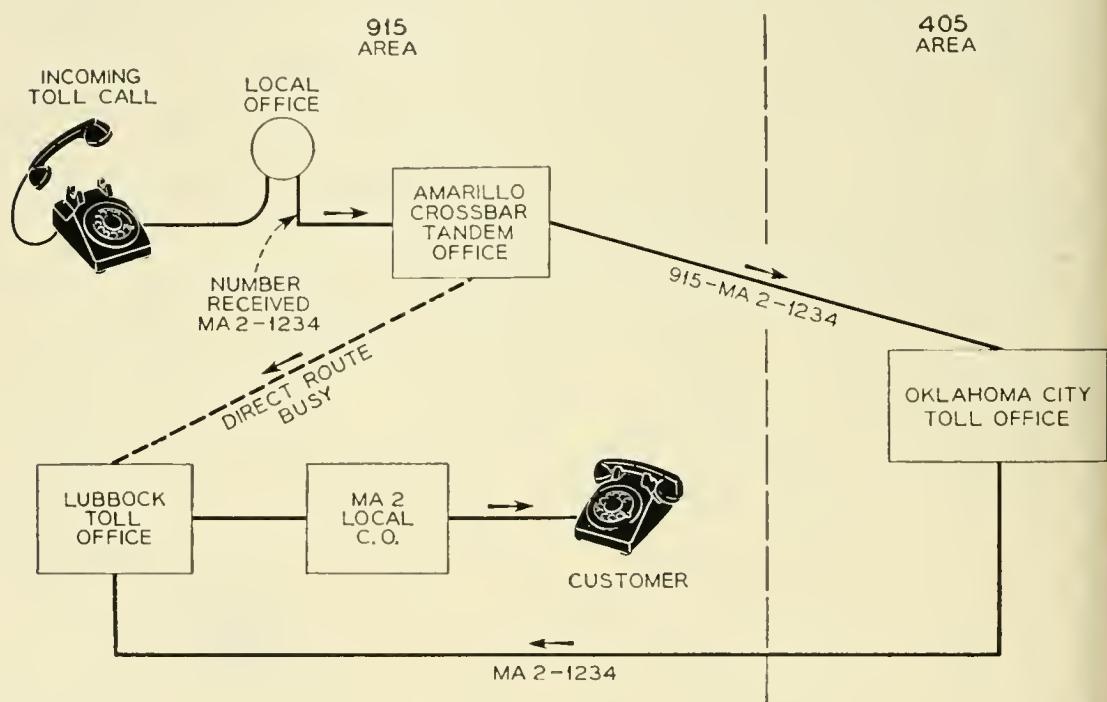


Fig. 7 — Prefixing.

and that the marker has informed the sender the called number is to be sent as received. The outpulsing control circuit is connected to each register in turn through the steering circuit S₁, S₂, etc. and sends the digits stored.

Case 2 illustrates a situation where the sender has stored ten digits in registers 1 to 10 and received information from the marker to delete the digits in registers 1 to 3 inclusive and to substitute the arbitrary digits stored in registers AA, AB and AC. The outpulsing circuit is first connected to register AA through steering circuit PS₁, then to AB through PS₂, continuing in a left to right sequence until all digits are outpulsed.

Case 3 covers a condition where the sender has stored seven digits and has obtained information from the marker to prefix the two digits stored in registers AB and AC. Outpulsing begins at the AB register through steering circuit PS₂ and then advances through steering circuit PS₃ to the AC register, continuing in a left to right sequence until all digits have been transmitted.

These are only a few of the many combinations that are used to give the crossbar tandem control switching equipment complete pulsing flexibility.

SIX-DIGIT TRANSLATION

Six-digit translation will be another feature added to the crossbar tandem system. When only three digits are translated, it is necessary to direct all calls to a foreign area over a single route. The ability to translate six digits permits the establishment of two or more routes from the switching center to or towards the foreign area. This is shown in Fig.

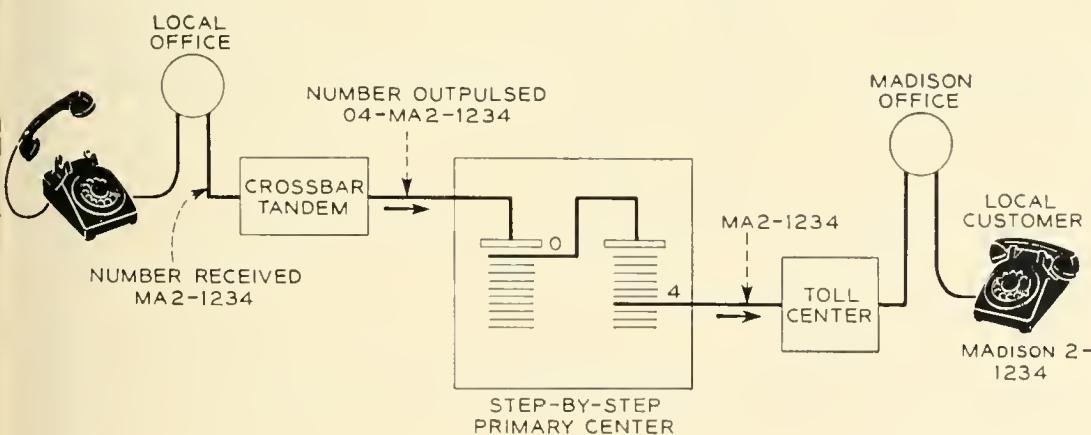


Fig. 8 -- Prefixing.

10 with Madison and Milwaukee, Wisconsin, in area 414 and Belle Plaine Crossbar Tandem in Chicago, Illinois, in area 312. An economical trunking plan may provide for direct circuits from Chicago to each place. If only three-digit translation were provided in the Chicago switching equipment, the route to both places would be selected as a result of the translation of the 414 area code alone and, therefore, calls to central offices reached through Madison, would need to be routed via Milwaukee. This involves not only the extra trunk mileage, but also the use of an extra switching point. With six-digit translation, both the area code and the central office code are analyzed, making it possible to select the direct route to either city.

Six-digit translation in crossbar tandem will involve primarily the use of a foreign area translator and a marker. The translator will have a capacity for translation of five foreign areas and for 60 routes to each area. Since the translator holding time is very short, one translator is sufficient to handle all of the calls requiring six-digit translation, but two are always provided for hazard and maintenance reasons.

On a call requiring six-digit translation the first three digits are

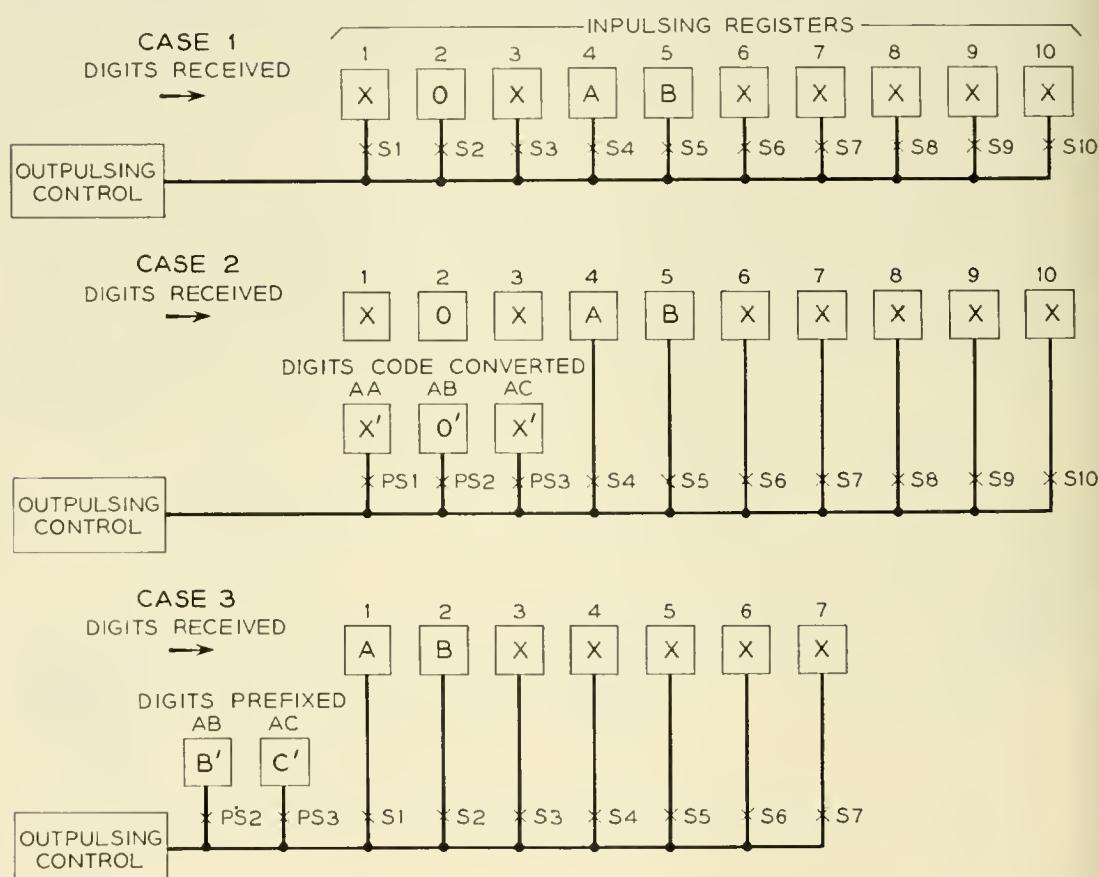


Fig. 9 — Method used for outpulsing digits.

translated in the marker and the second three digits in a foreign area translator which is associated with the marker. Fig. 11 shows, in simplified form, how this translation is accomplished.

The first three digits, corresponding to the area code, are received by a relay code tree in the marker which translates it into one of a thousand code points. This code point is cross-connected to the particular relay of the five area relays A0-A4 which has been assigned to the called area. A foreign area translator is now connected to the marker and a corresponding area relay is operated in it. The translator also receives the called office code from the sender via the marker and by means of a relay code tree similar to that in the marker translates the office code to one of a thousand code points. This code point plus the area relay is sufficient to determine the actual route to be used. As shown on the sketch, wires from each of the code points are threaded through transformers, two for each area. When the marker is ready to receive the route information, a surge of current is sent through one of these threaded wires which produces a voltage in the output winding to ionize the T- and U- tubes. Only the tubes associated with the area involved in the translation pass current to operate one each of the eight T- and U-relays. This information is passed to the marker and registered on corresponding tens and units relays. These operate a route relay which

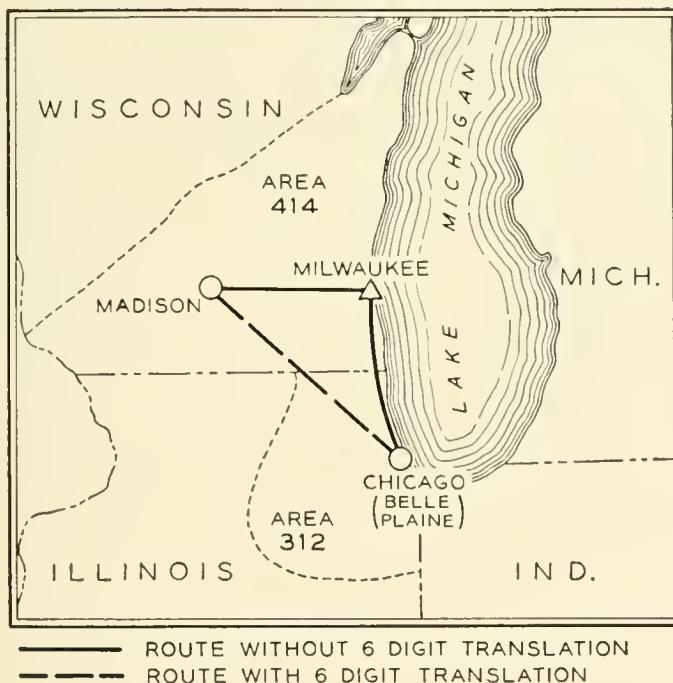


Fig. 10 — Six-digit translation.

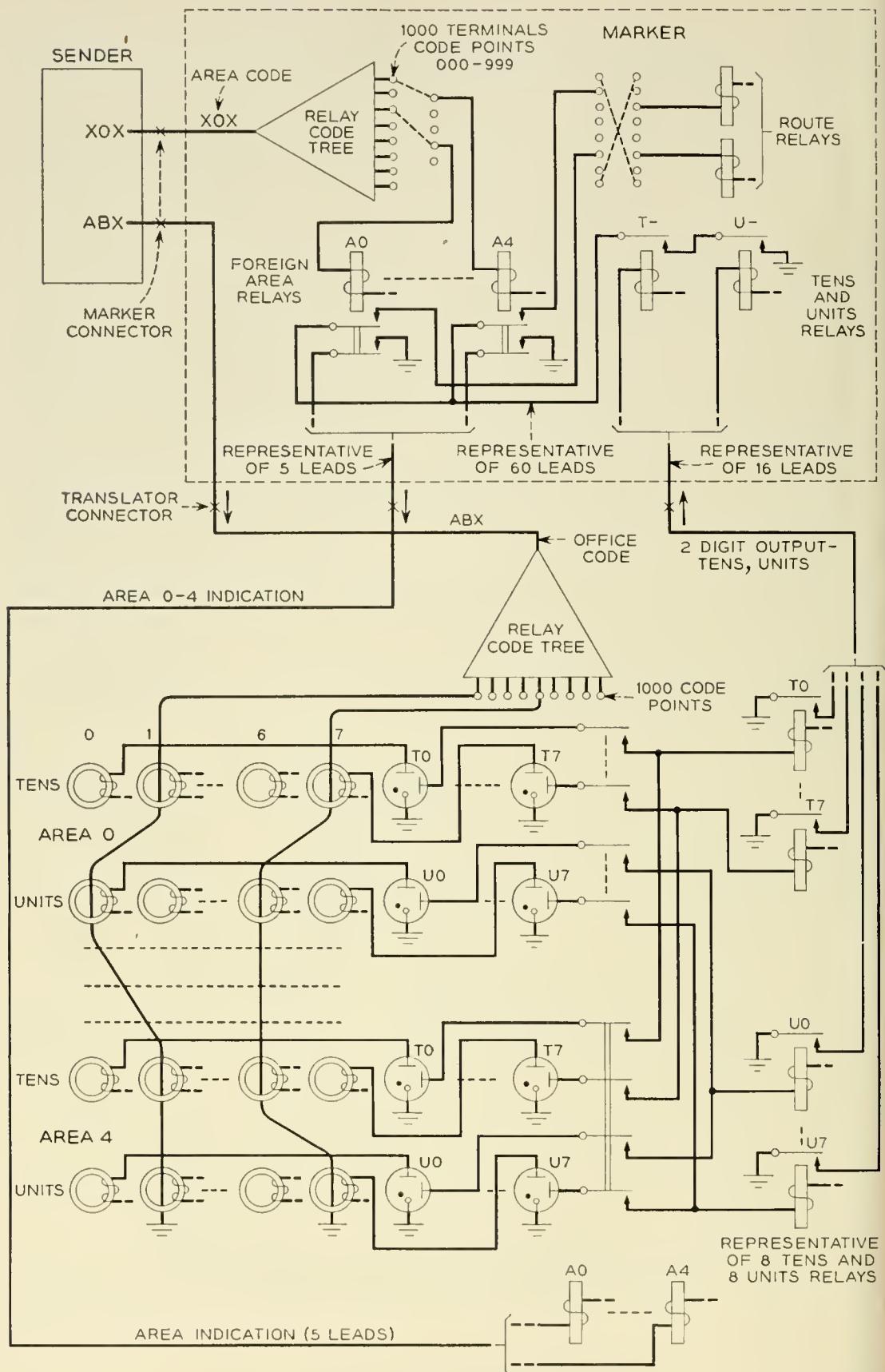


Fig. 11 — Method used for foreign area translation.

provides all the information necessary for routing the call to the central office involved.

CUSTOMER DIRECT DISTANCE DIALING

Crossbar tandem will provide arrangements permitting customers in step-by-step offices to dial their own calls anywhere in the country. Centralized automatic message accounting previously mentioned will be used for charging purposes. While the basic plan for direct distance dialing provides for the dialing of either seven or ten digits, it will be necessary for the customer in step-by-step areas to prefix a three-digit directing code, such as 112, to the called number. This directing code is required to direct the call through the step-by-step switches to the crossbar tandem office so that the seven or ten digit number can be registered in the crossbar tandem office.

When a customer in a step-by-step office originates a call to a distant customer whose national number is 915-CH3-1234, he first dials the directing code 112 and then the ten-digit number. The dialing of 112 causes the selectors in the step-by-step office to select an outgoing trunk to the crossbar tandem office. The incoming trunk in the crossbar tandem office has quick access to a three-digit register. The register must be connected during the interval between the last digit of the directing code and the first digit of the national number to insure registration of this number. This arrangement is used to permit the customer to dial all digits without delay and avoids the use of a second dial tone. If this arrangement were not used, the customer would be required to wait after dialing the 112 until the trunk in the tandem crossbar office could gain access to a sender through the sender link circuit which would then signal the customer to resume dialing by returning dial tone.

After recording the 915 area code digits in the case assumed, the CH3-1234 portion of the number is registered directly in the tandem sender which has been connected to the trunk while the customer was dialing 915. When the sender is attached to the trunk, it signals the three-digit register to transfer the 915 area code digits to it via a connector circuit. Thus when dialing is complete, the entire number 915-CH3-1234 is registered in the sender.

Crossbar tandem is being arranged to serve customers of panel and No. 1 crossbar offices for direct distance dialing. At the present time, ten digit direct distance dialing is not available to these customers because the digit storing equipments in these offices are limited to eight digits. Developments now under way, will provide arrangements for expanding the digit capacity in the local offices so that ultimately

calls from customers in panel and No. 1 crossbar offices may be routed through crossbar tandem or other equivalent offices to telephones anywhere in the country.

CONCLUSION

The new features developed for crossbar tandem will adapt it to switching all types of traffic at many important switching centers of the nationwide toll network. Of the 225 important toll switching centers now contemplated, it is expected that about 80 of these will be equipped with crossbar tandem.

REFERENCES

1. Collis, R. E., Crossbar Tandem System, A.I.E.E. Trans., **69**, pp. 997-1004, 1950.
2. King, G. V., Centralized Automatic Message Accounting, B.S.T.J., **33**, pp. 1331-1342, 1952.
3. Nunn, W. H., Nationwide Numbering Plan, B.S.T.J., **31**, pp. 851-859, 1952.
4. Pilliod, J. J., Fundamental Plans for Toll Telephone Plant, B.S.T.J., **31**, pp. 832-850, 1952.
5. Shipley, F. F., Automatic Toll Switching Systems, B.S.T.J., **31**, pp. 860-882, 1952.
6. Truitt, C. J., Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks, B.S.T.J., **33**, pp. 277-302, 1954.
7. Clos, C., Automatic Alternate Routing of Telephone Traffic, Bell Laboratories Record, **32**, pp. 51-57, Feb. 1954.

Growing Waves Due to Transverse Velocities

By J. R. PIERCE and L. R. WALKER

(Manuscript received March 30, 1955)

This paper treats propagation of slow waves in two-dimensional neutralized electron flow in which all electrons have the same velocity in the direction of propagation but in which there are streams of two or more velocities normal to the direction of propagation. In a finite beam in which electrons are reflected elastically at the boundaries and in which equal dc currents are carried by electrons with transverse velocities $+u_1$ and $-u_1$, there is an antisymmetrical growing wave if

$$\omega_p^2 \sim (\pi u_1/W)^2$$

and a symmetrical growing wave if

$$\omega_p^2 \sim \frac{4}{3}(\pi u_1/W)^2$$

Here ω_p is plasma frequency for the total charge density and W is beam width.

INTRODUCTION

It is well-known that there can be growing waves in electron flow when the flow is composed of several streams of electrons having different velocities in the direction of propagation of the waves.¹⁻⁵ While Birdsall⁶ considers the case of growing waves in electron flow consisting of streams which cross one another, the growing waves which he finds apparently occur when two streams have different components of velocity in the direction of propagation.

This paper shows that there can be growing waves in electron flow consisting of two or more streams with the same component of velocity in the direction of wave propagation but with different components of velocity transverse to the direction of propagation. Such growing waves can exist when the electric field varies in strength across the flow. Such waves could result in the amplification of noise fluctuations in electron flow. They could also be used to amplify signals.

Actual electron flow as it occurs in practical tubes can exhibit transverse velocities. For instance, in Brillouin flow,^{7, 8} if we consider electron motion in a coordinate system rotating with the Larmor frequency we see that electrons with transverse velocities are free to cross the beam repeatedly, being reflected at the boundaries of the beam. The transverse velocities may be completely disorganized thermal velocities, or they may be larger and better-organized velocities due to aberrations at the edges of the cathode or at lenses or apertures. Two-dimensional Brillouin flow allows similar transverse motions.

It would be difficult to treat the case of Brillouin or Brillouin-like flow with transverse velocities. Here, simpler cases with transverse velocities will be considered. The first case treated is that of infinite ion-neutralized two-dimensional flow with transverse velocities. The second case treated is that of two-dimensional flow in a beam of finite width in which the electrons are elastically reflected at the boundaries of the beam. Growing waves are found in both cases, and the rate of growth may be large.

In the case of the finite beam both an antisymmetric mode and a symmetric mode are possible. Here, it appears, the current density required for a growing wave in the symmetric mode is about $\frac{4}{3}$ times as great as the current density required for a growing wave in the anti-symmetric mode. Hence, as the current is increased, the first growing waves to arise might be antisymmetric modes, which could couple to a symmetrical resonator or helix only through a lack of symmetry or through high-level effects.

1. Infinite two-dimensional flow

Consider a two-dimensional problem in which the potential varies sinusoidally in the y direction, as $\exp(-j\beta z)$ in the z direction and as $\exp(j\omega t)$ with time. Let there be two electron streams, each of a negative charge ρ_0 and each moving with the velocity u_0 in the z direction, but with velocities u_1 and $-u_1$ respectively in the y direction. Let us denote ac quantities pertaining to the first stream by subscripts 1 and ac quantities pertaining to the second stream by subscripts 2. The ac charge density will be denoted by ρ , the ac velocity in the y direction by \dot{y} , and the ac velocity in the z direction by \dot{z} . We will use linearized or small-signal equations of motion.⁹ We will denote differentiation with respect to y by the operator D .

The equation of continuity gives

$$j\omega\rho_1 = -D(\rho_1 u_1 + \rho_0 \dot{y}_1) + j\beta(\rho_1 u_0 + \rho_0 \dot{z}_1) \quad (1.1)$$

$$j\omega\rho_2 = -D(-\rho_2 u_1 + \rho_0 \dot{y}_2) + j\beta(\rho_2 u_0 + \rho_0 \dot{z}_2) \quad (1.2)$$

Let us define

$$d_1 = j(\omega - \beta u_0) + u_1 D \quad (1.3)$$

$$d_2 = j(\omega - \beta u_0) - u_1 D \quad (1.4)$$

We can then rewrite (1.1) and (1.2) as

$$d_1 \rho_1 = \rho_0 (-D \dot{y}_1 + j\beta \dot{z}_1) \quad (1.5)$$

$$d_2 \rho_2 = \rho_0 (-D \dot{y}_2 + j\beta \dot{z}_2) \quad (1.6)$$

We will assume that we are dealing with slow waves and can use a potential V to describe the field. We can thus write the linearized equations of motion in the form

$$d_1 \dot{z}_1 = -j \frac{e}{m} \beta V \quad (1.7)$$

$$d_2 \dot{z}_2 = -j \frac{e}{m} \beta V \quad (1.8)$$

$$d_1 \dot{y}_1 = \frac{e}{m} DV \quad (1.9)$$

$$d_2 \dot{y}_2 = \frac{e}{m} DV \quad (1.10)$$

From (1.5) to (1.10) we obtain

$$d_1^2 \rho_1 = -\frac{e}{m} \rho_0 (D^2 - \beta^2) V \quad (1.11)$$

$$d_2^2 \rho_2 = -\frac{e}{m} \rho_0 (D^2 - \beta^2) V \quad (1.12)$$

Now, Poisson's equation is

$$(D^2 - \beta^2) V = -\frac{\rho_1 + \rho_2}{\epsilon} \quad (1.13)$$

From (1.11) to (1.13) we obtain

$$(D^2 - \beta^2) V = -\frac{1}{2} \omega_p^2 \left(\frac{1}{d_1^2} + \frac{1}{d_2^2} \right) (D^2 - \beta^2) V \quad (1.14)$$

$$\omega_p^2 = \frac{-2 \frac{e}{m} \rho_0}{\epsilon} \quad (1.15)$$

Here ω_p is the plasma frequency for the charge of both beams.

Either

$$(D^2 - \beta^2)V = 0 \quad (1.16)$$

or else

$$1 = \frac{-\omega_p^2}{2} \frac{(d_1^2 + d_2^2)}{d_1^2 d_2^2} \quad (1.17)$$

We will consider this second case.

We should note from (1.3) and (1.4) that

$$d_1^2 = u_1^2 D^2 - (\omega - \beta u_0)^2 + 2jD(\omega - \beta u_0)u_1 \quad (1.18)$$

$$d_2^2 = u_1^2 D^2 - (\omega - \beta u_0)^2 - 2jD(\omega - \beta u_0)u_1 \quad (1.19)$$

$$d_1^2 + d_2^2 = 2[u_1^2 D^2 - (\omega - \beta u_0)^2] \quad (1.20)$$

$$d_1^2 d_2^2 = [u_1^2 D^2 + (\omega - \beta u_0)^2]^2 \quad (1.21)$$

Thus, (1.17) becomes

$$1 = \frac{-\omega_p^2 [u_1^2 D^2 - (\omega - \beta u_0)^2]}{[u_1^2 D^2 + (\omega - \beta u_0)^2]^2} \quad (1.22)$$

If the quantities involved vary sinusoidally with y as $\cos \gamma y$ or $\sin \gamma y$, then

$$D^2 = -\gamma^2 \quad (1.23)$$

Our equation becomes

$$1 = \frac{\omega_p^2}{\gamma^2 u_1^2} \frac{\left[1 + \left(\frac{\omega - \beta u_0}{\gamma u_1}\right)^2\right]}{\left[1 - \left(\frac{\omega - \beta u_0}{\gamma u_1}\right)^2\right]} \quad (1.24)$$

What happens if we have many transverse velocities? If we refer back to (1.14) we see that we will have an equation of the form

$$1 = \sum -\frac{1}{2} \omega_{pn}^2 \left(\frac{d_{1n}^2 + d_{2n}^2}{d_{1n}^2 d_{2n}^2} \right) \quad (1.25)$$

Here ω_{pn}^2 is a plasma frequency based on the density of electrons having transverse velocities $\pm u_n$. Equation (1.25) can be written

$$1 = \sum \frac{\omega_{pn}^2}{\gamma^2 u_n^2} \frac{\left[1 + \frac{(\omega - \beta u_0)^2}{\gamma^2 u_n^2}\right]}{\left[1 - \frac{(\omega - \beta u_0)^2}{\gamma^2 u_n^2}\right]^2} \quad (1.26)$$

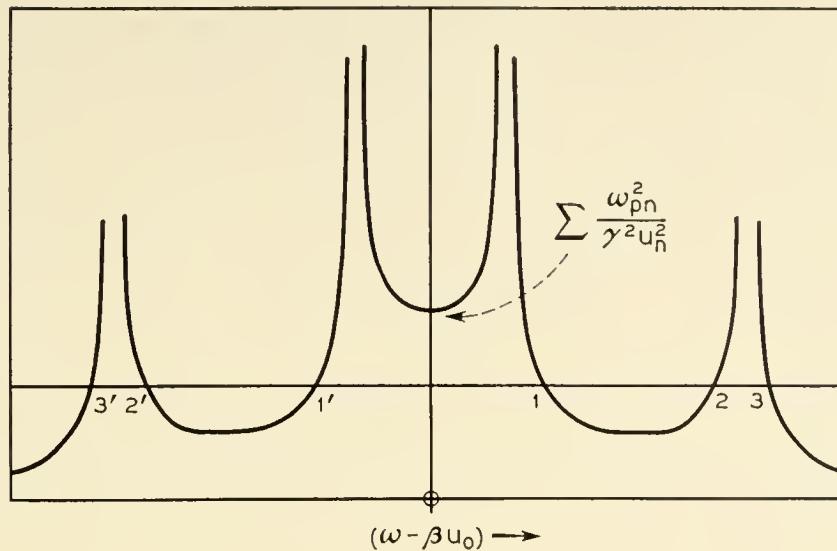


Fig. 1

Suppose we plot the left-hand and the right-hand sides of (1.26) versus $(\omega - \beta u_0)$. The general appearance of the left-hand and right-hand sides of (1.26) is indicated in Fig. 1 for the case of two velocities u_n . There will always be two unattenuated waves at values of $(\omega - \beta u_0)^2 > \gamma^2 u_e^2$ where u_e is the extreme value of u_n ; these correspond to intersections 3 and $3'$ in Fig. 2. The other waves, two per value of u_n , may be unattenuated or a pair of increasing and decreasing waves, depending on the values of the parameters. If

$$\sum \frac{\omega_{pn}^2}{\gamma^2 u_n^2} > 1$$

there will be at least one pair of increasing and decreasing waves.

It is not clear what will happen for a Maxwellian distribution of velocities. However, we must remember that various aberrations might give a very different, strongly peaked velocity distribution.

Let us consider the amount of gain in the case of one pair of transverse velocities, $\pm u_1$. The equation is now

$$\frac{\gamma^2 u_1^2}{\omega_p^2} = \frac{\left[1 + \left(\frac{\omega - \beta u_0}{\gamma u_1} \right)^2 \right]}{\left[1 - \left(\frac{\omega - \beta u_0}{\gamma u_1} \right)^2 \right]^2} \quad (1.27)$$

Let

$$\beta = \frac{\omega}{u_0} + j \frac{\gamma u_1 \epsilon}{u_0} \quad (1.28)$$

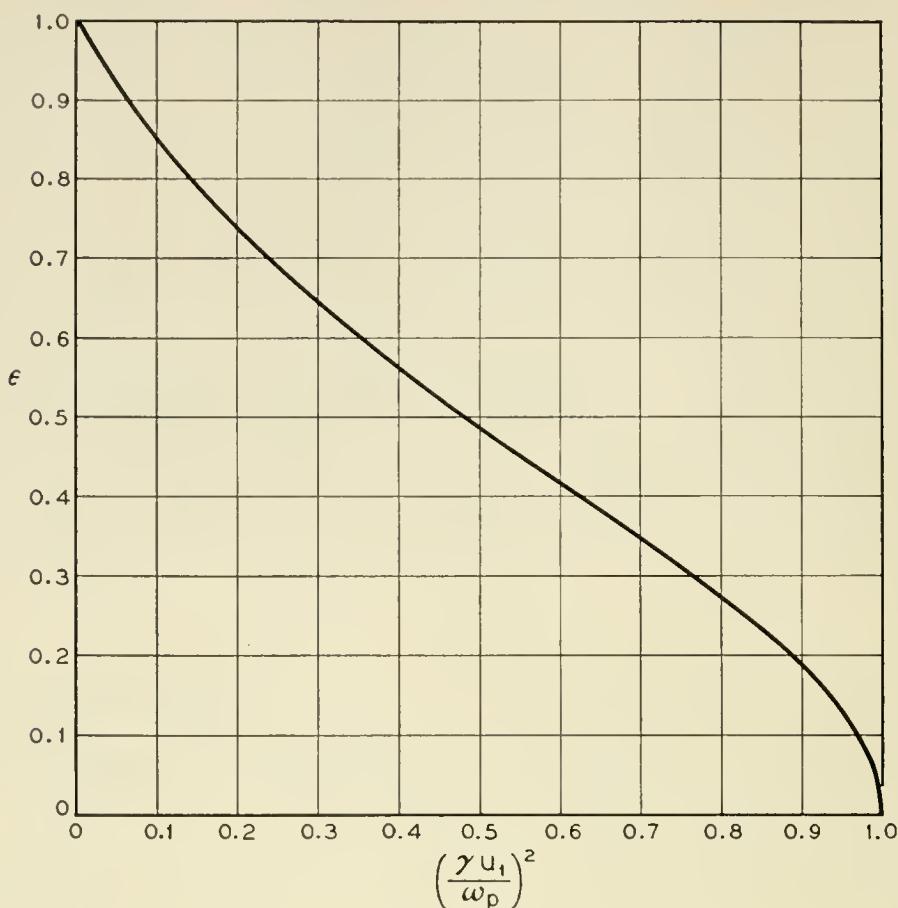


Fig. 2

This relation defines ϵ . Equation (1.27) becomes

$$\frac{\gamma^2 u_1^2}{\omega_p^2} = \frac{1 - \epsilon^2}{(1 + \epsilon^2)^2} \quad (1.29)$$

In Fig. 2, ϵ is plotted versus the parameter $\gamma^2 u_1^2 / \omega_p^2$. We see that as the parameter falls below unity, ϵ increases, at first rapidly, and then more slowly, reaching a value of ± 1 as the parameter goes to zero (as ω_p^2 goes to infinity, for instance).

It will be shown in Section 2 of this paper that these results for infinite flow are in some degree an approximation to the results for flow in narrow beams. It is therefore of interest to see what results they yield if applied to a beam of finite width.

If the beam has a length L , the voltage gain is

$$e^{\epsilon \gamma(u_1/u_0)L} \quad (1.30)$$

The gain G in db is

$$G = 8.7 \frac{\gamma u_1 L}{u_0} \epsilon \text{ db} \quad (1.31)$$

Let the width of the beam be W . We let

$$\gamma = \frac{n\pi}{W} \quad (1.32)$$

Thus, for $n = 1$, there is a half-cycle variation across the beam. From (1.31) and (1.32)

$$G = 27.3 \left(\frac{u_1}{u_0} \frac{L}{d} \right) n\epsilon \text{ db} \quad (1.33)$$

Now L/u_0 is the time it takes the electrons to go from one end of the beam to the other, while W/u_1 is the time it takes the electrons to cross the beam. If the electrons cross the beam N times

$$N = \frac{u_1}{u_0} \frac{L}{W} \quad (1.34)$$

Thus,

$$G = 27.3 N n\epsilon \text{ db} \quad (1.35)$$

While for a given value of ϵ the gain is higher if we make the phase vary many times across the beam, i.e., if we make n large, we should note that to get any gain at all we must have

$$\begin{aligned} \omega_p^2 &> \gamma^2 u_1^2 \\ \omega_p^2 &> \left(\frac{n\pi u_1}{W} \right)^2 \end{aligned} \quad (1.36)$$

If we increase ω_p^2 , which is proportional to current density, so that ω_p^2 passes through this value, the gain will rise sharply just after ω_p^2 passes through this value and will rise less rapidly thereafter.

2. A Two-Dimensional Beam of Finite Width.

Let us assume a beam of finite width in the y -direction; the boundaries lying at $y = \pm y_0$. It will be assumed also that electrons incident upon these boundaries are elastically reflected, so that electrons of the incident stream (1 or 2) are converted into those of the other stream (2 or 1). The condition of elastic reflection implies that

$$\dot{y}_1 = -\dot{y}_2 \quad (2.1)$$

$$\dot{z}_1 = \dot{z}_2 \quad \text{at } y = \pm y_0 \quad (2.2)$$

and, in addition, that

$$\rho_1 = \rho_2 \quad \text{at } y = \pm y_0 \quad (2.3)$$

since there is no change in the number of electrons at the boundary.

The equations of motion and of continuity (1.7–1.12) may be satisfied by introducing a single quantity, ψ , such that

$$V = d_1^2 d_2^2 \psi \quad (2.4)$$

$$\dot{z}_1 = -j \frac{e}{m} \beta d_1 d_2^2 \psi \quad (2.5)$$

$$\dot{z}_2 = -j \frac{e}{m} d_2^2 d_1 \psi \quad (2.6)$$

$$\dot{y}_1 = \frac{e}{m} d_1 d_2 D\psi \quad (2.7)$$

$$\dot{y}_2 = \frac{e}{m} d_1^2 d_2 D\psi \quad (2.8)$$

$$\rho_1 = -\frac{e}{m} \rho_0 (D^2 - \beta^2) d_2^2 \psi \quad (2.9)$$

$$\rho_2 = -\frac{e}{m} \rho_0 (D^2 - \beta^2) d_1^2 \psi \quad (2.10)$$

Then, if we introduce the symbol, Ω , for $\omega - \beta u_0$

$$\dot{y}_1 + \dot{y}_2 = 2j \frac{e}{m} d_1 d_2 D\Omega \psi \quad (2.11)$$

$$\dot{z}_1 - \dot{z}_2 = 2j \frac{e}{m} d_1 d_2 u_1 D\psi \quad (2.12)$$

$$\rho_1 - \rho_2 = 2j \frac{e}{m} \rho_0 (D^2 - \beta^2) u_1 \Omega D\psi \quad (2.13)$$

It is clear that if

$$D\psi = D^3\psi = 0 \quad y = \pm y_0 \quad (2.14)$$

the conditions for elastic reflection will be satisfied. The equation satisfied by ψ may now be found from Poisson's equation, (1–13), and is

$$(D^2 - \beta^2) d_1^2 d_2^2 \psi = \frac{e\rho_0}{m\epsilon} (D^2 - \beta^2) (d_1^2 + d_2^2) \psi$$

or

$$(D^2 - \beta^2)[(u_1^2 D^2 + \Omega^2)^2 + \omega_p^2 (u_1^2 D^2 - \Omega^2)] = 0 \quad (2.15)$$

which is of the sixth degree in D . So far four boundary conditions have been imposed. The remaining necessary pair arise from matching the

internal fields to the external ones. For $y > y_0$

$$V = V_0 e^{-j\beta z} \cdot e^{-\beta y} \quad (2.16)$$

and

$$\frac{\partial V}{\partial y} + \beta V = 0 \quad \text{at } y = y_0$$

Similarly

$$\frac{\partial V}{\partial y} - \beta V = 0 \quad \text{at } y = -y_0 \quad (2.17)$$

The most familiar procedure now would be to look for solutions of (2.15) of the form, e^{cy} . This would give the sextic for c

$$(c^2 - \beta^2)[(u_1^2 c^2 + \Omega^2)^2 + \omega_p^2(u_1^2 c^2 - \Omega^2)] = 0 \quad (2.18)$$

with the roots $c = \pm\beta, \pm c_1, \pm c_2$, let us say. We could then express ψ as a linear combination of these six solutions and adjust the coefficients to satisfy the six boundary equations. In this way a characteristic equation for β would be obtained. From the symmetry of the problem this has the general form $F(\beta, c_1) = F(\beta, c_2)$, where c_1 and c_2 are found from (2.18). The discussion of the problem in these terms is rather laborious and, if we are concerned mainly with examining qualitatively the onset of increasing waves, another approach serves better.

From the symmetry of the equations and of the boundary conditions we see that there are solutions for ψ (and consequently for V and ρ) which are even in y and again some which are odd in y . Consider first the even solutions. We will assume that there is an even function, $\psi_1(y)$, periodic in y with period $2y_0$, which coincides with $\psi(y)$ in the open interval, $-y_0 < y < y_0$ and that $\psi_1(y)$ has a Fourier cosine series representation:

$$\psi_1(y) = \sum_1^{\infty} c_n \cos \lambda_n y \quad \lambda_n = \frac{n\pi}{y_0} \quad n = 0, 1, 2, \dots \quad (2.19)$$

ψ inside the interval satisfies (2.15), so we assume that $\psi_1(y)$ obeys

$$(D^2 - \beta^2)[(u_1^2 D^2 + \Omega^2)^2 + \omega_p^2(u_1^2 D^2 - \Omega^2)]\psi_1 = \sum_{m=-\infty}^{+\infty} \delta(y - 2m + 1)y_0 \quad (2.20)$$

where δ is the familiar δ -function. Since $D\psi$ and $D^3\psi$ are required to vanish at the ends of the interval and $\psi, D^2\psi$ and $D^4\psi$ are even it follows that all

of these functions are continuous. We assume that $\psi_1 = \psi$, $D\psi_1 = D\psi$, $D^2\psi_1 = D^2\psi$, $D^3\psi_1 = D^3\psi$ and $D^4\psi_1 = D^4\psi$ at the ends of the intervals. From (2.20), $u_1^4 D^5\psi_1 \rightarrow -\frac{1}{2}$ as $y \rightarrow y_0$.

Since

$$\sum_{-\infty}^{+\infty} \delta(y - 2m + 1)y_0) = \frac{1}{2y_0} + \frac{1}{y_0} \sum_1^{\infty} (-1)^n \cos \lambda_n y \quad (2.21)$$

we obtain from (2.20)

$$2y_0\psi_1 = -\left(\frac{1}{\beta^2 \Omega^2 (\Omega^2 - \omega_p^2)} \right. \\ \left. + 2 \sum_1^{\infty} (-1)^n \frac{\cos \lambda_n y}{(\beta^2 + \lambda_n^2)[(\Omega^2 - u_1^2 \lambda_n^2)^2 - \omega_p^2 (\Omega^2 + u_1^2 \lambda_n^2)]} \right) \quad (2.22)$$

Since

$$\frac{\partial V}{\partial y} + \beta V = (D + \beta)(u_1^2 D^2 + \Omega^2)^2 \psi,$$

using (2.4), the condition for matching to the external field,

$$\frac{\partial V}{\partial y} + \beta V = 0,$$

yields, using $D\psi = D^3\psi = 0$ and $u_1^4 D^5\psi = -\frac{1}{2}$, the relation

$$(u_1^2 D^2 + \Omega^2)^2 \psi_1 = \frac{1}{2}\beta \quad \text{at } y = y_0.$$

Applying this to (2.22), we then obtain, finally,

$$\frac{y_0}{\beta} = \frac{1}{\beta^2 [\Omega^2 - \omega_p^2]} \\ + 2 \sum_1^{\infty} \frac{(\Omega^2 - u_1^2 \lambda_n^2)^2}{(\beta^2 + \lambda_n^2)[(\Omega^2 - u_1^2 \lambda_n^2)^2 - \omega_p^2 (\Omega^2 + u_1^2 \lambda_n^2)]} \quad (2.23)$$

For the odd solution we use a function, $\psi_2(y)$, equal to $\psi(y)$ in $-y_0 < y < y_0$ and representable by a sine series. To ensure the vanishing of $D\psi$ and $D^3\psi$ at $y = \pm y_0$ it is appropriate to use the functions, $\sin \mu_n y$, where $\mu_n = (n + \frac{1}{2})\pi/y_0$. The period is now $4y_0$ and we define $\psi_2(y)$ in $y_0 < y < 3y_0$ by the relation $\psi_2(y) = \psi(2y_0 - y)$ and in $-3y_0 < y < -y_0$ by $\psi_2(y) = \psi(-2y_0 - y)$. Thus, we write

$$\psi_2(y) = \sum_0^{\infty} d_n \sin \mu_n y \quad \mu_n = (n + \frac{1}{2})\pi/y_0$$

$\psi_2(y)$ will be supposed to satisfy

$$(D^2 - \beta^2)[(u_1^2 D^2 + \Omega^2)^2 + \omega_p^2(u_1^2 D^2 - \Omega^2)]\psi_2 \\ = \sum_{m=-\infty}^{+\infty} [\delta(y - \overline{4m+1}y_0) - \delta(y - \overline{4m-1}y_0)] \quad (2.24)$$

The extended definition of ψ_2 (outside $-y_0 < y < y_0$) is such that we may again take $\psi_1 = \psi, \dots, D^4\psi_1 = D^4\psi$ at the ends of the interval. $u_1^4 D^5\psi_1$ is still equal to $-\frac{1}{2}$ at $y = y_0$. Now

$$\sum_{-\infty}^{+\infty} [\delta(y - \overline{4m+1}y_0) - \delta(y - \overline{4m-1}y_0)] \\ = \frac{1}{y_0} \sum (-1)^n \sin \mu_n y \quad (2.25)$$

so from (2.24) we may find

$$y_0\psi_2 = - \sum_0^{\infty} \frac{(-1)^n \sin \mu_n y}{(\beta^2 + \mu_n^2)[(\Omega^2 - u_1^2 \mu_n^2)^2 - \omega_p^2(\Omega^2 + u_1^2 \mu_n^2)]} \quad (2.26)$$

Matching to the external field as before gives

$$(u_1^2 D^2 + \Omega^2)^2 \psi_2 = \frac{1}{2\beta} \quad \text{at } y = y_0$$

and applied to (2.26) we have

$$-\frac{y_0}{2\beta} = \sum_0^{\infty} \frac{(\Omega^2 - u_1^2 \mu_n^2)^2}{(\beta^2 + \mu_n^2)[(\Omega^2 - u_1^2 \mu_n^2)^2 - \omega_p^2(\Omega^2 + u_1^2 \mu_n^2)]} \quad (2.27)$$

The equations (2.23) and (2.27) for the even and odd modes may be rewritten using the following reduced variables.

$$z = \frac{\beta y_0}{\pi} \\ k = \frac{\omega y_0}{\pi u_1} - \frac{u_0}{u_1} z \\ \delta^2 = \frac{\omega_p^2 y_0^2}{\pi^2 u_1^2}$$

(2.23) becomes

$$\frac{k^2}{k^2 - \delta^2} + 2 \sum_{n=1}^{\infty} \frac{z^2}{z^2 + n^2} \cdot \frac{(n^2 - k^2)^2}{(n^2 - k^2)^2 - \delta^2(n^2 + k^2)} = -\pi z \quad (2.28)$$

and (2.27) transforms to

$$2 \sum_{n=0}^{\infty} \frac{z^2}{z^2 + (n + \frac{1}{2})^2} \cdot \frac{[(n + \frac{1}{2})^2 - k^2]^2}{[(n + \frac{1}{2})^2 - k^2]^2 - \delta^2[(n + \frac{1}{2})^2 + k^2]} \\ = -\pi z \quad (2.29)$$

We shall assume in considering (2.28) and (2.29) that the beam is sufficiently wide for the transit of an electron from one side to the other to take a few RF cycles. The number of cycles is in fact, $\omega y_0/\pi u_1$, and, hence, from the definition of z , we see that for values of k less than 2, perhaps, z is certainly positive.

Let us consider (2.29) first since it proves to be the simpler case. If we transfer the term πz to the right hand side, it follows from the observation that z is positive (for modest values of k), that it is necessary to make the sum negative. The sum may be studied qualitatively by sketching in the $k^2 - \delta^2$ plane the lines on which the individual terms go to infinity, given by

$$\delta^2 = \frac{[(n + \frac{1}{2})^2 - k^2]^2}{(n + \frac{1}{2})^2 + k^2} \quad (2.30)$$

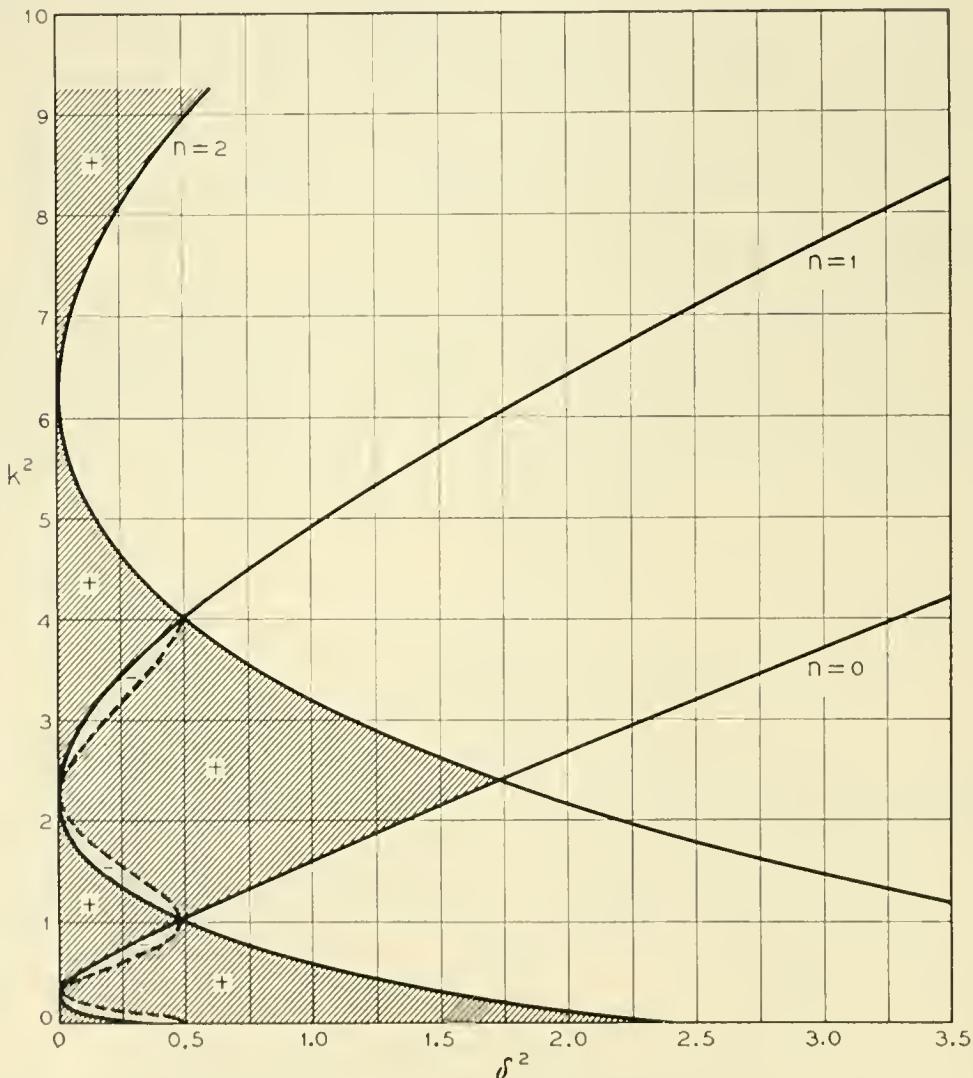


Fig. 3

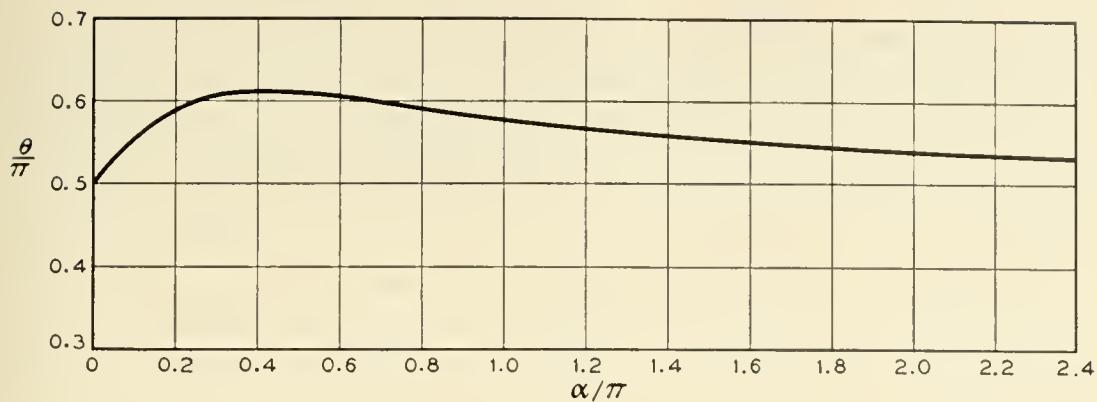


Fig. 4

Fig. 3 shows a few such curves ($n = 0, 1, 2$). To the right of such curves the individual term in question is negative, except on the line, $k^2 = (n + \frac{1}{2})^2$, where it attains the value of zero. Approaching the curves from the right the terms go to $-\infty$. On the left of the curves the function is positive and goes to $+\infty$ as the curve is approached from the

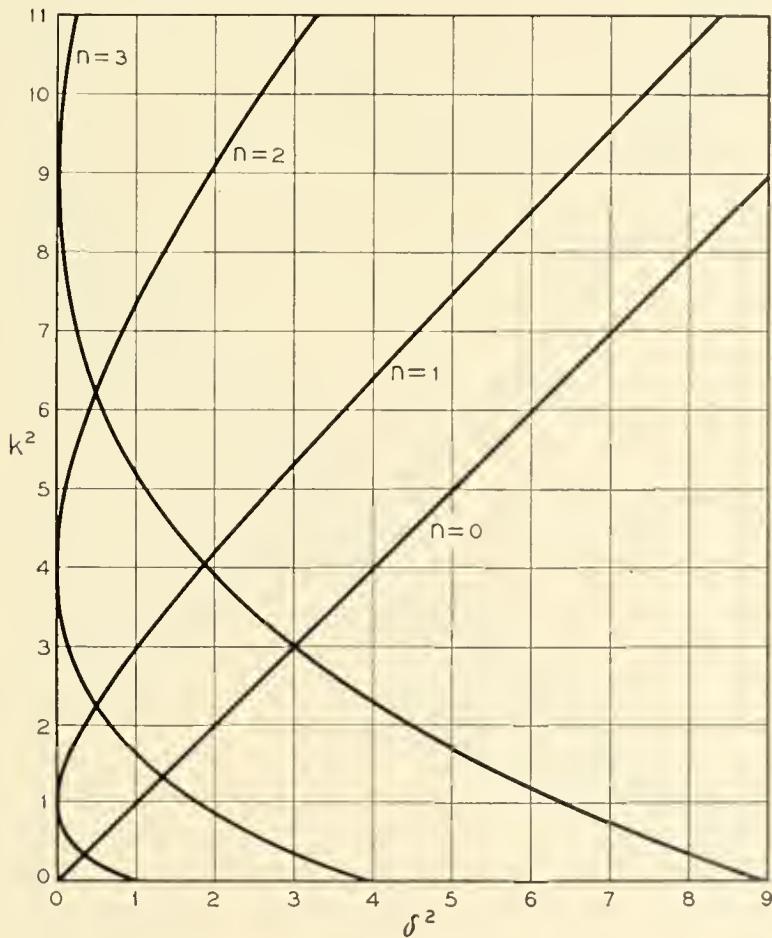


Fig. 5

left. Clearly in the regions marked + which lie to the left of every curve given by (2.30), the sum is positive and we cannot have roots. Let us examine the sum in the region to the right of the $n = 0$ curve and to the left of all others. On the line, $k^2 = \frac{1}{4}$, the sum is positive, since the first term is zero. On any other line, $k^2 = \text{constant}$, the sum goes from $+\infty$ at the $n = 1$ curve monotonically to $-\infty$ at the $n = 0$ curve, so that somewhere it must pass through 0. This enables us to draw the zero-sum contours qualitatively in this region and they are indicated in Fig. 3. We are now in a position to follow the variation in the sum as k varies at fixed δ^2 . It is readily seen that for $\delta^2 < 0.25$, because $-\pi z$ is negative in the region under consideration, there will be four real roots, two for positive, two for negative k . For δ^2 slightly greater than 0.25, the sum has

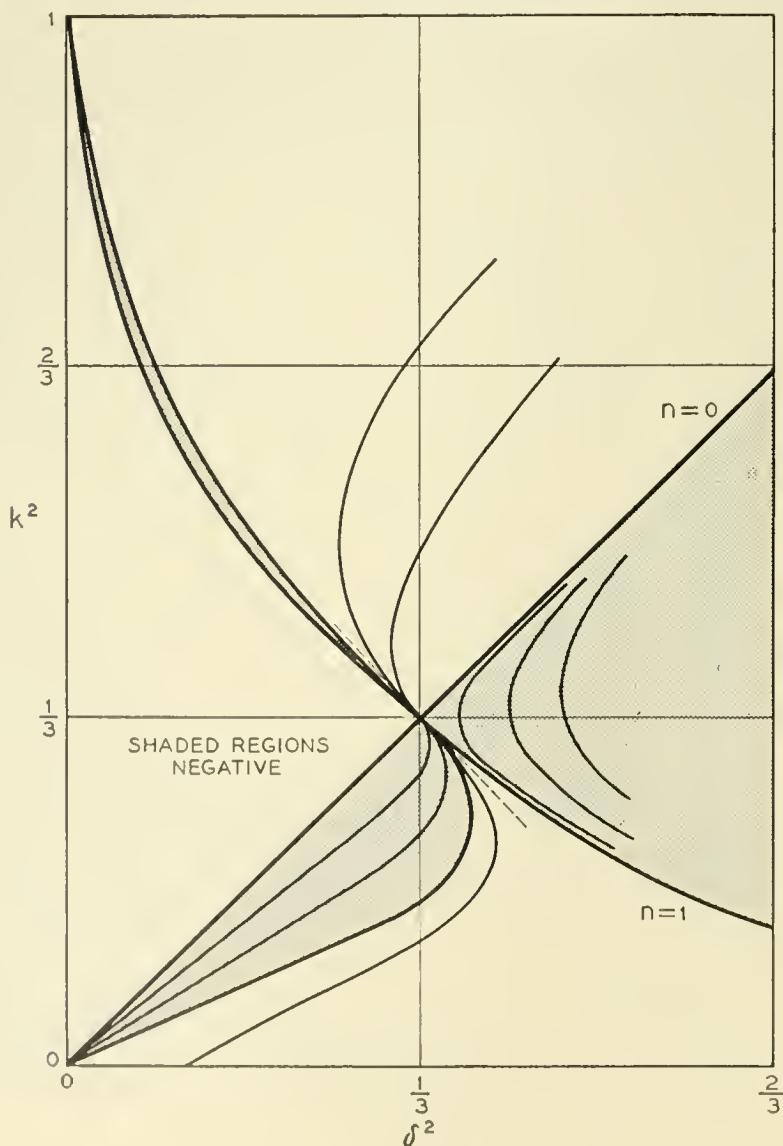


Fig. 6A

a deep minimum for $k = 0$, so that there are still four real roots unless z is very large. For z fixed, as δ^2 increases, the depth of the minimum decreases and there will finally occur a δ^2 for which the minimum is so shallow that two of the real roots disappear. Call $z(0)$ the value of z for $k = 0$, write the sum as $\Sigma(\delta^2, k^2)$ and suppose that $\Sigma(\delta_0^2, 0) = -\pi z(0)$, then for small k we have

$$\begin{aligned}\Sigma(\delta^2, k^2) &= -\pi z(0) + (\delta^2 - \delta_0^2) \frac{\partial \Sigma}{\partial \delta^2} + k^2 \frac{\partial \Sigma}{\partial k^2} = -\pi z(0) - \frac{u_1}{u_0} k \\ k^2 - \frac{u_1/u_0}{\frac{\partial \Sigma}{\partial k^2}} k &= \frac{\frac{\partial \Sigma}{\partial \delta^2}}{\frac{\partial \Sigma}{\partial k^2}} (\delta^2 - \delta_0^2) \\ k &= \frac{u_1/u_0}{2 \frac{\partial \Sigma}{\partial k^2}} \pm \sqrt{\frac{\frac{\partial \Sigma}{\partial \delta^2}}{\frac{\partial \Sigma}{\partial k^2}} (\delta^2 - \delta_0^2) + \left(\frac{u_1/u_0}{2 \frac{\partial \Sigma}{\partial k^2}} \right)^2}\end{aligned}$$

The roots become complex when

$$\delta^2 = \delta_0^2 - \frac{(u_1/u_0)^2}{4 \frac{\partial \Sigma}{\partial \delta^2} \frac{\partial \Sigma}{\partial k^2}}$$

Since u_1/u_0 may be considered small (say 10 per cent) it is sufficient to look for the values of δ_0^2 .

When $k^2 = 0$ we have

$$\begin{aligned}-\pi z &= 2 \sum \frac{z^2}{z^2 + (n + \frac{1}{2})^2} \cdot \frac{(n + \frac{1}{2})^2}{(n + \frac{1}{2})^2 - \delta^2} \\ &= \frac{2z^2}{z^2 + \delta^2} \sum_0^\infty \left(\frac{\delta^2}{(n + \frac{1}{2})^2 - \delta^2} + \frac{z^2}{(n + \frac{1}{2})^2 + z^2} \right) \\ &= \frac{\pi z^2}{z^2 + \delta^2} (\delta \tan \pi \delta + z \tanh \pi z)\end{aligned}$$

Fig. 4 shows the solution of this equation for various $z(0)$ or $\omega y_0/\pi u_0$. Clearly the threshold δ is rather insensitive to variations in $\omega y_0/\pi u_0$.

Equation (2.28) may be examined by a similar method, but here some complications arise. Fig. 5 shows the infinity curves for $n = 0, 1, 2, 3$; the $n = 0$ term being of the form $k^2/k^2 - \delta^2$. The lowest critical region in δ^2 is the neighborhood of the point $k^2 = \delta^2 = \frac{1}{3}$, which is the intersection of the $n = 0$ and $n = 1$ lines. To obtain an idea of the behavior of

the left hand side (l.h.s.) of (2.28) in this area we first see how the point $k^2 = \delta^2 = \frac{1}{3}$ can be approached so that the l.h.s. remains finite. If we put $k^2 = \frac{1}{3} + \varepsilon$ and $\sigma^2 = \frac{1}{3} + c\varepsilon$ and expand the first two dominant terms of (2.28), then adjust c to keep the result finite as $\varepsilon \rightarrow 0$ we find

$$c = \frac{1}{4} \frac{3z^2 - 5}{3z^2 + 1}$$

c varies from $-\frac{5}{4}$ to $\frac{1}{4}$ as z goes from 0 to ∞ , changing sign at $z^2 = \frac{5}{3}$. Every curve for which the l.h.s. is constant makes quadratic contact with the line $\delta^2 - \frac{1}{3} = c(k^2 - \frac{1}{3})$ at $k^2 = \delta^2 = \frac{1}{3}$. If we remember that the l.h.s. is positive for $k^2 = 0, 0 < \delta^2 < 1$ and for $k^2 = 1, 0 < \delta^2 < 1$,

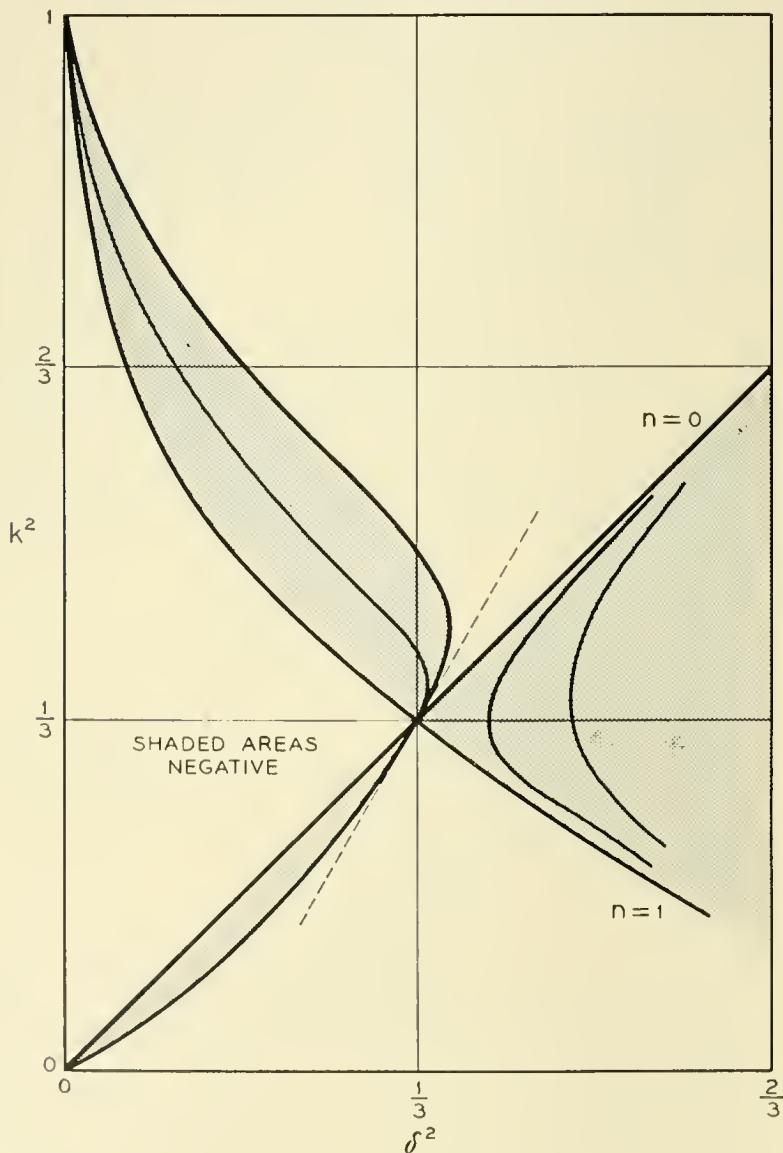
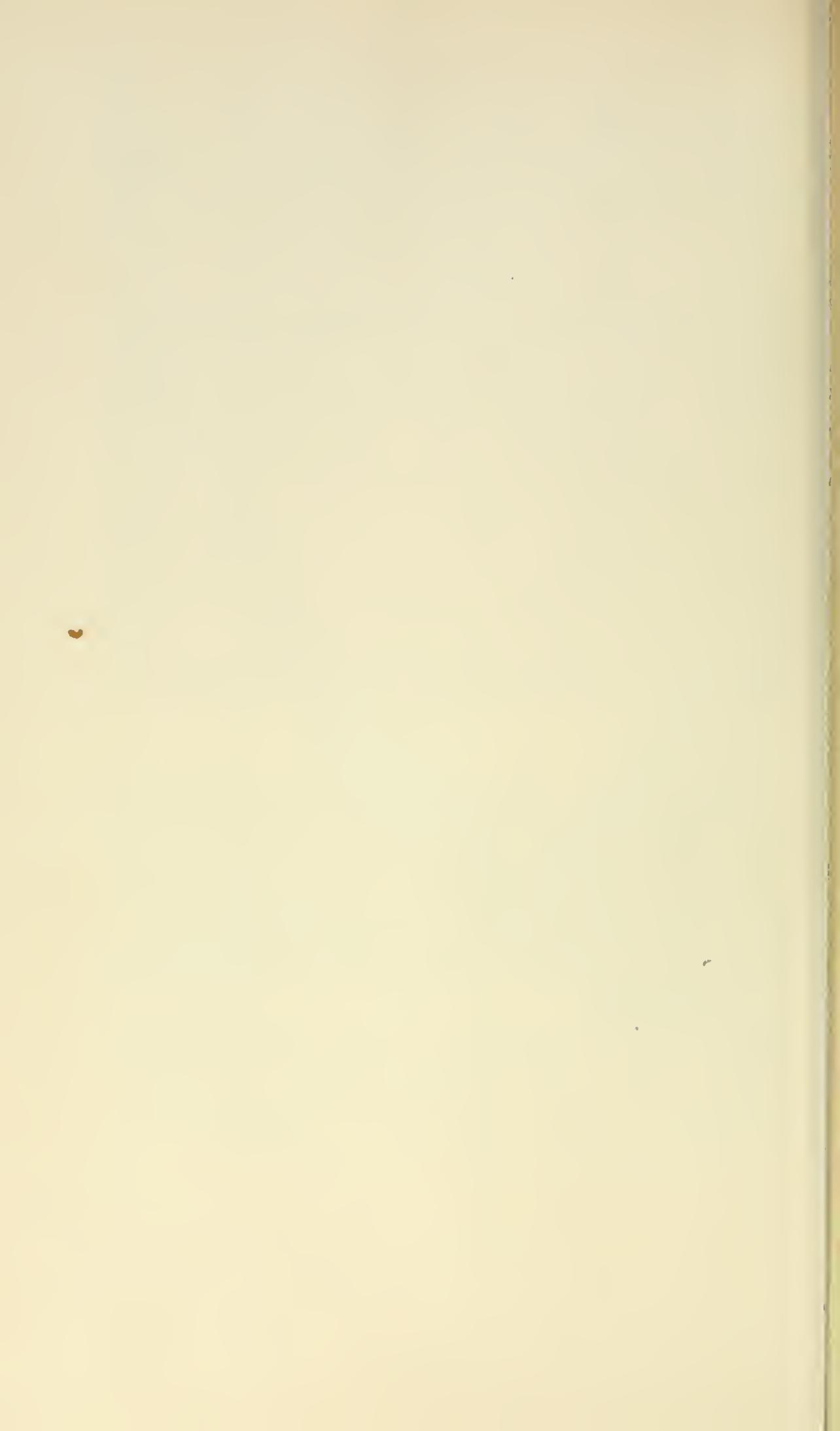


Fig. 6B

since there are no negative terms in the sum for these ranges and again that the l.h.s. must change sign between the $n = 0$ and $n = 1$ lines for any k^2 in the range $0 < k^2 < 1$ (since it varies from $\mp\infty$ to $\pm\infty$), this information may be combined with that about the immediate vicinity of $\delta^2 = k^2 = \frac{1}{3}$ to enable us to draw a line on which the l.h.s. is zero. This is indicated in Figs. 6A and 6B for small z and large z respectively. It will be seen that the zero curve and, in fact, all curves on which the l.h.s. is equal to a negative constant are required to have a vertical tangent at some point. This point may be above or below $k^2 = \frac{1}{3}$ (depending upon the sign of c or the size of z) but always at a $\delta^2 > \frac{1}{3}$. For $\delta^2 < \frac{1}{3}$ there are no regions where roots can arise as we can readily see by considering how the l.h.s. varies with k^2 at fixed δ^2 . For a fixed $\delta^2 > \frac{1}{3}$ we have, then, either for $k^2 > \frac{1}{3}$ or $k^2 < \frac{1}{3}$, according to the size of z , a negative minimum which becomes indefinitely deep as $\delta^2 \rightarrow \frac{1}{3}$. Thus, since the negative terms on the right-hand side are not sensitive to small changes in δ^2 , we must expect to find, for a fixed value of the l.h.s., two real solutions of (2.28) for some values of δ^2 and no real solutions for some larger value of δ^2 , since the negative minimum of the l.h.s. may be made as shallow as we like by increasing δ^2 . By continuity then we expect to find pairs of complex roots in this region. Rather oddly these roots, which will exist certainly for δ^2 sufficiently close to $\frac{1}{3} + 0$, will disappear if δ^2 is sufficiently increased.

REFERENCES

1. L. S. Nergaard, Analysis of a Simple Model of a Two-Beam Growing-Wave Tube, *RCA Review*, **9**, pp. 585–601, Dec., 1948.
2. J. R. Pierce and W. B. Hebenstreit, A New Type of High-Frequency Amplifier, *B. S. T. J.*, **28**, pp. 23–51, Jan., 1949.
3. A. V. Haeff, The Electron-Wave Tube — A Novel Method of Generation and Amplification of Microwave Energy, *Proc. I.R.E.*, **37**, pp. 4–10, Jan., 1949.
4. G. G. Macfarlane and H. G. Hay, Wave Propagation in a Slipping Stream of Electrons, *Proc. Physical Society Sec. B*, **63**, pp. 409–427, June, 1950.
5. P. Guénard and H. Huber, Étude Expérimentale de L'Interaction par Ondes de Charge d'Espace au Sein d'Un Faisceau Électronique se Déplaçant dans Des Champs Electrique et Magnétique Croisés, *Annales de Radioélectricité*, **7**, pp. 252–278, Oct., 1952.
6. C. K. Birdsall, Double Stream Amplification Due to Interaction Between Two Oblique Electron Streams, Technical Report No. 24, Electronics Research Laboratory, Stanford University.
7. L. Brillouin, A Theorem of Larmor and Its Importance for Electrons in Magnetic Fields, *Phys. Rev.*, **67**, pp. 260–266, 1945.
8. J. R. Pierce, Theory and Design of Electron Beams, 2nd Ed., Chapter 9, Van Nostrand, 1954.
9. J. R. Pierce, Traveling-Wave Tubes, Van Nostrand, 1950.



Coupled Helices

By J. S. COOK, R. KOMPNER and C. F. QUATE

(Received September 21, 1955)

An analysis of coupled helices is presented, using the transmission line approach and also the field approach, with the objective of providing the tube designer and the microwave circuit engineer with a basis for approximate calculations. Devices based on the presence of only one mode of propagation are briefly described; and methods for establishing such a mode are given. Devices depending on the simultaneous presence of both modes, that is, depending on the beat wave phenomenon, are described; some experimental results are cited in support of the view that a novel and useful class of coupling elements has been discovered.

CONTENTS

1. Introduction	129
2. Theory of Coupled Helices	132
2.1 Introduction	132
2.2 Transmission Line Equations	133
2.3 Solution for Synchronous Helices	135
2.4 Non-Synchronous Helix Solutions	137
2.5 A Look at the Fields	139
2.6 A Simple Estimate of b and x	141
2.7 Strength of Coupling versus Frequency	142
2.8 Field Solutions	144
2.9 Bifilar Helix	146
2.10 Effect of Dielectric Material between Helices	148
2.11 The Conditions for Maximum Power Transfer	151
2.12 Mode Impedance	152
3. Applications of Coupled Helices	154
3.1 Excitation of Pure Modes	156
3.1.1 Direct Excitation	156
3.1.2 Tapered Coupler	157
3.1.3 Stepped Coupler	158
3.2 Low Noise Transverse Field Amplifier	159
3.3 Dispersive Traveling Wave Tube	159
3.4 Devices Using Both Modes	161
3.4.1 Coupled Helix Transduceer	161
3.4.2 Coupled-Helix Attenuator	165
4. Conclusion	167
Appendix	
I Solution of Field Equations	168
II Finding \tilde{r}	173
III Complete Power Transfer	175

GLOSSARY OF SYMBOLS

<i>a</i>	Mean radius of inner helix
<i>b</i>	Mean radius of outer helix
<i>b</i>	Capacitive coupling coefficient
$B_{10, 20}$	shunt susceptance of inner and outer helices, respectively
$B_{1, 2}$	Shunt susceptance plus mutual susceptance of inner and outer helices, respectively, $B_{10} + B_m, B_{20} + B_m$
B_m	Mutual susceptance of two coupled helices
<i>c</i>	Velocity of light in free space
<i>d</i>	Radial separation between helices, $b-a$
<i>D</i>	Directivity of helix coupler
<i>E</i>	Electric field intensity
<i>F</i>	Maximum fraction of power transferable from one coupled helix to the other
$F(\gamma a)$	Impedance parameter
$I_{1, 2}$	RF current in inner and outer helix, respectively
<i>K</i>	Impedance in terms of longitudinal electric field on helix axis and axial power flow
<i>L</i>	Minimum axial distance required for maximum energy transfer from one coupled helix to the other, $\lambda_b/2$
<i>P</i>	Axial power flow along helix circuit
<i>r</i>	Radial coordinate
\bar{r}	Radius where longitudinal component of electric field is zero for transverse mode (about midway between <i>a</i> and <i>b</i>)
<i>R</i>	Return loss
<i>s</i>	Radial separation between helix and adjacent conducting shield
<i>t</i>	Time
$V_{1, 2}$	RF potential of inner and outer helices, respectively
<i>x</i>	Inductive coupling coefficient
$X_{10, 20}$	Series reactance of inner and outer helices, respectively
$X_{1, 2}$	Series reactance plus mutual reactance of inner and outer helices, respectively, $X_{10} + X_m, X_{20} + X_m$
X_m	Mutual reactance of two coupled helices
<i>z</i>	Axial coordinate
$Z_{1, 2}$	Impedance of inner and outer helix, respectively
$\alpha_{1, 2}$	Attenuation constant of inner and outer helices, respectively
β	General circuit phase constant; or mean circuit phase constant, $\sqrt{\beta_1\beta_2}$
β_0	Free space phase constant
$\beta_{10, 20}$	Axial phase constant of inner and outer helices in absence of coupling, $\sqrt{B_{10}X_{10}}, \sqrt{B_{20}X_{20}}$

β_1, β_2	May be considered as axial phase constant of inner and outer helices, respectively
β_b	Beat phase constant
β_c	Coupling phase constant, (identical with β_b when $\beta_1 = \beta_2$)
$\beta_c\epsilon$	Coupling phase constant when there is dielectric material between the helices
β_d	Difference phase constant, $ \beta_1 - \beta_2 $
β_ϵ	Axial phase constant of single helix in presence of dielectric
$\beta_{t, \ell}$	Axial phase constant of transverse and longitudinal modes, respectively
γ	Radial phase constant
$\gamma_{t, \ell}$	Radial phase constant of transverse and longitudinal modes, respectively
Γ	Axial propagation constant
$\Gamma_{t, \ell}$	Axial propagation constant for transverse and longitudinal coupled-helix modes, respectively
ϵ	Dielectric constant
ϵ'	Relative dielectric constant, ϵ/ϵ_0
ϵ_0	Dielectric constant of free space
λ	General circuit wavelength; or mean circuit wavelength, $\sqrt{\lambda_1\lambda_2}$
λ_0	Free space wavelength
$\lambda_{1, 2}$	Axial wavelength on inner and outer helix, respectively
λ_b	Beat wavelength
λ_c	Coupling wavelength (identical with λ_b when $\beta_1 = \beta_2$)
ψ	Helix pitch angle
$\psi_{1, 2}$	Pitch angle of inner and outer helix, respectively
ω	Angular frequency

1. INTRODUCTION

Since their first appearance, traveling-wave tubes have changed only very little. In particular, if we divide the tube, somewhat arbitrarily, into circuit and beam, the most widely used circuit is still the helix, and the most widely used transition from the circuits outside the tube to the circuit inside is from waveguide to a short stub or antenna which, in turn, is attached to the helix, either directly or through a few turns of increased pitch. Feedback of signal energy along the helix is prevented by means of loss, either distributed along the whole helix or localized somewhere near the middle. The helix is most often supported along its whole length by glass or ceramic rods, which also serve to carry a conducting coating ("aquadag"), acting as the localized loss.

We therefore find the following circuit elements within the tube envelope, fixed and inaccessible once and for all after it has been sealed off:

1. The helix itself, determining the beam voltage for optimum beam-circuit interaction;
2. The helix ends and matching stubs, etc., all of which have to be positioned very precisely with relation to the waveguide circuits in order to obtain a reproducible match;
3. The loss, in the form of "aquadag" on the support rods, which greatly influences the tube performance by its position and distribution.

In spite of the enormous bandwidth over which the traveling-wave tube is potentially capable of operating — a feature new in the field of microwave amplifier tubes — it turns out that the positioning of the tube in the external circuits and the necessary matching adjustments are rather critical; moreover the overall bandwidths achieved are far short of the obtainable maximum.

Another fact, experimentally observed and well-founded in theory, rounds off the situation: The electro-magnetic field surrounding a helix, i.e., the slow wave, under normal conditions, does not radiate, and is confined to the close vicinity of the helix, falling off in intensity nearly exponentially with distance from the helix. A typical traveling-wave tube, in which the helix is supported by ceramic rods, and the whole enclosed by the glass envelope, is thus practically inaccessible as far as RF fields are concerned, with the exception of the ends of the helix, where provision is made for matching to the outside circuits. Placing objects such as conductors, dielectrics or distributed loss close to the tube is, in general, observed to have no effect whatsoever.

In the course of an experimental investigation into the propagation of space charge waves in electron beams it was desired to couple into a long helix at any point chosen along its length. Because of the feebleness of the RF fields outside the helix surrounded by the conventional supports and the envelope, this seemed a rather difficult task. Nevertheless, if accomplished, such a coupling would have other and even more important applications; and a good deal of thought was given to the problem.

Coupled concentric helices were found to provide the solution to the problem of coupling into and out of a helix at any particular point, and to a number of other problems too.

Concentric coupled helices have been considered by J. R. Pierce,¹ who has treated the problem mainly with transverse fields in mind. Such fields were thought to be useful in low-noise traveling-wave tube devices. Pierce's analysis treats the helices as transmission lines coupled uniformly over their length by means of mutual distributed capacitance and inductance. Pierce also recognized that it is necessary to wind the

two helices in opposite directions in order to obtain well defined transverse and axial wave modes which are well separated in respect to their velocities of propagation.

Pierce did not then give an estimate of the velocity separation which might be attainable with practical helices, nor did anybody (as far as we are aware) then know how strong a coupling one might obtain with such helices.

It was, therefore, a considerable (and gratifying) surprise^{2, 3} to find that concentric helices of practically realizable dimensions and separations are, indeed, very strongly coupled when, and these are the important points,

(a) They have very nearly equal velocities of propagation when uncoupled, and when

(b) They are wound in opposite senses.

It was found that virtually complete power transfer from outer to inner helix (or vice versa) could be effected over a distance of the order of *one* helix wavelength (normally between $\frac{1}{10}$ and $\frac{1}{20}$ of a free-space wavelength).

It was also found that it was possible to make a transition from a coaxial transmission line to a short (outer) helix and thence through the glass surrounding an inner helix, which was fairly good over quite a considerable bandwidth. Such a transition also acted as a directional coupler, RF power coming from the coaxial line being transferred to the inner helix predominantly in one direction.

Thus, one of the shortcomings of the "conventional" helix traveling-wave tube, namely the necessary built-in accuracy of the matching parameters, was overcome by means of the new type of coupler that might evolve around coupled helix-to-helix systems.

Other constructional and functional possibilities appeared as the work progressed, such as coupled-helix attenuators, various types of broadband couplers, and schemes for exciting pure transverse (slow) or longitudinal (fast) waves on coupled helices.

One central fact emerged from all these considerations: by placing part of the circuit outside the tube envelope with complete independence from the helix terminations inside the tube, coupled helices give back to the circuit designer a freedom comparable only with that obtained at much lower frequencies. For example, it now appears entirely possible to make one type of traveling wave tube to cover a variety of frequency bands, each band requiring merely different couplers or outside helices, the tube itself remaining unchanged.

Moreover, one tube may now be made to fulfill a number of different

functions; this is made possible by the freedom with which couplers and attenuators can be placed at any chosen point along the tube.

Considerable work in this field has been done elsewhere. Reference will be made to it wherever possible. However, only that work with which the authors have been intimately connected will be fully reported here. In particular, the effect of the electron beam on the wave propagation phenomena will not be considered.

2. THEORY OF COUPLED HELICES

2.1 *Introduction*

In the past, considerable success has been attained in the understanding of traveling wave tube behavior by means of the so-called "transmission-line" approach to the theory. In particular, J. R. Pierce used it in his initial analysis and was thus able to present the solution of the so-called traveling-wave tube equations in the form of 4 waves, one of which is an exponentially growing forward traveling wave basic to the operation of the tube as an amplifier.

This transmission-line approach considers the helix — or any slow-wave circuit for that matter — as a transmission line with distributed capacitance and inductance with which an electron beam interacts. As the first approximation, the beam is assumed to be moving in an RF field of uniform intensity across the beam.

In this way very simple expressions for the coupling parameter and gain, etc., are obtained, which give one a good appreciation of the physically relevant quantities.

A number of factors, such as the effect of space charge, the non-uniform distribution of the electric field, the variation of circuit impedance with frequency, etc., can, in principle, be calculated and their effects can be superimposed, so to speak, on the relatively simple expressions deriving from the simple transmission line theory. This has, in fact, been done and is, from the design engineer's point of view, quite satisfactory.

However, physicists are bound to be unhappy over this state of affairs. In the beginning was Maxwell, and therefore the proper point to start from is Maxwell.

So-called "Field" theories of traveling-wave tubes, based on Maxwell's equation, solved with the appropriate boundary conditions, have been worked out and their main importance is that they largely confirm the results obtained by the inexact transmission line theory. It is, however, in the nature of things that field theories cannot give answers in terms of

simple closed expressions of any generality. The best that can be done is in the form of curves, with step-wise increases of particular parameters. These can be of considerable value in particular cases, and when exactness is essential.

In this paper we shall proceed by giving the "transmission-line" type theory first, together with the elaborations that are necessary to arrive at an estimate of the strength of coupling possible with coaxial helices. The "field" type theory will be used whenever the other theory fails, or is inadequate. Considerable physical insight can be gotten with the use of the transmission-line theory; nevertheless recourse to field theory is necessary in a number of cases, as will be seen.

It will be noted that in all the calculations to be presented the presence of an electron beam is left out of account. This is done for two reasons: Its inclusion would enormously complicate the theory, and, as will eventually be shown, it would modify our conclusions only very slightly. Moreover, in practically all cases which we shall consider, the helices are so tightly coupled that the velocities of the two normal modes of propagation are very different, as will be shown. Thus, only when the beam velocity is very near to either one or the other wave velocity, will growing-wave interaction take place between the beam and the helices. In this case conventional traveling wave tube theory may be used.

A theory of coupled helices in the presence of an electron beam has been presented by Wade and Rynn,⁴ who treated the case of weakly coupled helices and arrived at conclusions not at variance with our views.

2.2 Transmission Line Equations

Following Pierce we describe two lossless helices by their distributed series reactances X_{10} and X_{20} and their distributed shunt susceptances B_{10} and B_{20} . Thus their phase constants are

$$\beta_{10} = \sqrt{B_{10}X_{10}}$$

$$\beta_{20} = \sqrt{B_{20}X_{20}}$$

Let these helices be coupled by means of a mutual distributed reactance X_m and a mutual susceptance B_m , both of which are, in a way which will be described later, functions of the geometry.

Let waves in the coupled system be described by the factor

$$e^{j\omega t} e^{-\Gamma z}$$

where the Γ 's are the propagation constants to be found.

The transmission line equations may be written:

$$\begin{aligned}\Gamma I_1 - jB_1 V_1 + jB_m V_2 &= 0 \\ \Gamma V_1 - jX_1 I_1 + jX_m I_2 &= 0 \\ \Gamma I_2 - jB_2 V_2 + jB_m V_1 &= 0 \\ \Gamma V_2 - jX_2 I_a + jX_m I_1 &= 0\end{aligned}\tag{2.2.1}$$

where

$$B_1 = B_{10} + B_m$$

$$X_1 = X_{10} + X_m$$

$$B_2 = B_{20} + B_m$$

$$X_2 = X_{20} + X_m$$

I_1 and I_2 are eliminated from the (2.2.1) and we find

$$\frac{V_2}{V_1} = \frac{+(\Gamma^2 + X_1 B_1 + X_m B_m)}{X_1 B_m + B_2 X_m}\tag{2.2.2}$$

$$\frac{V_1}{V_2} = \frac{+(\Gamma^2 + X_2 B_2 + X_m B_m)}{X_2 B_m + B_1 X_m}\tag{2.2.3}$$

These two equations are then multiplied together and an expression for Γ of the 4th degree is obtained:

$$\begin{aligned}\Gamma^4 + (X_1 B_1 + X_2 B_2 + 2X_m B_m)\Gamma^2 \\ + (X_1 X_2 - X_m^2)(B_1 B_2 - B_m^2) = 0\end{aligned}\tag{2.2.4}$$

We now define a number of dimensionless quantities:

$$\frac{B_m^2}{B_1 B_2} = b^2 = (\text{capacitive coupling coefficient})^2$$

$$\frac{X_m^2}{X_1 X_2} = x^2 = (\text{inductive coupling coefficient})^2$$

$$B_1 X_1 = \beta_1^2, \quad B_2 X_2 = \beta_2^2$$

$$X_1 B_1 X_2 B_2 = \beta^4 = (\text{mean phase constant})^4$$

With these substitutions we obtain the general equation for Γ^2

$$\begin{aligned}\Gamma^2 = \beta^2 \left[-\frac{1}{2} \left(\frac{\beta^2}{\beta_2^2} + \frac{\beta^2}{\beta_1^2} + 2bx \right) \right. \\ \left. \pm \sqrt{\frac{1}{4} \left(\frac{\beta^2}{\beta_2^2} + \frac{\beta^2}{\beta_1^2} + 2bx \right)^2 - (1 - x^2)(1 - b^2)} \right]\end{aligned}\tag{2.2.5}$$

If we make the same substitutions in (2.2.2) we find

$$\frac{V_2}{V_1} = \sqrt{\frac{Z_2}{Z_1}} \left[\frac{\Gamma^2 + \beta_1^2 + \beta^2 bx}{\beta(\beta_1 b + \beta_2 x)} \right] \quad (2.2.6)$$

where the Z 's are the impedances of the helices, i.e.,

$$Z_n = \sqrt{X_n/B_n}$$

2.3 Solution for Synchronous Helices

Let us consider the particular case where $\beta_1 = \beta_2 = \beta$. From (2.2.5) we obtain

$$\Gamma^2 = -\beta^2[1 + xb \pm (x + b)] \quad (2.3.1)$$

Each of the above values of Γ^2 characterizes a normal mode of propagation involving both helices. The two square roots of each Γ^2 represent waves going in the positive and negative directions. We shall consider only the positive roots of Γ^2 , denoted Γ_t and Γ_ℓ , which represent the forward traveling waves.

$$\Gamma_{t,\ell} = j\beta\sqrt{1 + xb \pm (x + b)} \quad (2.3.2)$$

If $x > 0$ and $b > 0$

$$|\Gamma_t| > |\beta|, |\Gamma_\ell| < |\beta|$$

Thus Γ_t represents a normal mode of propagation which is slower than the propagation velocity of either helix alone and can be called the "slow" wave. Similarly Γ_ℓ represents a "fast" wave. We shall find that, in fact, x and b are numerically equal in most cases of interest to us; we therefore write the expressions for the propagation constants

$$\begin{aligned} \Gamma_t &= j\beta[1 + \frac{1}{2}(x + b)] \\ \Gamma_\ell &= j\beta[1 - \frac{1}{2}(x + b)] \end{aligned} \quad (2.3.3)$$

If we substitute (2.3.3) into (2.2.6) for the case where $\beta_1 = \beta_2 = \beta$ and assume, for simplicity, that the helix self-impedances are equal, we find that for $\Gamma = \Gamma_t$

$$\frac{V_2}{V_1} = -1$$

for $\Gamma = \Gamma_\ell$

$$\frac{V_2}{V_1} = +1$$

Thus, the slow wave is characterized by equal voltages of unlike sign on the two helices, and the fast wave by equal voltages of like sign. It follows that the electric field in the annular region between two such coupled concentric helices will be transverse for the slow wave and longitudinal for the fast. For this reason the slow and fast modes are often referred to as the transverse and longitudinal modes, respectively, as indicated by our subscripts.

It should be noted here that we arbitrarily chose b and x positive. A different choice of signs cannot alter the fact that the transverse mode is the slower and the longitudinal mode is the faster of the two.

Apart from the interest in the separate existence of the fast and slow waves as such, another object of interest is the phenomenon of the simultaneous existence of both waves and the interference, or spatial beating, between them.

Let V_2 denote the voltage on the outer helix; and let V_1 , the voltage on the inner helix, be zero at $z = 0$. Then we have, omitting the common factor $e^{j\omega t}$,

$$\begin{aligned} V_1 &= V_{t1}e^{-\Gamma_t z} + V_{\ell1}e^{-\Gamma_\ell z} \\ V_2 &= V_{t2}e^{-\Gamma_t z} + V_{\ell2}e^{-\Gamma_\ell z} \end{aligned} \quad (2.3.4)$$

Since at $z = 0$, $V_1 = 0$, $V_{t1} = -V_{\ell1}$. For the case we have considered we have found $V_{t1} = -V_{t2}$ and $V_{\ell1} = V_{\ell2}$. We can write (2.3.4) as

$$\begin{aligned} V_1 &= \frac{V}{2} (e^{-\Gamma_t z} - e^{-\Gamma_\ell z}) \\ V_2 &= \frac{V}{2} (e^{-\Gamma_t z} + e^{-\Gamma_\ell z}) \end{aligned} \quad (2.3.5)$$

V_2 can be written

$$\begin{aligned} V_2 &= \frac{V}{2} e^{-1/2(\Gamma_t + \Gamma_\ell)z} [e^{+1/2(\Gamma_t - \Gamma_\ell)z} + e^{-1/2(\Gamma_t - \Gamma_\ell)z}] \\ &= Ve^{-1/2(\Gamma_t + \Gamma_\ell)z} \cos [-j\frac{1}{2}(\Gamma_t - \Gamma_\ell)z] \end{aligned}$$

In the case when $x = b$, and $\beta_1 = \beta_2 = \beta$

$$V_2 = Ve^{-j\beta z} \cos [\frac{1}{2}(x + b)\beta z] \quad (2.3.6)$$

Correspondingly, it can be shown that the voltage on the inner helix is

$$V_1 = jVe^{-j\beta z} \sin [\frac{1}{2}(x + b)\beta z] \quad (2.3.7)$$

The last two equations exhibit clearly what we have called the spatial beat phenomenon, a wave-like transfer of power from one helix to the

other and back. We started, arbitrarily, with all the voltage on the outer helix at $z = 0$, and none on the inner; after a distance, z' , which makes the argument of the cosine $\pi/2$, there is no voltage on the outer helix and all is on the inner.

To conform with published material let us define what we shall call the "coupling phase-constant" as

$$\beta_c = \beta(b + x) \quad (2.3.8)$$

From (2.3.3) we find that for $\beta_1 = \beta_2 = \beta$, and $x = b$,

$$\Gamma_t - \Gamma_\ell = j\beta_c$$

2.4 Non-Synchronous Helix Solutions

Let us now go back to the more general case where the propagation velocities of the (uncoupled) helices are not equal. Equation (2.2.5) can be written:

$$\Gamma^2 = -\beta^2 [1 + (1/2)\Delta + xb \pm \sqrt{(1 + xb)\Delta + (1/4)\Delta^2 + (b + x)^2}] \quad (2.4.1)$$

where

$$\Delta \equiv \left[\frac{\beta_2 - \beta_1}{\beta} \right]^2$$

In the case where $x = b$, (2.4.1) has an exact root.

$$\Gamma_{t,\ell} = j\beta [\sqrt{1 + \Delta/4} \pm 1/2 \sqrt{\Delta + (x + b)^2}] \quad (2.4.2)$$

We shall be interested in the difference between Γ_t and Γ_ℓ ,

$$\Gamma_t - \Gamma_\ell = j\beta \sqrt{\Delta + (x + b)^2} \quad (2.4.3)$$

Now we substitute for Δ and find

$$\Gamma_t - \Gamma_\ell = j \sqrt{(\beta_1 - \beta_2)^2 + \beta^2 (b + x)^2} \quad (2.4.4)$$

Let us define the "beat phase-constant" as:

$$\beta_b = \sqrt{(\beta_1 - \beta_2)^2 + \beta^2 (b + x)^2}$$

so that

$$\Gamma_t - \Gamma_\ell = j\beta_b \quad (2.4.5)$$

Further, let us define

$$\beta_d = |\beta_1 - \beta_2|$$

and call this the "difference phase-constant," i.e., the hase constant corresponding to two uncoupled waves of the same frequency but differing phase velocities. We can thus state the relation between these phase constants:

$$\beta_b^2 = \beta_d^2 + \beta_c^2 \quad (2.4.6)$$

This relation is identical (except for notation) with expression (33) in S. E. Miller's paper.⁵ In this paper Miller also gives expressions for the voltage amplitudes in two coupled transmission systems in the case of unequal phase velocities. It turns out that in such a case the power transfer from one system to the other is necessarily incomplete. This is of particular interest to us, in connection with a number of practical schemes. In our notation it is relatively simple, and we can state it by saying that the maximum fraction of power transferred is

$$F = \left(\frac{\beta_c}{\beta_b} \right)^2 \quad (2.4.7)$$

or, in more detail,

$$F = \frac{\beta_c^2}{\beta_d^2 + \beta_c^2} = \frac{\beta^2(b + x)^2}{(\beta_1 - \beta_2)^2 + \beta^2(b + x)^2}$$

This relationship can be shown to be a good approximation from (2.2.6), (2.3.4), (2.4.2), on the assumption that $b \approx x$ and $Z_1 \approx Z_2$, and the further assumption that the system is lossless; that is,

$$|V_2|^2 + |V_1|^2 = \text{constant} \quad (2.4.8)$$

We note that the phase velocity difference gives rise to two phenomena: It reduces the coupling wavelength and it reduces the amount of power that can be transferred from one helix to the other.

Something should be said about the case where the two helix impedances are not equal, since this, indeed, is usually the case with coupled concentric helices. Equation (2.4.8) becomes:

$$\frac{|V_2|^2}{Z_2} + \frac{|V_1|^2}{Z_1} = \text{constant} \quad (2.4.9)$$

Using this relation it is found from (2.3.4) that

$$\frac{V_2}{V_1} \sqrt{\frac{Z_1}{Z_2}} = \pm \sqrt{\frac{1}{F}} (1 \pm \sqrt{1 - F}) \quad (2.4.10)$$

When this is combined with (2.2.6) it is found that the impedances drop out with the voltages, and that "F" is a function of the β 's only. In other

words, complete power transfer occurs when $\beta_1 = \beta_2$ regardless of the relative impedances of the helices.

The reader will remember that β_{10} and β_{20} , not β_1 and β_2 , were defined as the phase constants of the helices in the absence of each other. If the assumption that $b \approx x$ is maintained, it will be found that all of the derived relationships hold true when β_{n0} is substituted for β_n . In other words, throughout the paper, β_1 and β_2 may be treated as the phase constants of the inner and outer helices, respectively. In particular it should be noted that if these quantities are to be measured experimentally each helix must be kept in the same environment as if the helices were coupled; only the other helix may be removed. That is, if there is dielectric in the annular region between the coupled helices, β_1 and β_2 must each be measured in the presence of that dielectric.

Miller also has treated the case of lossy coupled transmission systems. The expressions are lengthy and complicated and we believe that no substantial error is made in simply applying his conclusions to our case.

If the attenuation constants α_1 and α_2 of the two transmission systems (helices) are equal, no change is required in our expressions; when they are unequal the total available power (in both helices) is most effectively reduced when

$$\frac{\alpha_1 - \alpha_2}{\beta_c} \approx 1 \quad (2.4.11)$$

This fact may be made use of in designing coupled helix attenuators.

2.5 A Look at the Fields

It may be advantageous to consider sketches of typical field distributions in coupled helices, as in Fig. 2.1, before we go on to derive a quantitative estimate of the coupling factors actually obtainable in practice.

Fig. 2.1(a) shows, diagrammatically, electric field lines when the coupled helices are excited in the fast or "longitudinal" mode. To set up this mode only, one has to supply voltages of like sign and equal amplitudes to both helices. For this reason, this mode is also sometimes called the "(++) mode."

Fig. 2.1(b) shows the electric field lines when the helices are excited in the slow or "transverse" mode. This is the kind of field required in the transverse interaction type of traveling wave tube. In order to excite this mode it is necessary to supply voltages of equal amplitude and opposite signs to the helices and for this reason it is sometimes called the "(+-) mode." One way of exciting this mode consists in connecting one

helix to one of the two conductors of a balanced transmission line ("Lecher"-line) and the other helix to the other.

Fig. 2.1(c) shows the electric field configuration when fast and slow modes are both present and equally strongly excited. We can imagine the two helices being excited by a voltage source connected to the outer

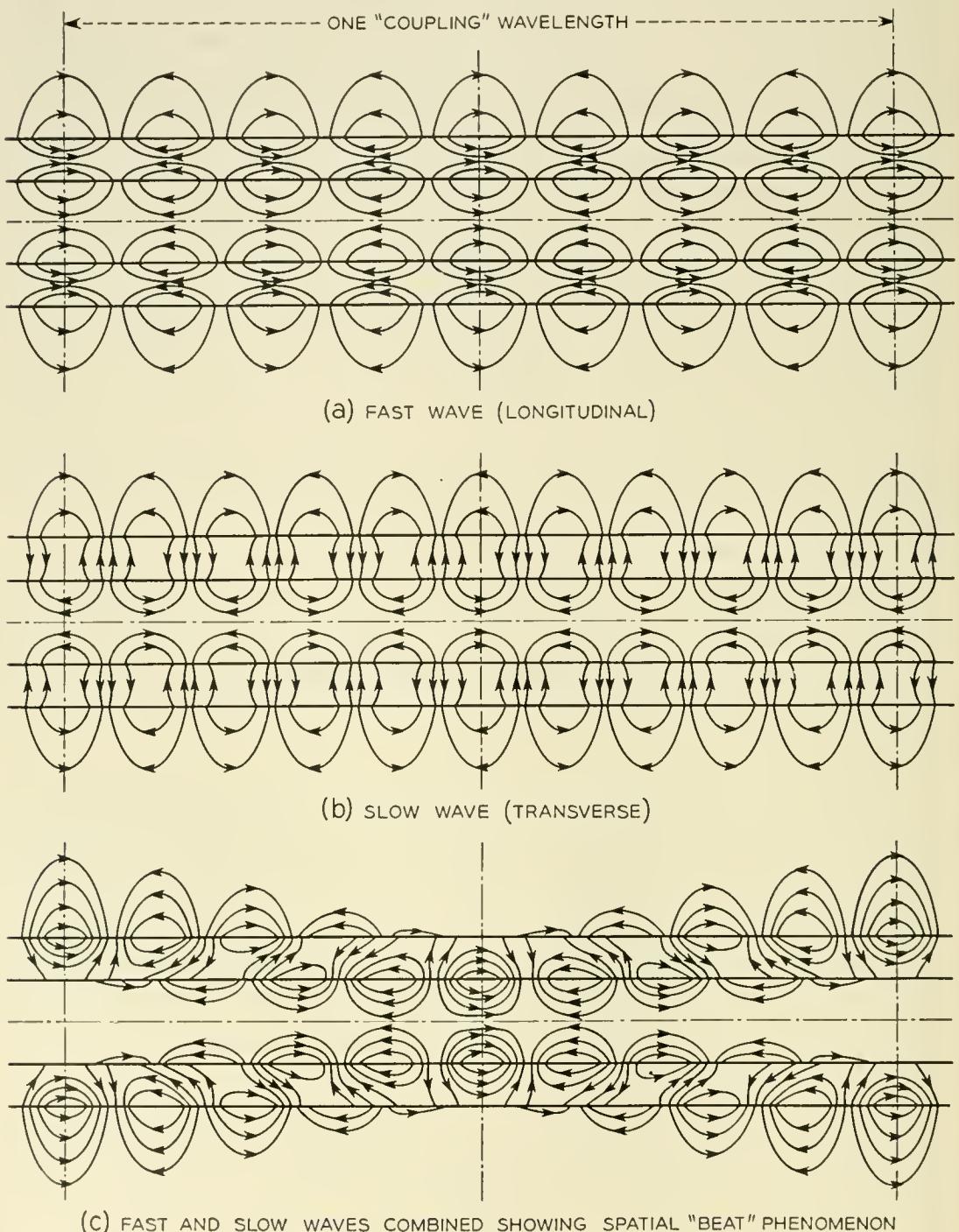


Fig. 2.1 — Typical electric field distributions in coupled coaxial helices when they are excited in: (a) the in-phase or longitudinal mode, (b) the out-of-phase or transverse mode, and (c) both modes equally.

helix only at the far left side of the sketch. One, perfectly legitimate, view of the situation is that the RF power, initially all on the outer helix, leaks into the inner helix because of the coupling between them, and then leaks back to the outer helix, and so forth.

Apart from noting the appearance of the stationary spatial beat (or interference) phenomenon these additional facts are of interest:

1) It is a simple matter to excite such a beat-wave, for instance, by connecting a lead to either one or the other of the helices, and

2) It should be possible to discontinue either one of the helices, at points where there is no current (voltage) on it, without causing reflections.

2.6 A Simple Estimate of b and x

How strong a coupling can one expect from concentric helices in practice? Quantitatively, this is expressed by the values of the coupling factors x and b , which we shall now proceed to estimate.

A first crude estimate is based on the fact that slow-wave fields are known to fall off in intensity somewhat as $e^{-\beta r}$ where β is the phase constant of the wave and r the distance from the surface guiding the slow wave. Thus a unit charge placed, say, on the inner helix, will induce a charge of opposite sign and of magnitude

$$e^{-\beta(b-a)}$$

on the outer helix. Here b = mean radius of the outer helix and a = mean radius of the inner. We note that the shunt mutual admittance coupling factor is negative, irrespective of the directions in which the helices are wound. Because of the similarity of the magnetic and electric field distributions a current flowing on the inner helix will induce a similarly attenuated current, of amplitude

$$e^{-\beta(b-a)}$$

on the outer helix. The direction of the induced current will depend on whether the helices are wound in the same sense or not, and it turns out (as one can verify by reference to the low-frequency case of coaxial coupled coils) that the series mutual impedance coupling factor is negative when the helices are oppositely wound.

In order to obtain the greatest possible coupling between concentric helices, both coupling factors should have the same sign. This then requires that the helices should be wound in opposite directions, as has been pointed out by Pierce.

When the distance between the two helices goes to zero, that is to say,

if they lie in the same surface, it is clear that both coupling factors b and x will go to unity.

As pointed out earlier in Section 2.3, the choice of sign for b is arbitrary. However, once a sign for b has been chosen, the sign of x is necessarily the opposite when the helices are wound in the same direction, and vice versa. We shall choose, therefore,

$$\begin{aligned} b &= +e^{-\beta(b-a)} \\ x &= \mp e^{-\beta(b-a)} \end{aligned} \quad (2.6.1)$$

the sign of the latter depending on whether the helices are wound in the same direction or not.

In the case of unequal velocities, β , the propagation constant, would be given by

$$\beta = \sqrt{\beta_1 \beta_2} \quad (2.6.2)$$

2.7 Strength of Coupling versus Frequency

The exponential variation of coupling factors with respect to frequency (since $\beta = \omega/v$) has an important consequence. Consider the expression for the coupling phase constant

$$\beta_c = \beta(b + x) \quad (2.3.8)$$

or

$$|\beta_c| = 2\beta e^{-\beta(b-a)} \quad (2.7.1)$$

The coupling wavelength, which is defined as

$$\lambda_c = \left| \frac{2\pi}{\beta_c} \right| \quad (2.7.2)$$

is, therefore,

$$\lambda_c = \frac{\pi}{\beta} e^{\beta(b-a)}$$

or

$$\lambda_c = \frac{\lambda}{2} e^{(2\pi/\lambda)(b-a)} \quad (2.7.3)$$

where λ is the (slowed-down) RF wavelength on either helix. It is convenient to multiply both sides of (2.7.1) with a , the inner helix radius, in order to obtain a dimensionless relation between β_c and β :

$$\beta_c a = 2\beta a e^{-\beta a((b/a)-1)} \quad (2.7.4)$$

This relation is plotted on Fig. 2.2 for several values of b/a .

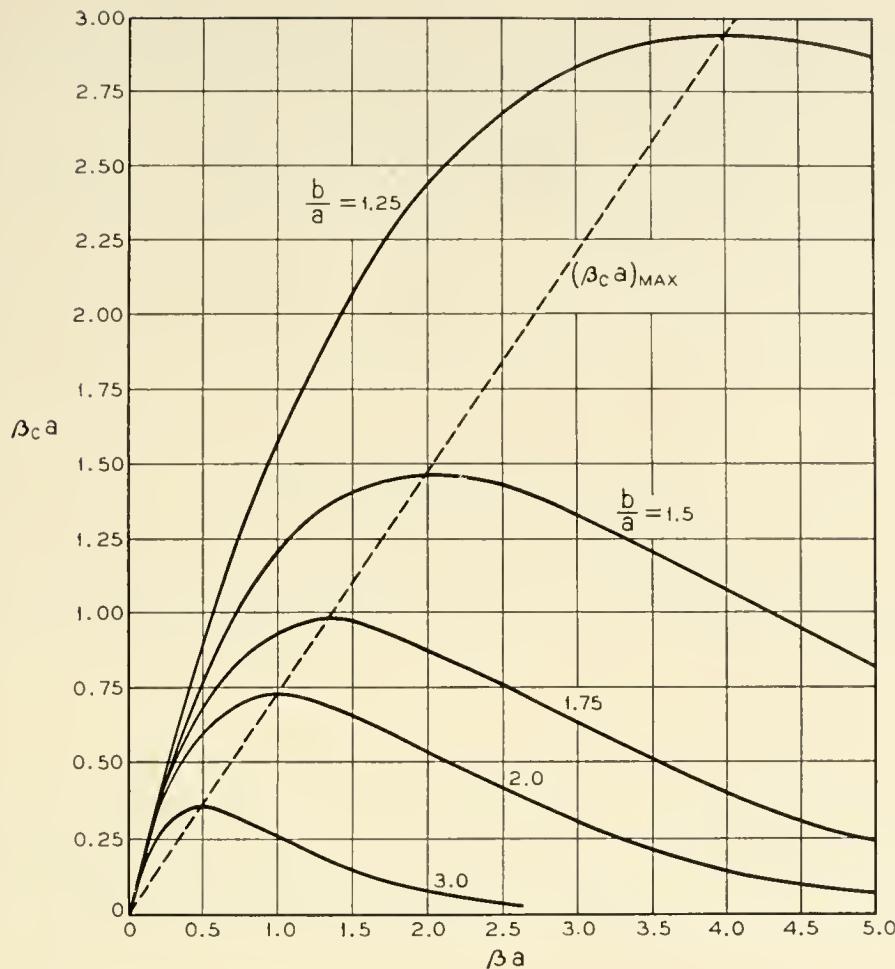


Fig. 2.2 — Coupling phase-constant plotted as a function of the single helix phase-constant for synchronous helices for several values of b/a . These curves are based on simple estimates made in Section 2.7.

There are two opposing tendencies determining the actual physical length of a coupling beat-wavelength:

- 1) It tends to grow with the RF wavelength, being proportional to it in the first instance;
- 2) Because of the tighter coupling possible as the RF wavelength increases in relation to the helix-to-helix distance, the coupling beat-wavelength tends to shrink.

Therefore, there is a region where these tendencies cancel each other, and where one would expect to find little change of the coupling beat-wavelength for a considerable change of RF frequency. In other words, the "bandwidth" over which the beat-wavelength stays nearly constant can be large.

This is a situation naturally very desirable and favorable for any device in which we rely on power transfer from one helix to the other by

means of a length of overlap between them an integral number of half beat-wavelengths long. Obviously, one will design the helices in such a way as to take advantage of this situation.

Optimum conditions are easily obtained by differentiating β_c with respect to β and setting $\partial\beta_c/\partial\beta$ equal to zero. This gives for the optimum conditions

$$\beta_{\text{opt}} = \frac{1}{b - a} \quad (2.7.5)$$

or

$$\beta_{c\text{ opt}} = \frac{2e^{-1}}{b - a} = 2e^{-1}\beta_{\text{opt}} \quad (2.7.6)$$

Equation (2.7.5), then, determines the ratio of the helix radii if it is required that deviations from a chosen operating frequency shall have least effect.

2.8 Field Solutions

In treating the problem of coaxial coupled helices from the transmission line point of view one important fact has not been considered, namely, the dispersive character of the phase constants of the separate helices, β_1 and β_2 . By dispersion we mean change of phase velocity with frequency. If the dispersion of the inner and outer helices were the same it would be of little consequence. It is well known, however, that the dispersion of a helical transmission line is a function of the ratio of helix radius to wavelength, and thus becomes a parameter to be considered. When the theory of wave propagation on a helix was solved by means of Maxwell's equations subject to the boundary condition of a helically conducting cylindrical sheath, the phenomenon of dispersion first made its appearance. It is clear, therefore, that a more complete theory of

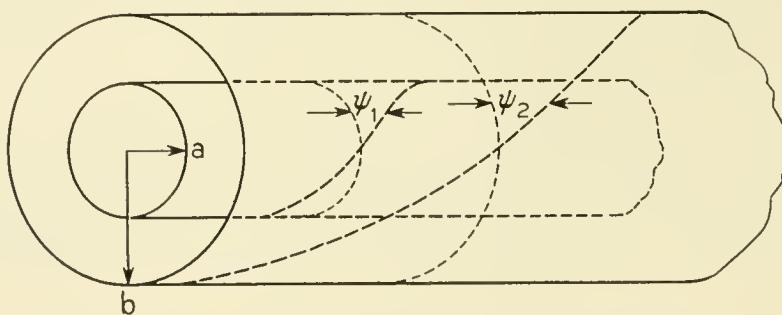


Fig. 2.3 — Sheath helix arrangement on which the field equations are based.

coupled helices will require similar treatment, namely, Maxwell's equations solved now with the boundary conditions of two cylindrical helically conducting sheaths. As shown on Fig. 2.3, the inner helix is specified by its radius a and the angle ψ_1 made by the direction of conductivity with a plane perpendicular to the axis; and the outer helix by its radius b (not to be confused with the mutual coupling coefficient b) and its corresponding pitch angle ψ_2 . We note here that oppositely wound helices require opposite signs for the angles ψ_1 and ψ_2 ; and, further, that helices with equal phase velocities will have pitch angles of about the same absolute magnitude.

The method of solving Maxwell's equations subject to the above mentioned boundary conditions is given in Appendix I. We restrict ourselves here to giving some of the results in graphical form.

The most universally used parameter in traveling-wave tube design is a combination of parameters:

$$\beta_0 a \cot \psi_1$$

where $\beta_0 = 2\pi/\lambda_0$, λ_0 being the free-space wavelength, a the radius of the inner helix, and ψ_1 the pitch angle of the inner helix. The inner helix is chosen here in preference to the outer helix because, in practice, it will be part of a traveling-wave tube, that is to say, inside the tube envelope. Thus, it is not only less accessible and changeable, but determines the important aspects of a traveling-wave tube, such as gain, power output, and efficiency.

The theory gives solutions in terms of radial propagation constants which we shall denote γ_t and γ_ℓ (by analogy with the transverse and longitudinal modes of the transmission line theory). These propagation constants are related to the axial propagation constants β_t and β_ℓ by

$$\gamma_n = \sqrt{\beta_n^2 - \beta_0^2}$$

Of course, in transmission line theory there is no such thing as a radial propagation constant. The propagation constant derived there and denoted Γ corresponds here to the axial propagation constant $j\beta$. By analogy with (2.4.5) the beat phase constant should be written

$$\beta_b = \beta_t - \beta_\ell$$

However, in practice β_0 is usually much smaller than β and we can therefore write with little error

$$\beta_b = \gamma_t - \gamma_\ell$$

for the beat phase constant. For practical purposes it is convenient to

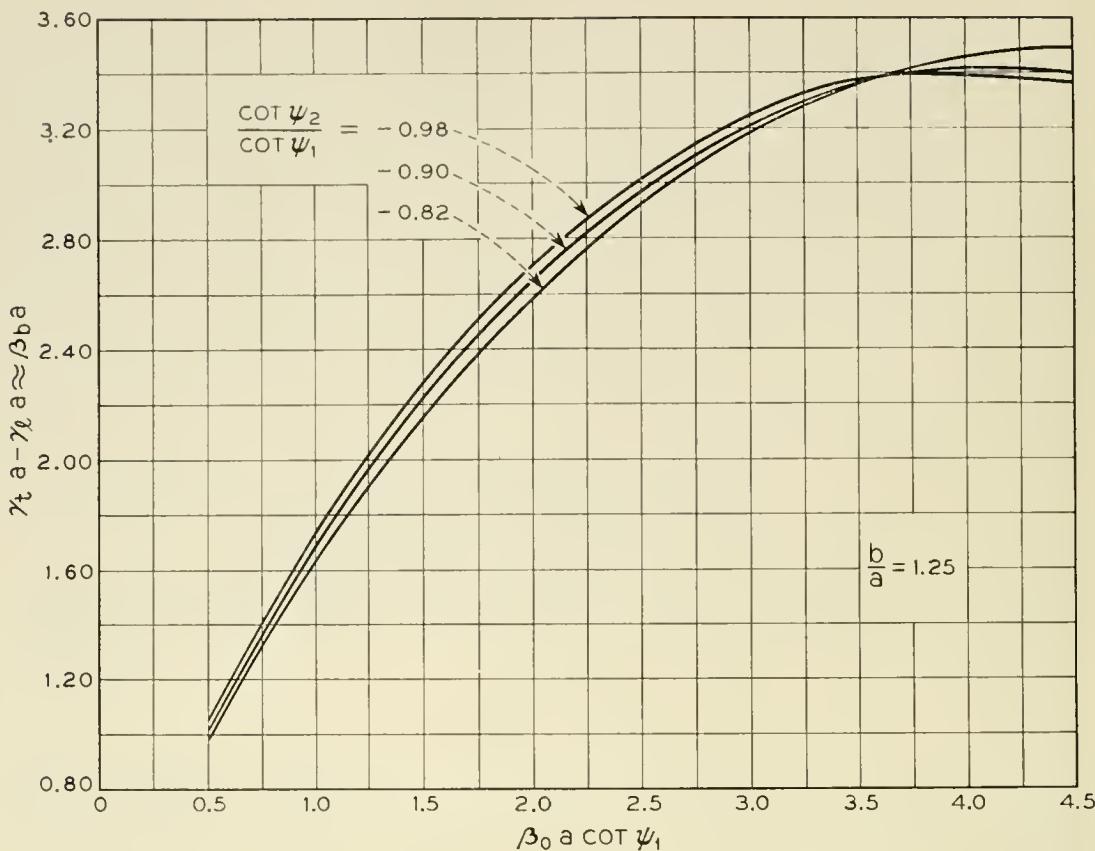


Fig. 2.4.1 — Beat phase-constant plotted as a function of $\beta_0 a \cot \psi_1$. These curves result from the solution of the field equations given in the appendix. For $b/a = 1.25$.

normalize in terms of the inner helix radius, a :

$$\beta_b a = \gamma_t a - \gamma_b a$$

This has been plotted as a function of $\beta_0 a \cot \psi_1$ in Fig. 2.4, which should be compared with Fig. 2.2. It will be seen that there is considerable agreement between the results of the two methods.

2.9 Bifilar Helix

The failure of the transmission line theory to take into account dispersion is well illustrated in the case of the bifilar helix. Here we have two identical helices wound in the same sense, and at the same radius. If the two wires are fed in phase we have the normal mode characterized by the sheath helix model whose propagation constant is the familiar Curve A of Fig. 2.5. If the two wires of the helix are fed out of phase we have the bifilar mode; and, since that is a two wire transmission system, we shall have a TEM mode which, in the absence of dielectric, propagates along the wire with the velocity of light. Hence, the propagation constant for this mode is simply $\beta_0 a \cot \psi$ and gives rise to the horizontal

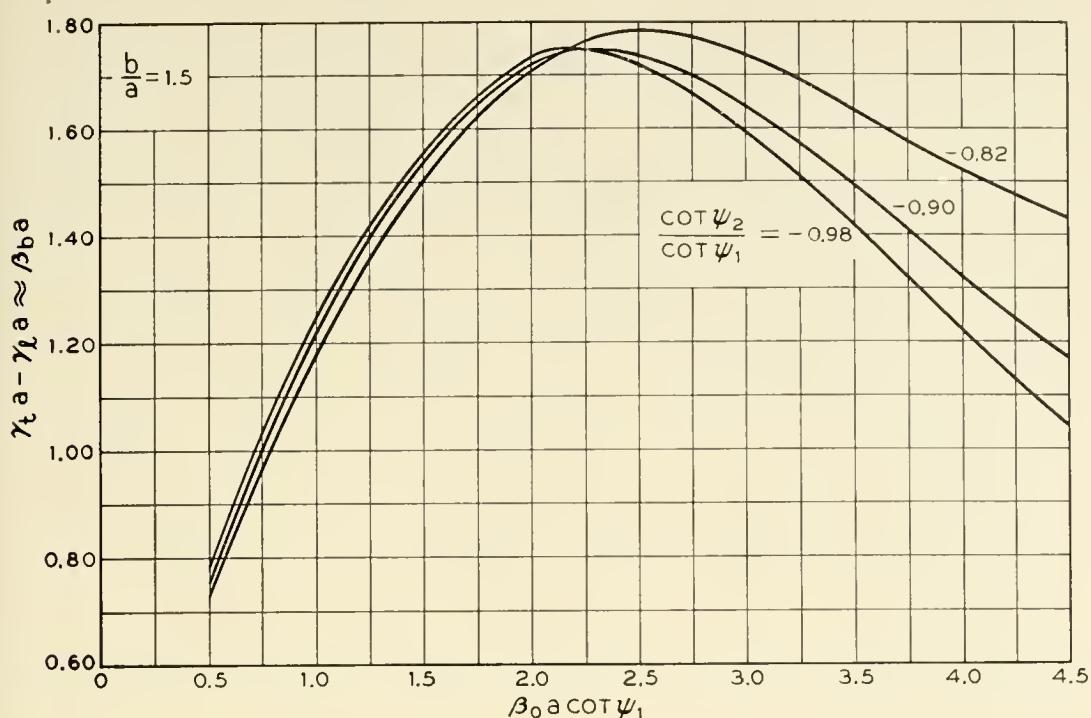


Fig. 2.4.2 — Beat phase-constant plotted as a function of $\beta_0 a \cot \psi_1$. These curves result from the solution of the field equations given in the appendix. For $b/a = 1.5$.

line of Curve *B* in Fig. 2.5. Again the coupling phase constant β_c is given by the difference of the individual phase constants:

$$\beta_c a = \beta_0 a \cot \psi - \gamma a \quad (2.9.1)$$

which is plotted in Fig. 2.6. Now note that when $\beta_0 \ll \gamma$ this equation is accurate, for it represents a solution of the field equations for the helix.

From the simple unsophisticated transmission line point of view no coupling between the two helices would, of course, have been expected, since the two helices are identical in every way and their mutual capacity and inductance should then be equal and opposite.

Experiments confirm the essential correctness of (2.9.1). In one experiment, which was performed to measure the coupling wavelength for the bifilar helices, we used helices with a $\cot \psi = 3.49$ and a radius of 0.036 cm which gave a value, at 3,000 me, of $\beta_0 a \cot \psi = 0.51$. In these experiments the coupling length, L , defined by

$$(\beta_0 a \cot \psi - \gamma a) \frac{L}{a} = \pi$$

was measured to be $15.7a$ as compared to a value of $13.5a$ from Fig. 2.6. At 4,000 me the measured coupling length was $14.6a$ as compared to

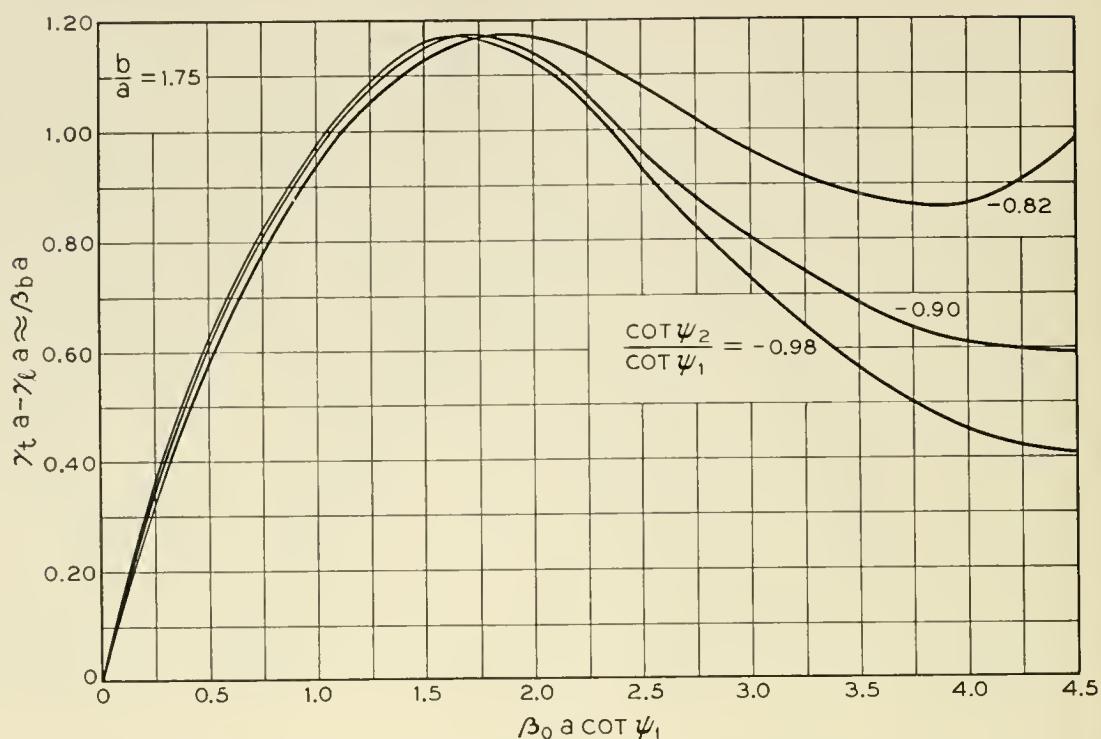


Fig. 2.4.3 — Beat phase-constant plotted as a function of $\beta_0 a \cot \psi_1$. These curves result from the solution of the field equations given in the appendix. For $b/a = 1.75$.

12.6a computed from Fig. 2.6, thus confirming the theoretical prediction rather well. The slight increase in coupling length is attributable to the dielectric loading of the helices which were supported in quartz tubing. The dielectric tends to decrease the dispersion and hence reduce β_e . This is discussed further in the next section.

2.10 Effect of Dielectric Material between Helices

In many cases which are of interest in practice there is dielectric material between the helices. In particular when coupled helices are used with traveling-wave tubes, the tube envelope, which may be of glass, quartz, or ceramic, all but fills the space between the two helices.

It is therefore of interest to know whether such dielectric makes any difference to the estimates at which we arrived earlier. We should not be surprised to find the coupling strengthened by the presence of the dielectric, because it is known that dielectrics tend to rob RF fields from the surrounding space, leading to an increase in the energy flow through the dielectric. On the other hand, the dielectric tends to bind the fields closer to the conducting medium. To find a qualitative answer to this question we have calculated the relative coupling phase constants for two sheath helices of infinite radius separated by a distance "d" for 1)

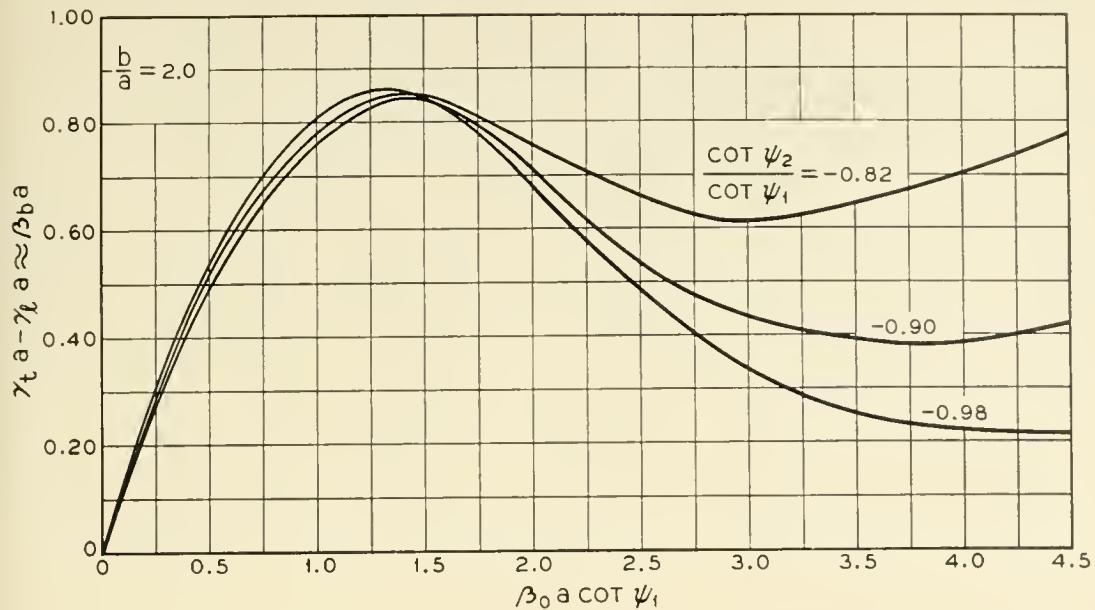


Fig. 2.4.4 — Beat phase-constant plotted as a function of $\beta_0 a \cot \psi_1$. These curves result from the solution of the field equations given in the appendix. For $b/a = 2.0$.

the case with dielectric between them having a relative dielectric constant $\epsilon' = 4$, and 2) the case of no dielectric. The pitch angles of the two helices were ψ and $-\psi$, respectively; i.e., the helices were assumed to be synchronous, and wound in the opposite sense.

Fig. 2.7 shows a plot of the ratio of $\beta_{c\epsilon}/\beta_\epsilon$ to β_c/β versus $\beta_0 (d/2) \cot \psi$,

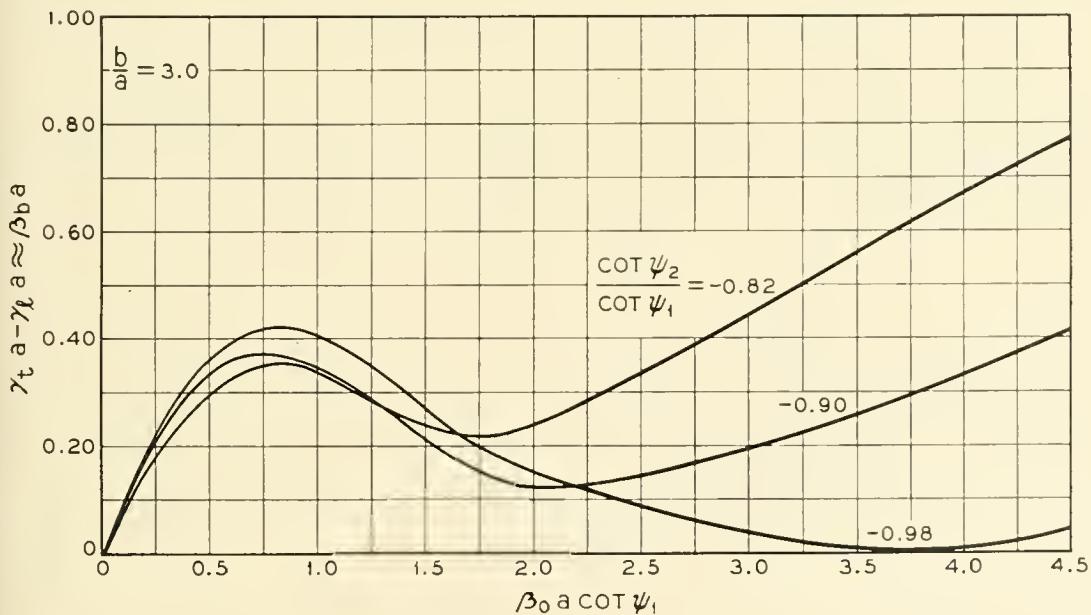


Fig. 2.4.5 — Beat phase-constant plotted as a function of $\beta_0 a \cot \psi_1$. These curves result from the solution of the field equations given in the appendix. For $b/a = 3.0$.

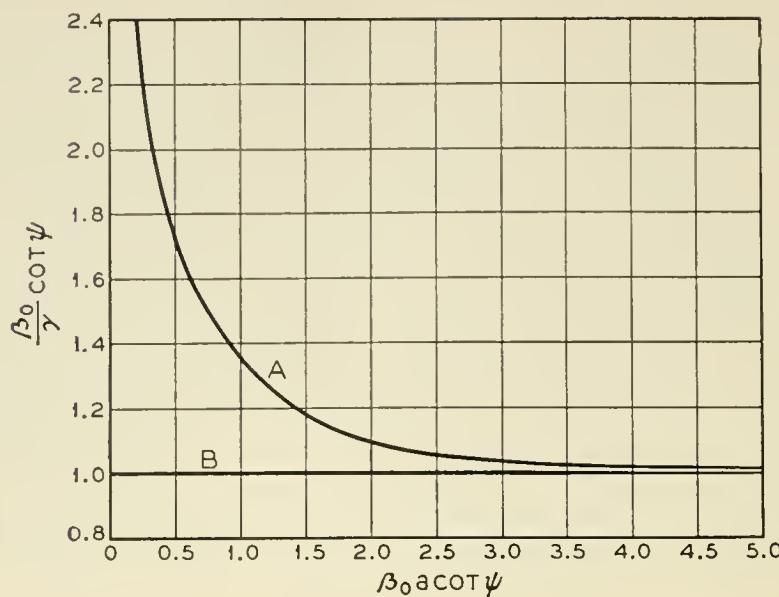


Fig. 2.5 — Propagation constants for a bifilar helix plotted as a function of $\beta_0 a \cot \psi$. The curves illustrate, (A) the dispersive character of the in-phase mode and, (B) the non-dispersive character of the out-of-phase mode.

where $\beta_{c\epsilon}$ is the coupling phase-constant in the presence of dielectric, β_ϵ is the phase-constant of each helix alone in the presence of the same dielectric, β_c is the coupling phase-constant with no dielectric, and β is the phase constant of each helix in free space. In many cases of interest $\beta_0(d/2) \cot \psi$ is greater than 1.2. Then

$$\frac{\beta_{c\epsilon}/\beta_\epsilon}{\beta_c/\beta} = \left[\frac{3\epsilon' + 1}{2\epsilon' + 2} \right] e^{-(\sqrt{2\epsilon'+2}-2)\beta_0(d/2) \cot \psi} \quad (2.10.1)$$

Appearing in the same figure is a similar plot for the case when there is a conducting shield inside the inner helix and outside the outer, and separated a distance, "s," from the helices. Note that

$$d \equiv b - a.$$

It appears from these calculations that the effect of the presence of dielectric between the helices depends largely on the parameter $\beta_0(d/2) \cot \psi$. For values of this parameter larger than 0.3 the coupling wavelength tends to increase in terms of circuit wavelength. For values smaller than 0.3 the opposite tends to happen. Note that the curve representing (2.10.1) is a fair approximation down to $\beta_0(d/2) \cot \psi = 0.6$ to the curve representing the exact solution of the field equations. J. W. Sullivan, in unpublished work, has drawn similar conclusions.

2.11 The Conditions for Maximum Power Transfer

The transmission line theory has led us to expect that the most efficient power transfer will take place if the phase velocities on the two helices, prior to coupling, are the same. Again, this would be true were it not for the dispersion of the helices. To evaluate this effect we have used the field equation to determine the parameter of the coupled helices which gives maximum power transfer. To do this we searched for combinations of parameters which give an equal current flow in the helix sheath for either the longitudinal mode or the transverse mode. This was suggested by L. Stark, who reasoned that if the currents were equal for the individual modes the beat phenomenon would give points of zero RF current on the helix.

The values of $\cot \psi_2 / \cot \psi_1$ which are required to produce this condition are plotted in Fig. 2.8 for various values of b/a . Also there are shown values of $\cot \psi_2 / \cot \psi_1$ required to give equal axial velocities for the helices before they are coupled. It can be seen that the uncoupled velocity of the inner helix must be slightly slower than that of the outer.

A word of caution is necessary for these curves have been plotted without considering the effects of dielectric loading, and this can have a rather marked effect on the parameters which we have been discussing. The significant point brought out by this calculation is that the optimum

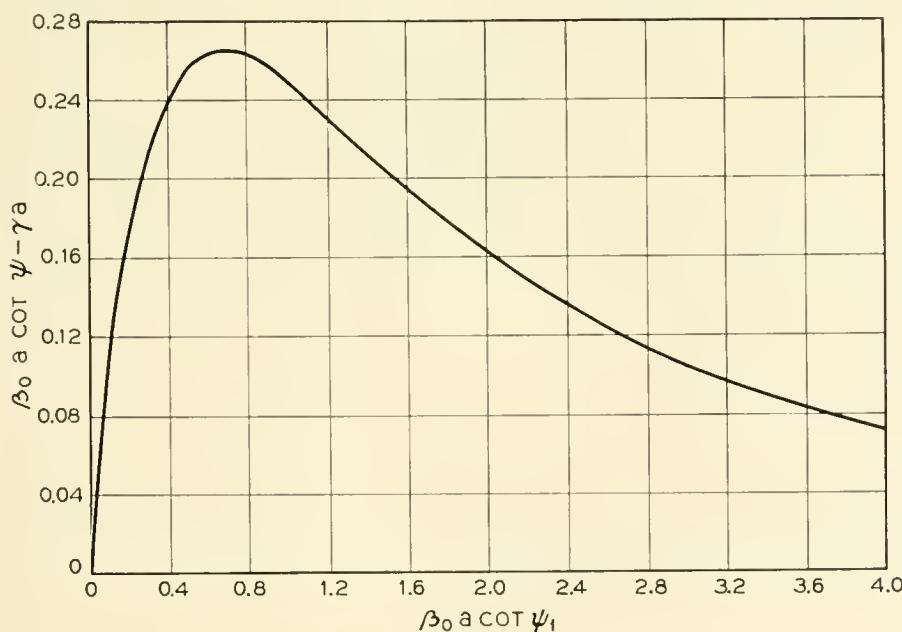


Fig. 2.6 — The coupling phase-constant which results from the two possible modes of propagation on a bifilar helix shown as a function of $\beta_0 a \cot \psi_1$.

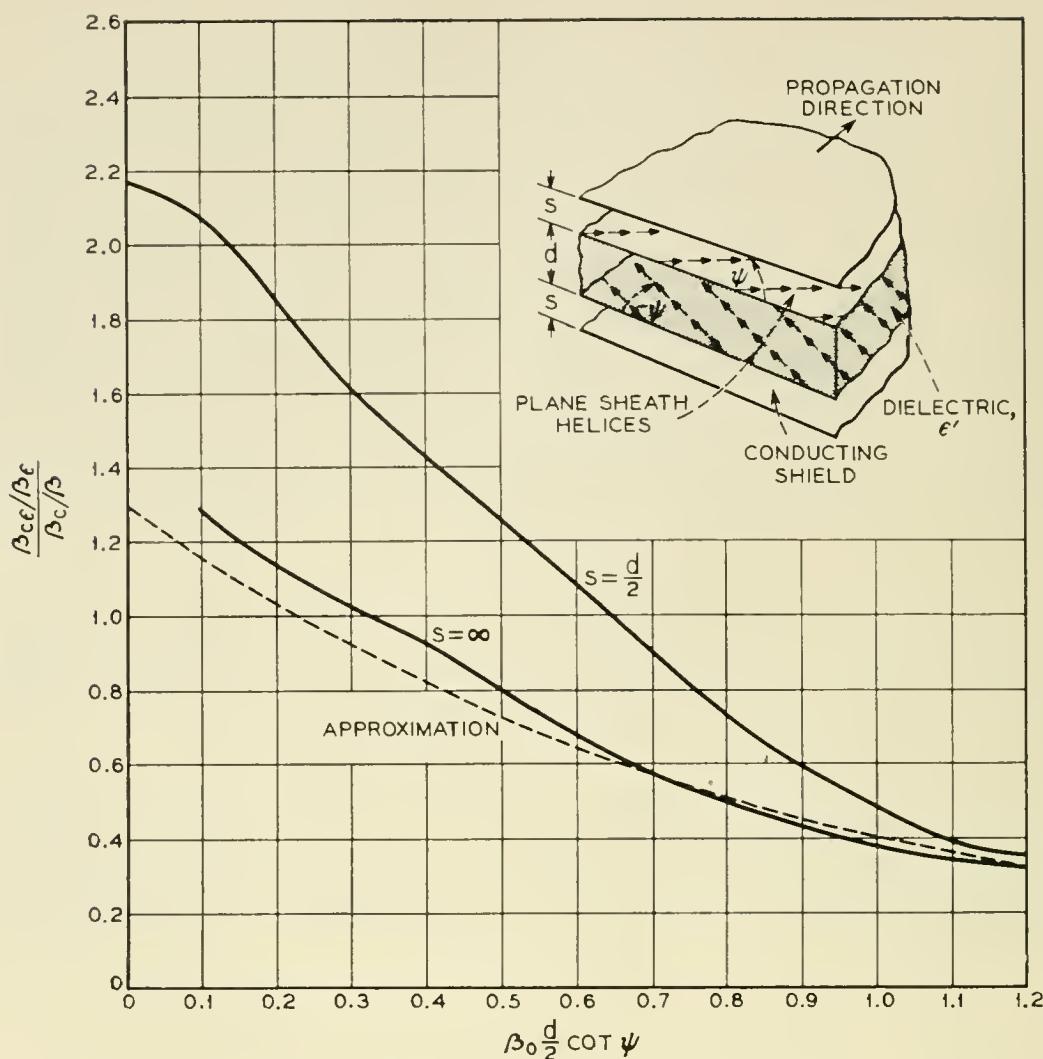


FIG. 2.7 — The effect of dielectric material between coupled infinite radius sheath helices on their relative coupling phase-constant shown as a function of $\beta_0 d/2 \cot \psi_1$. The effect of shielding on this relation is also indicated.

condition for coupling is not necessarily associated with equal velocities on the uncoupled helices.

2.12 Mode Impedance

Before leaving the general theory of coupled helices something should be said regarding the impedance their modes present to an electron beam traveling either along their axis or through the annular space between them. The field solutions for cross wound, coaxially coupled helices, which are given in Appendix I, have been used to compute the impedances of the transverse and longitudinal modes. The impedance, K , is defined, as usual, in terms of the longitudinal field on the axis and the power flow along the system.

$$K = \frac{E_z^2}{\beta^2 P} = \left(\frac{\beta}{\beta_0}\right)^{1/3} \left(\frac{\gamma}{\beta}\right)^{4/3} F(\gamma a)$$

In Fig. 2.9, $F(\gamma a)$, for various ratios of inner to outer radius, is plotted for both the transverse and longitudinal modes together with the value of $F(\gamma a)$ for the single helix ($b/a = \infty$). We see that the longitudinal mode has a higher impedance with cross wound coupled helices than does a single helix. We call attention here to the fact that this is the same phenomenon which is encountered in the contrawound helix⁶, where the structure consists of two oppositely wound helices of the same radius.

As defined here, the transverse mode has a lower impedance than the single helix. This, however, is not the most significant comparison; for it is the transverse field midway between helices which is of interest in the transverse mode. The factor relating the impedance in terms of the transverse field between helices to the longitudinal field on the axis is $E_r^2(\bar{r})/E_z^2(0)$, where \bar{r} is the radius at which the longitudinal component of the electric field E_z , is zero for the transverse mode. This factor, plotted in Fig. 2.10 as a function of $\beta_0 a \cot \psi_r$, shows that the impedance in terms of the transverse field at \bar{r} is interestingly high.

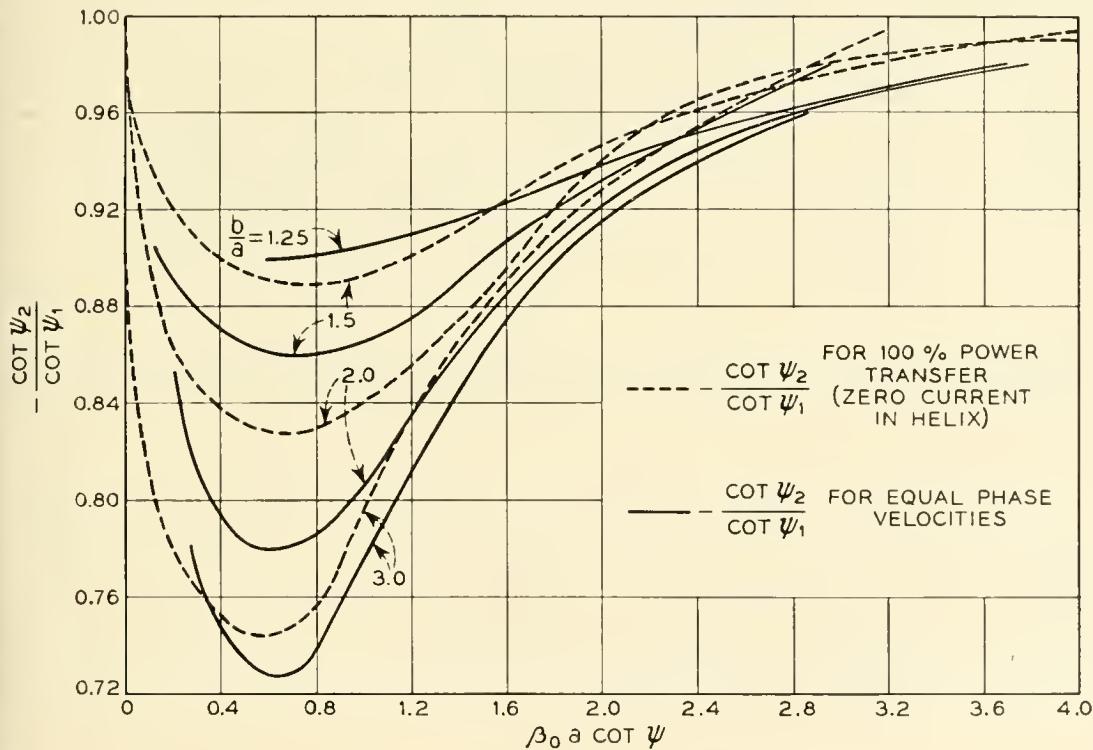


Fig. 2.8 — The values of $\cot \psi_2/\cot \psi_1$ required for complete power transfer plotted as a function of $\beta_0 a \cot \psi_1$ for several values of b/a . For comparison, the value of $\cot \psi_2/\cot \psi_1$ required for equal velocities on inner and outer helices is also shown.

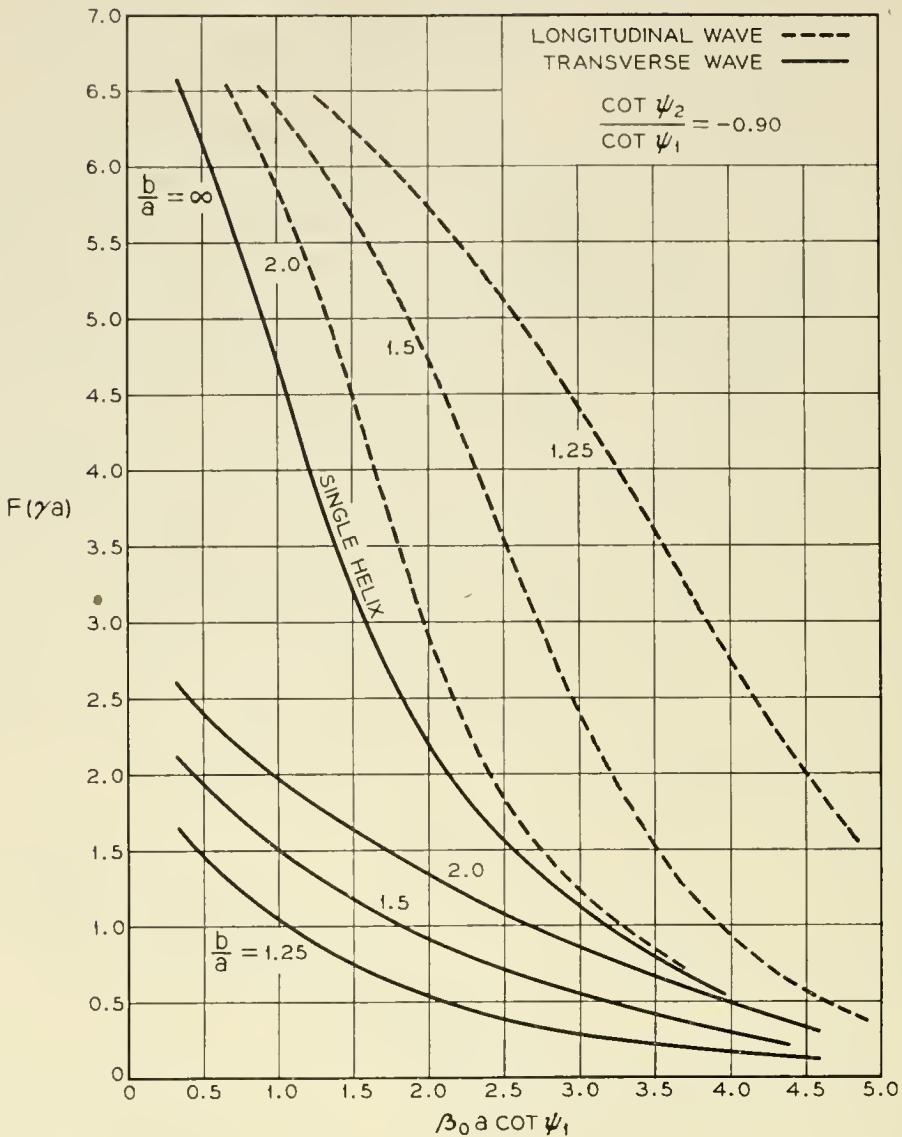


Fig. 2.9 — Impedance parameter, $F(\gamma a)$, associated with both transverse and longitudinal modes shown for several values of b/a . Also shown is $F(\gamma a)$ for a single helix.

It is also of interest to consider the impedance of the longitudinal mode in terms of the longitudinal field between the two helices. The factor, $E_z^2(\bar{r})/E_z^2(0)$, relating this to the axial impedance is plotted in Fig. 2.11. We see that rather high impedances can also be obtained with the longitudinal field midway between helices. This, in conjunction with a hollow electron beam, should provide efficient amplification.⁷

3. APPLICATION OF COUPLED HELICES

When we come to describe devices which make use of coupled helices we find that they fall, quite naturally, into two separate classes. One

class contains those devices which depend on the presence of only one of the two normal modes of propagation. The other class of devices depends on the simultaneous presence, in roughly equal amounts, of *both* normal modes of propagation, and is, in general, characterized by the words "spatial beating." Since spatial beating implies energy surging to and fro between inner and outer helix, there is no special problem in exciting both modes simultaneously. Power fed exclusively to one or the other

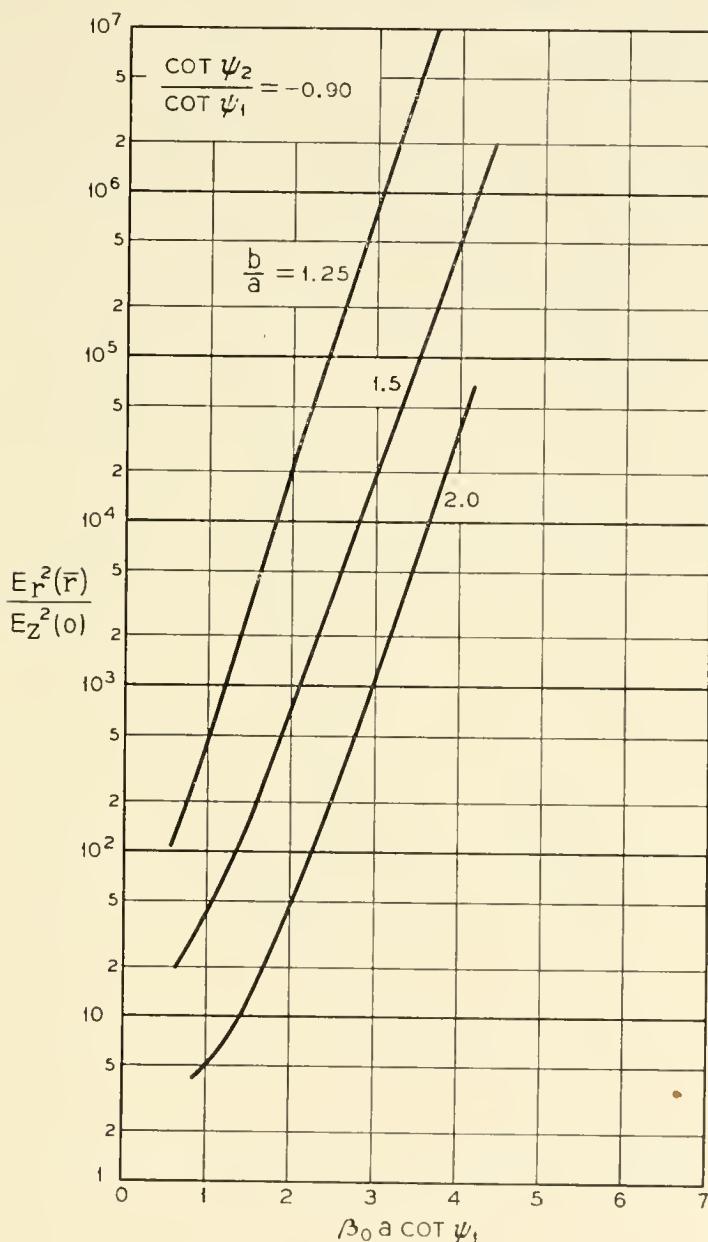


Fig. 2.10 — The relation between the impedance in terms of the transverse field between coupled helices excited in the out-of-phase mode, and the impedance in terms of the longitudinal field on the axis shown as a function of $\beta_0 a \cot \psi_1$.

helix will inevitably excite both modes equally. When it is desired to excite one mode exclusively a more difficult problem has to be solved. Therefore, in section 3.1 we shall first discuss methods of exciting one mode only before going on to discuss in sections 3.2 and 3.3 devices using one mode only.

In section 3.4 we shall discuss devices depending on the simultaneous presence of both modes.

3.1 Excitation of Pure Modes

3.1.1 Direct Excitation

In order to set up one or the other normal mode on coupled helices, voltages with specific phase and amplitudes (or corresponding currents)

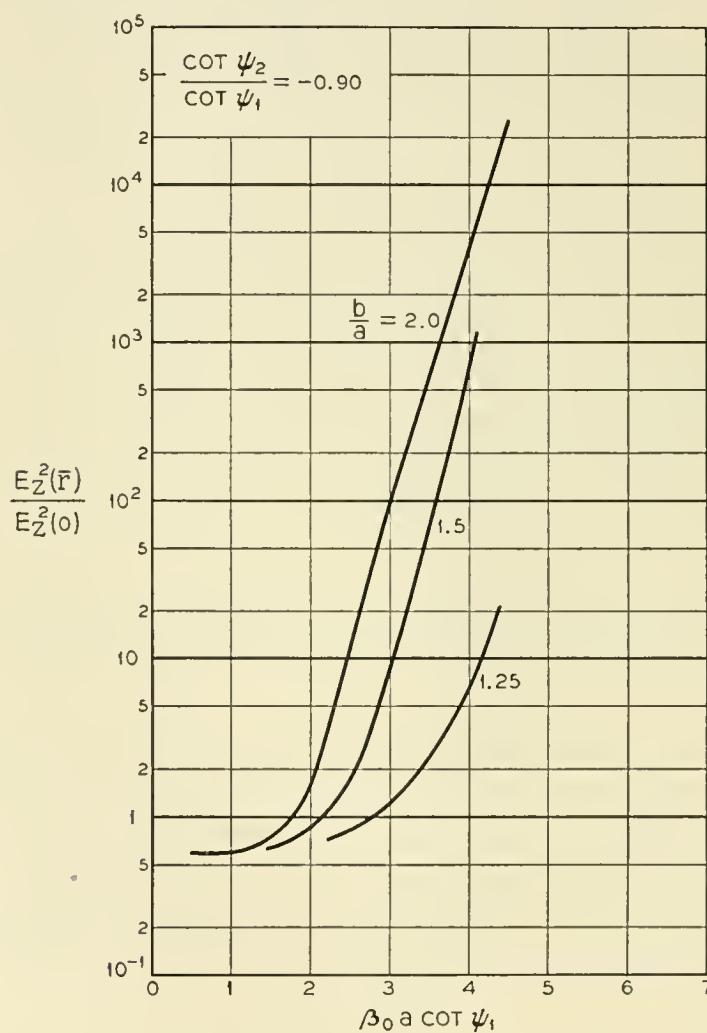


Fig. 2.11 — The relation between the impedance in terms of the longitudinal field between coupled helices excited in the in-phase mode, and the impedance in terms of the longitudinal field on the axis shown as a function of $\beta_0 a \cot \psi_1$.

have to be supplied to each helix at the input end. A natural way of doing this might be by means of a two-conductor balanced transmission line (Lecher-line), one conductor being connected to the inner helix, the other to the outer helix. Such an arrangement would cause something like the transverse (+-) mode to be set up on the helices. If the two conductors and the balanced line can be shielded from each other starting some distance from the helices then it is, in principle, possible to introduce arbitrary amounts of extra delay into one of the conductors. A delay of one half period would then cause the longitudinal (++) mode to be set up in the helices. Clearly such a coupling scheme would not be broad-band since a frequency-independent delay of one half period is not realizable.

Other objections to both of these schemes are: Balanced lines are not generally used at microwave frequencies; it is difficult to bring leads through the envelope of a TWT without causing reflection of RF energy and without unduly encumbering the mechanical design of the tube plus circuits; both schemes are necessarily inexact because helices having different radii will, in general, require different voltages at either input in order to be excited in a pure mode.

Thus the practicability, and success, of any general scheme based on the existence of a pure transverse or a pure longitudinal mode on coupled helices will depend to a large extent on whether elegant coupling means are available. Such means are indeed in existence as will be shown in the next sections.

3.1.2 *Tapered Coupler*

A less direct but more elegant means of coupling an external circuit to either normal mode of a double helix arrangement is by the use of the so-called "tapered" coupler.^{8, 9, 10} By appropriately tapering the relative propagation velocities of the inner and outer helices, outside the interaction region, one can excite either normal mode by coupling to one helix only.

The principle of this coupler is based on the fact that any two coupled transmission lines support two, and only two, normal modes, regardless of their relative phase velocities. These normal modes are characterized by unequal wave amplitudes on the two lines if the phase velocities are not equal. Indeed the greater the phase velocity difference and/or the smaller the coupling coefficient between the lines, the more their wave amplitudes diverge. Furthermore, the wave amplitude on the line with the slower phase velocity is greater for the out-of-phase or transverse normal mode, and the wave amplitude on the faster line is greater

for the longitudinal normal mode. As the ratio of phase constant to coupling constant approaches infinity, the ratio of the wave amplitudes on the two lines does also. Finally, if the phase velocities of, or coupling between, two coupled helices are changed gradually along their length the normal modes existing on the pair roughly maintain their identity even though they change their character. Thus, by properly tapering the phase velocities and coupling strength of any two coupled helices one can cause the two normal modes to become two separate waves, one existing on each helix.

For instance, if one desires to extract a signal propagating in the in-phase, or longitudinal, normal mode from two concentric helices of equal phase velocity, one might gradually increase the pitch of the outer helix and decrease that of the inner, and at the same time increase the diameter of the outer helix to decrease the coupling, until the longitudinal mode exists as a wave on the outer helix only. At such a point the outer helix may be connected to a coaxial line and the signal brought out.

This kind of coupler has the advantage of being frequency insensitive; and, perhaps, operable over bandwidths upwards of two octaves. It has the disadvantage of being electrically, and sometimes physically, quite long.

3.1.3 *Stepped Coupler*

There is yet a third way to excite only one normal mode on a double helix. This scheme consists of a short length at each end of the outer helix, for instance, which has a pitch slightly different from the rest. This has been called a "stepped" coupler.

The principle of the stepped coupler is this: If two coupled transmission lines have unlike phase velocities then a wave initiated in one line can never be completely transferred to the other, as has been shown in Section 2.4. The greater the velocity difference the less will be the maximum transfer. One can choose a velocity difference such that the maximum power transfer is just one half the initial power. It is a characteristic of incomplete power transfer that at the point where the maximum transfer occurs the waves on the two lines are exactly either in-phase or out-of-phase, depending on which helix was initially excited. Thus, the conditions for a normal mode on two equal-velocity helices can be produced at the maximum transfer point of two unlike velocity helices by initiating a wave on only one of them. If at that point the helix pitches are changed to give equal phase velocities in both helices, with equal current or voltage amplitude on both helices, either one or the other of the two normal modes will be propagated on the two helices from there on. Although the

pitch and length of such a stepped coupler are rather critical, the requirements are indicated in the equations in Section 2.4.

The useful bandwidth of the stepped coupler is not as great as that of the tapered variety, but may be as much as an octave. It has however the advantage of being very much shorter and simpler than the tapered coupler.

3.2 Low-Noise Transverse-Field Amplifier

One application of coupled helices which has been suggested from the very beginning is for a transverse field amplifier with low noise factor. In such an amplifier the RF structure is required to produce a field which is purely transverse at the position of the beam. For the transverse mode there is always such a cylindrical surface where the longitudinal field is zero and this can be obtained from the field equation of Appendix II. In Fig. 3.1 we have plotted the value of the radius \bar{r} at which the longitudinal field is zero for various parameters. The significant feature of this plot is that the radius which specifies zero longitudinal field is not constant with frequency. At frequencies away from the design frequency the electron beam will be in a position where interaction with longitudinal components might become important and thus shotnoise power will be introduced into the circuit. Thus the bandwidth of the amplifier over which it has a good noise factor would tend to be limited. However, this effect can be reduced by using the smallest practicable value of b/a .

Section 2.12 indicates that the impedance of the transverse mode is very high, and thus this structure should be well suited for transverse field amplifiers.

3.3 Dispersive Traveling-Wave Tube

Large bandwidth is not always essential in microwave amplifiers. In particular, the enormous bandwidth over which the traveling-wave tube is potentially capable of amplifying has so far found little application, while relatively narrow bandwidths (although quite wide by previous standards) are of immediate interest. Such a relatively narrow band, if it is an inherent electronic property of the tube, makes matching the tube to the external circuits easier. It may permit, for instance, the use of non-reciprocal attenuation by means of ferrites in the ferromagnetic resonance region. It obviates filters designed to deliberately reduce the band in certain applications. Last, but not least, it offers the possibility of trading bandwidth for gain and efficiency.

A very simple method of making a traveling-wave tube narrow-band

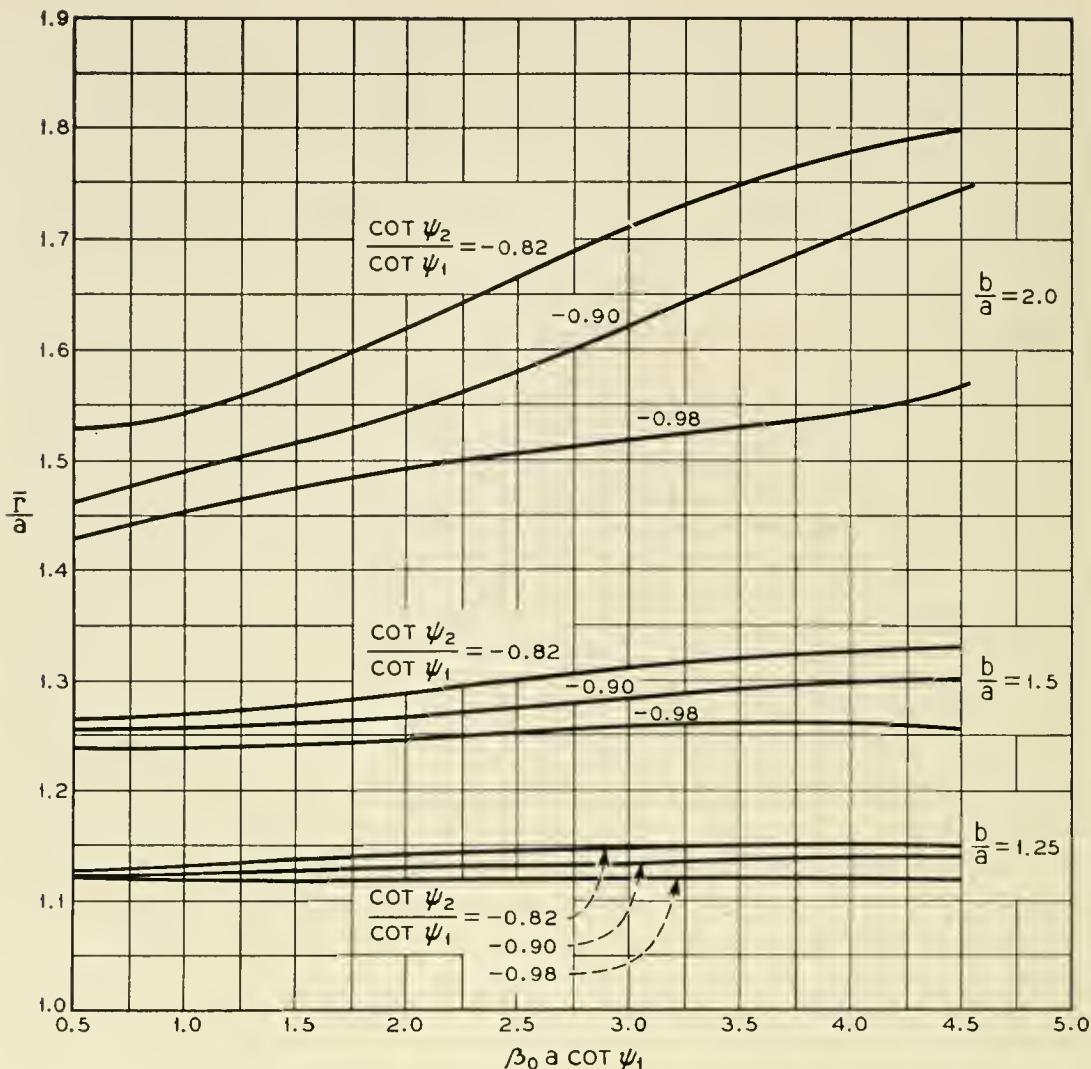


Fig. 3.1 — The radius \bar{r} at which the longitudinal field is zero for transversely excited coupled coaxial helices.

is by using a dispersive circuit, (i.e. one in which the phase velocity varies significantly with frequency). Thus, we obtain an amplifier that can be tuned by varying the beam voltage; being dispersive we should also expect a low group velocity and therefore higher circuit impedance.

Calculations of the phase velocities of the normal modes of coupled concentric helices presented in the appendix show that the fast, longitudinal or (++) mode is highly dispersive. Given the geometry of two such coupled helices and the relevant data on an electron beam, namely current, voltage and beam radius, it is possible to arrive at an estimate of the dependence of gain on frequency.

Experiments with such a tube showed a bandwidth 3.8 times larger than the simple estimate would show. This we ascribe to the presence

of the dielectric between the helices in the actual tube, and to the neglect of power propagated in the form of spatial harmonics.

Nevertheless, the tube operated satisfactorily with distributed non-reciprocal ferrite attenuation along the whole helix and gave, at the center frequency of 4,500 mc/s more than 40 db stable gain.

The gain fell to zero at 3,950 mc/s at one end of the band and at 4,980 mc/s at the other. The forward loss was 12 db. The backward loss was of the order of 50 db at the maximum gain frequency.

3.4 Devices Using Both Modes

In this section we shall discuss applications of the coupled-helix principle which depend for their function on the simultaneous presence of both the transverse and the longitudinal modes. When present in substantially equal magnitude a spatial beat-phenomenon takes place, that is, RF power transfers back and forth between inner and outer helix.

Thus, there are points, periodic with distance along each helix, where there is substantially no current or voltage; at these points a helix can be terminated, cut-off, or connected to external circuits without detriment.

The main object, then, of all devices discussed in this section is power transfer from one helix to the other; and, as will be seen, this can be accomplished in a remarkably efficient, elegant, and broad-band manner.

3.4.1 Coupled-Helix Transducer

It is, by now, a well known fact that a good match can be obtained between a coaxial line and a helix of proportions such as used in TWT's. A wire helix in free space has an effective impedance of the order of 100 ohms. A conducting shield near the helix, however, tends to reduce the helix impedance, and a value of 70 or even 50 ohms is easily attained. Provided that the transition region between the coaxial line and the helix does not present too abrupt a change in geometry or impedance, relatively good transitions, operable over bandwidths of several octaves, can be made, and are used in practice to feed into and out of tubes employing helices such as TWT's and backward-wave oscillators.

One particularly awkward point remains, namely, the necessity to lead the coaxial line through the tube envelope. This is a complication in manufacture and requires careful positioning and dimensioning of the helix and other tube parts.

Coupled helices offer an opportunity to overcome this difficulty in the form of the so-called coupled-helix transducer, a sketch of which is shown in Fig. 3.2. As has been shown in Section 2.3, with helices having

the same velocity an overlap of one half of a beat wavelength will result in a 100 per cent power transfer from one helix to the other. A signal introduced into the outer helix at point A by means of the coaxial line will be all on the inner helix at point B, nothing remaining on the outer helix. At that point the outer helix can be discontinued, or cut off; since there is no power there, the seemingly violent discontinuity represented by the "open" end of the helix will cause no reflection of power. In practice, unfortunately, there are always imperfections to consider, and there will often be some power left at the end of the coupler helix. Thus, it is desirable to terminate the outer helix at this point non-reflectively, as, for instance, by a resistive element of the right value, or by connecting to it another matched coaxial line which in turn is then non-reflectively terminated.

It will be seen, therefore, that the coupled-helix transducer can, in principle, be made into an efficient device for coupling RF energy from a coaxial line to a helix contained in a dielectric envelope such as a glass tube. The inner helix will be energized predominantly in one direction, namely, the one away from the input connection. Conversely, energy traveling initially in the inner helix will be transferred to the outer, and made available as output in the respective coaxial line. Such a coupled-helix transducer can be moved along the tube, if required. As long as the outer helix completely overlaps the inner, operation as described above should be assured. By this means a new flexibility in design, operation and adjustment of traveling-wave tubes is obtained which could not be achieved by any other known form of traveling-wave tube transducer.

Naturally, the applications of the coupled-helix transducer are not restricted to TWT's only, nor to 100 per cent power transfer. To obtain

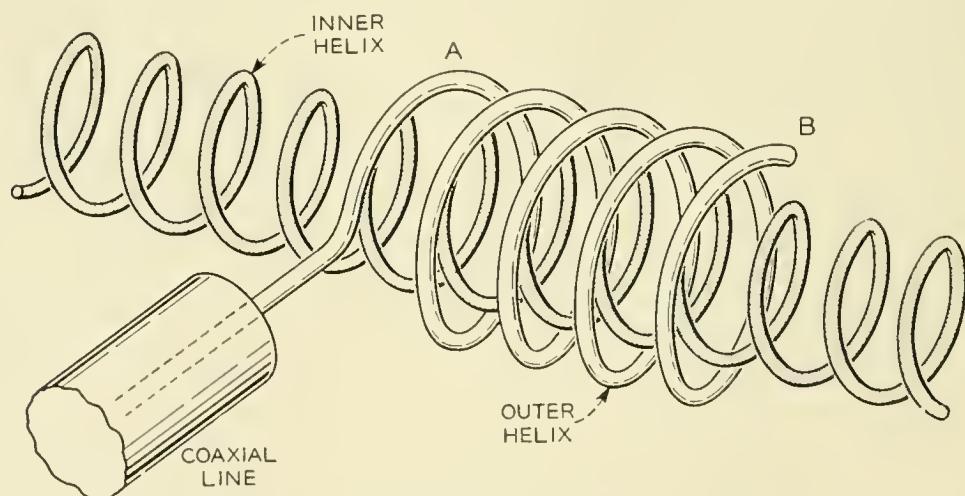


Fig. 3.2 — A simple coupled helix transducer.

power transfer of proportions other than 100 per cent two possibilities are open: either one can reduce the length of the synchronous coupling helix appropriately, or one can deliberately make the helices non-synchronous. In the latter case, a considerable measure of broad-banding can be obtained by making the length of overlap again equal to one half of a beat-wavelength, while the fraction of power transferred is determined by the difference of the helix velocities according to 2.4.7. An application of the principle of the coupled-helix transducer to a variable delay line has been described by L. Stark¹² in an unpublished memorandum.

Turning again to the complete power transfer case, we may ask: How broad is such a coupler?

In Section 2.7 we have discussed how the radial falling-off of the RF energy near a helix can be used to broad-band coupled-helix devices which depend on relative constancy of beat-wavelength as frequency is varied. On the assumption that there exists a perfect broad-band match between a coaxial line and a helix, one can calculate the performance of a coupled-helix transducer of the type shown in Fig. 3.2.

Let us define a center frequency ω , at which the outer helix is exactly one half beat-wavelength, λ_b , long. If ω is the frequency of minimum beat wavelength then at frequencies ω_1 and ω_2 , larger and smaller, respectively, than ω , the outer helix will be a fraction δ shorter than $\frac{1}{2}\lambda_b$, (Section 2.7). Let a voltage amplitude, V_2 , exist at the point where the outer helix is joined to the coaxial line. Then the magnitude of the voltage at the other end of the outer helix will be $|V_2 \cdot \sin(\pi\delta/2)|$ which means that the power has not been completely transferred to the inner helix. Let us assume complete reflection at this end of the outer helix. Then all but a fraction of the reflected power will be transferred to the inner helix in a *reverse* direction. Thus, we have a first estimate for the "directivity" defined as the ratio of forward to backward power (in db) introduced into the inner helix:

$$D = \left| 10 \log \sin^2 \left(\frac{\pi\delta}{2} \right) \right| \quad (3.4.1.1)$$

We have assumed a perfect match between coaxial line and outer helix; thus the power reflected back into the coaxial line is proportional to $\sin^4(\pi\delta/2)$. Thus the reflectivity defined as the ratio of reflected to incident power is given in db by

$$R = 10 \log \sin^4 \left(\frac{\pi\delta}{2} \right) \quad (3.4.1.2)$$

For the sake of definiteness, let us choose actual figures: let $\beta a = 2.0$, and $b/a = 1.5$. And let us, arbitrarily, demand that R always be less than -20 db.

This gives $\sin(\pi\delta/2) < 0.316$ and $\pi\delta/2 < 18.42^\circ$ or 0.294 radians, $\delta < 0.205$. With the optimum value of $\beta_c a = 1.47$, this gives the minimum permissible value of $\beta_c a$ of $1.47/(1 + 0.205) = 1.22$. From the graph on Fig. 2.2 this corresponds to values of βa of 1.00 and 3.50. Therefore, the reflected power is down 20 db over a frequency range of $\omega_2/\omega_1 = 3.5$ to one. Over the same range, the directivity is better than 10 to one. Suppose a directivity of better than 20 db were required. This requires $\sin(\pi\delta/2) = 0.10$, $\delta = 0.0638$ and is obtained over a frequency range of approximately two to one. Over the same range, the reflected power would be down by 40 db.

In the above example the full bandwidth possibilities have not been used since the coupler has been assumed to have optimum length when $\beta_c a$ is maximum. If the coupler is made longer so that when $\beta_c a$ is maximum it is electrically short of optimum to the extent permissible by the quality requirements, then the minimum allowable $\beta_c a$ becomes even smaller. Thus, for $b/a = 1.5$ and directivity 20 db or greater the realizable bandwidth is nearly three to one.

When the coupling helix is non-reflectively terminated at both ends, either by means of two coaxial lines or a coaxial line at one end and a resistive element at the other, the directivity is, ideally, infinite, irrespective of frequency; and, similarly, there will be no reflections. The power transfer to the inner helix is simply proportional to $\cos^2(\pi\delta/2)$. Thus, under the conditions chosen for the example given above, the coupled-helix transducer can approach the ideal transducer over a considerable range of frequencies.

So far, we have inspected the performance and bandwidth of the coupled-helix transducer from the most optimistic theoretical point of view. Although a more realistic approach does not change the essence of our conclusions, it does modify them. For instance, we have neglected dispersion on the helices. Dispersion tends to reduce the maximum attainable bandwidth as can be seen if Fig. 2.4.2 rather than Fig. 2.2 is used in the example cited above. The dielectric that exists in the annular region between coupled concentric helices in most practical couplers may also affect the bandwidth.

In practice, the performance of coupled-helix transducers has been short of the ideal. In the first place, the match from a coaxial line to a helix is not perfect. Secondly, a not inappreciable fraction of the RF power on a real wire helix is propagated in the form of spatial harmonic

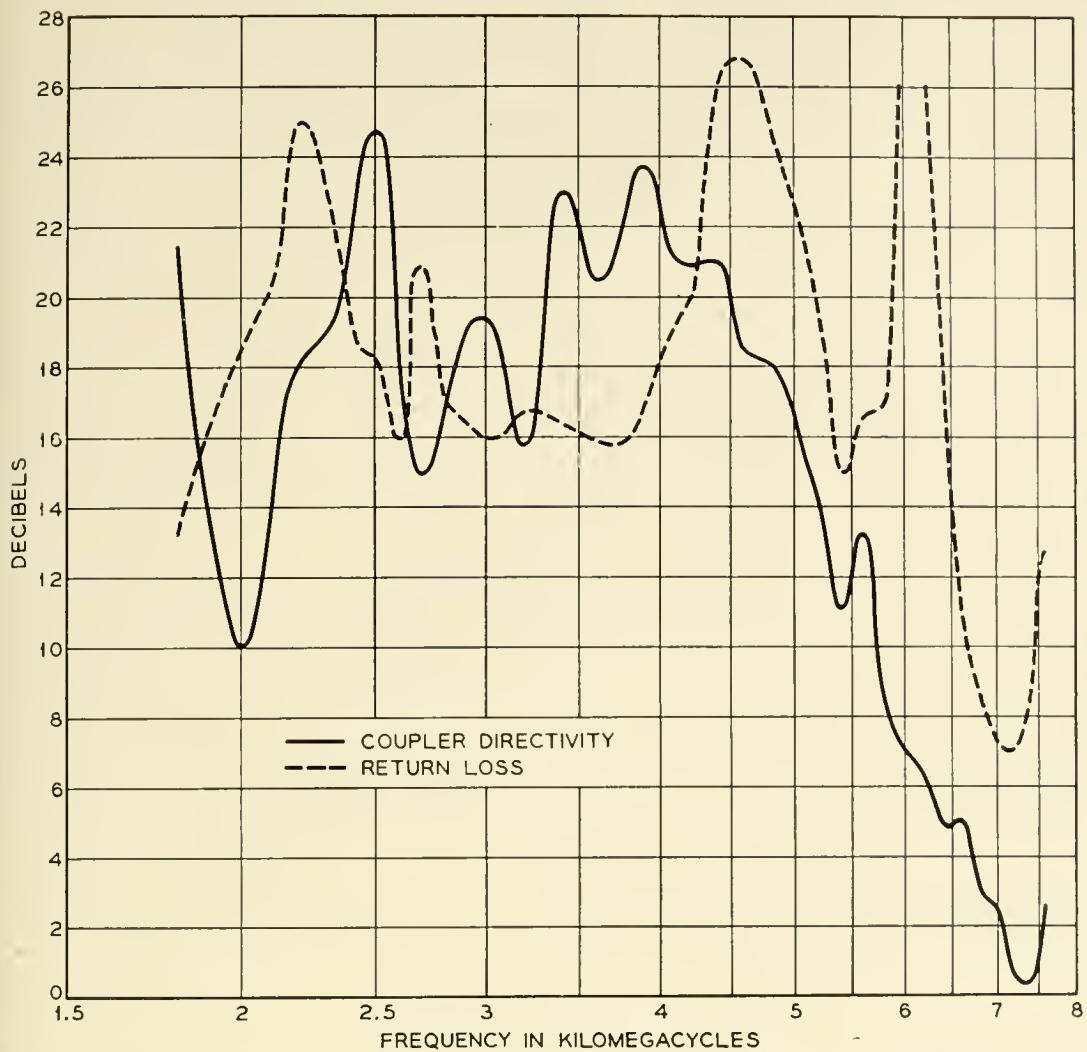


Fig. 3.3 — The return loss and directivity of an experimental 100 per cent coupled-helix transducer.

wave components which have variations with angle around the helix-axis, and coupling between such components on two helices wound in opposite directions must be small. Finally, there are the inevitable mechanical inaccuracies and misalignments.

Fig. 3.3 shows the results of measurements on a coupled-helix transducer with no termination at the far end.

3.4.2 Coupled-Helix Attenuator

In most TWT's the need arises for a region of heavy attenuation somewhere between input and output; this serves to isolate input and output, and prevents oscillations due to feedback along the circuit. Because of the large bandwidth over which most TWT's are inherently capable of amplifying, substantial attenuation, say at least 60 db, is

required over a bandwidth of maybe 2 octaves, or even more. Furthermore, such attenuation should present a very good match to a wave on the helix, particularly to a wave traveling backwards from the output of the tube since such a wave will be amplified by the output section of the tube.

Another requirement is that the attenuator should be physically as short as possible so as not to increase the length of the tube unnecessarily.

Finally, such attenuation might, with advantage, be made movable during the operation of the tube in order to obtain optimum performance, perhaps in respect of power output, or linearity, or some other aspect.

Coupled-helix attenuators promise to perform these functions satisfactorily.

A length of outer helix (synchronous with the inner helix) one half of a beat wavelength long, terminated at either end non-reflectively, forms a very simple, short, and elegant solution of the coupled-helix attenuator problem. A notable weakness of this form of attenuator is its relatively narrow bandwidth. Proceeding, as before, on the assumption that the attenuator is a fraction δ larger or smaller than half a beat wavelength at frequencies ω_1 and ω_2 on either side of the center frequency ω , we find that the fraction of power transferred from the inner helix to the attenuator is then given by $(1 - \sin^2(\pi\delta/2))$. The attenuation is thus simply

$$A = \sin^2\left(\frac{\pi\delta}{2}\right)$$

For helices of the same proportions as used before in Section 3.4.1, we find that this will give an attenuation of at least 20 db over a frequency band of two to one. At the center frequency, ω_0 , the attenuation is infinite; — in theory.

Thus to get higher attenuation, it would be necessary to arrange for a sufficient number of such attenuators in tandem along the TWT. Moreover, by properly staggering their lengths within certain ranges a wider attenuation band may be achieved. The success of such a scheme largely depends on the ability to terminate the helix ends non-reflectively. Considerable work has been done in this direction, but complete success is not yet in sight.

Another basically different scheme for a coupled-helix attenuator rests on the use of distributed attenuation along the coupling helix. The difficulty with any such scheme lies in the fact that *unequal* attenuation in the two coupled helices reduces the coupling between them and the more they differ in respect to attenuation, the less the coupling. Naturally, one

would wish to have as little attenuation as practicable associated with the inner helix (inside the TWT). This requires the attenuating element to be associated with the outer helix. Miller⁵ has shown that the maximum total power reduction in coupled transmission systems is obtained when

$$\left| \frac{\alpha_1 - \alpha_2}{\beta_b} \right| \approx 1$$

where α_1 and α_2 are the attenuation constants in the respective systems, and β_b the beat phase constant. If the inner helix is assumed to be lossless, the attenuation constant of the outer helix has to be effectively equal to the beat wave phase constant. It turns out that 60 db of attenuation requires about 3 beat wavelengths (in practice 10 to 20 helix wavelengths). The total length of a typical TWT is only 3 or 4 times that, and it will be seen, therefore, that this scheme may not be practical as the only means of providing loss.

Experiments carried out with outer helices of various resistivities and thicknesses by K. M. Poole (then at the Clarendon Laboratory, Oxford, England) tend to confirm this conclusion. P. D. Lacy¹¹ has described a coupled helix attenuator which uses a multifilar helix of resistance material together with a resistive sheath between the helices.

Experiments were performed at Bell Telephone Laboratories with a TWT using a resistive sheath (graphite on paper) placed between the outer helix and the quartz tube enclosing the inner helix. The attenuations were found to be somewhat less than estimated theoretically. The attenuator helix was movable in the axial direction and it was instructive to observe the influence of attenuator position on the power output from the tube, particularly at the highest attainable power level. As one might expect, as the power level is raised, the attenuator has to be moved nearer to the input end of the tube in order to obtain maximum gain and power output. In the limit, the attenuator helix has to be placed right close to the input end, a position which does not coincide with that for maximum low-level signal gain. Thus, the potential usefulness of the feature of mobility of coupled-helix elements has been demonstrated.

4. CONCLUSION

In this paper we have made an attempt to develop and collect together a considerable body of information, partly in the form of equations, partly in the form of graphs, which should be of some help to workers in the field of microwave tubes and devices. Because of the crudity of the assumptions, precise agreement between theory and experiment has not

been attained nor can it be expected. Nevertheless, the kind of physical phenomena occurring with coupled helices are, at least, qualitatively described here and should permit one to develop and construct various types of devices with fair chance of success.

ACKNOWLEDGEMENTS

As a final note the authors wish to express their appreciation for the patient work of Mrs. C. A. Lambert in computing the curves, and to G. E. Korb for taking the experimental data.

APPENDIX I

I. SOLUTION OF FIELD EQUATIONS

In this section there is presented the field equations for a transmission system consisting of two helices aligned with a common axis. The propagation properties and impedance of such a transmission system are discussed for various ratios of the outer helix radius to the inner helix radius. This system is capable of propagating two modes and as previously pointed out one mode is characterized by a longitudinal field midway between the two helices and the other is characterized by a transverse field midway between the two helices.

The model which is to be treated and shown in Fig. 2.3 consists of an inner helix of radius a and pitch angle ψ_1 which is coaxial with the outer helix of radius b and pitch angle ψ_2 . The sheath helix model will be treated, wherein it is assumed that helices consist of infinitely thin sheaths which allow for current flow only in the direction of the pitch angle ψ .

The components of the field in the region inside the inner helix, between the two helices and outside the outer helix can be written as follows — inside the inner helix

$$H_{z_1} = B_1 I_0(\gamma r) \quad (1)$$

$$E_{z_2} = B_2 I_0(\gamma r) \quad (2)$$

$$H_{\varphi_2} = j \frac{\omega \epsilon}{\gamma} B_2 I_1(\gamma r) \quad (3)$$

$$H_{r_1} = \frac{j\beta}{\gamma} B_1 I_1(\gamma r) \quad (4)$$

$$E_{\varphi_1} = -j \frac{\omega \mu}{\gamma} B_1 I_1(\gamma r) \quad (5)$$

$$E_{r_2} = \frac{j\beta}{\gamma} B_2 I_1(\gamma r) \quad (6)$$

and between the two helices

$$H_{z_3} = B_3 I_0(\gamma r) + B_4 K_0(\gamma r) \quad (7)$$

$$E_{z_5} = B_5 I_0(\gamma r) + B_6 K_0(\gamma r) \quad (8)$$

$$H_{\varphi_5} = \frac{j\omega\epsilon}{\gamma} [B_5 I_1(\gamma r) - B_6 K_1(\gamma r)] \quad (9)$$

$$H_{r_3} = \frac{j\beta}{\gamma} [B_3 I_1(\gamma r) - B_4 K_1(\gamma r)] \quad (10)$$

$$E_{\varphi_3} = -j \frac{\omega\mu}{\gamma} [B_3 I_1(\gamma r) - B_4 K_1(\gamma r)] \quad (11)$$

$$E_{r_5} = \frac{j\beta}{\gamma} [B_5 I_1(\gamma r) - B_6 K_1(\gamma r)] \quad (12)$$

and outside the outer helix

$$H_{z_y} = B_7 K_0(\gamma r) \quad (13)$$

$$E_{z_8} = B_8 K_0(\gamma r) \quad (14)$$

$$H_{\varphi_8} = -j \frac{\omega\epsilon}{\gamma} B_8 K_1(\gamma r) \quad (15)$$

$$H_{r_7} = \frac{-j\beta}{\gamma} B_7 K_1(\gamma r) \quad (16)$$

$$E_{\varphi_7} = j \frac{\omega\mu}{\gamma} B_7 K_1(\gamma r) \quad (17)$$

$$E_{r_8} = \frac{-j\beta}{\gamma} B_8 K_1(\gamma r) \quad (18)$$

With the sheath helix model of current flow only in the direction of wires we can specify the usual boundary conditions that at the inner and outer helix radius the tangential electric field must be continuous and perpendicular to the wires, whereas the tangential component of magnetic field parallel to the current flow must be continuous. These can be written as

$$E_z \sin \psi + E_\varphi \cos \psi = 0 \quad (19)$$

E_z , E_φ and $(H_z \sin \psi + H_\varphi \cos \psi)$ be equal on either side of the helix.

By applying these conditions to the two helices the following equations are obtained for the various coefficients,

First, we will define a more simple set of parameters. We will denote

$$I_0(\gamma a) \text{ by } I_{01} \quad \text{and} \quad I_0(\gamma b) \text{ by } I_{02}, \text{ etc.}$$

Further let us use the notation introduced by Humphrey, Kite and James¹¹ in his treatment of coaxial helices.

$$\begin{aligned} P_{01} &\equiv I_{01}K_{01} & P_{02} &\equiv I_{02}K_{02} & R_0 &\equiv I_{01}K_{02} \\ P_{11} &\equiv I_{11}K_{11} & P_{12} &\equiv I_{12}K_{12} & R_1 &\equiv I_{11}K_{12} \end{aligned} \quad (20)$$

and define a common factor (C.F.) by the equation

$$\begin{aligned} \text{C.F.} = - \left[\frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{01}P_{02} - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 R_0 \right. \\ \left. + R_0^2 - P_{01}P_{02} \right] \end{aligned} \quad (21)$$

With all of this we can now write for the coefficients of equations 1 through 18:

$$\frac{B_1}{B_2} = -j \sqrt{\frac{\epsilon}{\mu}} \frac{\gamma a}{\beta_0 a \cot \psi_1} \frac{I_{01}}{I_{02}} \quad (22)$$

$$\frac{B_3}{B_2} = -i \sqrt{\frac{\epsilon}{\mu}} \frac{\beta_0 a \cot \psi_1}{\gamma a} \frac{I_{01}K_{12}}{\text{C.F.}} \left[\frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} R_1 - \frac{\cot \psi_2}{\cot \psi_1} R_0 \right] \quad (23)$$

$$\frac{B_4}{B_2} = -j \sqrt{\frac{\epsilon}{\mu}} \frac{\beta_0 a \cot \psi_1}{\gamma a} \frac{I_{01}I_{11}}{\text{C.F.}} \left[\frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} - P_{02} \right] \quad (24)$$

$$\frac{B_5}{B_2} = -\frac{R_0}{\text{C.F.}} \left[R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a^2)} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right] \quad (25)$$

$$\frac{B_6}{B_2} = -\frac{I_{01}^2}{\text{C.F.}} \left[\frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} - P_{02} \right] \quad (26)$$

$$\frac{B_7}{B_2} = j \sqrt{\frac{\epsilon}{\mu}} \frac{\beta_0 a \cot \psi_1}{\gamma a} \frac{1}{\text{C.F.}} \frac{I_{01}}{K_{12}} \left[P_{02}R_1 - \frac{\cot \psi_2}{\cot \psi_1} P_{12}R_0 \right] \quad (27)$$

$$\frac{B_8}{B_2} = \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} \frac{I_{01}^2}{\text{C.F.} R_0} \left[P_{02}R_1 - \frac{\cot \psi_2}{\cot \psi_1} P_{12}R_0 \right] \quad (28)$$

The last equation necessary for the solution of our field problem is the transcendental equation for the propagation constant, γ , which can be

written

$$\left[R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right]^2 = \left[P_{02} - \frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} \right] \left[P_{01} - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} P_{11} \right] \quad (29)$$

The solutions of this equation are plotted in Fig. 4.1.

There it is seen that there are two values of γ , one, γ_t , denoting the slow mode with transverse fields between helices and the other, γ_ℓ , denoting the fast mode with longitudinal fields midway between the two helices.

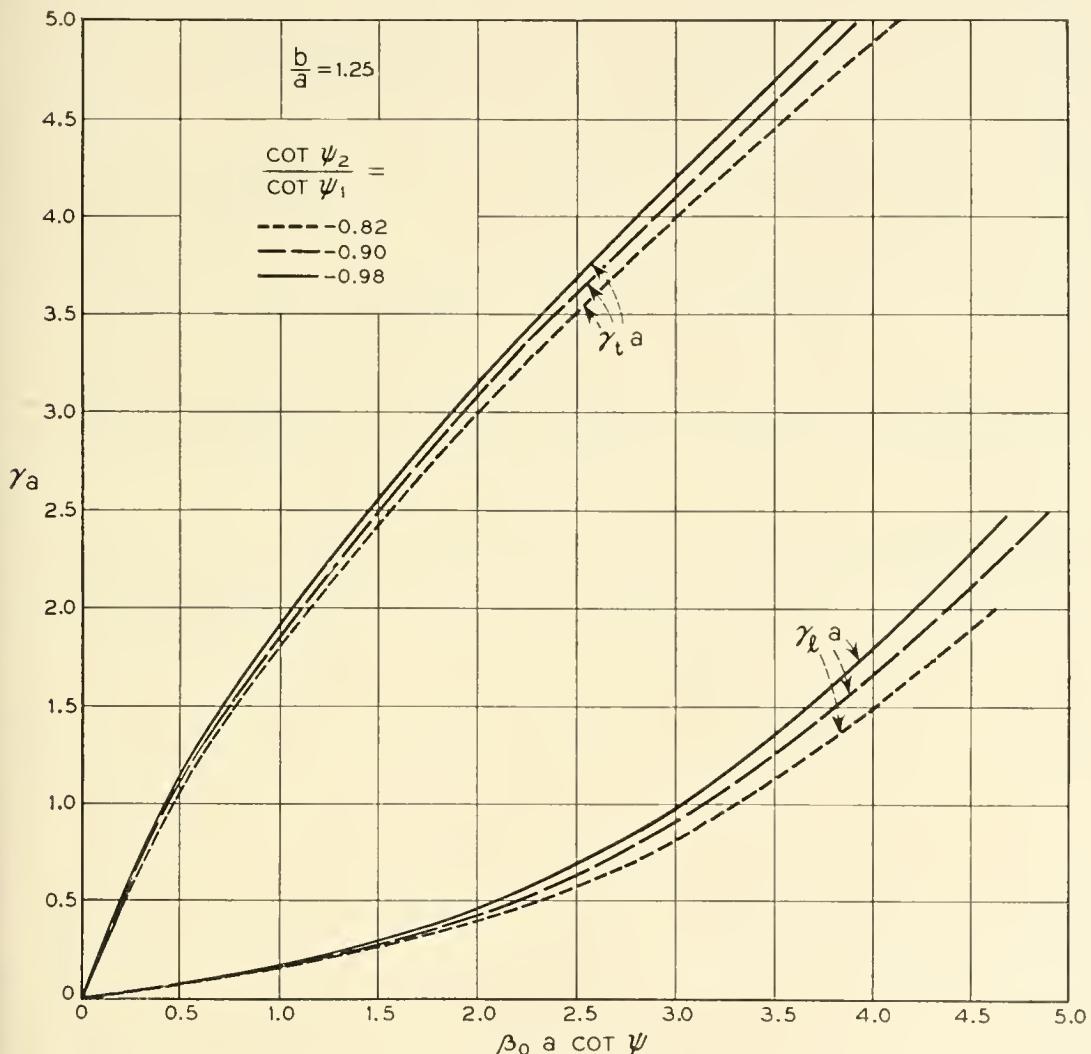


Fig. 4.1.1 — The radial propagation constants associated with the transverse and longitudinal modes on coupled coaxial sheath helices given as a function of $\beta_0 a \cot \psi_1$ for several values of $b/a = 1.25$.

These equations can now be used to compute the power flow as defined by

$$P = \frac{1}{2} \operatorname{Re} \int E \times H^* dA \quad (30)$$

which can be written in the form

$$\left[\frac{E_z^2(0)}{\beta^2 P} \right]^{\frac{1}{2}} = \frac{\beta}{\beta_0} \left(\frac{\gamma}{\beta} \right)^4 F(\gamma a, \gamma b) \quad (31)$$

where

$$\begin{aligned} [F(\gamma a, \gamma b)]^{-3} &= \frac{(\gamma a)^2}{240} \frac{I_{01}^2}{(\text{C.F.})^2} \left\{ \frac{\left(I_{01}^2 + \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} I_{11}^2 \right)}{(I_{11}^2 - I_{01} I_{21})(\text{C.F.})^2} \right. \\ &\quad - \left(K_{02}^2 + \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} K_{12}^2 \right) \left(R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right)^2 \\ &\quad - \left[\left(\frac{b}{a} \right)^2 (I_{02} I_{22} - I_{12}^2) + (I_{11}^2 - I_{01} I_{21}) \right] \\ &\quad + \left(R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right)^2 \left(P_{02} - \frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} \right) \\ &\quad \left. \left[\left(\frac{b}{a} \right)^2 (2I_{12} K_{12} + I_{02} K_{22} + I_{22} K_{02}) - (2I_{11} K_{11} + I_{01} K_{21} + I_{21} K_{01}) \right] \right\} \quad (32) \\ &\quad - \left[I_{01}^2 + \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} I_{11}^2 \right] \left[P_{02} - \frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} \right]^2 \\ &\quad \left[\left(\frac{b}{a} \right)^2 (K_{02} K_{22} - K_{12}^2) - (K_{01} K_{21} - K_{11}^2) \right] \\ &\quad + \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2 K_{12}^2 R_0^2} \left(\frac{b}{a} \right)^2 \left[R_0^2 + \frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} I_{01}^2 K_{12}^2 \right] \\ &\quad \left. \left[P_{02} R_1 - \frac{\cot \psi_2}{\cot \psi_1} P_{12} R_0 \right]^2 [K_{02} K_{22} - K_{12}^2] \right\} \end{aligned}$$

In (32) we find the power in the transverse mode by using values of

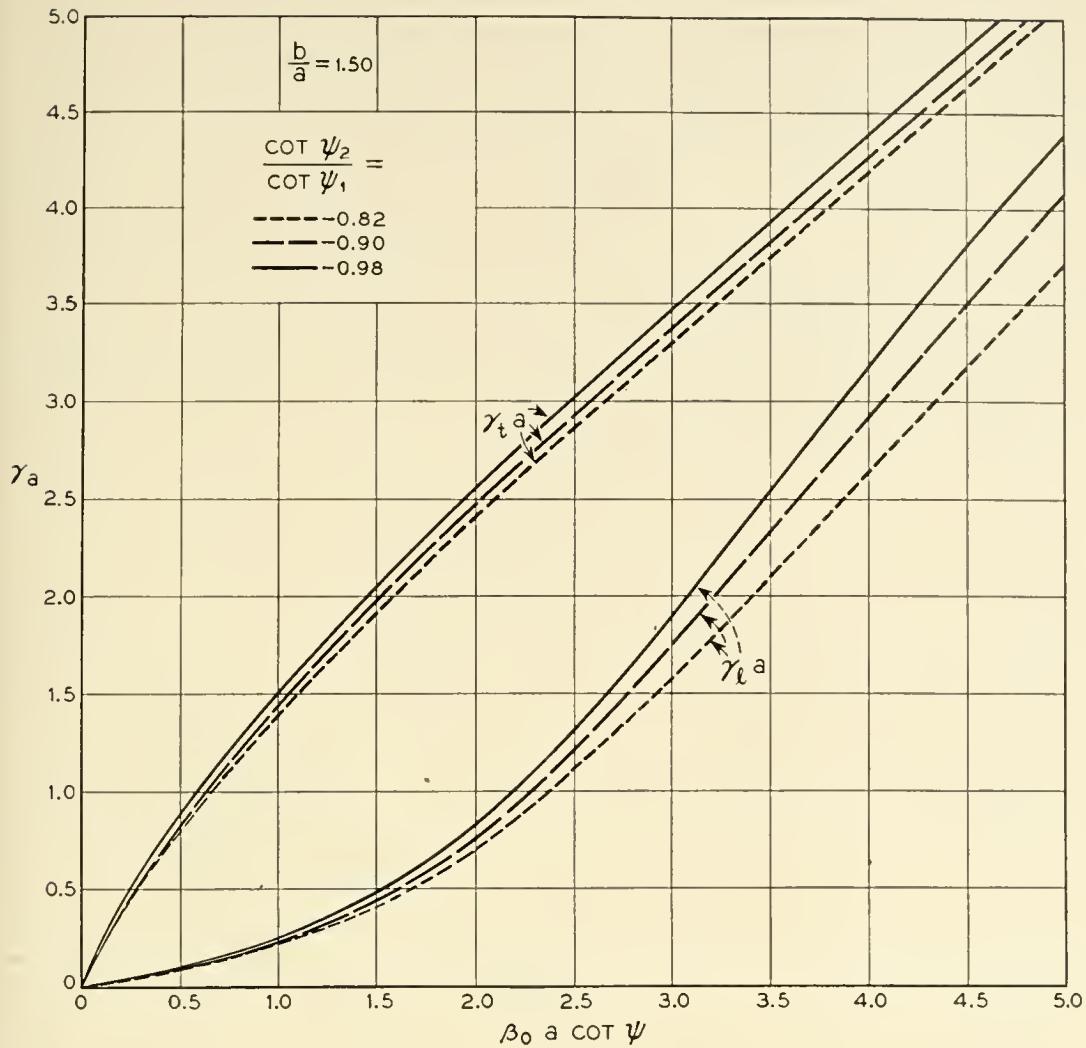


Fig. 4.1.2 — The radial propagation constants associated with the transverse and longitudinal modes on coupled coaxial sheath helices given as a function of $\beta_0 a \cot \psi_1$ when $b/a = 1.50$.

γ_t obtained from (29) and similarly the power in the longitudinal mode is found by using values of γ_ℓ .

II. FINDING \bar{r}

When coaxial helices are used in a transverse field amplifier, only the transverse field mode is of interest and it is important that the helix parameters be adjusted such that there is no longitudinal field at some radius, \bar{r} , where the cylindrical electron beam will be located. This condition can be expressed by equating E_z to zero at $r = \bar{r}$ and from (8)

$$B_5 I_0(\gamma \bar{r}) + B_6 K_0(\gamma \bar{r}) = 0 \quad (33)$$

which can be written with (25) and (26) as

$$K_{02} \left[R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right] I_0(\gamma \bar{r}) \\ = I_{01} \left[P_{02} - \frac{(\beta_0 a \cot \psi_2)^2}{(\gamma a)^2} P_{12} \right] K_0(\gamma \bar{r}) \quad (34)$$

This equation together with (29) enables one to evaluate \bar{r}/a versus $\beta_0 a \cot \psi_1$ for various ratios of b/a and $\cot \psi_2/\cot \psi_1$. The results of these calculations are shown in Fig. 3.1.

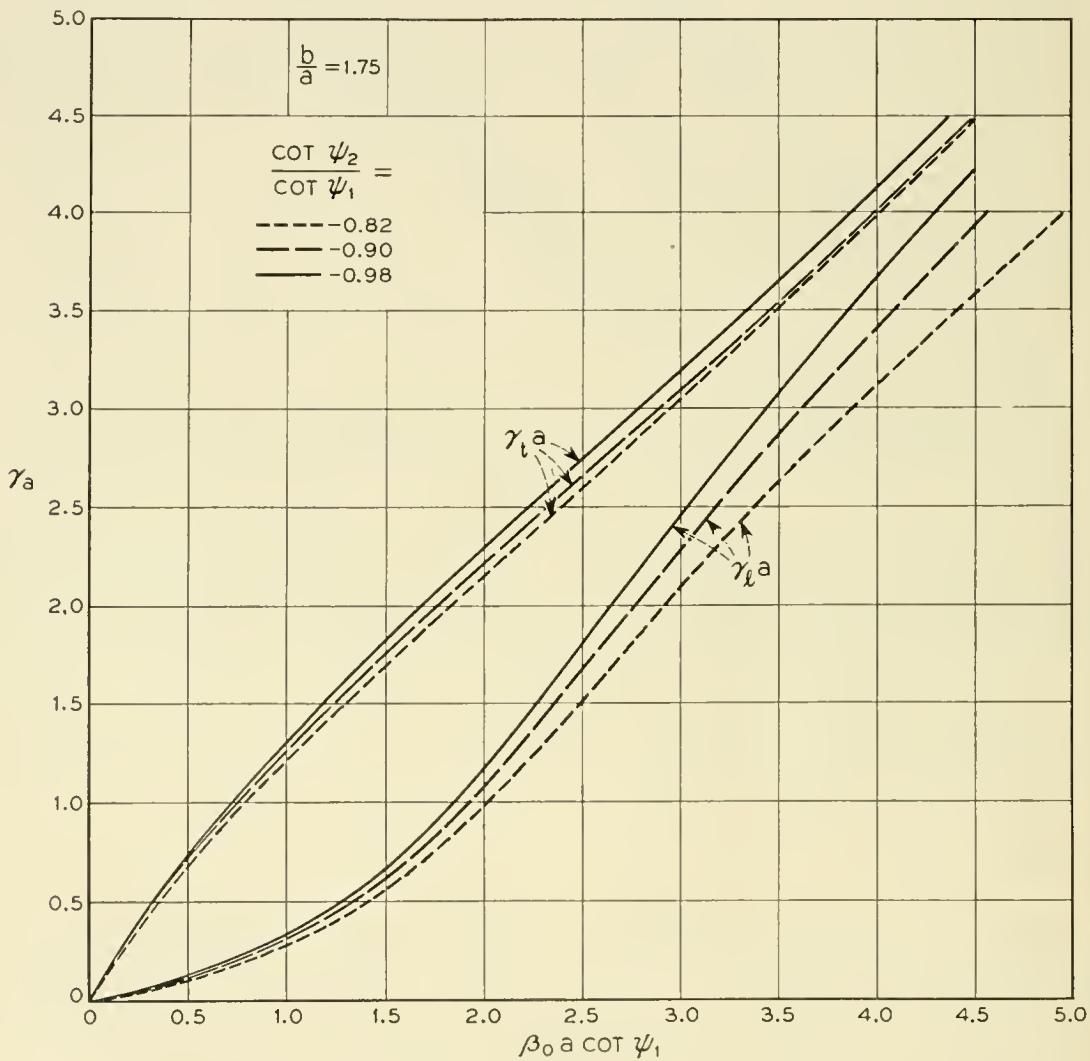


Fig. 4.1.3 — The radial propagation constants associated with the transverse and longitudinal modes on coupled coaxial sheath helices given as a function of $\beta_0 a \cot \psi_1$ when $b/a = 1.75$.

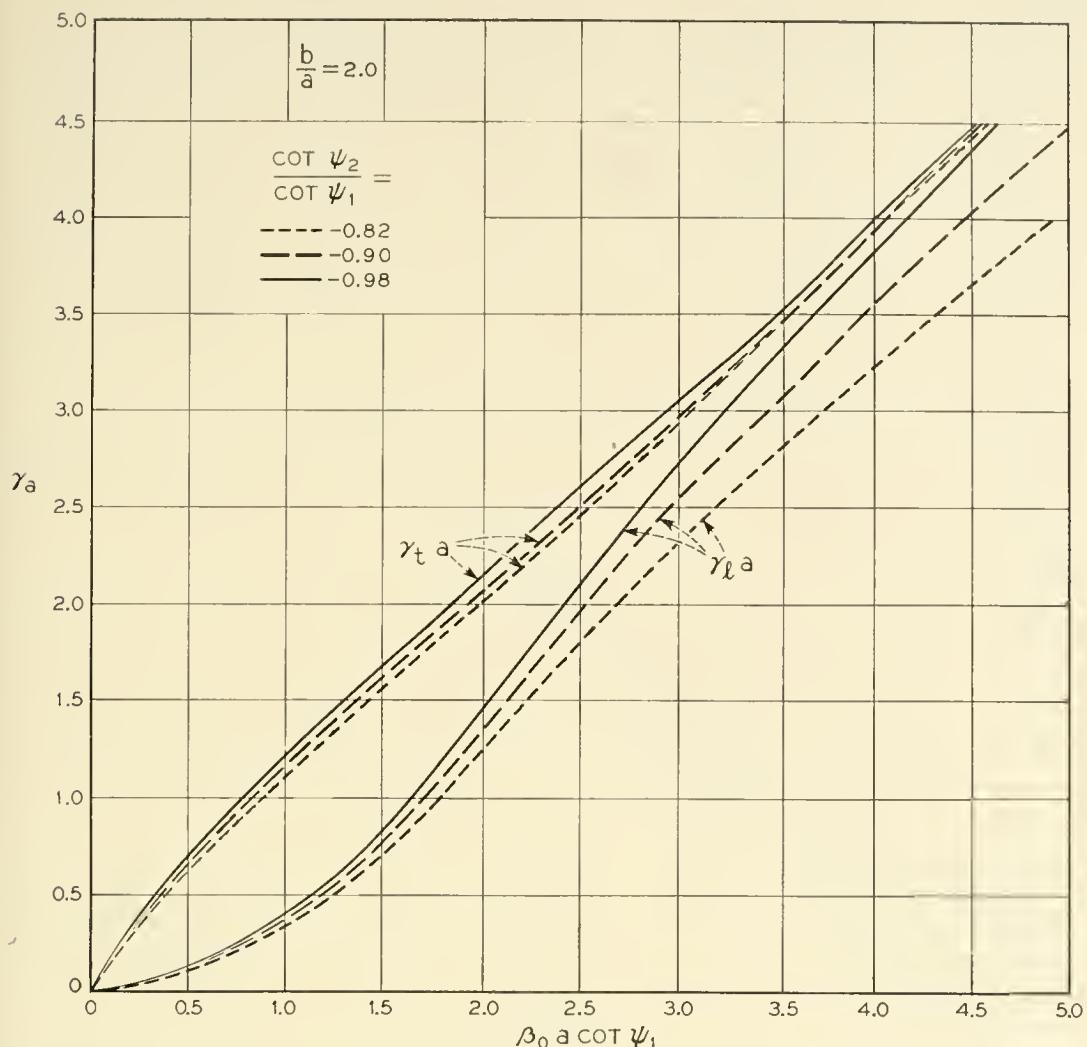


Fig. 4.1.4 — The radial propagation constants associated with the transverse and longitudinal modes on coupled coaxial sheath helices given as a function of $\beta_0 a \cot \psi_1$ when $b/a = 2.0$.

III. COMPLETE POWER TRANSFER

For coupled helix applications we require the coupled helix parameters to be adjusted so that RF power fed into one helix alone will set up the transverse and longitudinal modes equal in amplitude. For this condition the power from the outer helix will transfer completely to the inner helix. The total current density can be written as the sum of the current in the longitudinal mode and the transverse mode. Thus for the inner helix we have

$$J_a = J_{at} e^{-j\beta_t z} + J_{al} e^{-j\beta_l z} \quad (35)$$

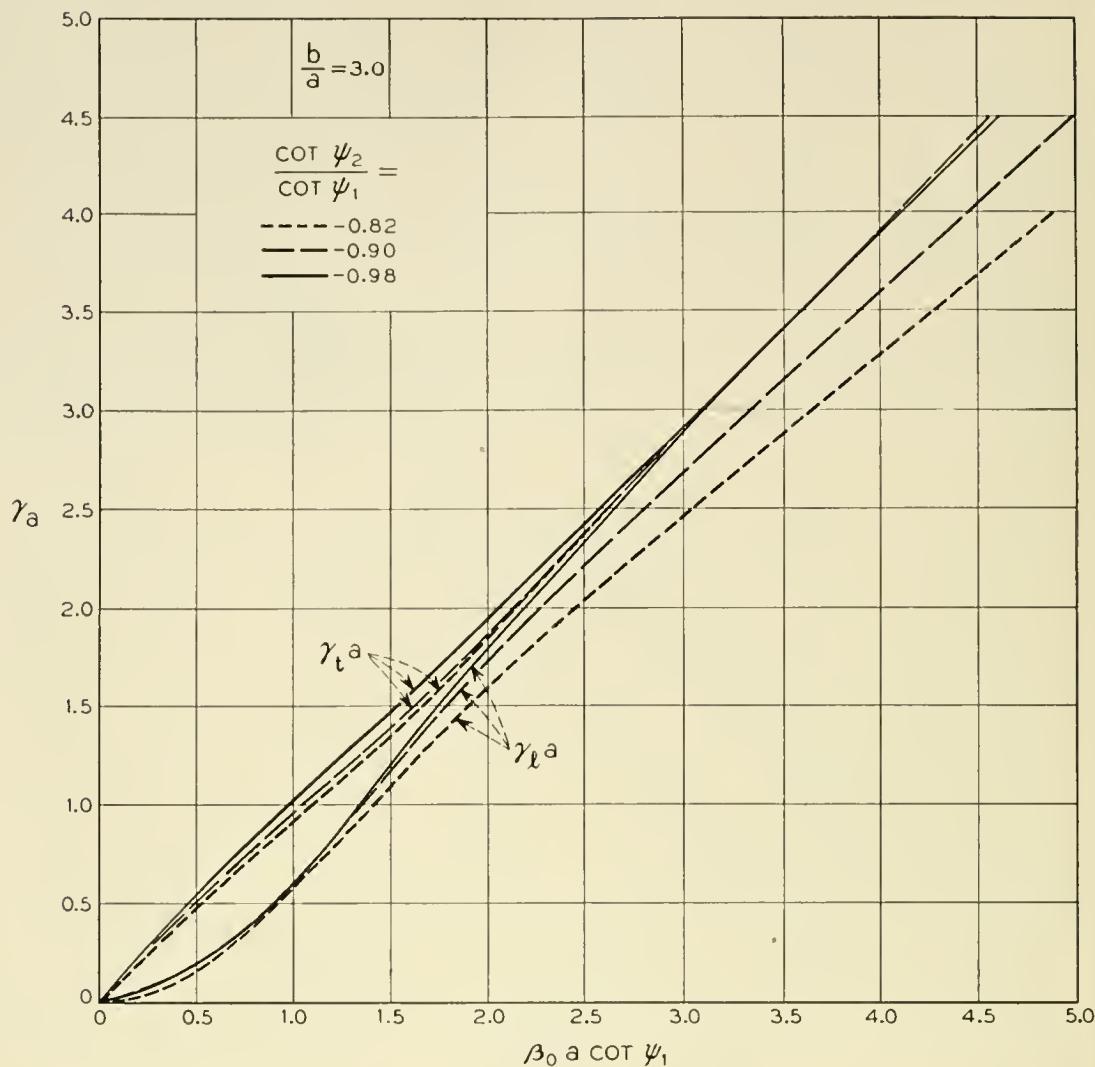


Fig. 4.1.5 — The radial propagation constants associated with the transverse and longitudinal modes on coupled coaxial sheath helices given as a function of $\beta_0 a \cot \psi_1$ when $b/a = 3.0$.

and for the outer helix

$$J_b = J_{b\ell} e^{-j\beta_\ell z} + J_{bt} e^{-j\beta_t z} \quad (36)$$

For complete power transfer we ask that

$$J_{b\ell} = J_{bt}$$

when J_a is zero at the input ($z = 0$)

$$J_{a\ell} = -J_{at}$$

or

$$\frac{J_{b\ell}}{J_{a\ell}} = -\frac{J_{bt}}{J_{at}} \quad (37)$$

Now $J_{a\ell}$ is equal to the discontinuity in the tangential component of magnetic field which can be written at $r = a$

$$J_{a\ell} = (H_{z3} \cos \psi_1 - H_{\varphi 5} \sin \psi_1) - (H_{z1} \cos \psi_1 - H_{\varphi 2} \sin \psi_1)$$

which can be written as

$$J_{a\ell} = -(H_{z1} - H_{z3})_{a\ell} (\cot \psi_1 + \tan \psi_1) \sin \psi_1 \quad (38)$$

and similarly at $r = b$

$$J_{b\ell} = -(H_{z7} - H_{z3})_{b\ell} (\cot \psi_2 + \tan \psi_2) \sin \psi_2 \quad (39)$$

Equations (38) and (39) can be combined with (37) to give as the condition for complete power transfer

$$A_\ell = -A_t \quad (40)$$

where

$$A = \frac{(I_{12}K_{02} + I_{02}K_{12}) \left(P_{01} - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} P_{11} \right)}{(I_{01}K_{11} + I_{11}K_{01}) \left(R_0 - \frac{(\beta_0 a \cot \psi_1)^2}{(\gamma a)^2} \frac{\cot \psi_2}{\cot \psi_1} R_1 \right)} \quad (41)$$

In (40) A_ℓ is obtained by substituting γ_ℓ into (41) and A_t is obtained by substituting γ_t into (41).

The value of $\cot \psi_2 / \cot \psi_1$ necessary to satisfy (40) is plotted in Fig. 2.8.

In addition to $\cot \psi_2 / \cot \psi_1$ it is necessary to determine the interference wavelength on the helices and this can be readily evaluated by considering (36) which can now be written

$$J_b = J_{b\ell} (e^{-j\beta_\ell z} + e^{-j\beta_t z})$$

or

$$J_b = J_{b\ell} e^{-j((\beta_\ell + \beta_t)z/2)} \cos \frac{(\beta_t - \beta_\ell)}{2} z \quad (48)$$

and

$$J_b = J_{b\ell} e^{-j((\beta_\ell + \beta_t)z/2)} \cos \frac{1}{2} \beta_b z \quad (49)$$

where we have defined

$$\beta_b a = (\gamma_t a - \gamma_\ell a) \quad (50)$$

This value of β_b is plotted versus $\beta_0 a \cot \psi_1$ in Fig. 2.4.

BIBLIOGRAPHY

1. J. R. Pierce, Traveling Wave Tubes, p. 44, Van Nostrand, 1950.
2. R. Kompfner, Experiments on Coupled Helices, A. E. R. E. Report No. G/M98, Sept., 1951.
3. R. Kompfner, Coupled Helices, paper presented at I. R. E. Electron Tube Conference, 1953, Stanford, Cal.
4. G. Wade and N. Rynn, Coupled Helices for Use in Traveling-Wave Tubes, I.R.E. Trans. on Electron Devices, Vol. ED-2, p. 15, July, 1955.
5. S. E. Miller, Coupled Wave Theory and Waveguide Applications, B.S.T.J., **33**, pp. 677-693, 1954.
6. M. Chodorow and E. L. Chu, The Propagation Properties of Cross-Wound Twin Helices Suitable for Traveling-Wave Tubes, paper presented at the Electron Tube Res. Conf., Stanford Univ., June, 1953.
7. G. M. Branch, A New Slow Wave Structure for Traveling-Wave Tubes, paper presented at the Electron Tube Res. Conf., Stanford Univ., June, 1953.
G. M. Branch, Experimental Observation of the Properties of Double Helix Traveling-Wave Tubes, paper presented at the Electron Tube Res. Conf., Univ. of Maine, June, 1954.
8. J. S. Cook, Tapered Velocity Couplers, B.S.T.J. **34**, p. 807, 1955.
9. A. G. Fox, Wave Coupling by Warped Normal Modes, B.S.T.J., **34**, p. 823, 1955.
10. W. H. Louisell, Analysis of the Single Tapered Mode Coupler, B.S.T.J., **34**, p. 853.
11. B. L. Humphreys, L. V. Kite, E. G. James, The Phase Velocity of Waves in a Double Helix, Report No. 9507, Research Lab. of G.E.C., England, Sept., 1948.
12. L. Stark, A Helical-Line Phase Shifter for Ultra-High Frequencies, Technical Report No. 59, Lincoln Laboratory, M.I.T., Feb., 1954.
13. P. D. Lacy, Helix Coupled Traveling-Wave Tube, Electronics, **27**, No. 11, Nov., 1954.

Statistical Techniques for Reducing the Experiment Time in Reliability Studies

By MILTON SOBEL

(Manuscript received September 19, 1955)

Given two or more processes, the units from which fail in accordance with an exponential or delayed exponential law, the problem is to select the particular process with the smallest failure rate. It is assumed that there is a common guarantee period of zero or positive duration during which no failures occur. This guarantee period may be known or unknown. It is desired to accomplish the above goal in as short a time as possible without invalidating certain predetermined probability specifications. Three statistical techniques are considered for reducing the average experiment time needed to reach a decision.

1. One technique is to increase the initial number of units put on test. This technique will substantially shorten the average experiment time. Its effect on the probability of a correct selection is generally negligible and in some cases there is no effect.

2. Another technique is to replace each failure immediately by a new unit from the same process. This replacement technique adds to the book-keeping of the test, but if any of the population variances is large (say in comparison with the guarantee period) then this technique will result in a substantial saving in the average experiment time.

3. A third technique is to use an appropriate sequential procedure. In many problems the sequential procedure results in a smaller average experiment time than the best non-sequential procedure regardless of the true failure rates. The amount of saving depends principally on the "distance" between the smallest and second smallest failure rates.

For the special case of two processes, tables are given to show the probability of a correct selection and the average experiment time for each of three types of procedures.

Numerical estimates of the relative efficiency of the procedures are given by computing the ratio of the average experiment time for two procedures of different type with the same initial sample size and satisfying the same probability specification.

INTRODUCTION

This paper is concerned with a study of the advantages and disadvantages of three statistical techniques for reducing the average duration of life tests. These techniques are:

1. Increasing the initial number of units on test.
2. Using a replacement technique.
3. Using a sequential procedure.

To show the advantages of each of these techniques, we shall consider the problem of deciding which of two processes has the smaller failure rate. Three different types of procedures for making this decision will be considered. They are:

- R_1 , A nonsequential, nonreplacement type of procedure
- R_2 , A nonsequential, replacement type of procedure
- R_3 , A sequential, replacement type of procedure

Within each type we will consider different values of n , the initial number of units on test for each process. The effect of replacement is shown by comparing the average experiment time for procedures of type 1 and 2 with the same value of n and comparable probabilities of a correct selection. The effect of using a sequential rule is shown by comparing the average experiment time for procedures of type 2 and 3 with the same value of n and comparable probabilities of a correct selection.

ASSUMPTIONS

1. It is assumed that failure is clearly defined and that failures are recognized without any chance of error.
2. The lifetime of individual units from either population is assumed to follow an exponential density of the form

$$f(x; \theta, g) = \frac{1}{\theta} e^{-(x-g)/\theta} \quad \text{for } x \geq g \quad (1)$$

$$f(x; \theta, g) = 0 \quad \text{for } x < g$$

where the location parameter $g \geq 0$ represents the common guarantee period and the scale parameter $\theta > 0$ represents the *unknown* parameter which distinguishes the two different processes. Let $\theta_1 \geq \theta_2$ denote the *ordered* values of the unknown parameter θ for the two processes; then the *ordered* failure rates are given by

$$\lambda_1 = 1/(\theta_1 + g) \leq \lambda_2 = 1/(\theta_2 + g) \quad (2)$$

3. It is not known which process has the parameter θ_1 and which has the parameter θ_2 .

4. The parameter g is assumed to be the same for both processes. It may be known or unknown.
5. The initial number n of units put on test is the same for both processes.
6. All units have independent lifetimes, i.e., the test environment is not such that the failure of one unit results in the failure of other units on test.
7. Replacements used in the test are assumed to come from the same population as the units they replace. If the replacement units have to sit on a shelf before being used then it is assumed that the replacements are not affected by shelf-aging.

CONCLUSIONS

1. Increasing the initial sample size n has at most a negligible effect on the probability of a correct selection. It has a substantial effect on the average experiment time for all three types of procedures. If the value of n is doubled, then the average time is reduced to a value less than or equal to half of its original value.
2. The technique of replacement always reduces the average experiment time. This reduction is substantial when $g = 0$ or when the population variance of either process is large compared to the value of g . This decrease in average experiment time must always be weighed against the disadvantage of an increase in bookkeeping and the necessity of having the replacement units available for use.
3. The sequential procedure enables the experimenter to make rational decisions as the evidence builds up without waiting for a predetermined number of failures. It has a shorter average experiment time than non-sequential procedures satisfying the same specification. This reduction brought about by the sequential procedure increases as the ratio α of the two failure rates increases. In addition the sequential procedure always terminates with a decision that is clearly convincing on the basis of the observed results, i.e., the *a posteriori* probability of a correct selection is always large at the termination of the experiment.

SPECIFICATION OF THE TEST

Each of the three types of procedures is set up so as to satisfy the same specification described below. Let α denote the true value of the ratio θ_1/θ_2 which by definition must be greater than, or equal to, one. It turns out that in each type of procedure the probability of a correct selection depends on θ_1 and θ_2 only through their ratio α .

1. The experimenter is asked to specify the smallest value of α (say it is $\alpha^* > 1$) that is worth detecting. Then the interval $(1, \alpha^*)$ represents a zone of indifference such that if the true ratio α lies therein then we would still like to make a correct selection, but the loss due to a wrong selection in this case is negligible.

2. The experimenter is also asked to specify the minimum value $P^* > \frac{1}{2}$ that he desires for the probability of a correct selection whenever $\alpha \geq \alpha^*$. In each type of procedure the rules are set up so that the probability of a correct selection for $\alpha = \alpha^*$ is as close to P^* as possible without being less than P^* .

The two constants $\alpha^* > 1$ and $\frac{1}{2} < P^* < 1$ are the only quantities specified by the experimenter. Together they make up the *specification* of the test procedure.

EFFICIENCY

If two procedures of different type have the same value of n and satisfy the same specification then we shall regard them as comparable and their relative efficiency will be measured by the ratio of their average experiment times. This ratio is a function of the true α but we shall consider it only for selected values of α , namely, $\alpha = 1$, $\alpha = \alpha^*$ and $\alpha = \infty$.

PROCEDURES OF TYPE R_1 — NONSEQUENTIAL, NONREPLACEMENT

"The same number n of units are put on test for each of the two processes. Experimentation is continued until either one of the two samples produces a predetermined number r ($r \leq n$) of failures. Experimentation is then stopped and the process with fewer than r failures is chosen to be the better one."

TABLE I — PROBABILITY OF A CORRECT SELECTION — PROCEDURE TYPE R_1

($\alpha = 2$, any $g \geq 0$, to be used to obtain r for $\alpha^* = 2$)

n	$r = 1$	$r = 2$	$r = 3$	$r = 4$
1	0.667	—	—	—
2	0.667	0.733	—	—
3	0.667	0.738	0.774	—
4	0.667	0.739	0.784	0.802
10	0.667	0.741	0.789	0.825
20	0.667	0.741	0.790	0.826
∞	0.667	0.741	0.790	0.827

Note: The value for $r = 0$ is obviously 0.500 for any n .

We shall assume that the number n of units put on test is determined by non-statistical considerations such as the availability of units, the availability of sockets, etc. Then the only unspecified number in the above procedure is the integer r . This can be determined from a table of probabilities of a correct selection to satisfy any given specification (α^*, P^*) . If, for example, $\alpha^* = 2$ then we can enter Table I. If n is given to be 4 and we wish to meet the specification $\alpha^* = 2, P^* = 0.800$ then we would enter Table I with $n = 4$ and select $r = 4$, it being the smallest value for which $P \geq P^*$.

The table above shows that for the given specification we would also have selected $r = 4$ for any value of n . In fact, we note that the probability of a correct selection depends only slightly on n . The given value of n and the selected value of r then determine a particular procedure of type R_1 , say, $R_1(n, r)$.

The average experiment time for each of several procedures $R_1(n, r)$ is given in Table II for the three critical values of the true ratio α , namely, $\alpha = 1, \alpha = \alpha^*$ and $\alpha = \infty$. Each of the entries has to be multiplied by θ_2 , the smaller of the two θ values, and added to the common guarantee period g . For $n = \infty$ the entry should be zero ($+g$) but it was found convenient to put in place of zero the leading term in the asymptotic expansion of the expectation in powers of $1/n$. Hence the entry for $n = \infty$ can be used for any large n , say, $n \geq 25$ when $r \leq 4$.

We note in Table II the undesirable feature that for each procedure the average experiment time increases with α for fixed θ_2 . For the sequential procedure we shall see later that the average experiment time is greater at $\alpha = \alpha^*$ than at either $\alpha = 1$ or $\alpha = \infty$. This is intuitively more desirable since it means that the procedure spends more time when the choice is more difficult to make and less time when we are indifferent or when the choice is easy to make.

PROCEDURES OF TYPE R_2 — NONSEQUENTIAL, REPLACEMENT

"Such procedures are carried out exactly as for procedures of R_1 except that failures are immediately replaced by new units from the same population."

To determine the appropriate value of r for the specification $\alpha^* = 2, P^* = 0.800$ when $g = 0$ we use the last row of Table I, i.e., the row marked $n = \infty$, and select $r = 4$. The probability of a correct selection for procedures of type R_2 is exactly the same for all values of n and depends only on r . Furthermore, it agrees with the probability for procedures of type R_1 with $n = \infty$ so that it is not necessary to prepare a separate table.

TABLE II — AVERAGE EXPERIMENT TIME — PROCEDURE TYPE R_1
 (Multiply entry by θ_2 and add g)

n	$r = 1$				$r = 2$				$r = 3$				$r = 4$			
	$\alpha = 1$		$\alpha = 2$		$\alpha = 1$		$\alpha = 2$		$\alpha = 1$		$\alpha = 2$		$\alpha = \infty$		$\alpha = 1$	
	$\alpha = 1$	$\alpha = \infty$	$\alpha = 1$	$\alpha = \infty$	$\alpha = 1$	$\alpha = \infty$										
1	0.500	0.667	1.000	1.500	0.917	1.200	1.500	1.833	1.217	1.572	1.833	1.833	—	—	—	—
2	0.250	0.333	0.500	0.500	0.517	0.675	0.833	0.944	0.735	0.944	1.083	1.449	1.854	—	—	—
3	0.167	0.222	0.333	0.333	0.363	0.474	0.583	0.735	0.211	0.231	0.297	0.336	0.347	1.449	2.083	—
4	0.125	0.167	0.250	0.250	0.100	0.132	0.172	0.211	0.211	0.231	0.297	0.336	0.347	0.439	0.479	—
10	0.050	0.067	0.067	0.067	0.050	0.064	0.084	0.103	0.109	0.139	0.158	0.157	0.157	0.200	0.217	—
20	0.025	0.033	0.050	0.064	0.050	0.064	0.084	0.103	0.109	0.139	0.158	0.157	0.157	0.200	0.217	—
∞	0.500/n	0.667/n	1.000/n	1.250/n	1.630/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n	2.000/n

TABLE III—VALUE OF r REQUIRED TO MEET THE SPECIFICATION
 (α^*, P^*) FOR PROCEDURES OF TYPE R_2 ($g = 0$)

P^*	α^*												
	1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50	2.00	2.50	3.00
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0
0.55	14	4	2	2	1	1	1	1	1	1	1	1	1
0.60	55	15	7	5	3	3	2	2	2	1	1	1	1
0.65	126	33	16	10	7	5	4	3	3	3	1	1	1
0.70	232	61	29	17	12	9	7	6	5	4	2	1	1
0.75	383	101	47	28	19	14	11	9	7	6	3	2	1
0.80	596	157	73	43	29	21	17	13	11	9	4	2	2
0.85	903	238	111	65	44	32	25	20	16	14	5	3	3
0.90	1381	363	169	100	67	49	37	30	25	21	8	5	4
0.95	2274	597	278	164	110	80	61	49	40	34	12	7	5
0.99	4549	1193	556	327	219	160	122	98	80	68	24	14	10

It is also unnecessary to prepare a separate table for the average experiment time for procedures of type R_2 since for $g = 0$ the exact values can be obtained by substituting the appropriate value of n in the expressions appearing in Table II in the row marked $n = \infty$. For example, for $n = 2$, $r = 1$ and $\alpha = 1$ the exact value for $g = 0$ is $0.500 \theta_2/2 = 0.250 \theta_2$, and for $n = 3$, $r = 4$, $\alpha = \infty$ the exact value for $g = 0$ is $4.000 \theta_2/3 = 1.333 \theta_2$. It should be noted that for procedures of type R_2 we need not restrict our attention to the cases $r \leq n$ but can also consider $r > n$.

Table III shows the value of r required to meet the specification (α^*, P^*) with a procedure of type R_2 for various selected values of α^* and P^* .

PROCEDURES OF TYPE R_3 —SEQUENTIAL, REPLACEMENT

Let $D(t)$ denote the absolute difference between the number of failures produced by the two processes at any time t . The sequential procedure is as follows:

“Stop the test as soon as the inequality

$$D(t) \geq \frac{\ln [P^*/(1 - P^*)]}{\ln \alpha^*} \quad (3)$$

is satisfied. Then select the population with the smaller number of failures as the better one.”

To get the best results we will choose (α^*, P^*) so that the right hand member of the inequality (3) is an integer. Otherwise we would be operating with a higher value of P^* (or a smaller value of α^*) than was specified.

TABLE IV—AVERAGE EXPERIMENT TIME AND PROBABILITY OF A
CORRECT SELECTION—PROCEDURE TYPE R_3
($\alpha^* = 2, P^* = 0.800, g = 0$)
(Multiply each average time entry by θ_2)

n	$\alpha = 1$	$\alpha = 2$	$\alpha = \infty$
1	2.000	2.400	2.000
2	1.000	1.200	1.000
3	0.667	0.800	0.667
4	0.500	0.600	0.500
10	0.200	0.240	0.200
20	0.100	0.120	0.100
∞	$2.000/n$	$2.400/n$	$2.000/n$
Probability	0.500	0.800	1.000

For example, we might choose $\alpha^* = 2$ and $P^* = 0.800$. For procedures of type R_3 the probability of a correct selection is again completely independent of n ; here it depends only on the true value of the ratio α . The average experiment time depends strongly on n and only to a limited extent on the true value of the ratio α . Table IV gives these quantities for $\alpha = 1, \alpha = 2$, and $\alpha = \infty$ for the particular specification $\alpha^* = 2, P^* = 0.800$ and for the particular value $g = 0$.

EFFICIENCY

We are now in a position to compare the efficiency of two different types of procedures using the same value of n . The efficiency of R_1 relative to R_2 is the reciprocal of the ratio of their average experiment time. This is given in Table V for $\alpha^* = 2, P^* = 0.800, r = 4$ and $n = 4, 10, 20$ and ∞ . By Table I the value $P^* = 0.800$ is not attained for $n < 4$.

In comparing the sequential and the nonsequential procedures it was found that the slight excesses in the last column of Table I over 0.800

TABLE V—EFFICIENCY OF TYPE R_1 RELATIVE TO
TYPE R_2
($\alpha^* = 2, P^* = 0.800, r = 4, g = 0$)

n	$\alpha = 1$	$\alpha = 2$	$\alpha = \infty$
4	0.501	0.495	0.480
10	0.837	0.836	0.835
20	0.925	0.917	0.922
∞	1.000	1.000	1.000

TABLE VI—EFFICIENCY OF ADJUSTED R_1 RELATIVE TO R_3
 $(\alpha^* = 2, P^* = 0.800, g = 0)$

n	$\alpha = 1$	$\alpha = 2$	$\alpha = \infty$
4	0.615	0.575	0.419
10	0.754	0.708	0.528
20	0.818	0.768	0.573
∞	0.873	0.822	0.612

had an effect on the efficiency. To make the procedures more comparable the values for $r = 3$ and $r = 4$ in Table I were averaged with values p and $1 - p$ computed so as to give a probability of *exactly* 0.800 at $\alpha = \alpha^*$. The corresponding values for the average experiment time were then averaged with the same values p and $1 - p$. The nonsequential procedures so altered will be called "adjusted procedures." The efficiency of the adjusted R_1 relative to R_3 is given in Table VI.

In Table VI the last row gives the efficiency of the adjusted procedure R_2 relative to R_3 . Thus we can separate out the advantage due to the replacement feature and the advantage due to the sequential feature. Table VII gives these results in terms of percentage reduction of average experiment time.

We note that the reduction due to the replacement feature alone is greatest for small n and essentially constant with α while the reduction

TABLE VII—PER CENT REDUCTION IN AVERAGE EXPERIMENT TIME
DUE TO STATISTICAL TECHNIQUES
 $(\alpha^* = 2, P^* = 0.800, g = 0)$

α	n	Reduction due to Replacement Feature Alone	Reduction due to Sequential Feature Alone	Reduction due to both Replacement and Sequential Features
1	4	29.5	12.7	38.5
	10	13.7	12.7	24.6
	20	6.3	12.7	18.2
	∞	0.0	12.7	12.7
2	4	30.1	17.8	42.5
	10	13.9	17.8	29.2
	20	6.6	17.8	23.2
	∞	0.0	17.8	17.8
∞	4	31.5	38.8	58.1
	10	13.6	38.8	47.2
	20	6.3	38.8	42.7
	∞	0.0	38.8	38.8

due to the sequential feature alone is greatest for large α and is independent of n . Hence if the initial sample size per process n is large we can disregard the replacement technique. On the other hand the true value of α is not known and hence the advantage of sequential experimentation should not be disregarded.

The formulas used to compute the accompanying tables are given in Addendum 2.

ACKNOWLEDGEMENT

The author wishes to thank Miss Marilyn J. Huyett for considerable help in computing the tables in this paper. Thanks are also due to J. W. Tukey and other staff members for constructive criticism and numerical errors they have pointed out.

ADDENDUM 1

In this addendum we shall consider the more general problem of selecting the best of k exponential populations treated on a higher mathematical level. For $k = 2$ this reduces to the problem discussed above.

DEFINITIONS AND ASSUMPTIONS

There are given k populations Π_i ($i = 1, 2, \dots, k$) such that the lifetimes of units taken from any of these populations are independent chance variables with the exponential density (1) with a common (known or unknown) location parameter $g \geq 0$. The distributions for the k populations are identical except for the unknown scale parameter $\theta > 0$ which may be different for the k different populations. We shall consider three different cases with regard to g .

- Case 1: The parameter g has the value zero ($g = 0$).
- Case 2: The parameter g has a positive, known value ($g > 0$).
- Case 3: The parameter g is unknown ($g \geq 0$).

Let the *ordered values* of the k scale parameters be denoted by

$$\theta_1 \geq \theta_2 \geq \dots \geq \theta_k \quad (4)$$

where equal values may be regarded as ordered in any arbitrary manner. At any time t each population has a certain number of failures associated with it. Let the *ordered values* of these integers be denoted by $r_i = r_i(t)$ so that

$$r_1 \leq r_2 \leq \dots \leq r_k \quad (5)$$

For each unit the life beyond its guarantee period will be referred to as its Poisson life. Let $L_i(t)$ denote the total amount of Poisson life observed up to time t in the population with r_i failures ($i = 1, 2, \dots, k$). If two or more of the r_i are equal, say $r_i = r_{i+1} = \dots = r_{i+j}$, then we shall assign r_i and L_i to the population with the largest Poisson life, r_{i+1} and L_{i+1} to the population with the next largest, \dots, r_{i+j} and L_{i+j} to the population with the smallest Poisson life. If there are two or more *equal pairs* (r_i, L_i) then these should be ordered by a random device giving equal probability to each ordering. Then the subscripts in (5) as well as those in (4) are in one-to-one correspondence with the k given populations. It should be noted that $L_i(t) \geq 0$ for all i and any time $t \geq 0$. The complete set of quantities $L_i(t)$ ($i = 1, 2, \dots, k$) need not be ordered. Let $\alpha = \theta_1/\theta_2$ so that, since the θ_i are ordered, $\alpha \geq 1$.

We shall further assume that:

1. The initial number n of units put on test is the same and the starting time is the same for each of the k populations.
2. Each replacement is assumed to be a new unit from the same population as the failure that it replaces.
3. Failures are assumed to be clearly recognizable without any chance of error.

SPECIFICATIONS FOR CASE 1: $g = 0$

Before experimentation starts the experimenter is asked to specify two constants α^* and P^* such that $\alpha^* > 1$ and $\frac{1}{2} < P^* < 1$. The procedure $R_3 = R_3(n)$, which is defined in terms of the specified α^* and P^* , has the property that it will correctly select the population with the largest scale parameter with probability at least P^* whenever $\alpha \geq \alpha^*$. The initial number n of units put on test may either be fixed by nonstatistical considerations or may be determined by placing some restriction on the average experiment time function.

Rule R_3 :

“Continue experimentation with replacement until the inequality

$$\sum_{i=2}^k \alpha^{*(r_i-r_1)} \leq (1 - P^*)/P^* \quad (6)$$

is satisfied. Then stop and select the population with the smallest number of failures as the one having the largest scale parameter.”

Remarks

1. Since $P^* > \frac{1}{2}$ then $(1 - P^*)/P^* < 1$ and hence no two populations can have the same value r_1 at stopping time.
2. For $k = 2$ the inequality (6) reduces to the inequality (3).
3. The procedure R_3 terminates only at a failure time, never between failures, since the left member of (6) depends on t only through the quantities $r_i(t)$.
4. After experimentation is completed one can make, at the $100P$ per cent confidence level, the *confidence statement*

$$\theta_s \leq \theta_1 \leq \alpha^* \theta_s \quad (\text{or } \theta_1/\alpha^* \leq \theta_s \leq \theta_1) \quad (7)$$

where θ_s is the scale parameter of the selected population.

Numerical Illustrations

Suppose the preassigned constants are $P^* = 0.95$ and $\alpha^* = 19^{1/4} = 2.088$ so that $(1 - P^*)/P^* = \frac{1}{19}$. Then for $k = 2$ the procedure is to stop when $r_2 - r_1 \geq 4$. For $k = 3$ it is easy to check that the procedure reduces to the simple form: "Stop when $r_2 - r_1 \geq 5$ ". For $k > 3$ either calculations can be carried out as experimentation progresses or a table of stopping values can be constructed before experimentation starts. For $k = 4$ and $k = 5$ see Table VIII.

In the above form the proposed rule is to stop when, for at least one

TABLE VIII — SEQUENTIAL RULE FOR $P^* = 0.95$, $\alpha^* = 19^{1/4}$

$k = 4$

$r_2 - r_1$	$r_3 - r_1$	$r_4 - r_1$	
5	5	9	
5	6	6	
6	6	6	*

$k = 5$

$r_2 - r_1$	$r_3 - r_1$	$r_4 - r_1$	$r_5 - r_1$	
5	5	9	10	
5	5	10	10	*
5	6	6	8	
5	6	7	7	
5	7	7	7	*
6	6	6	6	

* Starred rows can be omitted without affecting the test since every integer in these rows is at least as great as the corresponding integer in the previous row. They are shown here to illustrate a systematic method which insures that all the necessary rows are included.

row (say row j) in the table, the observed row vector ($r_2 - r_1$, $r_3 - r_1, \dots, r_k - r_1$) is such that each component is at least as large as the corresponding component of row j .

Properties of R_3 for $k = 2$ and $g = 0$

For $k = 2$ and $g = 0$ the procedure R_3 is an example of a Sequential Probability Ratio test as defined by A. Wald in his book.⁵ The Average Sample Number (ASN) function and the Operating Characteristics (OC) function for R_3 can be obtained from the general formulae given by Wald. Both of these functions depend on θ_1 and θ_2 only through their ratio α . In our problem there is no excess over the boundary and hence Wald's approximation formulas are *exact*. When our problem is put into the Wald framework, the symmetry of our problem implies equal probabilities of type 1 and type 2 errors. The OC function takes on complementary values for any point $\alpha = \theta_1/\theta_2$ and its reciprocal θ_2/θ_1 . We shall therefore compute it only for $\alpha \geq 1$ and denote it by $P(\alpha)$. For $\alpha > 1$ the quantity $P(\alpha)$ denotes the probability of a correct selection for the true ratio α .

The equation determining Wald's h function⁵ is

$$\frac{(\alpha^*)^h}{1 + \alpha} + \frac{\alpha(\alpha^*)^{-h}}{1 + \alpha} = 1 \quad (8)$$

for which the non-zero solution in h is easily computed to be

$$h(\alpha) = \frac{\ln \alpha}{\ln \alpha^*} \quad (9)$$

Hence we obtain from Wald's formula (3:43) in Reference 5

$$P(\alpha) = \frac{\alpha^s}{\alpha^s + 1} \quad (10)$$

where s is the smallest integer greater than or equal to

$$S = \ln [P^*/(1 - P^*)]/\ln \alpha^* \quad (11)$$

In particular, for $\alpha = 1^+$, α^* and ∞ we have

$$P(1^+) = \frac{1}{2}, \quad P(\alpha^*) \geq P^*, \quad P(\infty) = 1 \quad (12)$$

We have written $P(1^+)$ above for $\lim P(x)$ as $x \rightarrow 1$ from the right. The procedure becomes more efficient if we choose P and α^* so that S is an integer. Then $s = S$ and $P(\alpha^*) = P^*$.

Letting F denote the total number of observed failures required to

terminate the experiment we obtain for the ASN function

$$E(F; \alpha) = s \left(\frac{\alpha + 1}{\alpha - 1} \right) \left(\frac{\alpha^s - 1}{\alpha^s + 1} \right) \quad \text{for } \alpha > 1 \quad (13)$$

and, in particular, for $\alpha = 1, \infty$

$$E(F; 1) = s^2 \quad \text{and} \quad E(F; \infty) = s \quad (14)$$

It is interesting to note that for $s = 1$ we obtain

$$E(F; \alpha) = 1 \quad \text{for all } \alpha \geq 1 \quad (15)$$

and that this result is exact since for $s = 1$ the right-hand member S of (3) is at most one and hence the procedure terminates with certainty immediately after the first failure.

As a result of the exponential assumption, the assumption of replacement and the assumption that $g = 0$ it follows that the intervals between failures are independently and identically distributed. For a single population the time interval between failures is an exponential chance variable. Hence, for two populations, the time interval is the minimum of two exponentials which is again exponential. Letting τ denote the (chance) duration of a typical interval and letting T denote the (chance) total time needed to terminate the procedure, we have

$$E(T; \alpha, \theta_2) = E(F; \alpha)E(\tau; \alpha, \theta_2) = E(F; \alpha) \left(\frac{\theta_2}{n} \right) \left(\frac{\alpha}{1 + \alpha} \right) \quad (16)$$

Hence we obtain from (13) and (14)

$$E(T; \alpha, \theta_2) = \frac{\theta_2}{n} \frac{s\alpha}{\alpha - 1} \frac{\alpha^s - 1}{\alpha^s + 1} \quad \text{for } \alpha > 1 \quad (17)$$

$$E(T; 1, \theta_2) = \frac{\theta_2 s^2}{2n} \quad \text{and} \quad E(T; \infty, \theta_2) = \frac{\theta_2 s}{n} \quad (18)$$

For the numerical illustration treated above with $k = 2$ we have

$$P(\alpha) = \frac{\alpha^4}{1 + \alpha^4} \quad (19)$$

$$P(1^+) = \frac{1}{2}; \quad P(2.088) = 0.95; \quad P(\infty) = 1 \quad (20)$$

$$E(F; \alpha) = 4 \frac{\alpha + 1}{\alpha - 1} \frac{\alpha^4 - 1}{\alpha^4 + 1} = 4 \frac{(\alpha + 1)^2(\alpha^2 + 1)}{\alpha^4 + 1} \quad (21)$$

$$E(F; 1) = 16.0; \quad E(F; 2.088) = 10.2; \quad E(F; \infty) = 4 \quad (22)$$

$$\begin{aligned} E(T; 1, \theta_2) &= \frac{8\theta_2}{n}; & E(T; 2.088, \theta_2) &= \frac{6.9\theta_2}{n}; \\ E(T; \infty, \theta_2) &= \frac{4\theta_2}{n} \end{aligned} \quad (23)$$

For $k > 2$ the proposed procedure is an application of a general sequential rule for selecting the best of k populations which is treated in [1]. Proof that the probability specification is met and bounds on the probability of a correct decision can be found there.

CASE 2: COMMON KNOWN $g > 0$

In order to obtain the properties of the sequential procedure R_3 for this case it will be convenient to consider other sequential procedures. Let $\beta = 1/\theta_2 - 1/\theta_1$ so that, since the θ_i are ordered, $\beta \geq 0$. Let us assume that the experimenter can specify three constants α^* , β^* and P^* such that $\alpha^* > 1$, $\beta^* > 0$ and $\frac{1}{2} < P^* < 1$ and a procedure is desired which will select the population with the largest scale parameter with probability at least P^* whenever we have both

$$\alpha \geq \alpha^* \text{ and } \beta \geq \beta^*$$

The following procedure meets this specification.

Rule R_3' :

“Continue experimentation with replacement until the inequality

$$\sum_{i=2}^k \alpha^{*(r_i - r_1)} e^{-\beta^*(L_1 - L_i)} \leq (1 - P^*)/P^* \quad (24)$$

is satisfied. Then stop and select the population with the smallest number of failures as the one having the largest scale parameter. If, at stopping time, two or more populations have the same value r_1 then select that particular one of these with the largest Poisson life L_1 .”

Remarks

1. For $k = 2$ the inequality reduces to

$$(r_2 - r_1) \ln \alpha^* + (L_1 - L_2) \beta^* \geq \ln [P^*/(1 - P^*)] \quad (25)$$

If $g = 0$ then $L_1 \equiv L_i$ for all t and the procedure R_3' reduces to R_3 .

2. The procedure R_3' may terminate not only at failures but also between failures.

3. The same inequality (24) can also be used if experimentation is carried on *without replacement*, one advantage of the latter being that there is less bookkeeping involved. In this case there is a possibility that the units will all fail before the inequality is satisfied so that the procedure is not yet completely defined for this case. One possibility in such a situation is to continue experimentation with new units from each population until the inequality is satisfied. Such a procedure will terminate in a finite time with probability one, i.e., $\text{Prob}\{T > T_0\} \rightarrow 0$ as $T_0 \rightarrow \infty$, and the probability specification will be satisfied.

4. A procedure $R_3' (n_1, n_2, \dots, n_k, t_1, t_2, \dots, t_k)$ using the same inequality (24) but based on different initial sample sizes and/or on different starting times for the initial samples also satisfies the above probability specification. In the case of different starting times it is required that the experimenter wait at least g units of time after the last initial sample is put on test before reaching any decision.

5. One disadvantage of R_3' is that there is some (however remote) possibility of terminating while $r_1 = r_2$. This can be avoided by adding the condition $r_2 > r_1$ to (24) but, of course, the average experiment time is increased. Another way of avoiding this is to use the procedure R_3 which depends only on the number of failures; the effect of using R_3 when $g > 0$ will be considered below.

6. The terms of the sum in (24) represent likelihood ratios. If at any time each term is less than unity then we shall regard the decision to select the population with r_1 failures and L_1 units of Poisson life as optimal. Since $(1 - P^*)/P^* < 1$ then each term must be less than unity at termination.

Properties of Procedure R_3' for $k = 2$

The OC and ASN functions for R_3' will be approximated by comparing R_3' with another procedure R_3'' defined below. We shall assume that P^* is close to unity and that g is small enough (compared to θ_2) so that the probability of obtaining two failures within g units of time is small enough to be negligible. Then we can write approximately at termination

$$L_i \cong nT - r_i g \quad (i = 1, 2, \dots, k) \quad (26)$$

and

$$L_1 - L_i \cong (r_i - r_1)g \quad (i = 2, 3, \dots, k) \quad (27)$$

Substituting this in (24) and letting

$$\delta^* = \alpha^* e^{\beta^* g} \quad (28)$$

suggests a new rule, say R_3'' , which we now define.

Rule R₃"

"Continue experimentation with replacement until the inequality

$$\sum_{i=2}^k \delta^{*-r_i-r_1} \leq (1 - P^*)/P^* \quad (29)$$

is satisfied. Then stop and select the population with r_1 failures as the one with the largest scale parameter."

For rule $R_3"$ the experimenter need only specify P^* and the smallest value δ^* of the *single* parameter

$$\delta = \frac{\theta_1}{\theta_2} e^{g((1/\theta_2) - (1/\theta_1))} = \alpha e^{g\beta} \quad (30)$$

that he desires to detect with probability at least P^* .

We shall approximate the OC and ASN function of $R_3"$ for $k = 2$ by computing them under the assumption that (27) holds at termination. The results will be considered as an approximation for the OC and ASN functions respectively of R_3' for $k = 2$. The similarity of (29) and (6) immediately suggests that we might replace α^* by δ^* and α by δ in the formulae for (6). To use the resulting expressions for R_3' we would compute δ^* as a function of α^* and β^* by (28) and δ as a function of α and β by (30).

The similarity of (29) and (6) shows that Z_n (defined in Reference 5, page 170) under (27) with $g > 0$ is the same function of δ^* and δ as it is of α^* and α when $g = 0$. To complete the justification of the above result it is sufficient to show that the individual increment z of Z_n is the same function of δ^* and δ under (27) with $g > 0$ as it is of α^* and α when $g = 0$. To keep the increments independent it is necessary to associate each failure with the Poisson life that follows rather than with the Poisson life that precedes the failure. Neglecting the probability that any two failures occur within g units of time we have two values for z , namely

$$z = \log \frac{\frac{n}{\theta_1} e^{-(nt-g)/\theta_1} e^{-nt/\theta_2}}{\frac{n}{\theta_2} e^{-(nt-g)/\theta_2} e^{-nt/\theta_1}} = -\log \delta \quad (31)$$

and, interchanging θ_1 and θ_2 , gives $z = \log \delta$. Moreover

$$\text{Prob } \{z = -\log \delta\} = \frac{\int_g^\infty \int_g^\infty \frac{n}{\theta_2} e^{-(nx-g)/\theta_2} e^{-ny/\theta_1} dx dy}{\frac{\theta_2}{n} e^{-g[\theta_2(n-1)+\theta_1 n]/\theta_1 \theta_2} + \frac{\theta_1}{n} e^{-g[\theta_2 n+\theta_1(n-1)]/\theta_1 \theta_2}} \quad (32)$$

$$= \frac{\delta}{1 + \delta}$$

Thus the OC and ASN functions under (27) with $g > 0$ bear the same relation to δ^* and δ as they do to α^* and α when $g = 0$. Hence, letting w denote the smallest integer greater than or equal to

$$W = \frac{\ln [P^*/(1 - P^*)]}{\ln \delta^*} = \frac{\ln [P^*/(1 - P^*)]}{g\beta^* + \ln \alpha^*} \quad (33)$$

we can write (omitting P^* in the rule description)

$$P\{\delta; R_3'(\alpha^*, \beta^*)\} \cong P\{\delta; R_3''(\delta^*)\} \cong \frac{\delta^w}{\delta^w + 1} \quad (34)$$

$$E\{F; R_3'(\alpha^*, \beta^*)\} \cong E\{F; R_3''(\delta^*)\}$$

$$\cong \begin{cases} w \left(\frac{\delta + 1}{\delta - 1} \right) \left(\frac{\delta^w - 1}{\delta^w + 1} \right) & \text{for } \delta > 1 \\ w^2 & \text{for } \delta = 1 \end{cases} \quad (35)$$

We can approximate the average time between failures by

$$E\{\tau; \theta_1, \theta_2, g\} \cong \frac{(g + \theta_1)(g + \theta_2)}{n(\theta_1 + \theta_2 + 2g)} \leq g + \frac{\theta_2}{n} \left(\frac{\alpha}{1 + \alpha} \right) \quad (36)$$

and the average experiment time by

$$E\{T; R_3'(\alpha^*, \beta^*)\} \cong E\{F; R_3'(\alpha^*, \beta^*)\} \frac{(g + \theta_1)(g + \theta_2)}{n(\theta_1 + \theta_2 + 2g)} \quad (37)$$

Since $\delta \geq 1$ then $\delta^w/(1 + \delta^w)$ is an increasing function of w and by (33) it is a non-increasing function of δ^* . By (28) $\delta^* \geq \alpha^*$ and hence, if we disregard the approximation (34),

$$P\{\delta; R_3''(\alpha^*)\} \cong \frac{[P^*/(1 - P^*)]^{\ln \delta / \ln \alpha^*}}{1 + [P^*/(1 - P^*)]^{\ln \delta / \ln \alpha^*}} \geq P\{\delta; R_3''(\delta^*)\} \quad (38)$$

Clearly the rules $R_3(\alpha^*, P^*)$ and $R_3''(\alpha^*, P^*)$ are equivalent so that for $g > 0$ we have

$$P\{\delta; R_3(\alpha^*)\} \equiv P\{\delta; R_3''(\alpha^*)\} \quad (39)$$

and hence, in particular, letting $\delta = \delta^*$ in (38) we have

$$P\{\delta^*; R_3(\alpha^*)\} \geq P\{\delta^*; R_3''(\delta^*)\} \geq P^* \quad (40)$$

since the right member of (34) reduces to P^* when W is an integer and $\delta = \delta^*$. The error in the approximations above can be disregarded when g is small compared to θ_2 . Thus we have shown that for small values of g/θ_2 the probability specification based on (α^*, β^*, P^*) is satisfied in the sense of (40) if we use the procedure $R_3(\alpha^*, P^*)$, i.e., if we proceed as if $g = 0$.

It would be desirable to show that we can proceed as if $g = 0$ for all values of g and P^* . It can be shown that for sufficiently large n the rule $R_3(\alpha^*, P^*)$ meets its specification for all g . One effect of increasing n is to decrease the average time $E(\tau)$ between failures and to approach the corresponding problem without replacement since $g/E(\tau)$ becomes large. Hence we need only show that $R_3(\alpha^*, P^*)$ meets its specification for the corresponding problem without replacement. If we disregard the information furnished by Poisson life and rely solely on the counting of failures then the problem reduces to testing in a single binomial whether $\theta = \theta_1$ for population Π_1 and $\theta = \theta_2$ for population Π_2 or vice versa. Letting p denote the probability that the next failure arises from Π_1 then we have formally

$$H_0:p = \frac{1}{1 + \alpha} \quad \text{versus} \quad H_1:p = \frac{\alpha}{1 + \alpha}$$

For preassigned constants $\alpha^* > 1$ and P^* ($\frac{1}{2} < P^* < 1$) the appropriate sequential likelihood test to meet the specification:

"Probability of a Correct Selection $\geq P^*$ whenever $\alpha \geq \alpha^{**}$ " (41) then turns out to be precisely the procedure $R_3(\alpha^*, P^*)$. Hence we may proceed as if $g = 0$ when n is sufficiently large.

The specifications of the problem may be given in a different form. Suppose $\theta_1^* > \theta_2^*$ are specified and it is desired to have a probability of a correct selection of at least P^* whenever $\theta_1 \geq \theta_1^* > \theta_2^* \geq \theta_2$. Then we can form the following sequential likelihood procedure R_3^* which is more efficient than $R_3(\alpha^*, P^*)$.

Rule R_3^ :*

"Continue experimentation without replacement until a time t is reached at which the inequality

$$\sum_{i=2}^k \left[\frac{e^{t/\theta_2^*} - 1}{e^{t/\theta_1^*} - 1} \right]^{-(r_i - r_1)} \leq \frac{1 - P^*}{P^*} \quad (42)$$

is satisfied. Then stop and select the population with r_1 failures as the population with $\theta = \theta_1^*$.

It can be easily shown that the greatest lower bound of the bracketed quantity in (42) is θ_1^*/θ_2^* . Hence for $\theta_1^*/\theta_2^* = \alpha^*$ and $P^* > \frac{1}{2}$ the time required by $R_3^*(\theta_1^*, \theta_2^*, P^*)$ will always be less than the time required by $R_3(\alpha^*, P^*)$.

Another type of problem is one in which we are given that $\theta = \theta_1^*$ for one population and $\theta = \theta_2^*$ for the $k - 1$ others where $\theta_1^* > \theta_2^*$ are specified. The problem is to select the population with $\theta = \theta_1^*$. Then (42) can again be used. In this case the parameter space is discrete with k points only one of which is correct. If Rule R_3^* is used then the probability of selecting the correct point is at least P^* .

Equilibrium Approach When Failures Are Replaced

Consider first the case in which all items on test are from the same exponential population with parameters (θ, g) . Let T_{nj} denote the length of the time interval between the j^{th} and the $j + 1^{\text{st}}$ failures, ($j = 0, 1, \dots$), where n is the number of items on test and the 0^{th} failure denotes the starting time. As time increases to infinity the expected number of failures per unit time clearly approaches $n/(\theta + g)$ which is called the equilibrium failure rate. The inverse of this is the expected time between failures at equilibrium, say $E(T_{n\infty})$. The question as to how the quantities $E(T_{nj})$ approach $E(T_{n\infty})$ is of considerable interest in its own right. The following results hold for any fixed integer $n \geq 1$ unless explicitly stated otherwise. It is easy to see that

$$E(T_{n1}) \leqq E(T_{n\infty}) \leqq E(T_{n0}) \quad (43)$$

since the exact values are respectively

$$\frac{\theta}{n-1} \left(1 - \frac{e^{-(n-1)g/\theta}}{n} \right) \leqq \frac{g+\theta}{n} \leqq g + \frac{\theta}{n} \quad (44)$$

In fact, since all units are new at starting time and since at the time of the first failure all units (except the replacement) have passed their guarantee period with probability one then

$$E(T_{n1}) \leqq E(T_{nj}) \leqq E(T_{n0}) \quad (j \geq 0) \quad (45)$$

If we compare the case $g > 0$ with the special case $g = 0$ we obtain

$$E(T_{nj}) \geqq \frac{\theta}{n} \quad (j = 1, 2, \dots) \quad (46)$$

and if we compare it with the non-replacement case (g/θ is large) we obtain

$$E(T_{nj}) \leq \frac{\theta}{n-j} \quad (j = 1, 2, \dots, n-1). \quad (47)$$

These comparisons show that the difference in (46) is small when g/θ is small and for $j < n$ the difference in (47) is small when g/θ is large.

It is possible to compute $E(T_{nj})$ exactly for $g \geq 0$ but the computation is extremely tedious for $j \geq 2$. The results for $j = 1$ and 0 are given in (44). For $j = 2$

$$\begin{aligned} E(T_{n2}) &= \frac{\theta}{n-2} \left[1 - \frac{(n+2)(n-1)}{n^2} e^{-(n-2)g/\theta} \right. \\ &\quad \left. + \frac{n-2}{n-1} e^{-(n-1)g/\theta} - \frac{n-2}{n^2(n-1)} e^{-2(n-1)g/\theta} \right] \quad (n > 2) \end{aligned} \quad (48)$$

and

$$E(T_{22}) = g - \frac{\theta}{4} [1 - 4e^{-g/\theta} + e^{-2g/\theta}] \quad (49)$$

For the case of two populations with a common guarantee period g we can write similar inequalities. We shall use different symbols a, b for the initial sample size from the populations with scale parameters θ_1, θ_2 respectively even though our principal interest is in the case $a = b = n$ say. Let $T_{a,b,j}$ denote the interval between the j^{th} and $j + 1^{\text{st}}$ failures in this case and let $\lambda_i = 1/\theta_i$ ($i = 1, 2$). We then have for all values of a and b

$$\begin{aligned} [a\lambda_1 + b\lambda_2]^{-1} &\leq E(T_{a,b,j}) \leq E(T_{a,b,0}) \\ &= g + [a\lambda_1 + b\lambda_2]^{-1} \quad (j = 0, 1, 2, \dots, \infty) \end{aligned} \quad (50)$$

$$E(T_{a,b,\infty}) = \frac{(\theta_1 + g)(\theta_2 + g)}{a(\theta_2 + g) + b(\theta_1 + g)} \quad (51)$$

The result for $E(T_{a,b,1})$ corresponding to that in (43) does not hold if the ratio θ_1/θ_2 is too large; in particular it can be shown that

$$\begin{aligned} E(T_{a,b,1}) &= \left(\frac{a\lambda_1}{a\lambda_1 + b\lambda_2} \right) \left(\frac{1}{(a-1)\lambda_1 + b\lambda_2} \right) \left[1 - \frac{\lambda_1 e^{-g[(a-1)\lambda_1 + b\lambda_2]}}{a\lambda_1 + b\lambda_2} \right] \\ &\quad + \left(\frac{b\lambda_2}{a\lambda_1 + b\lambda_2} \right) \left(\frac{1}{a\lambda_1 + (b-1)\lambda_2} \right) \left[1 - \frac{\lambda_2 e^{-g[a\lambda_1 + (b-1)\lambda_2]}}{a\lambda_1 + b\lambda_2} \right] \end{aligned} \quad (52)$$

is larger than $E(T_{a,b,\infty})$ for $a = b = 1$ when $g/\theta_1 = 0.01$ and $g/\theta_2 = 0.10$

so that $\theta_1/\theta_2 = 10$. The expression (52) reduces to that in (44) if we set $\theta_1 = \theta_2 = \theta$ and replace a and b by $n/2$ in the resulting expression.

Corresponding exact expressions for $E(T_{a,b,j})$ for $j > 1$ are extremely tedious to derive and unwieldy although the integrations involved are elementary. If we let $g \rightarrow \infty$ then we obtain expressions for the non-replacement case which are relatively simple. They are best expressed as a recursion formula.

$$\begin{aligned} E(T_{a,b,j}) &= \frac{a\lambda_1}{a\lambda_1 + b\lambda_2} ET_{a-1,b,j-1} \\ &\quad + \frac{b\lambda_2}{a\lambda_1 + b\lambda_2} ET_{a,b-1,j-1} \quad (j \geq 1) \end{aligned} \quad (53)$$

$$\begin{aligned} E(T_{a,b,1}) &= \frac{a\lambda_1}{a\lambda_1 + b\lambda_2} \frac{1}{(a-1)\lambda_1 + b\lambda_2} \\ &\quad + \frac{b\lambda_2}{a\lambda_1 + b\lambda_2} \frac{1}{a\lambda_1 + (b-1)\lambda_2} \quad (a, b \geq 1) \end{aligned} \quad (54)$$

$$E(T_{a,0,j}) \leq g + \theta_1/a \quad \text{for } j \geq a \text{ and } j = 0 \quad (55)$$

$$E(T_{a,0,j}) = \theta_1/(a-j) \quad \text{for } 1 \leq j \leq a-1 \quad (56)$$

Results similar to (55) and (56) hold for the case $a = 0$. The above results for $g = \infty$ provide useful approximations for $E(T_{a,b,j})$ when g is large. Upper bounds are given by

$$E(T_{a,b,j}) \leq [a\lambda_1 + (b-j)\lambda_2]^{-1} \quad (j = 1, 2, \dots, b) \quad (57)$$

$$E(T_{a,b,j+b}) \leq [(a-j)\lambda_1]^{-1} \quad (j = 1, 2, \dots, a-1). \quad (58)$$

Duration of the Experiment

For the sequential rule R_3' with $k = 2$ we can now write down approximations as well as upper and lower bounds to the expected duration $E(T)$ of the experiment. From (50)

$$\begin{aligned} g + \frac{E(F; \delta)}{n(\lambda_1 + \lambda_2)} &\leq E(T) = \sum_{j=0}^{c-1} E(T_{n,n,j}) \\ &\quad + [E(F; \delta) - c]E(T_{n,n,c}) \end{aligned} \quad (59)$$

where c is the largest integer less than or equal to $E(F; \delta)$. The right expression of (59) can be approximated by (53) and (54) if g is large. If $c < 2n$ then the upper bounds are given by (57) and (58). A simpler

upper bound, which holds for all values of c is given by

$$E(T) \leq E(F; \delta)E(T_{n,n,0}) = E(F; \delta) \left(g + \frac{\theta_1}{n} \right) \quad (60)$$

CASE 3: COMMON UNKNOWN LOCATION PARAMETER $g \geq 0$

In this case the more conservative procedure is to proceed under the assumption that $g = 0$. By the discussion above the probability requirement will in most problems be satisfied for all $g \geq 0$. The OC and ASN functions, which are now functions of the true value of g , were already obtained above. Of course, we need not consider values of g greater than the smallest observed lifetime of all units tested to failure.

ADDENDUM 2

For completeness it would be appropriate to state explicitly some of the formulas used in computing the tables in the early part of the paper. For the nonsequential, nonreplacement rule R_1 with $k = 2$ the probability of a correct selection is

$$P(\alpha; R_1) = \int_0^\infty \int_0^x f_r(y, \theta_2) f_r(x, \theta_1) dy dx \quad (61)$$

where

$$f_r(x, \theta) = \frac{r}{\theta} C_r^n (1 - e^{-x/\theta})^{r-1} e^{-x(n-r+1)/\theta}. \quad (r \leq n) \quad (62)$$

and C_r^n is the usual combinatorial symbol. This can also be expressed in the form

$$\begin{aligned} P(\alpha; R_1) &= 1 - (rC_r^n)^2 \sum_{j=1}^r \frac{(-1)^{j-1}}{n-r+j} \\ &\quad C_{j-1}^{r-1} \{B[r, n-r+1+\alpha(n-r+j)]\}^{-1} \end{aligned} \quad (63)$$

where $B[x, y]$ is the complete Beta function. Equation (66) holds for any $g \geq 0$.

For the rule R_1 the expected duration of the experiment for $k = 2$ is given by

$$E(T) = \int_0^\infty x \{f_r(x, \theta_1)[1 - F_r(x, \theta_2)] + f_r(x, \theta_2)[1 - F_r(x, \theta_1)]\} dx \quad (64)$$

where $f_r(x, \theta)$ is the density in (62) and $F_r(x, \theta)$ is its c.d.f. This can

also be expressed in the form

$$\theta_1 r(C_r^n)^2 \sum_{i=1}^r \sum_{j=1}^r \frac{(-1)^{i+j} C_{i-1}^{r-1} C_{j-1}^{r-1}}{(n - r + j)[i + n - r + \alpha(j + n - r)]^2} \quad (65)$$

plus another similar expression in which θ_1, α are replaced by θ_2, α^{-1} respectively. For $g > 0$ we need only add g to this result. This result was used to compute $E(T)$ in table 1A for $\alpha = 1$ and $\alpha = 2$. For $\alpha = \infty$ the expression simplifies to

$$E(T) = \theta_2 r C_r^n \sum_{j=1}^r C_{j-1}^{r-1} \frac{(-1)^{r-j}}{(n - j + 1)^2} \quad (66)$$

which can be shown to be equivalent to

$$E(T) = \theta_2 \sum_{j=1}^r \frac{1}{n - j + 1} \quad (67)$$

REFERENCES

1. Bechhofer, R. E., Kiefer, J. and Sobel, M., On a Type of Sequential Multiple Decision Procedures for Certain Ranking and Identification Problems with k Populations. To be published.
2. Birnbaum, A., Statistical methods for Poisson processes and exponential populations, *J. Am. Stat. Assoc.*, **49**, pp. 254-266, 1954.
3. Birnbaum, A., Some procedures for comparing Poisson processes or populations, *Biometrika*, **40**, pp. 447-49, 1953.
4. Girshick, M. A., Contributions to the theory of sequential analysis I, *Annals Math. Stat.*, **17**, pp. 123-43, 1946.
5. Wald, A., *Sequential Analysis*, John Wiley and Sons, New York, 1947.

A Class of Binary Signaling Alphabets

By DAVID SLEPIAN

(Manuscript received September 27, 1955)

A class of binary signaling alphabets called "group alphabets" is described. The alphabets are generalizations of Hamming's error correcting codes and possess the following special features: (1) all letters are treated alike in transmission; (2) the encoding is simple to instrument; (3) maximum likelihood detection is relatively simple to instrument; and (4) in certain practical cases there exist no better alphabets. A compilation is given of group alphabets of length equal to or less than 10 binary digits.

INTRODUCTION

This paper is concerned with a class of signaling alphabets, called "group alphabets," for use on the symmetric binary channel. The class in question is sufficiently broad to include the error correcting codes of Hamming,¹ the Reed-Muller codes,² and all "systematic codes".³ On the other hand, because they constitute a rather small subclass of the class of all binary alphabets, group alphabets possess many important special features of practical interest.

In particular, (1) all letters of the alphabets are treated alike under transmission; (2) the encoding scheme is particularly simple to instrument; (3) the decoder — a maximum likelihood detector — is the best possible theoretically and is relatively easy to instrument; and (4) in certain cases of practical interest the alphabets are the best possible theoretically.

It has very recently been proved by Peter Elias⁴ that there exist group alphabets which signal at a rate arbitrarily close to the capacity, C , of the symmetric binary channel with an arbitrarily small probability of error. Elias' demonstration is an existence proof in that it does not show *explicitly* how to construct a group alphabet signaling at a rate greater than $C - \varepsilon$ with a probability of error less than δ for arbitrary positive δ and ε . Unfortunately, in this respect and in many others, our understanding of group alphabets is still fragmentary.

In Part I, group alphabets are defined along with some related con-

cepts necessary for their understanding. The main results obtained up to the present time are stated without proof. Examples of these concepts are given and a compilation of the best group alphabets of small size is presented and explained. This section is intended for the casual reader.

In Part II, proofs of the statements of Part I are given along with such theory as is needed for these proofs.

The reader is assumed to be familiar with the paper of Hamming,¹ the basic papers of Shannon⁵ and the most elementary notions of the theory of finite groups.⁶

PART I — GROUP ALPHABETS AND THEIR PROPERTIES

1.1 INTRODUCTION

We shall be concerned in all that follows with communication over the symmetric binary channel shown on Fig. 1. The channel can accept either of the two symbols 0 or 1. A transmitted 0 is received as a 0 with probability q and is received as a 1 with probability $p = 1 - q$; a transmitted 1 is received as a 1 with probability q and is received as a 0 with probability p . We assume $0 \leq p \leq \frac{1}{2}$. The "noise" on the channel operates independently on each symbol presented for transmission. The capacity of this channel is

$$C = 1 + p \log_2 p + q \log_2 q \text{ bits/symbol} \quad (1)$$

By a *K-letter, n-place binary signaling alphabet* we shall mean a collection of K distinct sequences of n binary digits. An individual sequence of the collection will be referred to as a *letter* of the alphabet. The integer K is called the size of the alphabet. A letter is transmitted over the channel by presenting in order to the channel input the sequence of n zeros and ones that comprise the letter. A *detection scheme* or *detector* for

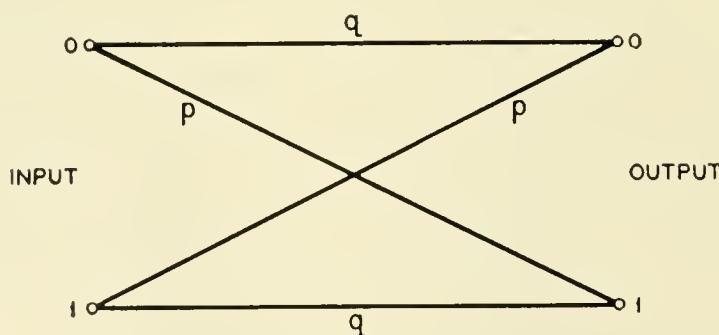


Fig. 1 — The symmetric binary channel.

a given K -letter, n -place alphabet is a procedure for producing a sequence of letters of the alphabet from the channel output.

Throughout this paper we shall assume that signaling is accomplished with a given K -letter, n -place alphabet by choosing the letters of the alphabet for transmission independently with equal probability $1/K$.

Shannon⁵ has shown that for sufficiently large n , there exist K -letter, n -place alphabets and detection schemes that signal over the symmetric binary channel at a rate $R > C - \varepsilon$ for arbitrary $\varepsilon > 0$ and such that the probability of error in the letters of the detector output is less than any $\delta > 0$. Here C is given by (1) and is shown as a function of p in Fig. 2. No algorithm is known (other than exhaustive procedures) for the construction of K -letter, n -place alphabets satisfying the above inequalities for arbitrary positive δ and ε except in the trivial cases $C = 0$ and $C = 1$.

1.2 THE GROUP B_n

There are a totality of 2^n different n -place binary sequences. It is frequently convenient to consider these sequences as the vertices of a cube of unit edge in a Euclidean space of n -dimensions. For example the 5-place sequence 0, 1, 0, 0, 1 is associated with the point in 5-space whose

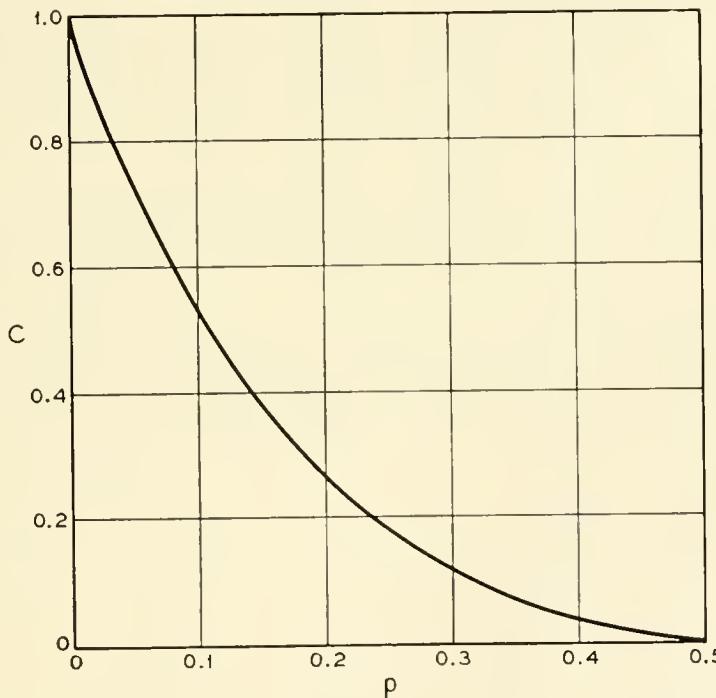


Fig. 2 — The capacity of the symmetric binary channel.

$$C = 1 + p \log_2 p + (1 - p) \log_2 (1 - p)$$

coordinates are $(0, 1, 0, 0, 1)$. For convenience of notation we shall generally omit commas in writing a sequence. The above 5-place sequence will be written, for example, 01001.

We define the *product* of two n -place binary sequences, $a_1a_2 \cdots a_n$ and $b_1b_2 \cdots b_n$ as the n -place binary sequence

$$a_1 \dot{+} b_1, \quad a_2 \dot{+} b_2, \cdots, a_n \dot{+} b_n$$

Here the a 's and b 's are zero or one and the $\dot{+}$ sign means addition modulo 2. (That is $0 \dot{+} 0 = 1 \dot{+} 1 = 0$, $0 \dot{+} 1 = 1 \dot{+} 0 = 1$) For example, $(01101)(00111) = 01010$. With this rule of multiplication the 2^n n -place binary sequences form an Abelian group of order 2^n . The elements of the group, denoted by T_1, T_2, \dots, T_{2^n} , say, are the n -place binary sequences; the identity element I is the sequence $000 \cdots 0$ and

$$IT_i = T_iI = T_i; \quad T_iT_j = T_jT_i; \quad T_i(T_jT_k) = (T_iT_j)T_k;$$

the product of any number of elements is again an element; every element is its own reciprocal, $T_i = T_i^{-1}$, $T_i^2 = I$. We denote this group by B_n .

All subgroups of B_n are of order 2^k where k is an integer from the set $0, 1, 2, \dots, n$. There are exactly

$$\begin{aligned} N(n, k) &= \frac{(2^n - 2^0)(2^n - 2^1)(2^n - 2^2) \cdots (2^n - 2^{k-1})}{(2^k - 2^0)(2^k - 2^1)(2^k - 2^2) \cdots (2^k - 2^{k-1})} \\ &= N(n, n - k) \end{aligned} \quad (2)$$

distinct subgroups of B_n of order 2^k . Some values of $N(n, k)$ are given in Table I.

TABLE I — SOME VALUES OF $N(n, k)$, THE NUMBER OF SUBGROUPS OF B_n OF ORDER 2^k . $N(n, k) = N(n, n - k)$

$n \setminus k$	0	1	2	3	4	5
2	1	3	1			
3	1	7	7	1		
4	1	15	35	15	1	
5	1	31	155	155	31	1
6	1	63	651	1395	651	63
7	1	127	2667	11811	11811	2667
8	1	255	10795	97155	200787	97155
9	1	511	43435	788035	3309747	3309747
10	1	1023	174251	6347715	53743987	109221651

1.3 GROUP ALPHABETS

An n -place *group alphabet* is a K -letter, n -place binary signaling alphabet whose letters form a subgroup of B_n . Of necessity the size of an n -place group alphabet is $K = 2^k$ where k is an integer satisfying $0 \leq k \leq n$. By an (n, k) -*alphabet* we shall mean an n -place group alphabet of size 2^k . Example: the $N(3, 2) = 7$ distinct $(3, 2)$ -alphabets are given by the seven columns

(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	
000	000	000	000	000	000	000	
100	100	100	010	010	001	110	
010	001	011	001	101	110	011	
110	101	111	011	111	111	101	

1.4 STANDARD ARRAYS

Let the letters of a specific (n, k) -alphabet be $A_1 = I = 00 \cdots 0$, A_2, A_3, \dots, A_μ , where $\mu = 2^k$. The group B_n can be developed according to this subgroup and its cosets:

$$\begin{aligned}
 &I, \quad A_2, \quad A_3, \quad \dots, A_\mu \\
 &S_2, \quad S_2A_2, \quad S_2A_3, \dots, S_2A_\mu \\
 &S_3, \quad S_3A_2, \quad S_3A_3, \dots, S_3A_\mu \\
 &\vdots \\
 &B_n = \quad \vdots \\
 &\quad \vdots \\
 &S_\nu, \quad S_\nu A_2, \quad S_\nu A_3, \dots, S_\nu A_\mu \\
 &\mu = 2^k, \quad \nu = 2^{n-k}.
 \end{aligned} \tag{4}$$

In this array every element of B_n appears once and only once. The collection of elements in any row of this array is called a *coset* of the (n, k) -alphabet. Here S_2 is any element of B_n not in the first row of the array, S_3 is any element of B_n not in the first two rows of the array, etc. The elements S_2, S_3, \dots, S_ν appearing under I in such an array will be called the *coset leaders*.

If a coset leader is replaced by any element in the coset, the same coset will result. That is to say the two collections of elements

$$S_i, \quad S_iA_2, \quad S_iA_3, \dots, S_iA_\mu$$

and

$$S_iA_k, \quad (S_iA_k)A_2, \quad (S_iA_k)A_3, \dots, (S_iA_k)A_\mu$$

are the same.

We define the *weight* $w_i = w(T_i)$ of an element, T_i , of B_n to be the number of ones in the n -place binary sequence T_i .

Henceforth, unless otherwise stated, we agree in dealing with an array such as (4) to adopt the following convention:

the leader of each coset shall be taken to be an element of minimal weight in that coset. (5)

Such a table will be called a *standard array*.

Example: B_4 can be developed according to the (4, 2)-alphabet 0000, 1100, 0011, 1111 as follows

0000	1100	0011	1111
1010	0110	1001	0101
1110	0010	1101	0001
1000	0100	1011	0111

(6)

According to (5), however, we should write, for example

0000	1100	0011	1111
1010	0110	1001	0101
0010	1110	0001	1101
1000	0100	1011	0111

(7)

The coset leader of the second coset of (6) can be taken as any element of that row since all are of weight 2. The leader of the third coset, however, should be either 0010 or 0001 since these are of weight one. The leader of the fourth coset should be either 1000 or 0100.

1.5 THE DETECTION SCHEME

Consider now communicating with an (n, k) -alphabet over the symmetric binary channel. When any letter, say A_i , of the alphabet is transmitted, the received sequence can be of any element of B_n . We agree to use the following detector:

if the received element of B_n lies in column i of the array (4), the detector prints the letter A_i , $i = 1, 2, \dots, \mu$. The array (4) is to be constructed according to the convention (5). (8)

The following propositions and theorems can be proved concerning signaling with an (n, k) -alphabet and the detection scheme given by (8).

1.6 BEST DETECTOR AND SYMMETRIC SIGNALING

Define the *probability* $\ell_i = \ell(T_i)$ of an element T_i of B_n to be $\ell_i = p^{w_i}q^{n-w_i}$ where p and q are as in (1) and w_i is the weight of T_i . Let

$Q_i, i = 1, 2, \dots, \mu$ be the sum of the probabilities of the elements in the i th column of the standard array (4).

Proposition 1. The probability that any transmitted letter of the (n, k) -alphabet be produced correctly by the detector is Q_1 .

Proposition 2. The equivocation⁵ per symbol is

$$H_y(x) = -\frac{1}{n} \sum_{i=1}^{\mu} Q_i \log_2 Q_i$$

Theorem 1. The detector (8) is a maximum likelihood detector. That is, for the given alphabet no other detection scheme has a greater average probability that a transmitted letter be produced correctly by the detector.

Let us return to the geometrical picture of n -place binary sequences as vertices of a unit cube in n -space. The choice of a K -letter, n -place alphabet corresponds to designating K particular vertices as letters. Since the binary sequence corresponding to any vertex can be produced by the channel output, any detector must consist of a set of rules that associates various vertices of the cube with the vertices designated as letters of the alphabet. We assume that every vertex is associated with some letter. The vertices of the cube are divided then into disjoint sets, W_1, W_2, \dots, W_K where W_i is the set of vertices associated with i th letter of the signaling alphabet. A maximum likelihood detector is characterized by the fact that every vertex in W_i is as close to or closer to the i th letter than to any other letter, $i = 1, 2, \dots, K$. For group alphabets and the detector (8), this means that no element in the i th column of array (4) is closer to any other A than it is to $A_i, i = 1, 2, \dots, \mu$.

Theorem 2. Associated with each (n, k) -alphabet considered as a point configuration in Euclidean n -space, there is a group of $n \times n$ orthogonal matrices which is transitive on the letters of the alphabet and which leaves the unit cube invariant. The maximum likelihood sets W_1, W_2, \dots, W_μ are all geometrically similar.

Stated in loose terms, this theorem asserts that in an (n, k) -alphabet every letter is treated the same. Every two letters have the same number of nearest neighbors associated with them, the same number of next nearest neighbors, etc. The disposition of points in any two W regions is the same.

1.7 GROUP ALPHABETS AND PARITY CHECKS

Theorem 3. Every group alphabet is a systematic³ code: every systematic code is a group alphabet.⁷

We prefer to use the word "alphabet" in place of "code" since the latter has many meanings. In a *systematic alphabet*, the places in any letter can be divided into two classes: the information places — k in number for an (n, k) -alphabet — and the check positions. All letters have the same information places and the same check places. If there are k information places, these may be occupied by any of the 2^k k -place binary sequences. The entries in the $n - k$ check positions are fixed linear (mod 2) combinations of the entries in the information positions. The rules by which the entries in the check places are determined are called *parity checks*. Examples: for the $(4, 2)$ -alphabet of (6), namely 0000, 1100, 0011, 1111, positions 2 and 3 can be regarded as the information positions. If a letter of the alphabet is the sequence $a_1a_2a_3a_4$, then $a_1 = a_2$, $a_4 = a_3$ are the parity checks determining the check places 1 and 4. For the $(5, 3)$ -alphabet 00000, 10001, 01011, 00111, 11010, 10110, 01100, 11101 places 1, 2, and 3 (numbered from the left) can be taken as the information places. If a general letter of the alphabet is $a_1a_2a_3a_4a_5$, then $a_4 = a_2 + a_3$, $a_5 = a_1 + a_2 + a_3$.

Two group alphabets are called *equivalent* if one can be obtained from the other by a permutation of places. Example: the 7 distinct $(3, 2)$ -alphabets given in (3) separate into three equivalence classes. Alphabets (i), (ii), and (iv) are equivalent; alphabets (iii), (v), (vi), are equivalent; (vii) is in a class by itself.

Proposition 3. Equivalent (n, k) -alphabets have the same probability Q_1 of correct transmission for each letter.

Proposition 4. Every (n, k) -alphabet is equivalent to an (n, k) -alphabet whose first k places are information places and whose last $n - k$ places are determined by parity checks over the first k places.

Henceforth we shall be concerned only with (n, k) -alphabets whose first k places are information places. The parity check rules can then be written

$$a_i = \sum_{j=1}^k \gamma_{ij} a_j, \quad i = k + 1, \dots, n \quad (9)$$

where the sums are of course mod 2. Here, as before, a typical letter of the alphabet is the sequence $a_1a_2 \dots a_n$. The γ_{ij} are $k(n - k)$ quantities, zero or one, that serve to define the particular (n, k) -alphabet in question.

1.8 MAXIMUM LIKELIHOOD DETECTION BY PARITY CHECKS

For any element, T , of B_n we can form the sum given on the right of (9). This sum may or may not agree with the symbol in the i th place of

T . If it does, we say T satisfies the i th-place parity check; otherwise T fails the i th-place parity check. When a set of parity check rules (9) is given, we can associate an $(n - k)$ -place binary sequence, $R(T)$, with each element T of B_n . We examine each check place of T in order starting with the $(k + 1)$ -st place of T . We write a zero if a place of T satisfies the parity check; we write a one if a place fails the parity check. The resultant sequence of zeros and ones, written from left to right is $R(T)$. We call $R(T)$ the *parity check sequence* of T . Example: with the parity rules $a_4 = a_2 + a_3$, $a_5 = a_1 + a_2 + a_3$ used to define the $(5, 3)$ -alphabet in the examples of Theorem 3, we find $R(11000) = 10$ since the sum of the entries in the second and third places of 11001 is not the entry of the fourth place and since the sum of $a_1 = 1$, $a_2 = 1$, and $a_3 = 0$ is $0 = a_5$.

Theorem 4. Let I, A_2, \dots, A_μ be an (n, k) -alphabet. Let $R(T)$ be the parity check sequence of an element T of B_n formed in accordance with the parity check rules of the (n, k) -alphabet. Then $R(T_1) = R(T_2)$ if and only if T_1 and T_2 lie in the same row of array (4). The coset leaders can be ordered so that $R(S_i)$ is the binary symbol for the integer $i - 1$.

As an example of Theorem 4 consider the $(4, 2)$ -alphabet shown with its cosets below

0000	1011	0101	1110
0100	1111	0001	1010
0010	1001	0111	1100
1000	0011	1101	0110

The parity check rules for this alphabet are $a_3 = a_1$, $a_4 = a_1 + a_2$. Every element of the second row of this array satisfies the parity check in the third place and fails the parity check in the 4th place. The parity check sequence for the second row is 01. The parity check for the third row is 10, and for the fourth row 11. Since every letter of the alphabet satisfies the parity checks, the parity check sequence for the first row is 00. We therefore make the following association between parity check sequences and coset leaders

$$\begin{aligned} 00 &\rightarrow 0000 = S_1 \\ 01 &\rightarrow 0100 = S_2 \\ 10 &\rightarrow 0010 = S_3 \\ 11 &\rightarrow 1000 = S_4 \end{aligned}$$

1.9 INSTRUMENTING A GROUP ALPHABET

Proposition 4 attests to the ease of the encoding operation involved

with the use of an (n, k) -alphabet. If the original message is presented as a long sequence of zeros and ones, the sequence is broken into blocks of length k places. Each block is used as the first k places of a letter of the signaling alphabet. The last $n-k$ places of the letter are determined by fixed parity checks over the first k places.

Theorem 4 demonstrates the relative ease of instrumenting the maximum likelihood detector (8) for use with an (n, k) -alphabet. When an element T of B_n is received at the channel output, it is subjected to the $n-k$ parity checks of the alphabet being used. This results in a parity check sequence $R(T)$. $R(T)$ serves to identify a unique coset leader, say S_i . The product $S_i T$ is then formed and produced as the detector output. The probability that this be the correct letter of the alphabet is Q_1 .

1.10 BEST GROUP ALPHABETS

Two important questions regarding (n, k) -alphabets naturally arise. What is the maximum value of Q_1 possible for a given n and k and which of the $N(n, k)$ different subgroups give rise to this maximum Q_1 ? The answers to these questions for general n and k are not known. For many special values of n and k the answers are known. They are presented in Tables II, III and IV, which are explained below.

The probability Q_1 that a transmitted letter be produced correctly by the detector is the sum, $Q_1 = \sum_i \ell(S_i)$ of the probabilities of the coset leaders. This sum can be rewritten as $Q_1 = \sum_{i=0}^n \alpha_i p^i q^{n-i}$ where α_i is the number of coset leaders of weight i . One has, of course, $\sum \alpha_i = v = 2^{n-k}$ for an (n, k) -alphabet. Also $\alpha_i \leq \binom{n}{i} = \frac{n!}{i!(n-i)!}$ since this is the number of elements of B_n of weight i .

The α_i have a special physical significance. Due to the noise on the channel, a transmitted letter, A_i , of an (n, k) -alphabet will in general be received at the channel output as some element T of B_n different from A_i . If T differs from A_i in s places, i.e., if $w(A_i T) = s$, we say that an s -tuple error has occurred. For a given (n, k) -alphabet, α_i is the number of i -tuple errors which can be corrected by the alphabet in question, $i = 0, 1, 2, \dots, n$.

Table II gives the α_i corresponding to the largest possible value of Q_1 for a given k and n for $k = 2, 3, \dots, n-1$, $n = 4, \dots, 10$ along with a few other scattered values of n and k . For reference the binomial coefficients $\binom{n}{i}$ are also listed. For example, we find from Table II that the best group alphabet with $2^4 = 16$ letters that uses $n = 10$ places has a

probability of correct transmission $Q_1 = q^{10} + 10q^9p + 39q^8p^2 + 14q^7p^3$. The alphabet corrects all 10 possible single errors. It corrects 39 of the possible $\binom{10}{2} = 45$ double errors (second column of Table II) and in addition corrects 14 of the 120 possible triple errors. By adding an additional place to the alphabet one obtains with the best (11, 4)-alphabet an alphabet with 16 letters that corrects all 11 possible single errors and all 55 possible double errors as well as 61 triple errors. Such an alphabet might be useful in a computer representing decimal numbers in binary form.

For each set of α 's listed in Table II, there is in Table III a set of parity check rules which determines an (n, k) -alphabet having the given α 's. The notation used in Table III is best explained by an example. A (10, 4)-alphabet which realizes the α 's discussed in the preceding paragraph can be obtained as follows. Places 1, 2, 3, 4 carry the information. Place 5 is determined to make the mod 2 sum of the entries in places 3, 4, and 5 equal to zero. Place 6 is determined by a similar parity check on places 1, 2, 3, and 6; place 7 by a check on places 1, 2, 4, and 7, etc.

It is a surprising fact that for all cases investigated thus far an (n, k) -alphabet best for a given value of p is uniformly best for all values of p , $0 \leq p \leq \frac{1}{2}$. It is of course conjectured that this is true for all n and k .

It is a further (perhaps) surprising fact that the best (n, k) -alphabets are not necessarily those with greatest nearest neighbor distance between letters when the alphabets are regarded as point configurations on the n -cube. For example, in the best (7, 3)-alphabet as listed in Table III, each letter has two nearest neighbors distant 3 edges away. On the other hand, in the (7, 3)-alphabet given by the parity check rules 413, 512, 623, 7123 each letter has its nearest neighbors 4 edges away. This latter alphabet does not have as large a value of Q_1 , however, as does the (7, 3)-alphabet listed on Table III.

The cases $k = 0, 1, n - 1, n$ have not been listed in Tables II and III. The cases $k = 0$ and $k = n$ are completely trivial. For $k = 1$, all $n > 1$ the best alphabet is obtained using the parity rule $a_2 = a_3 = \dots = a_n = a_1$. If $n = 2j$,

$$Q_1 = \sum_{i=0}^{j-1} \binom{n}{i} p^i q^{n-i} + \frac{1}{2} \binom{n}{j} p^j q^j. \text{ If } n = 2j + 1, Q_1 = \sum_{i=0}^j \binom{n}{i} p^i q^{n-i}.$$

For $k = n - 1, n > 1$, the maximum Q_1 is $Q_1 = q^{n-1}$ and a parity rule for an alphabet realizing this Q_1 is $a_n = a_1$.

If the α 's of an (n, k) -alphabet are of the form $\alpha_i = \binom{n}{i}$, $i = 0, 1,$

TABLE II—PROBABILITY OF NO ERROR WITH BEST ALPHABETS, $Q_1 = \sum \alpha_i p^i q^{n-i}$

$2, \dots, j, \alpha_{j+1} = r$ some integer, $\alpha_{j+2} = \alpha_{j+3} = \dots = \alpha_n = 0$, then there does not exist a 2^k -letter, n -place alphabet of any sort better than the given (n, k) -alphabet. It will be observed that many of the α 's of Table II are of this form. It can be shown that

Proposition 5 if $n + \binom{n-k}{2} + \binom{n-k}{3} \geq 2^{n-k} - 1$ there exists no 2^k -letter, n -place alphabet better than the best (n, k) -alphabet.

When the inequality of proposition 5 holds the α 's are either $\alpha_0 = 1, \alpha_1 = 2^{n-k} - 1$, all other $\alpha = 0$; or $\alpha_0 = 1, \alpha_1 = \binom{n}{1}, \alpha_2 = 2^{n-k} - 1 - \binom{n}{1}$ all other $\alpha = 0$; or the trivial $\alpha_0 = 1$ all other $\alpha = 0$ which holds when $k = n$. The region of the $n - k$ plane for which it is known that (n, k) -alphabets cannot be exceeded by any other is shown in Table IV.

1.11 A DETAILED EXAMPLE

As an example of the use of (n, k) -alphabets consider the not unrealistic case of a channel with $p = 0.001$, i.e., on the average one binary digit per thousand is received incorrectly. Suppose we wish to transmit messages using 32 different letters. If we encode the letters into the 32 5-place binary sequences and transmit these sequences without further encoding, the probability that a received letter be in error is $1 - (1 - p)^5 = 0.00449$. If the best $(10, 5)$ -alphabet as shown in Tables II and III is used, the probability that a letter be wrong is $1 - Q_1 = 1 - q^{10} - 10q^9p - 21q^8p^2 = 24p^2 - 72p^3 + \dots = 0.000024$. Thus by reducing the signaling rate by $\frac{1}{2}$, a more than *one hundredfold* reduction in probability of error is accomplished.

A $(10, 5)$ -alphabet to achieve these results is given in Table III. Let a typical letter of the alphabet be the 10-place sequence of binary digits $a_1a_2 \dots a_9a_{10}$. The symbols $a_1a_2a_3a_4a_5$ carry the information and can be any of 32 different arrangements of zeros and ones. The remaining places are determined by

$$\begin{aligned} a_6 &= a_1 \dotplus a_3 \dotplus a_4 \dotplus a_5 \\ a_7 &= a_1 \dotplus a_2 \dotplus a_4 \dotplus a_5 \\ a_8 &= a_1 \dotplus a_2 \dotplus a_3 \dotplus a_5 \\ a_9 &= a_1 \dotplus a_2 \dotplus a_3 \dotplus a_4 \\ a_{10} &= a_1 \dotplus a_2 \dotplus a_3 \dotplus a_4 \dotplus a_5 \end{aligned}$$

To design the detector for this alphabet, it is first necessary to determine the coset leaders for a standard array (4) formed for this alphabet.

TABLE III — PARITY CHECK RULES FOR BEST ALPHABETS

	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
$n = 4$	3 2								
	4 1 2								
$n = 5$	3 1 2	4 1 2							
	4 2	5 1 3							
$n = 6$	5 1								
	6 1								
$n = 7$	3 2	4 1 2	5 1 2 3	6 1 2 4					
	4 1	5 1 3	6 2 3						
$n = 8$	6 1	7 1 2	8 1 2 3	9 1 2 3					
	7 2								
$n = 9$	3 1	4 1 3	5 1 3 4	6 1	7 1	8 1	9 1	8 1	9 1
	4 1	5 1 2	6 1 2 4	7 1 2 4	8 1 2 3				
$n = 10$	5 2	6 1 3	7 1 2 3	8 1 2 3 4					
	6 2	7 2 3	8 1 2 3						
$n = 11$	7 1 2								
	8 1 2								
$n = 12$	3 1	4 1	5 1 3 4	6 1 3 4 5	7 1 3 4				
	4 1	5 2	6 1 2 4	7 1 2 4 5	8 1 2 4				
$n = 13$	5 1	6 1 2	7 1 2 3	8 1 2 3 5	9 1 2 3				
	6 2	7 1 3	8 1 2 3	9 1 2 3					
$n = 14$	7 2								
	8 1 2								
$n = 15$	9 1 2								

$n = 10$	3 1	4 1	5 3 4	6 1 3 4 5	7 1 3 4 5	8 1 3 4	9 1
	4 1	5 2	6 1 2 3	7 1 2 4 5	8 1 2 4 5	9 1 2 4	10 1
	5 1	6 3	7 1 2 4	8 1 2 3 5	9 1 2 3 5 6	10 1 2 3	
	6 2	7 1 2	8 1 3 4	9 1 2 3 4	10 1 2 3 4 5		
	8 1 2	9 2 3	10 1 2 3 4				
	9 1 2	10 1 2 3					
	10 1 2						
$n = 11$	3 1	4 3	5 1 3	7 1 3 4 5 6	8 1 3 4 5 6	9 1 3 4	10 1
	4 1	5 3	6 2 4	7 1 4	8 1 2 4 5 6	9 1 2 4 5 7	11 1
	5 1	6 2	7 1 3	8 2 3	9 1 2 3 5 6	10 1 2 3 5 6	12 1
	6 2	7 1 3	8 1 3	9 1 3 4	10 1 2 3 4 6	11 1 2 3 4 6 7	
	7 2	8 1 2	9 1 2	10 2 3 4	11 1 2 3 4 5		
	8 2	9 1 2	10 1 2 3	11 1 2 3 4			
	9 1 2	10 1 2	11 1 2 3				
$n = 12$	10 1 2	11 1 2					
	11 1 2						

TABLE IV—REGION OF THE $n-k$ PLANE FOR WHICH IT IS KNOWN
THAT (n, k) -ALPHABETS CANNOT BE EXCELLED

k	0	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	22	24	26	28	30	n
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
.
0	0	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	22	24	26	28	30	n

This can be done by a variety of special methods which considerably reduce the obvious labor of making such an array. A set of best S 's along with their parity check symbols is given in Table V.

A maximum likelihood detector for the $(10, 5)$ -alphabet in question forms from each received sequence $b_1 b_2 \dots b_{10}$ the parity check symbol $c_1 c_2 c_3 c_4 c_5$ where

$$\begin{aligned}c_1 &= b_6 + b_1 + b_3 + b_4 + b_5 \\c_2 &= b_7 + b_1 + b_2 + b_4 + b_5 \\c_3 &= b_8 + b_1 + b_2 + b_3 + b_5 \\c_4 &= b_9 + b_1 + b_2 + b_3 + b_4 \\c_5 &= b_{10} + b_1 + b_2 + b_3 + b_4 + b_5\end{aligned}$$

According to Table V, if $c_1 c_2 c_3 c_4 c_5$ contains less than three ones, the detector should print $b_1 b_2 b_3 b_4 b_5$. The detector should print $(b_1 + 1) b_2 b_3 b_4 b_5$ if the parity check sequence $c_1 c_2 c_3 c_4 c_5$ is either 11111 or 11110; the de-

TABLE V—COSET LEADERS AND PARITY CHECK SEQUENCES
FOR (10, 5)-ALPHABET

$c_1c_2c_3c_4c_5$	\leftrightarrow	S	$c_1c_2c_3c_4c_5$	\leftrightarrow	S
00000		0000000000	11100		0000100001
10000		0000010000	11010		0001000001
01000		0000001000	11001		0001000010
00100		0000000100	10110		0010000001
00010		0000000010	10101		0010000010
00001		0000000001	10011		0010000100
11000		0000011000	01110		0100000001
10100		0000010100	01101		0100000010
10010		0000010010	01011		0100000100
10001		0000010001	00111		0100001000
01100		0000001100	11110		1000000001
01010		0000001010	11101		0000100000
01001		0000001001	11011		0001000000
00110		0000000110	10111		0010000000
00101		0000000101	01111		0100000000
00011		0000000011	11111		1000000000

detector should print $b_1(b_2 + 1)b_3b_4b_5$ if the parity check sequence is 01111, 00111, 01011, 01101, or 01110; the detector should print $b_1b_2(b_3 + 1)b_4b_5$ if the parity check sequence is 10111, 10011, 10101, or 10110; the detector should print $b_1b_2b_3(b_4 + 1)b_5$ if the parity check sequence is 11011, 11001, 11010; and finally the detector should print $b_1b_2b_3b_4(b_5 + 1)$ if the parity check sequence is 11101 or 11100.

Simpler rules of operation for the detector may possibly be obtained by choice of a different set of S 's in Table V. These quantities in general are not unique. Also there may exist non-equivalent alphabets with simpler detector rules that achieve the same probability of error as the alphabet in question.

PART II—ADDITIONAL THEORY AND PROOFS OF THEOREMS OF PART I

2.1 THE ABSTRACT GROUP C_n

It will be helpful here to say a few more words about B_n , the group of n -place binary sequences under the operation of addition mod 2. This group is simply isomorphic with the abstract group C_n generated by n commuting elements of order two, say a_1, a_2, \dots, a_n . Here $a_i a_j = a_j a_i$ and $a_i^2 = I$, $i, j = 1, 2, \dots, n$, where I is the identity for the group. The eight distinct elements of C_3 are, for example, $I, a_1, a_2, a_3, a_1a_2, a_1a_3, a_2a_3, a_1a_2a_3$. The group C_n is easily seen to be isomorphic with the n -fold direct product of the group C_1 with itself.

It is a considerable saving in notation in dealing with C_n to omit the symbol "a" and write only the subscripts. In this notation for example, the elements of C_4 are $I, 1, 2, 3, 4, 12, 13, 14, 23, 24, 34, 123, 124, 134, 234, 1234$. The product of two or more elements of C_n can readily be written down. Its symbol consists of those numerals that occur an odd number of times in the collection of numerals that comprise the symbols of the factors. Thus, $(12)(234)(123) = 24$.

The isomorphism between C_n and B_n can be established in many ways. The most convenient way, perhaps, is to associate with the element $i_1 i_2 i_3 \cdots i_k$ of C_n the element of B_n that has ones in places i_1, i_2, \dots, i_k and zeros in the remaining $n - k$ places. For example, one can associate 124 of C_4 with 1101 of B_4 ; 14 with 1001, etc. In fact, the numeral notation afforded by this isomorphism is a much neater notation for B_n than is afforded by the awkward strings of zeros and ones. There are, of course, other ways in which elements of C_n can be paired with elements of B_n so that group multiplication is preserved. The collection of all such "pairings" makes up the group of automorphisms of C_n . This group of automorphisms of C_n is isomorphic with the group of non-singular linear homogenous transformations in a field of characteristic 2.

An element T of C_n is said to be *dependent* upon the set of elements T_1, T_2, \dots, T_j of C_n if T can be expressed as a product of some elements of the set T_1, T_2, \dots, T_j ; otherwise, T is said to be *independent* of the set. A set of elements is said to be independent if no member can be expressed solely in terms of the other members of the set. For example, in C_8 , 1, 2, 3, 4 form a set of independent elements as do likewise 2357, 12357, 14. However, 135 depends upon 145, 3457, 57 since $135 = (145)(3457)(57)$. Clearly any set of n independent elements of C_n can be taken as generators for the group. For example, all possible products formed of 12, 123, and 23 yield the elements of C_3 .

Any k independent elements of C_n serve as generators for a subgroup of order 2^k . The subgroup so generated is clearly isomorphic with C_k . All subgroups of C_n of order 2^k can be obtained in this way.

The number of ways in which k independent elements can be chosen from the 2^n elements of C_n is

$$F(n, k) = (2^n - 2^0)(2^n - 2^1)(2^n - 2^2) \cdots (2^n - 2^{k-1})$$

For, the first element can be chosen in $2^n - 1$ ways (the identity cannot be included in a non-trivial set of independent elements) and the second element can be chosen in $2^n - 2$ ways. These two elements determine a subgroup of order 2^2 . The third element can be chosen as any element of the remaining $2^n - 2^2$ elements. The 3 elements chosen determine a

subgroup of order 2^3 . A fourth independent element can be chosen as any of the remaining $2^n - 2^3$ elements, etc.

Each set of k independent elements serves to generate a subgroup of order 2^k . The quantity $F(n, k)$ is not, however, the number of distinct subgroups of C_n of this order, for, a given subgroup can be obtained from many different sets of generators. Indeed, the number of different sets of generators that can generate a given subgroup of order 2^k of C_n is just $F(k, k)$ since any such subgroup is isomorphic with C_k . Therefore the number of subgroups of C_n of order 2^k is $N(n, k) = F(n, k)/F(k, k)$ which is (2). A simple calculation gives $N(n, k) = N(n, n - k)$.

2.2 PROOF OF PROPOSITIONS 1 AND 2

After an element A of B_n has been presented for transmission over a noisy binary channel, an element T of B_n is produced at the channel output. The element $U = AT$ of B_n serves as a record of the noise during the transmission. U is an n -place binary sequence with a one at each place altered in A by the noise. The channel output, T , is obtained from the input A by multiplication by U : $T = UA$. For channels of the sort under consideration here, the probability that U be any particular element of B_n of weight w is $p^w q^{n-w}$.

Consider now signaling with a particular (n, k) -alphabet and consider the standard array (4) of the alphabet. If the detection scheme (8) is used, a transmitted letter A_i will be produced without error if and only if the received symbol is of the form $S_j A_i$. That is, there will be no error only if the noise in the channel during the transmission of A_i is represented by one of the coset leaders. (This applies for $i = 1, 2, \dots, \mu = 2^k$). The probability of this event is Q_1 (Proposition 1, Section 1.6). The convention (5) makes Q_1 as large as is possible for the given alphabet.

Let X refer to transmitted letters and let Y refer to letters produced by the detector. We use a vertical bar to denote conditions when writing probabilities. The quantity to the right of the bar is the condition. We suppose the letters of the alphabet to be chosen independently with equal probability 2^{-k} .

The equivocation $h(X | Y)$ obtained when using an (n, k) -alphabet with the detector (8) can most easily be computed from the formula

$$h(X | Y) = h(X) - h(Y) + h(Y | X) \quad (10)$$

The entropy of the source is $h(X) = k/n$ bits per symbol. The probability that the detector produce A_j when A_i was sent is the probability that the noise be represented by $A_i A_j S_\ell$, $\ell = 1, 2, \dots, \nu$. In symbols,

$$Pr(Y \rightarrow A_j | X \rightarrow A_i) = \sum_{\ell} Pr(N \rightarrow A_i A_j S_{\ell}) = Q(A_i A_j)$$

where $Q(A_i)$ is the sum of the probabilities of the elements that are in the same column as A_i in the standard array. Therefore

$$\begin{aligned} Pr(Y \rightarrow A_j) &= \sum_i Pr(Y \rightarrow A_j | X \rightarrow A_i) Pr(X \rightarrow A_i) = \frac{1}{2^k} \sum_i Q(A_i A_j) \\ &= \frac{1}{2^k}, \quad \text{since } \sum_i Q(A_i A_j) = \sum_i Q(A_i) = 1. \end{aligned}$$

This last follows from the group property of the alphabet. Therefore

$$h(Y) = -\frac{1}{n} \sum_i Pr(Y \rightarrow A_j) \log Pr(Y \rightarrow A_j) = \frac{k}{n} \text{ bits/symbol.}$$

It follows then from (10) that

$$h(X | Y) = h(Y | X)$$

The computation of $h(Y | X)$ follows readily from its definition

$$\begin{aligned} h(Y | X) &= \sum_i Pr(X \rightarrow A_i) h(Y | X \rightarrow A_i) \\ &= -\sum_{ij} Pr(X \rightarrow A_i) Pr(Y \rightarrow A_j | X \rightarrow A_i) \\ &\quad \log Pr(Y \rightarrow A_j | X \rightarrow A_i) \\ &= -\frac{1}{2^k} \sum_{ij} \sum_{\ell} Pr(N \rightarrow A_i S_{\ell} A_j) \log \sum_m Pr(N \rightarrow A_i S_m A_j) \\ &= -\frac{1}{2^k} \sum_{ij} Q(A_i A_j) \log Q(A_i A_j) \\ &= -\sum_i Q(A_i) \log Q(A_i) \end{aligned}$$

Each letter is n binary places. Proposition 2, then follows.

2.3 DISTANCE AND THE PROOF OF THEOREM 1

Let A and B be two elements of B_n . We define the *distance*, $d(A, B)$, between A and B to be the weight of their product,

$$d(A, B) = w(AB) \tag{11}$$

The distance between A and B is the number of places in which A and B differ and is just the "Hamming distance."¹ In terms of the n -cube, $d(A, B)$ is the minimum number of edges that must be traversed to go

from vertex A to vertex B . The distance so defined is a monotone function of the Euclidean distance between vertices.

It follows from (11) that if C is any element of B_n then

$$d(A, B) = d(AC, BC) \quad (12)$$

This fact shows the detection scheme (8) to be a maximum likelihood detector. By definition of a standard array, one has

$$d(S_i, I) \leq d(S_i A_j, I) \quad \text{for all } i \text{ and } j$$

The coset leaders were chosen to make this true. From (12),

$$\begin{aligned} d(S_i, I) &= d(S_i A_m S_i, I A_m S_i) = d(S_i A_m, A_m) \\ d(S_i A_j, I) &= d(S_i A_j S_i A_m, I S_i A_m) = d(A_j A_m, S_i A_m) \\ &= d(S_i A_m, A_\ell) \end{aligned}$$

where $A_\ell = A_j A_m$. Substituting these expressions in the inequality above yields

$$d(S_i A_m, A_m) \leq d(S_i A_m, A_\ell) \quad \text{for all } i, m, \ell$$

This equation says that an arbitrary element in the array (4) is at least as close to the element at the top of its column as it is to any other letter of the alphabet. This is the maximum likelihood property.

2.4 PROOF OF THEOREM 2

Again consider an (n, k) -alphabet as a set of vertices of the unit n -cube. Consider also n mutually perpendicular hyperplanes through the centroid of the cube parallel to the coordinate planes. We call these planes "symmetry planes of the cube" and suppose the planes numbered in accordance with the corresponding parallel coordinate planes.

The reflection of the vertex with coordinates $(a_1, a_2, \dots, a_i, \dots, a_n)$ in symmetry plane i yields the vertex of the cube whose coordinates are $(a_1, a_2, \dots, a_i + 1, \dots, a_n)$. More generally, reflecting a given vertex successively in symmetry planes i, j, k, \dots yields a new vertex whose coordinates differ from the original vertex precisely in places i, j, k, \dots . Successive reflections in hyperplanes constitute a transformation that leaves distances between points unaltered and is therefore a "rotation." The rotation obtained by reflecting successively in symmetry planes i, j, k , etc. can be represented by an n -place symbol having a one in places i, j, k , etc. and a zero elsewhere.

We now regard a given (n, k) -alphabet as generated by operating on the vertex $(0, 0, \dots, 0)$ of the cube with a certain collection of 2^k ro-

tation operators. The symbols for these operators are identical with the sequences of zeros and ones that form the coordinates of the 2^k points. It is readily seen that these rotation operators form a group which is transitive on the letters of the alphabet and which leave the unit cube invariant. Theorem 2 then follows.

Theorem 2 also follows readily from consideration of the array (4). For example, the maximum likelihood region associated with I is the set of points $I, S_2, S_3, \dots, S_\nu$. The maximum likelihood region associated with A_i is the set of points $A_i, A_iS_2, A_iS_3, \dots, A_iS_\nu$. The rotation (successive reflections in symmetry planes of the cube) whose symbol is the same as the coordinate sequence of A_i sends the maximum likelihood region of I into the maximum likelihood region of A_i , $i = 1, 2, \dots, \mu$.

2.5 PROOF OF THEOREM 3

That every systematic alphabet is a group alphabet follows trivially from the fact that the sum mod 2 of two letters satisfying parity checks is again a letter satisfying the parity checks. The totality of letters satisfying given parity checks thus constitutes a finite group.

To prove that every group alphabet is a systematic code, consider the letters of a given (n, k) -alphabet listed in a column. One obtains in this way a matrix with 2^k rows and n columns whose entries are zeros and ones. Because the rows are distinct and form a group isomorphic to C_k , there are k linearly independent rows (mod 2) and no set of more than k independent rows. The rank of the matrix is therefore k . The matrix therefore possesses k linearly independent (mod 2) columns and the remaining $n - k$ columns are linear combinations of these k . Maintaining only these k linearly independent columns, we obtain a matrix of k columns and 2^k rows with rank k . This matrix must, therefore, have k linearly independent rows. The rows, however, form a group under mod 2 addition and hence, since k are linearly independent, all 2^k rows must be distinct. The matrix contains only zeros and ones as entries; it has 2^k distinct rows of k entries each. The matrix must be a listing of the numbers from 0 to $2^k - 1$ in binary notation. The other $n - k$ columns of the original matrix considered are linear combinations of the columns of this matrix. This completes the proof of Theorem 3 and Proposition 4.

2.6 PROOF OF THEOREM 4

To prove Theorem 4 we first note that the parity check sequence of the product of two elements of B_n is the mod 2 sum of their separate

parity check sequences. It follows then that all elements in a given coset have the same parity check sequence. For, let the coset be $S_i, S_i A_2, S_i A_3, \dots, S_i A_\mu$. Since the elements $I, A_2, A_3, \dots, A_\mu$ all have parity check sequence $00 \dots 0$, all elements of the coset have parity check $R(S_i)$.

In the array (4) there are 2^{n-k} cosets. We observe that there are 2^{n-k} elements of B_n that have zeros in their first k places. These elements have parity check symbols identical with the last $n - k$ places of their symbols. These elements therefore give rise to 2^{n-k} different parity check symbols. The elements must be distributed one per coset. This proves Theorem 4.

2.7 PROOF OF PROPOSITION 5

If

$$n \geq 2^{n-k} - \binom{n-k}{2} - \binom{n-k}{3} - 1$$

we can explicitly exhibit group alphabets having the property mentioned in the paragraph preceding Proposition 5. The notation of the demonstration is cumbersome, but the idea is relatively simple.

We shall use the notation of paragraph 2.1 for elements of B_n , i.e., an element of B_n will be given by a list of integers that specify what places of the sequence for the element contain ones. It will be convenient furthermore to designate the first k places of a sequence by the integers $1, 2, 3, \dots, k$ and the remaining $n - k$ places by the "integers" $1', 2', 3', \dots, \ell'$, where $\ell = n - k$. For example, if $n = 8, k = 5$, we have

$$\begin{aligned} 10111010 &\leftrightarrow 13452' \\ 10000100 &\leftrightarrow 11' \\ 00000101 &\leftrightarrow 1'3' \end{aligned}$$

Consider the group generated by the elements $1', 2', 3', \dots, \ell'$, i.e. the 2^ℓ elements $I, 1', 2', \dots, \ell', 1'2', 1'3', \dots, 1'2'3' \dots \ell'$. Suppose these elements listed according to decreasing weight (say in decreasing order when regarded as numbers in the decimal system) and numbered consecutively. Let B_i be the i th element in the list. Example: if $\ell = 3$, $B_1 = 1'2'3', B_2 = 2'3', B_3 = 1'3', B_4 = 1'2', B_5 = 3', B_6 = 2', B_7 = 1'$.

Consider now the (n, k) -alphabet whose generators are

$$1B_1, 2B_2, 3B_3, \dots, kB_k$$

We assert that if

$$n \geq 2^{n-k} - \binom{n-k}{2} - \binom{n-k}{3} - 1$$

this alphabet is as good as any other alphabet of 2^k letters and n places.

In the first place, we observe that every letter of this (n, k) -alphabet (except I) has unprimed numbers in its symbols. It follows that each of the 2^ℓ letters $I, 1', 2', \dots, \ell', 1'2', \dots, 1'2'\dots\ell'$ occurs in a different coset of the given (n, k) -alphabet. For, if two of these letters appeared in the same coset, their product (which contains only primed numbers) would have to be a letter of the (n, k) alphabet. This is impossible since every letter of the (n, k) alphabet has unprimed numbers in its symbol. Since there are precisely 2^ℓ cosets we can designate a coset by the single element of the list $B_1, B_2, \dots, B_{2^\ell} = I$ which appears in the coset.

We next observe that the condition

$$n \geq 2^{n-k} - \binom{n-k}{2} - \binom{n-k}{3} - 1$$

guarantees that B_{k+1} is of weight 3 or less. For, the given condition is equivalent to

$$k \geq 2^\ell - \binom{\ell}{0} - \binom{\ell}{1} - \binom{\ell}{2} - \binom{\ell}{3}$$

We treat several cases depending on the weight of B_{k+1} .

If B_{k+1} is of weight 3, we note that for $i = 1, 2, \dots, k$, the coset containing B_i also contains an element of weight one, namely the element i obtained as the product of B_i with the letter iB_i of the given (n, k) -alphabet. Of the remaining $(2^\ell - k)$ B 's, one is of weight zero, ℓ are of weight one, $\binom{\ell}{2}$ are of weight 2 and the remaining are of weight 3. We have, then $\alpha_0 = 1$, $\alpha_1 = \ell + k = n$. Now every B of weight 4 occurs in the list of generators $1B_1, 2B_2, \dots, kB_k$. It follows that on multiplying this list of generators by any B of weight 3, at least one element of weight two will result. (E.g., $(1'2'3')(j1'2'3'4') = j4'$) Thus every coset with a B of weight 2 or 3 contains an element of weight 2 and $\alpha_2 = 2^\ell - \alpha_0 - \alpha_1$.

The argument in case B_{k+1} is of weight two or one is similar.

2.8 MODULAR REPRESENTATIONS OF C_n

In order to explain one of the methods used to obtain the best (n, k) -alphabets listed in Tables II and III, it is necessary to digress here to present additional theory.

It has been remarked that every (n, k) -alphabet is isomorphic with C_k . Let us suppose the elements of C_k listed in a column starting with I and proceeding in order $I, 1, 2, 3, \dots, k, 12, 13, \dots, (k-1)k, 123, \dots, 123\dots k$. The elements of a given (n, k) -alphabet can be paired off with these abstract elements so as to preserve group multiplication. This can be done in many different ways. The result is a matrix with elements zero and one with n columns and 2^k rows, these latter being labelled by the symbols $I, 1, 2, \dots$ etc. What can be said about the columns of this matrix? How many different columns are possible when all (n, k) -alphabets and all methods of establishing isomorphism with C_k are considered?

In a given column, once the entries in rows $1, 2, \dots, k$ are known, the entire column is determined by the group property. There are therefore only 2^k possible different columns for such a matrix. A table showing these 2^k possible columns of zeros and ones will be called a *modular representation* table for C_k . An example of such a table is shown for $k = 4$ in Table VI.

It is clear that the columns of a modular representation table can also be labelled by the elements of C_k , and that group multiplication of these column labels is isomorphic with mod 2 addition of the columns. The table is a symmetric matrix. The element with row label A and column label B is one if the symbols A and B have an odd number of different numerals in common and is zero otherwise.

Every (n, k) -alphabet can be made from a modular representation table by choosing n columns of the table (with possible repetitions) at least k of which form an independent set.

TABLE VI — MODULAR REPRESENTATION TABLE FOR GROUP C_4

	I	1	2	3	4	12	13	14	23	24	34	123	124	134	234	1234
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1
2	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
3	0	0	0	1	0	0	1	0	1	0	1	1	0	1	1	1
4	0	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1
12	0	1	1	0	0	0	1	1	1	1	0	0	0	1	1	0
13	0	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0
14	0	1	0	0	1	1	1	0	0	1	1	1	0	0	1	0
23	0	0	1	1	0	1	1	0	0	1	1	0	1	1	0	0
24	0	0	1	0	1	1	0	1	1	0	1	1	0	1	0	0
34	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0
123	0	1	1	1	0	0	0	1	0	1	1	1	0	0	0	1
124	0	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1
134	0	1	0	1	1	1	0	0	1	1	0	0	0	1	0	1
234	0	0	1	1	1	1	1	0	0	0	0	0	0	1	1	1
1234	0	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0

We henceforth exclude consideration of the column I of a modular representation table. Its inclusion in an (n, k) -alphabet is clearly a waste of 1 binary digit.

It is easy to show that every column of a modular representation table for C_k contains exactly 2^{k-1} ones. Since an (n, k) -alphabet is made from n such columns the alphabet contains a total of $n2^{k-1}$ ones and we have

Proposition 6. The weights of an (n, k) -alphabet form a partition of $n2^{k-1}$ into $2^k - 1$ non-zero parts, each part being an integer from the set $1, 2, \dots, n$.

The identity element always has weight zero, of course.

It is readily established that the product of two elements of even weight is again an element of even weight as is the product of two elements of odd weight. The product of an element of even weight with an element of odd weight yields an element of odd weight.

The elements of even weight of an (n, k) -alphabet form a subgroup and the preceding argument shows that this subgroup must be of order 2^k or 2^{k-1} . If the group of even elements is of order 2^{k-1} , then the collection of even elements is a possible $(n, k - 1)$ -alphabet. This $(n, k - 1)$ alphabet may, however, contain the column I of the modular representation table of C_{k-1} . We therefore have

Proposition 7. The partition of Proposition 6 must be either into $2^k - 1$ even parts or else into 2^{k-1} odd parts and $2^{k-1} - 1$ even parts. In the latter case, the even parts form a partition of $\alpha 2^{k-2}$ where α is some integer of the set $k - 1, k, \dots, n$ and each of the parts is an integer from the set $1, 2, \dots, n$.

2.9 THE CHARACTERS OF C_k

Let us replace the elements of B_n (each of which is a sequence of zeros and ones) by sequences of +1's and -1's by means of the following substitution

$$\begin{aligned} 0 &\leftrightarrow 1 \\ 1 &\leftrightarrow -1. \end{aligned} \tag{13}$$

The multiplicative properties of elements of B_n can be preserved in this new notation if we define the product of two +1, -1 symbols to be the symbol whose i th component is the ordinary product of the i th components of the two factors. For example, 1011 and 0110 become respectively -11 -1 -1 and 1 -1 -11. We have

$$(-11 -1 -1)(1 -1 -11) = (-1 -11 -1)$$

corresponding to the fact that

$$(1011) (0110) = (1101)$$

If the $+1$, -1 symbols are regarded as shorthand for diagonal matrices, so that for example

$$\begin{matrix} -1 & 1 & -1 & -1 \end{matrix} \leftrightarrow \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix}$$

then group multiplication corresponds to matrix multiplication.

(While much of what follows here can be established in an elementary way for the simple group at hand, it is convenient to fall back upon the established general theory of group representations⁸ for several propositions.)

The substitution (13) converts a modular representation table (column I included) into a square array of $+1$'s and -1 's. Each column (or row) of this array is clearly an irreducible representation of C_k . Since C_k is Abelian it has precisely 2^k irreducible representations each of degree one. These are furnished by the converted modular table. This table also furnishes then the characters of the irreducible representations of C_k and we refer to it henceforth as a *character table*.

Let $\chi^\alpha(A)$ be the entry of the character table in the row labelled A and column labelled α . The orthogonality relationship for characters gives

$$\sum_{A \in C_k} \chi^\alpha(A) \chi^\beta(A) = 2^k \delta_{\alpha\beta}$$

$$\sum_{\alpha \in C_k} \chi^\alpha(A) \chi^\alpha(B) = 2^k \delta_{AB}$$

where δ is the usual Kronecker symbol. In particular

$$\sum_{A \in C_k} \chi^I(A) \chi^\beta(A) = \sum_{A \in C_k} \chi^\beta(A) = 0, \quad \beta \neq I$$

Since each $\chi^\beta(A)$ is $+1$ or -1 , these must occur in equal numbers in any column $\beta \neq I$. This implies that each column except I of the modular representation table contains 2^{k-1} ones, a fact used earlier.

Every matrix representation of C_k can be reduced to its irreducible components. If the trace of the matrix representing the element A in an arbitrary matrix representation of C_k is $\chi(A)$, then this representation contains the irreducible representation having label β in the character table d_β times where

$$d_\beta = \frac{1}{2^k} \sum_{A \subseteq C_k} \chi(A) \chi^\beta(A) \quad (14)$$

Every (n, k) -alphabet furnishes us with a matrix representation of C_k by means of (13) and the procedure outlined below (13). The trace $\chi(A)$ of the matrix representing the element A of C_k is related to the weight of the letter by

$$\chi(A) = n - 2w(A) \quad (15)$$

Equations (14) and (15) permit us to compute from the weights of an (n, k) -alphabet what irreducible representations are present in the alphabet and how many times each is contained. It is assumed here that the given alphabet has been made isomorphic to C_k and that the weights are labelled by elements of C_k .

Consider the converse problem. Given a set of numbers w_1, w_2, \dots, w_{2^k} that satisfy Propositions 6 and 7. From these we can compute quantities $\chi_i = n - 2w_i$ as in (15). It is clear that the given w 's will constitute the weights of an (n, k) -alphabet if and only if the $2^k \chi_i$ can be labelled with elements of C_k so that the 2^k sums (14) (β ranges over all elements of C_k) are non-negative integers. The integers d_β tell what representations to choose to construct an (n, k) -alphabet with the given weights w_1 .

2.10 CONSTRUCTION OF BEST ALPHABETS

A great many different techniques were used to construct the group alphabets listed in Tables II and III and to show that for each n and k there are no group alphabets with smaller probability of error. Space prohibits the exhibition of proofs for all the alphabets listed. We content ourselves here with a sample argument and treat the case $n = 10, k = 4$ in detail.

According to (2) there are $N(10, 4) = 53,743,987$ different $(10, 4)$ -alphabets. We now show that none is better than the one given in Table III. The letters of this alphabet and weights of the letters are

I	0
1 6 7 8 10	5
2 6 7 9 10	5
3 5 6 8 9 10	6
4 5 7 8 9 10	6
1 2 8 9	4
1 3 5 7 9	5

1 4 5 6 9	5
2 3 5 7 8	5
2 4 5 6 8	5
3 4 6 7	4
1 2 3 5 7 9	6
1 2 4 5 7 10	6
1 3 4 8 10	5
2 3 4 9 10	5
1 2 3 4 6 7 8 9	8

The notation is that of Section 2.1. By actually forming the standard array of this alphabet, it is verified that

$$\alpha_0 = 1, \quad \alpha_1 = 10, \quad \alpha_2 = 39, \quad \alpha_3 = 14.$$

Table II shows $\binom{10}{2} = 45$, whereas $\alpha_2 = 39$, so the given alphabet does not correct all possible double errors. In the standard array for the alphabet, 39 coset leaders are of weight 2. Of these 39 cosets, 33 have only one element of weight 2; the remaining 6 cosets each contain two elements of weight 2. This is due to the two elements of weight 4 in the given group, namely 1289 and 3467. A portion of the standard array that demonstrates these points is

I	1289	3467
.	.	.
.	.	.
12	89	.
18	29	.
19	28	.
34	.	67
36	.	47
37	.	46
.	.	.
.	.	.

In order to have a smaller probability of error than the exhibited alphabet, it is necessary that a $(10, 4)$ -alphabet have an $\alpha_2 > 39$. We proceed to show that this is impossible by consideration of the weights of the letters of possible $(10, 4)$ -alphabets.

We first show that every $(10, 4)$ -alphabet must have at least one element (other than the identity, I) of weight less than 5. By Propositions 6 and 7, Section 2.8, the weights must form a partition of $10 \cdot 8 = 80$ into 15 positive parts. If the weights are all even, at least two must be less than 6 since $14 \cdot 6 = 84 > 80$. If eight of the weights are odd, we see from $8 \cdot 5 + 7 \cdot 6 = 82 > 80$ that at least one weight must be less than 5.

An alphabet with one or more elements of weight 1 must have an $\alpha_2 \leq 36$, for there are nine elements of weight 2 which cannot possibly be coset leaders. To see this, suppose (without loss of generality) that the alphabet contains the letter 1. The elements 12, 13, 14, \dots 1 10 cannot possibly be coset leaders since the product of any one of them with the letter 1 yields an element of weight 1.

An alphabet with one or more elements of weight 2 must have an $\alpha_2 \leq 37$. Suppose for example, the alphabet contained the letter 12. Then 13 and 23 must be in the same coset, 14 and 24 must be in the same coset, \dots , 1 10 and 2 10 must be in the same coset. There are at least eight elements of weight two which are not coset leaders.

Each element of weight 3 in the alphabet prevents three elements of weight 2 from being coset leaders. For example, if the alphabet contains 123, then 12, 13, and 23 cannot be coset leaders. We say that the three elements of weight 2 are "blocked" by the letter of weight 3. Suppose an alphabet contains at least three letters of weight three. There are several cases: (A) if three letters have no numerals in common, e.g., 123, 456, 789, then nine distinct elements of weight 2 are blocked and $\alpha_2 \leq 36$; (B) if no two of the letters have more than a single numeral in common, e.g., 123, 345, 789, then again nine elements of weight 2 are blocked and $\alpha_2 \leq 36$; and (C) if two of the letters of weight 3 have two numerals in common, e.g., 123, 234, then their product is a letter of weight 2 and by the preceding paragraph $\alpha_2 \leq 37$. If an alphabet contains exactly two elements of weight 3 and no elements of weight 2, the elements of weight 3 block six elements of weight 2 and $\alpha_2 \leq 39$.

The preceding argument shows that to be better than the exhibited alphabet a (10, 4)-alphabet with letters of weight 3 must have just one such letter. A similar argument (omitted here) shows that to be better than the exhibited alphabet, a (10, 4)-alphabet cannot contain more than one element of weight 4. Furthermore, it is easily seen that an alphabet containing one element of weight 3 and one element of weight 4 must have an $\alpha_2 \leq 39$.

The only new contenders for best (10, 4)-alphabet are, therefore, alphabets with a single letter other than I of weight less than 5, and this letter must have weight 3 or 4. Application of Propositions 6 and 7 show that the only possible weights for alphabets of this sort are: 35^76^7 and 5^846^6 where 5^7 means seven letters of weight 5, etc. We next show that there do not exist (10, 4)-alphabets having these weights.

Consider first the suggested alphabet with weights 35^76^7 . As explained in Section 2.9, from such an alphabet we can construct a matrix representation of C_4 having the character $\chi(I) = 10$, one matrix of trace 4,

seven of trace 0 and seven of trace -2 . The latter seven matrices correspond to elements of even weight and together with I must represent a subgroup of order 8. We associate them with the subgroup generated by the elements 2, 3, and 4. We have therefore

$$\begin{aligned}\chi(I) &= 10, \quad \chi(2) = \chi(3) = \chi(4) = \chi(23) \\ &\quad = \chi(24) = \chi(34) = \chi(234) = -2.\end{aligned}$$

Examination of the symmetries involved shows that it doesn't matter how the remaining χ_i are associated with the remaining group elements. We take, for example

$$\begin{aligned}\chi(1) &= 4, \quad \chi(12) = \chi(13) = \chi(14) = \chi(123) \\ &\quad = \chi(124) = \chi(134) = \chi(1234) = 0.\end{aligned}$$

Now form the sum shown in equation (14) with $\beta = 1234$ (i.e., with the character χ^{1234} obtained from column 1234 of the Table VI by means of substitution (13). There results $d_{1234} = \frac{1}{2}$ which is impossible. Therefore there does not exist a $(10, 4)$ -alphabet with weights $35^7 6^7$.

The weights $5^8 4^6$ correspond to a representation of C_4 with character $\chi(I) = 10, 0^8, 2, (-2)^6$. We take the subgroup of elements of even weight to be generated by 2, 3, and 4. Except for the identity, it is clearly immaterial to which of these elements we assign the character 2. We make the following assignment: $\chi(I) = 10, \chi(2) = 2, \chi(3) = \chi(4) = \chi(23) = \chi(24) = \chi(34) = \chi(234) = -2, \quad \chi(1) = \chi(12) = \chi(13) = \chi(14) = \chi(123) = \chi(124) = \chi(134) = \chi(1234) = 0$. The use of equation (14) shows that $d_2 = \frac{1}{2}$ which is impossible.

It follows that of the 53,743,987 $(10, 4)$ -alphabets, none is better than the one listed on Table III.

Not all the entries of Table III were established in the manner just demonstrated for the $(10, 4)$ -alphabet. In many cases the search for a best alphabet was narrowed down to a few alphabets by simple arguments. The standard arrays for the alphabets were constructed and the best alphabet chosen. For large n the labor in making such a table can be considerable and the operations involved are highly liable to error when performed by hand.

I am deeply indebted to V. M. Wolontis who programmed the IBM CPC computer to determine the α 's of a given alphabet and who patiently ran off many such alphabets in course of the construction of Tables II and III. I am also indebted to Mrs. D. R. Fursdon who evaluated many of the smaller alphabets by hand.

REFERENCES

1. R. W. Hamming, B.S.T.J., **29**, pp. 147-160, 1950.
2. I. S. Reed, Transactions of the Professional Group on Information Theory, PGIT-4, pp. 38-49, 1954.
3. See section 7 of R. W. Hamming's paper, loc. cit.
4. I.R.E. Convention Record, Part 4, pp. 37-45, 1955 National Convention, March, 1955.
5. C. E. Shannon, B.S.T.J., **27**, pp. 379-423 and pp. 623-656, 1948.
6. Birkhoff and MacLane, A Survey of Modern Algebra, Macmillan Co., New York, 1941. Van der Waerden, Modern Algebra, Ungar Co., New York, 1953. Miller, Blichfeldt, and Dickson, Finite Groups, Stechert, New York, 1938.
7. This theorem has been previously noted in the literature by Kiyasu-Zen'iti, Research and Development Data No. 4, Ele. Comm. Lab., Nippon Tele. Corp. Tokyo, Aug., 1953.
8. F. D. Murnaghan, Theory of Group Representations, Johns Hopkins Press, Baltimore, 1938. E. Wigner, Gruppentheorie, Edwards Brothers, Ann Arbor, Michigan, 1944.

Bell System Technical Papers Not Published in This Journal

ALLEN, L. J., see Fewer, D. R.

ALLISON, H. W., see Moore, G. E.

BAKER, W. O., see Winslow, F. H.

BARSTOW, J. M.¹

Color TV — How it Works, I.R.E. Student Quarterly, **2**, pp. 11–16, Sept., 1955.

BASSECHES, H.¹ and McLEAN, D. A.¹

Gassing of Liquid Dielectrics Under Electrical Stress, Ind. & Engg. Chem., **47**, pp. 1782–1794, Sept., 1955.

BECK, A. C.¹

Measurement Techniques for Multimode Waveguides, Proc. I.R.E., MRI, **4**, pp. 325–6, Oct. 1, 1955.

BECKER, J. A.¹

The Life History of Adsorbed Atoms, Ions, and Molecules, N. Y. Acad. Sci. Ann., **58**, pp. 723–740, Sept. 15, 1955.

BLACKWELL, J. H., see Fewer, D. R.

BOORSE, H. A., see Smith, B.

BOZORTH, R. M.,¹ GETLIN, B. B.,¹ GALT, J. K.,¹ MERRITT, F. R.,¹ and YAGER, W. A.¹

Frequency Dependence of Magnetocrystalline Anisotropy, Letter to the Editor, Phys. Rev., **99**, p. 1898, Sept. 15, 1955.

1. Bell Telephone Laboratories, Inc.

BOZORTH, R. M.¹, TILDEN, E. F.,¹ and WILLIAMS, A. J.¹

Anisotropy and Magnetostriction of Some Ferrites, *Phys. Rev.*, **99**, pp. 1788-1798, Sept. 15, 1955.

BRIDGERS, H. E.,¹ and KOLB, E. D.¹

Rate-Grown Germanium Crystals for High-Frequency Transistors, Letter to the Editor, *J. Appl. Phys.*, **26**, pp. 1188-1189, Sept., 1955.

BULLINGTON, K.¹

Characteristics of Beyond-the-Horizon Radio Transmission, *Proc. I.R.E.*, **43**, pp. 1175-1180, Oct., 1955.

BULLINGTON, K.¹ INKSTER, W. J.,⁵ and DURKEE, A. L.¹

Results of Propagation Tests at 505 Mc and 4,090 Mc on Beyond-Horizon Paths, *Proc. I.R.E.*, **43**, pp. 1306-1316, Oct., 1955.

CALBICK, C. J.¹

Surface Studies with the Electron Microscope, *N. Y. Acad. Sci. Ann.*, **58**, pp. 873-892, Sept. 15, 1955.

CASS, R. S., see Fewer, D. R.

DURKEE, A. L., see Bullington, K.

FEWER, D. R.,¹ BLACKWELL, J. H.,⁴ ALLEN, L. J.,⁴ and CASS, R. S.⁴

Audio-Frequency Circuit Model of the 1-Dimensional Schroedinger Equation and Its Sources of Error, *Canadian J. of Phys.*, **33**, pp. 483-491, Aug., 1955.

FRANCOIS, E. E., see Law, J. T.

DAVIS, J. L., see Suhl, H.

GALT, J. K., see Bozorth, R. M., and Yager, W. A.

GARN, P. D.,¹ and HALLINE, MRS. E. W.¹

Polarographic Determination of Phthalic and Anhydride Alkyd Resins, *Anal Chem.*, **27**, pp. 1563-1565, Oct., 1955.

1. Bell Telephone Laboratories, Inc.

4. University of Western Ontario, London, Canada

5. Bell Telephone Company of Canada, Montreal

GETLIN, B. B., see Bozorth, R. M.

GIANOLA, U. F.¹

Application of the Wiedemann Effect to the Magnetostrictive Coupling of Crossed Coils, J. Appl. Phys., **26**, pp. 1152–1157, Sept., 1955.

Goss, A. J., see Hassion, F. X.

GREEN, E. I.¹

The Story of Q, American Scientist, **43**: pp. 584–594, Oct., 1955.

HALLINE, MRS. E. W., see Garn, P. D.

HARROWER, G. A.¹

Measurement of Electron Energies by Deflection in a Uniform Electric Field, Rev. Sci. Instr., **26**, pp. 850–854, Sept., 1955.

HASSION, F. X.,¹ Goss, A. J.,¹ and TRUMBORE, F. A.¹

The Germanium-Silicon Phase Diagram, J. Phys. Chem., **59**, p. 1118, Oct., 1955.

HASSION, F. X.,¹ THURMOND, C. D.,¹ and TRUMBORE, F. A.¹

On the Melting Point of Germanium, J. Phys. Chem., **59**, p. 1076, Oct., 1955.

HINES, M. E.,¹ HOFFMAN, G. W.,¹ and SALOOM, J. A.¹

Positive-Ion Drainage in Magnetically Focused Electron Beams, J. Appl. Phys., **26**, pp. 1157–1162, Sept., 1955.

HOFFMAN, G. W., see Hines, M. E.

INKSTER, W. J., see Bullington, K.

KELLY, M. J.¹

Training Programs of Industry for Graduate Engineers, Elec. Engg., **74**, pp. 866–869, Oct., 1955.

KOLB, E. D., see Bridgers, H. E.

1. Bell Telephone Laboratories, Inc.

LAW, J. T.,¹ and FRANCOIS, E. E.¹

Adsorption of Gasses and Vapors on Germanium, N. Y. Acad. Sci. Ann., **58**, pp. 925-936, Sept. 15, 1955.

LOVELL, Miss L. C., see Pfann, W. G.

MATREYEK, W., see Winslow, F. H.

MCLEAN, D. A., see Basseches, H.

MERRITT, F. R., see Bozorth, R. M., and Yager, W. A.

MEYER, F. T.¹

An Improved Detached-Contact Type of Schematic Circuit Drawing, A.I.E.E. Commun. & Electronics, **20**, pp. 505-513, Sept., 1955.

MILLER, B. T.²

Telephone Merchandising, Telephony, **149**, pp. 116-117, Oct. 22, 1955.

MILLER, S. L.¹

Avalanche Breakdown in Germanium, Phys. Rev., **99**, pp. 1234-1241, Aug. 15, 1955.

MOORE, G. E.,¹ and ALLISON, H. W.¹

Adsorption of Strontium and of Barium on Tungsten, J. Chem. Phys., **23**, pp. 1609-1621, Sept., 1955.

NEISSER, W. R.,¹

Liquid Nitrogen Coal Traps, Rev. Sci. Instr., **26**, p. 305, Mar., 1955.

OSTERGREN, C. N.²

Some Observations on Liberalized Tax Depreciation, Telephony, **149**, pp. 16-23-37, Oct. 1, 1955.

OSTERGREN, C. N.²

Depreciation and the New Law, Telephony, **149**, pp. 96-100-104-108, Oct. 22, 1955.

PAPE, N. R., see Winslow, F. H.

1. Bell Telephone Laboratories, Inc.

2. American Telephone and Telegraph Co.

PEDERSEN, L.¹

Aluminum Die Castings for Carrier Telephone Systems, A.I.E.E. Commun. & Electronics, **20**, pp. 434-439, Sept., 1955.

PETERS, H.¹

Hard Rubber, Ind. and Engg. Chem., Part II, pp. 2220-2222, Sept. 20, 1955.

PFANN, W. G.¹

Temperature-Gradient Zone-Melting, J. Metals, **7**, p. 961, Sept., 1955.

PFANN, W. G.¹ and LOVELL, Miss L. C.¹

Dislocation Densities in Intersecting Lineage Boundaries in Germanium, Letter to the Editor, Acta Met., **3**, pp. 512-513, Sept., 1955.

PIERCE, J. R.¹

Orbital Radio Relays, Jet Propulsion, **25**, pp. 153-157, Apr., 1955.

POOLE, K. M.¹

Emission from Hollow Cathodes, J. Appl. Phys., **26**, pp. 1176-1179, Sept., 1955.

SALOOM, J. A., see Hines, M. E.

SLICHTER, W. P.¹

Proton Magnetic Resonance in Polyamides, J. Appl. Phys., **26**, pp. 1099-1103, Sept., 1955.

SMITH, B.,¹ and BOORSE, H. A.⁶

Helium II Film Transport. II. The Role of Surface Finish, Phys. Rev. **99**, pp. 346-357, July 15, 1955.

SMITH, B.,¹ and BOORSE, H. A.⁶

Helium II Film Transport. IV. The Role of Temperature, Phys. Rev., **99**, pp. 367-370, July 15, 1955.

SUHL, H.,¹ VAN UITERT, L. G.,¹ and DAVIS, J. L.¹

Ferromagnetic Resonance in Magnesium-Manganese Aluminum Ferrite Between 160 and 1900 Mc., Letter to the Editor, J. Appl. Phys., **26**, pp. 1181-1182, Sept., 1955.

1. Bell Telephone Laboratories, Inc.

6. Columbia University, New York City

THURMOND, C. D., see Hassion, F. X.

TIDD, W. H.¹

Demonstration of Bandwidth Capabilities of Beyond-Horizon Tropospheric Radio Propagation, Proc. I.R.E., **43**, pp. 1297–1299, Oct., 1955.

TIEN, P. K.¹ and WALKER, L. R.¹

Large Signal Theory of Traveling-Wave Amplifiers, Proc. I.R.E., **43**, p. 1007, Aug., 1955.

TILDEN, E. F., see Bozorth, R. M.

TRUMBORE, F. A., see Hassion, F. X.

UHLIR, A., JR.¹

Micromachining with Virtual Electrodes, Rev. Sci. Instr., **26**, pp. 965–968, Oct., 1955.

ULRICH, W., see Yokelson, B. J.

VAN UITERT, L. G., see Suhl, H.

WALKER, L. R., see Tien, P. K.

WEIBEL, E. S.¹

Vowel Synthesis by Means of Resonant Circuits, J. Acous. Soc., **27**, pp. 858–865, Sept., 1955.

WILLIAMS, A. J., see Bozorth, R. M.

WINSLOW, F. H.¹ BAKER, W. O.¹ and YAGER, W. A.¹

Odd Electrons in Polymer Molecules, Am. Chem. Soc., **77**, pp. 4751–4756, Sept. 20, 1955.

WINSLOW, F. H.¹ BAKER, W. O.¹ PAPE, N. R.¹ and MATREYEK, W.¹
Formation and Properties of Polymer Carbon, J. Polymer Science, **16**, p. 101, Apr., 1955.

YAGER, W. A., see Bozorth, R. M.

1. Bell Telephone Laboratories, Inc.

YAGER, W. A.¹ GALT, J. K.¹ and MERRITT, F. R.¹

Ferromagnetic Resonance in Two-Nickel-Iron Ferrites, Phys. Rev., **99**, pp. 1203-1209, Aug. 15, 1955.

YOKELSON, B. J.¹ and ULRICH, W.¹

Engineering Multistage Diode Logic Circuits, A.I.E.E. Commun. & Electronics, **20**, pp. 466-475, Sept., 1955.

1. Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ARNOLD, W. O., and HOEFLER, R. R.

A System Plan for Air Traffic Control, Monograph 2483.

BECK, A. C.

Measurement Techniques for Multimode Waveguides, Monograph 2421.

BECKER, J. A., and BRANDES, R. G.

Adsorption of Oxygen on Tungsten as Revealed in Field Emission Microscope, Monograph 2493.

BOYLE, W. S., see Germer, L. H.

BRANDES, R. G., see Becker, J. A.

BRATTAIN, W. H., see Garrett, C. G. B.

GARRETT, C. G. B., and BRATTAIN, W. H.

Physical Theory of Semiconductor Surfaces, Monograph 2453.

GERNER, L. H., BOYLE, W. S., and KISLIUK, P.

Discharges at Electrical Contacts — II, Monograph 2499.

HOEFLER, R. R., see Arnold, W. O.

KISLIUK, P., see Germer, L. H.

LINVILL, J. G.

Nonsaturating Pulse Circuits Using Two Junction Transistors, Monograph 2475.

* 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.

MASON, W. P.

Relaxations in the Attenuation of Single Crystal Lead, Monograph 2454.

MEYER, F. T.

An Improved Detached-Contact-Type of Schematic Circuit Drawing, Monograph 2456.

VOGEL, F. L., JR.

Dislocations in Low-Angle Boundaries in Germanium, Monograph 2455.

WALKER, L. R.

Generalizations of Brillouin Flow, Monograph 2432.

WARNER, A. W.

Frequency Aging of High-Frequency Plated Crystal Units, Monograph 2474.

WEIBEL, E. S.

On Webster's Horn Equation, Monograph 2450.

Contributors to This Issue

A. C. BECK, E.E., Rensselaer Polytechnic Institute, 1927; Instructor, Rensselaer Polytechnic Institute, 1927–1928; Bell Telephone Laboratories, 1928 –. With the Radio Research Department he was engaged in the development and design of short-wave and microwave antennas. During World War II he was chiefly concerned with radar antennas and associated waveguide structures and components. For several years after the war he worked on development of microwave radio repeater systems. Later he worked on microwave transmission developments for broadband communication. Recently he has concentrated on further developments in the field of broadband communication using circular waveguides and associated test equipment.

J. S. COOK, B.E.E., and M.S., Ohio State University, 1952; Bell Telephone Laboratories, 1952 –. Mr. Cook is a member of the Research in High-Frequency and Electronics Department at Murray Hill and has been engaged principally in research on the traveling-wave tube. Mr. Cook is a member of the Institute of Radio Engineers and belongs to the Professional Group on Electron Devices.

O. E. DELANGE, B.S. University of Utah, 1930; M.A. Columbia University, 1937; Bell Telephone Laboratories, 1930 –. His early work was principally on the development of high-frequency transmitters and receivers. Later he worked on frequency modulation and during World War II was concerned with the development of radar. Since that time he has been involved in research using broadband systems including microwave and baseband. Mr. DeLange is a member of the Institute of Radio Engineers.

R. KOMPFNER, Engineering Degree, Technische Hochschule, Vienna, 1933; Ph.D., Oxford, 1951; Bell Telephone Laboratories, 1951 –. Between 1941–1950 he did work for the British Admiralty at Birmingham University and Oxford University in the Royal Naval Scientific Service. He invented the traveling-wave tube and for this achievement Dr. Kompfner received the 1955 Duddell Medal, bestowed by the Physical Society of England. In the Laboratories' Research in High Frequency

and Electronics Department, he has continued his research on vacuum tubes, particularly those used in the microwave region. He is a Fellow of the Institute of Radio Engineers and of the Physical Society in London.

CHARLES A. LEE, B.E.E., Rensselaer Polytechnic Institute, 1943; Ph.D., Columbia University, 1953; Bell Telephone Laboratories, 1953-. When Mr. Lee joined the Laboratories he became engaged in research concerning solid state devices. In particular he has been developing techniques to extend the frequency of operation of transistors into the microwave range, including work on the diffused base transistor. During World War II, as a member of the United States Signal Corps, he was concerned with the determination and detection of enemy countermeasures in connection with the use of proximity fuses by the Allies. He is a member of the American Physical Society and the American Institute of Physics. He is also a member of Sigma Xi, Tau Beta Pi and Eta Kappa Nu.

JOHN R. PIERCE, B.S., M.S. and Ph.D., California Institute of Technology 1933, 1934 and 1936; Bell Telephone Laboratories, 1936-. Appointed Director of Research — Electrical Communications in August, 1955. Dr. Pierce has specialized in Development of Electron Tubes and Microwave Research since joining the Laboratories. During World War II he concentrated on the development of electronic devices for the Armed Forces. Since the war he has done research leading to the development of the beam traveling-wave tube for which he was awarded the 1947 Morris Liebmann Memorial Prize of the Institute of Radio Engineers. Dr. Pierce is author of two books: *Theory and Design of Electron Beams*, published in second edition last year, and *Traveling Wave Tubes* (1950). He was voted the "Outstanding Young Electrical Engineer of 1942" by Eta Kappa Nu. Fellow of the American Physical Society and the I.R.E. Member of the National Academy of Sciences, the A.I.E.E., Tau Beta Pi, Sigma Xi, Eta Kappa Nu, the British Interplanetary Society, and the Newcomen Society of North America.

C. F. QUATE, B.S., University of Utah 1944; Ph.D., Stanford University 1950; Bell Laboratories 1950-. Dr. Quate has been engaged in research on electron dynamics — the study of vacuum tubes in the microwave frequency range. He is a member of I.R.E.

DAVID SLEPIAN, University of Michigan, 1941-1943; M.A. and Ph.D., Harvard University, 1946-1949; Bell Telephone Laboratories, 1950-. Dr.

Slepian has been engaged in mathematical research in communication theory, switching theory and theory of noise. Parker Fellow in physics, Harvard University 1949-50. Member of I.R.E., American Mathematical Society, the American Association for the Advancement of Science and Sigma Xi.

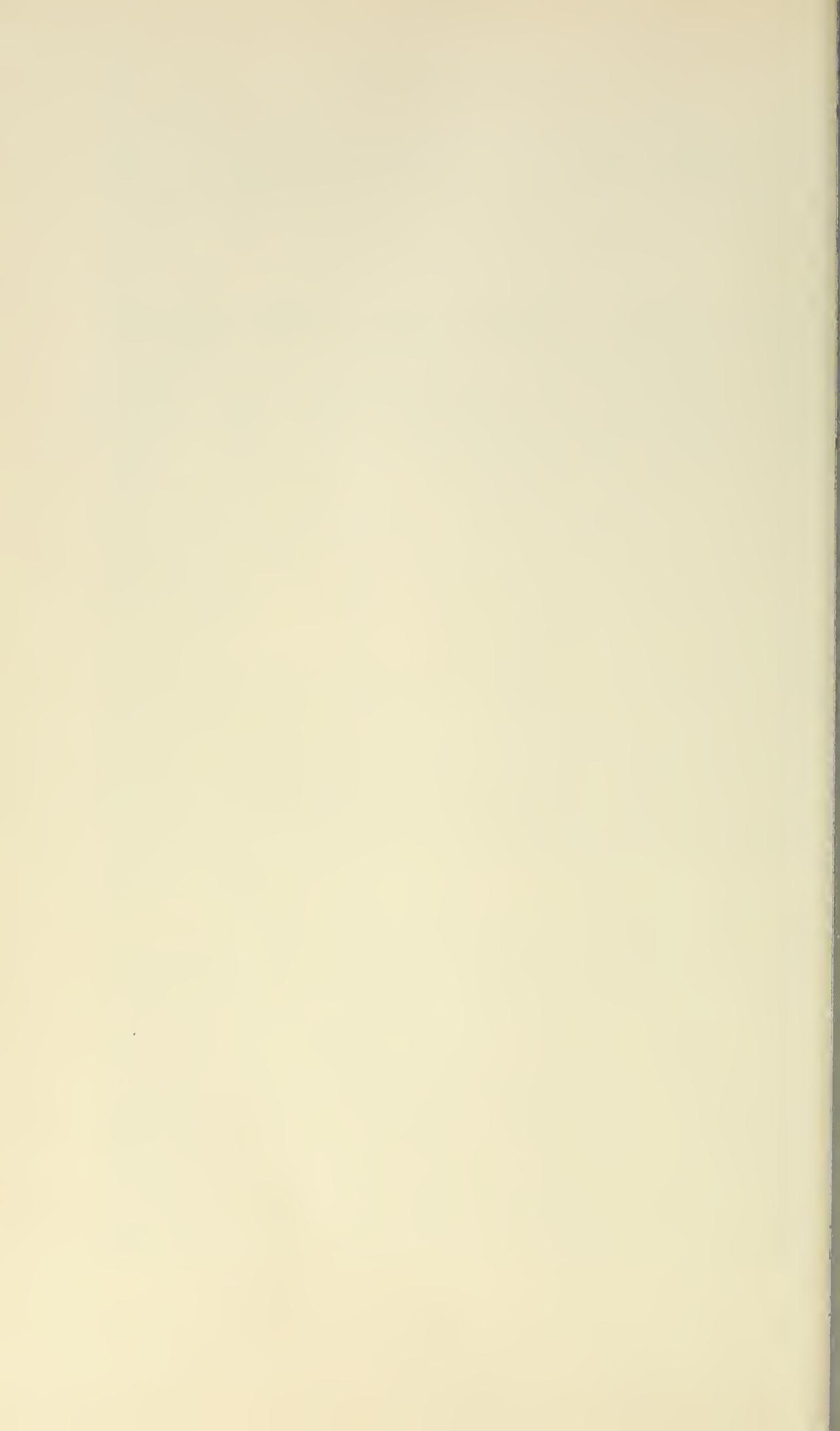
MILTON SOBEL, B.S., City College of New York, 1940; M.A., 1946 and Ph.D., 1951, Columbia University; U. S. Census Bureau, Statistician, 1940-41; U. S. Army War College, Statistician, 1942-44; Columbia University, Department of Mathematics, Assistant, 1946-48 and Research Associate 1948-50; Wayne University, Assistant Professor of Mathematics, 1950-52; Columbia University, Department of Mathematical Statistics, Visiting Lecturer, 1952; Cornell University, fundamental research in mathematical statistics, 1952-54; Bell Telephone Laboratories, 1954-. Dr. Sobel is engaged in fundamental research on life testing reliability problems with special application to transistors and is a consultant on many Laboratories projects. Member of Institute of Mathematical Statistics, American Statistical Association and Sigma Xi.

MORRIS TANENBAUM, A.B., Johns Hopkins University, 1949; M.A., Princeton University, 1950; Ph.D. Princeton University, 1952; Bell Telephone Laboratories, 1952-. Dr. Tanenbaum has been concerned with the chemistry and semiconducting properties of intermetallic compounds. At present he is exploring the semiconducting properties of silicon and the feasibility of silicon semiconductor devices. Dr. Tanenbaum is a member of the American Chemical Society and American Physical Society. He is also a member of Phi Lambda Upsilon, Phi Beta Kappa and Sigma Xi.

DONALD E. THOMAS, B.S. in E.E., Pennsylvania State College, 1929; M.A., Columbia University, 1932; Bell Telephone Laboratories, 1929-1942, 1946-. His first assignment at the Laboratories was in submarine cable development. Just prior to World War II he became engaged in the development of sea and airborne radar and continued in this work until he left for military duty in 1942. During World War II he was made a member of the Joint and Combined Chiefs of Staff Committees on Radio Countermeasures. Later he was a civilian member of the Department of Defense's Research and Development Board Panel on Electronic Countermeasures. Upon rejoining the Laboratories in 1946, Mr. Thomas was active in the development and installation of the first deep sea repeatered submarine telephone cable, between Key West and Havana,

which went into service in 1950. Later he was engaged in the development of transistor devices and circuits for special applications. At the present time he is working on the evaluation and feasibility studies of new types of semiconductors devices. He is a senior member of the I.R.E. and a member of Tau Beta Pi and Phi Kappa Phi.

LAURENCE R. WALKER, B.Sc. and Ph.D., McGill University, 1935 and 1939; University of California 1939-41; Radiation Laboratory, Massachusetts Institute of Technology, 1941-45; Bell Telephone Laboratories, 1945-. Dr. Walker has been primarily engaged in the development of microwave oscillators and amplifiers. At present he is a member of a physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.



H E B E L L S Y S T E M

Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXV

MARCH 1956

NUMBER 2

An Experimental Remote Controlled Line Concentrator

A. E. JOEL, JR. 249
1956

Transistor Circuits for Analog and Digital Systems

F. H. BLECHER 295

Electrolytic Shaping of Germanium and Silicon A. UHLIR, JR. 333

A Large Signal Theory of Traveling-Wave Amplifiers P. K. TIEN 349

A Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type W. E. DANIELSON, J. L. ROSENFELD AND J. A. SALOOM 375

Theories for Toll Traffic Engineering in the U.S.A. R. I. WILKINSON 421

Crosstalk on Open-Wire Lines

W. C. BABCOCK, ESTHER RENTROP AND C. S. THAELER 515

Bell System Technical Papers Not Published in This Journal 519

Recent Bell System Monographs 527

Contributors to This Issue 531

THE BELL SYSTEM TECHNICAL JOURNAL

A D V I S O R Y B O A R D

F. R. KAPPEL, *President Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Executive Vice President, American
Telephone and Telegraph Company*

E D I T O R I A L C O M M I T T E E

B. McMILLAN, *Chairman*

A. J. BUSCH

H. R. HUNTLEY

A. C. DICKIESON

F. R. LACK

R. L. DIETZOLD

J. R. PIERCE

K. E. GOULD

H. V. SCHMIDT

E. I. GREEN

C. E. SCHOOLEY

R. K. HONAMAN

G. N. THAYER

E D I T O R I A L S T A F F

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

MARCH 1956

NUMBER 2

Copyright 1956, American Telephone and Telegraph Company

An Experimental Remote Controlled Line Concentrator

By. A. E. JOEL, JR.

(Manuscript received June 30, 1955)

Concentration, which is the process of connecting a number of telephone lines to a smaller number of switching paths, has always been a fundamental function in switching systems. By performing this function remotely from the central office, a new balance between outside plant and switching costs may be obtained which shows promise of providing service more economically in some situations.

The broad concept of remote line concentrators is not new. However, its solution with the new devices and techniques now available has made the possibilities of decentralization of the means for switching telephone connections very promising.

Three models of an experimental equipment have been designed and constructed for service. The models have included equipment to enable the evaluation of new procedures required by the introduction of remote line concentrators into the telephone plant. The paper discusses the philosophy, devices, and techniques.

CONTENTS

1. Introduction	250
2. Objectives	251
3. New Devices Employed	252
4. New Techniques Employed	254
5. Switching Plan	257

6. Basic Circuits.....	261
a. Diode Gates.....	261
b. Transistor Bistable Circuit.....	262
c. Transistor Pulse Amplifier.....	263
d. Transistor Ring Counter.....	264
e. Crosspoint Operating Circuit.....	266
f. Crosspoint Relay Circuit.....	267
g. Pulse Signalling Circuit.....	268
h. Power Supply.....	269
7. Concentrator Operation.....	270
a. Line Scanning.....	270
b. Line Selection.....	272
c. Crosspoint Operation and Check.....	273
8. Central Office Circuits.....	274
a. Scanner Pulse Generator.....	279
b. Originating Call Detection and Line Number Registration.....	280
c. Line Selection.....	282
d. Trunk Selection and Identification.....	284
9. Field Trials.....	286
10. Miscellaneous Features of Trial Equipment.....	287
a. Traffic Recorder, b. Line Condition Tester.....	288
c. Simulator, d. Service Observing.....	290
e. Service Denial, f. Pulse Display Circuit.....	291

1. INTRODUCTION

The equipment which provides for the switching of telephone connections has always been located in what have been commonly called "central offices". These offices provide a means for the accumulation of all switching equipment required to handle the telephone needs of a community or a section of the community. The telephone building in which one or more central offices are located is sometimes referred to as the "wire center" because, like the spokes of a wheel, the wires which serve local telephones radiate in all directions to the telephones of the community.

A new development, made possible largely by the application of devices and techniques new to the telephone switching field, has recently been tried out in the telephone plant and promises to change much of the present conception of "central" offices and "wire" centers. It is known as a "line concentrator" and provides a means for reducing the amount of outside plant cables, poles, etc., serving a telephone central office by dispersing the switching equipment in the outside plant. It is not a new concept to reduce outside plant by bringing the switching equipment closer to the telephone customer but the technical difficulties of maintaining complex switching equipment and the cost of controlling such equipment at a distance have in the past been formidable obstacles to the development of line concentrators. With the invention of low power, small-sized, long-life devices such as transistors, gas tubes, and sealed relays, and their application to line concentrators, and with the development of new local switching systems with greater flexibility, it has been possible to make the progress described herein.

2. OBJECTIVES

Within the telephone offices the first switching equipment through which dial lines originate calls concentrates the traffic to the remaining equipment which is engineered to handle the peak busy hour load with the appropriate grade of service.¹ This concentration stage is different for different switching systems. In the step-by-step system² it is the line finder, and in the crossbar systems it is the primary line switch.³ Proposals for the application of remote line concentrators in the step-by-step system date back over 50 years.⁴ Continuing studies over the years have not indicated that any appreciable savings could be realized when such equipment is used within the local area served by a switching center.

When telephone customers move from one location to another within a local service area, it is desirable to retain the same telephone numbers. The step-by-step switching system in general is a unilateral arrangement where each line has two appearances in the switching equipment, one for originating call concentration (the line finder) and one for selection of the line on terminating calls (the connector). The connector fixes the line number and telephone numbers cannot be readily reassigned when moving these switching stages to out-of-office locations.

Common-control systems⁵ have been designed with flexibility so that the line number assignments on the switching equipment are independent of the telephone numbers. Furthermore, the first switching stage in the office is bilateral, handling both originating and terminating calls through the same facilities. The most recent common-control switching system in use in the Bell System, the No. 5 crossbar,⁶ has the further advantage of universal control circuitry for handling originating and terminating calls through the line switches. For these reasons, the No. 5 crossbar system was chosen for the first attempt to employ new techniques of achieving an economical remote line concentrator.

A number of assumptions were made in setting the design requirements. Some of these are influenced by the characteristics of the No. 5 crossbar system. These assumptions are as follows:

1. No change in customer station apparatus. Standard dial telephones to be used with present impedance levels, transmission characteristics, dial pulsing, party identification, superimposed ac-de ringing,⁶ and signaling and talking ranges.
2. Individual and two-party (full or semi-selective ringing) stations to be served but not coin or PBX lines.
3. Low cost could best be obtained by minimizing the per line equipment in the central office. AMA⁷ charging facilities could be used but to avoid per station equipment in the central office no message register operation would be provided.

4. Each concentrator would serve up to 50 lines with the central office control circuits common to a number of concentrators. (Experimental equipment described herein was designed for 60 lines to provide additional facilities for field trial purposes.) No extensive change would be made in central office equipment not associated with the line switches nor should concentrator design decrease call carrying capacities of existing central office equipment.

5. To provide data to evaluate service performance, automatic traffic recording facilities to be integrated with the design.

6. Remote equipment designed for pole or wall mounting as an addition to existing outside plant. Therefore, terminal distribution facilities would not be provided in the same cabinet.

7. Power to be supplied from the central office to insure continuity of telephone service in the event of a local power failure.

8. Concentrators to operate over existing types of exchange area facilities without change and with no decrease in station to central office service range.

9. Maintenance effort to be facilitated by plug-in unit design using the most reliable devices obtainable.

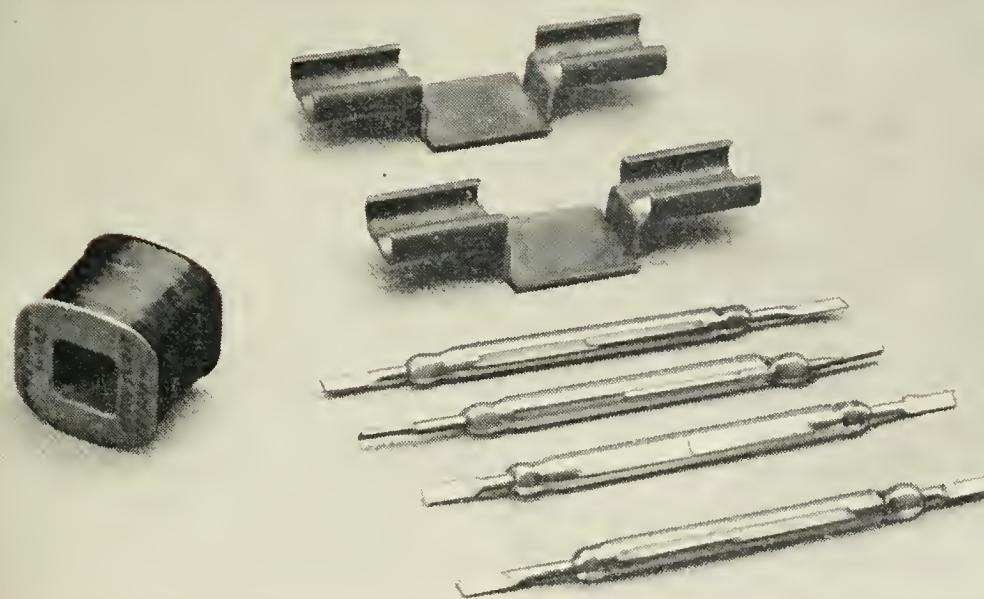
3. NEW DEVICES EMPLOYED

Numerous products of research and development were available for this new approach. Only those chosen will be described.

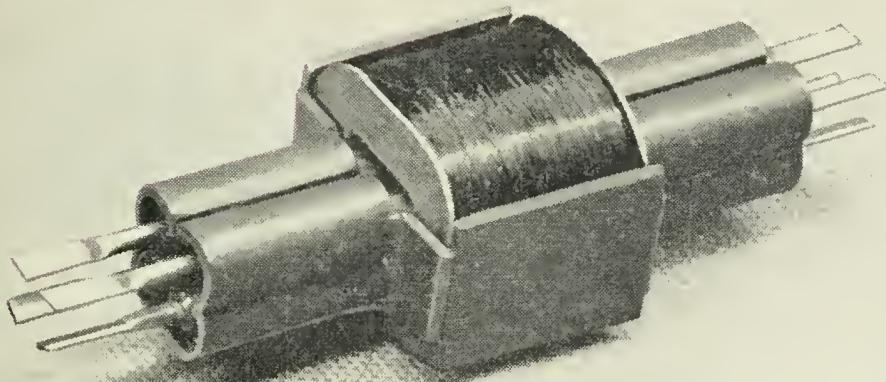
For the switching or "crosspoint" element itself, the sealed reed switch was chosen, primarily because of its imperviousness to dirt.⁸ A short coil magnet with magnetic shield for increasing sensitivity of the reed switches were used to form a relay per crosspoint (see Fig. 1).

A number of switching applications^{9,10} for crosspoint control using small gas diodes have been proposed by E. Bruee of our Switching Research Department. They are particularly advantageous when used in an "end marking" arrangement with reed relay crosspoints. Also, these diodes have long life and are low in cost. One gas diode is employed for operating each crosspoint (see Fig. 6). Its breakdown voltage is 125v \pm 10v. A different tube is used in the concentrator for detecting marking potentials when termination occurs. Its breakdown potential is 100v \pm 10v. One of these tubes is used on each connection.

Signaling between the remote concentrator and the central office control circuits is performed on a sequential basis with pulses indicative of the various line conditions being transmitted at a 500 cycle rate. This frequency encounters relatively low attenuation on existing exchange area wire facilities and yet is high enough to transmit and receive information at a rate which will not decrease call carrying capacity of the



A



B

Fig. 1 — Reed switch relay.

central office equipment. To accomplish this signaling and to process the information economically transistors appear most promising.

Germanium alloy junction transistors were chosen because of their improved characteristics, reliability, low power requirements, and margins, particularly when used to operate with relays.¹³ Both N-P-N and P-N-P transistors are used. High temperature characteristics are particularly important because of the ambient conditions which obtain on pole mounted equipment. As the trials of this equipment have progressed,

TABLE I—TRANSISTOR CHARACTERISTICS

Code No.	Type and Filling	Alpha	Max. I_{CO} at 28V and 65°C	Emitter Zener Voltage at 20 μ A
M1868	p-n-p Oxygen	0.9-1.0	150 μ A	>735
M1887	n-p-n Vacuum	0.5-.75	100 μ A	>735

considerable progress has been made in improving transistors of this type. Table I summarizes the characteristics of these transistors.

For directing and analyzing the pulses, the control employs semiconductor diode gate circuits.¹¹ The semiconductor diodes used in these circuits are of the silicon alloy junction type.¹⁵ Except for a few diodes operating in the gas tube circuits most diodes have a breakdown voltage requirement of 27v, a minimum forward current of 15 ma at 2v and a maximum reverse current at 22v of 2×10^{-8} amp.

4. NEW TECHNIQUES EMPLOYED

The concentrator represents the first field application in Bell System telephone switching systems which departs from current practices and techniques. These include:

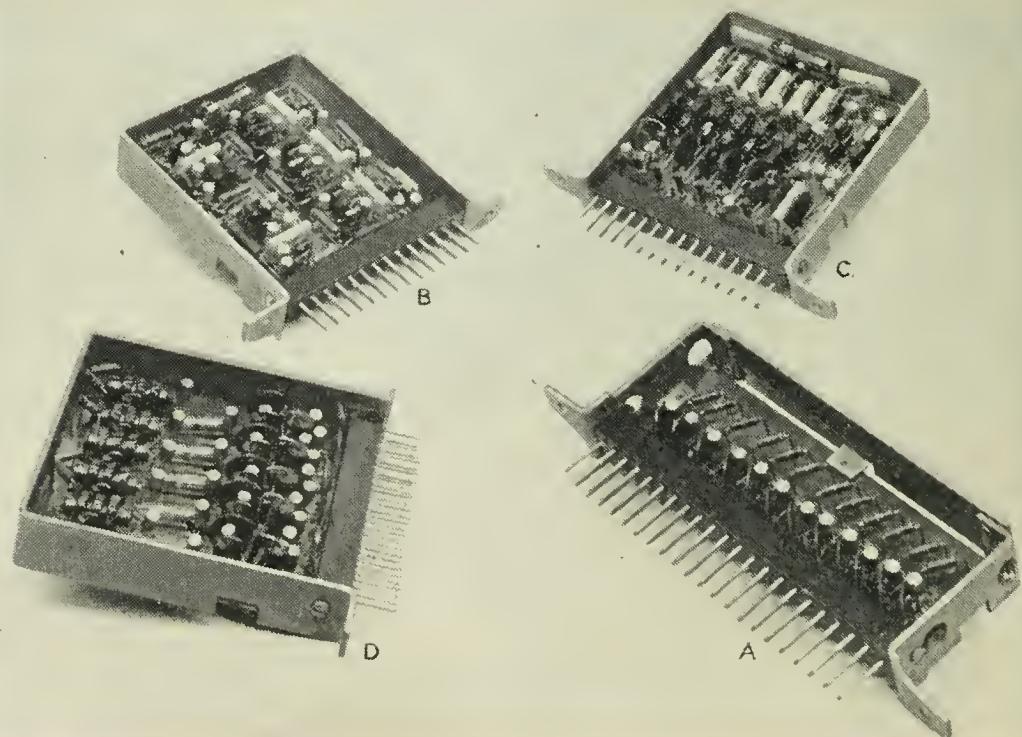


Fig. 2—Transistor packages. (a) Diode unit. (b) Transistor counter. (c) Transistor amplifiers and bi-stable circuits. (d) Five trunk unit.

1. High speed pulsing (500 pulses per second) of information between switching units.
2. The use of plug-in packages employing printed wiring and encapsulation. (Fig. 2 shows a representative group of these units.)
3. Line scanning for supervision with a passive line circuit. In present systems each line is equipped with a relay circuit for detecting call originations (service requests) and another relay (or switch magnet) for indicating the busy or idle condition of the line, as shown in Fig. 3(a). The line concentrator utilizes a circuit consisting of resistors and semiconductor diodes in pulse gates to provide these same indications. This circuit is shown in Fig. 3(b). Its operation is described later. The pulses for each line appear at a different time with respect to one another. These pulses are said to represent "time slots." Thus a different line is examined each .002 second for a total cycle time (for 60 lines) of .120 second. This process is known as "line scanning" and the portion of the circuit which produces these pulses is known as the scanner. Each of the circuits perform the same functions, viz., to indicate to the central office equipment when the customer originates a call and for terminating calls to indicate if the line is busy.
4. The lines are divided for control and identification purposes into twelve groups of five lines each. Each group of five lines has a different pattern of access to the trunks which connect to the central office. The ten trunks to the central office are divided into two groups as shown in Fig. 4. One trunk group, called the random access group, is arranged in a random multiple fashion, so that each of these trunks is available to approximately one-half of the lines. The other group, consisting of two trunks, is available to all lines and is therefore called the full access group. The control circuitry is arranged to first select a trunk of the random access group which is idle and available to the particular line to which a connection is to be made. If all of the trunks of this random access group are busy to a line to which a connection is desired, an attempt is then made to select a trunk of the full access group. The preference order for selecting cross-points in the random access group is different for each line group, as shown in the table on Fig. 4. By this means, each trunk serves a number of lines on a different priority basis. Random access is used to reduce by 40 per cent the number of individual reed relay crosspoints which would otherwise be needed to maintain the quality of service desired, as indicated by a theory presented some years ago.¹²
5. Built-in magnetic tape means for recording usage data and making call delay measurements. The gathering of this data is greatly facilitated by the line scanning technique.

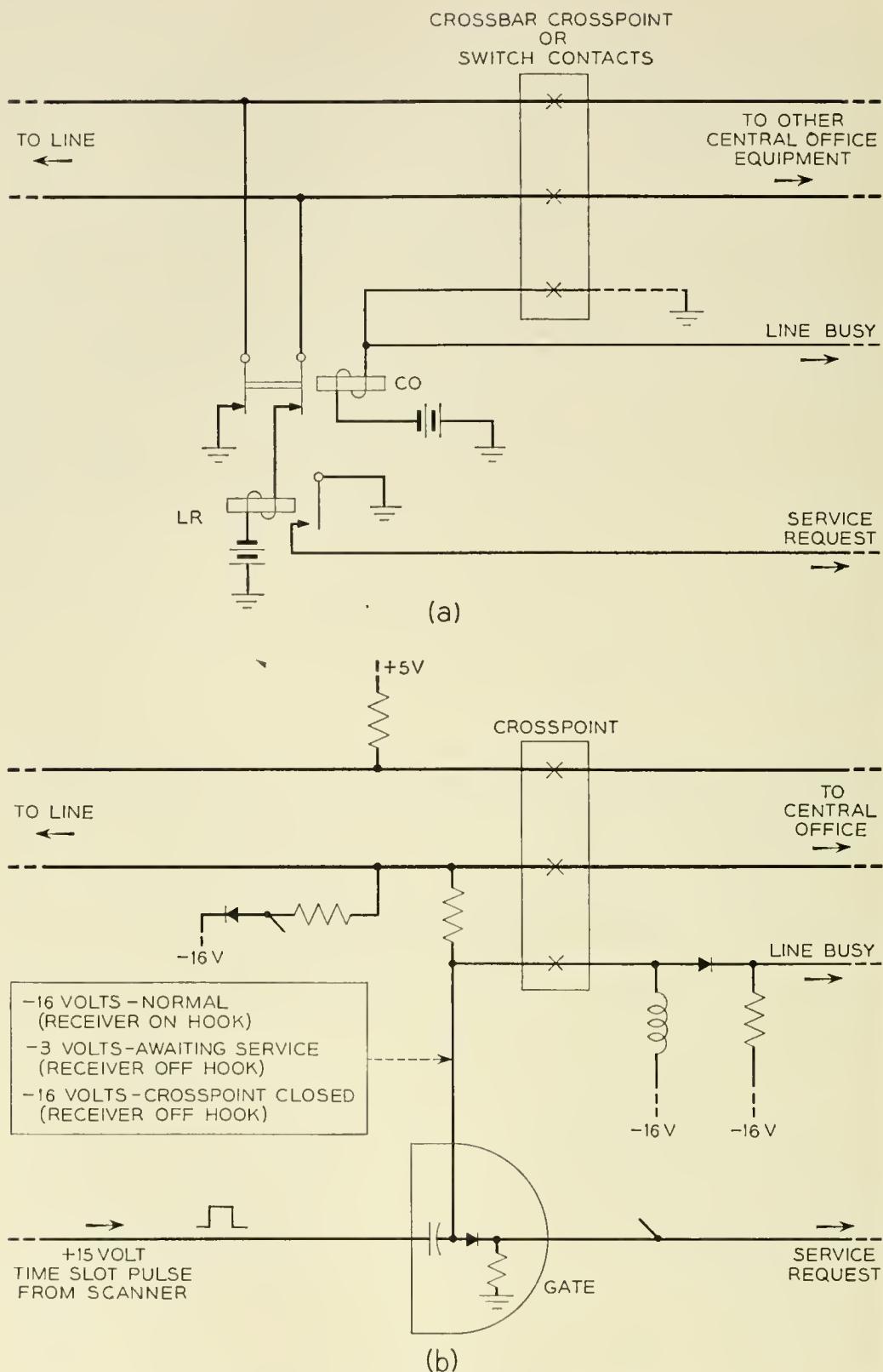


Fig. 3 — (a) Relay line circuit. (b) Passive line circuit.

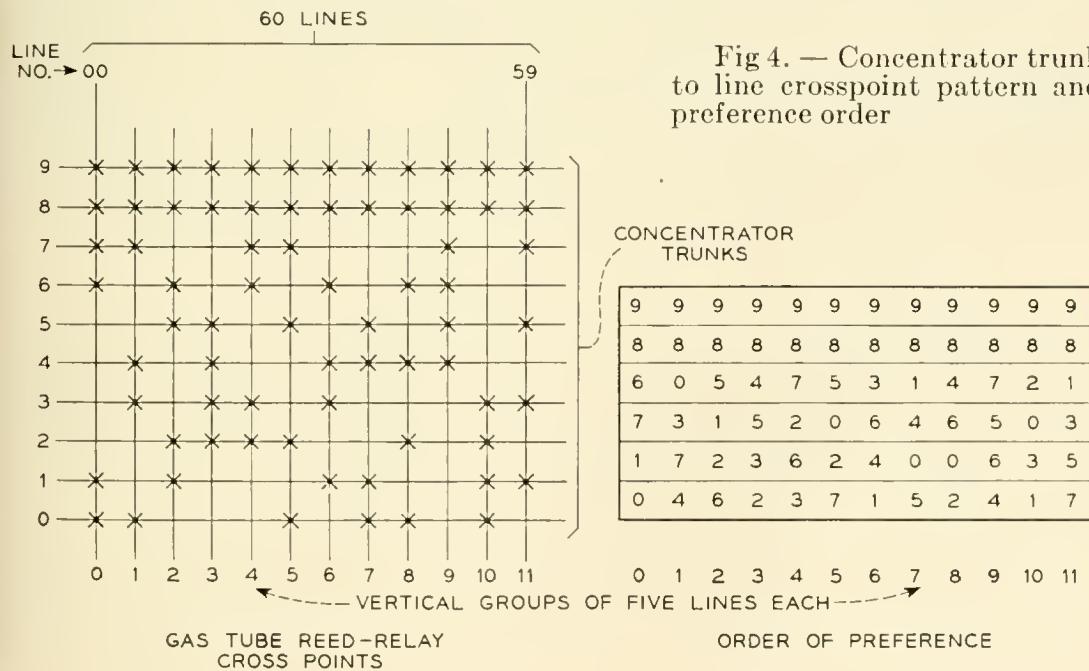
5. SWITCHING PLAN

The plan for serving lines directly terminating in a No. 5 Crossbar office is shown in Fig. 5(a). Each line has access through a primary line switch to 10 line links. The line links couple the primary and secondary switches together so that each line has access to all of the 100 junctors to the trunk link switching stage. Each primary line switch group accommodates from 19 to 59 lines (one line terminal being reserved for no-test calls). A line link frame contains 10 groups of primary line switches.¹⁴

The remote concentrator plan merely extends these line links as trunks to the remote location. However, an extra crossbar switching stage is introduced in the central office to connect the links to the secondary line switches with the concentrator trunks as shown in Fig. 5(b). Since each line does not have full access to the trunks, the path chosen by the marker to complete calls through the trunk link frame may then be independent of the selection of a concentrator trunk with access to the line. This arrangement minimizes call blocking, simplifies the selection of a matched path by the marker, and the additional crossbar switch hold magnet serves also as a supervisory relay to initiate the transmission of disconnect signals over the trunk.

In addition to the 10 concentrator trunks used for talking paths, 2 additional cable pairs are provided from each concentrator to the central office for signaling and power supply purposes. The use of these two pairs of control conductors is described in detail in Section 6g.

The concentrator acts as a slave unit under complete control of the central office. The line busy and service request signals originate at the



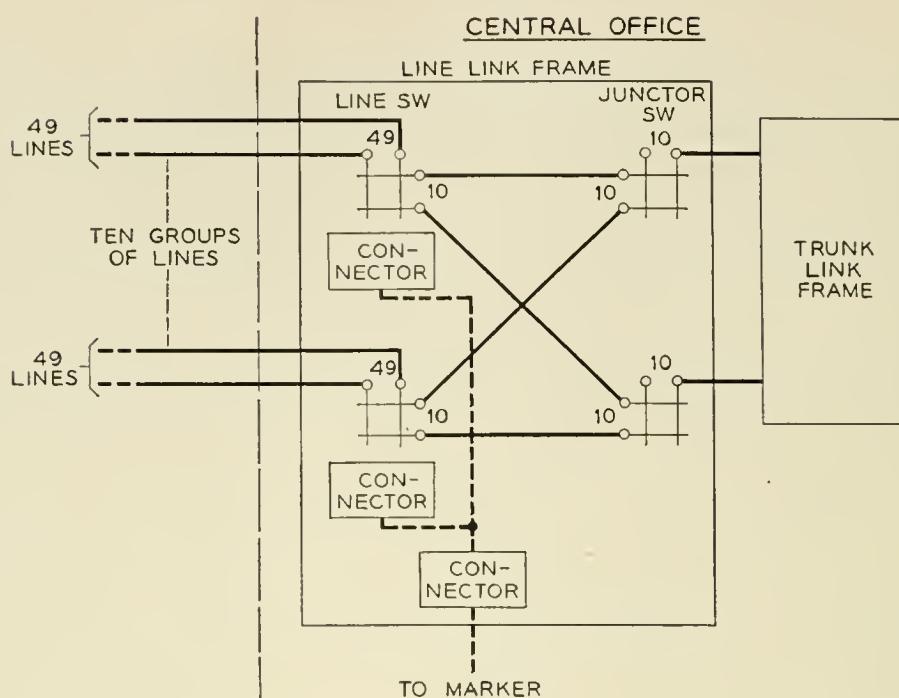


Fig. 5(a) — No. 5 crossbar system subscriber lines connected to line link frame.

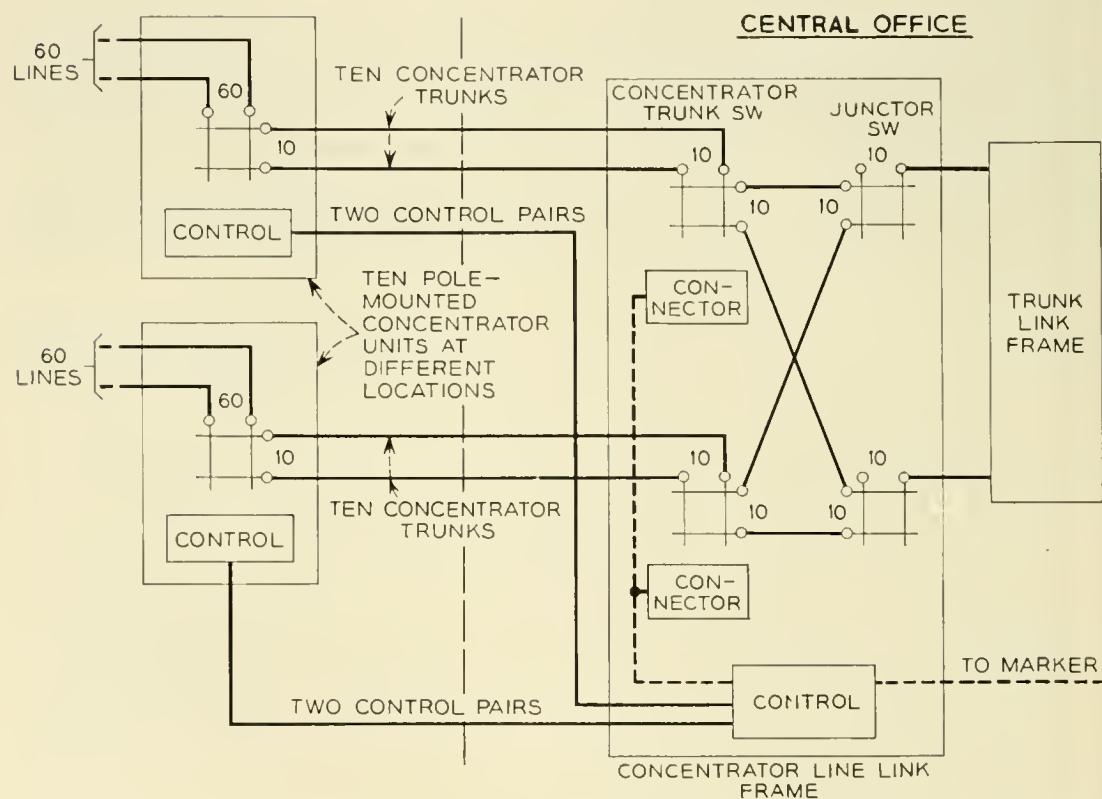


Fig. 5(b) — No. 5 crossbar system subscriber lines connected to remote line concentrators.

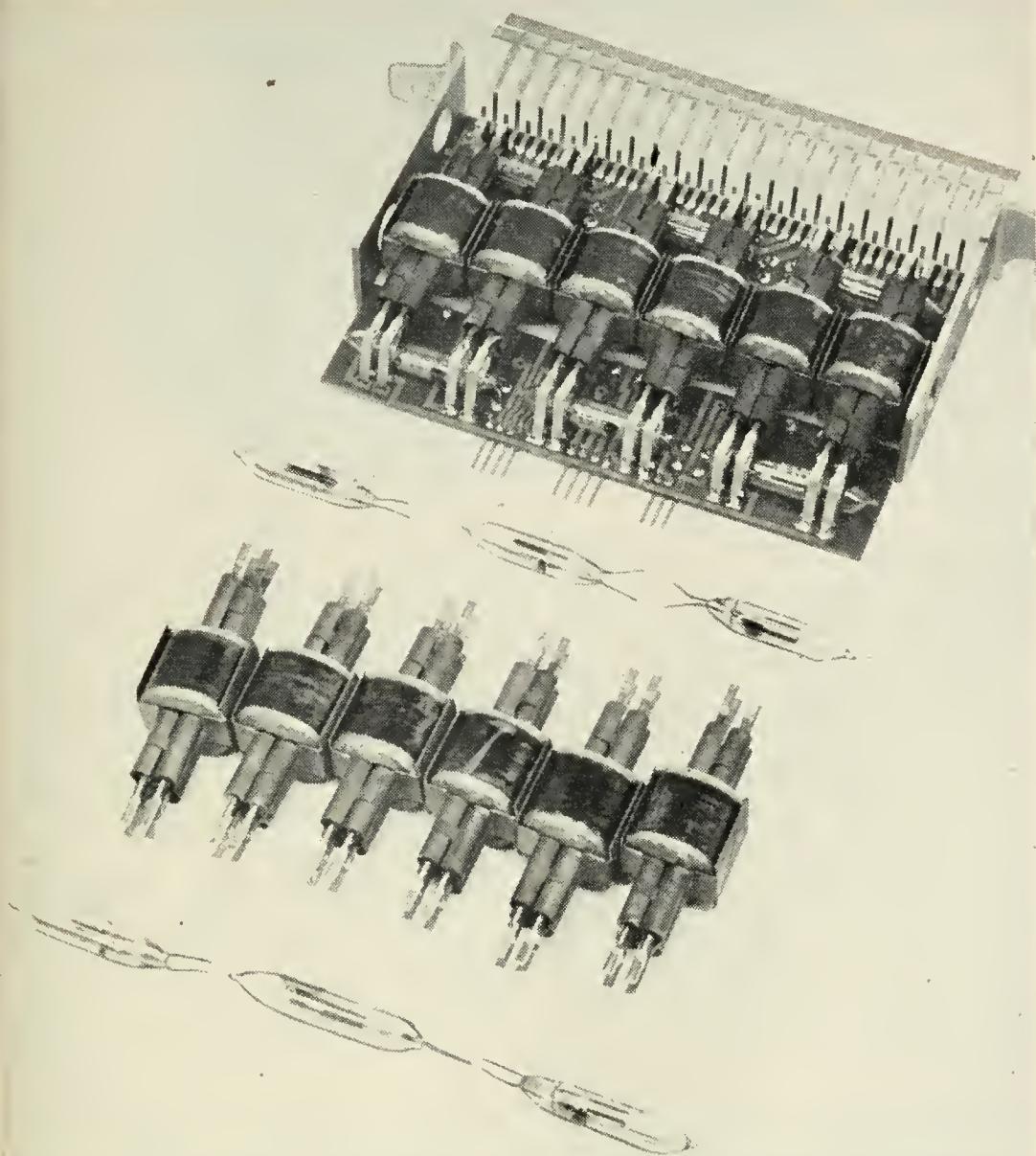


Fig. 6 — Line unit construction.

concentrator only in response to a pulse in the associated time slot or when a crosspoint operates (a line busy pulse is generated under this condition as a crosspoint closure check). The control circuit in the central office is designed to serve 10 remote line concentrators connected to a single line link frame. In this way the marker deals with a concentrator line link frame as it would with a regular line link frame and the marker modifications are minimized.

The traffic loading of the concentrator is accomplished by fixing the



Fig. 7(a) — Line unit.

number of trunks at 10 and equipping or reassigning lines as needed to obtain the trunk loading for the desired grade of service. The six cross-points, the passive line circuit and scanner gates individual to each line are packaged in one plug-in unit to facilitate administration. The cross-points are placed on a printed wiring board together with a comb of plug contacts as shown in Fig. 6. The entire unit is then dipped in rubber and encapsulated in epoxy resin, as shown in Fig. 7(a).

This portion of the unit is extremely reliable and therefore it may be considered as expendable, should a rare case of trouble occur. The passive line circuit and scanner gate circuit elements are mounted on a smaller second printed wiring plate (known as the "line scanner" plate, see Fig. 7(b) which fits into a recess in the top of the encapsulated line unit. Cir-

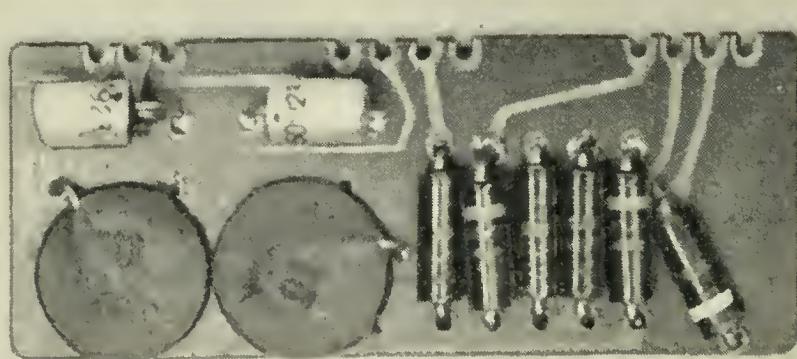


Fig. 7(b) — Scanner plate of the line unit shown in Fig. 7 (a).

cuit connection between printed wiring plates is through pins which appear in the recess and to which the smaller plate is soldered.

6. BASIC CIRCUITS

a. Diode Gates

All high speed signaling is on a pulse basis. Each pulse is positive and approximately 15 volts in amplitude. There is one basic type of diode gate circuit used in this equipment. By using the two resistors, one condenser and one silicon alloy junction diode in the gate configuration shown in Fig. 8, the equivalents of opened or closed contacts in relay circuits are obtained. These configurations are known respectively as enabling and inhibiting gates and are shown with their relay equivalents in Figs. 8(a) and 8(b).

In the enabling gate the diode is normally back biased by more than the pulse voltage. Therefore pulses are not transmitted. To enable or

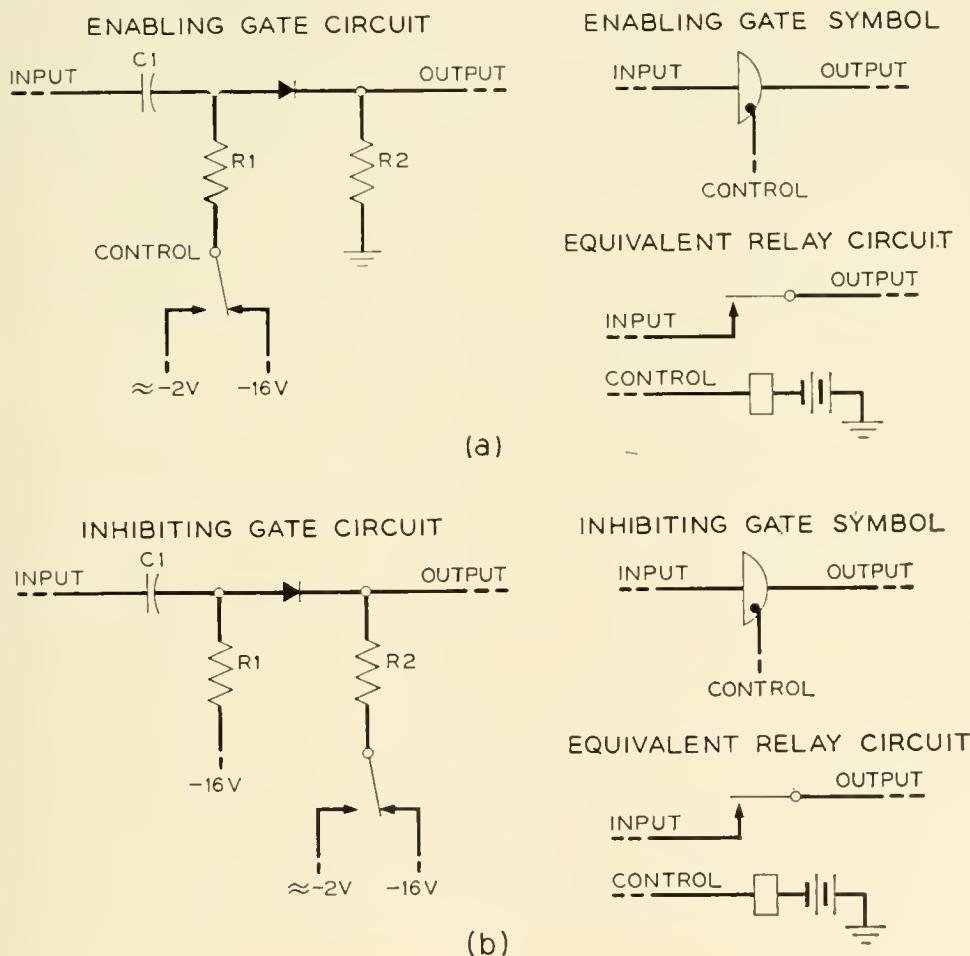


Fig. 8 — Gates and relay equivalents.

open the gate the back bias is reduced to a small reverse voltage which is more than overcome by the signal pulse amplitude of the pulse. The pulse thus forward biases the diode and is transmitted to the output.

The inhibiting gate has its diode normally in the conducting state so that a pulse is readily transmitted from input to output. When the bias is changed the diode is heavily back biased so that the pulse amplitude is insufficient to overcome this bias.

The elements of 12 gates are mounted on a single printed wiring board with plug-in terminals and a metal enclosure as shown in Fig. 2(a). All elements are mounted in one side of the board so that the opposite side may be solder dipped. After soldering the entire unit (except the plug) is dipped in a silicone varnish for moisture protection.

b. *Transistor Bistable Circuit*

Transistors are inherently well adapted to switching circuits using but two states, on (saturated) or off.¹⁶ In these circuits with a current gain greater than unity a negative resistance collector characteristic can be obtained which will enable the transistor to remain locked in its conducting state (high collector current flowing) until turned off (no collector current) by an unlocking pulse. At the time the concentrator development started only point contact transistors were available in quantity. Point contact transistors have inherently high current gains (>1) but the collector current flowing when in the normal or unlocked condition (I_{co}) was so great that at high ambient temperatures a relay once operated in the collector circuit would not release.

Junction transistors are capable of a much greater ratio of on to off current in the collector circuit. Furthermore their characteristics are amenable to theoretical design consideration.¹³ However, the alpha of a simple junction transistor is less than unity. To utilize them as one would a point contact transistor in a negative resistance switching circuit, a combination of n-p-n and p-n-p junction transistors may be employed, see Fig. 9(b). Two transistors combined in this manner constitute a "hooked junction conjugate pairs." This form of bi-stable circuit was used because it requires fewer components and uses less power than an Eccles-Jordan bistable circuit arrangement. It has the disadvantage of a single output but this was not found to be a shortcoming in the design of circuits employing pulse gates of the type described. In what follows the electrodes of the transistor will be considered as their equivalents shown in Fig. 9(b).

The basic bi-stable circuit employed is shown in Fig. 10. The set

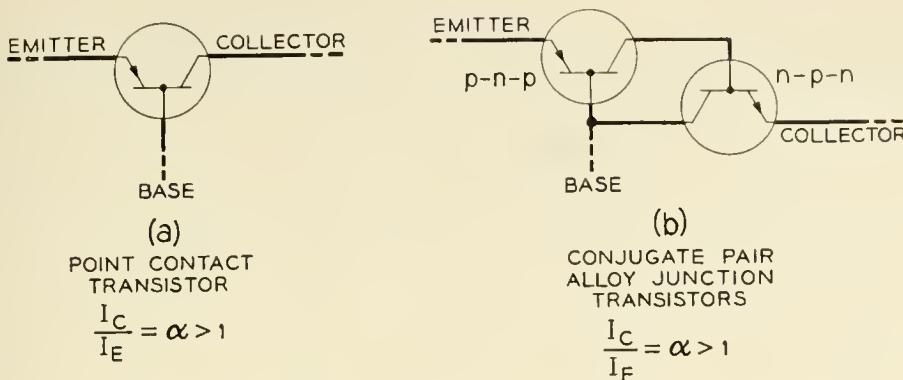


Fig. 9 — Point contact versus hooked conjugate pair.

pulse is fed into the emitter (of the pair) causing the emitter diode to conduct. The base potential is increased thus increasing the current flowing in the collector circuit. When the input pulse is turned off the base is left at about -2 volts thus maintaining the emitter diode conducting and continuing the increased current flow in the collector circuit. The diode in the collector circuit prevents the collector from going positive and thereby limits the current in the collector circuit. To reset, a positive pulse is fed into the base through a pulse gate. The driving of the base positive returns the transistor pair to the off condition.

c. Transistor Pulse Amplifier

This circuit (Fig. 11) is formed by making a bi-stable self resetting circuit. It is used to produce a pulse of fixed duration in response to a

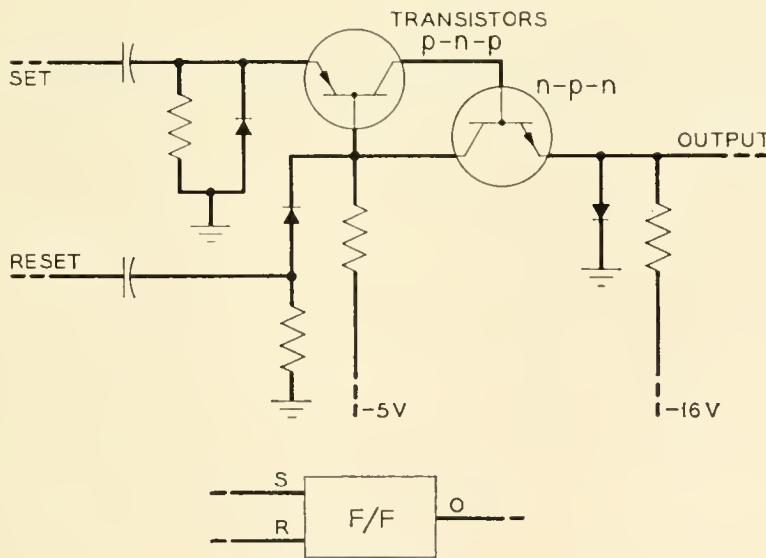


Fig. 10 — Transistor bi-stable circuit.

pulse of variable width (within limits) on the input. Normally the emitter is held slightly negative with respect to the base. The potential difference determines the sensitivity of the amplifier. When a positive input pulse is received, the emitter diode conducts causing an increase in collector current. The change in bias of the diode in the emitter circuit permits it to conduct and charge the condenser. With the removal of the input pulse the discharge of the condenser holds the transistor pair on. The time constant of the circuit determines the on time. When the emitter potential falls below the base potential, the transistor pair is turned off.

The amplifiers and bi-stable circuits or flip-flops, as they are called more frequently, are mounted together in plug-in packages. Each package contains 8 basic circuits divided 7-1, 6-2, or 2-6, between amplifiers and flip-flops. Fig. 2(c) shows one of these packages. They are smaller than the gate or line unit packages, having only 28 terminals instead of 42.

The transistors for the field trial model were plugged into small hearing aid sockets mounted on the printed wiring boards. For a production model it would be expected that the transistors would be soldered in.

d. Transistor Ring Counter

By combining bi-stable transistor and diode pulse gate circuits together in the manner shown in Fig. 12 a ring counter may be made, with

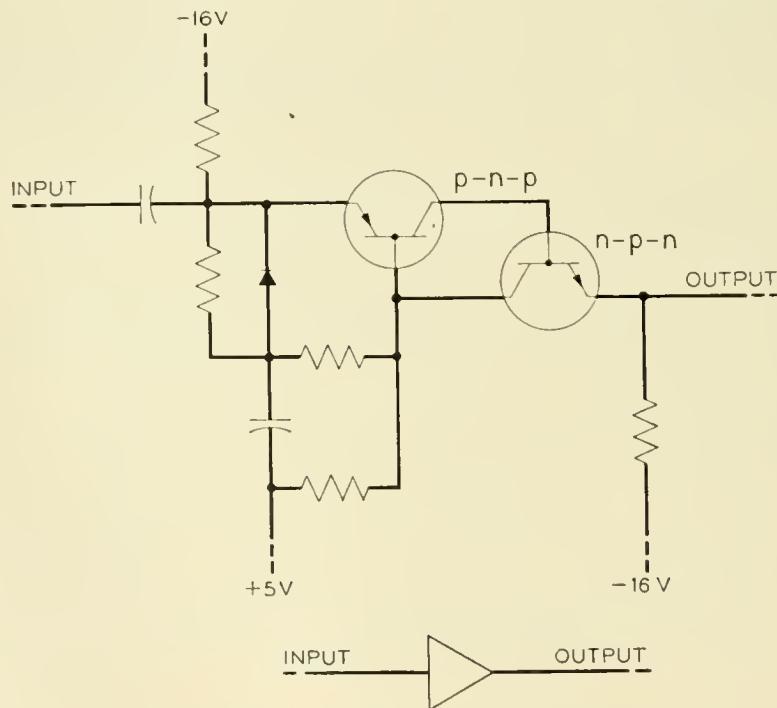


Fig. 11 — Transistor pulse amplifier.

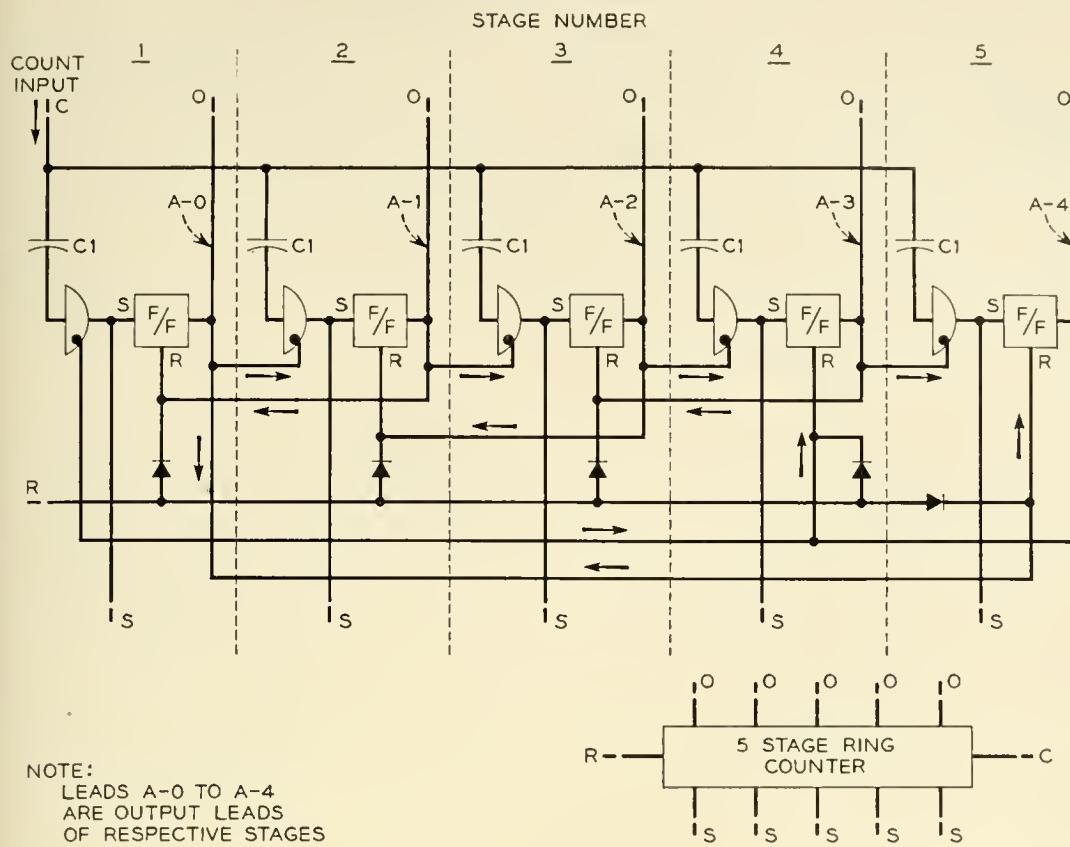


Fig. 12 — Ring counter schematic.

a bi-stable circuit per stage. The enabling gate for a stage is controlled by the preceding stage allowing it to be set by an input advance pulse. The output signal from a stage is fed back to the preceding stage to turn it off. An additional diode is connected to the base of each stage for resetting when returning the counter to a fixed reference stage.

A basic package of 5 ring counter stages is made up in the same framework and with the same size plug as the flip-flop and amplifier packages, see Fig. 2(b). A four stage ring counter is also used and is the same package with the components for one stage omitted. The input and output terminals of all stages are available on the plug terminals so that the stages may be connected in any combination and form rings of more than 5 stages. The reset lead is connected to all but the one stage which is considered the first or normal stage.

Other transistor circuits such as binary counters and square wave generators are used in small quantity in the central office equipment. They will not be described.

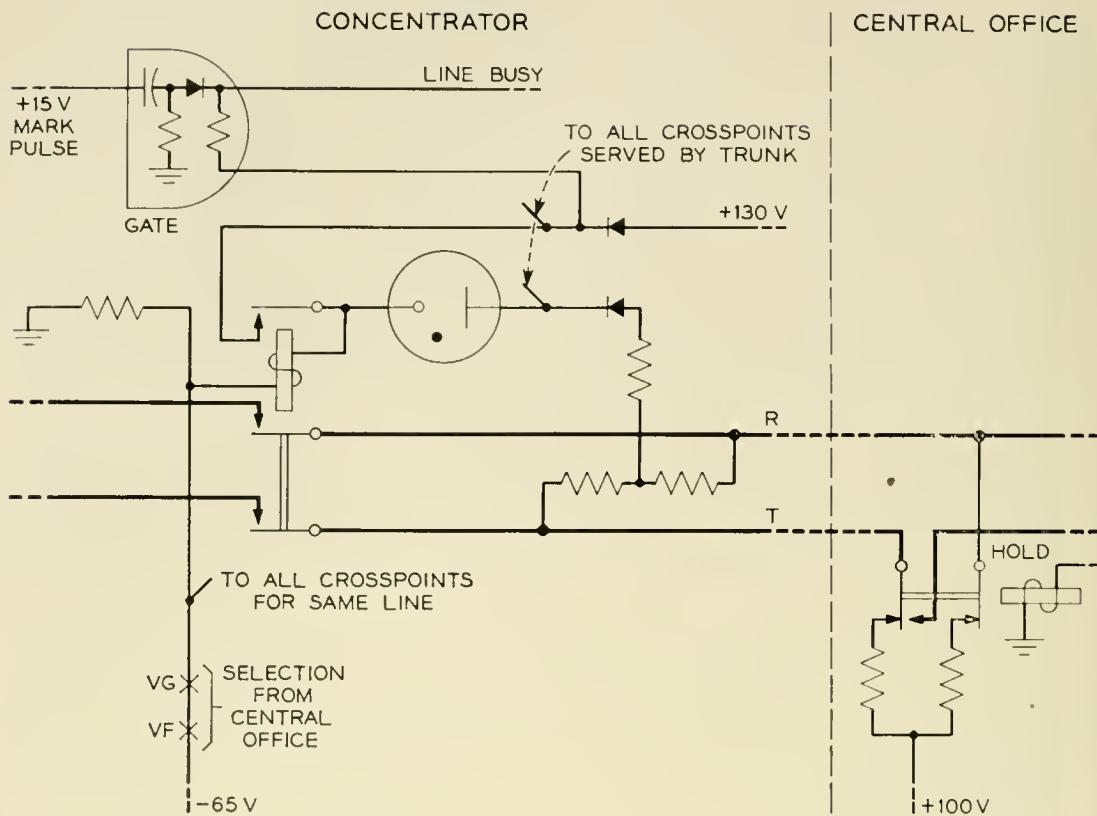


Fig. 13 — Crosspoint operating circuit.

e. Crosspoint Operating Circuit

The crosspoint consists of a reed relay with 4 reed switches and a gas diode (Fig. 1). The selection of a crosspoint is accomplished by marking with a negative potential (-65 volts) all crosspoints associated with a line, and marking with a positive potential (+100 volts) all crosspoints associated with a trunk (Fig. 13). The line is marked through a relay circuit set by signals sent over the control pair from the central office. The trunk is marked by a simplex circuit connected through the break contacts of the hold magnet of the crossbar switch associated with the trunk in the central office. Only one crosspoint at a time is exposed to 165 volts which is necessary and sufficient to break down the gas diode to its conducting state. The reed relay operates in series with the gas diode. A contact on the relay shunts out the gas diode. When the marking potentials are removed the relay remains energized in a local 30-volt circuit at the concentrator. The holding current is approximately 2.5 ma.

This circuit is designed so that ringing signals in the presence or absence of line marks will not falsely fire a crosspoint diode. Furthermore,

a line or trunk mark alone should not be able to fire a crosspoint diode on a busy line or trunk.

When the crosspoint operates, a gate which has been inhibiting pulses is forward biased by the -65 volt signal through the crosspoint relay winding. The pulse which initiates the mark operations at the concentrator then passes through the gate to return a line busy signal to the central office over this control pairs which is interpreted as a crosspoint closure check signal.

f. Crosspoint Release Circuit

The hold magnet of the central office crossbar switch operates, removing the +100-volt operate mark signal after the crosspoint check signal is received. A slow release relay per trunk is operated directly by the hold magnet. When the central office connection in the No. 5 crossbar system releases, the hold magnet is released. As shown in Fig. 14, with the hold magnet released and the slow release relay still operated, a -130-volt signal is applied in a simplex circuit to the trunk to break down a gas tube provided in the trunk circuit at the concentrator. This tube in

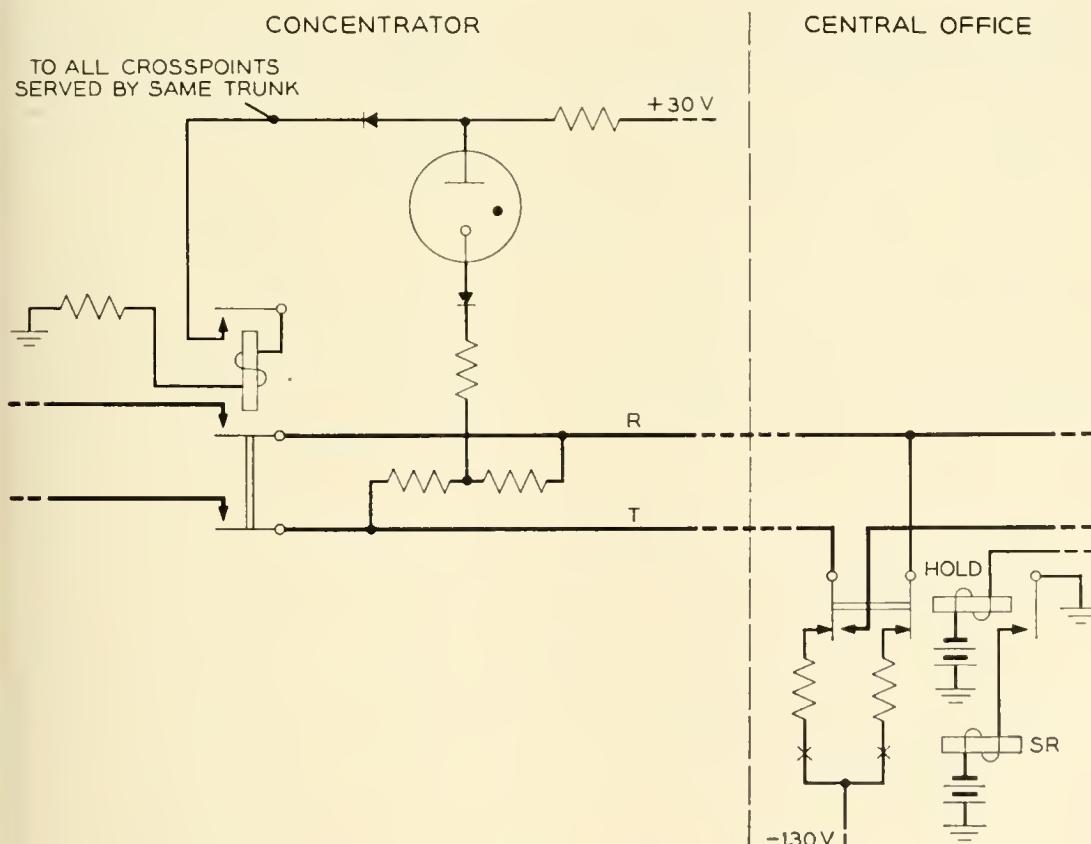


Fig. 14 — Crosspoint release circuit.

breaking down shunts the local holding circuit of the crosspoint causing it to release. The -130-volt disconnect signal is applied during the release time of the slow release relay which is long enough to insure the release of the crosspoint relay at the concentrator.

The release circuit is individual to the trunk and independent of the signal sent over the control pairs.

g. Pulse Signalling Circuits

To control the concentrator four distinct pulse signals are transmitted from the central office. Two of these at times must be transmitted simultaneously, but these and the other two are transmitted mutually exclusively. In addition, service request and line busy signals are transmitted from the concentrator to the central office. The two way transmission of information is accomplished on each pair by sending signals in each direction at different times and inhibiting the receipt of signals when others are being transmitted.

To transmit four signals over two such pairs, both positive and nega-

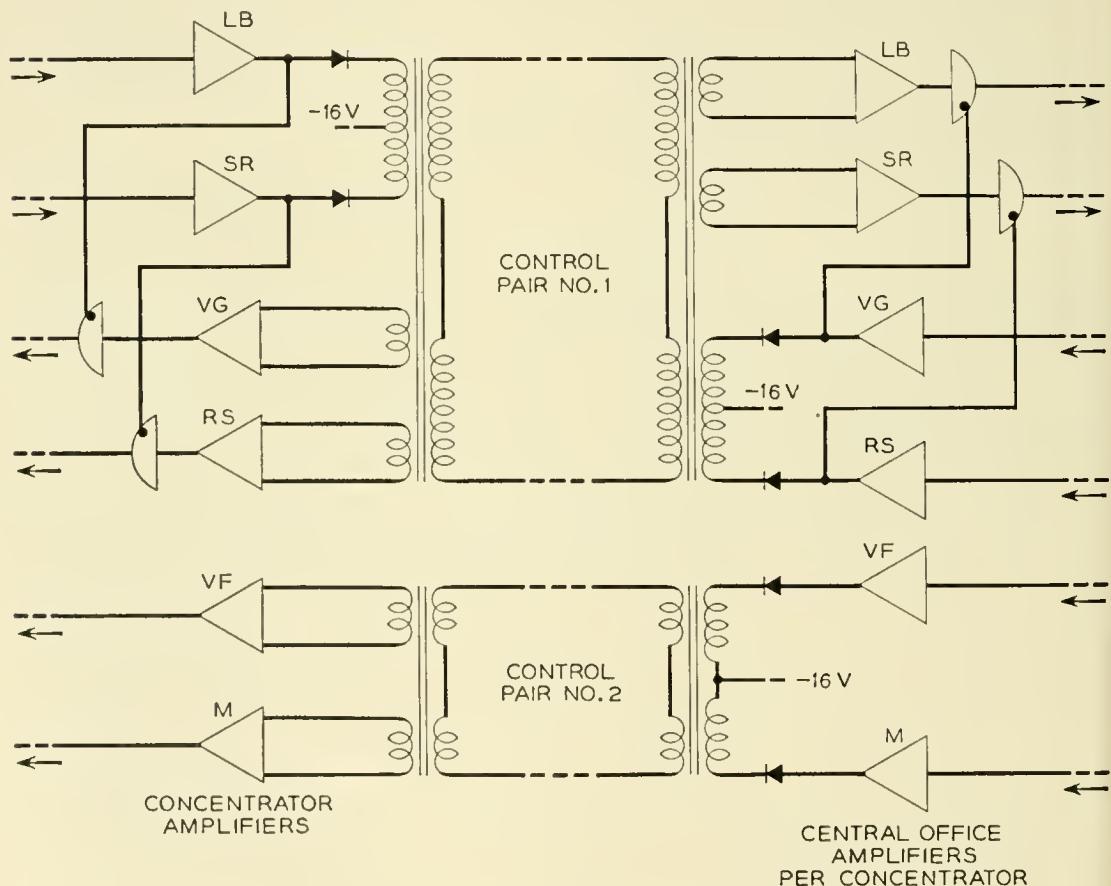


Fig. 15. — Signal transmission circuit.

tive pulses are employed. Diodes are placed in the legs of a center tapped transformer, as shown in Fig. 15, to select the polarity of the transmitted pulses. At the receiving end the desired polarity is detected by taking the signal as a positive pulse from a properly poled winding of a transformer. The amplifier, as described in Section 6c responds only to positive pulses. If pulses of the same polarity are transmitted in the other direction over the same pair, as for control pair No. 1, the outputs of the receiving amplifier for the same polarity pulse are inhibited whenever a pulse is transmitted.

As shown in Fig. 15, the service request and line busy signals are transmitted from the concentrator to the central office over one pair of conductors as positive and negative pulses respectively. The transmission of these pulses gates the outputs of two of the receiving amplifiers at the concentrator to permit the receipt of the polarized signals from the central office. This prevents the pulses from being used at the sending end. A similar gating arrangement is used with respect to the signals when sent over this control pair from the central office. The pulses designated VG or RS never occur when a pulse designated SR or LB is sent in the opposite direction. The transmission of the VF pulse over control pair No. 2 is processed by the concentrator circuit and becomes the SR or LB pulses. In section 7 the purpose of these pulses is described.

The signaling range objective is 1,200 ohms over regular exchange area cable including loaded facilities from station to central office.

h. Power Supply

Alternating current is supplied to the concentrator from a continuous service bus in the central office. The power supply path is a phantom circuit on the two control pairs as shown in Fig. 16. The power transformer has four secondary windings used for deriving from bridge rectifiers four basic dc voltages. These voltages and their uses are as follows: -16 volts (regulated) for transistor collector circuits and gate biases, +5 volts (regulated) for transistor base biases, +30 volts (regulated) for crosspoints holding circuits and -65 volts for the marking and operating of the line crosspoints. For this latter function a reference to the central office applied +100 volt trunk mark is necessary. The reference ground for the concentrator is derived from ground applied to a simplex circuit on the power supply phantom circuit. Series transistors and shunt silicon diodes with fixed reference breakdown voltages are used to regulate dc voltages.

Total power consumption of the concentrator is between 5 and 8 watts depending upon the number of connections being held.

7. CONCENTRATOR OPERATION

a. Line Scanning

The sixty lines are divided into 12 groups of 5 lines each. These groupings are designated VG and VF respectively corresponding to the vertical group and file designations used in the No. 5 crossbar system. Each concentrator corresponds to a horizontal group in that system.

To scan the lines two transistor ring counters, one of 12 stages and one of 5 stages, are employed as shown in Fig. 17. These counters are driven from pulses supplied from the central office control circuits and only one stage in each is on at any one time. The steps and combinations of these counters correspond to the group and file designation of a particular line. Each 0.002 second the five stage counter (VF) takes a step and between the fifth and sixth pulse the 12-stage counter (VG) is stepped. Thus the 5-stage counter receives 60 pulses or re-cycles 12 times in 120 milliseconds while the 12-stage counter cycles but once.

Each line is provided with a scanner gate. The collector output of each each stage of the VG counter biases this gate to enable pulses which are generated by the collector circuit of the 5-stage counter to pass on

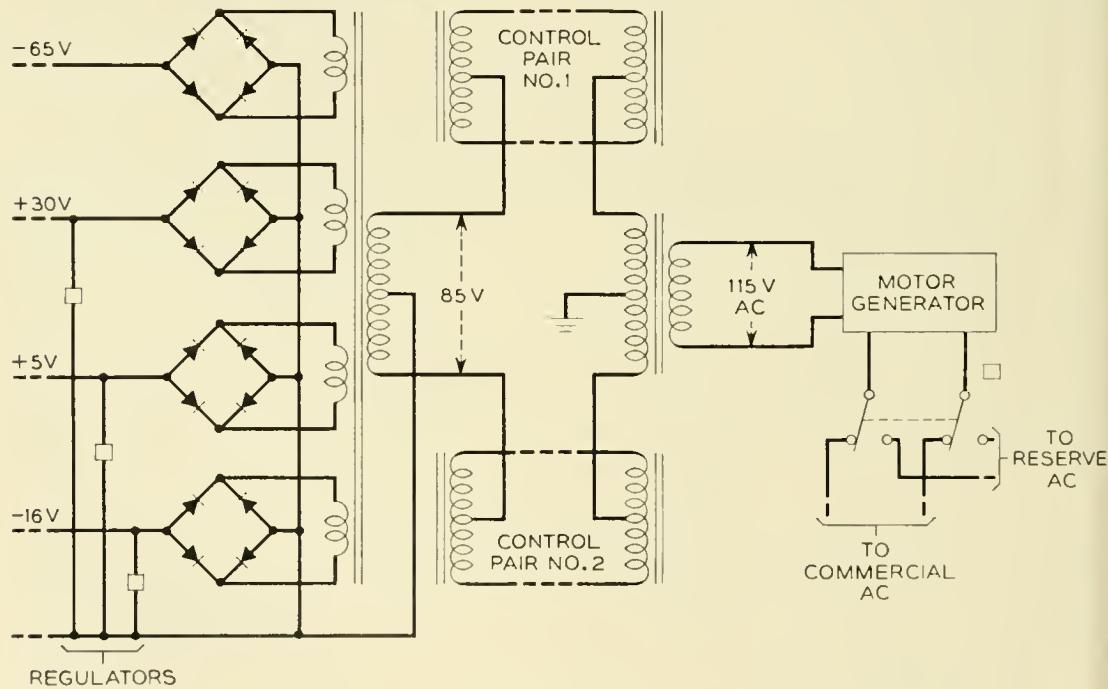


Fig. 16 — Power supply transmission circuit.

to the gate of the passive line circuit, Fig. 3(b). If the line is idle the pulses are inhibited. If the receiver is off-hook requesting service (no crosspoint closed) then the gate is enabled, the pulse passes to the service request amplifier and back to the central office in the same time slot as the pulse which stepped the VF counter. If the line has a receiver off-hook and is connected to a trunk the pulse passes through a contact of the crosspoint relay to the line busy amplifier and then to the central office in the same time slot.

At the end of each complete cycle a reset pulse is sent from the central office. This pulse instead of the VG pulse places the 12-stage counter in its first position. It also repulses the 5 stage VF counter to its fifth stage so that the next VF pulse will turn on its first stage to start the next cycle. The reset pulse insures that, in event of a lost pulse or defect in a counter stage, the concentrator will attempt to give continuous service without dependence on maintaining synchronism with the central office scanner pulse generator. Fig. 18(a) shows the normal sequence of line scanning pulses.

When a service request pulse is generated, the central office circuits

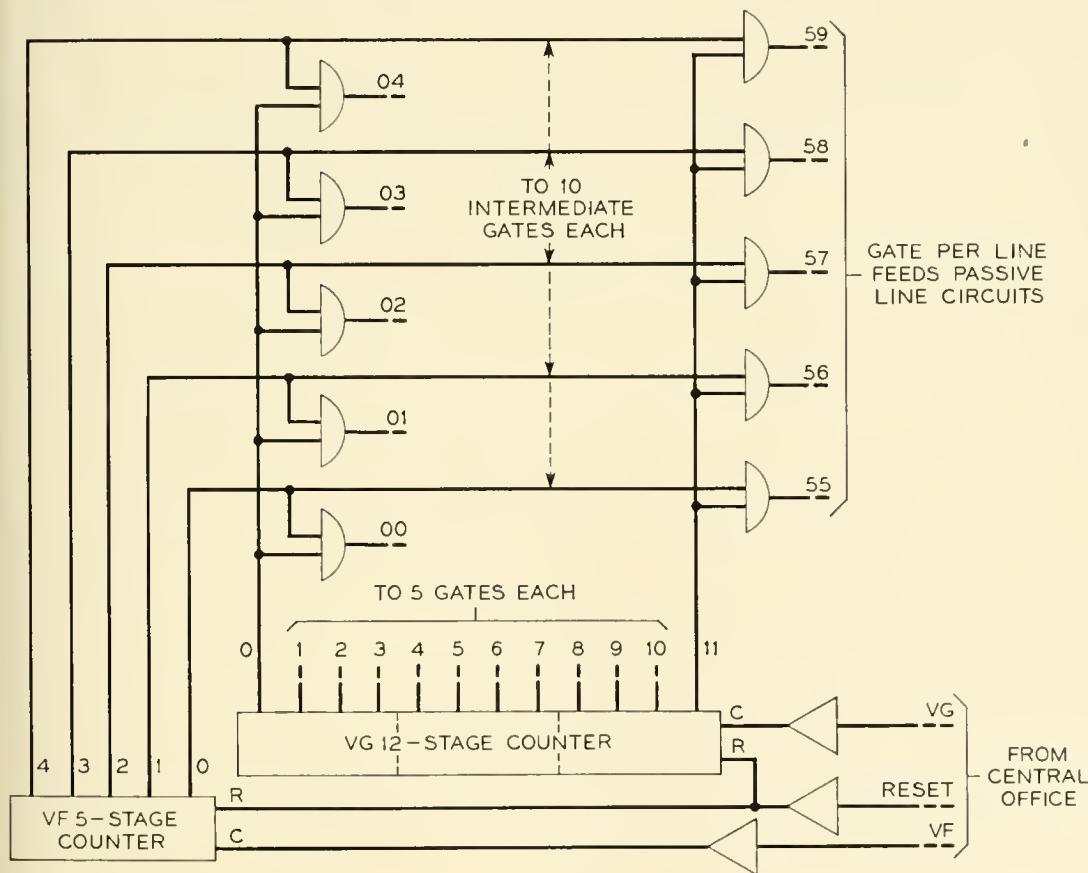


Fig. 17 — Diode matrix for scanning lines.

common to 10 concentrators interrupt the further transmission of the vertical group pulse so that the line scanning is confined to the 5 lines in the vertical group in which the call originated. In this way the central office will receive a service request pulse at least every 0.010 sec as a check that the call has not been abandoned while awaiting service. Fig. 18(b) shows the detection of a call origination and the several short scan cycles for abandoned call detection.

b. Line Selection

When the central office is ready to establish a connection at the concentrator a reset pulse is sent to return the counters to normal. In general, the vertical group and vertical file pulses are sent simultaneously to reduce holding time of the central office equipment and to minimize marker delays caused by this operation. For this reason the VG and VF pulses are each transmitted over different control pairs from the central office. The same polarity is used.

On originating calls it is desirable to make one last check that the call has not been abandoned, while on terminating calls it is necessary

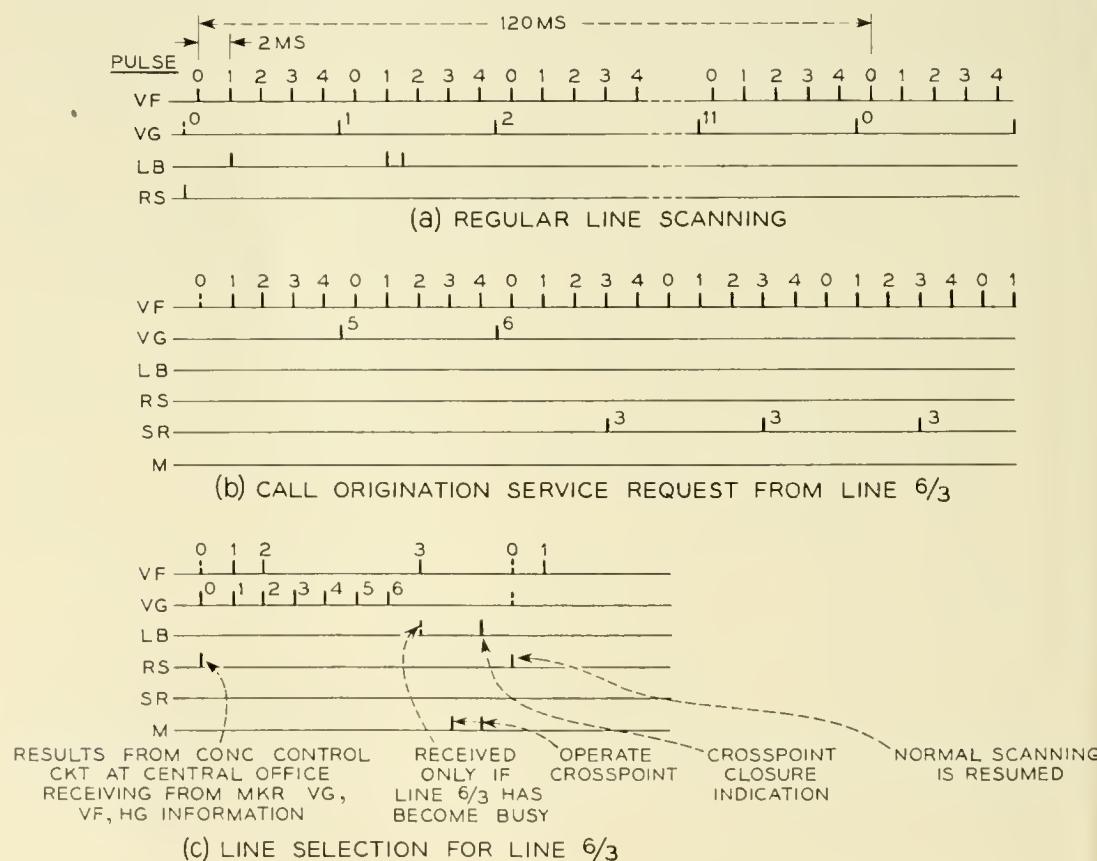


Fig. 18 — Pulse sequences. (a) Regular. (b) Call origination. (c) Line selection.

to determine if the line is busy or idle. These conditions are determined in the same manner as described for line scanning since a service request condition would still prevail on the line if the call was not abandoned. If the line was busy, a line busy condition would be detected. However to detect these conditions a VF pulse must be the last pulse transmitted since the stepping of the VF counter generates the pulse which is transmitted through an enabled line selection and passive line circuit gates. Fig. 18(c) shows a typical line selection where the number of VF pulses is equal to or less than the number of VG pulses. In all other cases there is no conflict and the sending of the last VF pulse need not be delayed. On terminating calls, the line busy indication is returned to the central office within 0.002 sec after the selection is complete. During selections the central office circuits are gated to ignore any extraneous service request or line busy pulses produced as a result of steps of the VF counter prior to its last step.

c. Crosspoint Operation and Check

Associated with each concentrator transistor counter stage is a reed relay. These relays are connected to the transistor collector circuits through diodes of the counter stages when relay M operates. The contacts of these reed relays are arranged in a selection circuit as shown in Fig. 19 and apply the -65 volt mark potential to the crosspoint relays of the selected line.

After a selection is made as described above a "mark" pulse is sent from the central office. This pulse is transmitted as a pulse of a different polarity over the same control pair as the VF pulses. The received pulse after amplification actuates a transistor bistable circuit which has the M reed relay permanently connected in its collector circuit. The bi-stable circuit holds the M relay operated during the crosspoint operation to maintain one VF and one VG relay operated, thereby applying -65 volts to mark and operate one of the 6 crosspoint relays of the selected line as described in section 6e, and shown on Fig. 13.

The operation and locking of the crosspoint relay with the marking potentials still applied enables a pulse gate associated with the holding circuit of the crosspoint relays in each trunk circuit. The mark pulses are sent out continuously. This does not affect the bi-stable transistor circuit once it has triggered but the mark pulse is transmitted through the enabled crosspoint closure check gate shown in Fig. 20 and back to the central office as a line busy signal.

With the receipt of the crosspoint closure check signal the sending

of the mark pulses is stopped and a reset pulse is sent to the concentrator to return the mark bi-stable circuit, counters and all operated selector relays to normal. The concentrator remains in this condition until it is resynchronized with the regular line scanning cycle.

A complete functional schematic of the concentrator integrating the circuits described above is shown in Fig. 21. Fig. 22(a) and (b) show an experimental concentrator built for field tests.

S. CENTRAL OFFICE CIRCUITS

The central office circuits for controlling one or more concentrators are composed of wire spring relays as well as transistors, diode and reed

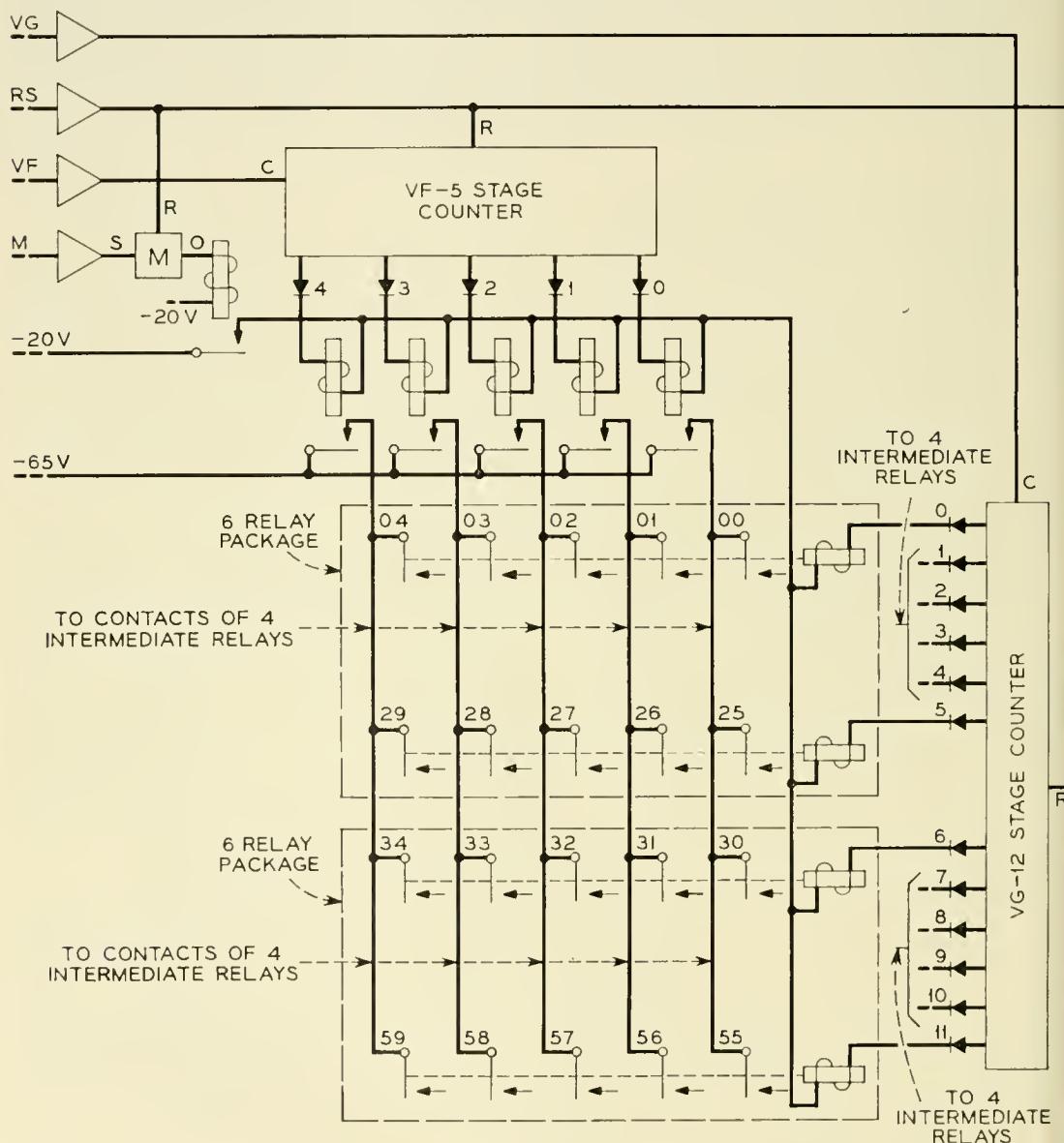


Fig. 19 — Line selection and marking

relay packages similar to those used in the concentrator. The reed relays are energized by transistor bi-stable circuits in the same manner as described in Section 7c. The reed relay contacts in turn operate wire spring relays or send the dc signals directly to the regular No. 5 crossbar marker and line link marker connector circuits.

Fig. 23 shows a block diagram of the central office circuits. A small amount of circuitry is provided for each concentrator. It consists of the following:

1. The trunk connecting crossbar switch and associated slow relays for disconnect control.
2. The concentrator control trunk circuits and associated pulse amplifiers.
3. An originating call detector to identify which concentrator among the ten served by the frame is calling.
4. A multicontact relay to connect the circuits individual to each concentrator with the common control circuits associated with the line link frame and markers.

The circuits associated with more than one concentrator are blocked out in the lower portion of Fig. 23. Much of this circuitry is similar to the relay circuits now provided on regular line link frames in the No. 5 crossbar system.³ Only those portions of these blocks which employ the new techniques will be covered in more detail. These portions consist of the following:

1. The scanner pulse generator.
2. The originating line number register.

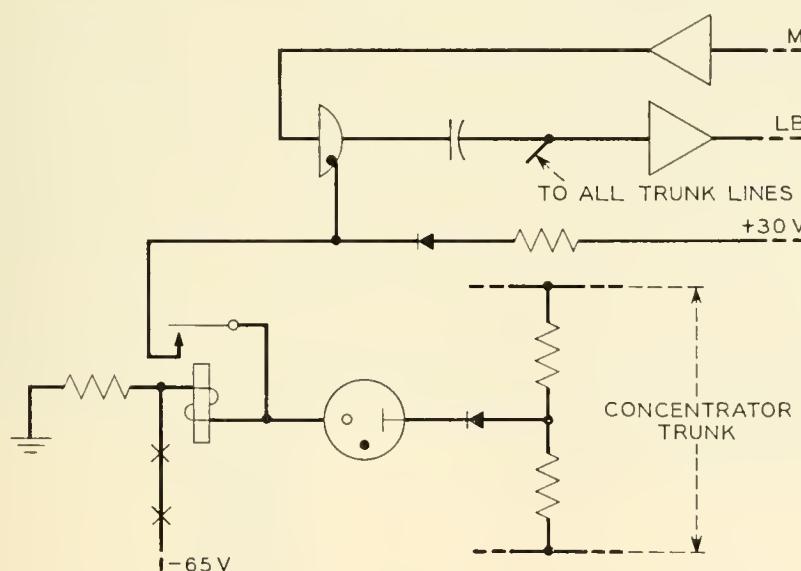
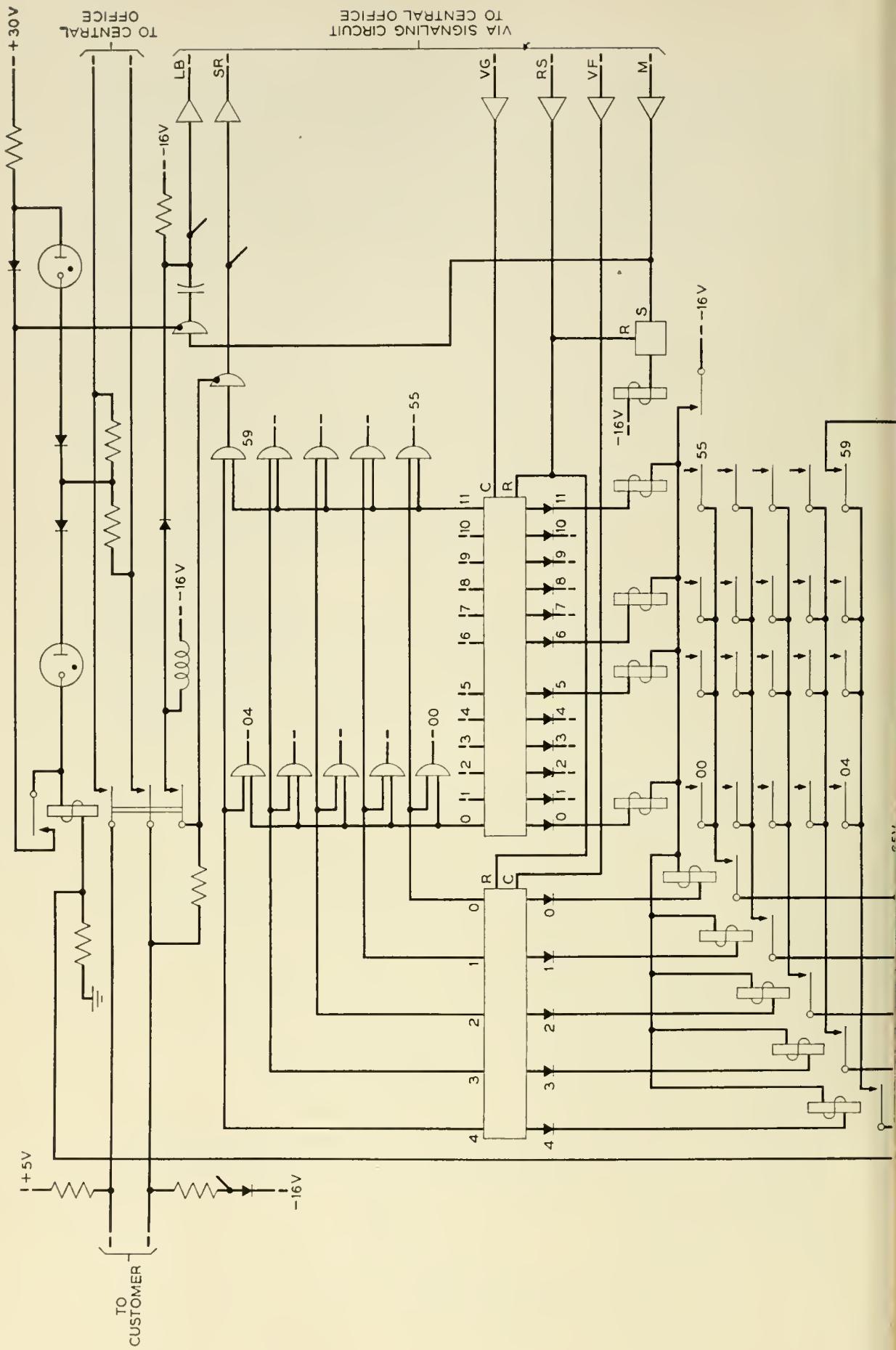


Fig. 20 — Crosspoint closure check.



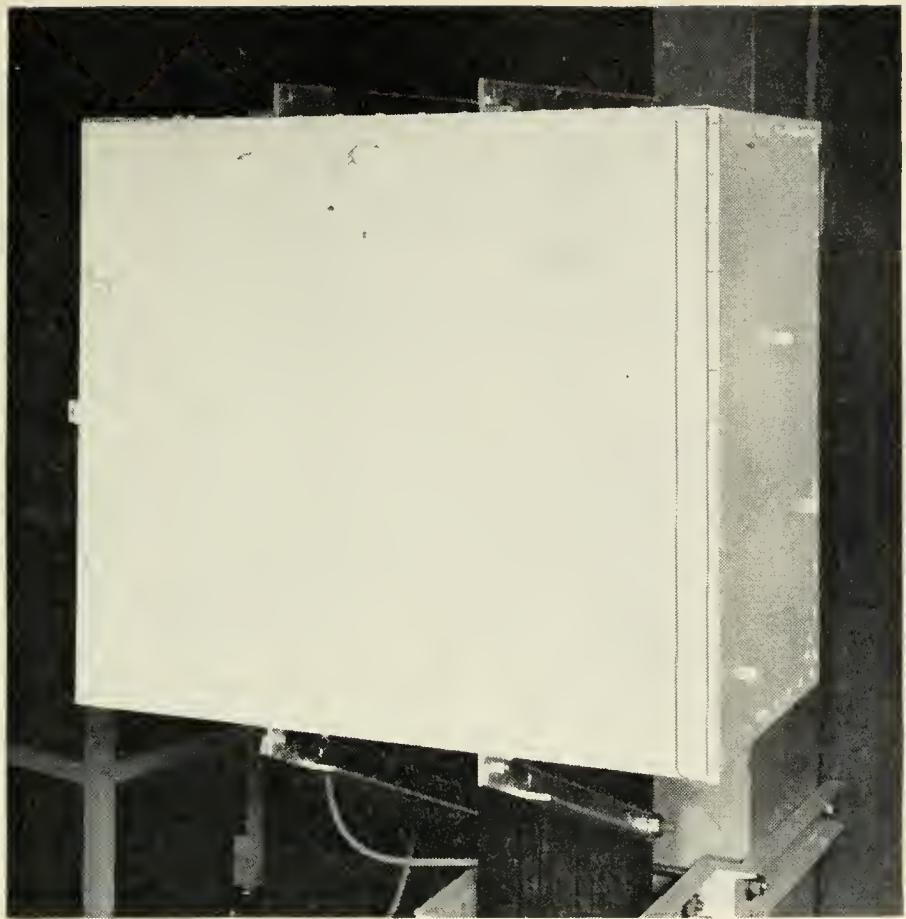


Fig. 22(a) — Complete line concentrator unit.

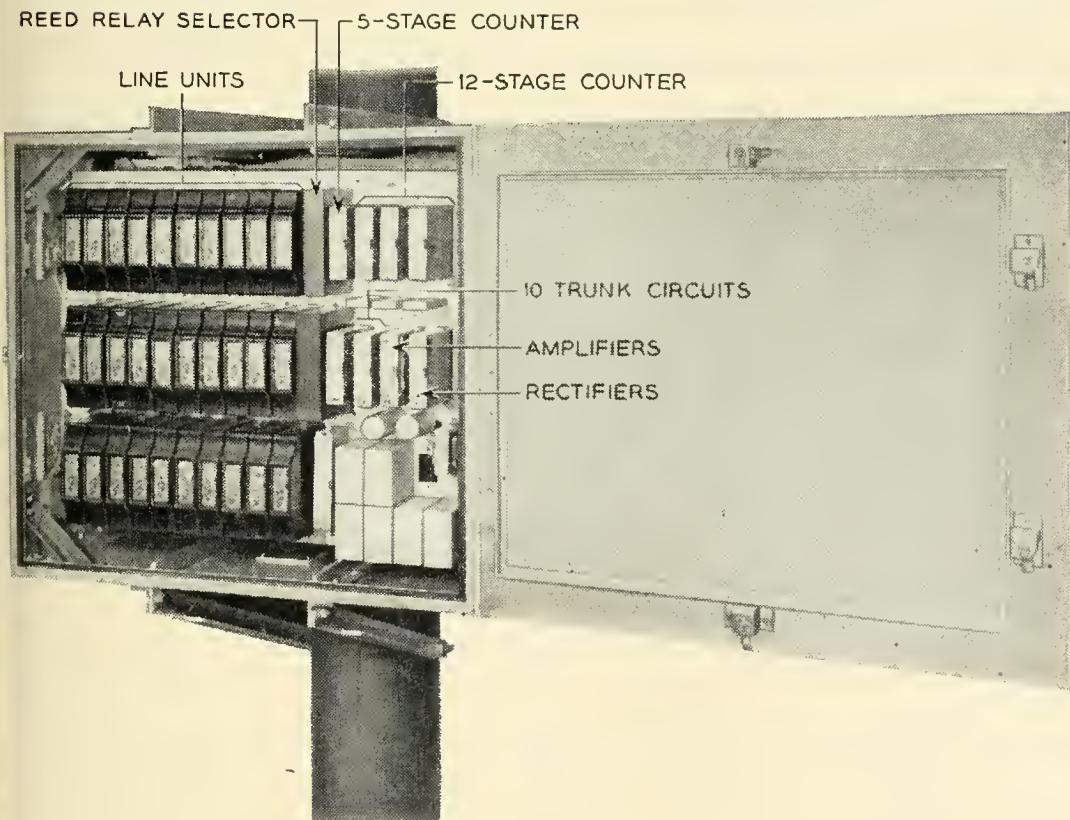
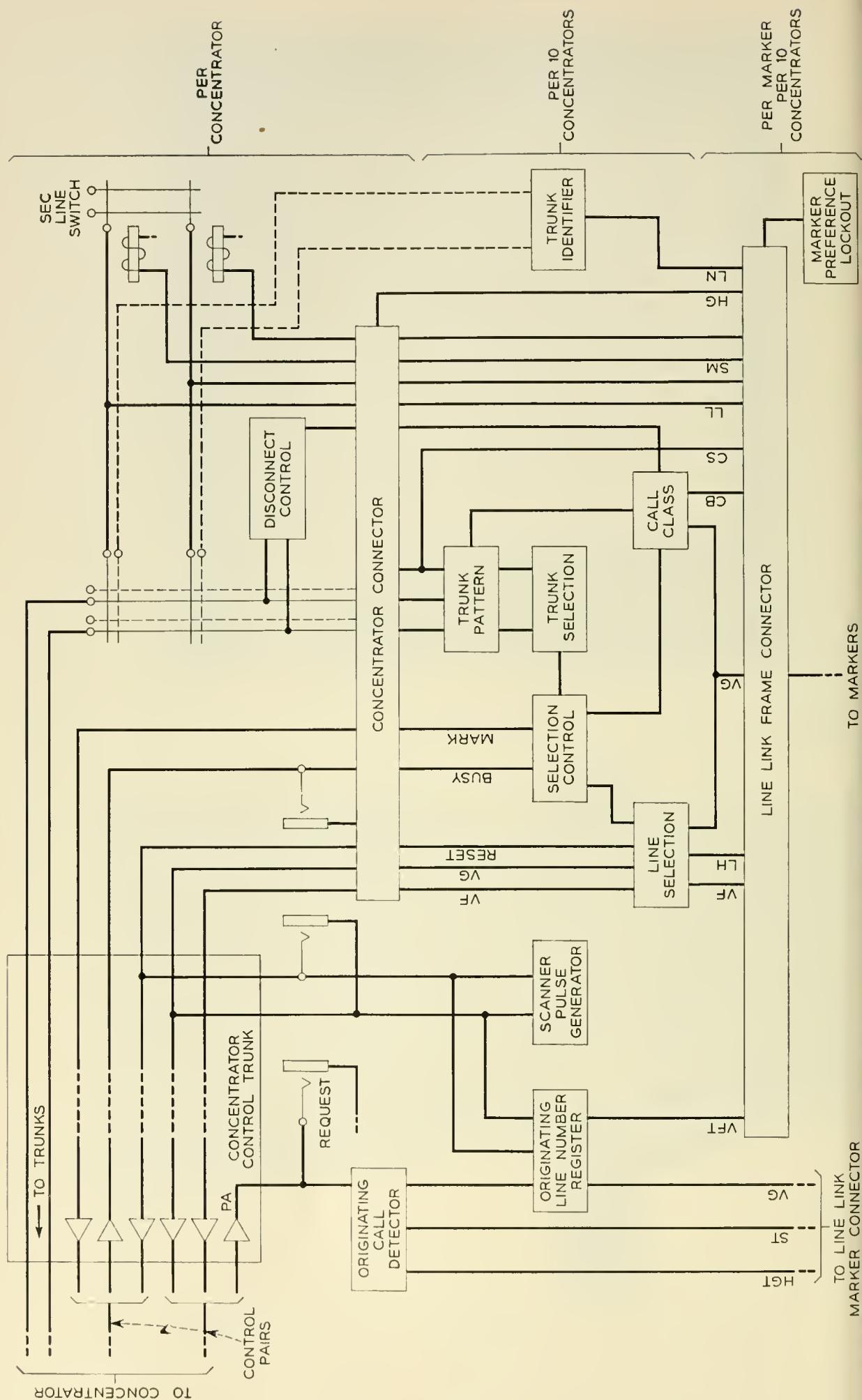


Fig. 22(b) — Identification of units within the line concentrator.



3. The line selection circuit.

4. The trunk identifier and selection relay circuits.

(For an understanding of how these frame circuits work through the line link marker connector and markers in the No. 5 system, the reader should consult the references.)

The common central office circuits will be described first.

a. *Scanner Pulse Generator*

The scanner pulse generator, shown in Fig. 24, produces continuously the combination of VG, VF and RS or reset pulses, described in connection with Fig. 18(a), required to drive the scanners for a number of concentrators. The primary pulse source is a 1,000-cycle transistor oscillator. This oscillator drives a transistor bi-stable circuit arranged as a binary counter such that on each cycle of the oscillator output it alternately assumes one of its states. Pulses produced by one state drive a 5-stage counter. Pulses produced by the other state through gates drive a 12-stage counter.

The pulses which drive the 5-stage counter are the same pulses which are used for the VF pulses to drive scanners. Each time the first stage of the 5-stage counter is on, a gate is opened to allow a pulse to drive the 12-stage counter. The pulses which drive the 12-stage counter are also the pulses used as the VG pulses for driving the scanners. They are out of phase with the VF pulses.

When the last stage of the 12-stage counter is on, the gate which

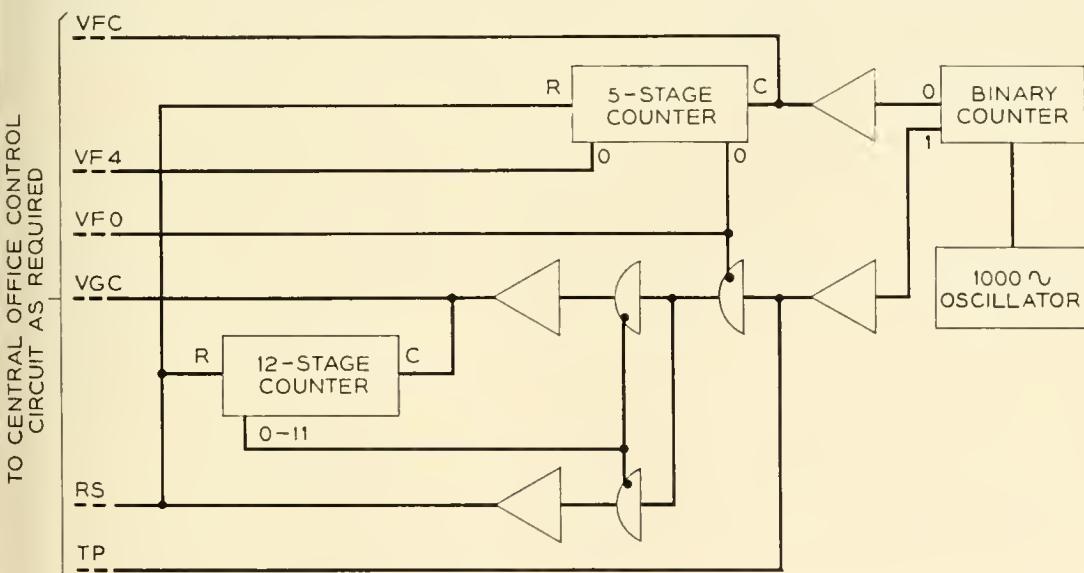


Fig. 24 — Scanner pulse generator.

transmits pulses to the 12-stage counter is closed and another gate is opened which produces the reset pulse. The reset pulse is thereby transmitted to the scanners in place of the first vertical group pulse. At the same time the 5 and 12-stage counters in the scanner pulse generator are reset to enable the starting of a new cycle.

In the central office control circuits, out of phase pulses on lead TP similar to those which drive the VG counters at the concentrator are used for various gating operations.

b. The Originating Call Detection and Line Number Registration

The originating call detector (Fig. 25) and the originating line number register (Fig. 26) together receive the information from the line concentrator used to identify the number of the line making a service request. The receipt of the service request pulse from a concentrator in a particular time slot will set a transistor bi-stable circuit HGT of Fig. 25 associated with that concentrator if no other originating call is being served by the frame circuits at this time.

The originating line number register consists of a 5 and 12-stage counter. These counters are normally driven through gates in synchronism with the scanning counters at concentrators with pulses supplied from the scanner pulse generator. When a service request pulse is received from any of the concentrators served by a line link frame, a pulse is sent to the originating line number register which operates a bi-stable circuit over a lead RH in Fig. 26. This bi-stable circuit then closes the gates through which the 5- and 12-stage counters are being driven, and also closes a gate which prevents them from being reset.

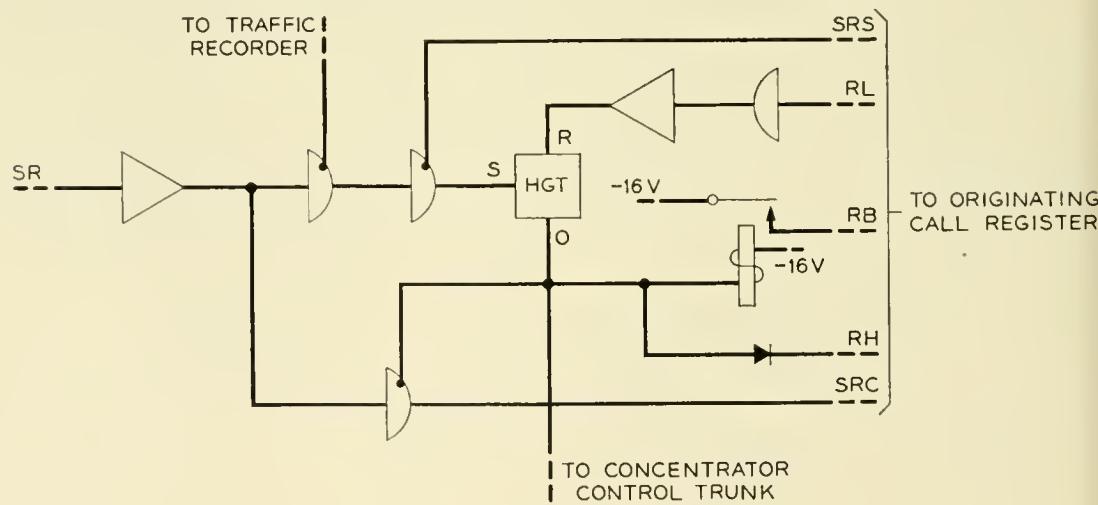


Fig. 25 — Originating call detector

In this way, the number of the line which originated a service request is locked into these counters until the bi-stable circuit is restored to normal.

The HGT bi-stable circuit of Fig. 25 indicates which particular concentrator has originated a service request. A relay in the collector circuit has contacts which pass this information on to the other central office control circuits to indicate the number of the concentrator on the frame which is requesting service. This is the same as a horizontal group on a regular line link frame and hence the horizontal group designation is used to identify a concentrator.

With the operation of this relay, relays associated with the counters of the originating line number register are operated. These relays indicate to the other central office circuits the vertical file and vertical group identification of the calling line. Contacts on the vertical group relays are used to set a bi-stable circuit associated with lead RL of Fig. 25 each time the scanner pulse generator generates a pulse corresponding to the vertical file of the calling line number registered.

The operation of the HGT bi-stable circuit inhibits in the concentrator control trunk circuit (Fig. 27) the transmission of further VG and

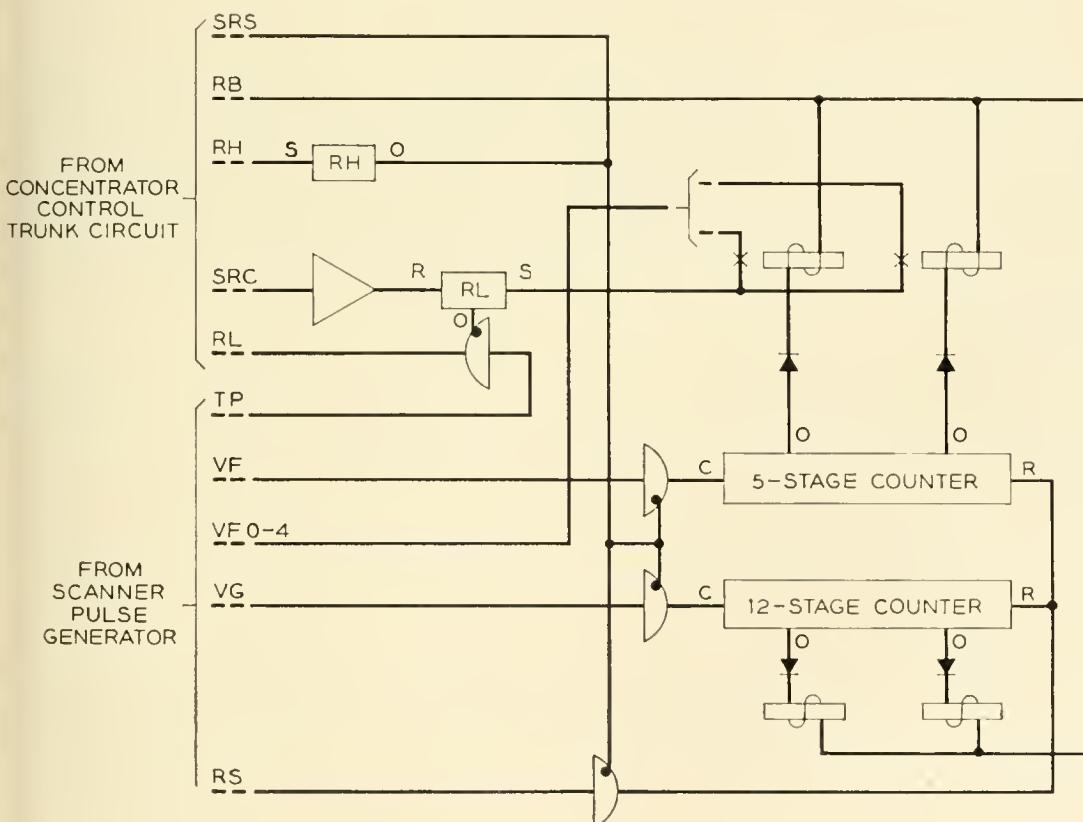


Fig. 26 — Originating line number register.

reset pulses to the concentrator so that, as described in Section 7a, only the VF counter continues to step once each 0.010 sec. So long as the line continues to request service this service request pulse is gated to reset the RL bi-stable circuit within the same time slot that it was set. If, however, a request for service is abandoned the RL bi-stable circuit of Fig. 26 will remain on and permit a TP pulse from the scanner pulse generator to reset the HGT bi-stable circuit which initiated the service request action.

Whenever the RH bi-stable circuit of Fig. 26 is energized it closes a gate over lead SRS for each concentrator to prevent any further service request pulses from being recognized until the originating call which has been registered is served. The resetting of the RH bi-stable circuit occurs once the call has been served. When more than one line concentrator is being served it is possible that the HGT bi-stable circuit of more than one concentrator will be set simultaneously as a result of coincidence in service requests from correspondingly numbered lines in these concentrators. The decision as to which concentrator is to be served is left to the marker, as it would normally decide which horizontal group to serve.

e. Line Selection

On all calls, originating and terminating, the marker transmits to the frame circuits the complete identity of the line which it will serve. In the case of originating calls it has received this information in the manner described in Section 8b. In either case, it operates wire spring relays VGO-11 and VFO-4, which enable gates so that the information may be stored in the 5- and 12-stage counters of the line selection circuit shown in Fig. 28.

The process of reading into the line selection counters starts when selection information has been received by the actuation of the HGS bi-stable circuit in the concentrator control trunk circuit of Fig. 27. This action stops the regular transmission of scanner pulses if they have not been stopped as a result of a call origination. At the same time it enables gates for transmission of information from the line selection circuit, Fig. 28.

The ST bi-stable circuit of the line selection circuit is also enabled to start the process of setting the line selection counters. The next TP pulse sets the R1 bi-stable circuit. This bi-stable circuit enables a gate which permits the next TP pulse to set the counters and transmit a reset pulse to the concentrator through pulse amplifier R1A. At the same time bi-stable circuit ST is reset to prevent the further read-in or reset

pulses and to permit pulses through amplifier OPA to start the out-pulsing of line selections. These pulses pass to the VGP and VFP leads as long as the VG and VF line selection counters have not reached their first and last stages respectively. The output pulses to the concentrator are also fed into the drive leads of these counters so that, as the counters in the concentrator are stepped up, the counters in the central office line selection circuit are stepped down. When the first stage of the VF counter goes on, the VF pulses are no longer transmitted until the first stage of the VG counter goes on. This insures that a VF pulse is the last to be transmitted. Also this pulse is not transmitted until the other frame circuits have successfully completed selections of an idle concentrator trunk. Then bi-stable circuit VFLD is energized,

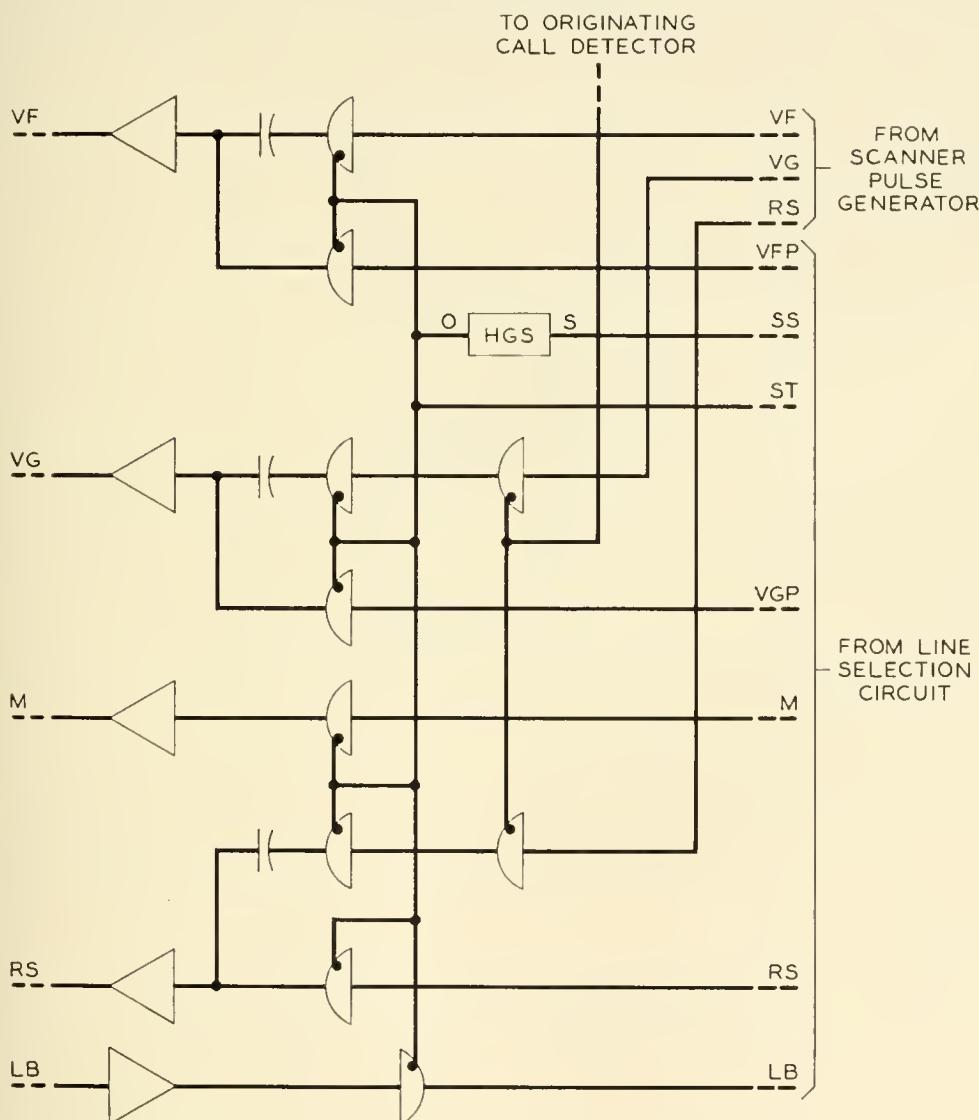


Fig. 27 — Concentrator control trunk circuit.

producing, during its transition, the last VF pulse for transmission to the concentrator.

d. Trunk Selection and Identification

The process of selecting an idle concentrator trunk to which the line has access utilizes familiar relay circuit techniques.¹⁹ This circuit, in Fig. 29, will not be described in detail. One trunk selection relay, TS, is operated indicating the preferred idle trunk serving a line in the particular vertical group being selected as indicated by the VG relay which has been operated by the marker.

The TS4 and TS5 relays select trunks 8 and 9 which are available to each line while the 4 trunks available to only half of the lines are selected by relays TS0-TS3. The busy or idle condition of each trunk is indicated by a contact on the hold magnet associated with each trunk through

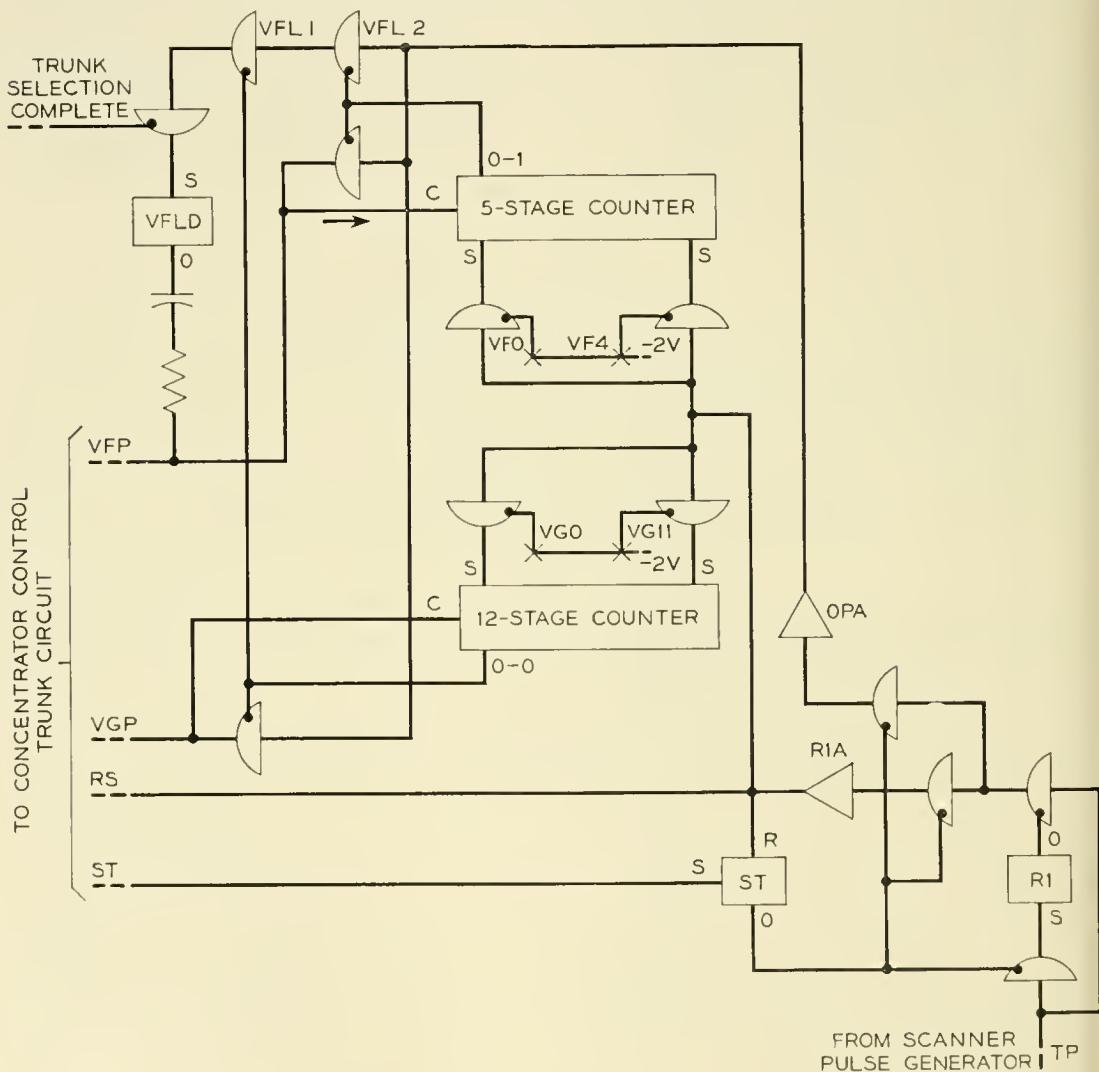


Fig. 28 — Line selection circuit.

relay HG which operates on all originating and terminating calls to the particular concentrator served by these trunks. The end chain relay TC of the lockout trunk selection circuit¹⁹ connects battery from the SR relay windings of idle trunks to the windings of the TS relays to permit one of the latter relays to operate and to steer circuits, not shown on Fig. 29, to the hold magnet of the trunk and to the tip-and-ring conductors of the trunk to apply the selection voltages shown on Figs. 13 and 14.

The path for operating the hold magnet originates in the marker. The path looks like that which the marker uses on the line hold magnet when setting up a call on a regular line link frame. For this reason and other similar reasons this concentrator line link frame concept has been nicknamed the "fool-the-marker" scheme.

Should a hold magnet release while a new call is being served the ground from the TC relay normal or the TS relay winding holds relay

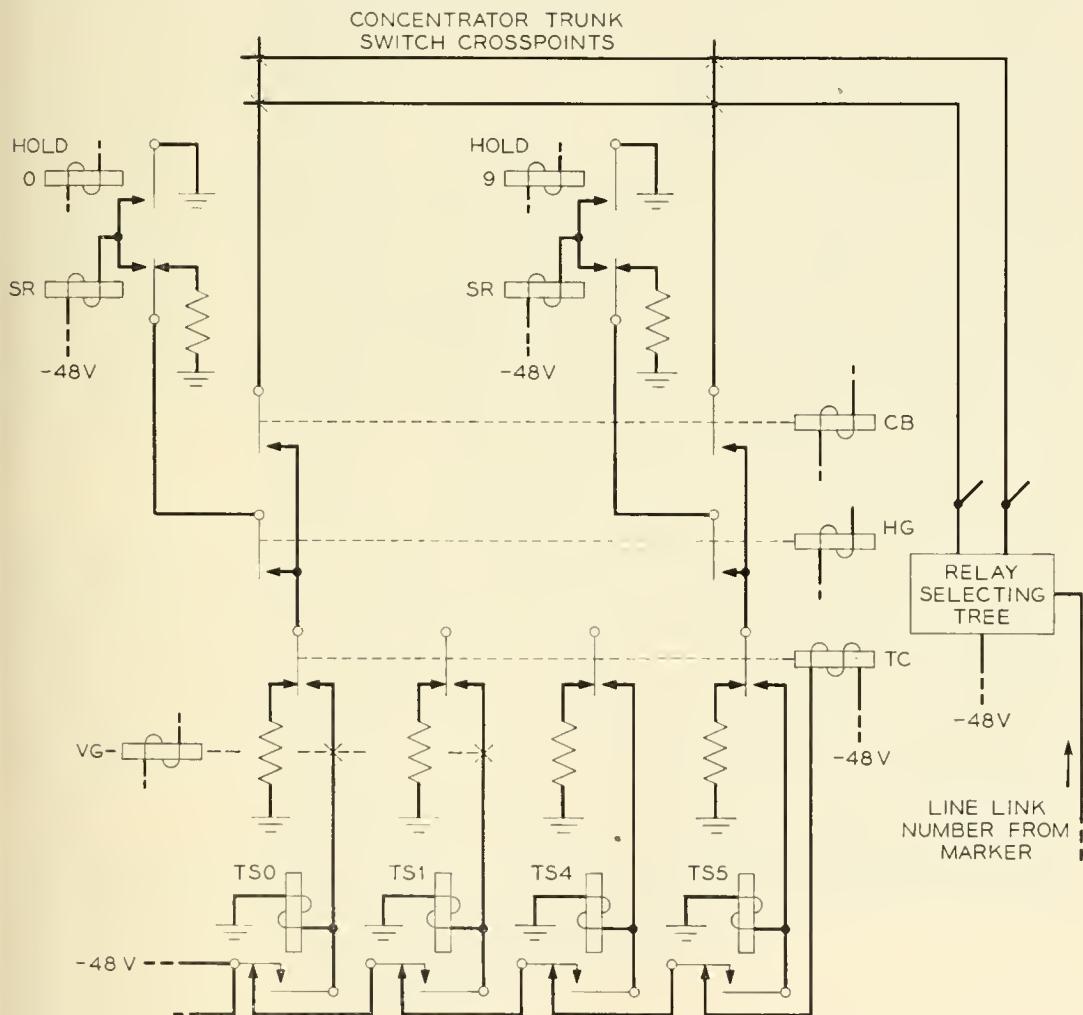


Fig. 29 — Trunk selection and identification.

SR operated through its own contact until the new call has been set up. This prevents interference of disconnect pulses applied to the trunk when a selection is being made and insures that a disconnect pulse is transmitted before the trunk is reused.

A characteristic of the No. 5 crossbar system is that the originating connection to a call register including the line hold magnet is released and a new connection, known as the "call back connection", is established to connect the line to a trunk circuit after dialing is completed.

With concentrator operation the concentrator trunk switch connection is released but the disconnect signal is not sent to the concentrator as a result of holding the SR relay as described above. However, the marker does not know to which trunk the call back connection is to be established. For this reason the frame circuits include an identification process for determining the number of the concentrator trunk to be used on call back prior to the release of the originating register connection.

Identification is accomplished by the marker transmitting to the frame circuits the number of the link being used on the call. This information is already available in the No. 5 system. The link being used is marked with -48 volts by a relay selecting tree²⁰ to operate the TS relay associated with the trunk to which the call back connection is to be established. Relay CB (Fig. 29) is operated on this type of call instead of relay HG. The circuits for reoperating the proper hold magnet are already available on the TS relay which was operated, thereby re-selecting the trunk to which the customer is connected. The concentrator connection is not released when the hold magnet releases and again the marker operates as it would on a regular line link frame call.

9. FIELD TRIALS

Three sets of the experimental equipment described here have been constructed and placed in service in various locations. The equipment for these trials is the forerunner of a design for production which will incorporate device, circuit and equipment design changes based on the trial experiences. Fig. 30 shows the cabinet mounted central office trial equipment with the designation of appropriate parts.

For the field trials described, the line links on a particular horizontal level of existing line link frames were extended to a separate cross-bar switch provided for this purpose in the trial equipment. The regular line link connector circuits were modified to work with the trial control circuits whenever a call was originated or terminated on this level. No lines were terminated in the regular primary line switches for this level.

10. MISCELLANEOUS FEATURES OF TRIAL EQUIPMENT

There are a number of auxiliary circuits provided with the trial equipment to aid in the solutions of problems brought about by the concepts of concentrator service. One of the purposes of the trials was to determine the way in which the various traffic, plant and commercial ad-

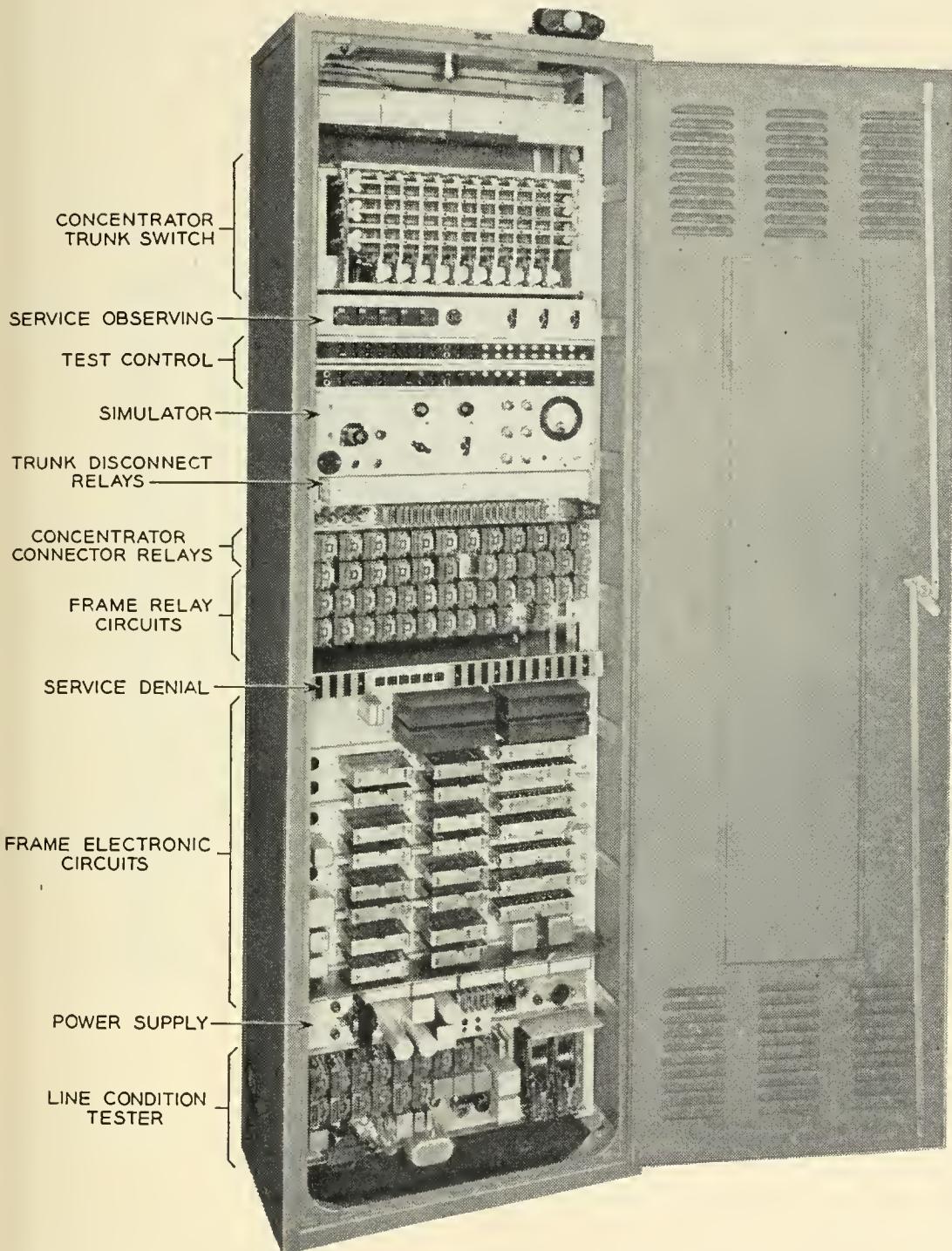


Fig. 30 — Trial central office equipment.

ministrative functions could be economically performed when concentrators become common telephone plant facilities. The more important of these miscellaneous features are discussed under the following headings:

a. *Traffic Recording*

To measure the amount and characteristics of the traffic handled by the concentrator a magnetic tape recorder, Fig. 31, was provided for each trial. The number of the lines and trunks in use each 15 seconds during programmed periods of each day were recorded in coded form with polarized pulses on the 3-track magnetic tape moving at a speed of $7\frac{1}{2}$ " per second. Combinations of these pulses designate trunks busy on intra-concentrator connections and reverting calls.

The line busy indications were derived directly from the line busy information received during regular scanning at the concentrator. During one cycle in each 15 seconds new service requests were delayed to insure that a complete scan cycle would be recorded. Terminating calls were not delayed since marker holding time is involved. Trunk conditions are derived for a trunk scanner provided in the recorder.

In addition to recording the line and trunk usage, recordings were made on the tape for each service request detected during a programmed period to measure the speed with which each call received dial tone and the manner in which the call was served. In this type of operation the length of the recording for each request made at a tape speed of only $\frac{1}{4}$ " per second is a measure of service delay time.

As may be observed from Fig. 31 the traffic recorder equipment was built with vacuum tubes and hence required a rather large power supply. It is expected that a transistorized version of this traffic recorder serving all concentrators in a central office will be included in the standard model of the line concentrator equipment. With this equipment, traffic engineers will know more precisely the degree to which each concentrator may be loaded and hence insure maximum utilization of the concentrator equipment.

b. *Line Condition Tester*

It has been a practice in more modern central office equipment to include automatic line testing equipment.²¹ An attempt has been made to include similar features with the concentrator trial equipment. The line condition tester (see Fig. 30) provides a means for automatically connecting a test circuit to each line in turn once a test cycle has been

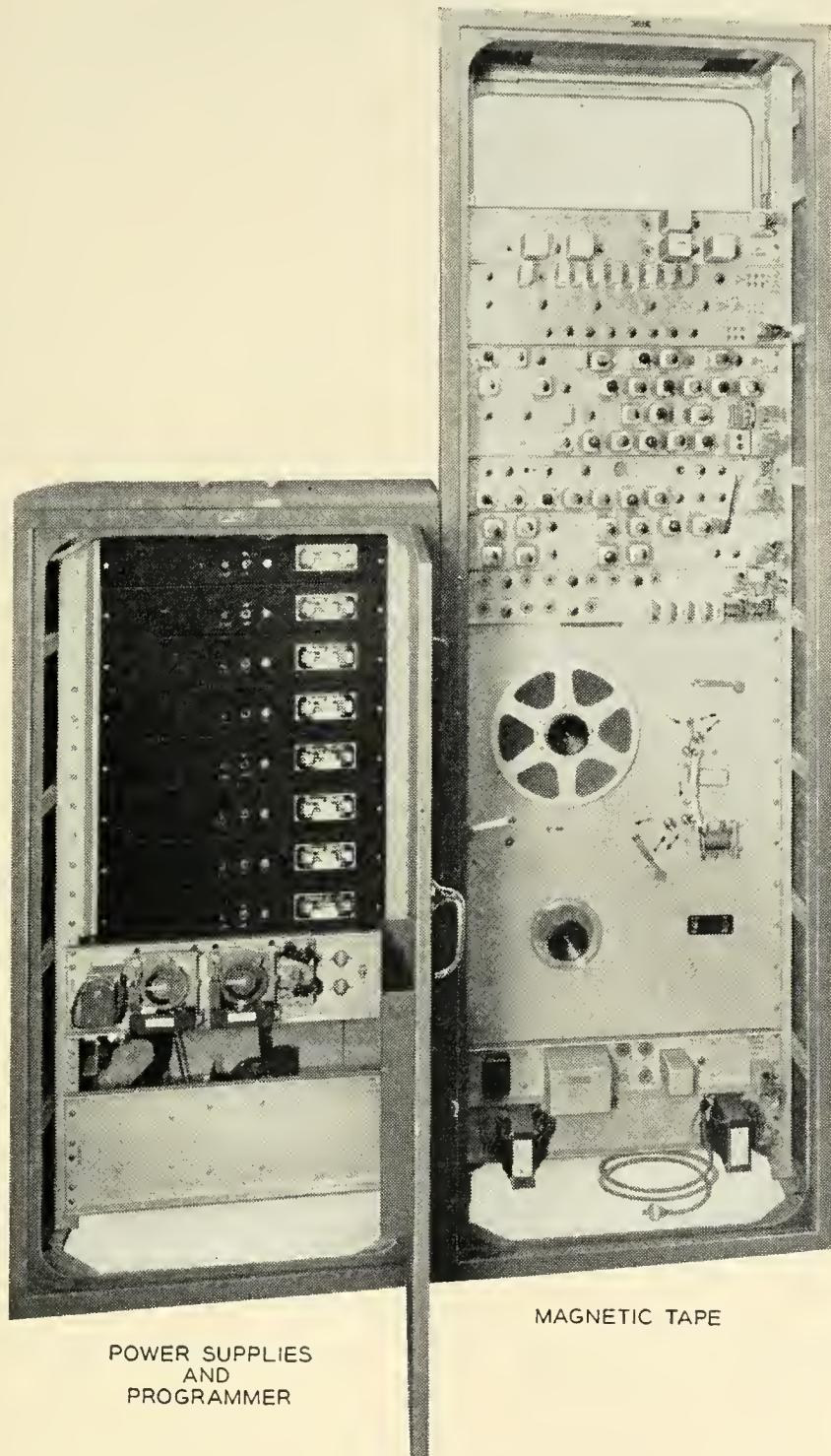


Fig. 31 — Traffic recorder.

manually initiated. This test is set up on the basis of the known concentrator passive line circuit capabilities. Should a line fail to pass this test, the test circuit stops its progress and brings in an alarm to summon central office maintenance personnel. The facilities of the line tester are also used to establish, under manual control, calls to individual lines as required to carry out routine tests.

e. *Simulator*

As the central office sends out scanner control pulses either no signal, a line busy or service request pulse is returned to the central office in each time slot. The simulator test equipment, shown in Fig. 30, was designed to place pulses in a specific time slot to simulate a line under test at the concentrator.

In addition to transmitting the equivalent of concentrator output pulses the simulator can receive the regular line selection pulses transmitted to the concentrator for purposes of checking central office operations. It is possible by combined use of the line tester and simulator to observe the operation of the concentrator and to determine the probable cause when a fault occurs.

d. *Service Observing*

The removal of the line terminals from the central office poses a number of problems in conjunction with the administration of central office equipment. One of these is service observing.

To maintain a check on the quality of service being rendered by the telephone system, service observing taps are made periodically on telephone lines. This is normally done by placing special connector shoes on line terminations in the central office.

To place such shoes at the remote concentrator point would lead to administrative difficulties and added expense. Therefore, a method was devised to permit service observing equipment to be connected to concentrator trunks on calls from specific lines which were to be observed. This method consisted of manual switches on which were set the number of the line to be observed in terms of vertical group and vertical file. Whenever this line originated a call and the call could be placed over the first preferred trunk, automatic connection was made to the service observing desk in the same manner as would occur for a line terminated directly in the central office.

In addition, facilities were provided for trying a new service observing technique where calls originating over a particular concentrator

trunk would be observed without knowledge of the originating line number. For this purpose a regular line observing shoe was connected to one of the ten concentrator trunk switch verticals in the trial equipment and from here connected to the service observing desk in the usual manner.

The basic service observing requirements in connection with line concentrator operation have not as yet been fully determined. However, it appears at this time that the trunk observing arrangement may be preferable.

e. *Service Denial*

In most systems denial of originating service for non-payment of telephone service charges, for trouble interception and for permanent signals caused by cable failures or prolonged receiver-off-hook conditions may be treated by the plant forces at the line terminals or by blocking the line relay. To avoid concentrator visits and to enable the prompt clearing of trouble conditions which tie up concentrator trunks, a service denial feature has been included in the design of the central office circuits.

This feature consists of a patch-panel with special gate cords which respond to particular time slots and inhibit service request signals produced by a concentrator during this period. In this way service requests can be ignored and prevent originating call service on particular lines until a trouble locating or other administrative procedure has been invoked.

f. *Display Circuit*

A special electronic switch was developed for an oscilloscope. This arrangement permitted the positioning of line busy and service request pulses in fixed positions representing each of the 60 lines served. Line busy pulses were shown as positive and service request pulses as negative. This plug connected portable aid, see Fig. 32, was useful in tracing calls and identifying lines to which service may be denied, due to the existence of permanent signals.

Other circuits and features, too detailed to be covered in this paper, have been designed and used in the field trials of remote line concentrators. Much has been learned from the construction and use of this equipment which will aid in making the production design smaller, lighter, economical, serviceable and reliable.

Results from the field trials have encouraged the prompt undertaking

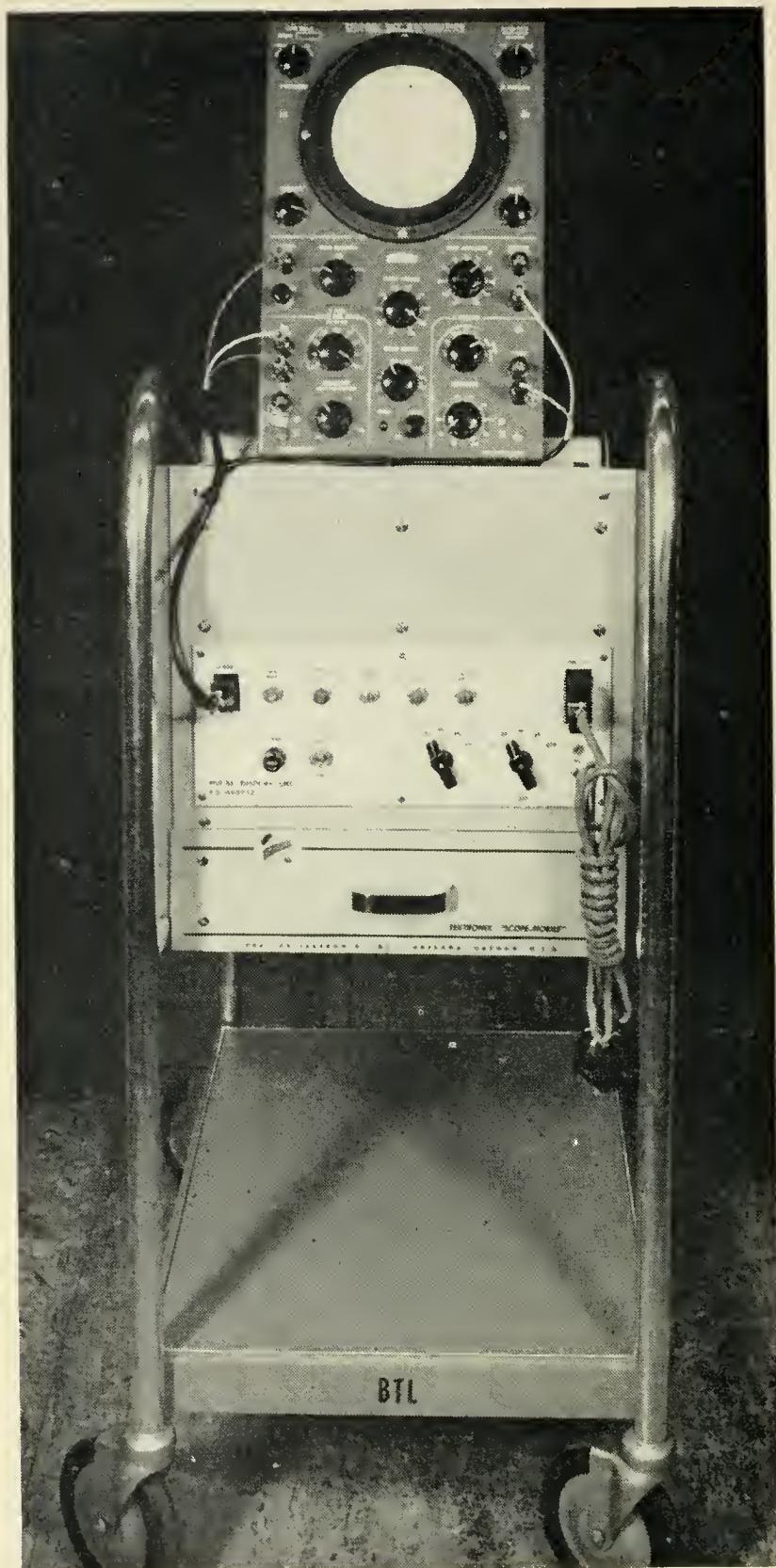


Fig. 32 — Pulse display oscilloscope.

of development of a remote line concentrator for quantity production. The cost of remote line concentrator equipment will determine the ultimate demand. In the meantime, an effort is being made to take advantage of the field trial experiences to reduce costs commensurate with insuring reliable service.

The author wishes to express his appreciation to his many colleagues at Bell Telephone Laboratories whose patience and hard work have been responsible for this new adventure in exploratory switching development. An article on line concentrators would not be complete without mention of C. E. Brooks who has encouraged this development and under whose direction the engineering studies were made.

BIBLIOGRAPHY

1. E. C. Molina, The Theory of Probabilities Applied to Telephone Trunking Problems, B.S.T.J., **1**, pp. 69-81, Nov., 1922.
2. Strowger Step-by-Step System, Chapter 3, Vol. 3, Telephone Theory and Practice by K. B. Miller, McGraw-Hill 1933.
3. F. A. Korn and J. G. Ferguson, Number 5 Crossbar Dial Telephone Switching System, Elec. Engg., **69**, pp. 679-684, Aug., 1950.
4. U. S. Patent 1,125,965.
5. O. Myers, Common Control Telephone Switching Systems, B.S.T.J., **31**, pp. 1086-1120, Nov., 1952.
6. L. J. Stacy, Calling Subscribers to the Telephone, Bell Labs. Record, **8**, pp. 113-119, Nov., 1929.
7. J. Meszar, Fundamentals of the Automatic Telephone Message Accounting System, A. I. E. E. Trans., **69**, pp. 255-268, (Part 1), 1950.
8. O. M. Hovgaard and G. E. Perreault, Development of Reed Switches and Relays, B.S.T.J., **34**, pp. 309-332, Mar., 1955.
9. W. A. Malthaner and H. E. Vaughan, Experimental Electronically Controlled Automatic Switching System, B. S.T.J., **31**, pp. 443-468, May, 1952.
10. S. T. Brewer and G. Hecht, A Telephone Switching Network and its Electronic Controls, B.S.T.J., **34**, pp. 361-402, Mar., 1955.
11. L. W. Hussey, Semiconductor Diode Gates, B.S.T.J., **32**, pp. 1137-54, Sept., 1953.
12. U. S. Patent 1,528,982.
13. J. J. Ebers and S. L. Miller, Design of Alloyed Junction Germanium Transistor for High-Speed Switching, B.S.T.J., **34**, pp. 761-781, July, 1955.
14. W. B. Graupner, Trunking Plan for No. 5 Crossbar System, Bell Labs. Record, **27**, pp. 360-365, Oct., 1949.
15. G. L. Pearson and B. Sawyer, Silicon p-n Junction Alloy Diodes, I.R.E. Proc., **42**, pp. 1348-1351, Nov., 1952.
16. A. E. Anderson, Transistors in Switching Circuits, B.S.T.J., **31**, pp. 1207-1249, Nov., 1952.
17. J. J. Ebers and J. L. Moll, Large-Signal Behavior of Junction Transistors, I. R. E. Proc., **42**, pp. 1761-1784, Dec., 1954.
18. J. J. Ebers, Four-Terminal p-n-p-n Transistors, I. R. E. Proc., **42**, pp. 1361-1364, Nov., 1952.
19. A. E. Joel, Relay Preference Lockout Circuits in Telephone Switching, Trans. A. I. E. E., **67**, pp. 720-725, 1948.
20. S. H. Washburn, Relay "Trees" and Symmetric Circuits, Trans. A. I. E. E., **68**, pp. 571-597, 1949.
21. J. W. Dehn and R. W. Burns, Automatic Line Insulation Testing Equipment for Local Crossbar Systems, B.S.T.J., **32**, pp. 627-646, 1953.

Transistor Circuits for Analog and Digital Systems*

By FRANKLIN H. BLECHER

(Manuscript received November 17, 1955)

This paper describes the application of junction transistors to precision circuits for use in analog computers and the input and output circuits of digital systems. The three basic circuits are a summing amplifier, an integrator, and a voltage comparator. The transistor circuits are combined into a voltage encoder for translating analog voltages into equivalent time intervals.

1.0. INTRODUCTION

Transistors, because of their reliability, small power consumption, and small size find a natural field of application in electronic computers and data transmission systems. These advantages have already been realized by using point contact transistors in high speed digital computers.¹ This paper describes the application of junction transistors to precision circuits which are used in dc analog computers and in the input and output circuits of digital systems. The three basic circuits which are used in these applications are a summing amplifier, an integrator, and a voltage comparator. A general procedure for designing these transistor circuits is given with particular emphasis placed on new design methods that are necessitated by the properties of junction transistors. The design principles are illustrated by specific circuits. The fundamental considerations in the design of transistor operational amplifiers are discussed in Section 2.0. In Section 3.0 an illustrative summing amplifier is described, which has a dc accuracy of better than one part in 5,000 throughout an operating temperature range of 0 to 50°C. The feedback in this amplifier is maintained over a broad enough frequency band so that full accuracy is attained in about 100 microseconds.

The design of a specific transistor integrator is presented in Section

* Submitted in partial fulfillment of the requirements for the degree of Doctor of Electrical Engineering at the Polytechnic Institute of Brooklyn.

4.0. The integrator can be used to generate a voltage ramp which is linear to within one part in 8,000. By means of an automatic zero set (AZS) circuit which uses a magnetic detector, the slope of the voltage ramp is maintained constant to within one part in 8,000 throughout a temperature range of 20°C to 40°C.

The voltage comparator, described in Section 5.0, is an electrical device which indicates the instant of time an input voltage waveform passes through a predetermined reference level. By taking advantage of the properties of semiconductor devices, the comparator can be designed to have an accuracy of ± 5 millivolts throughout a temperature range of 20°C to 40°C.

In Section 6.0, the system application of the transistor circuits is demonstrated by assembling the summing amplifier, the integrator, and the voltage comparator into a voltage encoder. The encoder can be used to translate an analog input voltage into an equivalent time interval with an accuracy of one part in 4,000. This accuracy is realized throughout a temperature range of 20°C to 40°C for the particular circuits described.

2.0. FUNDAMENTAL CONSIDERATIONS IN THE DESIGN OF OPERATIONAL AMPLIFIERS

The basic active circuit used in dc analog computers is a direct coupled negative feedback amplifier. With appropriate input and feedback networks, the amplifier can be used for multiplication by a constant coefficient, addition, integration, or differentiation as shown in Figure 1.² The accuracy of an operational amplifier depends only on the passive components used in the input and feedback circuits provided that there is sufficient negative feedback (usually greater than 60 db). The time that is required for the amplifier to perform a calculation is an inverse function of the bandwidth over which the feedback is maintained. Thus a fundamental problem in the design of an operational amplifier is the development of sufficient negative feedback over a reasonably broad frequency range. The associated problem is the realization of satisfactory stability margins. Finally there is the problem of reducing the drift which is inherent in direct coupled amplifiers and particularly troublesome for transistors because of the variation in their characteristics with temperature.

The first step in the design is the blocking out of the configuration for the forward gain circuit (designated A in Fig. 1). Three primary requirements must be satisfied:

- (1) Stages must be direct coupled.

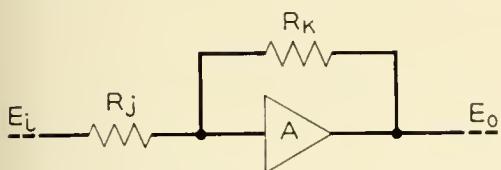
(2) Amplifier must provide one net phase reversal.

(3) Amplifier must have enough current gain to meet accuracy requirements.

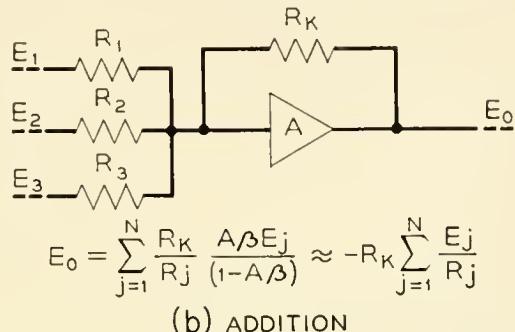
Three possible transistor connections are available:³ (a) the common base connection which may be considered analogous to the common grid vacuum tube connection; (b) the common emitter connection which is analogous to the common cathode connection; and (c) the common collector connection which is analogous to the cathode follower connection. These three configurations together with their approximate equivalent circuits are shown in Fig. 2. It has been shown⁴ that for most junction transistors the circuit element a is given by the expression

$$a = \text{sech} \left[\frac{W}{L_m} (1 + p\tau_m)^{1/2} \right] \quad (1)^*$$

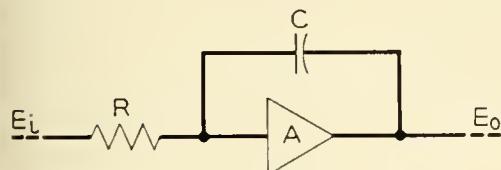
where W is the thickness of the transistor base region, L_m is the diffusion length and τ_m the lifetime of minority charge carriers in the base region,



(a) MULTIPLICATION BY A CONSTANT COEFFICIENT

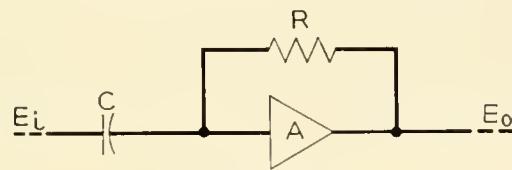


(b) ADDITION



$$\mathcal{L}[E_o] = \frac{A\beta}{(1-A\beta)} \frac{\mathcal{L}[E_i]}{pRC} \approx -\frac{\mathcal{L}[E_i]}{pRC}$$

(c) INTEGRATION



$$\mathcal{L}[E_o] = \frac{A\beta pRC \mathcal{L}[E_i]}{1-A\beta} \approx -pRC \mathcal{L}[E_i]$$

(d) DIFFERENTIATION

NOTE: $\mathcal{L}[E_o]$ = LAPLACE TRANSFORM OF OUTPUT VOLTAGE

$\mathcal{L}[E_i]$ = LAPLACE TRANSFORM OF INPUT VOLTAGE

$$p = j\omega$$

Fig. 1 — Summary of operational amplifiers.

* This expression assumes that the injection factor γ and the collector efficiency α_c are both unity. This is a good approximation for all alloy junction transistors and most grown junction transistors.

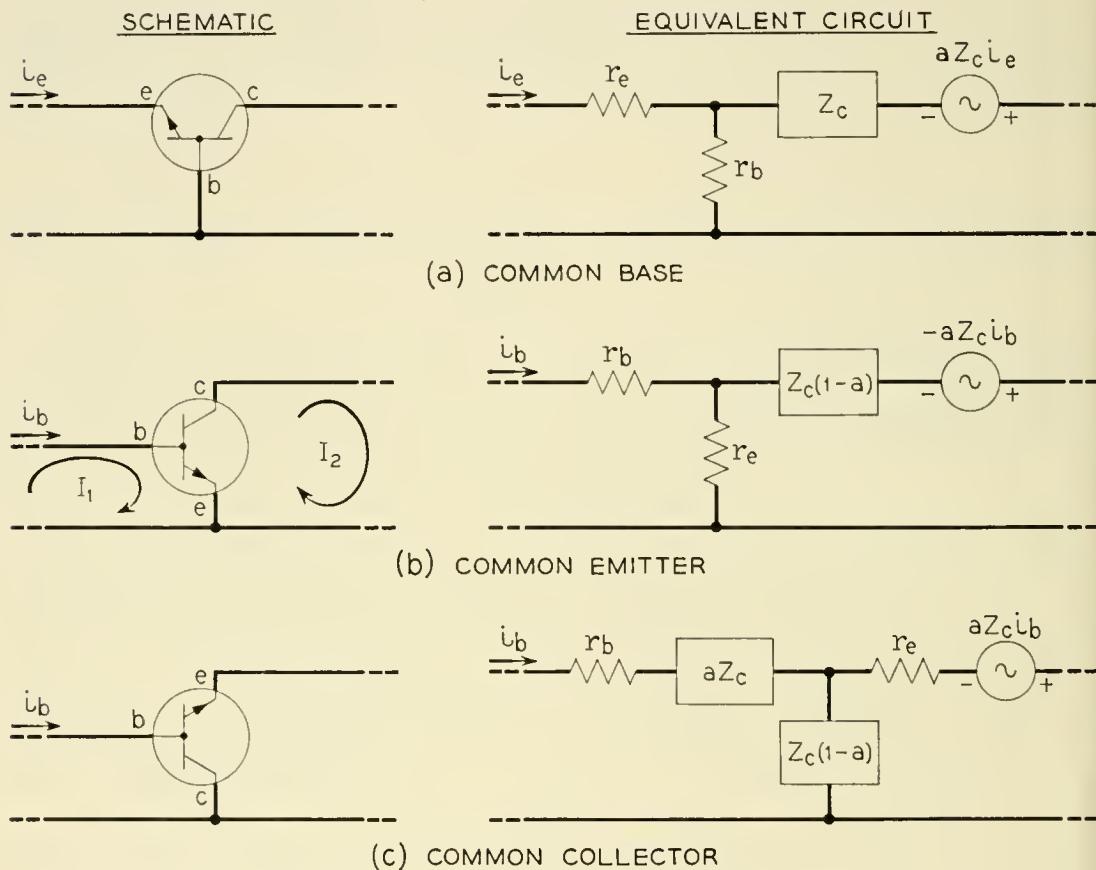
and $p = j\omega$. At frequencies less than $\omega_a/2\pi$, (1) can be approximated by

$$a = \frac{a_0}{1 + \frac{p}{\omega_a}} \quad (2)^5$$

where a_0 is the low frequency value of

$$a \approx 1 - \frac{1}{2} \left(\frac{W}{L_m} \right)^2, \quad \text{and} \quad \omega_a = \frac{2.4D_m}{W^2}$$

(D_m is the diffusion constant for the minority charge carriers in the base region). A readily measured parameter called alpha (α), the short circuit current gain of a junction transistor in the common base connec-



$$Z_c = \frac{r_c}{1 + pr_c C_c}$$

r_c = COLLECTOR RESISTANCE

$$a = \frac{a_0}{1 + \frac{p}{\omega_a}}$$

C_c = COLLECTOR CAPACITANCE

$$p = j\omega$$

$\frac{\omega_a}{2\pi} \approx$ ALPHA-CUTOFF FREQUENCY

Fig. 2 — Basic transistor connections.

tion, is related to a by the equation

$$\alpha = \frac{aZ_e + r_b}{Z_e + r_b} \quad (3)$$

For most junction transistors the base resistance, r_b , is much smaller than the collector impedance $|Z_c|$, at frequencies less than $\omega_a/2\pi$. Therefore, $\alpha \approx a$ and $\omega_a/2\pi$ is very nearly equal to the alpha-cutoff frequency, the frequency at which $|\alpha|$ is down by 3 db.

The transistor parameters r_e and r_b are actually frequency sensitive and should be represented as impedances.⁶ However, good agreement between theory and experiment is obtained at frequencies less than $\omega_a/2\pi$ with r_e and r_b assumed constant.

The choice of an appropriate transistor connection for a direct coupled, negative feedback amplifier, is based on the following reasoning. The common base connection may be ruled out immediately because this connection does not provide current gain unless a transformer interstage is used. The common emitter connection provides short circuit current gain and a phase reversal for each stage. Thus if the amplifier is composed of an odd number of common emitter stages, all three requirements previously listed, are satisfied. A common emitter cascade has the additional practical advantage, that by alternating n-p-n and p-n-p types of transistors, the stages can be direct coupled with practically zero interstage loss.⁷

The common collector connection provides short circuit current gain but no phase reversal. Consequently, the dc amplifier cannot consist entirely of common collector stages and operate as a negative feedback amplifier. This paper will consider only the common emitter connection since, in general, for the same number of transistor stages, the common emitter cascade provides more current gain than a cascade composed of both common collector and common emitter stages.

2.1 Evaluation of External Voltage Gain

Since the equivalent circuit of the junction transistor is current activated, it is convenient to treat feedback in a single loop transistor amplifier as a loop current transmission (refer to Appendix I) instead of as a loop voltage transmission which is commonly used for single loop vacuum tube amplifiers.⁸ Fig. 3 shows a single loop feedback amplifier in which a fraction of the output current is fed back to the input. A is defined as the short circuit current gain of the amplifier without feedback, and β is defined as the fraction of the short circuit output current (or Norton

equivalent circuit current) fed back to the input summing node. With these definitions,

$$I_{SC} = AI_{IN}' \quad (4)$$

$$I_\beta = \beta I_{SC} \quad (5)$$

where I_{SC} is the Norton equivalent short circuit current.

From Kirchhoff's first law

$$I_{IN}' = I_{IN} + I_\beta \quad (6)$$

Combining relations (4) to (6) yields

$$\frac{I_{SC}}{I_{IN}} = \frac{A}{1 - A\beta} \quad (7)$$

Expression (7) provides a convenient method for evaluating the external

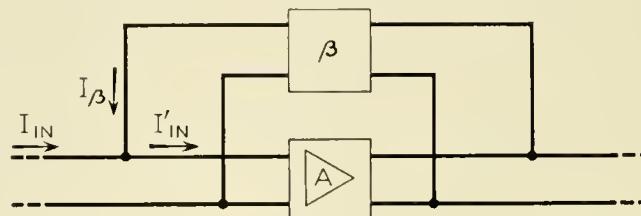


Fig. 3 — Single loop feedback amplifier.

voltage gain of an operational amplifier. Fig. 4 shows a generalized operational amplifier with N inputs. With this configuration,

$$I_{IN} = \sum_{j=1}^N \left[\frac{E_j - \frac{I_{SC}}{A} Z_{IN}'}{Z_j} \right] \quad (8)$$

where $E_j, j = 1, 2, \dots, N$, are the N input voltages referred to the ground node.

$Z_j, j = 1, 2, \dots, N$, are the N input impedances

Z_{IN}' is the input impedance of the amplifier measured at the summing node with the feedback loop opened.

$$I_\beta = \frac{E_{OUT} - Z_{IN}' \frac{I_{SC}}{A}}{Z_K} \quad (9)$$

$$E_{OUT} = \frac{I_{SC} - I_\beta}{\frac{1}{R_L} + \frac{1}{Z_{OUT}'}} \quad (10)$$

where Z_{OUT}' is the output impedance of the amplifier measured with the feedback loop opened. The expression for the output voltage is obtained by combining (7), (8), (9), and (10).

$$E_{\text{OUT}} = \sum_{j=1}^N E_j \frac{Z_K}{Z_j} \left[\frac{A\beta + \frac{Z_{\text{IN}}'}{Z_K}}{1 - A\beta + \sum_{j=1}^N \frac{Z_{\text{IN}}'}{Z_j}} \right] \quad (11)^*$$

where

$$A\beta = A \left[\frac{1 - \frac{Z_{\text{IN}}'}{A} \left(\frac{1}{R_L} + \frac{1}{Z_{\text{OUT}}'} \right)}{1 + \frac{Z_K}{R_L} + \frac{Z_K}{Z_{\text{OUT}}'}} \right]$$

$A\beta$ is equal to the current returned to the summing node when a unit

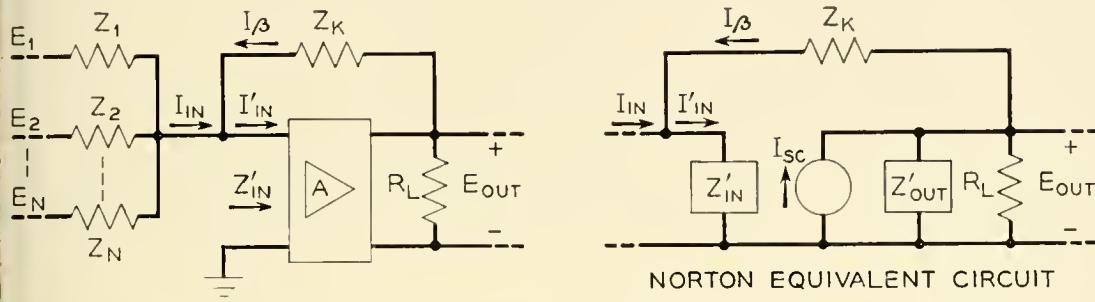


Fig. 4 — Generalized operational amplifier.

current is placed into the base of the first transistor stage ($I_{\text{IN}}' = 1$). If $|A\beta|$ is much greater than $|Z_{\text{IN}}'/Z_K|$ and

$$1 + \sum_{j=1}^N \left| \frac{Z_{\text{IN}}'}{Z_j} \right|,$$

then

$$E_{\text{OUT}} = - \sum_{j=1}^N E_j \frac{Z_K}{Z_j} \quad (12)$$

The accuracy of the operational amplifier depends on the magnitude of $A\beta$ and the precision of the components used in the input and feedback networks as can be seen from (11). There is negligible interaction between the input voltages because the input impedance at the summing node is equal to Z_{IN}' divided by $(1 - A\beta)$.⁹ This impedance is usually negligibly small compared to the impedances used in the input circuit.

* In general, E_j and E_{OUT} are the Laplace transforms of the input and output voltages, respectively.

2.2. Methods Used to Shape the Loop Current Transmission

An essential consideration in the design of a feedback amplifier is the provision of adequate margins against instability. In order to accomplish this objective, it is necessary to choose a criterion of stability. In Appendix I it is shown that it is convenient and valid to base the stability of single loop transistor feedback amplifiers on the loop current transmission. In order to calculate the loop current transmission of the de amplifier, the feedback loop is opened at a convenient point in the circuit, usually at the base of one of the transistors, and a unit current is injected into the base (refer to Fig. 24). The other side of the opened loop is connected to ground through a resistance ($r_e + r_b$) and voltage $r_e I_4$. In many instances, the voltage $r_e I_4$ can be neglected. If $|Z_K|$ and

$$\frac{1}{\left| \sum_{j=1}^N \frac{1}{Z_j} \right|}$$

are much greater than $|Z_{IX'}|$, then $A\beta$ is very nearly equal to the loop current transmission. For absolute stability¹⁰ the amplitude of the loop current transmission must be less than unity before the phase shift (from the low frequency value) exceeds 180° . Consequently, this characteristic must be controlled or properly shaped over a wide frequency

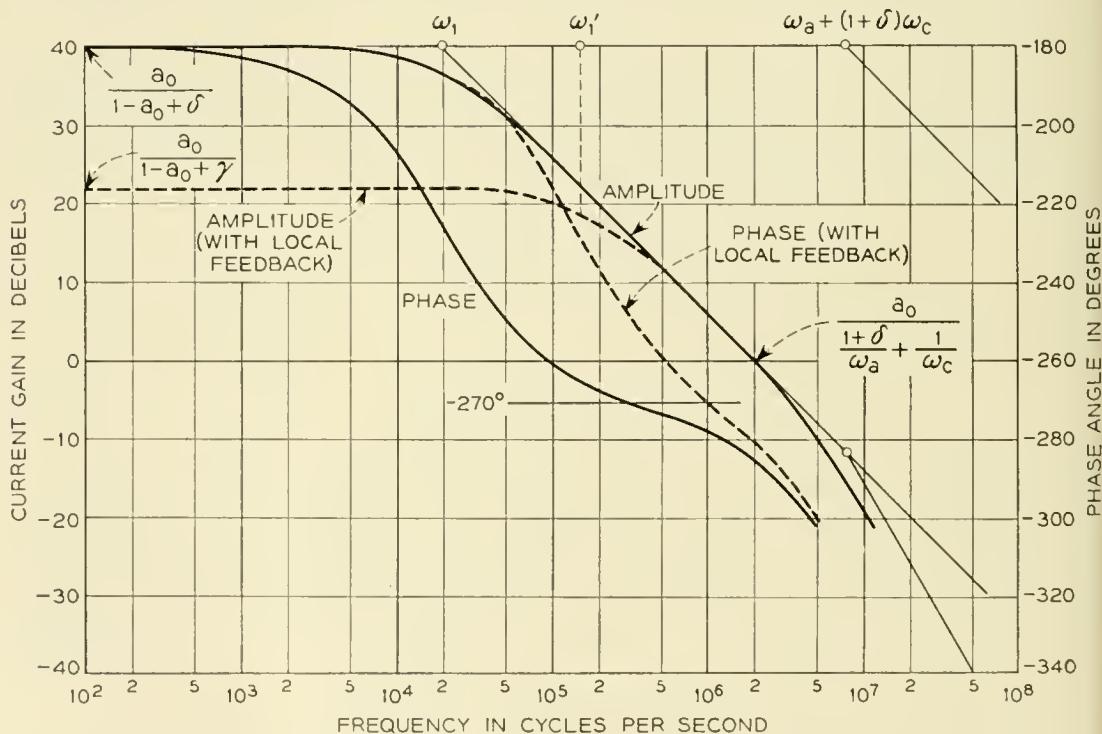


Fig. 5 — Current transmission of a common emitter stage.

band. In addition, it is desirable that the feedback fall off at a rate equal to or less than 9 db per octave in order to insure that the dc amplifier has a satisfactory transient response.

Three methods of shaping are described in this paper; local feedback shaping, interstage network shaping, and β circuit shaping. Local feedback shaping will be described first. The analysis starts by considering the current transmission of a common emitter stage, equivalent circuit shown in Fig. 2(b). If the stage operates into a load resistance R_L , then to a good approximation the current transmission is given by

$$G_I = \frac{I_2}{I_1} = \frac{\frac{a_0}{1 - a_0 + \delta}}{1 + \frac{p}{\omega_1} + \frac{p^2}{\omega_a \omega_c (1 - a_0 + \delta)}} \quad (13)^*$$

where

$$\begin{aligned} \delta &= \frac{R_L + r_e}{r_c} \\ \omega_1 &= \frac{(1 - a_0 + \delta)}{\frac{1 + \delta}{\omega_a} + \frac{1}{\omega_c}} \\ \frac{\omega_a}{2\pi} &\approx \text{alpha-cutoff frequency} \\ \omega_c &= \frac{1}{(R_L + r_e)C_c} \end{aligned}$$

It is apparent from expression (13) that if $(1 - a_0 + \delta)$ is less than 0.1, then the current gain of the common emitter stage falls off at a rate of 6 db per octave with a corner frequency at ω_1 .† A second 6 db per octave cutoff with a corner frequency at $[\omega_a + (1 + \delta)\omega_c]$ is introduced by the p^2 term in the denominator of (13). A typical transmission characteristic is shown in Fig. 5. The current gain of the common emitter stage is unity at a frequency equal to

$$\frac{a_0}{\frac{1 + \delta}{\omega_a} + \frac{1}{\omega_c}}$$

* Expressions (13) and (14) are poor approximations at frequencies above $\omega_a/2\pi$.

† Strictly speaking the corner frequency is equal to $\omega_1/2\pi$. However, for simplicity, corner frequencies will be expressed as radian frequencies.

Since the phase crossover of $A\beta^*$ is usually placed below this frequency, the principal effect of the second cutoff is to introduce excess phase. This excess phase can be minimized by operating the stage into the smallest load resistance possible, thus maximizing ω_c .

An undesirable property of the common emitter transmission characteristic is that the corner frequency ω_1 occurs at a relatively low frequency. However, the corner frequency can be increased by using local feedback as shown in Fig. 6(a). Shunt feedback is used in order to provide a low input impedance for the preceding stage to operate into. The amplitude and phase of the current transmission is controlled principally by the impedances Z_1 and Z_2 . If $|A\beta|$ is much greater than one, and if $\beta \approx Z_1/Z_2$, then from (7) the current transmission of the stage is approximately equal to $-Z_2/Z_1$. Because of the relatively small size of $A\beta$ for a single stage, this approximation is only valid for a very limited range of values of Z_1 and Z_2 . If Z_1 and Z_2 are represented as resistances R_1 and R_2 , then the current transmission of the circuit is given to a good approximation by

$$G_I = \frac{I_2}{I_1} = -\frac{R_2}{(R_2 + r_b)} \cdot \frac{\frac{a_0}{1 - a_0 + \gamma}}{\left[1 + \frac{p}{\omega_1'} + \frac{p^2}{\omega_a \omega_c (1 - a_0 + \gamma)} \right]} \quad (14)$$

where

$$\gamma = \frac{R_1 + r_e}{(R_2 + r_b)r_e} \approx \frac{R_1 + r_e}{R_2 + r_b}$$

$$\omega_1' = \frac{(1 + a_0 + \gamma)}{\frac{1 + \gamma}{\omega_a} + \frac{1}{\omega_c}}$$

$$\omega_c = \frac{1}{(R_1 + r_e)C_c}$$

By comparing (14) with (13), it is evident that the negative feedback has reduced the low-frequency current gain from $a_0/(1 - a_0)$ (δ may usually be neglected) to

$$\left(\frac{R_2}{R_2 + r_b} \right) \left(\frac{a_0}{1 - a_0 + \gamma} \right) \approx \frac{R_2}{R_1 + r_e} \quad (\text{if } \gamma > 1 - a_0)$$

* The phase crossover of $A\beta$ is equal to the frequency at which the phase shift of $A\beta$ from its low-frequency value is 180° .

The half power frequency, however, has been increased from

$$\frac{1 - a_0}{\frac{1}{\omega_a} + \frac{1}{\omega_c}} \quad \text{to} \quad \frac{1 - a_0 + \gamma}{\frac{1 + \gamma}{\omega_a} + \frac{1}{\omega_c}}$$

as shown by the dashed curves in Fig. 5.*

The bandwidth of the common emitter stage can be increased without reducing the current gain at dc and low-frequencies by representing Z_1 by a resistance R_1 , and Z_2 by a resistance R_2 in series with a condenser C_2 . If $1/R_2C_2$ is much smaller than ω_1' , then the current transmission of the stage is given by (14) multiplied by the factor

$$\frac{\left(1 + \frac{\omega_2}{p}\right)}{\left(1 + \frac{\omega_4}{p}\right)} \quad (15)$$

where

$$\omega_2 = \frac{1}{R_2 C_2}$$

$$\omega_4 = \frac{1 - a_0 + \frac{R_1 + r_e}{r_e}}{C_2(R_2 + r_b)(1 - a_0 + \gamma)}$$

The current transmission for this case is plotted in Fig. 6(b). The condenser C_2 introduces a rising 6 db per octave asymptote with a corner frequency at ω_2 . At dc the current gain is equal to

$$\frac{a_0}{1 - a_0 + \delta}$$

A second method of shaping the loop current transmission characteristic of a feedback amplifier is by means of interstage networks. These networks are usually used for reducing the loop current gain at relatively low frequencies while introducing negligible phase lag near the gain† and phase crossover frequencies. Interstage networks should be designed to take advantage of the variable transistor input impedance. The input impedance of a transistor in the common emitter connection

* In Figs. 5 and 6(b), the factor $R_2/(R_2 + r_b)$ is assumed equal to unity. This is a good approximation since in practice R_2 is equal to several thousand ohms while r_b is equal to about 100 ohms.

† The gain crossover frequency is equal to the frequency at which the magnitude of $A\beta$ is unity.

is given by the expression

$$Z_{\text{INPUT}} = r_b + r_e(1 - G_I) \quad (16)$$

where G_I is the current transmission given by (13). If G_I at dc is much greater than 1, then the input impedance and the current transmission of the common emitter stage fall off at about the same rate and with approximately the same corner frequency (ω_1). The input impedance finally reaches a limiting value equal to $r_e + r_b$.

A particularly useful interstage network is shown in Fig. 7(a). This network is analyzed in Appendix II and Fig. 7(b) shows a plot of the

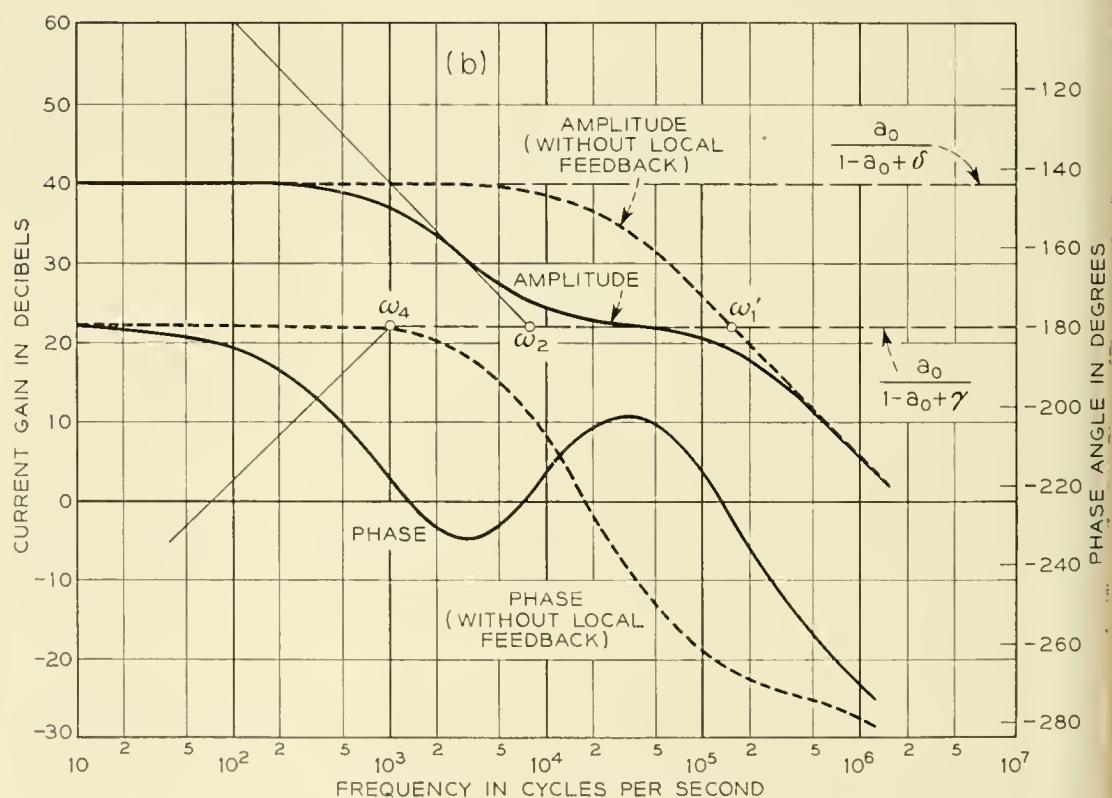
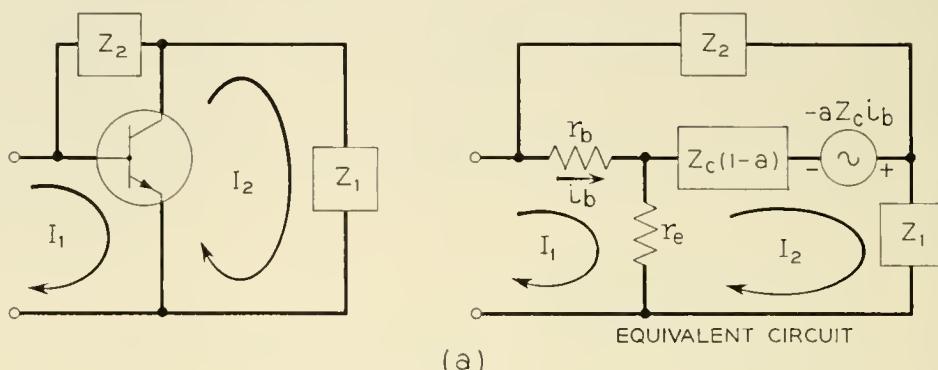


Fig. 6 — Negative feedback applied to a common emitter stage.

resulting current transmission. The amplitude of the transmission falls off at a rate of 6 db per octave with the corner frequency ω_5 determined by C_3 and the low frequency value of the transistor input impedance. The inductance L_3 introduces a 12 db per octave rising asymptote with a corner frequency at $\omega_3 = 1/\sqrt{L_3 C_3}$. The corner frequencies ω_3 and ω_5 are selected in order to obtain a desirable loop current transmission characteristic (specific transmission characteristics are presented in Sections 3.0 and 4.0). The half power frequency of the current transmission of the transistor, ω_1 , does not appear directly in the transmission characteristic of the circuit because of the variation in the transistor input impedance with frequency.

The overall β circuit of the feedback amplifier can also be used for

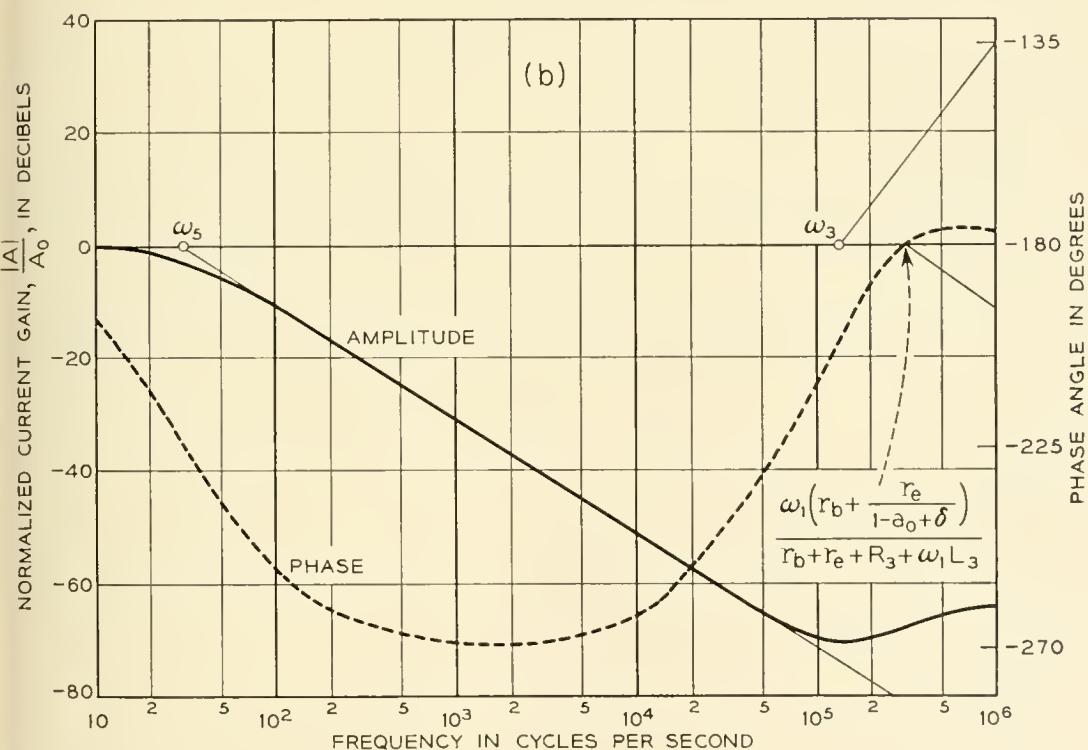
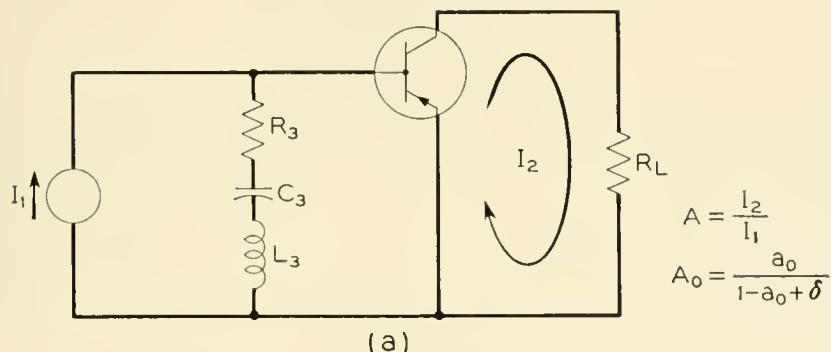


Fig. 7 — Interstage shaping network.

shaping the loop current transmission. If the feedback impedance Z_K (Fig. 4) consists of a resistance R_K and condenser C_K in parallel, then the loop current transmission is modified by the factor

$$\frac{\left(1 + \frac{p}{\omega_7}\right)}{\left(1 + \frac{p}{\omega_8}\right)} \quad (17)$$

where

$$\omega_7 = \frac{1}{R_K C_K}$$

$$\omega_8 = \frac{(R_L + R_K)}{R_L R_K C_K}$$

Since Z_K affects the external voltage gain of the operational amplifier, (11), the corner frequency ω_7 must be located outside of the useful frequency band. Usually it is placed near the gain crossover frequency in order to improve the phase margin and the transient response of the amplifier.

In Sections 3.0 and 4.0, the above shaping techniques are used in the design of specific operational amplifiers.

3.0. THE SUMMING AMPLIFIER

3.1. Circuit Arrangement

The schematic diagram of a dc summing amplifier is shown in Fig. 8. From the discussion in Section 2.0 it is apparent that each common emitter stage will contribute more than 90 degrees of high-frequency phase lag. Consequently, while the magnitude of the low-frequency feedback increases with the number of stages, this is at the expense of the bandwidth over which the negative feedback can be maintained. It is possible to develop 80 db of negative feedback at dc with three common emitter stages. This corresponds to a dc accuracy of one part in 10,000. In addition, the feedback can be maintained over a broad enough band in order to permit full accuracy to be attained in about 100 microseconds. Thus it is evident that the choice of three stages represents a satisfactory compromise between accuracy and bandwidth objectives.

The output stage of the amplifier is designed for a maximum power dissipation of 75 milliwatts and maximum voltage swing of ± 25 volts

when operating into an external load resistance equal to or greater than 50,000 ohms. A p-n-p transistor is used in the second stage and n-p-n transistors are used in the first and third stages. This circuit arrangement makes it possible to connect the collector of one transistor directly to the base of the following transistor without introducing appreciable interstage loss. "Shot" noise¹¹ and dc drift are minimized by operating the first stage at the relatively low collector current of 0.25 milliamperes. The 110,000-ohm resistor provides the collector current for the first stage, and the 4,700-ohm resistor provides 3.8 milliamperes of collector current for the second stage. The series 6,800-ohm resistor between the second and third stages, reduces the collector to emitter potential of the second stage to about 4.5 volts.

The loop current transmission is shaped by use of local feedback applied to the second stage, by an interstage network connected between the second and third stages, and by the overall β circuit. The 200-ohm resistor in the collector circuit of the second stage is, with reference to Fig. 6(a), Z_1 . The impedance of the interstage network can be neglected since it is small compared to 200 ohms at all frequencies for which the local feedback is effective. The interstage network is connected between the second and third stages in order to minimize the output noise voltage. With this circuit arrangement, practically all of the output noise voltage

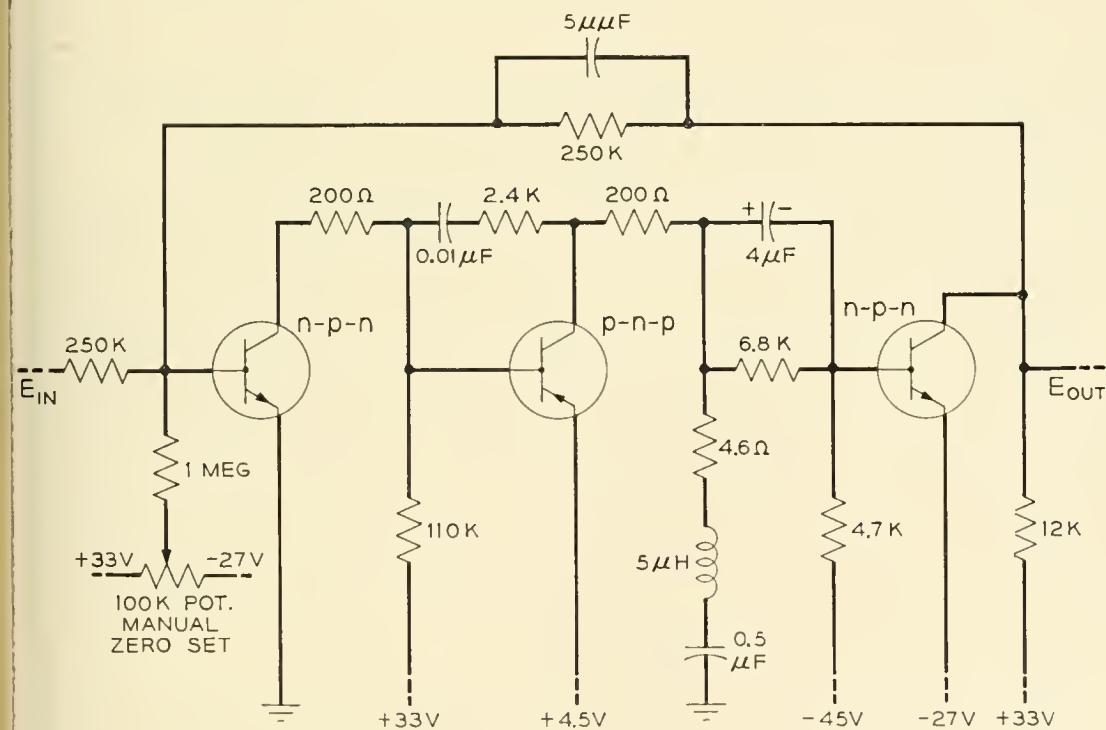


Fig. 8 — DC summing amplifier.

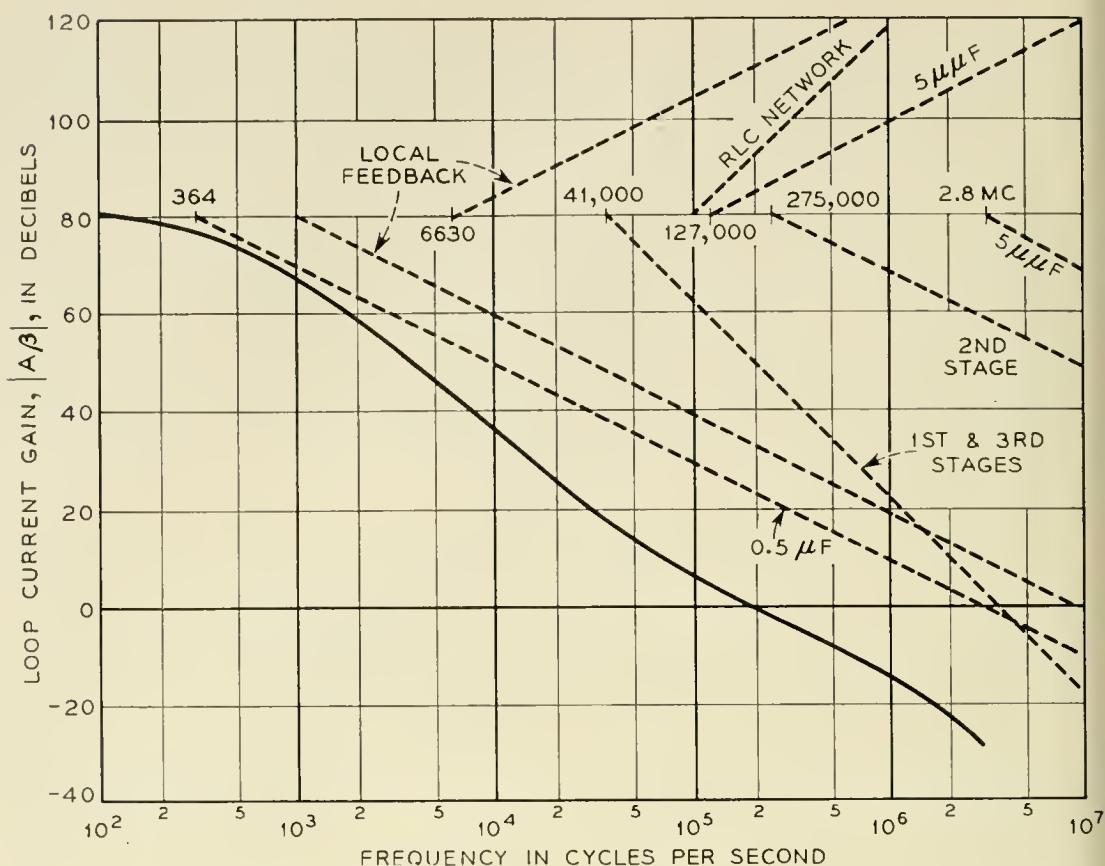


Fig. 9 — Gain-frequency asymptotes for summing amplifier.

is generated in the first transistor stage. If the transistor in the first stage has a noise figure less than 10 db at 1,000 cycles per second, then the RMS output noise voltage is less than 0.5 millivolts.

Fig. 9 shows a plot of the gain-frequency asymptotes for the summing amplifier determined from (13), (14), (15), (17), and (A6) under the assumption that the alphas and alpha-cutoff frequencies of the transistors are 0.985 and 3 mc, respectively. The corner frequencies introduced by the 0.5 microfarad condenser in the interstage network, the local feedback circuit, and the cutoff of the first and third stages are so located that the current transmission falls off at an initial rate of about 9 db per octave. This slope is joined to the final asymptote of the loop transmission by means of a step-type of transition.¹² The transition is provided by 3 rising asymptotes due to the interstage shaping network, and the overall β circuit. An especially large phase margin is used in order to insure a good transient performance.

Fig. 10 shows the amplitude and phase of the loop current transmission. When the amplitude of the transmission is 0 db, the phase angle is -292° , and when the phase angle is -360° , the amplitude is 27.5 db.

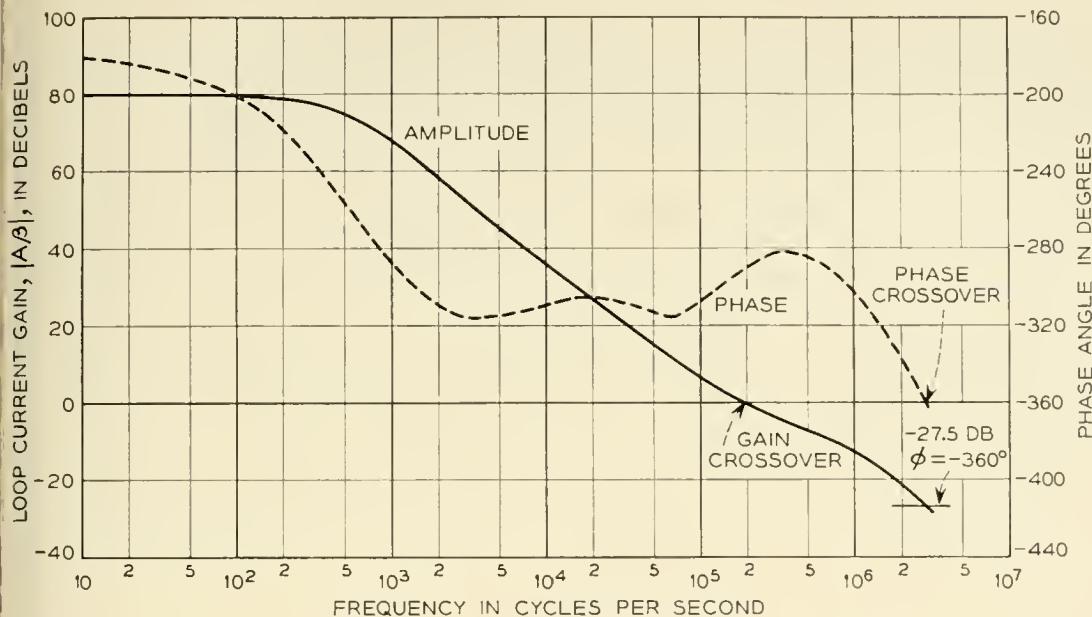


Fig. 10 — Loop current transmission of the summing amplifier.

below 0 db. The amplifier has a 68° phase margin and 27.5 db gain margin. In order to insure sufficient feedback at dc and adequate margins against instability, the transistors used in the amplifier should have alphas in the range 0.98 to 0.99 and alpha-cutoff frequencies equal to or greater than 2.5 mc.

3.2. Automatic Zero Set of the dc Summing Amplifier

The application of germanium junction transistors to dc amplifiers does not eliminate the problem of drift normally encountered in vacuum tube circuits. In fact, drift is more severe due principally to the variation of the transistor parameters alpha and saturation current with temperature variation. Even though the amplifier has 80 db of negative feedback at dc, this feedback does not eliminate the drift introduced by the first transistor stage. Because of the large amount of dc feedback, the collector current of the first stage is maintained relatively constant. The collector current of the transistor is related to the base current by the equation

$$I_c = \frac{I_{co}}{1-a} + \frac{a}{1-a} I_b \quad (18)$$

The saturation current, I_{co} , of a germanium junction transistor doubles approximately for every 11°C increase in temperature. The factor $a/(1-a)$ increases by as much as 6 db for a 25°C increase in tempera-

ture. Consequently, the base current of the first stage, I_b , and the output voltage of the amplifier must change with temperature in order to maintain I_c constant. The drift due to the temperature variation in a can be reduced by operating the first stage at a low value of collector current. With a germanium junction transistor in the first stage operating at a collector current of 0.25 milliamperes, the output voltage of the amplifier drifts about ± 1.5 volts over a temperature range of 0°C to 50°C . It is possible to reduce the dc drift by using temperature sensitive elements in the amplifier.^{13, 14} In general, temperature compensation of a transistor dc amplifier requires careful selection of transistors and critical adjustment of the dc biases. However, even with the best adjustments, temperature compensation cannot reduce the drift in the amplifier to within typical limits such as ± 5 millivolts throughout a temperature range of 0 to 50°C . In order to obtain the desired accuracy it is necessary to use an automatic zero set (AZS) circuit.

Fig. 11 shows a dc summing amplifier and a circuit arrangement for reducing any dc drift that may appear at the output of the amplifier. The output voltage is equal to the negative of the sum of the input voltages, where each input voltage is multiplied by the ratio of the feedback resistor to its input resistor. In addition, an undesirable dc drift voltage is also present in the output voltage. The total output voltage is

$$E_{\text{out}} = - \sum_{j=1}^N E_j \frac{R_K}{R_j} + E_{\text{drift}} \quad (19)$$

In order to isolate the drift voltage, the N input voltages and the output voltage are applied to a resistance summing network composed of resistors $R_0, R'_1, R'_2, \dots, R'_N$. The voltage across R_s is equal to

$$E_s = \frac{R_s}{R_0} E_{\text{drift}} \quad (20)$$

if

$$R_s \ll R_0, R'_j; \quad j = 1, 2, \dots, N$$

and

$$R_0 R_j = R_K R'_j; \quad j = 1, 2, \dots, N$$

The voltage E_s is amplified in a relatively drift-free narrow band dc amplifier and is returned as a drift correcting voltage to the input of the dc summing amplifier. If the gain of the AZS circuit is large, the drift voltage at the output of the summing amplifier can be made very small.

Fig. 12 shows the circuit diagram of a summing amplifier which uses a mechanical chopper in the AZS circuit.¹⁵ The AZS circuit consists of a

resistance summing network, a 400-cycle synchronous chopper, and a tuned 400-cycle amplifier. Any drift in the summing amplifier will produce a dc voltage E_s at the output of the summing network. The chopper converts the dc voltage into a 400 cycles per second waveform. The fundamental frequency in the waveform is amplified by a factor of about 400,000 by the tuned amplifier. The synchronous chopper rectifies the sinusoidal output voltage and preserves the original dc polarity of E_s . The rectified voltage is filtered and fed back to the summing amplifier as an additional input current. The loop voltage gain of the AZS circuit at dc is about 54 db. Any dc or low-frequency drift in the summing amplifier is reduced by a factor of about 500 by the AZS circuit. The drift throughout a temperature range of 0 to 50°C is reduced to ± 3 millivolts.

Since the drift in the summing amplifier changes at a relatively slow rate, the loop voltage gain of the AZS circuit can be cutoff at a relatively low frequency. In this particular case the loop voltage gain is zero db at about 10 cycles per second.

4.0. THE INTEGRATOR

4.1. Basic Design Considerations

The design principles previously discussed are illustrated in this section by the design of a transistor integrator for application in a voltage

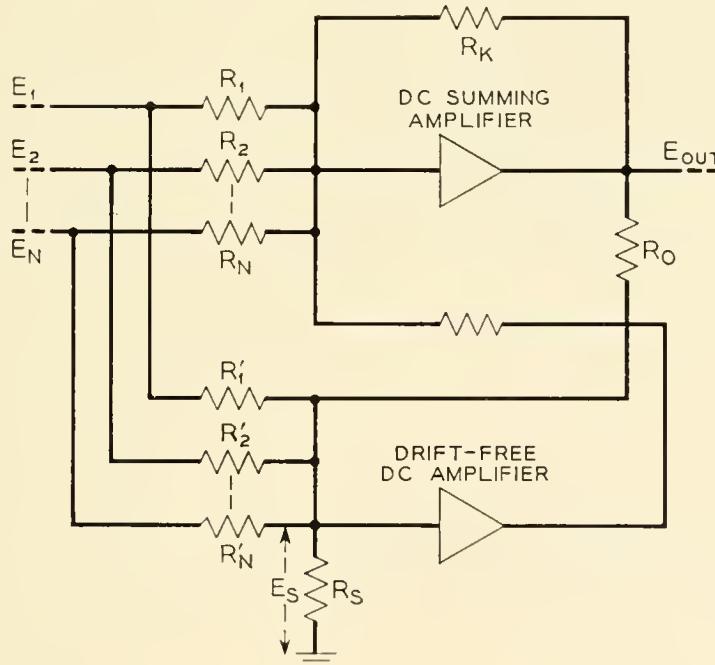


Fig. 11 — DC summing amplifier with automatic zero set.

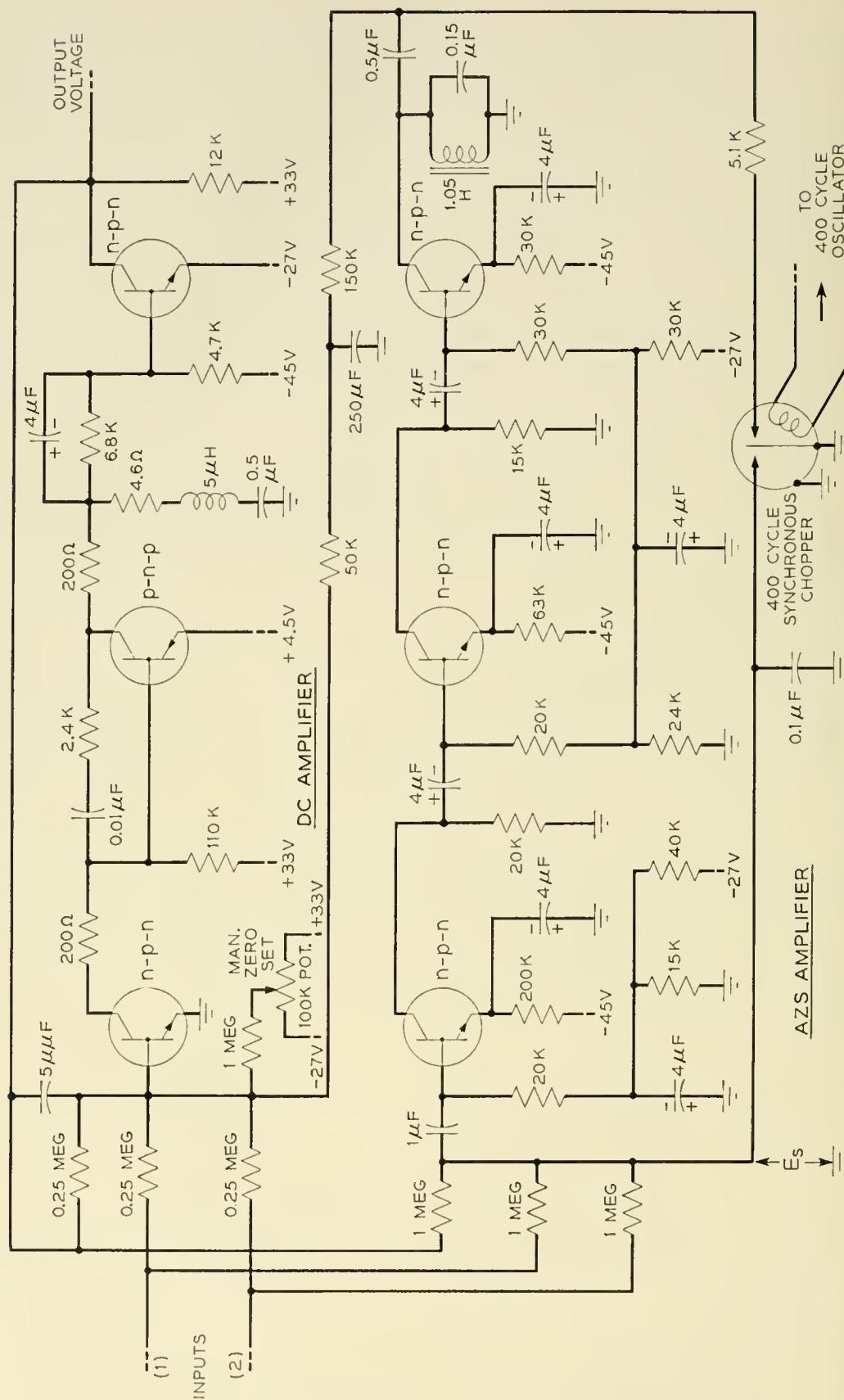


Fig. 12 — Summing amplifier.

encoder. The integrator is required to generate a 15-volt ramp which is linear and has a constant slope to within one part in 8,000. This ramp is to have a slope of 5 millivolts per microsecond for an interval of 3,000 microseconds.

The first step in the design is to determine the bandwidth over which the negative feedback must be maintained in order to realize the desired output voltage linearity. The relationship between the output and input voltage of the integrator can be obtained from expression (11) by substituting $(1/pc)$ for Z_K and R for Z_j (refer to Fig. 1).

$$\mathcal{L}[E_{\text{OUT}}] = \frac{\mathcal{L}[E_{\text{IN}}]}{pRC} \left[\frac{A\beta + Z_{\text{IN}}' pC}{1 - A\beta + \frac{Z_{\text{IN}}'}{R}} \right] \quad (21)$$

where $\mathcal{L}[E_{\text{OUT}}]$ and $\mathcal{L}[E_{\text{IN}}]$ are the Laplace transforms of the output and input voltages, respectively. In order to generate the voltage ramp, a step voltage of amplitude E is applied to the input of the integrator. The term Z_{IN}'/R is negligible compared to unity at all frequencies. Therefore,

$$\mathcal{L}[E_{\text{OUT}}] = \frac{E}{p^2 RC} \left[\frac{A\beta}{1 - A\beta} \right] + \frac{EZ_{\text{IN}}'}{pR} \left[\frac{1}{1 - A\beta} \right] \quad (22)$$

It will be assumed that $A\beta$ is given by the expression

$$A\beta = \frac{-K}{\left(1 + \frac{\omega_1}{p}\right)\left(1 + \frac{p}{\omega_2}\right)^2} \quad (23)$$

Expression (23) implies that $A\beta$ falls off at a rate of 6 db per octave at low frequencies and 12 db per octave at high frequencies. The output voltage of the integrator, as a function of time, is readily evaluated by substituting (23) into (22) and taking the inverse Laplace transform of the results. A good approximation for the output voltage is

$$E_{\text{OUT}} = -\frac{E}{RC} \left[t - \frac{\omega_1 t^2}{2K} - \frac{e^{-[(2\omega_2 + \omega_1)t/2]} \sin \sqrt{K\omega_2}t}{\sqrt{K\omega_2}} \right] + \frac{ER_{\text{IN}}'}{R} [1 - e^{-(\omega_1 t/K)} + e^{-[(2\omega_2 + \omega_1)t/2]} \cos \sqrt{K\omega_2}t] \quad (24)*$$

The linear voltage ramp is expressed by the term $-(Et/RC)$. The additional terms introduce nonlinearities. The voltage ramp has a slope of 5 millivolts per microsecond for $E = -21$ volts, $R = 42,000$ ohms,

* In evaluating E_{OUT} it was assumed that Z_{IN}' was equal to a fixed resistance R_{IN}' , the low frequency input resistance to the first common emitter stage. A complete analysis indicates that this assumption makes the design conservative.

and $C = 0.1$ microfarads. For these circuit values, and $K = 10,000$ (corresponding to 80 db of feedback) the nonlinear terms are less than $1/8,000$ of the linear term (evaluated when $t = 4 \times 10^{-3}$ seconds) if $f_1 \leq 30$ cycles per second, $f_2 \geq 800$ cycles per second, and if the first 1000 microseconds of the voltage ramp are not used. Consequently, 80 db of negative feedback must be maintained over a band extending from 30 to 800 cycles per second in order to realize the desired output voltage linearity.

4.2. Detailed Circuit Arrangement

Fig. 13 shows the circuit diagram of the integrator. The method of biasing is the same as is used in the summing amplifier. The 200,000-ohm resistor provides approximately 0.5 milliamperes of collector current for the first stage. The 40,000-ohm resistor provides approximately 0.9 milliamperes of collector current for the second stage. The output stage is designed for a maximum power dissipation of 120 milliwatts and for an output voltage swing between -5 and +24 volts when operating into a load resistance equal to or greater than 40,000 ohms.

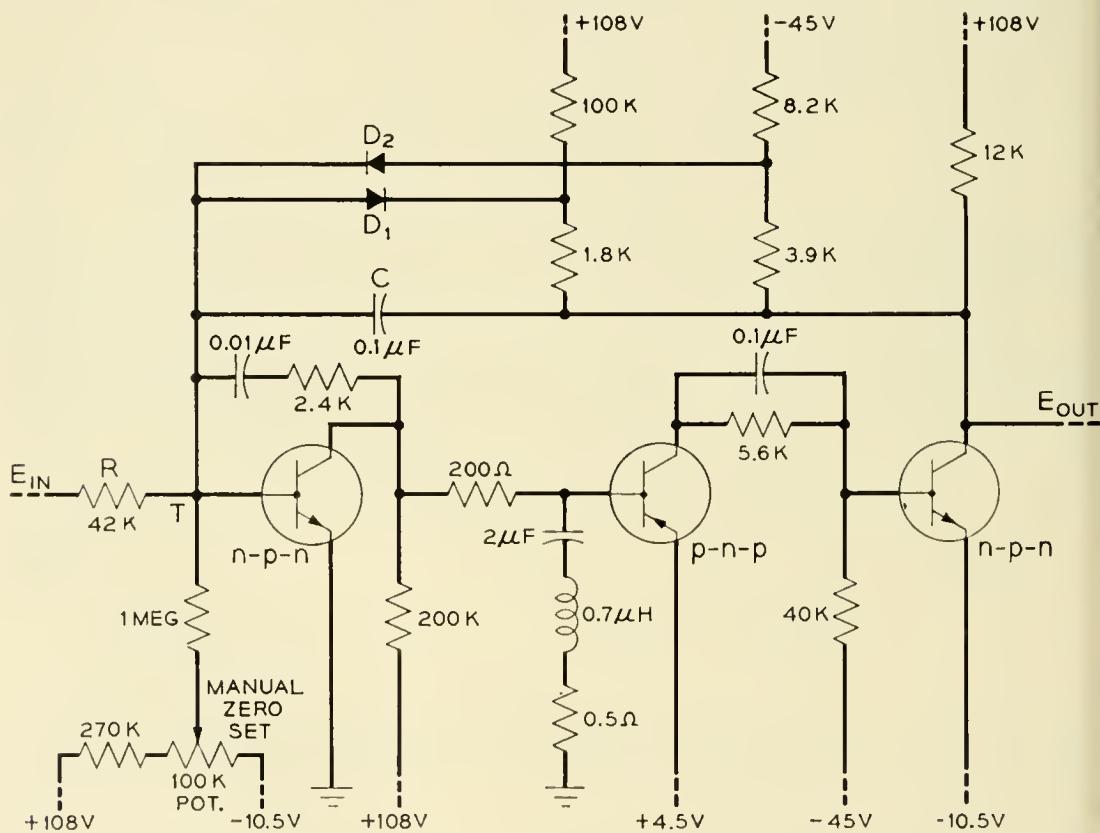


Fig. 13 — Integrator.

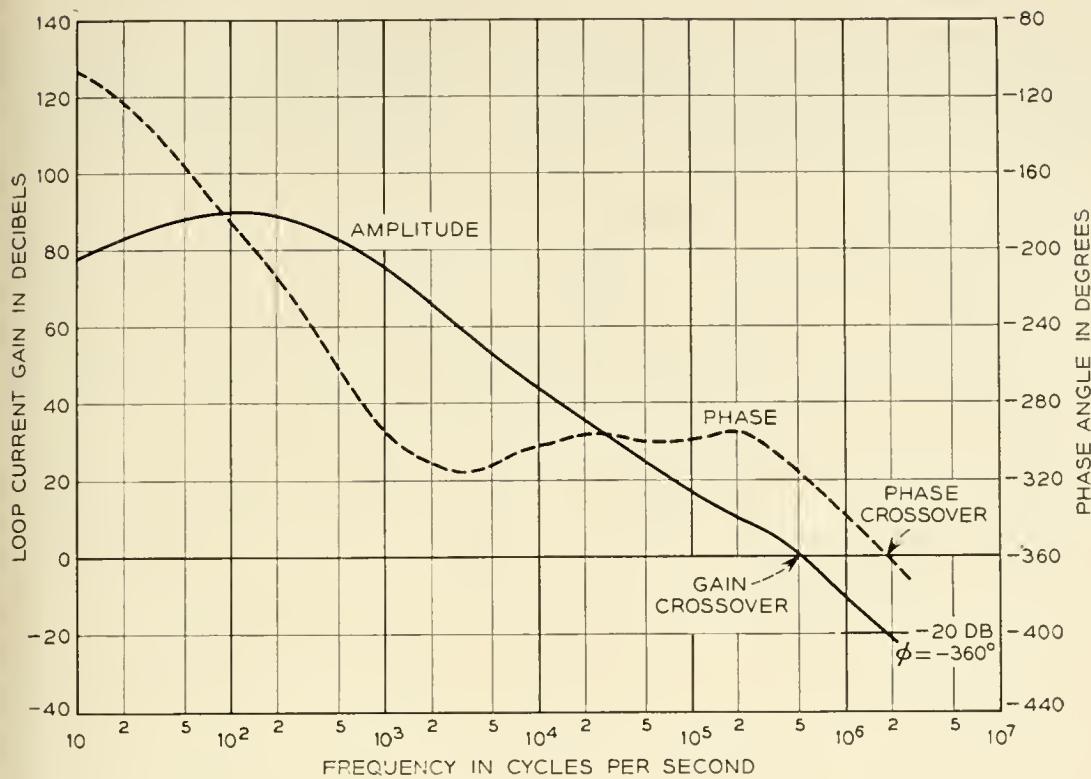


Fig. 14 — Loop current transmission of the integrator.

The negative feedback in the integrator has been shaped by means of local feedback and interstage networks as described in Section 2.2. The loop current transmission has been calculated from (13), (14), (15), and (A6) and is plotted in Fig. 14. The transmission is determined under the assumption that the alphas of the transistors are 0.985 and the alpha-cutoff frequencies are three megacycles. Since the feedback above 800 cycles per second falls off at a rate of 9 db per octave, the analysis in Section 4.1 using (23), is conservative. The integrator has a 44° phase margin and a 20 db gain margin. In order to insure sufficient feedback between 30 and 800 cycles per second and adequate margins against instability, the transistors used in the integrator should have alphas in the range 0.98 to 0.99 and alpha-cutoff frequencies equal to or greater than 2.5 megacycles.

The silicon diodes D_1 and D_2 are required in order to prevent the integrator from overloading. For output voltages between -4.0 and 21 volts the diodes are reverse biased and represent very high resistances, of the order of 10,000 megohms. If the output voltage does not lie in this range, then one of the diodes is forward biased and has a low resistance, of the order of 100 ohms. The integrator is then effectively a de amplifier with a voltage gain of approximately 0.1. The silicon diodes affect the

linearity of the voltage ramp slightly due to their finite reverse resistances and variable shunt capacities. If the diodes have reverse resistances greater than 1000 megohms, and if the maximum shunt capacity of each diode is less than 10 micromicrofarads (capacity with minimum reverse voltage), then the diodes introduce negligible error.

As stated earlier, the integrator generates a voltage ramp in response to a voltage step. This step is applied through a transistor switch which is actuated by a square wave generator capable of driving the transistor well into current saturation. Such a switch is required because the equivalent generator impedance of the applied step voltage must be very small. A suitable circuit arrangement is shown in Fig. 15. For the particular application under discussion the switch S is closed for 5,000 microseconds. During this time, the voltage $E = -21V$ appears at the input of the integrator. At the end of this time interval, the transistor switch is opened and a reverse current is applied to the feedback condenser C , returning the output voltage to -4.0 volts in about 2500 microseconds. An alternate way of specifying a low impedance switch is to say that the voltage across it be close to zero. For the transistor switch, connected as shown in Fig. 15, this means that its collector voltage be within

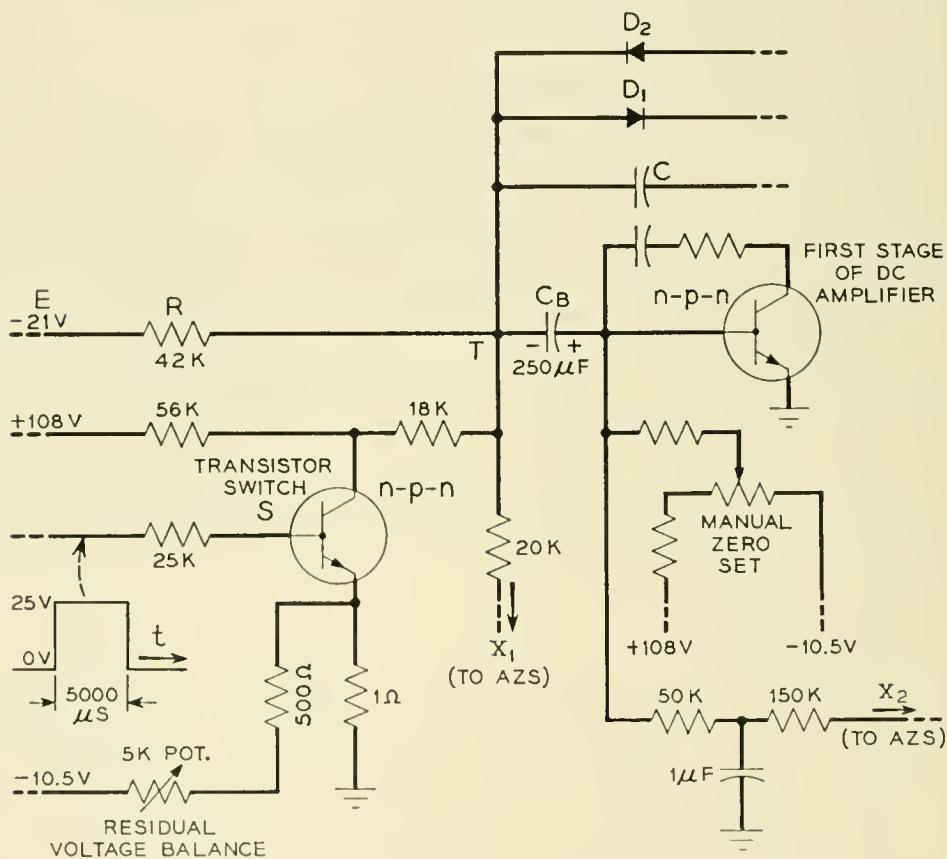


Fig. 15 — Input circuit arrangement of the integrator.

one millivolt of ground potential during the time the transistor is in saturation. Now, it has been shown¹⁶ that when a junction transistor in the common emitter connection is driven into current saturation, the minimum voltage between collector and emitter is theoretically equal to

$$\frac{kT}{q} \ln \frac{1}{\alpha_i} \quad (25)$$

where k is the Boltzmann constant, T is the absolute temperature, q is the charge of an electron ($(kT/q) = 26$ millivolts at room temperature), and α_i is the inverse alpha of the transistor, i.e., the alpha with the emitter and collector interchanged. There is an additional voltage drop across the transistor due to the bulk resistance of the collector and emitter regions (including the ohmic contacts). A symmetrical alloy junction transistor with an alpha close to unity is an excellent switch because both the collector to emitter voltage and the collector and emitter resistances are very small.

At the present time, a reasonable value for the residual voltage* between the collector and emitter is 5 to 10 millivolts. This voltage can be eliminated by returning the emitter of the transistor switch to a small negative potential. This method of balancing is practical because the voltage between the collector and emitter of the transistor does not change by more than 1.0 millivolt over a temperature range of 0°C to 50°C.

4.3. Automatic Zero Set of the Integrator

A serious problem associated with the transistor integrator is drift. The drift is introduced by two sources; variations in the base current of the first transistor stage and variations in the base to emitter potential of the first stage with temperature. In order to reduce the drift, the input resistor R and the feedback condenser C must be dissociated from the base current and base to emitter potential of the first transistor stage. This is accomplished by placing a blocking condenser C_B between point T and the base of the first transistor as shown in Fig. 15. An automatic zero set circuit is required to maintain the voltage at point T equal to zero volts. This AZS circuit uses a magnetic modulator known as a "magnetotor."¹⁷

A block diagram of the AZS circuit is shown in Fig. 16. The dc drift current at the input of the amplifier is applied to the magnetotor. The carrier current required by the magnetotor is supplied by a local transistor

* The inverse alphas of the transistors used in this application were greater than 0.95.

oscillator. The useful output of the magnetotor is the second harmonic of the carrier frequency. The amplitude of the second harmonic signal is proportional to the magnitude of the dc input current and the phase of the second harmonic signal is determined by the polarity of the dc input current. The output voltage of the magnetotor is applied to an active filter which is tuned to the second harmonic frequency. The signal is then amplified in a tuned amplifier and applied to a diode gating circuit. Depending on the polarity of the dc input current, the gating circuit passes either the positive or negative half cycle of the second harmonic signal. In order to accomplish this, a square wave at a repetition rate equal to that of the second harmonic signal is derived from the carrier oscillator and actuates the gating circuit.

A circuit diagram of the AZS circuit is shown in Figs. 17(a) and 17(b). The various sections of the circuit are identified with the blocks shown in Fig. 16. The active filter is adjusted for a Q of about 300, and the gain of the active filter and tuned amplifier is approximately 1000. The AZS circuit provides ± 1.0 volt of dc output voltage for ± 0.05 microamperes of dc input current. The maximum sensitivity of the circuit is limited to ± 0.005 microamperes because of residual second harmonic generation in the magnetotor with zero input current.

When the transistor integrator is used together with the magnetotor AZS circuit, the slope of the voltage ramp is maintained constant to within one part in 8,000 over a temperature range of 20°C to 40°C.

5.0. The Voltage Comparator

The voltage comparator is one of the most important circuits used in analog to digital converters. The comparator indicates the exact time that an input waveform passes through a predetermined reference level. It has been common practice to use a vacuum tube blocking oscillator as a voltage comparator.¹⁸ Due to variations in the contact potential, heater voltage, and transconductance of the vacuum tube, the maximum

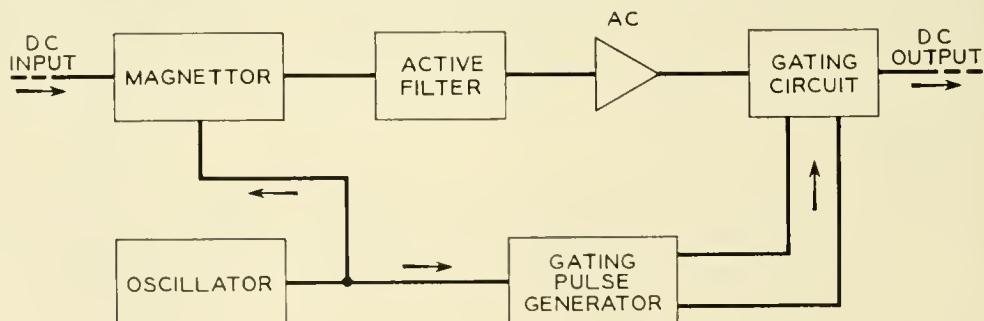


Fig. 16 — Block diagram of AZS circuit.

accuracy of the circuit is limited to about ± 100 millivolts. By taking advantage of the properties of semiconductor devices, the transistor blocking oscillator comparator can be designed to have an accuracy of ± 5 millivolts throughout a temperature range of 20°C to 40°C .

5.1. General Description of the Voltage Comparator

Fig. 18 shows a simplified circuit diagram of the voltage comparator. Except for the silicon junction diode D_1 , this circuit is essentially a transistor blocking oscillator. For the purpose of analysis, assume that the reference voltage V_{ee} is set equal to zero. When the input voltage V_i is large and negative, the silicon diode D_1 is an open circuit and the junction transistor has a collector current determined by R_b and E_{bb} [Expression (18)]. The base of the transistor resides at approximately -0.2 volts. As the input voltage V_i approaches zero, the reverse bias across the diode D_1 decreases. At a critical value of V_i (a small positive potential), the dynamic resistance of the diode is small enough to permit the circuit to become unstable. The positive feedback provided by transformer T_1 forces the transistor to turn off rapidly, generating a sharp output pulse across the secondary of transformer T_2 . When V_i is large and positive, the diode D_1 is a low impedance and the transistor is maintained cutoff. In order to prevent the comparator from generating more than one output pulse during the time that the circuit is unstable, the natural period of the circuit as a blocking oscillator must be properly chosen. Depending on this period, the input voltage waveform must have a certain minimum slope when passing through the reference level in order to prevent the circuit from misfiring.

The comparator has a high input impedance except during the switching interval.* When V_i is negative with respect to the reference level, the input impedance is equal to the impedance of the reverse biased silicon diode. When V_i is positive with respect to the reference level, the input impedance is equal to the impedance of the reverse biased emitter and collector junctions in parallel. This impedance is large if an alloy junction transistor is used. During the switching interval the input impedance is equal to the impedance of a forward biased silicon diode in series with the input impedance of a common emitter stage (approximately 1,000 ohms). This loading effect is not too serious since for the circuit described, the switching interval is less than 0.5 microseconds.

The voltage comparator shown in Fig. 18 operates accurately on voltage waveforms with positive slopes. The voltage comparator will operate accurately on waveforms with negative slopes if the diode and

* The switching interval is the time required for the transistor to turn off.

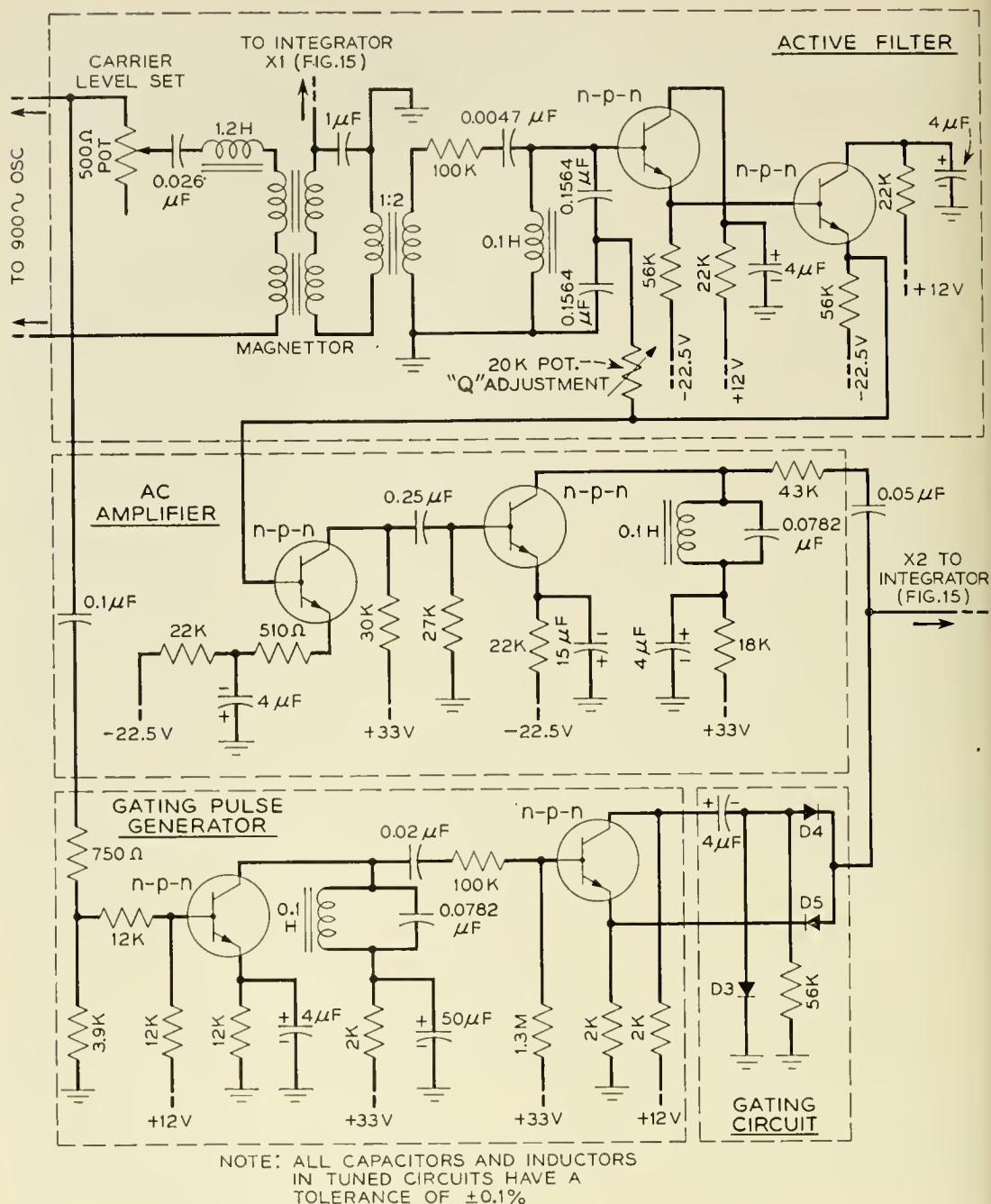


Fig. 17(a) — AZS circuit.

battery potentials are reversed and if an n-p-n junction transistor is used.

5.2. Factors Determining the Accuracy of the Voltage Comparator

Fig. 19 shows the ac equivalent circuit of the voltage comparator. In the equivalent circuit R_1 is the dynamic resistance of the diode D_1 , R_g is the source resistance of the input voltage, and R_2 is the impedance of

the load R_L as it appears at the primary of the transformer T_2 . R_1 is a function of the dc voltage across the diode D_1 . At a prescribed value of R_1 , the comparator circuit becomes unstable and switches. The relationship between this critical value of R_1 and the transistor and circuit parameters is obtained by evaluating the characteristic equation for the circuit and by determining the relationship which the coefficients of the equation must satisfy in order to have a root of the equation lie in the right hand half of the complex frequency plane. To a good approximation, the critical value of R_1 is given by the expression

$$R_1 + R_g + r_b = \frac{Ma_0}{R_2C_c + \frac{a_0}{\omega_a}} \quad (26)$$

where M is the mutual inductance of transformer T_1 and $R_2 = N'^2 R_L$. Since the transistor parameters which appear in expression (26) have only a small variation with temperature, the critical value of R_1 is independent of temperature (to a first approximation).

It will now be shown that the comparator can be designed for an accuracy of ± 5 millivolts throughout a temperature range of 20°C to 40°C . In order to establish this accuracy it will be assumed that the critical value of R_1 is equal to 30,000 ohms. This assumption is based on the

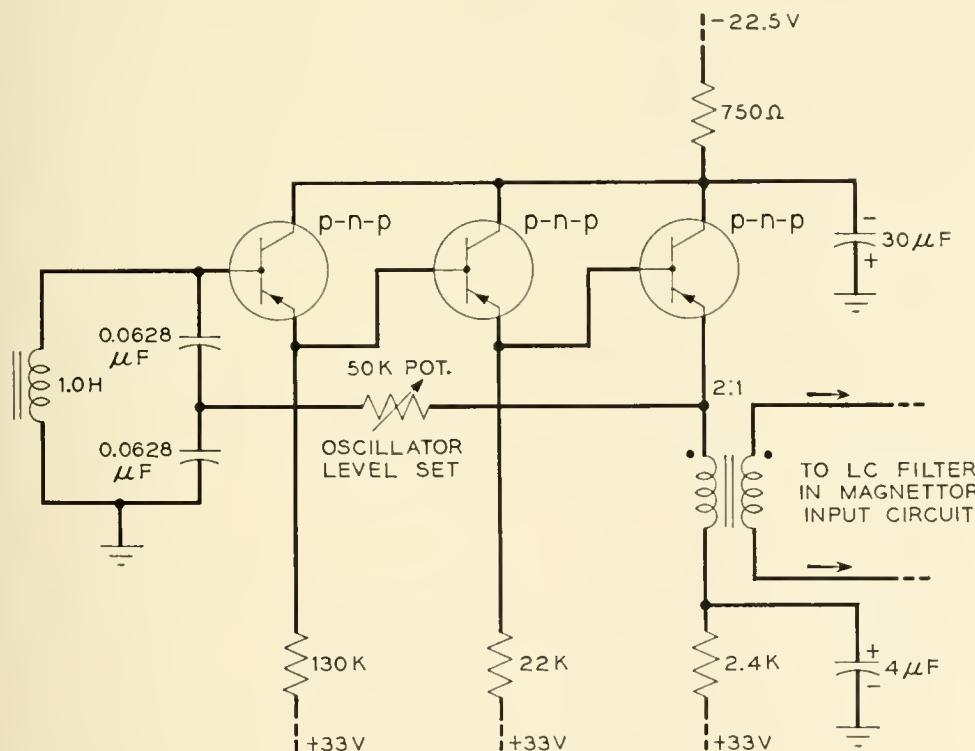


Fig. 17(b), 900-cycle carrier oscillator.

data displayed in Fig. 20 which gives the volt-ampere characteristics of a silicon diode measured at 20°C and 40°C. Throughout this temperature range, the diode voltage corresponding to the critical resistance of 30,000 ohms changes by about 30 millivolts. Fortunately, part of this voltage variation with temperature is compensated for by the variation in voltage V_{b-e} between the base and emitter of the junction transistor. From Fig. 18,

$$V_i = V_D - V_{b-e} + V_{ee} \quad (27)$$

For perfect compensation (V_i independent of temperature), V_{b-e} should have the same temperature variation as the diode voltage V_D . Experi-

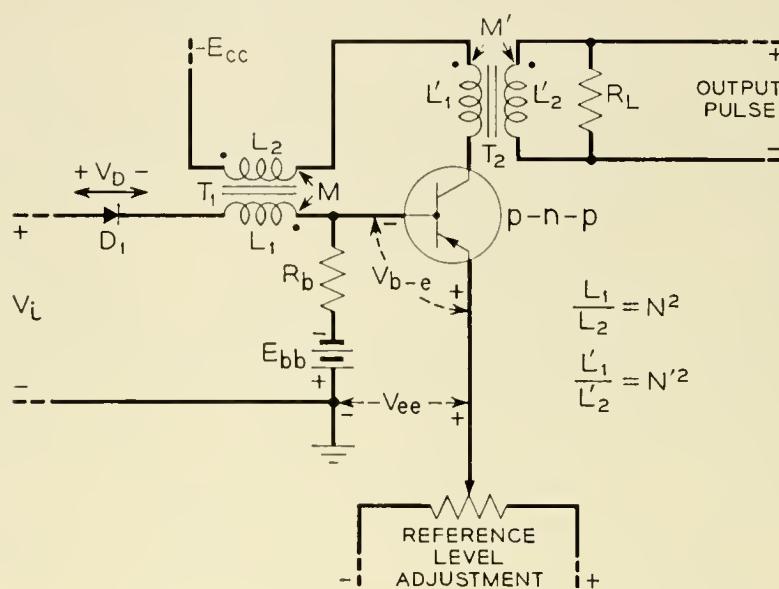


Fig. 18 — Simplified circuit diagram of voltage comparator.

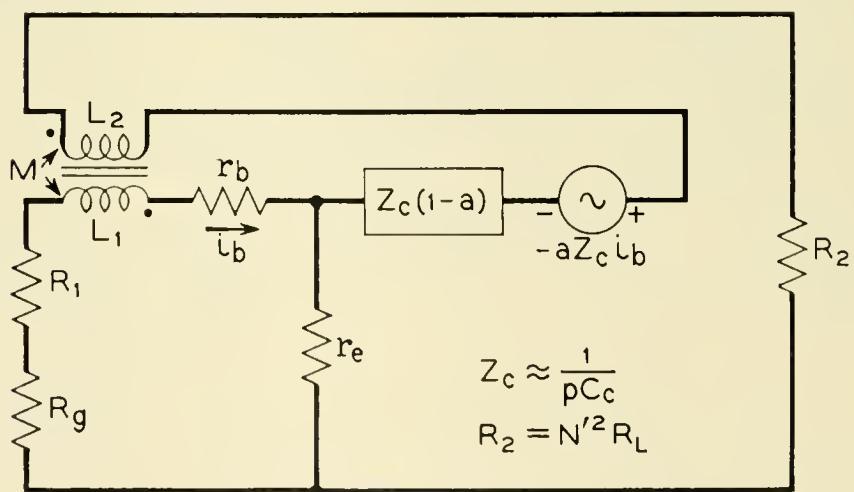


Fig. 19 — Equivalent circuit of voltage comparator.

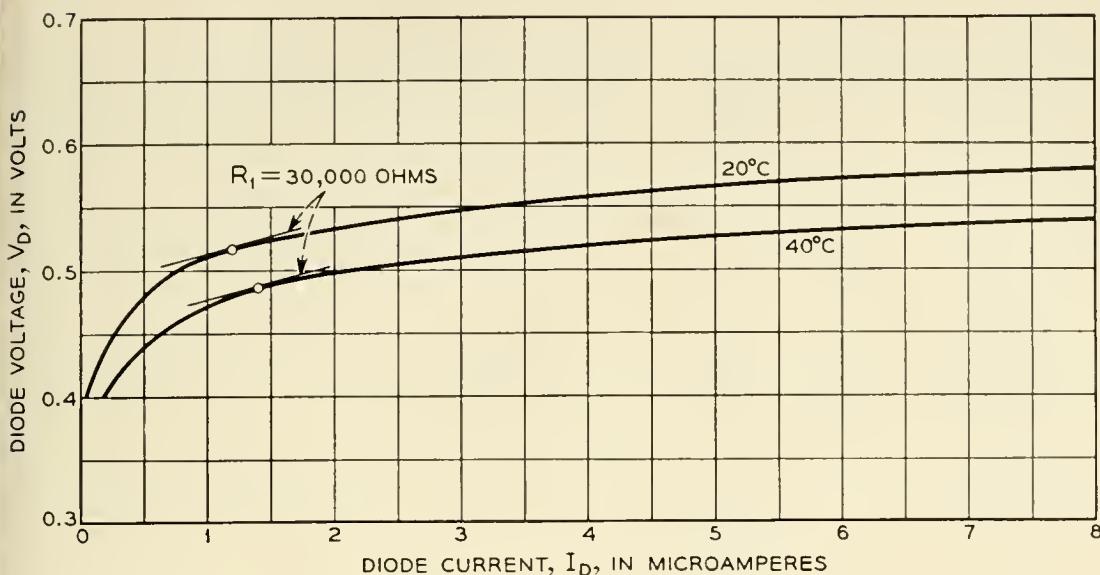


Fig. 20 — Volt-ampere-characteristic of a silicon junction diode.

mentally it is found that V_{b-e} for germanium junction transistors varies by about 20 millivolts throughout the temperature range of 20°C to 40°C . Consequently, the variation in V_i at which the circuit switches is ± 5 millivolts.

It is apparent from Fig. 20 that the accuracy of the comparator increases slightly for critical values of R_1 greater than 30,000 ohms, but decreases for smaller values. For example, the accuracy of the comparator is ± 10 millivolts for a critical value of R_1 equal to 5,000 ohms. In general, the critical value of R_1 should be chosen between 5,000 and 100,000 ohms.

5.3. A Practical Voltage Comparator

Fig. 21 shows the complete circuit diagram of a voltage comparator. The circuit is designed to generate a sharp output pulse* when the input voltage waveform passes through the reference level (set by V_{ee}) with a positive slope. The pulse is generated by the transistor switching from the "on" state to the "off" state. To a first approximation the amplitude of the output pulse is proportional to the transistor collector current during the "on" state. When the input voltage waveform passes through the reference level with a negative slope an undesirable negative pulse is generated. This pulse is eliminated by the point contact diode D_2 .

The voltage comparator is an unstable circuit and has the properties

* For the circuit values shown in Fig. 21, the output pulse has a peak amplitude of about 6 volts, a rise time of 0.5 microseconds, and a pulse width of about 2.0 microseconds.

of a free running blocking oscillator after the input voltage V_i passes through the reference level. After a period of time the transistor will return to the "on" state unless the voltage V_i is sufficiently large at this time to prevent switching. In order to minimize the required slope of the input waveform the time interval between the instant V_i passes through the reference level and the instant the transistor would naturally switch to the "on" state must be maximized. This time interval can be controlled by connecting a diode D_3 across the secondary winding of transformer T_1 . When the transistor turns off, the current which was flowing through the secondary of transformer T_1 (I_c) continues to flow through the diode D_3 so that L_2 and D_3 form an inductive discharge circuit. The point contact diode D_3 has a forward dynamic resistance of less than 10 ohms and a forward voltage drop of 0.3 volt. If the small forward resistance of the diode is neglected, the time required for the current in the circuit to fall to zero is

$$T = \frac{I_c L_2}{0.3} \quad (28)$$

During the inductive transient, 0.3 volt is induced into the primary of transformer T_1 (since $N = 1$) maintaining the transistor cutoff. The duration of the inductive transient can be made as long as desired by increasing L_2 . However, there is the practical limitation that increasing L_2 also increases the leakage inductance of transformer T_1 , and in turn,

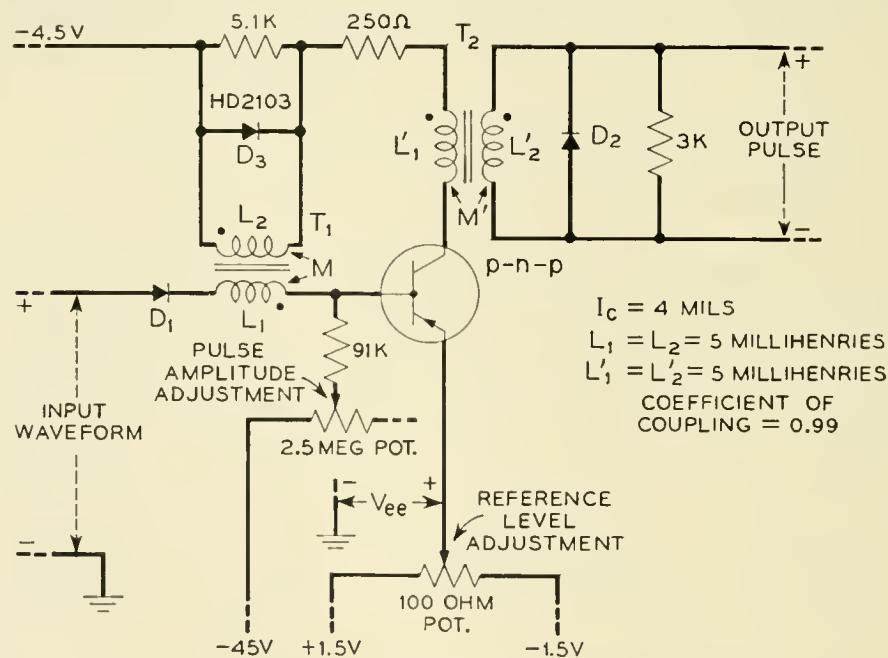


Fig. 21 — Voltage comparator.

increases the switching time. The circuit shown in Figure 21 does not misfire when used with voltage waveforms having slopes as small as 25 millivolts per microsecond, at the reference level.

6.0. A TRANSISTOR VOLTAGE ENCODER

6.1. Circuit Arrangement

The transistor circuits previously described can be assembled into a voltage encoder for translating analog voltages into equivalent time intervals. This encoder is especially useful for converting analog information (in the form of a dc potential) into the digital code for processing in a digital system. Fig. 22 shows a simplified block diagram of the encoder. The voltage ramp generated by the integrator is applied to amplitude selector number one and to one input of a summing amplifier. The amplitude selector is a dc amplifier which amplifies the voltage ramp in the vicinity of zero volts. Voltage comparator number one, which follows the amplitude selector, generates a sharp output pulse at the exact instant of time that the voltage ramp passes through zero volts.

The analog input voltage, which has a value between 0 and -15 volts,* is applied to the second input of the summing amplifier. The output voltage of the summing amplifier is zero whenever the ramp

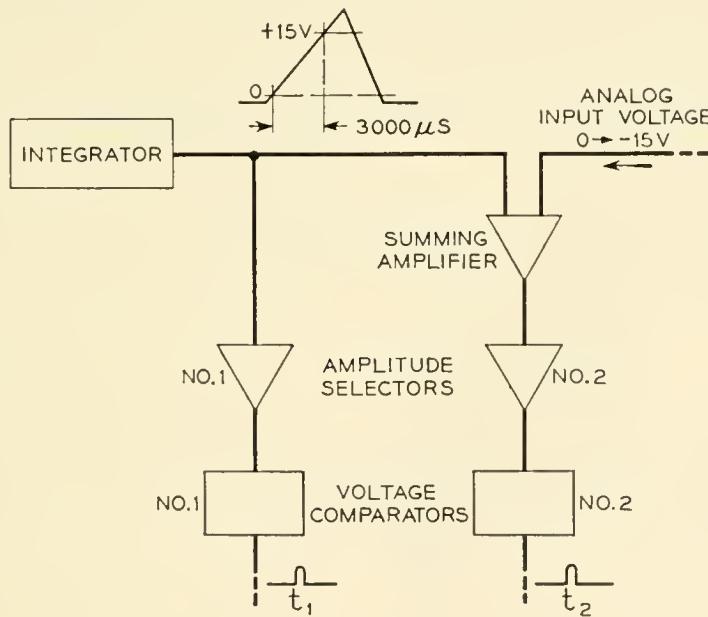


Fig. 22 — Simplified block diagram of voltage encoder.

* If the analog input voltage does not lie in this range, then the voltage gain of the summing amplifier must be set so that the analog voltage at the output of the summing amplifier lies in the voltage range between 0 and +15 volts.

voltage is equal to the negative of the input analog voltage. At this instant of time the second voltage comparator generates a sharp output pulse. The time interval between the two output pulses is proportional to the analog input voltage if the voltage ramp is linear and has a constant slope at all times.

6.2. The Amplitude Selector

The amplitude selector increases the slope of the input voltage waveform (in the vicinity of zero volts) sufficiently for proper operation of the voltage comparator. The amplitude selector consists of a limiter and a dc feedback amplifier as shown in Fig. 23. The two oppositely poled silicon diodes D_1 and D_2 , limit the input voltage of the dc amplifier to about ± 0.65 volts. The dc amplifier has a voltage gain of thirty, and so the maximum output voltage of the amplitude selector is limited to about ± 19.5 volts. The net voltage gain between the input and output of the amplitude selector is ten.

The principal requirement placed on the dc amplifier is that the input current and the output voltage be zero when the input voltage is zero. This is accomplished by placing a blocking condenser C_B between point T and the base of the first transistor stage, and by using an AZS circuit to maintain point T at zero volts. The dc and AZS amplifiers are identical in configuration to the amplifiers shown in Fig. 12. The dc amplifier is

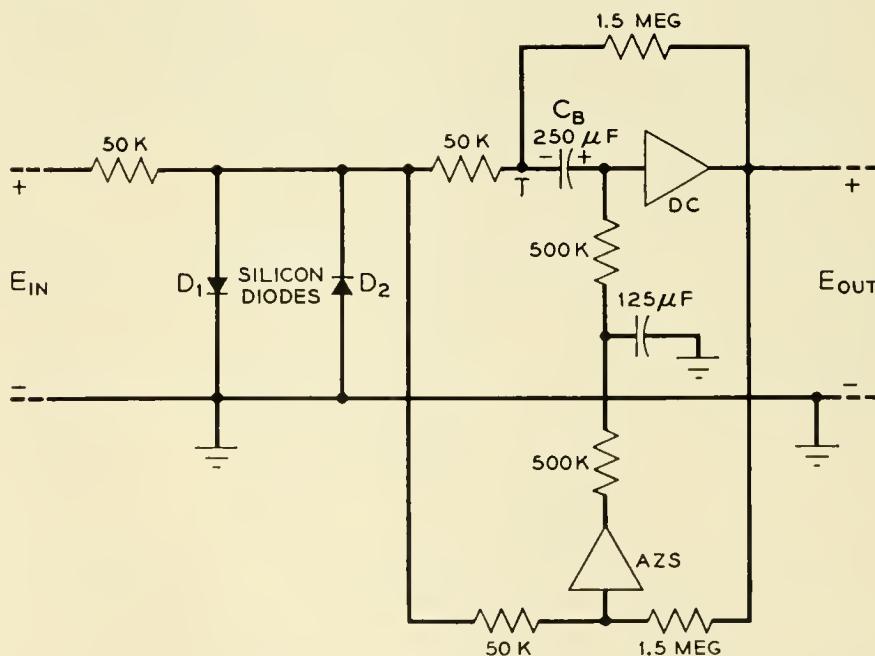


Fig. 23 — Block diagram of the amplitude selector.

designed to have about 15.6 db less feedback than that shown in Fig. 10 since this amount is adequate for the present purpose.

The bandwidth of the dc amplifier is only of secondary importance because the phase shifts introduced by the two amplitude selectors in the voltage encoder tend to compensate each other.

6.3. Experimental Results

The accuracy of the voltage encoder is determined by applying a precisely measured voltage to the input of the summing amplifier and by measuring the time interval between the two output pulses. The maximum error due to nonlinearities in the summing amplifier and the voltage ramp is less than ± 0.5 microseconds for a maximum encoding time of 3,000 microseconds. An additional error is introduced by the noise voltage generated in the first transistor stage of the summing amplifier. The RMS noise voltage at the output of the summing amplifier is less than 0.5 millivolts. This noise voltage produces an RMS jitter of 0.25 microseconds in the position of the second voltage comparator output pulse. The over-all accuracy of the voltage encoder is one part in 4,000 throughout a temperature range of 20°C to 40°C.

ACKNOWLEDGEMENTS

The author wishes to express his appreciation to T. R. Finch for the advice and encouragement received in the course of this work. D. W. Grant and W. B. Harris designed and constructed the magnetron used in the AZS circuit of the integrator.

APPENDIX I

RELATIONSHIP BETWEEN RETURN DIFFERENCE AND LOOP CURRENT TRANSMISSION

In order to place the stability analysis of the transistor feedback amplifier on a sound basis, it is desirable to use the concept of return difference.⁸ It will be shown that a measurable quantity, called the loop current transmission, can be related to the return difference of aZ_e with reference r_e .^{*,†} In Fig. 24, N represents the complete transistor network exclusive of the transistor under consideration. The feedback loop is broken at the input to the transistor by connecting all of the feedback paths to

* In this appendix it is assumed that the transistor under consideration is in the common emitter connection. The discussion can be readily extended to the other transistor connections.

† This fact was pointed out by F. H. Tendick, Jr.

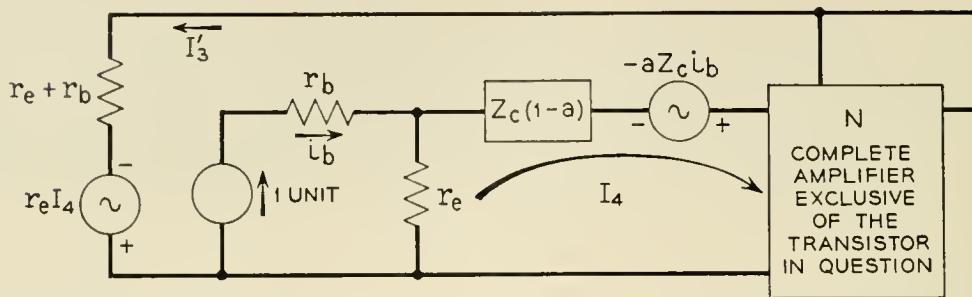


Fig. 24 — Measurement of loop current transmission.

ground through a resistance $(r_e + r_b)$ and a voltage $r_e I_4$. Using the nomenclature given in Reference 8, the input of the complete circuit is designated as the first mesh and the output of the complete circuit is designated as the second mesh. The input and output meshes of the transistor under consideration are designated 3 and 4, respectively. The loop current transmission is equal to I_3' , the total returned current when a unit input current is applied to the base of the transistor.

The return difference for reference r_e is equal to the algebraic difference* between the unit input current and the returned current I_3' . I_3' is evaluated by multiplying the open circuit voltage in mesh 4 (produced by the unit base current) by the backward transmission from mesh 4 to mesh 3 with zero forward transmission through the transistor under consideration. The open circuit voltage in mesh 4 is equal to $(r_e - aZ_c)$. The backward transmission is determined with the element aZ_c , in the fourth row, third column of the circuit determinant, set equal to r_e . Hence, the return difference is expressed as

$$F_{r'e} = 1 + (aZ_c - r_e) \frac{\Delta_{43}}{\Delta^{r'e}} \quad (A1)\dagger$$

$$F_{r'e} = \frac{\Delta^{r'e} + (aZ_c - r_e)\Delta_{43}}{\Delta^{r'e}} \quad (A2)$$

$$F_{r'e} = \frac{\Delta}{\Delta^{r'e}} = 1 + T_{r'e} \quad (A3)$$

The relative return ratio $T_{r'e}$ is equal to the negative of the loop current transmission and can be measured as shown in Fig. 24. The voltage $r_e I_4$ takes into account the fact that the junction transistor is not perfectly

* The positive direction for the returned current is chosen so that if the original circuit is restored, the returned current flows in the same direction as the input current.

† $\Delta^{r'e}$ is the network determinant with the element aZ_c in the fourth row, third column of the circuit determinant set equal to r_e .

unilateral. Fortunately, in many applications, this voltage can be neglected even at the gain and phase crossover frequencies.

In the case of single loop feedback amplifiers, $\Delta^{r'e}$ will not have any zeros in the right hand half of the complex frequency plane. A study of the stability of the amplifier can then be based on $F_{r'e}$ or $T_{r'e}$.

APPENDIX II

INTERSTAGE NETWORK SHAPING

This appendix presents the analysis of the circuit shown in Fig. 7(a). The input impedance of the common emitter connected junction transistor is given by the expression

$$Z_{\text{INPUT}} = r_b + r_e(1 - G_I) \quad (\text{A4})$$

where G_I is the current transmission of the common emitter stage, expression (13). The current transmission A of the complete circuit is equal to

$$A = \frac{I_2}{I_1} = \frac{Z_3}{Z_3 + Z_{\text{INPUT}}} \cdot G_I \quad (\text{A5})$$

where $Z_3 = R_3 + pL_3 + (1/pC_3)$. Combining (13), (A4), and (A5) yields

$$A = \frac{-\frac{a_0}{1 - a_0 + \delta} \left[\left(1 + \frac{p}{\omega_3}\right)^2 + p \left(R_3C_3 - \frac{2}{\omega_3}\right) \right]}{\left(1 + \frac{p}{\omega_5}\right) \left\{ 1 + p \frac{C_3\omega_5}{\omega_1} (r_b + r_e + R_3 + \omega_1 L_3) + p^2 \left[\frac{\omega_5}{\omega_1 \omega_3^2} + \frac{C_3\omega_5(r_b + r_e + R_3)}{\omega_a \omega_c (1 - a_0 + \delta)} \right] + \frac{p^3 \omega_5}{\omega_3^2 \omega_a \omega_c (1 - a_0 + \delta)} \right\}} \quad (\text{A6})$$

where

$$\delta = \frac{R_L + r_e}{r_c}$$

$$\omega_1 = \frac{(1 - a_0 + \delta)}{\frac{1 + \delta}{\omega_a} + \frac{1}{\omega_c}}$$

$$\omega_c = \frac{1}{(R_L + r_e)C_c}$$

$$\omega_3 = \frac{1}{\sqrt{L_3 C_3}}$$

$$\omega_5 = \frac{1}{\left[r_b + \frac{r_e}{(1 - a_0 + \delta)} \right] C_3}$$

Expression (A6) is valid if $1/\omega_5 \gg 1/\omega_1 + R_3 C_3$. The denominator of the expression indicates a falling 6 db per octave asymptote with a corner frequency at ω_5 . The second factor in the denominator can be approximated by a falling 6 db per octave asymptote with a corner frequency at

$$\frac{\omega_1 \left[r_b + \frac{r_e}{(1 - a_0 + \delta)} \right]}{r_b + r_e + R_3 + \omega_1 L_3}$$

plus additional phase and amplitude contributions at higher frequencies due to the p^2 and p^3 terms. If

$$\frac{1}{\omega_3 C_3 R_3} = \frac{1}{2}$$

then the circuit has a rising 12 db per octave asymptote with a corner frequency at ω_3 . Fig. 7(b) shows the amplitude and phase of the current transmission.

REFERENCES

1. Felker, J. H., Regenerative Amplifier for Digital Computer Applications, Proc. I.R.E., pp. 1584-1596, Nov., 1952.
2. Korn, G. A., and Korn, T. M., Electronic Analog Computers, McGraw-Hill Book Company, pp. 9-19.
3. Wallace, R. L. and Pietenpol, W. J., Some Circuit Properties and Applications of n-p-n Transistors, B. S.T.J., **30**, pp. 530-563, July, 1951.
4. Shockley, W., Sparks, M. and Teal, G. K., The p-n Junction Transistor, Physical Review, **83**, pp. 151-162, July, 1951.
5. Pritchard, R. L., Frequency Variation of Current-Amplification for Junction Transistors, Proc. I.R.E., pp. 1476-1481, Nov., 1952.
6. Early, J. M., Design Theory of Junction Transistors, B.S.T.J., **32**, pp. 1271-1312, Nov., 1953.
7. Sziklai, G. C., Symmetrical Properties of Transistors and Their Applications, Proc. I.R.E., pp. 717-724, June, 1953.
8. Bode, H. W., Network Analysis and Feedback Amplifier Design, Van Nostrand Co., Inc., Chapter IV.
9. Bode, H. W., Op. Cit., pp. 66-69.
10. Bode, H. W., Op. Cit., pp. 162-164.
11. Bargellini, P. M. and Herscher, M. B., Investigation of Noise in Audio Frequency Amplifiers Using Junction Transistors, Proc. I.R.E., pp. 217-226, Feb., 1955.
12. Bode, H. W., Op. Cit., pp. 464-468, and pp. 471-473.
13. Keonian, E., Temperature Compensated DC Transistor Amplifier, Proc. I.R.E., pp. 661-671, April, 1954.
14. Kretzmer, E. R., An Amplitude Stabilized Transistor Oscillator, Proc. I.R.E., pp. 391-401, Feb., 1954.
15. Goldberg, E. A., Stabilization of Wide-Band Direct-Current Amplifiers for Zero and Gain, R.C.A. Review, June, 1950.
16. Ebers, J. J. and Moll, J. L., Large Signal Behavior of Junction Transistors, Proc. I.R.E., pp. 1761-1772, Dec., 1954.
17. Manley, J. M., Some General Properties of Magnetic Amplifiers, Proc. I.R.E., March, 1951.
18. M.I.T., Waveforms, Volume 19 of the Radiation Laboratories Series. McGraw-Hill Book Company, pp. 342-344.

Electrolytic Shaping of Germanium and Silicon

By A. UHLIR, JR.

(Manuscript received November 9, 1955)

Properties of electrolyte-semiconductor barriers are described, with emphasis on germanium. The use of these barriers in localizing electrolytic etching is discussed. Other localization techniques are mentioned. Electrolytes for etching germanium and silicon are given.

INTRODUCTION

Mechanical shaping techniques, such as abrasive cutting, leave the surface of a semiconductor in a damaged condition which adversely affects the electrical properties of p-n junctions in or near the damaged material. Such damaged material may be removed by electrolytic etching. Alternatively, all of the shaping may be done electrolytically, so that no damaged material is produced. Electrolytic shaping is particularly well suited to making devices with small dimensions.

A discussion of electrolytic etching can conveniently be divided into two topics — the choice of electrolyte and the method of localizing the etching action to produce a desired shape. It is usually possible to find an electrolyte in which the rate at which material is removed is accurately proportional to the current. For semiconductors, just as for metals, the choice of electrolyte is a specific problem for each material; satisfactory electrolytes for germanium and silicon will be described.

The principles of localization are the same, whatever the electrolyte used. Electrolytic etching takes place where current flows from the semiconductor to the electrolyte. Current flow may be concentrated at certain areas of the semiconductor-electrolyte interface by controlling the flow of current in the electrolyte or in the semiconductor.

LOCALIZATION IN ELECTROLYTE

Localization techniques involving the electrolytic current are applicable to both metals and semiconductors. In some of these techniques,

the localization is so effective that the barrier effects found with n-type semiconductors can be ignored; if not, the barrier can be overcome by light or heat, as will be described below.

If part of the work is coated with an insulating varnish, electrolytic etching will take place only on the uncoated surfaces. This technique, often called "masking," has the limitation that the etching undercuts the masking if any considerable amount of material is removed. The same limitation applies to photoengraving, in which the insulating coating is formed by the action of light.

The cathode of the electrolytic cell may be limited in size and placed close to the work (which is the anode). Then the etching rate will be greatest at parts of the work that are nearest the cathode. Various shapes can be produced by moving the cathode with respect to the work, or by using a shaped cathode. For example, a cathode in the form of a wire has been used to slice germanium.¹

Instead of a true metallic cathode, a "virtual cathode" may be used to localize electrolysis.² In this technique, the anode and true cathode are separated from each other by a nonconducting partition, except for a small opening in the partition. As far as localization of current to the anode is concerned, the small opening acts like a cathode of equal size and so is called a virtual cathode. The nonconducting partition may include a glass tube drawn down to a tip as small as one micron diameter but nevertheless open to the flow of electrolytic current. With such a tip as a virtual cathode, micromachining can be conducted on a scale comparable to the wavelength of visible light. A general advantage of the virtual cathode technique is that the cathode reaction (usually hydrogen evolution) does not interfere with the localizing action nor with observation of the process.

In the jet-etching technique, a jet of electrolyte impinges on the work.^{3, 4} The free streamlines that bound the flowing electrolyte are governed primarily by momentum and energy considerations. In turn, the shape of the electrolyte stream determines the localization of etching. A stream of electrolyte guided by wires has been used to etch semiconductor devices.⁵ Surface tension has an important influence on the free streamlines in this case.

PROPERTIES OF ELECTROLYTE-SEMICONDUCTOR BARRIERS

The most distinctive feature of electrolytic etching of semiconductors is the occurrence of rectifying barriers. Barrier effects for germanium will be described; those for silicon are qualitatively similar.

The voltage-current curves for anodic n-type and p-type germanium

in 10 per cent KOH are shown in Fig. 1. The concentration of KOH is not critical and other electrolytes give similar results. The voltage drop for the p-type specimen is small. For anodic n-type germanium, however, the barrier is in the reverse or blocking direction as evidenced by a large voltage drop. The fact that n-type germanium differs from p-type germanium only by very small amounts of impurities suggests that the barrier is a semiconductor phenomenon and not an electrochemical one. This is confirmed by the light sensitivity of the n-type voltage-current characteristic. Fig. 2 is a schematic diagram of the arrangement for obtaining voltage-current curves. A mercury-mercuric oxide-10 per cent KOH reference electrode was used at first, but a gold wire was found equally satisfactory. At zero current, a voltage V_0 exists between the germanium and the reference electrode; this voltage is not included in Fig. 1.

The saturation current I_s , measured for the n-type barrier at a moderate reverse voltage (see Fig. 1), is plotted as a function of temperature in Fig. 3. The saturation current increases about 9 per cent per degree, just as for a germanium p-n junction, which indicates that the

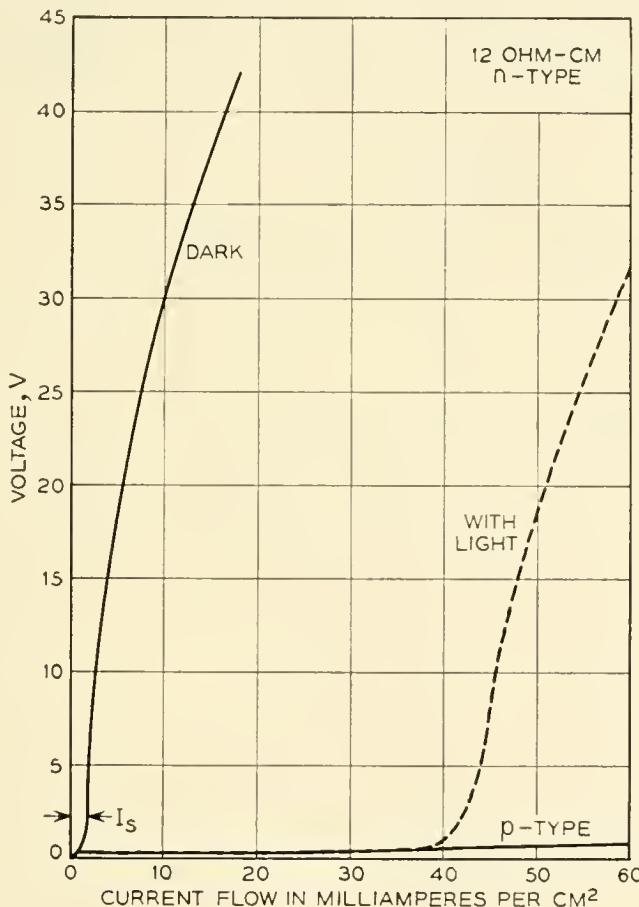


Fig. 1 — Anodic voltage-current characteristics of germanium.

current is proportional to the equilibrium density of minority carriers (holes). The same conclusion may be drawn from Fig. 4, which shows that the saturation current is higher, the higher the resistivity of the n-type germanium. But the breakdown voltages are variable and usually much lower than one would expect for planar p-n junctons made, for example, by alloying indium into the same n-type germanium.

Breakdown in bulk junctions is attributed to an avalanche multiplication of carriers in high fields.⁶ The same mechanism may be responsible for breakdown of the germanium-electrolyte barrier; low and variable breakdown voltages may be caused by the pits described below.

The electrolyte-germanium barrier exhibits a kind of current multiplication that differs from high-field multiplication in two respects: it occurs at much lower reverse voltages and does not vary much with voltage.⁷ This effect can be demonstrated very simply by comparison with a metal-germanium barrier, on the assumption that the latter has a current multiplication factor of unity. This assumption is supported by experiments which indicate that current flows almost entirely by hole flow, for good metal-germanium barriers.⁸

The experimental arrangement is indicated in Fig. 5(a) and (b). The voltage-current curves for an electrolyte barrier and a plated barrier on the same slice of germanium are shown in Fig. 5(c).* The curves for the

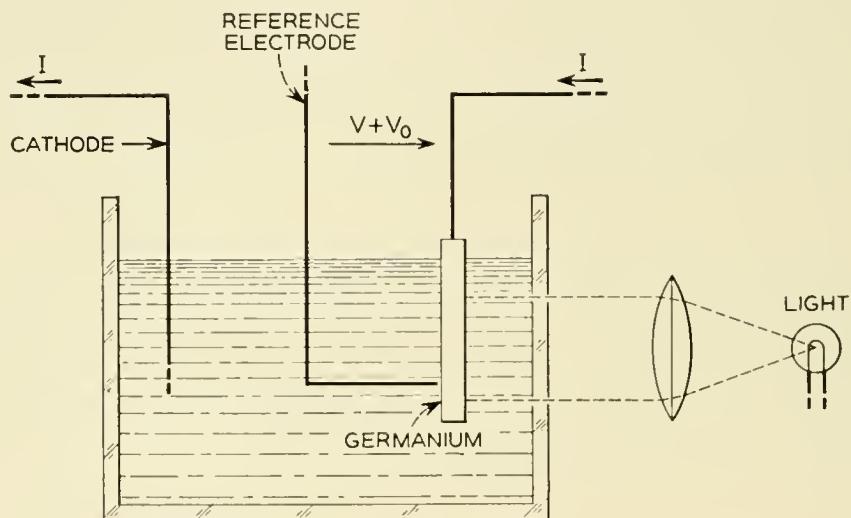


Fig. 2—Arrangement for obtaining voltage-current characteristics.

* In Fig. 5 the dark current for the plated barrier is much larger than can be explained on the basis of hole current; it is even higher than the dark current for the electrolyte barrier, which should be at least 1.4 times the hole current. This excess dark current is believed to be leakage at the edges of the plated area and probably does not affect the intrinsic current multiplication of the plated barrier as a whole.

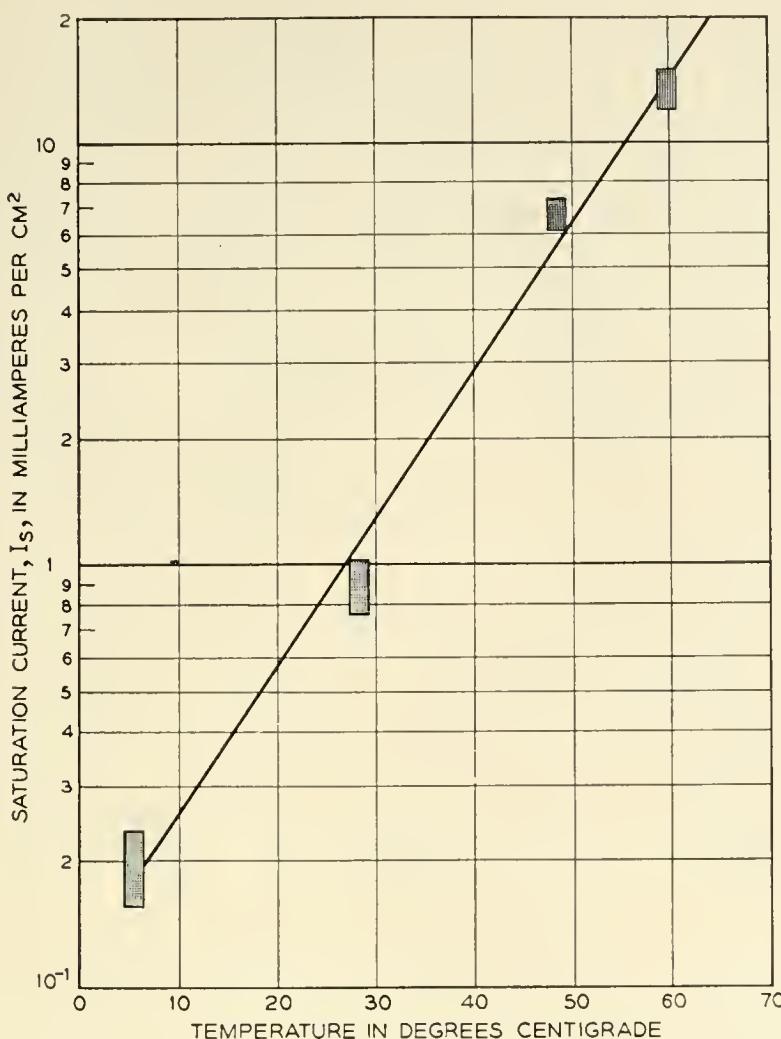


Fig. 3 — Temperature variation of the saturation current of a barrier between 5.5 ohm-cm n-type germanium and 10 per cent KOH solution.

illuminated condition were obtained by shining light on a dry face of a slice while the barriers were on the other face. The difference between the light and dark currents is larger for the electrolyte-germanium barrier than for the metal-germanium barrier, by a factor of about 1.4.

The transport of holes through the slice is probably not very different for the two barriers. Therefore, a current multiplication of 1.4 is indicated for the electrolyte barrier. About the same value was found for temperatures from 15°C to 60°C, KOH concentrations from 0.01 per cent to 10 per cent, n-type resistivities of 0.2 ohm-cm to 6 ohm-cm, light currents of 0.1 to 1.0 ma/cm², and for 0.1N indium sulfate.

Evidently the flow of holes to the electrolyte barrier is accompanied by a proportionate return flow of electrons, which constitutes an additional electric current. Possible mechanisms for the creation of the electrons will be discussed in a forthcoming article.⁹

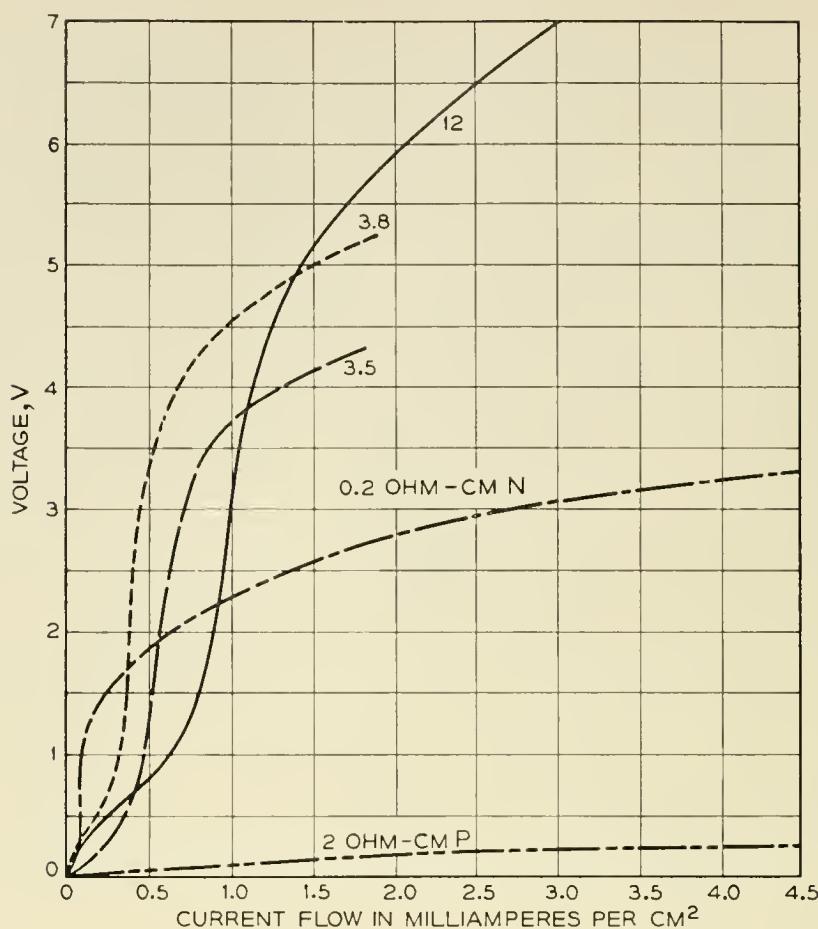


Fig. 4. — Anodic voltage-current curves for various resistivities of germanium.

SCRATCHES AND PITTING

The voltage-current curve of an electrolyte-germanium barrier is very sensitive to scratches. The curves given in the illustrations were obtained on material previously etched smooth in CP-4, a chemical etch.*¹⁰

If, instead, one starts with a lapped piece of n-type germanium, the electrolyte-germanium barrier is essentially "ohmic;" that is, the voltage drop is small and proportional to the current. A considerable reverse voltage can be attained if lapped n-type germanium is electrolytically etched long enough to remove most of the damaged germanium. However, a pitted surface results and the breakdown voltage achieved is not as high as for a smooth chemically-etched surface.

The depth of damage introduced by typical abrasive sawing and lapping was investigated by noting the voltage-current curve of the

* Five parts HNO_3 , 3 parts 48 per cent HF, 3 parts glacial acetic acid, $\frac{1}{10}$ part Br_2 .

electrolyte-germanium barrier after various amounts of material had been removed by *chemical* etching. After 20 to 50 microns had been removed, further chemical etching produced no change in the barrier characteristic. This amount of material had to be removed even if the lapping was followed by polishing to a mirror finish. The voltage-current curve of the electrolyte-germanium barrier will reveal localized damage. On the other hand, the photomagnetoelectric (PME) measurement of

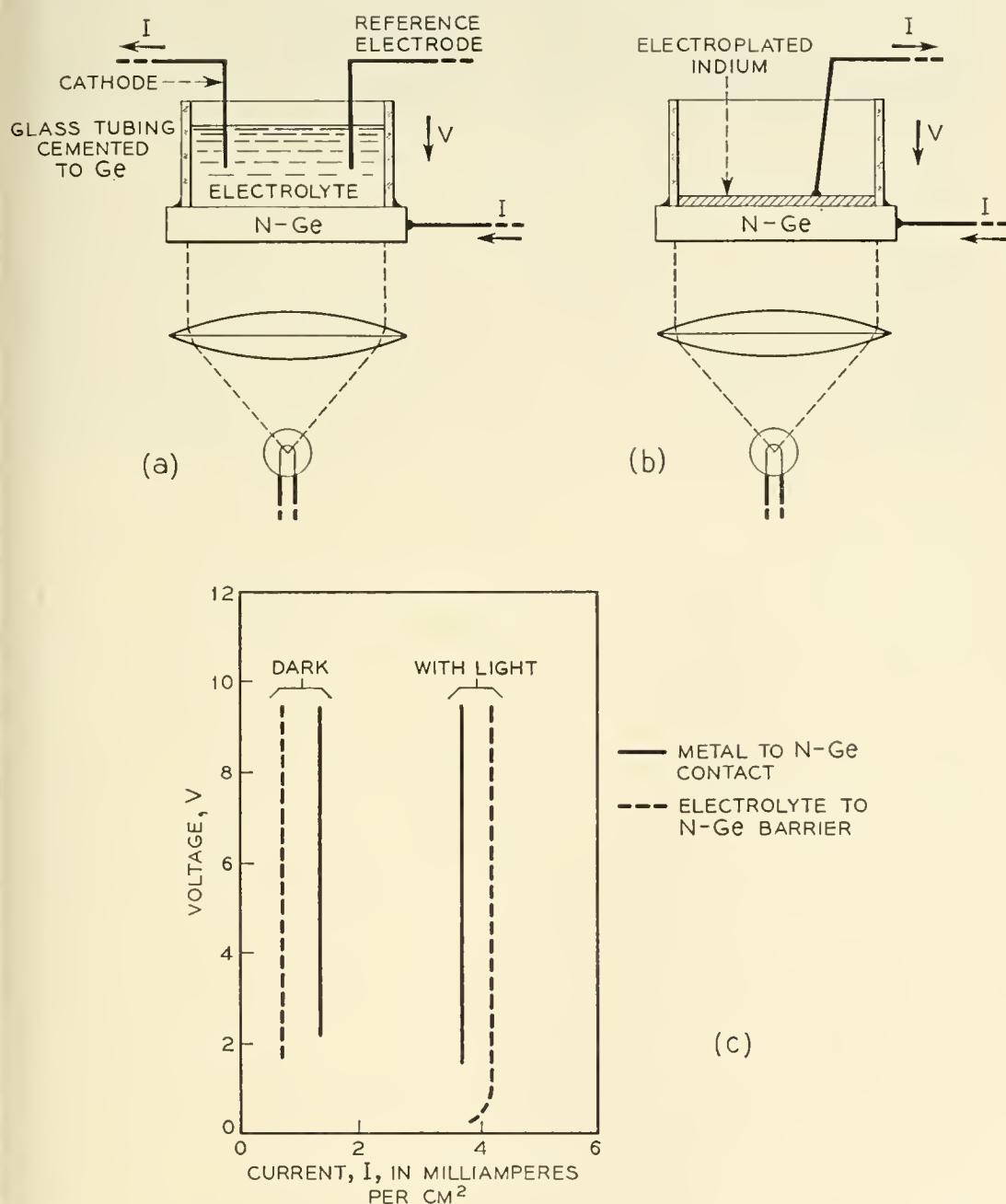


Fig. 5 — Determination of the current multiplication of the barrier between 6 ohm-cm n-type germanium and an electrolyte.



Fig. 6 -- Electrolytic etch pits on two sides of 0.02-inch slice of n-type germanium. Half of the slice was in contact with the electrolyte.

surface recombination velocity gives an evaluation of the average condition of the surface.¹¹ A variation of the PME method has been used to study the depth of abrasion damage; the damage revealed by this method extends only to a depth comparable to the abrasive size.¹²

A scratch is *sufficient* to start a pit that increases in size without limit if anodic etching is prolonged. However, a scratch is not *necessary*. Pits are formed even when one starts with a smooth surface produced by chemical etching. A drop in the breakdown voltage of the barrier is noticed when one or more pits form. The breakdown voltage can be restored by masking the pits with polystyrene cement.

Evidence that the spontaneous pits are caused by some features of the crystal, itself, was obtained from an experiment on single-crystal n-type germanium made by an early version of the zone-leveling process. A slice of this material was electrolytically etched on both sides, after preliminary chemical etching. Photographs of the two sides of the slice are shown in Fig. 6. Only half of the slice was immersed in the electrolyte. The electrolytic etch pits are concentrated in certain regions of the slice — the same general regions on both sides of the slice. It is interesting that radioautographs and resistivity measurements indicate high donor concentrations in these regions. Improvements, including more intensive stirring, were made in the zone-leveling process, and the electrolytic etch pit distribution and the donor radioautographs have been much more uniform for subsequent material.

Several pits on a (100) face are shown in Fig. 7. The pits grow most rapidly in $\langle 100 \rangle$ directions and give the spiked effect seen in the illustration. After prolonged etching, the spikes and their branches form a complex network of caverns beneath the surface of the germanium.

High-field carrier generation may be responsible for pitting. A locally

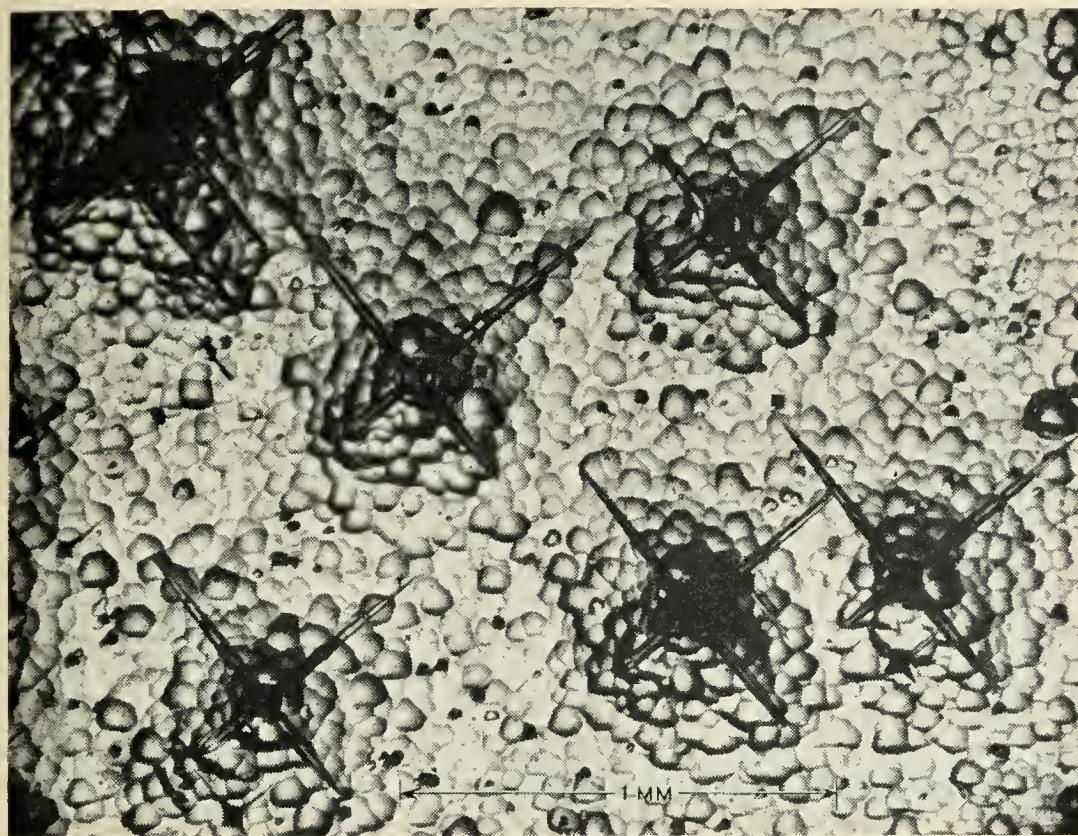


Fig. 7 — Electrolytic etch pits on n-type germanium.

high donor concentration would favor breakdown, as would any concavity of the germanium surface (which would cause a higher field for a given voltage). Very high fields must occur at the points of spikes such as those shown in Fig. 7. The continued growth of the spikes is thus favored by their geometry.

Microscopic etch pits arising from *chemical* etching have been correlated with the edge dislocations of small-angle grain boundaries.¹³ A specimen of n-type germanium with chemical etch pits was photomicrographed and then etched electrolytically. The etch pits produced electrolytically could not be correlated with the chemical etch pits, most of which were still visible and essentially unchanged in appearance. Also, no correlation could be found between either kind of etch pit and the locations at which copper crystallites formed upon immersion in a copper sulfate solution. Microscopic electrolytic etch pits at dislocations in p-type germanium have been reported in a recent paper that also mentions the deep pits produced on n-type germanium.¹⁴

Electrolytic etch pits are observed on n-type and high-resistivity silicon. These etch pits are more nearly round than those produced in germanium.

In spite of the pitting phenomenon, electrolytic etching is success-

fully used in the fabrication of devices involving n-type semiconductors. Pitting can be reduced relative to "normal" uniform etching by any agency that increases the concentration of holes in the semiconductor. Thus, elevated temperatures, flooding with light, and injection of holes by an emitter all favor smooth etching.

SHAPING BY MEANS OF INJECTED CARRIERS

Hole-electron pairs are produced when light is absorbed by semiconductors. Light of short wavelength is absorbed in a short distance, while long wavelength light causes generation at considerable depths. The holes created by the light move by diffusion and drift and increase the current flow through an anodic electrolyte-germanium barrier at whatever point they happen to encounter the barrier. In general, more holes will diffuse to a barrier, the nearer the barrier is to the point at which the holes are created. For n-type semiconductors, the current due to the light can be orders of magnitude greater than the dark current, so that the shape resulting from etching is almost entirely determined by the light. As shown in Fig. 3, the dark current can be made very small by lowering the temperature.

An example of the shaping that can be done with light is shown in Fig. 8. A spot of light impinges on *one* side of a wafer of n-type germanium or silicon. The semiconductor is made anodic with respect to an etching electrolyte. Accurately concentric dimples are produced on *both* sides of the wafer. Two mechanisms operate to transmit the effect to the opposite side. One is that some of the light may penetrate deeply before generating a hole-electron pair. The other is that a fraction of the carriers generated near the first surface will diffuse to the opposite side. By varying the spectral content of the light and the depth within the

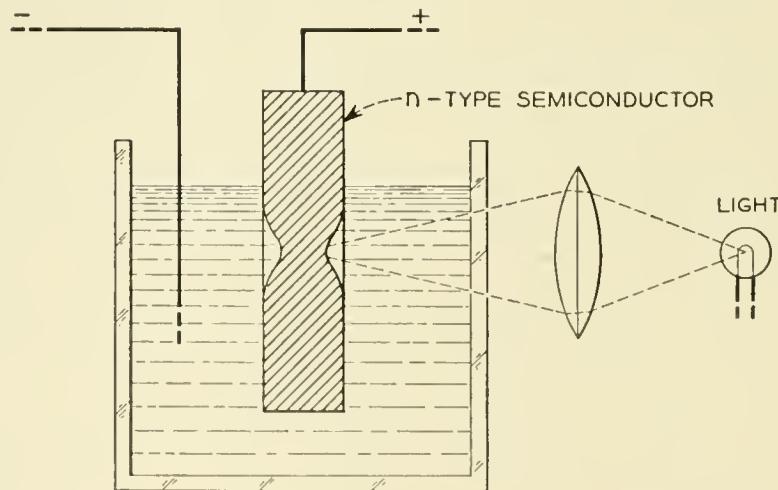


Fig. 8 — Double dimpling with light.

wafer at which the light is focused, one can produce dimples with a variety of shapes and relative sizes.

It is obvious that the double-dimpled wafer of Fig. 8 is desirable for the production of p-n-p alloy transistors. For such use, one of the most important dimensions is the thickness remaining between the bottoms of the two dimples. As has been mentioned in connection with the jet-etching process, a convenient way of monitoring this thickness to determine the endpoint of etching is to note the transmission of light of suitable wavelength.¹⁵ There is, however, a control method that is itself automatic. It is based on the fact that at a reverse-biased p-n junction or electrolyte-semiconductor barrier there is a space-charge region that is practically free of carriers.⁴ When the specimen thickness is reduced so that space-charge regions extend clear through it, current ceases to flow and etching stops in the thin regions, as long as thermally or optically generated carriers can be neglected. However, more pitting is to be expected in this method than when etching is conducted in the presence of an excess of injected carriers.

A p-n junction is a means of injecting holes into n-type semiconductors and is the basis of another method of dimpling, shown in Fig. 9. The p-n junction can be made by an alloying process such as bonding an acceptor-doped gold wire to germanium. The ohmic contact can be made by bonding a donor-doped gold wire and permits the injection of a greater excess of holes than would be possible if the current through the p-n junction were exactly equal to the etching current. Dimpling without the ohmic contact has been reported.¹⁴

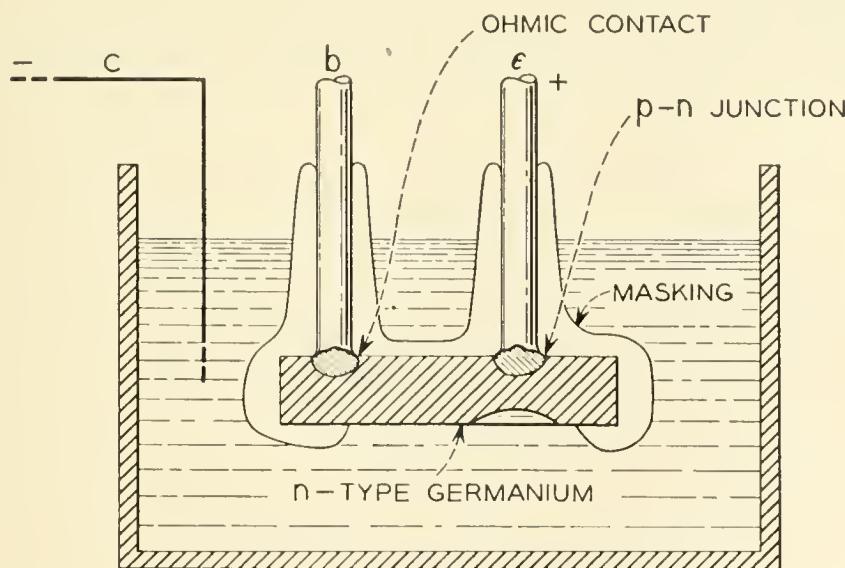


Fig. 9 -- Dimpling with carriers injected by a p-n junction.

CONTROL BY OHMIC CONDUCTION

The carrier-injection shaping techniques work very well for n-type material. It is also possible to inject a significant number of holes into rather high resistivity p-type material. But what can be done about p-type material in general, short of developing cathodic etches?

The ohmic resistivity of p-type material can be used as shown in Fig. 10. More etching current flows through surfaces near the small contact than through more remote surfaces. A substantial dimpling effect is observed when the semiconductor resistivity is equal to the electrolyte resistivity, but improved dimpling is obtained on higher resistivity semiconductor. This result is just what one might expect. But the mathematical solution for ohmic flow from a point source some distance from a planar boundary between semi-infinite materials of different conductivities shows that the current density distribution does not depend on the conductivities. An important factor omitted in the mathematical solution is the small but significant barrier voltage, consisting largely of electrochemical polarization in the electrolyte. The barrier voltage is approximately proportional to the logarithm of the current density while the ohmic voltage drops are proportional to current density. Thus, high current favors localization.

ELECTROLYTES FOR ETCHING GERMANIUM AND SILICON

The electrolyte usually has two functions in the electrolytic etching of an oxidizable substance. First, it must conduct the current necessary for the oxidation. Second, it must somehow effect removal of the oxidation product from the surface of the material being etched.

The usefulness of an electrolytic etch depends upon one or both of

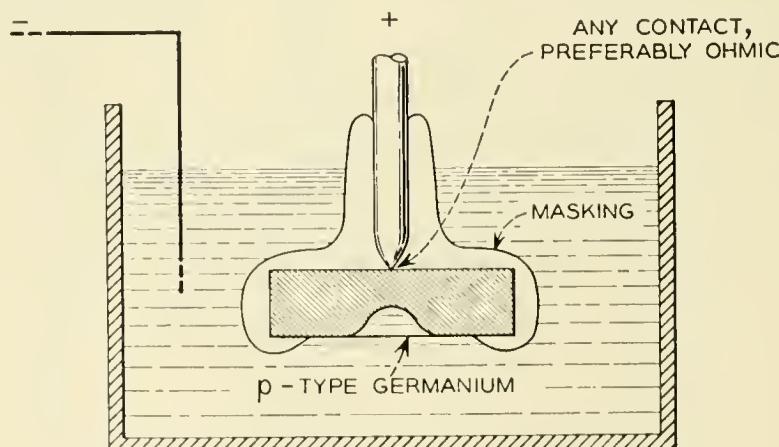


Fig. 10 — Dimpling by ohmic conduction.

the following situations — the electrolytic process accomplishes a reaction that cannot be achieved as conveniently in any other way or it permits greater control to be exercised over the reaction. Accordingly, chemical attack by the chosen electrolyte must be slight relative to the electrochemical etching.

A smooth surface is probably desirable in the neighborhood of a p-n junction, to avoid field concentrations and lowering of breakdown voltage. Therefore, a tentative requirement for an electrolyte is the production of a smooth, shiny surface on the p-type semiconductor. Such an electrolyte will give a shiny but possibly pitted surface on n-type specimens of the same semiconductor.

The effective valence of a material being electrolytically etched is defined as the number of electrons that traverse the circuit divided by the number of atoms of material removed. (The amount of material removed was determined by weighing in the experiments to be described.) If the effective valence turns out to be less than the valence one might predict from the chemistry of stable compounds, the etching is sometimes said to be "more than 100 per cent efficient." Since the anode reactions in electrolytic etching may involve unstable intermediate compounds and competing reactions, one need not be surprised at low or fractional effective valences.

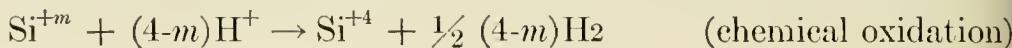
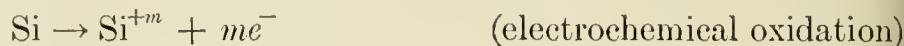
Germanium can be etched in many aqueous electrolytes. A valence of almost exactly 4 is found.¹⁶ That is, 4 electrons flow through the circuit for each atom of germanium removed. For accurate valence measurements, it is advisable to exclude oxygen by using a nitrogen atmosphere. Potassium hydroxide, indium sulfate, and sodium chloride solutions are among those that have been used. Sulfuric acid solutions are prone to yield an orange-red deposit which may be a suboxide of germanium.¹⁶ Similar orange deposits are infrequently encountered with potassium hydroxide.

Hydrochloric acid solutions are satisfactory electrolytes. The reaction product is removed in an unusual manner when the electrolyte is about 2N hydrochloric acid. Small droplets of a clear liquid fall from the etched regions. These droplets may be germanium tetrachloride, which is denser than the electrolyte. They turn brown after a few seconds, perhaps because of hydrolysis of the tetrachloride.

Etching of germanium in sixteen different aqueous electroplating electrolytes has been mentioned.⁸ Germanium can also be etched in the partly organic electrolytes described below for silicon.

One would expect that silicon could be etched by making it the anode in a cell with an aqueous hydrofluoric acid electrolyte. The seemingly

likely oxidation product, silicon dioxide, should react with the hydrofluoric acid to give silicon tetrafluoride, which could escape as a gas. In fact, a gas is formed at the anode and the silicon loses weight. But the gas is hydrogen and an effective valence of 2.0 ± 0.2 (individual determinations ranged from 1.3 to 2.7) was found instead of the value 4 that might have been expected. The quantity of hydrogen evolved is consistent with the formal reaction



where m is about two. The experiments were done in 24 per cent to 48 per cent aqueous solutions of HF at current densities up to 0.5 amp/cm^2 .

The suggestion that the electrochemical oxidation precedes the chemical oxidation is supported by the appearance and behavior of the etched surfaces. Instead of being shiny, the surfaces have a matte black, brown, or red deposit.

At $40\times$ magnification, the deposit appears to consist of flakes of a resinous material, tentatively supposed to be a silicon suboxide. A remarkable reaction can be demonstrated if the silicon is rinsed briefly in water and alcohol after the electrolytic etch, dried, and stored in air for as long as a year. Upon reimmersing this silicon in water, one can observe the liberation of gas bubbles at its surface. This gas is presumed to be hydrogen. To initiate the reaction it is sometimes necessary to dip the specimen first in alcohol, as water may otherwise not wet it. The specimens also liberate hydrogen from alcohol and even from toluene.

Thus, chemical oxidation can follow electrolytic oxidation. But chemical oxidation does not proceed at a significant rate before the current is turned on.

Smooth, shiny electrolytic etching of p-type silicon has been obtained with mixtures of hydrofluoric acid and organic hydroxyl compounds, such as alcohols, glycols, and glycerine. These mixtures may be anhydrous or may contain as much as 90 per cent water. The organic additives tend to minimize the chemical oxidation of the silicon. They also permit etching at temperatures below the freezing point of aqueous solutions. They lower the conductivity of the electrolyte.

For a given electrolyte composition, there is a threshold current density, usually between 0.01 and 0.1 amps/ cm^2 , for smooth etching. Lower current densities give black or red surfaces with the same hydrogen-liberating capabilities as those obtained in aqueous hydrofluoric acid.

In general, smooth etching of silicon seems to result when the effective valence is nearly 4 and there is little anodic evolution of gas. The electrical properties of the smooth surface appear to be equivalent to those of smooth silicon surfaces produced by chemical etching in mixtures of nitric and hydrofluoric acids. On the other hand, the reactive surface produced at a valence of about 2, with anodic hydrogen evolution, is capable of practically shorting-out a silicon p-n junction. The electrical properties of this surface tend to change upon standing in air.

ACKNOWLEDGEMENTS

Most of the experiments mentioned in this paper were carried out by my wife, Ingeborg. An exception is the double-dimpling of germanium by light, which was done by T. C. Hall. The dimpling procedures of Figs. 9 and 10 are based on suggestions by J. M. Early. The effect of light upon electrolytic etching was called to my attention by O. Loosme. W. G. Pfann provided the germanium crystals grown with different degrees of stirring.

REFERENCES

1. J. F. Barry, I.R.E.-A.I.E.E. Semiconductor Device Research Conference, Philadelphia, June, 1955.
2. A. Uhlir, Jr., Rev. Sci. Inst., **26**, pp. 965-968, 1955.
3. W. E. Bailey, U. S. Patent No. 1,416,929, May 23, 1922.
4. Bradley, *et al.* Proc. I.R.E., **24**, pp. 1702-1720, 1953.
5. M. V. Sullivan and J. H. Eigler, to be published.
6. S. L. Miller, Phys. Rev. **99**, p. 1234, 1955.
7. W. H. Brattain and C. G. B. Garrett, B.S.T.J., **34**, pp. 129-176, 1955.
8. E. H. Borneman, R. F. Schwarz, and J. J. Stickler, J. Appl. Phys., **26**, pp. 1021-1029, 1955.
9. D. R. Turner, to be submitted to the Journal of the Electrochemical Society.
10. R. D. Heidenreich, U. S. Patent No. 2,619,414, Nov. 25, 1952.
11. T. S. Moss, L. Pincherle, A. M. Woodward, Proc. Phys. Soc. London, **66B**, p. 743, 1953.
12. T. M. Buck and F. S. McKim, Cincinnati Meeting of the Electrochemical Society, May, 1955.
13. F. L. Vogel, W. G. Pfann, H. E. Corey, and E. E. Thomas, Phys. Rev., **90**, p. 489, 1953.
14. S. G. Ellis, Phys. Rev., **100**, pp. 1140-1141, 1955.
15. *Electronics*, **27**, No. 5, p. 194, May, 1954.
16. F. Jirsa, Z. f. Anorg. u. Allgemeine Chem., Bd. **268**, p. 84, 1952.



A Large Signal Theory of Traveling Wave Amplifiers

Including the Effects of Space Charge and Finite Coupling Between the Beam and the Circuit

By PING KING TIEN

Manuscript received October 11, 1955)

The non-linear behavior of the traveling-wave amplifier is calculated in this paper by numerically integrating the motion of the electrons in the presence of the circuit and the space charge fields. The calculation extends the earlier work by Nordsieck and the small-C theory by Tien, Walker and Wolontis, to include the space charge repulsion between the electrons and the effect of a finite coupling between the circuit and the electron beam. It however differs from Poulter's and Rowe's works in the methods of calculating the space charge and the effect of the backward wave.

The numerical work was done using 701-type I.B.M. equipment. Results of calculation covering QC from 0.1 to 0.4, b from 0.46 to 2.56 and k from 1.25 to 2.50, indicate that the saturation efficiency varies between 23 per cent and 37 per cent for C equal to 0.1 and between 33 per cent and 40 per cent for C equal to 0.15. The voltage and the phase of the circuit wave, the velocity spread of the electrons and the fundamental component of the charge-density modulation are either tabulated or presented in curves. A method of calculating the backward wave is provided and its effect fully discussed.

1. INTRODUCTION

Theoretical evaluation of the maximum efficiency attainable in a traveling-wave amplifier requires an understanding of the non-linear behavior of the device at various working conditions. The problem has been approached in many ways. Pierce,¹ and later Hess,² and Birdsall³ and Caldwell⁴ investigated the efficiency or the output power, using certain specific assumptions about the highly bunched electron beam. They either assume a beam in the form of short pulses of electrons, or, specify

an optimum ratio of the fundamental component of convection current to the average or d-c current. The method, although an abstract one, generally gives the right order of the magnitude. When the usual wave concept fails for a beam in which overtaking of the electrons arises, we may either overlook effects from overtaking, or, using the Boltzman's transport equation search for solutions in series form. This attack has been pursued by Parzen⁵ and Kiel,⁶ although their work is far from complete. The most satisfying approach to date is Nordsieck's analysis.⁷ Nordsieck followed a typical set of "electrons" and calculated their velocities and positions by numerically integrating a set of equations of motion. Poulter⁸ has extended Nordsieck equations to include space charge, finite C and circuit loss, although he has not perfectly taken into account the space charge and the backward wave. Recently Tien, Walker, and Wolontis⁹ have published a small C theory in which "electrons" are considered in the form of uniformly charged discs and the space charge field is calculated by computing the force exerted on one disc by the others. Results extended to finite C , have been reported by Rowe,¹⁰ and also by Tien and Walker.¹¹ Rowe, using a space charge expression similar to Poulter's, computed the space charge field based on the electron distribution in time instead of the distribution in space. This may lead to appreciable error in his space charge term, although its influence on the final results cannot be easily evaluated.

In the present analysis, we shall adopt the model described by Tien, Walker and Wolontis, but wish to add to it the effect of a finite beam to circuit coupling. A space charge expression is derived taking into account the fact that the a-c velocities of the electrons are no longer small compared with the average velocity. Equations are rewritten to retain terms involving C . As the backward wave becomes appreciable when C increases, a method of calculating the backward wave is provided and the effect of the backward wave is studied. Finally, results of the calculation covering useful ranges of design and operating parameters are presented and analyzed.

2. ASSUMPTIONS

To recapitulate, the major assumptions which we have made are:

1. The problem is considered to be one dimensional, in the sense that the transverse motions of the electrons are prohibited, and the current, velocity, and fields, are functions only of the distance along the tube and of the time.
2. Only the fundamental component of the current excites waves on the circuit.

3. The space charge field is computed from a model in which the helix is replaced by a conducting cylinder, and electrons are uniformly charged discs. The discs are infinitely thin, concentric with the helix and have a radius equal to the beam radius.

4. The circuit is lossfree.

These are just the assumptions of the Tien-Walker-Wolontis model. In addition, we shall assume a small signal applied at the input end of a long tube, where the beam entered unmodulated. What we are looking for are therefore the characteristics of the tube beyond the point at which the device begins to act non-linearly. Let us imagine a flow of electron discs. The motions of the discs are computed from the circuit and the space charge fields by the familiar Newton's force equation. The electrons, in turn, excite waves on the circuit according to the circuit equation¹² derived either from Brillouin's model¹³ or from Pierce's equivalent circuit.¹⁴ The force equation, the circuit equation, and the equation of conservation of charge in kinematics,¹⁵ are the three basic equations from which the theory is derived.

3. FORWARD AND BACKWARD WAVES

In the traveling-wave amplifier, the beam excites forward and backward waves on the circuit. (We mean by "forward" wave, the wave which propagates in the direction of the electron flow, and by "backward" wave, the wave which propagates in the opposite direction.) Because of phase cancellation, the energy associated with the backward wave is small, but increases with the beam to circuit coupling. It is therefore important to compute it accurately. In the first place, the waves on the circuit must satisfy the circuit equation¹²

$$\frac{\partial^2 V(z, t)}{\partial t^2} - v_0^2 \frac{\partial^2 V(z, t)}{\partial z^2} = v_0 Z_0 \frac{\partial^2 \rho_\omega(z, t)}{\partial t^2} \quad (1)$$

Here, V is the total voltage of the waves. v_0 and Z_0 are respectively the phase velocity and the impedance of the cold circuit. z is the distance along the tube and t , the time. ρ_ω is the fundamental component of the linear charge density. V and ρ_ω are functions of z and t . The complete solution of (1) is in the form

$$\begin{aligned} V(z) = & C_1 e^{-\Gamma_0 z} + C_2 e^{\Gamma_0 z} \\ & + e^{-\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_{z'}^z e^{\Gamma_0 z} \rho_\omega(z) dz \\ & + e^{\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_{z'}^z e^{-\Gamma_0 z} \rho_\omega(z) dz \end{aligned} \quad (2)$$

where the common factor $e^{j\omega t}$ is omitted. $\Gamma_0 = j(\omega/v_0)$, $j = \sqrt{-1}$ and ω is the angular frequency. C_1 and C_2 are arbitrary constants which will be determined by the boundary conditions at the both ends of the beam. The first two terms are the solutions of the homogeneous equation (or the complementary functions) and are just the cold circuit waves. The third and the fourth terms are functions of electron charge density and are the particular solution of the equation.

Let us consider a long traveling-wave tube in which the beam starts from $z = 0$ and ends at $z = D$. The motion of electrons observed at any particular position is periodic in time, though it varies from point to point along the beam. To simplify the picture, we may divide the beam along the tube into small sections and consider each of them as a current element uniform in z and periodic in time. Each section of beam, or each current element excites on the circuit a pair of waves equal in amplitudes, one propagating toward the right (i.e., forward) and the other, toward the left. One may in fact imagine that these are trains of waves supported by the periodic motion of the electrons in that section of the beam. Obviously, a superposition of these waves excited by the whole beam gives the actual electromagnetic field distribution on the circuit. One may thus compute the forward traveling wave at z by summing all the waves at z which come from the left. Stated more specifically, the forward traveling energy at z is contributed by the waves excited by the current elements at the left of the point z . Similarly the backward traveling energy, (or the backward wave) at z is contributed by the waves excited by the current elements at the right of the point z . It follows obviously from this picture that there is no forward wave at $z = 0$ (except one corresponding to the input signal), and no backward wave at $z = D$. (This implies that the output circuit is matched.) With these boundary conditions, (1) is reduced to

$$V(z) = V_{\text{input}} e^{-\Gamma_0 z} + e^{-\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_0^z e^{\Gamma_0 z} \rho_\omega(z) dz + e^{\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_z^D e^{-\Gamma_0 z} \rho_\omega(z) dz \quad (3)$$

Equations (2) and (3) have been obtained by Poulter.⁸ The first term of (3) is the wave induced by the input signal. It propagates as though the beam were not present. The second term is the voltage at z contributed by the charges between $z = 0$ and $z = z$. It is just the voltage of the forward wave described earlier. Similarly the third term which is the voltage at z contributed by the charges between $z = z$ and $z = D$ is the voltage of the backward wave at the point z . Denote F and B respec-

tively the voltages of the forward and the backward waves, we have

$$F(z) = V_{\text{input}} e^{-\Gamma_0 z} + e^{-\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_0^z e^{\Gamma_0 z} \rho_\omega(z) dz \quad (4)$$

$$B(z) = e^{\Gamma_0 z} \frac{\Gamma_0 v_0 Z_0}{2} \int_z^D e^{-\Gamma_0 z} \rho_\omega(z) dz \quad (5)$$

It can be shown by direct substitution that F and B satisfy respectively the differential equations

$$\frac{\partial F(z, t)}{\partial z} + \frac{1}{v_0} \frac{\partial F(z, t)}{\partial t} = \frac{Z_0}{2} \frac{\partial \rho_\omega(z, t)}{\partial t} \quad (6)$$

$$\frac{\partial B(z, t)}{\partial z} - \frac{1}{v_0} \frac{\partial B(z, t)}{\partial t} = -\frac{Z_0}{2} \frac{\partial \rho_\omega(z, t)}{\partial t} \quad (7)$$

We put (4) and (5) in the form of (6) and (7) simply because the differential equations are easier to manipulate than the integral equations. In fact, we should start the analysis from (6) and (7) if it were not for a physical picture useful to the understanding of the problem. Equations (6) and (7) have the advantage of not being restricted by the boundary conditions at $z = 0$ and D , which we have just imposed to derive (4) and (5). Actually, we can derive (6) and (7) directly from the Brillouin model¹³ in the following manner. Suppose V , I and Z_0 are respectively the voltage, current and the characteristic impedance of a transmission line system in the usual sense. $(V + IZ_0)$ must then be the forward wave and $(V - IZ_0)$ must be the backward wave. If we substituted F and B in these forms into (1) of the Brillouin's paper,¹³ we should obtain exactly (6) and (7).

It is obvious that the first and third terms of (2) are respectively the complementary function and the particular solution of (6), and similarly the second and the fourth terms of (2) are respectively the complementary function and the particular solution of (7). From now on, we shall overlook the complementary functions which are far from synchronism with the beam and are only useful in matching the boundary conditions. It is the particular solutions which act directly on the electron motion. With these in mind, it is convenient to put F and B in the form

$$F(z, t) = \frac{Z_0 I_0}{4C} [a_1(y) \cos \varphi - a_2(y) \sin \varphi] \quad (8)$$

$$B(z, t) = \frac{Z_0 I_0}{4C} [b_1(y) \cos \varphi - b_2(y) \sin \varphi] \quad (9)$$

where $a_1(y)$, $a_2(y)$, $b_1(y)$ and $b_2(y)$ are functions of y . y and φ are independent variables and have been used by Nordsieck to replace the variables, z and t , such as

$$y = C \frac{\omega}{u_0} z$$

$$\varphi = \omega \left(\frac{z}{v_0} - t \right)$$

Here as defined earlier, v_0 is the phase velocity of the cold circuit and u_0 the average velocity of the electrons. They are related by the parameter b defined by Pierce as

$$\frac{u_0}{v_0} = \frac{1}{(1 - bC)}$$

C is the gain parameter also defined by Pierce,

$$C^3 = \frac{Z_0 I_0}{4 V_0}$$

in which, V_0 and I_0 are respectively the beam voltage and current. Adding (6) to (7), we obtain an important relation between F and B , that is,

$$\frac{\partial F(z, t)}{\partial z} + \frac{1}{v_0} \frac{\partial F(z, t)}{\partial t} = -\frac{\partial B(z, t)}{\partial z} + \frac{1}{v_0} \frac{\partial B(z, t)}{\partial t} \quad (10)$$

Substituting (8) and (9) into (10) and carrying out some algebraic manipulation, we obtain

$$b_1(y) = -\frac{C}{2(1 + bC)} \frac{d}{dy} [a_2(y) + b_2(y)] \quad (11)$$

$$b_2(y) = \frac{C}{2(1 + bC)} \frac{d}{dy} [a_1(y) + b_1(y)]$$

or

$$B(z, t) = -\frac{Z_0 I_0}{4C} \frac{C}{2(1 + bC)} \cdot \left[\frac{d(a_2(y) + b_2(y))}{dy} \cos \varphi + \frac{d(a_1(y) + b_1(y))}{dy} \sin \varphi \right] \quad (12a)$$

For better understanding of the problem, we shall first solve (12a) approximately. Assuming for the moment that $b_1(y)$ and $b_2(y)$ are small compared with $a_1(y)$ and $a_2(y)$ and may be neglected in the right-hand

member of the equation, we obtain for the first order solution

$$B(z, t) \cong \frac{Z_0 I_0}{4C} \left(-\frac{C}{2(1 + bC)} \left[\frac{da_1(y)}{dy} \sin \varphi + \frac{da_2(y)}{dy} \cos \varphi \right] \right) \quad (12b)$$

Of course, the solution (12b) is justified only when $b_1(y)$ and $b_2(y)$ thus obtained are small compared with $a_1(y)$ and $a_2(y)$. The exact solution¹⁶ of B obtained by successive approximation reads

$$B(z, t) = \frac{Z_0 I_0}{4C} \left(-\frac{C}{2(1 + bC)} \left[\frac{da_1(y)}{dy} \sin \varphi + \frac{da_2(y)}{dy} \cos \varphi \right] + \frac{C^2}{4(1 + bC)^2} \left[-\frac{d^2 a_1(y)}{dy^2} \cos \varphi + \frac{d^2 a_2(y)}{dy^2} \sin \varphi \right] + \dots \right) \quad (12c)$$

It may be seen that the term involving

$$\frac{C^2}{4(1 + bC)^2}$$

and the higher order terms are neglected in our approximate solution. For C equal to few tenths, the difference between (12b) and (12c) only amounts to few per cent. We thus can calculate the backward wave by (12b) or (12c) from the derivatives of the forward wave. To obtain the complete solution of the backward wave, we should add to (12b) or (12c) a solution of the homogeneous equation. We shall return to this point later.

4. WORKING EQUATIONS

With this discussion of the backward wave, we are now in a position to derive the working equations on which our calculations are based. In Nordsieck's notation, each electron is identified by its initial phase. Thus, $\varphi(y, \varphi_0)$ and $C u_0 w(y, \varphi_0)$ are respectively the phase and the ac velocity of the electron which has an initial phase φ_0 . It should be remembered that y is equal to

$$C \frac{\omega}{u_0} z$$

and is used by Nordsieck as an independent variable to replace the variable z . Let us consider an electron which is at z_0 when $t = 0$ and is at z (or y) when $t = t$. Its initial phase is then

$$\varphi_0 = \frac{\omega z_0}{u_0}$$

and its phase at y is

$$\begin{aligned}\varphi(y, \varphi_0) &= \omega \left(\frac{z}{v_0} - t \right) \\ &= \omega \left(\frac{z}{u_0} - t \right) + by\end{aligned}$$

The velocity of the electron is expressed as

$$\frac{dz}{dt} = u_0[1 + Cw(y, \varphi_0)]$$

where u_0 is the average velocity of the electrons, and, $Cu_0w(y, \varphi_0)$ as mentioned earlier, is the ac velocity of the electron when it is at the position y . The electron charge density near an electron which has an initial phase φ_0 and which is now at y , can be computed by the equation of conservation of charge,¹⁵ it is

$$\rho(y, \varphi_0) = -\frac{I_0}{u_0} \left| \frac{d\varphi_0}{d\varphi(y, \varphi_0)} \right| \frac{1}{1 + Cw(y, \varphi_0)} \quad (13)$$

One should recall here that I_0 is the dc beam current and has been defined as a positive quantity. When several electrons with different initial phases are present at y simultaneously, a summation of

$$\left| \frac{d\varphi_0}{d(y, \varphi_0)} \right|$$

of these electrons should be used in (13). From (13), the fundamental component of the electron charge density is

$$\begin{aligned}\rho_\omega(z, t) &= -\frac{1}{\pi} \frac{I_0}{u_0} \left(\sin \varphi \int_0^{2\pi} d\varphi_0 \frac{\sin \varphi(y, \varphi_0)}{1 + Cw(y, \varphi_0)} \right. \\ &\quad \left. + \cos \varphi \int_0^{2\pi} d\varphi_0 \frac{\cos \varphi(y, \varphi_0)}{1 + Cw(y, \varphi_0)} \right) \quad (14)\end{aligned}$$

These are important relations given by Nordsieck and should be kept in mind in connection with later work. In addition, we shall frequently use the transformation

$$\frac{d}{dt} = \frac{dz}{dt} \frac{\partial}{\partial z} = C\omega(1 + Cw(y, \varphi_0)) \frac{\partial}{\partial y}$$

which is written following the motion of the electron. Let us start from the forward wave. It is computed by means of (6). After substituting (8) and (14) into (6), we obtain by equating the $\sin \varphi$ and the $\cos \varphi$

terms

$$\frac{da_1(y)}{dy} = -\frac{2}{\pi} \int_0^{2\pi} d\varphi_0 \frac{\sin \varphi(y, \varphi_0)}{1 + Cw(y, \varphi_0)} \quad (15)$$

$$\frac{da_2(y)}{dy} = -\frac{2}{\pi} \int_0^{2\pi} d\varphi_0 \frac{\cos \varphi(y, \varphi_0)}{1 + Cw(y, \varphi_0)} \quad (16)$$

Next we shall calculate the electron motion. We express the acceleration of an electron in the form

$$\frac{d^2z}{dt^2} = C\omega u_0(1 + Cw(y, \varphi_0)) \frac{dw(y, \varphi_0)}{dy}$$

and calculate the circuit field by differentiating F in (8) and B in (12c) with respect to z . One thus obtains from Newton's law

$$\begin{aligned} 2[1 + Cw(y, \varphi_0)] \frac{dw(y, \varphi_0)}{dy} &= (1 + bC)[a_1(y) \sin \varphi + a_2(y) \cos \varphi] \\ &\quad - \frac{C}{2} \left[\frac{da_1(y)}{dy} \cos \varphi - \frac{da_2(y)}{dy} \sin \varphi \right] \quad (17) \\ &\quad + \frac{C^2}{4(1 + bC)} \left[\frac{d^2a_1(y)}{dy^2} \sin \varphi + \frac{d^2a_2(y)}{dy^2} \cos \varphi \right] - \frac{2e}{u_0 m \omega C^2} E_s \end{aligned}$$

Here E_s is the space charge field, which will be discussed in detail later. Finally a relation between $w(y, \varphi_0)$ and $\varphi(y, \varphi_0)$ is obtained by means of (13)

$$\frac{d\varphi(y, \varphi_0)}{dy} - b = \frac{w(y, \varphi_0)}{1 + Cw(y, \varphi_0)} \quad (18)$$

Equations (15), (16), (17) and (18) are the four working equations which we have derived for finite C and including space charge.

Instead of writing the equations in the above form, Rowe,¹⁰ ignoring the backward wave, derives (15) and (16) directly from the circuit equation (1). He obtains an additional term

$$\frac{C}{2} \frac{d^2a_2}{dy^2}$$

for (15) and another term

$$\frac{C}{2} \frac{d^2a_1}{dy^2}$$

for (16). It is apparent that the backward wave, though generally a small quantity, does influence the terms involving C .

5. THE SPACE CHARGE EXPRESSION

We have mentioned earlier that the space charge field is computed from the disc-model suggested by Tien, Walker and Wolontis. In their calculation, the force excited on one disc by the other is approximated by an exponential function

$$F_s = \frac{q^2}{2\pi r_0^2 \epsilon_0} e^{-[\alpha(z' - z)/r_0]}$$

Here r_0 is the radius of the disc or the beam, q is the charge carried by each disc, and ϵ_0 is the dielectric constant of the medium. The discs are supposed to be respectively at z and z' . α is a constant and is taken equal to 2.

Consider two electrons which have their initial phases φ_0 and φ_0' and which reach the position y (or z) at times t and t' respectively. The time difference,

$$t - t' = \frac{1}{\omega} \left[\omega t - \frac{\omega}{v_0} z - \left(\omega t' - \frac{\omega}{v_0} z \right) \right] = \frac{1}{\omega} [\varphi(y, \varphi_0') - \varphi(y, \varphi_0)]$$

multiplied by the velocity of the electron $u_0[1 + Cw(y, \varphi_0')]$ is obviously the distance between the two electrons at the time t . Thus

$$(z' - z)_{t=t} = \frac{1}{\omega} [\varphi(y, \varphi_0') - \varphi(y, \varphi_0)] u_0 [1 + Cw(y, \varphi_0')] \quad (19a)$$

In this equation, we are actually taking the first term of the Taylor's expansion,

$$(z' - z)_{t=t} = \frac{dz(y, \varphi_0')}{dt} \Big|_{t=t} (t - t') + \frac{1}{2} \frac{d^2 z(y, \varphi_0')}{dt^2} \Big|_{t=t} (t - t')^2 \quad (19b)$$

$$+ \dots$$

It is clear that the electrons at y may have widely different velocities after having traveled a long distance from the input end, but changes in their velocities, in the vicinity of y and in a time-period of around 2π , are relatively small. This is why we must keep the first term of (19b) and may neglect the higher order terms. From (19a) the space charge field E_s in (17) is

$$\frac{2e}{m\omega C^2 u_0} E_s = \left(\frac{\omega_p}{\omega C} \right)^2 \int_{-\infty}^{+\infty} e^{-k|\varphi(y, \varphi_0 + \phi) - \varphi(y, \varphi_0)|[1 + Cw(y, \varphi_0 + \phi)]} d\phi \operatorname{sgn}(\varphi(\varphi_0 + \phi) - \phi(y, \varphi_0))$$

Here, e/m is the ratio of electron charge to mass, ω_p is the electron

angular plasma frequency for a beam of infinite extent, and k is

$$k = \frac{\alpha}{\frac{\omega}{u_0} r_0} = \frac{2}{\frac{\omega}{u_0} r_0} \quad (20)$$

In the small C theory, the distribution of electrons in time or in time-phase at z is approximately the same as the distribution in z (also expressed in the unit of time-phase) at the vicinity of z . This is, however, not true when C becomes finite. The difference between the time and space distributions is the difference between unity and the factor $(1 + Cw(y, \varphi_0'))$. We can show later that the error involved in considering the time phase as the space phase can easily reach 50 per cent or more, depending on the velocity spread of the electrons.

6. NUMERICAL CALCULATIONS

Although the process of carrying out numerical computations has been discussed in Nordsieck's paper, it is desirable to recapitulate here a few essential points including the new feature added. Using the working equations (15), (16), (17) and (18),

$$\frac{da_1}{dy}, \frac{da_2}{dy}, \frac{dw}{dy} \quad \text{and} \quad \frac{d\varphi}{dy}$$

are calculable from a_1 , a_2 , w and φ . The distance is divided into equal intervals of Δy , and the forward integrations of a_1 , a_2 , w and φ are performed by a central difference formula

$$a_1(y + \Delta y) = a_1(y) + \frac{da_1}{dy} \Big|_{y+1/2\Delta y} \cdot \Delta y$$

In addition,

$$\frac{d^2 a_1}{dy^2} \quad \text{and} \quad \frac{d^2 a_2}{dy^2}$$

in (17) are computed from the second difference formula such that

$$\frac{d^2 a_1}{dy^2} \Big|_{y=y} = \left[\frac{da_1}{dy} \Big|_{y+1/2\Delta y} - \frac{da_1}{dy} \Big|_{y-1/2\Delta y} \right] \div \Delta y$$

We thus calculate the behavior along the tube by forward integration made in steps of Δy , starting from $y = 0$. At $y = 0$ the initial conditions are determined from Pierce's linearized theory. Because of its complications in notation, this will be discussed in detail in Appendix I.

Numerical calculations were carried out using the 701-type I.B.M.

TABLE I

Case No.	QC	k	C	b	μ_1	μ_2	$y(\text{SAT.})$	$A(y)(\text{SAT.})$	$ \theta(y) - \mu_2 y (\text{SAT.})$
1	0.1	2.5	0.05	0.455	$\mu_1 \text{ max.}$ 0.795662	-0.748052	5.6	1.26	0.415
2	0.1	2.5	0.1	0.541	$\mu_1 \text{ max.}$ 0.827175	-0.787624	5.2	1.24	0.482
3	0.1	2.5	0.1	1.145	$0.941\mu_1 \text{ max.}$ 0.778535	-1.05370	5.6	1.31	0.820
4	0.1	2.5	0.1	1.851	$0.66\mu_1 \text{ max.}$ 0.550736	-1.37968	7.0	1.36	1.05
5	0.1	2.5	0.2	0.720	$\mu_1 \text{ max.}$ 0.900312	-0.873606	4.8	1.02	0.726
6	0.2	1.25	0.1	0.875	$\mu_1 \text{ max.}$ 0.769795	-1.04078	5.9	1.22	0.570
7	0.2	1.25	0.1	1.422	$0.951\mu_1 \text{ max.}$ 0.724527	-1.29469	6.0	1.30	0.803
8	0.2	1.25	0.1	2.072	$0.666\mu_1 \text{ max.}$ 0.512528	-1.60435	7.6	1.35	1.08
9	0.2	2.5	0.05	0.765	$\mu_1 \text{ max.}$ 0.731493	-0.973376	6.2	1.30	0.412
10	0.2	2.5	0.1	0.875	$\mu_1 \text{ max.}$ 0.769795	-1.04078	5.8	1.22	0.490
11	0.2	2.5	0.1	1.422	$0.941\mu_1 \text{ max.}$ 0.724527	-1.29469	6.0	1.26	0.720
12	0.2	2.5	0.1	2.072	$0.666\mu_1 \text{ max.}$ 0.512528	-1.60435	7.2	1.25	0.92
13	0.2	2.5	0.1	2.401	$0.300\mu_1 \text{ max.}$ 0.230930	-1.76243	12.4	1.24	1.36
14	0.2	2.5	0.15	0.976	$\mu_1 \text{ max.}$ 0.812900	-1.10656	5.4	1.11	0.572
15	0.2	2.5	0.15	1.549	$0.941\mu_1 \text{ max.}$ 0.765101 <small>0.765101</small>	-1.37540	5.8	1.14	1.03
16	0.2	2.5	0.15	2.2311	$0.666\mu_1 \text{ max.}$ 0.541234	-1.70180	7.0	1.12	1.22
17	0.2	2.5	0.15	2.575	$0.300\mu_1 \text{ max.}$ 0.243864	-1.86844	10.8	1.04	1.34
18	0.4	2.5	0.05	1.25	$\mu_1 \text{ max.}$ 0.653014	-1.36746	7.6	1.26	0.315
19	0.4	2.5	0.1	1.38	$\mu_1 \text{ max.}$ 0.701470	-1.47477	6.6	1.11	0.674
20	0.4	2.5	0.1	1.874	$0.941\mu_1 \text{ max.}$ 0.660223	-1.71341	7.8	1.19	1.05
21	0.4	2.5	0.1	2.458	$0.666\mu_1 \text{ max.}$ 0.467038	-1.99840	8.6	1.09	1.25

equipment. The problem was programmed by Miss D. C. Legaus. The cases computed are listed in Table I in which μ_1 and μ_2 are respectively Pierce's x_1 and y_1 , and $A, (\theta - \mu_2 y)$ and y at saturation will be discussed later. All the cases were computed with $\Delta y = 0.2$ using a model based on 24 electron discs per electronic wavelength. To estimate the error involved in the numerical work, Case (10) has been repeated for 48 electrons and Cases (10) and (19) for $\Delta y = 0.1$. The results obtained by using different numbers of electrons are almost identical and those obtained by varying the interval Δy indicate a difference in $A(y)$ less than 1 per cent for Case (10) and about 6 per cent for Case (19). As error generally increases with QC and C the cases listed in this paper are limited to $QC = 0.4$ and $C = 0.15$. For larger QC or C , a model of more electrons or a smaller interval of integration, or both should be used.

7. POWER OUTPUT AND EFFICIENCY

Define

$$\begin{aligned} A(y) &= \sqrt{\frac{1}{4} a_1(y)^2 + a_2(y)^2} \\ -\theta(y) &= \tan^{-1} \frac{a_2(y)}{a_1(y)} + b y \end{aligned} \quad (21)$$

We have then

$$F(z, t) = \frac{Z_0 I_0}{C} A(y) \cos \left[\frac{\omega z}{u_0} - \omega t - \theta(y) \right] \quad (22)$$

The power carried by the forward wave is therefore

$$\left(\frac{F^2}{Z_0} \right)_{\text{average}} = 2CA^2 I_0 V_0 \quad (23)$$

and the efficiency is

$$\text{Eff.} = \frac{2CA^2 I_0 V_0}{I_0 V_0} = 2CA^2 \quad \text{or} \quad \frac{\text{Eff.}}{C} = 2CA^2 \quad (24)$$

In Table I, the values of $A(y)$, $\theta(y)$ and y at the saturation level are listed for every case computed. We mean by the saturation level, the distance along the tube or the value of y at which the voltage of the forward wave or the forward traveling power reaches its first peak. The $\text{Eff.}/C$ at the saturation level is plotted in Fig. 1 versus QC , for $k = 2.5$, b for maximum small-signal gain and $C = \text{small}, 0.05, 0.1, 0.15$ and 2. It is also plotted versus b in Fig. 2 for $QC = 0.2$, $k = 2.5$ and $C = \text{small}, 0.1$ and 0.15, and in Fig. 3 for $QC = 0.2$, $C = 0.1$ and $k = 1.25$ and 2.50. In Fig. 2 the dotted curves indicate the values of b at which

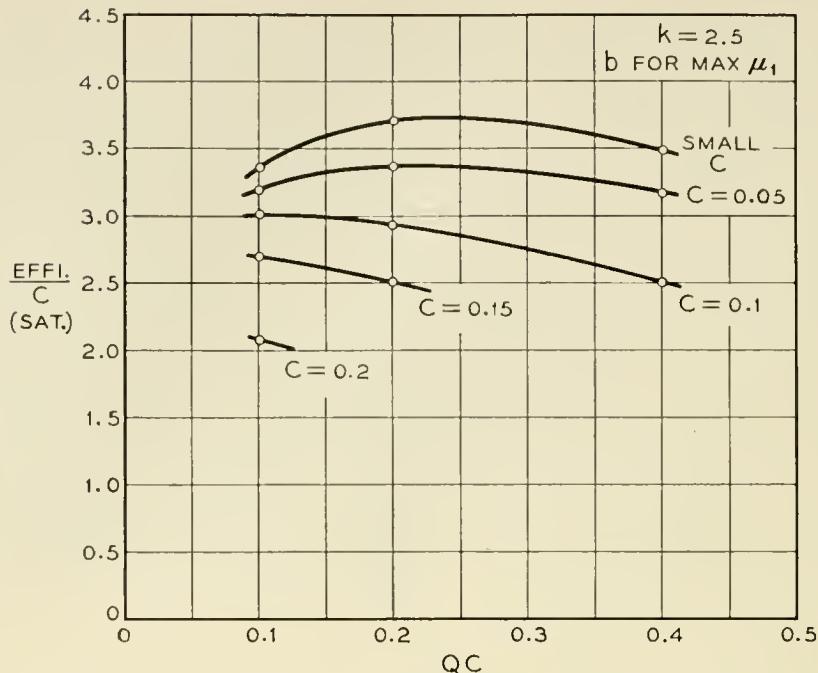


Fig. 1 — The saturation eff./C versus QC , for $k = 2.5$, b for maximum small-signal gain and $C = \text{small}, 0.1, 0.15$ and 0.2 .

$\mu_1 = \mu_1(\text{max})$, $0.94 \mu_1(\text{max})$, $0.67 \mu_1(\text{max})$ and $0.3 \mu_1(\text{max})$, respectively. It is seen that Eff./C decreases as C increases particularly when b is large. It is almost constant between $k = 1.25$ and 2.50 and decreases slowly for large values of C when QC increases.

The (Eff./C) at saturation is also plotted versus QC in Fig. 4(a) for small C , and in Fig. 4(b) for $C = 0.1$. It should be noted that for $C = 0.1$ the values of Eff./C fall inside a very narrow region say between 2.5 to 3.5, whereas for small C they vary widely.

8. VELOCITY SPREAD

In a traveling-wave amplifier, when electrons are decelerated by the circuit field, they contribute power to the circuit, and when electrons are accelerated, they gain kinetic energy at the expense of the circuit power. It is therefore of interest to plot the actual velocities of the fastest and the slowest electrons at the saturation level and find how they vary with the parameters QC , C , b and k . This is done in Fig. 5. These velocities are also plotted versus y for Case 10 in Fig. 6, in which, the $A(y)$ curve is added for reference.

9. THE BACKWARD WAVE AND THE FUNDAMENTAL COMPONENT OF THE ELECTRON CHARGE DENSITY

Our calculation of efficiency has been based on the power carried by the forward wave only. One may, however, ask about the actual power

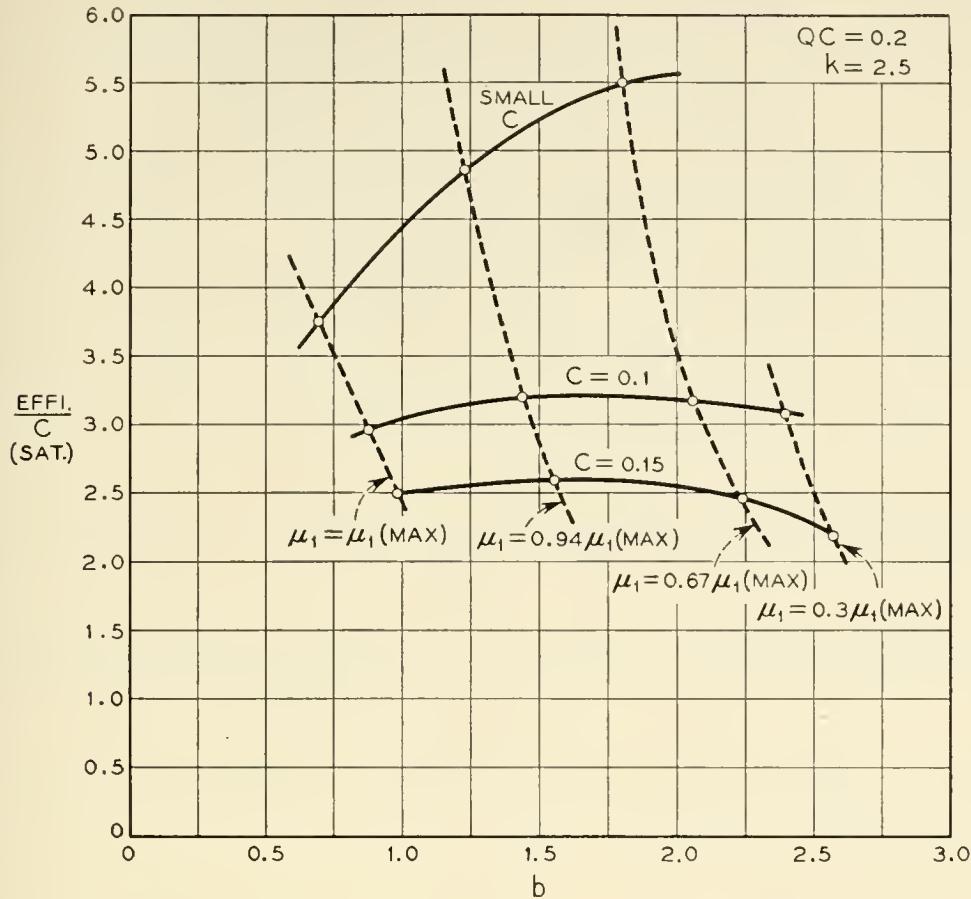


Fig. 2 — The saturation eff./ C versus b , for $k = 2.5$, $QC = 0.2$, and $C = \text{small}$, 0.1 and 0.15 . The dotted curves indicate the values of b for $\mu_1 = 1$, 0.94 , 0.67 , and 0.3 of $\mu_1(\text{max})$ respectively.

output in the presence of the backward wave. For simplicity, we shall use the approximate solution (12b) which can be written in the form

$B(z, t) \cong \text{Real Component of}$

$$\left(\frac{Z_0 I_0}{4C} \frac{C}{2(1 + bC)} \sqrt{\left(\frac{da_1(y)}{dy} \right)^2 + \left(\frac{da_2(y)}{dy} \right)^2} e^{j\omega t - \Gamma_0 z - by + j\xi} \right) \quad (12d)$$

with

$$\tan \xi = \left(-\frac{da_1(y)}{dy} \right) / \left(-\frac{da_2(y)}{dy} \right)$$

As mentioned earlier that the complete solution of (6) is obtained by adding to (12b) a complementary function such that

$$B(z, t) = C_1 e^{-j\omega t + \Gamma_0 z}$$

$$+ \frac{Z_0 I_0}{4C} \frac{C}{2(1 + bC)} \sqrt{\left(\frac{da_1}{dy'} \right)^2 + \left(\frac{da_2}{dy} \right)^2} e^{j\omega t - \Gamma_0 z - by + j\xi} \quad (25)$$

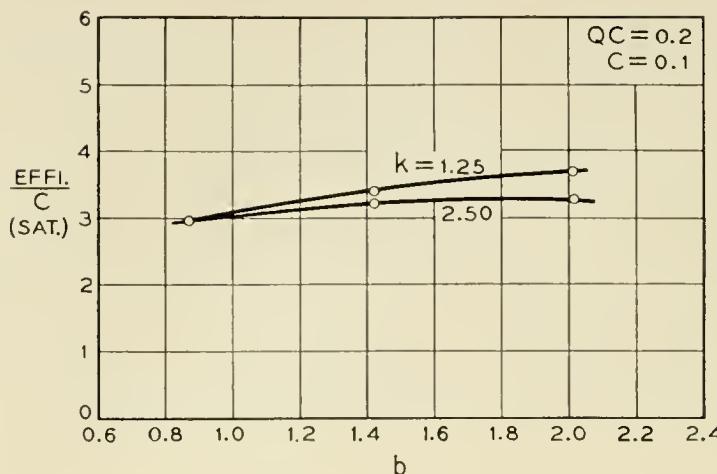


Fig. 3 — The saturation eff./C versus b , for $QC = 0.2$ $C = 0.1$ and $k = 1.25$ and 2.50 .

If the output circuit is matched by cold measurements, the backward wave must be zero at the output end, $z = D$. This determines C_1 , that is,

$$C_1 = -\frac{Z_0 I_0}{4C} \frac{C}{2(1 + bC)} \sqrt{\left(\frac{da_1(y)}{dy}\right)_{z=D}^2 + \left(\frac{da_2(y)}{dy}\right)_{z=D}^2} e^{\Gamma_0(2+bc)D+j\xi}$$

or

$$C_1 e^{j\omega t + \Gamma_0 z} = -\frac{Z_0 I_0}{4C} \frac{C}{2(1 + bC)} \sqrt{\left(\frac{da_1(y)}{dy}\right)_{z=D}^2 + \left(\frac{da_2(y)}{dy}\right)_{z=D}^2} \cdot e^{\Gamma_0(2+bc)D+j\xi} e^{j\omega t + \Gamma_0 z} \quad (26)$$

The backward wave therefore consists of two components. One compo-

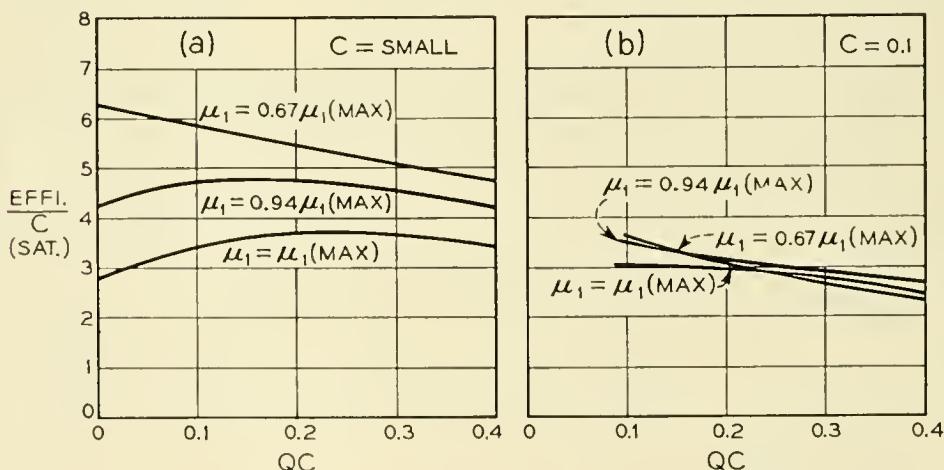


Fig. 4 — The saturation eff./C versus QC for b corresponding $\mu_1 = 1$, 0.94 and 0.67 of $\mu_1(\max)$, (a) for $C = \text{small}$, (b) for $C = 0.1$.

ment is coupled to the beam and has an amplitude equal to

$$\frac{Z_0 J_0}{4C} \frac{C}{2(1 + bC)} \sqrt{\left(\frac{da_1}{dy}\right)^2 + \left(\frac{da_2}{dy}\right)^2}$$

which generally grows with the forward wave. It thus has a much larger amplitude at the output end than at the input end. The other component is a wave of constant amplitude, which travels in the direction opposite to the electron flow with a phase velocity equal to that of the cold circuit. At the output end, $z = D$, both components have the same amplitude but are opposite in sign. One thus realizes that there exists a reflected wave of noticeable amplitude, in the form of (26), even though the output circuit is properly matched by cold measurements. Under such circumstances, the voltage at the output end is the voltage of the forward wave and the power output is the power carried by the forward wave only. This is computed in (23).

Since (26) is a cold circuit wave it may be eliminated by properly ad-

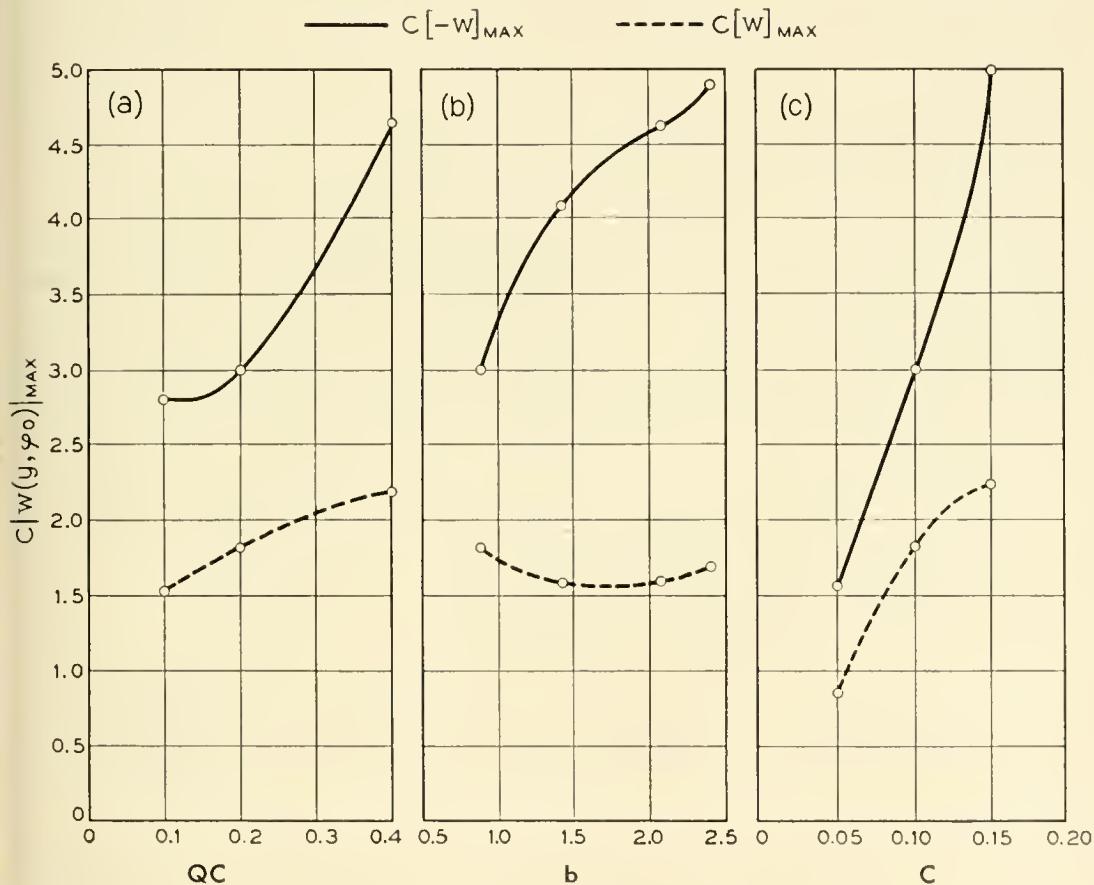


Fig. 5 — $Cw(y, \varphi_0)$ of the fast and the slowest electrons at the saturation level. (a) versus QC for $k = 2.5$, $C = 0.1$ and b for maximum small-signal gain; (b) versus b for $k = 2.50$, $C = 0.1$ and $QC = 0.2$; and (c) versus C for $k = 2.50$, $QC = 0.2$ and b for maximum small-signal gain.

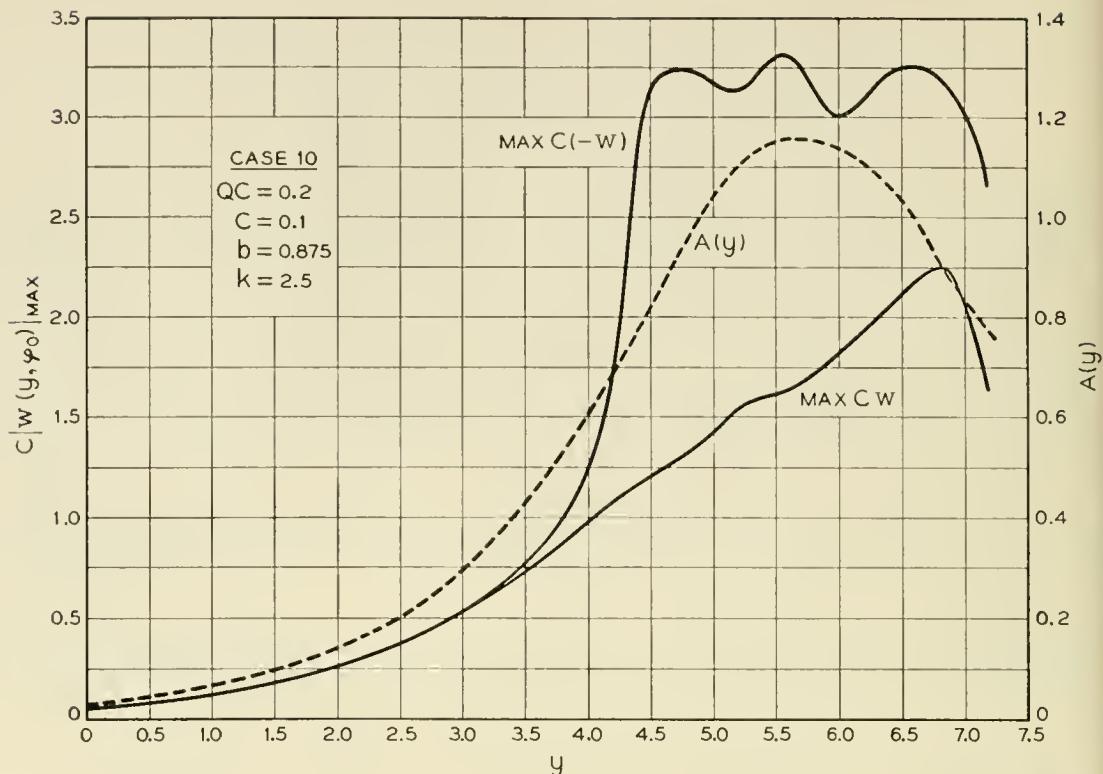


Fig. 6 — $C_w(y, \varphi_0)$ of the fast and the slowest electrons versus y for Case (10). $A(y)$ is also plotted in dotted lines for reference.

justing the impedance of the output circuit. This may be necessary in practice for the purpose of avoiding possible regenerative oscillation. In doing so, the voltage at $z = D$ is the sum of the voltage of the forward wave and that of the particular solution of the backward wave. In every case, the output power is always equal to the square of the net voltage actually at the output end divided by the impedance of the output circuit.

We find from (14), (15) and (16) that the fundamental component of electron charge density may be written as

$$\begin{aligned} \rho_\omega(z, t) &= \frac{1}{2} \frac{I_0}{u_0} \left(\sin \varphi \frac{da_1(y)}{dy} + \cos \varphi \frac{da_2(y)}{dy} \right) \\ &= \text{Real component of} \left(-\frac{1}{2} \frac{I_0}{u_0} \sqrt{\left(\frac{da_1(y)}{dy} \right)^2 + \left(\frac{da_2(y)}{dy} \right)^2} e^{j\omega - \Gamma_0 z - by + j\xi} \right) \quad (26) \end{aligned}$$

where $-I_0/u_0$ is the dc electron charge density, ρ_0 .

If (26) is compared with (12d) or (12c), it might seem surprising that the particular solution of the backward wave is just equal to the funda-

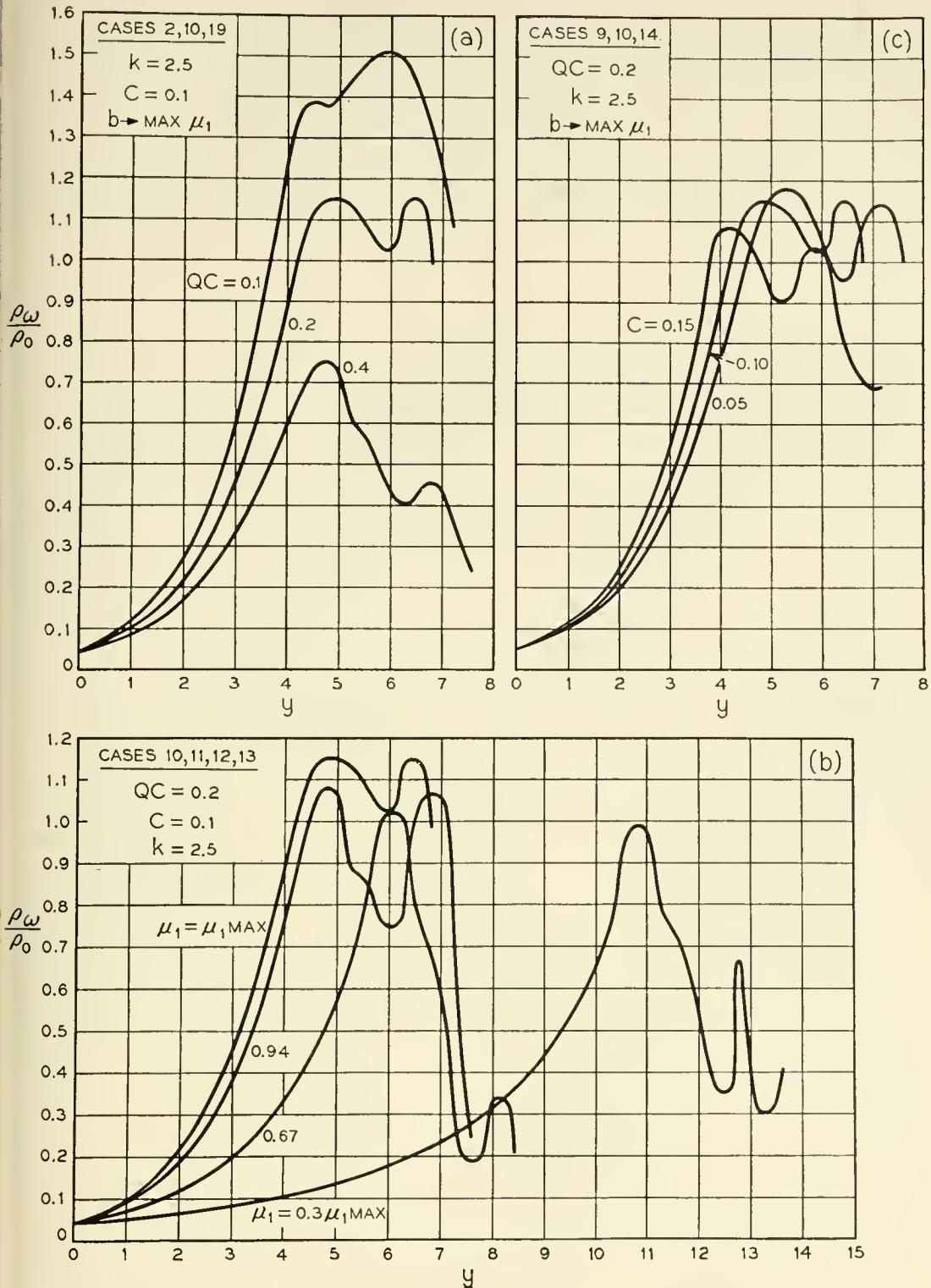


Fig. 7(a) — ρ_ω/ρ_0 versus y , (a) using QC as the parameter, for $k = 2.5$, $C = 0.1$, and b for maximum small-signal gain (Cases 2, 10, and 19); (b) using b as the parameter, for $k = 2.50$, $C = 0.1$ and $QC = 0.2$ (Cases 10, 11, 12 and 13); and (c) using C as the parameter, for $k = 2.50$, $QC = 0.2$ and b for maximum small-signal gain (Cases 9, 10 and 14).

mental component of the electron charge density of the beam multiplied by a constant

$$\left(-\frac{Z_0 J_0}{4C} \frac{C}{2(1 + bC)} \frac{2u_0}{I_0} \right) \quad (27)$$

The ratio of the electron charge density to the average charge density,

$$\frac{\rho_\omega(z)}{\rho_0}$$

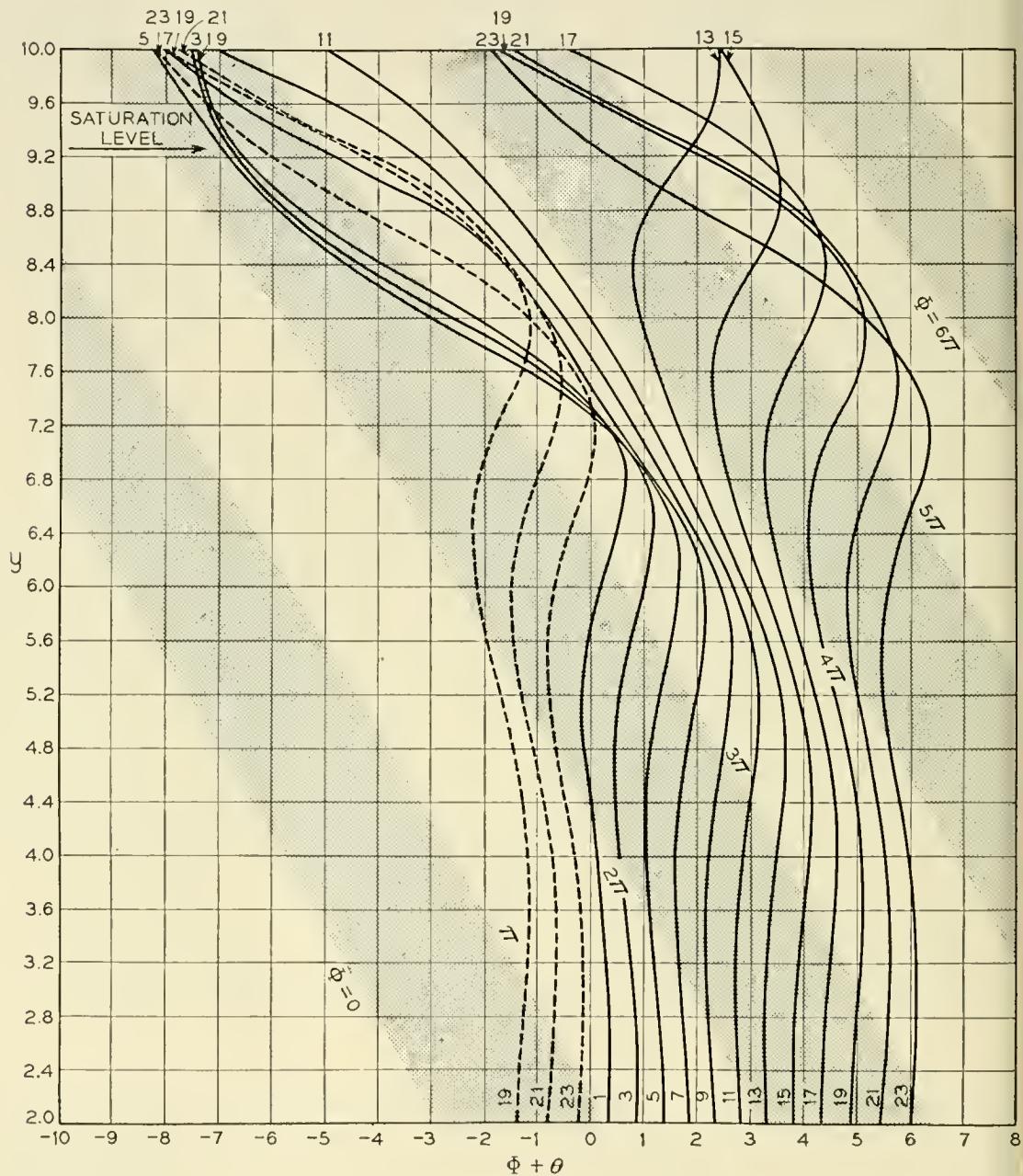


Fig. 8(a) — y versus $\varphi = by$ for $QC = 0.2$, $k = 2.5$, b for $\mu_1 = 0.67\mu_1(\text{max})$ and $C = \text{small}$.

is plotted in Fig. 7 versus y , using QC , b and C , as the parameters. They are also the curves for the backward wave (the component which is coupled to the beam) when multiplied by the proportional constant given in (27). It is interesting to see that the maximum values of ρ_ω/ρ_0 are between 1.0 and 1.2 for $QC = 0.2$ and decrease as QC increases. The peaks of the curves do not occur at the saturation values of y .

10. y VERSUS $(\varphi - by)$ DIAGRAMS

To study the effect of C , b , and QC on efficiency y versus $(\varphi - by)$ diagrams are plotted in Figs. 8(b), (c), (d) and (e) for Cases (21), (16), (10) and (21), respectively. $(\varphi - by)$ here is $(\Phi + \theta)$ in Nordsieck's notation. In these diagrams, the curves numbered from 1 to 24 correspond to the 24 electrons used in the calculation with each curve for one electron. Only odd numbered electrons are presented to avoid possible confusion arisen from too many lines. The reciprocal of the slope of the curve as

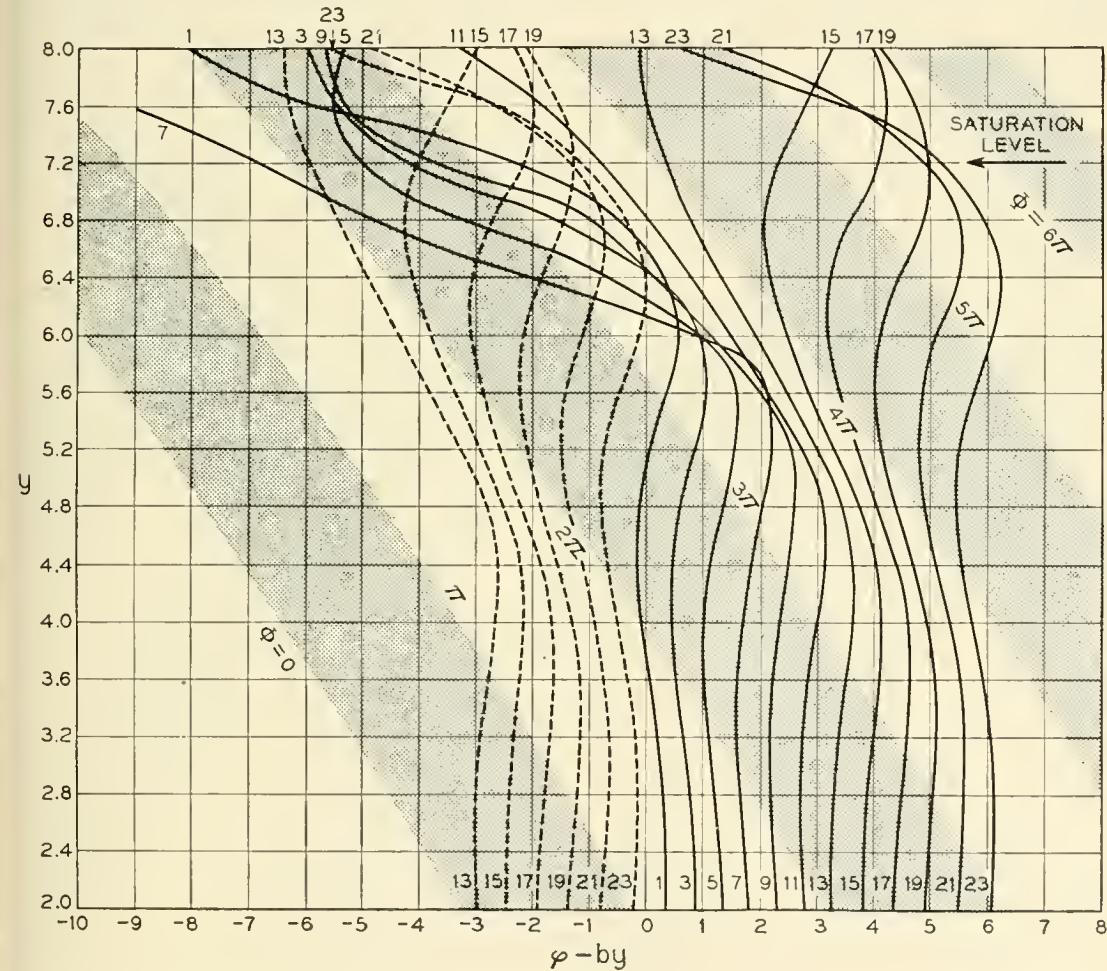


Fig. 8(b) — y versus $\varphi - by$ for $QC = 0.2$, $k = 2.5$, b for $\mu_1 = 0.67\mu_1(\text{max})$ and $C = 0.1$ (Case 12).

given by (18) is proportional to the ac displacement of electron per unit of y . (In small- C theory it is proportional to the ac velocity of the electron.) Concentration of curves is obviously proportional to the charge-density distribution of the beam. In the shaded regions, the axially directed electric field of the circuit is negative and thus accelerates electrons in the positive z direction. Electrons are decelerated in the unshaded regions where the circuit field is positive. The boundaries of these regions are constant phase contours of the circuit wave. (They are constant Φ contours in Nordsieck's notation.)

These figures are actually the "space-time" diagrams which unfold the history of every electron from the input to the output ends. The effect of C can be clearly seen by comparing Figs. 8(a), (b) and (c). These diagrams are plotted for $QC = 0.2$, $k = 2.5$, b for $\mu_1 = 0.67\mu_1(\text{max})$ and for Fig. 8(a), $C = \text{small}$, for Fig. 8(b), $C = 0.1$, and for Fig. 8(c), $C = .15$. It may be seen that because of the velocity spread of the electrons, the saturation level in Fig. 8(a) is 9.3 whereas in Figs. 8(b) and (c), it is 7.2 and 7.0, respectively. It is therefore not surprising that $\text{Eff./}C$ decreases as C increases.

The effects of b and QC may be observed by comparing Figs. 8(d) and (b), and Figs. 8(b) and (e), respectively. The details will not be described here. It is however suggested to study these diagrams with those given in the small- C theory.

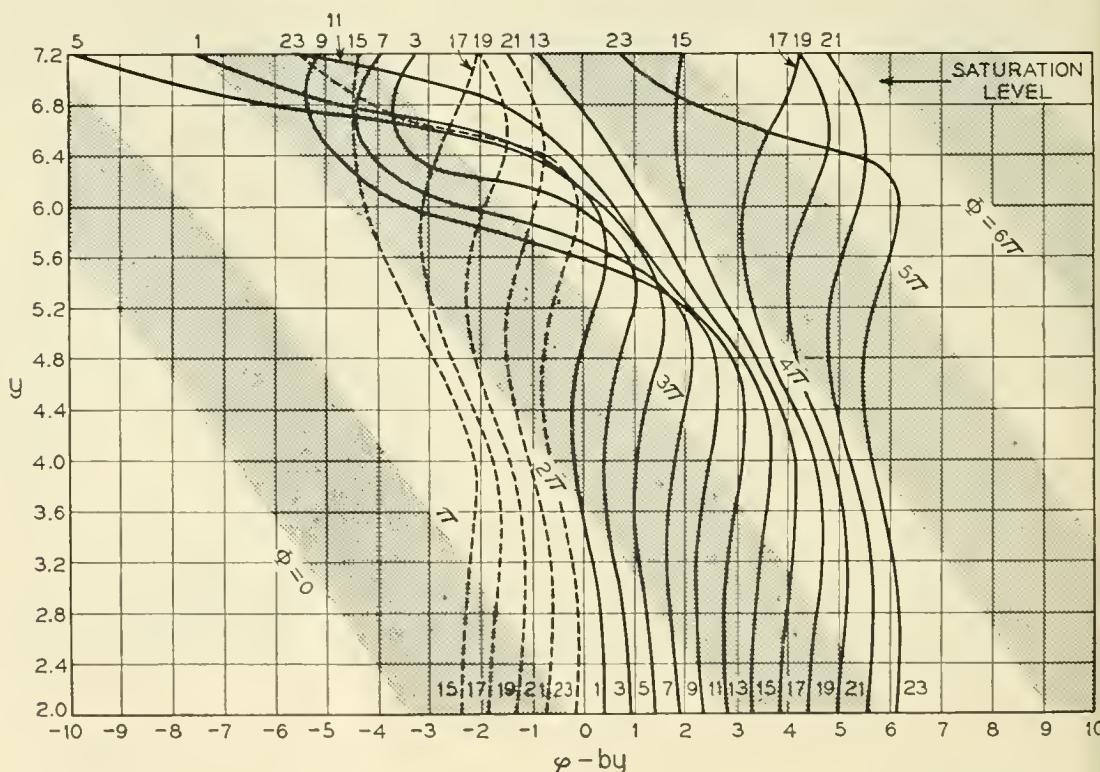


Fig. 8(c) — y versus $\varphi - by$ for $QC = 0.2$, $k = 2.5$, b for $\mu_1 = 0.67\mu_1(\text{max})$ and $C = 0.15$ (Case 16).

11. A QUALITATIVE PICTURE AND CONCLUSIONS

We have exhibited in the previous sections the most important nonlinear characteristics of the traveling wave amplifier. Numerical computations based on a model of 24 electrons have been carried out for more than twenty cases covering useful ranges of design and operating parameters. The results obtained for the saturation Eff./C may be summarized as follows:

- (1) It decreases with C particularly at large values of QC .
- (2) For $C = 0.1$, it varies roughly from 3.7 for $QC = 0.1$ to 2.3 for $QC = 0.4$, and only varies slightly with b .
- (3) For $C = 0.15$, it varies from 2.7 to 2.5 for QC from 0.1 to 0.2 and b corresponding to the maximum small-signal gain. It varies slightly with b for $QC = 0.2$.
- (4) It is almost constant between $k = 1.25$ and 2.50.

In order to understand the traveling-wave tube better, it is important to have a simplified qualitative picture of its operation. It is obvious that to obtain higher amplification, more electrons must travel in the region where the circuit field is positive, that is, in the region where electrons

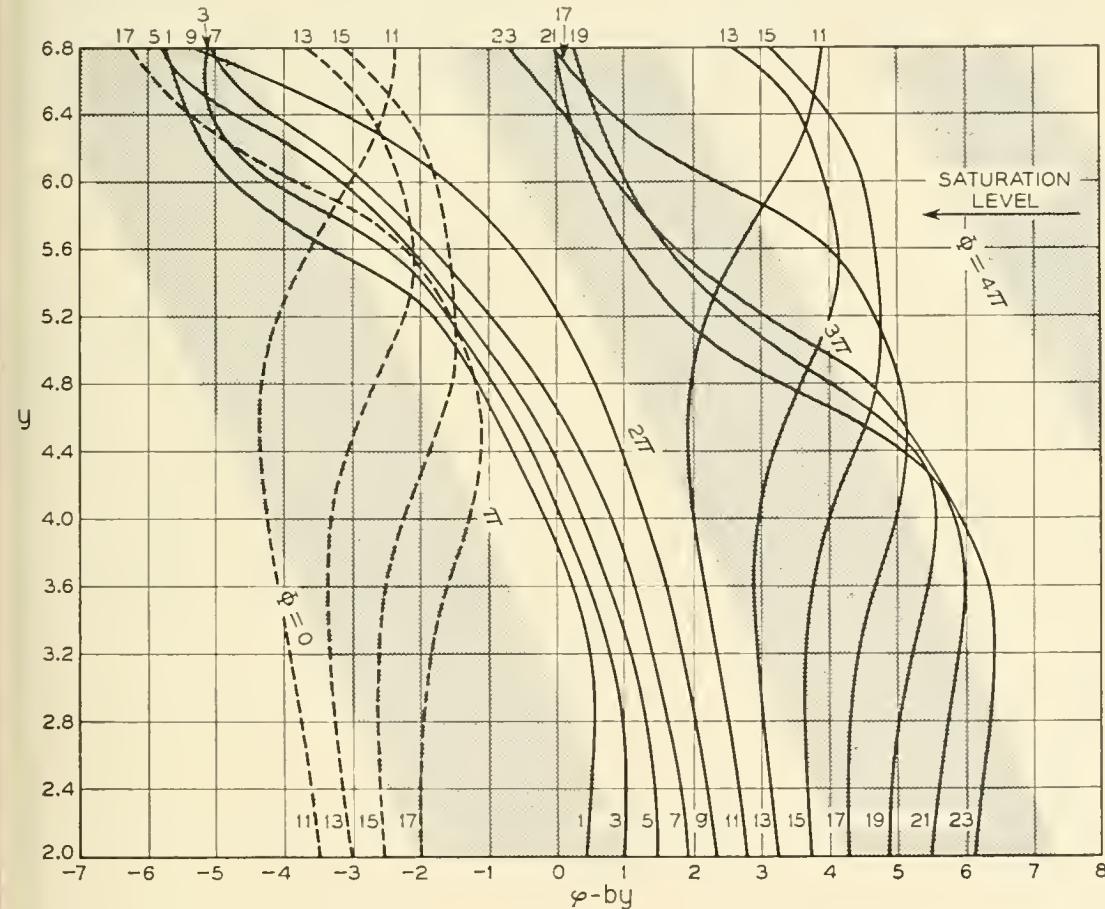


Fig. 8(d) — y versus $\varphi - by$ for $QC = 0.2$, $k = 2.5$, b for $\mu_1 = \mu_1(\max)$ and $C = 0.1$ (Case 10).

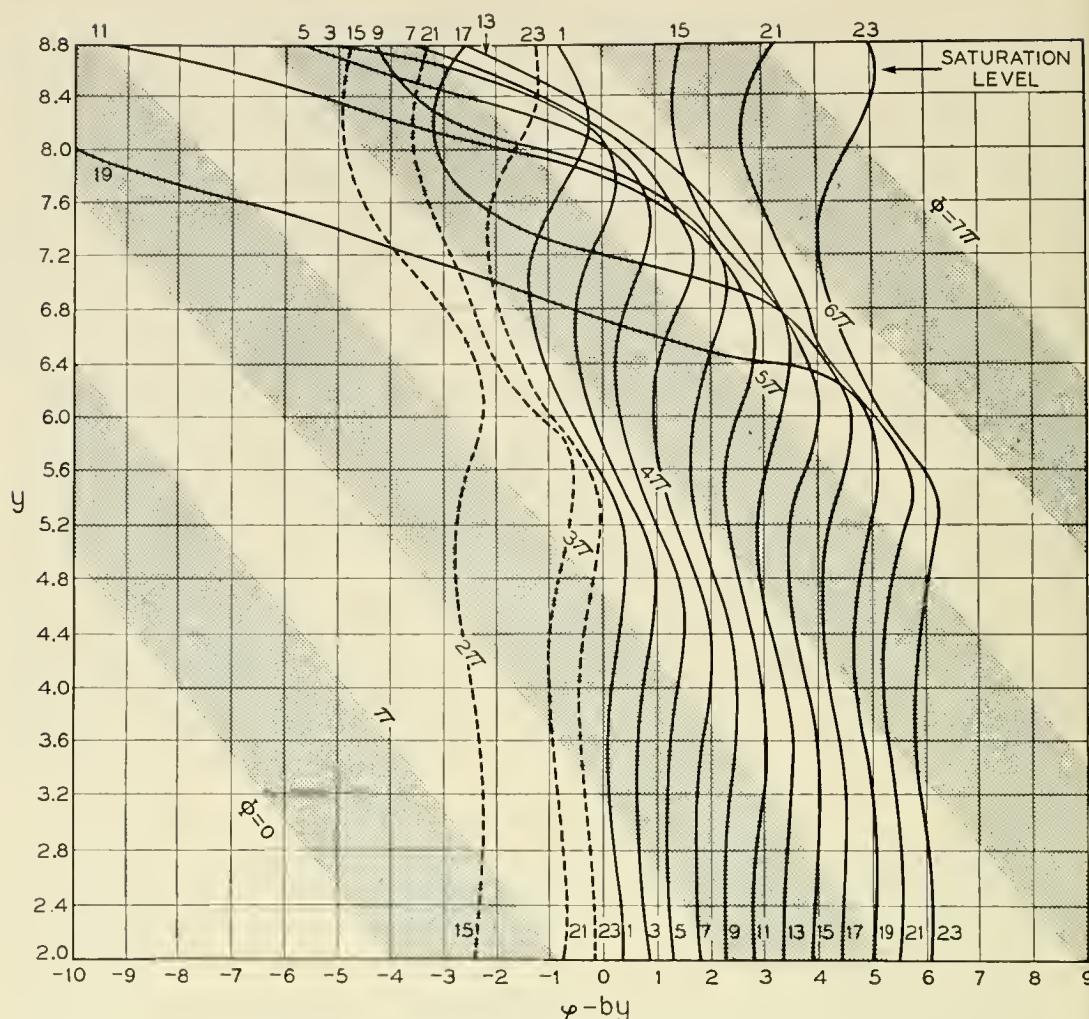


Fig. 8(e) — y versus $\varphi - by$ for $QC = 0.4$, $k = 2.5$, b for $\mu_1 = 0.67\mu_1(\text{max})$ and $C = 0.1$ (Case 21).

are decelerated by the circuit field. At the input end of the tube, electrons are uniformly distributed both in the accelerating and decelerating field regions. Bunching takes place when the accelerated electrons push forward and the decelerated ones press backward. The center of a bunch of electrons is located well inside the decelerating field region because the circuit wave travels slower than the electrons on the average (b is positive). The effectiveness of the amplification, or more specifically the saturation efficiency, therefore depends on (1), how tight the bunching is, and (2), how long a bunch travels inside the decelerating field region before its center crosses the boundary between the accelerating and decelerating fields.

For small- C , the ac velocities of the electrons are small compared with the dc velocity. The electron bunch stays longer with the decelerating circuit field before reaching the saturation level when b or QC is larger. On the other hand, the space charge force, or large QC or k tends to distort the bunching. As the consequence, the saturation efficiency increases with b , and decreases as k or QC increases. When C becomes finite how-

ever, the ac velocities of the electrons are no longer small as compared with their average speed. The velocity spread of the electrons becomes an important factor in determining the efficiency. Its effect is to loosen the bunching, and consequently it lowers the saturation level and reduces the limiting efficiency. It is seen from Figs. 5 and 6 that the velocity spread increases sharply with C and also steadily with b and QC . This explains the fact that in the present calculation the saturation Eff./ C decreases with C and is almost constant with b whereas in the small- C theory it is constant with C and increases steadily with b .

12. ACKNOWLEDGEMENTS

The writer wishes to thank J. R. Pierce for his guidance during the course of this research, and L. R. Walker for many interesting discussions concerning the working equations and the method of calculating the backward wave. The writer is particularly grateful to Miss D. C. Leagus who, under the guidance of V. M. Wolontis, has carried out the numerical work presented with endless effort and enthusiasm.

APPENDIX

The initial conditions at $y = 0$ are computed from Pierce's linearized theory. For small-signal, we have

$$a_1(y) = 4A(y) \cos(b + \mu_2)y \quad (\text{A-1})$$

$$a_2(y) = -4A(y) \sin(b + \mu_2)y \quad (\text{A-2})$$

$$A(y) = \epsilon e^{\mu_1 y} \quad (\text{A-3})$$

Here ϵ is taken equal to 0.03, a value which has been used in Tien-Walker-Wolontis' paper. Define

$$\frac{\partial X}{\partial y} = w(y, \varphi_0) \quad (\text{A-4})$$

$$X = pe^{-j\varphi_0} + p^*e^{j\varphi_0} \quad (\text{A-5})$$

where p^* is the conjugate of p . After substituting (A-1) to (A-5) into the working equations (15) to (18) and carrying out considerable algebraic work, we obtain exactly Pierce's equation.¹⁶

$$\mu^2 = \frac{(1 + jC\mu)(1 + bC)}{(j - \frac{1}{2}C\mu + j\frac{1}{2}bC)(\mu + jb)} - 4QC(1 + jC\mu)^2 \quad (\text{A-6})$$

provided that

$$-\left(\frac{\omega_p}{\omega C}\right)^2 \int_{-0}^{+\infty} e^{-k[\varphi(y, \varphi_0 + \phi) - \varphi(y, \varphi_0)] [1 + Cw(y, \varphi_0 + \phi)]} \cdot d\phi \operatorname{sgn}(\varphi(y, \varphi_0 + \phi) - \varphi(y, \varphi_0)) = 8\epsilon QC \quad (\text{A-7})$$

$$\cdot |(1 + jC\mu)(\mu + jb)| e^{\mu_1 y} \cos(\arg[(1 + jC\mu)(\mu + jb)] + \mu_2 y - \varphi_0)$$

Here $\mu = \mu_1 + j\mu_2$ or Pierce's $x_1 + jy_1$. From (A-7) the value of ω_p is determined for a given QC . The ac velocities of the electrons are derived from (A-4), such as,

$$w(y, \varphi_0)$$

$$= -2\epsilon \left| \mu \frac{\mu + jb}{1 + jc\mu} \right| e^{\mu_1 y} \cos \left(\arg \left[\mu \left(\frac{\mu + jb}{1 + jc\mu} \right) \right] + \mu_2 y - \varphi_0 \right) \quad (\text{A-8})$$

(A-1), (A-2), (A-7) and (A-8) are the expressions used to calculate the initial conditions at $y = 0$, when μ_1 and μ_2 are solved from Pierce's equation (A-6).

From (12c), the particular solution of the backward wave at small-signal is found to be

$$B(z, t) = -2\epsilon \left| \frac{-2jC(1 + jc\mu)(\mu + jb)}{2j - c\mu + icb} \right| e^{\mu_1 y} \cos \left[\arg \left[\frac{-2jC(1 + jc\mu)(\mu + jb)}{2j - c\mu + icb} \right] + \mu_2 y - \varphi_0 \right]$$

which agrees with Pierce's analysis.¹⁷

REFERENCES

1. J. R. Pierce, Traveling-Wave Tubes, D. Van Nostrand Co., N.Y., 1950, p. 160.
2. R. L. Hess, Some Results in the Large-Signal Analysis of Traveling-Wave Tubes, Technical Report Series No. 60, Issue No. 131, Electronic Research Laboratory, University of California, Berkeley, California.
3. C. K. Birdsall, unpublished work.
4. J. J. Caldwell, unpublished work.
5. P. Parzen, Nonlinear Effects in Traveling-Wave Amplifiers, TR/AF-4, Radiation Laboratory, The Johns Hopkins University, April 27, 1954.
6. A. Kiel and P. Parzen, Non-linear Wave Propagation in Traveling-Wave Amplifiers, TR/AF-13, Radiation Laboratory, The Johns Hopkins University, March, 1955.
7. A. Nordsieck, Theory of the Large-Signal Behavior of Traveling-Wave Amplifiers, Proc. I.R.E., **41**, pp. 630-637, May, 1953.
8. H. C. Poulter, Large Signal Theory of the Traveling-Wave Tube, Tech. Report No. 73, Electronics Research Laboratory, Stanford University, California, Jan., 1954.
9. P. K. Tien, L. R. Walker and V. M. Wolontis, A Large Signal Theory of Traveling-Wave Amplifiers, Proc. I.R.E., **43**, pp. 260-277 March, 1955.
10. J. E. Rowe, A Large Signal Analysis of the Traveling-Wave Amplifier, Tech. Report No. 19, Electron Tube Laboratory, University of Michigan, Ann Arbor, April, 1955.
11. P. K. Tien and L. R. Walker, Correspondence Section, Proc. I.R.E., **43**, p. 1007, Aug., 1955.
12. Nordsieck, op. cit., equation (1).
13. L. Brillouin, The Traveling-Wave Tube (Discussion of Waves for Large Amplitudes), J. Appl. Phys., **20**, p. 1197, Dec., 1949.
14. Pierce, op. cit., p. 9.
15. Nordsieck, op. cit., equation (4).
16. Pierce, op. cit., equation (7.13).
17. J. R. Pierce, Theory of Traveling-Wave Tube, Appendix A, Proc. I.R.E. **35**, p. 121, Feb., 1947.

A Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type

By W. E. DANIELSON, J. L. ROSENFELD,* and J. A. SALOOM

(Manuscript received November 10, 1955)

The theory of Cutler and Hines is extended in this paper to permit an analysis of beam-spreading in electron guns of high convergence. A lens correction for the finite size of the anode aperture is also included. The Cutler and Hines theory was not applicable to cases where the effects of thermal velocities are large compared with those of space charge and it did not include a lens correction. Gun design charts are presented which include all of these effects. These charts may be conveniently used in choosing design parameters to produce a prescribed beam.

CONTENTS

1. Introduction	377
2. Present Status of Gun Design; Limitations	378
3. Treatment of the Anode Lens Problem	379
A. Superposition Approach	379
B. Use of a False Cathode	382
C. Calculation of Anode Lens Strength by the Two Methods	383
4. Treatment of Beam Spreading, Including the Effect of Thermal Electrons	388
A. The Gun Region	388
B. The Drift Region	392
5. Numerical Data for Electron Gun and Beam Design	402
A. Choice of Variables	402
B. Tabular Data	402
C. Graphical Data, Including Design Charts and Beam Profiles	402
D. Examples of Gun Design Using Design Charts	403
6. Comparison of Theory with Experiment	413
A. Measurement of Current Densities in the Beam	413
B. Comparison of the Experimentally Measured Spreading of a Beam with that Predicted Theoretically	416
C. Comparison of Experimental and Theoretical Current Density Distributions where the Minimum Beam Diameter is Reached	418
D. Variation of Beam Profile with Γ	418
7. Some Additional Remarks on Gun Design	418

* Mr. Rosenfeld participated in this work while on assignment to the Laboratories as part of the M.I.T. Cooperative Program.

GLOSSARY OF SYMBOLS

$A_1, 2$	anode designations
B, C	anode potentials
$C_1, 2$	functions used in evaluating σ_{+}'
dA	increment of area
dl, dz	increments of length
e	electronic charge, base of natural logarithms
E_n	electric field normal to electron path
F	modified focal length of the anode lens
F_D	focal length of the anode lens as given by Davisson ⁴
F_n	force acting normal to an electronic path
$F_{r, \sigma}$	fraction of the total current which would flow through a circle of radius r, σ
I, I_D	total beam current
I_r	beam current within a radius, r , of the center
J	current density
k	Boltzman's constant
K	a quantity proportional to gun permeance
m	electronic mass
P	gun permeance
$P(r)$	probability that a thermal electron has a radial posi- tion between r and $r + dr$
r	radial distance from beam axis
r_a, c	anode, cathode radii
r_e	distance from beam axis to path of an electron emitted with zero velocity at the edge of the cathode
r_{95}	radius of circle through which 95% of the beam cur- rent would pass
\bar{r}	distance from center of curvature of cathode; hence, \bar{r}_c is the cathode radius of curvature and $(\bar{r}_c - \bar{r}_a)$ is the distance from cathode to anode
r_{e+}'	slope of edge nonthermal electron path on drift side of anode lens
r_{e-}'	slope of edge nonthermal electron path on gun side of anode lens
R	a dummy integration variable
t	time
T	cathode temperature in degrees K
u	longitudinal electron velocity
$v_{c, x, y}$	transverse electron velocities
$V, V_{a, f, z}$	beam voltages with cathode taken as ground

$V(\bar{r}, r)$, $V_c(\bar{r}, r)$, potential distributions used in the anode lens study

etc.

V'	voltage gradient
z	distance along the beam from the anode lens
z_{\min}	distance to the point where r_{95} is a minimum
$(-\alpha)$	Langmuir potential parameter for spherical cathode-anode gun geometry
γ	slope of an electron's path after coming into a space charge free region just beyond the anode lens
Γ	the factor which divides F_D to give the modified anode focal length
δ	dimensionless radius parameter
ϵ_0	dielectric constant of free space
ζ	dimensionless voltage parameter
θ	slope of an electron's path in the gun region
η	charge to mass ratio for the electron
μ	normalized radial position in a beam
σ	the radial position of an electron which left the cathode center with "normal" transverse velocity
σ_+'	slope of σ -electron on drift side of anode lens
σ_-'	slope of σ -electron on gun side of anode lens
ψ	electric flux

1. INTRODUCTION

During the past few years there have been several additions to the family of microwave tubes requiring long electron beams of small diameter and high current density. Due to the limited electron current which can be drawn from unit area of a cathode surface with some assurance of long cathode operating life, high density electron beams have been produced largely through the use of convergent electron guns which increase markedly the current density in the beam over that at the cathode surface.

An elegant approach to the design of convergent electron guns was provided by J. R. Pierce¹ in 1940. Electron guns designed by this method are known as *Pierce guns* and have found extensive use in the production of long, high density beams for microwave tubes.

More recent studies, reviewed in Section 2, have led to a better understanding of the influence on the electron beam of (a) the finite velocities with which electrons are emitted from the cathode surface, and (b) the defocusing electric fields associated with the transition from the accelerating region of the gun to the drift region beyond. Although these two effects have heretofore been treated separately, it is in many cases

necessary to produce electron beams under circumstances where both effects are important and so must be dealt with simultaneously and more precisely than has until now been possible. It is the purpose of this paper to provide a simple design procedure for typical Pierce guns which includes both effects. Satisfactory agreement has been obtained between measured beam contours and those predicted for several guns having perveances (i.e., ratios of beam current to the $\frac{3}{2}$ power of the anode voltage) from 0.07×10^{-6} to 0.7×10^{-6} amp (volt) $^{-3/2}$.

2. PRESENT STATUS OF GUN DESIGN — LIMITATIONS

Gun design techniques of the type originally suggested by J. R. Pierce were enlarged in papers by Samuel² and by Field³ in 1945 and 1946. Samuel's work did not consider the effect of thermal velocities on beam shape and, although Field pointed out the importance of thermal velocities in limiting the theoretically attainable current density, no method for predicting beam size and shape by including thermal effects was suggested. The problem of the divergent effect of the anode lens was treated in terms of the Davisson⁴ electrostatic lens formula, and no corrections were applied.*

More recently, Cutler and Hines⁶ and also Cutler and Saloom⁷ have presented theoretical and experimental work which shows the pronounced effects of the thermal velocity distribution on the size and shape of beams produced by Pierce guns. Cutler and Saloom also point to the critical role of the beam-forming electrode in minimizing beam distortion due to improper fields in the region where the cathode and the beam-forming electrode would ideally meet. With regard to the anode lens effect, these authors also show experimental data which strongly suggest a more divergent lens than given by the Davisson formula. The Hines and Cutler thermal velocity calculations have been used^{6, 7} to predict departures in current density from that which should prevail in ideal beams where thermal electrons are absent. Their theory is limited, however, by the assumption that the beam-spreading caused by thermal velocities is small compared to the nominal beam size.

In reviewing the various successes of the above mentioned papers in affording valuable tools for electron beam design, it appeared to the present authors that significant improvement could be made, in two respects, by extensions of existing theories. First, a more thorough in-

* It is in fact erroneously stated in Reference 5 that the lens action of an actual structure must be somewhat weaker than predicted by the Davisson formula so that the beam on leaving the anode hole is more convergent than would be calculated by the Davisson method. This question is discussed further in Section 3.

vestigation of the anode lens effect was called for; and second, there was a need to extend thermal velocity calculations to include cases where the percentage increase in beam size due to thermal electrons was as large as 100 per cent or 200 per cent. Some suggestions toward meeting this second need have been included in a paper by M. E. Hines.⁸ They have been applied to two-dimensional beams by R. L. Schrag.⁹ The particular assumptions and methods of the present paper as applied to the two needs cited above are somewhat different from those of References 8 and 9, and are fully treated in the sections which follow.

3. TREATMENT OF THE ANODE LENS PROBLEM

Using thermal velocity calculations of the type made in Reference 6, it can easily be shown that at the anode plane of a typical moderate permeance Pierce type electron gun, the average spread in radial position of those electrons which originate from the same point of the cathode is several times smaller than the beam diameter. For guns of this type, then, we may look for the effect of the anode aperture on an electron beam for the idealized case in which thermal velocities are absent and confidently apply the correction to the anode lens formula so obtained to the case of a real beam.

Several authors have been concerned with the diverging effect of a hole in an accelerating electrode where the field drops to zero in the space beyond,¹⁰ but these treatments do not include space charge effects except as given by the Davisson formula for the focal length, F_D , of the lens:

$$F_D = -\frac{4V}{V'} \quad (1)$$

where V' would be the magnitude of the electric field at the aperture if it were gridded, and V would be the voltage there.

In attempting to describe the effect of the anode hole with more accuracy than (1) affords, we have combined analytical methods with electrolytic tank measurements in two rather different ways. The first method to be given is more rigorous than the second, but a modification of the second method is much easier to use and gives essentially the same result.

A. Superposition Approach to the Anode Lens Problem

Special techniques are required for finding electron trajectories in a space charge limited Pierce gun having a non-gridded anode. M. E.

Hines has suggested* that a fairly accurate description of the potential distribution in such guns can be obtained by a superposition method as follows:

By the usual tank methods, find suitable beam forming electrode and anode shapes for conical space charge limited flow in a diode having cathode and anode radii of curvature given by \bar{r}_c and \bar{r}_{a1} , respectively, as shown in Fig. 1(a). Using the electrolytic tank with an insulator along the line which represents the beam edge, trace out an equipotential which intersects the insulator at a distance \bar{r}_{a2} from the cathode center of curvature. Let the cathode be at ground potential and let the voltage on anode A_1 be called B . Suppose, now, that we are interested in electron trajectories in a non-gridded gun where the edge of the anode hole is a distance \bar{r}_{a2} from the center of curvature of the cathode. Let the voltage, C , for this anode be chosen the same as the value of the equipotential traced out above for the case of cathode at ground potential and A_1 at potential B . If we consider the space charge limited flow from a cathode which is followed by the apertured anode, A_2 , and the full anode, A_1 , at potentials C and B , respectively, it is clear that a conical flow of the type which would exist between concentric spheres will result. The flow for such cases was treated by Langmuir,¹⁴ and the associated potentials are commonly called the "Langmuir potentials."

If we operate both A_1 and A_2 at potential C , however, the electrons will pass through the aperture in anode A_2 into a nearly field-free region. If the distance, $\bar{r}_{a2} - \bar{r}_{a1}$, from A_2 to A_1 is greater than the diameter of the aperture in A_2 , the flow will depend very little on the shape of A_1 and the electron trajectories and associated equipotentials will be of the type we wish to consider except in a small region near A_1 . We will shortly make use of the fact that the space charge between cathode and A_2 is not changed much when the voltage on A_1 is changed from B to C , but first we will define a set of potential functions which will be needed.

In order to obtain the potential at arbitrary points in any axially symmetric gun when space charge is not neglected, we may superpose potential solutions to 3 separate problems where, in each case, the boundary condition that each electrode be an equipotential is satisfied. We will follow the usual notation in using \bar{r} for the distance of a general point from the cathode center of curvature, and r for its radial distance from the axis of symmetry. Let $V_a(\bar{r}, r)$, $V_b(\bar{r}, r)$ and $V_{sc}(\bar{r}, r)$ be the three potential solutions where: (1) $V_a(\bar{r}, r)$ is the solution for the case of no space charge with A_1 and cathode at zero potential and A_2 at potential C , (2) $V_b(\bar{r}, r)$ is the solution for the case of no space charge with A_2

* Verbal disclosure.

and cathode at zero potential and A_1 at potential B , and (3) $V_{sc}(\bar{r}, r)$ is the solution when space charge is present but when A_1 , A_2 , and cathode are all grounded.

If the configuration of charge which contributes to $V_{sc}(\bar{r}, r)$ is that corresponding to ideal Pierce type flow, then we can use the principle of superposition to give the Langmuir potential, $V_L(\bar{r}, r)$:

$$V_L(\bar{r}, r) = V_a(\bar{r}, r) + V_b(\bar{r}, r) + V_{sc}(\bar{r}, r) \quad (2)$$

Furthermore, the potential configuration for the case where A_1 and A_2 are at potential C can be written

$$V = V_a + \frac{C}{B} V_b + V_{(sc)}, \quad (3)$$

where the functional notation has been dropped and $V_{(sc)}$ is the potential due to the new space charge when A_1 and A_2 are grounded. We are now ready to use the fact that $V_{(sc)}$ may be well approximated by V_{sc} which is easily obtained from (2). This substitution may be justified by noting that the space charge distribution in a gun using a voltage C for A_1 does not differ significantly from the corresponding distribution when A_1 is at voltage B except in the region near and beyond A_2 where the charge density is small anyway (because of the high electron velocities there). Substituting V_{sc} as given by (2) for $V_{(sc)}$ in (3) then gives

$$V \approx V_L - \left(1 - \frac{C}{B}\right) V_b \quad (4)$$

We have thus obtained an expression, (4), for the potential at an arbi-

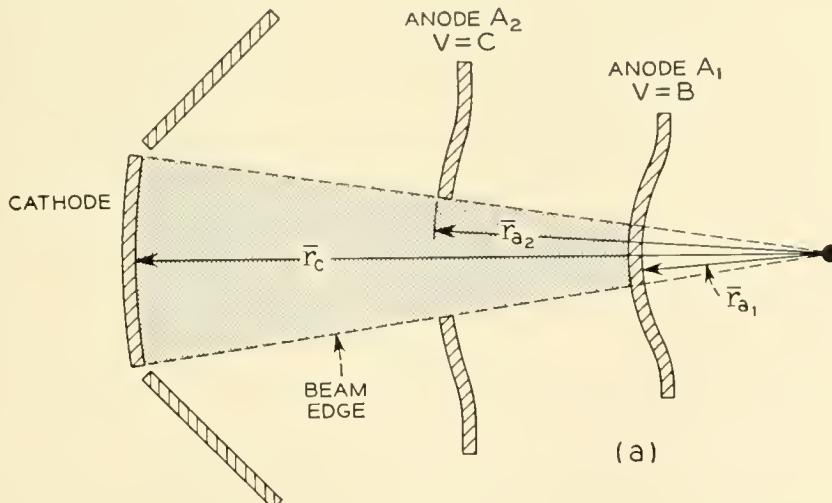


Fig. 1(a) — Electrode configuration for anode lens evaluation in Section 3A.

trary point in our gun in terms of the well known solution for space charge limited flow between two concentric spheres, V_L , and a potential distribution, V_b , which does not depend on space charge and can therefore be obtained in the electrolytic tank. Once the potential distribution is found, electron trajectories may be calculated, and an equivalent lens system found. Equation (4) is used in this way in Part C as one basis for estimating a correction to the Davisson equation. (It will be noted that (4) predicts a small but finite negative field at the cathode. This is because the space charge density associated with V_{sc} is slightly greater near the cathode than that associated with $V_{(sc)}$, and it is this latter space charge which will make the field zero at the cathode under real space charge limited operation. Equation (4), as applied in Part C of this section, is used to give the voltage as a function of position at all points except near the cathode where the voltage curves are extended smoothly to make the field at the cathode vanish.)

B. Use of a False Cathode in Treating the Anode Lens Problem

Before evaluating the lens effect by use of (4), it will be useful to develop another approach which is a little simpler. The evaluation of the lens effect predicted by both methods will then be pursued in Part C where the separate results are compared.

In Part A we noted that no serious error is made in neglecting the difference between the two space charge configurations considered there because these differences were mainly in the very low space charge region near and beyond A_2 . It similarly follows that we can, with only a small decrease in accuracy, ignore the space charge in the region near and beyond A_2 so long as we properly account for the effect of the high space charge regions closer to the cathode. To place the foregoing observations on a more quantitative basis, we may graph the Langmuir potential (for space charge limited flow between concentric spheres) versus the distance from cathode toward anode, and then superpose a plot of the potential from LaPlace's equation (concentric spheres; no space charge) which will have the same value and slope at the anode. The LaPlace curve will depart significantly from the Langmuir in the region of the cathode, but will adequately represent it farther out.¹¹ Our experience has shown that the representation is "adequate" until the difference between the two potentials exceeds about 2 per cent of the anode voltage. Then, since space charge is not important in the region near the anode for the case of a gridded Pierce gun, corresponding to space charge limited flow between concentric spheres, it can be expected to be similarly unimportant for cases where the grid is replaced by an aperture. Let us

therefore consider a case where electrons are emitted perpendicularly and with finite velocity from what would be an appropriate spherical equipotential between cathode and anode in a Pierce type gun. So long as (a) there is good agreement between the LaPlacee and Langmuir curves at this artificial cathode and (b) the distance from this artificial cathode to the anode hole is somewhat greater than the hole diameter, we will find that the divergent effect of the anode hole will be very nearly the same in this conected space charge free case as in the actual case where space charge is present. (The quantitative support for this last statement comes largely from the agreement between calculations based on this method and calculations by method A.) The electrode configuration is shown in Fig. 1(b), and the potential distribution in this space charge free anode region can now be easily obtained in the electrolytic tank. This potential distribution will be used in the next section to provide a second basis for estimating a correction to the Davisson equation.

C. Calculation of Anode Lens Strength by the Two Methods

The Davisson equation, (1), may be derived by assuming that none of the electric field lines which originate on charges in the cathode-anode region leave the beam before reaching the ideal anode plane where the voltage is V , and that all of these field lines leave the beam symmetrically and radially in the immediate neighborhood of the anode. Electrons are thus considered to travel in a straight line from cathode to anode, and then to receive a sudden radial impulse as they cross radially diverging electric field lines at the anode plane. A discontinuous change in

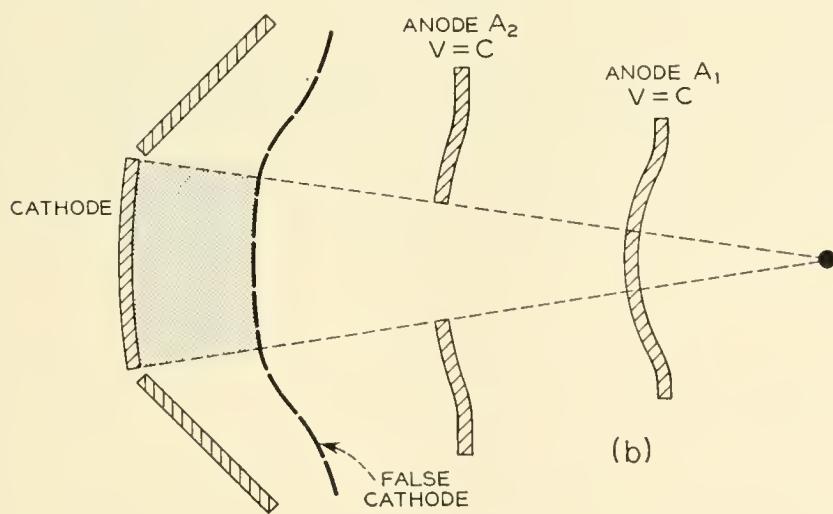


Fig. 1(b) — The introduction of a false cathode at the appropriate potential allows the effect of space charge on the potential near the anode hole to be satisfactorily approximated as discussed in Section 3B.

slope is therefore produced as is common to all thin lens approximations. The diverging effect of electric field lines which originate on charges which have passed the anode plane is then normally accounted for by the universal beam spread curve.¹² In our attempt to evaluate the lens effect more accurately, we will still depend upon using the universal beam spread curve in the region following the lens and on treating the equivalent anode lens as thin. Consequently our improved accuracy must come from a mathematical treatment which allows the electric field lines originating in the cathode-anode region to leave the beam gradually, rather than a treatment where all of these flux lines leave the beam at the anode plane. In practice the measured perveances, $P (= I/V^{3/2})$, of active guns of the type considered here have averaged within 1 or 2 per cent of those predicted for corresponding gridded Pierce guns. Therefore the total space charge between cathode and anode is much the same with and without the use of a grid, even though the charge distribution is not the same in the two cases. The total flux which must leave our beam is therefore the same as that which will leave the corresponding idealized beam and we may write

$$\psi = \int E_n dA = \pi r_a^2 V_{\text{ideal}}' \quad (5)$$

where E_n is the electric field normal to the edge of the beam, $r_a = r_c(\bar{r}_a/\bar{r}_c)$ is the beam radius at the anode lens, and V_{ideal}' is the magnitude of the field at the corresponding gridded Pierce gun anode.

To find the appropriate thin lens focal length we will now find the total integrated transverse impulse which would be given to an electron which follows a straight-line path on both sides of the lens (see Fig. 2), and we will equate this impulse to $m\Delta u$ where Δu is the transverse velocity given to the electron as it passes through the equivalent thin lens. In this connection we will restrict our attention to paraxial electrons and evaluate the transverse electric fields from (4) and from the tank plot outlined in Section B, respectively. The total transverse impulse experienced by an electron can be written

$$\int_{\text{Path}} F_n dt = e \int_{\text{Path}} \frac{E_n}{u} dl \quad (6)$$

where u is the velocity along the path and F_n is the force normal to the path.

We will usually find that the correction to (1) is less than about 20 per cent. It will therefore be worthwhile to put (6) in a form which in effect allows us to calculate *deviations* from F_D as given by (1) instead

of deriving a completely new expression for F . In accomplishing this purpose, it will be helpful to define a dimensionless function of radius, δ , by

$$\frac{r_a}{r} = 1 + \delta, \quad (7a)$$

and a dimensionless function of voltage, ξ , by

$$\sqrt{\frac{V_x}{V}} = 1 + \xi, \quad (7b)$$

where r_a is the radius at the anode lens when the lens is considered thin, and V_x is a constant voltage to be specified later. (Note that the quantities δ and ξ are not necessarily small compared to 1.) Using $u = \sqrt{2\eta V}$, and substituting for \sqrt{V} from (7b) we obtain

$$e \int \frac{E_n dl}{u} = \frac{4}{r_a \sqrt{2\eta V_x}} \int E_n r (1 + \xi + \delta + \xi \delta) dl \quad (8)$$

where use has also been made of (7a) in the form $1 \equiv r(1 + \delta)/r_a$. Now, as outlined above, we equate this impulse to $m\Delta u$, and we obtain

$$\Delta u = \frac{e/m}{r_a \sqrt{2\eta V_x}} \left(\int E_n r dl + \int E_n r (\xi + \delta + \xi \delta) dl \right) \quad (9)$$

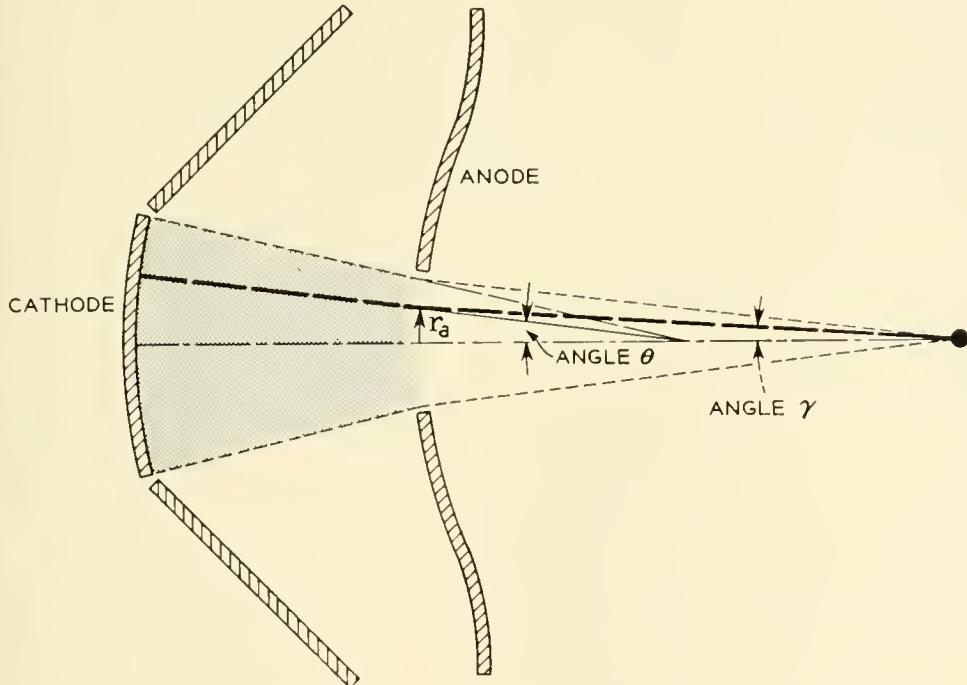


Fig. 2 — The heavy line represents an electron's path when the effect of the anode hole may be represented by a thin lens, and when space charge forces are absent in the region following the anode aperture. For paraxial electrons, the (negative) focal length is related to the indicated angles by ($\gamma = \theta + r_a/F$).

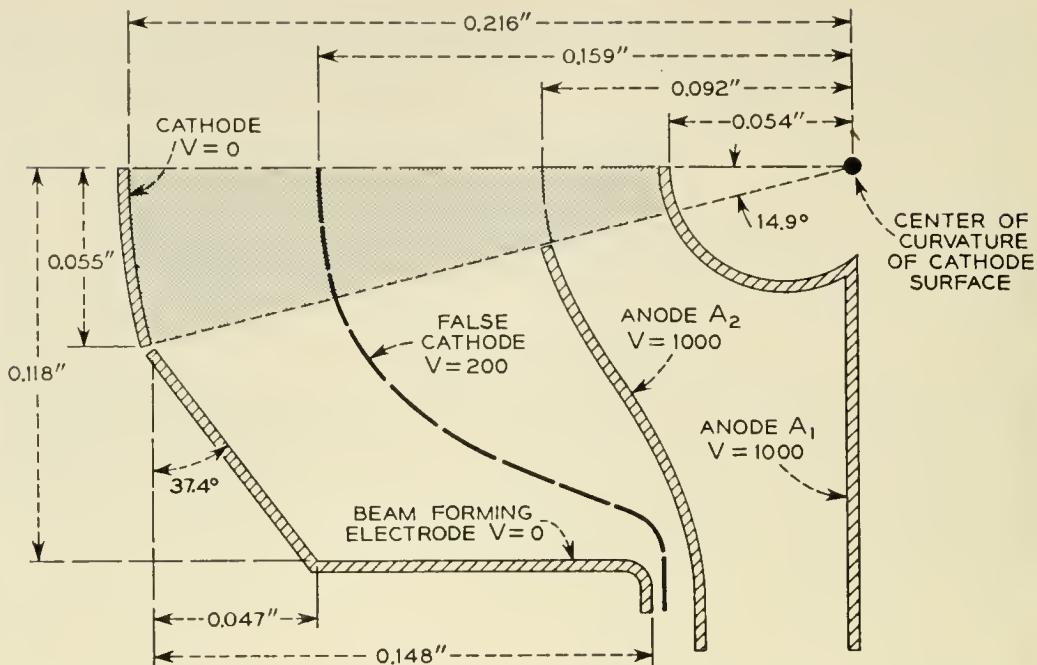


Fig. 3 — The gun parameters used in Section 3C for comparing two methods of evaluating the effect of the anode lens.

The first integral can be obtained from (5); hence, if we are able to choose V_x so that the second integral vanishes, we may write:

$$\Delta u = \frac{\eta}{r_a \sqrt{2\eta V_x}} \left(\frac{r_a^2 V_{\text{ideal}}'}{2} \right)$$

The reciprocal of the thin lens focal length is therefore

$$\frac{1}{F} = -\frac{\Delta u}{r_a u_f} = -\frac{V'}{4\sqrt{V_x V_f}} \quad (10)$$

where u_f and V_f are the final velocity and voltage of the electron after it leaves the lens region.

The real task, then, is to use the potential distribution in the gun as obtained by the methods of Part A or Part B above to find the value of V_x which causes the last integral in (9) to vanish: To compare the two focal lengths obtained by the methods of Part A and B respectively, a specific tank design of the type indicated in Fig. 1 was carried out. The relevant gun parameters are indicated in Fig. 3. Approximate voltages on and near the beam axis were obtained as indicated in Parts A and B, above, with the exception that in the superposition method, A, special techniques were used to subtract the effect of the space charge lying in the post-anode region (because the effect of this space charge is accounted for separately as a divergent force in the drift region*). From these data,

* See Section 4B.

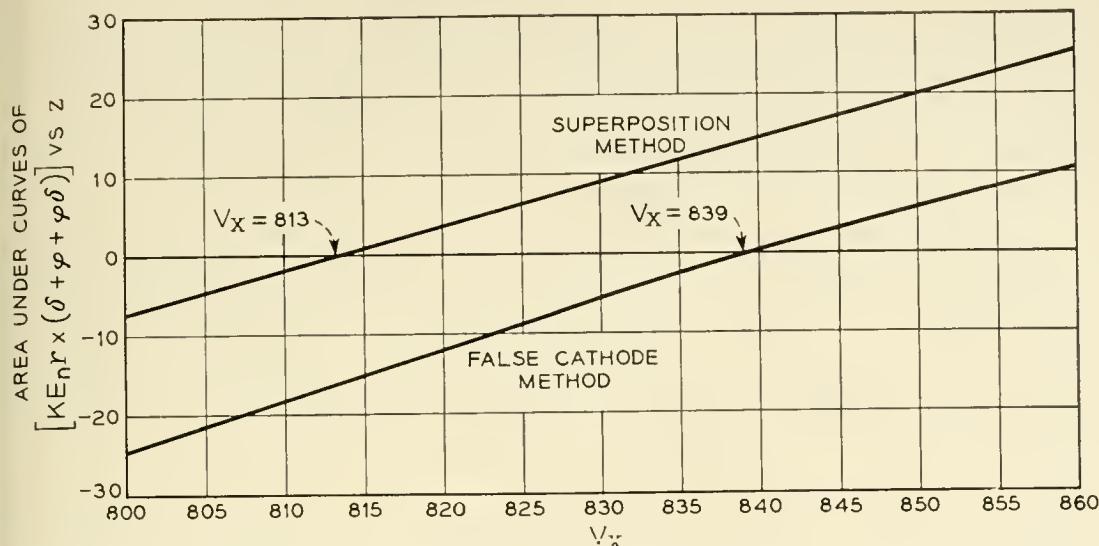


Fig. 4 — Curves for finding the value of V_x to be used in equation (10) for the set of gun parameters of Fig. 3.

both the direction and magnitude of the total electric field near the beam axis were (with much labor) determined. Once these data had been obtained, a trial value was selected for V_x , and the corresponding focal length was calculated by (10). This enabled the electron's path through the associated thin lens to be specified so that, at this point in the procedure, both r and V were known functions of ℓ , and the quantities δ and ξ were then obtained as functions of ℓ from (7). Finally the second integral in (9) was evaluated for the particular V_x chosen, and then the process was repeated for other values of V_x . Fig. 4 shows curves whose ordinates are proportional to this second integral and whose abscissae are trial values for V_x . As noted above, the appropriate value for V_x is that value which makes the ordinate vanish, so that we obtain $V_x = 813$ and 839 for methods A and B, respectively. The percentage difference in the focal lengths obtained by the two methods is thus only 1.6 per cent, and the reasonableness of making calculations as outlined in Part B is thus put on a more quantitative basis.

Even calculations based on the method of Part B are tedious, and we naturally look for simpler methods of estimating the lens effect. In this connection we have found that V_x is usually well approximated by the value of the potential at the point of intersection between the beam axis and the ideal anode sphere. The specific values of the potential at this point as obtained by the methods of Parts A and B were 814 and 827, respectively. It will be noted that these values agree remarkably well with the values obtained above. Furthermore, very little extra effort is required to obtain the potential at this intersection in the false cathode case:

Electrolytic tank measurements are normally made in the cathode-anode region to give the potential variation along the outside edge of the electron beam (for comparison with the Langmuir potential); hence, by tracing out a suitable equipotential line, the shape of the false cathode can easily be obtained. With the false cathode in place and at the proper potential, the approximate value for V_x is then obtained by a direct tank measurement of the potential at an axial point whose distance from the true cathode center is $(\bar{r}_c - \bar{r}_a)$ as outlined above. Although finite electron emission velocities typically do not much influence the trajectory of an electron at the anode, they do nevertheless significantly alter the beam in the region beyond. It is in this affected region where experimental data can be conveniently taken. We must, therefore, postpone a comparison of lens theory with experiment until the effect of thermal velocities has been treated. At that time theoretical predictions combining the effects of both thermal velocities and the anode lens can be made and compared with experiment. Such a comparison is made in Section 6.

4. TREATMENT OF BEAM SPREADING, INCLUDING THE EFFECT OF THERMAL ELECTRONS

In Section 2 the desirability of having an approach to the thermal spreading of a beam which would be applicable under a wide variety of conditions was stressed. In particular, there was a need to extend thermal velocity calculations to include the effects of thermal velocities even when electrons with high average transverse velocities perturb the beam size by as much as 100 or 200 per cent. Furthermore, a realistic mathematical description which would allow electrons to cross the axis seemed essential. The method described below is intended adequately to answer these requirements.

A. *The Gun Region*

The Hines-Cutler⁶ method of including the effect of thermal velocities on beam size and shape leads one to conclude that, for usual anode voltages and gun perveance, the beam density profile in the plane of the anode hole is not appreciably altered by thermal velocities of emission. (This statement will be verified and put on a more quantitative basis below.) Under these conditions, the beam at the anode is adequately described by the Hines-Cutler treatment. We will therefore find it convenient to adopt their notation where possible, and it will be worthwhile to review their approach to the thermal problem.

It is assumed that electrons are emitted from the cathode of a thermionic gun with a Maxwellian distribution of transverse velocities

$$dJ_c = J_c \frac{m}{2\pi kT} e^{-(m/2kT)(v_x^2 + v_y^2)} dv_x dv_y \quad (11)$$

where J_c is the cathode current density in the z direction, T is the cathode temperature, and v_x and v_y are transverse velocities. The number of electrons emitted per second with radially directed voltages between V and $V + dV$ is then

$$dJ_e = J_c e^{-(Ve/kT)} d\left(\frac{Ve}{kT}\right) \quad (12)$$

Now in the accelerating region of an ideal Pierce gun (and more generally in any beam exhibiting laminar flow and having constant current density over its cross section) the electric field component perpendicular to the axis of symmetry must vary linearly with radius. Consequently Hines and Cutler measure radial position in the electron beam as a fraction, μ , of the outer beam radius (r_e) at the same longitudinal position,

$$r = \mu r_e \quad (13)$$

The laminar flow assumption for constant current densities and small beam angles implies a radius of curvature for laminar electrons which also varies linearly with radius at any given cross section so that

$$\frac{d^2r}{dt^2} = \mu \frac{d^2r_e}{dt^2} \quad (14)$$

Substituting for r from (13), (14) becomes

$$\frac{d^2\mu}{dt^2} + \left(\frac{2}{r_e} \frac{dr_e}{dt} \right) \frac{d\mu}{dt} = 0 \quad (15)$$

where r_e and dr/dt can be easily obtained from the ideal Langmuir solution. Since the equation is linear in μ , we are assured that the radial position of a non-ideal electron that is emitted with finite transverse velocity from the cathode center (where $\mu = 0$) will, at any axial point, be proportional to $d\mu/dt$ at the cathode.

Let us now define a quantity " σ " such that $\mu = \sigma/r_e$ is the solution to (15) with the boundary conditions $\mu_c = 0$ and

$$\left(\frac{d\mu}{dt} \right)_c = \frac{1}{r_c} \sqrt{\frac{kT}{m}}$$

where the subscript c denotes evaluation at the cathode surface, k is

Boltzman's constant, T is the cathode temperature in degrees Kelvin, and m is mass of the electron. For the case $\mu_c = 0$, but with arbitrary initial transverse velocity, we will then have

$$\mu = \frac{\sigma}{r_e} \frac{1}{\sqrt{\frac{kT}{m}}} \left(\frac{d\mu}{dt} \right)_c \quad (16)$$

Hence we can express σ in terms of the thermal electron's radial position (r), and its initial transverse velocity, v_c ,

$$\sigma = \frac{\mu r_e \sqrt{\frac{kT}{m}}}{\left(\frac{d(\mu r_e)}{dt} \right)} \equiv \frac{r \sqrt{\frac{kT}{m}}}{v_c} \quad (17)$$

The quantity σ can now be related to the radial spread of thermal electrons (emitted from a given point on the cathode) with respect to an electron with no initial velocity: By (11) we see that the number of electrons leaving the cathode with $d\mu/dt = v_c/r_e$ is proportional to $v_c \exp -v_c^2 m/2kT$. Suppose many experiments were conducted where all electrons except one at the cathode center had zero emission velocity, and suppose the number of times the initial transverse velocity of the single thermal electron were chosen as v_c , is proportional to $v_c \exp -v_c^2 m/2kT$. Then the probability, $P(r)$, that the thermal electron would have a radial position between r and $r + dr$ when it arrived at the transverse plane of interest would be proportional to $v_c \exp -v_c^2(m/2kT)$. Here v_c is the proper transverse velocity to cause arrival at radius r , and by (17) we have

$$v_c = \frac{r}{\sigma} \sqrt{\frac{kT}{m}}$$

so that the probability becomes

$$P(r) = J_c e^{-(r^2/2\sigma^2)} d\left(\frac{r^2}{2\sigma^2}\right) \quad (18)$$

We therefore identify σ with the standard deviation in a normal or Gaussian distribution of points in two dimensions. At the real cathode, thermal electrons are simultaneously being emitted from the cathode surface with a range of transverse velocities. However, if σ as defined above is small in comparison with r_e , the forces experienced by a thermal electron when other thermal electrons are present will be very nearly

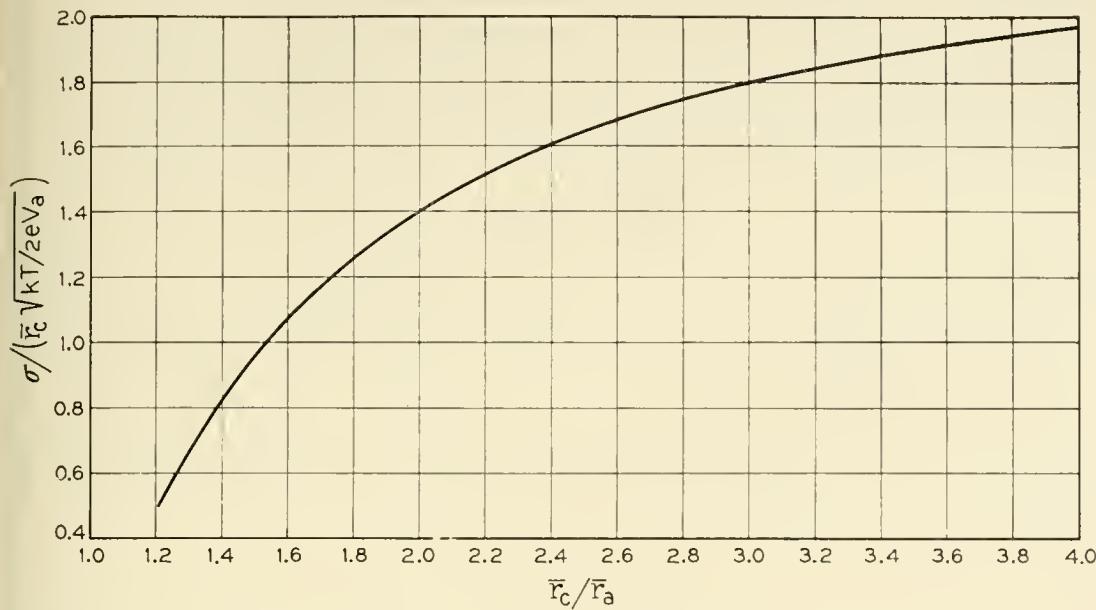


Fig. 5 — Curves useful in finding the transverse displacement of electron trajectories at the anode of Pierce-type guns.

the same as the forces involved in the equations above. Thus if $\sigma \ll r_e$, (18) may be taken as the distribution, in a transverse plane, of those electrons which were simultaneously emitted at the cathode center. Furthermore, the nature of the Pierce gun region is such that electrons emitted from any other point on the cathode will be similarly distributed with respect to the path of an electron emitted from this other point with zero transverse velocity (so long as they stay within the confines of the ideal beam). Hines and Cutler have integrated (15) with $\mu_c = 0$ and $(d\mu/dt)_c = 1$ to give $\sigma/(\bar{r}_c \sqrt{kT/2eV_a})$ at the anode as a function of \bar{r}_c/\bar{r}_a . This relationship is included here in graphical form as Fig. 5.

For a large class of magnetically shielded Pierce-type electron guns, including all that are now used in our traveling wave tubes, r_e/σ at the anode is indeed found to be greater than 5 (in most cases, greater than 10) so that evaluation of σ at the anode of such guns can be made with considerable accuracy by the methods outlined above. One source of error lies in the assumption that electrons which are emitted from a point at the cathode edge become normally distributed about the corresponding non-thermal (no transverse velocity of emission) electron's path, and with the same standard deviation as calculated for electrons from the cathode center. In the gun region where r_e/σ tends to be large this difference between representative σ -values for the peripheral and central parts of the beam is unimportant, but it must be re-examined in the drift region following the anode.

We have already investigated the region of the anode hole in some detail in Section 3 and have found it worth while to modify the ideal Davisson expression for focal length of an equivalent anode lens. In particular, let us define a quantity Γ by

$$F = \text{focal length} = F_D/\Gamma \quad (19)$$

where F_D is the Davisson focal length. Thus Γ represents a corrective factor to be applied to F_D to give a more accurate value for the focal length. In so far as any thin lens is capable of describing the effects of diverging fields in the anode region, we may then use the appropriate optical formulas to transfer our knowledge of the electron trajectories (calculated in the anode region as outlined above) to the start of the drift region. In particular,

$$\left(\frac{dr}{dz}\right)_2 = \left(\frac{dr}{dz}\right)_1 - \frac{r}{F} \quad (20)$$

where $(dr/dz)_1$ and $(dr/dz)_2$ are the slopes of the path just before and just after the lens, and r is the distance from the axis to the point where the ideal path crosses the lens plane.

B. The Drift Region

Although r_e/σ was found to be large at the anode plane for most guns of interest, this ratio often shrinks to 1 or less at an axial distance of only a few beam diameters from the lens. Therefore, the assumption that electron trajectories may be found by using the space charge forces which would exist in the absence of thermal velocities of emission (i.e., forces consistent with the universal beam spread curve) may lead to very appreciable error. For example, if equal normal (Gaussian) distributions of points about a central point are superposed so that the central points are equally dense throughout a circle of radius r_e , and if the standard deviation for each of the normal distributions is $\sigma = r_e$, the relative density of points in the center of the circle is only about 39 per cent of what it would be with $\sigma < (r_e/5)$.

In order to minimize errors of this type we have modified the Hines-Cutler treatment of the drift space in two ways: (1) The forces influencing the trajectories of the non-thermal electrons are calculated from a progressive estimation of the actual space charge configuration as modified by the presence of thermal electrons. (2) Some account is taken of the fact that, as the space charge density in the beam becomes less uniform as a function of radius, the spread of electrons near the center of the beam increases more rapidly than does the corresponding spread

farther out. Since item (1) is influenced by item (2), the specific assumptions involved in the latter case will be treated first.

When current density is uniform across the beam and its cross section changes slowly with distance, considerations of the type outlined above for the gun region show that those thermal electrons which remain within the beam will continue to have a Gaussian distribution with respect to a non-thermal electron emitted from the same cathode point. When current density is not uniform over the cross section, we would still like to preserve the mathematical simplicity of obtaining the current density as a function of beam radius merely by superposing Gaussian distributions which can be associated with each non-thermal electron. To lessen the error involved in this simplified approach, we will arrive at a value for the standard deviation, σ (which specifies the Gaussian distribution), in a rather special way. In particular, σ at any axial position, z , will be taken as the radial coordinate of an electron emitted from the center of the cathode with a transverse velocity of emission given by,

$$v_c = \sqrt{\frac{kT}{m}} \quad (21)$$

It is clear from (17) that for such an electron, $r = \sigma$ in the gun region. From (18), the fraction of the electrons from a common point on the cathode which will have $r \leq \sigma$ in the gun region is

$$\text{fraction} = \int_0^\sigma e^{-(r^2/2\sigma^2)} d \frac{r^2}{2\sigma^2} = 1 - e^{-1/2} = 0.393 \quad (22)$$

If r_e denotes the radial position of the outermost non-thermal electron and if $\sigma > r_e$, the " σ -electron" will be moving in a region where the space charge density is significantly lower than at the axis. We could, of course, have followed the path of an electron with initial velocity equal to say 0.1 or 10 times that given in (21) and called the corresponding radius 0.1σ or 10σ . The reason for preferring (21) is that about 0.4 or nearly half of the thermal electrons emitted from a common cathode point will have wandered a distance less than σ from the path of a non-thermal electron emitted from the same cathode point, while other thermal electrons will have wandered farther from this path; consequently, the current density in the region of the σ -electron is expected to be a reasonable average on which beam spreading due to thermal velocities may be based. With this understanding of how σ is to be calculated, we can proceed to the calculation of non-thermal electron trajectories as suggested in item (1).

The non-thermal paths remain essentially laminar, and with r_e denoting the radial coordinate of the outermost non-thermal electron, we will make little error in assuming that the current density of non-thermal electrons is constant for $r < r_e$. Consequently, if equal numbers of thermal electrons are assumed to be normally distributed about the corresponding non-thermal paths, the longitudinal current density as a function of radius can be found in a straightforward way¹³ by using (18). The result is

$$\frac{J_r}{J_D} = e^{-(r^2/2\sigma^2)} \int_0^{r_e/\sigma} \frac{R}{\sigma} e^{-(R^2/2\sigma^2)} I_0\left(\frac{rR}{\sigma^2}\right) d\left(\frac{R}{\sigma}\right) \quad (23)$$

where I_0 is the zero order modified Bessel function and the total current is $J_D = \pi r_e^2 J_D$. Equation (23) was integrated to give a plot of J_r/J_D versus r/σ , with r_e/σ as a parameter and is given as Fig. 6 in Reference 6. It is reproduced here as Fig. 6. Since the only forces acting on electrons in the drift region are due to space charge, we may write the equation of motion as

$$\frac{d^2r}{dt^2} = \eta E_r \quad (24)$$

where E_r is the radial electrical field acting on an electron with radial coordinate r . Since the beam is long and narrow, all electric lines of force may be considered to leave the beam radially so that E_r is simply obtained from Gauss' law. Equation (24) therefore becomes

$$\begin{aligned} \frac{d^2r}{dt^2} &= \frac{\eta}{2\pi\epsilon_0 r} \int_0^r 2\pi\rho dr = \frac{\eta}{2\pi\epsilon_0 r} \int_0^r \frac{J(r)}{\sqrt{2\eta V_a}} 2\pi r dr \\ &= \frac{\sqrt{\eta/(2V_a)}}{2\pi\epsilon_0 r} \int_0^r J(r) 2\pi r dr \end{aligned} \quad (25)$$

From (23) we note that the fraction of the total current within any radius depends only on r_e/σ and r/σ :

$$\frac{I_r}{I_D} = \frac{\int_0^r J(r) 2\pi r dr}{\int_0^\infty J(r) 2\pi r dr} = 2 \left(\frac{\sigma}{r_e}\right)^2 \int_0^{r_e/\sigma} e^{-(r^2/2\sigma^2)} \quad (26)$$

$$\times \left[\int_0^{r_e/\sigma} \frac{R}{\sigma} e^{-(R^2/2\sigma^2)} I_0\left(\frac{rR}{\sigma^2}\right) d\left(\frac{R}{\sigma}\right) \right] \frac{r}{\sigma} d\left(\frac{r}{\sigma}\right) \equiv F\left(\frac{r}{\sigma}, \frac{r_e}{\sigma}\right)$$

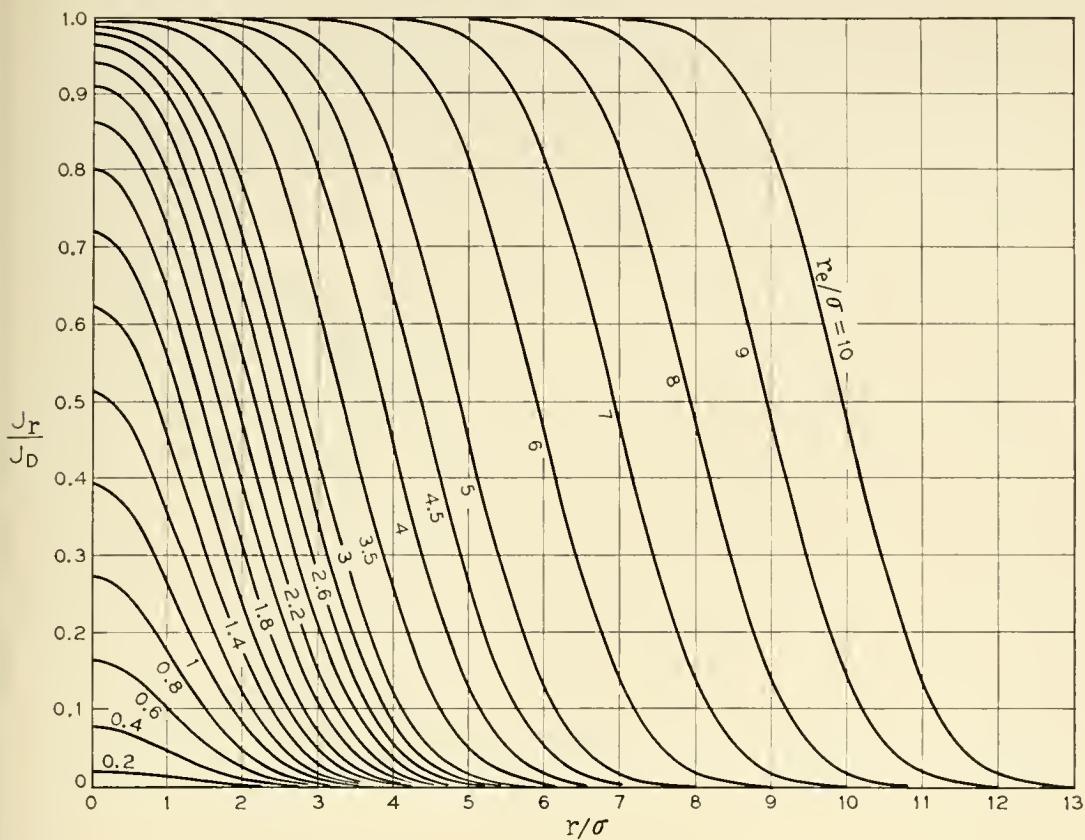


Fig. 6 — Curves showing the current density variation with radius in a beam which has been dispersed by thermal velocities. Here r_e is the nominal beam radius, r is the radius variable, and σ is the standard deviation defined in equation 17.

A family of curves with this ratio, F_r , as parameter has been reproduced from the Hines-Cutler paper and appears here as Fig. 7. Using this notation, (25) becomes

$$\frac{d^2r}{dt^2} = \frac{\sqrt{\eta/(2V_a)}}{2\pi\epsilon_0} I_D \frac{F_r}{r}$$

or

$$\frac{d^2r}{dz^2} = \frac{\eta}{2\pi\epsilon_0} \frac{I_D}{(2\eta V_a)^{3/2}} \frac{F_r}{r} \equiv K \frac{F_r}{r} \quad (27)$$

where we have made use of the dc electron drift velocity to make distance the independent variable instead of time, and have defined a quantity K which is proportional to gun perveance. We can now apply (27) to the motion of both the outer (edge) non-thermal electron and the σ -electron. From (26) we see that F_{r_e} and F_σ depend only on r_e/σ ;

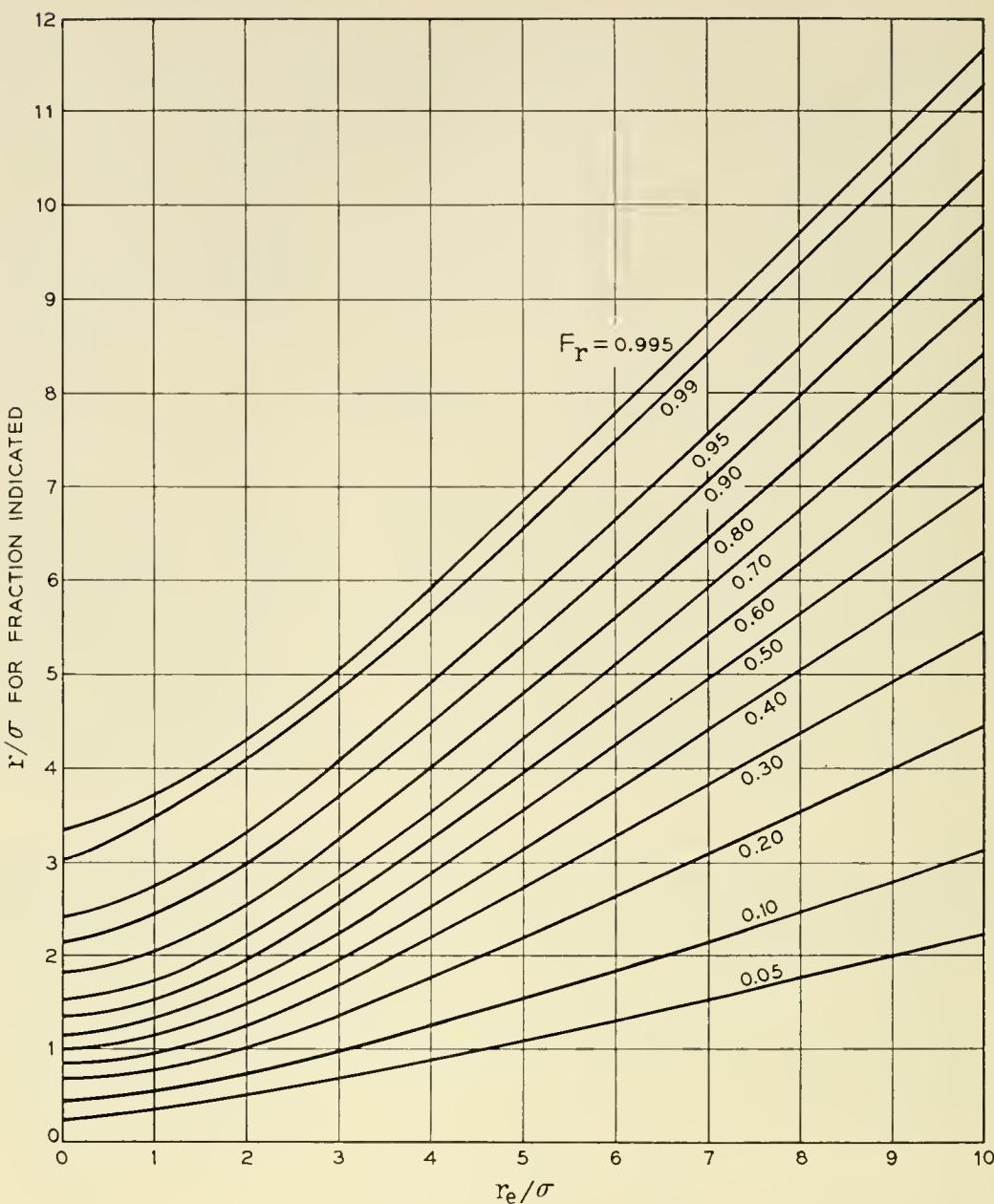


Fig. 7 — Curves showing the fraction, F_r , of the total beam current to be found within any given radius in a beam dispersed by thermal velocities as in Fig. 6.

consequently the continuous solution for r_e and r_σ ($= \sigma$) as one moves axially along the drifting beam involves the simultaneous solution of two equations:

$$\frac{d^2 r_e}{dz^2} = K F_{r_e} / r_e \quad (28)$$

$$\frac{d^2 \sigma}{dz^2} = K F_\sigma / \sigma$$

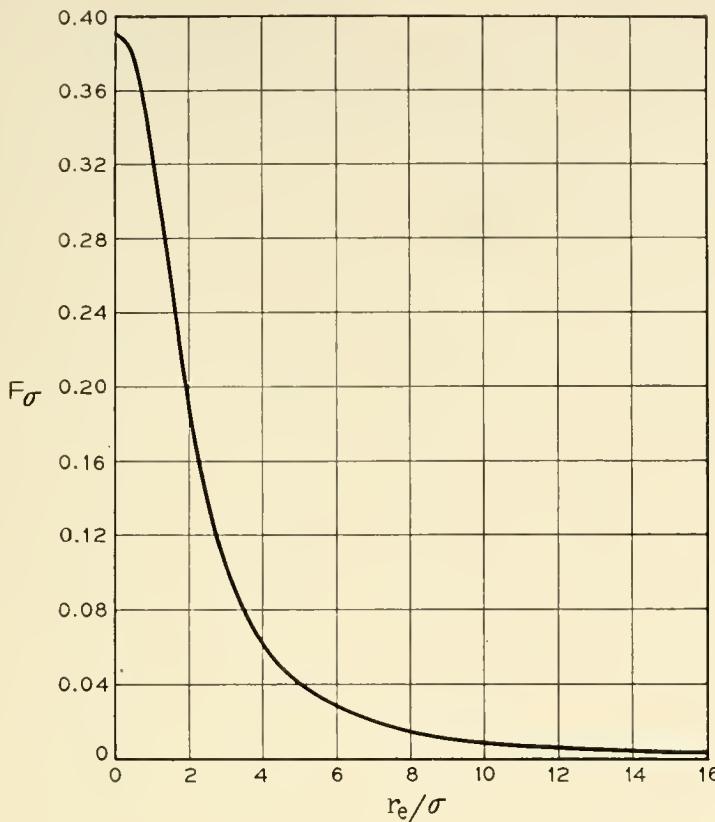


Fig. 8 — A curve showing the effect of a quantity related to the space charge force (in the drift region) on a thermal electron with standard deviation σ . (See equation 28.)

which are related by the mutual dependence of F_{r_e} and F_σ on r_e/σ . F_σ and F_{r_e}/r_e are plotted in Figs. 8 and 9.

We may summarize the treatment of the drift region, then, as follows:

(a) The input values of r_e and r_{e-}' at the entrance to the anode lens are obtained from the Pierce gun parameters r_a and θ , while the value of σ and σ_-' at the lens entrance can be obtained as mentioned above by integrating (15) from the cathode, where $\mu_c = 0$ and $(d\mu/dt)_e = 1$, to the anode plane. (The minus subscripts on r' and σ' indicate that these slopes are being evaluated on the gun side of the lens; a plus subscript will be used to indicate evaluation on the drift region side of the lens.) The values of r_e and σ on leaving the lens will of course be their entrance values in the drift region, and the effect of the lens on r_e' and σ' is simply found in terms of the anode lens correction factor Γ by use of (20). The value of σ at the anode can be obtained from (17) if μ is known there. In this regard, (15) can be integrated once to give

$$d\mu = \frac{1}{r_e} \left(\frac{dr}{dt} \right)_e \frac{dt}{(r_e/r_e)^2} \quad (29)$$

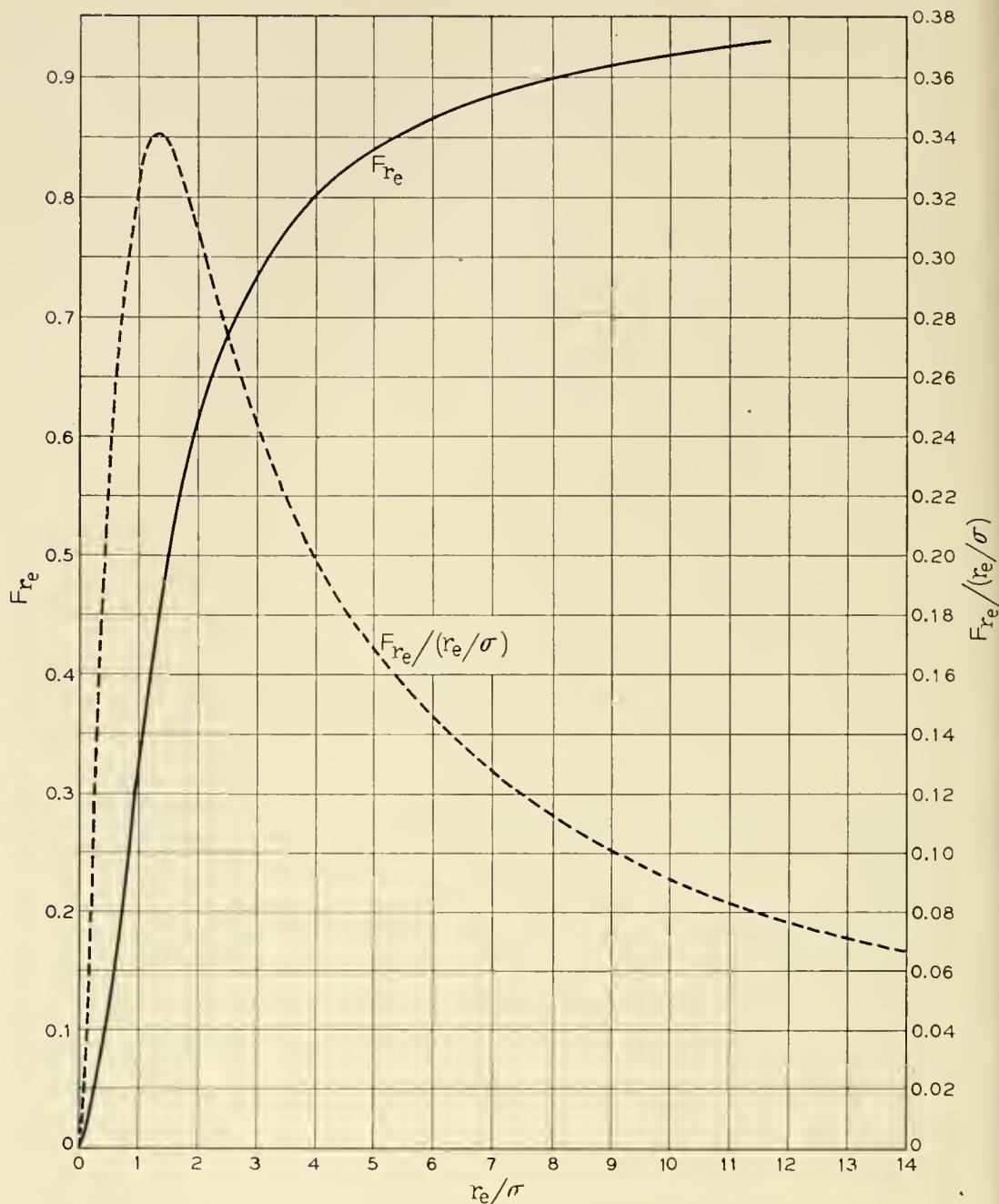


Fig. 9 — Showing quantities related to the effect of the space charge force in the drift region on the outermost non-thermal electron. (See equation 28.)

We can now substitute for transit time in terms of distance and Langmuir's well known potential function,¹⁴ $-\alpha$. The value of this parameter, for the case of spherical cathode-anode geometry in which we are interested, depends only on the ratio \bar{r}_c/\bar{r} which is equal to r_c/r_e . (Because of their frequent use in gun design, certain functions of $-\alpha$ are included here as Table I.) In terms of $-\alpha$, then, the potential in the gun region

TABLE I—TABLE OF FUNCTIONS OF $-\alpha$ OFTEN USED IN ELECTRON GUN DESIGN

\bar{r}_c/\bar{r}	$(-\alpha)^2$	$(-\alpha)^{4/3}$	$(-\alpha)^{2/3}$	$\int_1^{\bar{r}_c/\bar{r}} \frac{d\left(\frac{\bar{r}_c}{\bar{r}}\right)}{(-\alpha)^{2/3}}$	$\frac{d(-\alpha)^{4/3}}{d(\bar{r}_c/\bar{r})}$
1.0	0.0000	0.0000	0.0000	0.0000	
1.025	0.0006	0.0074			
1.05	0.0024	0.0179	0.134		
1.075	0.0052	0.0306	0.173		
1.10	0.0096	0.0452	0.212	1.392	0.590
1.15	0.0213	0.0768	0.277		
1.20	0.0372	0.1114	0.334	1.767	0.716
1.25	0.0571	0.1483	0.385		
1.30	0.0809	0.1870	0.432	2.031	0.790
1.35	0.1084	0.2273	0.476		
1.40	0.1396	0.2691	0.519	2.243	0.874
1.45	0.1740	0.3117	0.558		
1.50	0.2118	0.3553	0.596	2.423	0.886
1.60	0.2968	0.4450	0.667	2.583	0.915
1.70	0.394	0.5374	0.733	2.725	0.939
1.80	0.502	0.6316	0.795	2.855	0.954
1.90	0.621	0.7279	0.853	2.975	0.970
2.00	0.750	0.8255	0.908	3.087	0.982
2.10	0.888	0.9239	0.961	3.192	0.993
2.20	1.036	1.024	1.012	3.292	1.003
2.30	1.193	1.125	1.061	3.388	1.012
2.40	1.358	1.226	1.107	3.481	1.020
2.50	1.531	1.328	1.152	3.570	1.028
2.60	1.712	1.431	1.196	3.655	1.034
2.70	1.901	1.535	1.239	3.738	1.039
2.80	2.098	1.639	1.280	3.817	1.044
2.90	2.302	1.743	1.320	3.894	1.048
3.00	2.512	1.848	1.359	3.968	1.052
3.1	2.729	1.953	1.397	4.040	1.056
3.2	2.954	2.059	1.435	4.111	1.059
3.3	3.185	2.164	1.471	4.180	1.062
3.4	3.421	2.270	1.507	4.247	1.064
3.5	3.664	2.376	1.541	4.315	1.066
3.6	3.913	2.483	1.576	4.377	1.068
3.7	4.168	2.590	1.609	4.441	1.070
3.8	4.429	2.697	1.642	4.501	1.072
3.9	4.696	2.804	1.674	4.563	1.074
4.0	4.968	2.912	1.706	4.621	1.076

may be written

$$V = V_a(-\alpha)^{4/3}/(-\alpha_a)^{4/3} \quad (30)$$

$$dt = -\frac{d\bar{r}}{\sqrt{2\eta V}} = -\frac{d\bar{r}}{\sqrt{2\eta V_a}} \frac{(-\alpha_a)^{2/3}}{(-\alpha)^{2/3}} \quad (31)$$

so that upon substitution from (29) and (31), (17) becomes

$$\sigma = r_e \sqrt{\frac{T}{V_a}} \sqrt{\frac{k}{2e} \frac{\bar{r}_c}{r_c} (-\alpha_a)^{2/3}} \int_1^{\bar{r}_c/\bar{r}} (-\alpha)^{-2/3} d\left(\frac{\bar{r}_c}{\bar{r}}\right) \quad (32)$$

Fig. 5, which has been referred to above, shows

$$\frac{\sigma_a}{\bar{r}_c} \sqrt{\frac{2eV_a}{kT}}$$

as a function of (\bar{r}_c/\bar{r}_a) as obtained from (32), and allows σ_a to be determined easily. Using (20), the value of r_{e+}' is given by

$$r_{e+}' = -\frac{r_{ea}}{F} + r_{e-}' = -\frac{r_{ea}}{F} - \theta_e = -\frac{\Gamma r_{ea}}{F_D} - \theta_e = \theta_e \left(-\frac{\Gamma \bar{r}_a}{F_D} - 1 \right) \quad (33)$$

where θ_e is the half-angle of the cathode (and hence the initial angle which the path of a non-thermal edge electron makes with the axis). We may write for $1/F_D$

$$-\frac{1}{F_D} = \frac{V'}{4V} = \frac{\bar{r}_c}{4(-\alpha_a)^{4/3}\bar{r}_a^2} \left(\frac{d(-\alpha)^{4/3}}{d(\bar{r}_c/\bar{r})} \right)_a \quad (34)$$

In Fig. 10 we plot $-\bar{r}_a/F_D$ as a function of \bar{r}_c/\bar{r}_a for easy evaluation of r_{e+}' in (33). Taking the first derivative of (32) with respect to z , we obtain an expression for σ_-' . Using this in conjunction with (20) and (34) we find

$$\sigma_{+}' = \sqrt{\frac{T}{V_a}} (\Gamma C_1 + C_2) \quad (35)$$

where

$$C_1 = \sqrt{\frac{k}{2e}} \frac{\bar{r}_c/\bar{r}_a}{4(-\alpha)^{2/3}} \left(\frac{d(-\alpha)^{4/3}}{d(\bar{r}_c/\bar{r})} \right)_a \int_1^{\bar{r}_c/\bar{r}_a} \frac{d(\bar{r}_c/\bar{r})}{(-\alpha)^{2/3}}$$

and

$$C_2 = \sqrt{\frac{k}{2e}} \left[\frac{\bar{r}_c}{\bar{r}_a} - (-\alpha_a)^{2/3} \int_1^{\bar{r}_c/\bar{r}_a} \frac{d(\bar{r}_c/\bar{r})}{(-\alpha)^{2/3}} \right]$$

C_1 and C_2 are plotted as functions of \bar{r}_c/\bar{r}_a in Fig. 11.

(b) After choosing a specific value for Γ and evaluating $K = \eta I_D/$

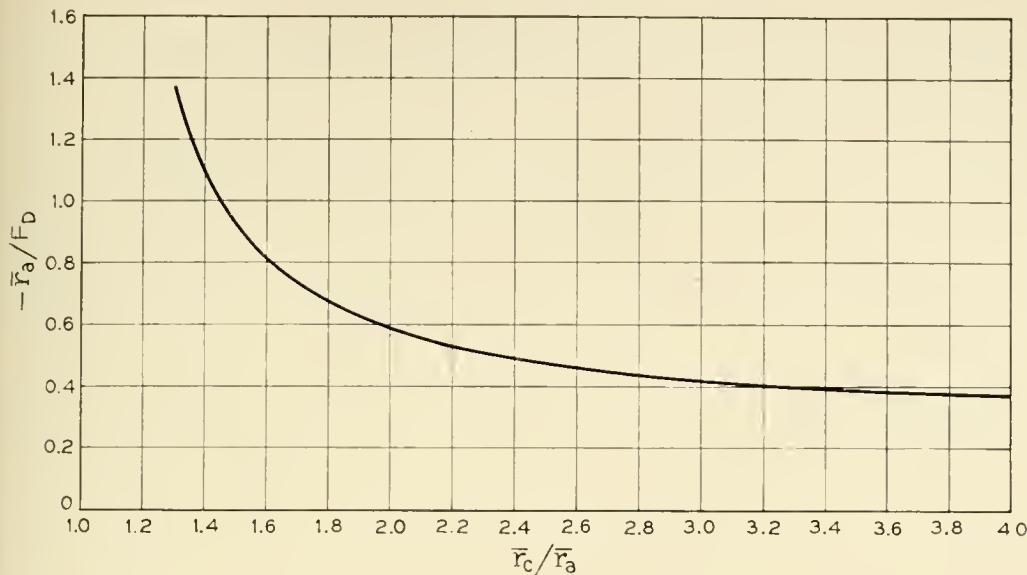


Fig. 10 — Curve used in finding r_{e+}' , the direction of a nonthermal edge electron as it enters the drift region. (See equation 33.)

$(2\pi\epsilon_0(2\eta V_a)^{3/2})$, (28) is integrated numerically using the BTL analog computer to obtain σ and r_e as functions of axial distance along the beam.

(c) Knowing σ and r_e , other beam parameters such as current distribution and the radius of the circle which would encompass a given percentage of the total current can be found from Figs. 6 and 7.

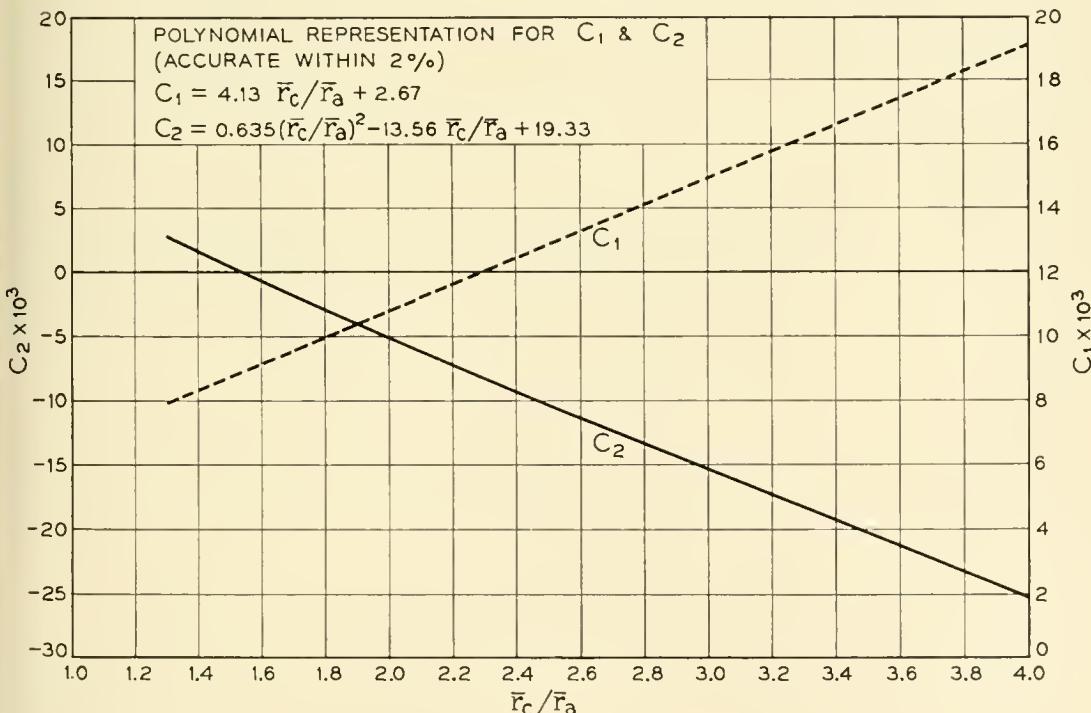


Fig. 11 — Curves used in evaluating σ_{+}' , the slope of the trajectory of a thermal electron with standard deviation σ as it enters the drift region. (See equation 35.)

5. NUMERICAL DATA FOR ELECTRON GUN AND BEAM DESIGN

A. Choice of Variables

Except for a scaling parameter, the electrical characteristics of an ideal Pierce electron gun are completely determined when three parameters are specified, e.g., \bar{r}_c/\bar{r}_a , perveance, and V_a/T . Also, for the simplest case Γ is equal to 1 so that (since K depends only on gun perveance) in this case no additional parameter is needed. This implies that normalized values of r_e' , σ , σ' , and K at the drift side of the anode lens are not independent. If, however, the value of Γ at the anode lens is taken as an additional variable, four parameters plus simple scaling are required before complete predictions of beam characteristics can be made. In assembling analog computer data which would adequately cover values of \bar{r}_c/\bar{r}_a , perveance, and V_a/T which are likely to be of interest to us in designing future guns, we chose to present the major part of our data with Γ fixed at 1.1. This has seemed to be a rather typical value for Γ , and by choosing a specific value we decrease the total number of significant variables from 4 to 3. (The effect of variations in Γ on the minimum radius which contains 95 per cent of the beam is, however, included in Fig. 16 for particular values of V_a/T and perveance.) Although the boundary conditions for our mathematical description of the beam in a drift space are simplest when expressed in terms of r_e , r_e' , σ and σ' , we have attempted to make the results more usable by expressing all derived parameters in terms of \bar{r}_c/\bar{r}_a , $\sqrt{V_a/T}$, and the perveance, P .

B. Tabular Data

The rather extensive data obtained from the analog computer for the $\Gamma = 1.1$ case and for practical ranges in perveance, V_a/T , and \bar{r}_c/\bar{r}_a are summarized in Tables II A to E where the parameters r_e and σ which specify the beam cross section are given as functions of axial distance from the anode plane. Some feeling for the decrease in accuracy to be expected as the distance from the anode plane increases can be obtained by reference to Section 6B where experiment and theory are compared over a range of this axial distance parameter.

C. Graphical Data, Including Design Charts and Beam Profiles

In typical cases, the designer of Pierce electron guns is much more concerned with the beam radius at the axial position where it is smallest (and in the axial position of this minimum) than he is in the general

spreading of the beam with distance. This is true because, in microwave beam tubes, the beam from a magnetically shielded Pierce gun normally enters a strong axial magnetic field near a point where the radius is a minimum, so that magnetic focusing forces largely determine the beam's subsequent behavior. The analog computer data has therefore been re-processed to stress the dependence of the beam's minimum diameter and the corresponding axial position of the minimum on the basic design parameters \bar{r}_c/\bar{r}_a , perveance, and $\sqrt{V_a/T}$. As a first step in this direction, the radius, r_{95} , of a circle which includes 95 per cent of the beam current is obtained as a function of axial position along the beam. Such data are shown graphically in Fig. 12. Finally, the curves of Fig. 12 are used in conjunction with the tabular data to obtain the "Design Curves" of Fig. 13 where all of the pertinent information relating to the beam at its minimum diameter is presented.

D. Example of Gun Design Using Design Charts

Assume that we desire an electron gun with the following properties: anode voltage $V_A = 1,080$ volts, cathode current $I_D = 7.1$ ma, and minimum beam diameter $2(r_{95})_{\min} = 0.015$ inches. Let us further assume a cathode temperature $T = 1080^\circ$ Kelvin, an available cathode emission density of 190 ma per square cm, and an anode lens correction factor of $\Gamma = 1.1$. From these data we find $\sqrt{V_a/T} = 1.0$, perveance $P = 0.2 \times 10^{-6}$ amps/(volts)^{3/2} and $(r_{95})_{\min}/r_c = 0.174$. Reference to the design chart, Fig. 13, now gives us the proper value for \bar{r}_c/\bar{r}_a : using the upper set of curves in the column for $\sqrt{V_a/T} = 1.0$ we note the point of intersection between the horizontal line for $(r_{95})_{\min}/r_c = 0.174$ and the perveance line $P = 0.2$, and read the value of $\bar{r}_c/\bar{r}_a (= 2.8)$ as the corresponding abscissa. The convergence angle of the gun, θ_e , is now simply determined from the equation¹⁵

$$\theta_e = \cos^{-1} \left(1 - \frac{(-\alpha_a)^2 P}{14.67} \times 10^6 \right) \quad (37)$$

(θ_e is found to be 13.7° in this example) and the potential distribution in the region of the cathode can be obtained from (30).

When this point has been reached, the gun design is complete except for the shapes of the beam forming electrode and the anode, which are determined with the aid of an electrolytic tank in the usual way.¹⁶ The radius of the anode hole which will give a specified transmission can be found by obtaining $(r_e/\sigma)_a$ through the use of Fig. 5, and then choosing the anode radius from Fig. 7. In practical cases where $(r_e/\sigma)_a > 3.0$,

TABLE II A - SUMMARY OF ANALOG COMPUTER DATA FOR AN ANODE LENS CORRECTION OF $\Gamma = 1.1$ FOR PERVEANCE = 0.05×10^{-6} AMPS/(VOLTS) $^{3/2}$

$\tilde{r}_c/\tilde{r}_a = 1.515$				$\tilde{r}_c/\tilde{r}_a = 2.0$				$\tilde{r}_c/\tilde{r}_a = 2.5$				$\tilde{r}_c/\tilde{r}_a = 3.0$				$\tilde{r}_c/\tilde{r}_a = 3.5$				$\tilde{r}_c/\tilde{r}_a = 4.0$				
$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\varepsilon}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_e}{\sigma}$	
0	0.2	2.02	0.495	0	1.95	0.513	0	0.527	0	1.90	0.545	0	1.85	0.540	0	1.81	0.552	0	1.75	0.565	0	1.44	0.549	
2	4	1.90	1.76	0.535	1	1.86	0.528	2	1.78	0.541	4	1.45	0.560	4	1.35	0.560	4	1.19	0.548	4	1.07	0.531	6	0.519
6	8	1.65	1.65	0.620	3	1.67	0.556	6	1.24	0.580	6	1.04	0.600	8	0.82	0.565	8	0.56	0.548	6	0.69	0.505	10	-0.12
12	14	1.55	1.55	0.669	4	1.60	0.570	8	1.04	0.621	10	0.86	0.621	10	0.58	0.575	10	0.20	0.552	10	0.498	0.490	$\sqrt{V_a/T} = 0.5$	
16	18	1.44	1.44	0.601	6	1.44	0.601	12	0.69	0.645	12	0.35	0.589	12	-0.10	0.558	12	-0.56	0.490	14	-1.01	0.485	$\sqrt{V_a/T} = 0.5$	
20	22	1.30	1.18	0.669	8	1.30	0.669	14	0.52	0.670	14	0.13	0.602	14	-0.40	0.563	14	-1.48	0.482	16	-1.92	0.482	$\sqrt{V_a/T} = 0.5$	
24	26	1.05	1.05	0.701	10	1.18	0.669	16	0.38	0.700	16	-0.10	0.619	16	-0.70	0.575	16	-1.0	0.585	18	-2.38	0.485	$\sqrt{V_a/T} = 0.5$	
30	32	1.2	12	0.701	18	0.25	0.725	20	0.12	0.760	24	-0.47	0.657	18	-1.0	0.60	20	-1.25	0.615	22	-1.53	0.615	$\sqrt{V_a/T} = 0.5$	
36	38	0	22	0	0	0.790	26	0	0.790	26	-0.96	0.729	22	-1.0	0.60	20	-1.25	0.615	28	-1.53	0.615	$\sqrt{V_a/T} = 0.5$		
40	42	0	0	0	0	0.256	0	0	0.264	0	3.78	0.270	0	3.70	0.270	0	3.62	0.276	0	3.56	0.281	$\sqrt{V_a/T} = 0.5$		
44	46	0.267	0.267	0.270	2	3.55	0.270	2	3.32	0.273	2	3.15	0.273	2	3.0	0.273	2	2.85	0.272	4	2.11	0.265		
50	52	0.288	0.288	0.285	4	3.20	0.285	4	2.87	0.281	4	2.62	0.277	4	2.35	0.273	4	2.74	0.274	6	1.32	0.260		
56	58	0.310	0.310	0.302	6	2.89	0.302	6	2.46	0.292	6	2.1	0.282	6	1.71	0.274	6	0.54	0.259	8	0.15	0.260		
64	66	0.325	0.325	0.322	8	2.61	0.318	8	2.08	0.308	8	1.6	0.289	8	1.06	0.280	8	0.286	0.286	10	-0.25	0.262		
72	74	0.358	0.358	0.335	10	2.35	0.338	10	1.71	0.318	10	1.1	0.299	10	0.45	0.292	10	-1.0	-1.0	12	-1.71	0.282		
80	82	0.382	0.382	0.358	12	2.11	0.358	12	1.38	0.334	12	0.67	0.311	12	0.15	0.292	10	-1.36	0.297	16	-2.94	0.312		
88	90	0.408	0.408	0.408	14	1.90	0.379	13	1.23	0.344	13	0.46	0.32	12	-0.12	0.299	12	-1.0	-1.0	18	-3.44	0.33		
96	98	0.443	0.443	0.402	16	1.71	0.402	14	1.09	0.352	14	0.25	0.329	14	-0.63	0.314	14	-1.71	0.282	20	-2.87	0.33		
104	106	0.492	0.492	0.456	20	1.40	0.456	16	0.83	0.374	16	-0.10	0.349	16	-1.1	0.332	16	-2.36	0.297	24	-4.31	0.368		
112	114	0.553	0.553	0.516	24	1.14	0.516	20	0.38	0.426	18	-0.42	0.372	20	-1.86	0.380	18	-2.94	0.312	28	-4.31	0.368		
120	122	0.588	0.588	0.582	28	0.95	0.582	24	0.06	0.489	20	-0.68	0.40	24	-2.43	0.433	20	-3.44	0.33	32	-4.31	0.368		
128	130	0.535	0.535	0.535	32	0.20	0.561	28	-0.20	0.463	24	-1.13	0.463	28	-1.47	0.535	28	-1.47	0.535	36	-4.31	0.368		

TABLE IIB—SUMMARY OF ANALOG COMPUTER DATA FOR ANODE LENS CORRECTION OF $\Gamma = 1.1$ FOR
 $\text{PERVEANCE} = 0.1 \times 10^{-6}$ AMPS/(VOLTS) $^{3/2}$

TABLE II B (CONTINUED)

$\bar{r}_c/\bar{r}_a = 1.515$				$\bar{r}_c/\bar{r}_a = 2.0$				$\bar{r}_c/\bar{r}_a = 2.5$				$\bar{r}_c/\bar{r}_a = 3.0$				$\bar{r}_c/\bar{r}_a = 3.5$				$\bar{r}_c/\bar{r}_a = 4.0$			
$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{z}{r_a}$	$\frac{r_c}{\sigma}$	$\frac{\sigma}{r_a}$	
0	11.35	0.088	0	11.0	0.091	0	10.75	0.093	0	10.40	0.096	0	10.2	0.098	0	10.10	0.099	0	10.10	0.099	0	10.10	0.099
2	10.05	0.096	2	9.4	0.095	2	8.90	0.095	2	8.20	0.095	2	7.65	0.092	2	7.20	0.094	2	7.20	0.094	2	7.20	0.094
4	9.15	0.106	4	8.1	0.103	4	7.20	0.101	4	6.20	0.100	4	5.12	0.091	4	4.20	0.092	4	4.20	0.092	4	4.20	0.092
6	8.39	0.118	6	7.0	0.111	6	5.67	0.107	6	4.20	0.101	6	2.65	0.100	6	2.60	0.091	6	2.60	0.091	6	2.60	0.091
8	7.75	0.130	8	6.08	0.120	8	4.33	0.115	8	2.40	0.110	7	1.45	0.103	6	1.00	0.083	6	1.00	0.083	6	1.00	0.083
10	7.22	0.143	10	5.3	0.132	10	3.15	0.126	9	1.60	0.115	8	0.40	0.111	7	-0.28	0.100	7	-0.28	0.100	7	-0.28	0.100
12	6.80	0.156	12	4.6	0.145	12	2.12	0.141	10	0.90	0.124	9	-0.39	0.124	8	-1.47	0.113	8	-1.47	0.113	8	-1.47	0.113
16	6.15	0.186	14	4.03	0.160	13	1.72	0.151	11	0.32	0.138	10	-1.05	0.142	10	-3.05	0.149	10	-3.05	0.149	10	-3.05	0.149
20	5.65	0.218	16	3.55	0.176	14	1.37	0.164	12	-0.08	0.154	12	-1.98	0.186	12	-3.95	0.189	12	-3.95	0.189	12	-3.95	0.189
24	5.25	0.254	18	3.15	0.194	16	0.80	0.197	14	-0.68	0.200	14	-2.5	0.237	14	-4.62	0.230	14	-4.62	0.230	14	-4.62	0.230
28	4.95	0.293	20	2.80	0.216	20	0.20	0.295	16	-1.08	0.259	16	-2.85	0.292	16	-4.95	0.274	16	-4.95	0.274	16	-4.95	0.274
32	4.75	0.336	24	2.25	0.270	24	-0.05	0.427	20	-1.43	0.393	18	-3.08	0.350	18	-5.22	0.337	18	-5.22	0.337	18	-5.22	0.337
34	4.65	0.357	28	1.93	0.339	28	-0.18	0.580	24	-1.65	0.548	20	-3.2	0.408	20	-5.45	0.385	20	-5.45	0.385	20	-5.45	0.385

 $\sqrt{V_a/T} = 1.0$ $\sqrt{V_a/T} = 2.0$

$\tilde{r}_c/\tilde{r}_a = 1.515$			1.97			2.54			3.15			3.46			4.0		
$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$
0	4.05	0.247	0	3.9	0.258	0	3.75	0.266	0	3.63	0.276	0	3.55	0.282	0	3.48	0.287
2	3.55	0.290	2	3.2	0.290	2	2.85	0.282	2	2.52	0.278	2	2.37	0.282	2	2.05	0.270
4	3.12	0.335	4	2.62	0.322	4	2.05	0.305	4	1.42	0.285	4	1.14	0.282	4	0.45	0.261
6	2.77	0.380	6	2.15	0.360	6	1.35	0.331	6	0.45	0.308	5	0.55	0.285	4.5	0.05	0.265
8	2.51	0.432	8	1.75	0.408	8	0.78	0.370	8	-0.38	0.341	6	0.00	0.296	5	-0.31	0.266
10	2.30	0.490	10	1.45	0.459	10	0.34	0.420	10	-1.02	0.390	8	-1.00	0.325	6	-1.06	0.275
12	2.12	0.550	12	1.21	0.520	12	0.00	0.480	12	-1.51	0.450	10	-1.78	0.370	8	-2.42	0.300
14	1.98	0.614	14	1.01	0.580	14	-0.25	0.549	14	-1.88	0.515	12	-2.39	0.420	10	-3.52	0.330
				16	0.655	16	-0.46	0.625	16	-2.15	0.588				12	-4.44	0.365
0	8.1	0.123	0	7.75	0.129	0	7.5	0.133	0	7.25	0.138	0	7.1	0.141	0	6.95	0.144
2	7.0	0.145	2	6.30	0.147	2	5.7	0.142	2	5.00	0.140	2	4.7	0.140	2	4.1	0.134
4	6.15	0.166	4	5.20	0.163	4	4.1	0.153	4	2.85	0.145	4	2.25	0.140	4	0.95	0.135
6	5.5	0.192	6	4.25	0.182	6	2.7	0.170	6	0.92	0.160	6	0.0	0.148	4.5	0.2	0.138
8	5.0	0.216	8	3.5	0.205	8	1.59	0.195	7	0.12	0.178	8	-1.5	0.201	5	0.5	0.141
10	4.6	0.244	10	2.95	0.234	10	0.75	0.237	8	-0.42	0.200	10	-2.35	0.264	6	1.72	0.160
12	4.3	0.274	12	2.45	0.268	12	0.25	0.299	10	-1.27	0.261	12	-2.85	0.332	8	3.45	0.205
14	4.02	0.306	14	2.10	0.301	14	-0.10	0.375	12	-1.76	0.339	14	-3.19	0.408	10	4.49	0.255
16	3.8	0.34	16	1.80	0.358	16	-0.30	0.470	14	-2.08	0.423				12	5.15	0.310
18	3.65	0.378	20	1.42	0.475	18	-0.45	0.570	16	-2.25	0.513						
20	3.5	0.42	24	1.20	0.625	20	-0.52	0.680	18	-2.40	0.610						
0	16.0	0.063	0	15.5	0.065	0	15.0	0.067	0	14.5	0.069	0	14.2	0.071	0	13.9	0.072
2	13.6	0.073	2	12.5	0.074	2	11.6	0.070	2	9.9	0.070	2	9.2	0.071	2	8.1	0.068
4	12.1	0.084	4	10.3	0.080	4	8.3	0.075	4	5.7	0.072	4	4.4	0.071	4	1.9	0.068
6	10.8	0.096	6	8.5	0.090	6	5.5	0.083	6	2.0	0.081	5	2.1	0.074	4.5	0.4	0.071
8	9.8	0.110	8	7.1	0.101	8	3.4	0.095	7	0.5	0.094	6	0.2	0.085	5	-0.8	0.080
10	9.1	0.124	10	6.0	0.115	10	1.8	0.119	8	-0.4	0.119	7	-1.0	0.108	6	-2.5	0.102
12	8.5	0.139	12	5.1	0.131	12	0.8	0.165	10	-1.4	0.196	8	-1.8	0.140	8	-4.2	0.162
14	8.05	0.156	14	4.45	0.151	14	0.3	0.235	12	-1.8	0.291	10	-2.58	0.220	10	-4.95	0.225
16	7.65	0.174	16	3.9	0.175	16	0.1	0.330	14	-1.95	0.395	12	-2.9	0.305	12	-5.38	0.290
18	7.35	0.193	20	3.2	0.236	18	-0.1	0.435	16	-2.05	0.505	14	-3.05	0.397	14	-5.60	0.360
20	7.1	0.214	24	2.75	0.316	20	-0.1	0.558	18	-2.1	0.621	16	-3.15	0.492			
22	6.85	0.236	28	2.45	0.415	22	-0.15	0.685									
24	6.7	0.258	32	2.26	0.535	24	-0.2	0.820									

 $\sqrt{V_a/T} = 0.5$ $\sqrt{V_a/T} = 1.0$ $\sqrt{V_a/T} = 2.0$

TABLE IID—SUMMARY OF ANALOG COMPUTER DATA FOR AN ANODE LENS CORRECTION OF $\Gamma = 1.1$ FOR
 PERVEREANCE = 0.4×10^{-6} AMPS./(VOLTS) $^{3/2}$

$\bar{r}_c/\bar{r}_a = 1.515$				2.04				2.44				2.86				3.6				4.0				
$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	$\frac{z}{r_a}$	$\frac{r_e}{\sigma}$	$\frac{\sigma}{r_a}$	
0	5.69	0.176	0	5.45	0.184	0	5.32	0.188	0	5.19	0.193	0	5	0.200	0	4.85	0.206	0	4.85	0.206	0	4.85	0.206	
2	4.65	0.220	2	4.00	0.210	2	3.60	0.200	2	3.20	0.198	1	3.7	0.195	1	3.40	0.198	1	3.40	0.198	1	3.40	0.198	
4	3.92	0.268	4	2.95	0.242	4	2.29	0.230	4	1.50	0.220	2	2.35	0.192	2	1.86	0.192	2	1.86	0.192	2	1.86	0.192	
6	3.41	0.321	6	2.15	0.290	6	1.20	0.270	5	0.75	0.225	3	0.97	0.196	3	0.25	0.196	3	0.25	0.196	3	0.25	0.196	
8	3.09	0.382	8	1.60	0.350	8	0.49	0.341	6	0.18	0.270	4	-0.27	0.210	3.5	-0.45	0.201	3.5	-0.45	0.201	3.5	-0.45	0.201	
10	2.80	0.452	10	1.20	0.430	10	0.07	0.450	8	-0.65	0.360	5	-1.32	0.231	4	-1.15	0.210	4	-1.15	0.210	4	-1.15	0.210	
12	2.59	0.526	12	0.91	0.525	12	-0.18	0.570	10	-1.15	0.460	6	-2.15	0.265	5	-2.40	0.232	5	-2.40	0.232	5	-2.40	0.232	
14	2.44	0.604	14	0.75	0.640	14	-0.32	0.705	12	-1.41	0.575	7	-2.78	0.300	6	-3.35	0.260	6	-3.35	0.260	6	-3.35	0.260	
16	2.32	0.688	16	0.61	0.760	16	-0.45	0.860	14	-1.60	0.715	8	-3.26	0.340	7	-4.10	0.290	7	-4.10	0.290	7	-4.10	0.290	
18	2.22	0.782	18	0.61	0.836	18	-0.45	0.860	16	-1.60	0.715	9	-3.65	0.380	8	-4.74	0.320	8	-4.74	0.320	8	-4.74	0.320	
20	2.14	0.888	20	0.61	0.902	20	0	10.9	0.092	0	10.7	0.094	0	10.3	0.097	0	10.0	0.100	0	9.75	0.103	0	9.75	0.103
22	2.07	0.108	22	8.1	0.106	22	7.2	0.100	2	6.6	0.099	1	7.5	0.098	1	6.8	0.097	1	6.8	0.097	1	6.8	0.097	
24	1.99	0.132	24	6.0	0.124	24	4.5	0.110	4	3.2	0.11	2	4.85	0.096	2	3.7	0.095	2	3.7	0.095	2	3.7	0.095	
26	1.92	0.156	26	4.5	0.147	26	2.4	0.135	6	0.5	0.15	3	2.1	0.099	3	0.50	0.098	3	0.50	0.098	3	0.50	0.098	
28	1.86	0.188	28	3.4	0.179	28	1.1	0.185	8	-0.6	0.25	3.5	0.80	0.104	4	-1.95	0.121	4	-1.95	0.121	4	-1.95	0.121	
30	1.81	0.220	30	2.6	0.220	30	0.42	0.275	10	-1.05	0.34	4	-0.32	0.115	5	-3.45	0.156	5	-3.45	0.156	5	-3.45	0.156	
32	1.76	0.256	32	2.08	0.280	32	0.18	0.40	12	-1.3	0.54	5	-1.9	0.15	6	-4.37	0.195	6	-4.37	0.195	6	-4.37	0.195	
34	1.72	0.293	34	1.7	0.352	34	0.00	0.54	14	-1.4	0.70	6	-2.8	0.195	7	-4.98	0.236	7	-4.98	0.236	7	-4.98	0.236	
36	1.68	0.336	36	1.5	0.445	36	-0.10	0.70	8	-3.7	0.292	8	-5.4	0.278	8	-5.4	0.278	8	-5.4	0.278	8	-5.4	0.278	
38	1.64	0.380	38	1.25	0.675	38	-0.15	0.88	10	-4.1	0.398	9	-5.68	0.320	9	-5.68	0.320	9	-5.68	0.320	9	-5.68	0.320	
40	1.60	0.434	40	21.8	0.046	0	21.3	0.047	0	20.7	0.048	0	20	0.05	0	19.5	0.052	0	19.5	0.052	0	19.5	0.052	
42	1.57	0.052	42	17	0.050	2	14.7	0.050	2	12.5	0.050	1	14.6	0.049	1	13.4	0.050	1	13.4	0.050	1	13.4	0.050	
44	1.54	0.064	44	12.5	0.060	4	9.3	0.055	4	5.7	0.056	2	9.2	0.048	2	7.4	0.048	2	7.4	0.048	2	7.4	0.048	
46	1.51	0.078	46	9.5	0.070	6	5.2	0.070	6	1.0	0.082	3	3.8	0.049	3	1.3	0.049	3	1.3	0.049	3	1.3	0.049	
48	1.48	0.092	48	7.25	0.086	8	2.5	0.098	8	-0.50	0.174	3.5	1.3	0.052	4	-2.6	0.076	4	-2.6	0.076	4	-2.6	0.076	
50	1.45	0.108	50	5.6	0.110	10	1.2	0.16	10	-0.80	0.236	4	-0.60	0.062	5	-4.1	0.120	5	-4.1	0.120	5	-4.1	0.120	
52	1.42	0.126	52	4.6	0.140	12	0.75	0.265	12	-0.95	0.52	5	-2.45	0.106	6	-4.7	0.167	6	-4.7	0.167	6	-4.7	0.167	
54	1.39	0.146	54	3.9	0.175	14	0.60	0.405	14	-1.0	0.72	6	-3.15	0.160	7	-5.1	0.215	7	-5.1	0.215	7	-5.1	0.215	
56	1.36	0.168	56	3.5	0.222	16	0.49	0.57	16	-0.70	0.75	8	-3.5	0.220	8	-5.5	0.265	8	-5.5	0.265	8	-5.5	0.265	
58	1.33	0.190	58	2.90	0.350	18	0.45	0.75	18	-0.70	0.75	8	-3.65	0.281	9	-5.5	0.315	9	-5.5	0.315	9	-5.5	0.315	

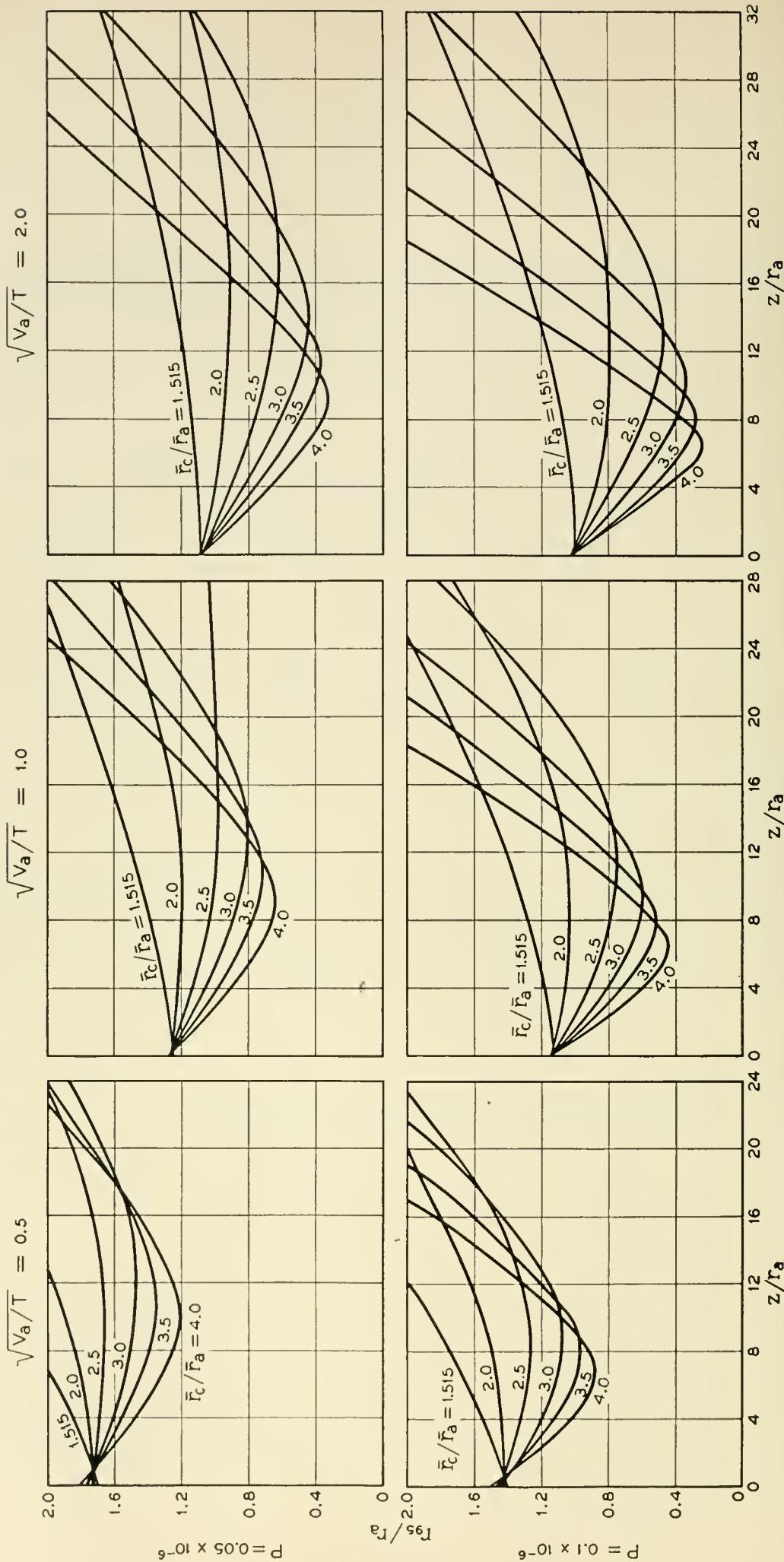


Fig. 12A — Curves showing normalized beam radius (95 per cent) versus distance from gun anode for variations in permeance (P), \bar{r}_c/\bar{r}_a and $\sqrt{V_a/T}$.

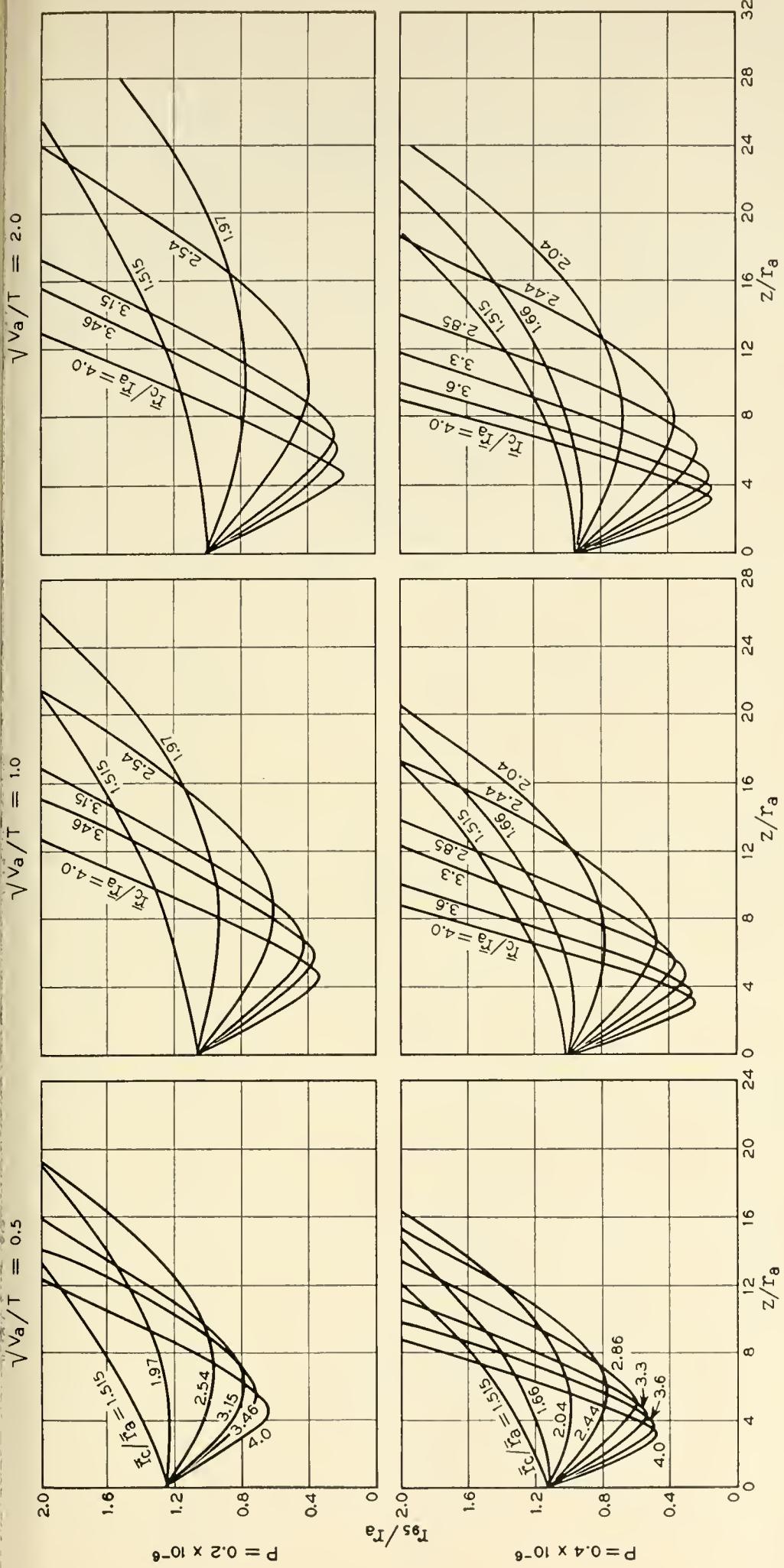


Fig. 12B — Curves showing normalized beam radius (95 per cent) versus distance from gun anode for variations in permeance (P), \bar{r}_c/\bar{r}_a and $\sqrt{V_a}/T$.

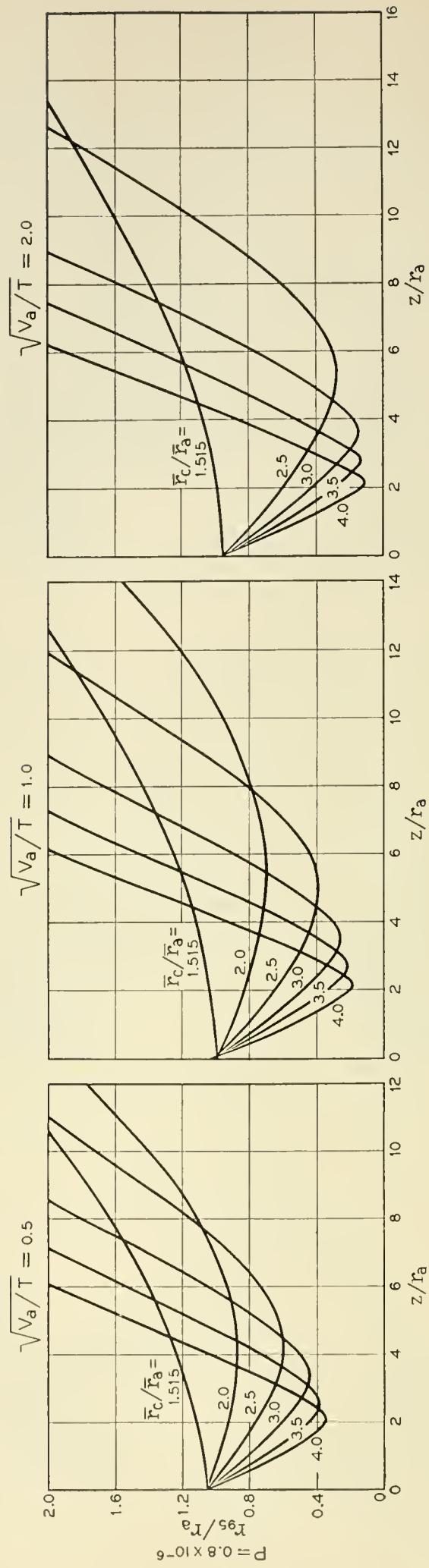


Fig. 12C — Curves showing normalized beam radius (95 per cent) versus distance from gun anode for variations in permeance (P), \bar{r}_c/\bar{r}_a and $\sqrt{V_a/T}$.

we find less than 1 per cent anode interception if

$$\text{anode hole radius} = 0.93 r_{ea} + 2\sigma_a \quad (38)$$

Additional information about the axial position of $(r_{95})_{\min}$ and the current density distribution in the corresponding transverse plane is contained in Fig. 13. The second set of curves in the $\sqrt{V_a/T} = 1$ column gives $z_{\min}/r_c = 2.42$ for this example, so that we would predict

$$z_{\min} \equiv \text{distance from anode to } (r_{95})_{\min} = 0.104''$$

The remaining 3rd and 4th sets of curves in the $\sqrt{V_a/T} = 1$ column allow us to find σ and r_e/σ at z_{\min} . In particular we obtain $\sigma = 0.0029''$ and $r_e/\sigma = 0.8$, and use Fig. 6 to give the current density distribution at z_{\min} .* Section VI contains experimental data which indicate a somewhat larger value for z_{\min} than that obtained here. However the parameter of greatest importance, $(r_{95})_{\min}$, is predicted with embarrassing precision.

For those cases in which additional information is required about the beam shape at axial points other than z_{\min} , the curves of Fig. 12 or the data of Table II may be used.

6. COMPARISON OF THEORY WITH EXPERIMENT

In order to check the general suitability of the foregoing theory and the usefulness of the design charts obtained, several scaled-up versions of Pierce type electron guns, including the gun described in Section 5D, were assembled and placed in the double-aperture beam analyzer described in Reference 7.

A. Measurement of Current Densities in the Beam

Measurements of the current density distributions in several transverse planes near z_{\min} were easily obtained with the aid of the beam analyzer. The resulting curve of relative current density versus radius at the experimental z_{\min} is given in Fig. 14 for the gun of Section 5D. (This curve is further discussed in Part C below.) For this case, as well as for all others, special precautions were taken to see that the gun was functioning properly: In addition to careful measurement of the size and position of all gun parts, these included the determination that the distribution of transverse velocities at the center of the beam was smooth

* When $r_e/\sigma < 0.5$, the current density distribution depends almost entirely on σ , and, in only a minor way, on the ratio r_e/σ so that in such cases this ratio need not be accurately known.

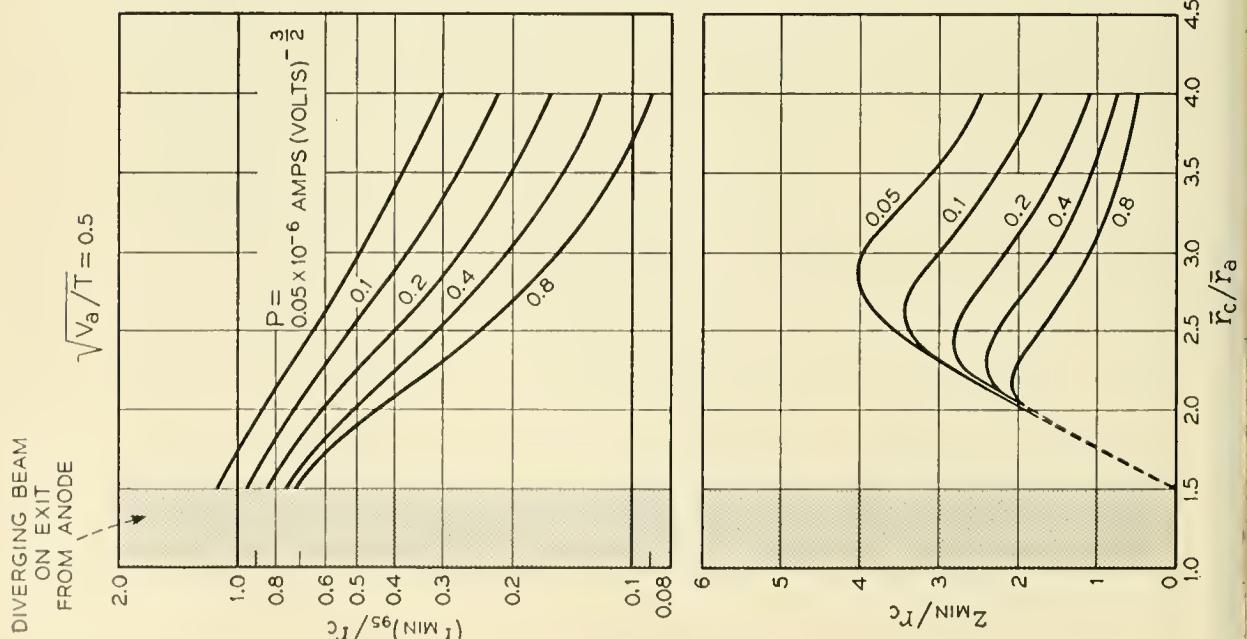
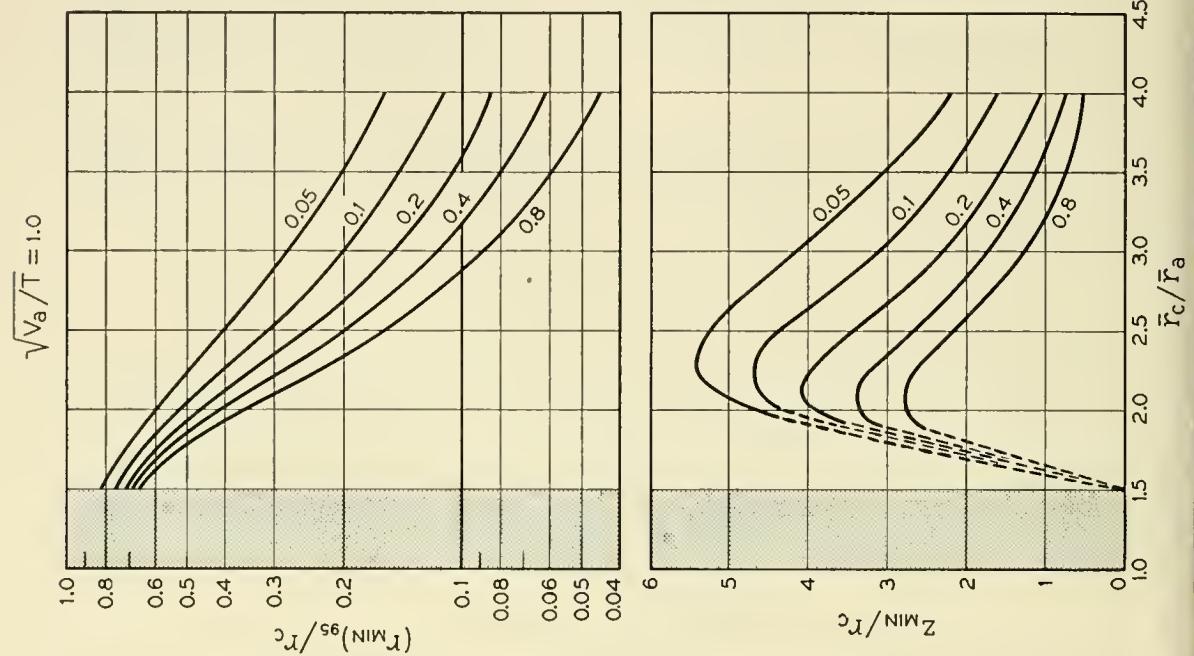
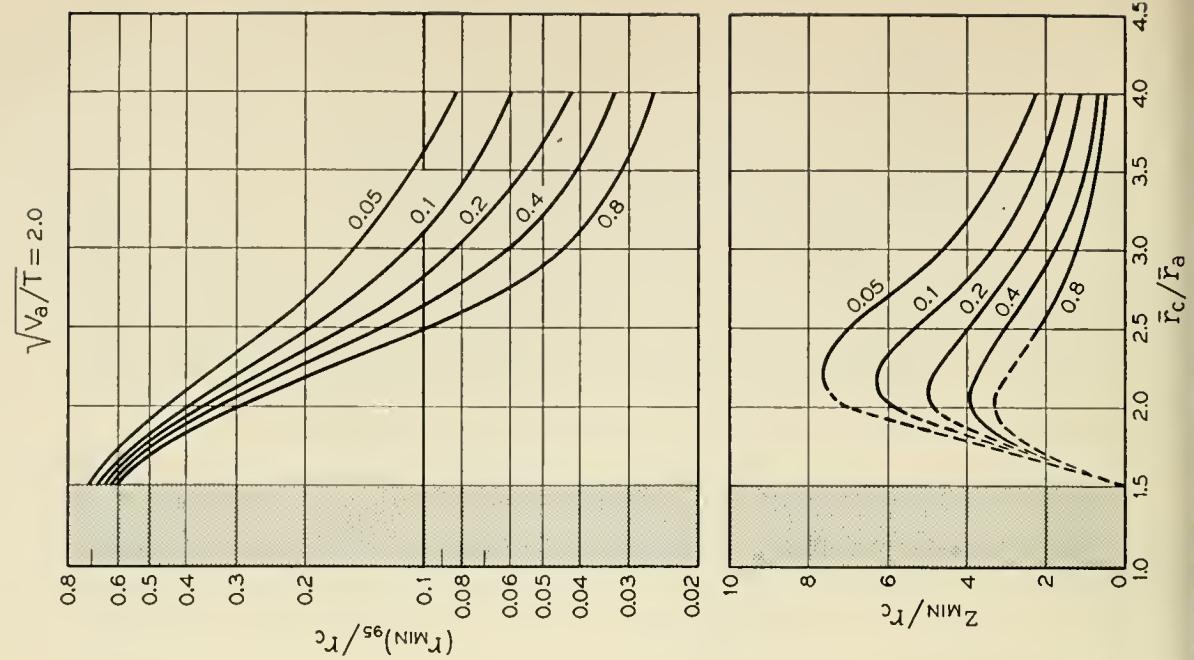
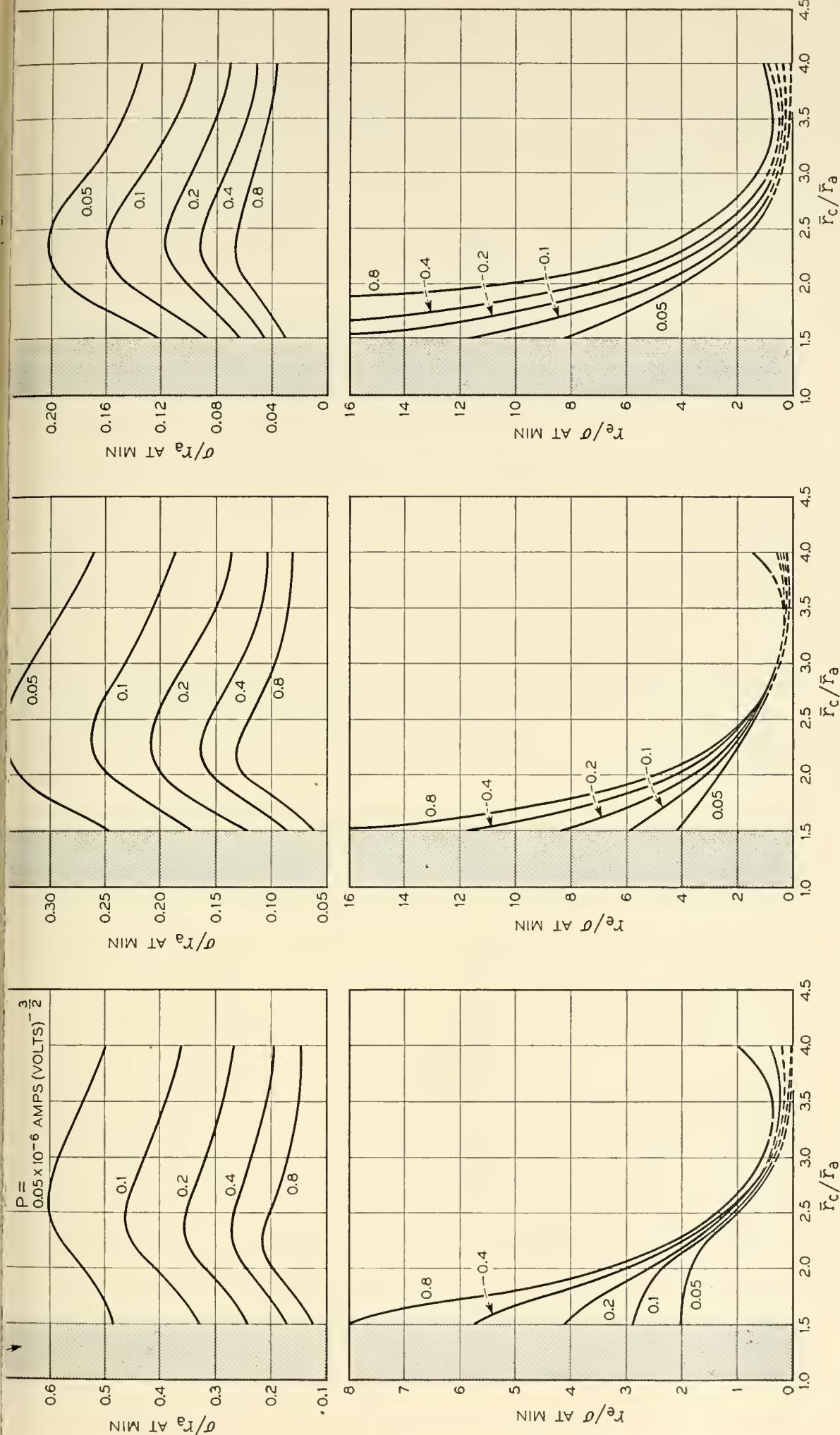


Fig. 13A — Design curves for Pierce-type electron gun considering transverse thermal velocities of electrons on emission from the cathode and

Fig. 13B — Design curves for Pierce-type electron guns considering transverse thermal velocities of electrons on emission from the cathode and an anode lens correction of $T = 1.1$. (for values of $\tau_e/\sigma < 0.5$, see footnote to Section 5D.)



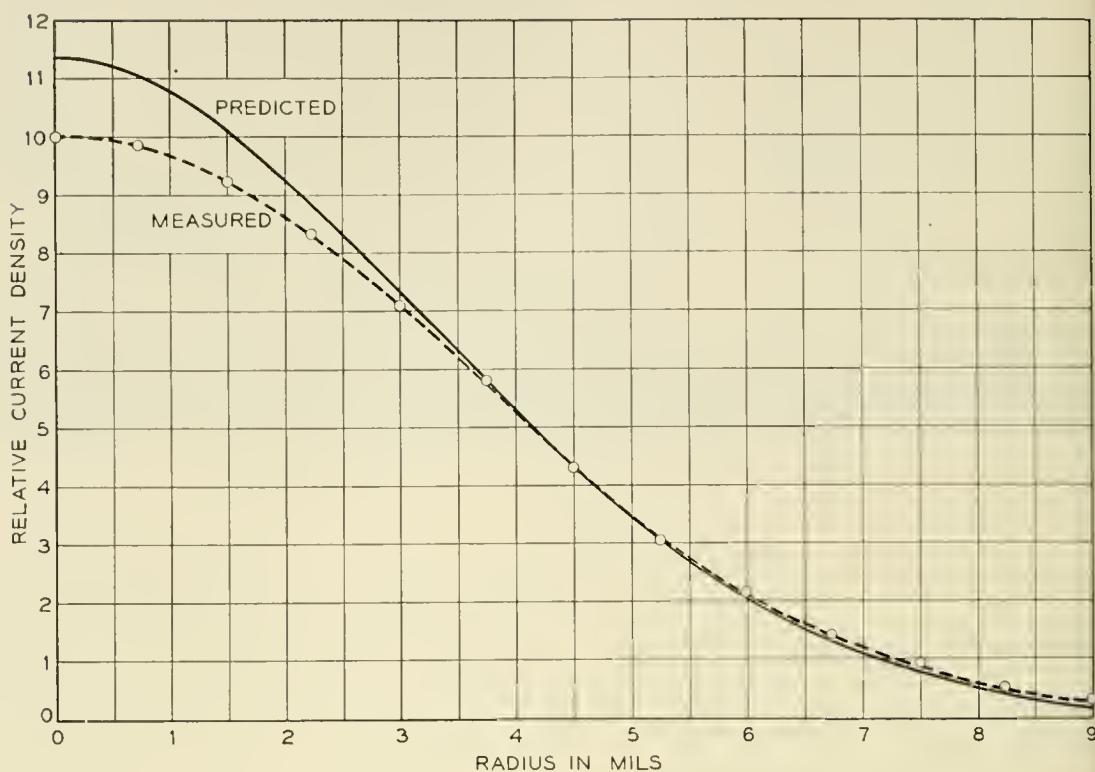


Fig. 14 — Current density distribution in a transverse plane located where the 95 per cent radius is a minimum. The predicted and measured curves are normalized to contain the same total current. (The corresponding prediction from the universal beam spread curve would show a step function with a constant relative current density of 64.2 for $r < 1.2$ mils and zero beyond.) The gun parameters are given in Section 5D.

and generally Gaussian in form, thereby indicating uniform cathode emission and proper boundary conditions at the edge of the beam near the cathode. The effect of positive ions on the beam shape was in every case reduced to negligible proportions, either by using special pulse techniques,⁷ or by applying a small voltage gradient along the axis of the beam.

B. Comparison of the Experimentally Measured Spreading of a Beam with that Predicted Theoretically

From the experimentally obtained plots of current density versus radius at several axial positions along the beam, we have obtained at each position (by integrating to find the total current within any radius) a value for the radius, r_{95} , of that circle which encompasses 95 per cent of the beam. For brevity, we call the resulting plots of r_{95} versus axial distance, "beam profiles". The experimental profile for the gun described in Section 5D is shown as curve A in Fig. 15(a). Curve B shows the profile as predicted by the methods of this paper and obtained from Fig. 12. Curve C is the corresponding profile which one obtains by the Hines-Cutler method,⁶ and Curve D represents r_{95} as obtained from the

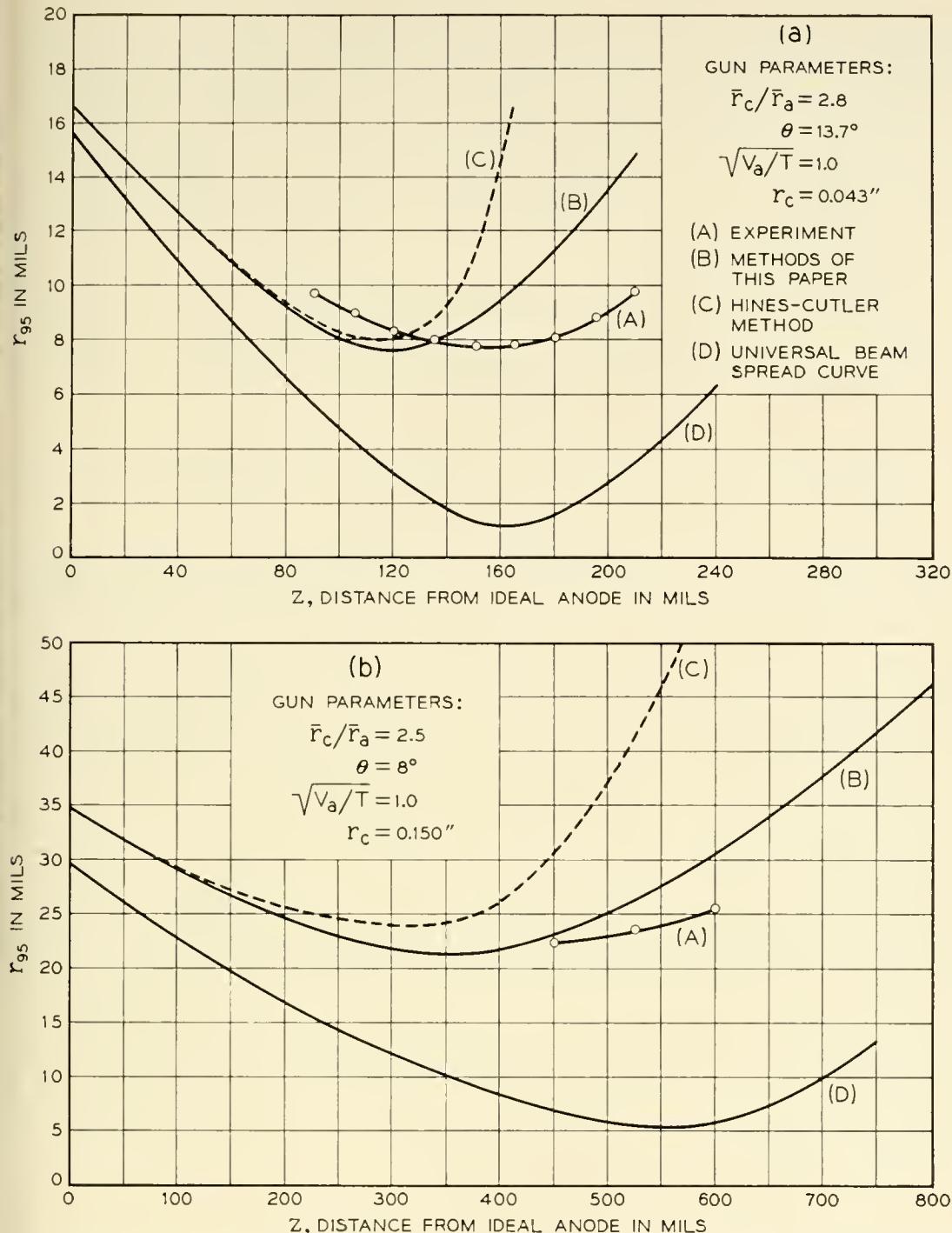


Fig. 15 — Beam profiles (using an anode lens correction of $\Gamma = 1.1$ and the gun parameters indicated) as obtained (A) from experiment, (B) by the methods of this paper, (C) Hines-Cutler method, (D) by use of the universal beam spread curve.

universal beam spread curve¹² (i.e., under the assumption of laminar flow and gradual variations of beam radius with distance). Note that in each case a value of 1.1 has been used for the correction factor, Γ , representing the excess divergence of the anode lens. The agreement in $(r_{95})_{\min}$ as obtained from Curves A and B is remarkably good, but the axial position of $(r_{95})_{\min}$ in Curve A definitely lies beyond the correspond-

ing minimum position in Curve B. Fortunately, in the gun design stage, one is usually more concerned with the value of $(r_{95})_{\min}$ than with its exact axial location. The principal need for knowing the axial location of the minimum is to enable the axial magnetic field to build up suddenly in this neighborhood. However, since this field is normally adjusted experimentally to produce best focusing, an approximate knowledge of z_{\min} is usually adequate.

In Fig. 15b we show a similar set of experimental and theoretical beam profiles for another gun. The relative profiles are much the same as in Fig. 15a, and all of several other guns measured yield experimental points similarly situated with respect to curves of Type B.

C. Comparison of Experimental and Theoretical Current Density Distributions where the Minimum Beam Diameter is Reached

In Fig. 14 we have plotted the current density distribution we would have predicted in a transverse plane at z_{\min} for the example introduced in Section 5D. Here the experimental and theoretical curves are normalized to include the same total currents in their respective beams. The noticeable difference in predicted and measured current densities at the center of the beam does not appreciably alter the properties such a beam would have on entering a magnetic field because so little total current is actually represented by this central peak.

D. Variation of Beam Profile with Γ

All of the design charts have been based on a value of $\Gamma = 1.1$, which is typical of the values obtained by the methods of Section 3. When appreciably different values of Γ are appropriate, we can get some feeling for the errors involved, in using curves based on $\Gamma = 1.1$, by reference to Fig. 16. Here we show beam profiles as obtained by the methods of this paper for three values of Γ . The calculations are again based on the gun of Section 5D, and a value of just over 1.1 for Γ gives the experimentally obtained value for $(r_{95})_{\min}$.

7. SOME ADDITIONAL REMARKS ON GUN DESIGN

In previous sections we have not differentiated between the voltage on the accelerating anode of the gun and the final beam voltage. It is important, however, that the separate functions of these two voltages be kept clearly in mind: The accelerating anode determines the total current drawn and largely controls the shaping of the beam; the final beam voltage is, on the other hand, chosen to give maximum interaction between the electron beam and the electromagnetic waves traveling along the slow wave circuit. As a consequence of this separation of func-

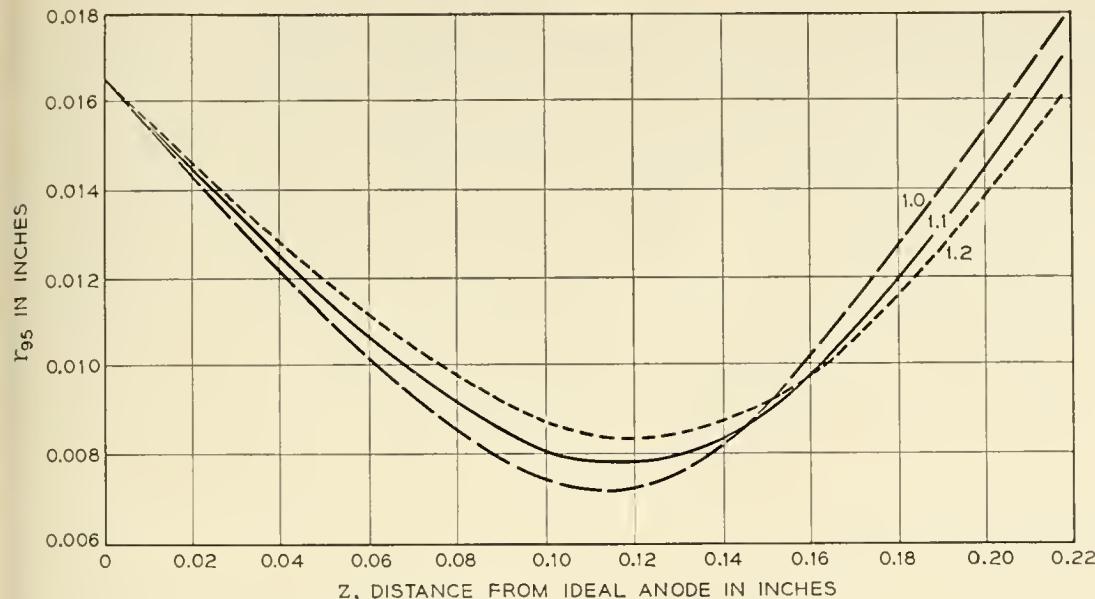


Fig. 16 — Beam profiles as obtained by the methods of this paper for the gun parameters given in Section 5D. Curves are shown for three values of the anode lens correction, viz. $\Gamma = 1.0, 1.1$, and 1.2 .

tions, it is found that some beams which are difficult or impossible to obtain with a single Pierce-gun acceleration to final beam voltage may be obtained more easily by using a lower voltage on the gun anode. The acceleration to final beam voltage is then accomplished after the beam has entered a region of axial magnetic field.

Suppose, for example, that one wishes to produce a 2-ma, 4-kv beam with $(r_{95}/r_c) = 0.25$. If the cathode temperature is 1000°K , and the gun anode is placed at a final beam voltage of 4 kv, we have $\sqrt{V_a/T} = 2$ and $P = 0.008$. From the top set of curves under $\sqrt{V_a/T} = 2$ in Fig. 13, we find (by using a fairly crude extrapolation from the curves shown) that a ratio of $\bar{r}_c/\bar{r}_a \approx 3.5$ is required to produce such a beam. The value of (r_e/σ) at z_{\min} is therefore less than about 0.2 so that there is little semblance of laminar flow here. On the other hand we might choose $V_z = 250$ volts so that $\sqrt{V_a/T} = 0.5$ and $P = 0.51$. From Fig. 13 we then obtain $\bar{r}_c/\bar{r}_a = 2.6$ and $(r_e/\sigma)_{\min} = 0.8$ for the same ratio of $r_{95}/r_c (= 0.25)$. While the flow could still hardly be called laminar, it is considerably more ordered than in the preceding case. Here we have included no correction for the (convergent) lens effect associated with the post-anode acceleration to the final beam voltage, $V = 4$ kv.

Calculations of the Hines-Cutler type will always predict, for a given set of gun parameters and a specified anode lens correction, a minimum beam size which is larger than that predicted by the methods of this paper. Nevertheless, in many cases the difference between the minimum sizes predicted by the two theories is negligible so long as the same anode lens correction is used. The extent to which the two theories agree ob-

viously depends on the magnitude of r_e/σ . When r_e/σ as calculated by the Hines-Cutler method (with a lens correction added) remains greater than about 2 throughout the range of interest, the difference between the corresponding values obtained for r_{95} will be only a few per cent. For these cases where r_e/σ does not get too small, the principal advantages of this paper are in the inclusion of a correction to the anode lens formula and in the comparative ease with which design parameters may be obtained. In other cases r_e/σ may become less than 1, and the theory presented in this paper has extended the basic Hines-Cutler approach so that one may make realistic predictions even under these less ideal conditions where the departure from a laminar-type flow is quite severe.

ACKNOWLEDGMENT

We wish to thank members of the Mathematical Department at B.T.L., particularly H. T. O'Neil and Mrs. L. R. Lee, for their help in programming the problem on the analog computer and in obtaining the large amount of computer data involved. In addition, we wish to thank J. C. Irwin for his help in the electrolytic tank work and both Mr. Irwin and W. A. L. Warne for their work on the beam analyzer.

REFERENCES

- Pierce, J. R., Rectilinear Flow in Beams, *J. App. Phys.*, **11**, pp. 548-554, Aug., 1940.
- Samuel, A. L., Some Notes on the Design of Electron Guns, *Proc. I.R.E.*, **33**, pp. 233-241, April, 1945.
- Field, L. M., High Current Electron Guns, *Rev. Mod. Phys.*, **18**, pp. 353-361, July, 1946.
- Davisson, C. J., and Calbick, C. J., Electron Lenses, *Phys. Rev.*, **42**, p. 580, Nov., 1932.
- Helm, R., Spangenburg, K., and Field, L. M., Cathode-Design Procedure for Electron Beam Tubes, *Elec. Comm.*, **24**, pp. 101-107, March, 1947.
- Cutler, C. C., and Hines, M. E., Thermal Velocity Effects in Electron Guns, *Proc. I.R.E.*, **43**, pp. 307-314, March, 1955.
- Cutler, C. C., and Saloom, J. A., Pin-hole Camera Investigation of Electron Beams, *Proc. I.R.E.*, **43**, pp. 299-306, March, 1955.
- Hines, M. E., Manuscript in preparation.
- Private communication.
- See for example, Zworykin, V. K., et al., *Electron Optics and the Electron Microscope*, Chapter 13, Wiley and Sons, 1945, or Klemperer, O., *Electron Optics*, Chapter 4, Cambridge Univ. Press, 1953.
- Brown, K. L., and Süsskind, C., The Effect of the Anode Aperture on Potential Distribution in a "Pierce" Electron Gun, *Proc. I.R.E.*, **42**, p. 598, March, 1954.
- See, for example, Pierce, J. R., *Theory and Design of Electron Beams*, p. 147, Van Nostrand Co., 1949.
- See Reference 6, p. 5.
- Langmuir, I. L., and Blodgett, K., Currents Limited by Space Charge Between Concentric Spheres, *Phys. Rev.*, **24**, p. 53, July, 1924.
- See Reference 12, p. 177.
- See Reference 12, Chap. X.

Theories for Toll Traffic Engineering in the U. S. A.*

By ROGER I. WILKINSON

(Manuscript received June 2, 1955)

Present toll trunk traffic engineering practices in the United States are reviewed, and various congestion formulas compared with data obtained on long distance traffic. Customer habits upon meeting busy channels are noted and a theory developed describing the probable result of permitting subscribers to have direct dialing access to high delay toll trunk groups.

Continent-wide automatic alternate routing plans are described briefly, in which near no-delay service will permit direct customer dialing. The presence of non-random overflow traffic from high usage groups complicates the estimation of correct quantities of alternate paths. Present methods of solving graded multiple problems are reviewed and found unadaptable to the variety of trunking arrangements occurring in the toll plan.

Evidence is given that the principal fluctuation characteristics of overflow-type of non-random traffic are described by their mean and variance. An approximate probability distribution of simultaneous calls for this kind of non-random traffic is developed, and found to agree satisfactorily with theoretical overflow distributions and those seen in traffic simulations.

A method is devised using "equivalent random" traffic, which has good loss predictive ability under the "lost calls cleared" assumption, for a diverse field of alternate route trunking arrangements. Loss comparisons are made with traffic simulation results and with observations in exchanges.

Working curves are presented by which multi-alternate route trunking systems can be laid out to meet economic and grade of service criteria. Examples of their application are given.

TABLE OF CONTENTS

1. Introduction.....	422
2. Present Toll Traffic Engineering Practice.....	423

* Presented at the First International Congress on the Application of the Theory of Probability in Telephone Engineering and Administration, Copenhagen, June 21, 1955.

3. Customers Dialing on Groups with Considerable Delay	431
3.1. Comparison of Some Formulas for Estimating Customers' NC Service on Congested Groups	434
4. Service Requirements for Direct Distance Dialing by Customers	436
5. Economics of Toll Alternate Routing	437
6. New Problems in the Engineering and Administration of Intertoll Groups Resulting from Alternate Routing	441
7. Load-Service Relationships in Alternate Route Systems	442
7.1. The "Peaked" Character of Overflow Traffic	443
7.2. Approximate Description of the Character of Overflow Traffic	446
7.2.1. A Probability Distribution for Overflow Traffic	452
7.2.2. A Probability Distribution for Combined Overflow Traffic Loads .	457
7.3. Equivalent Random Theory for Prediction of Amount of Traffic Over- flowing a Single Stage Alternate Route, and Its Character, with Lost Calls Cleared	461
7.3.1. Throwdown Comparisons with Equivalent Random Theory on Simple Alternate Routing Arrangements with Lost Calls Cleared	468
7.3.2. Comparison of Equivalent Random Theory with Field Results on Simple Alternate Routing Arrangements	470
7.4. Prediction of Traffic Passing Through a Multi-Stage Alternate Route Network	475
7.4.1. Correlation of Loss with Peakedness of Components of Non- Random Offered Traffic	481
7.5. Expected Loss on First Routed Traffic Offered to Final Route	482
7.6. Load on Each Trunk, Particularly the Last Trunk, in a Non-Slipped Alternate Route	486
8. Practical Methods for Alternate Route Engineering	487
8.1. Determination of Final Group Size with First Routed Traffic Offered Directly to Final Group	490
8.2. Provision of Trunks Individual to First Routed Traffic to Equalize Service	491
8.3. Area in Which Significant Savings in Final Route Trunks are Real- ized by Allowing for the Preferred Service Given a First Routed Traffic Parcel	494
8.4. Character of Traffic Carried on Non-Final Routes	495
8.5. Solution of a Typical Toll Multi-Alternate Route Trunking Arrange- ment: Bloomsburg, Pa	500
9. Conclusion	505
Acknowledgements	506
References	506
Abridged Bibliography of Articles on Toll Alternate Routing	507
Appendix I: Derivation of Moments of Overflow Traffic	507
Appendix II: Character of Overflow when Non-Random Traffic is Offered to a group of Trunks	511

1. INTRODUCTION

It has long been the stated aim of the Bell System to make it easily and economically possible for any telephone customer in the United States to reach any other telephone in the world. The principal effort in this direction by the American Telephone and Telegraph Company and its associated operating companies is, of course, confined to interconnecting the telephones in the United States, and to providing communication channels between North America and the other countries of the world. Since the United States is some 1500 miles from north to south and 3000 miles from east to west, to realize even the aim of fast

and economical service between customers is a problem of great magnitude; it has engaged our planning engineers for many years.

There are now 52 million telephones in the United States, over 80 per cent of which are equipped with dials. Until quite recently most telephone users were limited in their direct dialing to the local or immediately surrounding areas and long distance operators were obliged to build up a circuit with the aid of a "through" operator at each switching point.

Both speed and economy dictated the automatic build-up of long toll circuits without the intervention of more than the originating toll operator. The development of the No. 4-type toll crossbar switching system with its ability to accept, translate, and pass on the necessary digits (or equivalent information) to the distant office made this method of operation possible and feasible. It was introduced during World War II, and now by means of it and allied equipment, 55 per cent of all long distance calls (over 25 miles) are completed by the originating operator.

As more elaborate switching and charge-recording arrangements were developed, particularly in metropolitan areas, the distances which customers themselves might dial measurably increased. This expansion of the local dialing area was found to be both economical and pleasing to the users. It was then not too great an effort to visualize customers dialing to all other telephones in the United States and neighboring countries, and perhaps ultimately across the sea.

The physical accomplishment of nationwide direct distance dialing which is now gradually being introduced has involved, as may well be imagined, an immense amount of advance study and fundamental planning. Adequate transmission and signalling with up to eight intertoll trunks in tandem, a nationwide uniform numbering plan simple enough to be used accurately and easily by the ordinary telephone caller, provision for automatic recording of who called whom and how long he talked, with subsequent automatic message accounting, are a few of many problems which have required solution. How they are being met is a romantic story beyond the scope of the present paper. The references given in the bibliography at the end contain much of the history as well as the plans for the future.

2. PRESENT TOLL TRAFFIC ENGINEERING PRACTICE

There are today approximately 116,000 intertoll trunks (over 25 miles in length) in the Bell System, apportioned among some 13,000 trunk groups. A small segment of the 2,600 toll centers which they interconnect is shown in Fig. 1. Most of these intertoll groups are presently traffic engineered to operate according to one of several so-called T-schedules: T-8, T-15, T-30, T-60, or T-120. The number following T (T for Toll) is

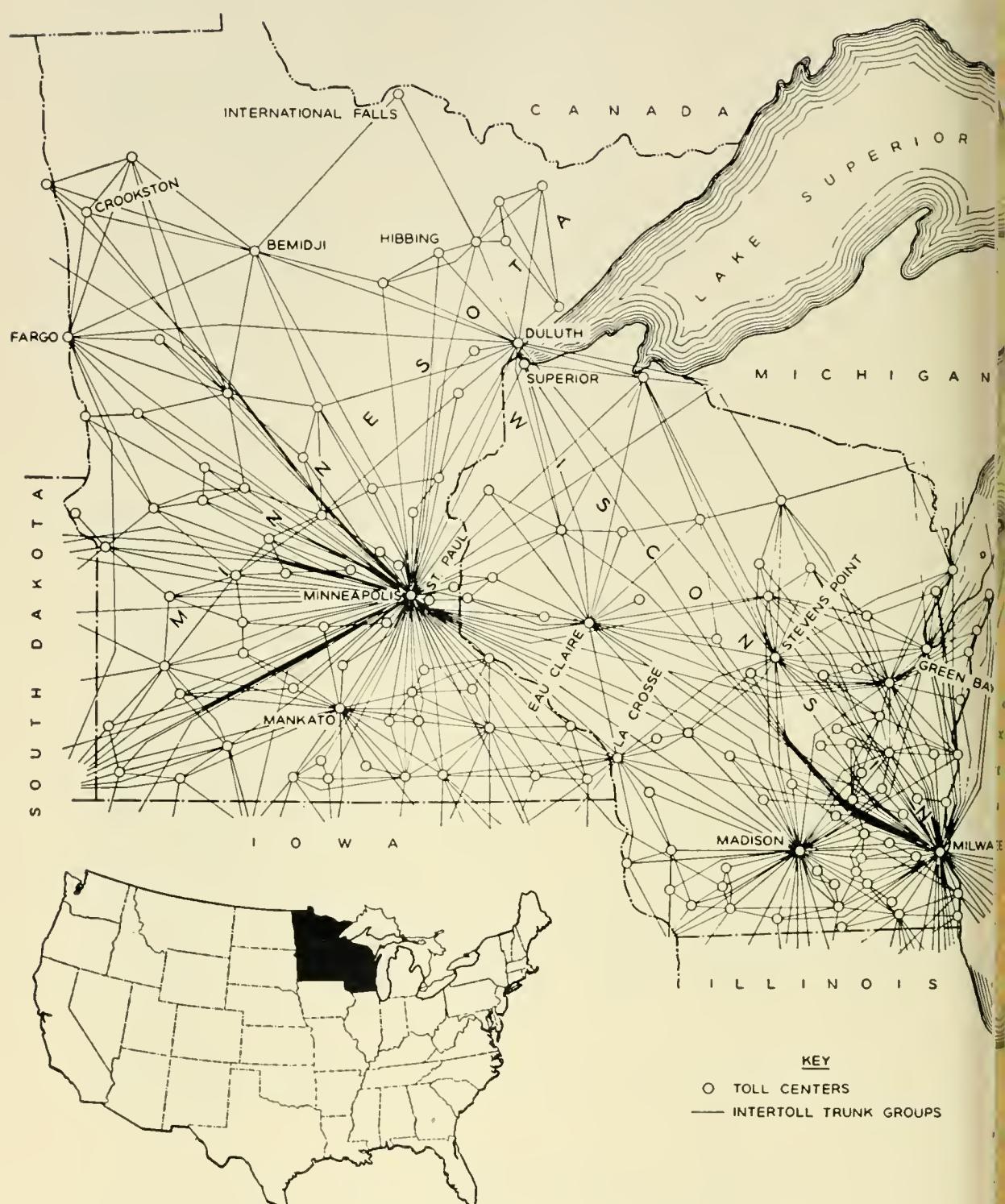


Fig. 1 — Principal intertoll trunk groups in Minnesota and Wisconsin.

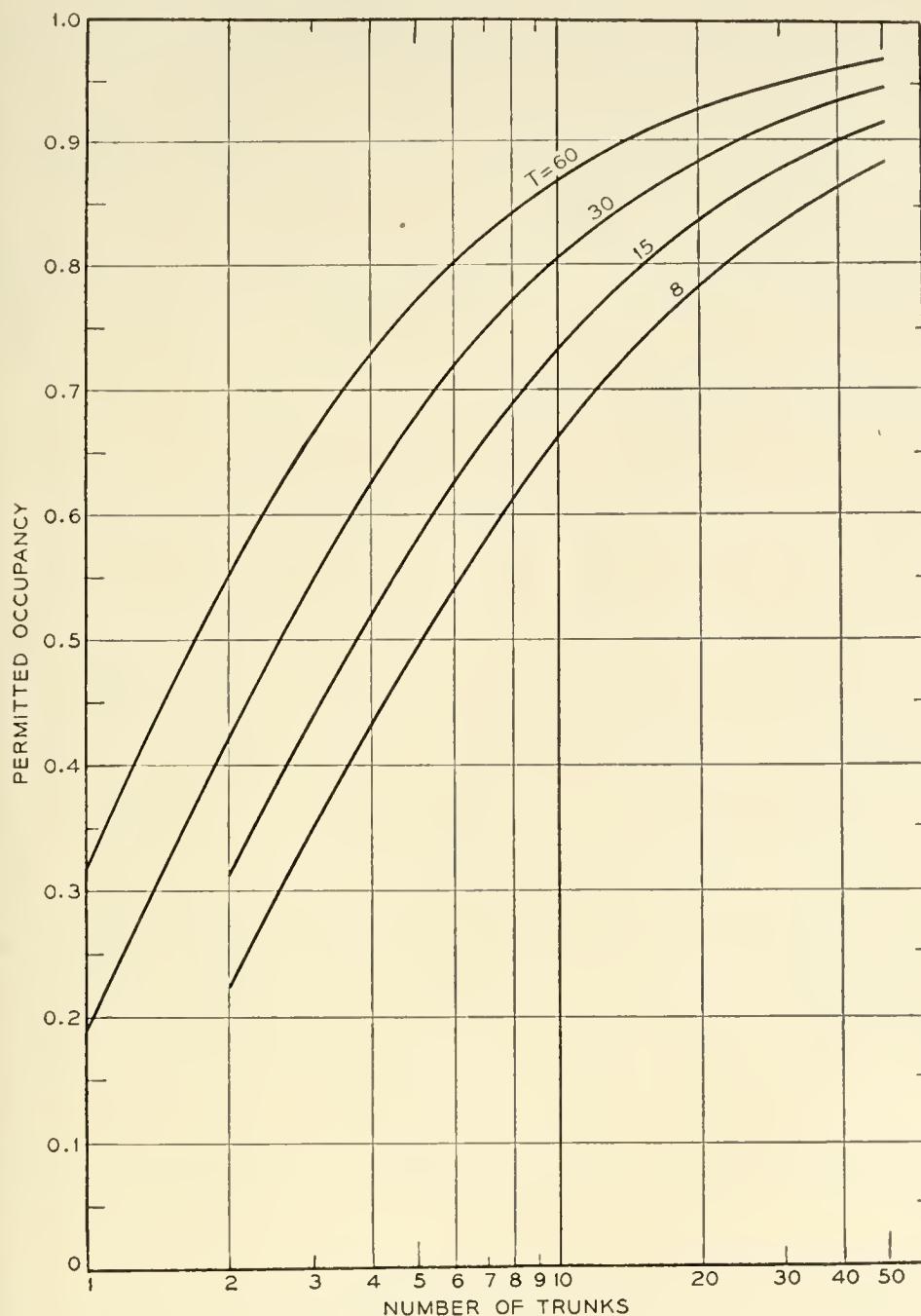


Fig. 2 — Permitted intertoll trunk occupancy for a 6.5-minute usage time per message.

the expected, or average, delay in seconds for calls to obtain an idle trunk in that group during the average Busy Season Busy Hour. In 1954 the system "average trunk speed" was approximately 30 seconds, resulting from operating the majority of the groups at a busy-hour trunking efficiency of 75 to 85 per cent in the busy season.

The T-engineering tables show permissible call minutes of use for a

wide range of group sizes, and several selections of message holding times. They were constructed following summarization of many observations of load and resultant average delays on ringdown (non-dial) intertoll trunks.¹ Fig. 2 shows the permissible occupancy (efficiency) of various trunk group sizes for 6.5 minutes of use per message, for a variety of T-schedules. It is perhaps of some interest that the best fitting curves relating average delay and load were found to be the well-known Pollaczek-Crommelin delay curves for constant holding time — this in spite of the fact that the circuit holding times were far indeed from having a constant value.

A second, and probably not uncorrelated, observation was that the per cent "No-Circuit" (NC) reported on the operators' tickets showed consistently lower values than were measured on group-busy timing devices. Although not thoroughly documented, this disparity has generally been attributed to the reluctance of an operator to admit immediately the presence of an NC condition. She exhibits a certain tolerance (very difficult to measure) before actually recording a delay which would require her to adopt a prescribed procedure for the subsequent handling of the call.* There are then two measures of the No-Circuit condition which are of some interest, the "NC encountered" by operators, and the "NC existing" as measured by timing devices.

It has long been observed that the distribution of numbers n of simultaneous calls found on T-engineered ringdown intertoll groups is in remarkable agreement with the individual probability terms of the Erlang "lost calls" formula,

$$f(n) = \frac{\frac{a'^n e^{-a'}}{n!}}{\sum_{n=0}^c \frac{a'^n e^{-a'}}{n!}} \quad (1)$$

where c = number of paths in the group,

a' = an enhanced average load submitted such that

$a'[1 - E_{1,c}(a')] = L$, the actual load carried, and

$E_{1,c}(a') = f(c) =$ Erlang loss probability (commonly called Erlang B in America).

An example of the agreement of observations with (1) is shown in Fig. 3, where the results of switch counts made some years ago on many ringdown circuit groups of size 3 are summarized. A wide range of "sub-

* Upon finding No-Circuit, an operator is instructed to try again in 30 seconds and 60 seconds (before giving an NC report to the customer), followed by additional attempts 5 minutes and 10 minutes later if necessary.

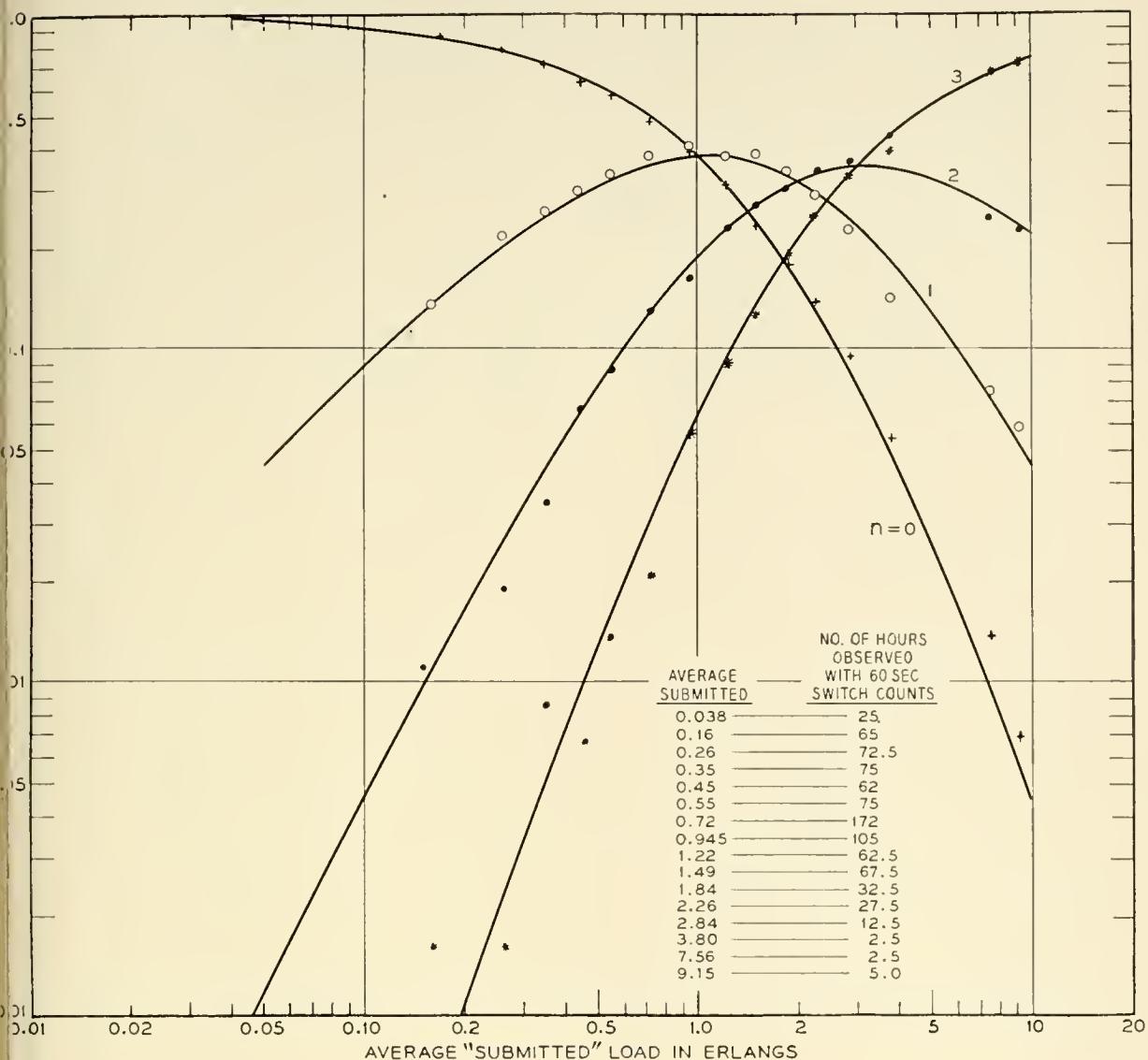


Fig. 3 — Distributions of simultaneous calls on three-trunk toll groups at Albany and Buffalo.

mitted" loads a' to produce the observed carried loads is required. On Fig. 4 are shown the corresponding comparisons of theory and observations for the proportions of time all paths are busy ("NC Existing") for 2-, 4-, 5-, 7-, and 9-circuit groups. Good agreement has also been observed for circuit groups up to 20 trunks. This has been found to be a stable relationship, in spite of the considerable variation in the actual practices in ringdown operation on the resubmission of delayed calls. Since the estimation of traffic loads and the subsequent administration of ringdown toll trunks has been performed principally by means of Group Busy Timers (which cumulate the duration of NC time), the Erlang relationship just described has been of great importance.

With the recent rapid increase in operator dialed intertoll groups, it might be expected that the above discrepancy between "% NC encountered" and "% NC existing" would disappear—for an operator now initiates each call unaware of the momentary state of the load on any particular intertoll group. By the use of peg count meters (which count calls offered) and overflow call counters, this change has in fact been observed to occur. Moreover, since the initial re-trial intervals are commonly fairly short (30 seconds) subsequent attempts tend to find some of the previous congestion still existing, so that the ratio of overflow to peg count readings now exceeds slightly the "% NC existing." This situation is illustrated in Fig. 5, which shows data taken on an operator-

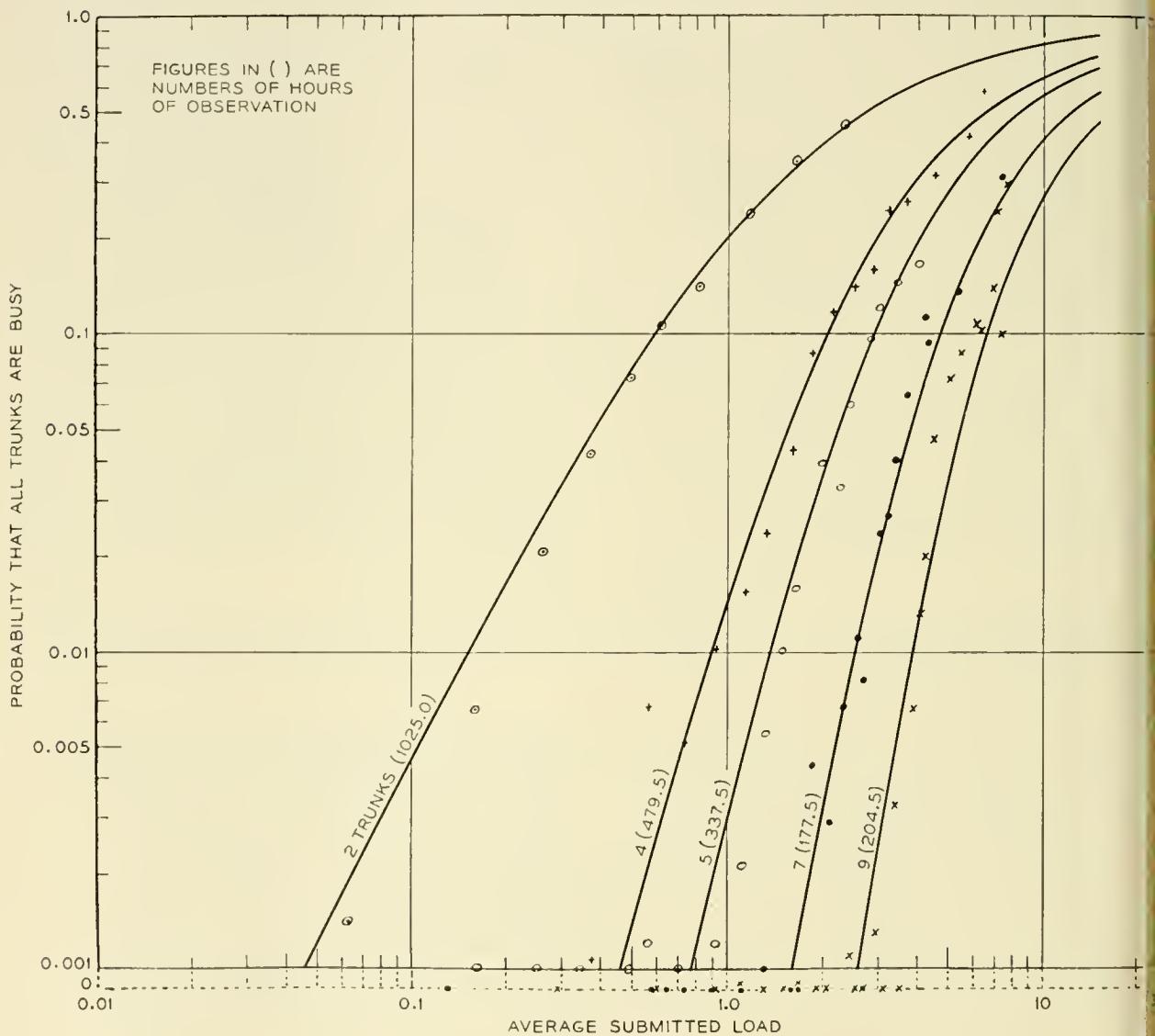


Fig. 4—Observed proportions of time all trunks were busy on Albany and Buffalo groups of 2, 4, 5, 7, and 9 trunks.

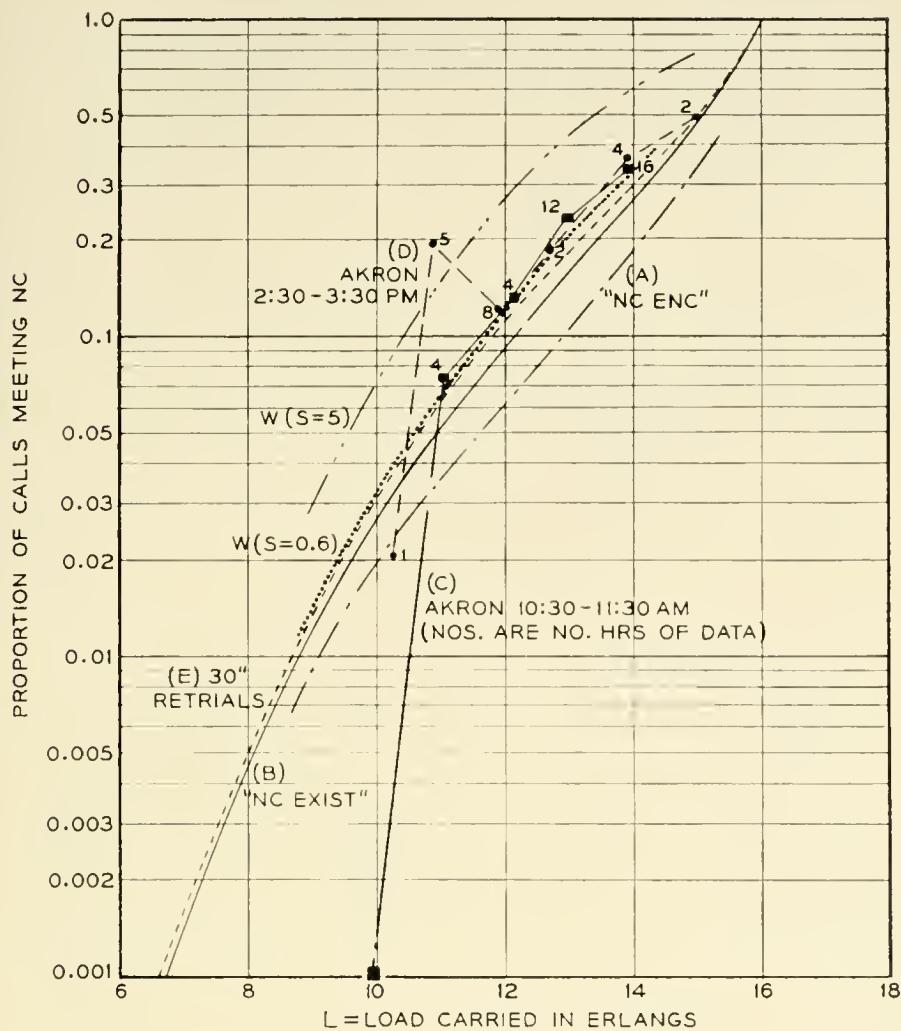
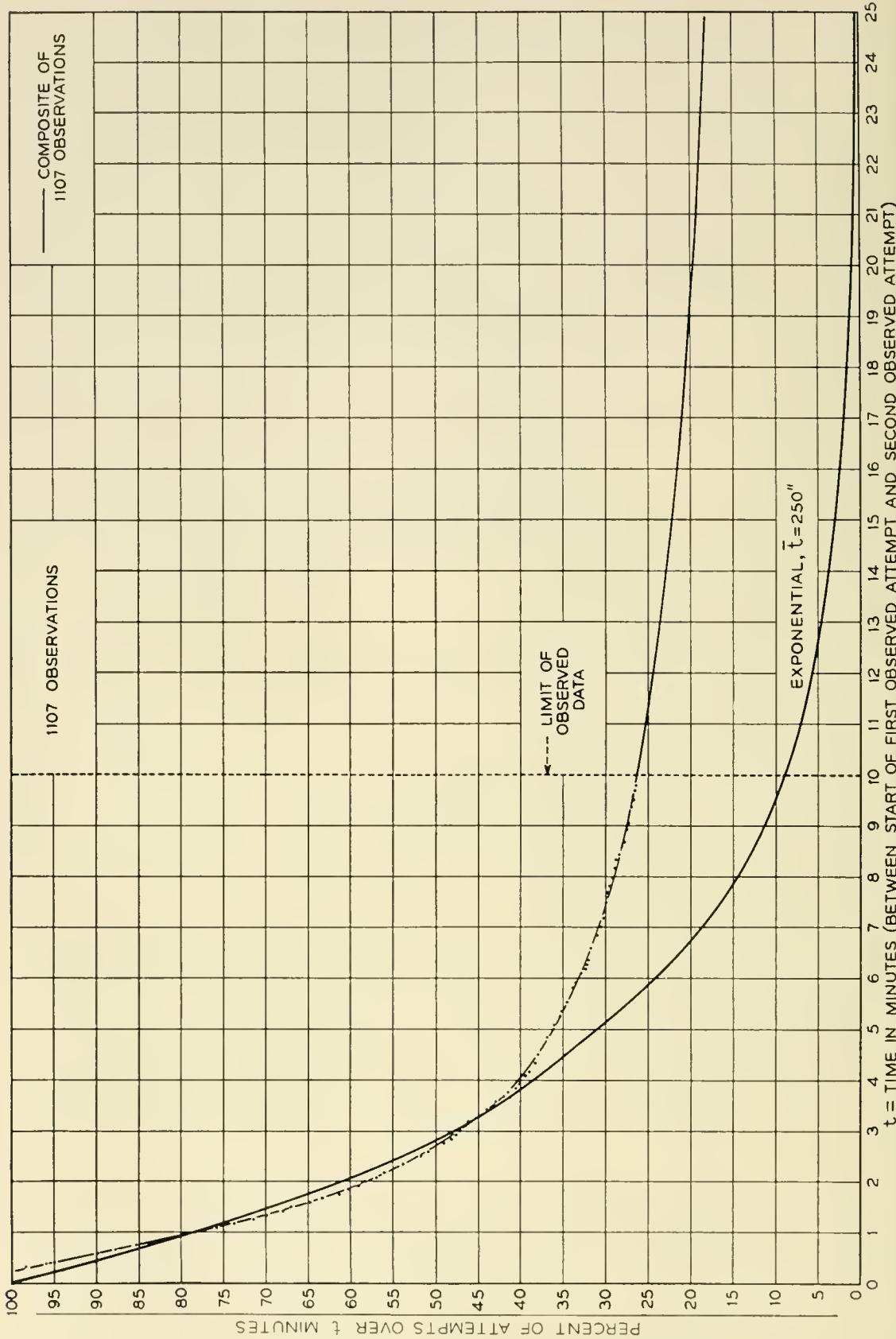


Fig. 5.—Comparison of NC data on a 16-trunk T-engineered toll group with various load versus NC theories.

dialed T-engineered group of 16 trunks between Newark, N. J., and Akron, Ohio. Curve A shows the empirically determined "NC encountered" relationship described above for ringdown operation; Curve B gives the corresponding theoretical "NC existing" values. Lines C and D give the operator-dialing results, for morning and afternoon busy hours. The observed points are now seen generally to be significantly above Curve B.*

At the same time as this change in the "NC encountered" was occurring, due to the introduction of operator toll dialing, there seems to have been little disturbance to the traditional relationship between load

* The observed point at 11 erlangs which is clearly far out of agreement with the remainder of the data was produced by a combination of high-trend hours and an hour in which an operator apparently made many re-trials in rapid succession.



carried and “% NC existing.” C. J. Truitt of the A.T. & T. Co. studied a number of operator-dialed T-engineered groups at Newark, New Jersey, in 1954 with a traffic usage recorder (TUR) and group-busy timers, and found the relationship of equation (1) still good. (This analysis has not been published.)

A study by Dr. L. Kosten has provided an estimate of the probability that when an NC condition has been found, it will also appear at a time τ later.² When this modification is made, the expected load-versus-NC relationship is shown by Curve E on Fig. 5. (The re-trial time here was taken as the operators’ nominal 30 seconds; with 150-second circuit-use time the return is 0.2 holding time.) The observed NC’s are seen to lie slightly above the E-curve. This could be explained either on the basis that Kosten’s analysis is a lower limit, or that the operators did not strictly observe the 30-second return schedule, or, more probably, a combination of both.

3. CUSTOMERS DIALING ON GROUPS WITH CONSIDERABLE DELAY

It is not to be expected that customers could generally be persuaded to wait a designated constant or minimum re-trial time on their calls which meet the NC condition. Little actual experience has been accumulated on customers dialing long distance calls on high-delay circuits. However, it is plausible that they would follow the re-trial time distributions of customers making local calls, who encounter paths-busy or line-busy signals (between which they apparently do not usually distinguish). Some information on re-trial times was assembled in 1944 by C. Clos³ by observing the action of customers who received the busy signal on 1,100 local calls in the City of New York. As seen in Fig. 6, the return times, after meeting “busy,” exhibit a marked tendency toward the exponential distribution, after allowance for a minimum interval required for re-dialing.

An exponential distribution with average of 250 seconds has been fitted by eye on Fig. 6, to the earlier — and more critical — customer return times. This may seem an unexpectedly long wait in the light of individual experience; however it is probably a fair estimate, especially since, following the collection of the above data, it has become common practice for American operating companies in their instructional literature to advise customers receiving the busy signal to “hang up, wait a few minutes, and try again.”

The mathematical representation of the situation assuming exponential return times is easily formulated. Let there be x actual trunks, and

imagine y waiting positions, where y is so large that few calls are rejected.* Assume that the offered load is a erlangs, and that the calls have exponential conversation holding times of unit average duration. Finally let the average return time for calls which have advanced to the waiting positions, be $1/s$ times that of the unit conversation time. The statistical equilibrium equation can then be written for the probability $f(m, n)$ that m calls are in progress on the x trunks and n calls are waiting on the y storage positions:

$$\begin{aligned} f(m, n) = & af(m - 1, n) dt + s(n + 1) f(m - 1, n + 1) dt \\ & + (m + 1)f(m + 1, n) dt + af(x, n - 1) dt \star \\ & + [1 - (a\star\star\star + sn\star\star) dt - m dt]f(m, n) \end{aligned} \quad (2)$$

where $0 \leq m \leq x$, $0 \leq n \leq y$, and the special limiting situations are recognized by:

- ★ Include term only when $m = x$
- ★★ Omit sn when $m = x$
- ★★★ Omit a when $m = x$ and $n = y$

Equation (2) reduces to

$$\begin{aligned} (a\star\star\star + sn\star\star + m)f(m, n) = & af(m - 1, n) \\ & + s(n + 1)f(m - 1, n + 1) \\ & + (m + 1)f(m + 1, n) + af(x, n - 1)\star, \end{aligned} \quad (3)$$

Solution of (3) is most easily effected for moderate values of x and y by first setting $f(x, y) = 1.000000$ and solving for all other $f(m, n)$ in terms of $f(x, y)$. Normalizing through $\sum_{m=0}^x \sum_{n=0}^y f(m, n) = 1.0$, then gives the entire $f(m, n)$ array.

The proportion of time "NC exists," will, of course be

$$\sum_{n=0}^y f(x, n) \quad (4)$$

and the load carried is

$$L = \sum_{m=0}^x \sum_{n=0}^y mf(m, n) \quad (5)$$

The proportion of call attempts meeting NC, including all re-trials

* The quantity y can also be chosen so that some calls are rejected, thus roughly describing those calls abandoned after the first attempt.

will be

$$\begin{aligned}
 W(x, a, s) &= \frac{\text{Expected overflow calls per unit time}}{\text{Expected calls offered per unit time}} \\
 &= \frac{\sum_{n=0}^y (a + sn)f(x, n)}{\sum_{m=0}^x \sum_{n=0}^y (a + sn)f(m, n)} = \frac{s\bar{n} + af(x, y)}{a + s\bar{n}}
 \end{aligned} \tag{6}$$

in which $\bar{n} = \sum_{m=0}^x \sum_{n=0}^y nf(m, n)$. And when y is chosen so large that $f(x, y)$ is negligible, as we shall use it here,

$$L = a \tag{5'}$$

$$W(x, a, s) = \frac{s\bar{n}}{a + s\bar{n}} \tag{6'}$$

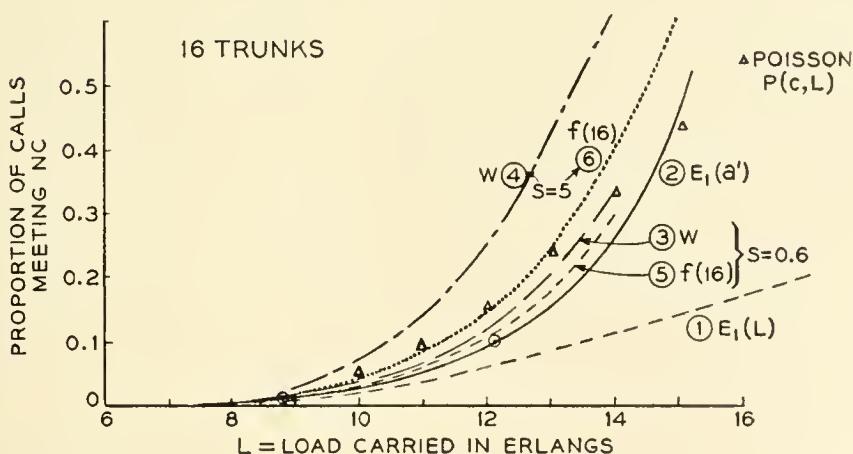
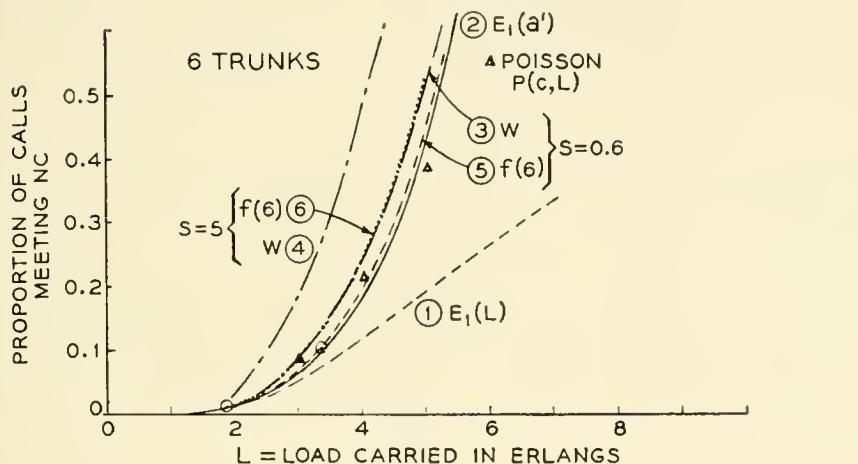


Fig. 7 — Comparison of trunking formulas.

This formula provides a means for estimating the grade of service which customers might be expected to receive if asked to dial their calls over moderate-delay or high-delay trunk groups. For a circuit use length of 150 seconds, and an average return time of 250 seconds (as on Fig. 6), both exponential, the load-versus-proportion-NC curves for 6 and 16 trunks are given as curves (3) on Fig. 7. For example with an offered (= carried) load of $a = 4.15$ erlangs on 6 trunks we should expect to find 27.5 per cent of the total attempts resulting in failure.

For comparison with a fixed return time of NC-calls, the W -formula curves for exponential returns of 30 seconds ($s = .5$) and 250 seconds ($s = 0.6$) averages are shown on Fig. 5. The first is far too severe an assumption for operator performance, giving NC's nearly double those actually observed (and those given by theory for a 30-second constant return time). The 250-second average return, however, lies only slightly above the 30-second constant return curve and is in good agreement with the data. Although not logically an adequate formula for interpreting Peg Count and Overflow registrations on T-engineered groups under operator dialing conditions, the W -formula apparently could be used for this purpose with suitable s -values determined empirically.

3.1. Comparison of Some Formulas for Estimating Customers' NC Service on Congested Groups

As has been previously observed, a large proportion of customers who receive a busy signal, return within a few minutes (on Fig. 6, 75 per cent of the customers returned within 10 minutes). It is well known too, that under adverse service conditions subscriber attempts (to reach a particular distant office for example) tend to produce an inflated estimate of the true offered load. A count of calls carried (or a direct measurement of load carried) will commonly be a closer estimate of the offered load than a count of attempts. An exception may occur when a large proportion of attempts is lost, indicating an offered load possibly in excess even of the number of paths provided. Under the latter condition it is difficult to estimate the true offered load by any method, since not all the attempts can be expected to return repeatedly until served; instead, a significant number will be abandoned somewhere through the trials. In most other circumstances, however, the carried load will prove a reasonably good estimate of the true offered load in systems not provided with alternate paths.

This is a matter of especial interest for both toll and local operation in America since principal future reliance for load measurement is ex-

pected to be placed on automatically processed TUR data, and as the TUR is a switch counting device the results will be in terms of load carried. Moreover, the quantity now obtained in many local exchanges is load carried.* Visual switch counting of line finders and selectors off-normal is widely practiced in step-by-step and panel offices; a variety of electromechanical switch counting devices is also to be found in crossbar offices. It is common to take load-carried figures as equal to load-offered when using conventional trunking tables to ascertain the proper provision of trunks or switches. Fig. 7 compares the NC predictions made by a number of the available load-loss formulas when load carried is used as the entry variable.

The lowest curves (1) on Fig. 7 are from the Erlang lost calls formula E_1 (or B) with load carried L used as the offered load a . At low losses, say 0.01 or less, either L or $a = L/[1 - E_1(a)]$ can be used indiscriminately as the entry in the E_1 formula. If however considerably larger losses are encountered and calls are not in reality "cleared" upon meeting NC, it will no longer be satisfactory to substitute L for a . In this circumstance it is common to calculate a fictitious load a' to submit to the c paths such that the load carried, $a'[1 - E_{1,c}(a')]$, equals the desired L . (This was the process used in Section 2 to obtain "% NC existing.") The curves (2) on Fig. 7 show this relation; physically it corresponds to an initially offered load of L erlangs (or L call arrivals per average holding time), whose overflow calls return again and again until successful but without disturbing the randomness of the input. Thus if the loss from this enhanced random traffic is E , then the total trials seen per holding time will be $L(1 + E + E^2 + \dots) = L/(1 - E) = a'$, the apparent arrival rate of new calls, but actually of new calls plus return attempts.

The random resubmission of calls may provide a reasonable description of operation under certain circumstances, presumably when re-trials are not excessive. Kosten² has discussed the dangers here and provided upper and lower limit formulas and curves for estimating the proportions of NC's to be expected when re-trials are made at any specified fixed return time. His lower bounds (lower bound because the change in congestion character caused by the returning calls is ignored) are shown by open dots on Fig. 7 for return times of 1.67 holding times. They lie above curves (2) (although only very slightly because of the relatively long return time) since they allow for the fact that a call shortly returning

* In fact, it is difficult to see how any estimate of offered load, other than carried load, can be obtained with useful reliability.

after meeting a busy signal will have a higher probability of again finding all paths busy, than would a randomly originated call.

The curves (3) show the W -formula previously developed in this section, which contemplates exponential return times on all NC attempts. The average return time here is also taken as 1.67 holding times. These curves lie higher than Kosten's values for two reasons. First, the altered congestion due to return calls is allowed for; and second, with exponential returns nearly two-thirds of the return times are shorter than the average, and of these, the shortest ones will have a relatively high probability of failure upon re-trying. If the customers were to return with exponential times after waiting an average of only 0.2 holding time (e.g., 30 seconds wait for 150-second calls) the W -curves would rise markedly to the positions shown by (4).

Curves (5) and (6) give the proportions of time that all paths are busy (equation 4) under the W -formula assumptions corresponding to NC curves (3) and (4) respectively; their upward displacement from the random return curves (2) reflects the disturbance to the group congestion produced by the non-random return of the delayed calls. (The limiting position for these curves is, of course, given by Erlang's E_2 (or C) delay formula.) As would be expected, curve (6) is above (5) since the former contemplates exponential returns with average of 0.2 holding time, as against 1.67 for curve (5). Neither the (5)-curves nor the open dots of constant 30-second return times show a marked increase over curves (2). This appears to explain why the relationship of load carried versus "NC existing" (as charted in Figs. 3 and 4) was found so insensitive to variable operating procedures in handling subsequent attempts in toll ring-down operation, and again, why it did not appreciably change under operator dialing.

Finally, through the two fields of curves on Fig. 7 is indicated the Poisson summation $P(c, L)$ with load carried L used as the entering variable. The fact that these values approach closely the (2) and (3) sets of curves over a considerable range of NC's should reassure those who have been concerned that the Poisson engineering tables were not useful for losses larger than a few per cent.*

4. SERVICE REQUIREMENTS FOR DIRECT DISTANCE DIALING BY CUSTOMERS

As shown by the W -curves (3) on Fig. 7, the attempt failures by customers resulting from their tendency to re-try shortly following an NC

* Reference may be made also to a throwdown by C. Clos (Ref. 3) using the return times of Fig. 6; his "% NC" results agreed closely with the Poisson predictions.

would be expected to exceed slightly the values for completely random re-trials. These particular curves are based on a re-trial interval of 1.67 times the average circuit-use time. Such moderation on the part of the customer is probably attainable through instructional literature and other means if the customer believes the "NC" or "busy" to be caused by the called party's actually using his telephone (the usual case in local practice). It would be considerably more difficult, however, to dissuade the customer from re-trying at a more rapid rate if the circuit NC's should generally approach or exceed actual called-party busies, a condition of which he would sooner or later become aware. His attempts might then be more nearly described by the (4) curves on Fig. 7 corresponding to an average exponential return of only 0.2 holding time—or even higher. Such a result would not only displease the user, but also result in the requirement of increased switching control equipment to handle many more wasted attempts.

If subscribers are to be given satisfactory direct dialing access to the intertoll trunk network, it appears then that the probability of finding NC even in the busy hours must be kept to a low figure. The following engineering objective has tentatively been selected: *The calls offered to the "final" group of trunks in an alternate route system should receive no more than 3 per cent NC(P.03) during the network busy season busy hour.* (If there are no alternate routes, the direct group is the "final" route.)

Since in the nationwide plan there will be a final route between each of some 2,600 toll centers and its next higher center, and the majority of calls offered to high usage trunks will be carried without trying their final route (or routes), the over-all point-to-point service, while not easy to estimate, will apparently be quite satisfactory for customer dialing.

5. ECONOMICS OF TOLL ALTERNATE ROUTING

In a general study of the economics of a nationwide toll switching plan, made some years ago by engineers of the American Telephone and Telegraph Company, it was concluded that a toll line plant sufficient to give the then average level of service (about T-40) with ordinary single-route procedures could, if operated on a multi-alternate route basis, give the desired P.03 service on final routes with little, if any, increase in toll line investment.* On the other hand to attain a similar P.03 grade of service by liberalizing a typical intertoll group of 10 trunks working presently

* This, of course, does not reflect the added costs of the No. 4 switching equipment.

at a T-40 grade of service and an occupancy of 0.81 would require an increase of 43 per cent (to 14.3 trunks), with a corresponding decrease in occupancy to 0.57. The possible savings in toll lines with alternate routing are therefore considerable in a system which must provide a service level satisfactory for customer dialing.

In order to take fullest advantage of the economies of alternate routing, present plans call for five classes of toll offices. There will be a large number of so-called End Offices, a smaller number of Toll Centers, and progressively fewer Primary Centers (about 150), Sectional Centers (about 40) and Regional Centers (9), one of which will be the National Center, to be used as the "home" switching point of the other eight Regional Centers.* Primary and higher centers will be arranged to perform automatic alternate routing and are called Control Switching Points (CSP's). Each class of office will "home" on a higher class of office (not necessarily the next higher one); the toll paths between them are called "final routes." As described in Section 4, these final routes will be provided to give low delays, so that between each principal toll point and every other one there will be available a succession of approximately P.03 engineered trunk groups. Thus if the more direct and heavily loaded interconnecting paths commonly provided are busy there will still be a good chance of making immediate connection over final routes.

Fig. 8 illustrates the manner in which automatic alternate routing will operate in comparison with present-day operator routing. On a call from Syracuse, N. Y., to Miami, Florida, (a distance of some 1,250 miles), under present-day operation, the Syracuse operator signals Albany, and requests a trunk to Miami. With T-schedule operation the Syracuse-Miami traffic might be expected to encounter as much as 25 per cent NC during the busy hour, and approximately 4 per cent NC for the whole day, producing perhaps a two-minute over-all speed of service in the busy season.

With the proposed automatic alternate routing plan, all points on the chart will have automatic switching systems.† The customer (or the operator until customer dialing arrangements are completed) will dial a ten-digit code (three-digit area code 305 for Florida plus the listed Miami seven-digit telephone number) into the machine at Syracuse. The various routes which then might conceivably be tried automatically

* See the bibliography (particularly Pilliod and Truitt) for details of the general trunking plan.

† The notation used on the diagram of Fig. 8 is: Open circle — Primary Center (Syracuse, Miami); Triangle — Sectional Center (Albany, Jacksonville); Square — Regional Center (White Plains, Atlanta, St. Louis; St. Louis is also the National Center).

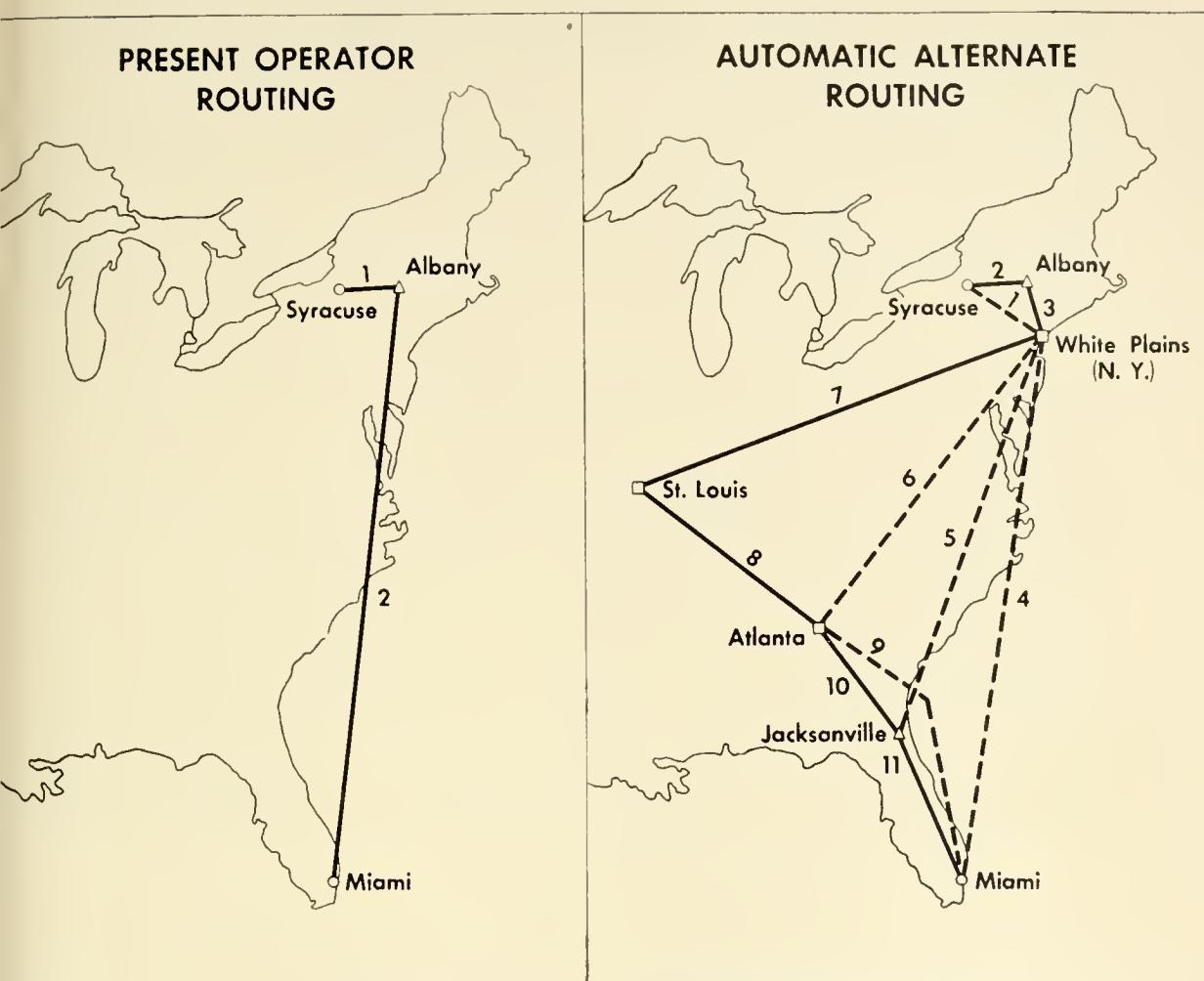


Fig. 8 — Present and proposed methods of handling a call from Syracuse, N. Y., to Miami, Florida.

are shown on the diagram numbered in the order of trial; in this particular layout shown, a maximum of eleven circuit groups could be tested for an idle path if each high usage group should be found NC. Dotted lines show the high usage routes, which if found busy will overflow to the final groups represented by solid lines. The switching equipment at each point upon finding an idle circuit passes on the required digits to the next machine.

While the routing possibilities shown are factual, only in rare instances would a call be completed over the final route via St. Louis. Even in the busy season busy hour just a small portion of the calls would be expected to be switched as many as three times. And only a fraction of one per cent of all calls in the busy hour should encounter NC. As a result the service will be fast. When calls are handled by a toll operator, the cus-

tomer will not ordinarily need to hang up when NC is obtained. When he himself dials, a second trial after a short wait following NC should have a high probability of success.

Not many situations will be as complex as shown in Fig. 8; commonly several of the links between centers will be missing, the particular ones retained having been chosen from suitable economic studies. A large number of switching arrangements will be no more involved than the illustrative one shown in Fig. 9(a), centering on the Toll Center of Bloomsburg, Pennsylvania. The dashed lines indicate high usage groups from Bloomsburg to surrounding toll centers; since Bloomsburg "homes" on Scranton this is a final route as denoted by the solid line. As an example of the operation, consider a call at Bloomsburg destined for Williamsport. Upon finding all direct trunks busy, a second trial is made via Harrisburg; and should no paths in the Harrisburg group be available, a third and final trial is made through the Scranton group.

In considering the traffic flow of a network such as illustrated at Bloomsburg it is convenient to employ the conventional form of a two-stage graded multiple having "legs" of varying sizes and traffic loads individual to each, as shown in Fig. 9(b). Here only the circuits immediately outgoing from the toll center are shown; the parcels of traffic

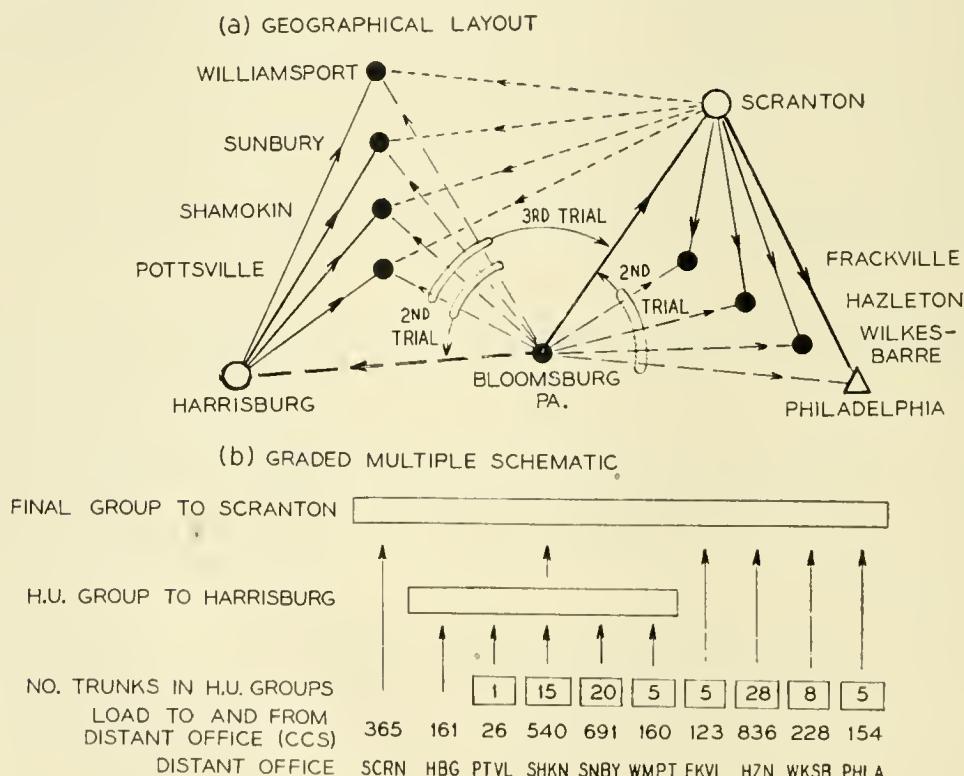


Fig. 9—Automatic alternate routing for direct distance dialing at Bloomsburg, Pa.

calculated for each further connecting route will be recorded as part of the offered load for consideration when the next higher switching center is engineered. It is implicitly assumed that a call which has selected one of the alternate route paths will be successful in finding the necessary paths available from the distant switching point onward. This is not quite true but is believed generally to be close enough for engineering purposes, and permits ignoring the return attempt problem.

6. NEW PROBLEMS IN THE ENGINEERING AND ADMINISTRATION OF INTER-TOLL GROUPS RESULTING FROM ALTERNATE ROUTING

With the greatly increased teamwork among groups of intertoll trunks which supply overflow calls to an alternate route, an unexpected increase or flurry in the offered load to any one can adversely affect the service to all. The high efficiency of the alternate route networks also reduces their overload carrying ability. Conversely, the influence of an underprovision of paths in the final alternate route may be felt by many groups which overflow to it. With non-alternate route arrangements only the single groups having these flurries would be affected.

Administratively, an alternate route trunk layout may well prove easier to monitor day by day than a large number of separate and independent intertoll groups, since a close check on the service given on the final routes only may be sufficient to insure that all customers are being served satisfactorily. When rearrangements are indicated, how-

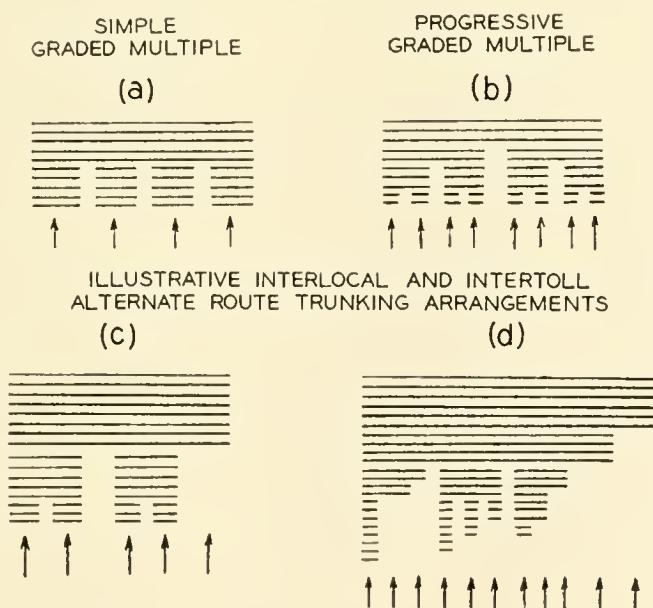


Fig. 10 — Graded multiples and alternate route trunking arrangements.

ever, the determination of the proper place to take action, and the desirable extent, may sometimes be difficult to determine. Suitable traffic measuring devices must be provided with these latter problems in mind.

For engineering purposes, it will be highly desirable:

(1) To be able to estimate the load-service relationships with any specified loads offered to a particular intertoll alternate routing network; and

(2) To know the day-to-day busy hour variations in the various groups' offered loads during the busy season, so that the general grade of service given to customers can be estimated.

The balance of this paper will review the studies which have been made in the Bell System toward a practicable method for predicting the grade of service given in an alternate route network under any given loads. Analyses of the day-to-day load variations and their effects on customer dialing service are currently being made, and will be reported upon later.

7. LOAD-SERVICE RELATIONSHIPS IN ALTERNATE ROUTE SYSTEMS

In their simplest form, alternate route systems appear as symmetrical graded multiples, as shown in Fig. 10(a) and 10(b). Patterns such as these have long been used in local automatic systems to partially overcome the trunking efficiency limitations imposed by limited access switches. The traffic capacity of these arrangements has been the subject of much study by theory and "throwdowns" (simulated traffic studies) both in the United States and abroad. Field trials have substantiated the essential accuracy of the trunking tables which have resulted.

In toll alternate route systems as contemplated in America, however, there will seldom be the symmetry of pattern found in local graded multiples, nor does maximum switch size generally produce serious limitation on the access. The "legs" or first-choice trunk groups will vary widely in size; likewise the number of such groups overflowing calls jointly to an alternate route may cover a considerable range. In all cases a given group, whether or not a link of an alternate route, will have one or more parcels of traffic for which it is the first-choice route. [See the right-hand parcel of offered traffic on Fig. 10(c).] Often this first routed traffic will be the bulk of the load offered to the group, which also serves as an alternate route for other traffic.

The simplest of the approximate formulas developed for solving the local graded multiple problems are hopelessly unwieldy when applied to such arrangements as shown in Fig. 10(d). Likewise it is impracticable

to solve more than a few of the infinite variety of arrangements by means of "throwdowns."

However, for both engineering (planning for future trunk provisions) and administration (current operating) of trunks in these multi-alternate routing systems, a rapid, simple, but reasonably accurate method is required. The basis for the method which has been evolved for Bell System use will be described in the following pages.

7.1. *The "Peaked" Character of Overflow Traffic*

The difficulty in predicting the load-service relationship in alternate route systems has lain in the non-random character of the traffic overflowing a first set of paths to which calls may have been randomly offered. This non-randomness is a well appreciated phenomenon among traffic engineers. If adequate trunks are provided for accommodating the momentary traffic peaks, the time-call level diagram may appear as in Fig. 11(a), (average level of 9.5 erlangs). If however a more limited number of trunks, say $x = 12$, is provided, the peaks of Fig. 11(a) will be clipped, and the overflow calls will either be "lost" or they may be handled on a subsequent set of paths y . The momentary loads seen on y then appear as in Fig. 11(b). It will readily be seen that a given average load on the y trunks will have quite different fluctuation characteristics than if it had been found on the x trunks. There will be more occurrences of large numbers of calls, and also longer intervals when few or no calls are present. This gives rise to the expression that overflow traffic is "peaked."

Peaked traffic requires more paths than does random traffic to operate at a specified grade of delayed or lost calls service. And the increase in paths required will depend upon the degree of peakedness of the traffic involved. A measure of peakedness of overflow traffic is then required which can be easily determined from a knowledge of the load offered and the number of trunks in the group immediately available.

In 1923, G. W. Kendrick, then with the American Telephone and Telegraph Company, undertook to solve the graded multiple problem through an application of Erlang's statistical equilibrium method. His principal contribution (in an unpublished memorandum) was to set up the equations for describing the existence of calls on a full access group of $x + y$ paths, arranged so that arriving calls always seek service first in the x -group, and then in the y -group when the x are all busy.

Let $f(m, n)$ be the probability that at a random instant m calls exist on the x paths and n calls on the y paths, when an average Poisson load

of a erlangs is submitted to the $x + y$ paths. The general state equation for all possible call arrangements, is

$$(a^* + m + n)f(m, n) = (m + 1)f(m + 1, n) \\ + (n + 1)f(m, n + 1) + af(m - 1, n) + af(x, n - 1)\ddagger \quad (7)$$

in which the term marked (\ddagger) is to be included only when $m = x$, and $*$ indicates that the a in this term is to be omitted when $m + n = x + y$. m and n may take values only in the intervals, $0 \leq m \leq x$; $0 \leq n \leq y$. As written, the equation represents the "lost calls cleared" situation.

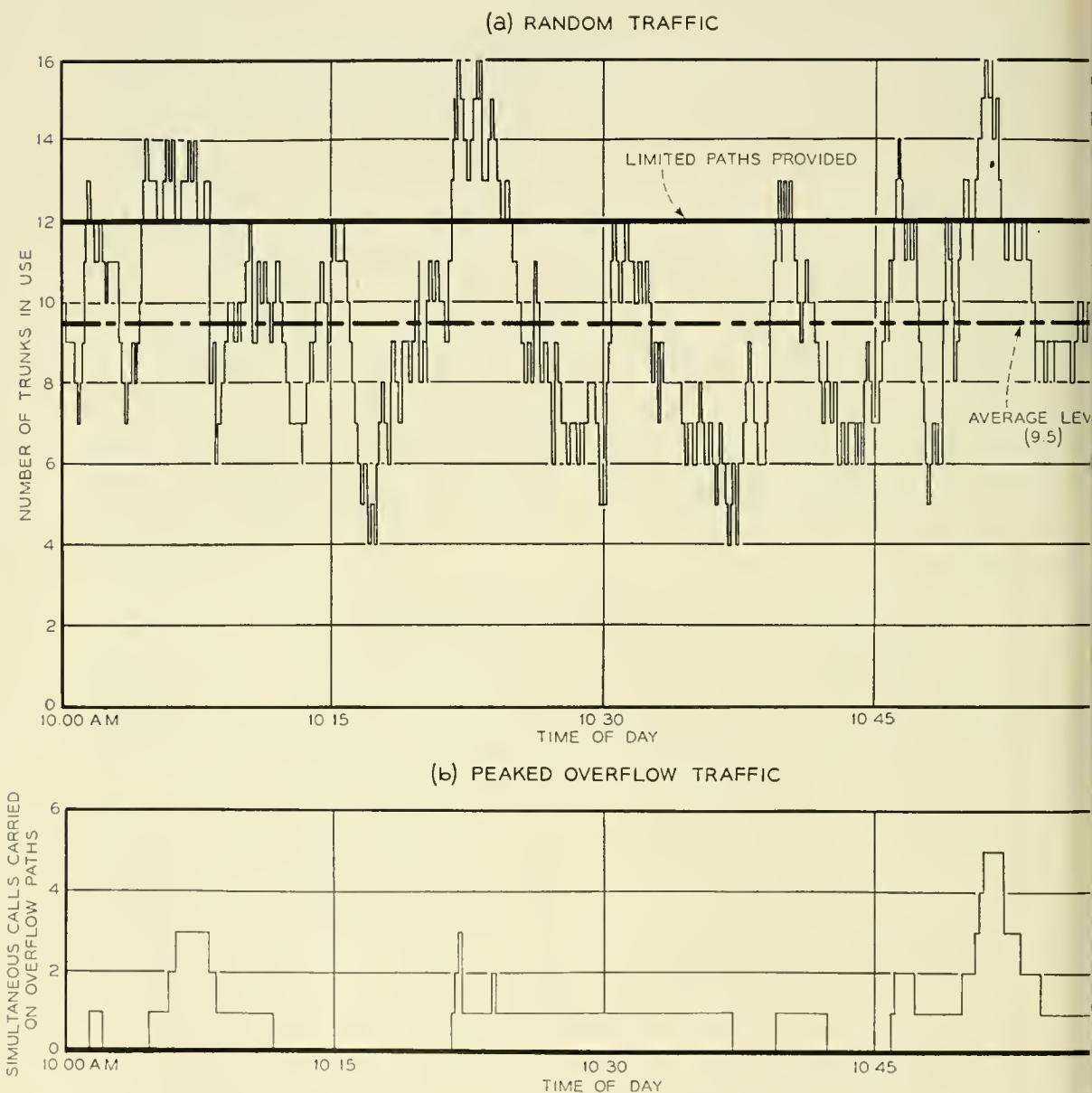


Fig. 11 — Production of peakedness in overflow traffic.

By choosing $x + y$ large compared with the submitted load a a "lost calls held" situation or infinite-overflow-trunks result can be approached as closely as desired.

Kendrick suggested solving the series of simultaneous equations (7) by determinants, and also by a method of continued fractions. However little of this numerical work was actually undertaken until several years later.

Early in 1935 Miss E. V. Wyckoff of Bell Telephone Laboratories became interested in the solution of the $(x + 1)(y + 1)$ lost calls cleared simultaneous equations leading to all terms in the $f(m, n)$ distribution. She devised an order of substituting one equation in the next which provided an entirely practical and relatively rapid means for the numerical solution of almost any set of these equations. By this method a considerable number of $f(m, n)$ distributions on x, y type multiples with varying load levels were calculated.

From the complete m, n matrix of probabilities, one easily obtains the distribution $\theta_m(n)$ of overflow calls when exactly m are present on the lower group of x trunks; or by summing on m , the $\theta(n)$ distribution without regard to m , is realized. A number of other procedures for obtaining the $f(m, n)$ values have been proposed. All involve lengthy computations, very tedious for solution by desk calculating machines, and most do not have the ready checks of the Wyckoff-method available at regular points through the calculations.

In 1937 Kosten⁴ gave the following expression for $f(m, n)$:

$$f(m, n) = (-1)^n \varphi_0(x) \sum_{i=0}^{\infty} \binom{i}{n} \frac{(-a)^i}{i!} \cdot \frac{\varphi_i(m)}{\varphi_{i+1}(x) \varphi_i(x)} \quad (8)$$

where

$$\varphi_0(x) = \frac{a^x e^{-a}}{x!}$$

and for $i > 0$,

$$\varphi_i(x) = e^{-a} \sum_{j=0}^x \binom{i+j-1}{j} \frac{a^{x-j}}{(x-j)!}$$

These equations, too, are laborious to calculate if the load and numbers of trunks are not small. It would, of course, be possible to program a modern automatic computer to do this work with considerable rapidity.

The corresponding application of the statistical equilibrium equations to the graded multiple problem was visualized by Kendrick who, however, went only so far as to write out the equation for the three-trunk

case consisting of two subgroups of one trunk each and one common overflow trunk.

Instead of solving the enormously elaborate system of equations describing all the calls which could simultaneously be present in a large multiple, several ingenious methods of convoluting the

$$\theta(n) = \sum_{m=0}^x f(m, n)$$

overflow distributions from the individual legs of a graded multiple have been devised. For example, for the multiple of Fig. 10(a), the probability of loss P_i as seen by a call entering subgroup number i , is approximately,

$$P_i = \sum_{r=0}^{y-1} \sum_{z=y}^{\infty} \theta_{x,i}(r) \cdot \psi(z - r) + \sum_{r=y}^{\infty} \theta_{x,i}(r) \quad (9)$$

in which $\psi(z - r)$ is the probability of exactly $z - r$ overflow calls being present, or wanting to be present, on the alternate route from all the subgroups except the i th, and with no regard for the numbers of calls present in these subgroups. The $\theta_{x,i}(r) = f_i(x_i, r)$ term, of course, contemplates all paths in the particular originating call's subgroup being occupied, forcing the new call arriving in subgroup i to advance to the alternate route. This corresponds to the method of solving graded multiples developed by E. C. Molina⁶ but has the advantage of overcoming the artificial "no holes in the multiple" assumption which he made. Similar calculating procedures have been suggested by Kosten.* These computational methods doubtless yield useful estimates of the resulting service, and for the limited numbers of multiple arrangements which might occur in within-office switching trains (particularly ones of a symmetrical variety) such procedures might be practicable. But it would be far too laborious to obtain the individual overflow distributions $\theta(n)$, and then convolute them for the large variety of loads and multiple arrangements expected to be met in toll alternate routing.

7.2. Approximate Description of the Character of Overflow Traffic

It was natural that various approximate procedures should be tried in the attempt to obtain solutions to the general loss formula sufficiently accurate for engineering and study purposes. The most obvious of these is to calculate the lower moments or semi-invariants of the loads overflowing the subgroups, and from them construct approximate fitting

* Kosten gives the above approximation (9), which he calls W_b^+ , as an upper limit to the blocking. He also gives a lower limit, W_b^- , in which $z = y$ throughout (References 4, 5).

distributions for $\theta(n)$ and $\theta_x(n)$. Since each such overflow is independent of the others, they may be combined additively (or convoluted), to obtain the corresponding total distribution of calls appearing before the alternate route (or common group). It may further be possible to obtain an approximate fitting distribution to the sum-distribution of the overflow calls.

The ordinary moments about the 0 point of the subgroup overflow distribution, when m of the x paths are busy, are found by

$$\mu_i'(m) = \sum_{n=0}^y n^i f(m, n) \quad (10)$$

When an infinite number of y -paths is assumed, the resulting expressions for the mean and variance are found to be:^{*}

Number of x -paths busy unspecified:[†]

$$\text{Mean} = \alpha = a \cdot E_{1,x}(a) \quad (11)$$

$$\text{Variance} = v = \alpha[1 - \alpha + a(x + 1 + \alpha - a)^{-1}] \quad (12)$$

All x -paths occupied

$$\text{Mean} = \alpha_x = a[x - a + 1 + aE_{1,x}(a)]^{-1} \quad (13)$$

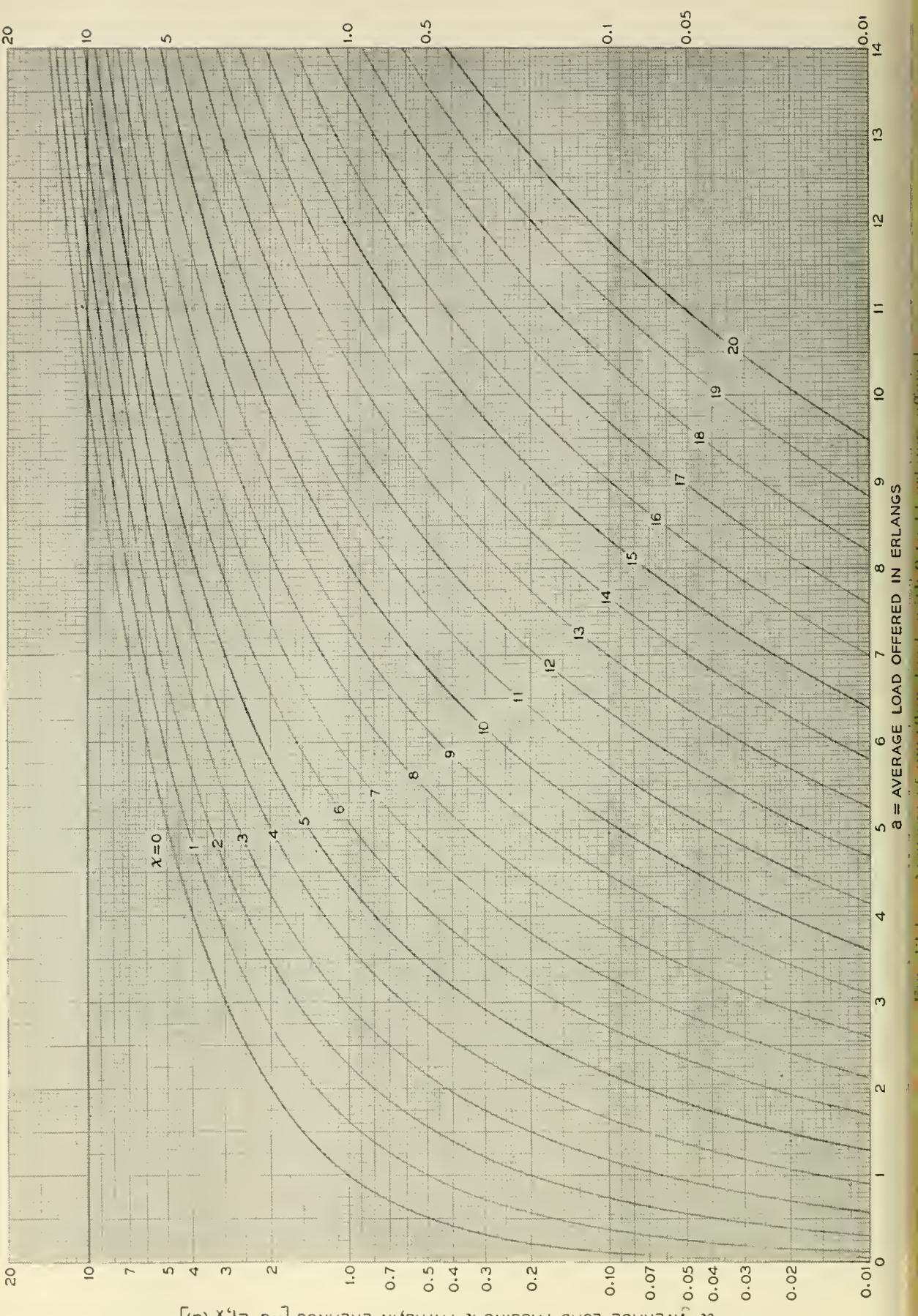
$$\text{Variance} = v_x = \alpha_x[1 - \alpha_x + 2a(x + 2 + \alpha_x - a)^{-1}] \quad (14)$$

Equations (11) and (12) have been calculated for considerable ranges of offered load a and paths x . Figs. 12 and 13 are graphs of these results. For example when a load of 4 erlangs is submitted to 5 paths, the average overflow load is seen to be $\alpha = 0.80$ erlang, the same value, of course, as determined through a direct application of the Erlang E_1 formula. During the time that all x paths are busy, however, the overflow load will tend to exceed this general level as indicated by the value of $\alpha_x = 1.41$ erlangs calculated from (13). Similarly the variance of the overflow load will tend to increase when the x -paths are fully occupied,

* The derivation of these equations is given in Appendix I.

† The skewness factor may also be of interest:

$$\begin{aligned} \sqrt{\beta_1} &= \frac{\mu_3}{\mu_2^{2/3}} \\ &= \frac{1}{v^{3/2}} \left[\frac{a}{x+1+\alpha-a} \left\{ \frac{2}{x+2} \left(\frac{(x+\alpha-a)a^2}{(x-a)^2+2(x-a)+x+2+(x+2-a)\alpha} + a \right) \right. \right. \\ &\quad \left. \left. + 3(1-\alpha) \right\} + \alpha(1-\alpha)(1-2\alpha) \right] \end{aligned} \quad (15)$$



$\alpha = \text{AVERAGE LOAD PASSING } x \text{ PATHS, IN ERLANGS } [= a \cdot E^{1/x} (a)]$

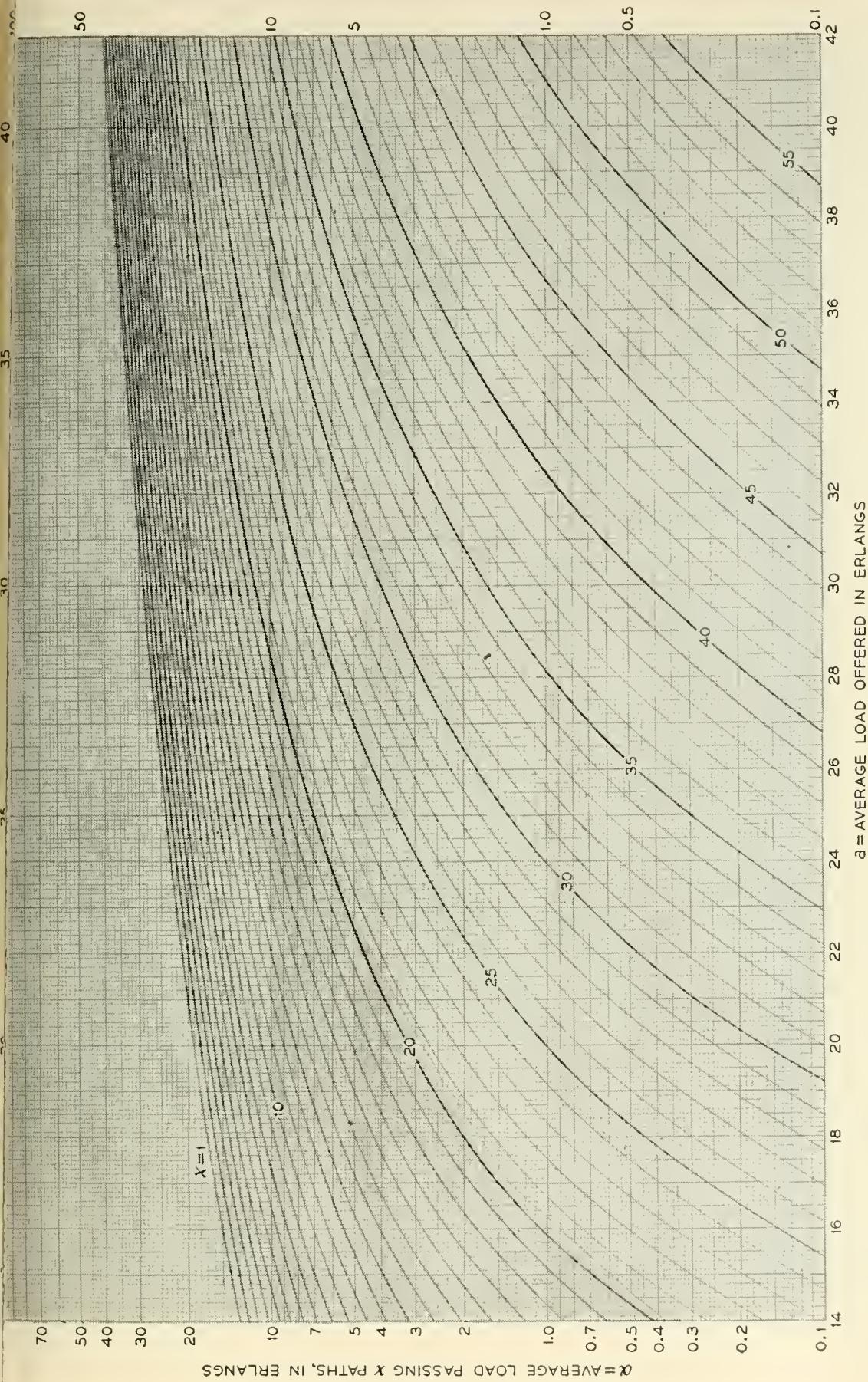
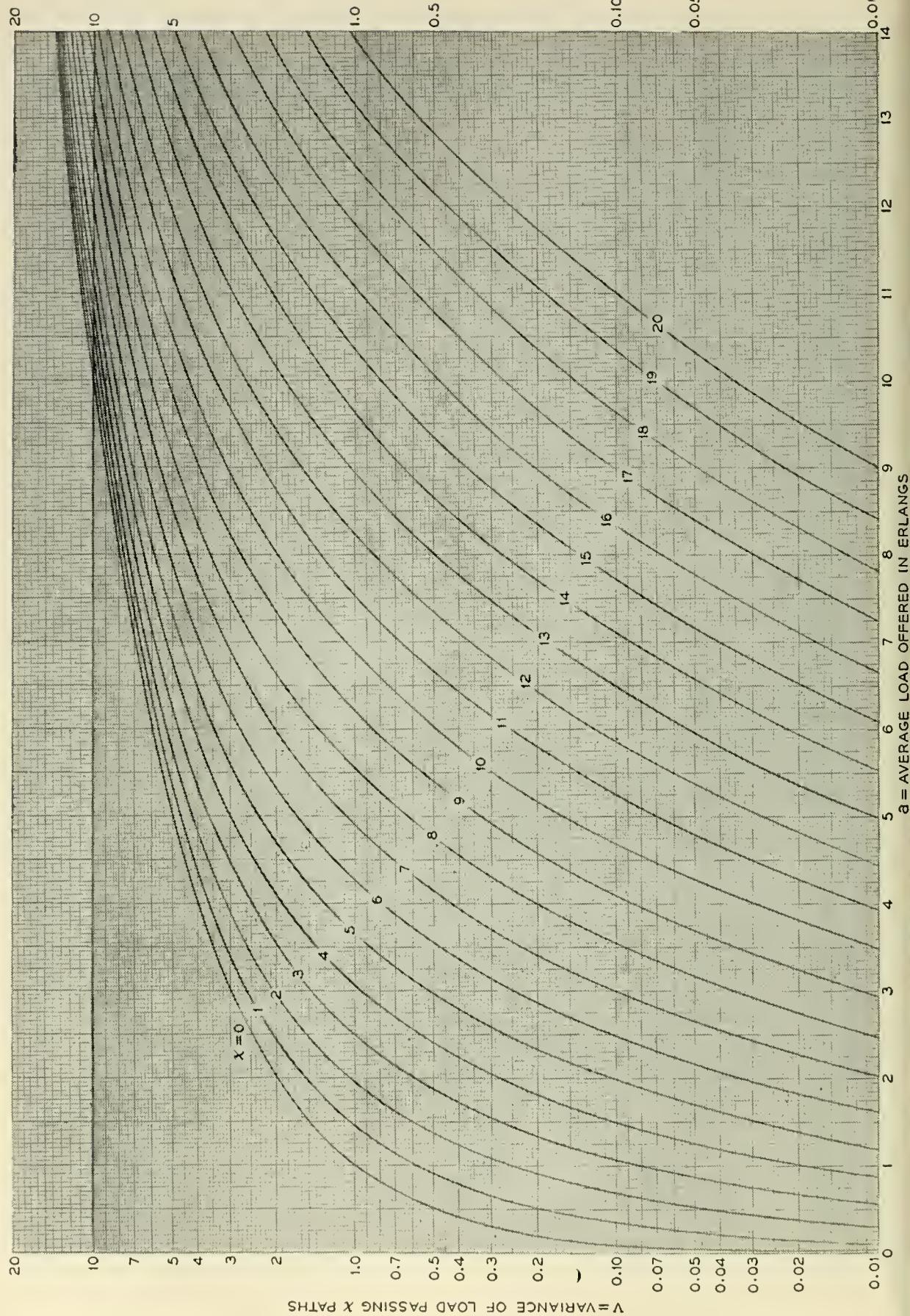


Fig. 12.2 — Average of overflow load, with 14 to 42 erlangs offered.



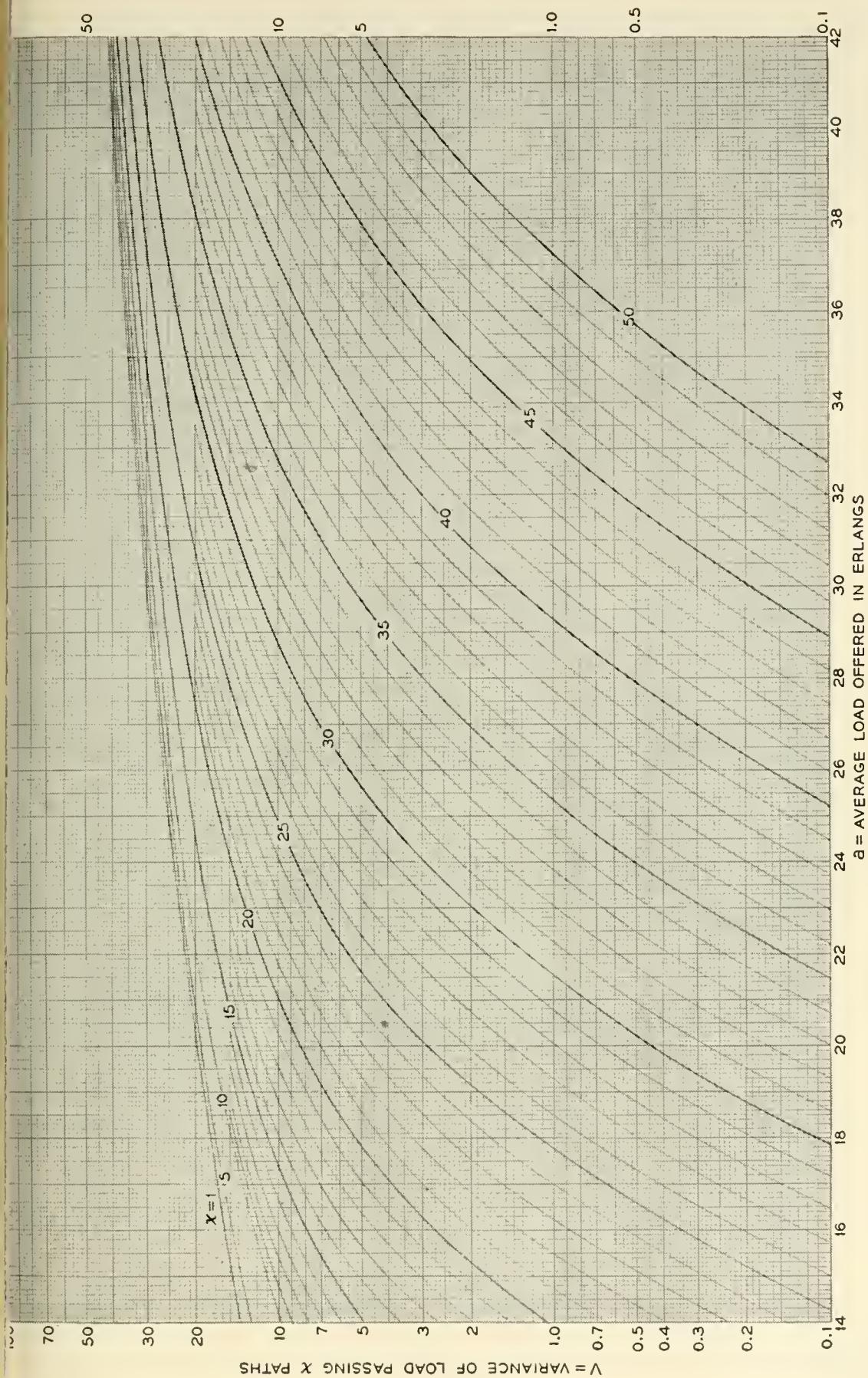


Fig. 13.2 — Variance of overflow load, with 14 to 42 erlangs offered.

as shown by $v = 1.30$, and $v_x = 1.95$. In all cases the variances v and v_x will exceed the variance of corresponding Poisson traffic (which would have variances of α and α_x respectively).

7.2.1. A Probability Distribution for Overflow Traffic

It would be of interest to be able, given the first several descriptive parameters of any traffic load (such as the mean and variance and skewness factors of the overflow from a group of trunks), to construct an approximate probability distribution $\theta(n)$ which would closely describe the true momentary distribution of simultaneous calls. Any proposed fitting distribution for the overflow from random traffic offered to x trunks, can, of course, be compared with

$$\theta(n) = \sum_{m=0}^x f(m, n)$$

determined from (7) or (8).

Suitable fitting curves should give probabilities for all positive integral values of the variable (including zero), and have sufficient unspecified constants to accommodate the parameters selected for describing the distribution. Moreover, the higher moments of a fitting distribution should not diverge too radically from those of the true distribution; that is, the "natural shapes" of fitting and true distributions should be similar. Particularly desirable would be a fitting distribution form derived with some attention to the physical circumstances causing the ebb and flow of calls in an overflow situation. The following argument and derivation undertake to achieve these desiderata.*

A Poisson distribution of offered traffic is produced by a random arrival of calls. The assumption is made or implied that the probability of a new arrival in the next instant of time is quite independent of the number currently present in the system. When this randomness (and corresponding independence) are disturbed the resulting distribution will no longer be Poisson. The first important deviation from the Poisson would be expected to appear in a change from variance = mean, to variance \neq

* A two-parameter function which has the ability to fit quite well a wide variety of true overflow distributions, has the form

$$\xi(n) = K(n + 1)^b e^{-c(n+1)}$$

in which K is the normalizing constant. The distribution is displaced one unit from the usual discrete generalized exponential form, so that $\xi(0) \neq 0$. The expression, however, has little rationale for being selected a priori as a suitable fitting function.

mean. Corresponding changes in the higher moments would also be expected.

What would be the physical description of a cause system with a variance smaller or larger than the Poisson? If the variance is smaller, there must be forces at work which retard the call arrival rate as the number of calls recently offered exceeds a normal, or average, figure, and which increase the arrival rate when the number recently arrived falls below the normal level. Conversely, the variance will exceed the Poisson's should the tendencies of the forces be reversed.* This last is, in fact, a rough description of the incidence rates for calls overflowing a group of trunks.

Since holding times are attached to and extend from the call arrival instants, calls are enabled to project their influence into the future; that is, the presence of a considerable number of calls in a system at any instant reflects their having arrived in recent earlier time, and now can be used to modify the current rate of call arrival.

Let the probability of a call originating in a short interval of time dt be

$$P_{o,n} = [a + (n - a)\omega(n)] dt$$

where n = number of calls present in the system at time t ,

a = base or average arrival rate of calls per unit time, and

$\omega(n)$ = an arbitrary function which regulates the modification in call origination rate as the number of calls rises above or falls below a .

Correspondingly, let the probability that one of n calls will end in the short interval of time dt be

$$P_{e,n} = n dt,$$

which will be satisfied in the case of exponential call holding times, with mean unity. Following the usual Erlang procedure, the general statistical equilibrium equation is

$$\begin{aligned} f(n) = f(n)[(1 - P_{o,n})(1 - P_{e,n})] &+ f(n-1)P_{o,n-1}(1 - P_{e,n-1}) \\ &+ f(n+1)(1 - P_{o,n+1})P_{e,n+1} \end{aligned} \quad (16)$$

which gives

$$(P_{o,n} + P_{e,n})f(n) = P_{o,n-1}f(n-1) + P_{e,n+1}f(n+1)$$

ignoring terms of order higher than the first in dt .

* The same thinking has been used by Vaulot⁷ for decreasing the call arrival rate according to the number momentarily present; and by Lundquist⁸ for both increasing and decreasing the arrival rate.

Or,

$$\begin{aligned} & [a + (n - a)\omega(n) + n]f(n) \\ &= [a + (n - a - 1)\omega(n - 1)]f(n - 1) + (n + 1)f(n + 1) \end{aligned} \quad (17)$$

The choice of $\omega(n)$ will determine the solution of (17). Most simply, $\omega(n) = k$, making the variation from the average call arrival rate directly proportional to the deviation in numbers of calls present from their average number. In this case, the solution for an unlimited trunk group becomes, with $a' = a(1 - k)$,

$$f(n) = \frac{\frac{a'(a' + k) \cdots [a' + (n - 1)k]}{n!}}{1 + a' + \frac{a'(a' + k)}{2!} + \frac{a'(a' + k)(a' + 2k)}{3!} + \dots} \quad (18)$$

which may also be written after setting $a'' = a'/k = a(1 - k)/k$, as

$$f(n) = \frac{\frac{a''(a'' + 1) \cdots [a'' + (n - 1)]k^n}{n!}}{(1 - k)^{-a''}} \quad (19)$$

The generating function (g.f.) of (19) is

$$\sum_{n=0}^{\infty} f(n)T^n = \frac{(1 - kT)^{-a''}}{(1 - k)^{-a''}}$$

which is recognized as that for the negative binomial, as distinguished from the g.f.,

$$(q + pT)^N = \frac{\left(1 + \frac{p}{q} T\right)^N}{(1/q)^N}$$

for the positive binomial.

The first four descriptive parameters of $f(n)$ are:

Order	Moment about Mean	Descriptive Parameter
1	$\mu_1 = 0$	Mean = $\bar{n} = a$ (20)
2	$\mu_2 = \text{variance}, v = a/(1 - k)$	Std Devn, $\sigma = [a/(1 - k)]^{1/2}$ (21)
3	$\mu_3 = \frac{a(1 + k)}{(1 - k)^2}$	Skewness, $\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \frac{1 + k}{a^{1/2}(1 - k)^{1/2}}$ (22)
4	$\mu_4 = \frac{3a^2(1 - k) + a(k^2 + 4k + 1)}{(1 - k)^3}$	Kurtosis, $\beta_2 = \frac{\mu_4}{\sigma^4} = 3 + \frac{k^2 + 4k + 1}{a(1 - k)}$ (23)

Since only two constants, a and k , need specification in (18) or (19), the mean and variance are sufficient to fix the distribution. That is, with the mean \bar{n} and variance v known,

$$a = \bar{n} \quad \text{or} \quad a' = \bar{n}(1 - k) = \bar{n}^2/v, \quad \text{or} \quad a'' = \bar{n}(1 - k)/k \quad (24)$$

$$k = 1 - a/v = 1 - \bar{n}/v. \quad (25)$$

The probability density distribution $f(n)$ is readily calculated from (19); the cumulative distribution $G(\geq n)$ also may be found through use of the Incomplete Beta Function tables since

$$\begin{aligned} G(\geq n) &= I_k(n - 1, a'') \\ &= I_k(n - 1, a(1 - k)/k) \end{aligned} \quad (26)$$

The goodness with which the negative binomial of (19) fits actual distributions of overflow calls requires some investigation. Perhaps a more elaborate expression for $\omega(n)$ than a constant k in (17) is required. Three comparisons appear possible: (1), comparison with a variety of $\theta_m(n)$ distributions with exactly m calls on the x trunks, or $\theta(n)$ with m unspecified, (obtained by solving the statistical equilibrium equations (7) for a divided group); (2), comparison with simulation or "throwdown" results; and (3), comparison with call distributions seen on actual trunk groups. These are most easily performed in the order listed.*

Comparison of Negative Binomial with True Overflow Distributions

Figs. 14 to 17 show various comparisons of the negative binomial distribution with true overflow distributions. Fig. 14 gives in cumulative form the cases of 5 erlangs offered to 1, 2, 5, and 10 trunks. The true

$$F(\geq n) = \sum_{j=n}^{\infty} \theta(j)$$

distributions (shown as solid lines) are obtained by solving the difference equations (7) in the manner described in Section 7.1. The negative binomial distributions (shown dashed) are chosen to have the same mean and variance as the several $F(\geq n)$ cases fitted. The dots shown on

* Comparison could also be made after equating means and variances respectively, between the higher moments of the overflow traffic beyond x trunks and the corresponding negative binomial moments: e.g., the skewness given by (15) can be compared with the negative binomial skewness of (22). The difficulty here is that one is unable to judge whether the disparity between the two distribution functions as described by differences in their higher parameters is significant or not for traffic engineering purposes.

the figure are for random (Poisson) traffic having the same mean values as the F distributions. The negative binomial provides excellent fits down to cumulated probabilities of 0.01, with a tendency thereafter to give somewhat larger values than the true ones. The Poisson agreement is good only for the overflow from a single trunk, as might have been anticipated, the divergence rapidly increasing thereafter.

Fig. 15 corresponds with the cases of Fig. 14 except that the true overflow $F_x(\geq n)$ distributions for the conditional situation of all x -paths busy, are fitted. Again the negative binomial is seen to give a good agreement down to 0.01 probability, with somewhat too-high estimates for larger values of the simultaneous overflow calls n .

Fig. 16 shows additional comparisons of overflow and negative binomial distributions. As before, the agreement is quite satisfactory to 0.01 probability, the negative binomial thereafter tending to give somewhat high values.

On Fig. 17 are compared the individual $\theta(n)$ density distributions for several cases. The agreement of the negative binomial with the true distribution is seen to be uniformly good. The dots indicate the random (Poisson) individual term distribution corresponding to the $a = 9.6$ case;

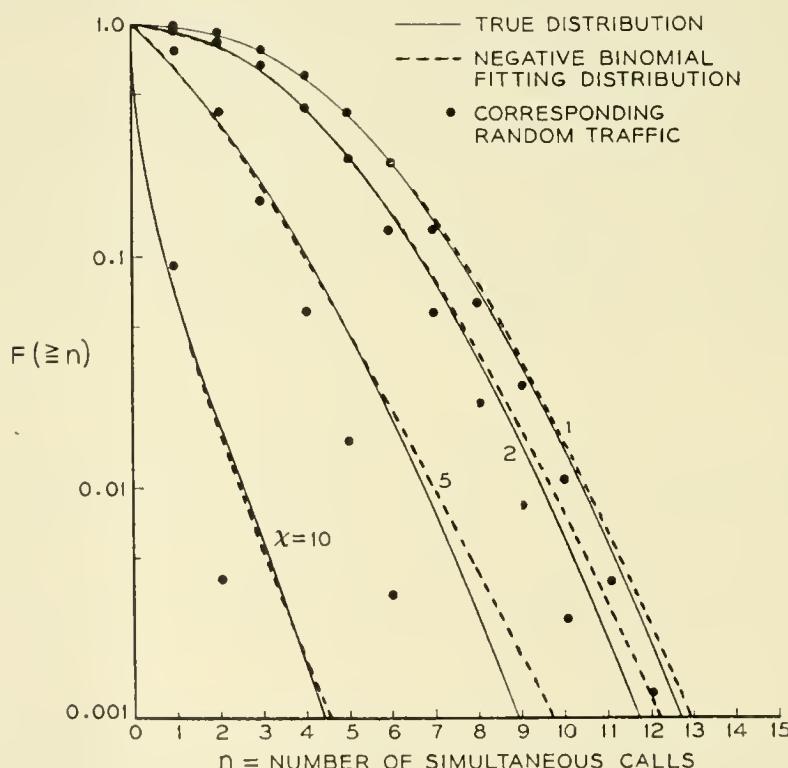


Fig. 14 — Probability distributions of overflow traffic with 5 erlangs offered to 1, 2, 5, and 10 trunks, fitted by negative binomial.

the agreement, of course, is poor since the non-randomness of the overflow here is marked, having an average of 1.88 and a variance of 3.84.

Comparison of Negative Binomial with Overflow Distributions Observed by Throwdowns and on Actual Trunk Groups

Fig. 18 shows a comparison of the negative binomial with the overflow distributions from four direct groups as seen in throwdown studies. The agreement over the range of group sizes from one to fifteen trunks is seen to be excellent. The assumption of randomness (Poisson) as shown by the dot values is clearly unsatisfactory for overflows beyond more than two or three trunks.

A number of switch counts made on the final group of an operating toll alternate routing system at Newark, New Jersey, during periods when few calls were lost, have also shown good agreement with the negative binomial distribution.

7.2.2. A Probability Distribution for Combined Overflow Traffic Loads

It has been shown in Section 7.2.1 that, at least for load ranges of wide interest, the negative binomial with but two parameters, chosen to agree

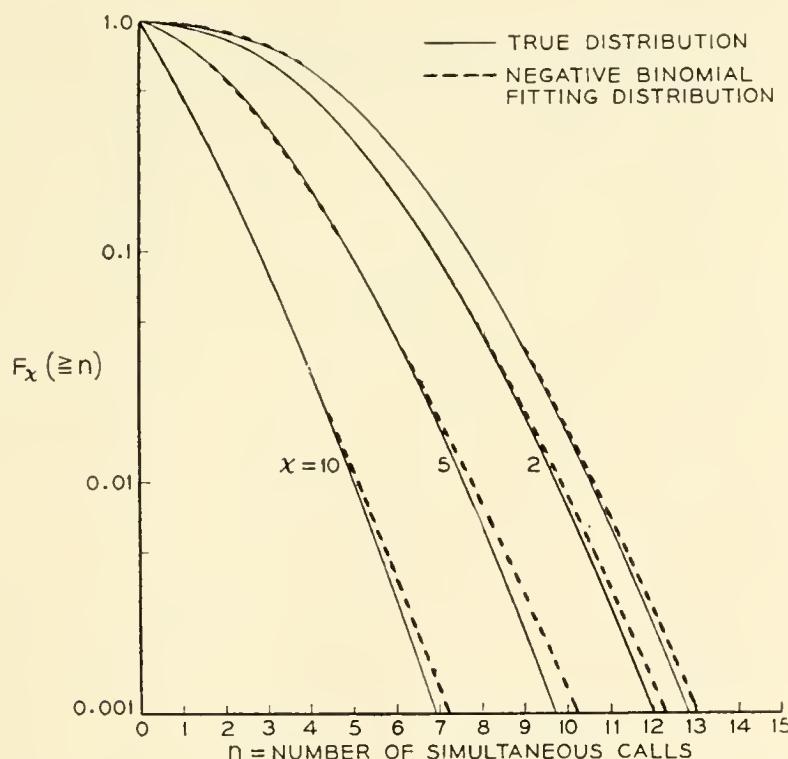


Fig. 15 — Probability distributions of overflow traffic with 5 erlangs offered to 1, 2, 5, and 10 trunks, when all trunks are busy; fitted by negative binomial.

with mean and variance, gives a satisfactory fit to the distribution of traffic overflowing a group of trunks. It is now possible, of course, to convolute the various overflows from any number of groups of varying sizes, to obtain a combined overflow distribution. This procedure, however, would be very clumsy and laborious since at each switching point in the toll alternate route system an entirely different layout of loads and high usage groups would require solution; it would be unfeasible for practical working.

We return again to the method of moments. Since the overflows of the several high usage groups will, in general, be independent of one another, the i th semi-invariants λ_i of the individual overflows can be combined to give the corresponding semi-invariants Λ_i of their total,

$$\Lambda_i = {}_1\lambda_i + {}_2\lambda_i + \dots \quad (27)$$

Or, in terms of the overflow means and variances, the corresponding parameters of the combined loads are

$$\text{Average} = A' = \alpha_1 + \alpha_2 + \dots \quad (28)$$

$$\text{Variance} = V' = v_1 + v_2 + \dots \quad (29)$$

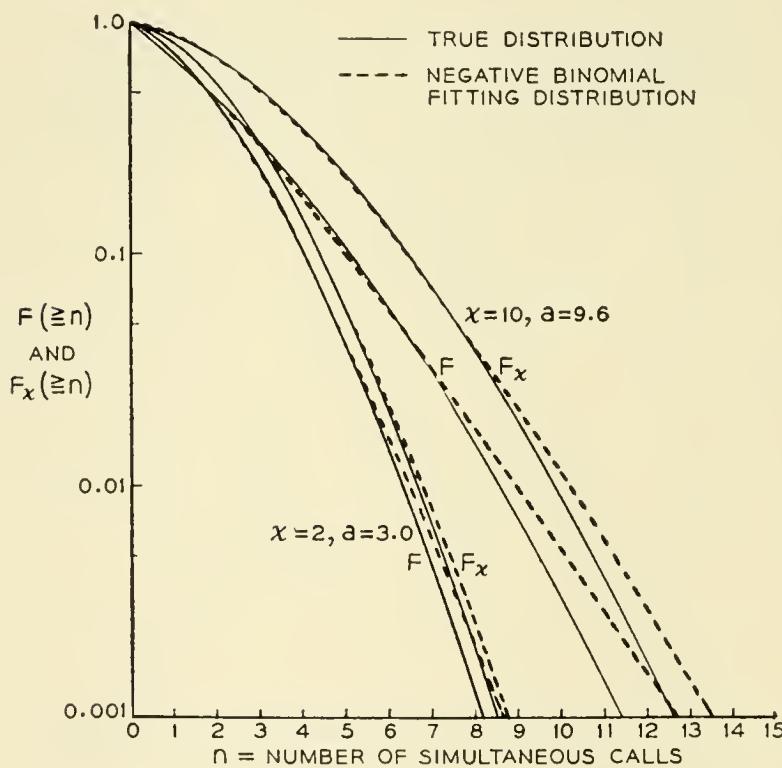


Fig. 16 — Probability distributions of overflow traffic: 3 erlangs offered to 2 trunks, and 9.6 erlangs offered to 10 trunks.

With the mean and variance of the combined overflows now determined, the negative binomial can again be employed to give an approximate description of the distribution of the simultaneous calls $\varphi(z)$ offered to the common, or alternate, group.

The acceptability of this procedure can be tested in various ways. One way is to examine whether the convolution of several negative binomials (representing overflows from individual groups) is sufficiently well fitted by another negative binomial with appropriate mean and variance, as found above.

It can easily be shown that the convolution of several negative binomials all with the same over-dispersion (variance-to-mean ratio) but not necessarily the same mean, is again a negative binomial. Shown in Table I are the distribution components and their parameters of two examples in which the over-dispersion parameters are not identical. The third and fourth semi-invariants of the fitted and fitting distributions, are seen to diverge considerably, as do the Pearsonian skewness and kurtosis factors. The test of acceptability for traffic fluctuation description comes in comparing the fitted and fitting distributions which are shown on Fig. 19. Here it is seen that, despite what might appear alarming dis-

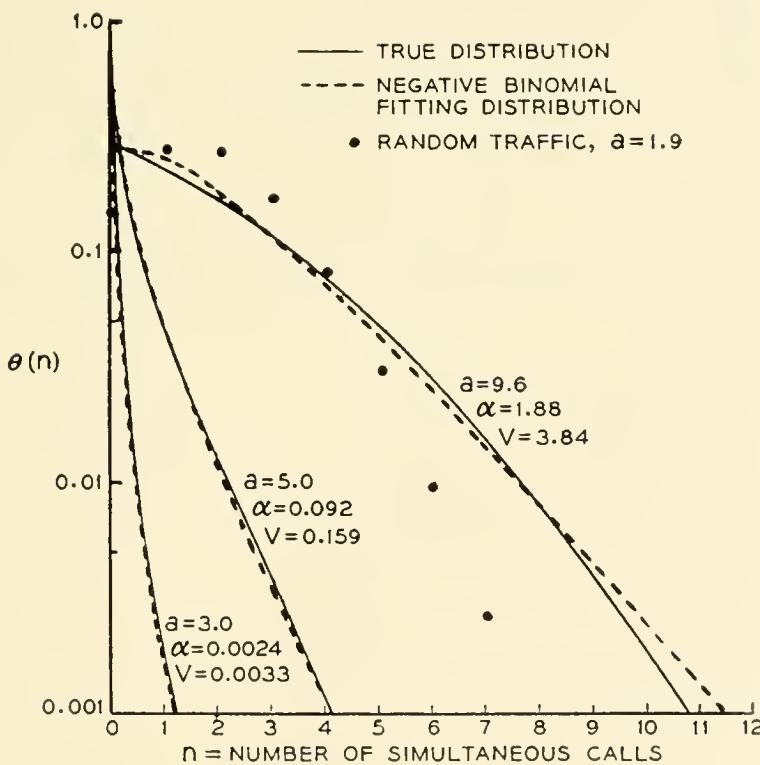


Fig. 17 — Probability density distributions of overflow traffic from 10 trunks, fitted by negative binomial.

parities in the higher semi-invariants, the agreement for practical traffic purposes is very good indeed.

Numerous throwdown checks confirm that the negative binomial employing the calculated sum-overflow mean and variance has a wide range over which the fit is quite satisfactory for traffic description purposes. Fig. 20 shows three such trunking arrangements selected from a considerable number which have been studied by the simulation method. Approximately 5,000, 3,500, and 580 calls were run through in the three examples, respectively. The overflow parameters obtained by experiment are seen to agree reasonably well with the theoretical ones from (28) and (29) when the numbers of calls processed is considered.

On Fig. 21 are shown, for the first arrangement of Fig. 20, distributions of simultaneous offered calls in each subgroup of trunks compared with the corresponding Poisson; the agreement is satisfactory as was to be expected. The sum distribution of the overflows from the eight subgroups is given at the foot of the figure. The superposed Poisson, of course, is a poor fit; the negative binomial, on the other hand, appears quite acceptable as a fitting curve.

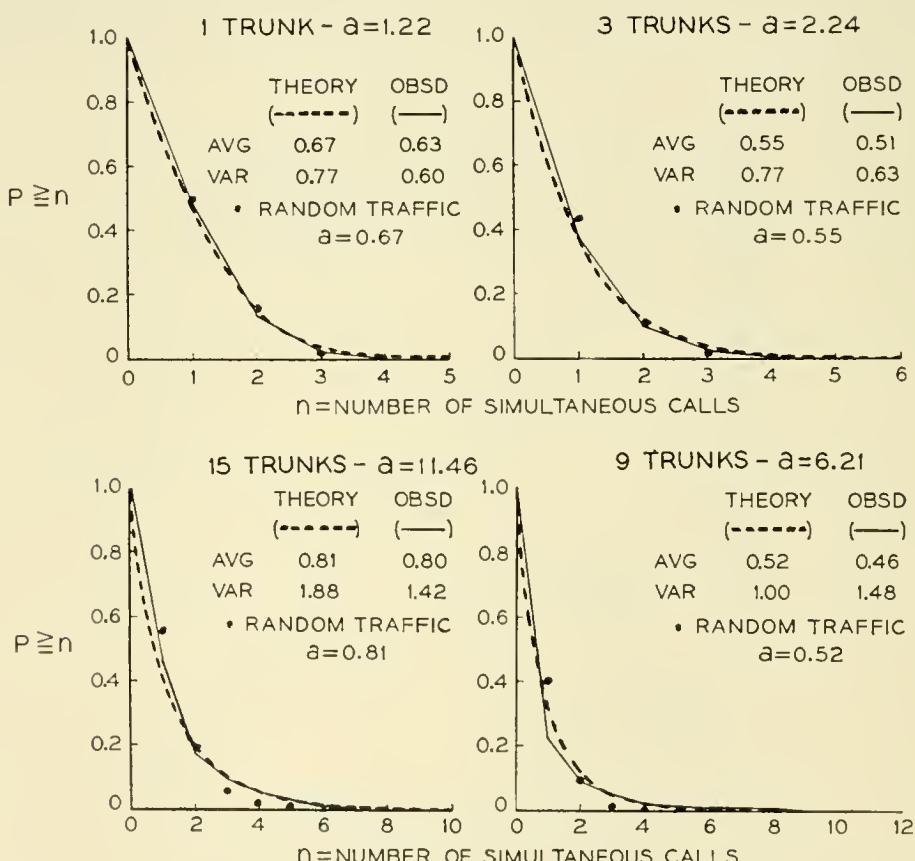


Fig. 18 — Overflow distributions from direct interoffice trunk groups; negative binomial theory versus throwdown observations.

TABLE I—COMPARISON OF PARAMETERS OF A FITTING
NEGATIVE BINOMIAL TO THE CONVOLUTION OF
THREE NEGATIVE BINOMIALS

Example No. 1			Example No. 2		
Component dist'n No.	Component parameters		Component dist'n No.	Component parameters	
	Mean	Variance		Mean	Variance
1	5	5	1	1	1
2	2	4	2	2	3
3	1	3	3	2	6
	—	—		—	—
	8	12		5	10

Semi-Invariants Λ , Skewness $\sqrt{\beta_1}$, and Kurtosis β_2 , of Sum Distributions

Parameter	Exact	Fitting	Parameter	Exact	Fitting
Λ_1	8	8	Λ_1	5	5
Λ_2	12	12	Λ_2	10	10
Λ_3	32	24	Λ_3	37	30
Λ_4	168	66	Λ_4	239.5	130
$\sqrt{\beta_1}$	0.770	0.577	$\sqrt{\beta_1}$	1.170	0.949
β_2	4.167	3.458	β_2	5.395	4.300

Fig. 22 shows the corresponding comparisons of the overflow loads in the other two trunk arrangements of Fig. 20. Again good agreement with the negative binomial is seen.

7.3. Equivalent Random Theory for Prediction of Amount of Traffic Overflowing a Single Stage Alternate Route, and Its Character, with Lost Calls Cleared

As discussed in Section 7.2, when random traffic is offered to a limited number of trunks x , the overflow traffic is well described (at least for traffic engineering purposes) by the two parameters, mean α and variance v . The result can readily be applied to a group divided (in one's mind) two or more times as in Fig. 23.

Employing the α and v curves of Figs. 12 and 13, and the appropriate numbers of trunks x_1 , $x_1 + x_2$, and $x_1 + x_2 + x_3$, the pairs of descriptive parameters, α_1 , v_1 , α_2 , v_2 and α_3 , v_3 can be read at once. It is clear then that if at some point in a straight multiple a traffic with parameters α_1 , v_1 is seen, and it is offered to x_2 paths, the overflow therefrom will have the characteristics α_2 , v_2 . To estimate the particular values of α_2 and v_2 , one would first determine the values of the equivalent random

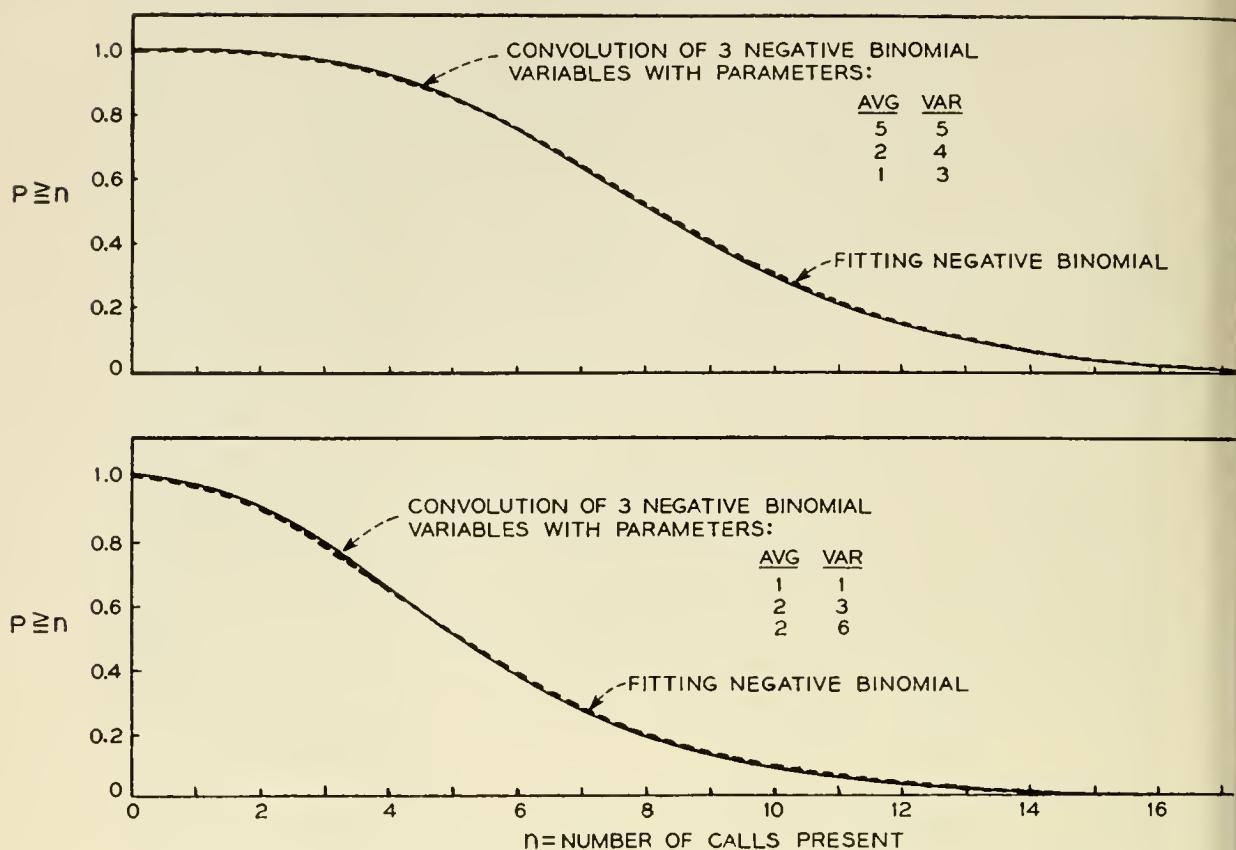


Fig. 19 — Fitting sums of negative binomial variables with a negative binomial.

traffic a and trunks x_1 which would have produced α_1 and v_1 . Then proceeding in the forward direction, using a and $x_1 + x_2$, one consults the α and v charts to find α_2 and v_2 . Thus, within the limitations of straight group traffic flow, the character (mean and variance) of any overflow load from x trunks can be predicted if the character (mean and variance) of the load submitted to them is known.

Curves could be constructed in the manner just described by which the overflow's α' and v' are estimated from a load, α and v , offered to x trunks. An illustrative fragment of such curves is shown in Appendix II, with an example of their application in the calculation of a straight trunk group loss by considering the successive overflows from each trunk as the offered loads to the next.

Enough, perhaps, has been shown in Section 7.2 of the generally excellent descriptions of a variety of non-random traffic loads obtainable by the use of only the two parameters α and v , to make one strongly suspect that most of the fluctuation information needed for traffic engineering purposes is contained in those two values. If this is, in fact, the case, we should then be able to predict the overflow α' , v' from x trunks

with an offered load α , v which has arisen in any manner of overflow from earlier high usage groups, as illustrated in Fig. 24.

This is found to be the case, as will be illustrated in several studies described in the balance of this section. In the determination of the characteristics of the overflow traffic α' , v' in the cases of non-full-access groups, such as Figs. 24(b) and 24(c), the equivalent straight group is visualized [Fig. 24(a)], and the Equivalent Random load A and trunks S are found.* Using A , and $S + C$, to enter the α and v curves of Figs. 12 and 13, α' and v' are readily determined. To facilitate the reading of A and S , Fig. 25† and Fig. 26† (which latter enlarges the lower left corner of Fig. 25) have been drawn. Since, in general, α and v will not have come from a simple straight group, as in Fig. 24(a), it is not to be expected that S ,

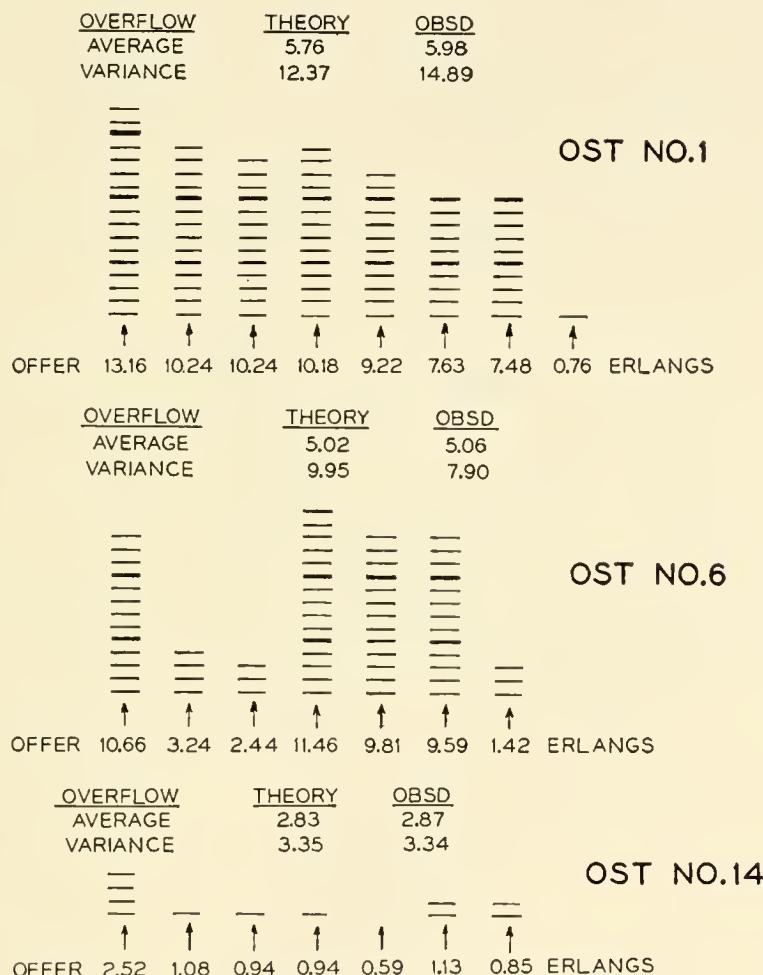


Fig. 20 — Comparison of joint-overflow parameters; theory versus throwdown.

* A somewhat similar method, commonly identified with the British Post Office, which uses one parameter, has been employed for solving symmetrical graded multiples (Ref. 9).

† Figs. 25 and 26 will be found in the envelope on the inside back cover.

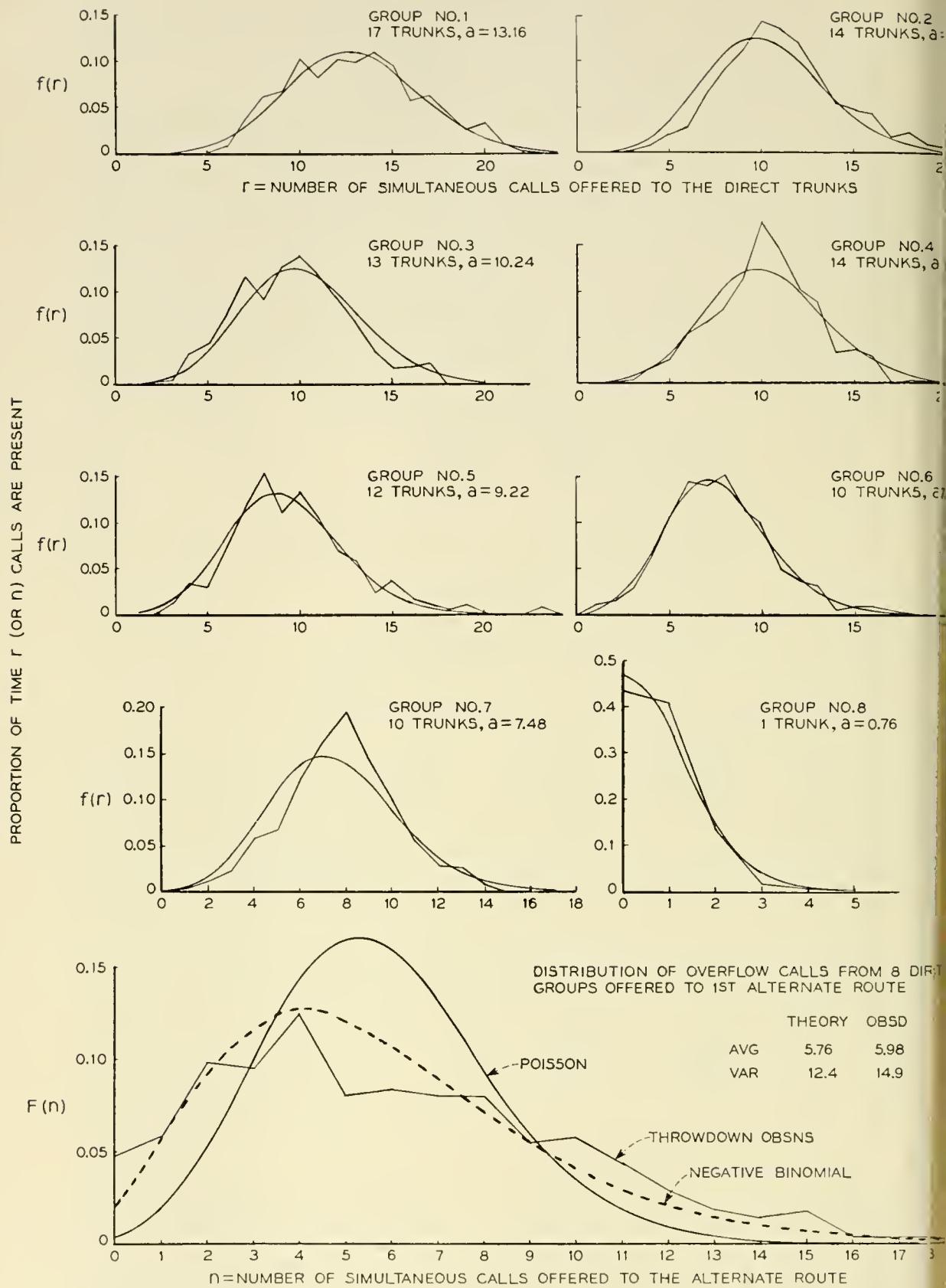


Fig. 21 — Comparison of theoretical and throwdown distributions of simultaneous calls offered to direct groups and to their first alternate route (OST No. 1).

read from Fig. 25, will be an integer. This causes no trouble and S should be carried along fractionally to the extent of the accuracy of result desired. Reading S to one-tenth of a trunk will usually be found sufficient for traffic engineering purposes.

Example 1: Suppose a simple graded multiple has three trunks in each of two subgroups, which overflow to C common trunks, where $C = 1$,

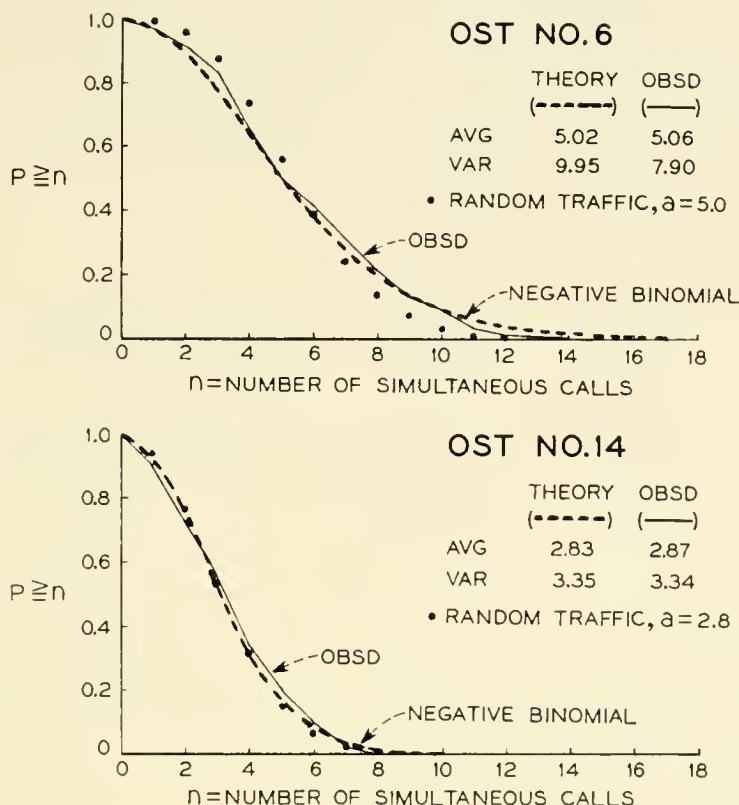


Fig. 22 — Combined overflow loads offered to alternate-route OST trunks from direct interoffice trunks; negative binomial theory vs throwdown observations.

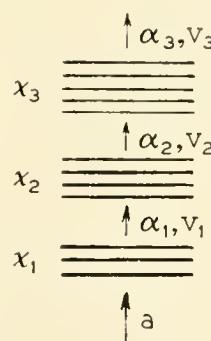


Fig. 23 — A full access group divided at several points to examine the traffic character at each point.

2 or 3. A load of a erlangs is submitted to each subgroup, a having the values 1, 2, 3, 4 or 5. What grade of service will be given?

Solution: The load overflowing each subgroup, when $a = 1$ for example, has the characteristics $\alpha = 0.0625$ and $v = 0.0790$. Then $A' = 2\alpha = 0.125$ and $V' = 2v = 0.158$. Reading on Fig. 26 gives the Equivalent Random values of $A = 1.04$ erlangs, $S = 2.55$ trunks. Reading on Fig. 12.1 with $C + S = 3.55$ when $C = 1$, and $A = 1.04$, we find $\alpha' = 0.0350$ and $\alpha'/(a_1 + a_2) = 0.0175$. We construct Table II in which loss values predicted by the Equivalent Random (ER) Theory are given in columns (3), (5) and (7). For comparison, the corresponding exact values given by Neovius* are shown in columns (2), (4) and (6). (Less exact loss

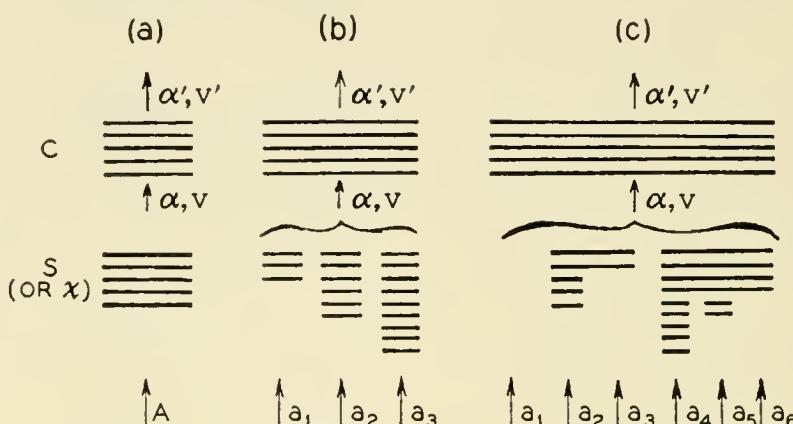


Fig. 24 — Various high usage trunk group arrangements producing the same total overflow α, v .

figures were given previously by Conny Palm¹⁰. The agreement is seen to be excellent for engineering needs for all values in the table.

Example 2: Suppose in Fig. 24(b) the random offered loads and paths are as given in Table III; we desire the proportion of overflow and the overflow load characteristics from an alternate route of 5 trunks.

Solution: The individual overflows α_1, v_1 ; α_2, v_2 ; and α_3, v_3 are read from Figs. 12 and 13 and recorded in columns (4) and (5) of the table. The α and v columns are totalled to obtain the sum-overflow average A' and variance V' . The Equivalent Random load A which, if submitted to S trunks would produce overflow A', V' , is found from Fig. 26. Finally, with A submitted to $S + C$ trunks the characteristics α' and v' , of the load overflowing the C trunks are found. The numerical values obtained

* Artificial Traffic Trials Using Digital Computers, a paper presented by G. Neovius at the First International Congress on the Application of the Theory of Probability on Telephone Engineering and Administration, Copenhagen, June, 1955.

TABLE II—CALCULATION OF LOSS IN A SIMPLE GRADED MULTIPLE
 $g = 2, x_1 = x_2 = 3, a_1 = a_2 = a = 1 \text{ to } 5, C = 1 \text{ to } 3$

Load Submitted to each Subgroup in Erlangs a	Proportion of Each Subgroup Load which Overflows $= \alpha'/(a_1 + a_2)$					
	C = 1		C = 2		C = 3	
	True	ER	True	ER	True	ER
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.01737	0.0175	0.00396	0.0045	0.00077	0.00088
2	0.11548	0.115	0.05630	0.057	0.02438	0.024
3	0.24566	0.246	0.16399	0.163	0.10212	0.103
5	0.35935	0.363	0.27705	0.279	0.20535	0.210
5	0.44920	0.445	0.37336	0.370	0.30308	0.305

for this example are shown in the lower section of Table III. As before, of course, the "lost" calls are assumed cleared, and do not reappear in the system.

Example 3: A load of 18 erlangs is offered through four groups of 10-point selector switches to twenty-two trunks which have been designated as "high usage" paths in an alternate route plan. Which of the trunk arrangements shown in Fig. 27 is to be preferred, and to what extent?

Solution: By successive applications of the Equivalent Random method the overflow percentages for each of the three trunk arrangements are determined. The results are shown in column 2 of Table IV. The difference in percentage overflow between the three trunk plans is small; however, plan 2 is slightly superior followed by plans 3 and 1 in

TABLE III—CALCULATION OF OVERFLOWS FROM A SIMPLE ALTERNATE ROUTE TRUNK ARRANGEMENT

Subgroup Number	Offered Load in Erlangs a	Number of Trunks x	Overflow Loads	
			α	v
1	3.5	3	1.41	1.98
2	5.7	6	1.39	2.40
3	6.0	9	0.45	0.85
	15.2		3.25	5.23

Description of load offered to alternate route: $A' = 3.25, V' = 5.23$.

Equivalent straight multiple: $S = 5.8$ trunks, $A = 8.00$ erlangs (from Fig. 26). Overflow from $C = 5$ alternate route trunks (enter Figs. 12 and 13 with $A = 8.0$ and $S + C = 10.8$): $\alpha' = 0.72, v' = 1.48$.

Proportion of load to commons which overflows = $0.72/3.25 = 0.22$.

Proportion of offered load which overflows = $0.72/15.2 = 0.0475$.

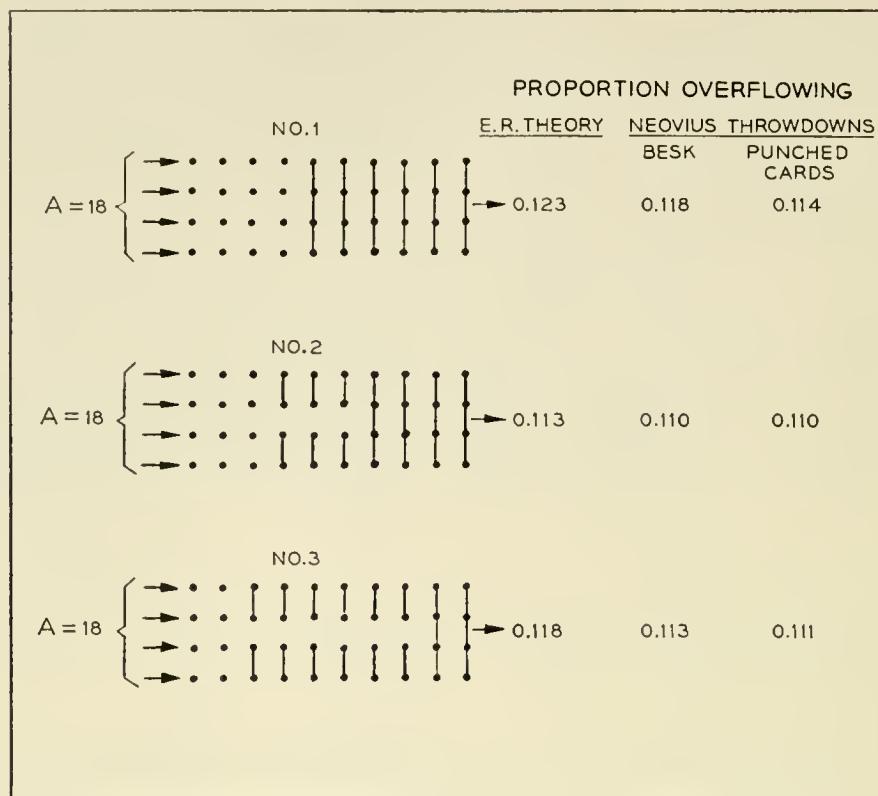


Fig. 27 — Comparison of losses on three graded arrangements of 22 trunks.

that order. The results of extensive simulations made by Neovius on the three trunk plans are available for comparison.* The values so obtained are seen to be very close to the ER theoretical ones; moreover the same order of preference among the three plans is indicated and with closely similar loss differentials between them.

7.3.1. Throwdown Comparisons with Equivalent Random Theory on Simple Alternate Routing Arrangements with Lost Calls Cleared

Results of manually run throwdowns on a considerable number of non-symmetrical single-stage alternate route arrangements are available. Some of these were shown in Fig. 20; they represent part of a projected multi-alternate route layout (to be described later) for outgoing calls from the local No. 1 crossbar Murray Hill-6 office in New York to all other offices in the metropolitan area. The paths hunted over initially are called direct trunks; they overflow calls to Office Selector Tandem (OST) groups, numbered from 1 to 17, which are located in widely dispersed central office buildings in the Greater New York area.

* Loc. cit.

TABLE IV—LOSS COMPARISON OF GRADED ARRANGEMENTS

Plan Number	Estimates of Percentage of Load Overflowing			
	ER Theory	Neovius Throwdowns		
		BESK Computer (262144 calls)	Punched Cards (10,000 calls)	
(1)	(2)	(3)	(4)	
1	12.3	11.81	11.4	
2	11.3	10.98	11.0	
3	11.8	11.25	11.1	

TABLE V—COMPARISON OF THEORY AND THROWDOWNS FOR THE PARAMETERS OF LOADS OVERFLOWING THE COMMON TRUNKS IN SINGLE-STAGE GRADED MULTIPLES

OST (Alternate) Route Group		No. of Groups of Direct Trunks	Total No. of Trunks in Direct Groups	Total Load Offered to Direct Trunks		Total Overflow Load from OST					
Group no.	No. of trunks			Erlangs	Approximate No. of Calls (in 2.7 hours)	Theory		Throwdown			
						α'	v'	α'	v'		
1	6	8	91	68.91	4950	2.00	5.50	2.36	6.52		
2	3	3	45	37.49	2690	2.10	5.60	2.05	6.36		
3	6	6	80	60.62	4355	1.50	4.00	1.30	5.67		
4	3	6	52	38.49	2765	2.30	5.20	2.08	6.43		
5	3	3	17	12.51	900	0.45	0.83	0.49	1.02		
6	4	7	64	48.62	3490	2.50	5.90	2.36	4.88		
7	8	12	78	57.42	4125	2.20	5.60	1.71	4.08		
8	6	9	16	12.96	930	0.82	1.63	0.81	1.11		
9	1	2	22	16.96	1220	1.30	2.60	1.02	1.73		
10	5	6	10	9.52	685	0.78	1.40	1.05	2.07		
11	8	13	16	16.43	1180	1.90	3.80	2.77	7.29		
12	8	9	2	6.88	495	0.70	1.30	0.81	1.83		
13	5	15	33	21.42	1540	1.75	3.30	1.16	2.01		
14	2	7	11	8.05	580	1.46	2.20	1.63	2.14		
15	9	15	8	11.97	860	1.60	3.25	1.55	4.12		
16	11	22	34	27.46	1970	1.75	4.00	1.34	2.26		
17	3	7	4	5.81	420	1.53	2.31	1.43	1.80		
						26.64	58.42	25.92	61.32		

In Table V are given certain descriptive data for the 17 OST trunk arrangements showing numbers of legs of direct trunks, total direct trunks, the offered erlangs and calls, and the mean and variance of the alternate routes' overflows, as obtained by the ER theory and by throwdowns.* The throwdown α' and v' values of the OST overflow

* Additional details of this simulation study are given in Section 7.4.

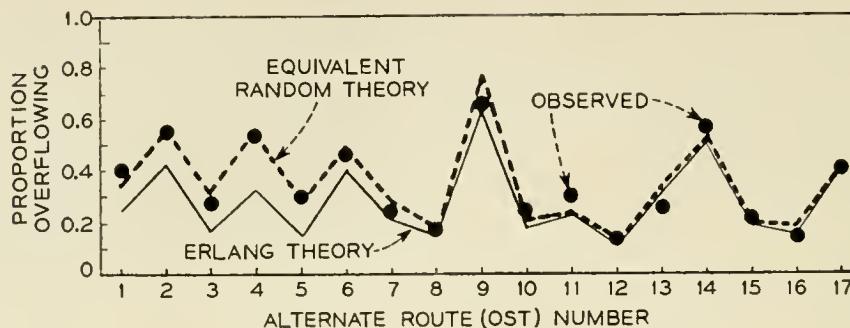


Fig. 28 — Comparison of theoretical and throwdown overflows from a number of first alternate routes.

were obtained by 36-second switch counts of those calls from each OST group which had come to rest on *subsequent* alternate routes.

On Fig. 28 is shown a summary of the observed and calculated proportions of "lost" to "offered" traffic at each OST alternate route group. As may be seen from the figure and the last four columns of Table V, the general agreement is quite good; the individual group variations are probably no more than to be expected in a simulation of this magnitude.

An assumption of randomness (which has sometimes been argued as returning when several overflows are combined) for the load offered to the OST's gives the Erlang E_1 loss curve on Fig. 28. This, as was to be expected, rather consistently understates the loss.

Since "switch-counts" were made on the calls overflowing each OST, the distributions of these overflows may be compared with those estimated by the Negative Binomial theory having the mean and variance predicted above for the overflow. Fig. 29 shows the individual and cumulative probability distributions of the overflow simultaneous calls from the first two OST alternate routes. As will be seen, the agreement is quite good even though this is traffic which has been twice "non-randomized." Comparison of the observed and calculated overflow means and variances in Table V indicates that similar agreement between observed and theoretical fitting distributions for most of the other OST's would be found.

7.3.2. Comparison of Equivalent Random Theory with Field Results on Simple Alternate Routing Arrangements

Data were made available to the author from certain measurements made in 1941 by his colleague C. Clos on the automatic alternate routing trunk arrangement in operation in the Murray Hill-2 central office in New York. Mr. Clos observed for one busy hour the load carried on

several of its OST alternate route groups (similar to those shown in Table V for the Murray Hill-6 office, but not identical) by means of an electromechanical switch-counter having a six-second cycle. During each hour's observation, numbers of calls offered and overflowing were also recorded.

Although the loads offered to the corresponding direct trunks which overflowed to the OST group under observation were not simultaneously measured, such measurements had been made previously for several hours so that the relative contribution from each direct group was closely known. In this way the loads offered to each direct group which produced the total arriving before each OST group could be estimated with considerable assurance. From these direct group loads the character (mean and variance) of the traffic offered to and overflowing the OST's was predicted. The observed proportion of offered traffic which overflowed is shown on Fig. 30 along with the Equivalent Random theory prediction. The general agreement is again seen to be fairly good although with some tendency for the ER theory to predict higher than observed losses in the lower loss ranges; perhaps the disparity on in-

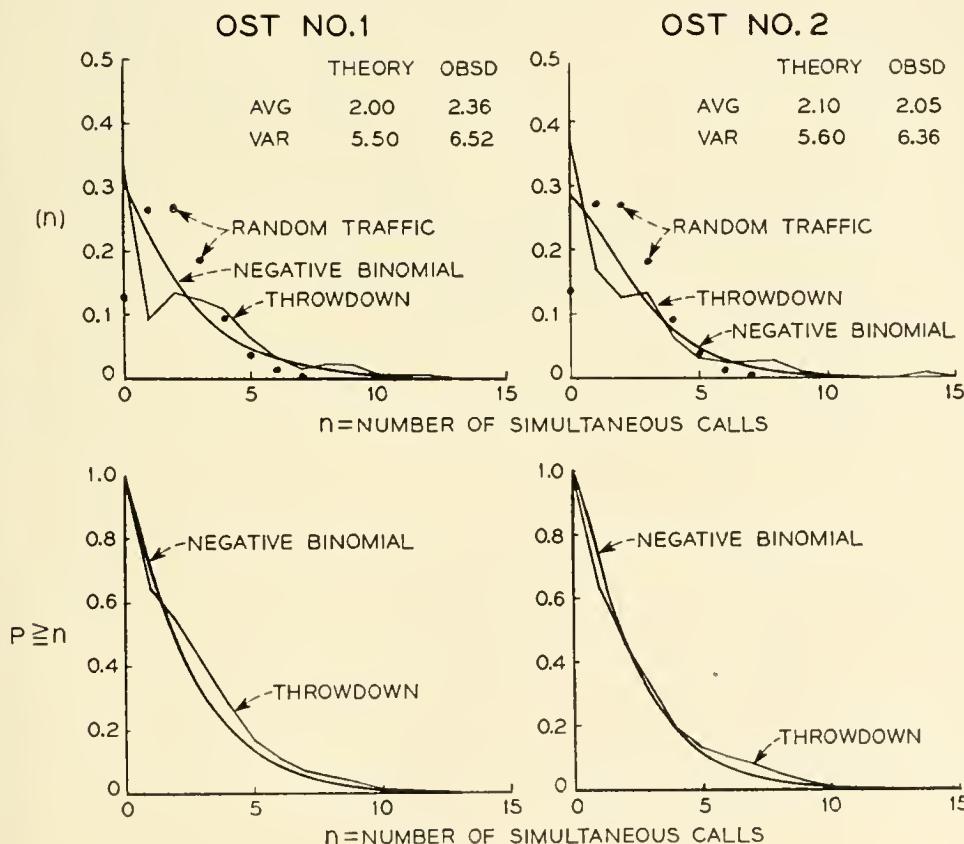


Fig. 29 — Distributions of loads overflowing from first alternate (OST) groups; negative binomial theory versus throwdown observations.

dividual OST groups is within the limits one might expect for data based on single-hour observations and for which the magnitudes of the direct group offered loads required some estimation. The assumption of random traffic offered to the OST gives, as anticipated, loss predictions (Erlang E_1) consistently below those observed.

More recently extensive field tests have been conducted on a working toll automatic alternate route system at Newark, New Jersey. High usage groups to seven distant large cities overflowed calls to the Newark-Pittsburgh alternate (final) route. Data describing the high usage groups and typical system busy hour loads are given in Table VI. (The loads, of course, varied considerably from day to day.) The size of the Pittsburgh route varied over the six weeks of the 1955 tests from 64 to 71 trunks. Altogether the system comprised some 255 intertoll trunks.

Observations were made at the Newark end of the groups by means of a Traffic Usage Recorder — making switch counts every 100 seconds — and by peg count and overflow registers. Register readings were photographically recorded by half-hourly, or more frequent, intervals. To

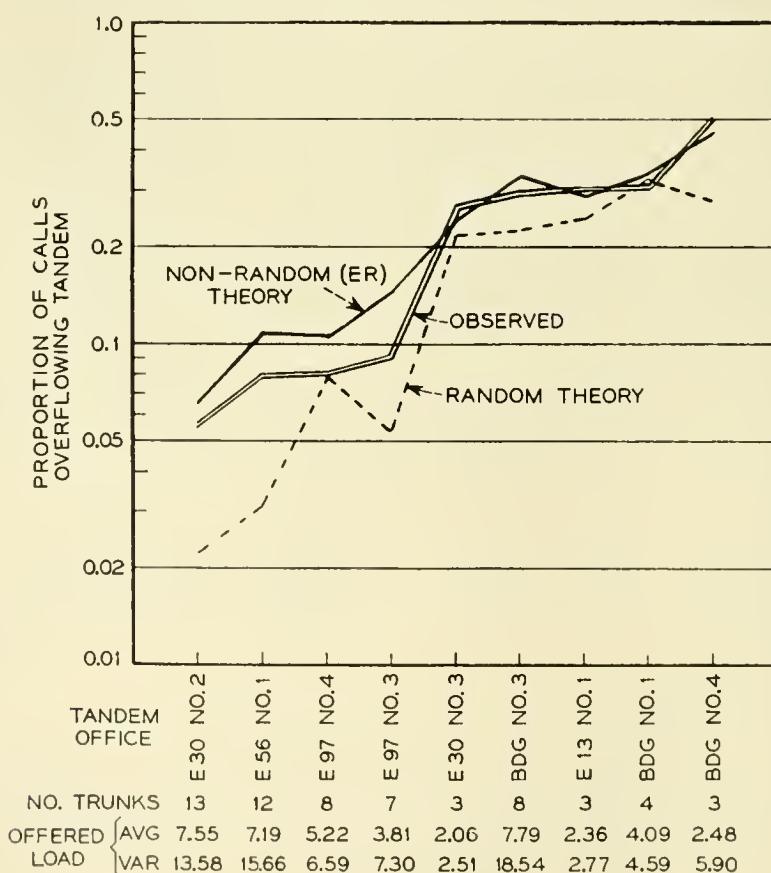


Fig. 30 — Observed tandem overflows in alternate route study at Murray Hill-2 (New York) 1940-1941.

TABLE VI—HIGH USAGE GROUPS AND TYPICAL SYSTEM
BUSY HOUR LOADS

High Usage Group, Newark to:	Length of Direct Route (Air Miles)	Nominal Size of Group (Number of Trunks)	Typical Offered Load (erlangs)
Baltimore.....	170	18	19
Cincinnati.....	560	42	43
Cleveland.....	395	27	26
Dallas.....	1375	33	34
Detroit.....	470	37	36
Kansas City.....	1100	26	23
New Orleans.....	1170	5	4

compare theory with the observed overflow from the final route, estimates of the offered load A' and its variance V' are required. In the present case, the total load offered to the final route in each hour was estimated as

$$A' = \frac{\text{Peg Count of Calls Offered to Pittsburgh Group}}{(\text{Peg Count of Offered Calls}) - (\text{Peg Count of Overflow Calls})} \times \frac{\text{Average Load Carried by Pittsburgh Group}}{\text{Average Load Carried by All Groups}}$$

The variance V' of the total load offered to the final route was estimated for each hour as

$$V' = \text{Variance of Offered Load} \\ = A' - \sum_{i=1}^7 \alpha_i + \sum_{i=1}^7 v_i$$

where α_i and v_i are, respectively, the average and variance of the load overflowing from the i th high usage group. (The expression, $A' = \sum_{i=1}^7 \alpha_i$, is an estimate of the average — and, therefore of the variance — of the first-routed traffic offered directly to the final route. Thus the total variance, V' , is taken as the sum of the direct and overflow components.) Using A' , V' and the actual number, C , of final route trunks in service, the proportion of offered calls expected to overflow was calculated for the traffic and trunk conditions seen for 25 system busy hours from February 17 to April 1, 1955 on the Pittsburgh route. The results are displayed on Fig. 31, where certain traffic data on each hour are given in the lower part of the figure. The hours are ordered — for convenience in plotting and viewing — by ascending proportions of calls overflowing the group; observed results are shown by the double line

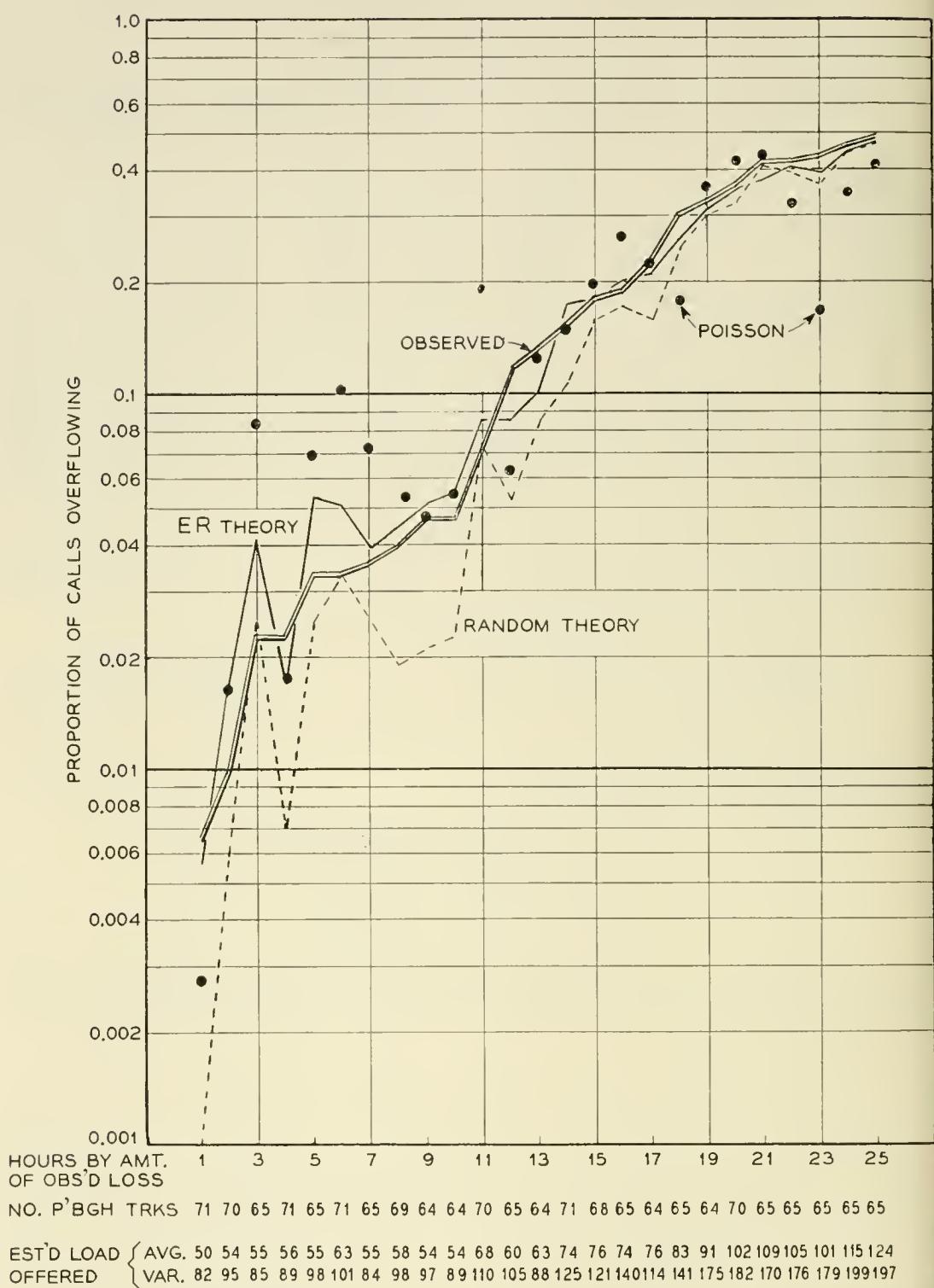


Fig. 31—Final route (Newark-Pittsburgh) overflows in 1955 toll alternate route study.

curve. The superposed single line is the corresponding estimate by ER theory of the hour-to-hour call losses. As may be seen, theory and observation are in good agreement both point by point and on the average over the range of losses from 0.01 to 0.50. The dashed line shows the prediction of final route loss for each hour on the assumption that the offered traffic A' was random. Such an assumption gives consistently low estimates of the existing true loss.

As of interest, a series of heavy dots is included on Fig. 31. These are the result of calculating the Poisson Summation, $P(C,L)$, where L is the average load carried on, rather than offered to, the C trunks. It is interesting that just as in earlier studies in this paper on straight groups of intertoll trunks (for example as seen on Fig. 7), the Poisson Summation with load carried taken as the load offered parameter, gives loss values surprisingly close to those observed. Also, as before, this summation has a tendency to give too-great losses at light loadings of a group and too-small losses at the heavier loadings.

7.4 Prediction of Traffic Passing Through a Multi-Stage Alternate Route Network

In the contemplated American automatic toll switching plan, wide advantage is expected to be taken of the efficiency gains available in multi-alternate routing. Thus any procedure for traffic analysis and prediction needs to be adaptable for the more complex multi-stage arrangements as well as the simpler single-stage ones so far examined. Extension of the Equivalent Random theory to successive overflows is easily done since the characterizing parameters, average and variance, of the load overflowing a group of paths are always available.

Since few cases of more than single-stage automatic alternate routing are yet in operation in the American toll plant, it is not readily possible to check an extension of the theory with actual field data. Moreover collecting and analyzing observations on a large operating multi-alternate route system would be a comparatively formidable experiment.

However, in New York city's local interoffice trunking there is a very considerable development of multi-alternate routing made possible by the flexibility of the marker arrangements in the No. 1 crossbar switching system. None of these overflow arrangements has been observed as a whole, simultaneously and in detail. The Murray Hill-2 data in OST groups reviewed in Section 7.3.2 were among the partial studies which have been made.

In connection with studies made just prior to World War II on these

TABLE VII — SUM OF DIRECT GROUP OVERFLOW LOADS,
OFFERED TO OST'S

	Theory	Observed
Average.....	86.06	87.12
Variance.....	129.5	127.4

local multi-alternate route systems, a throwdown was made in 1941 on a proposed trunk plan for the Murray Hill-6 office. The arrangement of trunks is shown on Fig. 32. Three successive alternate routes, Office Selector Tandems (OST), Crossbar Tandem (XBT), and Suburban Tandem (ST), are available to the large majority of the 123 direct trunk groups leading outward to 169 distant offices. (The remaining 46 parcels of traffic did not have direct trunks to distant offices but, as indicated on the diagram, offered their loads directly to a tandem group.) A total of 726 trunks is involved, carrying 475 erlangs of traffic.

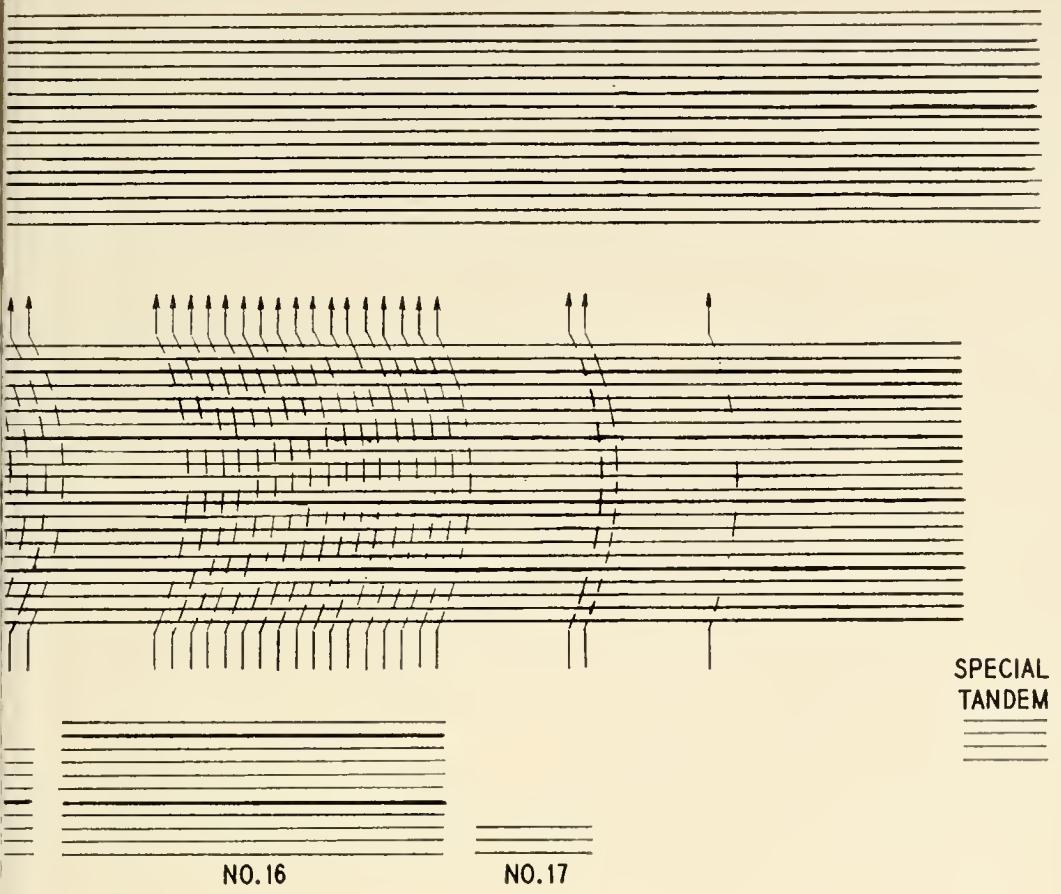
A throwdown of 34,001 offered calls corresponding to 2.7 hours of traffic was run. Calls had approximate exponential holding times, averaging 135 seconds. Records were kept of numbers of calls and the load from the traffic parcels offered to each direct group, as they were carried or passed beyond the groups of paths to which they had access. Loads carried by each trunk in the system were also observed by means of a 36-second "switch-count." (The results on the 17 OST groups reported in Section 7.3.1 were part of this study.)

Comparisons of observation and theory which are of interest include the combined loads to and overflowing the 17 OST's. Observed versus calculated parameters (starting with theory from the original direct group submitted loads) are given in Table VII. The agreement is seen to be very good.

The corresponding comparison of total load from all the OST's is given in Table VIII. Again the agreement is highly satisfactory.

Not all of the overflow from the OST's was offered to the 22 crossbar tandem trunks; for economic reasons certain parcels bypassed XBT and were sent directly to Suburban Tandem.* This posed the problem of breaking off certain portions of the overflow from the OST's, to be added again to the overflow from XBT. An estimate was needed of the contribution made by each parcel of direct group traffic to any OST's overflow. These were taken as proportional to the loads offered the OST by each direct group (this assumes that each parcel suffers the same over-

* In the toll alternate route system bypassing of this sort will not occur.



crossbar office.

SUBURBAN TANDEM

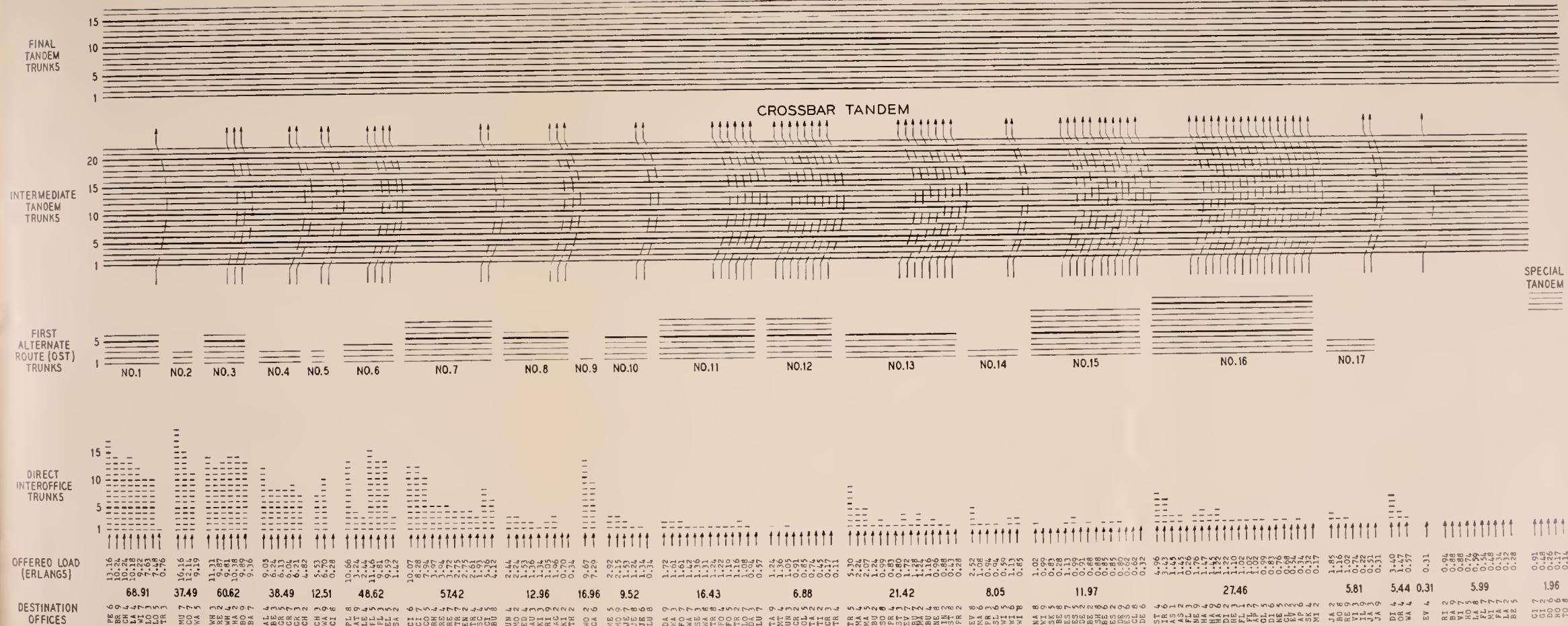


Fig. 32. — Multi-alternate route trunking arrangement at Murray Hill—6 (New York) local No. 1 crossbar office.

flow probability). The variance of this overflow portion by-passing XBT was estimated by assigning to it the same variance-to-average ratio as was found for the total load overflowing the OST. Subtracting the means and variances so estimated for all items by-passing XBT, left an approximate load for XBT from each OST. Combining these corrected overflows gave mean and variance values for offered load to XBT. Observed values

TABLE VIII — SUM OF LOADS OVERFLOWING OST's

	Theory	Observed
Average.....	26.64	25.92
Variance.....	58.42	61.32

TABLE IX — LOAD OFFERED TO CROSSBAR TANDEM

	Theory	Observed
Average.....	25.18	25.51
Variance.....	47.67	56.10

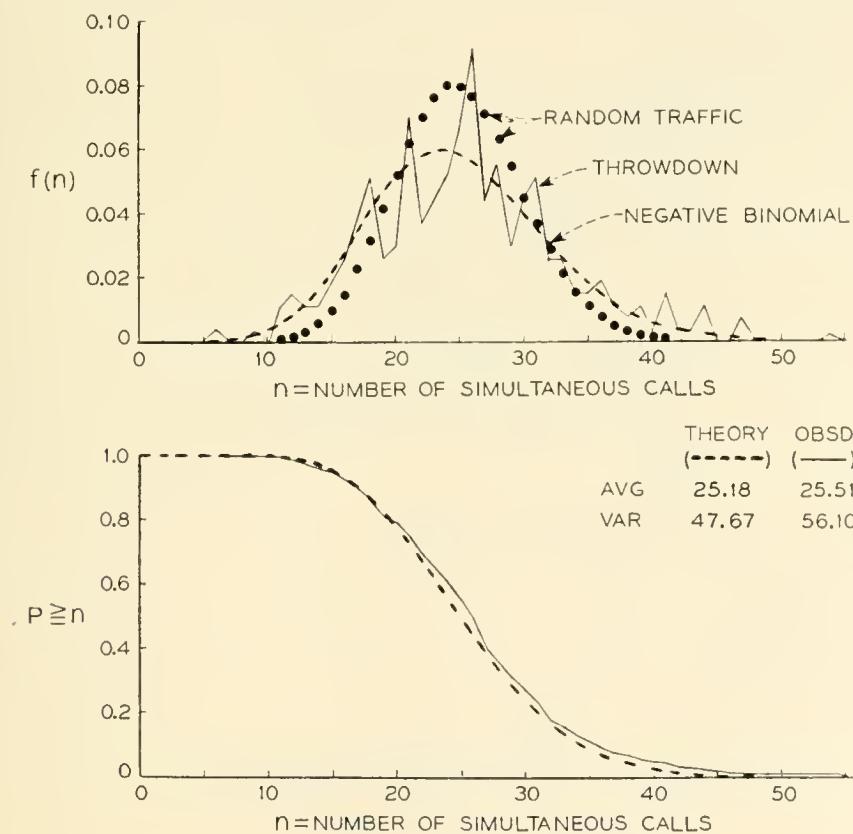


Fig. 33 — Distribution of load offered to crossbar tandem trunks; negative binomial theory versus throwdown observations.

flow probability). The variance of this overflow portion by-passing XBT was estimated by assigning to it the same variance-to-average ratio as was found for the total load overflowing the OST. Subtracting the means and variances so estimated for all items by-passing XBT, left an approximate load for XBT from each OST. Combining these corrected overflows gave mean and variance values for offered load to XBT. Observed values

TABLE VIII — SUM OF LOADS OVERFLOWING OST's

	Theory	Observed
Average.....	26.64	25.92
Variance.....	58.42	61.32

TABLE IX — LOAD OFFERED TO CROSSBAR TANDEM

	Theory	Observed
Average.....	25.18	25.51
Variance.....	47.67	56.10

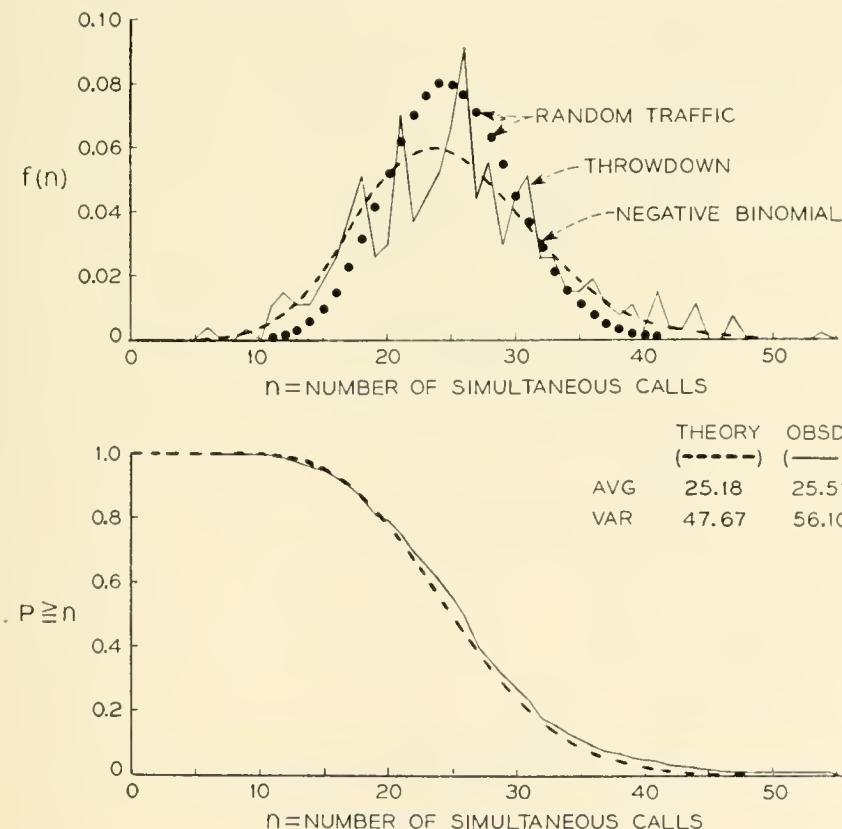


Fig. 33 — Distribution of load offered to crossbar tandem trunks; negative binomial theory versus throwdown observations.

TABLE X — LOAD OVERFLOWING CROSSBAR TANDEM

	Theory	Observed
Average.....	6.55	6.47
Variance.....	23.80	33.48

and those calculated (in the above manner) are given in Table IX. Fig. 33 shows the distribution of XBT offered loads, observed and calculated. The agreement is very satisfactory. The random traffic (Poisson) distribution, is of course, considerably too narrow.

In a manner exactly similar to previous cases, the Equivalent Random load method was applied to the XBT group to obtain estimated parameters of the traffic overflowing. Comparison of observation and theory at this point is given in Table X.

Fig. 34 shows the corresponding observed and calculated distributions

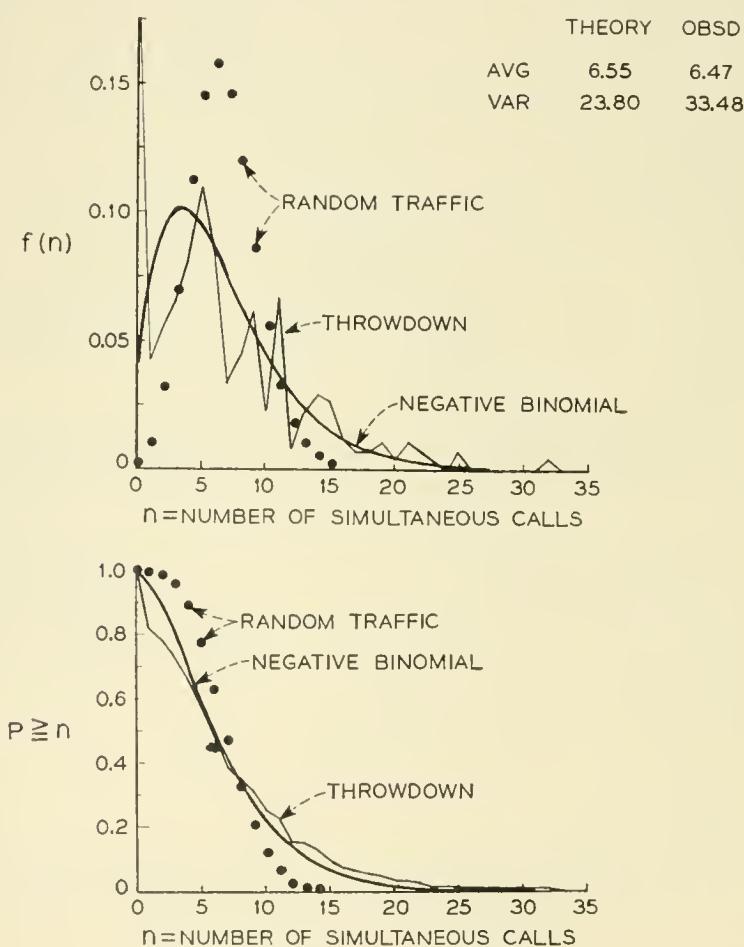


Fig. 34 — Distribution of calls from crossbar tandem trunks; negative binomial theory versus throwdown observations.

of simultaneous calls. The agreement again is reasonably good, in spite of the considerable disparity in variances.

The overflow from XBT and the load which by-passed it, as well as some other miscellaneous parcels of traffic, were now combined for final offer to the Suburban Tandem group of 17 trunks. The comparison of parameters here is again available in Table XI. On Fig. 35 are shown the observed and calculated distributions of simultaneous calls for the load offered to the ST trunks. The agreement is once again seen to be very satisfactory.

We now estimate the loss from the ST trunks for comparison with the actual *proportion of calls* which failed to find an idle path, and finally

TABLE XI — LOAD OFFERED TO SUBURBAN TANDEM

	Theory	Observed
Average.....	15.38	14.52
Variance.....	42.06	48.53

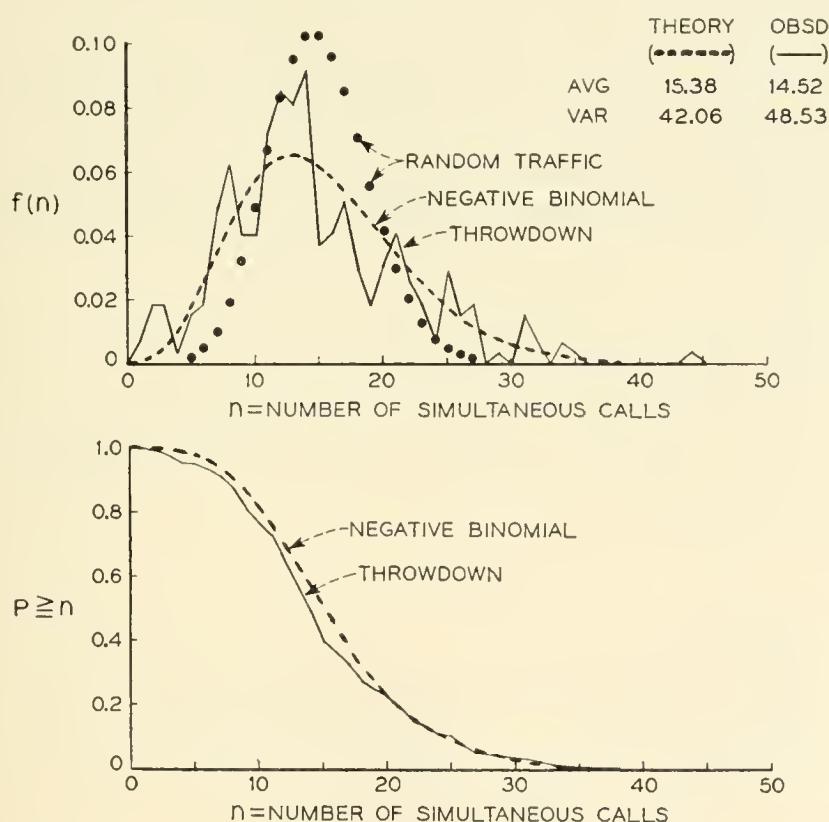


Fig. 35 — Distribution of load offered to suburban tandem trunks; negative binomial theory versus throwdown observations.

TABLE XII — GRADE OF SERVICE ON ST GROUP

	Theory	Observation	Observation
Load submitted (erlangs)	15.38	14.52	Number of calls submitted 1057
Load overflowing (erlangs)	3.20	2.63	Number of calls overflowing 200
Proportion load overflowing	0.209	0.181	Proportion of calls overflowing 0.189

TABLE XIII — GRADE OF SERVICE ON THE SYSTEM

	Theory	Observed
Total load submitted.....	475 erlangs	34,001 calls
Total load overflowing.....	3.20 erlangs	200 calls
Proportion of load not served.....	0.00674	0.00588

compare the proportions of all traffic offered the system which failed to find a trunk immediately. See Tables XII and XIII.

After these several and varied combinations of offered and overflowed loads to a system of one direct and three alternate routes it is seen that the final prediction of amount of load finally lost beyond the ST trunks is gratifyingly close to that actually observed in the throwdown. The prediction of the system grade of service is, of course, correspondingly good.

It is interesting in this connection to examine also the proportions overflowing the ST group when summarized by parcels contributed from the several OST groups. The individual losses are shown on Fig. 36; they appear well in line with the variation one would expect from group to group with the moderate numbers of calls which progressed this far through the multiple.

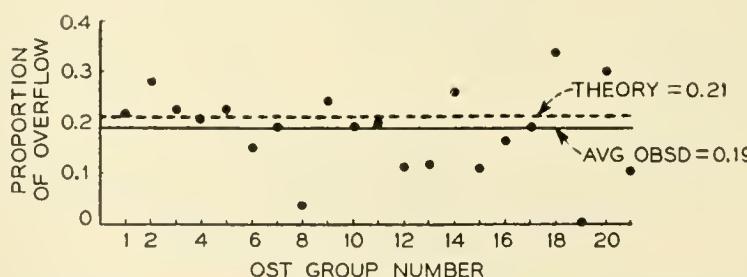


Fig. 36 — Overflow calls on third alternate (ST) route.

7.4.1 Correlation of Loss with Peakedness of Components of Non-Random Offered Traffic

Common sense suggests that if several non-random parcels of traffic are combined, and their joint proportion of overflow from a trunk group is P , the parcels which contain the more peaked traffic should experience overflow proportions larger than P , and the smoother traffic an overflow proportion smaller than P . It is by no means clear however, *a priori*, the extent to which this would occur. One might conjecture that if any one parcel's contribution to the total combined load is small, its loss would be caused principally by the aggregate of calls from the other parcels, and consequently its own loss would be at about the general average loss P , and hence not very much determined by its own peakedness. The Murray Hill-6 throwdown results may be examined in this respect. The mean and variance of each OST-parcel of traffic, for example, arriving at the final ST route was recorded, together with, as noted before, its own proportion of overflow from the ST trunks. The variance/mean overdispersion ratio, used as a measure of peakedness, is plotted for each parcel of traffic against its proportion of loss on Fig. 37. There is an undoubtedly, but only moderate, increase in proportion of overflow with increased peakedness in the offered loads.

It is quite possible, however, that by recognizing the differences between the service given various parcels of traffic, significant savings in final route trunks can be effected for certain combinations of loads and trunking arrangements. Of particular interest is the service given to a parcel of random traffic offered directly to the final route when compared

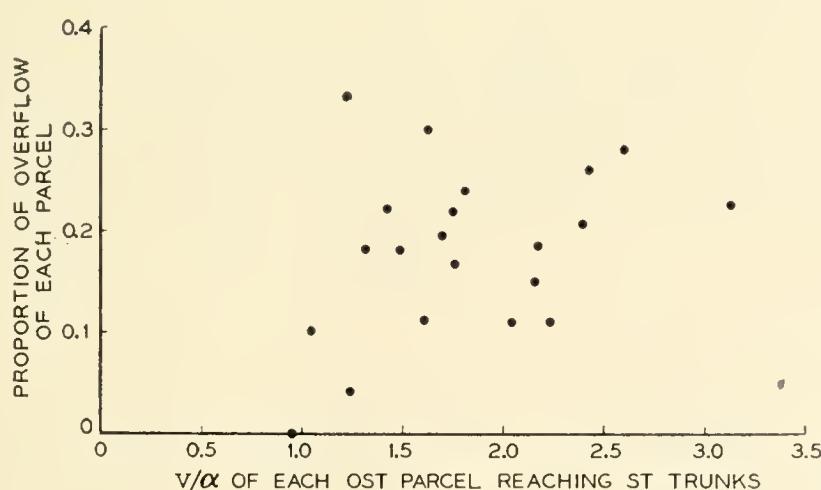


Fig. 37 — Effect of peakedness on overflow of a parcel of traffic reaching an alternate route.

with that received by non-random parcels overflowing to it from high usage groups.

7.5 Expected Loss on First Routed Traffic Offered to Final Route

The congestion experienced by the first-routed traffic offered to the final group in a complex alternate route arrangement [such as the right hand parcels in Figs. 10(c) and (d)] will be the same as encountered in a series of random tests of the final route by an independent observer, that is, it will be the proportion of time that all of the final trunks are busy. As noted before, the distribution of simultaneous calls n (and hence the congestion) on the C final trunks produced by some specific arrangement of offered load and high usage trunks can be closely simulated by that due to a single Equivalent Random load offered to a straight group of $S + C$ trunks. Then the proportion of time that the C trunks are busy in such an equivalent system provides an estimate of the corresponding time in the real system; and this proportion should be approximately the desired grade of service given the first routed traffic.

Brockmeyer¹¹ has given an expression (his equation 36) for the proportion of time, R_1 , in a simple $S + C$ system with random offer A , and "lost calls cleared," that all C trunks are busy, independent of the condition of the S -trunks:

$$\begin{aligned} R_1 &= f(S, C, A) \\ &= E_{1,s+c}(A) \frac{\sigma_{c+1}(S)}{\sigma_c(S)} \end{aligned} \quad (30)$$

where

$$\sigma_c(S) = \sum_{m=0}^S \binom{C - 1 + m}{m} \frac{A^{S-m}}{(S-m)!}$$

However, $\sigma_c(S)$ is usually calculated more readily step-by-step using the formula

$$\sigma_c(S) = \sigma_c(S-1) + \sigma_{c-1}(S),$$

starting with

$$\sigma_c(0) = 1 \quad \text{and} \quad \sigma_0(S) = A^S/S!$$

The average load carried on the C paths is clearly

$$L_C = A[E_{1,s}(A) - E_{1,s+c}(A)], \quad (31)$$

and the variance of the carried load can be shown to be*

$$V_c = AL_c \frac{\sigma_1(S)}{\sigma_2(S)} - ACE_{1,s+c}(A) + L_c - L_c^2 \quad (32)$$

On Fig. 38, R_1 values are shown in solid line curves for several combinations of A and C over a small range of S trunks. The corresponding losses R_2 for all traffic offered the final group, where $R_2 = \alpha'/A'$, are shown as broken curves on the same figure. The R_2 values are always above R_1 , agreeing with the common sense conclusion that a random component of traffic will receive better service than more peaked non-random components.

However, there are evidently considerable areas where the loss difference between the two R 's will not be large. In the loss range of principal interest, 0.01 to 0.10, there is less proportionate difference between the R 's as the $A = C$ paired values increase on Fig. 38. For example, at $R_2 = 0.05$, and $A = C = 10$, $R_2/R_1 = 0.050/0.034 = 1.47$; while for $A = C = 30$, $R_2/R_1 = 0.050/0.044 = 1.13$. Similarly for $A = 2C$, the R_2/R_1 ratios are given in Table XIV. Again the rapid decrease in the R_2/R_1 ratio is notable as A and C increase.

F. I. Tånge of the Swedish Telephone Administration has performed elaborate simulation studies on a variety of semi-symmetrical alternate route arrangements, to test the disparity between the R_1 and R_2 types of losses on the final route.† For example if g high-usage groups of 8 paths each, jointly overflow 2.0 erlangs to a final route which also serves 2.0 erlangs of first routed traffic, Tånge found the differences in losses between the two 2-erlang parcels, $R_{\text{high usage (h.u.)}} - R_1$, shown in column 9 of Table XV. The corresponding ER calculations are performed in columns 2 to 8, the last of which is comparable with the throwdown values of column 9. The agreement is not unreasonable considering the sensitiveness of determining the difference between two small probabilities of loss. A quite similar agreement was found for a variety of other loads and trunk arrangements.

* In terms of the first two factorial moments of n : V_c is given by

$$V_c = M_{(2)} + M_{(1)} - M_{(1)}^2, \quad \text{where } M_{(1)} = L_c$$

General expressions $M_{(i)}$ for the factorial moments of n are derived in an unpublished memorandum by J. Riordan.

† Optimal Use of Both-Way Circuits in Cases of Unlimited Availability, a paper by F. I. Tånge, presented at the First International Congress on the Application of the Theory of Probability in Telephone Engineering and Administration, June 1955, Copenhagen.

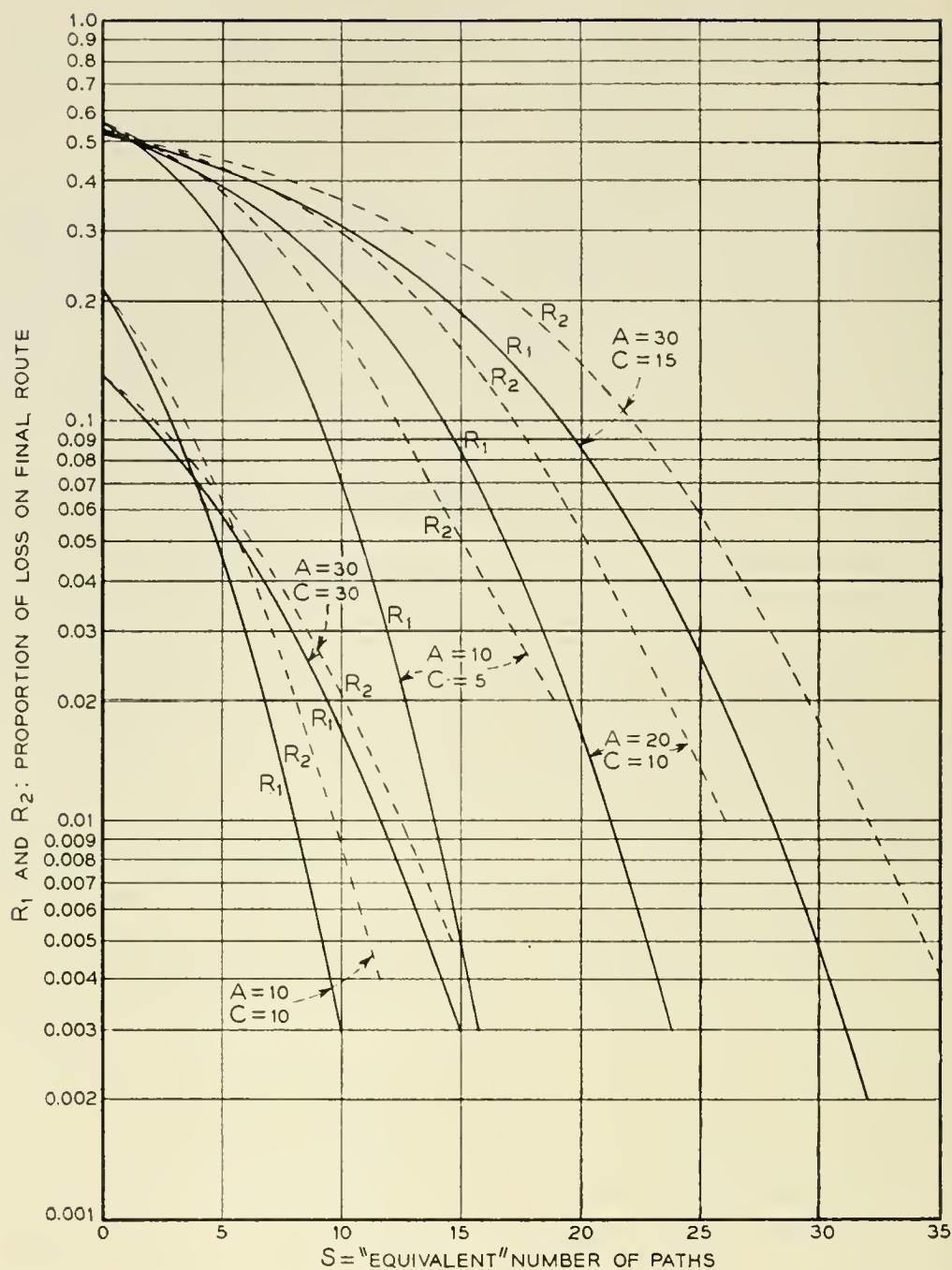


Fig. 38 — Comparison of R_1 and R_2 losses under various load and trunk conditions.

TABLE XIV—THE R_2/R_1 RATIOS FOR $A = 2C$

A	C	R_2/R_1 when $R_2 = 0.05$
10	5	10.6
20	10	3.25
30	15	2.44

TABLE XV—COMPARISON OF E.R. THEORY AND THROWDOWNS ON
DISPARITY OF LOSS BETWEEN HIGH USAGE OVERFLOW AND
RANDOM OFFER TO A FINAL GROUP

(8 trunks in each high usage group; 9 final trunks serving 2.0 erlangs
high usage overflow and 2.0 erlangs first routed traffic.)

Number of Groups of 8 High Usage Trunks	ER Theory ($A' = 4.0$)							T ^o ange Throwdown $R_{h.u.} - R_1$
	V'	A	S	$R_2 = \alpha'/A'$	R_1	$\frac{R_{h.u.}}{2R_2 - R_1}$	$\frac{R_{h.u.} - R_1}{2(R_2 - R_1)}$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	5.77	7.51	4.17	0.0375	0.0251	0.0499	0.0248	0.0180
2	5.80	7.50	4.25	0.0383	0.0255	0.0511	0.0256	0.0247
3	5.74	7.44	4.08	0.0369	0.0248	0.0490	0.0242	0.0286
4	5.68	7.30	3.91	0.0362	0.0247	0.0477	0.0230	0.0276
5	5.64	7.20	3.80	0.0355	0.0242	0.0468	0.0226	0.0245
6	5.58	7.06	3.64	0.0350	0.0240	0.0460	0.0220	0.0221
7	5.55	7.00	3.56	0.0345	0.0238	0.0452	0.0204	0.0202
8	5.51	6.91	3.45	0.0335	0.0236	0.0434	0.0198	0.0188
9	5.47	6.81	3.34	0.0325	0.0231	0.0419	0.0188	0.0177
10	5.45	6.76	3.29	0.0312	0.0225	0.0399	0.0174	0.0166

Limited data are available showing the disparity of R_1 and R_2 in actual operation in a range of load and trunk values well beyond those for which R_1 values have been calculated. Special peg count and overflow registers were installed for a time on the final route during the 1955 Newark alternate route tests. These gave separate readings for the calls from high usage groups, and for the first routed Newark to Pittsburgh calls. Comparative losses for 17 hours of operation over a wide range of loadings are shown on Fig. 39. The numbers at each pair of points give the per cent of final route offered traffic which was first routed (random). In general, approximately equal amounts of the two types of traffic were offered.

In 6 of the hours almost identical loss ratios were observed, in 7 hours the overflow-from-high-usage calls showed higher losses, and in 4 hours lower losses, than the corresponding first routed calls. The non-random calls clearly enjoyed practically as good service as the random calls. This result is not in disagreement with what one might expect from theory. To compare directly with the Newark-Pittsburgh case we should need curves on Fig. 38 expanded to correspond to $A' = 70$, $V' = 120$, $R_1 = 0.0225$, $R_2 = 0.0312$. Examining the mid-range case of $C = 65$, $A' = 70$, $V' = 120$, we find $A \doteq 123$, $S \doteq 54$. Here A is approximately $2C$; extrapolating the $A = 2C$ curves of Fig. 38 to these much higher values of A and C suggests that R_2/R_1 would be but little different from unity.

It is clear from the above theory, throwdowns, and actual observation that there are certain areas where the service differences given first routed and high usage trunk overflow parcels of traffic are significant. In Section 8, where practical engineering methods are discussed, curves are presented which permit recognition of this fact in the determination of final trunk requirements.

7.6 Load on Each Trunk, Particularly the Last Trunk, in a Non-Slipped Alternate Route

In the engineering of alternate route systems it is necessary to determine the point at which to limit a high usage group of trunks and send the overflow traffic via an alternate route. This is an economic problem whose solution requires an estimate of the load which will be carried on

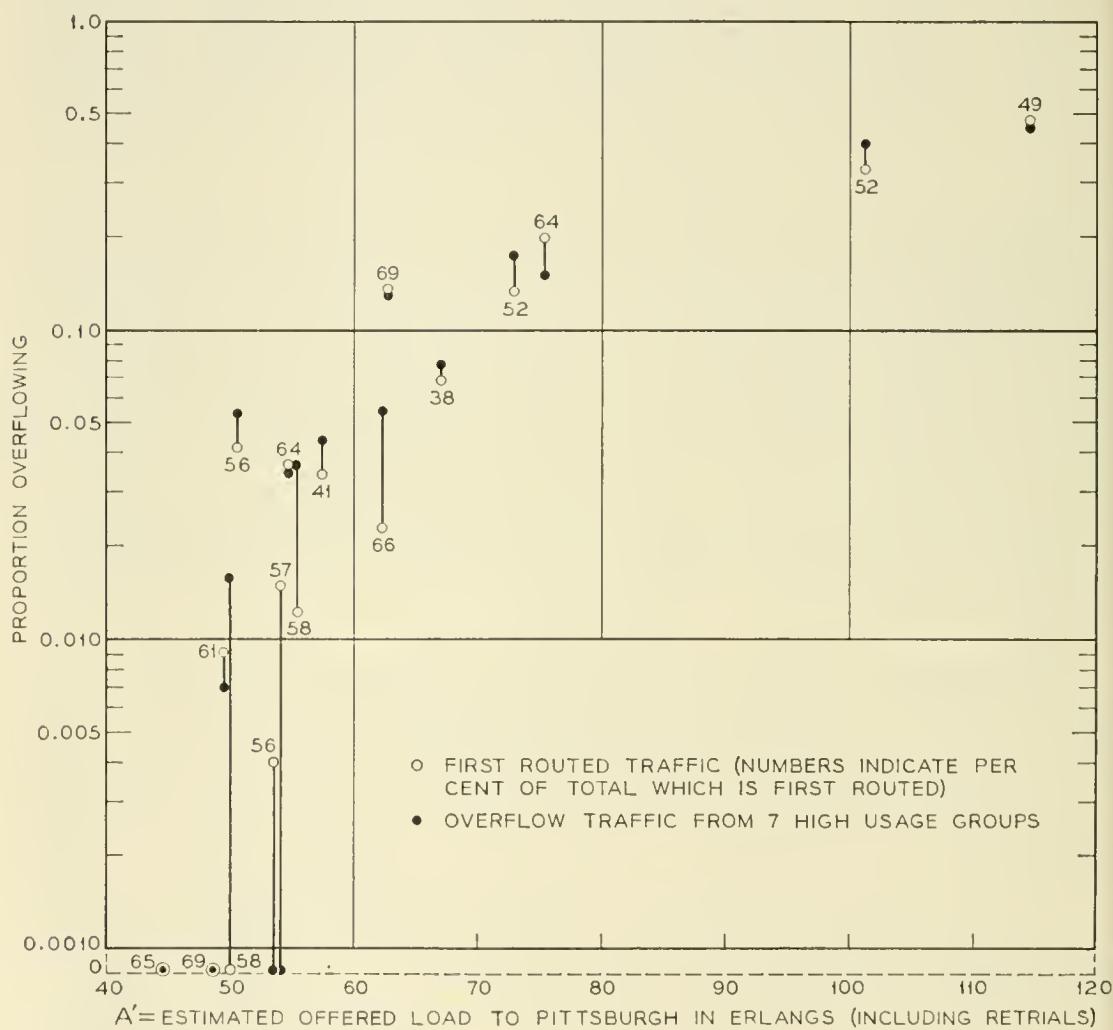


FIG. 39 — Comparison of losses on final route (Newark to Pittsburgh) for high usage overflow and first routed traffic.

the last trunk of a straight high usage group of any specified size, carrying either first or higher choice traffic or a mixture thereof.*

The Equivalent Random theory readily supplies estimates of the loads carried by any trunk in an alternate routing network. After having found the Equivalent Random load A offered to $S + C$ trunks which corresponds to the given parameters of the traffic offered to the C trunks, it is a simple matter to calculate the expected load ℓ on any one of the C trunks if they are not slipped or reversed. The load on the i th trunk in a simple straight multiple (or the $S + j$ th in a divided multiple of S lower and C upper trunks), is

$$\ell_i = l_{s+j} = A[E_{1,s+j-1}(A) - E_{1,s+j}(A)] \quad (33)$$

where $E_{1,n}(A)$ is the Erlang loss formula. A moderate range of values of ℓ_i versus load A is given on Figure 40.†

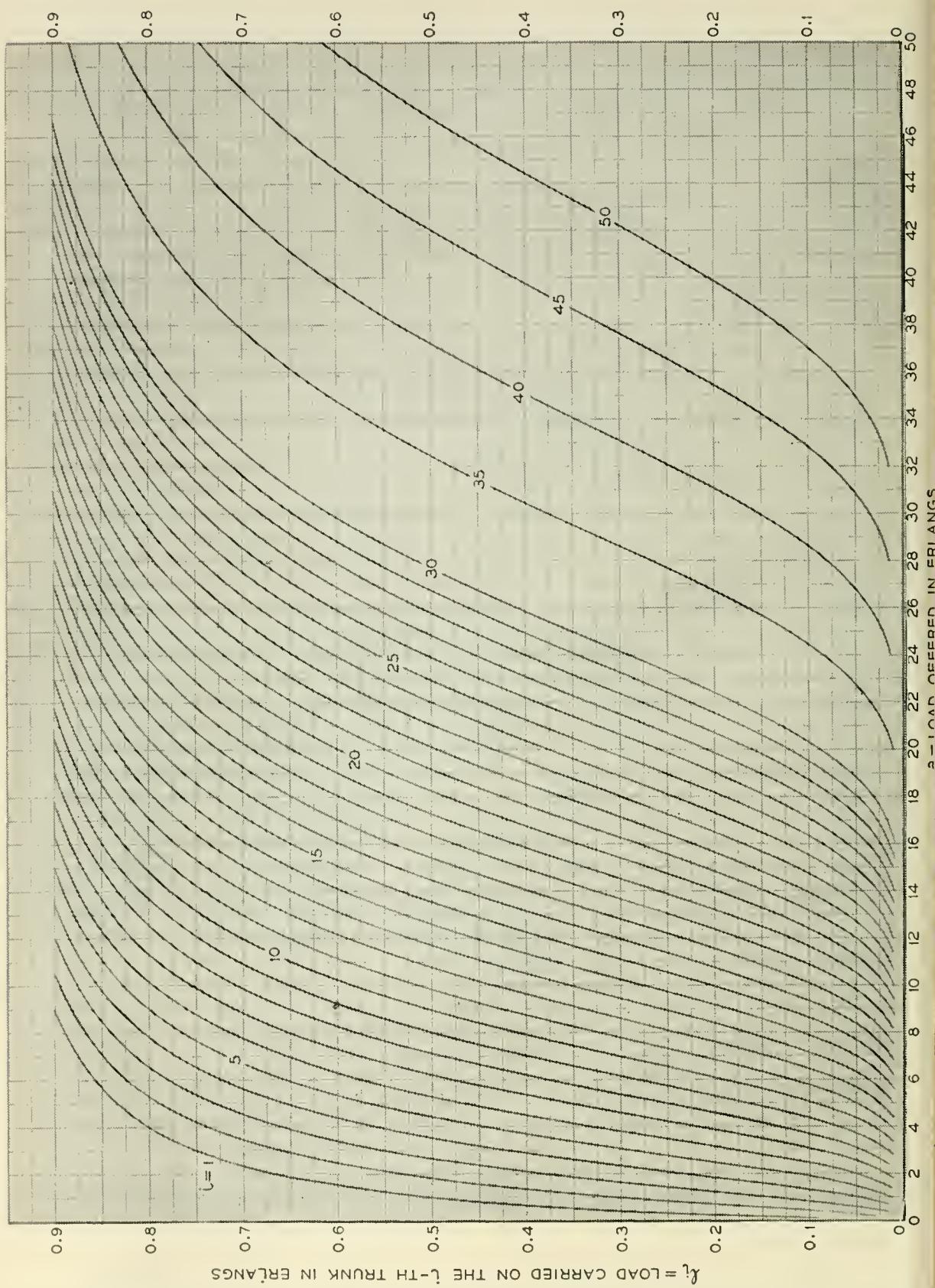
Using this method, selected comparisons of theoretical versus observed loads carried on particular trunks at various points in the Murray-Hill-6 throwdown are shown in Fig. 41; these include the loads on each of the trunks of the first two OST groups of Fig. 32, and on the second and third alternate routes, crossbar and suburban tandem, respectively. The agreement is seen to be fairly good, although at the tail end of the latter two groups the observed values drop away somewhat from the theoretical ones. There seems no explanation for this beyond the possibility that the throwdown load samples here are becoming small and might by chance have deviated this far from the true values (or the arbitrary breakdown of OST overflows into parcels offered to and bypassing XBT may well have introduced errors of sufficient amount to account for this disparity). As is well known, (33) gives good estimates of the loads carried by each trunk in a high usage group to which random (Poisson) traffic is offered; this relationship has long been used for the purpose in Bell System trunk engineering.

8. PRACTICAL METHODS FOR ALTERNATE ROUTE ENGINEERING

To reduce to practical use the theory so far presented for analysis of alternate route systems, working curves are needed incorporating the

* The proper selection point will be where the circuit annual charge per erlang of traffic carried on the last trunk, is just equal to the annual charge per erlang of traffic carried by the longer (usually) alternate route enlarged to handle the overflow traffic.

† A comprehensive table of ℓ_i is given by A. Jensen as Table IV in his book "Moe's Principle," Copenhagen, 1950; coverage is for $\ell \geq 0.001$ erlang, $i = 1(1)140$; $A = 0.1(0.1)10, 10(1)50, 50(4)100$. Note that $n + 1$, in Jensen's notation, equals i here.



pertinent load-loss relationships. The methods so far discussed, and proposed for use, will be briefly reviewed.

The dimensioning of each high usage group of trunks is expected to be performed in the manner currently in use, as described in Section 7.6. The critical figure in this method is the load carried on the last high usage trunk, and is chosen so as to yield an economic division of the offered load between high usage and alternate route trunks. Fig. 40 is one form of load-on-each-trunk presentation suitable for choosing economic high usage group size once the permitted load on the last trunk is established.

The character (average α and variance v) of the traffic overflowing each high usage group is easily found from Figs. 12 and 13 (or equivalent

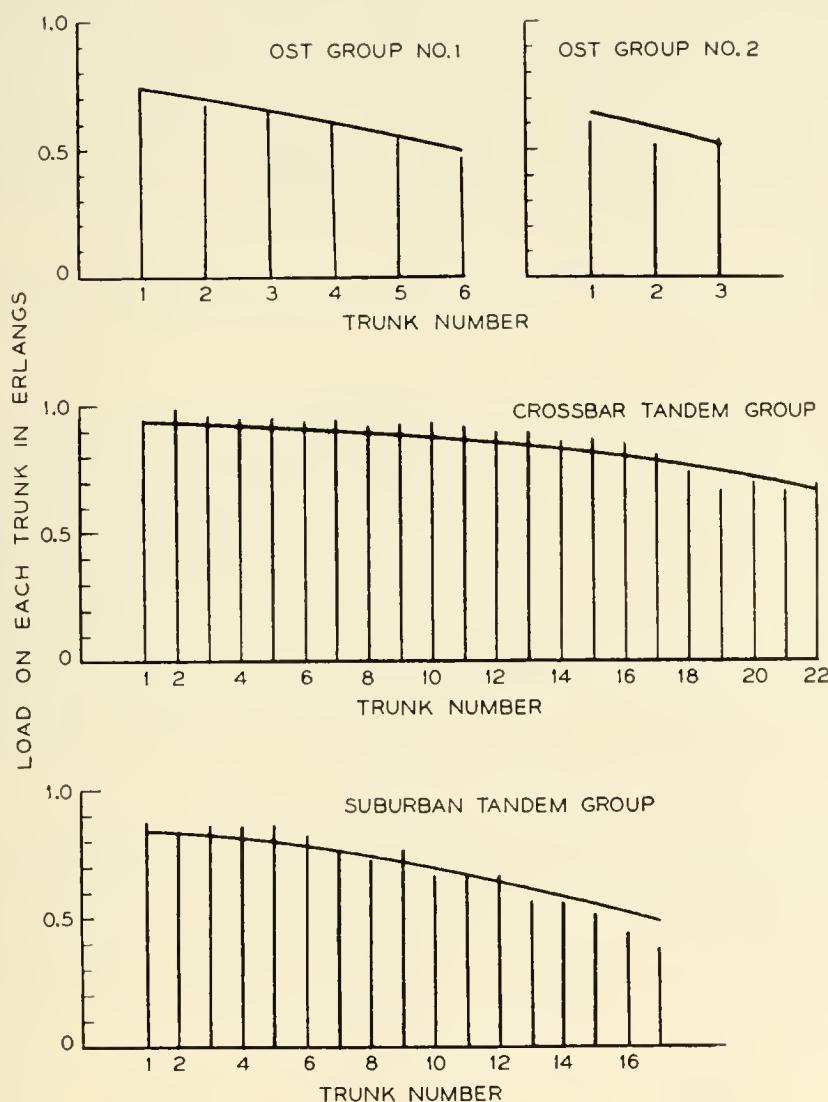


Fig. 41 — Comparison of load carried by each alternate route trunk; theory versus throwdowns.

tables). The respective sums of the overflow α 's and v 's, give A' and V' by (28) and (29); they provide the necessary statistical description of traffic offered to the alternate route.

According to the Equivalent Random method for estimating the alternate route trunks required to provide a specified grade of service to the overflow traffic A' , one next determines a random load A which when submitted to S trunks will yield an overflow with the same character (A' , V') as that derived from the complex system's high usage groups. An alternate route of C trunks beyond these S trunks is then imagined. The erlang overflow α' , with random offer A , to $S + C$ trunks is found from standard E_1 -formula tables or curves (such as Fig. 12).

The ratio $R_2 = \alpha'/A'$ is a first estimate of the grade of service given to each parcel of traffic offered to the alternate route. As discussed in Section 7.5, this service estimate, under certain conditions of load and trunk arrangement, may be significantly pessimistic when applied to a first routed parcel of traffic offered directly to the alternate route. An improved estimate of the overflow probability for such first routed traffic was found to be R_1 as given by (30).

8.1 Determination of Final Group Size with First Routed Traffic Offered Directly to the Final Group

When first routed traffic is offered directly to the final group, its service R_1 will nearly always be poorer than the *overall* service given to those other traffic parcels enjoying high usage groups. The first routed traffic's service will then be controlling in determining the final group size. Since R_1 is a function of S , C and A in the Equivalent Random solution (30), and there is a one-to-one correspondence of pairs of A and S values with A' and V' values, engineering charts can be constructed at selected service levels R_1 which show the final route trunks C required, for any given values of A' and V' . Figs. 42 to 45 show this relation at service levels of $R_1 = 0.01, 0.03, 0.05$ and 0.10 , respectively.*

* On Fig. 42 (and also Figs. 46-49) the low numbered curves assume, at first sight, surprising shapes, indicating that a load with given average and variance would require fewer trunks if the average were *increased*. This arises from the sensitivity of the tails of the distribution of offered calls, to the V'/A' peakedness ratio which, of course, decreases with increases in A' . For example, with $C = 4$ trunks and fixed $V' = 0.52$, the loss rapidly decreases with increasing A' :

A'	V'/A'	A	S	α'	α'/A'
0.28	1.86	6.1	10.	0.0155	0.055
0.33	1.58	3.0	5.0	0.0081	0.025
0.40	1.30	1.42	2.03	0.0036	0.009
0.52	1.00	0.52	0	0.0008	0.002

These four R_1 levels would appear to cover the most used engineering range. For example, if the traffic offered to the final route (including the first routed traffic) has parameters $A' = 12$ and $V' = 20$, reading on Fig. 43 indicates that to give $P = 0.03$ "lost calls cleared" service to the first routed traffic, $C = 19$ final route trunks should be provided. (For random traffic ($V' = A' = 12$), 17.8 trunks would be required.)

Other charts, of course, might be constructed from which R_1 could be read for specific values of A' , V' and C . They would become voluminous, however, if a wide range of all three variables were required.

8.2 Provision of Trunks Individual to First Routed Traffic to Equalize Service

If the difference between the service R_1 given the first routed parcel of traffic and the service given all of the other parcels, is material, it may be desirable to take measures to diminish these inequities. This may readily be accomplished by setting aside a number of the otherwise full access final route trunks, for exclusive and first choice use of the first routed traffic. High usage groups are now provided for all parcels of traffic. The alternate route then services their combined overflow. The overall grade of service given the i th parcel of offered traffic in a single stage alternate route system will then be approximately

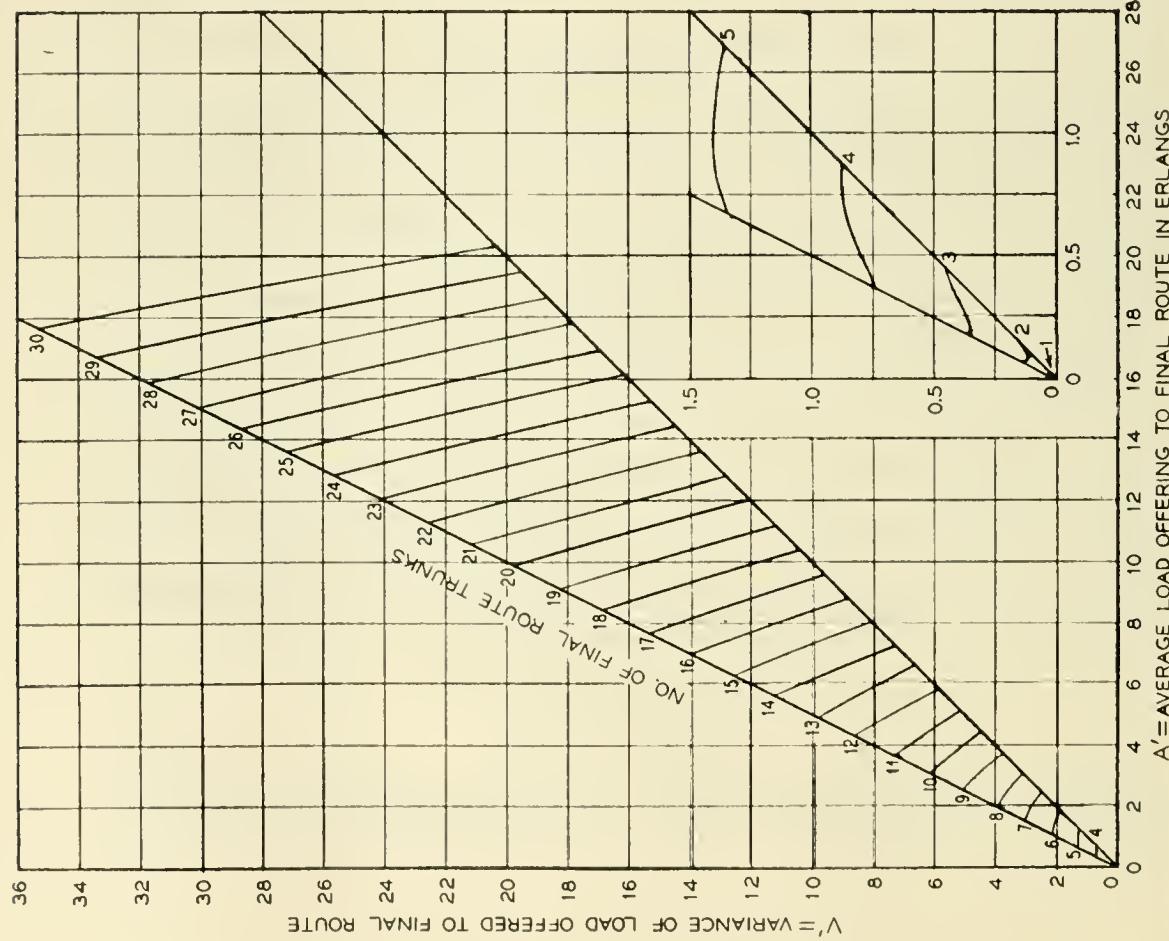
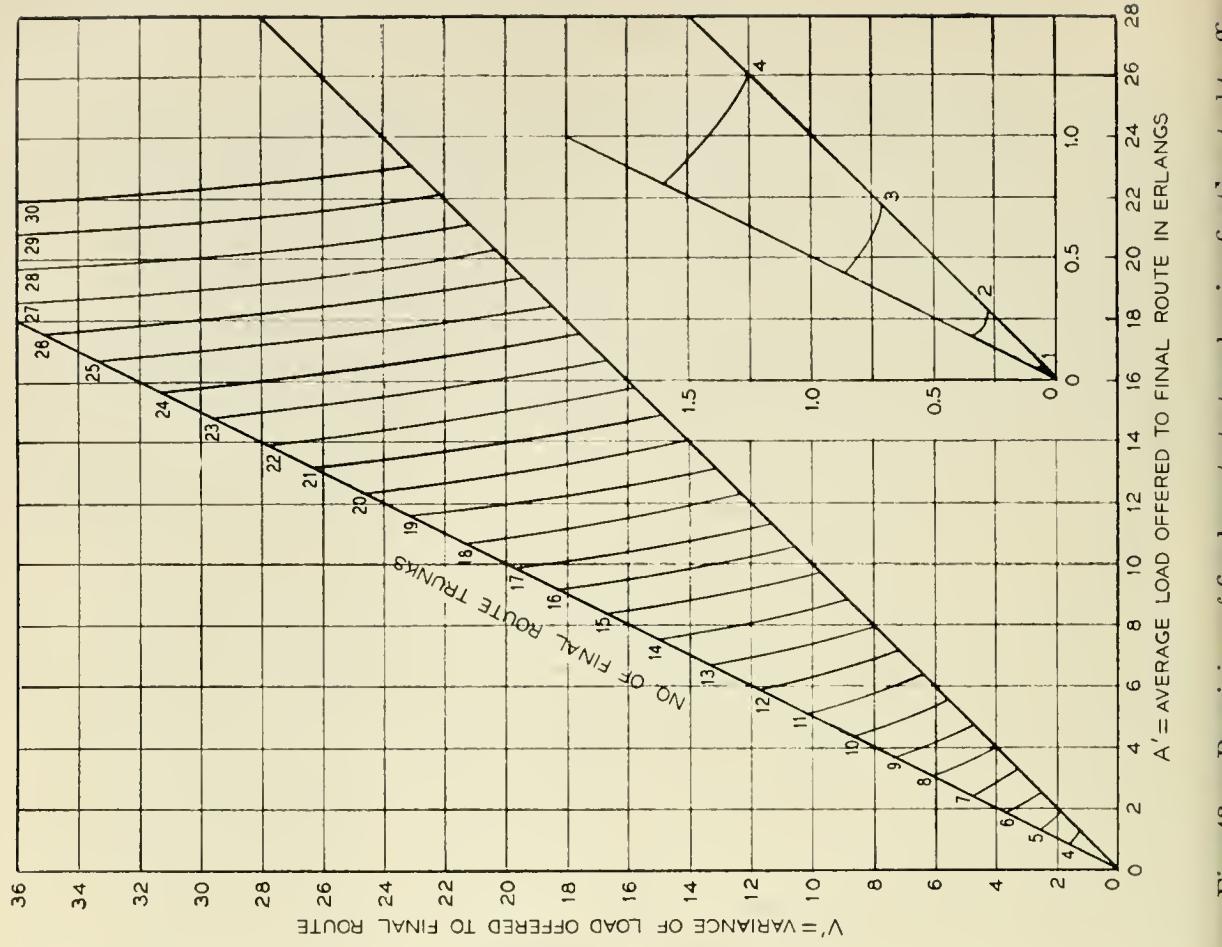
$$P_i \doteq E_{1,x_i}(a_i)R_2 = E_{1,x_i}(a_i) \frac{\alpha'}{A'}^* \quad (34)$$

Thus the service will tend to be uniform among the offered parcels when all send substantially identical proportions of their offered loads to the alternate route. And the natural provision of "individual" trunks for the exclusive use of the first routed traffic would be such that the same proportion should overflow as occurs in the associated high usage groups.

This procedure cannot be followed literally since high usage group size is fixed by economic considerations rather than any predetermined overflow value. The resultant overflow proportions will commonly vary over a considerable range. In this circumstance it would appear reasonable to estimate the objective overflow proportion to be used in establishing the individual group for the first routed traffic, as some weighted average \bar{b} of the overflow proportions of the several high usage groups. Thus with weights g and overflow proportions b ,

$$\bar{b} = \frac{g_1 b_1 + g_2 b_2 + \dots}{g_1 + g_2 + \dots} \quad (35)$$

* Although not exact, this equation can probably be accepted for most engineering purposes where high usage trunks are provided for each parcel of traffic.



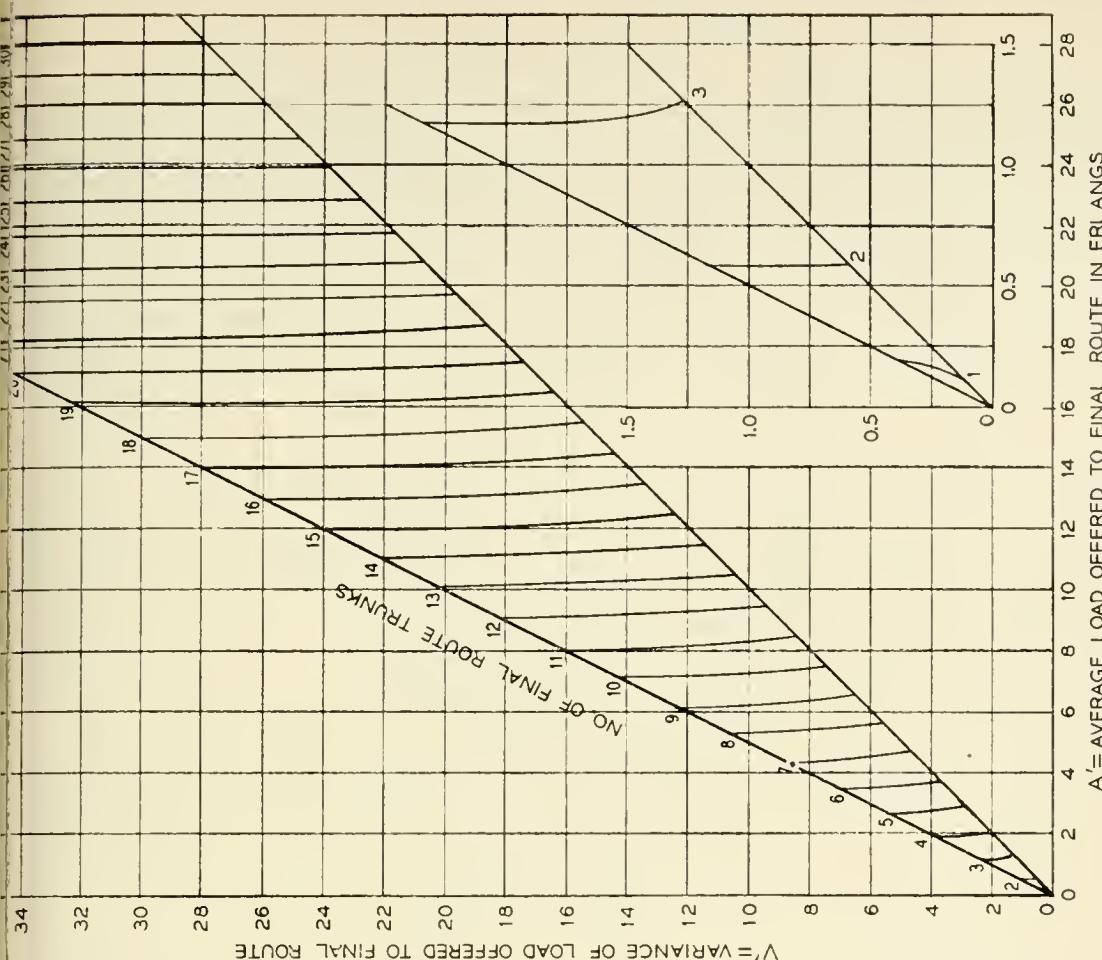


Fig. 44 — Provision of final route trunks to give first routed traffic service of $R_1 = 0.05$

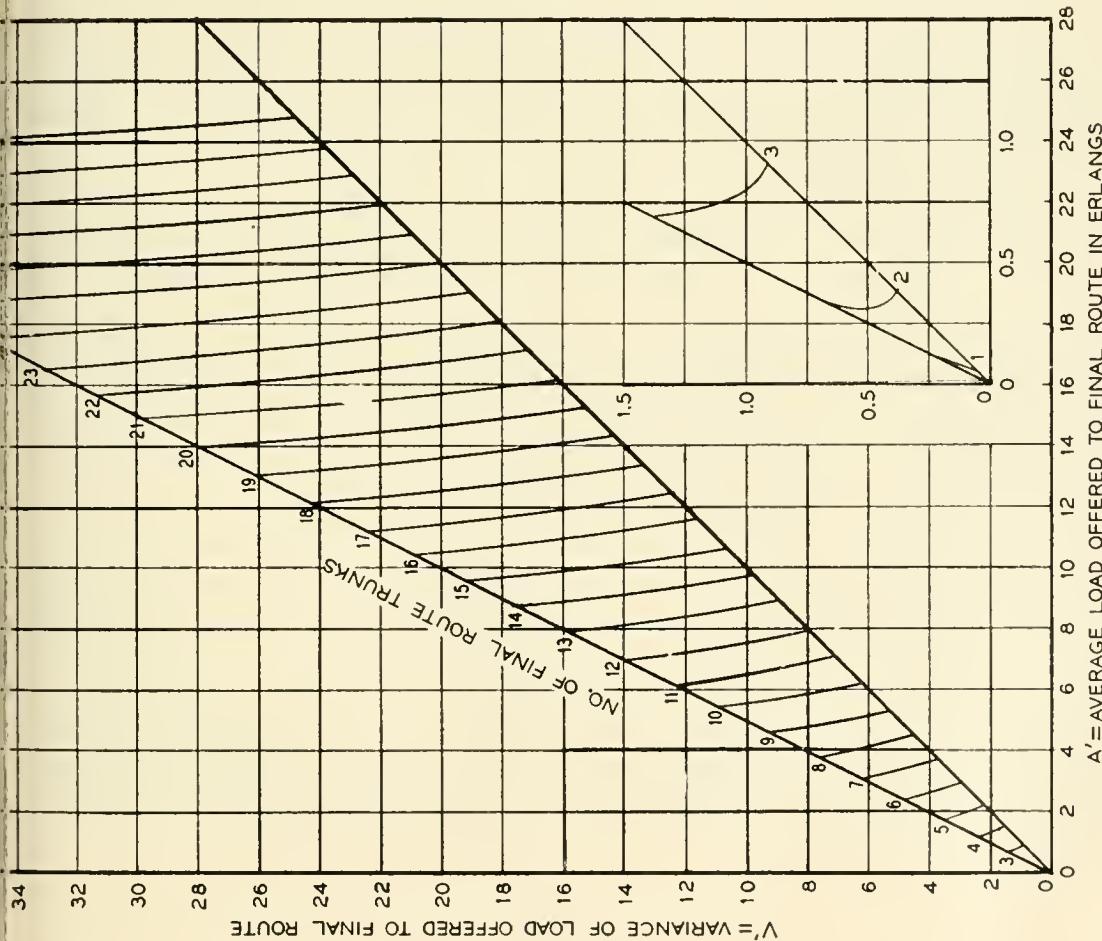


Fig. 45 — Provision of final route trunks to give first routed traffic service of $R_1 = 0.10$.

A choice of all weights g equal to unity will often be satisfactory for the present purpose. The desired high usage group size for the first routed traffic is then found from standard E_1 -tables showing trunks x required, as a function of offered traffic a and proportion overflow \bar{b} .

Since the different parcels of traffic have varying proportions b of their loads overflowing to the final route, by equation (34) the parcel with the largest proportion will determine the permitted value of R_2 . Thus

$$R_2 = P/b_{\max} \quad (36)$$

where P is the specified poorest overall service (say 0.03) for any parcel. It may be noted that on occasion some one parcel, perhaps a small one, may provide an outstandingly large b_{\max} value, which will tend to give a considerably better than required service to all the major traffic parcels. Some compromise with a literal application of a fixed poorest service criterion may be indicated in such cases.

An alternative and somewhat simpler procedure here is to use an average value \bar{b} in (36) instead of b_{\max} , with a compensating modification of P , so that substantially the same R_2 is obtained as before. The allowance in P will be influenced by the choice of weights g in (35). It will commonly be found in practice that overflow proportions to final groups for large parcels of traffic are lower than for small parcels. Choosing all weights, as unity, as opposed to weighting by traffic volumes for example, tends to insert a small element of service protection for those traffic parcels (often the smaller ones) with the higher proportionate high usage group overflows.

Having determined R_2 , a ready means is needed for estimating the required number of final route trunks. Curves for this purpose are provided on Figs. 46 to 49, within whose range, $R_2 = 0.01$ to 0.10, it will usually be sufficiently accurate to interpolate for trunk engineering purposes. These R_2 -curves exactly parallel the R_1 -curves for use when first routed traffic is offered directly to the final group without benefit of individual high usage trunks. If R_2 is well outside the charted range a run-through of the ER calculations may be required.

8.3 Area in Which Significant Savings in Final Route Trunks are Realized by Allowing for the Preferred Service Given a First Routed Traffic Parcel

Considerable effort has been expended by alternate route research workers in various countries to discover and evaluate those areas where first routed (random) traffic offered to a final route enjoys a substantial service advantage over competing parcels of traffic which have over-

flowed from high usage groups. A comparison of Figs. 42 to 45, (which indicate trunk provision for meeting a first routed traffic criterion R_1) with Figs. 46 to 49 (which indicate trunk provision for meeting a composite-load-offered-to-the-final-route criterion R_2) gives a means for deciding under what conditions in practice it is important to distinguish between the two criteria. Fig. 50 shows the borders of areas, defined in terms of A' and V' , the characterizing parameters of the total load offered to the final route, where a 2 and 5 per cent overprovision of final trunks would occur using R_2 for R_1 as the loss measure for first routed traffic. Thus in the alternate route examples displayed in Table XV, where $x = 8$, $g = 2$ to 10, $A' = 4.0$ and V' varies from 5.80 to 5.45, Fig. 50 shows that by failing to allow for the preferred position of the 2 erlang first routed parcel, we should at $R = 0.02$ engineered loss, provide a little over 5 per cent more final trunks than necessary. (Actually 10.2 and 9.9 versus 9.6 and 9.4 trunks for $g = 2$ and 10, respectively.)

The curves of Fig. 50 for final route loads larger than a few erlangs, are almost straight lines. At an objective engineering base of $R = 0.03$, for example, the 2 and 5 per cent trunk overprovision areas through using R_2 instead of R_1 are outlined closely by:

$$\begin{aligned} 2 \text{ per cent overprovision occurs at } V'/(A' - 1) &\doteq 1.4 \\ 5 \text{ per cent overprovision occurs at } V'/(A' - 1) &\doteq 1.8. \end{aligned}$$

Thus in the range of loads covered by Fig. 50, one might conclude that useful and determinable savings in final trunks can be achieved by use of the specialized R_1 -curves instead of the more general R_2 -curves, when the ratio $V'/(A' - 1)$ exceeds some figure in the 1.4 to 1.8 range, say 1.6. (In the examples just cited the $V'/(A' - 1)$ ratio is approximately 1.9.)

8.4. Character of Traffic Carried on Non-Final Routes

Telephone traffic which is carried by a non-final route will ordinarily be subjected to a peak clipping process which will depress the variance of the carried portion below that of the offered load. If this traffic terminates at the distant end of the route, its character, while conceivably affecting the toll and local switching trains in that office, will not require further consideration for intertoll trunk engineering. If, however, some or all of the route's load is to be carried on toll facilities to a more distant point (the common situation), the character of such parcels of traffic will be of interest in providing suitable subsequent paths. For this purpose it will be desirable to have estimates of the mean and variance of these carried parcels.

When a random traffic of "a" erlangs is offered to a group of "c" paths

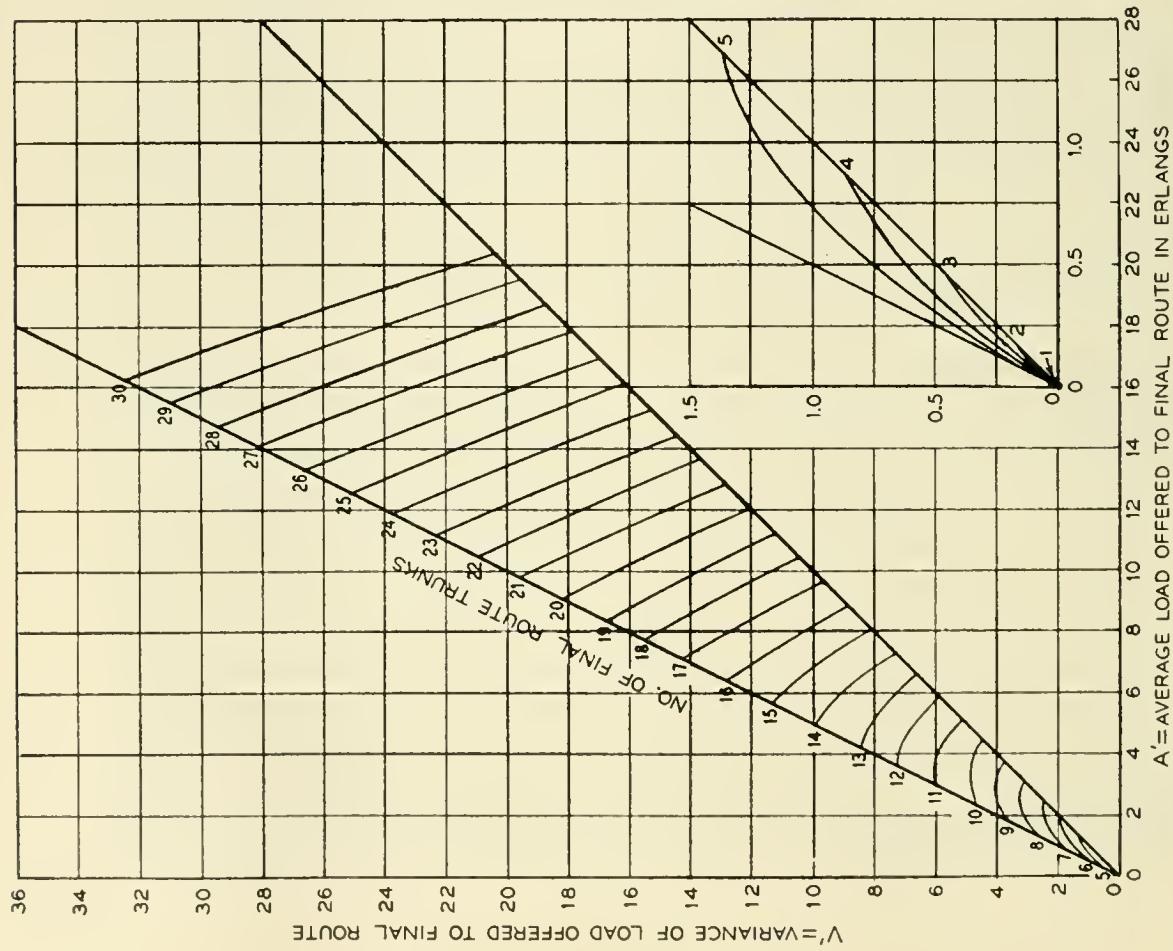
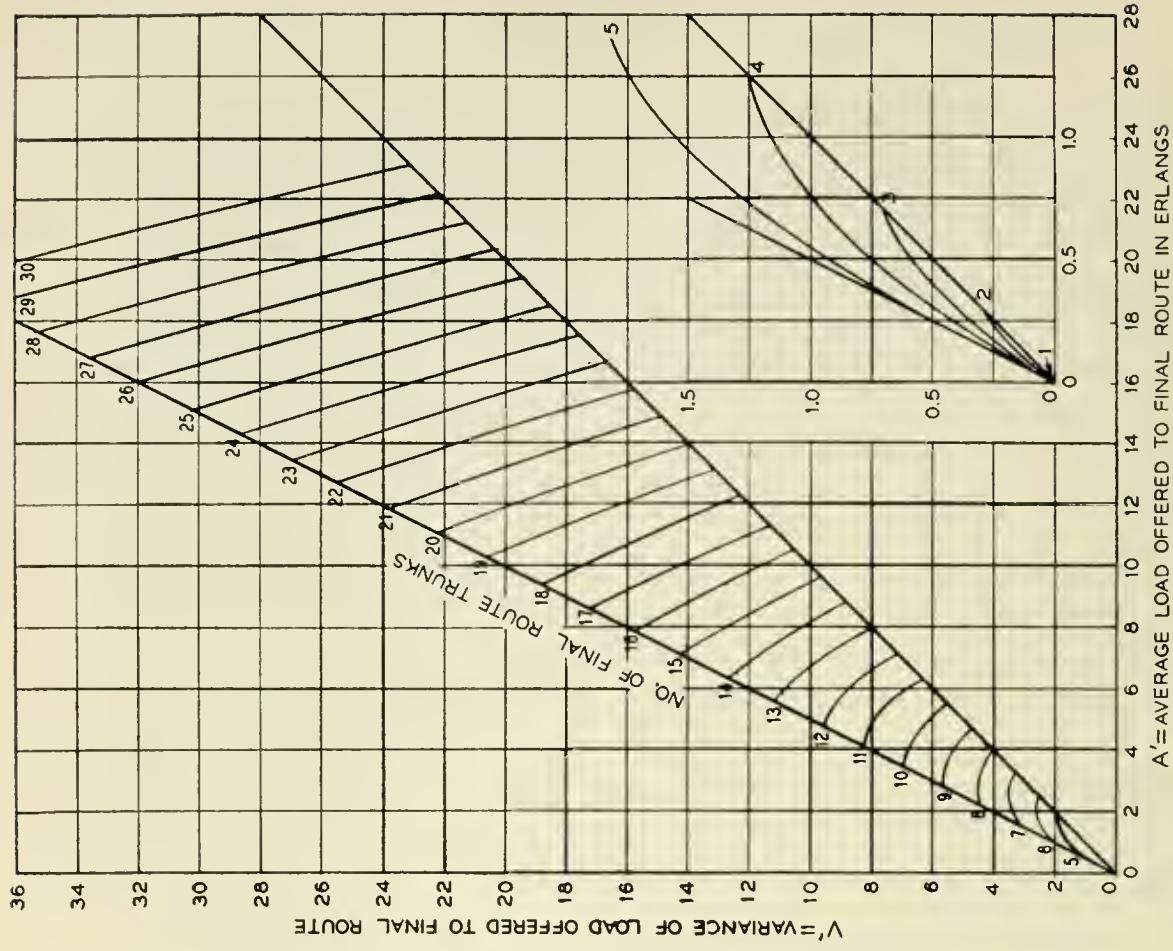


Fig. 47 — Provision of final route trunks to give combined offered load $A' = 1.0$

Fig. 47 — Provision of final route trunks to give combined offered load $A' = 0.5$

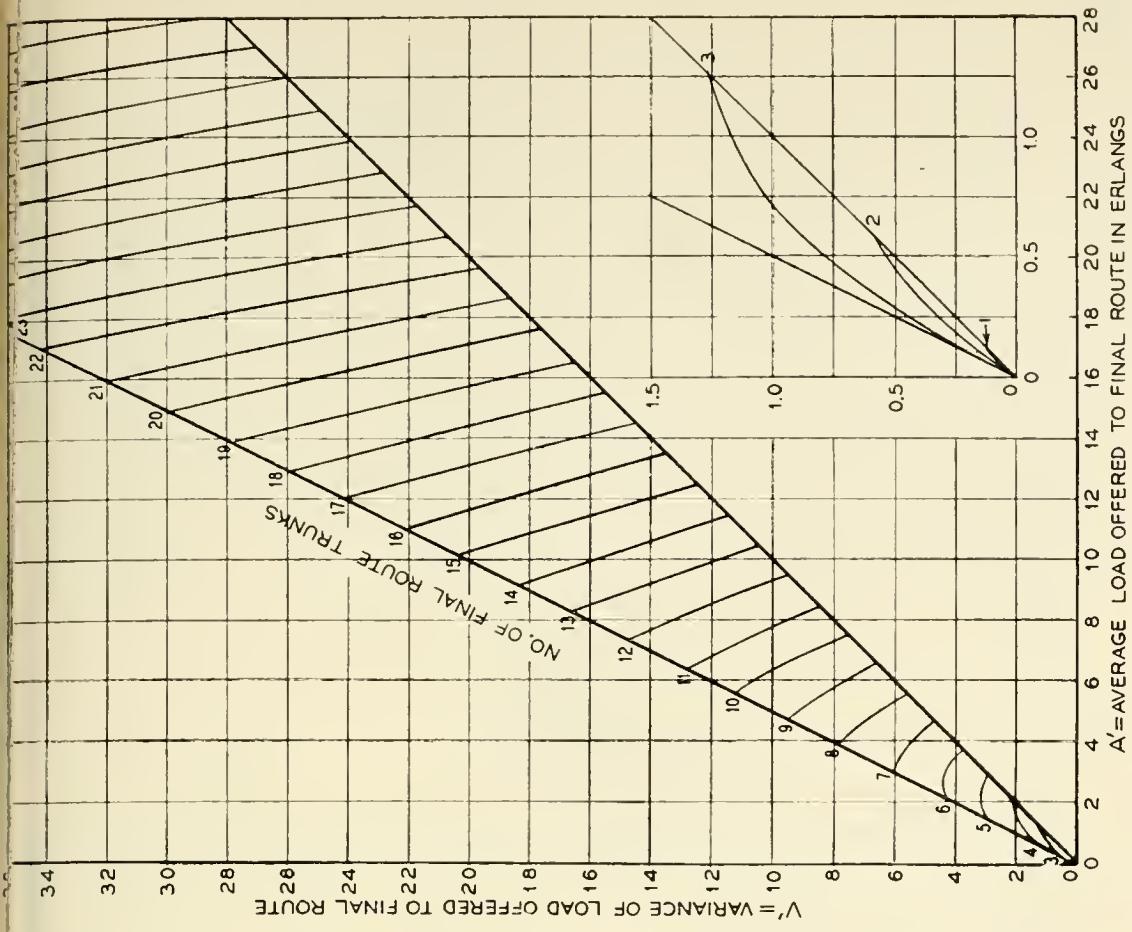


Fig. 48 — Provision of final route trunks to give combined offered load a service of $R_2 = 0.05$.

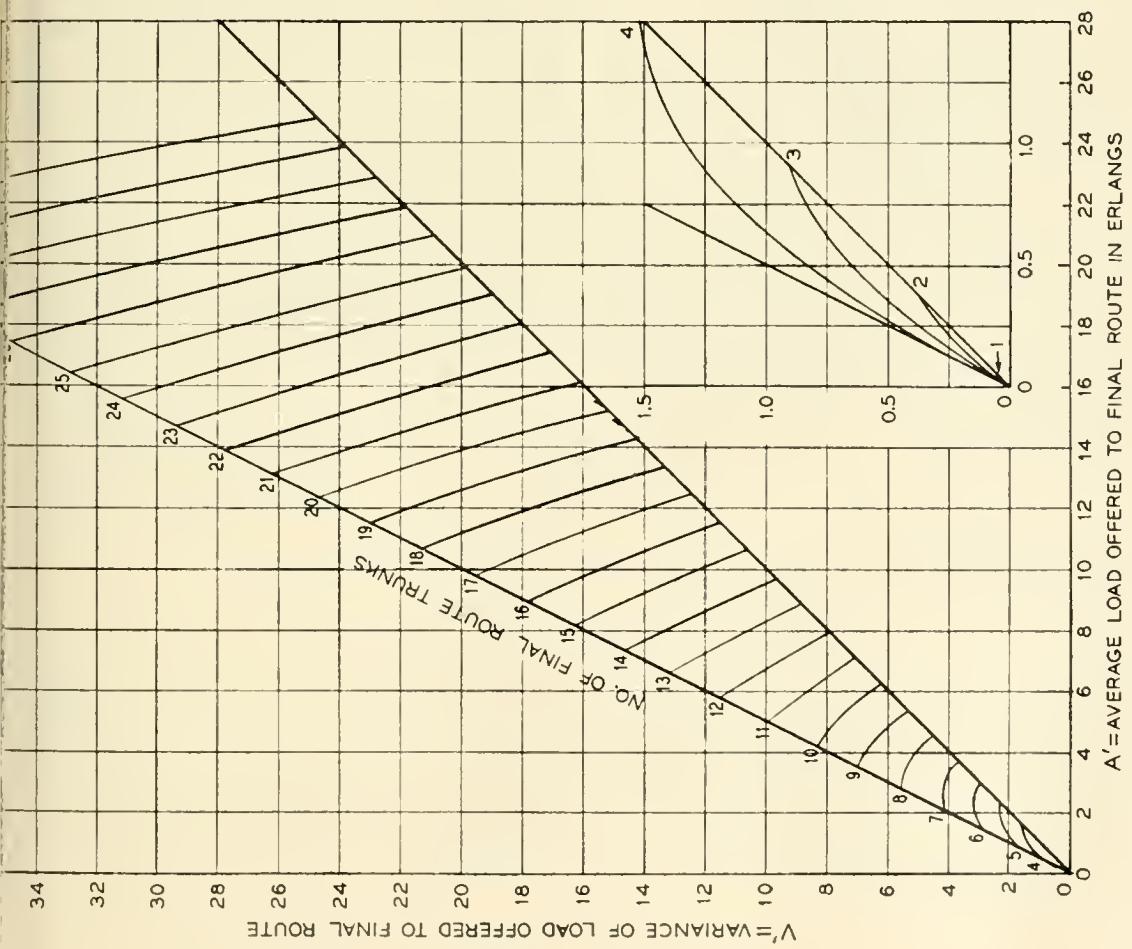


Fig. 49 — Provision of final route trunks to give combined offered load a service of $R_2 = 0.10$.

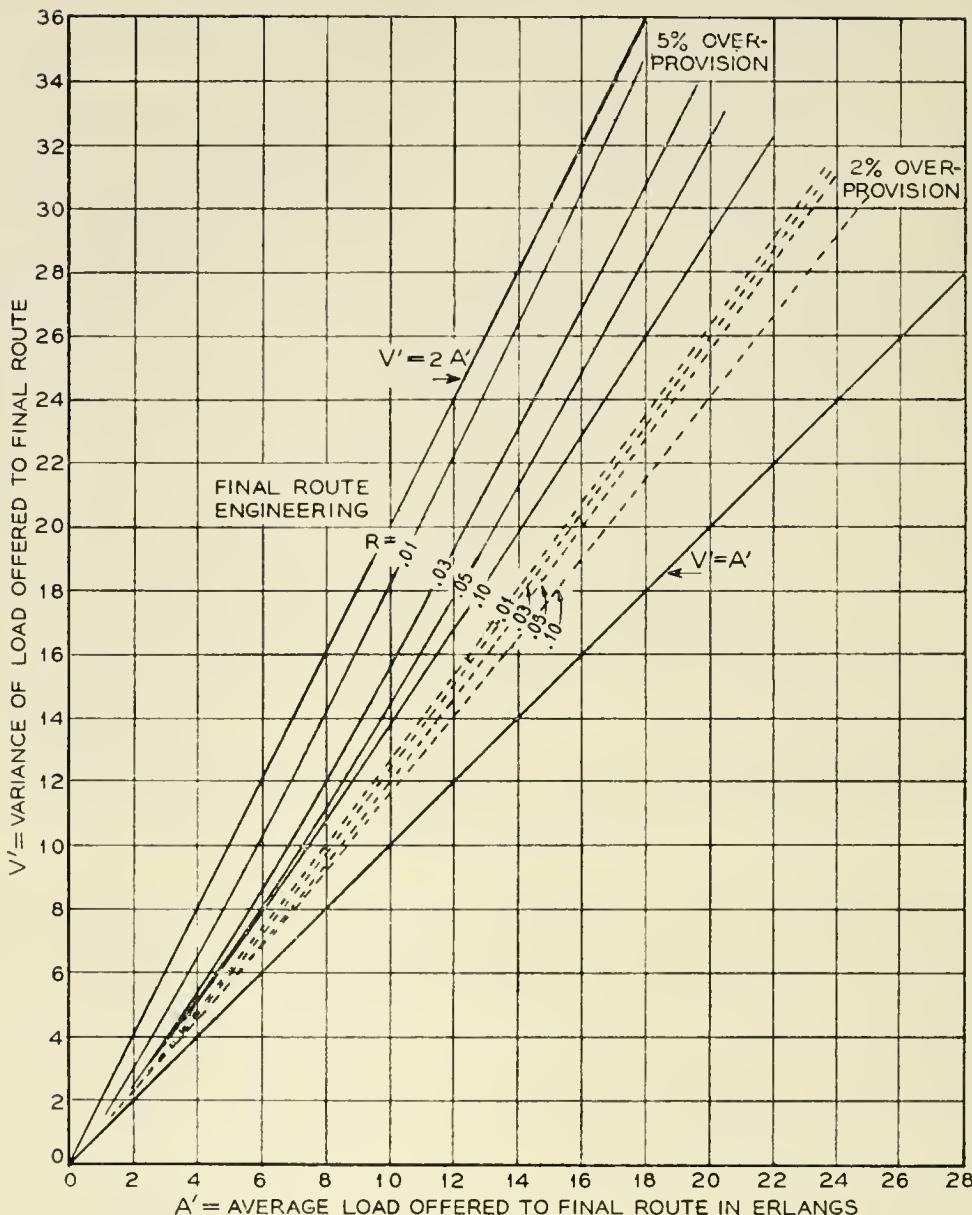


Fig. 50 — Overprovision of final route trunks when R_2 is used instead of R_1 as service to first routed traffic.

and overflowing calls do not return, the variance of the carried load is

$$V_{cd} = a[1 - E_{1,c}(a)] [1 + aE_{1,c}(a) - aE_{1,c-1}(a)]^* \quad (37)$$

and the ratio of variance to average of the carried load is

$$\begin{aligned} \frac{V_{cd}}{L} &= 1 - a [E_{1,c-1}(a) - E_{1,c}(a)]^* \\ &= 1 - \left(\frac{c}{L} - 1 \right) (a - L)^* \\ &= 1 - \ell_c \end{aligned} \quad (38)$$

* These particular forms are due to P. J. Burke.

From (38) it is easy to see that

$$\begin{aligned} V_{cd} &= L(1 - \ell_c) \\ &= (\text{Load carried by the group})(1 - \text{load on last trunk}) \end{aligned} \quad (39)$$

This is a convenient relationship since for high usage trunk study work, both the loads carried (in erlangs) on the group and on the last trunk will ordinarily be at hand.

If the high usage group's load is to be split in various directions at the distant point for re-offer to other groups, it would appear not unreasonable to assign a variance to each portion so as to maintain the ratio expressed in equation (38). That is, if a carried load L is divided into parts $\lambda_1, \lambda_2 \dots$ where $L = \lambda_1 + \lambda_2 \dots$, then the associated variances $\gamma_1, \gamma_2 \dots$ would be

$$\begin{aligned} \gamma_1 &= \lambda_1 (1 - \ell_c) \\ \gamma_2 &= \lambda_2 (1 - \ell_c) \\ &\dots \end{aligned} \quad (40)$$

If, however, the load offered to the group is non-random (e.g., the group is an intermediate route in a multi-alternate route system), the procedure is not quite so simple as in the random case just discussed. Equation (32) expresses the variance V_c of the carried load on a group of C paths whose offered traffic consists of the overflow from a first group of S paths to which a random load of A erlangs has been offered. V_c could of course be expressed in terms of A' , V' and C , and curves or tables constructed for working purposes. However, such are not available, and in any case might be unwieldy for practical use.

A simple alternative procedure can be used which yields a conservative (too large) estimate of carried load variance. With random load offered to a divided two stage multiple of x paths followed by y paths, a positive correlation exists between the numbers m and n of calls present simultaneously on the x and y paths, respectively. Then the variance V_{m+n} of the $m + n$ distribution is greater than the sum of the individual variances of m and n ,

$$V_{m+n} > V_m + V_n$$

or

$$V_m \leq V_{m+n} - V_n \quad (41)$$

Now n can be chosen arbitrarily, and if made very large, V_{m+n} becomes the offered load variance, and V_n the overflow load variance. Both of these are usually (or can be made) available. Their difference then, according to (41) gives an upper limit to V_m , the desired carried load

TABLE XVI—APPROXIMATE DETERMINATION OF THE VARIANCE
OF CARRIED LOADS;
 x lower paths, 8 upper paths; offer to upper paths = 3 erlangs

No. Lower Paths x	Lower Paths, x				Upper Paths, y				True variance of cd load (Brock- meyer)
	Random offered load A ($= V$)	Variance of overflow V_n	Estimated variance of carried load $V - V_n$	True variance of carried load Eq (37)	Variance of offer V' ($= V_n$) (Col 3)	Variance of overflow V''	Estimated variance of cd load $V' - V''$		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		(9)
0	3.00	3.00	0	0	3.00	0.035	2.97		2.853
3	5.399	4.05	1.35	0.60	4.05	0.121	3.93		3.664
6	7.856	4.98	2.88	1.418	4.95	0.236	4.74		4.175
12	12.882	6.22	6.66	3.538	6.22	0.520	5.70		4.790

variance. Corresponding reasoning yields the same conclusion when the offered load before the x paths is non-random.

A numerical example by Broekmeyer¹¹ while clearly insufficient to establish the degree of the inequality (41), indicates something as to the discrepancy introduced by this approximate procedure. Comparison with the true values is shown in Table XVI.

In the case of random offer to the 0, 3, 6, 12 "lower paths," the approximate method of equation (41) overestimates the variance of the carried load by nearly two to one (columns 4 and 5 of Table XVI). The exact procedure of (37) is then clearly desirable when it is applicable, that is when random traffic is being offered. For the 8 upper paths to which non-random load is offered (the non-randomness is suggested by comparing the variance of column 6 in Table XVI with the average offered load of 3 erlangs), the approximate formula (41) gives a not too extravagant overestimate of the true carried load variance. Until curves or tables are computed from equation (32), it would appear useful to follow the above procedure for estimating the carried load variance when non-random load is offered.

8.5. Solution of a Typical Toll Multi-Alternate Route Trunking Arrangement: Bloomsburg, Pa.

In Fig. 9 a typical, moderately complex, toll alternate route layout was illustrated. It is centered on the toll office at Bloomsburg, Pa. The loads to be carried between Bloomsburg and the ten surrounding cities are indicated in CCS (hundred call seconds per hour of traffic; 36 CCS = 1 erlang). The numbers of direct high usage trunks shown are assumed to have been determined by an economic study; we are asked to find

the number of trunks which should be installed on the Bloomsburg-Harrisburg route, so that the last trunk will carry approximately 18 CCS (0.50 erlang). Following this determination, (a) the number of final trunks from Bloomsburg to Scranton is desired so that the poorest service given to any of the original parcels of traffic will be no more than 3 calls in 100 meeting NC . Also (b) the modified Bloomsburg-Scranton trunk arrangement is to be determined when a high usage group is provided for the first routed traffic.

Solution (a): First Routed Traffic Offered Directly to Final Group

The offered loads in CCS to each distant point are shown in column (2) of Table XVII; the corresponding erlang values are in column (3). Consulting Figs. 12 and 13, the direct group overflow load parameters, average and variance, are read and entered in columns (5) and (6) respectively for the four groups overflowing to Harrisburg, and in columns (7) and (8) for the four groups directly overflowing to Scranton. The variance for the direct Bloomsburg-Harrisburg traffic equals its average; likewise for the direct Bloomsburg-Scranton traffic. They are so entered in the table. The parameters of the total load on the Harrisburg group are found by totalling, giving $A' = 11.19$, and $V' = 19.90$.

The required size C_1 of the Harrisburg group is now determined by the Equivalent Random theory. Entering Fig. 25 with A' and V' just determined, the ER values of trunks and load found are $S_1 = 13.55$, and $A_1 = 23.75$. C_1 is to be selected so that on a straight group of $S_1 + C_1$ trunks with offered load A , the last trunk will carry 0.50 erlang. Reading from Fig. 40, the load carried by the 26th trunk approximates this figure. Hence $C_1 = 26 - S_1 = 12.45$ trunks; or choose 12 trunks.

The overflow load's mean and variance from the Harrisburg group with 12 trunks, is now read from Figs. 12 and 13, entering with load $A_1 = 23.75$ and $C_1 + S_1 = 25.55$ trunks. The overflow values ($\alpha' = 2.50$ and $v' = 7.50$) are entered in columns (7) and (8) of the table. The total offered load to Scranton is now obtained by totalling columns (7) and (8), giving $A'' = 16.27$ and $V'' = 25.60$.

We desire now to know the number of trunks C_2 for the Scranton group which will provide NC 3 per cent of the time to the poorest service parcel of traffic, i.e., the first routed Bloomsburg-Scranton parcel. The $R_1 = 0.03$ and $R_2 = 0.03$ solutions are available, the former of course being more closely applicable. A check reference to Fig. 50 shows a difference of approximately 4 per cent in trunk provision would result from the two methods. Entering Figs. 43 and 47 with $A'' = 16.27$ and

TABLE XVII — ILLUSTRATIVE CALCULATION OF ALTERNATE ROUTE TRUNKS AT BLOOMSBURG, P.A.

Distant Office	Load Between Bloomsburg and Distant Office		Characteristics of Load to Harrisburg Group		Characteristics of Load to Scranton Group		Approximate Proportion of Original Offer Going To Final Route
	CCS	Erlangs	No. Trunks Between Bloomsburg and Distant Office <i>x</i>	α (Fig. 12)	β (Fig. 13)	(7)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pottsville	26	0.72	1	0.30	0.35		$\text{Col. } \frac{5}{3} \times \frac{2.50}{11.19}$
Shamokin	540	15.0	15	2.70	6.35		0.093
Sunbury	691	19.19	20	2.66	7.00		0.040
Williamsport	160	4.44	5	1.06	1.73		0.031
Harrisburg	161	4.47	C_1	$\frac{4.47}{A' = 11.19}$	$\frac{4.47}{V' = 19.90}$		0.053
							0.223
Frackville	123	3.42	5			0.50	0.81
Hazleton	836	23.22	28			1.28	0.146
Wilkes-Barre	228	6.33	8			0.89	3.90
Philadelphia	154	4.28	5			0.96	0.055
Scranton	365	10.14	C_2			1.57	0.141
							0.224
							$\bar{b} = 0.112$
						$A'' = 16.27$	$V'' = 25.60$ = unweighted average

$A_1 = 23.75$ $C_1 = 12$ $S_1 = 13.55$
Final route trunks required = 24.1. (Read from Fig. 43 using $A''' = 16.57$, $V''' = 25.90$ for 3 per cent first routed retrials.)

$V'' = 25.60$, we obtain the trunk requirements:

$$\begin{aligned} R_1 \text{ Method} &\dots\dots\dots 23.8 \text{ trunks} \\ R_2 \text{ Method} &\dots\dots\dots 24.8 \text{ trunks} \end{aligned}$$

Thus the more precise method of solution here yields a reduction of 1.0 in 25 trunks, a saving of 4 per cent, as had been predicted.

The above calculation is on a Lost Calls Cleared basis. Since the overflow direct traffic calls will return to this group to obtain service, to assure their receiving no more than 3 per cent NC , the provision of the final route would theoretically need to be slightly more liberal. An estimate of the allowance required here may be made by adding the expected erlangs loss Δ for the direct traffic (most of the final route overflow calls which come from high usage routes will be carried by their respective groups on the next retrial) to both the A'' and V'' values previously obtained, and recalculating the trunks required from that point onward. (In fact this could have been included in the initial computation.) Thus:

$$\begin{aligned} \Delta &= 0.03 \times 10.14 = 0.30 \text{ erlang} \\ A''' &= 16.27 + 0.30 = 16.57 \text{ erlangs} \\ V''' &= 25.60 + 0.30 = 25.90 \text{ erlangs} \end{aligned}$$

Again consulting Figs. 43 and 47 gives the corresponding final trunk values

$$\begin{aligned} R_1 \text{ Method} &\dots\dots\dots 24.1 \text{ trunks} \\ R_2 \text{ Method} &\dots\dots\dots 25.1 \text{ trunks} \end{aligned}$$

Of the above four figures for the number of trunks in the Scranton route, the R_1 -Method with retrials, i.e., 24.1 trunks, would appear to give the best estimate of the required trunks to give 0.03 service to the poorest service parcel.

Solution (b): With High Usage Group Provided for First Routed Traffic

Following the procedure outlined in Section 8.2, we obtain an average of the proportions overflowing to the final route for all offered load parcels. The individual parcel overflow proportion estimates are shown in the last column of Table XVII; their unweighted average is 0.112. With a first routed offer to Scranton of 10.14 erlangs, a provision of 12 high usage trunks will result in an overflow of $\alpha = 1.26$ erlangs, or a proportion of 0.125 which is the value most closely attainable to the objective 0.112. With 12 trunks the overflow variance is found to be 2.80.

Replacing 10.14 in columns 7 and 8 of Table XVII with 1.26 and 2.80, respectively, gives new estimates characterizing the offer to the final route, $A'' = 7.39$ and $V'' = 18.26$. We now proceed to insure that the poorest service parcel obtains 0.03 service. This occurs on the Philadelphia and Harrisburg groups, which overflow to the final group approximately 0.224 of their original offered loads. The final group must then, according to equation (34) be engineered for

$$R_2 = 0.03/0.224 = 0.134 \text{ service.}$$

This value lies above the highest R_2 engineering chart (Fig. 49, $R_2 = 0.10$), so an ER calculation is indicated.

The Equivalent Random average is 28.6 erlangs, and $S = 23.5$ trunks. We determine the total trunks $S + R$ which, with 28.6 erlangs offered, will overflow $0.134(7.39) = 0.99$ erlang. From Fig. 12.2, 35.6 trunks are required. Then the final route provision should be $C = 35.6 - 23.5 = 12.1$ trunks; and a total of $12 + 12.1$ or 24.1 Scranton trunks is indicated.

Simplified Alternative Solution: In Section 8.2 a simplified approximate procedure was described using a modified probability P' for the average overall service for all parcels of traffic, instead of P for the poorest service parcel. Suppose $P' = 0.01$ is chosen as being acceptable. Then

$$R_2 = \frac{P'}{\bar{b}} = \frac{0.01}{0.112} = 0.089$$

Interpolating between the $R_2 = 0.05$ and 0.10 curves (Figs. 48 and 49) gives with $A'' = 7.39$ and $V'' = 18.26$, $C = 13.4$, the number of final trunks required. Again the same result could have been obtained by making the suitable ER computation. It may be noted that if P' had been chosen as 0.015 (one-half of P), R_2 would have become 0.134, exactly the same value found in the poorest-service-parcel method. The final trunk provision, of course, would have again been 12.1 trunks.

Discussion

In the first solution above, 24.1 full access final trunks from Bloomsburg to Scranton were required. The service on the first routed traffic was 0.03; however, the service enjoyed by the offered traffic as a whole was markedly better than 0.03. The corresponding ER calculation shows ($A = 28.3$, $S + C = 12.3 + 24.1$) a total overflow of $\alpha'' = 0.72$ erlangs, or an overall service of $0.72/91.21 = 0.008$.

In the second solution, 12 high usage and 12.1 common final, or a total of 24.1, trunks were again required, to give 0.03 service to the poorest service parcels of offered load. The overall service here, however, was $0.99/91.21 = 0.011$. Thus, with the same number of paths provided, in the second solution (high usage arrangement) the overall call loss was 40 per cent larger than in the first solution.* However, it may well be desirable to accept such an average service penalty since by providing high usage trunks for the first routed traffic, the latter's service cannot be degraded nearly so readily should heavy overloads occur momentarily in the other parcels of traffic.

9. CONCLUSION

As direct distance dialing increases, it will be necessary to provide intertoll paths so that substantially no-delay service is given at all times. To do this economically, automatic multi-alternate routing will replace the present single route operation. Traffic engineering of these complicated trunking arrangements will be more difficult than with simple intertoll groups.

One of the new problems is to describe adequately the non-random character of overflow traffic. In the present paper this is proposed to be done by employing both mean and variance values to describe each parcel of traffic, instead of only the mean as used heretofore. Numerous comparisons are made with simulation results which indicate that adequate predictive reliability is obtained by this method for most traffic engineering and administrative purposes. Working curves are provided by which trunking arrangements of considerable complexity can readily be solved.

A second problem requiring further review is the day-to-day variation among the primary loads and their effect on the alternate route system's grade of service. A thorough study of these variations will permit a re-evaluation of the service criteria which have tentatively been adopted. A closely allied problem is that of providing the necessary kind and amounts of traffic measuring devices at suitable points in the toll alternate route systems. Requisite to the solution of both of these problems is an understanding of traffic flow character in a complex overflow-type

* The actual loss difference may be slightly greater than estimated here since in the first solution (complete access final trunks), an allowance was included for return attempts to the final route by first routed calls meeting an 0.03 loss, while in the second solution (high usage group for first routed traffic) no return attempts to the final route were considered. These would presumably be small since only 1 per cent of all calls would overflow and most of these upon retrial would be handled on their respective high usage groups.

of trunking plan, and a method for estimating quantitatively the essential fluctuation parameters at each point in such a system. The present paper has undertaken to shed some light on the former, and to provide an approximate yet sufficiently accurate method by which the latter can be accomplished. It may be expected then that these studies, as they are developed, will provide the basis for assuring an adequate direct distance dialing service at all times with a minimum investment in intertoll trunk facilities.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the technical and mathematical assistance of his associates, Mrs. Sallie P. Mead, P. J. Burke, W. J. Hall, and W. S. Hayward, in the preparation of this paper. Dr. Hall provided the material on the convolution of negative binomials leading to Fig. 19. Mr. Hayward extended Kosten's curve E on Fig. 5 to higher losses by a calculating method involving the progressive squaring of a probability matrix. The author's thanks are also due J. Riordan who has summarized some of the earlier mathematical work of H. Nyquist and E. C. Molina, as well as his own, in the study of overflow load characteristics; this appears as Appendix I.

The extensive calculations and chart constructions are principally the work of Miss C. A. Lennon.

REFERENCES

1. Rappleye, S. C., A Study of the Delays Encountered by Toll Operators in Obtaining an Idle Trunk, B.S.T.J., **25**, p. 539, Oct., 1946.
2. Kosten, L., Over de Invloed van Herhaalde Oproepen in de Theorie der Blokkeringskausen, De Ingenieur, **59**, p. E123, Nov. 21, 1947.
3. Clos, C., An Aspect of the Dialing Behavior of Subscribers and Its Effect on the Trunk Plant, B.S.T.J., **27**, p. 424, July, 1948.
4. Kosten, L., Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen, E.N.T., **14**, p. 5, Jan., 1937.
5. Kosten, L., Over Blokkeerings-en Wachtproblemen, Thesis, Delft, 1942.
6. Molina, E. C., Appendix to: Intereconnection of Telephone Systems — Graded Multiples (R. I. Wilkinson), B.S.T.J., **10**, p. 531, Oct., 1931.
7. Vaulot, A. E., Application du Calcul des Probabilités à l'Exploitation Téléphonique, Revue Gen. de l'Electricité, **16**, p. 411, Sept. 13, 1924.
8. Lundquist, K., General Theory for Telephone Traffic, Ericsson Technics, **9**, p. 111, 1953.
9. Berkeley, G. S., Traffic and Trunking Principles in Automatic Telephony, 2nd revised edition, 1949, Ernest Benn, Ltd., London, Chapter V.
10. Palm, C., Calcul Exact de la Perte dans les Groupes de Circuits Échelonnés, Ericsson Technics, **3**, p. 41, 1936.
11. Brockmeyer, E., The Simple Overflow Problem in the Theory of Telephone Traffic, Teleteknik, **5**, p. 361, December, 1954.

ABRIDGED BIBLIOGRAPHY OF ARTICLES ON TOLL ALTERNATE ROUTING

- Clark, A. B., and Osborne, H. S., Automatic Switching for Nationwide Telephone Service, A.I.E.E., Trans., **71**, Part I, p. 245, 1952. (Also B.S.T.J., **31**, p. 823, Sept., 1952.)
- Pilliard, J. J., Fundamental Plans for Toll Telephone Plant, A.I.E.E. Trans., **71**, Part I, p. 248, 1952. (Also B.S.T.J., **31**, p. 832, Sept., 1952.)
- Nunn, W. H., Nationwide Numbering Plan, A.I.E.E. Trans., **71**, Part I, p. 257, 1952. (Also B.S.T.J., **31**, p. 851, Sept., 1952.)
- Clark, A. B., The Development of Telephony in the United States, A.I.E.E. Trans., **71**, Part I, p. 348, 1952.
- Shipley, F. F., Automatic Toll Switching Systems, A.I.E.E. Trans., **71**, Part I, p. 261, 1952. (Also B.S.T.J., **31**, p. 860, Sept., 1952.)
- Myers, O., The 4A Crossbar Toll System for Nationwide Dialing, Bell Lab. Record, **31**, p. 369, Oct., 1953.
- Clos, C., Automatic Alternate Routing of Telephone Traffic, Bell Lab. Record, **32**, p. 51, Feb., 1954.
- Truitt, C. J., Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks, B.S.T.J., **33**, p. 277, March, 1954.
- Molnar, I., Some Recent Advances in the Economy of Routing Calls in Nationwide Dialing, A.E. Tech. Jl., **4**, p. 1, Dec., 1954.
- Jacobitti, E., Automatic Alternate Routing in the 4A Crossbar System, Bell Lab. Record, **33**, p. 141, April, 1955.

APPENDIX I*

DERIVATION OF MOMENTS OF OVERFLOW TRAFFIC

This appendix gives a derivation of certain factorial moments of the equilibrium probabilities of congestion in a divided full-access multiple used as a basis for the calculations in the text. These moments were derived independently in unpublished memoranda (1941) by E. C. Molina (the first four) and by H. Nyquist; curiously, the method of derivation here, which uses factorial moment generating functions, employs auxiliary relations from both Molina and Nyquist. Although these factorial moments may be obtained at a glance from the probability expressions given by Kosten in 1937, if it is remembered that

$$p(x) = \sum_{k=0}^{\infty} (-1)^{k-x} \binom{k}{x} \frac{M_{(k)}}{k!}, \quad (1.1)$$

where $p(x)$ is a discrete probability and $M_{(k)}$ is the k th factorial moment of its distribution, Kosten does not so identify the moments and it may be interesting to have a direct derivation.

Starting from the equilibrium formulas of the text for $f(m, n)$, the probability of m trunks busy in the specific group of x trunks, and n in

* Prepared by J. Riordan.

the (unlimited) common group, namely

$$\begin{aligned} (a + m + n)f(m, n) - (m + 1)f(m + 1, n) \\ - (n + 1)f(m, n + 1) - af(m - 1, n) &= 0 \\ (a + x + n)f(x, n) - af(x, n - 1) \\ - (n + 1)f(x, n + 1) - af(x - 1, n) &= 0 \end{aligned} \quad (1.2)$$

and

$$f(m, n) = 0, \quad m < 0 \quad \text{or} \quad n < 0 \quad \text{or} \quad m > x,$$

factorial moment generating function recurrences may be found and solved.

With m fixed, factorial moments of n are defined by

$$M_{(k)}(m) = \sum_{n=0}^{\infty} (n)_k f(m, n) \quad (1.3)$$

or alternatively by the factorial moment exponential generating function

$$M(m, t) = \sum_{k=0}^{\infty} M_{(k)}(m) t^k / k! = \sum_{n=0}^{\infty} (1+t)^n f(m, n) \quad (1.4)$$

In (1.3), $(n)_k = n(n - 1) \cdots (n - k + 1)$ is the usual notation for a falling factorial.

Using (1.4) in equations (1.2), and for brevity $D = d/dt$, it is found that

$$\begin{aligned} a + m + tD M(m, t) - (m + 1)M(m + 1, t) \\ - aM(m - 1, t) &= 0 \quad (1.5) \\ (x - at + tD)M(x, t) - aM(x - 1, t) &= 0 \end{aligned}$$

which correspond (by equating powers of t) to the factorial moment recurrences

$$\begin{aligned} (a + m + k)M_{(k)}(m) - (m + 1)M_{(k)}(m + 1) \\ - aM_{(k)}(m - 1) &= 0 \quad (1.6) \end{aligned}$$

$$(x + k)M_{(k)}(x) - akM_{(k-1)}(x) - aM_{(k)}(x - 1) = 0$$

Notice that the first of (1.6) is a recurrence in m , which suggests (following Molina) introducing a new generating function defined by

$$G_k(u) = \sum M_{(k)}(m) u^m \quad (1.7)$$

Using this in (1.5), it is found that

$$\left[(a + k - au + (u - 1) \frac{d}{du}) \right] G_k(u) = 0 \quad (1.8)$$

Hence

$$\frac{1}{G_k(u)} \frac{dG_k(u)}{du} = a + \frac{k}{1-u} \quad (1.9)$$

and, by easy integrations,

$$G_k(u) = ce^{au} (1-u)^{-k}, \quad (1.10)$$

with c an arbitrary constant, which is clearly identical with $G_k(0) = M_{(k)}(0)$.

Expansion of the right-hand side of (1.10) shows that

$$M_{(k)}(m) = M_{(k)}(0) \sum_{j=0}^m \binom{k+j-1}{j} \frac{a^{m-j}}{(m-j)!} = M_{(k)}(0) \sigma_k(m), \quad (1.11)$$

if

$$\sigma_0(m) = a^m/m! \quad \text{and,} \quad \sigma_k(m) = \sum_{j=0}^m \binom{k+j-1}{j} \frac{a^{m-j}}{(m-j)!} \quad (1.12)$$

The notation $\sigma_k(m)$ is copied from Nyquist; the functions are closely related to the $\varphi_x^{(n)}$ used by Kosten; indeed $\sigma_k(m) = e^a \varphi_m^{(k)}$. They have the generating function

$$g_k(u) = \sum_{m=0}^{\infty} \sigma_k(m) u^m = e^{au} (1-u)^{-k} \quad (1.13)$$

from which a number of recurrences are found readily. Thus

$$\begin{aligned} g_k(u) &= (1-u)g_{k+1}(u) \\ u \frac{dg_k(u)}{du} &= aug_k(u) + kug_{k+1}(u) \\ &= -ag_{k-1}(u) + (a-k)g_k(u) + kg_{k+1}(u) \end{aligned}$$

(the last by use of the first) imply

$$\begin{aligned} \sigma_k(m) &= \sigma_{k+1}(m) - \sigma_{k+1}(m-1) \\ m\sigma_k(m) &= a\sigma_k(m-1) + k\sigma_{k+1}(m-1) \\ &= -a\sigma_{k-1}(m) + (a-k)\sigma_k(m) + k\sigma_{k+1}(m) \end{aligned}$$

The first of these leads to

$$\sigma_k(0) + \sigma_k(1) + \cdots + \sigma_k(x) = \sigma_{k+1}(x) \quad (1.14)$$

and the last is useful in the form

$$k\sigma_{k+1}(m) = (m + k - a)\sigma_k(m) + a\sigma_{k-1}(m) \quad (1.15)$$

Also, the first along with $\sigma_0(m) = a^m/m!$ leads to a simple calculation procedure, as Kosten has noticed.

By (1.11) the factorial moments are now completely determined except for $M_{(k)}(0)$. To determine the latter, the second of (1.6) and the normalizing equation

$$\sum_{m=0}^x M_0(m) = 1 \quad (1.16)$$

are available.

Thus from the second of (1.6)

$$[(x + k)\sigma_k(x) - a\sigma_k(x - 1)]M_{(k)}(0) = ak\sigma_{k-1}(x)M_{(k-1)}(0) \quad (1.17)$$

Also

$$\begin{aligned} (x + k)\sigma_k(x) - a\sigma_k(x - 1) \\ &= (x + k - a)\sigma_k(x) + a[\sigma_k(x) - \sigma_k(x - 1)] \\ &= (x + k - a)\sigma_k(x) + a\sigma_{k-1}(x) \\ &= k\sigma_{k+1}(x), \end{aligned}$$

the last step by (1.15). Hence

$$M_{(k)}(0) = a \frac{\sigma_{k-1}(x)}{\sigma_{k+1}(x)} M_{(k-1)}(0) \quad (1.18)$$

and by iteration

$$M_{(k)}(0) = a^k \frac{\sigma_1(x)\sigma_0(x)}{\sigma_{k+1}(x)\sigma_k(x)} M_0(0) \quad (1.19)$$

From (1.11) and (1.16), and in the last step (1.14),

$$\sum_{m=0}^x M_0(m) = \sum_{m=0}^x M_0(0)\sigma_0(m) = M_0(0)\sigma_1(x) = 1 \quad (1.20)$$

Hence finally

$$\begin{aligned} M_{(k)}(m) &= M_{(k)}(0)\sigma_k(m) \\ &= a^k \frac{\sigma_0(x)\sigma_k(m)}{\sigma_{k+1}(x)\sigma_k(x)} \end{aligned} \quad (1.21)$$

and

$$M_{(k)} = \sum_{m=0}^x M_{(k)}(m) = a^k \sigma_0(x)/\sigma_k(x) \quad (1.22)$$

Ordinary moments are found from the factorial moments by linear relations; thus if m_k is the k th ordinary moment (about the origin)

$$\begin{aligned} m_0 &= M_{(0)} & m_1 &= M_{(1)} & m_2 &= M_{(2)} + M_{(1)} \\ m_3 &= M_{(3)} + 3M_{(2)} + M_{(1)} \end{aligned}$$

Thus

$$\begin{aligned} m_0(m) &= \sigma_0(m)/\sigma_1(x) \\ m_1(m) &= a\sigma_1(m)\sigma_0(x)/\sigma_1(x)\sigma_2(x) \\ m_2(m) &= a^2\sigma_2(m)\sigma_0(x)/\sigma_2(x)\sigma_3(x) + a\sigma_1(m)\sigma_0(x)/\sigma_1(x)\sigma_2(x) \end{aligned}$$

and, in particular, using notation of the text

$$\begin{aligned} m_0(x) &= \sigma_0(x)/\sigma_1(x) = E_{1,x}(a) \\ \alpha_x &= \frac{m_1(x)}{m_0(x)} = a \frac{\sigma_1(x)}{\sigma_2(x)} = \frac{a}{x - a + 1 + aE_{1,x}(a)} \end{aligned} \quad (1.23)$$

$$\begin{aligned} v_x &= \frac{m_2(x)}{m_0(x)} - \alpha_x^2 = \frac{a^2\sigma_1(x)}{\sigma_3(x)} + \alpha_x - \alpha_x^2 \\ &= \alpha_x[1 - \alpha_x + 2a(x + 2 + \alpha_x - a)^{-1}] \end{aligned} \quad (1.24)$$

Finally the sum moments: $m_k = \sum_0^x m_k(m)$ are

$$\begin{aligned} m_0 &= 1 \\ m_1 &= \alpha = a\sigma_0(x)/\sigma_1(x) = aE_{1,x}(a) \end{aligned} \quad (1.25)$$

$$m_2 = a^2\sigma_0(x)/\sigma_2(x) + m_1 = m_1[a(x + 1 + m_1 - a)^{-1} + 1] \quad (1.26)$$

$$v = m_2 - m_1^2 = m_1[1 - m_1 + a(x + 1 + m_1 - a)^{-1}]$$

In these, $E_{1,x}(a) = \sigma_0(x)/\sigma_1(x)$ is the familiar Erlang loss function.

APPENDIX II — CHARACTER OF OVERFLOW LOAD WHEN NON-RANDOM TRAFFIC IS OFFERED TO A GROUP OF TRUNKS

It has long been recognized that it would be useful to have a method by which the character of the overflow traffic could be determined when non-random traffic is offered to a group of trunks. Excellent agreement has been found in both throwdown and field observation over ranges of considerable interest with the "equivalent random" method of describ-

α_2 = AVERAGE OF OVERFLOW LOAD FROM X TRUNKS IN ERLANGS

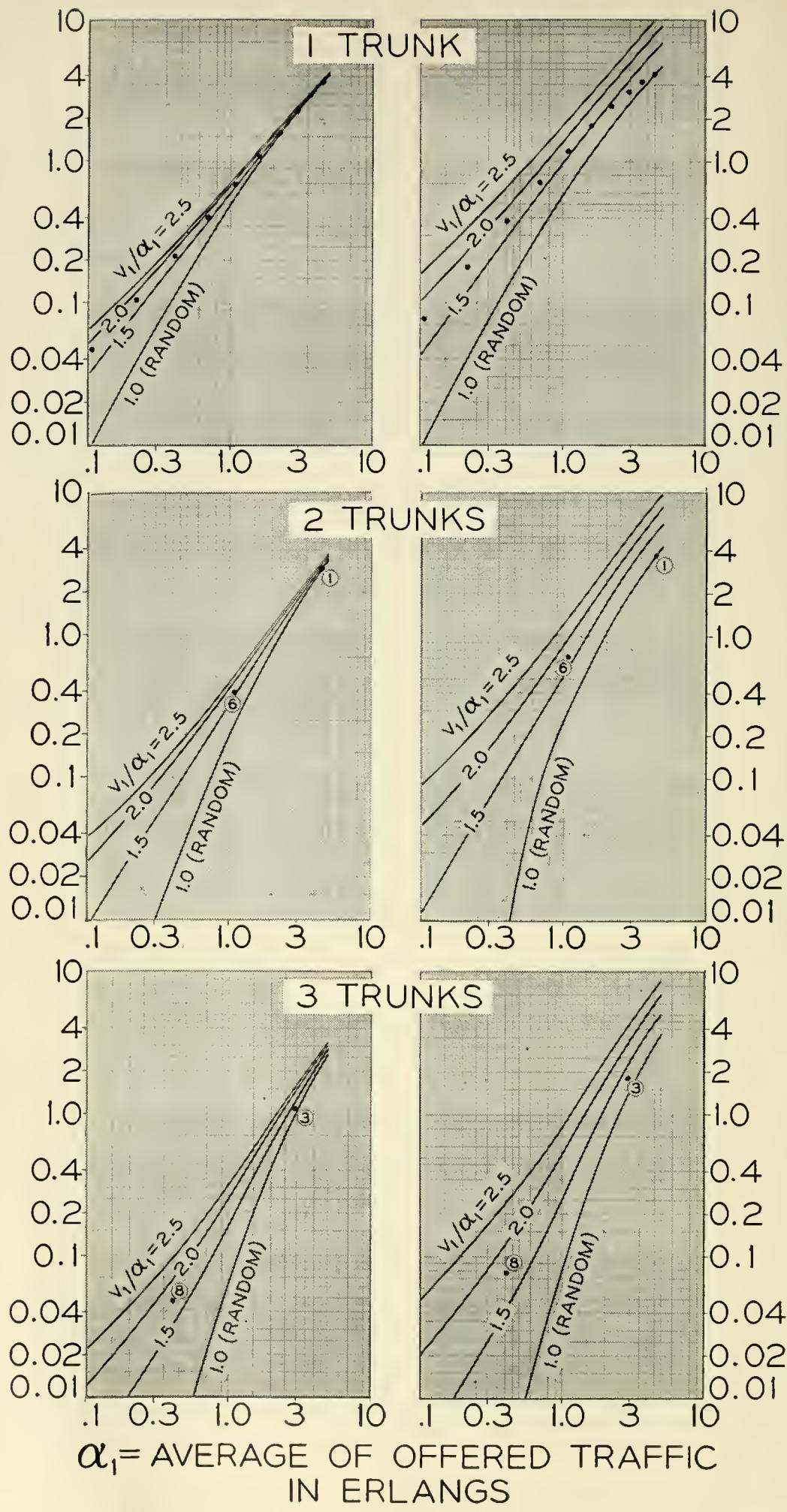
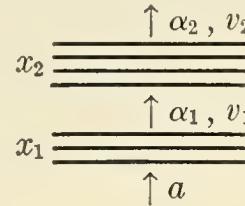


Fig. 51 — Mean and variance of overflow load when non-random traffic is offered to a group of trunks.

ing the character of non-random traffic. An approximate solution of the problem is offered based on this method.

Suppose a random traffic a is offered to a straight multiple which is divided into a lower x_1 portion and an upper x_2 portion, as follows:



From Nyquist's and Molina's work we know the mean and variance of the two overflows to be:

$$\begin{aligned}\alpha_1 &= a \cdot E_{1,x_1}(a) = a \frac{\frac{a^{x_1}}{x_1!}}{1 + a + \frac{a^2}{2!} + \cdots + \frac{a^{x_1}}{x_1!}} \\ v_1 &= \alpha_1 \left[1 - \alpha_1 + \frac{a}{x_1 - a + \alpha_1 + 1} \right] \\ \alpha_2 &= a \cdot E_{1,x_1+x_2}(a) \\ v_2 &= \alpha_2 \left[1 - \alpha_2 + \frac{a}{x_1 + x_2 - a + \alpha_2 + 1} \right]\end{aligned}$$

Since α_1 and v_1 completely determine a and x_1 , and these in turn, with x_2 , determine α_2 and v_2 , we may express α_2 and v_2 in terms of only α_1 , v_1 , and x_2 . The overflow characteristics (α_2 and v_2), are then given for a non-random load (α_1 and v_1) offered to x trunks as was desired.

Fig. 51 of this Appendix has been constructed by the Equivalent Random method. The charts show the expected values of α_2 and v_2 when α_1 , v_1 (or v_1/α_1), and x_2 , are given. The range of α_1 is only 0 to 5 erlangs, and v/α is given only from the Poisson unity relation to a peakedness value of 2.5. Extended and more definitive curves or tables could readily, of course, be constructed.

The use of the curves can perhaps best be illustrated by the solution of a familiar example.

Example: A load of 4.5 erlangs is submitted to 10 trunks; on the "lost calls cleared" basis; what is the average load passing to overflow?

Solution: Compute the load characteristics from the first trunk when 4.5 erlangs of random traffic are submitted to it. These values are found to be $\alpha_1 = 3.68$, $v_1 = 4.15$. Now using α_1 and v_1 (or $v_1/\alpha_1 = 4.15/3.68 = 1.13$) as the offered load to the second trunk, read on the chart the parameters of the overflow from the second trunk, and so on. The successive overflow values are given in Table XVIII.

The proportion of load overflowing the group is then $0.0472/4.50 = 0.0105$, which agrees, of course, with the Erlang $E_{1,10}(4.5)$ value. The successive overflow values are shown on the chart by the row of dots along the α_2 and v_2 1-trunk curves.

Instead of considering successive single-trunk overflows as in the example above, other numbers of trunks may be chosen and their overflows determined. For example suppose the 10 trunks are subdivided into $2 + 3 + 2 + 3$ trunks. The loads overflowing these groups are given in Table XIX.

Again the overflow is 0.0472 erlang, or a proportion lost of 0.0105, which is, as it should be, the same as found in the previous example. The values read in this example are indicated by the row of dots marked 1, 3, 6, 8 on the 2-trunk and 3-trunk curves.

The above procedure and curves should be of use in obtaining an estimate of the character of the overflow traffic when a non-random load is offered to a group of paths.

TABLE XVIII — SUCCESSIVE NON-RANDOM OVERFLOWS

Trunk Number <i>i</i>	Characteristics of Load Offered to Trunk No. <i>i</i> (same as overflow from previous trunk)		
	Average	Variance	Ratio of variance to average
1	4.50	4.50	1.00 (Random)
2	3.68	4.15	1.13
3	2.92	3.68	1.26
4	2.22	3.11	1.40
5	1.61	2.46	1.53
6	1.09	1.80	1.64
7	0.694	1.19	1.72
8	0.406	0.709	1.75
9	0.217	0.377	1.74
10	0.106	0.180	1.70
Overflow	0.0472	0.077	1.64

TABLE XIX — SUCESSIVE NON-RANDOM OVERFLOWS

Trunk Number <i>i</i>	No. Trunks in Next Bundle	Offered Load Characteristics (same as overflow from previous trunk)		
		Average	Variance	Ratio of variance to average
1	2	4.50	4.50	1.00 (Random)
3	3	2.92	3.68	1.26
6	2	1.09	1.80	1.64
8	3	0.406	0.709	1.75
Overflow		0.0472	0.077	1.64

Crosstalk on Open-Wire Lines

By W. C. BABCOCK, ESTHER RENTROP, and C. S. THAELER

(Manuscript received September 29, 1955)

Crosstalk on open-wire lines results from cross-induction between the circuits due to the electric and magnetic fields surrounding the wires. The limitation of crosstalk couplings to tolerable magnitudes is achieved by systematically turning over or transposing the conductors that comprise the circuits. The fundamental theory underlying the engineering of such transposition arrangements was presented by A. G. Chapman in a paper entitled *Open-Wire Crosstalk* published in the Bell System Technical Journal in January and April, 1934.

There is now available a Monograph (No. 2520) supplementing Mr. Chapman's paper which reflects a considerable amount of experience resulting from the application of these techniques and provides a basis for the engineering of open-wire plant. The scope of the material is indicated by the following:

TRANSPOSITION PATTERNS

This describes the basic transposition types which define the number and locations of transpositions applied to the individual open-wire circuits.

TYPES OF CROSSTALK COUPLING

Crosstalk occurs both within incremental segments of line and between such segments. Furthermore, the coupling may result from cross-induction directly from a disturbing to a disturbed circuit or indirectly by way of an intervening tertiary circuit. On the disturbed circuit the crosstalk is propagated both toward the source of the original signal and toward the distant terminal. A knowledge of the relative importance of the various types of coupling is valuable in establishing certain time-saving approximations which facilitate the analysis of the total crosstalk picture.

TYPE UNBALANCE CROSSTALK

Crosstalk is measured in terms of a current ratio between the disturbing and disturbed circuits at the point of observation. Crosstalk between open-wire circuits is also generally computed in terms of a current ratio (cu) but it is also convenient to refer to it in terms of a coupling loss (db). The coupling in crosstalk units (cu) is the product of three terms: a coefficient dependent on wire configuration; a type unbalance dependent on transposition patterns; and frequency. The coefficient represents the coupling between relatively untransposed circuits of a specified length (1 mile) at a specific frequency (1 kc). The type unbalance is a measure of the inability to completely cancel out crosstalk by introducing transpositions because of interaction effects between the two halves of the exposure and because of propagation effects, primarily phase shift. Type unbalance is expressed in terms of a residual unbalance in miles and the frequency is expressed in kilocycles.

The coefficients applicable to lines built in accordance with certain standardized specifications are available in tabular form. When it is desired to obtain coefficients for other types of line, it is possible to compute approximate values which may be modified by correction factors to indicate the relationship between the computed values and measurements on carefully constructed lines.

Expressions for near-end type unbalance for certain simple types of exposures are developed and the formulas for all types of exposures are given. In addition, the values for near-end type unbalance are tabulated at 30° line angle intervals for lines where the propagation angle is $2,880^\circ$ or less.

The principal component of far-end crosstalk between well transposed circuits results from compound couplings involving tertiary circuits. Again the expressions are developed for some of the exposures involving a few transpositions and the procedure for obtaining the formulas for any type of exposure is shown. Formulas are included for the types of exposures encountered in normal practice and the numerical values of far-end type unbalance are given at 30° intervals for line angles up to $2,880^\circ$.

SUMMATION OF CROSSTALK

The procedures referred to thus far evaluate the crosstalk occurring within a limited length of line known as a transposition section. In practice, however, a line is transposed as a series of sections. It is necessary, therefore, to determine how the crosstalk arising within the several

sections and that arising from interactions between the sections tend to combine. In a series of like transposition sections there is a tendency for the crosstalk to increase systematically, sometimes reaching intolerable magnitudes. This tendency can be controlled to a degree by introducing transpositions at the junctions between the sections, thus cancelling out some of the major components of the crosstalk. Complete cancellation is impossible because of interaction and propagation effects.

ABSORPTION

Since very significant couplings exist by way of tertiary circuits, it is possible for crosstalk to reappear on the disturbing circuit and thus strengthen or attenuate the original signal. This gives rise to the appearance of high attenuation known as absorption peaks in the line loss characteristic at certain critical frequencies. The evaluation of such pair-to-self coupling requires the use of coefficients which differ from those between different pairs and these are given for standard configurations.

STRUCTURAL IRREGULARITIES

It is impracticable to maintain absolute uniformity in the spacing between wires and in the spacing of transpositions. Thus there are unavoidable variations in the couplings between pairs from one transposition interval to the next. This in turn reduces the effectiveness of the measures to control the systematic or type unbalance crosstalk and produces what is known as irregularity crosstalk. Since the occurrence of structural irregularities tends to follow a random distribution, it is possible to evaluate it statistically and procedures for doing so are included. In addition to this direct effect of structural irregularities, there is a component of crosstalk resulting from the combination of systematic and random unbalances. A method is developed for estimating the magnitude of this important component of crosstalk.

EXAMPLES

In order to demonstrate how the procedures and data are used in solving practical problems, there is included the development of a transposition system to satisfy certain assumed conditions. This is carried through to the selection of transposition types for one transposition section and the selection of suitable junction transpositions.

Additional examples of transposition engineering are given in the form

of several transposition systems which have been widely used in the Bell System. These include:

- Exposed Line — for voice frequency service.
- C1 — for voice frequency and carrier service up to 30 kc.
- J5 — for voice frequency and carrier operation up to 143 kc.
- O1 — for voice frequency and compandored carrier operation up to 156 kc.
- R1C — suitable for exchange lines with a limited number of carrier assignments.

Altogether, the theory, explanatory material, formulas and comprehensive data included in the Monograph make it possible to estimate open-wire crosstalk couplings and provide the necessary background for the development of new transposition systems.

Bell System Technical Papers Not Published in This Journal

ALSBERG, D. A.¹

6-KMC Sweep Oscillator, I.R.E. Trans., **PGI-4**, pp. 32–39, Oct., 1955.

ANDERSON, J. R.,¹ BRADY, G. W.,¹ MERZ, W. J.,¹ and REMEIKA, J. P.¹

Effects of Ambient Atmosphere on the Stability of Barium Titanate, J. Appl. Phys., Letter to the Editor, **26**, pp. 1387–1388, Nov., 1955.

ANDERSON, O. L.,¹ and ANDREATCH, P.¹

Stress Relaxation in Gold Wire, J. Appl. Phys., **26**, pp. 1518–1519, Dec., 1955.

ANDERSON, P. W.,¹ and HASEGAWA, H.⁵

Considerations on Double Exchange, Phys. Rev., **100**, pp. 675–681, Oct. 15, 1955.

ANDERSON, P. W.¹

Electromagnetic Theory of Cyclotron Resonance in Metals, Phys. Rev., Letter to the Editor, **100**, pp. 749–750, Oct. 15, 1955.

ANDREATCH, P., see Anderson, O. L.

AUGUSTINE, C. F., see Slocum, A.

BARSTOW, J. M.¹

The ABC's of Color Television, Proc. I.R.E., **43**, pp. 1574–1579, Nov., 1955.

BARTLETT, C. A.²

Closed-Circuit Television in the Bell System, Elec. Engg., **75**, pp. 34–37, Jan., 1956.

1. Bell Telephone Laboratories, Inc.

2. American Telephone and Telegraph Company.

5. University of Tokyo, Japan.

BECKER, J. A.¹

Adsorption on Metal Surfaces and Its Bearing on Catalysis, Advances in Catalysis, 1955, Nov., 1955.

BÖMMEL, H. E.¹

Ultrasonic Attenuation in Superconducting and Normal-Conducting Tin at Low Temperatures, Phys. Rev., Letter to the Editor, 100, pp. 758-759, Oct. 15, 1955.

BEMSKI, G.¹

Lifetime of Electrons in p-Type Silicon, Phys. Rev., 100, pp. 523-524, Oct. 15, 1955.

BENNETT, W. R.¹

Steady State Transmission Through Networks Containing Periodically Operated Switches, Trans. I.R.E., PGCT., 2, pp. 17-21, Mar., 1955.

BÖMMEL, H. E.,¹ MASON, W. P.,¹ and WARNER, A. W., JR.¹

Experimental Evidence for Dislocation in Crystalline Quartz, Phys. Rev., Letter to the Editor, 99, pp. 1895-1896, Sept. 15, 1955.

BRADLEY, W. W., see Compton, K. G.

BRATTAIN, W. H., see Buck, T. M., and Pearson, G. L.

BRADY, G. W., see Anderson, J. R.

BROWN, W. L.¹

Surface Potential and Surface Charge Distribution from Semiconductor Field Effect Measurements, Phys. Rev., 100, pp. 590-591, Oct. 15, 1955.

BUCK, T. M.,¹ and BRATTAIN, W. H.¹

Investigations of Surface Recombination Velocities on Germanium by the Photoelectric Magnetic Method, J. Electrochem. Soc., 102, pp. 636-640, Nov., 1955.

CETLIN, B. B., see Galt, J. K.

CHARNES, A., see Jacobson, M. J.

1. Bell Telephone Laboratories, Inc.

COMPTON, K. G.,¹ MENDIZZA, A.,¹ and BRADLEY, W. W.¹

Atmospheric Galvanic Couple Corrosion, Corrosion, **11**, pp. 35-44, Sept., 1955.

CORENZWIT, E., see Matthias, B. T.

DAIL, H. W., JR., see Galt, J. K.

DILLON, J. F., JR.,¹ GESCHWIND, S.,¹ and JACCARINO, V.¹

Ferromagnetic Resonance in Single Crystals of Manganese Ferrite, Phys. Rev., Letter to the Editor, **100**, pp. 750-752, Oct. 15, 1955.

DODGE, H. F.¹

Chain Sampling Inspection Plan, Ind. Quality Control, **11**, pp. 10-13, Jan., 1955.

DODGE, H. F.¹

Skip-lot Sampling Plan, Ind. Quality Control, **11**, pp. 3-5, Feb., 1955.

FAGEN, R. E.,¹ and RIORDAN, J.¹

Queueing Systems for Single and Multiple Operation, J. S. Ind. Appl. Math., **3**, pp. 73-79, June, 1955.

FINE, M. E.¹

Erratum: Elastic Constants of Germanium Between 1.7° and 80°K J. Appl. Phys., Letter to the Editor, **26**, p. 1389, Nov., 1955.

FLASCHEN, S. S.¹

A Barium Titanate Synthesis from Titanium Esters, J. Am. Chem. Soc., **77**, p. 6194, Dec., 1955.

FLETCHER, R. C.,¹ YAGER, W. A.,¹ and MERRITT, F. R.¹

Observation of Quantum Effects in Cyclotron Resonance, Phys. Rev., Letter to the Editor, **100**, pp. 747-748, Oct. 15, 1955.

FRANKE, H. C.¹

Noise Measurement on Telephone Circuits, Tele-Tech., **14**, pp. 85-97, Mar., 1955.

1. Bell Telephone Laboratories, Inc.

GALT, J. K.,¹ YAGER, W. A.,¹ MERRITT, F. R.,¹ CETLIN, B. B.,¹ and DAIL, H. W., JR.¹

Cyclotron Resonance in Metals: Bismuth, Phys. Rev., Letter to the Editor, **100**, pp. 748-749, Oct. 15, 1955.

GELLER, S.,¹ and THURMOND, C. D.¹

On the Question of a Crystalline SiO, Am. Chem. Soc. J., **77**, pp. 5285-5287, Oct. 20, 1955.

GESCHWIND, S., see Dillon, J. F.

HARKER, K. J.¹

Periodic Focusing of Beams from Partially Shielded Cathodes, I.R.E. Trans., ED-2, pp. 13-19, Oct., 1955.

HASEGAWA, H., see Anderson, P. W.

HAYNES, J. R.,¹ and HORNBECK, J. A.¹

Trapping of Minority Carriers in Silicon II: n-type Silicon, Phys. Rev., **100**, pp. 606-615, Oct. 15, 1955.

HORNBECK, J. A., see Haynes, J. R.

ISRAEL, J. O.,¹ MECHLINE, E. B.,¹ and MERRILL, F. F.¹

A Portable Frequency Standard for Navigation, I.R.E. Trans., PGI-4, pp. 116-127, Oct., 1955.

JACCARINO, V., see Dillon, J. F.

JACOBSON, M. J.,¹ CHARNES, A., and SAIBEL, E.¹

The Complete Journal Bearing With Circumferential Oil Inlet, Trans. A.S.M.E., **77**, pp. 1179-1183, Nov., 1955.

JAMES, D. B., see Neilson, G. C.

KOHN, W.,¹ and SCHECHTER, D.⁴

Theory of Acceptor Levels in Germanium, Phys. Rev., Letter to the Editor, **99**, pp. 1903-1904, Sept. 15, 1955.

1. Bell Telephone Laboratories, Inc.

4. Carnegie Institute.

LAW, J. T.,¹ and MEIGS, P. S.¹

The Effect of Water Vapor on Grown Germanium and Silicon n-p Junction Units, *J. Appl. Phys.*, **26**, pp. 1265-1273, Oct., 1955.

LEWIS, H. W.¹

Search for the Hall Effect in a Superconductor: II — Theory, *Phys. Rev.*, **100**, pp. 641-645, Oct. 15, 1955.

LINVILL, J. G.,¹ and MATTSON, R. H.¹

Junction Transistor Blocking Oscillators, *Proc. I.R.E.*, **43**, pp. 1632-1639, Nov., 1955.

LOGAN, R. A.¹

Precipitation of Copper in Germanium, *Phys. Rev.*, **100**, pp. 615-617, Oct. 15, 1955.

LOGAN, R. A.,¹ and SCHWARTZ, M.¹

Restoration of Resistivity and Lifetime in Heat Treated Germanium, *J. Appl. Phys.*, **26**, pp. 1287-1289, Nov., 1955.

MCCALL, D. W., see Shulman, R. G.

MASON, W. P., see Bömmel, H. E.

MATTHIAS, B. T.,¹ and CORENZWIT, E.¹

Superconductivity of Zirconium Alloys, *Phys. Rev.*, **100**, pp. 626-627, Oct. 15, 1955.

MATTSON, R. H., see Linvill, J. G.

MAYS, J. M., see Shulman, R. G.

MECHLINE, E. B., see Israel, J. O.

MEIGS, P. S., see Law, J. T.

MENDIZZA, A., see Compton, K. G.

MERRILL, F. F., see Israel, J. O.

MERRITT, F. R., see Fletcher, R. C., and Galt, J. K.

MERZ, W. J., see Anderson, J. R.

1. Bell Telephone Laboratories, Inc.

MOLL, J. L.¹

Junction Transistor Electronics, Proc. I.R.E., **43**, pp. 1807-1818, Dec., 1955.

MUMFORD, W. W.,¹ and SCHAFERSMAN, R. L.¹

Data on Temperature Dependence of X-Band Fluorescent Lamp Noise Sources, I.R.E. Trans., PGI-4, pp. 40-46, Oct., 1955.

NEILSON, G. C.,⁶ and JAMES, D. B.¹

Time of Flight Spectrometer for Fast Neutrons, Rev. Sci. Instr., **26**, pp. 1018-1023, Nov., 1955.

NESBITT, E. A.,¹ and WILLIAMS, H. J.¹

New Facts Concerning the Permanent Magnet Alloy, Alnico 5, Conf. on Magnetism and Magnetic Materials, T-78, pp. 205-209, Oct., 1955.

NESBITT, E. A.,¹ and WILLIAMS, H. J.¹

Shape and Crystal Anisotropy of Alnico 5, J. Appl. Phys., **26**, pp. 1217-1221, Oct., 1955.

OWNES, C. D.¹

Stability of Molybdenum Permalloy Powder Cores, Conf. on Magnetism and Magnetic Materials, T-78, pp. 334-339, Oct., 1955.

PEARSON, G. L.,¹ and BRATTAIN, W. H.¹

History of Semiconductor Research, Proc. I.R.E., **43**, pp. 1794-1806, Dec., 1955.

PEDERSON, L.¹

Aluminum Die Castings in Carrier Telephone Systems, Modern Metals, **11**, pp. 65, 68, 70, Sept., 1955.

PRINCE, M. B.¹

High-Frequency Silicon Aluminum Alloy Junction Diode, Trans. I.R.E., ED-2, pp. 8-9, Oct., 1955.

REMEIKA, J. P., see Anderson, J. R.

RIORDAN, J., see Fagen, R. E.

1. Bell Telephone Laboratories, Inc.

6. University of British Columbia, Vancouver, Canada.

SAIBEL, E., see Jacobson, M. J.

SCHAFFERSMAN, R. L., see Mumford, W. W.

SCHECHTER, D., see Kohn, W.

SCHELKUNOFF, S. A.¹

On Representation of Electromagnetic Fields in Cavities in Terms of Natural Modes of Oscillation, *J. Appl. Phys.*, **26**, pp. 1231-1234, Oct., 1955.

SCHWARTZ, M., see Logan, R. A.

SHULMAN, R. G.,¹ MAYS, J. M.,¹ and McCALL, D. W.¹

Nuclear Magnetic Resonance in Semiconductors: I—Exchange Broadening in InSb and GaSb, *Phys. Rev.*, **100**, pp. 692-699, Oct. 15, 1955.

SLOCUM, A.,¹ and AUGUSTINE, C. F.¹

6-KMC Phase Measurement System For Traveling Wave Tube, *Trans. I.R.E.*, **PGI-4**, pp. 145-149, Oct., 1955.

THURMOND, C. D., see Geller, S.

UHLIR, A., JR.¹

Micromachining with Virtual Electrodes, *Rev. Sci. Instr.*, **26**, pp. 965-968, Oct., 1955.

ULRICH, W., see Yokelson, B. J.

VAN UITERT, L. G.¹

DC Resistivity in the Nickel and Nickel Zinc Ferrite System, *J. Chem. Phys.*, **23**, pp. 1883-1887, Oct., 1955.

VAN UITERT, L. G.¹

Low Magnetic Saturation Ferrites for Microwave Applications, *J. Appl. Phys.*, **26**, pp. 1289-1290, Nov., 1955.

WANNIER, G. H.¹

Possibility of a Zener Effect, *Phys. Rev. Letter to the Editor*, **100**, p. 1227, Nov., 15, 1955.

1. Bell Telephone Laboratories, Inc.

WANNIER, G. H.¹

Threshold Law for Multiple Ionization, Phys. Rev., **100**, pp. 1180,
Nov. 15, 1955.

WARNER, A. W., JR., see Bömmel, H. E.

WILLIAMS, H. J., see Nesbitt, E. A.

YAGER, W. A., see Fletcher, R. C., and Galt, J. K.

YOKELSON, B. J.,¹ and ULRICH, W.¹

Engineering Multistage Diode Logic Circuits, Elec. Engg., **74**, p. 1079,
Dec., 1955.

1. Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ALLISON, H. W., see Moore, G. E.

BAKER, W. O., see Winslow, F. H.

BASSECHES, H., and MCLEAN, D. A.

Gassing of Liquid Dielectrics Under Electrical Stress, Monograph 2448.

BOZORTH, R. M., TILDEN, E. F., and WILLIAMS, A. J.

Anistropy and Magnetostriction of Some Ferrites, Monograph 2513.

BRADLEY, W. W., see Compton, K. G.

COMPTON, K. G., MENDIZZA, A., and BRADLEY, W. W.

Atmospheric Galvanic Couple Corrosion, Monograph 2470.

DAVIS, J. L., see Suhl, H.

FAGEN, R. E., and RIORDAN, JOHN

Queueing Systems for Single and Multiple Operation, Monograph 2506.

FINE, M. E.

Elastic Constants of Germanium Between 1.7° and 80°K, Monograph 2479.

FORSTER, J. H., see Miller, L. E.

GALT, J. K., see Yager, W. A.

GEBALLE, T. H., see Morin, F. J.

* 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.

GIANOLA, U. F.

Use of Wiedemann Effect for Magnetostrictive Coupling of Crossed Coils, Monograph 2492.

GREEN, E. I.

The Story of Q, Monograph 2491.

GULDNER, W. G., see Wooten, L. A.

HARROWER, G. A.

Measurement of Electron Energies by Deflection in a Uniform Electric Field, Monograph 2495.

HAUS, H. A., and ROBINSON, F. N. H.

The Minimum Noise Figure of Microwave Beam Amplifiers, Monograph 2468.

HINES, M. E., HOFFMAN, G. W., and SALOOM, J. A.

Positive-ion Drainage in Magnetically Focused Electron Beams, Monograph 2481.

HOFFMAN, G. W., see Hines, M. E.

KELLY, M. J.

Training Programs of Industry for Graduate Engineers, Monograph 2512.

LAW, J. T., and MEIGS, P. S.

Water Vapor on Grown Germanium and Silicon n-p Junction Units, Monograph 2500.

McAfee, K. B., Jr.

Attachment Coefficient and Mobility of Negative Ions by a Pulse Technique, Monograph 2471.

MCLEAN, D. A., see Basseches, H.

MEIGS, P. S., see Law, J. T.

MENDIZZA, A., see Compton, K. G.

MERRITT, F. R., see Yager, W. A.

MILLER, L. E., and FORSTER, J. H.

Accelerated Power Aging with Lithium-Doped Point Contact Transistors, Monograph 2482.

MILLER, S. L.

Avalanche Breakdown in Germanium, Monograph 2477.

MOORE, G. E., see Wooten, L. A.

MOORE, G. E., and ALLISON, H. W.

Adsorption of Strontium and of Barium on Tungsten, Monograph 2498.

MORIN, F. J., and GEBALLE, T. H.

Electrical Conductivity and Seebeck Effect in $\text{Ni}_{0.80}\text{Fe}_{2.20}\text{O}_4$, Monograph 2514.

MORRISON, J., see Wooten, L. A.

NESBITT, E. A., and WILLIAMS, H. J.

Shape and Crystal Anisotropy of Alnico 5, Monograph 2502.

OLMSTEAD, P. S.

Quality Control and Operations Research, Monograph 2530.

PEARSON, G. L., see Read, W. T., Jr.

PFANN, W. G.

Temperature Gradient Zone Melting, Monograph 2451.

POOLE, K. M.

Emission from Hollow Cathodes, Monograph 2480.

READ, W. T., JR., and PEARSON, G. L.

The Electrical Effects of Dislocations in Germanium, Monograph 2511.

RIORDAN, JOIN, see Fagen, R. E.

ROBINSON, F. N. H., see Haus, H. A.

SALOOM, J. A., see Hines, M. E.

SCHELKUNOFF, S. A.

Electromagnetic Fields in Cavities in Terms of Natural Modes of Oscillation, Monograph 2505.

SEARS, R. W.

A Regenerative Binary Storage Tube, Monograph 2527.

SLICITER, W. P.

Proton Magnetic Resonance in Polyamides, Monograph 2490.

SUHL, H., VAN UITERT, L. G., and DAVIS, J. L.

Ferromagnetic Resonance in Magnesium-Manganese Aluminum Ferrite Between 160 and 1900 mc, Monograph 2472.

TILDEN, E. F., see Bozorth, R. M.

TREUTING, R. G.

Some Aspects of Slip in Germanium, Monograph 2485.

UHLIR, A., JR.

Micromachining with Virtual Electrodes, Monograph 2515.

VAN UITERT, L. G., see Suhl, H.

WALKER, L. R.

Power Flow in Electron Beams, Monograph 2469.

WILLIAMS, A. J., see Bozorth, R. M.

WILLIAMS, H. J., see Nesbitt, E. A.

WINSLOW, F. H., BAKER, W. O., YAGER, W. A.

Odd Electrons in Polymer Molecules, Monograph 2486.

WOOTEN, L. A., MOORE, G. E., GULDNER, W. G., and MORRISON, J.

Excess Barium in Oxide-Coated Cathodes, Monograph 2497.

YAGER, W. A., see Winslow, F. H.

YAGER, W. A., GALT, J. K., and MERRITT, F. R.

Ferromagnetic Resonance in Two Nickel-Iron Ferrites, Monograph 2478.

Contributors to This Issue

ARMAND O. ADAM,* New York Telephone Company, 1917-1920; Western Electric Company, 1920-24; Bell Telephone Laboratories; 1925-. Mr. Adam tested local dial switching systems before turning to design on the No. 1 and toll crossbar systems. From 1942 to 1945 he was associated with the Bell Laboratories School For War Training. Since then he has been concerned with the design and development of the marker for the No. 5 crossbar system. Currently he is supervising a group doing common control circuit development work for the crossbar tandem switching system.

WALLACE C. BABCOCK, A.B., Harvard University, 1919; S.B., Harvard University, 1922. U.S. Army, 1917-1919. American Telephone and Telegraph Company, 1922-1934; Bell Telephone Laboratories, 1934-. Mr. Babcock was engaged in crosstalk studies until World War II. Afterward he was concerned with radio countermeasure problems for the N.D.R.C. Since then he has been working on antenna development for mobile radio and point-to-point radio telephone systems and military projects. Member of I.R.E. and Harvard Engineering Society.

FRANKLIN H. BLECHER, B.E.E., 1949, M.E.E., 1950 and D.E.E., 1955, Brooklyn Polytechnic Institute; Polytechnic Research and Development Company, June, 1950 to July, 1952; Bell Telephone Laboratories 1952-. Dr. Blecher has been engaged in transistor network development. His principal interest has been the application of junction transistors to feedback amplifiers used in analog and digital computers. He is a member of Tau Beta Pi, Eta Kappa Nu and Sigma Xi and is an associate member of the I.R.E.

W. E. DANIELSON, B.S., 1949, M.S., 1950, Ph.D, 1952, California Institute of Technology; Bell Laboratories 1952-. Dr. Danielson has been engaged in microwave noise studies with application to traveling-wave tubes and he has been in charge of development of traveling-wave tubes

* Inadvertently, Mr. Adam's biography was omitted from the January issue of the JOURNAL in which his article, "Crossbar Tandem as a Long Distance Switching Equipment," appeared.

for use at 11,000 megacycles since June of 1954. He is the author of articles published by the Journal of Applied Physics, Proceedings of the I.R.E., and the B.S.T.J., and he is a Member of the American Physical Society, Tau Beta Pi, and Sigma Xi.

AMOS E. JOEL, JR., B.S., Massachusetts Institute of Technology, 1940; M.S., 1942; Bell Telephone Laboratories, 1940-. Mr. Joel's first assignment was in relay engineering. He then worked in the crossbar test laboratory and later conducted fundamental development studies. During World War II, he made studies of communications projects and from 1944 to 1945 designed circuits for a relay computer. Later he prepared text and taught a course in switching design. The next two years were spent designing AMA computer circuits, and since 1949 Mr. Joel has been engaged in making fundamental engineering studies and directing exploratory development of electronic switching systems. He was appointed Switching Systems Development Engineer in 1954. Member of A.I.E.E., I.R.E., Association for Computing Machinery, and Sigma Xi.

ESTHER M. RENTROP, B.S., 1926, Louisiana State Normal College. Miss Rentrop joined the transmission group of the Development and Research Department of the American Telephone and Telegraph Company in 1928, and transferred to Bell Laboratories in 1934. In both companies she has been concerned principally with control of crosstalk, both in field studies and transposition design work. During World War II, she assisted in problems of the Wire Section, Eatontown Signal Corps Laboratory at Fort Monmouth, and later she worked on other military projects at the Laboratories for the duration of the war. Miss Rentrop is presently a member of the noise and crosstalk studies group of the Outside Plant Engineering Department and is engaged in studies of interference prevention.

JACK L. ROSENFELD is a student in electrical engineering at the Massachusetts Institute of Technology. He will receive the S.M. and S.B. degrees in 1957. He has been with Bell Telephone Laboratories on co-operative assignments in microwave tube development and electronic central office during 1954 and 1955. He is a student member of the I.R.E. and a member of Tau Beta Pi and Eta Kappa Nu.

JOSEPH A. SALOOM, JR., B.S., 1948, M.S., 1949, and Ph.D., 1951, all in Electrical Engineering, University of Illinois. He joined Bell Laboratories in 1951. Mr. Saloom worked on electron tube development at

Murray Hill until 1955 with particular emphasis on electron beam studies. He is now at the Allentown, Pa., laboratory where he is engaged in the development of microwave oscillators. Member of the Institute of Radio Engineers, Sigma Xi, Eta Kappa Nu, Pi Mu Epsilon.

CHARLES S. THAELER, Moravian College, 1923-25, Lehigh University 1925-28, E.E., 1928. During the summer of 1927 he was employed by the Bell Telephone Company of Pennsylvania, returning there after graduation, where he was concerned with transmission engineering and the Toll Fundamental Plan. In 1943 he was on loan to the Operating and Engineering Department of the A.T.&T. Co., working on toll transmission studies. From 1944 to the present he has been with the Operating and Engineering Department and is currently engaged in toll circuit noise and crosstalk problems on open wire and cable systems. Mr. Thaeler is an Associate Member of A.I.E.E., and member of Phi Beta Kappa, Tau Beta Pi, and Eta Kappa Nu.

PING KING TIEN, B. S., National Central University, China, 1942; M.S., 1948, Ph.D., 1951, Stanford University; Stanford Microwave Laboratory, 1949-50; Stanford Electronics Research Laboratory, 1950-52; Bell Telephone Laboratories, 1952-. Since joining the Laboratories, Dr. Tien has been concerned with microwave tube research, particularly traveling-wave tubes. In the course of this research he has engaged in studies of space charge wave amplifiers, helix propagation, electron beam focusing, and noise. He is a member of Sigma Xi.

ARTHUR UHLIR, JR., B.S., M.S. in Ch.E., Illinois Institute of Technology, 1945, 1948; S.M. and Ph.D. in Physics, University of Chicago, 1950, 1952. Dr. Uhrlir has been engaged in many phases of transistor development since joining the Laboratories in 1951, including electrochemical techniques and semiconductor device theory. Since 1952 he has participated in the Laboratories' Communications Development Training Program, giving instruction in semiconductors. Member of American Physical Society, Sigma Xi, Gamma Alpha, and the Institute of Radio Engineers.

ROGER I. WILKINSON, B.S. in E.E., 1924, Prof. E.E., 1950, Iowa State College; Northwestern Bell Telephone Company, 1920-21; American Telephone and Telegraph Company, 1924-34; Bell Telephone Laboratories, 1934-. As a member of the Development and Research Department of the A.T.&T. Co., Mr. Wilkinson specialized in the applications of the mathematical theory of probability to telephone problems.

Since transferring to Bell Telephone Laboratories in 1934, he has continued in the same field of activity and is at present Traffic Studies Engineer responsible for probability studies and traffic research. For two years during World War II, in a civilian capacity, he engaged in operations analysis studies for the Far East Air Forces in the South Pacific, for which he received the Medal for Merit. He has also served as a consultant to the Air Force, the Navy and the Air Navigation Development Board. Mr. Wilkinson is a member of A.I.E.E., American Society for Engineering Education, American Statistical Association, Institute of Mathematical Statistics, Operations Research Society of America, American Society for Quality Control, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi and Pi Mu Epsilon.



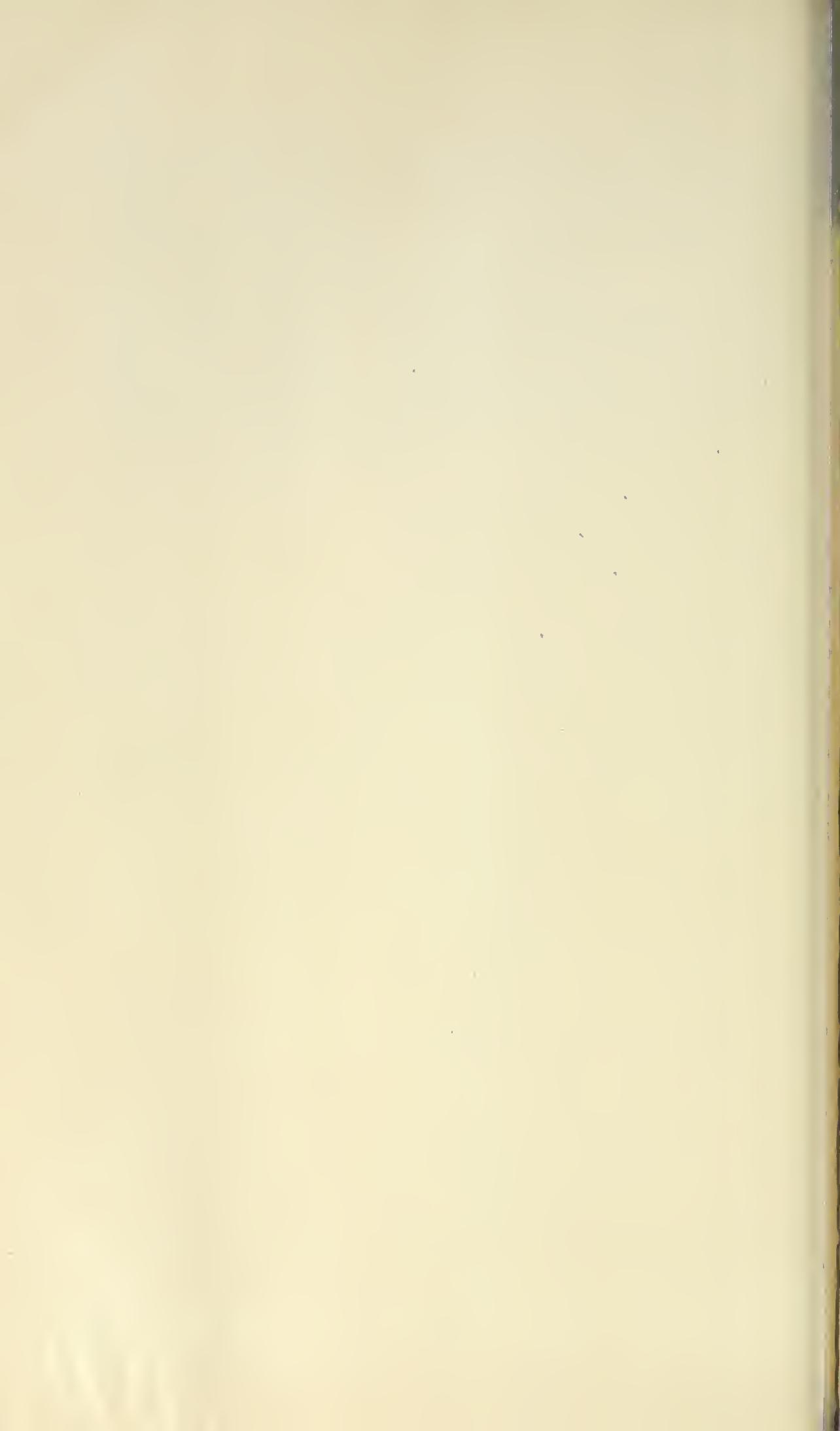




FIG. 25 EQUIVALENT RANDOM LOAD A AND TRUNKS S, FROM NON-RANDOM LOAD A',V'

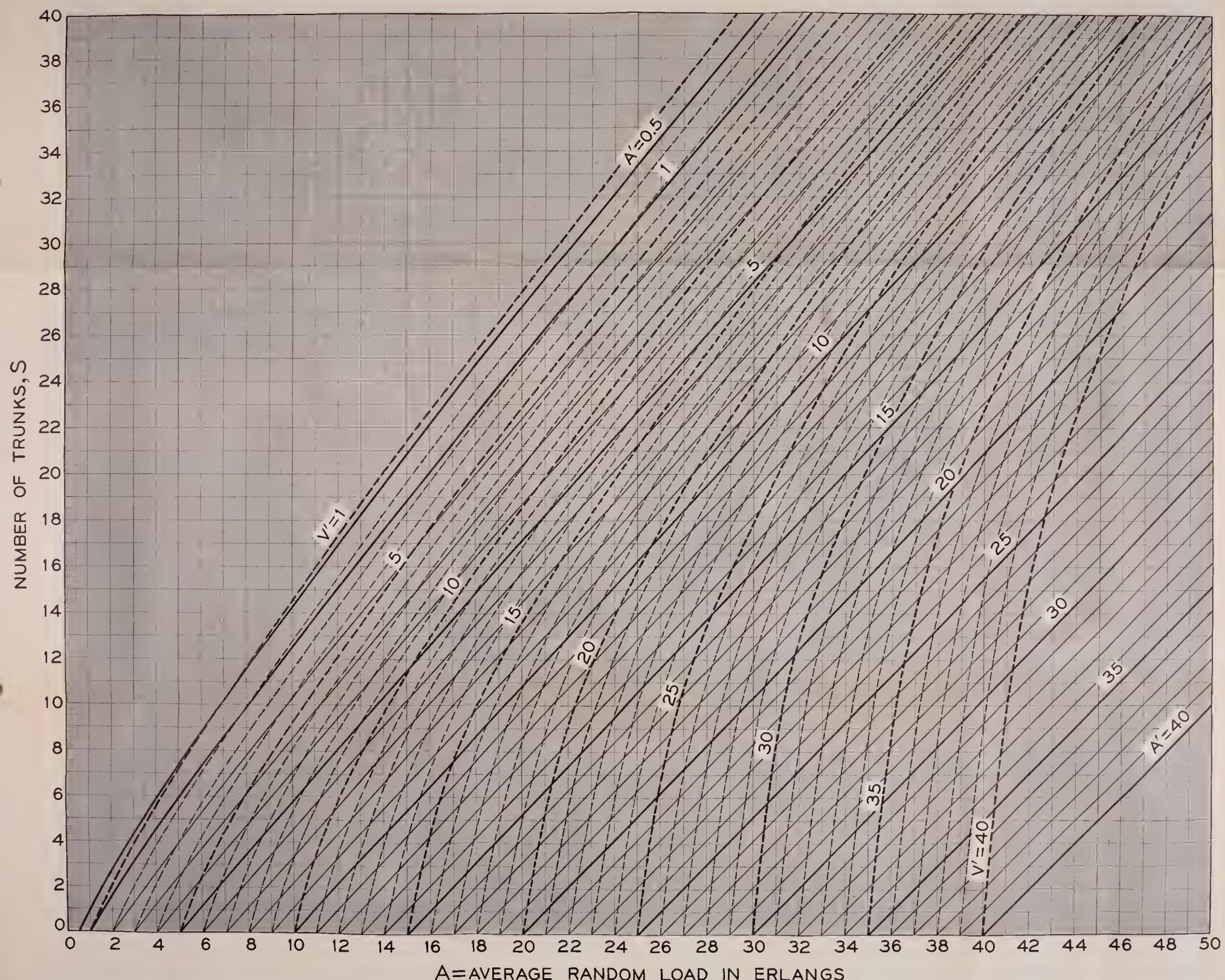
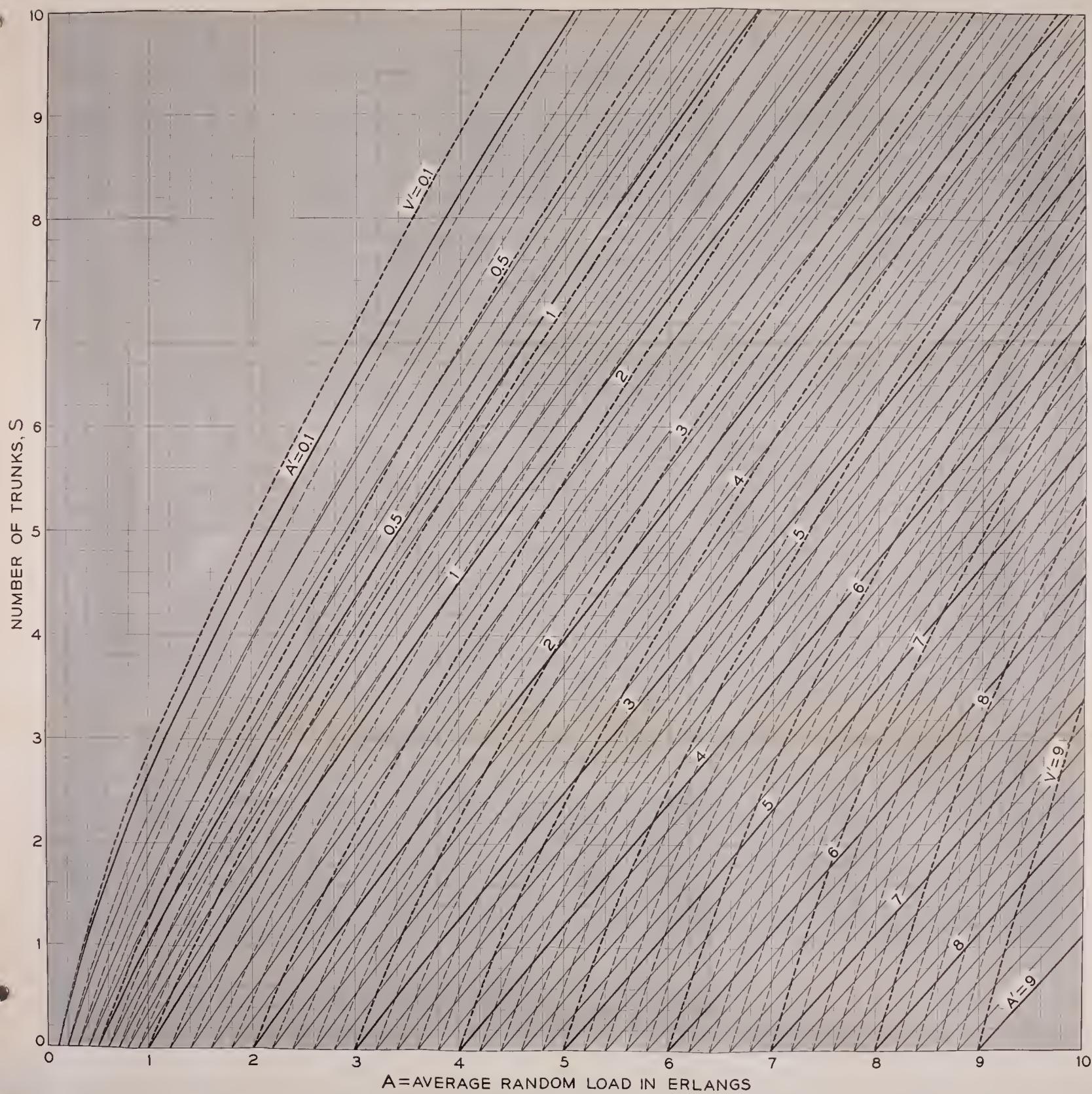
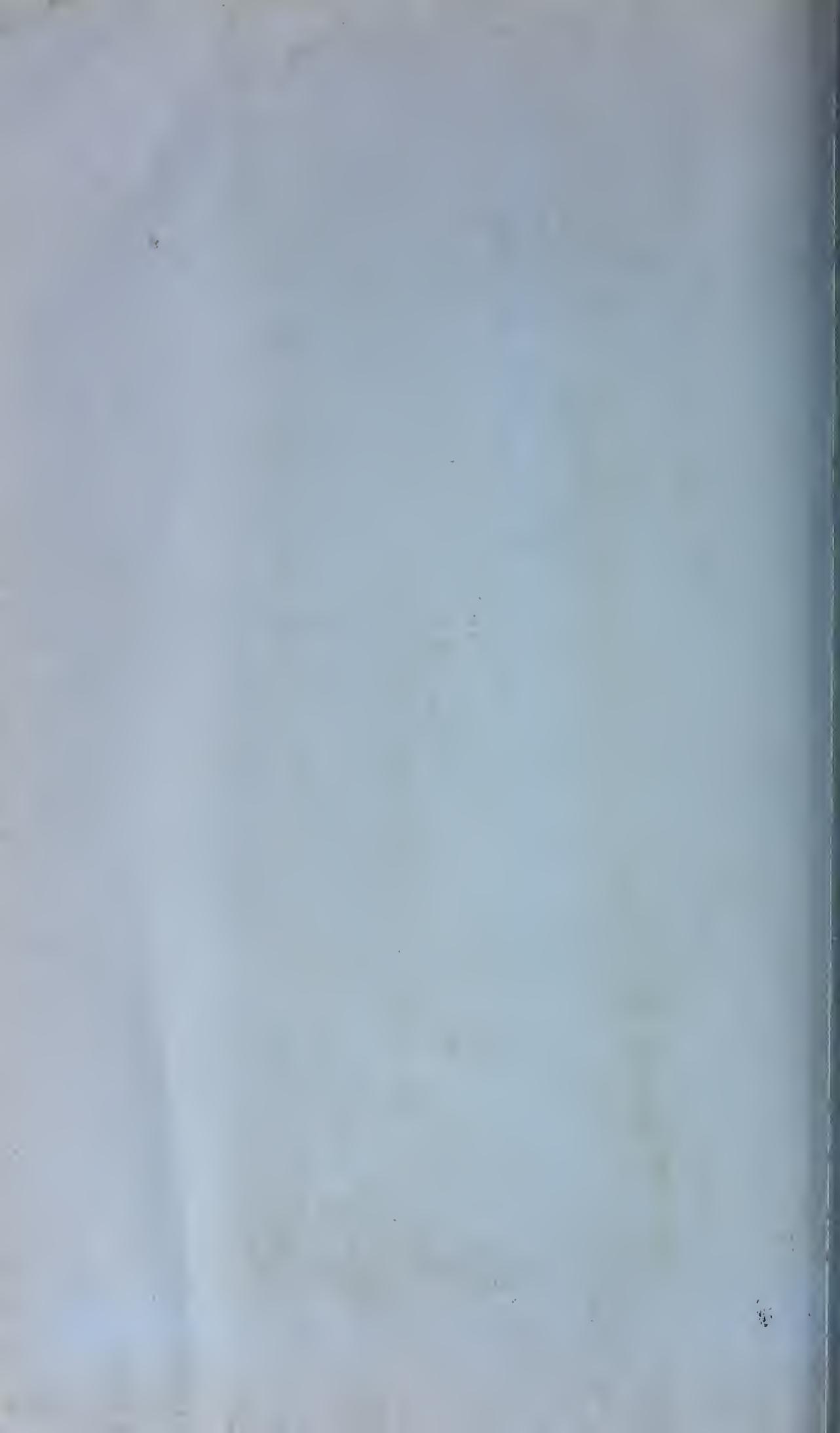


FIG. 26 EQUIVALENT RANDOM LOAD A AND TRUNKS S, FROM NON-RANDOM LOAD A', V'





THE BELL SYSTEM
Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING

PECTS OF ELECTRICAL COMMUNICATION

KANSAS CITY, MO.
PUBLIC LIBRARY

VOLUME XXXV

MAY 1956

JUN 5 1956

NUMBER 3

Chemical Interactions Among Defects in Germanium and Silicon

H. REISS, C. S. FULLER AND F. J. MORIN 535

Single Crystals of Exceptional Perfection and Uniformity by Zone
Leveling

D. C. BENNETT AND B. SAWYER 637

Diffused p-n Junction Silicon Rectifiers

M. B. PRINCE 661

The Forward Characteristic of the PIN Diode

D. A. KLEINMAN 685

A Laboratory Model Magnetic Drum Translator for Toll Switching Offices

F. J. BUHRENDORF, H. A. HENNING AND O. J. MURPHY 707

Tables of Phase of a Semi-Infinite Unit Attenuation Slope

D. E. THOMAS 747

Bell System Technical Papers Not Published in This Journal 751

Recent Bell System Monographs 759

Contributors to This Issue 762

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

B. McMILLAN, *Chairman* R. K. HONAMAN

A. J. BUSCH H. R. HUNTLEY

A. C. DICKIESON F. R. LACK

R. L. DIETZOLD J. R. PIERCE

K. E. GOULD H. V. SCHMIDT

E. I. GREEN G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

MAY 1956

NUMBER 3

Copyright 1956, American Telephone and Telegraph Company

Chemical Interactions Among Defects in Germanium and Silicon

By HOWARD REISS, C. S. FULLER, and F. J. MORIN

Interactions among defects in germanium and silicon have been investigated. The solid solutions involved bear a strong resemblance to aqueous solutions insofar as they represent media for chemical reactions. Such phenomena as acid-base neutralization, complex ion formation, and ion pairing, all take place. These phenomena, besides being of interest in themselves, are useful in studying the properties of the semiconductors in which they occur. The following article is a blend of theory and experiment, and describes developments in this field during the past few years.

CONTENTS

I. Introduction.....	536
II. Electrons and Holes as Chemical Entities.....	537
III. Application of the Mass Action Principle.....	546
IV. Further Applications of the Mass Action Principle.....	550
V. Complex Ion Formation.....	557
VI. Ion Pairing.....	565
VII. Theories of Ion Pairing.....	567
VIII. Phenomena Associated with Ion Pairing in Semiconductors.....	575
IX. Pairing Calculations.....	578
X. Theory of Relaxation.....	582
XI. Investigation of Ion Pairing by Diffusion.....	591

XII. Investigation of Ion Pairing by Its Effect on Carrier Mobility	601
XIII. Relaxation Studies	607
XIV. The Effect of Ion Pairing on Energy Levels	610
XV. Research Possibilities	611
Acknowledgements	613
Appendix A — The Effect of Ion Pairing on Solubility	613
Appendix B — Concentration Dependence of Diffusivity in the Presence of Ion Pairing	617
Appendix C — Solution of Boundary Value Problem for Relaxation	619
Appendix D — Minimization of the Diffusion Potential	623
Appendix E — Calculation of Diffusivities from Conductances of Diffusion Layers	626
Glossary of Symbols	630
References	634

I. INTRODUCTION

The effort of Wagner¹ and his school to bring defects in solids into the domain of chemical reactants has provided a framework within which various abstruse statistical phenomena can be viewed in terms of the intuitive principle of mass action.² Most of the work to date in this field has been performed on oxide and sulfide semiconductors or on ionic compounds such as silver chloride. In these materials the control of defects (impurities are to be regarded as defects) is not all that might be desired, and so with a few exceptions, experiments have been either semiquantitative or even qualitative.

With the emergence of widespread interest in semi-conductors, culminating in the perfection of the transistor, quantities of extremely pure single crystal germanium and silicon have become available. In addition the physical properties, and even the quantum mechanical theory of the behavior of these substances have been widely investigated, so that a great deal of information concerning them exists. Coupled with the fact that defects in them, especially impurities, are particularly susceptible to control, these circumstances render germanium and silicon ideal substances in which to test many of the concepts associated with defect interactions.

This view was adopted at Bell Telephone Laboratories a few years ago when experimental work was first undertaken. Not only has it been possible to demonstrate quantitatively the validity of the mass action principle applied to defects, but new kinds of interactions have been discovered and studied. Furthermore new techniques of measurement have been developed which we feel open the way for broader investigation of a still largely unexplored field.

In fact solids (particularly semiconductors like germanium and silicon)

appear in every respect to provide a medium for chemical reactivity similar to liquids, particularly water. Such phenomena as acid-base reactions, complex ion formation, and electrolyte phenomena such as Debye Hückel effects, ion pairing, etc., all seem to take place.

Besides the experiments theoretical work has been done in an attempt to define the limits of validity of the mass action principle, to furnish more refined electrolyte theories, and most importantly, to provide firm theoretical bases for entirely new phenomena such as ion pair relaxation processes.

The consequence is that the field of diamond lattice³ semiconductors which has previously engaged the special interests of physicists threatens to become important to chemists. Semiconductor crystals are of interest, not only because of the specific chemical processes occurring in these substances, but also because they serve as proving grounds for certain ideas current among chemists, such as electrolyte theory. On the other hand renewed interest is induced on the part of physicists because chemical effects like ion pairing engender new physical effects.

The purpose of this paper is to present the field of defect interaction as it now stands, in a manner intelligible to both physicists and chemists. However, this is not a review paper. Most of the experimental results, and particularly the theories which are fully derived in the text or the appendices are entirely new. Some allusion will be made to published work, particularly to descriptions of the results of some previous theories, in order to round out the development.

The governing theme of the article lies in the analogy between semiconductors and aqueous solutions. This analogy is useful not so much for what it explains, but for the experiments which it suggests. More than once it has stimulated us to new investigations.

In our work we have made extensive use of lithium as an impurity. This is so because lithium can be employed with special ease to demonstrate most of the concepts we have in mind. This specialization should not obscure the fact that other impurities although not well suited to the performance of accurate measurements, will exhibit much of the same behavior.

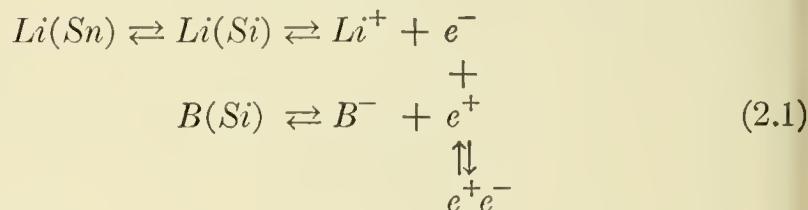
II. ELECTRONS AND HOLES AS CHEMICAL ENTITIES

Since electrons and holes⁴ are obvious occupants of semiconductors like germanium and silicon, and are intimately associated with the presence of donor and acceptor impurities,⁴ it is fitting to inquire into the roles they may play in chemical interactions between donors and ac-

ceptors. This question has been discussed in two papers,^{5, 6} and only its principle aspects will be considered.

To gain perspective it is convenient to consider a system representing the prototype of most systems to be discussed here. Consider a single crystal of silicon containing substitutional boron atoms. Boron, a group III element, is an acceptor, and being substitutional cannot readily diffuse⁷ at temperatures much below the melting point of silicon. If this crystal is immersed in a solution containing lithium, e.g., a solution of lithium in molten tin, lithium will diffuse into it and behave as a donor. Evidence suggests that lithium dissolves interstitially in silicon, thereby accounting for the fact that it possesses a high diffusivity⁸ at a temperature where boron is immobile, for example, below 300°C. When the lithium is uniformly distributed throughout the silicon its solubility in relation to the external phase can be determined. Throughout this process boron remains fixed in the lattice.

If both lithium and boron were inert impurities the solubility of the former would not be expected to depend on the presence or absence of the latter, for the level of solubility is low enough to render (under ordinary circumstances) the solid solution ideal.⁹ On the other hand the impurities exhibit donor and acceptor behaviors respectively, and some unusual effects might exist. We shall first speculate on the simplest possibility in this direction, with the assistance of the set of equilibrium reactions diagrammed below.*



At the left lithium in tin is shown as $Li(Sn)$. It is in reversible equilibrium with $Li(Si)$, un-ionized lithium dissolved in silicon. The latter, in turn, ionizes to yield a positive Li^+ ion and a conduction electron, e^- . Boron, confined to the silicon lattice as $B(Si)$ ionizes as an acceptor to give B^- and a positive hole, e^+ . The conduction electron, e^- , may fall into a valence band hole, e^+ , to form a recombined hole-electron pair, e^+e^- . This process and its reverse are indicated by the vertical equilibrium at the right.

All of the reactions in (2.1), occurring within the silicon crystal are describable in terms of transitions between states in the energy band dia-

* A glossary of symbols is given at the end of this article.

gram of silicon, exhibited in Fig. 1. The conduction band, the valence band, and the forbidden gap are shown. Lithium and boron both introduce localized energy states in the range of forbidden energies. The state for lithium lies just below the bottom of the conduction band while that for boron lies just over the top of the valence band. The separations in energy between most donors or acceptors and their nearest bands are of the order of hundredths of an electron volt while the breadths of the forbidden gaps in germanium or silicon are of the order of one electron volt.

Process 1 in Fig. 1 involving a transition between the donor level and conduction band corresponds to the ionization of lithium in (2.1). Process 2 is the ionization of boron while process 3 represents hole-electron recombination and generation. The various energies of transition are the heats of reaction of the chemical-like changes in (2.1).

Proceeding in the chemists fashion one might argue as follows concerning (2.1). If e^+e^- is a stable compound, as it is at fairly low temperatures, then its formation should exhaust the solution of electrons, forcing the set of lithium equilibria to the right. In this way the presence of boron, supplying holes toward the formation of e^+e^- , increases the solubility of lithium. In fact if e^+ is regarded as the solid state analogue of the hydrogen ion in aqueous solution, and e^- as the counterpart of the hydroxyl ion, then the donor, lithium, may be considered a base while boron, may be considered an acid. Furthermore e^+e^- must correspond to water. Thus the scheme in (2.1) is analogous to a neutralization reaction in which the weakly ionized substance is e^+e^- .

If the immobile boron atoms were replaced by immobile donors, e.g., phosphorus atoms, a reduction, rather than an increase, in the solubility

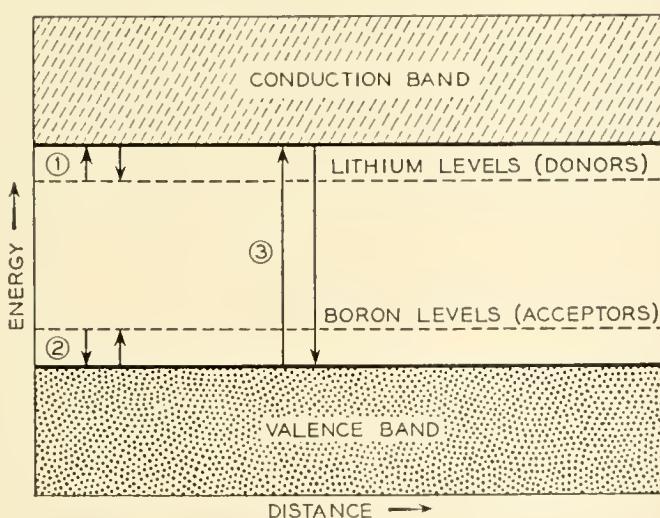


Fig. 1 — Energy band diagram showing the chemical equilibria of (2.1).

of lithium might be expected on the basis of an oversupply of electrons (i.e., by the common ion effect¹⁰). In that case we would have a base displacing another base from solution.

The intimate comparison between this kind of solution and an aqueous solution is worth emphasizing not so much for what it adds to one's understanding of the situation but rather for the further effects it suggests along the lines of analogy. These additional phenomena have been looked for and found, and will be discussed later in this article.

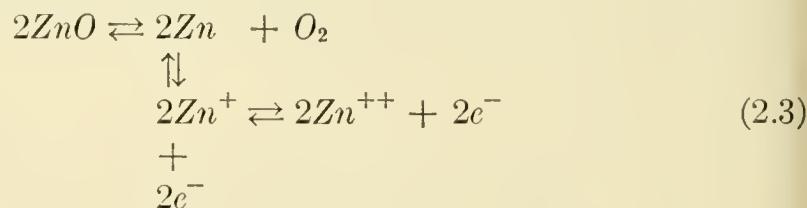
The scheme shown in (2.1) should be applicable, in principle, to other donors and acceptors and to germanium and other semiconductors as well as silicon. Furthermore the external phase may be any one of a suitable variety, and need not even be liquid. Other systems, however, are not as convenient, especially in regard to the ease of equilibration of an impurity over the parts of an heterogeneous system. The lengths to which one can go in comparing electrolytes and semiconductors are discussed in a recent paper.¹¹

In order to quantify the scheme of (2.1) it seems natural to invoke the law of mass action.² Treatments in which holes and electrons are involved in mass action expressions are not new, although systems forming such perfect analogies to aqueous solutions do not seem to have been discussed in the past. For example, in connection with the oxidation of copper Wagner¹² writes



in which \Box^- is a negatively charged cation vacancy in the Cu_2O lattice, and e^+ is a hole. Wagner proceeds to invoke the law of mass action in order to compute the oxygen pressure dependence in this system.

In another example Baumbach and Wagner¹³ and others have investigated oxygen pressure over non-stoichiometric zinc oxide. They consider the possible reactions



and apply the law of mass action. In (2.3) the various states of Zn are presumably interstitial.

Kroger and Vink¹⁴ have recently considered the problem in oxides and sulfides in a rather general way. However in none of the oxide-sulfide systems has it been possible to achieve really quantitative results. In

contrast silicon and germanium offer possibilities of an entirely new order. The advent of the transistor has not only provided large supplies of pure single crystal material, but it has also made available a store of fundamental information concerning the physical properties of these substances. For example, data exists on their energy band diagrams including impurity states — also on resistivity — impurity density curves, diffusivities of impurities, etc. Furthermore, the amount of ionizable impurities can be controlled within narrow limits, and can be changed at will and measured accurately. Consequently it is reasonable to assume that experiments on germanium and silicon will be more successful than similar investigations using other materials.

At this point it is in order to examine whether or not the treatment of electrons and holes as normal chemical entities satisfying the law of mass action is altogether simple and straightforward. This problem has been investigated by Reiss⁵ who found the treatment permissible only as long as the statistics satisfied by holes and electrons remain classical. The validity of this contention can be seen in a very simple manner. Consider a system like that in (2.1). Let the total concentration of donor (ionized and un-ionized) be N_D , the concentration of ionized donor be D^+ , the concentration of conduction electrons be n , and that of valence band holes be p . Let N_A and A^- denote the concentrations of total acceptor and acceptor ions respectively. Finally, let α be the thermodynamic activity¹⁵ of the donor (lithium in (2.1)) in the external phase.

Then, corresponding to the heterogeneous equilibrium in which lithium distributes itself between the two phases we can write

$$\frac{N_D - D^+}{\alpha} = K_0 \quad (2.4)$$

in which K_0 depends on temperature, but not on composition. This assumes the semiconductor to be dilute enough in donor so that the activity of un-ionized donor can be replaced by its concentration, $N_D - D^+$. For the ionization of the donor we can write the mass action relation,

$$\frac{D^+ n}{N_D - D^+} = K_D \quad (2.5)$$

and for the acceptor,

$$\frac{A^- p}{N_A - A^-} = K_A \quad (2.6)$$

while for the electron-hole recombination equilibrium

$$np = K_1 \quad (2.7)$$

In (2.5), (2.6), and (2.7) all the K 's are independent of composition. To these equations is added the charge neutrality condition,

$$D^+ + p = A^- + n \quad (2.8)$$

Equations (2.4) through (2.8) are enough to determine N_D in its dependence on N_A , α , and the various K 's. Together they represent the mass action approach. To demonstrate their validity it is necessary to appeal to statistical considerations.

Thus $N_D - D^+$, the concentration of un-ionized donor is really the density of electrons in the donor level of the energy diagram for the semiconductor. According to Fermi statistics this density is given by⁵

$$N_D - D^+ = N_D / \{1 + \frac{1}{2} \exp [(E_D - F)/kT]\} \quad (2.9)$$

in which E_D is the energy of the donor level, F is the Fermi level,¹⁶ k , the Boltzmann constant, and T , the temperature. Furthermore, according to Fermi statistics, n , the total density of electrons in the conduction band is

$$n = \sum_i g_i / \{1 + \exp [(E_i - F)/kT]\} \quad (2.10)$$

where g_i is the density of levels of energy, E_i , in the conduction band, and the sum extends over all states in that band. Similar expressions are available for the occupation of the acceptor level and the valence band. F is usually determined by summing over all expressions like (2.9) and (2.10) and equating the result to the total number of electrons in the system. This operation corresponds exactly to applying the conservation condition, (2.8). It is obvious from the manner of its determination that F depends upon $N_D - D^+$, n , etc.

If we now form the expression on the left of (2.5) by substituting for each factor in it from (2.9) and (2.10), it is obvious that the result depends in a very complicated fashion upon F , and so cannot be the constant, K_D , independent of composition, since in the last paragraph F was shown to depend on composition. On the other hand if attention is confined to the limit in which classical statistics apply¹⁷ the unities in the denominators of (2.9) and (2.10) can be disregarded in comparison to the exponentials, and those equations become

$$N_D - D^+ = 2N_D e^{F/kT} e^{-E_D/kT} \quad (2.11)$$

and

$$n = e^{F/kT} \sum_i g_i e^{-E_i/kT} \quad (2.12)$$

respectively. Moreover, from (2.11)

$$D^+ = N_D [1 - 2e^{F/kT} e^{-E_D/kT}] = N_D \quad (2.13)$$

where the second term in brackets is ignored for the same reason as unity in the denominators of (2.9) and (2.10). Substituting (2.11) through (2.13) into (2.5) yields

$$\frac{D^+ n}{N_D - D^+} = \frac{\sum_i g_i e^{-E_i/kT}}{2e^{-E_D/kT}} \quad (2.14)$$

in which the right side is truly independent of composition, since F has cancelled out of the expression. Similar arguments hold for (2.6) and (2.7). Therefore in the classical limit the law of mass action is valid, at least insofar as internal equilibria are concerned.

We have next to examine the validity of (2.4) which is really the law of mass action applied to the heterogeneous equilibrium between phases. Substitution of (2.11) into (2.4) leads to the prediction

$$\alpha = \frac{2e^{-E_D/kT}}{K_0} \{e^{F/kT}\} N_D = K \{e^{F/kT}\} N_D \quad (2.15)$$

in the classical case, if (2.4) is valid. In order to confirm (2.15) it is necessary to evaluate the chemical potentials¹⁸ of the donor in the external phase and in the semiconductor, and equate the two. The resulting expression should be equivalent to (2.15).

Since α is the activity of the donor in the external phase its chemical potential in that phase is, by definition,

$$\mu = \mu^0(T, p) + kT \ln \alpha \quad (2.16)$$

where μ^0 , the chemical potential in the standard state, may depend on temperature and pressure, but not on composition. To compute the chemical potential in the semiconductor statistical methods must once more be invoked. Thus, according to (2.13), donor atoms are nearly totally ionized in the classical case, so that the addition of a donor atom to the semiconductor amounts to addition of two separate particles, the donor ion and the electron. The chemical potential of the added atom is therefore the sum of the potentials of the ion and the electron separately. Since the ions are supposedly present in low concentration the latter can serve as an activity,¹⁹ and in analogy to (2.16) we obtain for the ionic chemical potential

$$\mu_{D^+} = \mu_{D^+}^0(T, p) + kT \ln D^+ \quad (2.17)$$

Furthermore, it is well established²⁰ that the Fermi level plays the role of chemical potential, μ_e , for the electron

$$\mu_e = F \quad (2.18)$$

Thus the chemical potential for the donor atom is

$$\begin{aligned} \mu_D &= \mu_{D^+} + \mu_e = \mu_{D^+}^0 + kT \ln D^+ + F \\ &= \mu_{D^+}^0 + kT \ln N_D + F = \mu_{D^+}^0 + kT \ln \{e^{F/kT}\} N_D \end{aligned} \quad (2.19)$$

where (2.13) has been used to replace D^+ by N_D . We note that the activity of the donor atom must be

$$\{e^{F/kT}\} N_D \quad (2.20)$$

with $e^{F/kT}$ playing the role of an activity coefficient.²¹

Equating μ_D given by (2.19) to μ in (2.16) results in the equation

$$\alpha = \exp[(\mu_{D^+}^0 - \mu^0)/kT] \{e^{F/kT}\} N_D \quad (2.21)$$

which can be made identical to (2.15) by identifying

$$\exp[(\mu_{D^+}^0 - \mu^0)/kT]$$

with K of that expression. Thus in the classical case the law of mass action is applicable to the heterogeneous equilibrium.

When classical statistics no longer apply it is still possible to evaluate $N_D - D^+$, using the full expression (2.9). Therefore the solubility N_D , of the donor can still be determined if (2.4) remains valid. To decide this question it is necessary to evaluate μ_D , the chemical potential of the donor in the semiconductor under non-classical conditions. This problem is not as simple as those treated above, but it can be solved, and the detailed arguments can be found in Reference 5. Here we shall be content with quoting the results. However, before doing this the non-classical counterpart of (2.15) will be written by combining (2.9) with (2.4). The result is

$$\alpha = [K_0/\{1 + \frac{1}{2} \exp[(E_D - F)/kT]\}] N_D \quad (2.22)$$

and if (2.4) is valid (2.22) should be derivable by equating μ to the proper value of μ_D .

Since in the non-classical case a finite portion of the donor states are occupied by electrons, the introduction of an additional *average* donor atom is no longer equivalent to adding two independent particles whose chemical potentials can be summed. In the statistical derivation of μ_L it is therefore necessary to evaluate the total free energy of the semi-

conductor phase, and to differentiate this with respect to N_D , keeping temperature and pressure fixed.* The result is

$$\begin{aligned}\mu_D = & \mu_{D^+}^0 + kT \ln N_D \\ & + F - kT \ln \{1 + 2 \exp[-(E_D - F)/kT]\}\end{aligned}\quad (2.23)$$

in which it has been assumed that the concentration of impurity is sufficiently low so that the solution would be ideal if the impurity could not ionize. In the classical case the exponential in the logarithm is small compared to unity and (2.23) becomes identical with (2.19), as it should. In the totally degenerate case the exponential dominates the unity and we have

$$\begin{aligned}\mu_D = & \{\mu_{D^+}^0 + E_D - kT \ln 2\} + kT \ln N_D \\ = & \mu_D^0 + kT \ln N_D\end{aligned}\quad (2.24)$$

which is the chemical potential of an un-ionized component of a dilute

* An interesting by-product of this derivation (discussed in Reference 5) is the fact that the Fermi level, F , is hardly ever the Gibbs free energy per electron for the electron assembly, although it is always the electronic chemical potential, in the sense that it measures the direction of flow of electrons. This arises because the Gibbs free energy is not always a homogeneous function²² of the first degree in the mole numbers (electron numbers). Thus if the number of electrons in the assembly is N , the Gibbs free energy, G , is given by

$$G = NF + kT \sum_i \left[V \left\{ \frac{\partial \omega_i}{\partial V} \right\}_{T,N} - \omega_i \right] \ln \frac{\omega_i}{h_i}$$

where the sum is over all energy levels, j , referred to an invariant standard level. V is the volume of the system, ω_i is the total number of states at the j th level, and h_i is the number of unoccupied states (holes) at the j th level. For F to be the free energy per electron the term involving the sum must vanish so that

$$F = \frac{G}{N}$$

But this can only happen when

$$\omega_i = K_i V$$

where K_i is independent of V . This requirement is formally met in the case of the free electron gas where the electrons have been treated as independent particles in a box so that

$$\omega_i = [8m_0^{3/2} \pi E dE/2h^3]V$$

where m_0 is the electron mass, and h , Plank's constant. Since this is the case most frequently dealt with in thermodynamic problems it has been customary to think of F as the free energy per electron, although even here the truth of the contention depends on the assumption of particle in the box behavior.

At the other extreme, it is obvious that ω_i for a level corresponding to the deep closed shell states of the atoms forming a solid cannot depend at all on the external volume since they are essentially localized. In computing the free energy of the semiconductor phase it is necessary to understand carefully subtleties of this nature.

solution, as it should be for the degenerate case in which ionization is suppressed. Equating μ_D in (2.23) to μ in (2.16) yields

$$\alpha = \left\{ \frac{\frac{1}{2} \exp [(\mu_{D^+}^0 - \mu^0 + E_D)/kT]}{1 + \frac{1}{2} \exp [(E - F)/kT]} \right\} N_D \quad (2.25)$$

which is identical with (2.22) if K_0 is taken to be

$$\frac{1}{2} \exp [(\mu_{D^+}^0 - \mu^0 + E_D)/kT] \quad (2.26)$$

Thus one arrives at the conclusion that the law of mass action remains valid for the heterogeneous equilibrium even when it fails for the homogeneous internal equilibria.

This is a fairly important result since it implies that solubilities can give information on the behavior of the Fermi level and hence on the distribution of electronic energy levels, even under conditions of degeneracy.

The chemical potential specified by (2.23) is of course important in itself, for treating any equilibrium (external or internal) in which the donor may participate.

One last remark is in order. This concerns the treatment of heterogeneous equilibria involving some external phase, and the surface²³ rather than the body of a semiconductor. In such treatments it has been customary to compute the chemical potential of an ionizable adsorbed atom by summing the ion chemical potential and the Fermi level, as in (2.19). This is no more possible if the statistics of the surface states are non-classical, then it is possible when considering non-classical situations involving the body of the crystal. Care must therefore be exercised also in the treatment of surface equilibria.

The above discussion has shown that there are extensive ranges of conditions under which holes and electrons obey the law of mass action, and behave like chemical entities. In the next section some of the consequences of this fact will be developed.

III. APPLICATION OF THE MASS ACTION PRINCIPLE

Equations (2.4) through (2.8) will now be used to determine how, in the classical case, the solubility, N_D , of lithium in (2.1) depends upon N_A the concentration of boron in silicon. In the experiments to be described, the systems are classical, and the donors and acceptors therefore so thoroughly ionized that N_D can be replaced by D^+ and N_A by A^- . Insertion of (2.4) into (2.5) yields

$$D^+ n = \alpha K_D K_0 = K^* \quad (3.1)$$

since α is maintained constant. Furthermore (2.7) can be written as

$$np = K_1 = n_i^2 \quad (3.2)$$

where n_i is obviously the concentration of holes or electrons under the condition that the two are equal. It is called the intrinsic concentration²⁴ of holes or electrons. The values of n_i in germanium and silicon have been determined by Morin.^{25, 26} Fig. 2 gives plots of the logarithms of n_i in germanium and silicon versus the reciprocals of temperature. These results are necessary for subsequent calculations.

Since N_A and A^- are assumed equal, we may dispense with (2.6). The one remaining equation is then (2.8) which we adopt unchanged. These three relations, (3.1), (3.2), and (2.8) are sufficient to determine D^+ or N_D as a function of A^- or N_A . The only undetermined parameter in the set is K^* and this can be evaluated by measuring the solubility, D^+ , in the absence of acceptor, i.e., under the condition that A^- is zero. The symbol D_0^+ is used to designate this value of D^+ . In Reference 6 it is shown that

$$D_0^+ = K^*/(K^* + n_i^2)^{1/2}$$

or

$$K^* = (D_0^+)^2/2 + \{(D_0^+)^4/4 + n_i^2(D_0^+)^2\}^{1/2} \quad (3.3)$$

Eliminating K^* by the use of this relation it is further shown in Reference 6 that

$$D^+ = \frac{A^-}{1 + \sqrt{1 + (2n_i/D_0^+)^2}} + \left\{ \left[\frac{A^-}{1 + \sqrt{1 + (2n_i/D_0^+)^2}} \right]^2 + (D_0^+)^2 \right\}^{1/2} \quad (3.4)$$

which is the required relation between donor solubility and acceptor concentration.

Examination of (3.4) reveals several simple features, the more important of which we list below:

(1) When A^- (the acceptor doping) is sufficiently large so that $(D_0^+)^2$ in the second term can be ignored relative to the term in A^- , (3.4) reduces to that of a straight line with slope

$$D^+/A^- = \frac{2}{1 + \sqrt{1 + (2n_i/D_0^+)^2}} \quad (3.5)$$

Knowledge of this slope is equivalent to knowledge of D_0^+ .

(2) Where the straight line portion of the D^+ versus A^- curve is in-

volved, the temperature dependence of the solubility, D^+ , enters only through the ratio, n_i/D_0^+ . If this ratio is very small, then

$$D^+ \approx A^- \quad (3.6)$$

and the solubility is independent of temperature. In this condition D^+ may approximate A^- by being either slightly less or slightly greater than the latter. Details are given in Reference 6.

(3) Whereas D^+ at small values of doping may be an increasing function of temperature, it may, depending on the system, be a decreasing function of temperature at high dopings. Thus doping may change the sign of the temperature coefficient of solubility. Because of this, doping sometimes may prevent precipitation of a donor when a semiconductor is cooled, since the latter becomes an undersaturated rather than a supersaturated solution of impurity. Details are given in Reference 6.

(4) It is also shown in Reference 6 that for the acceptor to have any effect on the solubility of the donor the concentration of A^- should satisfy the following criterion

$$A^- > (D_0^+ \quad \text{or} \quad n_i) \quad (3.7)$$

D_0^+ or n_i being used depending on which is greater. Obviously at high

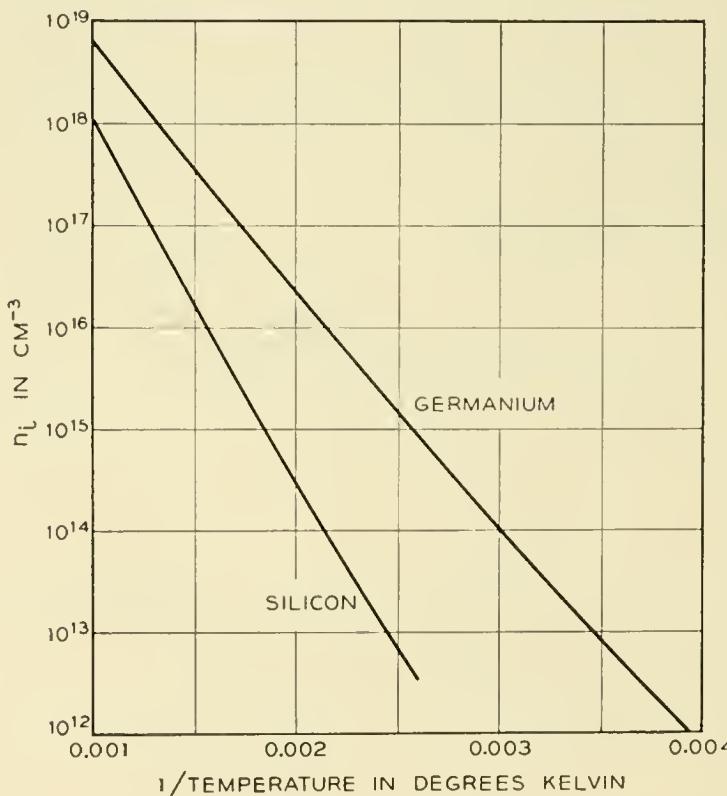


Fig. 2 — Temperature dependences of intrinsic carrier concentrations in germanium and silicon.

temperatures when n_i achieves a very large value it may not be possible to have A^- exceed n_i , and no effect due to the acceptor will be observable. This is simply a mathematical reflection of the fact that the hypothetical compound e^+e^- in (2.1) is highly dissociated at high temperatures so that the holes contributed by the acceptor cannot cause the exhaustion of electrons in the solution.

In Reference 6 the system described in (2.1) was investigated for the purpose of testing (3.4). The concentrations, D^+ and A^- , of lithium and boron respectively were determined by measuring the electrical resistivities of the crystal specimens before and after immersion in molten tin containing lithium. Some typical results of these experiments are shown in Fig. 3 which contains three D^+ versus A^- isotherms for the temperatures 249° , 310° , and 404°C . For the case shown the tin phase contained 0.18 per cent lithium by weight.

The points in the figure represent experimental findings, while the drawn curves are based on theory. The agreement between theory and experiment is very good, in fact the overall accuracy appears to be better than 1 per cent. These isotherms are only a few of a large group obtained at different temperatures and with differently proportioned external phases. The accuracy in all of these is of the same order.

Various of the features of (3.4) listed above are apparent in the curves of Fig. 3. For example at large values of A^- the curves are straight lines, thus validating (3.5). Also, the inversion of the temperature coefficient of solubility with doping is apparent for the curves cross one another, and whereas, at low dopings (low A^-) the solubility is an increasing function of temperature, at high dopings it decreases with increasing temperature. Finally we note that D^+ remains more or less independent of A^- until A^- exceeds n_i , confirming (3.7). Values of n_i appear in the Figure.

The possible increases in solubility above D_0^+ are really quite large. For example in Fig. 3 the largest increase is of the order of a factor of 10^3 . However in some experiments increases of 10^6 have been observed. These effects truly represent profound interactions between impurities which are present in highly attenuated form. Thus the number of atoms per cubic centimeter in crystal silicon is of the order of $5 \times 10^{22} \text{ cm}^{-3}$. Interactions at doping levels as low as 10^{15} cm^{-3} , as appear in Fig. 3, therefore take place at atom fraction levels of about 2×10^{-8} .

In Fig. 4 we show a curve of lithium solubility at room temperature in gallium-doped germanium. The curve is wholly experimental; no attempt has been made to apply theory. The symbols D^+ and A^- are once more used for the donor and acceptor. In this case the curve again exhibits some of the general features required by (3.4). The measure-

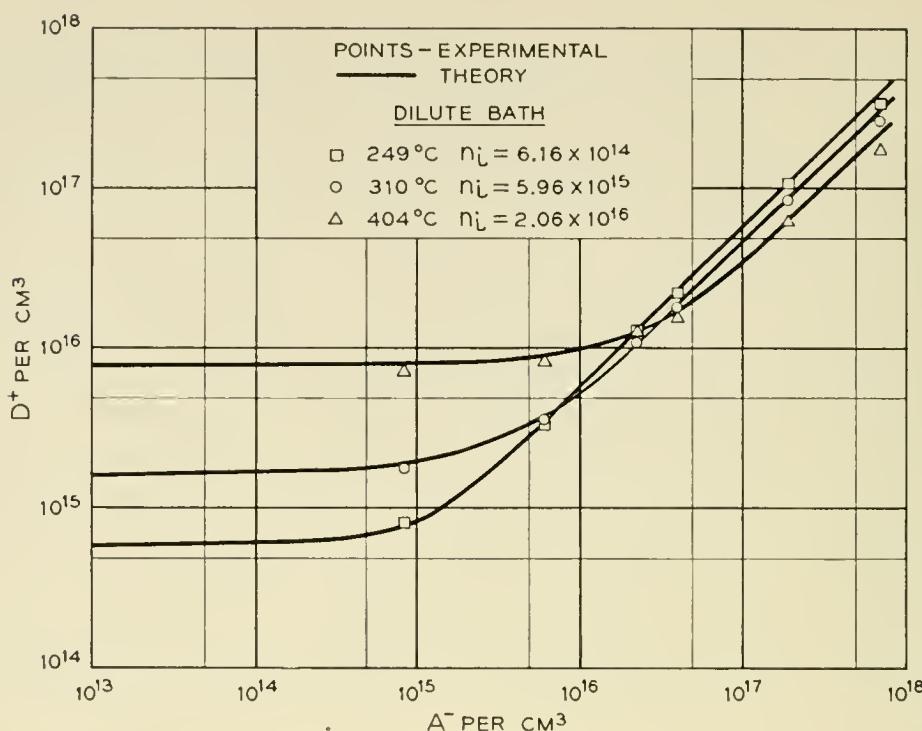


Fig. 3 — Isotherms showing the solubility of lithium D^+ , in silicon as a function of boron doping A^- , for an external phase of tin containing 0.18 per cent lithium.

ments were made by saturating gallium-doped germanium crystals with lithium by alloying lithium to the germanium surface at a high temperature, and letting it diffuse in. Following this the crystals were cooled and lithium was allowed to precipitate to equilibrium. In this case the external solution is the precipitate and is of unknown composition.

If the straight line portion of the curve is used to determine D^+/A^- appearing in (3.5), the value of D_0^+ associated with the precipitate as an external phase can be computed by using the value of n_i obtained from Fig. 2 for 25°C. The latter is $3 \times 10^{13} \text{ cm}^{-3}$, and the measured D^+/A^- is 0.85. Application of (3.5) then leads to a value of D_0^+ of $6.6 \times 10^{13} \text{ cm}^{-3}$ at 25°C. Since the highest value of D^+ measured in Fig. 4 is $5.5 \times 10^{18} \text{ cm}^{-3}$, the solubility increase here shows a factor of 10^5 . Interaction is already apparent at values of A^- as low as 10^{14} cm^{-3} , and since there are $4.4 \times 10^{22} \text{ cm}^{-3}$ atoms per cubic centimeter in pure germanium this represents interaction at levels of atom fraction as low as 2×10^{-9} .

IV. FURTHER APPLICATIONS OF THE MASS ACTION PRINCIPLE

In the last section the possibility was mentioned of inverting the sign of the temperature coefficient of solubility, and so preventing impurity

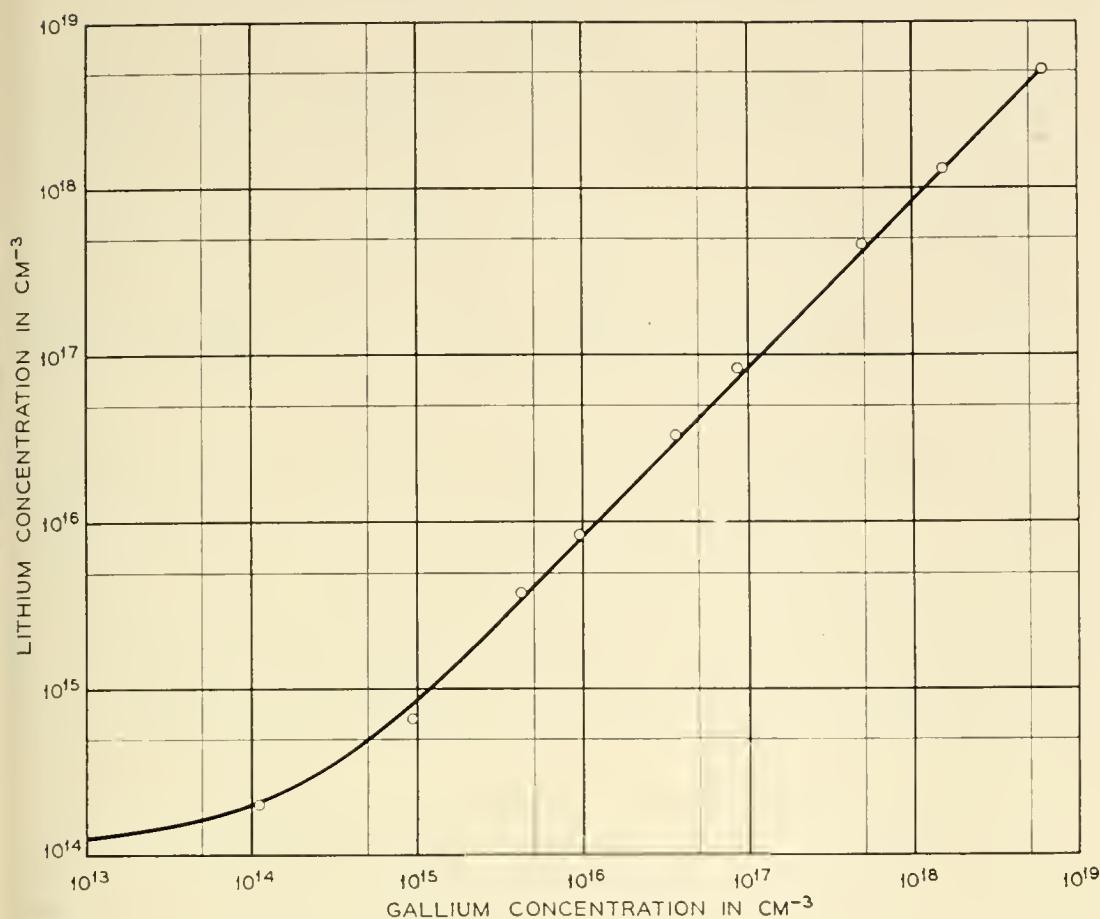


Fig. 4 — Room temperature isotherm showing the solubility of lithium in germanium as a function of gallium doping, the external phase being an alloy of lithium and germanium. The curve merely shows locus of experimental points.

precipitation which might normally occur upon cooling a crystal specimen. An experiment demonstrating this effect is described in Reference 6. Two specimens of germanium, one without added acceptor, and the other containing gallium at an estimated concentration of $1.3 \times 10^{19} \text{ cm}^{-3}$, were saturated with lithium. Table I compares the changes in lithium content observed in these samples with the passage of time. After 25 days no apparent precipitation had occurred in the gallium doped specimen, while precipitation was almost complete in the other.

This result suggests a practical scheme for measuring the concentration of lithium along the solidus curve of the lithium-germanium phase diagram, i.e., the solubility of lithium in solid germanium when the external phase is also composed of germanium and lithium, and probably represents the liquidus phase. This measurement, though desirable, has not been performed before because lithium, diffused into germanium at an elevated temperature, precipitates when the specimen is cooled.

TABLE I

Ga Conc. (cm^{-3})	Li Conc. after saturation (cm^{-3})	Li Conc. after 4 days at room Temp. (cm^{-3})	Li Conc. after 25 days at room Temp. (cm^{-3})
0	1.4×10^{16}	9.0×10^{15}	1.1×10^{15}
1.3×10^{19}	8.0×10^{18}	8.0×10^{18}	8.0×10^{18}

Resistivities then measure only the dissolved lithium although the true solubility at the temperature of saturation includes the precipitated material.

However, we have seen that germanium suitably doped with gallium will not lose lithium by precipitation. Therefore the experiment might be performed in doped germanium. The only difficulty with this suggestion lies in the fact that *doping changes the solubility*. This objection can be overcome through use of (3.4). In terms of that equation D^+ would be measured in the presence of gallium whereas D_0^+ , the solubility in undoped germanium, is required. But according to (3.4) if D^+ , n_i , and A^- (gallium concentration) are known D_0^+ can be computed. In fact solving (3.4) for D_0^+ yields

$$D_0^+ = \frac{\frac{D^+(D^+ - A^-)}{2} + \sqrt{\left[\frac{D^+(D^+ - A^-)}{2}\right]^2 + (D^+)^2 n_i^2}}{\sqrt{n_i^2 + \frac{D^+(D^+ - A^-)}{2} + \sqrt{\left[\frac{D^+(D^+ - A^-)}{2}\right]^2 + (D^+)^2 n_i^2}}} \quad (4.1)$$

The plan is therefore self-evident. Samples of germanium of known suitable gallium contents A^- are to be saturated with lithium at various temperatures. If a judicious choice of gallium content is made the lithium will not precipitate when the specimen is cooled. Therefore the value of D^+ characteristic of the saturation temperature can be determined through resistivity measurements performed at room temperature. Taking n_i from Fig. 2 it then becomes possible to calculate D_0^+ using (4.1).

The crystal specimens employed were cut in the form of small rectangular wafers of dimensions, approximately $1 \text{ cm} \times 0.4 \text{ cm} \times 0.1 \text{ cm}$. On the surfaces of these, small filings of lithium were distributed densely enough so that their average separation was less than the half thickness of the specimen's smallest dimension. The filings were alloyed to the germanium specimen by heating in dry helium for 30 seconds at 530°C . Then the crystals were permitted to saturate with lithium by diffusion from the alloy at some chosen lower temperature. After the period of saturation which ranged from one half hour to as long as 168 days, de-

TABLE II

T °C.	ρ_0 ohm cm	A^- (cm $^{-3}$)	ρ ohm (cm)	D^+ cm $^{-3}$	D_0^+ (cm $^{-3}$)
25					6.6×10^{13}
100	0.0523	2.2×10^{17}	0.0735	$.9 \times 10^{16}$	2.5×10^{14}
200	0.44	1.3×10^{16}	0.90	7.8×10^{15}	4.6×10^{15}
250	0.1494	4.7×10^{16}	0.652	3.9×10^{16}	2.6×10^{16}
300	0.042	2.9×10^{17}	0.108	2.15×10^{17}	7.3×10^{16}
500	0.00614	4.5×10^{18}	0.0340	4.13×10^{18}	1.7×10^{18}
608	0.00577	5.0×10^{18}	0.049	4.78×10^{18}	2.8×10^{18}
650	0.00584	4.3×10^{18}	0.0178	3.75×10^{18}	2.4×10^{18}

pending on the temperature, the specimen surface was lapped smooth with carborundum paper. Resistivities were then measured by means of a two point probe.

Table II collects the data showing T , the temperature of saturation in degrees centigrade, ρ_0 the resistivity before saturation, A^- the gallium concentration computed from ρ_0 , ρ the resistivity after saturation, and D^+ the lithium concentration computed from ρ . The final column shows D_0^+ computed using (4.1) and Fig. 2.

In Table II the 25°C value of D_0^+ has been taken as the value computed in section III in connection with Fig. 4. It might be thought (in view of a later section in this paper) that the 25° and 100°C values of D_0^+ are not as reliable as the others because at the low temperatures involved the solubility of lithium may be influenced by ion pairing as well as electron-hole equilibria. However, Appendix A shows that the possible error is small.

In Fig. 5 D_0^+ is plotted against temperature using these data. The plot is the curve labeled $Ga^- = 0$, and the open circles were obtained by inserting the measured D^+ values (crosses) into (4.1). We notice that the curve has a maximum in the neighborhood of 600°C. The occurrence of a maximum, is a necessity if D_0^+ is to pass to zero, as it must at the melting point of germanium. It is also worth noticing that D_0^+ near room temperature lies in the range of order 10^{13} cm $^{-3}$, but that its measurement has been effected at concentrations as high as 10^{18} cm $^{-3}$. This illustrates another application of the electron-hole equilibrium, namely in the determination of solubilities.

With D_0^+ in our possession it is interesting to return to (3.4) and to calculate D^+ as a function of temperature for various levels of A^- . This has been done for values of A^- equal to 10^{15} , 10^{16} , 10^{17} , and 10^{18} cm $^{-3}$. The curves so obtained appear in Fig. 5, labeled $Ga^- = 10^{15}$, 10^{16} , 10^{17} , 10^{18} cm $^{-3}$, respectively. Their most striking common feature is the minimum which appears below 200°C. This minimum introduces a new prob-

lem in preparing samples without precipitate. Thus consider the $A^- = 10^{16} \text{ cm}^{-3}$ curve. Suppose the specimen is saturated at 200°C . Then according to Fig. 5, if A^- for the specimen is 10^{16} cm^{-3} , D^+ after saturation will be $7 \times 10^{15} \text{ cm}^{-3}$. However, as the sample is cooled it will tend, at first, to become supersaturated. For example it will achieve its maximum supersaturation at about 140°C , where the minimum of the 10^{16} cm^{-3} curve appears. Thereafter it will return to its undersaturated state. In fact at 25°C a concentration of $9.3 \times 10^{15} \text{ cm}^{-3}$ could be supported, whereas the solution contains no more than $7 \times 10^{15} \text{ cm}^{-3}$ lithium atoms. Some of these may have precipitated as the cooling process passed through the minimum, so that sufficient time must be provided for the process of re-solution.

If the original saturation had taken place at 250°C , the concentration

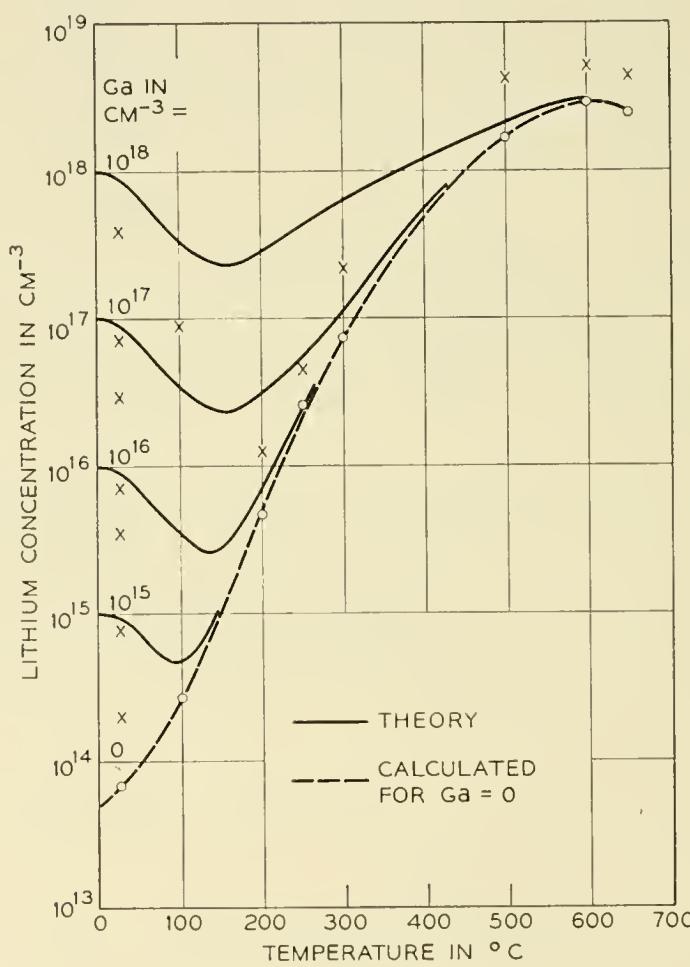


Fig. 5 — Solubility of lithium in germanium as a function of temperature for various gallium dopings. The external phase is an alloy of lithium and germanium. The broken line is the locus of the points (circles) calculated from equation (4.1) for zero gallium concentration. The values of A^- and D^+ , used in applying (4.1), correspond to the points shown by \times in the illustration. See Table II.

of lithium would have been $2.4 \times 10^{16} \text{ cm}^{-3}$. Since this exceeds the $9.3 \times 10^{15} \text{ cm}^{-3}$ supportable at 25°C, such a sample would have contained some precipitate. It was important to avoid these various pitfalls in preparing the specimens used in the above study. Care was taken to insure that this was the case.

We now turn to another application of the electron-hole equilibrium. It has been emphasized that just as a fixed acceptor will increase the solubility of lithium in silicon, a fixed donor should decrease it. In fact in a crystal containing a p-n junction²⁷ the solubility should be above normal on the *p* side and below normal on the *n* side. The built-in field²⁸ which exists at the junction is a reflection of this difference in solubility, for if it were not present the concentration gradient created by the disparity in solubilities would cause the lithium to diffuse from the *p* to the *n* side until its concentration was uniform throughout the crystal. Obviously this field is in such a direction as to cause lithium ions to move back to the *p* side.*

Now in both silicon and germanium the oxide layers on the surface can react readily with dissolved lithium. As a result the surface behaves as a sink, and at temperatures as low as room temperature lithium is lost to the surface from the body of the crystal. At higher temperatures the body of the crystal can be exhausted of lithium in a few minutes. There are many experiments which one would like to perform in which the crystal must be maintained without loss of lithium at an elevated temperature for long periods of time.

The application now to be discussed involves utilization of the built-in field at a p-n junction to prevent lithium from reaching the surface where

* The distribution of lithium in the space charge region of a p-n junction cannot be computed by the methods advanced thus far. This is because the charge neutrality condition (2.8) is no longer valid. Instead the concentration of lithium is determined by Boltzmann's law,²⁹ and is given by

$$D^+ = D_*^+ \exp [-qV/kT]$$

where q is the charge on a lithium ion, V is the local electrostatic potential, and D_*^+ is the concentration where V is zero.

V itself must be determined from Poisson's equation³⁰

$$\nabla^2 V = -\frac{4\pi\rho}{\kappa}$$

where ρ is the local charge density and κ is the dielectric constant of the medium. In semiconductors ρ is given in terms of V by³¹

$$\begin{aligned} \rho &= q[H + D^+ - 2n_i \sinh (qV/kT)] \\ &= q[H + D_*^+ \exp [-qV/kT] - 2n_i \sinh (qV/kT)] \end{aligned}$$

where H is the local density of fixed donors less the local density of fixed acceptors.

it can attack the oxide. Two specimens of 0.34 ohm cm *p*-type silicon doped with boron were cut from adjacent parts of a crystal. Each specimen was about 1 cm long, 0.2 cm wide, and 0.15-cm thick. The samples were lapped on No. 400 silicon carbide paper, etched in HF and HNO₃ and sealed in helium-flushed evacuated quartz tubes, one containing a small grain of P₂O₅. The tubes were then heated at 1,200°C. for 24 hours. This treatment introduced an *n*-type layer, highly doped with phosphorus and about 0.001-cm thick, into the surface regions of the specimen in the tube containing P₂O₅. Upon removal from the tube this specimen was lapped on the end to remove the *n*-skin. Complete removal was determined by testing with a thermal probe.

Small cubes of lithium (0.038 em on a side) were placed on the ends of both samples (the lapped end of phosphorus-doped one) and alloyed to the silicon for 30 seconds at 650°C in an atmosphere of dry helium. After this treatment the various junction contours should have looked like those in Fig. 6, in which the bottom crystal is shown with the phosphorus-doped skin (cross hatched). During the alloying process a small amount of spherical diffusion of lithium occurs so that small hemispherical *n*-regions form with the alloy beads as origins. These are shown in Fig. 6.

Next the specimens were heated in vacuum for about six hours at 400°C. Diffusion of lithium into the body of the crystal should occur during this period. However in the sample not protected by the *n*-type skin lithium should leak to the oxide sink on the surface so that the *n*-type region due to the lithium should have the pear-shaped contour shown in the upper part of Fig. 7. If the built-in field at the p-n junction

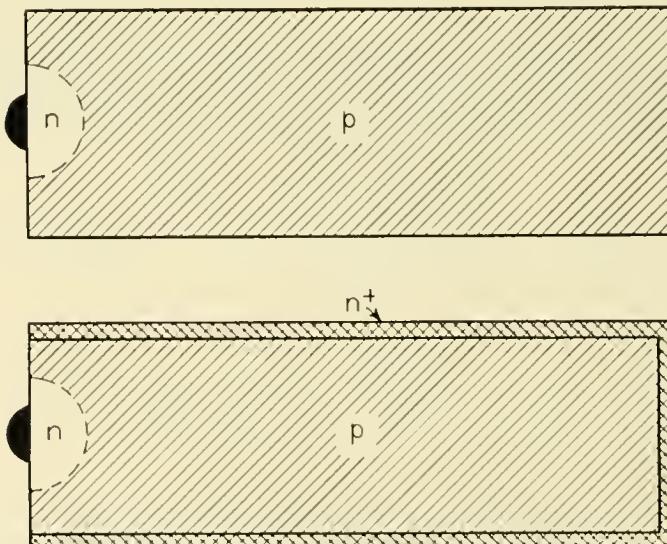


Fig. 6 — Initial stage following alloying in the diffusion experiment to demonstrate the impermeability to lithium of a heavily doped *n*-type skin on silicon.

formed by the phosphorus layer prevents lithium from reaching the surface, diffusion in the sample with the skin should be plane parallel with a straight front (except at the rear where the skin has been lapped off and lithium can leak out) as the p-n junction contour in the lower part of Fig. 7 indicates.

An acid staining technique³² which reveals the junction contours should then develop a picture resembling Fig. 7. The two specimens were cut along their long axes and the stain applied to the newly exposed surfaces. The result has been photographed and is shown in Fig. 8 where the crystal on the right has the n-skin. The p-regions show up dark and the n, light. The result agrees with Fig. 7.

In another experiment a crystal completely enclosed in a phosphorus skin was immersed in the tin bath described in Section III. It was discovered that lithium entered the crystal with no evident difficulty, just as though the skin were absent, but once in, could not be driven out by removal of the external source and continued heating. The implication is clear. The built-in field has a rectifying action permitting the lithium to enter the crystal but not to leave. In this sense it performs the same function for the mobile lithium ions as it does for holes in a p-n junction diode.³³

V. COMPLEX ION FORMATION

In the previous text processes involving the interaction of electrons and holes have been considered. In this section attention will be drawn,

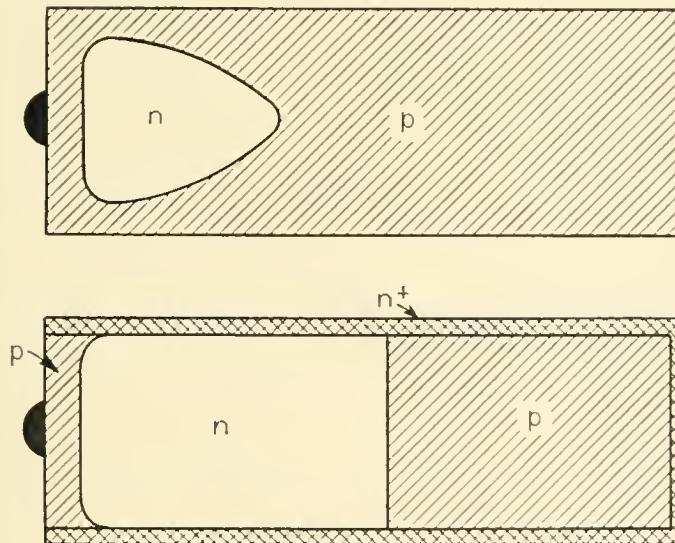


Fig. 7 — Distribution of lithium after an extended period of diffusion at a temperature lower than the alloying temperature — showing leakage out of the crystal in the one case (no-skin) and conservation in the other.

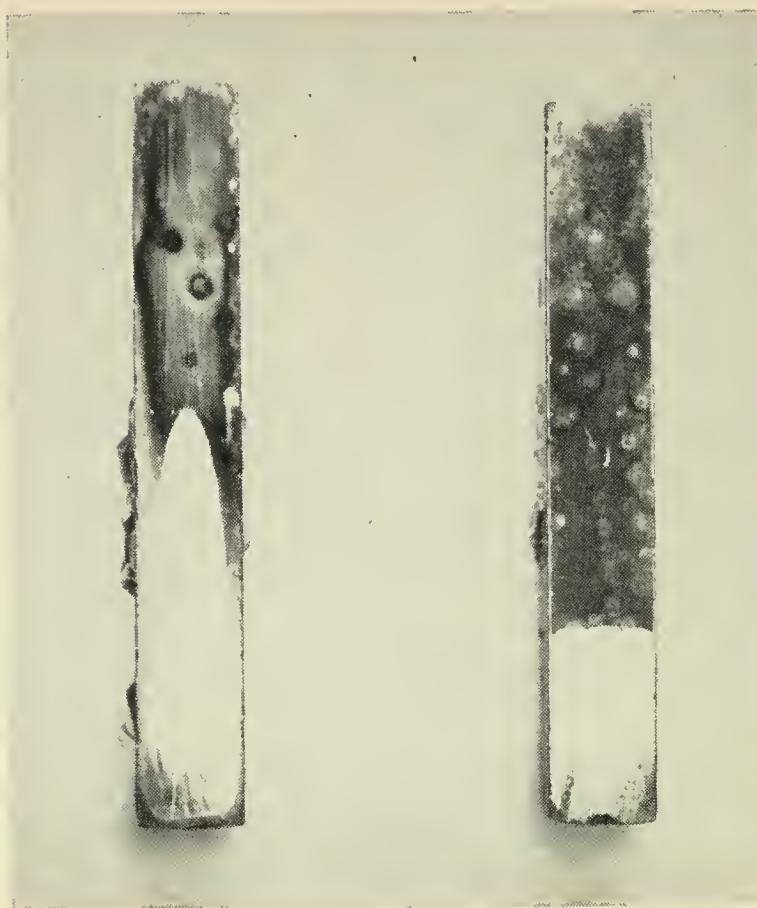


Fig. 8 — Photograph of experimental situation described schematically in Fig. 7.

to the possibility of interactions between the donor and acceptor ions themselves. For example, in (2.1) direct interaction of Li^+ and B^- above $600^\circ C$ may be possible, especially in view of the mobility of Li^+ . Such a reaction was indicated in the work of Reiss, Fuller, and Pietruszkiewicz.³⁴

Fig. 9 is of assistance in understanding the nature of these observations. In it are shown plots of the solubility of lithium in silicon. In this case the situation is similar to that involved in the germanium curves of Fig. 5 because the external phase is composed of silicon and lithium and is probably of the liquidus composition. It is formed by simply alloying lithium to the silicon surface. In Fig. 9, Curve A, illustrates how solubility depends on temperature when the silicon is undoped. Curve B, unlike A, is not an experimental plot, i.e., it is not supposed to represent the locus of the points through which it seems to pass. Instead it has been calculated from the theory expounded below. The points themselves are experimental and represent solubility measurements on silicon doped with boron to the level $1.9 \times 10^{18} \text{ cm}^{-3}$.

Curve A possesses a maximum (just as the D_0^+ curve of Fig. 5) in the neighborhood of 650°C. A marked disparity is apparent between solubilities in undoped and doped silicon, the solubility in the latter being greater. Below 500°C this disparity is easily understood. It stems from the electron-hole equilibrium considered previously. However the high solubility in doped silicon at high temperatures is not explicable on this basis since the crystal becomes intrinsic, and e^+e^- is mostly dissociated. To account for this phenomenon Reiss, Fuller, and Pietruszkiewicz invoked the idea of interaction between Li^+ and B^- . They presented the following argument.

At low temperatures lithium ions occupy the interstices of the silicon

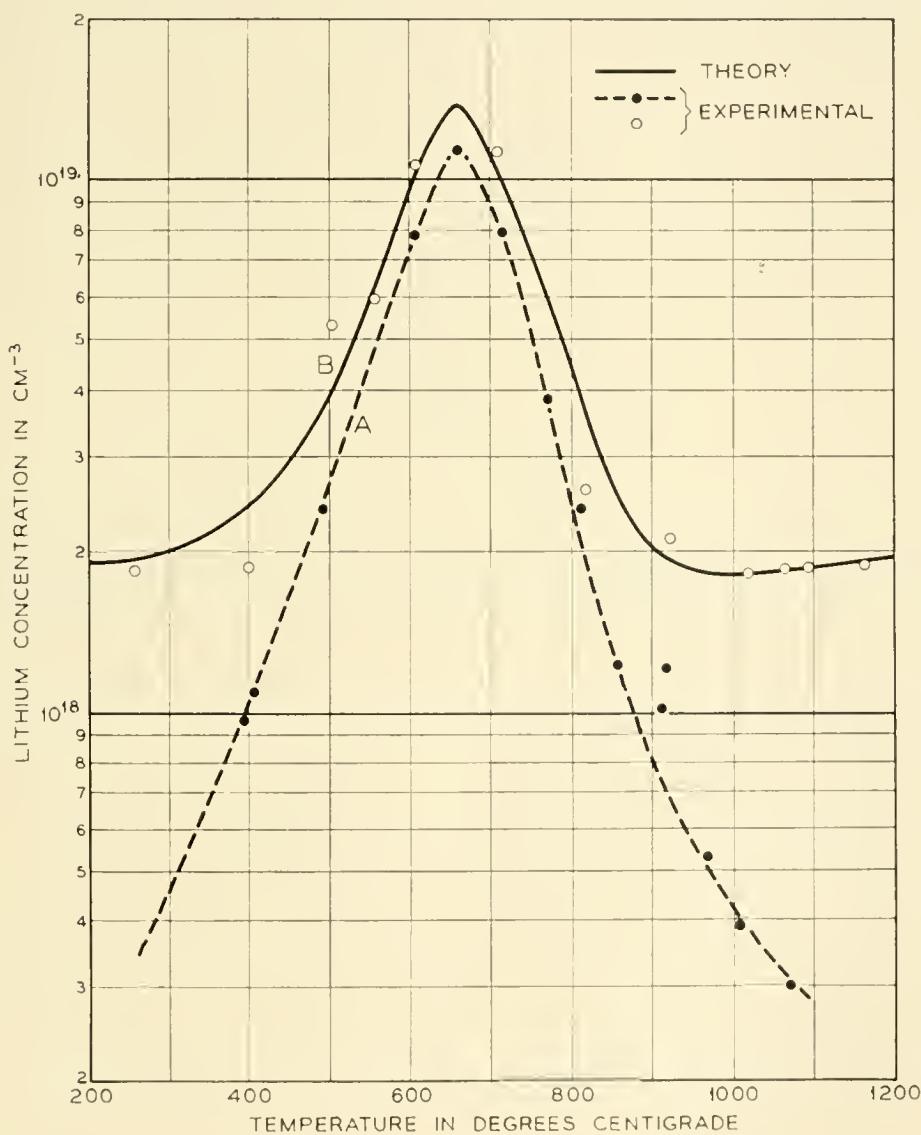


Fig. 9 — Plots showing the solubility of lithium in silicon as a function of temperature. The external phase is an alloy of lithium and silicon. Curve A is for undoped silicon. The locus of the points in B is for silicon doped with about $1.9 \times 10^{18} \text{ cm}^{-3}$ boron.

lattice as in Fig. 10. In an interstitial position lithium can approach an oppositely charged boron, but the interaction will be, at the most, coulombic so that an ion pair will form (see later sections). A covalent bond is unable to appear not only because there are no electrons available for it, but also because the lithium ion cannot move to a position where it can satisfy the tetrahedral symmetry inherent in sp^3 hybridization.³⁵ Calculations (of the sort appearing in the later sections of this paper) show that at high temperatures, at the ion densities involved, ion pairs of the kind depicted in Fig. 10 are completely dissociated.

Suppose, however, that as temperature is raised vacancies dissolve in the silicon lattice, and that one such vacancy occupies a position near

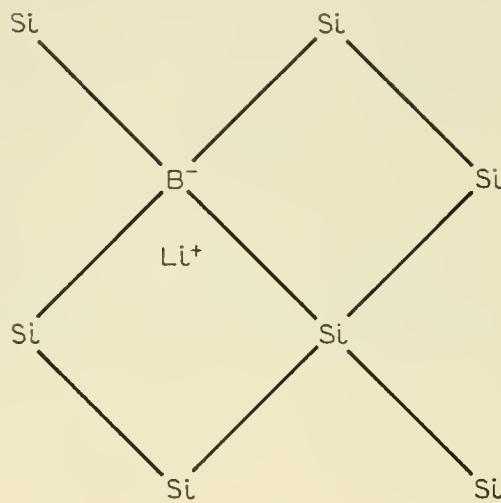


Fig. 10 — Schematic diagram of a silicon lattice showing a lithium ion in an interstitial position near a substitutional boron ion, as it occurs in an ion pair.

a boron ion, as in Fig. 11, a slight modification of Fig. 10 in which the dots represent electrons (dangling bonds). Unpaired electrons such as these might capture an electron from the valence band of silicon so that the vacancy acquires a negative charge and behaves like an acceptor. It is reasonable to suppose that the positive lithium ion will move into this negative vacancy, in the tetrahedral position, and form a covalent bond as in Fig. 11. The lithium-boron complex so formed retains a negative charge and is thus a complex ion. If the specimen were extrinsic at these high temperatures, there would still appear to be as many net acceptors as before the addition of lithium.*

If the LiB^- compound is stable enough (a question to which we shall

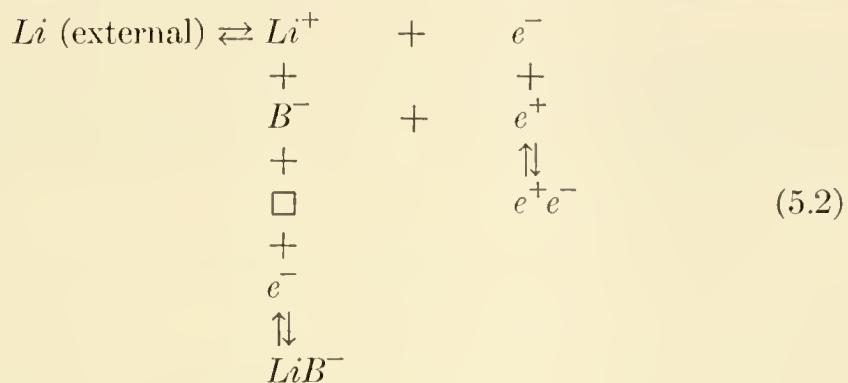
* It is possible that rapid cooling may quench some of these LiB acceptors into the crystal at room temperature. If this is so it should be possible to investigate the associated energy level by Hall measurements in the interval of time before the complexes anneal out. Similar phenomena might be observed in germanium.

return below) to hold the lithium atom, the solubility of lithium will be determined principally by the density of boron atoms. At low temperatures, vacancies are reabsorbed and the lithium atoms return to their interstitial positions, at quenched-in densities corresponding to the temperatures of equilibration. However, boron acceptors now appear to be compensated since interstitial lithium behaves as a donor. This renders it feasible to measure the concentration of lithium by the determination of resistivity.

The overall reaction may be written in the form



in which \square represents a vacancy. This equilibrium can be grafted onto (2.1) so that the latter becomes (ignoring un-ionized lithium and boron)



The original vertical equilibrium involving holes and electrons loses its significance at high temperatures, and the new vertical reaction becomes important, for both \square and e^- appear in increased concentrations. In this way a certain amount of symmetry, insofar as temperature is concerned, is introduced into the problem, i.e., as one equilibrium ceases to dominate

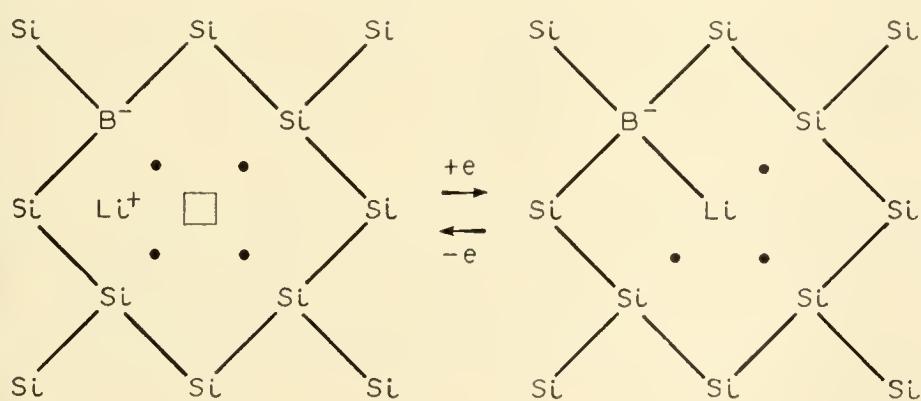


Fig. 11 — Schematic diagram illustrating the reaction in (5.1). The square represents the center of a vacancy and the dots, electrons left unpaired by the occurrence of the vacancy.

the system the other begins to take effect. This symmetry, of course, is necessary for explaining the symmetrical locus of the points around Curve B in Fig. 9.

The scheme (5.2) can be treated quantitatively by applying the mass action principle, but now the symbol D^+ can not be used for the solubility of lithium since the totality of dissolved lithium is distributed between LiB^- and Li^+ , and the symbol only applies to the latter. We therefore denote the total concentration of lithium by N_D , and the concentration of LiB^- by C . Then

$$N_D = D^+ + C \quad (5.3)$$

The same argument applies to boron, so that its total concentration will be designated by

$$N_A = A^- + C \quad (5.4)$$

The problem then reduces to specifying N_D as a function of N_A . To accomplish this, to (3.1) and (3.2) is added the mass action expression going with (5.1)

$$\frac{C}{D^+ A^- n} = \gamma e^{-(\beta/T)} = \pi \quad (5.5)$$

where γ and β are constants. It has been assumed that the vacancy concentration follows a temperature law of the form $\gamma^* \exp[-\beta^*/T]$ where γ^* and β^* like γ and β are constants. This permits the equilibrium constant when multiplied by the vacancy concentration to assume the form $\gamma \exp[-\beta/T]$ shown in (5.5). In place of (2.8) a new conservation condition,

$$D^+ + p = C + A^- + n \quad (5.6)$$

is introduced. The combination (3.1), (3.2), (5.3), (5.4), (5.5) and (5.6) can be solved so that N_D , the lithium solubility appears as a function of the total boron concentration N_A . Thus

$$N_D = \frac{N_A}{1 + \sqrt{1 + (2n_i/N_D^0)^2}} + \sqrt{\left\{ \frac{N_A}{1 + \sqrt{1 + (2n_i/N_D^0)^2}} \right\}^2 + (N_D^0)^2} + \frac{\pi N_A (N_D^0)^2 [1 + \sqrt{1 + (2n_i/N_D^0)^2}]}{2 + \pi (N_D^0)^2 [1 + \sqrt{1 + (2n_i/N_D^0)^2}]} \quad (5.7)$$

In this equation N_D^0 like D_0 in (3.4) is the solubility of lithium in undoped silicon, i.e., in silicon from which boron is absent.

All the parameters in (5.7) are independently measurable save π

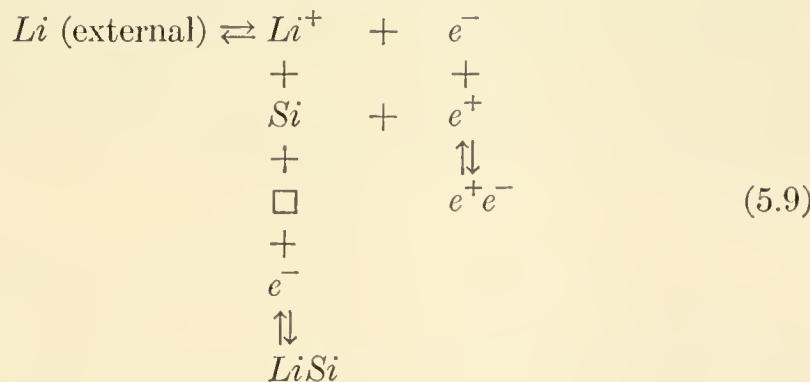
which can be known for all temperatures when γ and β have been determined. Reiss, Fuller, and Pietruszkiewicz used two of the points near Curve B in Fig. 9, above 1,000°C, to define values of N_D for use in (5.7). Then π was computed from (5.7) at these two temperatures. From these values of π , γ and β were determined, and from these, in turn, π was calculated for all temperatures down to 200°C. Using π , N_D was computed from (5.7) over the entire experimental range of temperature. The result is Curve B of Fig. 9 which fits the experimental points very well.

Another check on the validity of the theory (which has not yet been accomplished) would be the following. At high temperatures (5.7) reduces to

$$N_D = N_D^0 + \left\{ \frac{\pi(N_D^0)^2[1 + \sqrt{1 + (2n_i/N_D^0)^2}]}{2 + \pi(N_D^0)^2[1 + \sqrt{1 + (2n_i/N_D^0)^2}]} \right\} N_A \quad (5.8)$$

i.e., N_D is a linear function of N_A with the slope (in brackets) depending upon π . Measurement of this slope at one temperature would thus provide an independent evaluation of π .

A little thought concerning the scheme outlined in (5.2) leads one to wonder why the introduction of boron really increases the solubility of lithium because the same mechanism could be applied to the case in which boron is absent, i.e., to Curve A of Fig. 9. Thus, if B^- is replaced by a silicon atom in Figs. 10 and 11, the entire scheme can be adopted unchanged, except that Si replaces B^- . Thus



and one wonders why LiB^- should be more stable than $LiSi$. A possible answer is the following:

The tetrahedral covalent radius of boron is 0.88 Å.³⁶ This is to be contrasted with the tetrahedral radius of silicon which is 1.17 Å.³⁶ When boron is substituted in the silicon lattice it therefore produces considerable local compressive strain. This strain is partially relieved when a vacancy is formed adjacent to the boron. Thus the energy required to form a vacancy near a boron ion in silicon is less than is required for its

formation near a silicon atom. Hence the endothermal heat of formation of LiB^- in (5.2) is reduced substantially (by the amount of the released energy of elastic strain) below the heat of formation of $LiSi$. This accounts for the greater stability of the former.

The compressive strain around a substitutional boron in germanium is also illustrated by ion pairing studies to be described later in Section XII. Its action in that case keeps the ions which form a pair from approaching each other as closely as they otherwise might. Although really quantitative studies of pairing have not yet been performed in silicon, the lattice parameters of germanium and silicon are sufficiently close to render it fairly certain that the same strain exists in the latter as in the former. This lends support to the previous argument.

Before closing this section there is another related topic which is worth mentioning. This concerns part of the explanation of the retrograde solubility observable in the curves of Figs. 5 and 9, i.e., the occurrence of the maxima. The solubilities along these curves are given by (3.3) in the form

$$D_0^+ = K^*/(K^* + n_i^2)^{1/2}$$

Suppose that at low temperatures K^* is an increasing function of temperature and considerably larger than n_i . Then we have the approximation

$$D_0^+ = (K^*)^{1/2} \quad (5.10)$$

in which the solubility D_0^+ must increase with temperature. If n_i increases more rapidly than K^* with temperature, a point will be reached at which n_i^2 in the denominator of the (3.3) in its special form above, exceeds K^* by so much that the latter can be ignored. When this is so another approximation holds,

$$D_0^+ = \frac{K^*}{n_i} \quad (5.11)$$

in which D_0^+ decreases with temperature since n_i increases more rapidly than K^* . Since (5.10) predicts an increase in solubility with temperature at low temperatures and (5.11) a decrease at higher temperatures a maximum occurs somewhere between. The maximum may not be due to this cause alone, however. For example K^* contains the activity, α , in the external phase, and this may vary with temperature in an erratic manner.

In any event the influence of the electron-hole equilibrium on D_0^+ in both silicon and germanium cannot be ignored. The fact that the distribution coefficients of donors and acceptors in silicon are usually some

ten-fold greater than in germanium may be due to the smaller width of the forbidden gap in the latter. This makes for greater values of n_i and according to (3.3) smaller values of D_0^+ .

VI. ION PAIRING

The preceding text drew an analogy between semiconductors and aqueous solutions — phenomena such as neutralization, common ion effects, and complex formation have been discussed. Another feature of "wet" chemistry which has appealed to chemists concerns the influence of coulomb forces among ions on the properties of solutions. This subject is of peculiar interest because such forces are well understood, and considerable progress can be made in the quantitative prediction of their effects.

The first really successful theoretical treatment of coulomb forces in solution is the Debye-Hückel theory.³⁷ This treatment recognizes the long range character of coulomb forces, and endeavors to account for their effects in terms of a communal interaction involving all of the ions in solution. The theory has now been shown to include certain statistical inconsistencies³⁸ which, however, are of small consequence in dilute solutions where theory and experiment are in excellent agreement.

The central feature of the Debye-Hückel theory is the concept of the ionic atmosphere, i.e., the time average excess concentration of ions of opposite sign which accumulates in the neighborhood of a particular ion. The radius of this atmosphere is measured (order of magnitude-wise) by the now famous Debye length.

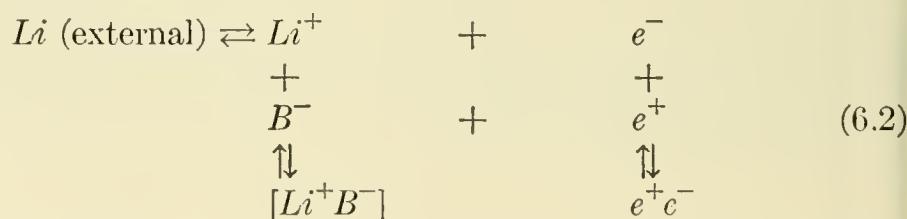
$$L = \sqrt{\frac{k\kappa T}{8\pi q^2 N}} \quad (6.1)$$

in which κ is the dielectric constant of the medium, q is the charge on an ion, and N is the (in this case identical) concentration of both positive and negative ions. As κ decreases or N increases, L becomes smaller so that the atmosphere is more tightly gathered in. As this process continues a stage is reached in which the atmospheres of some of the ions may be best thought of as being fully constituted by a single ion of opposite sign, i.e., an *ion pair* forms. This pair-wise interaction is so intense relative to the communal interaction mentioned above, that insofar as the paired ions are concerned it may be regarded as the only interaction influencing the distribution of the pairs themselves. Unpaired ions may still be treated by the communal Debye-Hückel theory but their concentration must be considered as the true concentration of ions reduced by the

concentration of pairs since the latter possess effectively no fields. In any event when pairing occurs the Debye-Hückel effects are relatively second order, since, even normally, they represent quite small deviations from ideal solution behavior. Under pairing conditions it is desirable, in the first approximation, to focus one's attention on the pairing interaction.

While developing the aqueous solution analogy inherent in our semiconductor model it is natural to inquire whether or not a system like (2.1), in which at least one of the ions can move, will show effects due to coulomb interaction. A preliminary calculation using (6.1) indicates that if coulomb effects are to be observed they are likely to be of the ion pairing variety rather than of the Debye-Hückel type because the dielectric constants of semiconductors are low relative to that of water, e.g., 12 for silicon³⁹ and 16 for germanium⁴⁰ as against 80 for water.⁴¹ The dominance of ion pairing stems, as it will become clear later, from still another feature peculiar to semiconductors. This is the closeness with which two ions of opposite sign can approach one another in semiconductors. In any event experiments are not yet at the stage of sensitivity necessary for the accurate measurement of the small Debye-Hückel effects so that we are virtually compelled to ignore such phenomena.

Fig. 10 is a picture of an ion pair in boron-doped silicon. Corresponding to this process one may sketch in another vertical equilibrium in (2.1) to yield (ignoring un-ionized Li)



where $[Li^+B^-]$ stands for the ion pair in which the individual ions maintain their polar identities and the binding energy is coulombic. The ion pair is a compound in a statistical sense since as will be seen later the distance between the ions of a pair is distributed over a range of values. The interaction between Li^+ and B^- is to be distinguished from that shown in (5.2). The latter occurs at high temperatures whereas the former is presumably limited to low temperatures, below 300°C.

The quantitative aspects of ion pairing were first considered by Bjerrum⁴² and later by Fuoss⁴³ who placed Bjerrum's theory on a somewhat more acceptable basis. Fuoss's theory, however, suffers from some of the same limitations as Bjerrum's. Nevertheless the Bjerrum-Fuoss theory is capable of satisfying experimental data over broad ranges of conditions. In the next section we present a brief resumé of this theory together with relevant criticism and its relation to a more refined theory due to Reiss.

VII. THEORIES OF ION PAIRING

Fuoss begins by considering a solution of dielectric constant κ , containing equal concentrations, N , of ions of opposite sign. When equilibrium has been achieved each negative ion will have another ion (most probably positive) as a nearest neighbor, a distance r away from it. Fuoss discounts the possibility that the nearest neighbor will be another negative ion, and proceeds to calculate what fraction of such nearest neighbors lies in spherical shells of volumes, $4\pi r^2 dr$, having the negative ions at their origins. If this fraction is denoted by $g(r) dr$, it may be evaluated as follows.

In order for the nearest neighbor to be located in the volume, $4\pi r^2 dr$, two events must take place simultaneously. First the volume, $4\pi r^3/3$, enclosed by the spherical shell must be devoid of ions, or else the ion in the shell would *not* be the nearest neighbor. Since $g(x)dx$ is the probability that a nearest neighbor lies in the shell, $4\pi x^2 dx$, the probability that a nearest neighbor does not lie in this shell is $1 - g(x)dx$. From this it is easily seen that the chance that the volume $4\pi r^3/3$ is empty is

$$E(r) = 1 - \int_a^r g(x) dx \quad (7.1)$$

where a is the distance separating the centers of the two ions of opposite sign when they have approached each other as closely as possible.

The second event which must take place is the occupation of the shell $4\pi r^2 dr$ by *any* positive ion. The chance of this event depends on the time average concentration of positive ions at r . This concentration is bound to exceed the normal concentration N by an amount depending on r , because of the attractive effect of the negative ion at the origin. It may be designated by $c(r)$. The probability in question is then

$$4\pi r^2 c(r) dr \quad (7.2)$$

The chance $g(r) dr$ that the nearest neighbor lies in the shell $4\pi r^2 dr$ is therefore the product of (7.1) by (7.2), i.e., the product of the probabilities of the two events required to occur simultaneously. This leads to the relation

$$g(r) = \left(1 - \int_a^r g(x) dx\right) 4\pi r^2 c(r) \quad (7.3)$$

an integral equation whose solution is

$$g(r) = \exp \left[-4\pi \int_a^r x^2 c(x) dx \right] 4\pi r^2 c(r) \quad (7.4)$$

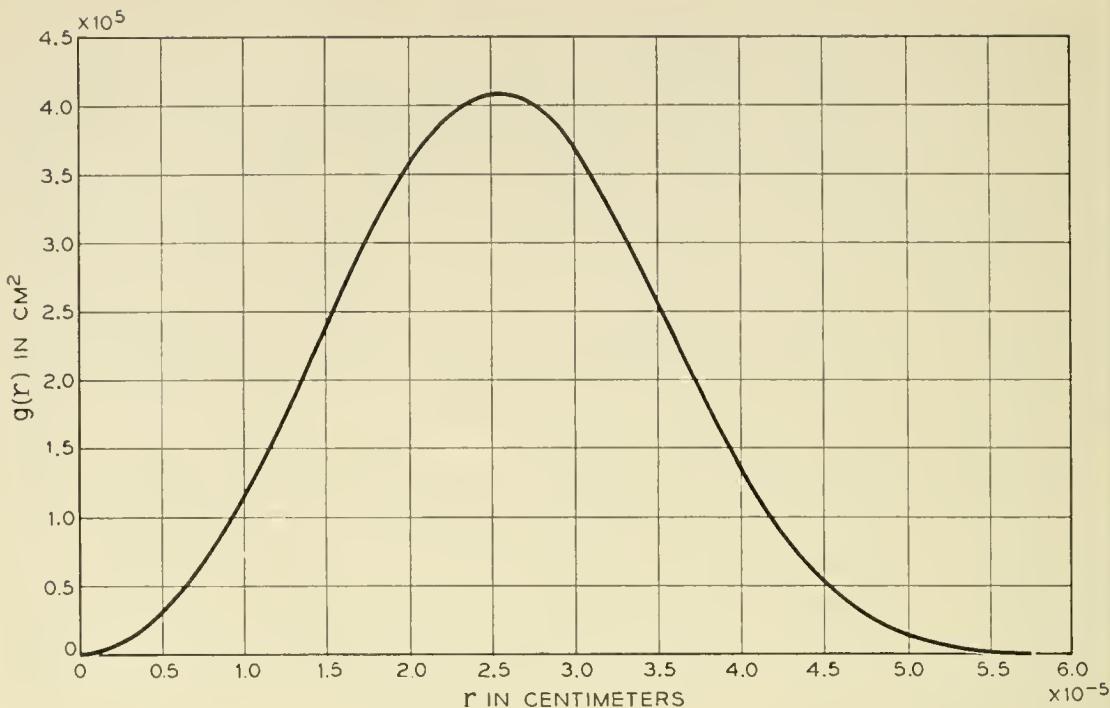


Fig. 12 — Distribution of nearest neighbors in a random assembly of particles for a concentration of 10^{16} cm^{-3} .

That (7.4) solves (7.3) is easily demonstrated by substitution of the latter into the former.

If there were no forces of attraction between ions then $c(r)$ would equal N , and if a is taken equal to zero (7.4) reduces to

$$g(r) = 4\pi r^2 N \exp(-4\pi r^3 N/3) \quad (7.5)$$

This function is plotted in Fig. 12 for the case $N = 10^{16} \text{ cm}^{-3}$. Note that the position of the maximum, the most probable distance of location of a nearest neighbor, occurs near the value of r equal to $(3/4\pi N)^{1/3}$. This is the radius of the average volume per particle when the concentration is N , i.e. the volume, $1/N$.

In order to write $g(r)$ for the case of coulombic interaction it is necessary to compute $c(r)$ under these conditions. Fuoss (after Bjerrum) reasoned as follows. If a theory can be constructed which depends only upon the characteristics of *near* nearest neighbors (nearest neighbors at small values of r) then the force of interaction experienced by the nearest neighbor can be assumed to originate completely in the coulomb field of the negative ion at the origin. This is predicated on the argument that both positive and negative ions develop atmospheres of opposite sign which are superposed when the two ions are close to one another. The result is a cancellation of the net atmosphere leaving nothing for the two

ions to interact with but themselves. Thus the potential energy of interaction, for near nearest neighbors will be

$$-\frac{q^2}{\kappa r} \quad (7.6)$$

For small values of r , therefore, $c(r)$ can be derived from Boltzmann's law and is given by

$$c(r) = h \exp [q^2/\kappa k T r] \quad (7.7)$$

where h is a constant. Guided by the requirement that $c(r)$ should equal N at infinite distance from the central negative ion, h was set equal to N giving, finally,

$$c(r) = N \exp [q^2/\kappa k T r] \quad (7.8)$$

The assumption that a theory could be developed depending only on near nearest neighbors proved reasonable, but the choice of $h = N$ in (7.8) leads to certain logical difficulties. Thus the average volume dominated by a given negative ion is evidently $1/N$. If (7.8) is summed over this volume the result, representing the number of positive ions in $1/N$, should be unity since there are equal numbers of positive and negative ions. Unfortunately, the result exceeds unity by very large amounts except for very small values of N , i.e., for very dilute solutions. We shall return to this point later.

If (7.8) is inserted into (7.4) the resulting $g(r)$ has the form typified by Fig. 13. First, there is an exponential maximum occurring at $r = a$, followed by a long low minimum, and this by another maximum which like the one in Fig. 12 occurs, not far from $r = (3/4\pi N)^{1/3}$, if N is not too large. For small values of N the minimum occurs at

$$r = b = q^2/2\kappa k T \quad (7.9)$$

The function $g(r)$ is actually normalized in (7.4) so that the area under the curve is unity. The second maximum corresponds to the most probable position for a nearest neighbor in a random assembly, i.e., to the maximum in Fig. 12. Essentially the first maximum has been grafted onto Fig. 12 by the interaction at close range which makes it probable that short range neighbors will exist. At high values of N the region under the first maximum becomes so great that enough area is drained (by the condition of normalization) from the second maximum to make it disappear entirely. At this point the minimum is replaced by a point of inflection. More will be said concerning this phenomenon later.

Fuoss chooses to define all sets of nearest neighbors inside the mini-

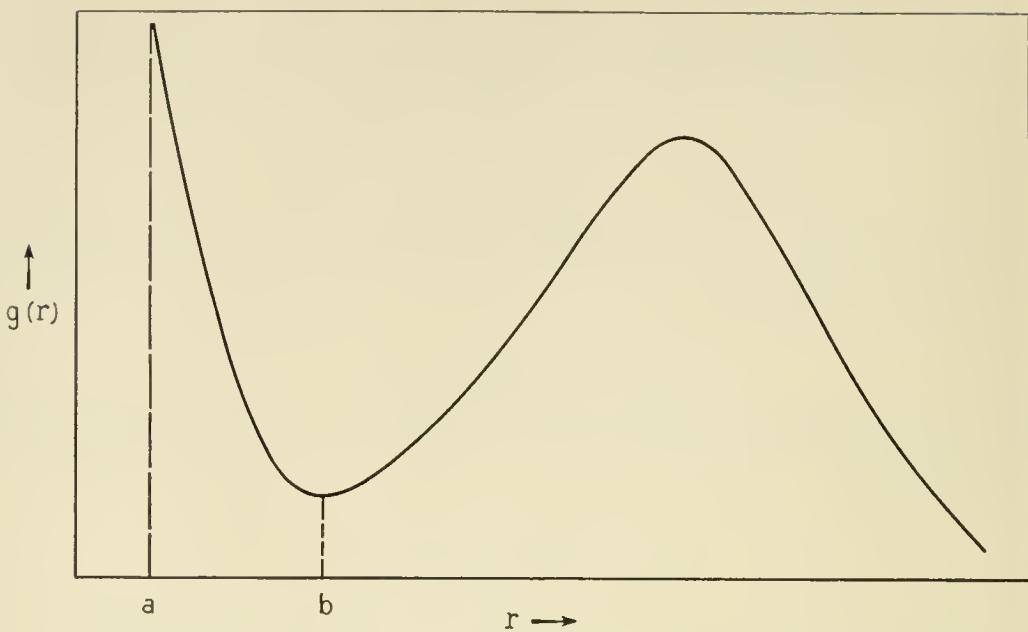


Fig. 13 — Schematic distribution of neighbors in an assembly of particles when forces of interaction are present. Repulsive forces are reflected in the appearance of a distance a , of closest approach of two particles, attractive forces by the exponential maximum at a .

mum, i.e., inside $b = q^2/2\kappa kT$, as ion pairs, and the rest as unpaired. No thought is given to the small fraction of nearest neighbors which involves ions of like sign, as it must be small inside $r = b$. Nor is any thought given to the possibility that a given positive nearest neighbor may be the nearest neighbor of two negative ions simultaneously. Such a coincidence would be very improbable at a distance short enough to be within $r = b$. Thus if the entire theory can be made to depend on what happens inside b , its foundations are reasonable, except for the choice of $h = N$.

To obviate this difficulty Fuoss had further to devise a means of performing all calculations under conditions where the choice of $h = N$ was not inconsistent. He assumed (following Bjerrum) that paired and unpaired ions were in dynamic equilibrium and that the law of mass action could be applied to this equilibrium. Thus if P represents the concentration of pairs, $N - P$ denotes the concentration of unpaired ions of one sign and the mass action expression is

$$\frac{P}{(N - P)^2} = \Omega \quad (7.10)$$

where Ω is an equilibrium constant independent of concentration. At infinite dilution, where the assignment $h = N$ is valid, Ω should be the same as at higher concentrations. Therefore (7.4) can be used to evalu-

ate Ω at infinite dilution, and the value so obtained employed at higher concentrations.

Besides the inconsistency of the choice, $h = N$, the form (7.4) contains another objectionable feature. This is revealed by a more rigorous treatment devised recently by Reiss,⁴⁴ and has to do with the factor,

$$\exp \left[-4\pi \int_a^r x^2 c(x) dx \right],$$

in (7.4). It can be shown that this factor is inconsistent with the supposition that the nearest neighbor to a given negative ion interacts only with that ion and no other. Fortunately, in Fuoss's scheme $g(r)$ given by (7.4) needs to be used only at infinite dilution, and then only for such values of r as lie inside b . Under this condition and in this range the exponential factor in question can be replaced by unity from which it deviates only slightly. Thus the form of $g(r)$ used eventually is

$$g(r) = 4\pi r^2 N \exp [q^2/\kappa kTr] \quad (7.11)$$

Ω is computed as follows. At infinite dilution P tends toward zero so that (7.10) becomes

$$\frac{P}{N} = \Omega N \quad (7.12)$$

But P/N is the fraction of ions paired which by definition is the fraction of nearest neighbors lying inside $r = b$. From the definition of $g(r)$, P/N is evidently given by

$$\frac{P}{N} = \int_a^b g(r) dr = 4\pi N \int_a^b r^2 \exp [q^2/\kappa kTr] dr \quad (7.13)$$

which upon substitution in (7.12) yields

$$\Omega = 4\pi \int_a^b r^2 \exp [q^2/\kappa kTr] dr \quad (7.14)$$

The evaluation of Ω in this way permits one to base the entire theory on the distribution of *near* nearest neighbors, so that all the assumptions which demand this procedure are validated.

Using the computed Ω in (7.10) P can be evaluated, and also $N - P$ which as the concentration of *free* ions of one species measures the thermodynamic activity of that species. In this manner it is possible to calculate the *equilibrium* effects of coulomb interaction insofar as solution properties are concerned. To treat transport phenomena such as ionic mobility in an applied electric field Fuoss assumes that paired ions repre-

senting neutral complexes are unable to respond to the applied field and so do not contribute to the overall mobility. The mobility of unpaired ions is assumed to be μ_0 , the mobility observable at infinite dilution. The apparent mobility μ at any finite concentration is then μ_0 reduced by the fraction P/N of ions paired. Thus

$$\mu = [1 - (P/N)]\mu_0 \quad (7.15)$$

The Bjerrum-Fuoss theory when applied to real systems reproduces the experimental data very well, although the parameter a , the distance of closest approach, needs to be determined from the data itself.

The concept of a pair defined in terms of the minimum occurring at b , becomes rather vague when that minimum vanishes in favor of a point of inflection. At this stage triplets and other higher order clusters form and the situation becomes very complicated.

In Reference 44, Reiss has developed a more refined theory of pairing. Instead of avoiding the use of an inconsistent $g(r)$ by introduction of the mass action principle, an attempt is made to provide a rigorous form for $g(r)$, which proves to be the following

$$g(r) = \exp [-4\pi r^3 N/3] 4\pi r^2 h \exp [q^2/\kappa kTr] \quad (7.16)$$

in which

$$h = 1 / \left[\int_a^\infty \exp [-4\pi r^3 N/3] 4\pi r^2 \exp [q^2/\kappa kTr] dr \right] \quad (7.17)$$

It is also shown that the activity of an ionic species, measured by $N - P$ in the Bjerrum-Fuoss theory, is measured by \sqrt{hN} in the more rigorous theory. The distribution (7.16) suffers neither from an inability to conserve charge in the volume $1/N$ (as does (7.4)) nor from any inconsistency involving the interaction of a nearest neighbor with other ions than the one to which it is nearest neighbor [as does (7.4)].

When \sqrt{hN} computed by (7.17) is compared with $(N - P)$ computed according to (7.10) and (7.14), for arbitrary values of κ , a , T , and N , the results are almost identical. This shows the virtue of the Bjerrum-Fuoss theory, and in fact, suggests that in most cases it should be used for calculation rather than the more refined theory, for the latter involves rather complicated numerical procedures.

The refined theory can also be adapted to the treatment of transport phenomena.⁴⁵ Thus in place of $g(r)$ it is possible to write a distribution function $\Gamma(\vec{r})$, specifying the fraction of nearest neighbors lying in the volume element $d\vec{r}$, in a system in the steady state rather than at equilibrium. In the presence of an applied field the distribution loses its spheri-

cal symmetry and it must be defined in terms of the volume element $d\vec{r}$, lying at the vector distance \vec{r} , rather than in terms of the spherical shell of volume, $4\pi r^2 dr$. In reference (44) it is shown that

$$\Gamma(\vec{r}) = \exp[-4\pi r^3 N/3] c(\vec{r}) \quad (7.18)$$

where $c(\vec{r})$ is the density function in the non-equilibrium case, and is determined by the equation

$$\frac{kT}{q} \nabla^2 c + c \nabla^2 \psi + \nabla c \cdot \nabla \psi = 0 \quad (7.19)$$

after suitable boundary conditions have been appended. The quantity ψ , designates the local electrostatic potential, determined by the ions as well as the applied field. These equations are restricted specifically to the semiconductor case in which the negative ion is unable to move.

The current carried by nearest neighbors in the volume element $d\vec{r}$ in unit volume of solution is

$$J(\vec{r}) = -\exp[-4\pi r^3 N/3] c(\vec{r}) \mu_0 \nabla [\psi + (kT/q) \ln c(\vec{r})] \quad (7.20)$$

Using these equations it proves possible in reference 45 to provide a more refined version of (7.15) in which the mobility of nearest neighbors inside $r = b$ need not be considered zero, nor those outside $r = b$ be considered perfectly free and possessed of the mobility μ_0 . In fact the average mobility of a nearest neighbor separated by a distance r from its immobile partner proves to be

$$\bar{\mu} = \frac{\mu_0}{2(1-F)} \left(\left[\frac{\varepsilon^2}{3r^2} + \frac{4\varepsilon}{3r} + 2 \right] \exp(-\varepsilon/r) + 2F \left(\frac{\varepsilon}{3r} - 1 \right) \right) \quad (7.21)$$

where

$$\varepsilon = q^2/\kappa kT \quad (7.22)$$

and

$$F = \left(\frac{\varepsilon^2}{2a^2} + \frac{\varepsilon}{a} + 1 \right) \exp(-\varepsilon/a) \quad (7.23)$$

For values of r greater than ε (7.21) can be approximated by

$$\frac{\bar{\mu}}{\mu_0} = \frac{1}{2} \left(\frac{\varepsilon^2}{3r^2} + \frac{4\varepsilon}{r} + 2 \right) \exp(-\varepsilon/r) \quad (7.24)$$

and is therefore a function of ε/r . Fuoss's b corresponds to $r = \varepsilon/2$ or to $\varepsilon/r = 2$. Fig. 14 contains a plot of $\bar{\mu}/\mu_0$ versus r for $T = 400^\circ\text{K}$, $a = 2.5 \times 10^{-8}$ cm, $q = 4.77 \times 10^{-10}$ statcoulombs, and $\kappa = 16$. Note that

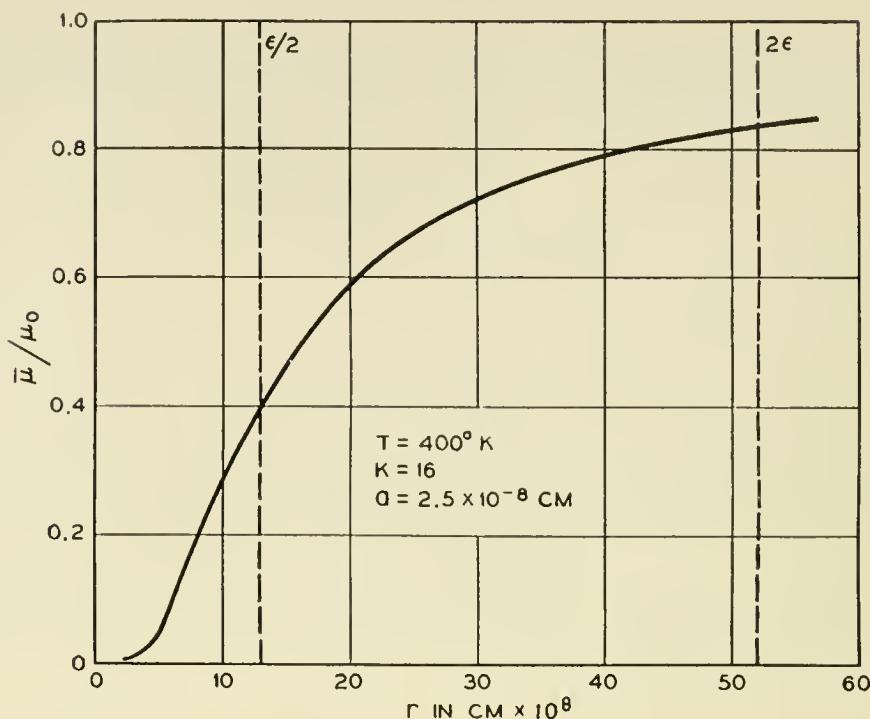


Fig. 14 — Average mobility (calculated from the refined theory of pairing) of a mobile ion in a pair as a function of the distance from its immobile neighbor. The example shown corresponds to a substance having $a = 2.5 \times 10^{-8} \text{ cm}$ $\kappa = 16$ at a temperature of 400°K .

at $r = \epsilon/2 = b$, $\bar{\mu}/\mu_0$ is near 0.5 which is the average value of Fuoss's $\bar{\mu}/\mu_0$ for ions taken from either side of $r = b$. Therefore a certain symmetry with respect to $r = b$ does exist, tending to justify Fuoss's model. According to (7.24) $\bar{\mu}/\mu_0$ is 0.8 by the time $r = 3\epsilon/2 = 3b$, independent of the value of a . In other words an ion located a short distance beyond b does have practically complete mobility as the Bjerrum-Fuoss theory assumes.

The refinement of (7.15) which occurs can be written as follows

$$\begin{aligned} \mu = & \left\{ \frac{2h}{(1 - F)} \int_a^\infty \left[\left(2r^2 + \frac{4\epsilon r}{3} + \frac{\epsilon^2}{3} \right) \right. \right. \\ & \left. \left. + 2F \left(\frac{\epsilon r}{3} - r^2 \right) \exp(\epsilon/r) \right] \exp(-4\pi r^3 N/3) dr \right\} \mu_0 \end{aligned} \quad (7.25)$$

Comparison of μ/μ_0 computed from (7.25) with $1 - (P/N)$ appearing in (7.15) over wide ranges of conditions again reveals an excellent correspondence and further substantiates the Bjerrum-Fuoss theory. Since calculations employing the latter are so much simpler it is expedient to regard the cruder theory as an accurate approximation to the more refined one. This practice will be followed from now on.

VIII. PHENOMENA ASSOCIATED WITH ION PAIRING IN SEMICONDUCTORS

In this section we shall discuss some of the phenomena which are to be expected in semiconductors when ion pairing takes place. At the time of writing several of these phenomena have been investigated quantitatively in germanium and casually in silicon. A report on these studies will be given in the later sections of this paper.

In the meantime it is fitting to inquire into the peculiarities which arise because a semiconducting medium rather than a dielectric liquid is involved. The possible means of detecting and measuring ion pairing in semiconductors are numerous, and many of them do not have counterparts in aqueous solution. This implies that a host of new phenomena are to be expected, many of which are peculiar to semiconductors.

Some distinctions between semiconductors and liquids are apparent at once. Thus ions are not always mobile in semiconductors at temperatures where ion pairing is pronounced. Lithium is exceptional in this respect, being mobile in germanium and silicon down to very low temperatures. In fact ion pairing has been observed in germanium containing lithium down to dry ice temperatures, and even below. Another difference is the low dielectric constant of semiconductors as compared with water. Furthermore, in semiconductors, charge balance need not be maintained by the ions themselves, but may be effected by the presence of holes or electrons. Although charged the latter entities need not be considered in pairing processes since, as particles, they possess effective radii of the order of their thermal wavelengths which may exceed 20 Angstroms at the temperatures involved. At these distances very little coulomb binding energy would be available. Under certain rare conditions the screening effect of these mobile carriers may make some contribution. This may be particularly the case when *relaxation processes* (to be discussed later) are carried out in poorly compensated specimens of semiconductor, since such processes involve phenomena between ions separated by large distances.

A very obvious distinction is the fact that ions in a semiconductor occupy a lattice, and cannot therefore move through a continuum of positions, as in the case of liquid solutions. Furthermore the lattice may introduce elastic strain energy into the binding energy of a pair. This influence will alter the value of a , the distance of closest approach, when the latter is chosen so as to achieve the best fit between theory and experiment. As the extent of pairing is extremely sensitive to the magnitude of a , its measurement provides a useful tool for exploring the state of strain in the neighborhood of an isolated impurity. We shall demonstrate

this application later in connection with the strain in the neighborhood of a substitutional boron in germanium.

Aside from its bearing on the minimum distance a , the existence of the lattice will be ignored in the following considerations.

The values of a , typical of semiconductors, are generally of the order of 2 Angstroms as against 6 to 8 Angstroms for ions in liquids. This results from the fact that liquid ions are generally solvated. The consequence to be expected, and indeed found, is that ion pairing will be far more pronounced in semiconductors than in liquids of comparable dielectric constant.

The fact that ions have limited mobilities in semiconductors can be turned to advantage by choosing a system such as lithium and boron in silicon in which only one species of ion, in the case mentioned, lithium, is mobile. Under these conditions it is possible to obviate the clustering phenomenon, mentioned previously, which appears in liquids at high ion concentrations. Clustering is prevented because the immobile ions are uniformly distributed in a random manner, having been grown into the crystals at high temperature where pairing and related processes are unimportant. The obvious complications attending cluster formation can therefore be avoided.

Of course, mobility, being limited to a single species of ion is also an advantage in the theory of the transport phenomena, in such systems.

It is convenient to list some of the effects due to pairing which are to be expected in semiconductors. We do so in the following compilation.

(A) Equilibrium Phase Relations

From (6.2) it is apparent that the pairing equilibrium should affect the solubility of lithium in silicon. The same must be true for germanium doped with an acceptor. Although such effects probably occur, they are accompanied by influences arising from the other possible equilibria. As a result the situation is somewhat complex and it is not easy (see Appendix A) to produce experimental conditions under which pairing will be evident. For this reason quantitative investigations along these lines have not yet been attempted.

(B) Variation of Energy Levels

When an ion pair is formed of a donor and acceptor, both the donor and acceptor levels are altered. Thus the proximity of the negative acceptor ion increases the difficulty of return to the donor state for an electron, (i.e. the donor level is raised). Likewise the acceptor level is

lowered. In ion pairs it is in fact to be expected that the donor level will be moved up into the conduction band and the acceptor level down into the valence band.* This change in energy level structure should be apparent in Hall coefficient measurements at low temperature. Experiments of this sort have been conducted and are reported in this paper. Under certain conditions this phenomenon may be useful for the elimination of trapping⁴⁶ levels from the forbidden gap.

(C) Change of Carrier Mobility

Ion pairs possess dipolar fields, and consequently, scattering cross-sections very much smaller than those of point charges. The addition of lithium to a sample under such conditions that more than half the added lithium becomes paired should therefore *increase* rather than *decrease* the mobility of holes. The latter effect is the one to be expected in the absence of pairing. In other words not only carriers but also the scatterers are removed by compensating the acceptor with donor. Experiments of this sort have been performed. They are described later in this paper. Since they allow us to measure the degree of pairing with good accuracy they have been very valuable in validating the theory, and also in exploring the nature of the potential function in the neighborhood of an isolated acceptor.

(D) Relaxation Times

A semiconductor containing unpaired donors and acceptors at one temperature can be cooled to a lower temperature, and the impurities should then pair. If the temperature is lowered sufficiently, the pairing process will be slow enough to be followed, kinetically, by observing any parameter (such as carrier mobility) sensitive to pairing. Experiments of this sort have been performed and will be described later.

The process of pairing can be characterized by a calculable relaxation time, which depends on the acceptor concentration, the diffusivity of the mobile donor, the dielectric constant, and the charges on the ions among other things. The measured time can therefore be used as a means of determining any one of these parameters.

(E) Diffusion

It is evident that pairing should reduce the diffusivity of a mobile donor. Studies of diffusion in the presence of an immobile acceptor should

* A rough calculation indicates that about 0.5 e.v. would be required to place an additional electron on an ion pair.

therefore reveal the action of pairing. Experiments of this sort have been performed and will also be described in this paper.

The reduction in the diffusivity of a donor such as lithium may be desirable in certain places.

(F) Direct Transport

Diffusion studies suffer from the defect that ion pairing produces a concentration dependent diffusivity. (See Appendix B). For this reason a very desirable measurement would involve determining the amount of a mobile donor like lithium transported by an electric field through a uniformly saturated specimen of semiconductor. This flux, together with information concerning the level of saturation, should provide a direct measure of the mobility of lithium under homogeneous conditions. Formula (7.15) or its refinement (7.25) could then be applied directly to the results.

The above list is by no means complete, for there are still other techniques available for measurement, for example nuclear and paramagnetic resonance. Enough has been given however to indicate the wide range of phenomena which ion pairing in solids can affect. In liquids, only A and F are of any consequence. It is important to realize that not only do these phenomena serve as tools for the study of ion pairing, but that ion pairing, when properly understood, can serve as a tool for the study of the phenomena themselves.

IX. PAIRING CALCULATIONS

The evaluation of Ω according to (7.14) presents somewhat of a problem because the integral must be arrived at numerically. Fortunately, the literature contains tables⁴⁷ of the integral in what amounts to dimensionless form. The transformation

$$\xi = q^2/\kappa kTr \quad (9.1)$$

is introduced and then Ω is shown to be given by

$$\Omega = 4\pi[q^2/\kappa kT]^3 Q(\alpha) \quad (9.2)$$

where

$$\alpha = q^2/\kappa kTa \quad (9.3)$$

and $\log_{10} Q(\alpha)$ is tabulated in Table III.

In a specimen in which the numbers of donors and acceptors are un-

TABLE III

α	$\log_{10} Q(\alpha)$	α	$\log_{10} Q(\alpha)$
2.0	$-\infty$	18.0	2.92
2.5	-0.728	20.0	3.59
3.0	-0.489	25.0	5.35
4.0	-0.260	30.0	7.19
5.0	-0.124	35.0	9.08
6.0	0.016	40.0	11.01
7.0	0.152	45.0	12.99
8.0	0.300	50.0	14.96
9.0	0.470	55.0	16.95
10.0	0.655	60.0	18.98
12.0	1.125	65.0	21.02
14.0	1.680	70.0	23.05
16.0	2.275	75.0	25.01
		80.0	27.15

equal* (7.10) may be written as

$$\frac{P}{(N_A - P)(N_D - P)} = \Omega \quad (9.4)$$

where N_A and N_D are, respectively, the total densities of acceptors and donors.

This equation has the following solution for P/N_D , the fraction of donors paired.

$$\frac{P}{N_D} = \frac{1}{2} \left(1 + \frac{1}{\Omega N_D} + \frac{N_A}{N_D} \right) - \sqrt{\frac{1}{4} \left(1 + \frac{1}{\Omega N_D} + \frac{N_A}{N_D} \right)^2 - \frac{N_A}{N_D}} \quad (9.5)$$

Inspection of (9.5) reveals that for given N_A and Ω , P/N_D is a decreasing function of increasing N_D .

Very often, P/N_D is measured in an experiment, and from this it is desired to calculate a , the distance of closest approach. For such purposes the form (9.5) is not very convenient. In fact an entirely different procedure is to be preferred. Suppose P/N_D is denoted by θ , and θ is substituted into (9.4), into which (9.2) has been inserted. We obtain

$$\log_{10} Q(\alpha) = \log_{10} \left[\frac{1}{4\pi} \left(\frac{\kappa k T}{q^2} \right)^3 \frac{\theta}{(N_A - \theta N_D)(1 - \theta)} \right] \quad (9.6)$$

A knowledge of θ thus suffices to determine $\log_{10} Q(\alpha)$, from which, in turn, α can be determined by interpolation in Table III. Then (9.3) can be used for the evaluation of a .

* This is a situation which cannot arise in liquids, since there, charge balance must be maintained by the ions themselves. It can occur when the ions are of different charge, but then things are complicated by the formation of triplets, etc., in addition to pairs.

TABLE IV

$T^{\circ}\text{K}$	$\Omega (\text{cm}^3)$	$T^{\circ}\text{K}$	$\Omega (\text{cm}^3)$
100	2.2×10^2	400	2.3×10^{-17}
150	6.45×10^{-7}	500	1.54×10^{-18}
200	3.42×10^{-11}	600	3.0×10^{-19}
225	1.28×10^{-12}	700	1.03×10^{-19}
250	8.79×10^{-14}	800	4.7×10^{-20}
300	1.61×10^{-15}		

Experiments which will be described later indicate that in germanium, gallium and lithium can approach as close as 1.7×10^{-8} cm. Using this value of a , and $\kappa = 16$, $q = 4.77 \times 10^{-10}$ statcoulombs, the values of Ω appearing in Table IV were computed from (9.2)

With these values, P/N_D , the fraction of donors paired can be computed from (9.5) as a function of temperature and N_A for the simplest case, i.e., the one for which $N_A = N_D$. Fig. 15 contains plots showing these dependences. It must be remembered that all other things remaining the same P/N_D will be greater than the values shown in Fig. 15 when $N_D < N_A$.

A rather important integral to which reference shall be made later is

$$I(r_2, r_1) = \int_{r_1}^{r_2} x^2 \exp(q^2/\kappa k T x) dx \quad (9.7)$$

The integral appearing in (7.14) is a special case of (9.7) with $r_1 = a$, and $r_2 = b$. $I(r_2, r_1)$ has been evaluated over a considerable range. To facilitate matters the transformation

$$x = (q^2/\kappa k T) \lambda \quad (9.8)$$

has been employed. In this notation r_1 and r_2 transform to ρ_1 and ρ_2 , and

$$I(r_2, r_1) = (q^2/\kappa k T)^3 \int_{\rho_1}^{\rho_2} \lambda^2 \exp(1/\lambda) d\lambda = (q^2/\kappa k T)^3 i(\rho_2, \rho_1) \quad (9.9)$$

Figs. 16 and 17 contain plots of $i(\rho_2, 0.05)$ out to $\rho_2 = 5$. The choice of ρ_1 equal to 0.05 was rather unfortunate since for $\kappa = 16$, and $T = 300^{\circ}\text{K}$ it corresponds to $\rho_1 = 2.5 \times 10^{-8}$ cm. Since acceptors like gallium possess values in respect to lithium as low as 1.7×10^{-8} cm $i(\rho_2, 0.05)$ is not much use in these cases. The choice 0.05 was made before the experimental data on gallium was available. Below we shall describe a method for extending $i(\rho_2, \rho_1)$ to cases where r_1 is less than 2.5×10^{-8} cm

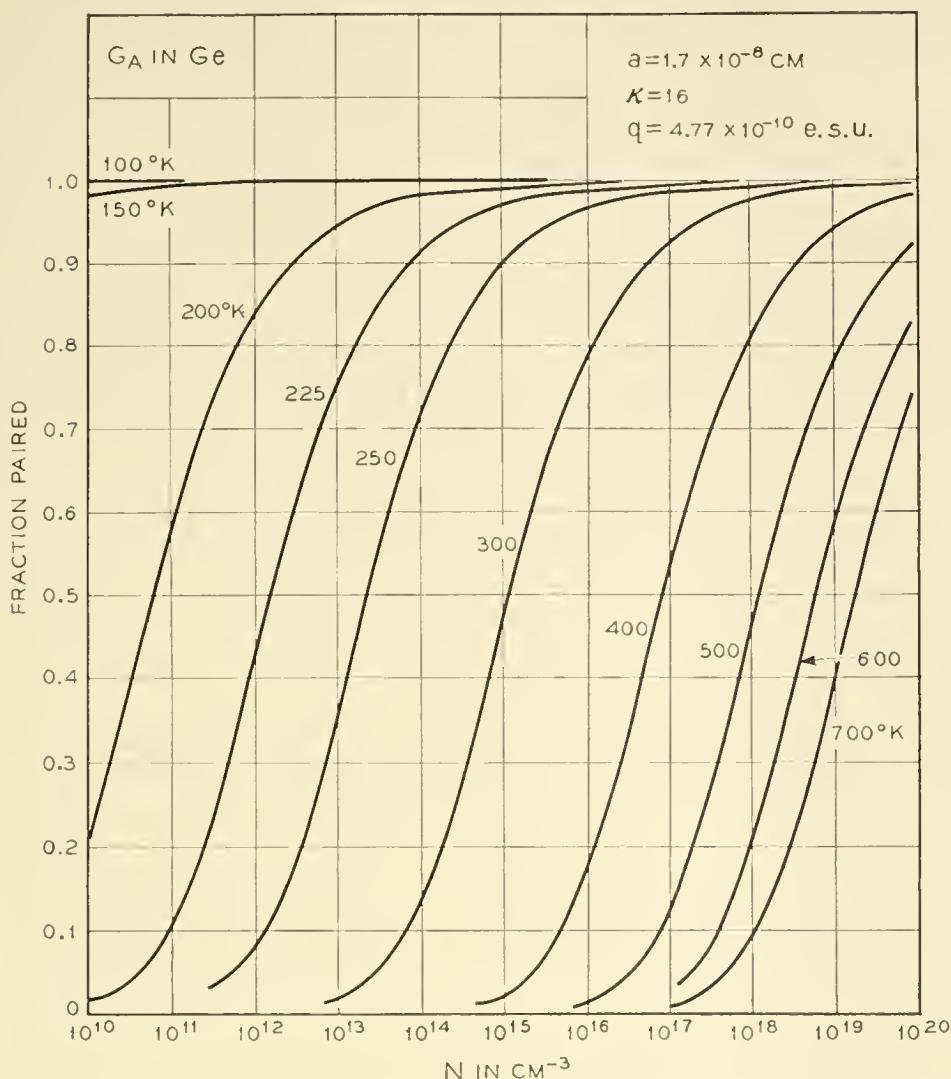


Fig. 15 — Fraction of ions paired, assuming equal densities of positive and negative ions, calculated as a function of temperature and concentration from equation (9.5). The situation illustrated might apply to gallium and lithium in germanium in view of the choice of a and κ .

Fig. 16 covers the range from $\rho_2 = 0.05$ to 0.08 and involves a logarithmic scale because of the sharp variation of i in this range. (This points up the sensitivity of the degree of pairing to the magnitude of a .) Fig. 17 extends the curve to $\rho_2 = 5$. When ρ_2 exceeds 5, $i(\rho_2, 0.05)$ can be obtained from the formula

$$i(\rho_2, 0.05) = 3865 + \frac{(\rho_2)^2}{2} + \frac{(\rho_2)^3}{3} \quad (9.10)$$

In order to determine $i(\rho_2, \rho_1)$ when $\rho_1 \geq 0.05$, the following formula may be used.

$$i(\rho_2, \rho_1) = i(\rho_2, 0.05) - i(\rho_1, 0.05) \quad (9.11)$$

Finally for cases in which $\rho_1 < 0.05$, Table III can be used. Thus

$$i(\rho_2, \rho_1) = Q(1/\rho_1) - Q(20) + i(\rho_2, 0.05) \quad (9.12)$$

where $1/\rho_1$, and 20 are α values in Table III.

X. THEORY OF RELAXATION

In Section VIII attention was drawn to the fact that ion pairing in semiconductors can be made to occur slowly enough so that its kinetics can be followed. It is possible to characterize these kinetics by a relaxation time τ , which we shall endeavor to calculate in the present section.

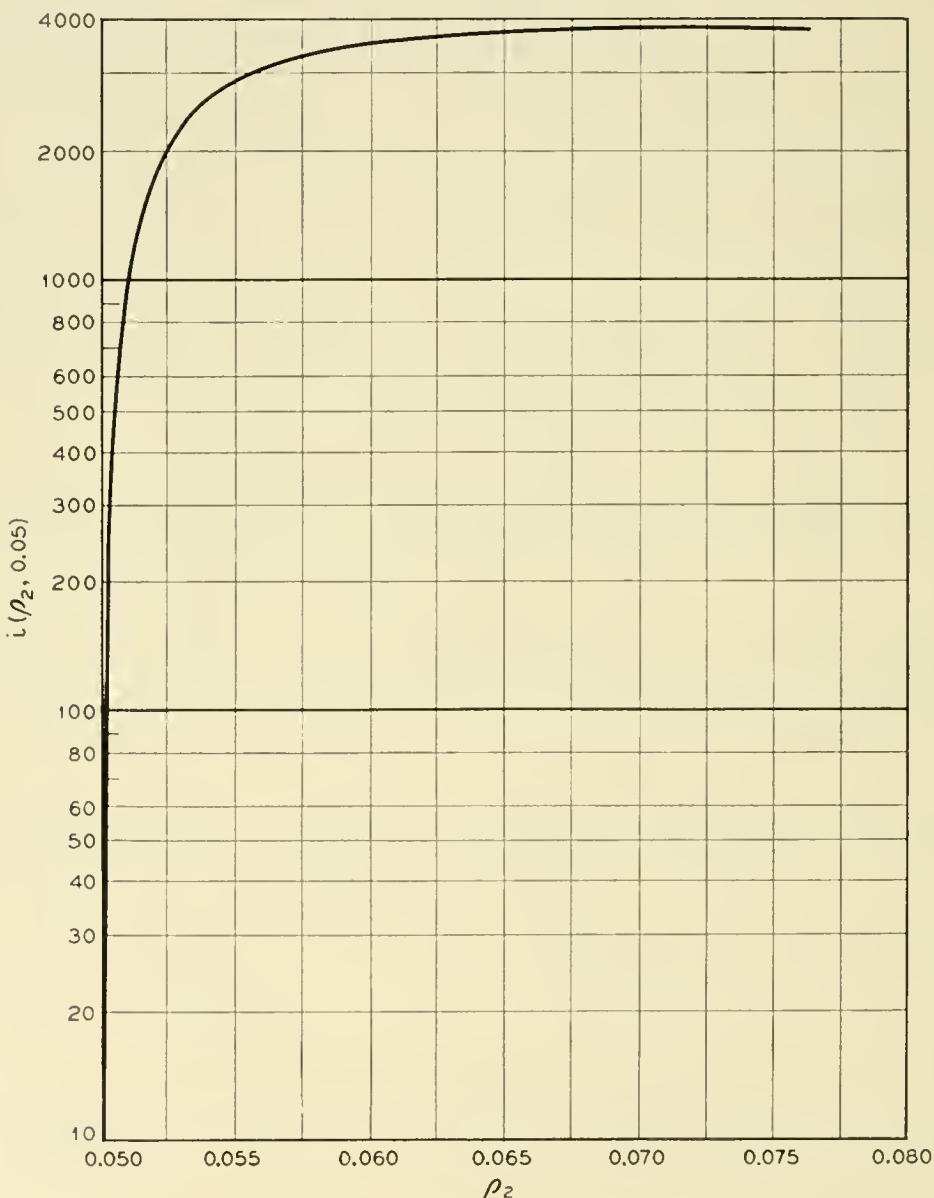


Fig. 16 — Plot, for small values of ρ_2 of $i(\rho_2, 0.05)$ from (9.9).

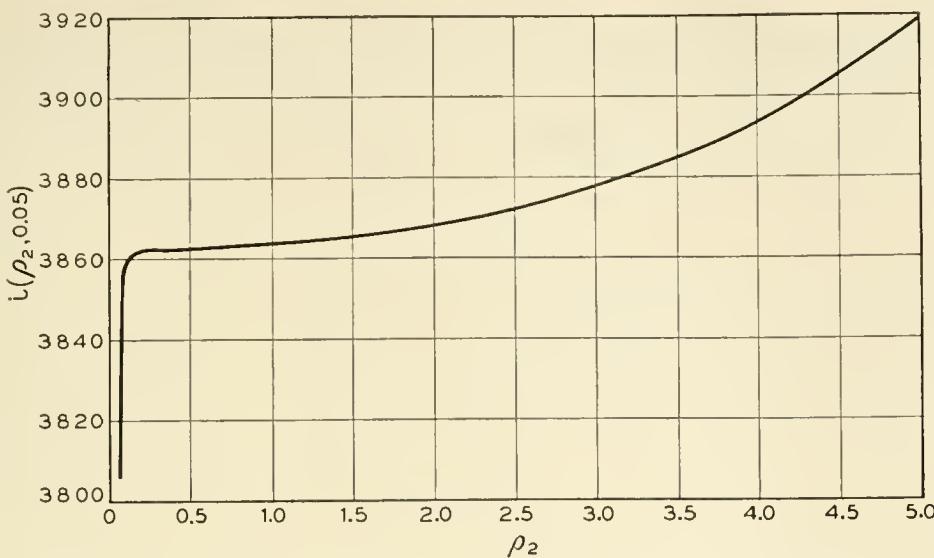


Fig. 17 — Plot, for larger values of ρ_2 , of $i(\rho_2, 0.05)$ from (9.9).

Suppose a system is first maintained at a temperature high enough to prevent pairing, and then, at an instant designated as zero time, is suddenly chilled to a temperature at which pairing takes place. One thereby has a system which would normally contain pairs but which finds itself with donors and acceptors which are uniformly and randomly distributed. Since the donors are assumed mobile, a process ensues whereby they drift toward acceptors until an equilibrium is established in which each acceptor develops an atmosphere of donors with density $c(r)$, given by (7.7).

This final state in which the atmosphere is fully developed is the paired state characteristic of the lower temperature. The relaxation time to be defined must measure the interval required for the near completion of the above process.

In order to acquire physical feeling for the phenomenon, we begin with some simple considerations. In particular a system will be dealt with containing equal numbers of positive and negative ions. This restriction can be lifted later.

Now, to a first approximation the pairing phenomenon may be regarded as a trapping process in which mobile, positive donor atoms are captured by the negative acceptors. Thus, suppose each acceptor is imagined to possess a sphere of influence of radius R , beyond which its force field may be considered negligible, and inside which a positive ion is to be regarded as captured. This picture immediately emphasizes certain subtleties which require discussion before further progress can be made.

In the crudest sense one might reason that the probability of an encounter between a positive ion and a negative trap would depend on the

product of the densities of both. These densities must be equal because when a positive ion is trapped the resulting ion pair is neutral so that a trap is eliminated simultaneously. If these equal densities are designated by n , we arrive at the second order rate law

$$-\frac{dn}{dt} = k_2 n^2 \quad (10.1)$$

where k_2 is a suitable constant, and t is time.

This law would be perfectly valid if the mean free path of a mobile positive ion were large compared to the distance between ions and the probability of sticking on a first encounter were small. The trapping cross-section rather than the movement prior to trapping would determine the trapping rate. In this case the rate would certainly depend on the concentrations of both the traps and the ions being trapped.

On the other hand, in our case, not only is the mean free path of a positive ion much smaller than the distance between ions, but the sticking probability is high. A given ion must *diffuse* or make many random jumps before encountering a trap and upon doing so is immediately captured. Therefore, the rate of reaction is diffusion controlled.

Because of the random jump process a given mobile ion is most likely to be captured by its *nearest neighbor* during the first half of relaxation, and relative to the degree of advancement of the trapping process, the density of traps may be considered constant. This leads to first order kinetics rather than second,* i.e., to

$$-\frac{dn}{dt} = k_1 n \quad (10.2)$$

where n is the density of untrapped ions.

By definition k_1 is the fraction of ions captured in unit time, i.e., the probability that one ion will be captured per unit time. Its reciprocal must be the average lifetime of an ion. This lifetime

$$\tau = \frac{1}{k_1} \quad (10.3)$$

shall be defined as the *relaxation time* for ion pairing. A rough calculation of τ can be made quickly. Thus, suppose that the initial concentrations of donors and acceptors are equally N . About each fixed acceptor can be described a sphere of volume, $1/N$. On the average this sphere should be occupied by one donor which according to what has been said above, will eventually be captured by the acceptor at the center. In the mind, all

* The phenomenon stems from the fact that first and second order processes are almost indistinguishable during the first half of the reaction, but also from the fact that the diffusion control prevents the process from being a true second order one, although its departure from second order may be small.

the spheres can be superposed so that an assembly of donors N in number is contained in the volume $1/N$, at the density N^2 . The problem of relaxation is then the problem of diffusion of these donors to the sink of radius R , at the center of the volume. The bounding shell of the sphere may be considered impermeable, thus enforcing the condition that each donor shall be trapped by its nearest neighbor. Since the diffusion problem has spherical symmetry the radius, r , originating at the center of the sink at the origin may be chosen as the position coordinate. At $r = R$, the density, ρ , of diffusant may be considered zero. The radius, L , of the volume, $1/N$, is so large compared to R , that in the initial stages of diffusion L may be regarded as infinite.

In spherical diffusion to a sink from an infinite field, a true steady state is possible, and this steady state is quickly arrived at when the radius, R , of the sink is small.⁴⁸ Under this condition concentration is described by

$$\rho = A - \frac{B}{r} \quad (10.4)$$

where A and B are constants. Furthermore at early times n is still N , the initial concentration at $r = L \approx \infty$, so that

$$\rho(\infty) = N^2 \quad (10.5)$$

In addition we know that

$$\rho(R) = 0 \quad (10.6)$$

These boundary conditions suffice to determine A and B in (10.4), and yield

$$\rho = N^2 \left[1 - \frac{R}{r} \right] \quad (10.7)$$

Now the rate of capture ($-(dn/dt)$ in (10.2)) is obviously measured by the flux of ions into the spherical shell of area, $4\pi R^2$, which marks the boundary of the sink. This flux is given according to Fick's law⁴⁹ by

$$4\pi R^2 D_0 \left(\frac{\partial \rho}{\partial r} \right)_{r=R} = - \frac{dn}{dt} \quad (10.8)$$

where D_0 is the diffusivity of the donor. Substituting (10.7) into (10.8) yields

$$4\pi N^2 R D_0 = - \frac{dn}{dt} \quad (10.9)$$

During the initial stages of trapping the right side of (10.2) may be

written as $k_1 N$, i.e.,

$$k_1 N = - \frac{dn}{dt} \quad (10.10)$$

Equating the left sides of (10.9) and (10.10) gives

$$k_1 = 4\pi NRD_0$$

or

$$\tau = \frac{1}{k_1} = \frac{1}{4\pi NRD_0} \quad (10.11)$$

It now remains to choose a value for the capture radius, R . A reasonable guess may be made as follows: Around each acceptor there is a coulomb potential well of depth

$$V = -q^2/\kappa r \quad . \quad (10.12)$$

Since the average thermal energy is kT , it seems reasonable to regard an ion as trapped when it falls to a depth kT in this well. Thus, inserting kT on the left of (10.12) and R for r on the right leads to

$$R = q^2/\kappa kT \quad (10.13)$$

and upon substitution in (10.11) we obtain

$$\tau \approx \frac{\kappa kT}{4\pi q^2 N D_0} \quad (10.14)$$

This result, obtained by crude reasoning, is actually quite close to the more rigorous value derived below. Furthermore, the above derivation is useful in providing insight into the physical meaning of the relaxation time.

The chief difficulty with the preceding lies in the arbitrary choice of R , and is a direct consequence of the long range nature of coulomb forces. Another difficulty arises because the distribution of donors about acceptors is eventually specified by (7.7) so that at $r = R = q^2/\kappa kT$

$$\frac{\partial c}{\partial r} = -\frac{h}{e} \left\{ \frac{\kappa kT}{q^2} \right\} \quad (10.15)$$

Since this slope has a negative value the trap exhibits some aspects of a source rather than a sink which could only produce a positive concentration gradient. This last objection will not be serious when h is very small since, then the final value of $c(r)$ beyond $r = q^2/\kappa kT = R$ will be effectively zero, as would be required for a perfect sink.

The last point raises still another question: What happens when the sink is not perfect, i.e. where the equilibrium state does not involve complete pairing?

All these difficulties can be removed by a more sophisticated treatment of the diffusion problem. Thus, retain the sphere of volume, $1/N$, enclosing N donors at the density N^2 . However, the equations of motion of these donors are altered to account for the fact that besides diffusing they drift in the field of the acceptor at the origin. Thus the flux density of donors will be given by

$$\begin{aligned} J^*(r, t) &= -D_0 \left\{ \frac{\partial \rho}{\partial r} + \left\{ \frac{q^2}{\kappa k T r^2} \right\} \rho \right\} \\ &= -D_0 \left\{ \frac{\partial \rho}{\partial r} + \frac{R}{r^2} \rho \right\} \end{aligned} \quad (10.16)$$

where R has been substituted for $q^2/\kappa k T$. Equation (10.16) is obtained by adding to the diffusion component,

$$-D_0 \frac{\partial \rho}{\partial r}$$

of the flux density, the drift component,

$$-\frac{\mu_0 q}{\kappa r^2} \rho,$$

where μ_0 is the mobility of a donor ion and $-q/\kappa r^2$ is the field due the acceptor at the origin. The Einstein relation⁵⁰

$$\mu_0 = q D_0 / k T \quad (10.17)$$

has also been used to replace μ_0 with D_0 .

The spherical shell bounding the volume, $1/N$, of radius

$$L = \left(\frac{3}{4\pi N} \right)^{1/3} \quad (10.18)$$

is regarded as impermeable, so we obtain the boundary condition

$$J^*(L, t) = 0. \quad (10.19)$$

Furthermore an arbitrary inner boundary, $r = R$, is no longer defined but use is made of the real boundary, $r = a$, i.e., the distance of closest approach, at which is applied the condition

$$J^*(a, t) = 0 \quad (10.20)$$

As before, the initial condition may be expressed as

$$\rho = N^2 \quad t = 0 \quad a < r < L \quad (10.21)$$

The continuity equation,⁵¹ in spherical coordinates takes the form

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left\{ r^2 J^* \right\} = - \frac{\partial \rho}{\partial t} \quad (10.22)$$

Substitution of (10.16) into (10.22) gives, finally,

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left\{ r^2 \frac{\partial \rho}{\partial r} + R\rho \right\} = \frac{1}{D_0} \frac{\partial \rho}{\partial t} \quad (10.23)$$

Equations (10.23), (10.21), (10.20) and (10.19) form a set defining a boundary value problem, the solution of which is $\rho(r, t)$, from which, in turn, $J^*(r, t)$ can be computed. It then remains to compute (dn/dt) in (10.2) from J^* . The former is not simply $4\pi R^2 J^*$ (as in (10.8)) because now J^* is not defined unambiguously, being a function of r . $J^*(R, t)$ might be employed but then the method is no less arbitrary than the simple one described above.

Fortunately, nature eliminates the dilemma. It is a peculiarity of spherical diffusion, when the sink radius is much smaller than the radius of the diffusion field, that after a brief transient period, $4\pi r^2 J^*(r)$, except near the boundaries of the field, becomes practically independent of r , and depends only on t . This feature is elaborated in Appendix C. Since in our case the radius of the field is of order, L , and the effective radius of the sink is of order, R , and $L \gg R$, it may be expected that this phenomenon will be observed. In fact its existence has been assumed previously in the derivation of (10.4).

Under such conditions it does not matter how the radius of the sink is defined so long as $4\pi R^2$ is multiplied by $J^*(R)$ and not the value of J^* at some other location.

The boundary value problem, (10.23), (10.21), (10.20), (10.19) is solved in Appendix C, and it is shown there that the value of $4\pi r^2 J^*(r)$ obtained after the transient has passed is closely approximated by

$$4\pi r^2 J^*(r) = - \frac{4\pi q^2 N^2 D_0}{\kappa k T} e^{-t/\tau} \quad (10.24)$$

with

$$\tau = \frac{\kappa k T (N - M)}{4\pi q^2 N^2 D_0} \quad (10.25)$$

where

$$M = 1/4\pi \int_a^L r^2 \exp [q^2 / \kappa k T r] dr \quad (10.26)$$

The close connection between M defined by (10.26) and h defined by (7.17) is apparent. Thus in (7.17) when $r = L$, $\exp[-4\pi r^3 N/3]$ is e^{-1} , and for larger values of r this exponential quickly forces the convergence of the integral. Therefore the values of h and M will be almost equal. This is not surprising since they are meant to be the same thing, i.e., the average concentration, $c(\infty)$, of donors at infinite distance in the equilibrium atmosphere of an acceptor. Both quantities are computed so as to conserve charge in this atmosphere.

At large values of N , M proves to be much smaller than N so that (10.25) reduces to (10.14), validating the crude treatment, for τ in (10.24) is obviously the relaxation time. This is easily seen by writing

$$-\frac{dn}{dt} = -4\pi r^2 J^*(r) = \frac{4\pi q^2 N^2 D_0}{\kappa k T} e^{-t/\tau} \quad (10.27)$$

from which one derives by integration

$$n = M + (N - M)e^{-t/\tau} \quad (10.28)$$

According to (10.28) at $t = 0$, $n = N$, the correct initial density for unpaired ions. At $t = \infty$, $n = M$, also the correct density, i.e., the density at large values of r , when equilibrium is achieved. Obviously τ plays the role of the relaxation time, since by differentiation of (10.28)

$$-\frac{d(n - M)}{dt} = \frac{(n - M)}{\tau} \quad (10.29)$$

which is to be compared with (10.2) and (10.3).

Values of M can be computed using formulas (9.10), (9.11), and (9.12) and Figs. 16 and 17 since the integral in (10.26) is one of the i integrals. Fig. 18 shows some values of M , computed in this way for the temperatures 206° , 225° , 250° , and 300° K, for a semiconductor where the value of $a = 2.5 \times 10^{-8}$ cm, $\kappa = 16$, and $q = 4.77 \times 10^{-10}$ statcoulombs. The plots are of M versus N . Note that the values of M are generally much less than N , the disparity increasing with lower temperatures and larger N .

It is also possible to calculate τ for the above system in its dependence upon N and T . To do this the value of D_0 must be known as a function of temperature. Fuller and Severiens⁵² have measured the diffusivities of lithium in germanium and silicon down to about 500° K. These data plot logarithmically against $1/T$ as excellent straight lines. In Fig. 19, we show an extrapolation of the line for lithium in germanium down to the neighborhood of 200° K. From this figure it is possible to read values of

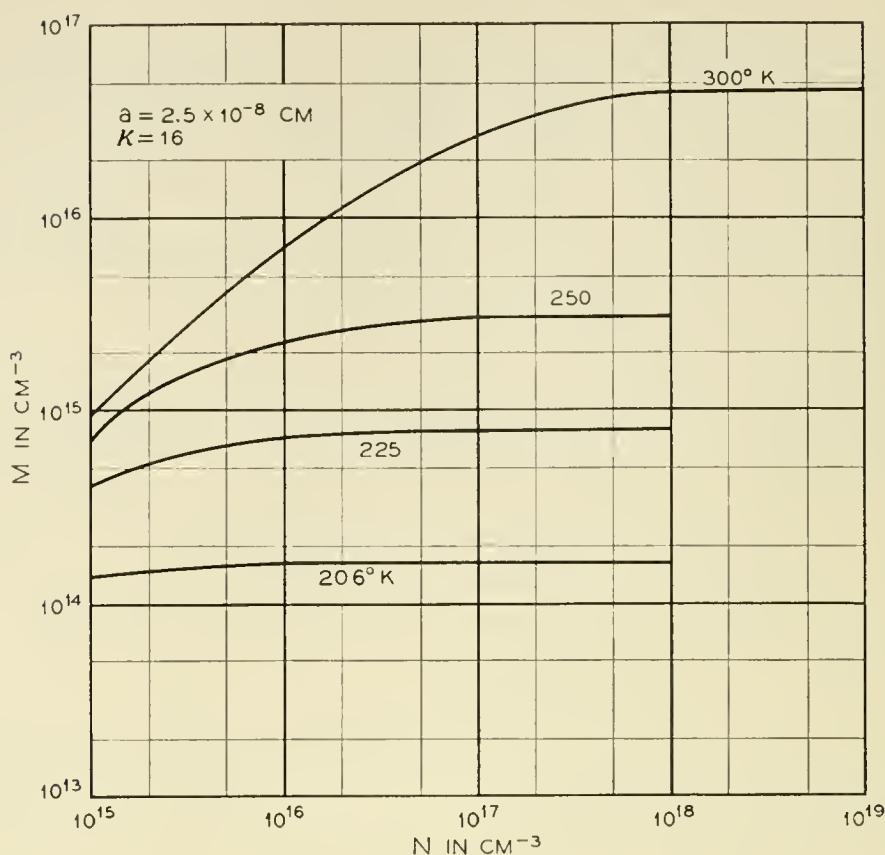


Fig. 18 — Dependence of constant M defined by (10.26) on temperature and concentration, for particular values of a and κ .

D_0 for germanium to which the system of Fig. 18 refers, since κ has been chosen at 16.

Using Figs. 18 and 19, Fig. 20 was computed. It shows τ plotted in seconds versus N for the same temperatures appearing in Fig. 18. These curves show that at values of N as low as 10^{15} cm^{-3} relaxation times are short enough to be observable down to 200°K, being at the most some 50 hours in extent. The value of N makes a big difference.¹ For example at 200°K the relaxation time is only 4 minutes with $N = 10^{18} \text{ cm}^{-3}$. Presumably, at 10^{18} cm^{-3} , relaxation could be observed down to much lower temperatures.

It is interesting to note that insofar as M hardly appears in τ , the latter is independent of the distance of closest approach, a . Since a is to some extent empirical this is a fortunate circumstance, and the measurement of τ may provide an accurate means of determining, N , D_0 , κ , or q , whichever parameter is regarded as unknown. Furthermore κ as a macroscopic parameter has real meaning in τ since the forces involved may be regarded as being applied over the many lattice parameters separating the drifting donor from its acceptor.

This section will be closed by indicating how the restriction to systems containing equal numbers of donors and acceptors might be lifted. Thus, suppose N_A exceeds N_D . Then there will be $N_A - N_D$ mobile holes maintaining charge neutrality. To a first approximation these will screen the $N_A - N_D$ uncompensated acceptor ions so that the N_D donors will see effectively only N_D acceptors. Thus in first approximation τ can be computed for this system by replacing N in the preceding formulas by N_D .

Of course it is possible that there will be a further effect. Thus the mobile holes will probably shield some of the compensated acceptors as well. This will lead to a further (probably small) reduction in τ , over and above that obtained by replacing N by N_D . We shall not go into this in the present paper, because in most of the experiments performed N_D was near N_A . In the few exceptions the crude correction, suggested above, can be used.

XI. INVESTIGATION OF ION PAIRING BY DIFFUSION

Most of the theoretical tools required for the study of ion pairing have now been provided, and attention will be turned to experiments which

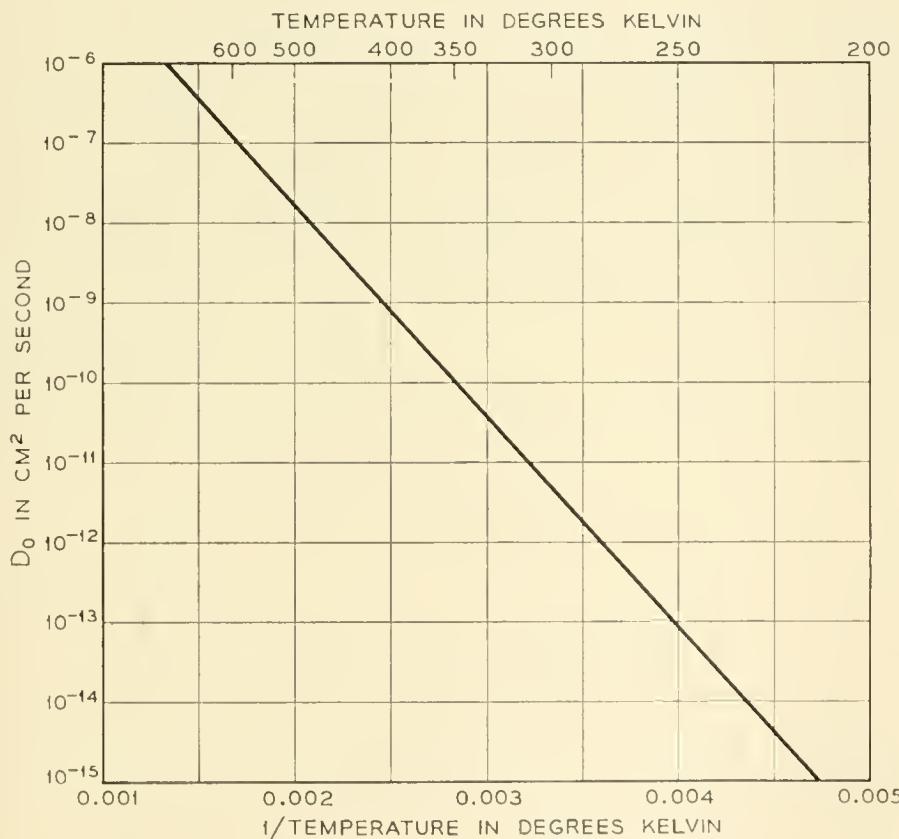


Fig. 19 — Diffusivity of lithium in germanium extrapolated from the data of Fuller and Severiens.

have been performed in this field. A fairly large group of these exist, and it remains to describe them in detail. We shall begin with the study of the diffusion of lithium in *p*-type germanium.

At the outset a matter having to do with the *diffusion potential* demands attention. This is the potential which arises, for example, in *p*-type material, because the mobility of a hole is so much greater than the mobility of a lithium ion. In consequence, holes diffuse into regions containing high concentrations of lithium more rapidly than lithium ions can diffuse out to maintain space charge neutrality. As a result such re-

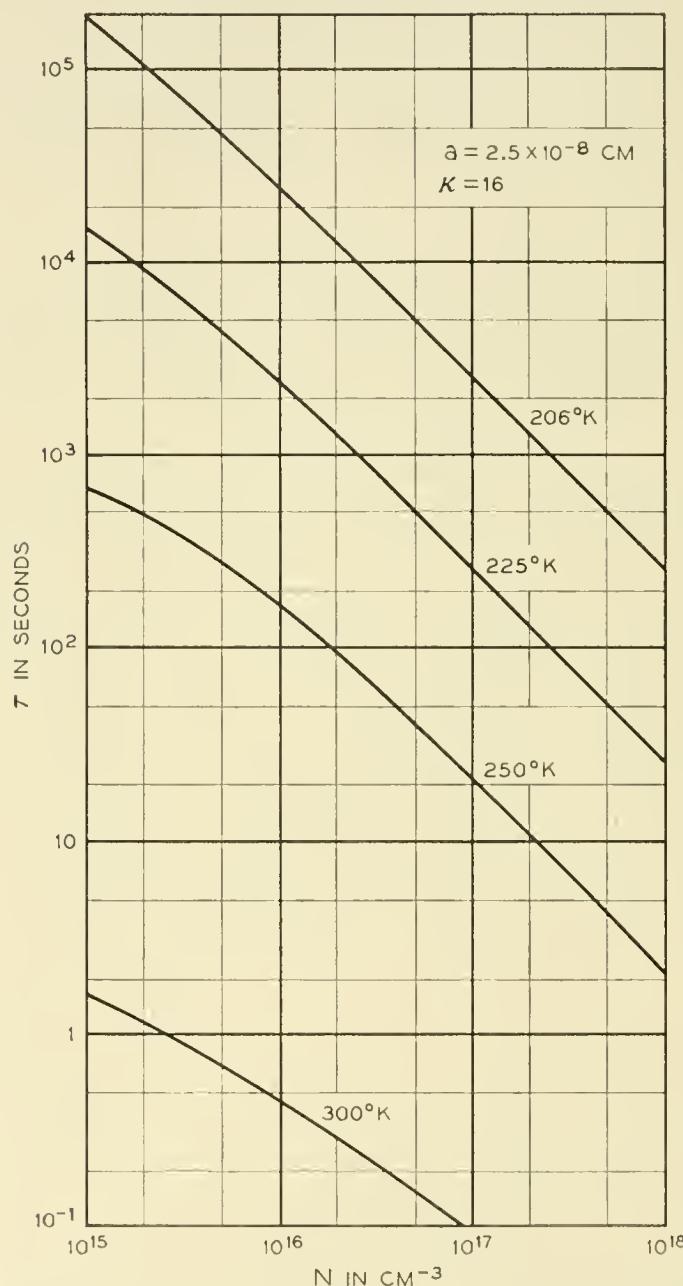


Fig. 20 — Relaxation time as a function of temperature and concentration computed from equation (10.25) using the data of Figs. 18 and 19.

gions develop positive potentials and a field exists tending to expel lithium. This causes the lithium to drift as well as diffuse so that Fick's law⁴⁹ is no longer valid.

The most that can be done toward the elimination of diffusion potentials is to minimize them so that no local space charge exists. At equilibrium, this corresponds to the condition⁵³

$$N_D - N_A = 2n_i \sinh(qV/kT) \quad (11.1)$$

where V is the local electrostatic potential. It is always permissible to assume that fast moving electrons and holes are in equilibrium relative to diffusing ions. If a material which is *p*-type everywhere is being considered, (11.1) can be simplified to

$$N_A - N_D = n_i \exp[-qV/kT] \quad (11.2)$$

In Appendix D it is proved that (11.2) will be valid everywhere within a region where N_A is constant and greater than N_D , provided that N_D does not fluctuate through ranges of the order N_A in a distance less than

$$\ell = \sqrt{\frac{\pi \kappa k T}{q^2 N_A}} \quad (11.3)$$

Under most conditions of experiment ℓ will be of the order of 10^{-5} cm. Unfortunately many of the experiments described in this section (particularly those performed at 25°C.) involve diffusion layers as thin as 10^{-6} cm. As a result space charge will exist and the diffusion potential will not always be minimized. Even if it is minimized so that (11.2) is satisfied the residual field will still aid diffusion and lead to higher apparent diffusivities. Therefore the effect cannot be ignored even when minimization has been achieved.

In the absence of space charge the drift component of flux density due to the field is easily computed. It will be given by

$$-\mu \frac{\partial V}{\partial x} N_D \quad (11.4)$$

According to (11.2)

$$-\frac{\partial V}{\partial x} = \frac{kT}{q(N_A - N_D)} \frac{\partial N_D}{\partial x} \quad (11.5)$$

so that (11.4) becomes

$$\begin{aligned} -\frac{\mu k T}{q} \left(\frac{N_D}{N_A - N_D} \right) \frac{\partial N_D}{\partial x} &= -\frac{\mu_0 k T}{q} \left(1 - \frac{P}{N_D} \right) \left(\frac{N_D}{N_A - N_D} \right) \frac{\partial N_D}{\partial x} \\ &= -D_0 \left(1 - \frac{P}{N_D} \right) \left(\frac{N_D}{N_A - N_D} \right) \frac{\partial N_D}{\partial x} \end{aligned} \quad (11.6)$$

where (7.15) and the Einstein relation⁵⁰ have been used, and D_0 is the diffusivity in the absence of pairing.

P/N_D in (11.6) can be evaluated using (9.5) so that the coefficient preceding $(\partial N_D / \partial x)$ contains N_D as the only variable.

In Appendix B it is shown that ion pairing itself leads to severe departures from Fick's law.⁴⁹ In fact the diffusion flux density in the presence of pairing is given by

$$-\frac{D_0}{2} \left(1 + \frac{\frac{1}{2} \left(N_D - N_A + \frac{1}{\Omega} \right)}{\sqrt{\frac{1}{4} \left(N_D - N_A - \frac{1}{\Omega} \right) + \frac{N_D}{\Omega}}} \right) \frac{\partial N_D}{\partial x} \quad (11.7)$$

Here again the diffusivity is specified by the factors preceding $(\partial N_D / \partial x)$ and, though variable, depends only on N_D , the local concentration of diffusant. Adding the two coefficients appearing in (11.6) and (11.7) the value of the diffusivity, D , in the presence of both pairing and diffusion potential is obtained. Thus

$$D = \frac{D_0}{2} \left(1 + \frac{\frac{1}{2} \left(N_D - N_A + \frac{1}{\Omega} \right)}{\sqrt{\frac{1}{4} \left(N_D - N_A - \frac{1}{\Omega} \right)^2 + \frac{N_D}{\Omega}}} \right. \\ \left. + 2 \left(1 - \frac{P}{N_D} \right) \left(\frac{N_D}{N_A - N_D} \right) \right) \quad (11.8)$$

It is obvious from (11.8) that even in the absence of space charge D is an extremely complicated function of N_D , and will be much more complex if space charge needs to be considered. When $N_D \ll N_A$ (11.8) reduces to

$$D = D_0 \left(\frac{1}{1 + \Omega N_A} \left(1 + \frac{N_D}{N_A} \right) \right) \approx \frac{D_0}{1 + \Omega N_A} \quad (11.9)$$

Comparison with equation (B15) shows that when (11.8) is true (i.e., in the absence of space charge) the diffusion potential may be ignored for $N_D \ll N_A$. Comparison of (B14) with (B15) shows how much D can vary with N_D when ion pairing occurs.

The proper study of diffusion in the presence of ion pairing should be augmented by a mathematical analysis, accounting for the concentration dependent diffusivity. Since this dependence is complicated the resulting boundary value problem must be solved numerically, and this

represents a formidable task. Although work along these lines is being done we shall content ourselves, in this article, with a less quantitative approach. The following plan has been followed.

A rectangular wafer of semiconductor uniformly doped with acceptor to the level, N_A , is uniformly saturated with lithium to a level, N_D , slightly less than N_A . Thus, the resulting specimen is well compensated but not converted to *n*-type. Lithium is then allowed to diffuse out of the specimen, and because of the thinness of the wafer, this process may be regarded as plane-parallel diffusion normal to its large surfaces. Low resistivity *p*-type layers therefore develop near the surfaces. If the thin ends of the wafer are put in contact with a source of current, current will flow parallel to its axis, so that the equipotential surfaces will be planes normal to this axis. The flow of current will be one dimensional because the inhomogeneity in lithium distribution occurs in the direction normal to its flow (see Fig. 21).

If two probe points are placed at a fixed distance apart on the broad surface of the wafer (see Fig. 21), then the conductance measured between them is a reflection of the total number of carriers in the low resistivity layers, i.e., a measure of the total amount of lithium which has diffused out. A more detailed connection between this conductance and diffusivity is derived in Appendix E. For the moment, however, attention will be confined to the description of the general plan of experiment.

According to the formulas derived in the early parts of this section, and also to (B14) and (B15), the diffusivity is something like $D_0/2$ in the

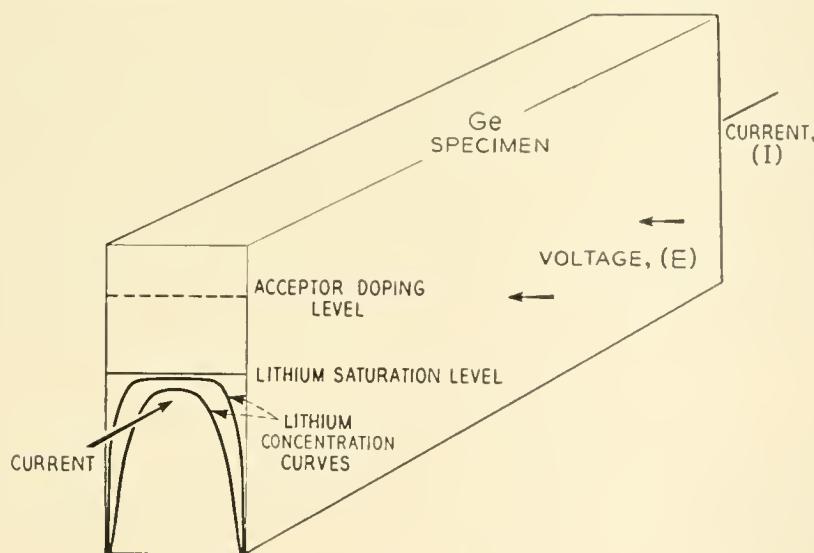


Fig. 21 — Diagram illustrating measurement of dependence of diffusivity on ion pairing (see Section XI).

bulk of the wafer where N_D almost equals N_A , but is as low as $D_0/(1 + \Omega N_A)$ near the surface where $N_D \ll N_A$. If ΩN_A is very much larger than unity as it will be under conditions where appreciable pairing occurs, the diffusivity will, therefore, be much smaller near the surface than at the high end of the diffusion curve, deeper within the specimen. The surface will then offer resistance to diffusion, and it may be expected that the measured value of the diffusivity will correspond more closely to the slow process near the surface rather than to the faster process occurring deeper in the semiconductor. Of course this cannot be entirely true because the resistance at the surface coupled with the lack of resistance inside the wafer will tend to steepen the concentration gradient near the surface. This will give the impression of a diffusivity somewhat higher than the one corresponding to the surface.

If the current flowing in the wafer under the conditions of measurement is I , and the potential measured between the points is V , then the conductance between the points is

$$\Sigma = I/V. \quad (11.10)$$

In Appendix E it is shown (under the assumption that D is constant) that

$$\Sigma/\Sigma_0 = 1 + \frac{2.256\vartheta\sqrt{D}}{d} \left(\frac{\Sigma_\infty N_D^\circ}{\Sigma_0 N_A} \right) \sqrt{t} \quad (11.11)$$

where Σ_0 is the conductance after the specimen is saturated with lithium, but before any lithium has diffused out, and Σ_∞ is the conductance before lithium has been added. N_A is the uniform concentration of acceptor, and N_D° is the initial uniform concentration of lithium, while d is the thickness of the wafer. ϑ is a correction factor which arises because the mobility of holes varies from point to point in the wafer, as the density of lithium varies. There are two extreme types of variation.

The first takes place in a specimen in which, at room temperature (where the conductance measurement is made) ion pairing is complete. Then the local density of impurity scatterers⁵⁴ will be $N_A - N_D$. At the other extreme no ion pairing occurs, and the density of scatterers is $N_A + N_D$.

The nature of ϑ depends on how much pairing is involved. In Fig. 22 ϑ has been evaluated in its dependence on N_D° for the extreme cases mentioned. Furthermore it has been assumed then that N_D is given by a Fick's law solution of the diffusion problem, and that diffusion begins in a nearly compensated specimen.

The first thing to notice is that ϑ is not very different from unity in

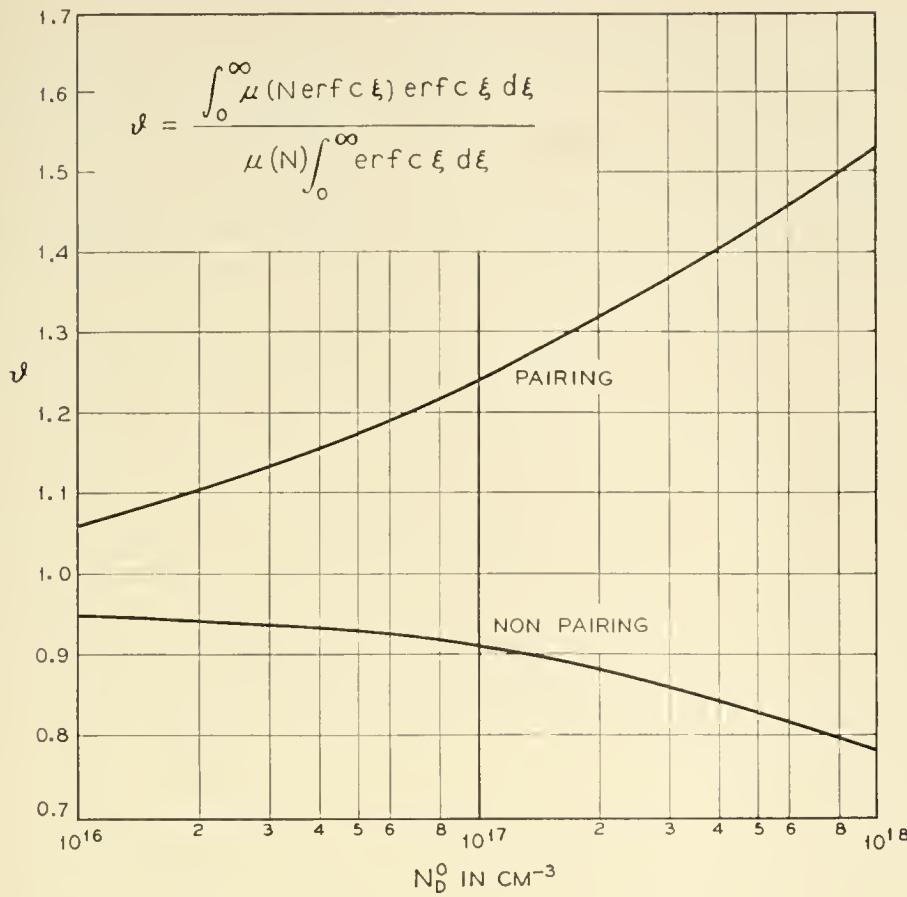


Fig. 22 — Plots of correction factor ϑ , required to compensate for the dependence of hole mobility on the density of scattering centers along a diffusion curve. ϑ is plotted against the initial density of donor and is shown for the two extreme cases of pairing and no pairing.

either extreme, and therefore closer to unity in some intermediate situation. In any event the correct value of ϑ can be read from Fig. 22 if the experiments involve either extreme at the measurement temperature. This has, in fact, been approximately the case in our experiments, in which pairing is almost complete at the temperature where conductances have been measured.

According to (11.11) a plot of Σ/Σ_0 against \sqrt{t} should be a straight line of slope

$$S = \frac{2.256 \vartheta \sqrt{D}}{d} \left(\frac{\Sigma_\infty N_D^0}{\Sigma_0 N_A} \right) \quad (11.12)$$

Measurement of S therefore affords a measure of D . Of course the apparent D obtained in this manner can never represent anything beyond some average quantity having the general significance of a diffusivity. This follows from the previous discussion concerning the non-constancy

of D . The only exception to this statement occurs in connection with high temperature experiments (above 200°C.) where both pairing and the diffusion potential are of little consequence. The mere fact that Σ/Σ_0 plots as a straight line against \sqrt{t} is not evidence for the constancy of D . In Appendix E it is shown that a straight line will result, even when ion pairing is important, provided that the diffusion potential is based on the no-space-charge condition, i.e. provided that D varies only through its dependence on N_D .

On the other hand, the last statement implies that the existence of a straight line relationship is evidence that the diffusion potential has at least been minimized.

The most careful experiments were performed in germanium doped to various levels with gallium, indium, and zinc as acceptors. The germanium specimens were cut in the form of rectangular wafers of approximate dimensions (1.25 cm \times 0.40 cm \times 0.15 cm). Fresh lithium filings, were evenly and densely spread on one surface of the wafer, and alloyed to the germanium by heating for 30 seconds at 530°C in an atmosphere of dry flowing helium. Then the other surface was subjected to similar treatment.

After this the specimen was sealed in an evacuated pyrex tube and heated at a predetermined temperature for a predetermined period of time. The temperature was chosen, according to Fig. 5, so that the saturated specimen would still be *p*-type and just barely short of being fully compensated. Also attention was paid to the problem of avoiding precipitation on cooling. The time of saturation was determined from an extrapolation of the known lithium diffusion data, in germanium, of Fuller and Severiens⁵² which is plotted in Figure 19 for the range extending from about 0° to 300°C.

After saturation the sealed tube was dropped into water and cooled. It was opened and the wafer ground on both sides, first with No. 600 Aloxite paper, and then with M 303½ American Optical corundum abrasive paper. The final thicknesses of the specimens ranged from 0.025 to 0.075 cm, the thinnest samples being used for the runs at the lowest temperature.

If the specimen is quite thin and highly compensated it is possible in principle to measure very small diffusivities (as low as 10^{-14} cm²/sec) within a period of several hours. This is so because the low resistivity layer formed near the surface, although thin, will carry a finite share of the current in thin compensated specimens. On the other hand, additional difficulties arise. Diffusion layers as small as 100 Å may be involved. If the surface is microscopically rough, diffusion will not be plane-parallel

and the measured diffusivity will appear larger than the real diffusivity. This condition can be partially corrected by etching the surface chemically until it is fairly smooth.

When dealing with such thin layers, the no-space-charge assumption becomes invalid and the diffusion potential ought really to be considered. Considering all the difficulties, i.e., concentration dependence of diffusion coefficient, possible existence of space charge, and roughness of surface, it is apparent that only qualitative effects are to be looked for in the diffusivities which have been measured.

The most that can be predicted is that for specimens containing a given amount of acceptor, the measured D (some average quantity) should be less than D_0 , the disparity increasing with decreasing temperature. At high temperatures D should converge on D_0 . Furthermore, at a given temperature D should decrease with an increase in concentration of acceptor. These tendencies are in line with the idea that reduction of temperature or increase of doping leads to an increase in pairing.

Runs were carried out on specimens etched with Superoxol⁵⁵ at the temperatures 25°, 100°, and 200°C. In the 25°C run the wafer was allowed to remain in the measuring apparatus under the two probe points in air, and Σ was measured from time to time. At 100°C the specimen was immersed in glycerine containing a few drops of HCl, the temperature of the bath being controlled. Periodic removal from the bath facilitated the measurement of Σ . At 200°C glycerine was again used as a sink for lithium, the sample being removed periodically for measurement.

Fig. 23 illustrates some typical plots of Σ/Σ_0 versus \sqrt{t} . They are all satisfactorily straight. Fig. 24 shows a plot of $\log D_0$ against $1/T$, extrapolated from the data of Fuller and Severiens.⁵² In this illustration, values of $\log D$ (obtained from the above measurements by determining the slopes S and employing (11.12)) are also plotted at the temperatures of diffusion. For the ease of complete pairing was assumed.

The first thing to note is that the points for $\log D$ all lie below $\log D_0$ except at 200°C and satisfy the qualitative requirement outlined above.* Moreover they drop further below $\log D_0$ as the temperature is reduced, while at 200°C they have almost converged on $\log D_0$.

The results for zinc are particularly interesting. Zinc is supposed to have a double negative charge in germanium.⁶⁴ Hence we would expect very intense pairing to occur. This is indicated in the diffusion data where the sample containing zinc at the rather low level, $N_A = 2.7$

* The long range nature of the interaction forces becomes evident when one considers that the diffusivities are being altered by impurity (acceptor) concentrations of the order of 1 part per million.

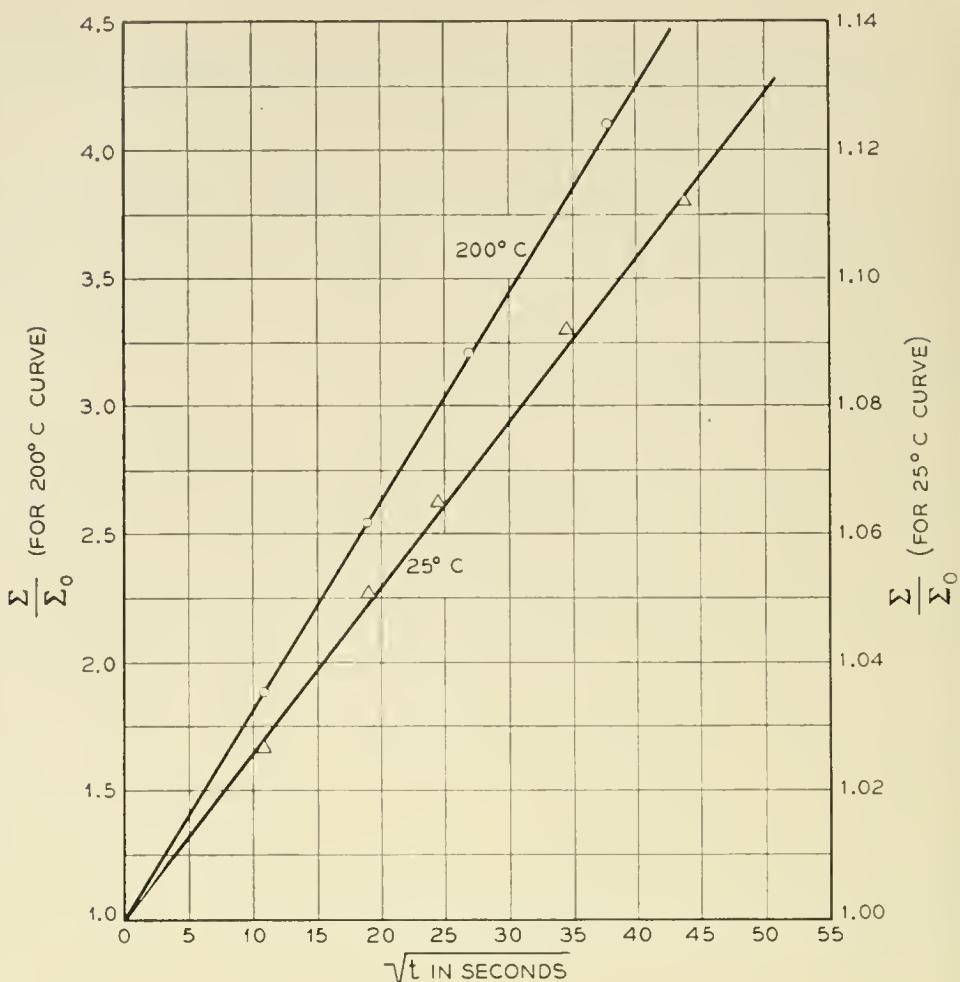


Fig. 23 — Curves illustrating the observed linear dependence of Σ/Σ_0 on the \sqrt{t} .

$\times 10^{16} \text{ cm}^{-3}$, shows a large reduction in diffusivity even at temperatures as high as 200°C.

The difficulties discussed in this section serve to emphasize the importance of a direct transport experiment in which lithium atoms *uniformly* distributed throughout germanium or silicon, uniformly doped with acceptor, are caused to migrate by an electric field, and their mobilities measured. Because of the uniform dispersion of solutes the mobility will be constant everywhere. Furthermore no diffusion potential will be involved, and also the refined formula (7.25) can be applied. There are, however, many difficulties associated with the performance of this type of measurement.

In closing it may be mentioned that a few much less careful experiments of the kind described here have been performed in boron-doped silicon. The results indicate ion pairing in a qualitative way but more definite experiments are needed.

XII. INVESTIGATION OF ION PAIRING BY ITS EFFECT ON CARRIER MOBILITY

In Section VIII attention was called to the fact that ion pairing should influence the mobility of holes, because each pair formed, reduces the number of charged impurities by two. Thus, a specimen previously doped with acceptor, might, if sufficient lithium is added, exhibit an increase in hole mobility, even though the addition of lithium implies the addition of more impurities. This effect has been observed in connection with the Hall mobility of holes in germanium.

Two specimens of germanium were cut from adjacent positions in a single crystal doped with gallium to the level $3 \times 10^{17} \text{ cm}^{-3}$. One of these was saturated with lithium through application of the same procedure

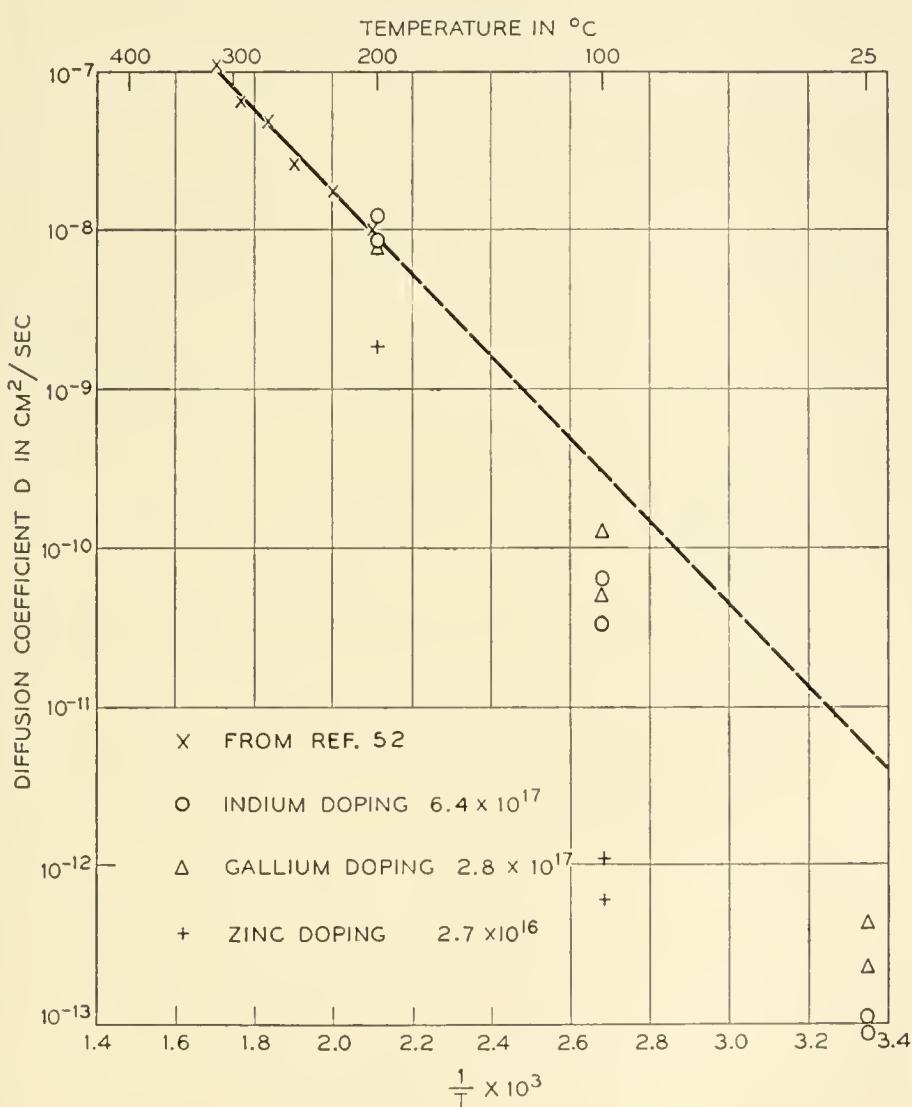


Fig. 24 — Plot of diffusivity of lithium in undoped germanium as a function of temperature — also showing points for apparent diffusivities of lithium in variously doped specimens.

employed in section V. Hall mobilities of the two specimens were measured⁶⁵ down to below 10°K. Cooling was carried out slowly to permit as much relaxation into the paired state as possible (see Section X). In Fig. 25 plots of the Hall mobilities versus temperature of both specimens are presented. Curve A is for the sample containing $2.8 \times 10^{17} \text{ cm}^{-3}$ lithium. It therefore contained about $5.8 \times 10^{17} \text{ cm}^{-3}$ total impurities as compared to the control sample whose curve is shown as B in Figure 25 and which contained only $3 \times 10^{17} \text{ cm}^{-3}$ impurities.

The lithium doped bridge exhibits by far the higher Hall mobility for holes (except at very low temperatures where poorly understood phenomena occur). In fact at 40°K the sample containing lithium shows a hole mobility 16 times greater than that of the control at the corresponding temperature. Rough analysis of the relative mobilities at $T = 100^\circ\text{K}$ indicate $\sim 2 \times 10^{17} \text{ cm}^{-3}$ scattering centers in the control sample and $5 \times 10^{15} \text{ cm}^{-3}$ scattering centers in the sample containing pairs.

This experiment has been repeated with other specimens doped to different levels with gallium and even with other acceptors, and leaves no doubt that a mechanism which is most reasonably assumed to be pairing, is removing charged impurities from the crystal.

The phenomenon we have just described suggests an excellent method for testing the ion pairing formula derived in Sections VII and XI, for it

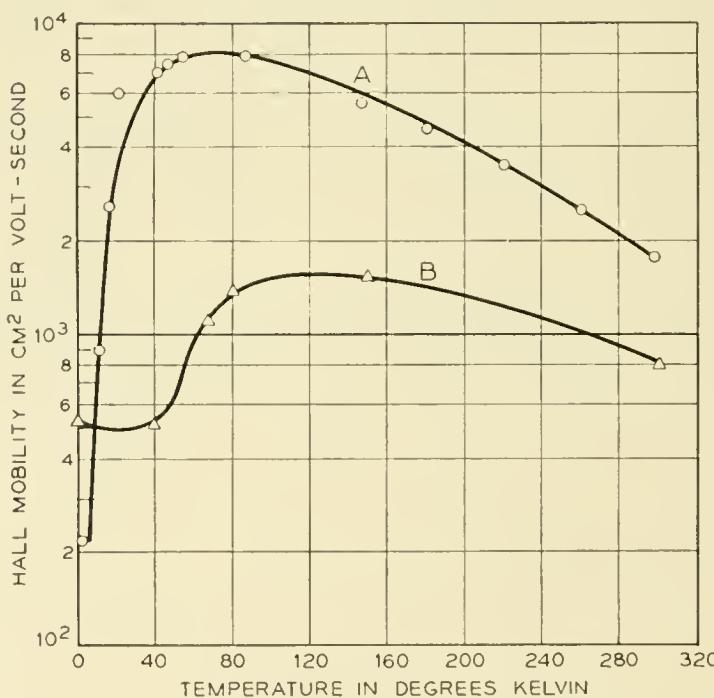


Fig. 25 — Plot of Hall mobility as a function of temperature for germanium containing $3 \times 10^{17} \text{ cm}^{-3}$ gallium. Curve A is for a sample containing $2.8 \times 10^{17} \text{ cm}^{-3}$ lithium.

enables us to determine at what temperature, at given values of N_A and N_D , P/N_D is exactly 0.5. Thus consider the fact that, all other things being equal, the control bridge and the one containing added lithium will exhibit equal Hall mobilities at a given temperature when the concentrations of charged impurities are identical in both of them. Now the concentration of such impurities in the control is simply N_A . The concentration in the bridge containing lithium is

$$N_A + N_D - 2P \quad (12.1)$$

The quantity $2P$ is removed from $N_A + N_D$, because each time a pair forms two charged scatterers are eliminated. The condition that the scattering densities in both bridges be equal is then simply

$$N_A = N_A + N_D - 2P$$

or

$$\frac{P}{N_D} = 0.5 \quad (12.2)$$

Therefore if plots of Hall mobilities versus temperature such as those appearing in Figure 25 are continued until they cross, the temperature of crossing marks the point at which P/N_D is 0.5.

In Fig. 26 typical crossings of this kind are shown. They are for two different gallium doped germanium crystals, one containing 3×10^{17} cm⁻³ gallium and the other 9×10^{15} cm⁻³. The curves for the controls and lithium saturated samples in each case are shown as plots of the logarithm of Hall mobility against logarithm of absolute temperature. The lines plotted in this manner are straight. The lithium content of the bridge containing 9×10^{15} cm⁻³ gallium was 6.1×10^{15} cm⁻³ while that in the bridge with 3×10^{17} cm⁻³ gallium was 2.8×10^{17} cm⁻³. All of these concentrations were obtained from Hall coefficient measurements in the controls and the lithium doped specimens.

As the temperature is increased the mobilities of the samples with lithium are reduced and approach the mobilities of the controls. This happens because pairs dissociate and more charged impurities appear. Finally when P/N_D is exactly 0.5 the curves cross. In Fig. 27 we notice that mobility measurements were not performed right up to the cross point, but that the straight lines have been extrapolated. This procedure was adopted of necessity, because of the high diffusivity of lithium. Thus, reference to Fig. 5 shows that the solubility in doped germanium decreases to a minimum as the temperature is raised from room temperature, and there is danger of precipitation. For this reason the measure-

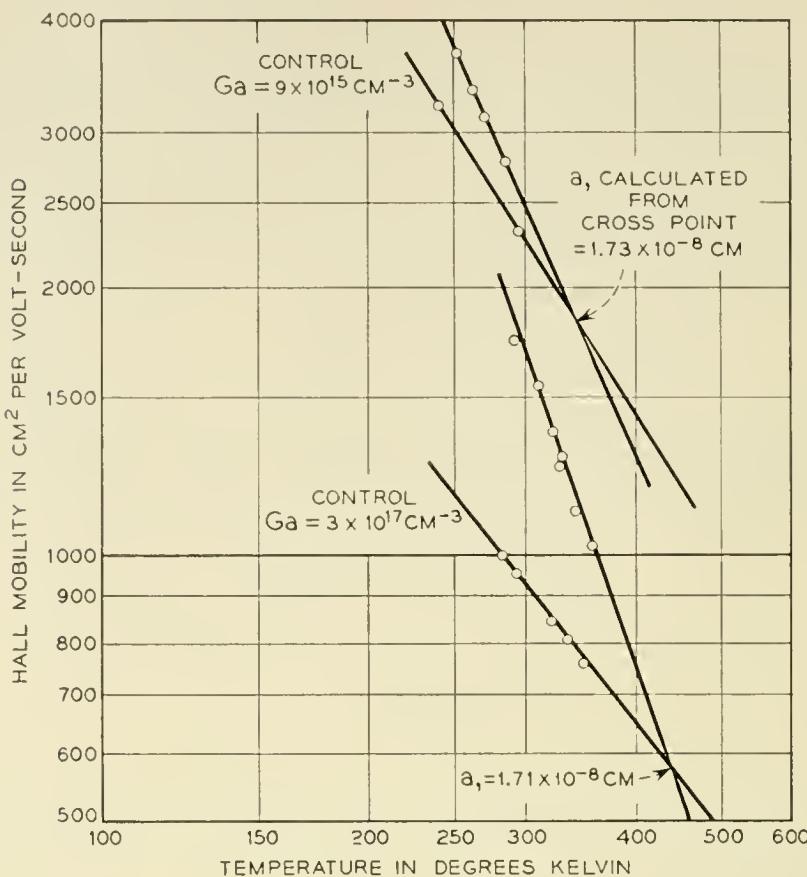


Fig. 26 — Illustration of cross over phenomenon for germanium samples containing gallium. Sample 314 contains $9 \times 10^{15} \text{ cm}^{-3}$ gallium and sample 302 contains $3 \times 10^{17} \text{ cm}^{-3}$. Samples 316 and 301 are the corresponding samples to which lithium has been added.

ments were not carried to high temperatures.* In addition the value of the Hall coefficient was carefully checked at each temperature to see if it had changed. Since the reciprocal of the Hall coefficient⁶⁶ measures the carrier density any reduction in its value would have implied loss of compensation, or precipitation of lithium.

Over the measured points no appreciable variation of Hall coefficient was noted. Fortunately, the pairing relaxation time is quite small (less than a second) at the high temperatures involved so that it wasn't necessary to hold the samples at these temperatures for long periods in order to achieve pairing equilibrium. The times involved were too short for the occurrence of phase equilibrium characterized by precipitation.

The above discussion points up some of the care that must be taken to obtain reliable measurements. Another factor which enters the picture is the possible existence of a precipitate in the lithium doped bridge.

* In boron-doped germanium the cross-over was actually observed — no extrapolation having been necessary, because the temperature of intersection was sufficiently low.

During the course of our experiments it was discovered that precipitates have a profound effect on carrier mobility, reducing it so severely, that the mobility of the lithium doped bridge may never even rise above that of the control. Great care must be exercised in the preparation of suitable bridges to avoid the presence of precipitated lithium. Thus it may be necessary to saturate the bridge at a very low temperature (see Section IV, Figure 5) so that it is somewhat undersaturated at room temperature. This means that diffusion periods of weeks may be involved.

In Fig. 26 the sample with $N_A = 9 \times 10^{15} \text{ cm}^{-3}$, and $N_D = 6.1 \times 10^{15} \text{ cm}^{-3}$ has $P/N_D = 0.5$ at 348°K , while the sample with $N_A = 3 \times 10^{17} \text{ cm}^{-3}$ and $N_D = 2.8 \times 10^{17} \text{ cm}^{-3}$ is half-paired at 440°K . This is to be expected, the more heavily doped specimen remaining paired up to higher temperatures. Using (9.6) and (9.3) it is possible to calculate a , the distance of closest approach of a gallium and lithium ion, from each of the measured cross points.

Thus in (9.6) we set $\theta = 0.5$, and N_A , N_D and T to correspond to each of the cases described. Having $\log_{10} Q(\alpha)$, α can be determined by interpolation in Table III and a then determined from (9.3). Of course κ is taken to be 16. Carrying through this procedure in connection with Fig. 26 leads to the satisfying result that $a = 1.71 \times 10^{-8} \text{ cm}$ for the heavily doped sample and $1.73 \times 10^{-8} \text{ cm}$ for the lightly doped one. The values of Ω appearing in Table IV based on $a = 1.7 \times 10^{-8} \text{ cm}$ therefore correspond to gallium.

Not only is this result satisfying because the two a 's agree so well even though the samples involved were so different in constitution, but also because it is expected on the basis of the addition of known particle radii. Thus according to Pauling³⁶ the tetrahedral covalent radius of gallium is $1.26 \times 10^{-8} \text{ cm}$ while the ionic radius of lithium is $0.6 \times 10^{-8} \text{ cm}$. Since gallium is presumably substitutional in a tetrahedral lattice we use its tetrahedral covalent radius, and since lithium is probably interstitial we use the ionic radius. The sum of the two is $1.86 \times 10^{-8} \text{ cm}$ which compares very favorably with the values of a quoted above.

This result constitutes good evidence that lithium is interstitial, for if it were somehow substitutional we might expect a to be something like a germanium-germanium bond length which is $2.46 \times 10^{-8} \text{ cm}$. Such a value of a would lead to profoundly different crossing temperatures (of the order of 100° lower) so that it is not very likely.

One further point needs mention. This is the fact that as the two ions approach very closely, the concept of the uniform macroscopic dielectric constant, κ , loses its meaning. In fact, the binding energy should be increased (as though κ were reduced). Crude estimates of the magnitude

of this effect based on a dielectric cavity model show it to be of the order of some 10 or 15 percent of the energy computed on the assumption of the dielectric continuum, the increased binding energy showing up as a reduced value of a . This may account for the fact that the observed a , at 1.7×10^{-8} cm is less than the theoretical value, 1.86×10^{-8} cm.

The above example shows the ion pairing phenomenon in action as a structural tool, useful in investigating isolated impurities. In particular the demonstration that lithium is interstitial is interesting. The values of a have much more meaning as independent parameters in solids than they have in liquids, where a given ion may be surrounded by a sheath of solvating solvent molecules. Under the latter conditions the value of a can only be determined through application of the ion pairing theory itself.

Of course, certain unusual situations arise in solids also, and values of a (determined from ion pairing) are valuable indications of structural peculiarities.

Similar experiments have been performed on specimens doped with indium and boron. The results of all our investigations on the cross-over phenomenon are tabulated in Table V. In the table the first column lists the acceptor involved, and the second and third the appropriate concentrations of impurities. The fourth column contains the cross-over temperature, while the fifth, the measured value of a determined from it. The last column lists the values of a to be expected on the basis of the addition of tetrahedral covalent radii to the ionic radius of lithium — all of which appear in Pauling.³⁶

The reliability of the measurements are in the order gallium, aluminum, boron, and indium. The principal reason for this is that the indium crystal was not grown specially for this work and was somewhat non-uniform. Of the two values obtained for a we tend to place more confi-

TABLE V

Acceptor	Acceptor conc. (cm ⁻²)	Lithium conc. (cm ⁻²)	Cross-over Temp. (°C.)	Measured a (cm)	Pauling a (cm)
B	7.0×10^{16}	5.9×10^{16}	338	2.05×10^{-8}	1.48×10^{-8}
B	7.0×10^{16}	5.54×10^{16}	320	2.27×10^{-8}	1.48×10^{-8}
B	7.0×10^{16}	5.85×10^{16}	330	2.16×10^{-8}	1.48×10^{-8}
Al	9.5×10^{15}	9.0×10^{15}	350	1.68×10^{-8}	1.86×10^{-8}
Ga	3.0×10^{17}	2.8×10^{17}	440	1.71×10^{-8}	1.86×10^{-8}
Ga	9.0×10^{15}	6.1×10^{15}	348	1.73×10^{-8}	1.86×10^{-8}
In	3.3×10^{17}	1.9×10^{17}	476	1.61×10^{-8}	2.04×10^{-8}
In	3.3×10^{17}	2.68×10^{17}	426	1.83×10^{-8}	2.04×10^{-8}

dence in 1.83×10^{-8} than in 1.61×10^{-8} cm. More work is necessary, however, before a real decision can be made.

A feature of Table V is the fact that gallium, aluminum, and indium exhibit orthodox behavior, i.e., the measured a 's are in both cases slightly less than those expected on the basis of the addition of radii. The internal consistency of the theory gains support from the fact that gallium and aluminum behave similarly as the Pauling a 's tabulated in Table V predict. In fact if 1.83×10^{-8} cm is taken as the more reliable indium value the three cases fail to match the Pauling radii by about the same amount, a result which implies that the disparity is due to the same cause, i.e., failure of the dielectric continuum concept.

Another feature of Table V is the fact that boron is out of line to the extent that the measured a exceeds the Pauling a by 50 per cent. A possible explanation is the following. The tetrahedral radii of boron and germanium are poorly matched (0.88 Å and 1.26 Å, respectively). The strain in the boron-germanium bond may appear as a distortion of the germanium atom in such a way as to increase the effective size of the boron ion. This strain was mentioned before in Section V where it was invoked to explain the stability of LiB^- complex in silicon.

XIII. RELAXATION STUDIES

The relaxation time discussed in Section X has been studied experimentally. The following procedure was used. A specimen was warmed to 350°K where a considerable amount of pair dissociation occurred, and then cooled quickly by plunging into liquid nitrogen. It was then rapidly transferred to a constant temperature bath, held at a temperature where pair formation took place at a reasonable rate, and the change in sample conductivity (as pairing took place) was measured as a function of time.

The principle upon which this measurement is based is the following. At a given temperature the occurrence of pairing does not change the carrier concentration, only the carrier mobility. As a result the measurement of conductivity is effectively a measurement of relative mobility. During relaxation the densities of charged impurities are changed, at the most, by amounts of the order of 50 per cent. Over this range, the mobility may be considered a linear function of scatterer density. The dependence of conductivity on time, as pairing takes place, must be of the form

$$\sigma = \sigma_\infty - \Phi e^{-t/\tau} \quad (13.1)$$

where σ_∞ is the conductivity when $t = \infty$, and τ is the relaxation time defined in section X while Φ is some unknown constant, depending among

other things on the initial state of the system. Equation (13.1) is based on the assumption that the number of charged scatterers decays as a first order process, and that σ is a linear function of this number, relative to the exponential dependence on time.

The first order character of pairing is fortunate for it renders the measurement of τ independent of a knowledge of Φ , i.e. independent of the initial state of the system. This is not only fortunate from the point of view of calculation but from experiment, since it is almost impossible to prepare a specimen in a well defined initial state.

The unimportance of Φ is best seen by plotting the logarithm of $\sigma_\infty - \sigma$ against time. According to (13.1) this plot is specified by

$$\log(\sigma_\infty - \sigma) = \log \Phi + \frac{t}{\tau} \quad (13.1)$$

Thus the reciprocal of its slope measures τ , and Φ is not involved. Fig. 27 illustrates the data for a typical run plotted in this manner. The sample is one containing about $9 \times 10^{15} \text{ cm}^{-3}$ gallium and the experiment was performed at 195°K (dry ice temperature). Notice that the curve is absolutely straight out to 3500 minutes, demonstrating beyond a doubt that the process is first order. The relaxation time computed from its slope is 1.51×10^5 seconds as against a value calculated by the methods

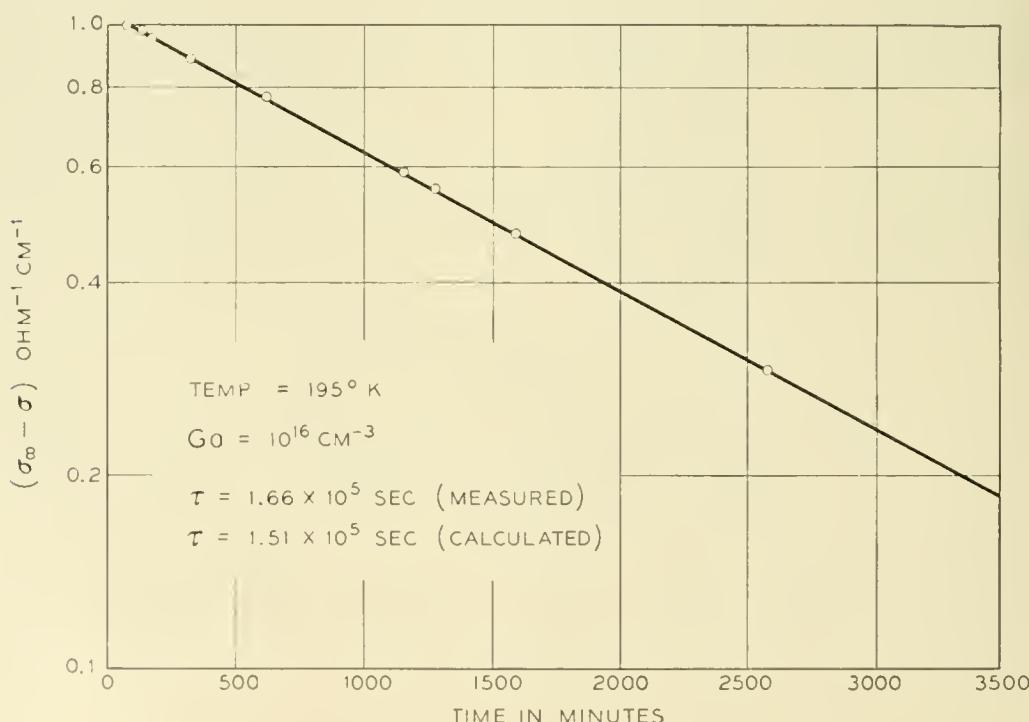


Fig. 27 — Plot of $\log(\sigma_\infty - \sigma)$ as a function of time showing first order kinetics of pairing.

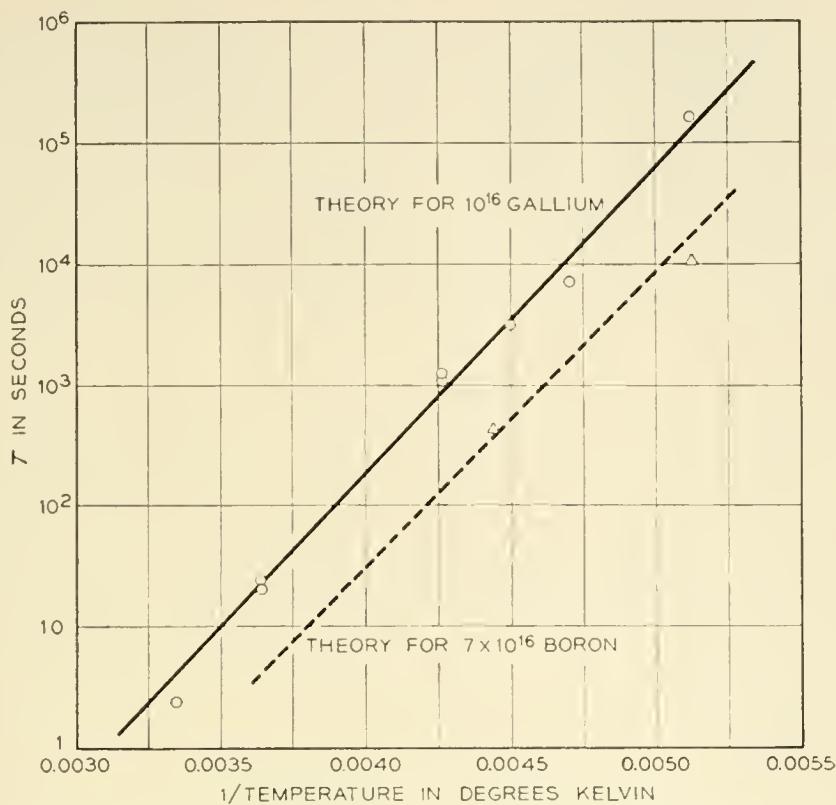


Fig. 28 — Plots of logarithm of relaxation time versus reciprocal temperature showing agreement between theory and experiment.

of section X of 1.66×10^5 seconds. The result is in good agreement with theory.

Studies of the kind illustrated in Fig. 27 have been carried out in samples doped to various levels and also at various temperatures. Boron and indium have been used as doping agents, as well as gallium. Relaxation times have been measured over the range extending from about a second to hundreds of thousands of seconds. In each case straight line plots were obtained and the agreement between calculated and measured τ 's has been as good as in the example illustrated by Fig. 28. Relaxation connected with dissociation has also been measured with equally satisfactory results.

Some of these data are shown in Fig. 29 where $\log \tau$ is plotted as a function of reciprocal temperature for gallium and boron at two different values of doping. The drawn curves are theoretical obtained from Fig. 20 while the points shown are experimental. It is seen that agreement is nearly perfect. The relaxation time, true to the demands of theory, does not seem to depend on the kind of acceptor used for doping, i.e., it is independent of a , the distance of closest approach.

The data in Fig. 28 actually can be used to measure the diffusivity of

lithium. As must be the case from the above mentioned agreement, the values of D_0 computed from them agree with the diffusion data of Fuller and Severiens⁶² almost perfectly. This is a very quick and sensitive method (also probably exceedingly accurate) for determining diffusivities. For example the work already completed, in effect, represents the determination of diffusivities of the order of 10^{-16} cm²/sec within a matter of an hour, and, no doubt, smaller diffusivities could be determined by doping more heavily with acceptor.

XIV. THE EFFECT OF ION PAIRING ON ENERGY LEVELS

It was predicted in Section VIII that ion pairing would drive the electronic energy states of donors and acceptors from the forbidden energy region. In this section it will be demonstrated by low temperature Hall effect measurements that the addition of lithium to gallium-doped germanium does indeed result in the removal of states from the forbidden gap rather than in the simple compensation which occurs when a non-mobile donor such as antimony is added.

At low temperatures where carrier concentration, p , is less than the donor concentration, it can be expressed in the form⁶⁷

$$p = \frac{N_A - N_D}{N_D} \left(\frac{2\pi m_p k T}{h^2} \right)^{3/2} \exp [-E_A/kT] \quad (14.1)$$

where N_A and N_D are the concentrations of acceptor and donor states, respectively, m_p , the effective mass of free holes, h , Plank's constant, and E_A the ionization energy of the acceptor. The values of m_p and E_A are known for the group III acceptors.⁶⁸

Lithium was added to a specimen of germanium known to contain 1.0×10^{16} cm⁻³ gallium atoms and a negligible amount of ordinary donors. Carrier concentrations for this specimen were determined from Hall coefficient measurements. The logarithm of this concentration is shown in Fig. 29 plotted against reciprocal temperature. The high temperature limit of this plot fixes $N_A - N_D$ at 1.15×10^{15} cm⁻³.

At low temperatures the curve exhibits an extended linear portion to which (14.1) should apply. Evaluating (14.1) with $p = 4.0 \times 10^{13}$ cm⁻³ at $1/T = 0.06$ deg⁻¹ and $N_A - N_D = 1.15 \times 10^{15}$ cm⁻³ we find that $N_D = 2.6 \times 10^{14}$ cm⁻³ and $N_A = 1.4 \times 10^{15}$ cm⁻³.

Therefore, the density of apparent acceptor states has been decreased by $1.0 \times 10^{16} - 1.4 \times 10^{15} = 8.6 \times 10^{15}$ cm⁻³. The added concentration of lithium was 1.0×10^{16} cm⁻³ - 1.15×10^{15} cm⁻³ = 8.85×10^{15} cm⁻³, *almost identical with the loss in concentration of acceptor states*. This implies (as would be expected) that the lithium is almost totally paired.

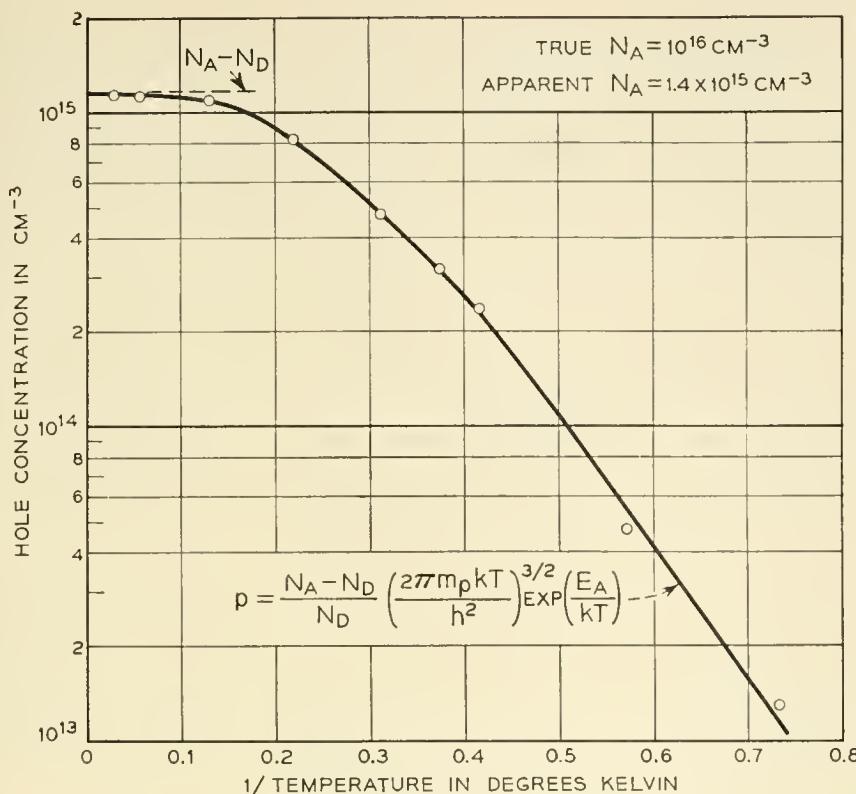


Fig. 29 — Plot of hole concentration as a function of reciprocal temperature for a sample containing ion pairs.

An even more striking result appears. From the above results the density of lithium atoms involved in pairs is $8.85 \times 10^{15} \text{ cm}^{-3} - 2.6 \times 10^{14} \text{ cm}^{-3} = 8.6 \times 10^{15} \text{ cm}^{-3}$, the same number by which the density of acceptors has been decreased! There can be little question that ion pairing is the mechanism responsible for the removal of states.

In closing it is worth pointing out that the density of unpaired lithiums $2.6 \times 10^{14} \text{ cm}^{-3}$, is certainly not characteristic of the low temperatures at which the above Hall measurements were performed. Obviously a density characteristic of some higher temperature has been quenched into the specimen. At the low temperature involved the unpaired density would be effectively zero.

XV. RESEARCH POSSIBILITIES

The fields described in the preceding text have been hardly touched, even by this long paper, and it does not seem fitting to close without some speculation concerning the possibilities of future work.

In the first place, there are other donors and acceptors besides lithium which are reasonably mobile in germanium or silicon, e.g. copper, iron,

zinc or gold. To some extent the methods of this paper can be applied to these. Furthermore, returning to lithium, there are impurities both mobile and immobile which introduce more than one energy state into the forbidden gap. The phase relations of lithium in the presence of these should be extremely interesting since the corresponding mass action equations are more complicated. Analogues of dibasic⁶⁹ acids and bases should exist.

In the case of ion pairing doubly charged acceptors like zinc in germanium⁶⁴ should be extremely interesting, since large amounts of pairing should persist up to very high temperatures. In fact such studies represent excellent means of testing for the existence of doubly charged ions. There is also the question of what happens to the two energy levels when an acceptor like zinc pairs with a single lithium ion. Are both levels driven from the forbidden gap or do they split under the perturbation?

Then there is the problem of ion *triplets* — a possibility with impurities like zinc — which is unexplored both theoretically and experimentally. Also, very strange diffusion effects must occur in the presence of doubly charged ions, to say nothing of the effect which uncompensated mobile holes might have on relaxation processes.

The field of ion pairing in silicon is relatively unexplored.

All of the phenomena discussed in this paper must occur in the group III-V compounds, more or less complicated by additional effects.

The question of the formation of the LiB⁻ complex in both germanium and silicon needs further study. It should behave as an acceptor and its electronic energy state might be revealed by suitable quenching techniques.

Non ionic reactions between group V donors and group III acceptors very likely occur, i.e., a real III-V covalent bond may be formed between such atoms dissolved in germanium or silicon at high temperatures. This possibility could be investigated by looking for changes in carrier mobility or impurity energy levels upon extended heating — in much the same way that ion pairing has been studied. If found, the phenomenon may provide an excellent technique for measuring the diffusivities of all classes of impurities even at fairly low temperatures.

Such compounds may possess strange energy levels and be responsible for unexplained traps and recombination centers.

The effect of stress on the extent of ion pairing may well be profound since there will be a tendency for such stress to concentrate at imperfections. Stress studies on ion pairing may therefore be useful for further investigating the strain about an isolated impurity.

Ion pairing between lithium ions and acceptor centers located in dislocations or vacancies should occur. In the first case the dislocation

would be the analogue of the polyelectrolyte molecule in the aqueous solution.

An interesting question, in the diffusion of substitutional acceptors, concerns whether the ion or the neutral atom is responsible for diffusion. It is possible that the neutral atom, less securely bonded to the lattice, is the chief agent. This might be determined by changing the ratio of neutral atoms to ions by suitably doping with other donors or acceptors.

Doping apparently affects the concentration of vacancies which have acceptor properties and therefore the rate of diffusion.^{70, 71}

Other interesting effects concerning the distribution of an impurity between two different kinds sites in the lattice⁷² are also possible.

These and many other fascinating fields still require exploration. We hope to investigate some of them in the near future.

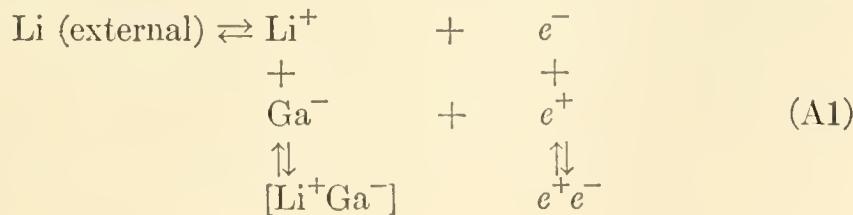
ACKNOWLEDGMENTS

The authors are greatly indebted to A. J. Pietruszkiewicz, Jr., for assistance in carrying out experimental work relating to solubility and diffusion and to J. P. Maita for help with experimental work on Hall effect and an ion-pair relaxation. Thanks are due N. B. Hannay for many helpful comments during the course of the work and during preparation of the manuscript. Thanks are also due Miss M. C. Gray for the evaluation of the integrals in Section VII and to F. G. Foster for the photograph of Fig. 8. Finally the authors would like to thank the editors of the Bell System Technical Journal for providing space so that all of the important features of our subject could be treated in one article.

APPENDIX A

THE EFFECT OF ION PAIRING ON SOLUBILITY

In Section VIII attention was called to the fact that ion pairing should have some effect on lithium solubility but that it would be difficult to achieve conditions under which the effect would be observable. Now, this point will be enlarged upon. Consider an equilibrium like (2.1) except imagine it to take place in germanium with gallium as the immobile acceptor. (This because germanium with gallium has been studied in ion pairing investigations.)



where $[Li^+Ga^-]$ represents an ion pair, whose concentration we denote by P . N_A and N_D will be the total densities of acceptor and donor respectively and A^- and D^+ the densities of acceptor and donor ions in the unpaired state.

As in the main text, n and p will represent the concentrations of holes and electrons. The following relations are then to be expected on the basis of definition, mass action, and charge balance.

$$N_A = A^- + P \quad (A2)$$

$$N_D = D^+ + P \quad (A3)$$

$$D^+n = K^* \quad (A4)$$

$$np = n_i^2 \quad (A5)$$

$$\frac{P}{A^+D^-} = \Omega \quad (A6)$$

$$D^+ + p = A^- + n \quad (A7)$$

Equations (A4), (A5), and (A7) are just reproductions of (3.1), (3.2), (2.8), while (A6) is the same as (9.4). The problem is to express the solubility of lithium, N_D , as a function of N_A . Manipulation of the preceding set of equations gives this result as

$$N_D = \frac{(N_A - A^-)(1 + \Omega A^-)}{\Omega A^-} \quad (A8)$$

with A^- given by the solution of

$$\begin{aligned} \frac{N_A - A^-}{\Omega A^-} &= \frac{A^-}{1 + \sqrt{1 + \left(\frac{2n_i}{D_0^+}\right)^2}} \\ &\quad + \sqrt{\left(\frac{A^-}{1 + \sqrt{1 + \left(\frac{2n_i}{D_0^+}\right)^2}}\right)^2 + (D_0^+)^2} \end{aligned} \quad (A9)$$

where D_0^+ is defined by (3.3). Equation (A9) generally needs to be solved numerically for A^- .

To see what these relations predict in a special case consider the solubility of lithium in gallium-doped germanium at 300°K. At this temperature the values of n_i and D_0^+ and Ω are

$$\begin{aligned} n_i &= 2.8 \times 10^{13} \text{ cm}^{-3} \\ D_0^+ &= 7 \times 10^{13} \text{ cm}^{-3} \\ \Omega &= 1.61 \times 10^{15} \text{ cm}^{-3}. \end{aligned} \quad (A10)$$

TABLE AI — TEMPERATURE = 300°K

N_A (cm^{-3})	N_D (cm^{-3})	N_D^* (cm^{-3})	$P = N_A - A^-$ (cm^{-3})
10^{14}	1.25×10^{14}	1.25×10^{14}	0.15×10^{14}
10^{15}	0.94×10^{15}	0.875×10^{15}	0.44×10^{15}
10^{16}	0.985×10^{16}	0.875×10^{16}	0.77×10^{16}
10^{17}	0.990×10^{17}	0.875×10^{17}	0.92×10^{17}
10^{18}	0.995×10^{18}	0.875×10^{18}	0.97×10^{18}

The value of n_i is taken from Figure 2, of D_0^+ , from Figure 5, and of Ω , from Table IV. Using (A10) together with (A9) and (A8) leads to the results tabulated in Table AI. In this table, N_D^* represents the solubility for the case $\Omega = 0$, i.e., the solubility if there were no ion pairing. The main feature to be obtained from the Table is that N_D is not very much larger than N_D^* , no matter how large the value of N_A . This is true in spite of the fact that the last column which lists P shows that at $N_A = 10^{18} \text{ cm}^{-3}$ P is about 98% of N_D so that pairing of the donor is virtually *complete*.

The result is not limited to the special conditions of doping and temperature chosen in compiling Table AI, but must be quite general. One can arrive at this conclusion in the following way.

By subtracting (A3) from (A2) we obtain

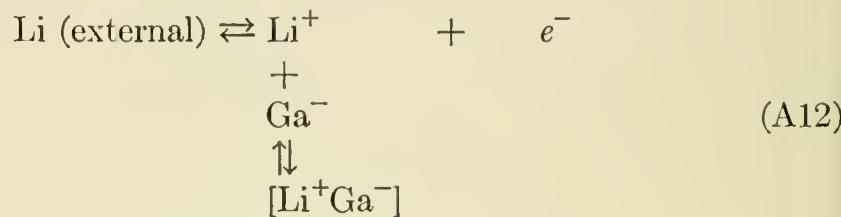
$$N_A - N_D = A^- - D^+. \quad (\text{A11})$$

The quantities A^- and D^+ appear in equations (A4) and (A7), while n and p , appearing in (A4) and (A7) are related by (A5). These three equations are sufficient for the determination of D^+ in its dependence on A^- . That this is the case is immediately obvious when (A4), (A5), and (A7) are recognized as reproductions of (3.1), (3.2) and (2.8). In fact this means that the desired relationship between D^+ and A^- is nothing more than equation (3.4) which itself is predicated on (3.1), (3.2), and (2.8). But then it is known according to (3.6), that D^+ can at the most be slightly greater than A^- , although most likely less. This assumes of course that we deal with dopings sufficiently high so that (3.5) applies. On the other hand at low dopings (3.4) tells us that D^+ will be D_0^+ . Therefore if we work with a system in which in the absence of pairing the electron-hole equilibrium has driven the value of N_D close to N_A (as it has in this system — see N_D^*) the introduction of pairing cannot drive it much higher, since according to (A11) if D^+ cannot get higher than A^- , N_D cannot exceed N_A . This is evident in Table AI where N_D comes very close to N_A but never exceeds it.

When N_A is very small so that D^+ equals D_0^+ and does exceed A^- by

a large amount, there can be no visible increment in solubility as a result of pairing because P can never exceed N_A which by definition is small.

The physical reason for these limitations is the following. Suppose N_D is driven close to N_A by the hole-electron equilibrium so that in terms of carriers (holes and electrons) the specimen is very closely compensated. Then if by the formation of pairs, additional donors are caused to enter the crystal, the electrons they donate cannot be absorbed by holes because very few of the latter are present. Thus the following two sketched equilibria will oppose each other



the one involving electrons attempting to drive lithium out of solution because of the build-up of electron concentration, and the pairing equilibrium attempting to bring lithium into solution in order to form pairs. Thus the pairing process will not be as efficient a solubilizer as might be thought at first.

This point can be illustrated by considering a situation in which the germanium crystal not only contains gallium to the level, N_A but also an immobile donor, to the level $N = 0.99 N_A$. Thus, the crystal is almost compensated before any lithium has been added. Nevertheless, there are still N_A gallium ions so that even though the hole-electron equilibrium, working on the differential, $0.01 N_A$, cannot increase the solubility of lithium, the pairing process might. To investigate this situation equations (A2) to (A7) can be adopted with the simple change that $(A^- - N)$ replaces A^- in (A7).

Taking the situation covered by (A10) at 300°K, Table AII was compiled. Here again N_D^* is the solubility for $\Omega = 0$.

If only the hole-electron effect were operative, then we could not expect to drive N_D much beyond $N_A - N$. In the 10^{16} case $N_A - N$ is 10^{14} cm^{-3} and in the 10^{17} case it is 10^{15} cm^{-3} . The values of N_D^* in Table AII thus confirm this argument. Furthermore, N_D is in neither case much greater than N_D^* showing that despite the fact that there were, respec-

TABLE AII — TEMPERATURE 300°K

N_A (cm^{-3})	N (cm^{-3})	N_D (cm^{-3})	N_D^* (cm^{-3})	P (cm^{-3})
10^{16}	0.99×10^{16}	3.2×10^{14}	1.26×10^{14}	3×10^{14}
10^{17}	0.99×10^{17}	1.6×10^{16}	0.88×10^{15}	1.6×10^{15}

tively, 10^{16} and 10^{17} cm^{-3} gallium ions available for pairing, the pairing process did very little to increase the solubility.

If the constant Ω is exceedingly large as is probably the case for a multiply charged acceptor, it is possible that ion paring will have a measurable effect on solubility.

APPENDIX B

CONCENTRATION DEPENDENCE OF DIFFUSIVITY IN THE PRESENCE OF ION PAIRING

In Section VIII it was mentioned that the diffusivity of a mobile donor like lithium is concentration dependent when the donor participates in a pairing equilibrium with an immobile acceptor. In this appendix we propose to investigate the nature of the dependence.

Consider a semiconductor, uniformly doped to the level, N_A , with acceptor. Let the local density of mobile donor be $N_D(x)$, x being the position coordinate. If $P(x)$ is the local pair concentration, then the local density of free diffusible ions is $(N_D - P)$. The flux of these diffusing ions then depends upon the gradient (assuming Fick's law⁴⁹) of $(N_D - P)$. Thus, if D_0 is the diffusivity of free donor, i.e. the diffusivity in the absence of pairing, then the flux density is

$$f = -D_0 \frac{\partial(N_D - P)}{\partial x} \quad (\text{B1})$$

If we apply (9.4) to the present case we can write

$$\Omega = \frac{P}{(N_A - P)(N_D - P)} = \frac{-(N_D - P) + N_D}{[(N_A - N_D) + (N_D - P)](N_D - P)} \quad (\text{B2})$$

from which it is possible to solve for $(N_D - P)$. Thus

$$N_D - P = \frac{1}{2} \left(N_D - N_A - \frac{1}{\Omega} \right) + \sqrt{\frac{1}{4} \left(N_D - N_A - \frac{1}{\Omega} \right)^2 + \frac{N_D}{\Omega}} \quad (\text{B3})$$

Substitution of (B3) into (B1) yields

$$f = -\frac{D_0}{2} \left[1 + \frac{\frac{1}{2} \left(N_D - N_A + \frac{1}{\Omega} \right)}{\sqrt{\frac{1}{4} \left(N_D - N_A - \frac{1}{\Omega} \right)^2 + \frac{N_D}{\Omega}}} \right] \frac{\partial N_D}{\partial x} \quad (\text{B4})$$

If ion pairing was not thought of, the flux density would have been written in terms of the gradient of the total concentration, N_D .

$$f = -D \frac{\partial N_D}{\partial x} \quad (\text{B5})$$

where D is the diffusivity. Comparison of (B5) with (B4) leads to the relation

$$D = \frac{D_0}{2} \left[1 + \frac{\frac{1}{2} \left(N_D - N_A + \frac{1}{\Omega} \right)}{\sqrt{\frac{1}{4} \left(N_D - N_A - \frac{1}{\Omega} \right)^2 + \frac{N_D}{\Omega}}} \right] \quad (\text{B6})$$

so that D depends on the local concentration, N_D , of diffusant.

It is interesting to explore the limiting forms of D when $N_D \ll N_A$ and when $N_D = N_A$. In the latter case (B6) reduces to

$$D = \frac{D_0}{2} \left[1 + \frac{\frac{1}{2\Omega}}{\sqrt{\frac{1}{4\Omega^2} + \frac{N_A}{\Omega}}} \right] \quad (\text{B7})$$

while (B3) becomes

$$N_A - P + \frac{1}{2\Omega} = \sqrt{\frac{1}{4\Omega^2} + \frac{N_A}{\Omega}}. \quad (\text{B8})$$

Substituting the left side of (B8) for the denominator involving the radical in (B7) leads to

$$D = \frac{D_0}{2} \left[1 + \frac{1}{2(N_A - P)\Omega + 1} \right] \quad (\text{B9})$$

But according to (B2), when $N_A = N_D$,

$$(N_A - P)\Omega = \frac{P}{N_A - P} \quad (\text{B10})$$

so that (B9) becomes

$$D = \frac{D_0}{2} \left[1 + \frac{1}{\frac{2P}{N_A - P} + 1} \right] \quad (\text{B12})$$

Now in case the degree of pairing is high (which is, of course, the case we are interested in) P will be almost equal to N_A so that

$$\frac{2P}{N_A - P} \quad (\text{B13})$$

will be a very large number. If this is so the second term in brackets in (B12) can be set equal to zero and we have

$$D = \frac{D_0}{2}. \quad (\text{B14})$$

In the other extreme with $N_D \ll N_A$ (B6) becomes

$$D = \frac{D_0}{2} \left[1 + \frac{\frac{1}{2} \left(\frac{1}{\Omega} - N_A \right)}{\sqrt{\frac{1}{4} \left(\frac{1}{\Omega} + N_A \right)^2}} \right] = \frac{D_0}{1 + \Omega N_A} \quad (\text{B15})$$

Since ΩN_A can exceed unity by a large amount it is evident that the relation in (B15) predicts a large reduction in diffusivity towards the front end of a diffusion curve where $N_D \ll N_A$, and (B14) a smaller reduction in D_0 where N_D may be close to N_A . That part of the medium near the front of the diffusion curve acts therefore like a region of high resistance, confining the diffusant to the back end where the resistance is low.

APPENDIX C

SOLUTION OF BOUNDARY VALUE PROBLEM FOR RELAXATION

In Section X equations (10.23), (10.21), (10.20), and (10.19) defined a boundary value problem which we reproduce here, except that (10.20) and (10.19) have been written more completely with the aid of (10.16). Thus

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \rho}{\partial r} + R\rho \right) = \frac{1}{D_0} \frac{\partial \rho}{\partial t} \quad (\text{C1})$$

$$\frac{\partial \rho}{\partial r} + \frac{R}{r^2} \rho = 0, \quad r = L, \quad r = a \quad (\text{C2})$$

$$\rho = N^2, \quad t = 0, \quad a < r < L \quad (\text{C3})$$

In principle this problem is soluble by separation of variables.⁵⁶ Thus we define

$$\rho(r, t) = G(r) S(t) \quad (\text{C4})$$

which upon substitution into (C1), yields the two ordinary differential equations

$$\frac{d}{dr} \left[r^2 \frac{dG}{dr} + RG \right] + \eta^2 G = 0 \quad (\text{C5})$$

$$\frac{d \ln S}{dt} + \eta^2 D_0 = 0 \quad (\text{C6})$$

where η^2 is an arbitrary positive parameter.

The allowable values of η are determined by (C2) which can now be

replaced by

$$\frac{dG}{dr} + \frac{R}{r^2} G = 0, \quad r = L, \quad r = a \quad (C7)$$

Equation (C6) can be solved immediately to give

$$S_\eta(t) = e^{-\eta^2 D_0 t} \quad (C8)$$

and if we assign the subscript η to the G going with η the most general solution of (C1) and (C2) will be

$$\rho = \sum_{\eta} A_{\eta} G_{\eta}(r) e^{-\eta^2 D_0 t} \quad (C9)$$

where the A_{η} are arbitrary constants so determined that (C3) is satisfied.

Equation (C9) shows that in reality there exists, for this problem, a spectrum of relaxation times, $1/\eta^2 D_0$. After a brief transient period many of the higher order terms will decay away and eventually only the first two terms will have to be considered. Finally when equilibrium is attained only the first term will survive.

The last statement implies that $\eta = 0$, is an allowable eigenvalue, i.e., that the first term is independent of time. That this is so can be proved by solving (C5) for $\eta = 0$, and substituting the result in (C7). Thus

$$G_0(r) = \exp\left(\frac{R}{r}\right) \quad (C10)$$

and this does satisfy (C7). ρ can then be approximated after the transient by

$$\rho = A_0 \exp\left(\frac{R}{r}\right) + A_1 G_1(r) e^{-\eta_1^2 D_0 t} \quad (C11)$$

from which it is obvious that the relaxation time dealt with in section X is

$$\tau = \frac{1}{\eta_1^2 D_0} \quad (C12)$$

In principle it should be possible to evaluate G_1 by the straightforward solution of (C5) and determination of the second eigenvalue through substitution of this solution in (C7). In fact this represents a rather unpleasant task since G is a confluent hypergeometric function.⁵⁷ Therefore we shall follow an alternative route based on the assumption that by the time (C11) applies the flux $4\pi r^2 J^*(r)$, where J^* is given by (10.16), is almost independent of r . The reader is referred to some related papers^{58, 59}

for the justification of this view. Briefly it is permissible, after a short transient period, in spherical diffusion, whenever the dimensions of the diffusion field are large compared to the dimension of the sink. This results from the fact that in spherical diffusion from an infinite field⁴⁸ a real steady state is reached after a brief transient period. In contrast, in plane-parallel diffusion to a sink from an infinite field,⁶⁰ a steady state is never reached.

Substituting (C11) into (10.16) then yields

$$J^* = -D_0 A_1 e^{-\eta_1^2 D_0 t} \left(\frac{dG_1}{dr} + \frac{R}{r^2} G_1 \right) \quad (\text{C13})$$

Multiplying J^* by $4\pi r^2$ and demanding that the product be independent of r , leads to the relation

$$r^2 \frac{dG_1}{dr} + RG_1 = \delta \quad (\text{C14})$$

where δ is constant. The solution of (C14) is

$$G_1 = \exp \left(\frac{R}{r} \right) + \frac{\delta}{R} \quad (\text{C15})$$

This is a sufficient approximation for G_1 .

The constants η_1 , A_0 , A_1 , and δ must now be determined. To accomplish this we note that (C2) which specifies that the boundaries at $r = a$ and $r = L$, are impermeable is equivalent to the condition that ions be conserved with the interval (a, L) , or that

$$4\pi \int_a^L r^2 \rho dr = N \quad (\text{C16})$$

After infinite time ρ is specified by the first term of (C11) and when this is inserted into (C16) the result is

$$A_0 = NM \quad (\text{C17})$$

where M is defined by (10.26).

Substitution of (C17) and (C15) into (C11) gives

$$\rho = NM \exp(R/r) + \left(A_1 \exp(R/r) + \frac{A_1 \delta}{R} \right) e^{-\eta_1^2 D_0 t} \quad (\text{C18})$$

Now (C3) applied to (C18) demands

$$NM + A_1 = 0 \quad (\text{C19})$$

$$\frac{A_1 \delta}{R} = N^2 \quad (\text{C20})$$

Of course this presumes that the approximation contained in (C18) is valid down to very small values of time. This assumption is well founded as the transient does vanish after a rather short time.

Inserting (C19) and (C20) in (C18) then gives us

$$\rho = NM \exp(R/r) + N[N - M \exp(R/r)]e^{-\eta_1^2 D_0 t} \quad (C21)$$

in which only η_1 remains to be determined.

Substitution of (C21) into (C16), recalling the definitions of M and L , shows that it already satisfies (C16) for any time, t . Thus (C16) cannot be used for determining η_1 .

On the other hand we note from (C21) that as soon as r becomes of order, R , ρ becomes almost independent of r , being given

$$\rho = N\{N + (N - M)e^{-\eta_1^2 D_0 t}\} \quad (C22)$$

Since L is of the order $10R$ or greater, this means that throughout most of the volume, $1/N$ (in fact throughout $0.999 1/N$) ρ is independent of r . Effectively, the entire volume $1/N$ has been drained of ions, i.e., they have been trapped. The total ion content at time t , may then be taken as the product of ρ , given by (C22), with $1/N$, that is,

$$N + (N - M)e^{-\eta_1^2 D_0 t} \quad (C23)$$

The time rate of change of this content must be given by the flux $4\pi r^2 J^*$.

$$\begin{aligned} \frac{d}{dt} [N + (N - M)e^{-\eta_1^2 D_0 t}] \\ &= -\eta_1^2 D_0 (N - M)e^{-\eta_1^2 D_0 t} = 4\pi r^2 J^*(r, t) \\ &= -4\pi R N^2 D_0 e^{-\eta_1^2 D_0 t} \end{aligned} \quad (C24)$$

in which (C21) has been substituted into (10.16) to pass from the third to the fourth expression. Comparing the second and fourth term of (C24) reveals

$$\eta_1^2 D_0 = \frac{4\pi R N^2 D_0}{(N - M)} = \frac{4\pi q^2 N^2 D_0}{\kappa k T (N - M)} \quad (C25)$$

or

$$\tau = \frac{1}{\eta_1^2 D_0} = \frac{\kappa k T (N - M)}{4\pi q^2 N^2 D_0} \quad (C26)$$

the value quoted in (10.25).

APPENDIX D

MINIMIZATION OF THE DIFFUSION POTENTIAL

In Section V the statement was made that equation (11.2) was a valid approximation everywhere within a *p* type region, provided that N_D did not fluctuate through ranges of order N_A in shorter distances than

$$\ell = \sqrt{\frac{\pi\kappa kT}{q^2 N_A}} \quad (\text{D1})$$

This statement will now be proved.

The electrostatic potential is determined by the space charge equation³¹

$$\frac{d^2V}{dx^2} = -\frac{4\pi q}{\kappa} [N_D(x) + p(x) - N_A] \quad (\text{D2})$$

where we assume that the material is everywhere *p*-type so that the electron density, n , does not enter the right side of (D2). Furthermore, the mobility of holes is so much greater than that of donor ions that the former may be considered to always be at equilibrium with respect to the distribution of the latter. Boltzmann's law²⁹ may then be applied to p . The result is

$$p = N_A \exp [-qV/kT] \quad (\text{D3})$$

where the potential is taken to be zero when $p = N_A$.

Choose an arbitrary point, x_0 , where the potential is V_0 and investigate (D2) in its neighborhood. We wish to determine the conditions under which the right side of (D2) may be approximated by zero, i.e., the "no-space-charge condition," in this neighborhood. The limits of the neighborhood will be defined such that

$$|V - V_0| = |u| \leq kT/2q \quad (\text{D4})$$

so that, in it, the exponential in (D3) can be linearized

$$p = N_A \exp [-qV_0/kT] \left(1 - \frac{qu}{kT}\right) \quad (\text{D5})$$

Then (D2) becomes

$$\begin{aligned} \frac{d^2u}{dx^2} = \frac{4\pi q}{\kappa} & \left\{ N_A [1 - \exp (-qV_0/kT)] - N_D(x) \right. \\ & \left. + \left[\frac{qN_A}{kT} \exp (-qV_0/kT) \right] u \right\} \end{aligned} \quad (\text{D6})$$

The no space charge condition in the defined region is therefore

$$u = \left(\frac{kT \exp(qV_0/kT)}{qN_A} \right) N_D(x) + \left(\frac{kT}{q} \right) \frac{\exp(-qV_0/kT) - 1}{\exp(-qV_0/kT)} \quad (\text{D7})$$

To simplify notation define

$$\exp[-qV_0/kT] = \gamma_0 \quad (\text{D8})$$

Next expand both N_D and u in Fourier series

$$N_D = \sum_{s=0}^{\infty} A_s \sin sx + B_s \cos sx \quad (\text{D9})$$

$$u = \sum_{s=0}^{\infty} \alpha_s \sin sx + \beta_s \cos sx \quad (\text{D10})$$

Substitution of (D9) and (D10) into (D6) and equating coefficients of like terms leads to the set of relations

$$\beta_0 = \frac{kT}{qN_A \gamma_0} [N_A(\gamma_0 - 1) + B_0] \quad (\text{D11})$$

$$\alpha_s = \frac{4\pi q}{\kappa} \left(\frac{\ell^2/4\pi^2 \gamma_0}{1 + (s^2 \ell^2/4\pi^2 \gamma_0)} \right) A_s \quad (\text{D12})$$

$$\beta_s = \frac{4\pi q}{\kappa} \left(\frac{\ell^2/4\pi^2 \gamma_0}{1 + (s^2 \ell^2/4\pi^2 \gamma_0)} \right) B_s \quad (\text{D13})$$

Now the wavelength of the s th component in (D9) is

$$\lambda_s = 2\pi/s \quad (\text{D14})$$

If N_D contains no important components of wavelength shorter than

$$\frac{\ell}{\sqrt{\gamma_0}} \quad (\text{D15})$$

the B_k for such components may be set equal to zero. But then the only terms which appear in (D12) and (D13) are terms where the denominators which (with the aid of (D14)) may be written as

$$\kappa \left(1 + \frac{\ell^2}{\gamma_0 \lambda_s^2} \right) \quad (\text{D16})$$

may be set equal to κ . Thus we have in place of (D12) and (D13)

$$\alpha_s = \frac{q\ell^2}{\kappa\pi} A_s = \frac{kT}{qN_A \gamma_0} A_s \quad (\text{D17})$$

$$\beta_s = \frac{q^2 \ell^2}{\kappa \pi} B_s = \frac{kT}{qN_A \gamma_0} B_s \quad (D18)$$

The requirement that N_D contain no Fourier terms of wavelength shorter than (D15) is obviously the condition that N_D never pass from its maximum to its minimum value in a distance shorter than D(15). Since we are assuming that N_D may at places be of order N_A , and at others, of order zero, this amounts to the condition that N_D does not fluctuate over ranges comparable with N_A in distances shorter than (D15).

The use of (D11), (D17), and (D18) in (D10) yields

$$\begin{aligned} u &= \frac{kT}{qN_A \gamma_0} \left[N_A(\gamma_0 - 1) + \sum_{s=0}^{\infty} (A_s \sin sx + B_s \cos sx) \right] \\ &= \frac{kT}{q} \frac{(\gamma_0 - 1)}{(\gamma_0)} + \frac{kT}{q \gamma_0} \frac{N_D}{N_A} \end{aligned} \quad (D19)$$

which by reference to the definition (D8) for γ_0 proves to be identical with (D7), the no-space-charge condition.

Equation (D19) is only true when N_D does not fluctuate through ranges of order, N_A , in distances smaller than $\ell/\sqrt{\gamma_0}$. This distance depends on γ_0 and thus on the point where $V = V_0$, whose neighborhood is being explored. Thus, we may say that there will be no space charge at all points whose V_0 is such as to fix γ_0 at a value such that

$$\gamma_0 > \frac{\ell^2}{\lambda_{\min}^2} \quad (D20)$$

where λ_{\min} is the minimum wavelength which needs to be considered in the Fourier expansion of N_D . In terms of the definition of γ_0 this means

$$V_0 < \frac{kT}{q} \ell n \frac{\lambda_{\min}^2}{\ell^2} \quad (D21)$$

Thus, at all points where V_0 is less than the right side of (D21) the no space charge approximation will hold. (D21) shows, that in the limit when λ_{\min} goes toward zero, i.e. when the infinite series must be used for N_D , the right side of (D21) will approach $-\infty$ and V_0 will satisfy (D21) hardly anywhere. Thus space charge will exist almost everywhere.

In most diffusion problems the extremes of potential will occur in regions where there is no space charge. Thus in one extreme N_D may equal $0.9 N_A$ and in the other it may equal zero. If there is no space charge in these extremes we may write for them

$$N_A - N_D = p = N_A \exp(-qV/kT) \quad (D22)$$

in which (D3) has been used. Setting N_D equal to zero in one extreme

yields $V = 0$. In the other extreme $N_D = 0.9 N_A$ so that we get

$$V = \frac{kT}{q} \ell n 10 \quad (D23)$$

This therefore is the largest value which V_0 may assume in our case. Inserting the expression in D21 in place of V_0 we end with the relation

$$10 < \frac{\lambda_{\min}^2}{\ell^2} \quad (D24)$$

Thus provided that in the distribution being considered

$$\lambda_{\min} > 3.5\ell \quad (D25)$$

there will be no space charge anywhere.

At high temperatures $0.1 N_A$ may be less than n_i . Under these conditions (D24) should be replaced by

$$\frac{N_A}{n_i} < \frac{\lambda_{\min}^2}{\ell^2} \quad (D26)$$

and in the limit that n_i becomes very large it is obvious that (D26) will always be satisfied. The rule to be enunciated for the cases we shall be interested in is the one given in section XI, i.e. that no space charge will exist provided that λ_{\min}^2 is no less than order, ℓ .

APPENDIX E

CALCULATION OF DIFFUSIVITIES FROM CONDUCTANCES OF DIFFUSION LAYERS

In this appendix equation (11.12) will be derived. In the first place we note that the dependence of N_D on position x , and time t , will be of the form $N_D(x/\sqrt{t})$ at any stage of the diffusion process. This results from a theorem due to Boltzmann⁶¹ that when the dependence of D upon x and t is of the form $D(N_D)$, i.e., the dependence is through N_D , and a semi-infinite region extending from $x = 0$ to $x = \infty$ is being considered, then, in the case of plane parallel diffusion, the only variable in the problem will be x/\sqrt{t} .

Although the wafers considered in Section XI are of finite thickness d , the stages of diffusion investigated are such that the two regions of loss near the surfaces have not contacted each other. As a result the system behaves like two semi-infinite regions backed against one another, and the preceding arguments hold. The conductance Σ , defined in section XI will be proportional to the integral of the product of the local carrier

density by the local mobility. Thus

$$\Sigma = \omega \int_0^{d/2} \mu(x, t) [N_A - N_D(x, t)] dx \quad (E1)$$

where ω is a proportionality constant and $\mu(x, t)$ is the local mobility. An upper limit of $d/2$ rather than d is used because of symmetry. The local mobility will vary because N_D , and therefore the local density of charged impurity scatterers,⁵⁴ varies. Let N_D^0 be the initial uniform density (before any diffusion out) of donors, and write (E1) as

$$\begin{aligned} \Sigma &= \omega \int_0^{d/2} \mu(x, t) [N_A - N_D^0 + N_D^0 - N_D(x, t)] dx \\ &= \omega \int_0^{d/2} \mu(x, t) [N_A - N_D^0] dx + \omega \int_0^{\infty} \mu(x, t) [N_D^0 \\ &\quad - N_D(x, t)] dx \end{aligned} \quad (E2)$$

The second integral on the right of (E2) is given the upper limit ∞ , because in the experiments we wish to perform $N_D^0 - N_D$ becomes zero long before x reaches $d/2$.

Now in the first integral on the right of (E2) we may set $\mu(x, t)$ equal to the constant value μ_0 , which it assumes in the bulk of the wafer, because the breadth of the depletion layer near the surface (in which $\mu(x, t)$ departs from μ_0) is small compared to $d/2$. The same thing cannot be done in the second integral since the integrand vanishes beyond the depletion layer and the total contribution comes from that layer. We thus obtain

$$\begin{aligned} \Sigma &= \omega \mu_0 (N_A - N_D^0) d/2 \\ &\quad + \omega \int_0^{\infty} \mu \left(\frac{x}{\sqrt{t}} \right) \left[N_D^0 - N_D \left(\frac{x}{\sqrt{t}} \right) \right] dx \end{aligned} \quad (E3)$$

In the integral in (E3) both μ and N_D are represented as functions of x/\sqrt{t} , the latter because of what has been said above, and the former, because it is a function of the latter. Defining

$$\nu = x/2\sqrt{Dt} \quad (E4)$$

in which D is constant, and substituting in (E3) gives finally

$$\Sigma = \omega \mu_0 (N_A - N_D^0) d/2 + 2\omega \sqrt{Dt} \int_0^{\infty} \mu(\nu) [N_D^0 - N_D(\nu)] d\nu \quad (E5)$$

Since the definite integral is a constant (E5) shows that Σ is a linear function of \sqrt{t} , a fact mentioned in section XI.

In order to make use of the measured dependence of Σ on \sqrt{t} to determine diffusivities, the functions $\mu(\nu)$ and $N_D(\nu)$ must be specified. For the latter we shall assume the Fick's law solution⁶²

$$N_D = N_D^0 \operatorname{erf} \nu \quad (\text{E6})$$

going with constant D , and $N_D = 0$ as a boundary condition at $x = 0$ at the surface. (In section XI the limitations of this assumption in the presence of ion pairing and diffusion potential are discussed.) The ν dependence of μ is more complicated. In general, we shall be concerned with electrical measurements in two extreme cases. In the first case ion pairing, under the condition of measurement, is everywhere complete so that the local density of scatterers will be given by

$$N_A = N_D(\nu) \quad (\text{E7})$$

In the other case ion pairing will be entirely absent, so that the local scatterer density, will be specified by

$$N_A + N_D(\nu) \quad (\text{E8})$$

In all experiments N_A will be only slightly greater than N_D^0 so that it may be replaced by this quantity. Doing this, and substituting (E6) into (E8) and (E9) gives

$$N_D^0 \operatorname{erfc} \nu = N(\nu) \quad (\text{E9})$$

for the scattering density in the ion pairing case, and

$$N_D^0(1 + \operatorname{erf} \nu) = N(\nu) \quad (\text{E10})$$

for the no pairing case.

Since almost all our experiments have been in germanium we now specialize our attention to that substance. However, the procedure invoked below can be applied to silicon as well.

The dependence of hole mobility, μ , on scattering density, N , for germanium at room temperature is shown in Fig. 30 taken from Prine's data.⁶³ The integral in (E5) assumes the form

$$N_D^0 \int_0^\infty \mu(N(\nu)) \operatorname{erfc} \nu d\nu. \quad (\text{E11})$$

Choosing $N(\nu)$ as either (E9) or (E10) and using Fig. 30 together with a table of error functions makes the numerical evaluation of (E11) possible. Since $N(\nu)$ given by (E9) or (E10) depends on N_D^0 , so will the integral.

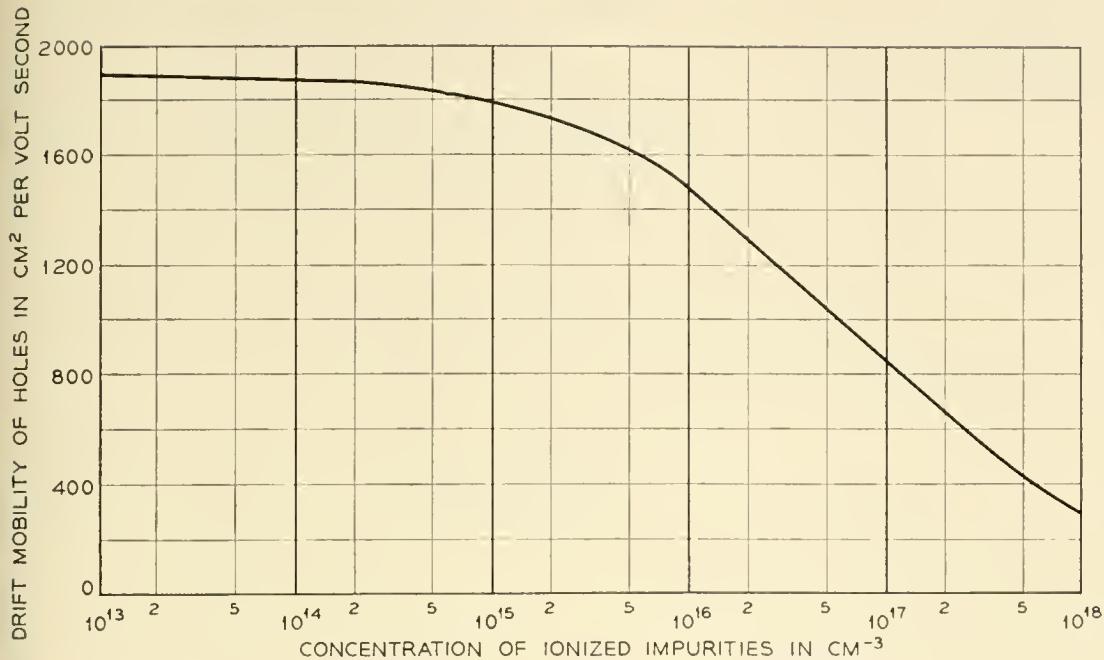


Fig. 30 — Plot of hole-drift mobility in germanium as a function of ionized impurity concentration after Prince.

The numerical evaluation has been performed for a range of N_D^0 in both the pairing and non-pairing cases. In this manner it has been possible to evaluate the "correction factor" ϑ defined by the following equation

$$\int_0^\infty \mu(\nu) \operatorname{erfc} \nu d\nu = \vartheta \mu_\infty \int_0^\infty \operatorname{erfc} \nu d\nu \quad (\text{E12})$$

$$= \vartheta \mu_\infty (0.563)$$

where μ_∞ is the mobility in the presence of N_A scatterers. Fig. 22 contains plots (for germanium) of $\vartheta(N_D^0)$ versus N_D^0 for both the pairing and non-pairing cases. It is seen that ϑ is never much different from unity.

Equation (E5) can now be written as

$$\Sigma = \omega \mu_0 (N_A - N_D^0) d/2 + \omega \mu_\infty [1.128 \vartheta N_D^0 \sqrt{D}] \sqrt{t} \quad (\text{E13})$$

Defining

$$\Sigma_0 = \omega \mu_0 (N_A - N_D^0) d/2 \quad (\text{E14})$$

$$\Sigma_\infty = \frac{\omega \mu_\infty N_A D}{2} \quad (\text{E15})$$

it is obvious that Σ_0 is the conductance before any donor has diffused

out and Σ_∞ after all the donor has been diffused out. With these definitions (E16) becomes

$$\frac{\Sigma}{\Sigma_0} = 1 + \frac{2.256 \vartheta \mu_\infty \sqrt{D}}{\mu_0 d} \left(\frac{N_D^0}{N_A - N_D^0} \right) \sqrt{t} \quad (\text{E17})$$

Calling the slope of this curve S leads to the result

$$S = \frac{2.256 \vartheta \mu_\infty \sqrt{D}}{\mu_0 d} \left(\frac{N_D^0}{N_A - N_D^0} \right) \quad (\text{E18})$$

or using (E14) and (E15)

$$D = \left(\frac{S d \Sigma_0 N_A}{2.256 \vartheta \Sigma_\infty N_D^0} \right)^2 \quad (\text{E19})$$

This is equivalent to equation (11.12).

GLOSSARY OF SYMBOLS

a	distance of closest approach of two ions of opposite sign
A	constant in expression for ρ in section on relaxation theory
A^-	concentration of ionized acceptors
A_0	A_η going with $\eta = 0$
A_1	A_η going with η_1
A_η	constant preceding the η th eigenfunction in solution of the relaxation problem
A_s	coefficient of $\sin sx$ in Fourier expression for N_D
b	$q^2/2\kappa kT$, position of minimum in $g(r)$
B	constant in expression for ρ in section on relaxation theory
B^-	boron ion
$B(Si)$	un-ionized boron in silicon
B_s	coefficient of $\cos sx$ in Fourier expression for N_D
$c(r)$	concentration of positive ions in atmosphere of a negative ion
C	concentration of LiB^-
d	thickness of wafer in diffusion experiment
D	diffusivity of donor ion in the most general sense
D_0	diffusivity of donor ion in the absence of pairing
D^+	concentration of ionized donors
D_0^+	value of D^+ in the absence of acceptor
D_*^+	concentration of mobile donor ions where $V = 0$
e^-	conduction band electron
e^+	valence band hole

e^+e^-	recombined hole-electron pair
E	energy level in electron gas
E_D	ionization energy of a donor
E_A	ionization energy of an acceptor
E_i	energy level in conduction band
$E(r)$	chance that volume $4\pi r^3/3$ will not contain an ion
f	flux density
F	Fermi level — also constant in equation (7.21)
g_i	density of states of energy E_i in conduction band
$g(r)$	nearest neighbor distribution function at equilibrium
G	Gibbs free energy of electron assembly
Ga^-	gallium ion in germanium
G_η	space dependent part of relaxation eigenfunction
G_0	G_η for $\eta = 0$
G_1	G_η for $\eta = \eta_1$
h	Plank's constant — also used for normalizing constant in $c(r)$
h_j	number of holes in the j th energy level
H	net local density of fixed donors
$i(\rho_2, \rho_1)$	$\epsilon^3 I(r_2, r_1)$
I	field current in diffusion measurement
$I(r_2, r_1)$	integral for ion pairing calculations taken between r_1 and r_2
$\vec{J}(\vec{r})$	current in the atmosphere of a nearest neighbor
J^*	flux density of ions being trapped
k	Boltzmann's constant
k_1	first order rate constant in relaxation theory
k_2	second order rate constant in relaxation theory
K_0	distribution coefficient of donor between semiconductor and external phase
K_1	electron-hole recombination equilibrium constant
K_A	ionization constant of acceptor
K_D	ionization constant of donor
K_j	constant relating ω_j to volume, V
K^*	product of K_D , K_0 , and α
ℓ	screening length for diffusion potential
L	Debye length — also used for radius of volume, $1/N$
Li^+	lithium ion
$Li(Sn)$	lithium in molten tin
$Li(Si)$	un-ionized lithium in silicon
$LiSi$	lithium-silicon complex
LiB	un-ionized LiB^-

LiB^-	lithium-boron complex ion in semiconductor
$[Li^+B^-]$	lithium-boron ion pair
$[Li^+Ga^-]$	lithium-gallium ion pair
m_0	normal mass of electron
m_p	effective mass of a hole
M	normalizing constant in relaxation theory
n	concentration of conduction electrons — also used for density of untrapped ions in relaxation
n_i	intrinsic concentration of electrons
N_A	total acceptor concentration
N_D	total donor concentration
N_D^0	total solubility of donor in undoped semiconductor—also used for initial density of donors in diffusion experiments
N	ion concentration in an electrolyte solution—also used for initial value of n in relaxation—also used for concentration of immobile donors in Appendix A
N_D^*	solubility of donor in absence of ion pairing in Appendix A
p	concentration of holes
P	concentration of ion pairs
q	charge on an ion
$Q(\alpha)$	tabulated integral for computing Ω
r	distance between positive and negative ions in a pair
r_1	a particular value of r
r_2	a particular value of r
R	capture radius of an ion in relaxation
S	slope of Σ/Σ_0 versus \sqrt{t} curve
S_η	time dependent part of relaxation eigenfunction belonging to eigenvalue η
t	time
T	temperature
u	$V - V_0$
V	electrostatic potential — also used for volume — also used for potential difference between probe points — also used for potential energy of a positive in neighborhood of negative ion
V_0	electrostatic potential where $x = x_0$
x	variable of integration — same as r also rectangular position coordinate
x_0	special value of x .

α	ε/a —also used for thermodynamic activity of donor in external phase
α_s	coefficient of $\sin s\bar{x}$ in Fourier expression for u
β	constant in exponential in LiB^- equilibrium constant
β^*	constant in exponential in expression for vacancy concentration
β_0	β_s for $s = 0$
β_s	coefficient of $\cos s\bar{x}$ in Fourier expression for u
γ	pre-exponential factor in LiB^- equilibrium constant
γ^*	pre-exponential factor in expression for vacancy concentration
γ_0	$\exp[-qV_0/kT]$
$\Gamma(\vec{r})$	non-equilibrium nearest neighbor distribution function
δ	constant appearing in Appendix C
ε	$q^2/\kappa kT$
r	eigenvalue in relaxation problem
η_1	second eigenvalue in set of η
θ	fraction of donor paired
ϑ	correction factor for variable carrier mobility
κ	dielectric constant
λ	x/ε
λ_s	$2\pi/s$, wavelength of s th component of Fourier series
λ_{\min}	wavelength of component of Fourier series for N_D , having minimum wavelength
μ	chemical potential of donor in an external phase — also used for mobility of donor ion — also used for local carrier mobility
μ^0	chemical potential of donor in external phase in standard state
μ_{D+}	chemical potential of donor ion
μ_{D+}^0	chemical potential of donor ion in the standard state
μ_e	chemical potential of an electron
μ_D	chemical potential of donor atom in semiconductor
μ_D^0	chemical potential of donor atom in standard state
μ_0	mobility of donor atom at infinite dilution — also used for carrier mobility in diffusion experiments before diffusion
μ_∞	carrier mobility in diffusion experiments after all diffusant has diffused out
v	$x/2\sqrt{Dt}$
ξ	ε/r .
π	LiB^- equilibrium constant
ρ	resistivity of gallium-doped germanium after saturation with

	lithium — also used for local charge density in Poisson's equation — also used for density of diffusing positive ions in relaxation
ρ_0	resistivity of gallium-doped germanium before saturation with lithium
ρ_1	r_1/ϵ
ρ_2	r_2/ϵ
σ	conductivity during relaxation
σ_∞	conductivity in relaxed state
Σ	conductance between probe points
Σ_0	conductance before diffusion begins in diffusion experiments
Σ_∞	conductance after diffusion is over in diffusion experiments
τ	relaxation time
Φ	constant in relaxation formula for conductivity
Ψ	local electrostatic potential in ionic atmosphere
ω	proportionality constant connecting conductance between probe points with integral over carrier concentration
ω_j	number of states in j th level of electronic energy diagram
Ω	ion pairing equilibrium constant
\square	vacant lattice site in covalent crystal
\square^-	negatively charged cation vacancy

REFERENCES

1. Wagner, C., Z. Phys. Chem., **B21**, p. 25, 1933, **B32**, p. 447, 1936.
2. Taylor, H. S., and Taylor, H. A., Elementary Physical Chemistry, p. 343, Van Nostrand, 1937.
3. Shockley, W., Electrons and Holes in Semiconductors, p. 6, Van Nostrand, 1950.
4. Shockley, W., Electrons and Holes in Semiconductors, Van Nostrand, 1950.
5. Reiss, H., J. Chem. Phys., **21**, p. 1209, 1953.
6. Reiss, H., and Fuller, C. S., J. Metals, **12**, p. 276, 1956.
7. Fuller, C. S., and Ditzenberger, J. A., J. App. Phys., **25**, p. 1439, 1954.
8. Fuller, C. S., and Ditzenberger, J. A., Phys. Rev., **91**, p. 193, 1953.
9. MacDougall, F. H., Thermodynamics and Chemistry, p. 143, Wiley, 1939.
10. Miller, F. W., Elementary Theory of Qualitative Analysis, p. 102, Century Company, New York, 1929.
11. Fuller, C. S., Record of Chemical Progress, **17**, No. 2, 1956.
12. Wagner, C., and Grünwald, K., Z. Phys. Chem., **B40**, p. 455, 1938.
13. von Baumbach, H. H., and Wagner, C., Z. Phys. Chem., **22B**, p. 199, 1933.
14. Kröger, F. A., and Vink, H. J., Physica, **20**, p. 950, 1954.
15. MacDougall, F. H., Thermodynamics and Chemistry, p. 258, Wiley, 1939.
16. Shockley, W., Electrons and Holes in Semiconductors, p. 231, Van Nostrand, 1950.
17. Mayer, J. E., and Mayer, M. G., Statistical Mechanics, p. 120, Wiley, 1940.
18. MacDougall, F. H., Thermodynamics and Chemistry, p. 137, Wiley, 1939.
19. Lewis, G. N., and Randall, M. C., Thermodynamics, p. 258, McGraw Hill, 1923.

20. Mayer, J. E., and Mayer, M. G., Statistical Mechanics, p. 121, Wiley, 1940.
21. MacDougall, F. H., Thermodynamics and Chemistry, p. 261, Wiley, 1939.
22. MacDougall, F. H., Thermodynamics and Chemistry, p. 25, Wiley, 1939.
23. Engell, H. J., and Houffe, K., *Z. Electrochem.*, **56**, p. 366, 1952, **57**, p. 762, 1953.
24. Shockley, W., Electrons and Holes in Semiconductors, p. 15, Van Nostrand, 1950.
25. Morin, F. J., and Maita, J. P., *Phys. Rev.*, **94**, p. 1525, 1954.
26. Morin, F. J., and Maita, J. P., *Phys. Rev.*, **96**, p. 28, 1954.
27. Shockley, W., Electrons and Holes in Semiconductors, p. 86, Van Nostrand, 1950.
28. Shockley, W., Electrons and Holes in Semiconductors, p. 88, Van Nostrand, 1950.
29. Fowler, R. H., Statistical Mechanics, p. 48, Cambridge, 1929.
30. Slater, J. C., and Frank, N. H., Introduction to Theoretical Physics, p. 212, McGraw Hill, 1933.
31. Shockley, W., *B.S.T.J.*, **28**, p. 435, 1949.
32. Fuller, C. S., and Ditzingerger, J. A., *J. App. Phys.*, May, 1956.
33. Shulman, R. G., and McMahon, M. E., *J. App. Phys.*, **24**, p. 1267, 1953.
34. Reiss, H., Fuller, C. S., and Pietruszkiewicz, A. J., *J. Chem. Phys.* (in press).
35. Eyring, H., Walter, J., and Kimball, G. E., Quantum Chemistry, p. 231, Wiley, 1946.
36. Pauling, L., The Nature of the Chemical Bond, p. 179, Cornell, 1942.
37. Debye, P., and Huckel, E., *Phys. Z.*, **24**, p. 195, 1923.
38. Kirkwood, J. G., *J. Chem. Phys.*, **2**, p. 767, 1934.
39. Briggs, H. B., *Phys. Rev.*, **77**, p. 287, 1950.
40. Briggs, H. B., *Phys. Rev.*, **77**, p. 287, 1950.
41. Wyman, *Phys. Rev.*, **35**, p. 623, 1930.
42. Bjerrum, N., *Kgle. Danske Vidensk. Selskab.*, **7**, No. 9, 1926.
43. Fuoss, R. M., *Trans. Faraday Soc.*, **30**, p. 967, 1934.
44. Reiss, H., *J. Chem. Phys.* (in press).
45. Reiss, H., *J. Chem. Phys.* (in press).
46. Shockley, W., and Read, W. T., Jr., *Phys. Rev.*, **87**, p. 835, 1952, Haynes, J. R., and Hornbeck, J. A., *Phys. Rev.*, **90**, p. 152, 1953, **97**, p. 311, 1955.
47. Harned and Owen, The Physical Chemistry of Electrolytes, p. 123, A. C. S. Monograph, 1950.
48. Carslaw, H. S., and Jaeger, J. C., Conduction of Heat in Solids, p. 209, Oxford, 1948.
49. Glasstone, S., Textbook of Physical Chemistry, p. 1231, Van Nostrand, 1940.
50. Shockley, W., Electrons and Holes in Semiconductors, p. 300, Van Nostrand, 1950.
51. Slater, J. C., and Frank, N. H., Introduction to Theoretical Physics, p. 186, McGraw Hill, 1933.
52. Fuller, C. S., and Severiens, J. C., *Phys. Rev.*, **95**, p. 21, 1954.
53. Shockley, W., *B.S.T.J.*, **28**, p. 435, 1949.
54. Shockley, W., Electrons and Holes in Semiconductors, p. 258, Van Nostrand, 1950.
55. Theuerer, H. C., U. S. Pat. No. 2542727.
56. Margeneau, H., and Murphy, G. M., The Mathematics of Physics and Chemistry, p. 213, Van Nostrand, 1943.
57. Margeneau, H., and Murphy, G. M., The Mathematics of Physics and Chemistry, p. 72, Van Nostrand, 1943.
58. Reiss, H., and La Mer, V. K., *J. Chem. Phys.*, **18**, p. 1, 1950.
59. Reiss, H., *J. Chem. Phys.*, **19**, p. 482, 1951.
60. Carslaw, H. S., and Jaeger, J. C., Conduction of Heat in Solids, p. 40, Oxford, 1948.
61. Boltzmann, L., *Ann. Phys.*, **53**, p. 959, 1894.
62. Carslaw, H. S., and Jaeger, J. C., Conduction of Heat in Solids, p. 41, Oxford, 1948.
63. Prince, M. B., *Phys. Rev.*, **92**, p. 681, 1953, **93**, p. 1204, 1954.
64. Tyler, W. W., and Woodbury, H. H., *Bull. Am. Phys. Soc.*, **30**, No. 7, p. 32, 1955.
65. Debye, P. P., Conwell, E. M., *Phys. Rev.*, **93**, p. 693, 1954.

66. Shockley, W., Electrons and Holes in Semiconductors, Chapter 8, Van Noststrand, 1950.
67. Shockley, W., Electrons and Holes in Semiconductors, Chapter 16, Van Noststrand, 1950.
68. Geballe, T. H., and Morin, F. J., Phys. Rev., **95**, p. 1085, 1954.
69. Conant, J. B., The Chemistry of Organic Compounds, p. 196, Macmillan, 1939.
70. Valenta, M., and Ramasastry, C., Symposium on Semiconductors, Meeting I.M.D., and A.I.M.E., Feb. 20, 1956.
71. Longini, R. E., and Green, R., Phys. Rev. (in press).

Single Crystals of Exceptional Perfection and Uniformity by Zone Leveling

By D. C. BENNETT and B. SAWYER

(Manuscript received January 23, 1956)

The zone-leveling process has been developed into a simple and effective tool, capable of growing large single crystals having high lattice perfection and containing an essentially uniform distribution of one or more desired impurities. Experimental work with germanium is discussed, and the possibility of broad application of the principles involved is indicated.

INTRODUCTION

The first publication describing the concept of zone melting appeared about four years ago.¹ As there defined, the term zone melting designates a class of solidification techniques, all of which involve the movement of one or more liquid zones through an elongated charge of meltable material. This simple concept has opened a whole new field of possibilities for utilizing the principles of melting and solidification.

The first zone melting technique to gain widespread usage was one for zone refining germanium by the passage of a number of liquid zones in succession through a germanium charge. This process may be quite properly compared to distillation, the essential difference being that the change in phase is from solid to liquid and back, instead of from liquid to vapor and back. The zone refining technique has been eminently successful in the purification of germanium. Harmful impurity concentrations are of the order of one part in 10^{10} . This is mainly because all the impurities whose segregation behavior in freezing germanium has been measured have segregation coefficients (see equation 1) differing from 1 by an order of magnitude or more.² During the zone refining operation, these impurities collect in the liquid zones and are swept with them to the ends of the charge, which may be later removed.

¹ Pfann, W. G., Trans. A.I.M.E., **194**, p. 747, 1952.

² Burton, J. A., Impurity centers in Germanium and Silicon, Physica, **20**, p. 845, 1954.

This paper deals with a second zone melting process, zone leveling,^{1, 3} which has gained usage somewhat more slowly than zone refining, but which has proved to be a highly effective tool for distributing desired impurities uniformly throughout a charge. For this process, only one liquid zone is used and its composition is adjusted to produce the desired impurity concentration in the material which is solidified from the liquid zone. Appropriate precautions are taken to insure the production of single crystals, if the material is desired in this form.

Since the invention of zone leveling, the process has been developed into a precision tool and as such it has become a preferred practical method for growing germanium single crystals of uniform donor or acceptor content. It is the purpose of this paper to discuss the technical development of this process, which has had two chief objectives: (1) the attainment of the greatest possible uniformity of donor and/or acceptor impurity distribution in the crystal; and (2) the attainment of a germanium crystal lattice with a minimum of imperfections of all kinds. The presentation will cover the principles involved, the means developed and results achieved toward these objectives in that order.

The first applications of the principles of zone melting have been in the field of semiconductor materials processing, chiefly because there are no other known refining techniques capable of meeting the extremely stringent purity requirements necessary for material to be used in semiconductor devices. Nevertheless, it is clear that these relatively simple and very effective zone melting techniques are beginning to find a wide variety of useful applications throughout the general fields of metallurgy and chemical engineering.

BASIC PRINCIPLES

The basic concept, theory and experimental confirmation of zone leveling have been well covered in previous publications.^{1, 3} Accordingly, the intention here is only to repeat the salient points of the theory with a special emphasis on the assumptions involved since it will be necessary to refer to them.

Fig. 1 is a schematic drawing of a zone leveling operation showing a liquid zone of constant volume containing a solute whose concentration is C_L . As the zone moves a distance Δx an increment of germanium is melted at the right end, and another is frozen at the left end. The concentration of solute in the newly frozen Δx of solid solution is C_S . The distribution coefficient k is now conveniently defined as the ratio

³ Pfann, W. G., and Olsen, K. M., Physical Review, **89**, p. 322, 1953.

of these solute concentrations:

$$k = \frac{C_s}{C_L} \quad (1)$$

When $k < 1$, the freezing interface may be regarded as a filter permitting only a fraction k of the solute concentration in the liquid to pass into the growing solid and rejecting the rest to remain in the liquid. If the unmelted charge of solvent is pure — that is, if no solute passes into the zone at the melting interface it is readily seen that the liquid zone will be gradually depleted of its solute impurity content during passage through the charge.

An expression for the solute concentration in the solid, C_s , deposited there by the passage of one zone, for the case of "starting charge into pure solvent" has been derived¹ based on the following assumptions:

- (1) The liquid volume is constant (both cross section of charge and zone length l are constant).
- (2) k is constant.
- (3) Mixing in the liquid is complete (i.e. concentration in the liquid is uniform).
- (4) Diffusion in the solid is negligible.

The expression is

$$C_s = kC_{L_0} e^{-kx/l} \quad (2)$$

where C_{L_0} is the initial concentration of impurities in the liquid, l is the zone length, and x is the distance moved by the solidifying interface. A set of C_s versus x/l curves is shown in Fig. 2 for various k 's. From this figure it is readily seen that when k is small the decay of C_s is slow (i.e., the depletion of C_L is slight).

Largely because of this consideration, most of the practical work reported in this paper has utilized solutes in germanium having low segregation coefficients.

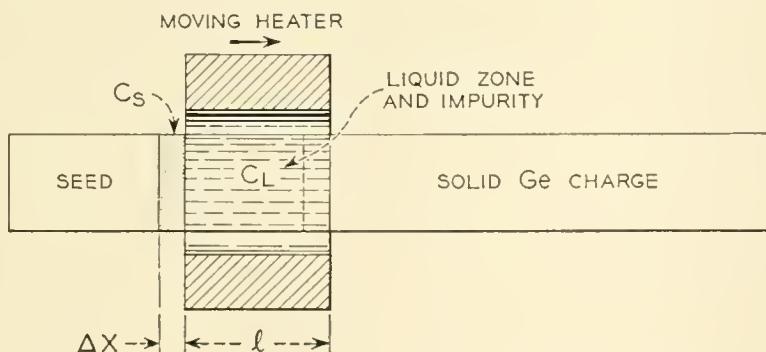


Fig. 1 — Schematic of zone leveling operation.

gation coefficients (usually antimony, whose $k = 0.003$ as donor, and indium whose $k = 0.001$ as acceptor). However, the principles of zone leveling are broad and capable of application to any solvent-solute system within the range of solubilities of its solid and liquid phases. The general method of attack¹ is first to find that composition of the liquid zone which will deposit the desired solid solution. Secondly, if one or more of the segregation coefficients involved is not small, the liquid zone must be maintained at its proper composition by admixing to the solid charge the same solutes that the zone will deposit in its product. Thus the solutes that are removed from the liquid zone at the freezing end will be replenished at the melting end.

The above mathematical treatment leads one to expect an essentially uniform solute distribution throughout a zone leveled crystal for the case under discussion in which k is small and the zone moves through a charge of pure solvent as indicated in Fig. 2. Irregular variations of C_s along the length or over the cross-section of the ingot are not predicted. The

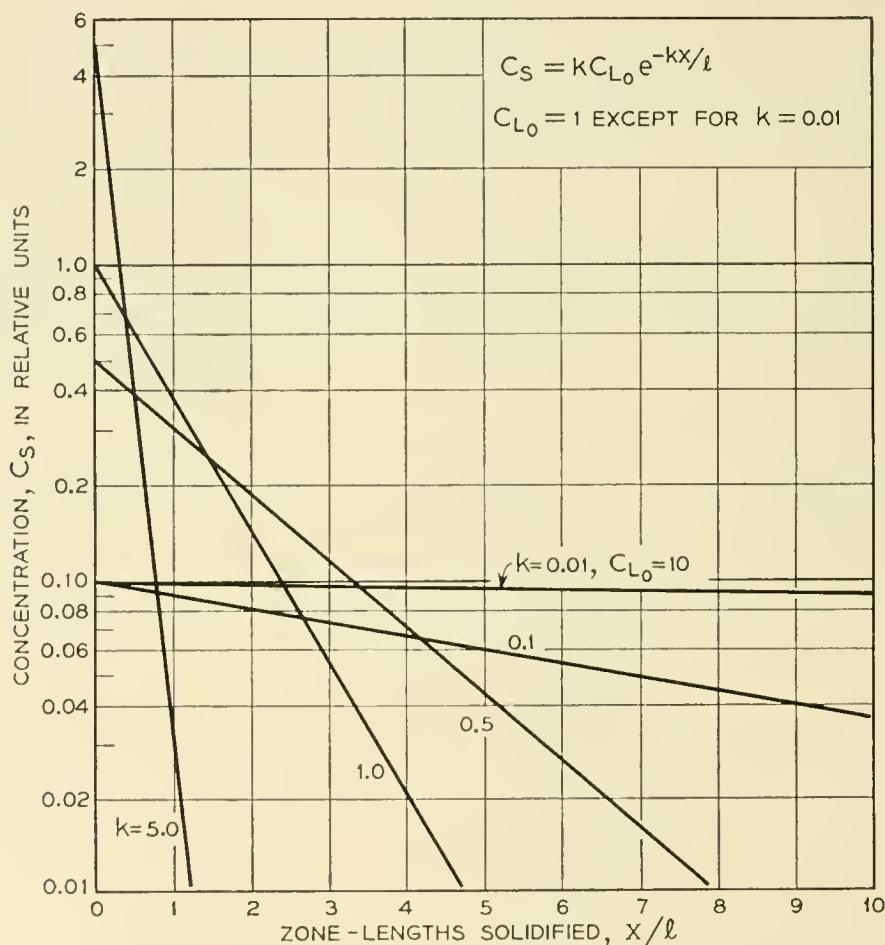


Fig. 2 — Solute concentration curves predicted for zone leveling with a starting charge of solute into pure solvent.

treatment is not concerned with lattice imperfections in the ingot such as dislocations, lineage, or grain boundaries. The predictions the theory does make have been well verified by experiment insofar as it has been possible to meet the assumptions enumerated above. However, as with most assumptions, their validity is sensitive to the experimental conditions, particularly in the cases of the first three. Much of the development effort, especially that toward improving resistivity uniformity, has been directed toward controlling the process so that these assumptions will be as nearly valid as possible.

Early experiments in zone leveling yielded crystals good enough to meet device requirements of that time. However, as semiconductor devices were designed to meet tighter design requirements, the demands on the germanium material grew more critical. Under these circumstances, it became necessary to examine the requirements on the product of the process and what precautions would be necessary to insure that its operation was under sufficient control. Accordingly, we shall discuss first the requirements on semiconductor material and then those critical aspects of the leveling operation which must be controlled to insure quality of the final product.

REQUIREMENTS ON GERMANIUM FOR SEMICONDUCTOR USES

The basic electrical bulk property of a germanium crystal is its conductivity or the reciprocal of that quantity, its resistivity. For a great majority of semiconductor uses, an extrinsic conductivity⁴ is required in addition to the $\frac{1}{50}$ ohm $^{-1}$ cm $^{-1}$ intrinsic conductivity that results at room temperature from thermal excitation of electron-hole pairs in pure germanium. An extrinsic conductivity may be either n-type or p-type. Both of these may be produced by trace impurities distributed throughout the crystal, the n-type by donor impurities and the p-type by acceptor impurities. At room temperature donors give rise to conduction electrons and the acceptors to conduction holes which are free to move within the germanium crystal. If both donors and acceptors are present in the same crystal, the resulting electrons and holes recombine, leaving essentially the extrinsic conductivity contributed by the excess of one over the other, that is by $|N_D - N_A|$.

The fundamental requirement is, then, to control the net donor and the acceptor balance, $|N_D - N_A|$, to a predetermined value throughout the crystal. For most applications, the conductivity is to be increased by one or two orders of magnitude above the 27°C intrinsic value. An idea of the donor or acceptor concentrations involved may be acquired

⁴ Shockley, W., Electrons and Holes in Semiconductors.

from noting that a conductivity of $\frac{1}{5}$ ohm $^{-1}$ cm $^{-1}$ (that is, a conductivity increased by one order of magnitude) corresponds to a $N_D - N_A$ concentration of 7 parts per billion.

The next most commonly measured bulk property of germanium is the lifetime of minority carriers,⁵ i.e., the time constant for decay by recombination of a surplus population of minority carriers artificially introduced into the crystal. Minority carriers are holes in n-type germanium or electrons in p-type germanium. This time constant may be regarded reasonably as a figure of merit for the crystal, being an indication of its freedom both from certain chemical impurities and from crystal faults, since these act as catalysts to the electron-hole recombination reaction. Normally, the highest possible lifetime is desired. Thus it becomes important to take extreme precautions during handling and processing of the germanium to avoid contamination, particularly by such known recombination center elements as nickel and copper⁶ and it is also important to avoid crystal lattice faults such as dislocations, lineage, and grain boundaries.

Another observable quantity has recently been gaining acceptance as a more definite indication of mechanical crystal perfection than the minority carrier lifetime measurement. This is the etch pit density count, ϵ , (see Fig. 3) which is observed microscopically on an oriented (111) surface of a Ge crystal that has been etched three minutes in an agitated CP-4 etch (20 parts by volume concentrated HNO₃, 12 parts concentrated HF, 12 parts concentrated acetic acid, and $\frac{1}{2}$ part Br₂). There is strong evidence⁷ that the etch pits are formed at the intersections of dislocations with the surface of the crystal. While an etch pit count probably indicates only certain edge dislocations which intersect the surface of the crystal, it is at least a relative indication of the total dislocation density, and thus appears to be a highly useful index of crystal lattice perfection.

In the last year, evidence of a strong correlation has been observed⁸ between certain electrical properties of alloy junctions, especially the breakdown voltage, and the etch pit density of the material on which the alloy junction is made. Accordingly, material to be used for alloy junction transistors is now selected on the basis of its maximum etch pit count and its freedom from lineage, twin, and grain boundaries.

The usual device test requirements on n- or p-type Ge material vary

⁵ Valdes, L. B., Proc. I.R.E., **40**, p. 1420, 1952.

⁶ Vogel, F. L., Read W. T., and Lovell, L. C., Phys. Rev., **94**, p. 1791, 1954.

⁷ Vogel, F. L., Pfann, W. G., Corey, H. E., Thomas, E. E., Physical Review, **90**, p. 489, 1953.

⁸ Zuk, P., and Westberg, R. W., private communication.



Fig. 3.—Microphotograph of Typical Etch Pits on (111) Plane.

from device to device, but may be summarized as follows:

- (1) *Composition* — The donor-acceptor balance $N_D - N_A$ must be accurately controlled so that the resistivity, ρ , of the crystal is uniform and falls within acceptable tolerance limits.
- (2) *Macro Perfection* — The crystal shall contain no grain boundaries, lineage, or twinning.
- (3) *Micro Perfection* — The etch pit density, ϵ , must be lower than a certain empirically determined maximum.

(4) *Lifetime of Minority Carriers* — τ , must usually be above a certain minimum, although in many cases this minimum may be as low as a few microseconds.

Assuming macro perfection a consideration of these requirements leads directly to the two general objectives mentioned in the introduction of this paper: composition uniformity and control, and crystal lattice perfection. A third objective, high chemical purity, might also be inferred from the lifetime requirement, but the results obtained by zone refining raw material and by fairly standard laboratory techniques of cleaning and baking of furnace parts at high temperature have been satisfactory. Hence this objective has required little development effort. We proceed to a discussion of critical aspects of zone leveling in the light of the two major development objectives.

COMPOSITION UNIFORMITY AND CONTROL

The experimental development work described in this paper has been concerned with the distribution of two trace impurities, indium and anti-

mony, in a pure element, germanium. The traces are generally desired in concentrations varying from 1 to 100 parts per billion, ($\rho = 35$ to 0.35 ohm cm). These amounts are too small to be detected by chemical or spectrographic means, but are readily detectable by electrical resistivity measurements. Although this application of zone leveling is very specific, it should be possible, as we have already suggested, to apply the experimental results to be described to more general systems. The subject of uniformity is conveniently discussed in two sections: (a) longitudinal resistivity uniformity, and (b) cross-sectional uniformity.

(a) *Longitudinal Composition Uniformity*

It has already been shown, by (2), that if the k is small, the variation in C_s over four or five zone lengths should be slight. This should be true either if a charge of pure germanium is used, or if a charge containing the same impurity present in the liquid zone is used, provided that the charge concentration of this impurity is of the same order of magnitude as that sought in the product. Where the solute has a small k , the leveling action of the zone is strong and the large C_L that is required is relatively unaffected by variations of the order of C_s .

The primary cause of observed variations in the *longitudinal* resistivity is fluctuation of the volume of the liquid zone. If this volume increases for any reason, the solute dissolved in it will be diluted. On the other hand, if the volume decreases, which can occur only when some of the liquid freezes and if k is small, most of the zone's solute will be concentrated in the smaller volume. Thus for small k 's the concentration of solute in the liquid zone, C_L , varies inversely with the zone's volume. If C_L is to be constant, the volume must be constant, i.e. assumption (1) must be valid.

Unfortunately, the zone volume is directly affected by many variables, namely temperature fluctuation and drift, fluctuation in growth rate, variation in the cross-section of the unmelted charge, variation in the inert gas flow, and even cracks in the unmelted charge. For optimum control of longitudinal resistivity uniformity, it is, therefore, necessary to control all of these variables. The remainder of this section will consider their control.

Toward minimizing the effect of temperature variation on the zone volume, it is important to consider both the means of overall temperature control and the design of the temperature field which melts the liquid zone. It is clear that variation of the temperature field as a whole will directly affect the length of the liquid zone. Accordingly, it will be important to use a precision temperature controller in order to maintain a

constant zone length. The controller used here is a servo system that cycles the power on and off about ten times a second, adjusting the on fraction of the cycle according to the demands of a control thermocouple. The sensitivity of the controller is $\pm 0.2^\circ\text{C}$ at 940°C . With a liquid zone about 4 centimeters long and a temperature gradient of about 10°C per centimeter at the solidification interface, this degree of control should introduce longitudinal resistivity variations no greater than ± 0.3 per cent.

When other requirements permit, it is possible to design a temperature contour to minimize the effects of control fluctuations. When the temperature gradients at the ends of the liquid zone are small, a slight change in the general temperature of the system will cause a relatively large change in the position of the solid-liquid interface. On the other hand, when the gradient is steep, the shift in position of the interface will be small. It is with this consideration in mind that a temperature gradient of about 130°C/cm is provided at the melting end of the liquid zone (Fig. 4). A steep gradient has the added advantage that it provides a large heat flux which is capable of supplying or removing the heat of solidification even at relatively fast leveling rates. Thus, a steep temperature gradient serves effectively to localize a solid-liquid interface. Other considerations, soon to be discussed, dictate that a small temperature gradient (about 10°C/cm) must be used at the freezing end of the zone. Accordingly, high precision of temperature control is required to properly stabilize the position of this solid-liquid interface.

Variation in the cross-section of the liquid zone may be controlled by using a boat with uniform cross-section, and by using as charge material which has been cast into a mold of controlled cross-section. Less precise control is obtained by using ingots from the zone refining process which were produced in a boat matched to the zone leveler boat. Even when care is used to maintain a uniform height of the zone refined ingot,⁹ the control is less precise than in a casting.

A constant and uniform growth rate is important toward obtaining uniform longitudinal resistivity because segregation coefficients vary with growth rate.¹⁰ This is especially true in the case of the *k* for antimony. Under steady state conditions, the growth rate is the rate at which the boat is pulled through the heater. A stiff pulling mechanism is required in order that the slow motion be steady. In the apparatus described here, a synchronous motor, operating through a gear reduction to drive a lead screw, has served to pull the boat smoothly over polished quartz rods.

⁹ Pfann, W. G., *J. Metals*, **5**, p. 1441, 1953.

¹⁰ Burton, J. A., Kolb, E. D., Slichter, W. P., Struthers, J. D., *J. Chem. Phys.*, **21**, p. 1991, Nov., 1953.

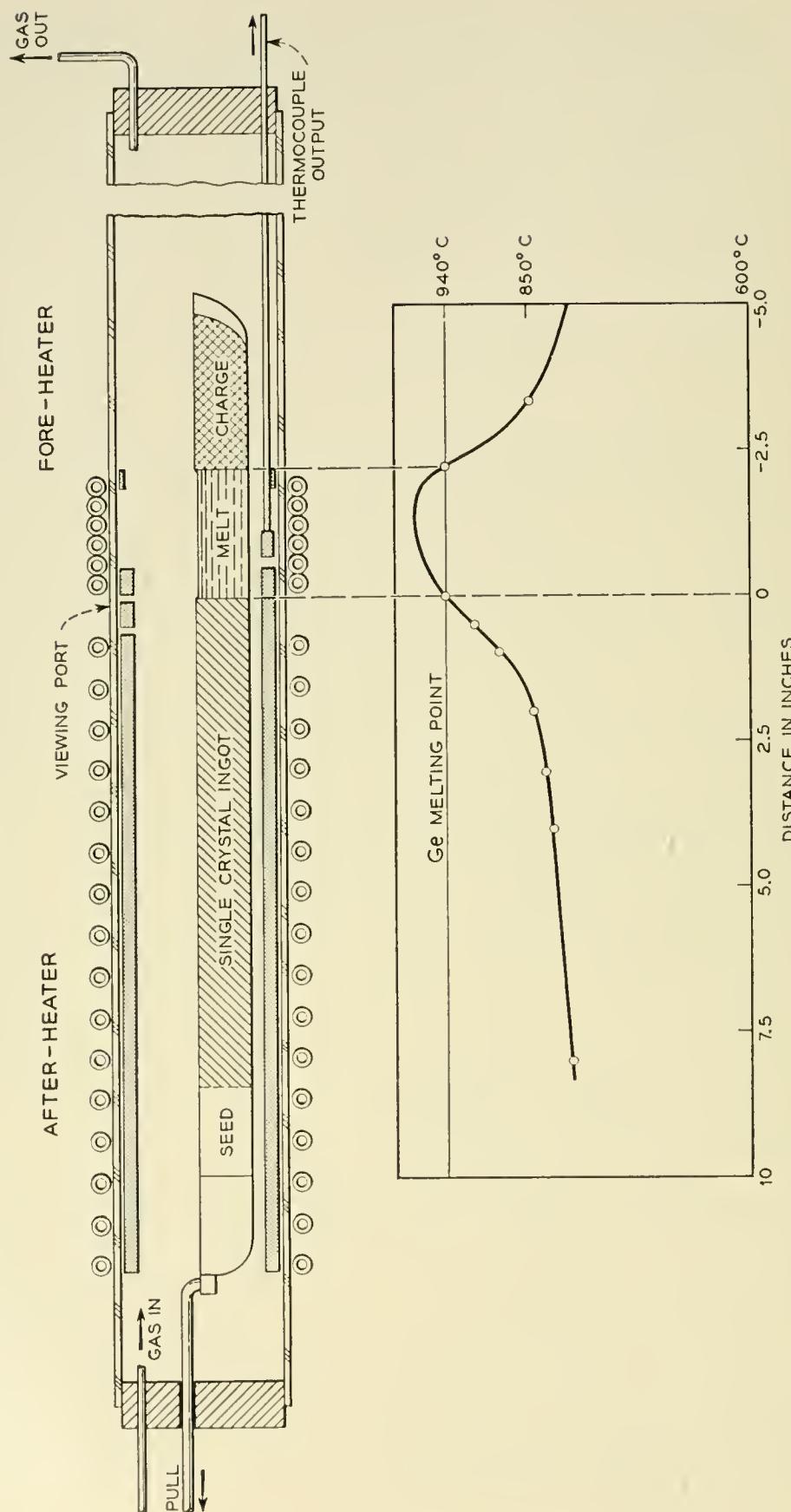


Fig. 4 — Axial temperature curve for the germanium zone leveler with after heater.

The true growth rate may be affected by factors that cause variations from steady state growth such as temperature and gas flow fluctuations. The need to control these variables has already been mentioned because of their effect on zone volume; their effect on growth rate is thus a second reason for their control.

Cracks or similar discontinuities in the unmelted charge act as barriers to heat flow. Thus they cause a local rise in temperature and lengthening of the liquid zone as the crack approaches the zone, until it is closed by melting. The resulting transient increase in liquid volume (and in ρ of the product) may be of the order of 10 per cent.

(b) *Cross-Sectional Composition Uniformity*

Difficulty may be expected in controlling the cross-sectional uniformity of the zone leveled ingot chiefly when the third assumption is invalid, i.e., when C_L throughout the liquid is non-uniform. As shown in the next paragraph, the true C_L must always rise locally near the solidifying interface due to the solute diffusion which is necessary when $k < 1$. However, it is possible to improve the validity of assumption 3 both by slowing the growth rate and by stirring the liquid zone.

One can form an estimate of a theoretically reasonable growth rate in terms of the rate of diffusion of impurities in liquid germanium. It should be noted that movement of a liquid zone containing a solute whose segregation coefficient is small implies a general movement by diffusion of essentially all the solute atoms away from the solidifying interface at a speed equal to the rate of motion of the zone. Even slow zone motion corresponds to a high diffusion flux of the solute through the liquid. As a consequence, the solute concentration must rise in front of the advancing solidification interface to a concentration $C_{L'}$ (see Fig. 5) until a concentration gradient is reached sufficient to provide a diffusion flux equal to the growth rate. Fick's Law of diffusion is useful here to calculate the extent of the rise in $C_{L'}$ at the growth interface, assuming the liquid to be at rest. The ratio of the maximum concentration to the bulk concentration may be taken from Fig. 5. If the maximum is to be no greater than 10 per cent above the mean, a maximum growth rate of 2×10^{-7} mils per second or 7×10^{-7} inches/hour would be required. Clearly, this rate is far too slow to provide an economical means of growing single crystals. For a practical process, it will be necessary to use non-equilibrium conditions at growth rates that must result in appreciable concentration differences within the liquid zone. Of course, the slower the growth rate the smaller will be the diffusion gradient and the higher will be the expected cross-sectional uniformity.

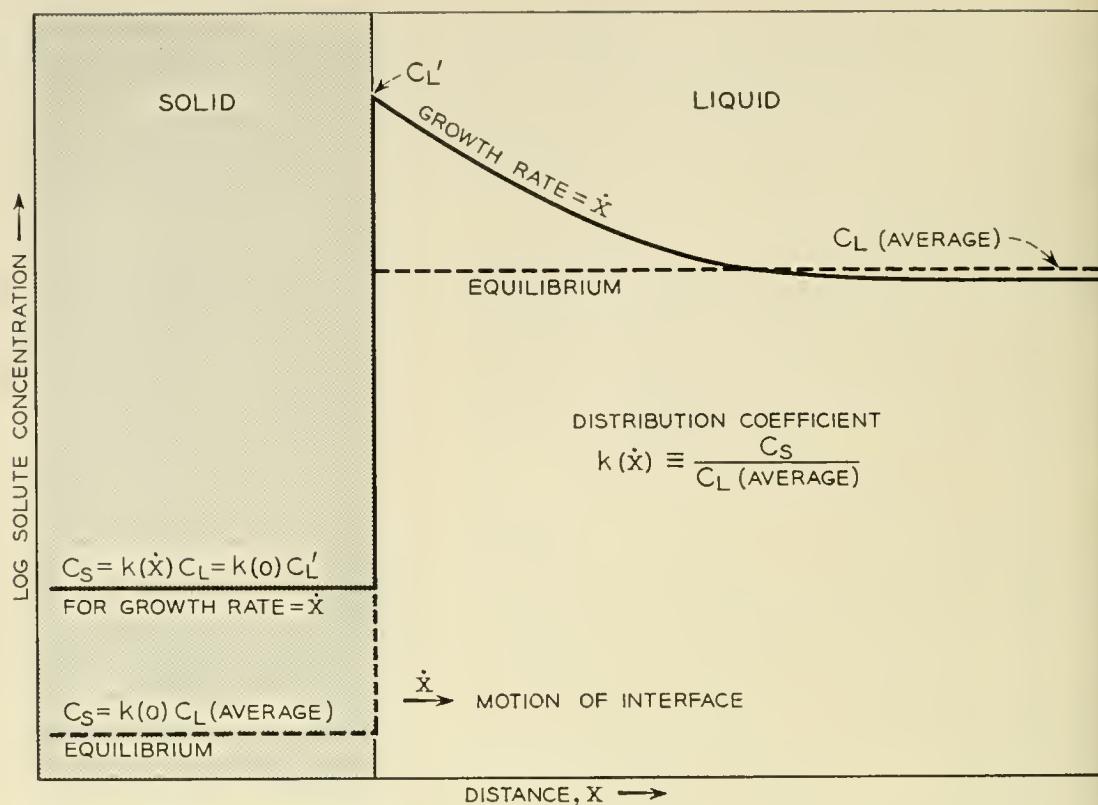


Fig. 5 — Solute concentration in solid and liquid at equilibrium and at finite growth rates.

If the liquid were static, that is, without any currents, it should be possible to obtain a uniform, controlled solute concentration in the solid even at appreciable growth rates, merely by adjusting the average concentration in the liquid to arrange that the C_L obtained at the growing interface will be the desired one. Instead of working with the equilibrium distribution coefficient k_0 , one works with an effective distribution coefficient $k(\dot{x})$ for the given growth rate, \dot{x} :

$$k(\dot{x}) = \frac{C_s}{C_L(\text{ave})} \quad (3)$$

In practice, however, the situation is complicated by the existence of convection currents in the liquid zone. It is true that these currents tend to stir the liquid zone and thereby to minimize the concentration gradient within it. However, the currents are not uniform over the growing interface and they carry liquid of varying concentrations past the interface, causing fluctuations in C_s . Since these convection currents cannot be eliminated, one turns to the alternative of using forced stirring of the liquid zone. Such a forced stirring is readily available when RF induction heating is used by allowing the RF field to couple directly with the

liquid zone.¹¹ The resulting stirring currents are shown schematically in Fig. 6. It is seen that the liquid is moved from the center of the zone along its axis toward both ends. There it passes radially outward across the interface and returns along the outside of the zone to its center. These stirring currents are faster than convection currents and tend to minimize the rise of C_L at the solidification interface and to improve the uniformity of C_L and of crystal growth conditions in general over the freezing interface.

CRYSTAL LATTICE PERFECTION

A single edge dislocation in germanium may be regarded as a line of free valence bonds. The dislocation line is believed to have about 4×10^6 potential acceptor centers per centimeter, producing a space charge in the neighboring germanium and strongly modifying its semiconductor properties.¹² A lineage boundary (a term found useful to designate a low angle grain boundary) is a set of regularly spaced dislocations, and may be regarded as a surface of p-type material. Since the basic electrical properties of a semiconductor, resistivity (and also minority carrier lifetime) are drastically out of control at dislocations and arrays of dislocations, it is easy to understand why these lattice imperfections are undesirable in crystals to be used for most semiconductor purposes.

The attainment of high perfection in germanium lattices may conveniently be discussed in two parts: first, the growth of a single crystal of high perfection and, second, the preservation of the crystal's perfection during its cooling to room temperature.

The problem of growing a single crystal in the zone leveler is basically one of arranging conditions so that the liquid germanium solidifies only

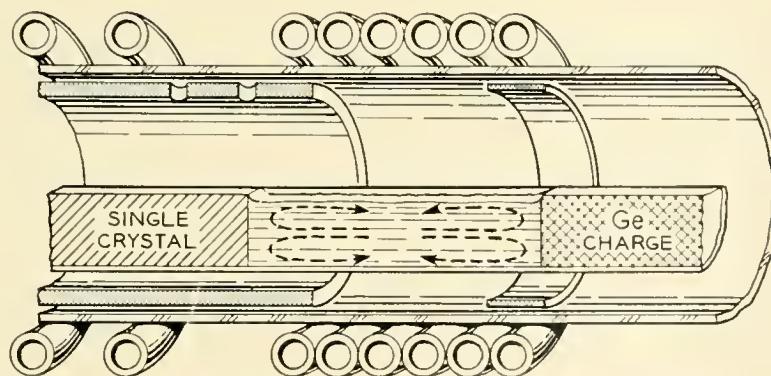


Fig. 6 — Stirring currents in liquid induced by RF induction heater.

¹¹ Brockmeir, K., Aluminium, **28**, p. 391, 1952.

¹² Read, W. T., Phil. Mag. **45**, p. 775, 1954.

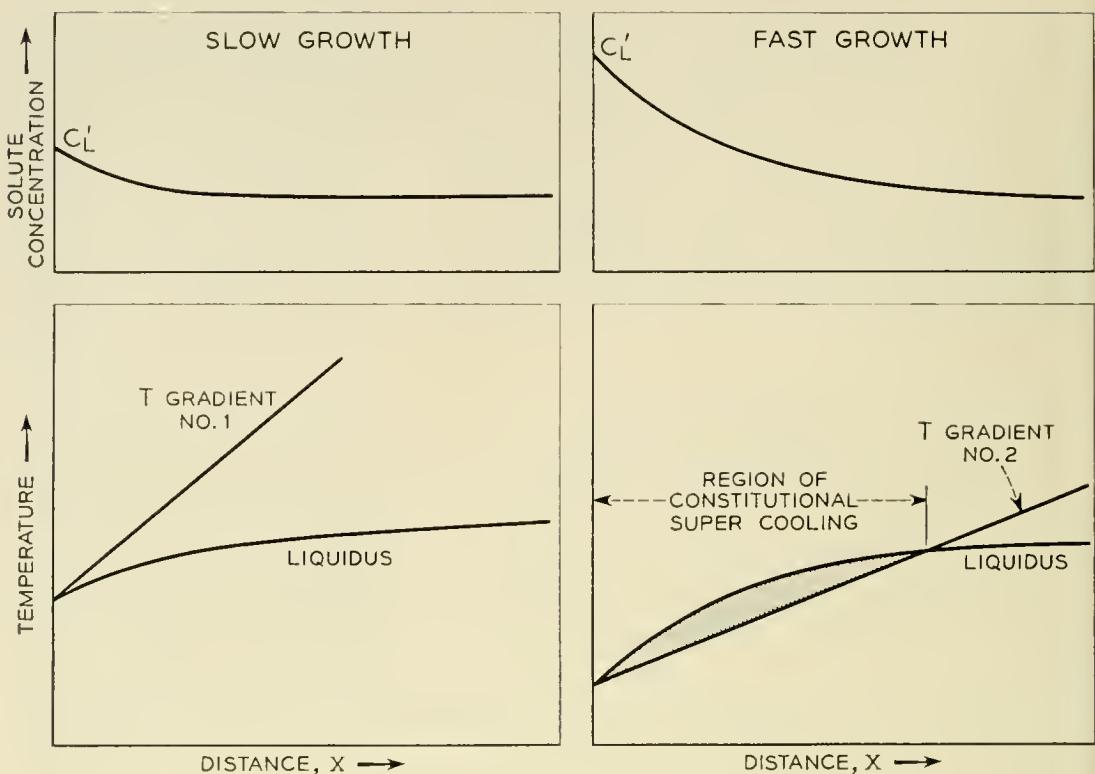


Fig. 7 — Schematic solute concentration and temperature curves in liquid, near freezing interface, illustrating constitutional supercooling. The left edge of each diagram represents the solid-liquid interface.

on the single crystal germanium seed. In order to achieve this situation, it is essential that no stable nuclei form. Thus, not only must the temperature of the liquid zone be above its freezing point everywhere except at the interface, but the liquid must also be free of foreign bodies that can act as nuclei. Furthermore, temperature fluctuations are to be avoided.

The requirement that the liquid temperature be above its freezing point necessitates a slow growth rate because of what has been termed "constitutional supercooling."¹³ This phenomenon can best be described with the aid of Fig. 7. The freezing point of a liquid is depressed by increasing concentration of solutes having k 's less than unity. Because of the rise in C_L near the solidifying interface, the freezing point is more depressed in this region than that in the bulk of the liquid zone as shown in Fig. 7.

It has also been shown¹⁴ for crystals growing in one dimension that the temperature gradient in the liquid decreases for increasing growth rates. The temperature gradients for two growth rates are plotted on Fig. 7. It can be seen that where the growth rate is slow and the temperature

¹³ Chalmers, B., J. Metals, **6**, No. 5, Section 1, May, 1954.

¹⁴ Burton, J. A., and Slichter, W. P., private communication.

gradient is steep, the temperature of the liquid is above its liquidus (freezing point curve) throughout the liquid, and no stable nuclei can form. However, increasing the growth rate decreases the temperature gradient, while it depresses the liquidus. If the temperature gradient is reduced to that indicated for fast growth, a region of constitutional supercooling will exist in front of the solidifying interface where nuclei can form and grow. The freezing of such a crystallite onto the growing crystal marks the end of single crystal growth.

A foreign body may also initiate polycrystalline growth. A natural site for nucleation by foreign bodies is the wall of the boat, close to the growth interface. Here the liquid germanium is in contact with foreign matter at temperatures approaching its freezing point. It was found by D. Dorsi that germanium single crystals could be grown satisfactorily in a smoked quartz boat, at growth rates up to 2 mils per second. However, uniformity considerations mentioned previously make it desirable to zone level at much slower rates.

It is believed that scattered dislocations may be produced in a single crystal germanium lattice by three chief mechanisms. They may be propagated from a seed into the new lattice as it grows; they may result from various possible growth faults; but probably the most important mechanism in this work is plastic deformation of the solid crystal. The first cause may be minimized by selecting the most nearly perfect seeds available, the second by using slow growth rates, and the third by minimizing stresses in the crystal.

The first hint that plastic deformation in the crystal might be an important source of dislocations came from the study of crystals pulled from the melt by the Teal-Little technique. Frequently when sections of crystals grown in the [111] direction were etched in CP_4 the pits were arrayed in a star pattern, Fig. 8(a), in which the pits appeared on lines—not randomly distributed. This coherent pattern suggested strongly that the lines were caused by dislocations in slip planes which had been active in plastic deformation of the crystal. The slip system of germanium has been determined to be the $\langle 110 \rangle$ directions on {111} planes.¹⁵ If the periphery of the crystal is assumed to be in tension, it is possible to calculate the relative shear stress pattern in each slip system of the 3 {111} planes which intersect the (111) section plane. The results of these calculations are summarized in Fig. 8(b) which shows a polar plot of the largest resolved shear stresses for these planes and also their traces in the section plane. The agreement with the observed star pattern is striking.

¹⁵ Treuting, R. G. Journal of Metals, 7, p. 1027, Sept., 1955.

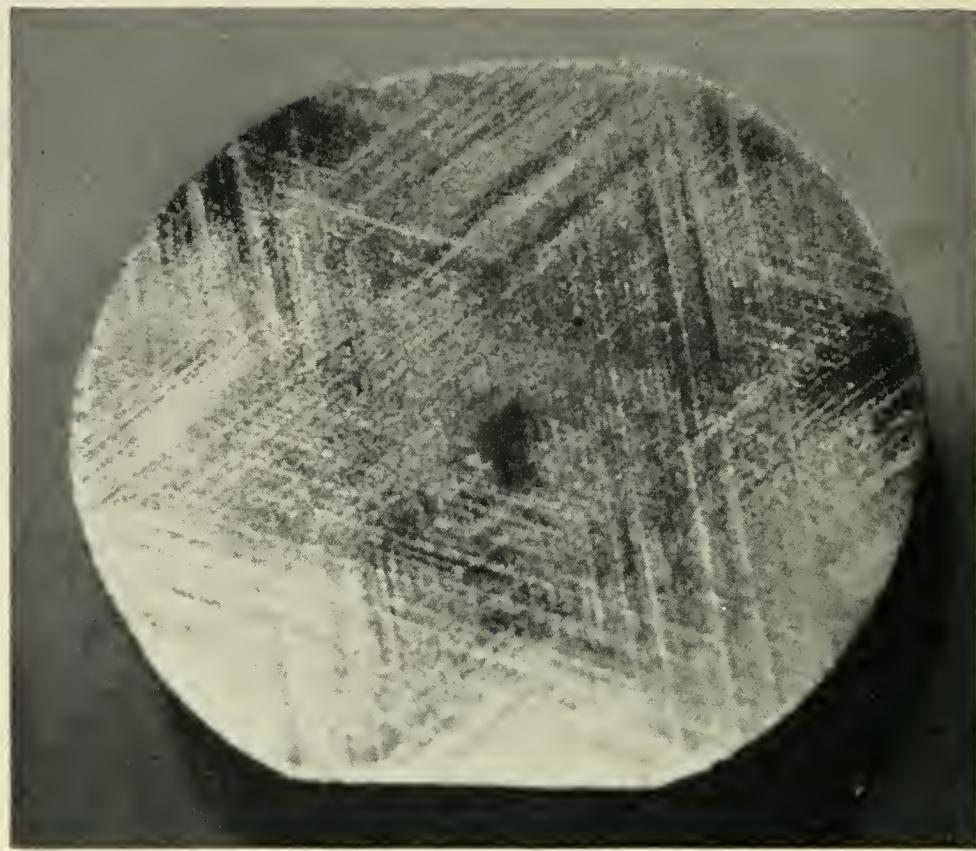


Fig. 8(a) — Star Pattern on (111) plane (etched cross-section of crystal pulled from melt).

The peripheral tension assumed in the above paragraph may be seen to be qualitatively reasonable upon consideration of the heat flow pattern of the crystal during growth. Heat must enter the crystal by conduction through its hottest surface, the growing interface, which is a 940°C isotherm. It must leave through all the other surfaces by radiation and conduction. Therefore, these surfaces must be cooler than their adjacent interiors, and cross-sections of the crystal must have cooler peripheries than cores because of the heat escaping from the peripheral surfaces. Due to thermal contraction the cooler periphery must be in tension and the core in compression.

In zone leveled crystals the distribution of etch pits on a (111) section was not dense or symmetric enough to display a star pattern. However, it was reasoned that since thermal contraction stresses appeared to play a major role in the production of dislocations in pulled crystals through plastic deformation in the available slip systems, the same mechanism might be playing a significant role in zone leveled crystals.

The only stresses in a zone leveled ingot other than those due to the weight of the crystal itself must be those due to non-uniformities in

thermal contraction. Consider a small increment of the length of a newly formed zone leveled crystal as heat flows through it from its hotter to its colder ends while the crystal moves slowly through the apparatus. Heat flows in by conduction from the higher temperature germanium adjacent to it. Heat leaves not only by conduction out the other end, but also by conduction and radiation from the ingot surface. Because of this latter heat loss, there is a radial component as well as a longitudinal component to the temperature gradient. The cooler surface contracts resulting, as above, in peripheral tension and internal compression. Clearly if the radial component of heat flow could be eliminated, there would be no peripheral contraction. Accordingly, the most desirable temperature distribution is one whose radial heat flow is zero, i.e., a case of purely axial or one dimensional heat flow, which implies a uniform temperature gradient along the axis of the ingot. In practice, it is difficult to obtain a uniform axial temperature gradient except for the special case of a very small one. This may be obtained fairly easily by the use of an ap-

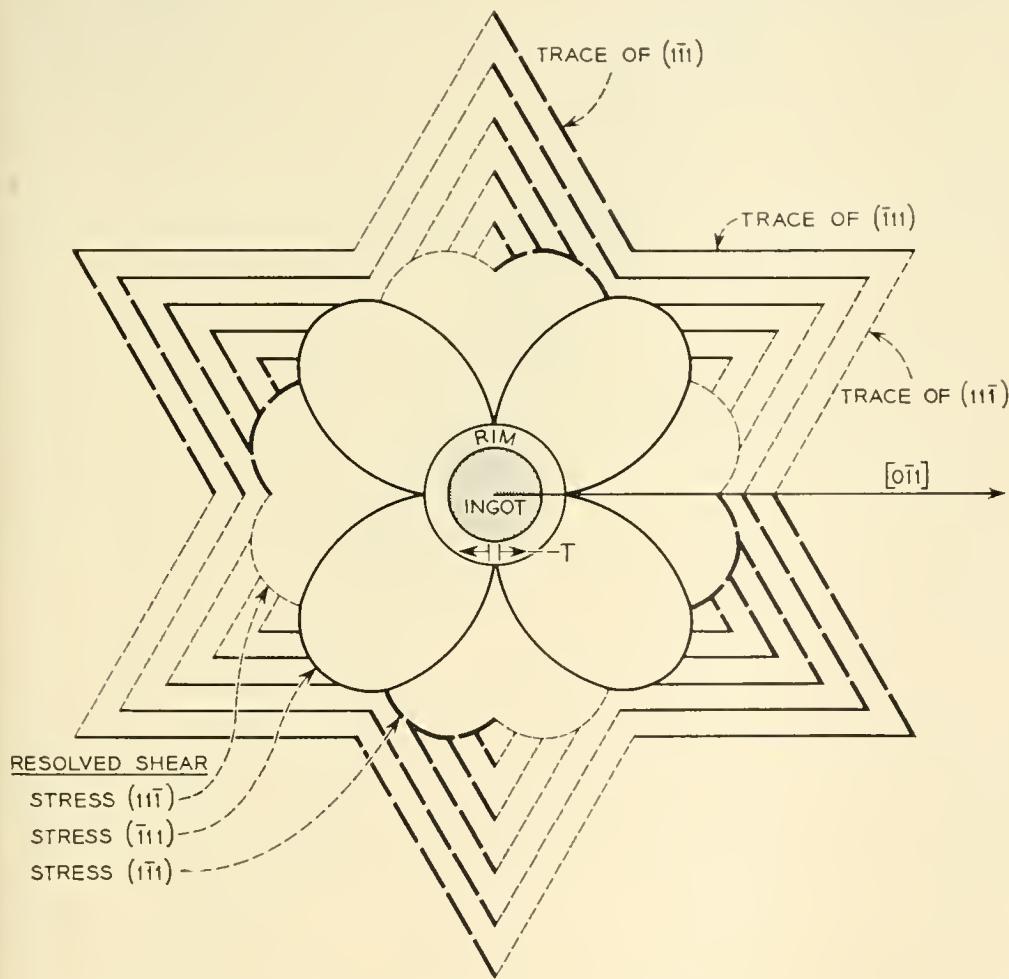
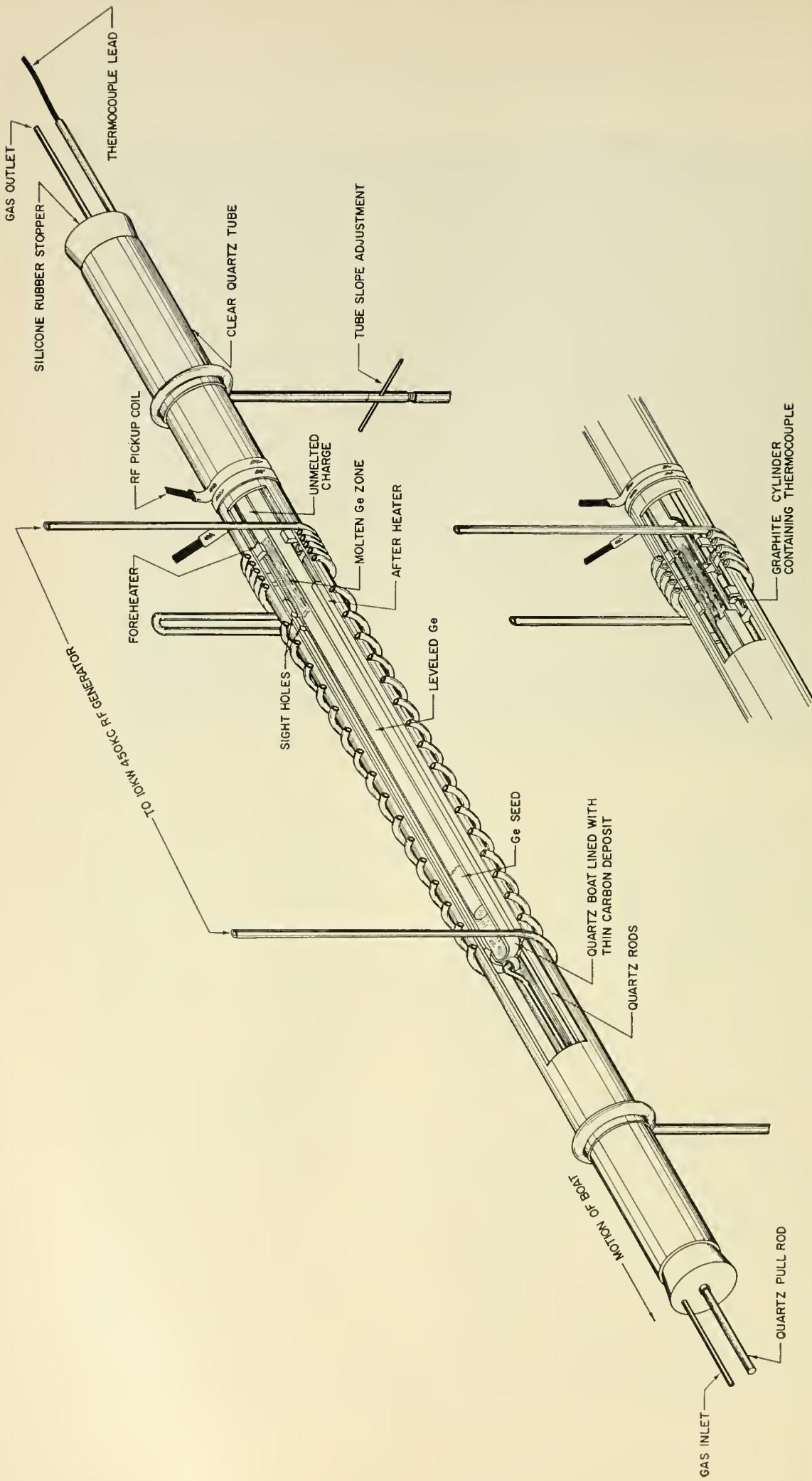


Fig. 8(b) — Resolved shear stress and slip-plane traces on (111) Plane.



R - WITHOUT AFTER HEATER

propriate heater. The heater designed for this purpose is called an after-heater and is shown in Figs. 4 and 9.

The after-heater reduces the heat loss by radiation and radial conduction from the crystal maintaining the entire crystal at a temperature only slightly below its melting point throughout its growth. After zone leveling has been completed, the entire ingot is cooled slowly and uniformly. Of course, a finite temperature gradient must exist at the liquid-solid interface. The gradient at the interface of the leveler shown in Figs. 4 and 9 is about 10°C per centimeter and the maximum gradient, about $\frac{1}{4}$ inch into the solid, is 30°C per centimeter. The gradient decreases slowly to nearly zero within the after-heater, as can be seen in the measured temperature curve of Fig. 4.

A ZONE LEVELING APPARATUS AND TECHNIQUE FOR GERMANIUM

The apparatus required for zone leveling is basically simple. A single crystal seed, the desired impurities, and a germanium charge, are held in a suitable container in an inert atmosphere. Provision is supplied for either moving a heater along the charge or the charge container through a heater. The heater may be either an electric resistance type or a radio frequency induction type. The resistance heater offers the advantage of economy while the induction heating offers the advantage of direct inductive stirring of the melted zone by the RF field, which, as mentioned previously, is helpful in attaining uniformity of impurity distribution, and is therefore to be preferred for critical work.

Schematic drawings of an RF powered zone leveler following in general the original design by K. M. Olsen are shown in Fig. 9 in two useful configurations. The outer clear quartz tube serves to support the inner members of the apparatus and also to contain the inert atmosphere for which nitrogen, hydrogen, helium, or argon, can serve. For this apparatus, a quartz boat is used to contain the germanium, since it permits inductive stirring of the liquid germanium by the RF field. The auxiliary fore and after heaters, which are made of graphite, have special purposes discussed in the two preceding sections. A typical boat used in this apparatus is about 16" long, is smoked on the inside, and is made of thin-walled clear quartz of 1" I.D. and of semi-circular cross-section. A normal charge of zone refined Ge and seed is about 12 inches long and weighs about 500 gm. A photograph of the assembled apparatus appears in Fig. 10.

For the best results in crystal perfection and resistivity uniformity, the apparatus is run with the full length after-heater and at a slow pull

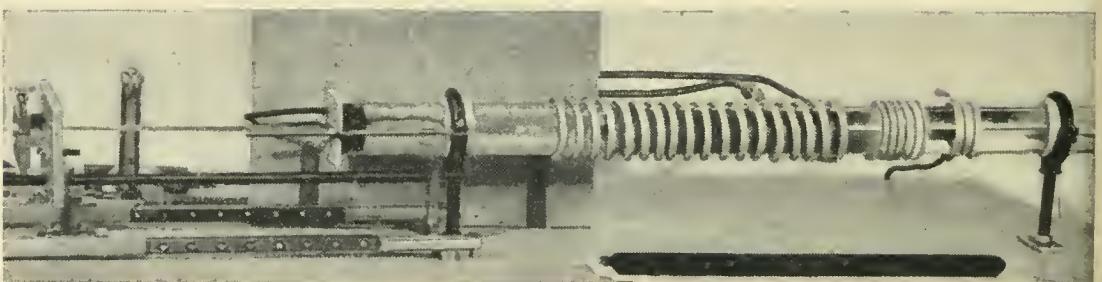
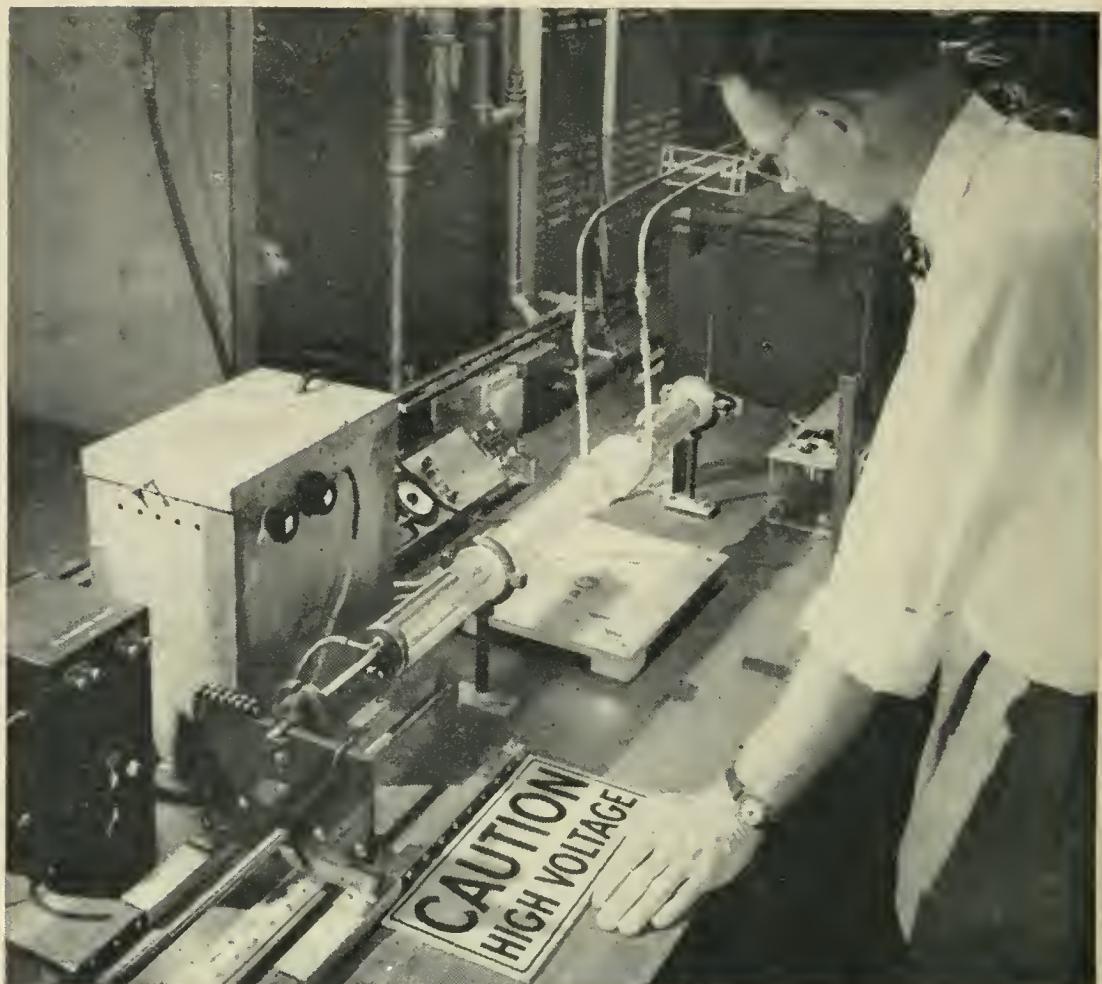


Fig. 10 — Zone leveler.

rate, 0.09 mils per second (approximately 1" in three hours). For somewhat less critical demands a pull rate 10 times faster is used, with a shortened after-heater or none at all.

If it is desired to reproduce a resistivity obtained in the zone leveler, it is very convenient to reuse the solidified zone containing the impurity addition that yielded the desired resistivity. This solid zone, if undamaged (when cut from the finished ingot), will contain all of the solute that was not deposited during the ingot run. When it is remelted next to a seed the solute will redissolve into the liquid to yield very nearly

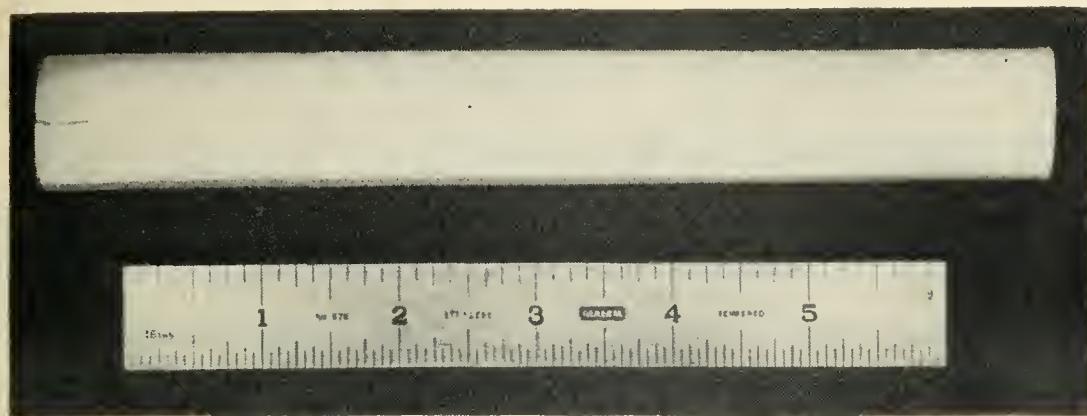


Fig. 11. — Photograph of zone leveled single crystal ingot.

the same C_L , provided that the zone volume is accurately reproduced. In this way it is readily possible to resume leveling as before and hence virtually to reproduce a desired resistivity. For the small k solutes, *In* and *Sb*, discussed in this paper the loss of C_L in one leveling run is so small as to be insignificant compared to other sources of error in this quantity.

PILOT PRODUCTION RESULTS

The capabilities of the zone leveling equipment and techniques just described may be evaluated with reasonably good accuracy on the basis of the measurement results obtained from more than 300 single crystal ingots so produced. Over 200 of these crystals were grown in the after-heater at the "slow" growth rate of 0.09 mils per second. The rest were grown with a short after-heater or none at all at a growth rate about ten times greater.

The ingots to be measured (see Fig. 11) were usually 4–6 inches long after removing seeds and solidified zones (i.e., 2–3 zone lengths), and were cut into 1 inch lengths. The ρ , τ , and ϵ measurements were taken on the flat ends of these segments. The results of the observations will be summarized and discussed in terms of the four device test requirements described earlier.

(1) Compositional Uniformity

The resistivity measurements were taken with a calibrated 4-point probe technique¹⁶ at five locations on each ingot cross-section (center, top, bottom and each side). The spacing between adjacent points of the probe was 50 mils. Accordingly, these measurements would be insensitive to ρ fluctuations in the material of this order or smaller. However, an investigation by potential probing techniques, of Ge filaments cut from zone leveled ingots¹⁷ indicates that ρ fluctuations in zone leveled material are

¹⁶ L. B. Valdes, Proc. I.R.E., 42, p. 420, 1954.

¹⁷ Erhart, D. L., private communication.

TABLE I—AVERAGE RESISTIVITY VARIATIONS
(A) Along length axis. Grand Length Average $\pm 10\%$.

Growth Rate Mils per Second	n-Type		p-Type		Average $\pm \%$
	$\pm \%$	No. of Ingots	$\pm \%$	No. of Ingots	
0.9	9.9	27	10.9	33	10.4
0.8	7.6	12	17.4	16	13.2
0.09	9.0	108	9.3	137	9.2

(B) Over Cross-Section

Growth Rate Mils per Second	n-Type		p-Type		Average $\pm \%$
	$\pm \%$	No. of Ingots	$\pm \%$	No. of Ingots	
0.9	9.5	22	8.5	30	8.9
0.8	8.3	12	6.9	14	7.5
0.09	4.3	93	2.3	122	3.2

generally coarse — changing over distances 2 to 5 times larger in dimension than the 50 mil dimension in question. Thus the ρ data summarized here should give a reasonably valid representation of the true ρ variations in the ingots measured.

Table I summarizes the resistivity variations recorded as percentages of the mean resistivity of each ingot. These variations are separated into those observed (a) along the length axis and (b) over the cross-section, for the different growth conditions and resistivity types.

It is readily seen that the average variation along the length, about ± 10 per cent, is larger than the average cross-sectional variation. The variations are not systematic along the length of the ingot and are chiefly due to fluctuation in the length of the liquid zone. An appreciable part of this variation is due to the effect, mentioned earlier, of discontinuities in the unmelted charges between 1 inch lengths of crystals that were being leveled. A smaller length variation of ρ , about ± 7 per cent, was observed in those ingots grown from continuous charges.

Part B of the table shows that the variation of ρ over the cross-section is sensitive to the growth rate in the range covered. For slow growth, it is small, and one would reasonably expect that if further improvement in ρ variation were required, it should first be sought by improving the control of the zone length.

(2) Macro Perfection

Macro perfection of the pilot production product is extremely high. There were essentially no cases of polycrystallinity, or twinning, except

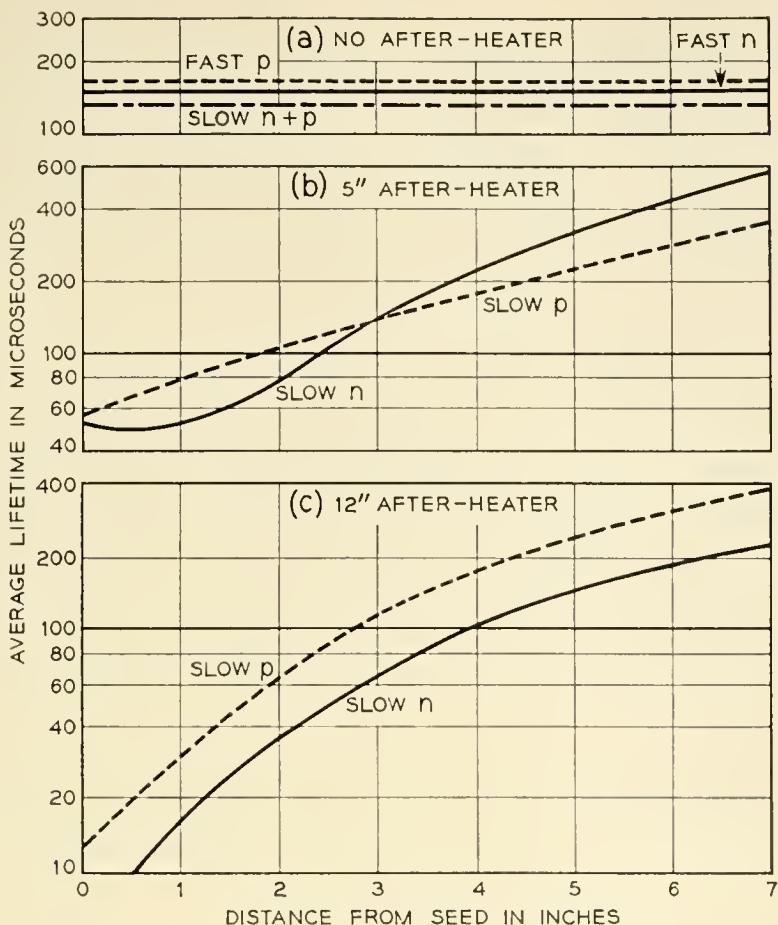


Fig. 12 — Average minority carrier lifetime plotted against distance from seed for 2-8 ohm cm crystals grown with 12", 5" and no after-heaters.

for clearly attributable causes such as power or equipment failure. There were few cases of lineage in the short after-heater and virtually none in the full after-heater, while lineage is not uncommon in ingots grown with no after-heater.

(3) Micro Perfection

Table II summarizes the etch pit density, ϵ , measurement results. In general, it can be seen that with the after-heater one can expect etch pit counts of the order of 1,500 pits per cm^2 which is lower than results without an after-heater by about an order of magnitude (and lower than

TABLE II — AVERAGE ETCH PIT DENSITIES, ϵ

	Growth Rate Mils per Second	ϵ Ave	σ	No. of Ingots
(12" after-heater)	0.09	1560	770	39
(5" after-heater)	0.09	3800	1600	3
	0.9	7000	1900	3
No after-heater	0.9	11000	6600	6

ϵ 's of pulled Ge crystals by about two orders of magnitude). The lowest average count that has been observed is 40 pits per cm^2 . This crystal was found to have the smallest X-Ray rocking-curve widths observed in germanium at Bell Telephone Laboratories — very nearly the theoretically ideal widths. The perfection indicated is exceptional — comparable to that of selected quartz crystals.

(4) *Lifetime of Minority Carriers*

τ data are summarized in Fig. 12 in which are plotted averages of the τ measurements on the ingot sections against distance from the seed. One sees a systematic rise in τ along the length axis of an ingot grown slowly in the after-heater. This is interpreted to indicate that the ingot is being slowly contaminated with chemical recombination centers during its long wait inside the after-heater at high temperatures. If improvement were needed in lifetime, it should be sought first by increasing the chemical cleanliness precautions, which were nonetheless strict in this work.

SUMMARY

A zone leveler has been developed to provide growth conditions suitable for the production of quality germanium single crystals. The crystals are nearly uniform and have exceptionally high lattice perfection. Similar levelers are in use in production.

The apparatus developed has been used to supply germanium single crystals for experiments and for the pilot production of a variety of point contact, alloy, and diffusion transistors. The machine operating at slow growth rate with an after-heater can produce one 6-inch 250-gm crystal per day. For less critical demands, it can produce several longer crystals per day.

Evaluation of the product indicates that resistivity variation on a cross-section of the ingot can be ± 3 per cent and that along the length axis it can be controlled to ± 7 per cent if a continuous charge is used. Furthermore, the crystals contain no grain boundaries or lineage and the scattered etch pit densities average about 1,500 per cm^2 . Thus, the zone leveling process has proved to be simple, efficient, and capable of more than meeting the present specifications for quality germanium single crystals.

ACKNOWLEDGMENTS

The authors are indebted for the help and cooperation of many people, especially that of L. P. Adda and D. L. Erhart who guided the evaluation of zone leveled material summarized above, and that of F. W. Bergwall through whose patient effort and suggestions the machine worked.

Diffused p-n Junction Silicon Rectifiers

By M. B. PRINCE

(Manuscript received December 12, 1955)

Diffused p-n junction silicon rectifiers incorporating the feature of conductivity modulation are being developed. These rectifiers are made by the diffusion of impurities into thin wafers of high-resistivity silicon. Three development models with attractive electrical characteristics are described which have current ratings from 0 to 100 amperes with inverse peak voltages greater than 200 volts. These devices are attractive from an engineering standpoint since their behavior is predictable, one process permits the fabrication of an entire class of rectifiers, and large enough elements can be processed so that power dissipation is limited only by the packaging and mounting of the unit.

1.0 INTRODUCTION

1.1 The earliest solid state power rectifier, the copper oxide rectifier, was introduced in the 1920's. It found some applications where efficiency, space, and weight requirements were not important. In 1940 the selenium rectifier was introduced commercially and overcame to a great extent the limitations of the copper oxide rectifier. As a result, the selenium rectifier has found wide usage. In early 1952 a large area germanium¹ junction diode was announced which showed further improvements in efficiency, size, and weight. In addition it shows promise of greater reliability and life as compared to the earlier devices. However, all of these devices have one drawback in that they cannot operate in ambient temperatures greater than about 100°C.

Also in 1952, the silicon alloy² junction diode was announced and was shown to be capable of operating at temperatures over 200°C. However it was a small area device and could not handle the large power that the other devices could rectify. During the past three years development has been carried on by several laboratories in improving the size and power capabilities of these alloy diodes. In early 1954 the gaseous diffu-

¹ Hall, R. N., Proc. I.R.E., 40, p. 1512, 1952.

² Pearson, G. L., and Sawyer, B., Proc. I.R.E., 40, p. 1348, 1952.

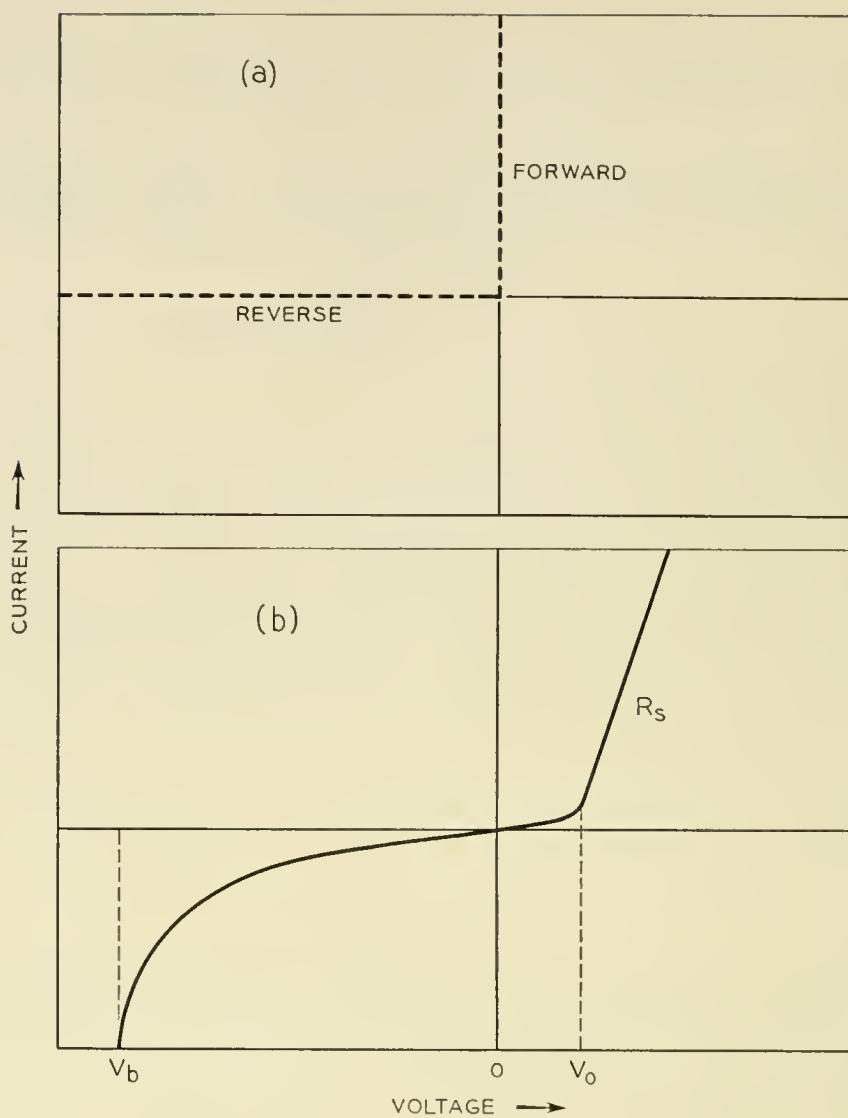


Fig. 1 — (a). Ideal rectifier. (b). Semiconductor rectifier. .

sion technique³ for producing large area junctions in silicon was announced. This technique lends itself very readily to controlling the position of junctions in silicon. An early rectifier³ made by this technique was one half cm^2 in area and conducted 8 amperes at one volt in the forward direction and about 2 milliamperes at 80 volts in the reverse direction. The series resistance of this device was approximately 0.07 ohms.

1.2 In order to understand quantitatively the problems associated with power rectifier development, consider Fig. 1(a) which shows what an engineer would like in the way of an ideal rectifier. It will pass a large amount of current in the forward direction without any voltage

³ Pearson, G. L., and Fuller, C. S., Proc. I.R.E., 42, No. 4., 1954.

drop and will pass no current for any applied voltage in the reverse direction. At present no device with this characteristic exists. A typical semiconductor rectifier has a characteristic of the type shown in Fig. 1(b). In these devices there is a forward voltage, V_0 , that must be developed before appreciable current will flow and a series resistance, R_s , thru which the current will flow. In the reverse biased direction there is a current that will flow due to body and surface leakage and that usually increases with reverse voltage. At some given reverse voltage, V_B , the device will break down and conduct appreciable currents. To have an efficient rectifier, V_0 and R_s should be as small as possible and V_B should be as large as can be made; also, the reverse leakage currents should be kept to a minimum. According to semiconductor theory, V_0 depends mainly upon the energy gap of the semiconductor, increasing with increasing energy gap. R_s consists of two parts; body resistance of the semiconductor and resistance due to the contacts to the semiconductor. The higher the resistivity of the semiconductor, the higher is the body resistance part of R_s . The leakage currents in the reverse direction depend to some extent on the energy gap of the semiconductor, being smaller with larger energy gap; and V_B depends most strongly on the resistivity of the semiconductor, being larger for higher resistivity material. Another factor that is important in the choice of the semiconductor is the ability of devices fabricated from the semiconductor to operate at high temperatures; high temperature operation of devices improves with larger energy gap semiconductors. Thus there are two compromises to be made in choosing the material (energy gap) and resistivity of the semiconductor.

1.3 This paper reports on a special class of rectifiers in which improved performance has been obtained. These devices are made by using the diffusion technique with silicon. The diffusion process permits both accurate geometric control and low resistance ohmic contacts, which in turn makes it possible to reduce R_s to very small values independent of the resistivity of the initial silicon. Therefore, high resistivity material can be used to obtain high V_B . An explanation of this result is given in Section 3. Silicon permits small reverse currents and high temperature operation. Its only drawback is that $V_0 \approx 0.6$ volts. Rectifiers made of silicon with the diffusion technique are able to pass hundreds of amperes per square centimeter continuously in the forward direction in areas up to 0.4 square centimeter. One type of device whose area is 0.06 cm^2 readily conducts ten amperes with less than one volt forward drop. The forward current voltage characteristic of this family of rectifiers follows an almost exponential characteristic indicating that

R_s is extremely small (<0.05 ohms). Although the measured reverse currents are greater than those predicted by theory for temperatures up to 100°C , the reverse losses are low and do not affect the efficiency appreciably.

1.4 The diodes made by the diffusion of silicon are very attractive from an engineering standpoint for several reasons. First of all, their behavior is predictable from the theory of semiconductor devices, as are junction transistors. This makes it possible to design rectifiers of given electrical, thermal, and mechanical characteristics. Secondly, rectifier elements of many sizes are available from the same diffused wafers making it possible to use the same diffusion process, material, and equipment for a range of devices. Thirdly, large enough elements can be processed so that the power dissipation in the unit is limited only by the thermal impedance of mount and package.

2.0 DIFFUSION PROCESS

2.1 It will be shown in 3.2 that the forward characteristic of these devices is practically independent of the type (n or p) and resistivity of the starting material. The reverse breakdown voltage of a silicon p-n junction depends primarily on the resistivity of the lightly doped region. With these two considerations in mind; that is, to fabricate rectifiers having the desirable excellent forward characteristic and at the same time high reverse breakdown voltage, high resistivity silicon is used as the starting material for the diffused barrier silicon rectifiers. Single crystal material has been found to give a better reverse characteristic than multicrystalline material. Also, it has been found that p-type material has yielded units with a better reverse characteristic than n-type material. Therefore, in the remainder of this paper, we will limit discussion to rectifiers made from high resistivity, single crystalline, p-type silicon. We will designate this material as π type silicon.

2.2 In addition to the fine control one has in the diffusion process (see 2.4), the process lends itself admirably to the semiconductor rectifier field in as much as the distribution of impurities in this process results in a gradual transition from a degenerate semiconductor at the surface of the material to a non-degenerate semiconductor a short distance below the surface. This condition permits low resistance ohmic metallic contacts to be made to the surfaces of the diffused silicon.

In order to create a p-n junction in the π silicon, it is necessary to diffuse donor impurities into one side of the slice. Although several donor type impurities have been diffused into silicon, all the devices discussed

in this paper were fabricated by using phosphorus as the donor impurity. In order to make the extremely low resistance contact to the π side of the junction that is desirable in rectifiers, acceptor impurities are diffused into the opposite side of the π silicon slice. Boron was selected from the several possible acceptor type impurities to use for the fabrication of these devices. A configuration of the diffused slice is shown in Figure 2.

2.3 It will be shown in Section 3 that there are limits to the thicknesses of the three regions, $N+$, π , $P+$, due to the nature of the operation of these rectifiers. With present techniques, it is necessary to keep

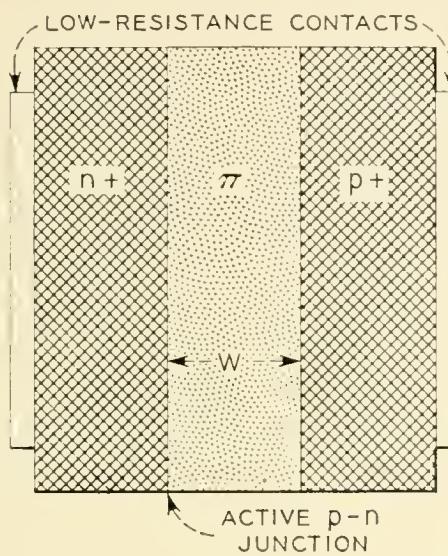


FIG. 2 — Diffused silicon rectifier configuration.

the thickness of the π region to the order of two or three mils (thousandths of an inch).

2.4 In the diffusion process of introducing impurities in silicon for the purpose of creating junctions or ohmic contacts, the diffusant is deposited on the silicon and serves as an infinite source. The resulting concentration of the diffusant is given by

$$C = C_0 \left[1 - \frac{2}{\sqrt{\pi}} \int_0^{x/\sqrt{4Dt}} e^{-y^2} dy \right] \quad (1)$$

$$= C_0 \operatorname{erfc} y$$

where C = concentration at distance x below surface

C_0 = concentration at surface

D = diffusion constant for impurity at temperature of diffusion

t = total time of diffusion

$$y = \frac{x}{\sqrt{4Dt}} = \text{variable of integration}$$

A plot of $C/C_0 = \text{erfc } y$ versus y is given in Fig. 3. C_0 is the surface solubility density and depends upon the temperature of the diffusion process.⁴ At some depth, x_j , the concentration C equals the original impurity concentration where the silicon will change conductivity type resulting in a junction. In order to obtain desirable depths of the diffused layers, $N+$ and $P+$, it is necessary to diffuse at temperatures in the range of 1000°C to 1300°C for periods of hours. With such periods it is obvious that the diffusion process lends itself to easy control and reproducibility.

3.0 CONDUCTIVITY MODULATION

3.1 It is well known that the series resistance of a power rectifier is the most important electrical parameter to control and should be made as small as possible for several reasons. The series resistance consists essentially of two parts; the body resistance of the semiconductor and the contact resistance to the semiconductor. In the early stages of rectifier development both parts of the series resistance contributed about equally to the total series resistance. However, methods were soon found to reduce the contact resistance. It then became apparent that in order to reduce the body resistance, the geometry would have to be changed and the resistivity chosen carefully. By going to larger, thinner wafers it was possible to reduce this body resistance. However, the cost of pure silicon made it important that conductivity modulation (described below) be incorporated in these devices as a method for reducing the body resistance. Our initial attempts were successful due to the fact that higher lifetime of minority carriers could be maintained in the extremely thin wafers that were used as compared to the lifetime remaining after the diffusion process in thicker wafers.

3.2 A complete mathematical description of the I-V characteristic for the conductivity modulated rectifier is practically impossible due to the fact that the equations are transcendental. However, it is easy to understand the operation of the device physically.

When the device is biased in the forward direction, electrons from the heavily doped $N+$ region are injected into the high resistivity π region. If the lifetime for these electrons in the π region is long enough, the electrons will diffuse across the π region and reach the $P+$ region

⁴ Fuller, C. S., and Ditzenberger, J. A., *J. Appl. Phys.*, **25**, p. 1439, 1954.

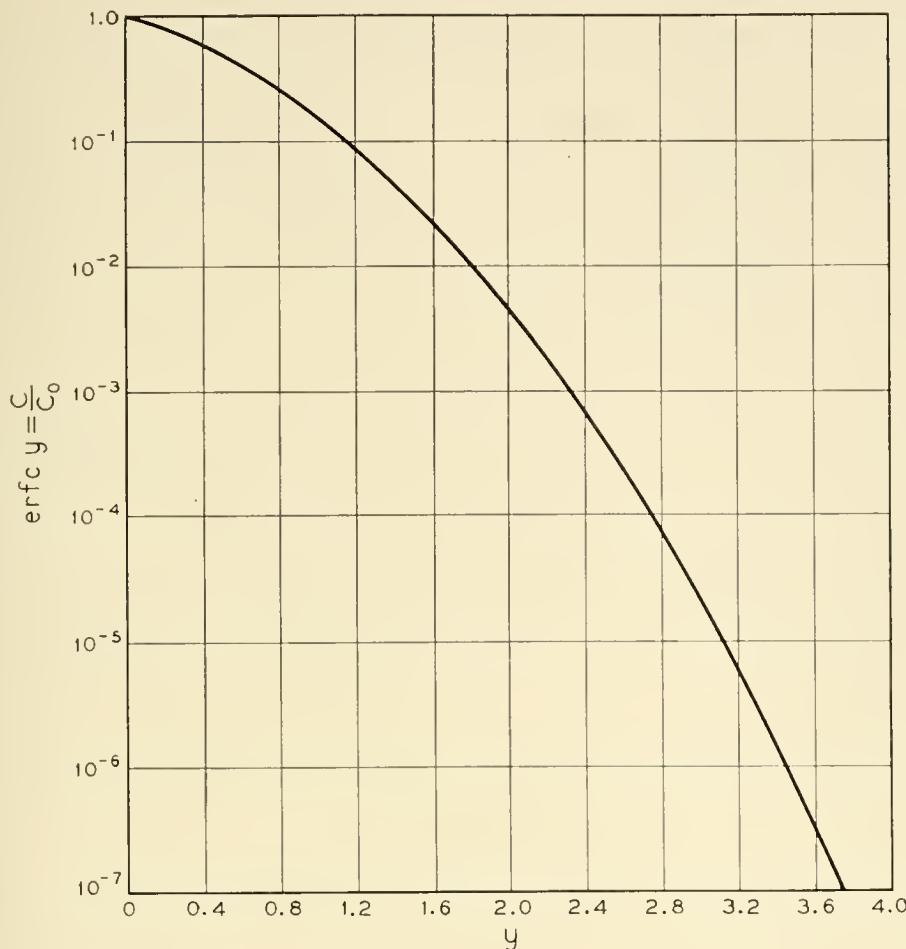


Fig. 3. — Error function complement.

with little recombination. To maintain electrical neutrality, holes are injected into the π region from the $P+$ region. These extra mobile carriers (both electrons and holes) reduce the effective resistance of the π layer and thus decrease the voltage drop across this layer. The higher the current density, the higher is the injected mobile carrier densities and therefore, the lower is the effective resistance. It is for this reason that the process is termed conductivity modulation. This effect tends to make the voltage drop across the π region almost independent of the current, resistivity, and semiconductor type.

When the junction is biased in the reverse direction, a normal reverse characteristic with an avalanche breakdown is expected and observed.

3.3 The forward characteristic of a typical unit is plotted semi-logarithmically in Fig. 4. The best fit to the low current data can be

expressed as

$$I = I_0 e^{qV/NkT} \quad (2)$$

where I = current thru unit

I_0 = constant

q = charge of electron

V = voltage across unit

k = Boltzmann's constant

T = absolute temperature

and $1 < N < 2$.

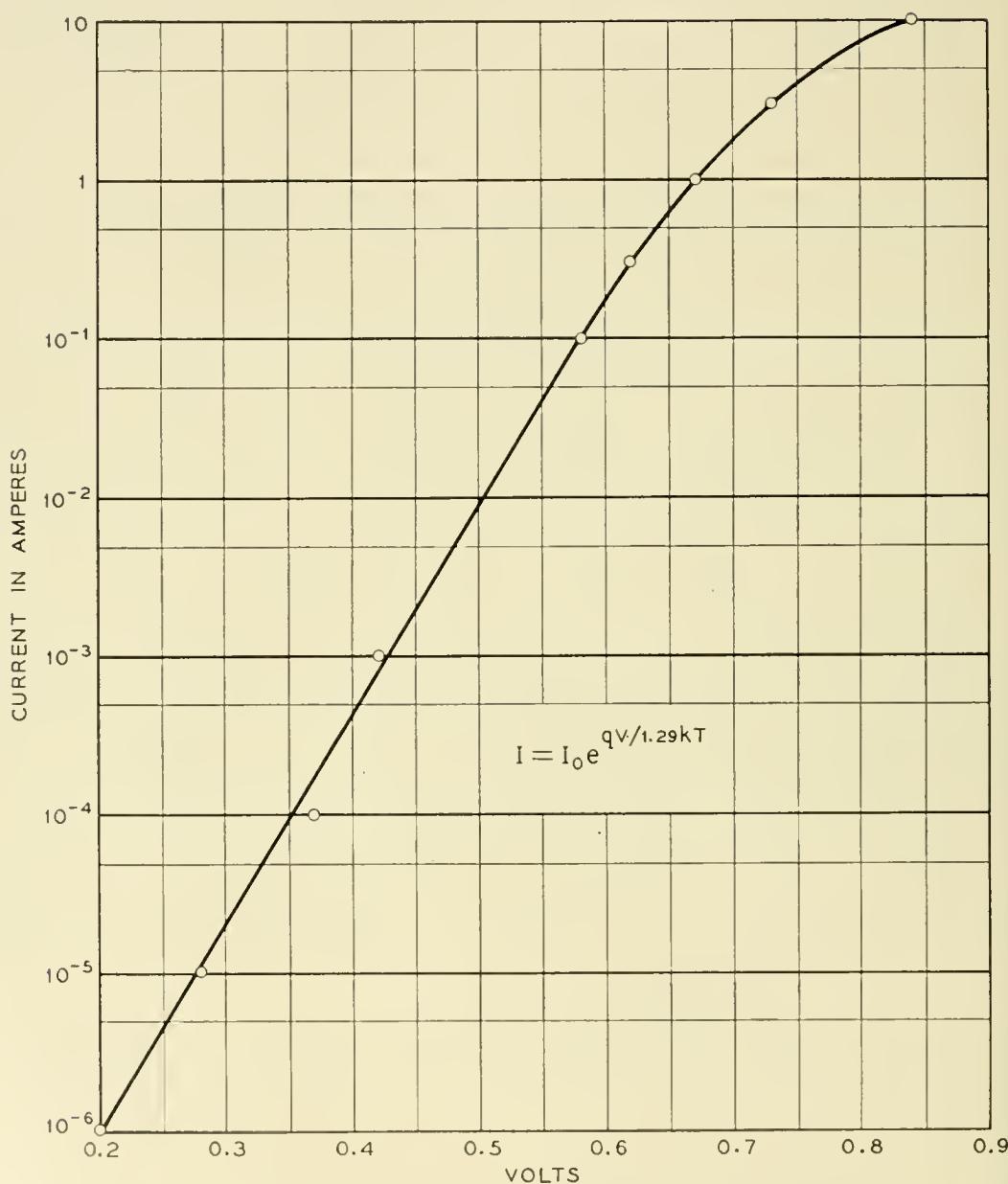


Fig. 4 — Forward characteristic of silicon power rectifier.

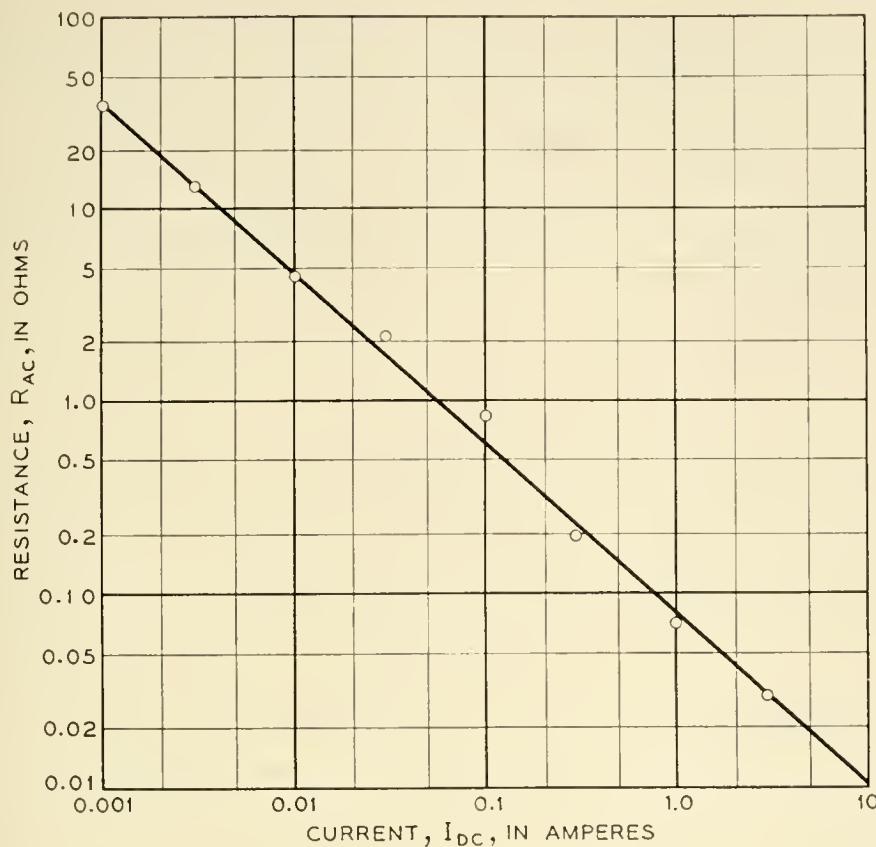


Fig. 5 — Small signal resistance versus dc forward current.

The departure of the high current data from the exponential characteristic is due to the contact resistance. Another interesting measurement of the forward characteristic is given in Fig. 5 where the small signal ac resistance is plotted as a function of the forward dc current for a typical rectifier element. The departure from the simple rectifier theory⁵ where $N = 1$ is not surprising inasmuch as *p-n* junctions made by various methods and of different materials almost always have $N > 1$. Several calculations have been carried out using different assumptions and all indicate that the forward characteristic is independent of the type and resistivity of the middle region as long as the diffusion length for minority carriers is the order of or larger than the thickness of the region.

3.4 In order to go to higher reverse breakdown voltages (> 500 volts) it is necessary to use still higher resistivity starting material. It might be expected that intrinsic silicon will be used for the highest reverse breakdown voltages when it becomes available. However, in this case

⁵ Shockley, W., B.S.T.J., 28, p. 435, 1949.

thick wafers are necessary since the reverse biased junction space charge region extends rapidly with voltage for almost intrinsic material, and high lifetime is necessary in order to get the conductivity modulation effect in these thick wafers. Therefore at present it is necessary to compromise the highest reverse breakdown voltages with the lowest forward voltage drops, in a similar manner to that discussed in Section 1. However this is now done at a different order of magnitude of voltage and current density.

4.0 FABRICATION OF MODELS

4.1 It has been pointed out in Section 1.2 that a low series resistance, R_s , is desirable and that it is composed of two parts; the body resistance and the contact resistance. In Section 3 a method for reducing the body resistance was described. The contact resistance can also be made very low. It has been found to be very difficult to solder low temperature solders (M.P. up to 325°C) to silicon with any of the standard commercial fluxes. However, it is quite easy to plate various metals to a surface of silicon from an electroplating bath or by an electro-less process⁶ to which leads can readily be soldered. Some metals used for plating contacts are rhodium, gold, copper, and nickel. This type of contact yields a low contact resistance. Another technique that has shown some promise for making the necessary extremely low resistance contact is the hydride fluxing method.⁷

4.2 A wafer which may be about one inch in diameter is ready to be diced after it is prepared for a soldering operation. Up to this point all the material may undergo the same processing. Now it is necessary to decide how the prepared material is to be used; whether low current (~ 1 amp) devices or medium or high current ($\sim 10\text{--}50$ amps) devices are desired. The common treatment of all material for the entire class of rectifiers is one reason these devices are highly attractive from a manufacturing point of view.

The dicing process may be one of several techniques; mechanical cutting with a saw, breaking along preferred directions, etching along given paths with chemical or electrical means after suitable masking methods, etc. In the case of mechanical damage to the exposed junctions, the dice should be etched to remove the damaged material. The dice are cleaned by rinses in suitable solvents and are then ready for

⁶ Brenner, A., and Riddell, Grace E. J., Proc. American Electroplaters' Society, 33, p. 16, 1946, 34, p. 156, 1947.

⁷ Sullivan, M. V., Hydrides as Alloying Agents on Silicon, Semiconductor Symposium of the Electrochemical Society, May 2-5, 1955.

assembly into the mechanical package designed for a given current rating.

4.3 The dice may be tested electrically before assembly by using pressure contacts to either side. Pressure contacts have been considered for packaging the units; however, this type of contact was dropped from development due to mechanical, chemical, and electrical instabilities.

4.4 The drawbacks of the pressure contact make it important to find a solder contact that does not have the same objections. The solder used should have a melting point above 300°C, be soft to allow for different coefficients of expansion of the silicon and the copper connections, wet the plated metal, and finally, be chemically inactive even at the high temperature operation of the device. These requirements are met with many solders in a package that is hermetically sealed. This combination of a solder and a hermetically sealed package has been adopted for the intermediate development of the diffused silicon power rectifiers.

5.0 ELECTRICAL PERFORMANCE CHARACTERISTICS

5.1 Before describing the electrical properties of these diodes, let us consider some of the physical properties of a few members of the class.

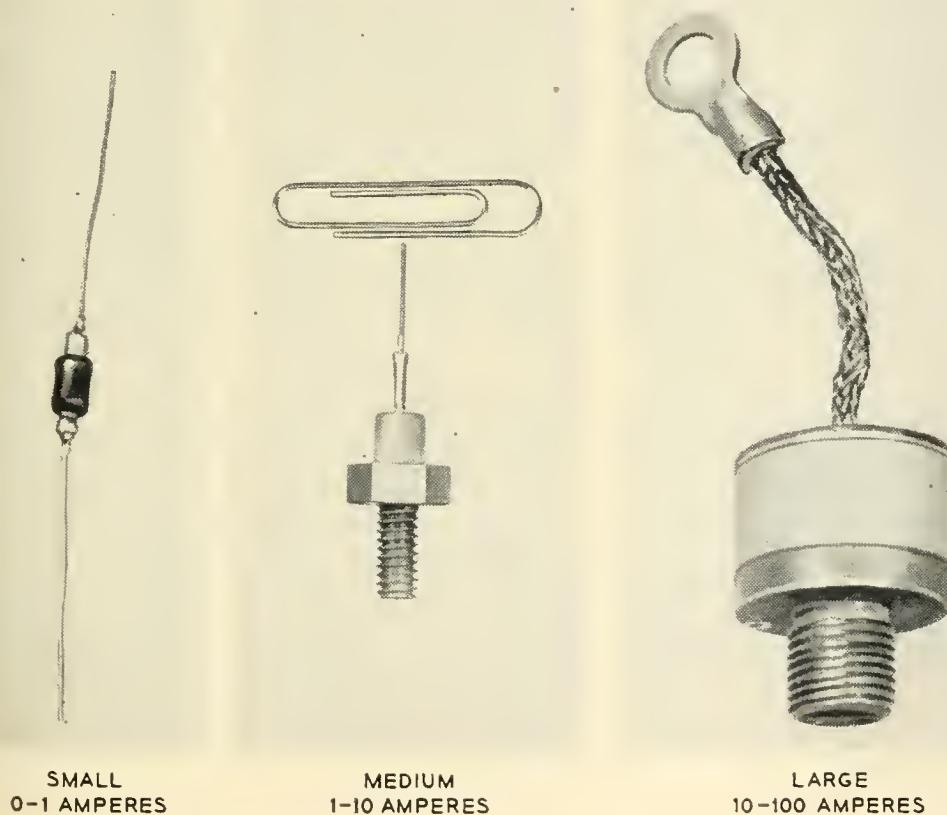


Fig. 6 — Development silicon rectifiers.

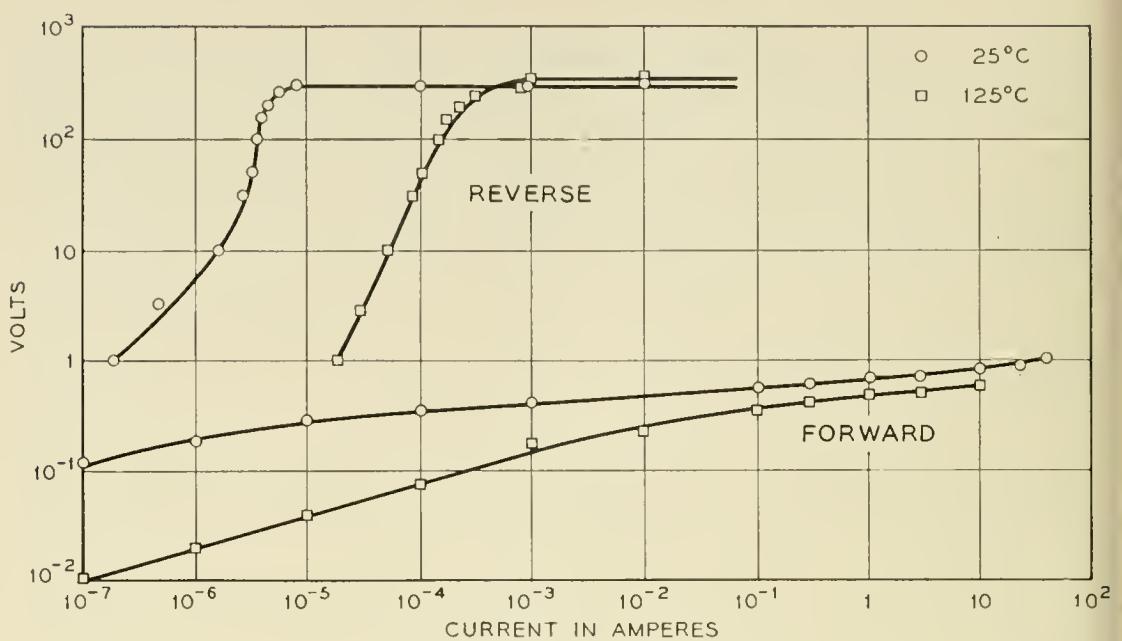


Fig. 7 — I-V characteristic of medium size rectifier.

Fig. 6 shows a picture of three sizes of units that will be discussed in this section together with the range of currents that these units can conduct. The actual current rating will depend upon the ability of the device to dispose of the heat dissipated in the unit. A description of how the rating is reached is given in Section 6.

The smallest device has a silicon die that is 0.030" by 0.030" in area

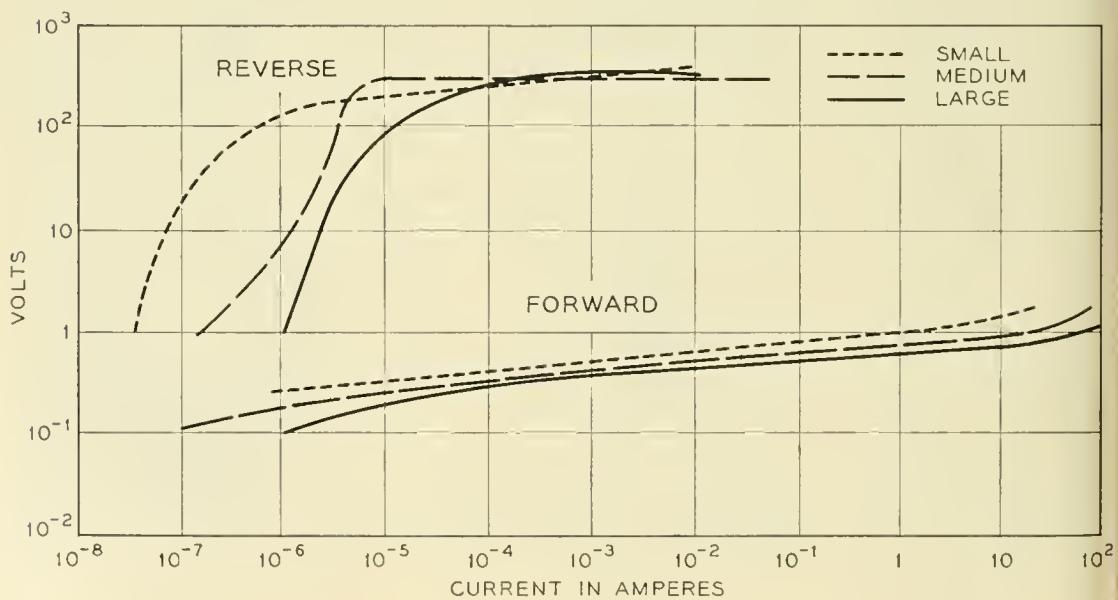


Fig. 8 — I-V characteristics of development rectifiers.

and all the units have dice about 0.005" thick. The medium size device has a wafer 0.100" by 0.100" in area. The largest device has a element 0.250" by 0.250" in area. It is obvious that a range of die size could have been chosen for any of these rectifiers. However, electrical and thermal considerations have dictated minimum sizes and economic considerations have suggested maximum sizes. The actual sizes are intermediate in value and appear to be satisfactory for the given ratings.

5.2 Of fundamental importance to users of these rectifiers are the forward and reverse current — voltage characteristics. These characteristics of the medium size unit are shown in Fig. 7 for two temperatures, 25°C and 125°C, using logarithmic scales. It can be seen that in the forward direction at room temperature, 25°C, more than 20 amperes are conducted with a one volt drop in the rectifier. At the higher temperature more current will be conducted for a given voltage drop. In the reverse direction, this particular unit can withstand inverse voltages as high as 300 volts before conducting appreciable currents (>1 ma) even at 125°C. A comparison of the current-voltage characteristics for the three different size units is shown in Fig. 8 where again the information is plotted on logarithmic scales. This information was obtained at 25°C. One can observe that the reverse leakage current varies directly as the area of the device and the forward voltage drop varies inversely as the area. These relations are to be expected; however, the reverse characteristics indicate that surface effects are probably effecting the exact shape of the curves. The changes in the forward characteristics can be attributed to the contacts and the internal leads of the packages. The breakdown voltage can be adjusted in any size device by the proper choice of starting material and therefore no significance should be placed on the different breakdown voltages in Fig. 8.



SILICON

GERMANIUM

SELENIUM

Fig. 9 — Semiconductor rectifiers of different materials.

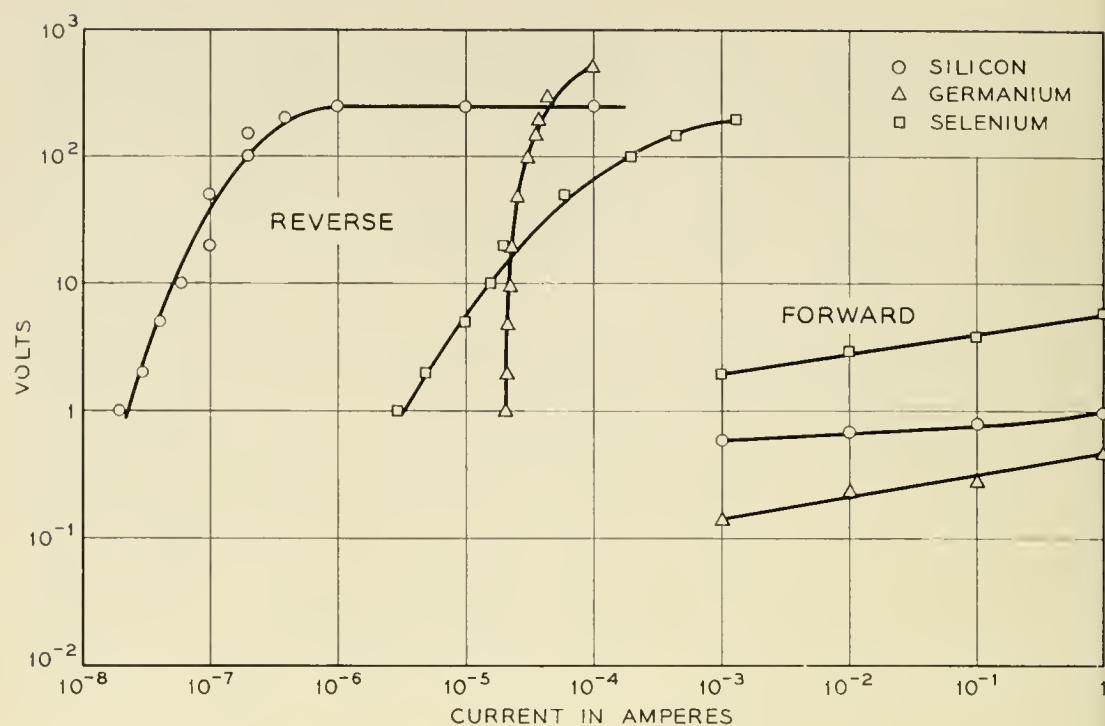


Fig. 10 — Rectifier characteristics at 25°C.

It is quite interesting to compare these units with germanium and selenium rectifiers that are commercially available. To make the comparison as realistic as one can, we have chosen to compare the smallest silicon unit with a commercially available germanium unit and a six element selenium rectifier stack rated at 100 milliamperes. The comparative size of these units can be seen in Fig. 9. Curves of the forward and reverse characteristics at 25°C are given in Fig. 10. Similar curves taken at 80°C are given in Fig. 11 and at 125°C in Fig. 12. It can be seen that the forward characteristic is best for the germanium device at all temperatures and that the reverse currents are least for the silicon rectifier. The selenium rectifier is a poor third in the forward direction. However, if one has to operate the device at 125°C, only the silicon device will be satisfactory in both the forward and reverse directions.

5.3 Capacitance measurements of all the silicon units have been made at different reverse voltages and temperatures. The temperature dependence is negligible. However, as expected in semiconductor rectifiers, the capacitance varies inversely with the voltage according to the relation $VC^N = \text{constant}$ where $2 < N < 3$. Measurements are given in Fig. 13 for a group of medium size units. The other units made from the same resistivity material have capacitances that vary directly as their areas.

5.4 The reverse breakdown voltage, V_B , of these devices is controlled by the choice of resistivity of the starting material and the depth of diffusion of the junction. By keeping the resistivity of the initial p-type silicon above 20 ohm-cm., it is possible to keep V_B above 200 volts. Units have been made with V_B greater than 1,000 volts. The deeper diffusion causes the junction to be more "graded"⁵ and therefore require a greater voltage for the breakdown characteristic. This is in line with the capacitance measurements where the exponent indicates that the junction is neither a purely abrupt junction which would result in an exponent of two nor a constant gradient junction which would result in an exponent of three.

5.5 Another interesting measurement, which is related to the lifetime of minority carriers in the high-resistivity region and the frequency response, is the recovery time of these devices. During a forward bias on a p-n junction, excess minority carriers are injected into either region. When the applied voltage polarity is reversed, these excess minority carriers flow out of these regions, giving rise initially to a large reverse current until the excess carriers are removed. The magnitude and time variation of this current will depend to some extent upon the level of the forward current but mostly upon the circuit resistance. If one adjusts the circuit resistance such that the maximum initial current in

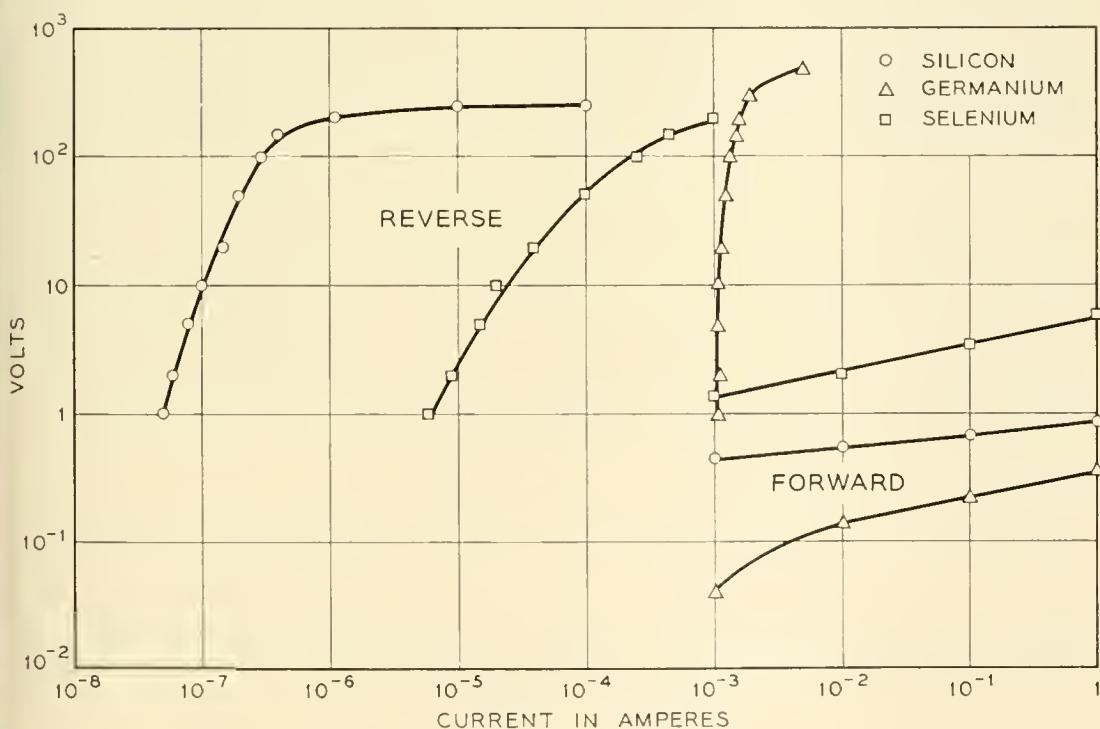


Fig. 11 — Rectifier characteristics at 80°C.

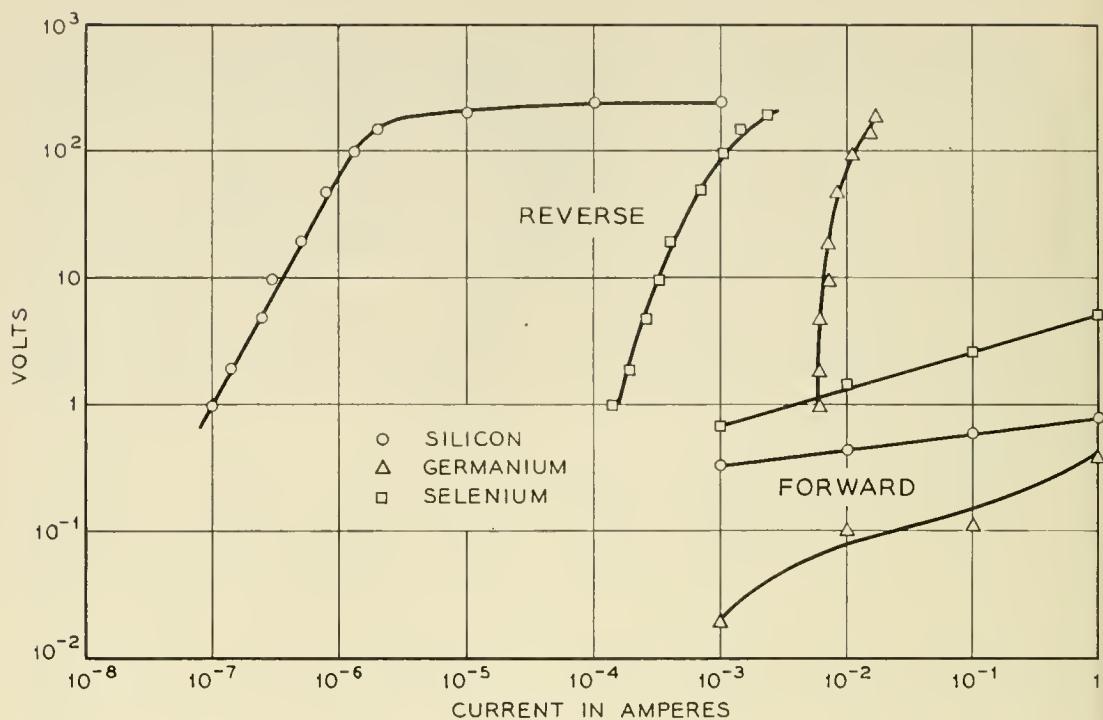


Fig. 12 — Rectifier characteristics at 125°C.

the reverse direction is equal to the forward current before reversing the polarity of the junction, then the reverse current will have a constant magnitude, limited by the circuit resistance, for a time known as the recovery time before it decays to a small steady-state value. Fig. 14 shows graphically this effect. The recovery time in diffused junctions is found to be in the range of less than 0.1 microsecond to more than 4

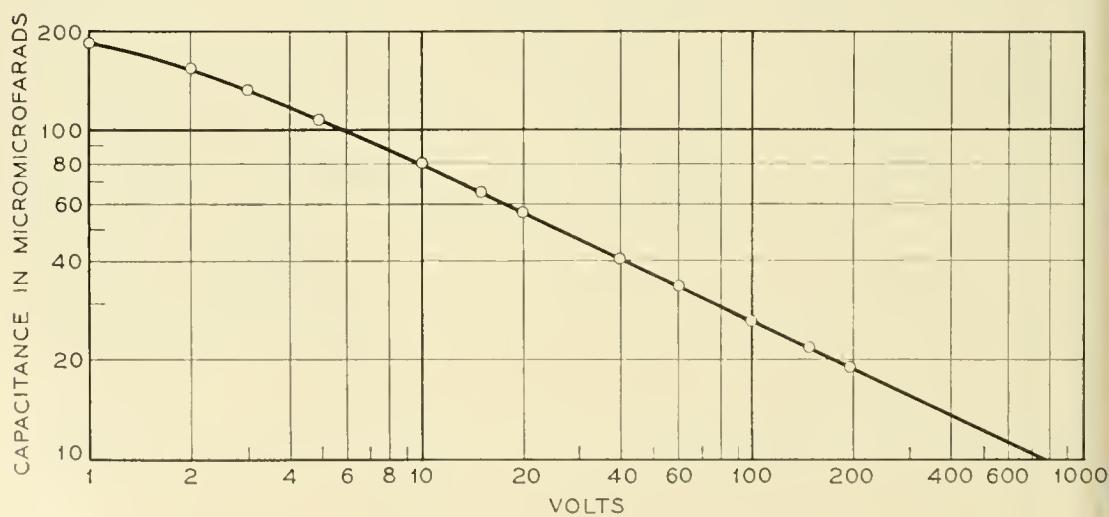


Fig. 13 — Capacitance versus reverse voltage in medium size rectifier.

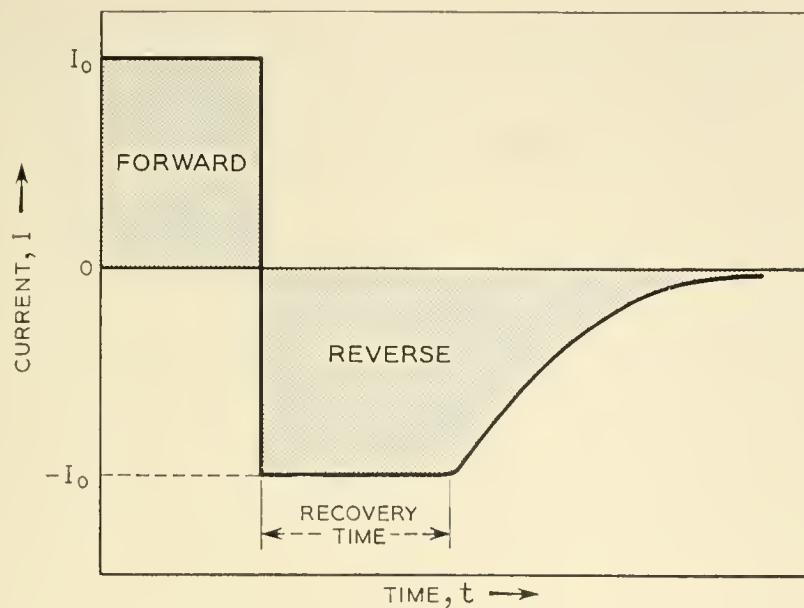


Fig. 14 — Recovery effect in silicon rectifiers.

microseconds. It can be shown that the longer recovery times are associated with higher lifetimes of minority carriers. More interesting, however, is the fact that these devices will have their excellent rectification characteristics to frequencies near the reciprocal of the recovery time. Measurements have been made of the rectification ability of typical small and medium size units by using the circuit shown in Fig. 15. The results of normalized rectified current versus frequency are given in Fig. 16 and it is seen that these units could be used to rectify power up to 1 kc/sec without any appreciable loss of efficiency.

5.6 It is interesting to note that many of the electrical measurements made with the diffused barrier silicon rectifiers are self-consistent and can be related to simple concepts of semiconductor theory. As an example, experimental measurements indicating variations of recovery time of units are related to variations in minority carrier lifetime which in turn are related to experimental variations in the forward characteristic

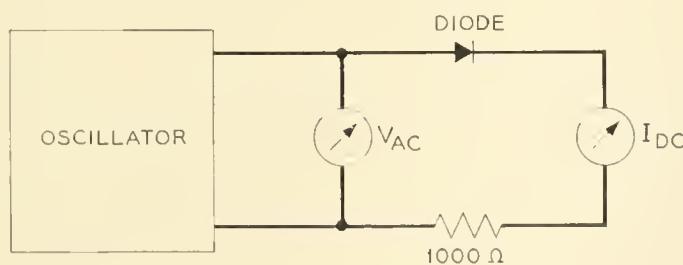


Fig. 15 — Rectification measuring circuit.

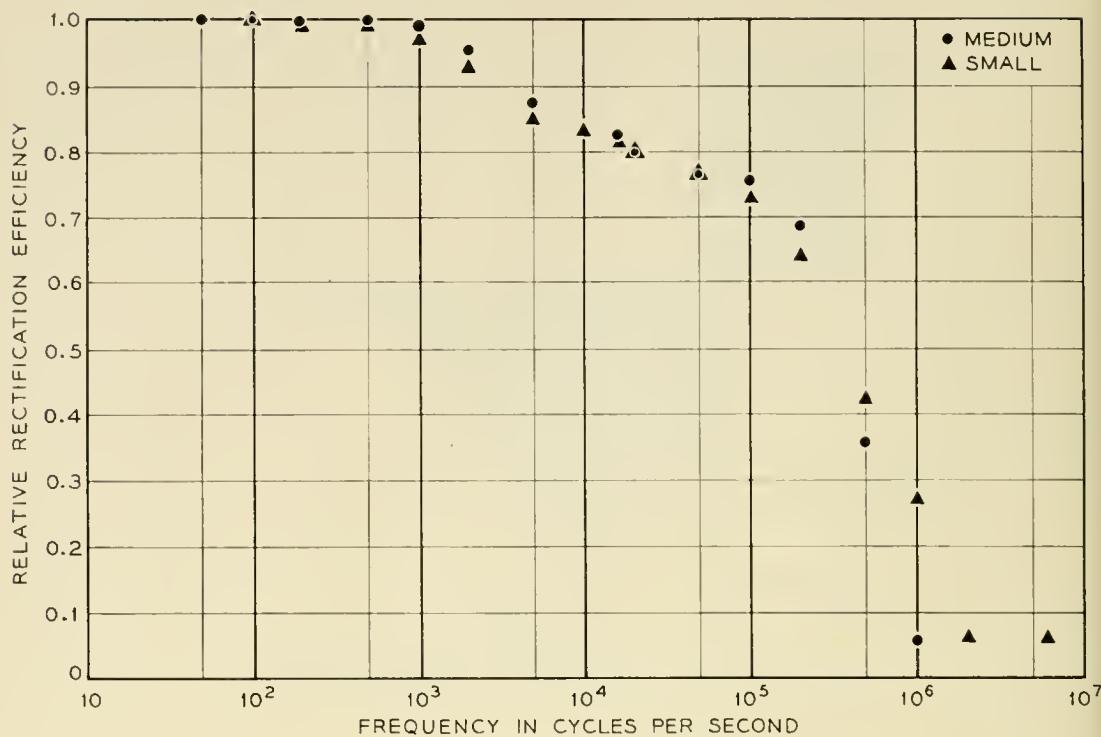


Fig. 16 — Relative rectification efficiency versus frequency.

of these same devices. Such relationships among the measurable parameters of these devices make it possible to design and control the electrical characteristics of the units and therefore make them extremely attractive from an engineering point of view.

6.0 MECHANICAL AND THERMAL DESIGN

6.1 In order to have a device that is usable for more than experimental purposes, it is necessary that it be packaged in a mechanically stable structure and that the heat generated in the combined unit should not lead to a condition where the device no longer has its desirable characteristics. In earlier sections of this paper several mechanical requirements of a satisfactory package have been suggested. These may be repeated at this point. First, pressure contacts are not satisfactory; second, oxidizing ambients are to be avoided; third, approximately one watt per ampere of forward current is generated and must be disposed; and fourth, the package must be electrically satisfactory. The first requirement is met by using soldered contacts. Since these rectifiers are usable at temperatures over 200°C, a solder was chosen that has a melting point over 300°C. The second requirement necessitated the use of a hermetic seal structure. If the seal is truly hermetic, no gases can

enter or leave the package and thus no changes of the device due to the enclosed gas should occur as long as the gas does not react with the silicon, solder or package. However, no seal is absolutely vacuum tight and thus care should be used in choosing a package design so that minimum effects should occur to the electrical properties during the use of the device. The third requirement of the disposal of the internally developed heat suggested the use of copper due to its high thermal conductivity. However, a small package alone is capable of dissipating only a small amount of heat without reaching a temperature that is too high for the device. This necessitates the use of cooling fins in conjunction with the device to make use of its electrical properties. This thermal requirement demands a package to which thermal fins can be attached. This is met by having the package contain a bolt terminal to which thermal fins can be attached or by which the unit can be mounted to a chassis for cooling. The fourth requirement consists of two parts; the package must have two leads that are electrically separated from one another and the leads must be sufficiently heavy to conduct the maximum currents. The first of these requirements is met by using glass-to-metal seals in the package and the second is met by using copper leads of sufficiently heavy cross-section. The resulting packages for the units discussed in this paper are shown in Fig. 6. It should be remembered that the packages are only intermediate development packages and that further work will probably alter these both in size and in shape. However, all the requirements mentioned will be applicable to any package.

6.2 The units pictured in Fig. 6 have a range of dc current ratings associated with them. The lower rating of each device corresponds to the maximum rating of the next smaller device. Of course, the larger units could be used for smaller current applications; however, such use would be like using a freight car to haul a pound of coal. The maximum rating of each device has been arbitrarily chosen for it to operate with a reasonable sized cooling fin at an ambient of 125°C and no forced air or water cooling. It is known that the ratings could be increased by either method of forced cooling. It has been found that a copper convection cooling fin is able to dissipate 8 milliwatts per square inch per degree centigrade. This cooling rate is obtained from the difference between the average temperature of the fin and the ambient temperature over the effective exposed area of the fin. For example, a copper fin $3\frac{1}{2}$ inches square when mounted so that both surfaces are effective for cooling will be able to dissipate ten watts and at the same time prevent the temperature of the fin from exceeding 50°C above the ambient temperature. Another thermal drop is found between the junction and the

base of the package. This temperature difference depends mostly on the material of the base and its geometry. In the devices presented this drop is not more than 15°C at the maximum rated current. Thus the largest drop in temperature occurs between the cooling fin and the ambient which means that the design of the cooling fin is the controlling factor in the operating junction temperature of the rectifier.

6.3 It is possible to use the devices without an attached cooling fin. In this case, the maximum current is limited essentially by the size of the package. The small rectifier package is designed for $\frac{1}{2}$ watt dissipation and therefore the maximum current that should be rectified is about 500 milliamperes. The medium size unit will comfortably rectify 1 ampere without any additional cooling and the large rectifier unit will conduct 3 amperes under the same conditions.

7.0 RELIABILITY AND LIFE MEASUREMENTS

7.1 One of the desired properties of any device is that it should operate satisfactorily at its rating for a long period of time. The above general statement contains many implications which should be made specific for the devices under consideration in this paper. By stating that these devices should operate satisfactorily we mean that they should not age during operation; that is, the forward and reverse characteristics at any temperature should not change with time. The statement implies that a rating has been established for the units. Furthermore, a "long period of time" has to be defined. There are applications where a few hours is considered a long time as in some military applications. However, in most Bell System applications, a long period of time may be 20 years or approximately 200,000 hours. Clearly, in the short time since these rectifiers have been developed, it is impossible to make a fair statement as to their reliability and their life expectancy. However, it is possible to present some results of some early experiments and describe where and how the units have lived and died. It is this information that we will present in this section. It is a common experience that during the early development of any new component, there are many units that do not satisfy all the requirements of the desired end product. These units will generally deteriorate very rapidly on life testing due to some electrical or mechanical instability. The units used for life testing have been screened to remove the above mentioned unstable devices.

7.2 The life tests consist of four types; shelf tests at room temperature and at 150°C , forward characteristic tests, reverse characteristic

tests, and load tests. The last tests are really the important tests; however, these require the dissipation of large quantities of power in the load to test only a few devices. Therefore only a few units were tested in this condition and the majority tested under other conditions. The several units under load test have been operating for six months with no noticeable change in their characteristics. These devices are the small and medium size development units. The large rectifiers would require about 10 kilowatts of dissipation each in a load to give them a fair load test.

The shelf tests at room temperature and at a temperature of 150°C have been running for six months and have indicated that most of the units remain practically constant. There have been some units that improve on standing but there is no method of predicting which ones will improve. Some units get worse on standing; however, most of these can be predicted from the initial tests since these units usually have a noisy reverse characteristic near the reverse breakdown voltage. The units that change differ only in their reverse characteristic; the forward characteristic changes are not detectable indicating that the contacts are stable. The changes in the reverse characteristic are probably due to the trapping of ions and vapors on the surface of the devices during the packaging operation. Another source of these variations is due to the non-hermeticity of the glass-to-metal seals allowing gases to diffuse into the package where they may cause changes in the reverse characteristic. These leaks have been found in many early units and new assemblies are being tried at present.

The forward characteristic life test was considered a good test since the device is subject to practically all the internal power dissipation without requiring the relatively high load dissipation. It is tests of this nature that allow one to rate the various size devices. The medium size rectifiers that ran at 15 amperes in this test failed after three months of testing; whereas no units running at 5 and 10 amperes have failed during the six months since the tests have started although their reverse characteristics have changed slightly. It should be noted that most of the change of reverse characteristic occurred during the first test period of two weeks. These changes are probably due to the causes mentioned in the above paragraph.

Reverse characteristic tests have been running for several months on a group of 10 small rectifiers which we feel have a better gas tight seal than the other development units. The voltage has been adjusted on these units such that they are pulsed into the breakdown region with a

maximum current of one milliampere. None of these units show any appreciable change.

7.3 All of those tests in the past sub-section had to do with continuous dc or ac power being supplied to the units under test. However, in actual operation the units may be subject to voltage pulses due to power line pulses, accidental shorts, etc. In order for the rectifier to be useful, it should be able to take an overload for a period of time sufficiently long to allow a protective device to operate. Pulse tests have been performed on the medium size rectifier. These devices are able to withstand over 300 amperes for times of the order of 50 microseconds. However, the fastest circuit breakers operate in about 20 milliseconds and for this period, these units can stand only approximately 50 amperes before failing. Since these units have such a low forward resistance at the operating currents (Fig. 7), any small increase in voltage across the diode will change the current through the device to a very large quantity. Therefore series protective resistances may be necessary where the possibility of short-circuiting the device is high. Such operation would reduce the efficiency of the unit and is to be avoided if possible. Another type of protection may be afforded through the use of a high impedance, high current inductor. This type of protection is quite bulky and heavy and suitable only for stationary apparatus. Another common possibility of burnout of the devices occurs when using a capacitance input in conjunction with the rectifier. When the circuit is turned on, large currents will flow to charge up the capacitors and consequently burn out the rectifiers. One possible protection from such operation is the use of a series resistance in conjunction with a time delay relay. The series resistance will limit the initial capacitor charging current and the time delay relay will short out the resistance after the capacitors have reached near their maximum charge.

7.4 Dissection of burned out units have indicated that the failure takes place through small spots on the device. This can be explained by the fact that some small areas of the device have slightly better forward characteristics. These areas will tend to conduct most of the forward current. Therefore most of the power will be dissipated there and these areas will become even more conducting leading to a channeling of the forward current through these spots with the consequent burnout. The best way to avoid such mishaps would be to make a more uniform device. Experiments are in process along this line. Another less satisfactory method would be the control of contact resistance such that the current would be limited in any particular area by the contact resistance. Similar ideas must be considered when paralleling these diffused junction

silicon rectifiers. It is possible to use these devices in parallel if one adjusts the lead resistances such that no one unit will be allowed to conduct much more than its share of the current.

7.5 As a conclusion to this section, it should be noted that these rectifiers are expected to have a long life when operated within their ratings. They are able to operate for short periods of time (seconds) at five times their rated currents. Since the rectifiers have an extremely small series resistance, they should be protected against accidental surges and turning on to a capacitance input filter.

8.0 SUMMARY

8.1 The development rectifiers described in the article are silicon diffused p-n junction rectifiers. These devices together with associated cooling fins can be used to rectify a complete range of currents from 0 to 50 amperes in a single phase, half wave rectifier circuit. They can be used in more complex rectification circuits to yield even more dc current. Also, they are able to withstand at least 200 volts peak in the inverse direction and operate satisfactorily at temperatures as high as 200°C. Furthermore, one process of diffusion and plating is sufficient for all the devices of the class. This makes it possible for one diffusion and plating line to feed material for all the rectifiers in a manufacturing operation.

8.2 The rectifiers discussed behave according to the theory of semiconductor devices which makes it possible to design them for given electrical, thermal, and mechanical characteristics. One failure to meet ideal theory of a p-n junction is with the forward characteristic.

8.3 The diffused silicon type of rectifier has been compared with germanium and selenium units and has better reverse characteristics at all temperatures. In the forward direction, the germanium units have a smaller voltage drop for any given current than the silicon rectifiers but the silicon devices are capable of operating at much higher temperatures, thereby permitting higher overall current densities than the germanium devices.

8.4 The diffused silicon rectifiers are capable of use in any rectifier application where dc currents up to the order of 100 amperes are required and where inverse peak voltages up to 200 volts are encountered. Another important use for these devices will be in the magnetic amplifier application where the low reverse currents of silicon will enable large amplification factors to be realized. Since the forward characteristics of these devices are so uniform, they can be used in voltage reference circuits that require voltages near 0.6 volts and in circuits uti-

lizing the exponential character of the forward characteristic. However, as is to be expected from devices with the characteristics described in this paper, the most immediate application will be found in power supplies.

ACKNOWLEDGMENTS

It is obvious that the work reported in this paper is not the result of one man's labor. Much of the stimulus and many of the ideas are those of K. D. Smith. Other members of the Semiconductor Device Department who have contributed considerably to the development of these devices are R. L. Johnston, R. Rulison, and R. C. Swenson. D. A. Kleinman, J. L. Moll and I. M. Ross have been most helpful in discussing the theoretical aspects of these devices. The author wishes to thank H. R. Moore for his suggestions on protecting the silicon rectifiers against large overloads.

The Forward Characteristic of the PIN Diode

By D. A. KLEINMAN

(Manuscript received January 18, 1956)

A theory is given for the forward current-voltage characteristic of the PIN diffused junction silicon diode. The theory predicts that the device should obey a simple PN diode characteristic until the current density approaches 200 amp/cm². At higher currents an additional potential drop occurs across the middle region proportional to the square root of the current. A moderate amount of recombination in the middle region has little effect on the characteristic. It is shown that the middle region cannot lead to anomalous characteristics at low currents.

INTRODUCTION

In some diode applications it is desirable to have a very low ohmic resistance as well as a high reverse breakdown voltage. A device meeting these requirements, in which the resistance is low because of heavily doped P^+ and N^+ contacts and the breakdown voltage is high because of a lightly doped layer between the contacts, has been described by M. B. Princee.¹ The device is shown schematically in Figure 1a and consists of three regions, the P^+ contact, the middle P layer, and the N^+ contact. The device is called a *PIN* diode because the density P of uncompensated acceptors in the middle region is much less than P^+ or N^+ and in normal forward operation much less than the injected carrier density.²

We shall let the edge of the P^+P junction in the middle region be $x = 0$, and the edge of the PN^+ junction in the middle region be $x = w$. Thus the region $0 \leq x \leq w$ is space charge neutral and bounded at each end by space charge regions whose width is of the order of the Debye length

¹ Princee, M. B., Diffused *p-n* Junction Silicon Rectifiers, B.S.T.J., page 661 of this issue.

² A device with similar geometry has been discussed by R. N. Hall, Proc. I.R.E., 40, p. 1512, 1952.

$$\lambda = (K/\beta eP)^{1/2} \sim 1.5 \times 10^{-5} \text{ cm.} \quad (1)$$

where K is the dielectric constant, e is the electronic charge, and β is the constant.

$$\beta = e/kT = \mu_n/D_n = \mu_P/D_P \quad (2)$$

which at room temperature is 38.7 volt^{-1} . We shall denote points in the P^+ and N^+ contacts on the edges of the space charge regions by oo and ww respectively. Thus n_{oo} is the electron density in the P^+ contact at the junction, and n_o is the electron density at the same junction in the middle region. Similarly p_{ww} is the hole density at the junction in the N^+ contact and p_w is the hole density at the junction in the middle region. We shall denote equilibrium carrier densities in the three regions by n_{P^+} , n_P , p_p , p_{N^+} . Typical values for the parameters characterizing

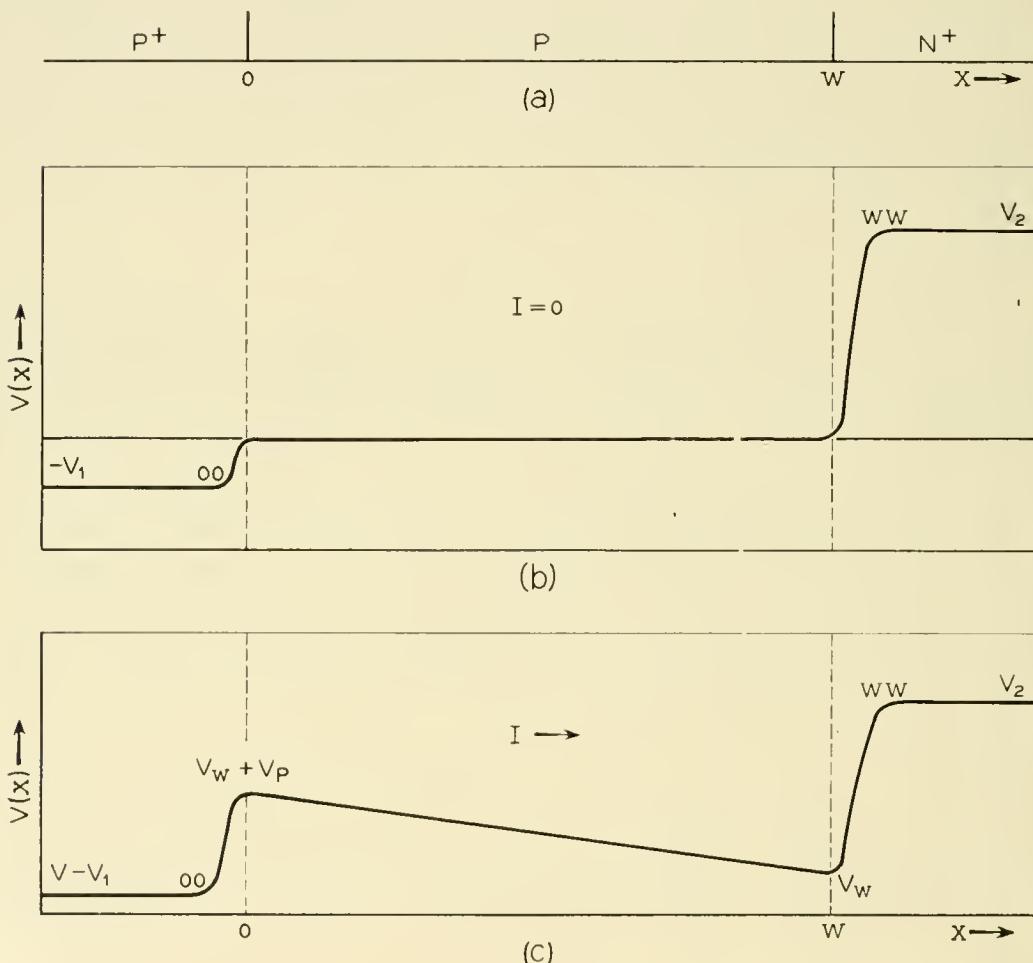


Fig. 1 — Schematic representation of the PIN diode with the P^+ and N^+ contacts regarded as extending to infinity. (b) shows the electrostatic potential in equilibrium and (c) shows the potential when a forward current flows.

the device are

$$\begin{aligned} w &\sim 2 \times 10^{-3} \text{ cm} \\ P &\sim 10^{15} \text{ cm}^{-3} \\ N^+, P^+ &\sim 10^{18} \text{ cm}^{-3} \\ L_n, L_p &\sim 10^{-4} \text{ cm} \end{aligned} \tag{3}$$

where L_n, L_p are minority carrier diffusion lengths in the contacts.

The present treatment makes three distinct approximations. The first is to neglect the voltage drop in the contacts. The highest currents ordinarily used are of the order of 500 amp/cm² which should produce an ohmic drop in the contacts of about 1 volt/cm. Since the entire diode has a length of about 0.01 cm we are neglecting only about 0.01 volts in this approximation.

The second approximation is to regard the Debye length as small compared to w and the diffusion lengths L_n, L_p . If L_n, L_p are as small as the typical values given in (3) the error made in this approximation is not completely negligible. Nevertheless, we use the approximation because it enables us to regard the device as three relatively large neutral regions and two relatively narrow space charge regions. The behavior of the device can then be determined by solving for the diffusion and drift of carriers in the neutral regions subject to boundary conditions connecting the carrier densities across the space charge layers.

The third approximation is to neglect any increase in majority carrier density in the contacts due to injection of minority carriers. This approximation is valid until the current density approaches 5×10^4 amp/cm², which is well above anticipated operating currents. It is conceivable that in some junctions all the current may flow through small active spots at which the current density is very high, perhaps exceeding the above figure. In such cases the current flow is two or three dimensional and the present analysis would not apply.

It is also necessary to assume some law for carrier recombination. We shall assume that recombination in the contacts is linear in the injected minority carrier density

$$\frac{dI_n}{dx} \sim \frac{n - n_{P^+}}{\tau} \tag{4}$$

Modification of the theory to suit other recombination laws is simple in principle, although considerable analytical complications might be encountered. It seems most likely that in silicon PN junctions the recombination actually is nonlinear. It can be shown that if the recombi-

nation follows some power ν of the injected density

$$\frac{dI_n}{dx} \sim n^\nu \quad (5)$$

- the forward characteristic of a simple PN junction is of the form

$$\exp [1/2\beta(\nu + 1)V] \quad (6)$$

Thus nonlinear recombination can account for the observation that in silicon diodes the slope of V versus $\log I$ is usually much less than β . Our purpose here is not to study this interesting effect, but to study those effects which are due to the presence of the middle region. Therefore, we assume linear recombination for the sake of simplicity. In the last section we give a brief consideration of what to expect in the case of nonlinear recombination in the contacts. Recombination in the middle region will also be assumed to be linear in the injected carrier density, but this assumption is not critical, since it turns out that a moderate amount of recombination in the middle region does not change the qualitative behavior of the device.

BASIC EQUATIONS

Fig. 1(b)³ shows the electrostatic potential $V(x)$ for the equilibrium case $I = 0$. The potential is constant except in the space charge layers. If we call the potential of the middle region zero, the P^+ and N^+ contacts are at the potentials $-V_1$ and V_2 respectively, where

$$\begin{aligned}\beta V_1 &= \ln (P^+/p_P) \\ \beta V_2 &= \ln (N^+/n_P)\end{aligned} \quad (7)$$

Figure 1c shows the potential when a forward current I flows and a forward bias V is produced across the device. We shall define the potential so that the N^+ contact remains at V_2 , which puts the P^+ contact at potential $V - V_1$. The potential at a point x is then given by

$$V(x) = V_2 - \int_{ww}^x E(x) dx \quad (8)$$

where $E(x)$ is the electric field assumed zero in the contact regions $x > ww$ and $x < oo$. The applied bias V consists of three terms

$$V = V_0 + V_P + V_w \quad (9)$$

³ This potential distribution has been discussed by A. Herlet and E. Spenke, Zeits. f. Ang. Phys., B7, H3, p. 149, 1955.

where V_0 is the forward bias across the junction at $x = 0$, V_P is the potential drop in the middle region, and V_w is the forward bias across the junction at $x = w$. In this notation $V(0) = V_w + V_P$ and $V(w) = V_w$.

The total current density is constant

$$I_n(x) + I_p(x) = I \quad (10)$$

We shall denote electric current densities by eI_n , eI_p , so that I_n , I_p , I have the dimensions of (particles/cm²-sec). At $x = 0$ and $x = w$ the minority carrier currents must flow into the contacts by diffusion, which gives the boundary conditions

$$\begin{aligned} I_p(w) &= I_{ps} \left\{ \frac{p_{ww}}{p_{N^+}} - 1 \right\} \\ I_n(0) &= I_{ns} \left\{ \frac{n_{oo}}{n_{P^+}} - 1 \right\} \end{aligned} \quad (11)$$

where I_{ps} , I_{ns} are saturation current densities

$$I_{ps} = \frac{p_N^+ D_p}{L_p}, \quad I_{ns} = \frac{n_P^+ D_n}{L_n} \quad (12)$$

The order of magnitude of the saturation current density is given by

$$e(I_{ns} + I_{ps}) \sim 3 \times 10^{-10} \text{ amp/cm}^2 \text{ in Si}$$

based on the typical values of (3). Equations (11) contain the assumptions of linear recombination and small injection into the contacts as discussed in the introduction.

In the middle region the current densities satisfy

$$\begin{aligned} I_p(x) &= D_p \left\{ -\frac{dp}{dx} + \beta p E \right\} \\ I_n(x) &= D_n \left\{ \frac{dn}{dx} + \beta n E \right\} \end{aligned} \quad (13)$$

Let us assume these equations remain valid in the space charge regions.⁴ Since these space charge regions are narrow I_n and I_p can be considered constant and the solution of (13) in the space charge regions is

$$\begin{aligned} p(x) &= e^{-\beta V(x)} \left\{ p_{ww} e^{\beta V_2} - \frac{I_p(w)}{D_p} \int_{ww}^x e^{\beta V(x)} dx \right\} \\ n(x) &= e^{\beta V(x)} \left\{ n_{oo} e^{-\beta(V-V_1)} + \frac{I_n(0)}{D_n} \int_{00}^x e^{-\beta V(x)} dx \right\} \end{aligned} \quad (14)$$

⁴ Shockley, W., B.S.T.J., 28, p. 435, 1949.

Since $\lambda/L_p \ll 1$ we can write for the junction at $x = w$

$$\begin{aligned} p(w) &= e^{-\beta V_w} \left\{ p_{ww} e^{\beta V_2} - \frac{(p_{ww} - p_N^+)}{L_p} \int_{ww}^w e^{\beta v} dx \right\} \\ &= p_{ww} e^{\beta(V_2 - V_w)} \left\{ 1 - O\left(\frac{\lambda}{L_p}\right) \right\} \end{aligned} \quad (15)$$

where $O(\lambda/L_p)$ means a term of order λ/L_p . Thus we see that if we may neglect λ/L_p and λ/L_n we have the following simple boundary conditions at the junctions

$$\begin{aligned} n_{oo} &= n_o (n_P^+ / n_P) e^{\beta V_0} \\ p_o &= p_P e^{\beta V_0} \\ n_w &= n_P e^{\beta V_w} \\ p_{ww} &= p_w (p_N^+ / p_P) e^{\beta V_w} \end{aligned} \quad (16)$$

It is clear that in order to divide the device into three neutral regions we must also be able to neglect λ/w .

Finally, we have the condition of space charge neutrality

$$p - n = P \quad (17)$$

It can be shown that the term $K^{-1} dE/dx$ is of order $(\lambda/L)^2$ or $(\lambda/w)^2$ and therefore negligible in our approximation. Therefore (17) is the Poisson equation for the middle region in our approximation. When we use (17) we are not saying that $E(x)$ is constant but only that $K^{-1} dE/dx$ is negligible compared to $p(x)$ and $n(x)$. The basic equations then are (10), (11), (13), (16), (17).

Large Injection, No Recombination

In this section we consider current densities of the order of magnitude of those that flow in normal operation of the diode as a power rectifier. These currents inject large densities of electrons and holes into the middle region greatly increasing its conductivity. The result is that the voltage drop V_P is small even though the normal resistivity of the middle region is high. For this reason the device has been called a conductivity modulated rectifier. Also in this section we shall neglect recombination in the middle region, which makes $I_n(x)$ and $I_p(x)$ constant and greatly simplifies the analysis. The effect of recombination is to remove carriers and increase the drop across the middle region. Therefore, it is desirable to keep recombination in the middle region as low as possible.

Under conditions of large injection we can say

$$\begin{aligned} n &\gg P, & p &\gg P \\ n_{oo} &\gg n_P^+ & p_{ww} &\gg p_N^+ \end{aligned} \quad (18)$$

so that (11) becomes

$$\begin{aligned} I_n &= I_{ns}(n_{oo}/n_P^+)^+ \\ I_p &= I_{ps}(p_{ww}/p_N^+)^+ \end{aligned} \quad (19)$$

and (17) becomes

$$n(x) = p(x) \quad 0 \leq x \leq w \quad (20)$$

Equation (16) becomes

$$\begin{aligned} n_{oo} &= n_o(n_P^+/n_P)e^{\beta V_0} \\ n_o &= p_P e^{\beta V_0} \\ n_w &= n_P e^{\beta V_w} \\ p_{ww} &= n_w(p_N^+/p_P)e^{\beta V_w} \end{aligned} \quad (21)$$

Equations (13) can be written

$$\begin{aligned} \beta E &= \frac{I_n + bI_p}{2D_n n} \\ \frac{dn}{dx} &= \frac{I_n - bI_p}{2D_n} \end{aligned} \quad (22)$$

where $b = D_n/D_p$. Combining (19) and (21) gives the equations

$$\begin{aligned} n_o &= n_i(I_n/I_{ns})^{1/2} \\ n_w &= n_i(I_p/I_{ps})^{1/2} \end{aligned} \quad (23)$$

where $n_i^2 = n_P p_P$ is a constant, and also

$$\begin{aligned} \beta V_0 &= \frac{1}{2} \ln \frac{n_P}{p_P} \frac{I_n}{I_{ns}} \\ \beta V_w &= \frac{1}{2} \ln \frac{p_P}{n_P} \frac{I_p}{I_{ps}} \end{aligned} \quad (24)$$

From the first equation (22) we have

$$\beta V_P = \frac{I_n + bI_p}{2D_n} \int_0^w \frac{dx}{n(x)} \quad (25)$$

Upon invoking the second equation of (22) we get

$$\beta V_p = \frac{I_n + bI_p}{I_n - bI_p} \ln \frac{n_w}{n_o} \quad (26)$$

and

$$n_w = n_o + \frac{I_n - bI_p}{2D_n} w. \quad (27)$$

We see that V_p is always positive in sign whatever the sign of $I_n - bI_p$.

We now define a parameter

$$\gamma \equiv n_o/n_w \quad (28)$$

and a device constant

$$R \equiv I_{ns}/I_{ps} \quad (29)$$

Then from (23) and (10)

$$\begin{aligned} I_n/I_p &= R\gamma^2 \\ I_n &= \frac{R\gamma^2}{1 + R\gamma^2} I \quad I_p = \frac{1}{1 + R\gamma^2} I \end{aligned} \quad (30)$$

Combining (23), (27) and (30) gives the equation for γ as a function of total current

$$\begin{aligned} \gamma &= 1 - \frac{I_n - bI_p}{2D_n} \frac{w}{n_w} \\ &= 1 - \sqrt{\frac{I}{I_0}} \frac{(\gamma/\gamma_\infty)^2 - 1}{\sqrt{1 + b(\gamma/\gamma_\infty)^2}} \end{aligned} \quad (31)$$

where

$$\gamma_\infty^2 \equiv b/R \quad (32)$$

and I_0 is a unit of (particle) current density characteristic of the device

$$I_0 = \frac{4D_p^2 n_i^2}{w^2 I_{ps}} = 4 \left(\frac{L_p}{w} \right)^2 \frac{N^+ D_p}{L_p} \quad (33)$$

A typical value for $e I_0$ in a silicon diode is

$$e I_0 \sim 200 \text{ amp/cm}^2 \quad (34)$$

based on (3).

From (26) the potential drop in the middle region can be written

$$\beta V_p = -\frac{\gamma^2 + \gamma_\infty^2}{\gamma^2 - \gamma_\infty^2} \ln \gamma \quad (35)$$

From (24) and (30)

$$\beta(V_0 + V_w) = \ln \frac{I}{I_0} + \ln \frac{\gamma}{1 + b(\gamma/\gamma_\infty)^2} + \ln \frac{I_0}{I_{ps}} \quad (36)$$

Thus the total applied bias V as a function of total current density I is given by

$$\beta V = \ln \frac{I}{I_0} - \frac{\gamma^2 + \gamma_\infty^2}{\gamma^2 - \gamma_\infty^2} \ln \gamma + \ln \frac{\gamma}{1 + b(\gamma/\gamma_\infty)^2} + \ln \frac{I_0}{I_{ps}} \quad (37)$$

where $\gamma(I)$ is the (positive) solution of (31).

Thus far we have referred the problem of the $V - I$ characteristic to the problem of calculating $\gamma(I)$ from (31). We see that in the limits of high and low current γ approaches the limits

$$\begin{aligned} \gamma &\rightarrow 1 & I \ll I_0 \\ \gamma &\rightarrow \gamma_\infty & I \gg I_0 \end{aligned} \quad (38)$$

and in general lies between these limits. A good approximate solution is readily obtained by replacing (31) with the quadratic equation

$$\begin{aligned} \gamma &= 1 - z[(\gamma/\gamma_\infty)^2 - 1] \\ z &= (I/I_0)^{1/2} (1 + b)^{-1/2} \end{aligned} \quad (39)$$

which has the solution

$$\gamma = \frac{\sqrt{\gamma_\infty^4 + 4(1+z)z\gamma_\infty^2} - \gamma_\infty^2}{2z} \quad (40)$$

A plot of this solution is shown in Fig. 2 as a function of z for $\gamma_\infty = 1/2$, $\gamma_\infty = 2$. Since $\gamma(I)$ is bounded by unity and γ_∞ , which usually will be of order unity, we can reject some of the dependence of V upon γ and retain only its essential dependence upon I . This appears in the first and second terms of (37). By means of (31) this second term can be written

$$\beta V_p = \left[\frac{\ln \gamma}{\gamma - 1} \frac{(\gamma/\gamma_\infty)^2 + 1}{\sqrt{1 + b(\gamma/\gamma_\infty)^2}} \right] \sqrt{\frac{I}{I_0}} \quad (41)$$

Retaining only the essential dependence on I we write this equation

$$\beta V_p = C(I/I_0)^{1/2} \quad (42)$$

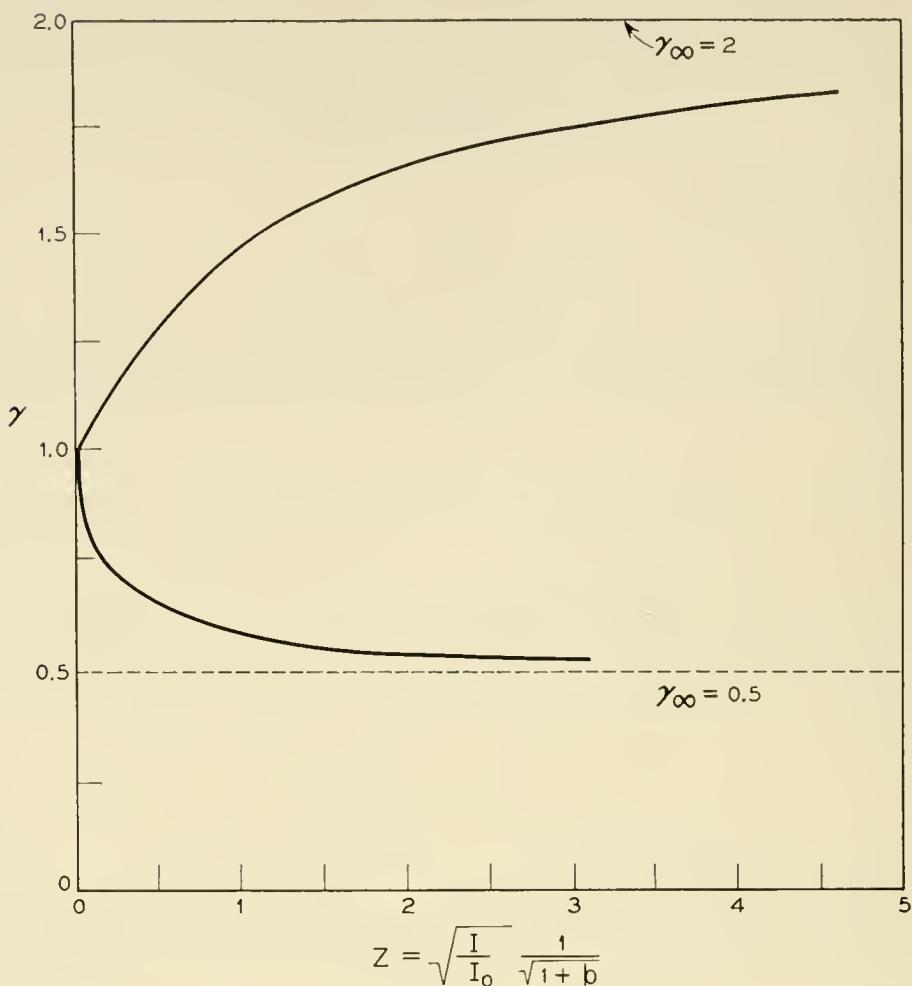


Fig. 2 — The function $\gamma(z)$ given by equation (40) for two choices of γ_∞ .

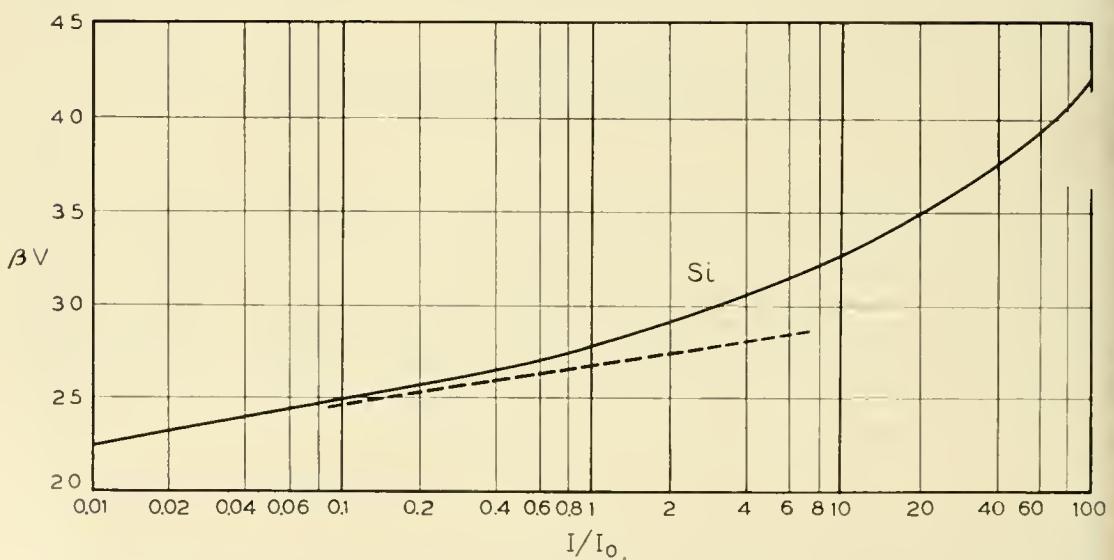


Fig. 3 — The voltage-current characteristic of the PIN diode according to equation (44). The dashed line represents an ideal PN diode and $eI_0 \sim 200$ amp/cm² in silicon.

where C is a constant representing the slowly varying coefficient of γ/I_o in (41). We choose C such that (42) becomes exact at high current density when βV_p is large

$$C = \frac{\ell n \gamma_\infty}{\gamma_\infty - 1} \frac{2}{\sqrt{b + 1}} \quad (43)$$

When we regard the third and fourth terms of (37) together as a constant βV_c we obtain the simplified voltage-current characteristic

$$\beta V = \ell n \frac{I}{I_0} + C \sqrt{\frac{I}{I_0}} + \beta V_c \quad (44)$$

In this approximation it is unnecessary to evaluate $\gamma(I)$ from (31).

Fig. 3 shows plots of βV versus I/I_o calculated from (44). For plotting the curves the value $C = 1.1$ was used. To choose a value for βV_c we put $\gamma = 1$, which gives

$$\ell n \frac{\gamma}{1 + b(\gamma/\gamma_\infty)^2} \rightarrow \ell n \frac{1}{1 + R} \gamma \rightarrow 1 \quad (45)$$

so that

$$\beta V_c \rightarrow \ell n[I_o/(I_{ns} + I_{ps})] \quad (46)$$

which has the value 27 in silicon according to the values in (3). The dotted line is the asymptote approached by the curve at low current densities

$$\beta V \rightarrow \ell n \frac{I}{I_{ns} + I_{ps}} \quad I \ll I_0 \quad (47)$$

This is the characteristic of a simple PN junction when

$$I \gg I_{ns} + I_{ps}.$$

We return now to the question of when the large injection conditions (18) are satisfied. Let us suppose I is much less than I_0 so that $\gamma \sim 1$, $I_n/I_p \approx R$. It follows from (30) and (23) that

$$n_o \approx n_w \approx n_i[I/(I_{ns} + I_{ps})]^{1/2} \quad (48)$$

Now let us set $n_o \gg P$ which gives a condition on the current density

$$I \gg (P/n_i)^2 (I_{ns} + I_{ps}). \quad (49)$$

Setting $n_{oo} \gg n_p^+$, $p_{ww} \gg p_N^+$ gives

$$I \gg I_{ns} + I_{ps}. \quad (50)$$

Usually $P \gg n_i$ so that (49) includes (50). When numbers are put in

from (3) we get the condition for large injection

$$eI \gg 0.07 \text{ amp/cm}^2 \text{ in } Si \quad (51)$$

Since this current in (51) is much less than eI_o , we may quite properly speak of large injection $n \gg P$ and small currents $I \ll I_o$ at the same time.

Let us denote by

$$I_{CM} = (P/n_i)^2 (I_{ns} + I_{ps}) \quad (52)$$

the current density at which conductivity modulation starts to be important. Then we may distinguish three ranges of current: (a) very small current $I < I_{CM}$ for which large injection analysis does not apply; (b) low current $I_{CM} < I < I_o$ for which large injection analysis applies, but the voltage drop V_p in the middle region is negligible; (c) large current $I > I_o$ for which V_p is sizable. The treatment of this section has covered ranges (b) and (c). Range (c) (as treated here) does not extend to infinity but only up to current densities of the order

$$\frac{eN^+ D_p}{L_p} \sim 8 \times 10^4 \text{ amp/cm}^2$$

so that the diffusion currents in the contacts may be treated as a small injection.

Small Injection, No Recombination

In this section, we shall cover ranges (a) and (b) in current density. We must go back to the basic equations, but we shall make use of two facts that have come out of the large injection analysis: (a) βV_p is negligible when $I \ll I_o$; (b) $\gamma = n_o/n_w \approx 1$ which means $n(x)$ and $p(x)$ are essentially constant in the middle region $0 \leq x \leq w$ when $I \ll I_o$. When we set

$$n_o = n_w, \quad p_o = p_w \quad (53)$$

equations (16) give us

$$n_{oo} = n_p^+ e^{\beta(V_0 + V_w)} \quad (54)$$

$$p_{ww} = p_n^+ e^{\beta(V_0 + V_w)}$$

Then (11) gives

$$I = I_n + I_p = (I_{ns} + I_{ps}) [e^{\beta(V_0 + V_w)} - 1] \quad (55)$$

Now $V_o + V_w$ is the total applied bias when V_p can be neglected; there-

fore we obtain the characteristic

$$\beta V = \ln \left(\frac{I}{I_{ns} + I_{ps}} + 1 \right) \quad (56)$$

which is valid until I approaches I_o . Of course we would not have obtained this ideal characteristic of a simple PN junction had we taken recombination into account; our result depends upon the constancy of $n(x)$ and $p(x)$ in the middle region. For the case of no recombination in the middle region (56) and (44) cover ranges (a), (b) and (c). Instead of (44) the more exact expression (37) could be used requiring the evaluation of $\gamma(I)$ from (31). It seems that the extra refinement is of no help in understanding the device and unnecessary in treating experimental data. Therefore, we shall adopt (44) and the approximations leading to it as a model for treating the more complicated recombination case. That is, we shall seek a generalization of (44) which takes recombination into account in a sufficiently good approximation.

Large Injection with Recombination

We are interested in determining the effect of recombination in the middle region upon the operating characteristics of the device. Therefore we go immediately to the large injection case $n = p$. Equation (16) become

$$\begin{aligned} n_w &= n_P e^{\beta V_w} & p_{ww} &= n_w (p_N^+ / p_P) e^{\beta V_w} \\ n_o &= p_P e^{\beta V_0} & n_{oo} &= n_o (n_P^+ / n_P) e^{\beta V_0} \end{aligned} \quad (57)$$

which gives

$$\beta(V_o + V_w) = \ln(n_w n_o / n_i^2) \quad (58)$$

We shall assume that recombination is linear in the injected carrier density to simplify the calculation. It will be possible, later to approximate bimolecular recombination by using an appropriate value for the lifetime τ corresponding to the injected carrier density. Therefore we write

$$\frac{dI_n}{dx} = -\frac{dI_p}{dx} = \frac{n}{\tau} \quad (59)$$

Eliminating $I_n(x)$ by use of (13) gives the equation for $n(x)$

$$\frac{d^2n}{dx^2} = \frac{n}{L^2} \quad (60)$$

where L is the effective diffusion length in the middle region

$$L = [2D_n \tau / (b + 1)]^{1/2} \quad (61)$$

The solution of (60) may be written

$$n(z) = \frac{n_0 \sinh(w - z) + n_w \sinh z}{\sinh \omega} \quad (62)$$

where $z = x/L$ is the position variable and $\omega = w/L$ is the length of the middle region in units of L . Fig. 4 shows several of these solutions for the case $n_0 = n_w$.

In equation (60) and the solution (62) we have neglected the equilibrium carrier densities n_p , p_p . The criterion for the validity of this approximation is

$$\sinh \frac{1}{2}\omega \ll (n_0/P), (n_w/P) \quad (63)$$

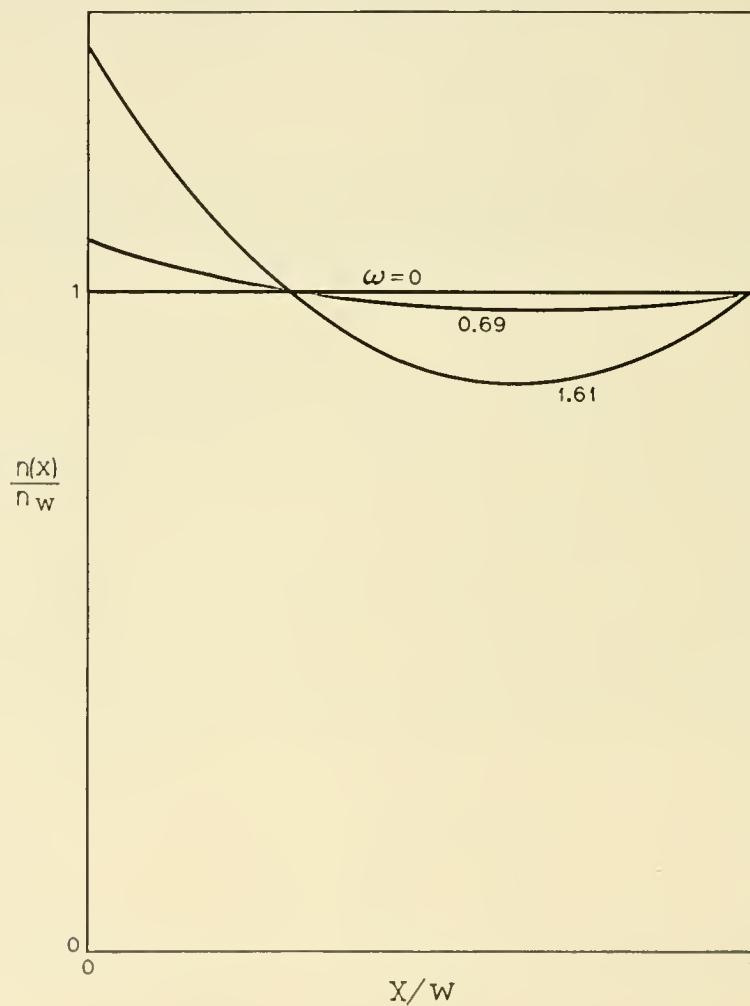


Fig. 4 — The carrier density according to equation (62) for the case $n_0 = n_w$ and several values of ω .

arrived at by considering the minima in the solutions for $\omega \gg 1$. This is really a criterion for conductivity modulation, so we shall assume henceforth that it is satisfied.

We now modify (13) by setting $n = p$ and eliminating $E(x)$ by use of (22)

$$\begin{aligned} I_n(x) &= \frac{bI + 2D_n n'(x)}{b + 1} \\ I_p(x) &= \frac{I - 2D_n n'(x)}{b + 1} \end{aligned} \quad (64)$$

where $n'(x) = dn/dx$. Inserting these currents into (22) gives $E(x)$ and integrating gives the potential drop V_P in the middle region

$$\beta V_p = \frac{bI}{(b + 1)D_n} \int_0^w \frac{dx}{n} - \frac{b - 1}{b + 1} \ln \frac{n_w}{n_0} \quad (65)$$

This is the generalization of (26) for linear recombination.

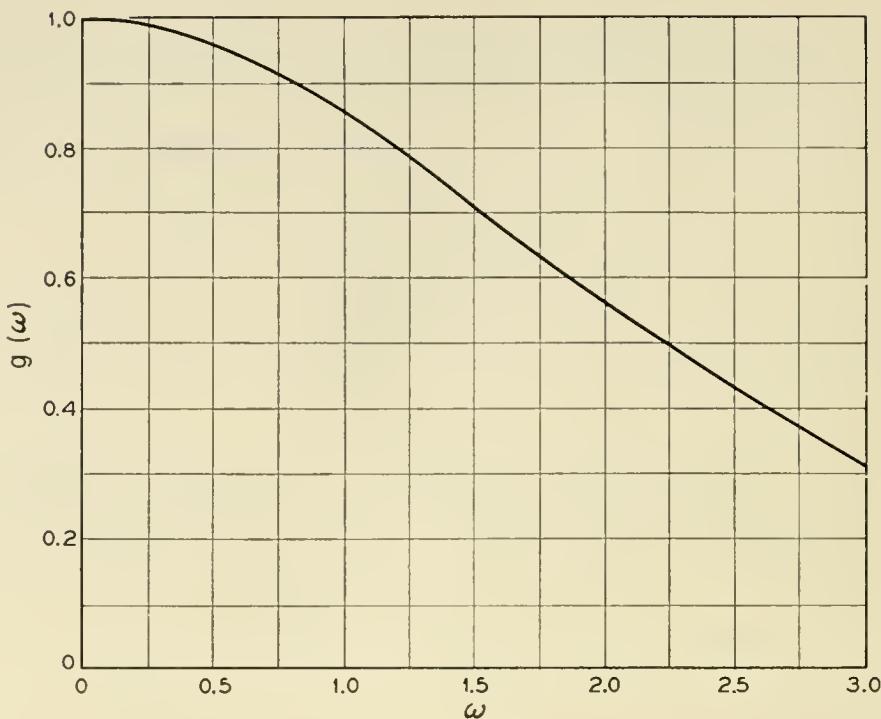
The direct evaluation of (58) and (65) in terms of the total current I leads to a very complicated expression for the applied voltage. It will be shown in the next section that this result reduces in its simplest approximate form retaining only the essential dependence on ω to the formula

$$\beta V \approx \ln \frac{I}{I_{ns} + I_{ps}} + C \sqrt{\frac{I}{I_0(\omega)}} \quad (66)$$

which is identical with (44) except that the characteristic current density is a function of ω

$$\begin{aligned} I(\omega) &= I_0 g(\omega) \\ g(\omega) &= \frac{(\omega/2)^2}{\left[\cosh \frac{\omega}{2} \tan^{-1} \left(\sinh \frac{\omega}{2} \right) \right]^2} \\ &= 1 - \frac{\omega^2}{6} + \frac{\omega^4}{48} - \dots \end{aligned} \quad (67)$$

Fig. 5 shows a plot of $g(\omega)$. These results show that if $\omega < 1$ as we might expect in a good diode recombination has no significant effect on the forward voltage-current characteristic in the conductivity modulation range of operation.

Fig. 5 — The function $g(\omega)$ of equation (67).*Analysis*

We denote

$$\xi = \frac{n_w}{n_i}, \quad \zeta = \frac{n_0}{n_i} \quad (68)$$

From (11) and (67)

$$I_p(\omega) = I_{ps}\xi^2, \quad I_n(0) = I_{ns}\zeta^2 \quad (69)$$

By means of (62) and (64) we eliminate I_n and I_p and obtain the equations

$$(b+1)I_{ps}\xi^2 = I - I_r(\xi \cosh \omega - \zeta) \quad (70)$$

$$(b+1)RI_{ps}\xi^2 = bI + I_r(\xi - \zeta \cosh \omega)$$

where I_r is a (particle) current density

$$I_r = \frac{2D_n n_i}{L \sinh \omega} \quad (71)$$

In principle we could solve (70) for ξ and ζ as functions of I with R and ω as parameters; this would determine βV through (58) and (65) and complete the problem. First we shall rewrite these equations in terms of γ as in the analysis of the second section.

If we eliminate I from equations (24) we get

$$\begin{aligned} bI_{ps} + \frac{I_r}{\xi} \frac{b \cosh \omega + 1}{b + 1} &= RI_{ps}\gamma^2 \\ &+ \gamma \frac{I_r}{\xi} \frac{\cosh \omega + b}{b + 1} \end{aligned} \quad (72)$$

which can be solved for ξ

$$\xi = \frac{I_r}{I_{ps}} \frac{\cosh \omega + b}{b + 1} \frac{\gamma_0 - \gamma}{R\gamma^2 - b} \quad (73)$$

where

$$\gamma_0 = \frac{b \cosh \omega + 1}{\cosh \omega + b} \quad (74)$$

Substituting (73) into (70) gives the equation satisfied by γ

$$\left(\gamma - \frac{R\gamma^2 \cosh \omega + 1}{R\gamma^2 + \cosh \omega} \right) (\gamma - \gamma_0) = \frac{I}{I_{00}} \frac{[(\gamma/\gamma_\infty)^2 - 1]^2}{R\gamma^2 + \cosh \omega} \quad (75)$$

where I_{00} is a characteristic (particle) current density

$$I_{00} = I_o \left[\frac{\omega}{\sinh \omega} \right]^2 \frac{\cosh \omega + b}{b + 1} \quad (76)$$

Now the solution of (75) has two branches which as $I \rightarrow 0$ approach values given by

$$\begin{aligned} \text{a)} \quad \gamma &\rightarrow \gamma_0 \\ \text{b)} \quad \gamma &\rightarrow \frac{R\gamma^2 \cosh \omega + 1}{R\gamma^2 + \cosh \omega} \end{aligned} \quad (77)$$

As I increases the first branch remains positive and approaches γ_∞ as $I \rightarrow \infty$. The second branch becomes negative and approaches $-\gamma_\infty$. Therefore, we choose that branch which satisfies

$$\begin{aligned} \gamma(0) &= \gamma_0 = \frac{b \cosh \omega + 1}{b + 1} \\ \gamma(\infty) &= \gamma_\infty = (b/R)^{1/2} \end{aligned} \quad (78)$$

$$\gamma > 0$$

On this branch γ always lies between γ_0 and γ_∞ , and γ never approaches the quantity in (77b). Therefore we replace $R\gamma^2$ by b (as if $\gamma = \gamma_\infty$) in

the first factor on the left of (75), and obtain the simpler form

$$\gamma - \gamma_0 = -\sqrt{\frac{I}{I_{00}}} \frac{(\gamma/\gamma_\infty)^2 - 1}{\sqrt{R\gamma^2 + \cosh \omega}} \quad (79)$$

which is the generalization of (31).

The drop βV_P in the middle region given by (65) can be written

$$\beta V_P = \frac{b-1}{b+1} \ln \gamma + \frac{2}{b+1} \sqrt{\frac{I}{I_{00}}} \sqrt{R\gamma^2 + \cosh \omega} F_\omega(\gamma) \quad (80)$$

where $F_\omega(\gamma)$ comes from $\int dx/n$ and is defined

$$\begin{aligned} F_\omega(\gamma) &= \int_0^1 \frac{\omega du}{\gamma \sinh [\omega(1-u)] + \sinh [\omega u]} \\ &= \frac{\ln \left| \frac{1+Q}{1+Q} \right| - \ln \left| \frac{1+e^\omega Q}{1-e^\omega Q} \right|}{\sqrt{1-2\gamma \cosh \omega + \gamma^2}} \end{aligned} \quad (81)$$

or

$$2 \frac{\tan^{-1} e^\omega Q - \tan^{-1} Q}{\sqrt{2\gamma \cosh \omega - 1 - \gamma^2}}$$

The first form applies when $\gamma > e^\omega$, or $\gamma < e^{-\omega}$, and the second applies when $e^{-\omega} < \gamma < e^\omega$, and Q is the quantity

$$Q = \frac{1 - \gamma e^{-\omega}}{\sqrt{|1-2\gamma \cosh \omega + \gamma^2|}} \quad (82)$$

It can readily be shown that when $\omega \rightarrow 0$

$$F_0(\gamma) = \frac{\ln \gamma}{\gamma - 1} \quad (83)$$

Thus when $\omega = 0$ (80) reduces to

$$\begin{aligned} \beta V_P &\rightarrow \frac{\ln \gamma}{\gamma - 1} \frac{b-1}{b+2} (\gamma - 1) + \frac{2}{b+1} \sqrt{\frac{I}{I_0}} \sqrt{R\gamma^2 + 1} \\ &= \left[\frac{\ln \gamma}{\gamma - 1} \frac{(\gamma/\gamma_\infty)^2 - 1}{R\gamma^2 + 1} \right] \sqrt{\frac{I}{I_0}} \end{aligned} \quad (84)$$

which is identical with (41). It is also clear that (79) reduces to (31) as the recombination goes to zero. Finally we write from (58)

$$\beta(V_0 + V_w) = \ln \gamma \xi^2 = \ln \frac{I}{I_{ps}} + \ln \frac{\gamma}{R\gamma^2 + \cosh \omega} \quad (85)$$

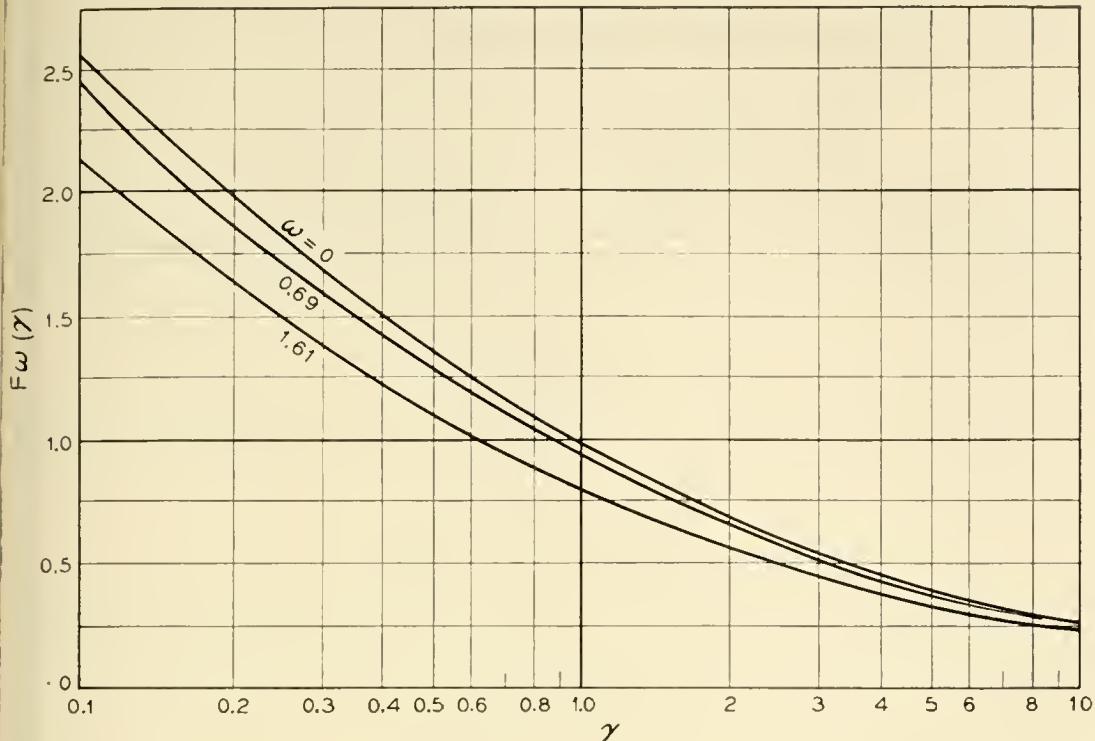


Fig. 6 — The function $F_\omega(\gamma)$ of equation (81) for several values of ω .

which reduces to (36) when $\omega = 0$. Thus the whole theory reduces correctly in the case $\omega = 0$.

The function $F_\omega(\gamma)$ is plotted in Fig. 6 for several values of ω including $\omega = 0$. The expansion of $F_\omega(\gamma)$ to order ω^2 is

$$\begin{aligned}
 F_\omega(\gamma) &= \frac{\ln \gamma}{\gamma - 1} - \frac{\omega^2}{4} f(\gamma) \\
 f(\gamma) &= \frac{(\gamma + 1) - 2\gamma \frac{\ln \gamma}{\gamma - 1}}{(\gamma - 1)^2} \\
 &= 1 - 2\gamma \ln \frac{1}{\gamma} + \dots
 \end{aligned} \tag{86}$$

Our next step is to eliminate from (80) and (85) unimportant dependencies on I which would be difficult or impossible to detect experimentally. If in (85) we let $\gamma = 1$, $\cosh \omega = 1$ we get

$$\beta(V_0 + V_w) = \ln \frac{I}{I_{ns} + I_{ps}} \tag{87}$$

In (80) we drop the first term (as if $\gamma = 1$) and in the second term we

put $R\gamma^2 = b$ (as if $\gamma = \gamma_\infty$) and $F_\omega(\gamma) = F_\omega(1)$,

$$\beta V_p \approx \frac{2}{b+1} \sqrt{\frac{I}{I_{00}}} \sqrt{b + \cosh \omega} F_\omega(1) \quad (88)$$

In this way we retain the correct form of dependence on ω , but throw out the dependence on I that comes from $\gamma(I)$. It can be shown from (81) that

$$\begin{aligned} F_\omega(1) &= \frac{\tan^{-1} \left(\sinh \frac{\omega}{2} \right)}{\sinh \frac{\omega}{2}} \\ &= 1 - \frac{\omega^2}{12} + \frac{\omega^4}{180} + \dots \end{aligned} \quad (89)$$

Thus we define the characteristic (particle) current density of the device

$$\begin{aligned} I_0(\omega) &= \frac{(b+1)I_{00}}{(b + \cosh \omega)F_\omega(1)^2} \\ &= I_0 \left[\frac{\omega}{F_\omega(1) \sinh \omega} \right]^2 = I_0 g(\omega) \end{aligned} \quad (90)$$

and (88) can be written

$$\beta V_p \approx \frac{2}{\sqrt{b+1}} \sqrt{\frac{I}{I_0(\omega)}} \quad (91)$$

This formula corresponds to (42) with $C = 2/\sqrt{b+1}$. In the spirit of the present theory the exact value of this constant is not important, so we may replace $2/\sqrt{b+1}$ in (91) by C . Then the sum of (87) and (91) gives the total applied bias (66).

Non Linear Recombination

In this section we shall consider the forward characteristic of a PIN diode in which the current densities at the contacts obey the law

$$\begin{aligned} I_n &= I_{ns} \left(\frac{n_{00}}{n_{p^+}} \right)^a \\ I_p &= I_{ps} \left(\frac{p_{w0}}{p_{N^+}} \right)^a \end{aligned} \quad (92)$$

where I_{ns} and I_{ps} are characteristic of the device and a is a number be-

tween 0 and 1. We see that (30) must be replaced by

$$\begin{aligned} I_n/I_p &= R\gamma^{2a} \\ I_p &= \frac{I}{I + R\gamma^{2a}} \quad I_n = \frac{R\gamma^{2a}I}{1 + R\gamma^{2a}} \end{aligned} \quad (93)$$

and (23) must be replaced by

$$\begin{aligned} n_0 &= n_i(I_n/I_{ns})^{1/2a} \\ n_w &= n_i(I_p/I_{ps})^{1/2a} \end{aligned} \quad (94)$$

The equation for γ is now

$$\gamma = 1 - \left(\frac{I}{I_1}\right)^{1-(1/2a)} \frac{(\gamma/\gamma_\infty')^{2a} - 1}{[1 + b(\gamma/\gamma_\infty')^{2a}]^{1-(1/2a)}} \quad (95)$$

where $\gamma_\infty' = (b/R)^{1/2a}$ and

$$I_1 = I_0(I_{ps}/I_0)^{(a-1/2a-1)} \quad (96)$$

is a characteristic (particle) current density of the device. We now obtain βV_p from (26)

$$\beta V_p \approx C'(I/I_1)^{1-(1/2a)} \quad (97)$$

where C' is a slowly varying function

$$C' = \frac{(\gamma/\gamma_\infty')^{2a} + 1}{[1 + b(\gamma/\gamma_\infty')^{2a}]^{1-(1/2a)}} \frac{\ln \gamma}{\gamma - 1} \quad (98)$$

similar to the coefficient in brackets in (41). From (21) and (94) we get

$$\beta(V_0 + V_w) = \frac{1}{2a} \ln \frac{I_n I_p}{I_{ns} I_{ps}} \quad (99)$$

If now $\gamma \sim 1$ we get

$$\frac{I}{I_{ns} + I_{ps}} = e^{\alpha \beta (V_0 + V_w)} \quad (100)$$

This shows how we must choose a to agree with the low current characteristic. On the basis of experience with silicon diodes we would choose $a \sim 0.6$, which would give

$$\beta V_p \sim C'(I/I_1)^{0.17} \quad (101)$$

The characteristic current density would be

$$eI_1 \sim 200 \times (I_{ps}/I_0)^{-2} \text{amp/cm}^2 \text{ in Si} \quad (102)$$

The value to use for I_{ps} is very uncertain, but it certainly is much less than I_0 , so $I_1 \gg I_0$. Thus we would not expect to observe βV_P , and the characteristic should have the form

$$I \sim I_s e^{a\beta} \quad (103)$$

up to the highest attainable currents.

We have shown in this section how the law of recombination in the contacts affects the dependence of V_P upon I . In particular if $a = \frac{1}{2}$ there is no dependence of V_P upon I , which means that the conductivity due to injection increases just as rapidly as the current. We may conclude from (97) that the smaller the value of a the more effective is conductivity modulation in keeping down the drop V_P in the middle region.

Discussion

We have considered the *PIN* structure of Fig. 1 having typical parameters given in (3). We find that the presence of the middle region causes no significant deviation in the voltage-current characteristic from that of a simple *PN* diode until very high current densities are reached, of the order of 200 amp/cm² in silicon. In particular the middle region is not responsible for an anomalous slope in the plot of V versus $\log I$. We find that recombination in the middle region can be accounted for by replacing the characteristic current density eI_0 of the device with $eI_0g(w/L)$ where $g(w/L) < 1$ is shown in Fig. 5. Thus qualitatively there is no change in the form of the voltage-current characteristic due to recombination in the middle region, although the effect of $g(w/L)$ is to make the voltage drop somewhat higher than if recombination were absent.

We have suggested that the anomalous slope of V versus $\log I$ usually observed in silicon diodes might be due to non-linear recombination. If the recombination obeys a power law chosen to give a typical (anomalous) $V - I$ characteristic for a *PN* diode, we have shown that the *PIN* diode should manifest the same characteristic up to extremely large current densities many times eI_0 . Thus the drop across the middle region should be even more negligible with non-linear than with linear recombination.

I am pleased to acknowledge my great benefit from discussions with M. B. Prince and I. M. Ross.

A Laboratory Model Magnetic Drum Translator for Toll Switching Offices

By F. G. BUHRENDORF, H. A. HENNING and O. J. MURPHY

(Manuscript received January 24, 1956)

A laboratory model magnetic drum translator, capable of serving as a one-to-one alternative to the card translator, has been built to study the problems arising from the prospective use of microsecond pulse apparatus in a telephone office environment. Electron tube amplifiers and germanium diode logic circuits supplement the drum information storage unit to provide the functional operations required. Results of preliminary laboratory tests indicate the feasibility of equipment of this kind for telephone switching control.

INTRODUCTION

The magnetic drum is one of the most widely used of the modern large-capacity digital-data storage devices. It is used as a memory unit in many of the present-day large-scale digital computers and in other applications such as inventory control of airline ticket reservations and traffic control of airplanes in flight. Two of the properties of drums as storage media have been considered particularly advantageous. One is the capacity to store up to several hundred thousand bits of information in a compact space at a low cost per bit; the other is the ability to keep the information in an easily alterable but nonvolatile form unaffected by power failure or other interruptions of operation. In terms of the speed with which information may be stored or recovered, drum memories fall near the middle of the present-day spectrum; they are very much faster than punched paper tape or groups of telephone relays but are considerably slower than cathode-ray tube or ferromagnetic-core storage devices. All of the information stored on a drum may be read out during the course of one complete revolution and, similarly, new information may be entered anywhere in the storage space within the time of one revolution; thus the access time is ordinarily of the order of a few tens of milliseconds.

It has already been pointed out¹ that automatic telephone switching

offices bear a generic resemblance to digital computers and it is therefore not surprising that the magnetic drum has engaged the attention of telephone engineers, since the speed and flexibility of such a device offers much promise in connection with forward-looking telephone office design. One system has already been described^{2, 3} involving the use of magnetic drums for telephone switching control applications in an entirely new form of telephone office; it is the purpose of this article to describe another application of less complexity which could function in cooperation with equipment in existing telephone offices.

The standards of reliability and ruggedness which must be met by any equipment proposed for Bell System use are in some respects a good deal higher than those imposed on other commercial systems such as digital computers. Thus when a new type of apparatus such as a magnetic drum and its associated electronic components is considered for a telephone job, it is necessary to determine whether the apparatus is capable of being designed to meet these stringent requirements. This was judged to be the most important objective of the undertaking about to be described, and it strongly influenced the choice of experimental application for the drum.

The program which the designers set for themselves to determine the possible suitability of the magnetic drum type of equipment might be summarized as follows:

- (1) Choose an existing telephone application in which a magnetic drum system can receive a satisfactory work-out without disordering the system.
- (2) Design a magnetic drum system to work cooperatively with existing office equipment, using existing power facilities. Assume that the design is aimed at practical application so that due regard is given to operating economies, and protection against power failures.
- (3) Construct a full-scale model following the design, and test the model in the chosen environment long enough to determine the failure rate and the reasons for each failure.
- (4) Evaluate the results in order to determine the sphere of usefulness, and the proper design philosophy for applying magnetic drum systems of any kind in existing telephone offices.

One telephone switching application which meets the qualifications of (1) above exists in the new No. 4A toll switching offices. Here, due to the demands of nationwide dialing, a large-scale translation function is required to convert destination codes into information which will properly route each call. The volume of information which must be stored for translation purposes, and the relatively rapid access desired, fall close to the optimum parameter values of magnetic drum systems. The action

takes place in cooperation with crossbar and other relay-type switching equipment typical of the present-day telephone office, thus providing an environment suitable for observing the behavior of fast pulse circuits in the presence of electrical disturbances. Finally, there exists a relatively new piece of apparatus which now performs the translation function, namely the card translator. Thus, if an exact one-to-one alternative for the card translator were constructed employing a magnetic drum, full advantage could be taken of the testing procedures already developed and a comparison could be made against a norm of performance; furthermore, a field trial would be possible, if desired, with a minimum of interference with normal operation of the telephone plant.

It was decided, therefore, to build a full-scale magnetic drum translator which could substitute for a card translator in order to obtain laboratory experience with apparatus of this type and to determine its adaptability to telephone standards and practices. The completed equipment is shown in Fig. 1. The equipment on the one frame illustrated is the equivalent in function and capacity of one card translator with its associated table. This magnetic drum apparatus is not aimed at replacing the card translator, which is a well-engineered device known to give satisfactory service in day-to-day operation. For evaluation purposes in this article, however, it is assumed to be competing with the card translator.

The following sections describe the design features and operating details of the translator which was constructed. A brief description of the card translator and that portion of the 4A office in which the drum translator must function has been included to provide the necessary background for the description. It will become evident that the requirement of interchangeability which necessitates a one-to-one equivalence with the card translator has imposed on the drum translator a number of restrictions which are not inherent in it. These tend to prevent full exploitation of the speed and code advantages which might be realized with the drum. Furthermore, the rapidity with which all of the information on the drum is presented on a continuous read-out basis would permit a type of centralized operation which will be touched on briefly and which would seem to offer apparatus economies not attained in the test model. None of these factors, however, impairs the usefulness of conclusions which may be drawn from test results concerning reliability.

SURVEY OF MAGNETIC RECORDING PRINCIPLES EMPLOYED IN THE TRANSLATOR

All magnetic drums have certain features in common: they consist of a means of moving a thin shell of magnetically-hard material rapidly

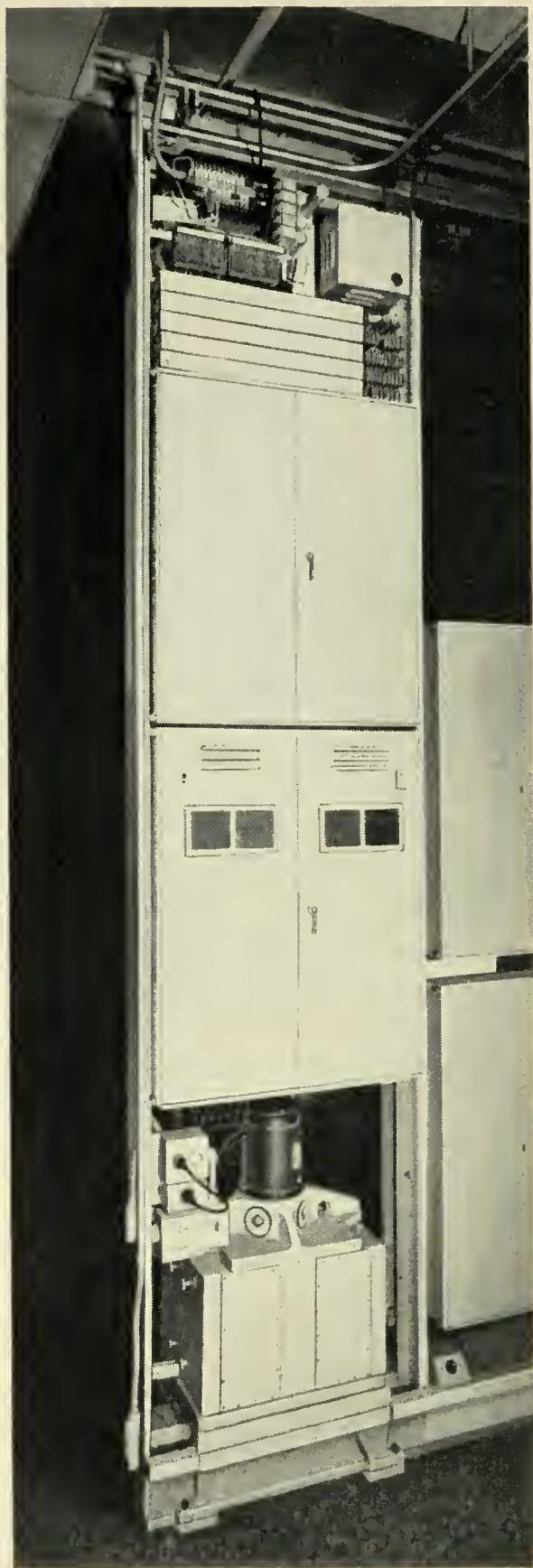


Fig. 1 — Magnetic drum translator, laboratory installation.

past one or more heads used for writing or reading digital data. Usually, as in the translator, the same head is used for both functions. In most drum-system designs the pole-tips of the heads are close to the recording surface but do not touch it, and the heads themselves bear a resemblance to those used in conventional magnetic sound recording, giving therefore, a "longitudinal" polarization to the medium as sketched diagrammatically in Fig. 2. There is very little further resemblance to sound recording, since digital information is stored in a binary or two-valued code which, on the translator drum, is represented by the two possible polarities of saturation of the magnetic medium. To one of these polarities is assigned the code value "0," and this condition prevails except where the opposite polarity is inserted to represent the code value "1."

It should be mentioned that several other systems have been devised which employ the two directions of saturation, sometimes accompanied by a general background of magnetic neutrality, to effect a greater concentration of digital information than that used in the translator. Systems other than the one chosen for this application were, for the most part, considered to be less reliable.

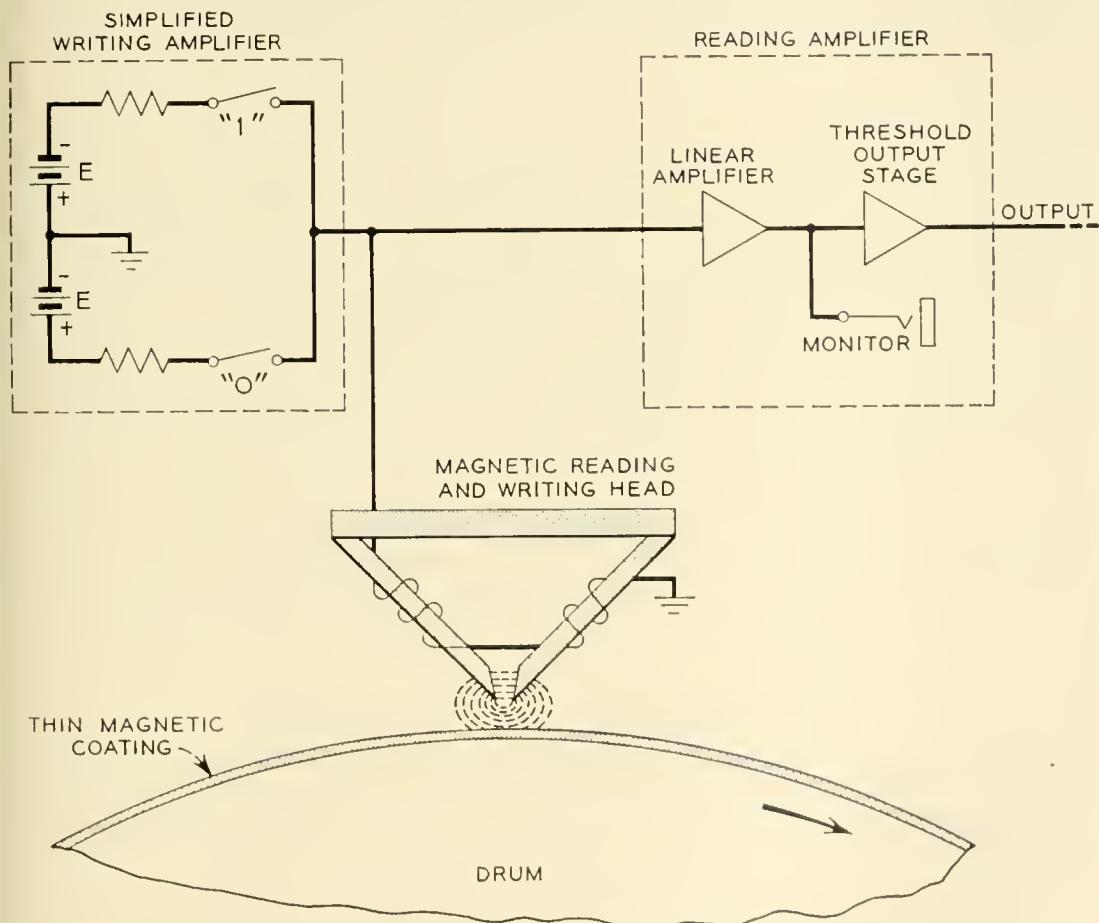


Fig. 2 — Simplified diagram of magnetic drum digital data storage system.

In order to facilitate an understanding of the action of the translator as a whole, a simplified account of the magnetic recording and reproducing process will now be given.

Magnetic Drum Geography

The circumferential strip of the drum surface which moves under the pole-tips of any magnetic head is commonly known as a *track*. On each track will be written magnetic perturbations or spots symbolizing "1's." It is essential that these spots be precisely located so that they may be readily removed or "altered." For this purpose a synchronizing track or some equivalent distribution of equally spaced identifying marks associated with the drum is provided. With the aid of the electronic circuits, the magnetic spots are restricted to a modular spacing defined by the synchronizing marks, and this module is spoken of as a "slot." On the drum surface, each intersection of track and slot is known as a "cell" and a cell may contain only one magnetic mark and therefore only one bit of information. As a matter of economies, the cell density should be as great as possible. The density which may be attained is determined by the degree of interference which can be tolerated among neighboring cells.

Writing Operations

The first step in preparing the drum to receive a recording is to uniformly magnetize the tracks to saturation in the polarity arbitrarily chosen to represent the code-value "0." This is a preconditioning operation required only when a drum is newly placed in service. Referring to Fig. 2, this may be done, for the typical head and track shown, by closing the switch marked "0" for the duration of at least one complete revolution of the drum. Enough current must flow through the windings of the head to establish the magnitude of fringing flux, from the pole-tips, required to saturate the thin magnetic coating. In the case of the translator drum, the coating is about $\frac{1}{3}$ milli-inch thick; the clearance between pole-tips and recording surface is about 2 milli-inches; the inter-pole gap is also about 2 milli-inches at the tips, and about 20 ampere-turns of energization are required.

With the track thus preconditioned, there is virtually no output voltage from the head since the magnetization is essentially uniform and there is no changing flux threading the head to induce a voltage in the windings.

Whenever a "1" is to be written, a pulse of current from an electronic writing amplifier (indicated, for convenience, on Fig. 2 as a switch) is

caused to flow through the windings of the head in a direction opposite to that taken by the preconditioning current. This pulse lasts for only two or three microseconds, and movement of the drum surface is negligibly small while the current persists. The peak value of the current pulse is sufficient to magnetize to saturation in the opposite direction that portion of the track which lies directly under the pole-tips at that instant. Areas of the track far-removed in each direction from the pole-tips of the head are, of course, unaffected by this operation, and remain at saturation in the original polarity. A region of transition in magnetization therefore extends in each direction along the track from the area directly under the pole-tips.

Fig. 3 illustrates some of the wave forms resulting from writing into and reading from four adjacent cells on one track of the drum. Line A shows the pulses of writing current which were applied to the windings on the head. These were caused to appear at precisely spaced distances along the track by the combined operation of the synchronizing system and an "administration" circuit. In cells 1 and 3 the writing current polarity is chosen so as to write "1's." Cell 2 remains in its original preconditioned state. In cell 4 a "1" was previously written but is now altered to a "0" by a writing current pulse of the same polarity as that chosen for the preconditioning operation.

Line B in Fig. 3 illustrates the resultant magnetic state of the drum surface as viewed by the reading head. The polarization portrayed as resulting from writing a "1" is a bell-shaped curve. When a "1" is selectively altered to a "0" the area of track directly under the pole-tips will be carried to saturation in the original preconditioned polarity. The whole cell area, however, cannot be affected so strongly, owing to the hysteresis properties of the coating material, and there will remain traces of the "1" type of magnetization near the cell edges, as indicated by the solid line in cell 4.

There is no difficulty in rewriting a "1" in a cell which has been subjected to the above described treatment. The procedure is that outlined for the original writing of a "1" and the results are practically indistinguishable from those obtained by writing in a virgin cell.

Reading Operations

On subsequent revolutions of the drum, the passage, under the pole-tips, of the magnetic irregularities created by writing "1's" will induce a change of flux through the windings of the head. The change is, of course, a function of distance along the drum surface but since the drum is rotating continuously at a substantially uniform speed the change

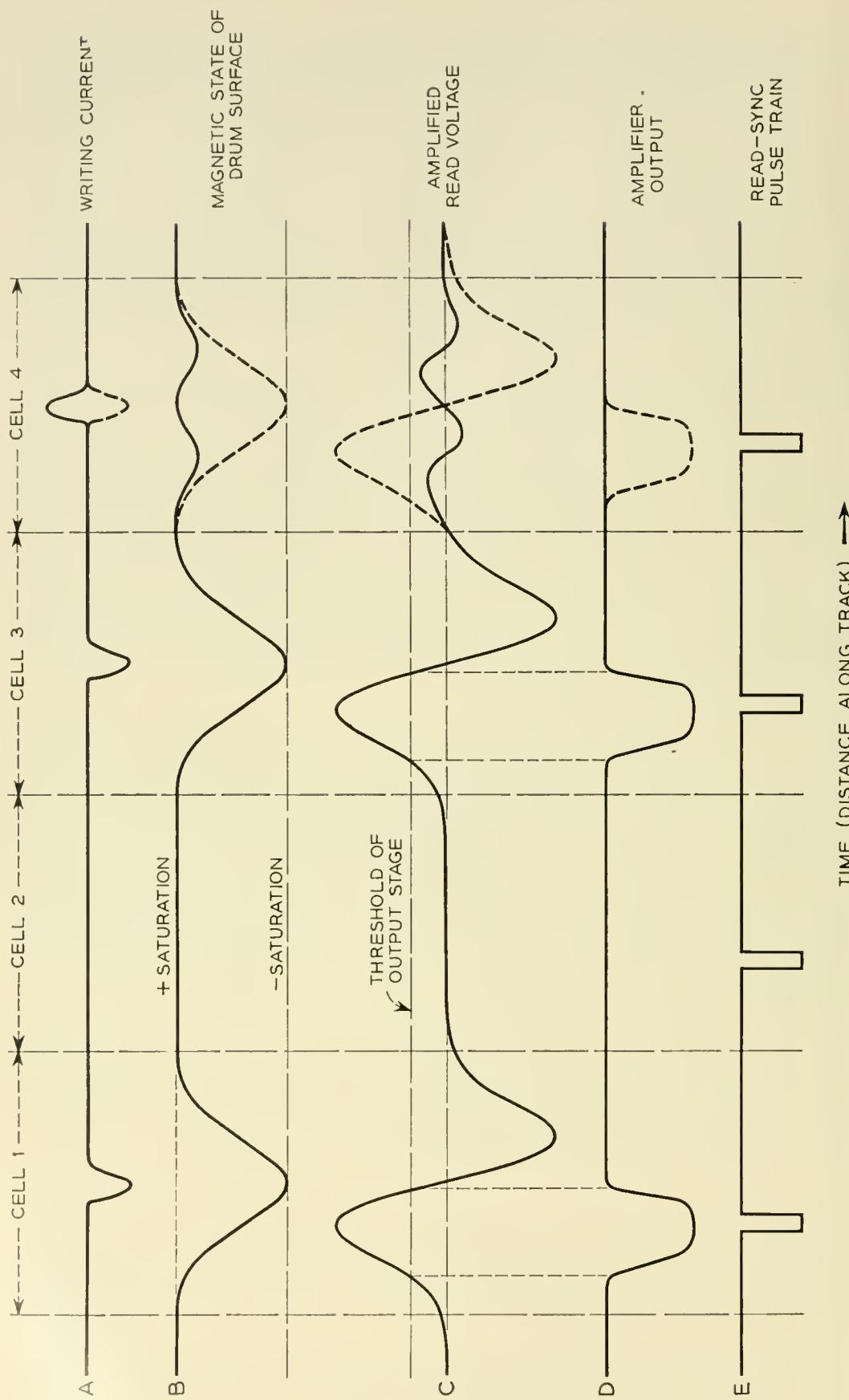


Fig. 3.—Magnetic writing and reading waveforms.

may also be represented as a function of time. This time-rate-of-change of flux within the coils of the head generates a voltage which is of the order of 50 millivolts peak-to-peak in the case of the translator. This voltage, after amplification, appears as shown in line C of Fig. 3. The trace shown is that which appears at the "linear output" monitor jack of a translator reading amplifier, and includes a phase inversion, characteristic of a three stage amplifier. Such a curve is readily recognized as being quite similar in shape to the first derivative of the normal error-function and hence we may infer that the magnetic condition of the drum surface, at least as interpreted by the head, may be portrayed by a bell-shaped curve, previously mentioned, similar to the error-function itself.

The residual magnetic irregularity pictured in cell 4 resulting from writing a "0" over a "1" will induce a voltage in the winding of the head having a different amplitude and wave shape from that occasioned by reading a "1." It is sketched out approximately to scale in Fig. 3 and is seen to be a smaller twinned-version of the "1" signal. Its amplitude ordinarily lies in the range of $\frac{1}{10}$ to $\frac{1}{3}$ of that of the "1" signal, and for about the middle third of the cell its instantaneous polarity is opposite to that which a "1" signal would have. These facts suggest at least two means of discriminating between the voltage signals obtained for the two code values: (a) on the basis of amplitude difference, and (b) on the basis of instantaneous polarity difference determined or sampled within a particular epoch in each cell.

The method adopted for the translator is that of simple amplitude threshold. The threshold value indicated by the dotted line in Fig. 3, is set so that the strongest of the residual signal outputs never exceeds it while, at the same time, the greatest possible proportion of the positive-going lobe of a "1" signal is allowed to produce an output. The threshold output stage of the amplifier is also arranged for limiting and this has the effect of blunting the peaks of the applied signals. The over-all result of these actions is shown by the shape of the signals in line D of Fig. 3.

Cell packing may be of major economic importance in a large installation. The general effect of making recordings closer and closer together is that the presence or absence of one of the recordings in a series has an increasing influence on the size and shape of the signals reproduced from its neighbors on either side. In the translator, the cells are spaced 20 milli-inches center-to-center along the track and the influence of action in one cell on the amplitude of reproduction from neighboring cells is never more than about 10 per cent. The trace of line C, Fig. 3, is drawn for this cell spacing and shows a slight inflection at the transition between the output voltage occasioned by reading cell 3, and the voltage obtained

from the "1" which was originally written in cell 4. In many applications a much larger "influence factor" may be tolerable, but this usually requires greater elaboration of the signal detecting devices. The cell size is also influenced by physical constants such as design of the head, properties of the medium, and dimensional clearances. A discussion of such factors is outside the scope of this paper but it is not unreasonable to hope for an improvement of two-to-one in packing factor in future designs.

Reading Synchronization

The magnetic drum used for the translator provides 80 tracks. About sixteen microseconds is required for each cell in a track to pass under its head. Information occupying the same slot on the drum (so-called because of its obvious relationship to the term "time-slot" commonly used in the digital computer field) is presented at the various heads essentially, but not exactly, simultaneously. Departure from exact simultaneity is occasioned by small variations in the shapes and amplitudes of the output waves shown typically as line C in Fig. 3, and by small time-variations occurring in the writing process, as applied to the various tracks.

To achieve exact simultaneity, as required for certain subsequent operations of the translator circuitry, narrow "Read Synchronizing" pulses are produced by the synchronizing circuit previously mentioned. These pulses are located, within the time boundaries of the cells, so that they fall approximately at the center of the broad output pulses from the reading amplifiers and thus permit the latter to be sampled. This relationship is indicated in lines D and E of Fig. 3. Similar pulses, slightly displaced in time, are used to control the writing operations, and are designated "Write Synchronizing" pulses. The necessity for the time-shift is apparent from an examination of lines A and E of Fig. 3.

This condensed explanation of the technology of magnetic drum digital data storage devices, particularly as applied to the translator drum, should serve as sufficient background for the description of the translator wherein the drum is but one part of a large ensemble of apparatus.

THE JOB WHICH THE CARD TRANSLATOR NOW DOES

It will be advantageous to examine very briefly the card translator and its functions in the No. 4A toll switching system so that the analogous operation of the magnetic drum equivalent may be more readily explained. A more detailed description is given in Reference 4.

The demands of nationwide toll dialing require a very extensive rep-

ertoire of translations between destination codes and routing instructions, and it must be possible to change the routing instructions with ease. The card translator fulfills these requirements. Each individual translation item is contained on a metallic card; the output code of routing instructions is in the form of selectively enlarged perforations in the perforated field of the card, arranged so as to be read by photoelectric means, and the input code, which identifies the card for purposes of selection, appears in the form of tabs projecting downward from the bottom edge. Each card is capable of holding a total of 154 bits of information, input and output, and somewhat over 1,000 cards are stacked in a bin in each card translator mechanism.

It is possible to classify the elements of any translator into three broad categories: the memory unit, the translation selecting unit, and the translation delivery unit. In the card translator the memory unit is, of course, the group of cards; the translation selecting unit consists of code bars, electro-mechanically actuated, for displacing a selected card sufficiently so that it may be "read." It also contains a network of relays which perform the function of checking the authenticity of the input codes applied to the code bars. The translation delivery unit consists, in the main, of a number of output channels, each originating with a light beam for probing one of the code elements (a bit of output information) on the card. Each output channel contains a photo-transistor, a transistor amplifier, a cold cathode gas tube circuit which has been designated a "channel output detector" and a register relay. The register relays perform work functions and therefore are located separately from the translator; some are in the decoders, others in the markers.

In the 4A office, the card translator is one of several items of common control equipment which cooperate to establish the talking connections. Other items are the sender, the decoder, and the marker. The sender receives and registers and subsequently transmits the decimal digits of the called designation; the decoder receives the code digits (from 3 to 6 in number) from the sender and submits them to the translator for conversion into information needed for the proper routing of the call; and the marker selects an outgoing trunk and establishes a transmission path by operating the crossbar switches. Since this common control equipment is associated with any one call for only the short interval necessary to establish the talking-circuit connection, its speed of operation is a matter of considerable importance.

It is obvious that the decoder is the intermediary between the translator and the remainder of the office. Each decoder, of which there are a maximum of 18 in a large office, has exclusively associated with itself a

card translator mechanism; each of these mechanisms contains an identical repertory of translations. Each decoder also has available, through connectors, a common pool of translators containing a large quantity of less-often used information. In order to better understand the duties that a magnetic drum translator must be expected to perform it will now be convenient to follow, in a highly abbreviated manner, a typical operation of the decoder and its associated card translator.

The first translation on an incoming call is performed using the first three decimal digits accumulated by a sender. As soon as three digits are available the sender connects to a decoder which immediately signals its individual translator to perform certain mechanical chores in preparation for selecting a card. There are several sequencing signals between the decoder and translator during the complete cycle of a translation (several of these signals must be synthesized by the drum translator); acting on one of these signals from the translator, the decoder passes the input code from the sender, adding certain supplemental information of its own.

The three decimal digits of the input code are in checkable combinations of two leads energized in each of three groups of five leads connected to the translator. The supplementary information supplied by the decoder is in a similar checkable combination on six leads. None of the remaining leads in the total of 38 is energized, since the translation being described involves only three code digits.

In the translator, the input code actuates the card selecting mechanism and also operates relays whose contacts are wired with a checking network which confirms that the input code, and the responsive operation of the code bars, is an authentic combination. This is done by establishing a path to operate a "code bar check" relay, CBK. (This relay retains the same identity in the magnetic drum translator.)

Acting upon the authenticity check, the card translator proceeds to select a card, and signals the decoder to begin timing for a possible non-appearance. When the card is in a position to be read, the decoder is signaled on two "index" channels, IND. The decoder now "reads" the card by applying 130 volt battery to the coils of its register relays; the required relays operate through the ionized cold-cathode gas tubes in the translator, and lock up, extinguishing the gas tubes.

The first card dropped may provide information sufficient for completing the connection; in this circumstance the decoder will then call in a marker. The first card, however, may specify that more digits are required and the decoder will so instruct the sender. The sender, unless it already has the necessary digits, is then dismissed by the decoder which also instructs the translator to restore itself to normal.

Six-digit translations are obtained in a manner similar to that described above except that the checking network on the relays is switched to check for six rather than three digits. In some instances the decoder must refer to one of the translators in the common pool of "foreign area translators" in order to obtain the required information. Frequently, several different cards must be dropped successively before a route is finally established for the outgoing call.

With the above description as a background, we may proceed to discuss the magnetic drum translator.

THE ANALOGOUS FUNCTIONS OF THE MAGNETIC DRUM TRANSLATOR

The magnetic drum translator is essentially a device which performs a translation by making a selection from a recurrent pattern of electrical pulses generated by a magnetic drum unit. A schematic diagram of the magnetic drum translator, as arranged for direct substitution for a card translator, is shown in Fig. 4. In this diagram, the system is divided into three principal functional components: (a) the drum memory assembly which produces (from the outputs of 80 reading amplifiers and a timing unit) a repetitive pattern of electrical pulses representing all the translations on the drum, both input codes and corresponding output codes; (b) the translation selecting unit which reads that portion of the pulse pattern representing input codes and acts to identify the unique code group which matches the incoming information from the decoder; (c) the translation delivery unit which, under control of the translation selecting unit, gates-out the particular pulses of the corresponding output code from the continuous stream of microsecond pulses, and converts them into signals capable of operating the register relays in the decoder.

To maintain direct interchangeability, two items of apparatus were adopted virtually without change from the card translator. These are the CODE CHECK RELAYS which accept and check input information, and the CHANNEL OUTPUT DETECTORS comprising cold-cathode gas tubes and associated transformers. This allows input and output terminal facilities to the decoder to be the same for both translators.

It should be noted that the magnetic drum memory assembly differs significantly in one functional respect from the binful of cards in the card translator. When a selected card is being read by the photo-electric cells in the output channels, no other cards are available. In the drum translator, all translations are continuously available and if a number of translation selecting and translation delivery circuits are employed, all may obtain translations from a common drum memory assembly at the same time without interference. This feature could not be demonstrated in the

test set-up as planned, but it would have been incorporated in any test which included more than one decoder in an office. In such an arrangement, the various units illustrated in Fig. 2, except the drum memory assembly, would be furnished to each decoder. One drum memory assembly (and an emergency standby) would supply the pattern of electrical

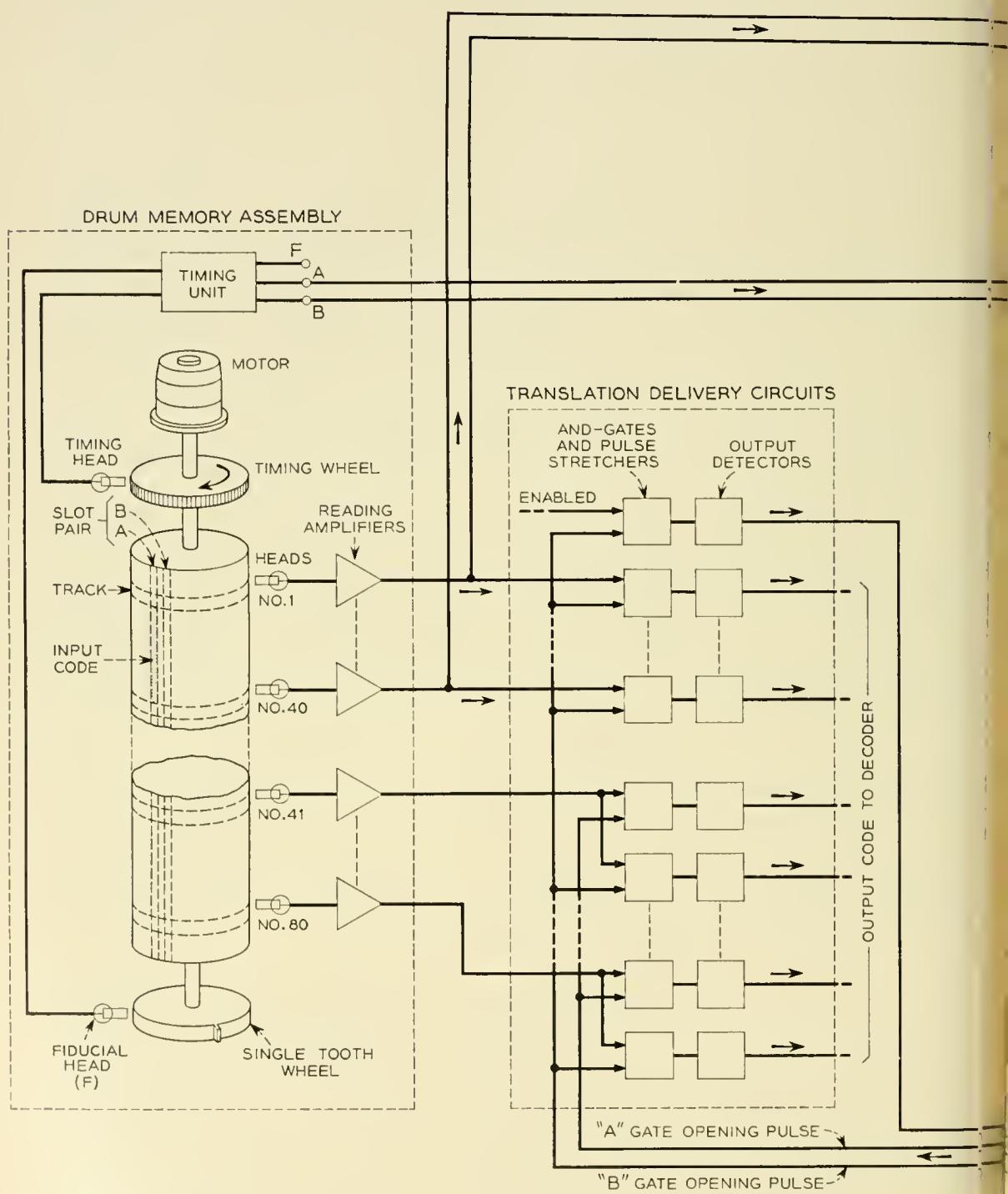
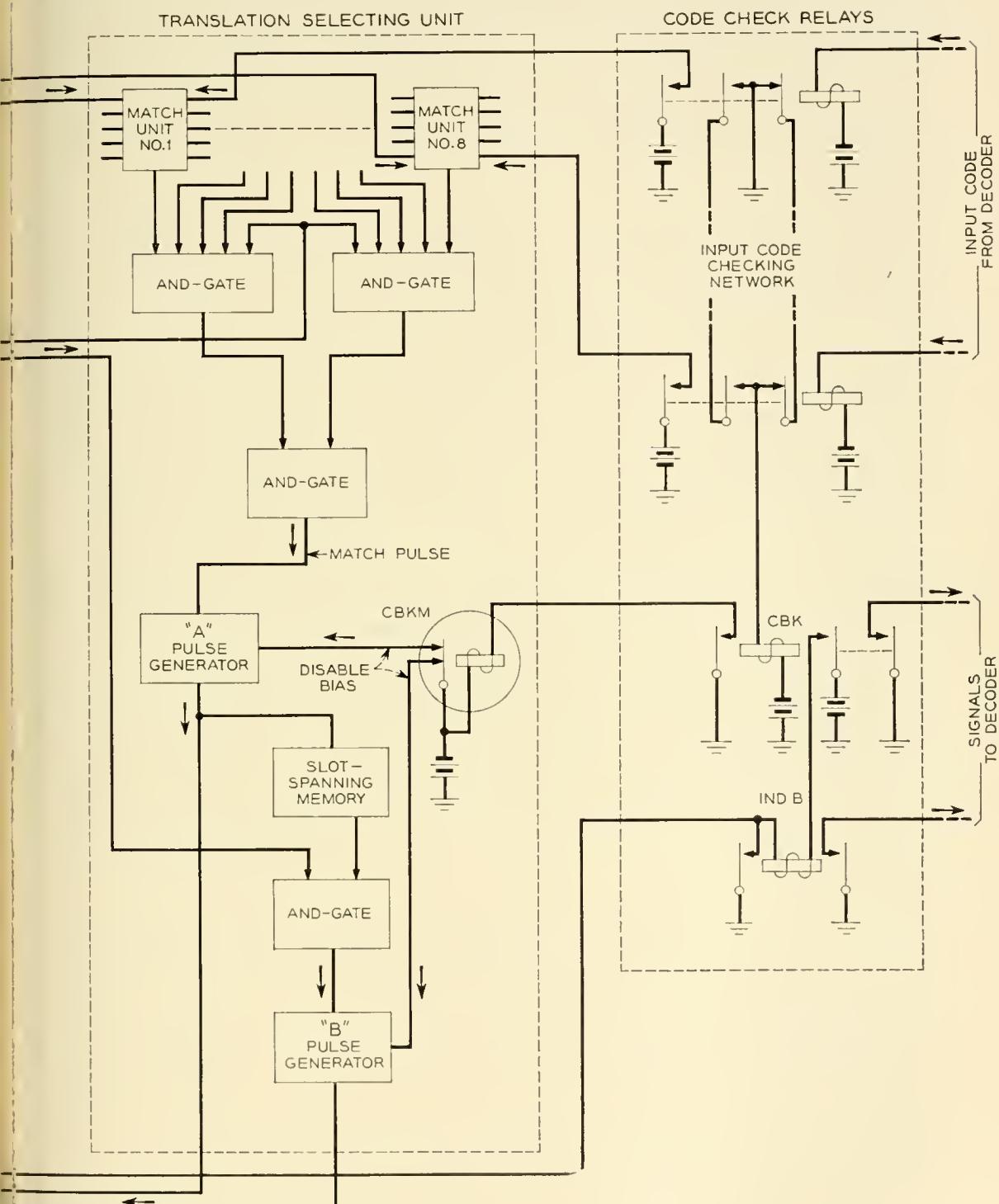


Fig. 4 — Magnetic drum

pulses to all translation selecting and translation delivery circuits in multiple. The object of such an arrangement, naturally, is to employ the magnetic drum system in the most economical manner. A further extension along the same lines would involve relay switching of the pulse circuits



istor block diagram.

to give access to the emergency drum memory, or to a "foreign area" memory where such extra memory capacity is necessary.

Let us now return to the discussion of Fig. 4 and consider the assignment of the translation information to the drum surface where it is stored. Recall that the drum surface is effectively divided into a grid by the coordinates of tracks, each passing under an individual write-read magnetic head, and "slots," each defined by the appearance of a timing pulse in a rhythmic train synchronized from the drum itself, and that the "cells," at the coordinate intersections, each accommodate one bit of code information.

Since each card in the card translator accommodates 38 bits of input code and 116 bits of output, about 160 cells, divided in the ratio of one cell for input to every three cells for output, must be assigned to each translation item. One simple and direct assignment would be to place the entire translation item in a single slot composed of 160 cells. With this layout the slot containing the desired translation would be identified by reading, or "matching" the input code, and during this same interval the output information in the same slot would be gated-out to the translation delivery circuits. A 1,000-translation drum would then be long and narrow, and far too many reading amplifiers would be required. Another evident arrangement would be to assign the entire input code to the first of each group of four slots proceeding under the heads, with the output code following in the next three slots. Such an allocation would require only 40 reading amplifiers but the drum necessary for the desired capacity, with the cell-spacing chosen, would have been larger in diameter than the mechanical designers cared to undertake in their first trial. A logical choice, therefore, was to place each translation item in a pair of adjacent slots, and this was done, although it was later recognized that other, more sophisticated, arrangements might offer certain advantages.

In Fig. 4, the apparent location of one translation item is sketched in relation to the drum surface. This sketch is not drawn to scale, since the slot width is actually only 0.020 inch, and the track width is comparable. It is also geographically inaccurate; actually the cells of any one slot are positioned in four quadrants on the drum, the associated heads being positioned in four stacks for mechanical reasons. However, all of the cells in a time slot pass under all of the heads at the same instant and the presentation of Fig. 4 was adopted for the sake of clarity.

Note, then, that the input code and one-third of the output code are recorded in the first or a slot of a slot-pair passing under the reading heads, and that the remaining two-thirds of the output code occupies

the B slot which immediately follows. The parallel (simultaneous) presentation of the entire input code to the translation selecting unit permits that unit to indicate, by a pulse, that the translation item is the one desired and to gate-out the output code in the same slot while it is still passing under the heads. Having thus identified the first slot of a translation item, it is a simple matter to provide the facility for gating-out the remaining information recorded in the next succeeding slot.

It will be seen, from the circuit arrangement shown, that the translation selecting unit also receives a portion of the output code recorded in the second slot of each pair. It is therefore necessary to distinguish between the A and B slots of a pair. This is most conveniently done by the Timing Unit, which is provided with two outputs, the pulses defining the slots appearing alternately at these outputs. One output lead is chosen to define all the A slots and it is routed to the translation selecting unit to provide a portion of the pulse-pattern required for complete and proper identification of an input code.

The action of the magnetic drum translator in making a translation may now be traced by following the block diagram of Fig. 4. The decoder, of course, gives the same preliminary signals as for the card translator, but these are ignored by the drum translator, because it is continuously presenting all 1024 translations at the rate of 30,000 per second and need not take any preparatory steps, provided its relays have returned to normal after the last translation. The normal state of the relays is checked by means of a circuit through their contacts; if this circuit is complete, the decoder receives the signal to apply the input code as soon as it seizes the translator. A more elaborate checking arrangement could have made this signal conditional upon other tests, such as a "standard translation," to determine that the electronic circuitry (in bulk) was functioning properly, but it was not considered worthwhile to do so in the system described here.

The decoder, then, furnishes the input code of the desired translation item, causing certain of the relays labeled CODE CHECK RELAYS in Fig. 4 to operate. Contacts on these relays are interwired to provide the same checking network as in the card translator, and a check on the authenticity of the input code will be evidenced by operation of the relay labeled CBK. This event is signaled to the decoder so that it may start its "no-card" timer action. When CBK closes, it also operates a chatter-free mercury-contact relay, CBKM, in the translation selecting unit, permitting that unit to produce an output at the appropriate time. Each code-check relay which operates applies a positive voltage to one of the input terminals of a "match" unit in the translation selecting unit. For each of

these input terminal there is a complementary terminal to which are applied negative-going pulses from one of the drum memory reading amplifiers. As will be explained later, advantage is taken of this complementary arrangement to obtain a signal indicating a match between either, (1) an operated code relay and a pulse from the reading amplifier, or (2) a nonoperated relay and no pulse from the reading amplifier. All of these signals, from 40 sections of the match units, are combined in a cascade of "AND" gates; when all indicate a match, the translation selecting unit delivers an output "match" pulse.

Since this match pulse is not strong enough to enable 40 gates in the output channels, it is passed to a "pulse generator" (a regenerative pulse repeater) which produces, virtually coincident in time, a powerful "A" gate-opening pulse. Note that both the "A" and the similar "B" pulse generators are enabled to operate only when the input code is authentic, as evidenced by the operated code check relay CBKM.

In an unrestricted magnetic drum translator design this identifying pulse would cause immediate registry of part of the desired information. Here, however, is evidenced one of the penalties for having a direct one-for-one substitution for a card translator. The decoder and card translator function in a definite sequence; one of the steps in this sequence is initiated by the IND signal from the translator which informs the decoder that the selected card is properly "indexed" so that it may be "read." Therefore, in the case of the drum translator, to preserve this sequence, the selected translation is permitted to pass unheeded, except that the IND signal is synthesized from the identifying B gate-opening pulse. This operation closes one relay, INDB, through a special output channel (top-most one in Fig. 4) provided for the purpose. The decoder, thus notified that the desired translation is available, applies battery to its register relays, and the output channels are completely enabled for a subsequent registry of the desired information.

The output information is usually registered during its next passage, one drum-revolution after initial identification of the item. The action of identifying the translation is again as described above, and there remains only to follow the operation in the output channels. Even before the translation selecting unit has initiated the identifying gate-opening pulse, reading amplifiers which are required to deliver an output code have each commenced delivery of a pulse to their corresponding gate terminals in the AND gate and pulse stretcher units. (See Fig. 4). When these pulse signals have reached a stable maximum, the gate-opening pulse (A or B depending on the slot which is being read at the moment) is free to pass through the gates and to trigger the pulse stretchers. The

latter devices, each containing a single transistor in a monostable circuit arrangement, deliver 12-volt pulses lasting about a millisecond. The pulse stretchers from which an output code is not required are not triggered, owing to the absence of pulses from the corresponding reading amplifiers.

The remainder of the output channel, as previously stated, is borrowed directly from the card translator, and the action is similar. In the output detector, a transformer steps-up the 12-volt pulse signal to a voltage more than sufficient to establish a discharge in the control gap of a cold-cathode gas tube. Since the decoder has applied voltage through a relay coil to the main gap, the discharge transfers, and the resultant current flow operates the relay. The operated relay, which may be in the decoder, registers the code and locks to ground through an auxiliary contact. This action also extinguishes the gas tube, thereby extending its life.

Except for relay operation, all of the activity described here for two drum revolutions repeats itself for every subsequent drum revolution for as long as the code check relay CBKM remains operated. However, once the code is registered, no further use is made of the pulses in the output channels.

When the decoder has made use of the translation, it transmits a signal which is used in the code-check relay system to indicate when all relays are properly restored. In the card translator this signal is also used to restore the selected card, but in the drum translator this operation, of course, is not required.

Administration Equipment

To utilize the magnetic drum translator as described above, it is obvious that some means for writing-in the translations is as necessary to the drum as a card punch is to the card translator. Although a selective writing, or "Administration Unit" was required, a highly efficient design was not essential to the experiment. Consequently there was constructed a separate, portable aggregation of essential basic electronic circuits, arranged for manual control, but designed with a view to possible extension to fully automatic operation. This equipment will be described in a later section.

EQUIPMENT AND CIRCUIT DESIGN DETAILS OF THE TRANSLATOR

General Description

The entire translator is mounted on an 11-foot by 32-inch bay and has been made to conform to telephone central office practices as far as pos-

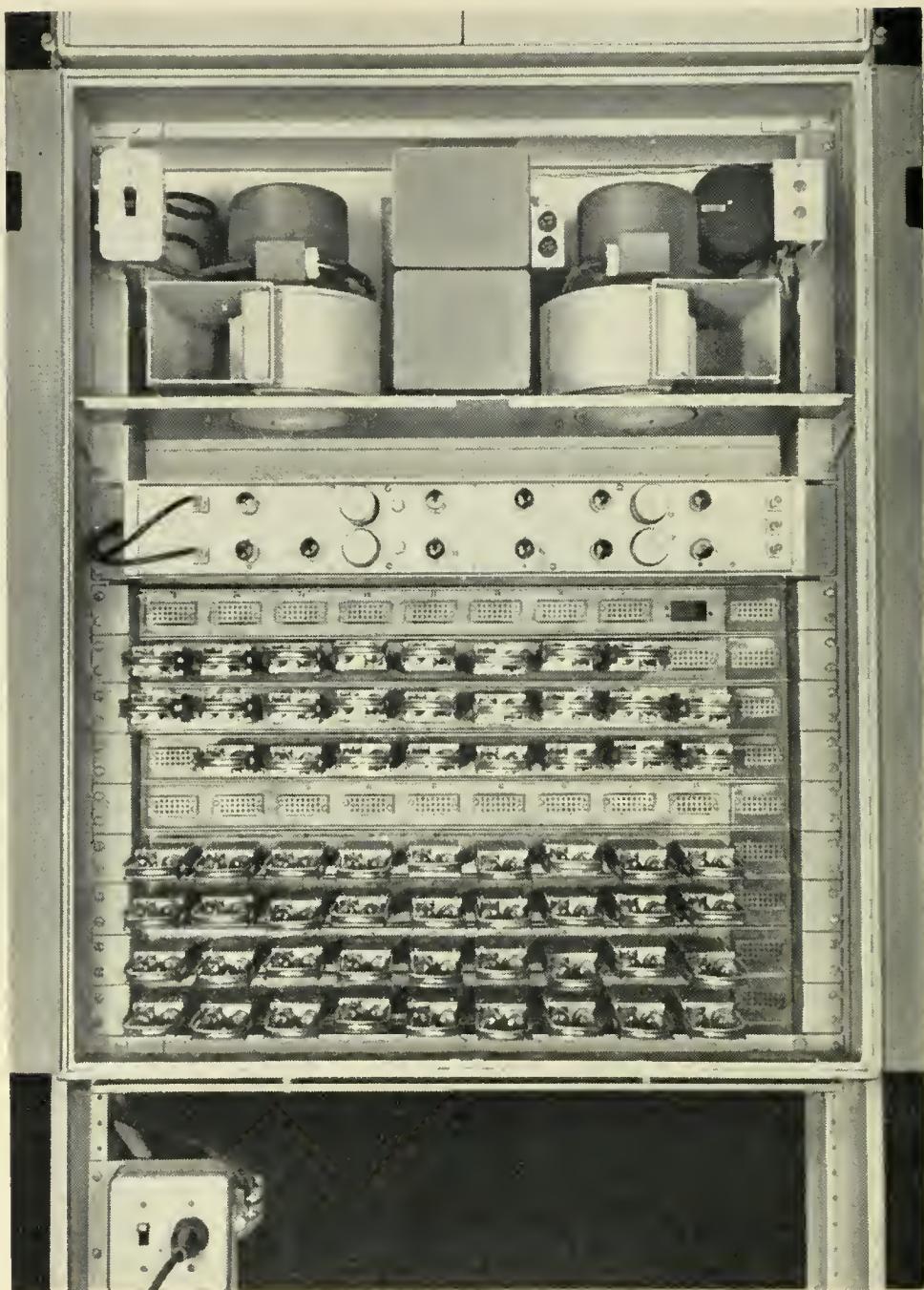


FIG. 5 — Lower casing containing partial complement of reading amplifiers, timing unit, filament transformers and blowers. Receptacle at right end of each amplifier mounting strip allows Administration unit to connect directly to magnetic heads associated with those amplifiers.

sible; except for the presence of the drum unit at the base of the rack, its appearance is not unlike that of other racks found in central offices.

Mounted directly above the drum unit is a casing of conventional design (shown open in Fig. 5) which houses the reading amplifiers, timing unit, filament transformers, and a self-contained forced-air ventilating

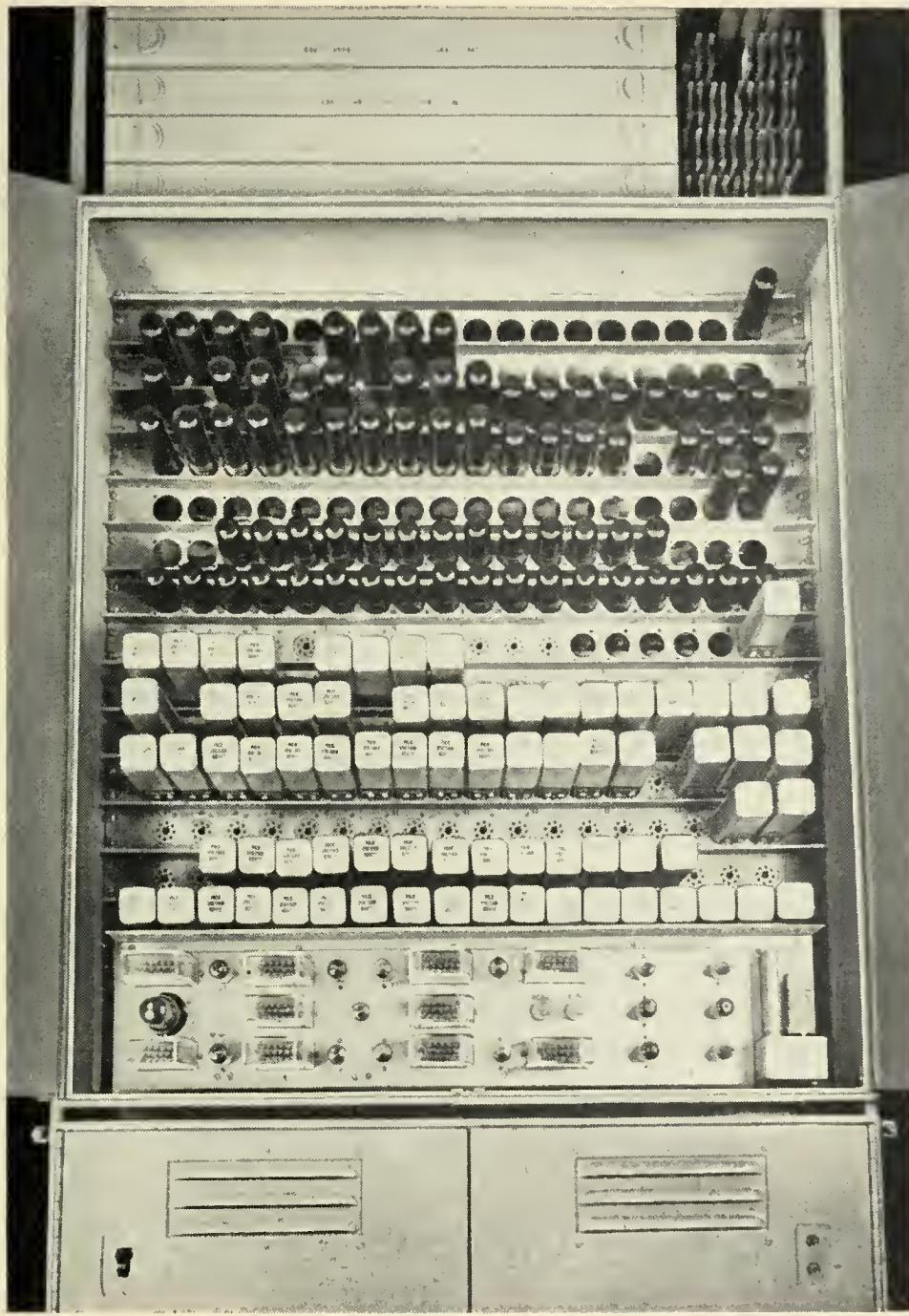


Fig. 6 — Upper casing containing translation selecting unit, and partial complement of pulse stretchers and channel detectors.

system. A second casing, (Fig. 6), located directly above the first, houses the translation selecting unit, pulse stretchers, and channel output detectors. The various plug-in components used in these sections are shown in Fig. 7. At the top of the rack are located the code-check input relays, fuses and terminal blocks.

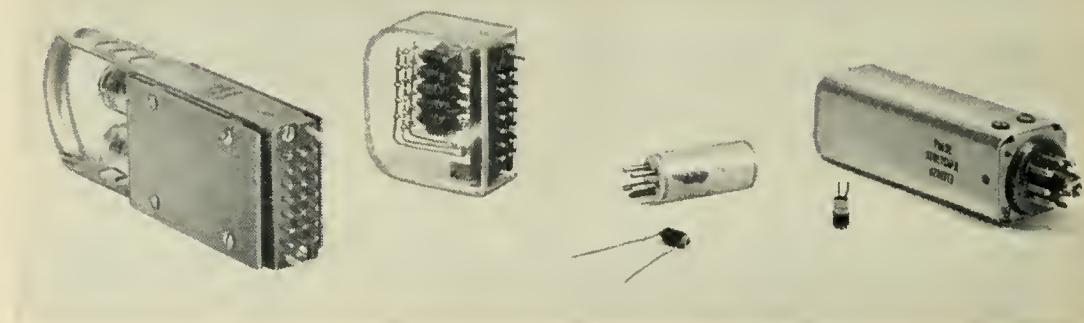


Fig. 7 — Plug-in units. Left to right, reading amplifier, match unit varistor cluster, individual varistor, match and-gate, transistor, and pulse stretcher.

In wiring the rack, use of individually-shielded conductors was held to a minimum. The cable between the drum unit and the reading amplifiers was composed of standard switchboard wire, shielded as a unit by removable sheet-metal enclosures, thus greatly reducing the bulk as compared to the usual bundle of coaxial cables.

The remainder of the wiring, which carries relatively high-level signals from unit to unit within the frame was also in the form of cables of switchboard wire; this type of wiring was tried as an experiment for micro-second pulse work, and was found to be successful in this instance.

Under normal conditions the entire translator, with the exception of the tube filaments and drum drive motor, operates from the standard plant batteries of +130 and -48 volts. Commercial 60-cycle power is normally used for filaments and motor; the motor is duplex and is designed to transfer automatically to the 48-volt plant battery in case of power failure, and the same provision would have to be made for the filaments in the event of a telephone plant installation.

Magnetic Drum Unit

The magnetic drum unit is located at the bottom of the rack, as shown in Fig. 1; a close-up view with one of the covers removed is shown in Fig. 8. A mounting casting supports the machine directly on the floor, straddling the lower member of the rack so that no load is imposed on the rack structure. The drum rotates about a vertical axis and is housed in two cast-iron end-bells spaced by a cast-iron shell. The end-bells carry the bearings for the drum, and serve to mount the motor, while the shell-casting rigidly locates the magnetic heads, each very close to the drum surface. This design requires a minimum of floor space, insures accurate bearing alignment, provides a convenient location for the magnetic heads, and permits the use of tightly-fitting gasketed covers to exclude

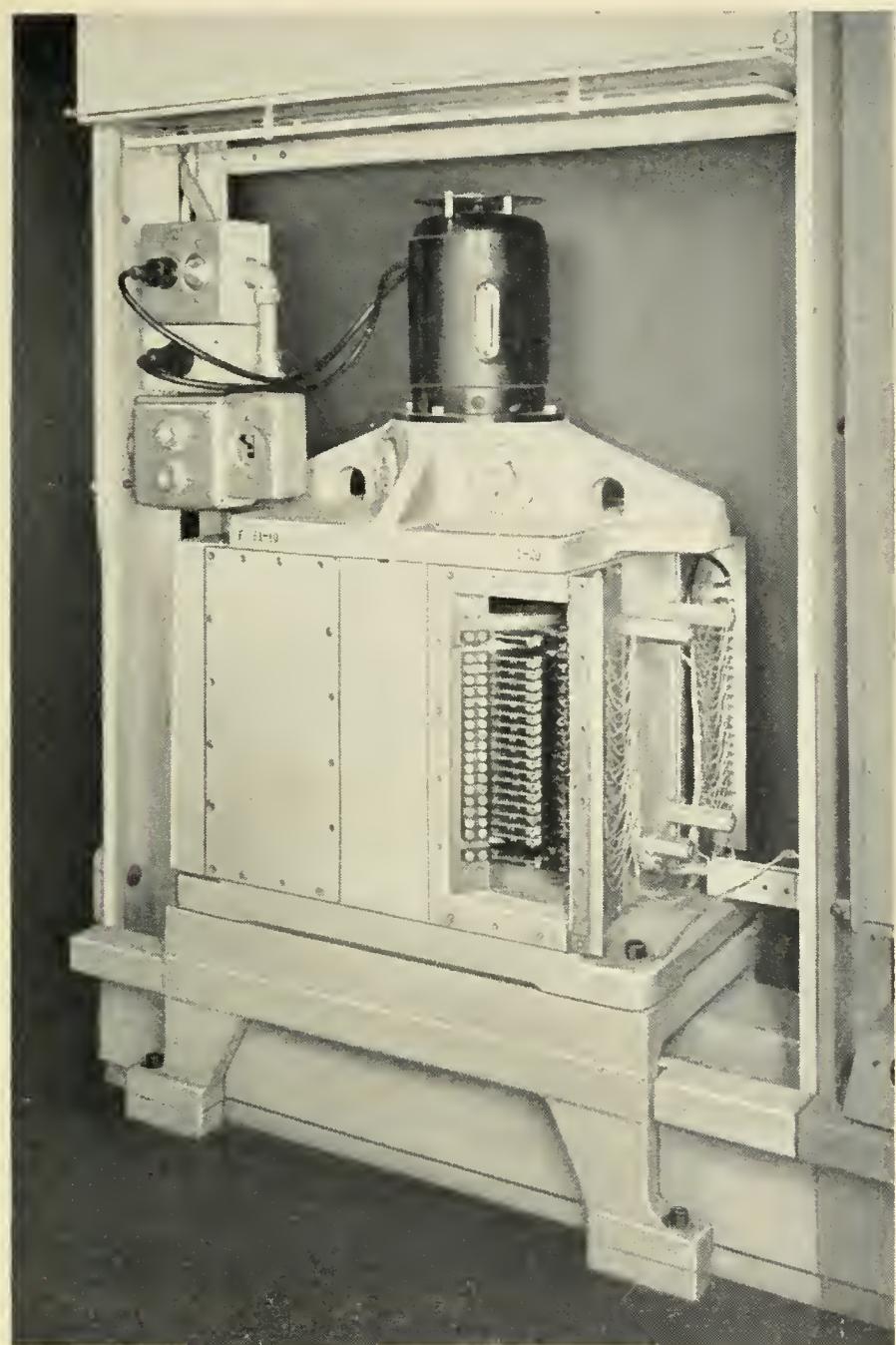


Fig. 8 — Magnetic drum unit partly uncovered to show magnetic heads and wiring terminals.

dirt and foreign material from the magnetic drum surface and the bearings. The $\frac{1}{16}$ -hp motor drives the drum through a spring-diaphragm coupling.

The drum is comprised of a stress-relieved iron casting of high dimensional stability, a press-fitted steel shaft, and a $\frac{1}{16}$ " thick brass outer shell which carries the magnetic recording medium. Since both drum and

housing are of similar materials, and have almost identical temperature-expansion coefficients, it is expected that pole-tip-to-drum clearance will remain unchanged under normal conditions of service. The drum, which is 12.8" in diameter, 10" long, and weighs 150 pounds, is dynamically balanced and runs without sensible vibration.

Commercial super-precision angular-contact ball bearings, two at each end, are used to mount the drum in its housing. The lower bearings are arranged to share the thrust load imposed by the weight of the drum, and the upper bearings are mounted opposing each other, and are pre-loaded one against the other. The upper bearings serve only as radial constraints, the outer races being free to move axially. This type of construction results in a finished unit having a total runout of only a few ten-thousandths of an inch without the necessity of machining the drum on its own bearings. For the experimental installation, the bearings were grease-packed at assembly and can be expected to function satisfactorily during any reasonable test period. If, however, such a drum unit were made a permanent part of the telephone plant, other provisions have been considered which would insure adequate lubrication over a much more extended period.

The magnetic coating used on the drum is an electro-deposited alloy of cobalt and nickel (90 per cent Co-10 per cent Ni) approximately 0.0003" thick. This coating was selected because of its hardness, strength, uniformity, and desirable magnetic characteristics. The thickness of the coating is such as to result in a satisfactory cell-size without undue sacrifice in output. The purpose of the brass sleeve mentioned previously is to form a nonmagnetic surface between the magnetic coating and the cast-iron core since, if the coating were applied directly to a ferro-magnetic material, its effectiveness would be greatly reduced by the shunting effect of the base material. The brass sleeve also serves to facilitate plating the drum, since brass, unlike cast-iron, is amenable to the electro-plating process.

Read-Write Heads

One of the read-write heads is shown in Fig. 9. The magnetic structure consists of three rectangular bars of laminated material, arranged in the form of a triangle (as schematically represented in Fig. 2). Two legs of this triangle carry single-layer coils which are series-connected. These two legs also serve as pole-tips, being pointed at the end and separated by an air gap. The third leg serves to complete the magnetic circuit and, in assembly, is butted tightly against the other members by means of a leafspring.



Fig. 9 — Magnetic head and mounting bracket showing means of adjustment.

The magnetic structure is assembled on a nickel-silver plate to which have been soldered two copper shoes which serve to locate the pole pieces and shield the pole-tips, thereby focusing the recording flux to some degree. After adjustment of the pole-tips, the assembly is clamped in a sandwich by means of a second, smaller nickel-silver plate. As is evident from the illustration, this magnetic assembly is in turn assembled to a mounting bracket which contains facilities for precisely adjusting the clearance between pole-tips and drum surface.

The pole-tips of the head are 0.050" wide and the tracks are on 0.10" centers, leaving a nominal value of 0.050" between tracks to allow for misalignment of heads and for flux-spreading. Heads which are physically adjacent in each of the four corner stacks are mounted on 0.40" centers, but the stacks are offset with respect to one another, thereby interlacing the tracks on the drum.

The read-write heads have been designed expressly for use in high-speed digital recording. Very thin laminations are used and this, coupled with carefully prescribed manufacturing techniques, results in a head having a satisfactory frequency response for the very short pulses employed. When used as a transducer to convert electrical pulses to mag-

netic flux, it is capable of responding faithfully to frequencies approaching ten megacycles per second.

The Timing Wheels and Associated Heads

The synchronizing pulses derived from the drum originate from a 512-tooth soft-steel gear mounted at the top end of the drum. In combination with a polarized reproducing head, the gear generates a timing signal which provides means for permanently locating the various cells used to store information on the drum surface. The polarized head differs from those used on the drum proper, being of a form which is conventional in tone-generators where, as in this instance, a sinusoidal output is desired.

A second gear is mounted at the bottom of the drum, carrying a single tooth of the same proportions as the teeth on the upper gear. In combination with a polarized reproducing head, otherwise quite similar to those used on the drum proper, this single tooth provides a signal once per revolution of the drum which (as will be shown later) is necessary for the operation of the administration unit.

The Reading Amplifier

One of the 80 plug-in reading amplifiers is pictured at the far left in Fig. 7. It employs two twin-triode vacuum tubes, and consists of a three-stage ac-coupled linear broad-band feedback amplifier, followed by a threshold output stage.

As shown in the circuit schematic of Fig. 10, the two halves of v1 and the left-hand half of v2 constitute the linear broad-band amplifier. A suitable choice of coupling elements insures that the amplification will diminish, with decreasing frequency, at a controlled rate for frequencies below a few hundred cycles per second. It is unnecessary to provide amplification at low frequencies, since the signals to be handled have no low-frequency components, and it is undesirable to do so from the standpoint of hum pickup. There is about 20db of feedback in the important part of the frequency range and the amplifier is thus substantially stabilized against variations of gain due to change in operating voltages and aging of tubes. The over-all operating voltage gain of the linear stages, with feedback, is about 56 db; the 3 db points are approximately 300 c/sec and 700 kc/sec.

The grid of the fourth stage of the reading amplifier is coupled to the output of the linear amplifier and is biased to about twice the plate-current cut-off value. The output signal from the plate of this stage, occa-

sioned by reading a "1", will be a negative-going pulse of approximately 40-volt amplitude from a standing potential equal to the plate supply, +130 volts. As a precaution against false signals, an externally-mounted plate-feed resistor is provided to establish at the output a condition corresponding to that of no signal present when the amplifier is removed from its receptacle.

Timing Unit

The timing unit accepts an approximately sinusoidal timing-wave signal from the upper timing head, and converts this signal into two pulse-trains, each having 1,024 narrow pulses per drum revolution, designated as A sync and B sync, alternating in time and available on separate outputs for controlling all the rest of the circuit action of the translator. A block-schematic indicating how the pulse trains are produced is shown in Fig. 11.

The general procedure for converting from a sine-wave to a synchronous train of short pulses, two per cycle of input, may be traced through the upper channel of the drawing. The signal, as represented by voltage trace 1, is amplified and clipped until a steep-sided square wave is obtained; this wave, trace 2, is applied to a push-pull phase inverter from which a pair of oppositely-phased outputs is obtained. Each of the two outputs is then differentiated by means of an R-C network, and the nega-

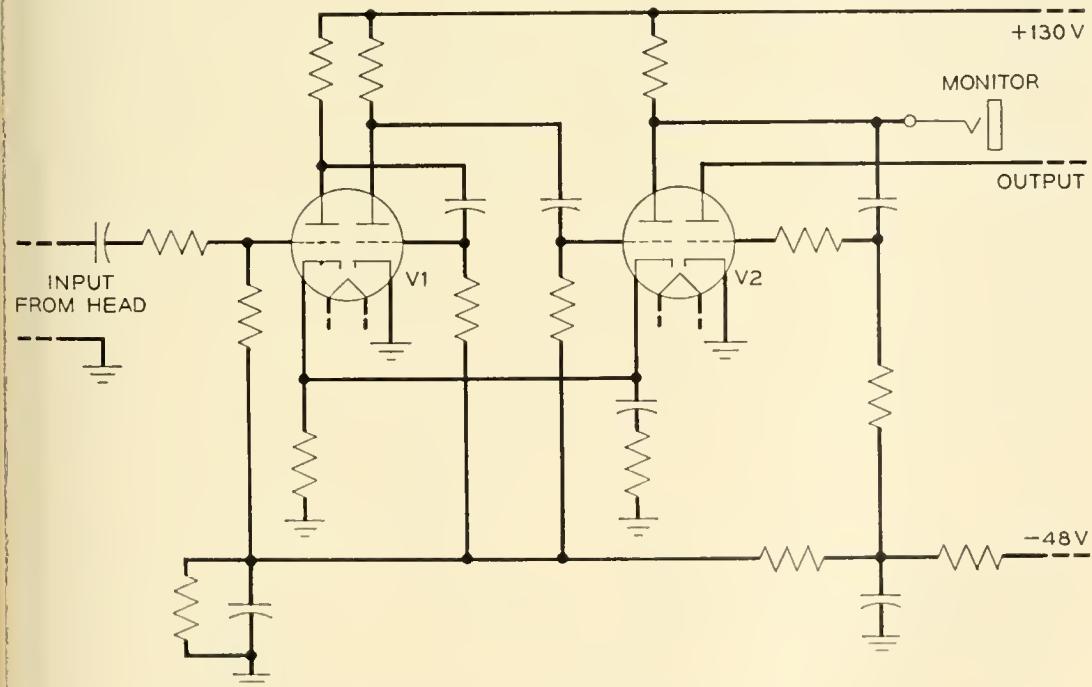


Fig. 10 — Reading amplifier circuit.

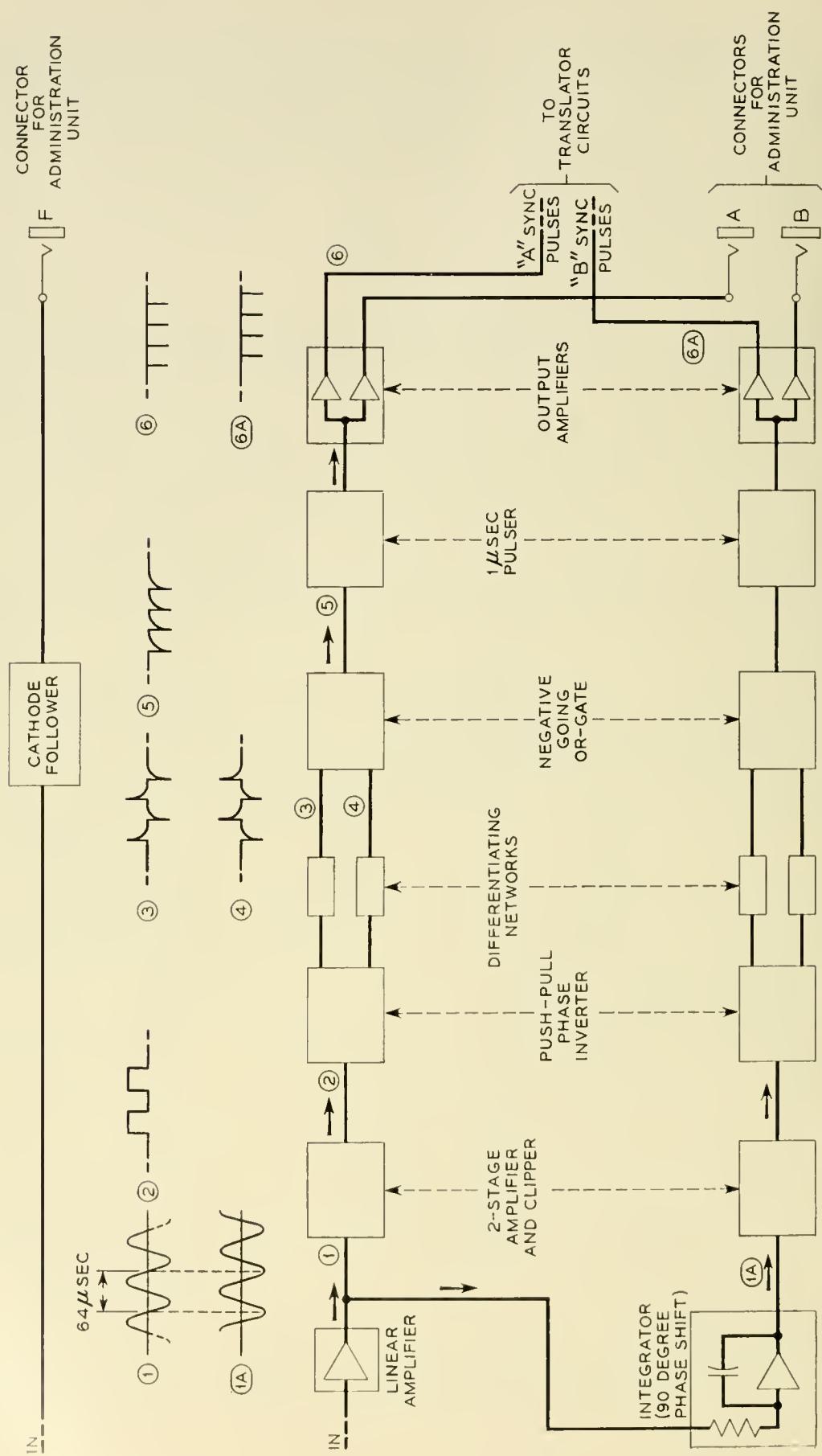


Fig. 11 — Timing unit block diagram.

tive-going spikes, traces 3 and 4, are combined in a negative-going OR gate of crystal diodes.

These spikes, trace 5, are used to trigger a cathode-coupled single-shot multivibrator, designed to give a rectangular pulse of about one microsecond duration. The multivibrator drives a pair of identical output stages: one furnishes the required A sync pulses to other equipment in the translator bay, and the other delivers its output to a coaxial connector so that, when required, the pulses may be furnished to the administration unit.

The B sync pulse-train is produced in the lower channel shown in Fig. 11. After some linear amplification, a part of the original input sine-wave is applied to a vacuum tube integrator circuit. The constants of the integrator are such that it provides very nearly a quarter-period of phase shift even if the drum varies from its nominal speed. The output of the integrator is then treated in the same manner as that described for the direct input, with the result that the required B sync pulses are produced.

The timing unit also contains a third channel which accepts the once-per-revolution signal from the special head adjacent to the single-tooth wheel. The output of this channel provides the fiducial signal, on a low-impedance basis, for administrative operations.

The Translation Selecting Unit

This unit, which appears as the bottom panel in the photograph, Fig. 6, performs a number of successive steps in making its selection. These are: (1) recognition of a match between input information from a decoder seeking a translation, and the unique corresponding information from the drum, selected from the flow of continuously-presented information; (2) production of a gate-opening pulse whose leading edge is substantially coincident in time with the leading edge of the particular A sync pulse corresponding to the entry for which the match occurred; (3) activation of a slot-spanning pulse circuit to bridge the time interval until the next-following B slot; (4) production, at a separate output, of another gate-opening pulse whose leading edge is substantially coincident in time with the leading edge of the identified B sync pulse. These actions will now be considered individually.

(1) *Recognition of Match*

Responsibility for this function is divided among a group of eight match-units operating with their associated differential amplifiers. Each match-unit is capable of comparing the inputs from five code-relays with the potentially-matching outputs of five reading amplifiers.

A circuit schematic of one of the units, with its associated differential

amplifier and some of the connected apparatus, is shown in Fig. 12. The uppermost channel on this diagram is typical of all five channels. Resistors R1 to R5 are proportioned so that the potential at point c assumes a value of +115 volts for either of the two acceptable conditions of match: (1) code-relay unoperated and reading amplifier not drawing plate current, or (2) code relay operated and reading amplifier drawing a pulse of plate current. Whenever either of the two possible conditions of mismatch exists, the potential at point c assumes a value about 15 volts higher or lower, depending on the nature of the mismatch. Resistor R6 is introduced for protective purposes only. Varistor VR1 limits the nega-

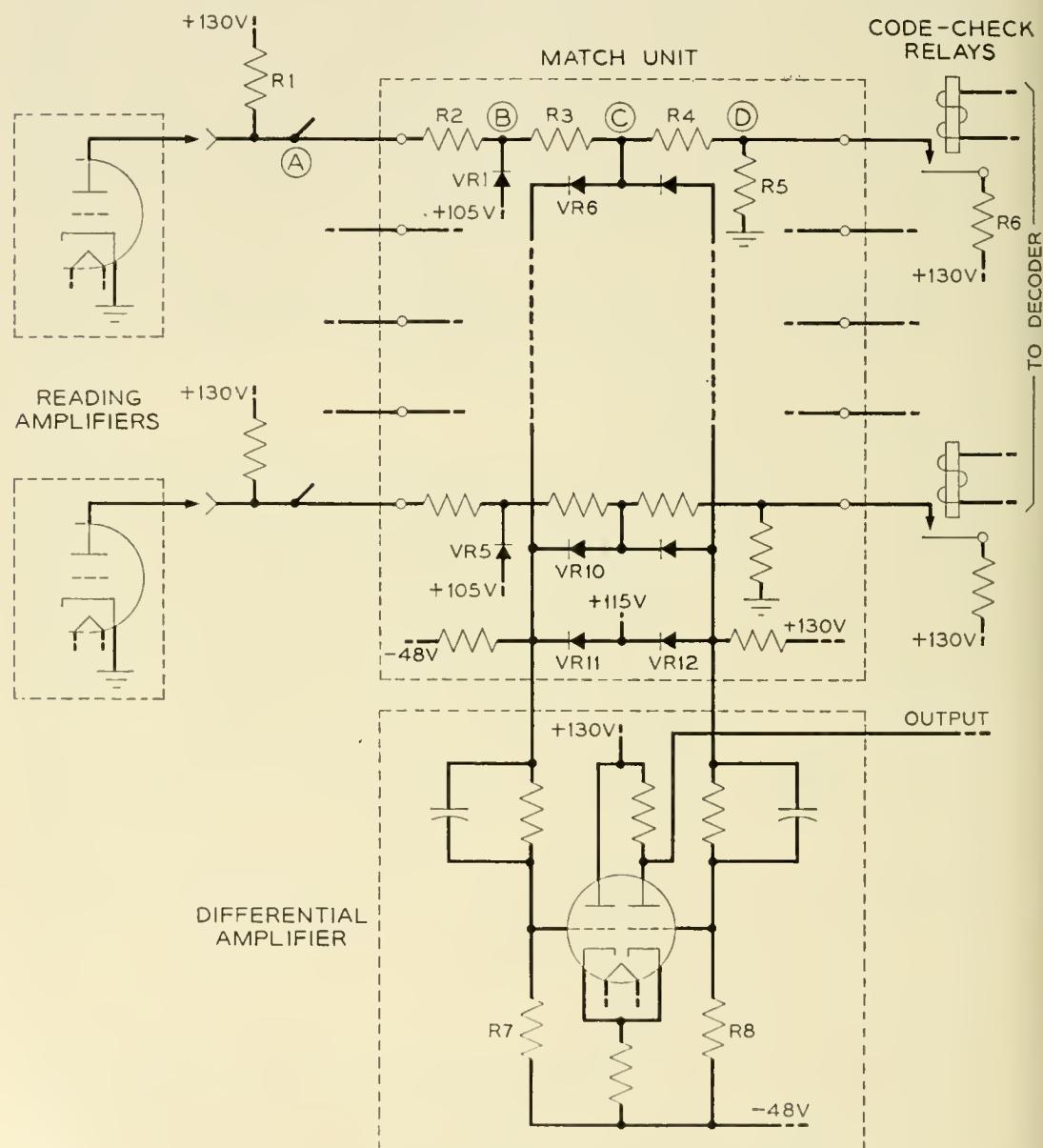


Fig. 12 — Match unit and differential amplifier circuit.

tive voltage excursion at point B, during a pulse, so that it never goes below +105 volts. This establishes the uniform pulse amplitude among the forty match channels which is necessary for proper functioning of the unit.

To detect and recognize the voltage conditions at the five junction points, two varistor gates and a differential amplifier are employed. One gate, comprising six varistors including VR6 and VR11, will transmit the type of mismatch signal which is more positive than +115 volts. This signal is dc-coupled to the left-hand grid of the differential amplifier as illustrated in Fig. 12. The type of mismatch signal which is less positive than +115 volts is blocked by this gate but is transmitted through the other gate to the right-hand grid. The threshold for this discriminating action is established by application of a fixed nominal potential of +115 volts to varistors VR11 and VR12.

At match, the output of each of the two gates presents a potential of +115 volts to the differential amplifier. The differential amplifier is biased (by inequality of R7 and R8) so that for this condition the right-hand triode is conducting, and the output potential is lower than the plate supply voltage. Positive-going mismatch signals on the left-hand grid, or negative-going signals on the right-hand grid are then equally effective in cutting off the right-hand triode, causing the output voltage to rise to plate supply potential signifying a mismatch.

The outputs from the differential amplifiers of the eight match units are combined with the A sync pulses in a system of AND gates, as illustrated in Fig. 4. A match-pulse output from this system thus signifies that conditions for match have been uniquely determined for 40 pairs of items. Thus the match unit, in total, is capable of distinguishing between all binary combinations of 40 bits or approximately 10^{12} items although when a self-checking code is employed, as in the translator application, many of these combinations are inadmissible.

(2) *The A Gate-Opening Pulse*

Occurrence of the match-pulse, as just described, indicates that the 40 items constituting one-half the contents of one of the A slots match the incoming input code; it is then desired to spill out from the other half of this same A slot the information which is also appearing at amplifier outputs at that instant. This is done by means of gates opened by the action of a gate-opening pulse, triggered by the match pulse.

The A gate-opening pulse is only a few microseconds in duration and normally is produced only once per revolution of the drum; a quiescent blocking-oscillator was chosen as the type of circuit best suited for this purpose. Whenever the code-check relays are operated in an authentic

code combination, relay CBKM is operated, removing a disabling bias from the driver stage of the blocking oscillator. When in this condition, each occurrence of the match pulse will trigger the blocking oscillator, thereby producing the A gate-opening pulse once per drum revolution.

(3) *Slot-Spanning Pulser*

Whenever an A gate-opening pulse has acted to permit read-out of information from half of the proper A slot, it is also desired to read out all the information from the next-following B slot. The first step toward doing this is to cause the A gate-opening pulse to trigger a single-shot multivibrator whose characteristic period is long enough to just bridge the time until the next slot appears. The output of this pulser is combined with the B sync pulses in an AND gate so that the selected B pulse, corresponding to the wanted B slot, can be used to trigger another gate-opening blocking-oscillator just as the match pulse was used to trigger the A gate-opening blocking-oscillator.

(4) *The B Gate-Opening Pulse*

The outputs of all the reading amplifiers must be gated for the B slot. Hence the B gate-opening pulse must operate twice as many gates as the A gate-opening pulse and must be correspondingly more powerful. This requirement is met by using the same circuit design with parallel output tubes.

Pulse Stretchers and Channel Detectors

Fig. 13 presents a simplified schematic of one of the translator output channels, together with certain of the relays in the decoder. Package-wise, the pulse stretchers combine two functions: that of an AND gate with two inputs and a threshold feature, and that of a single-shot multivibrator for amplifying and lengthening the short input pulse from the gate. A single point-contact transistor provides the necessary gain for the monostable action. The inputs to the AND gate come from sources which supply negative-going pulses from a standing potential of +130 volts. When one or the other, but not both, of these sources supplies a pulse, a larger portion of the current being supplied to resistor R1 must be drawn from the non-active source; this extra demand causes a small voltage drop which becomes evident at the gate output. The resultant weak false signal is prevented from affecting the transistor pulser by the action of threshold diode VR1 which is normally back-biased a few volts by the potential divider R2, R3. Small negative-going signals from the gate will not overcome the bias and will therefore be greatly attenuated; normal gate-output pulses, occasioned by coincidence of pulses at both inputs will,

however, overcome the bias and will be transmitted to the transistor monostable circuit.

When triggered at the base, the transistor delivers a pulse of about one millisecond duration to the load represented by the input transformer and the channel detector gas tube and thus provides the drive required to initiate ionization in the control gap of the gas tube. When brought into action, the transistor serves as a switch to connect capacitor C to collector supply resistor R6. The voltage change, occasioned by the resultant flow of current in R6, is communicated to the transformer primary through a blocking capacitor and a current limiting resistor. As capacitor C charges, the voltage at the transistor emitter will approach the collector supply potential at an approximately exponential rate. When the diminishing flow of emitter current can no longer maintain the transistor in its low-impedance mode, it reverts to its pre-triggered condition, and the timing capacitor C is then discharged, primarily through forward-conducting varistor VR2 and resistors R5 and R4.

Owing to the necessity of using early-production samples of the type of point-contact transistor chosen for this application, the associated circuitry for biasing the emitter into the normal non-conducting state is somewhat more elaborate than that which might have sufficed with later samples whose characteristics were more closely controlled.

The principal components of the channel detector are a step-up trans-

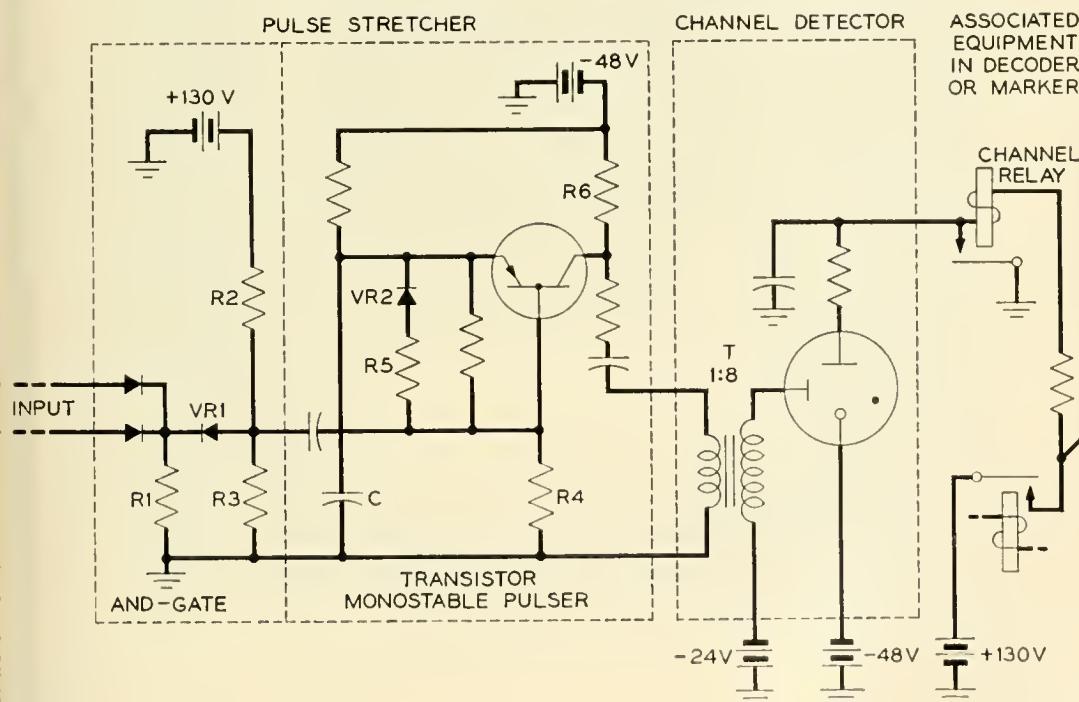


Fig. 13 — Pulse stretcher and channel detector circuit.

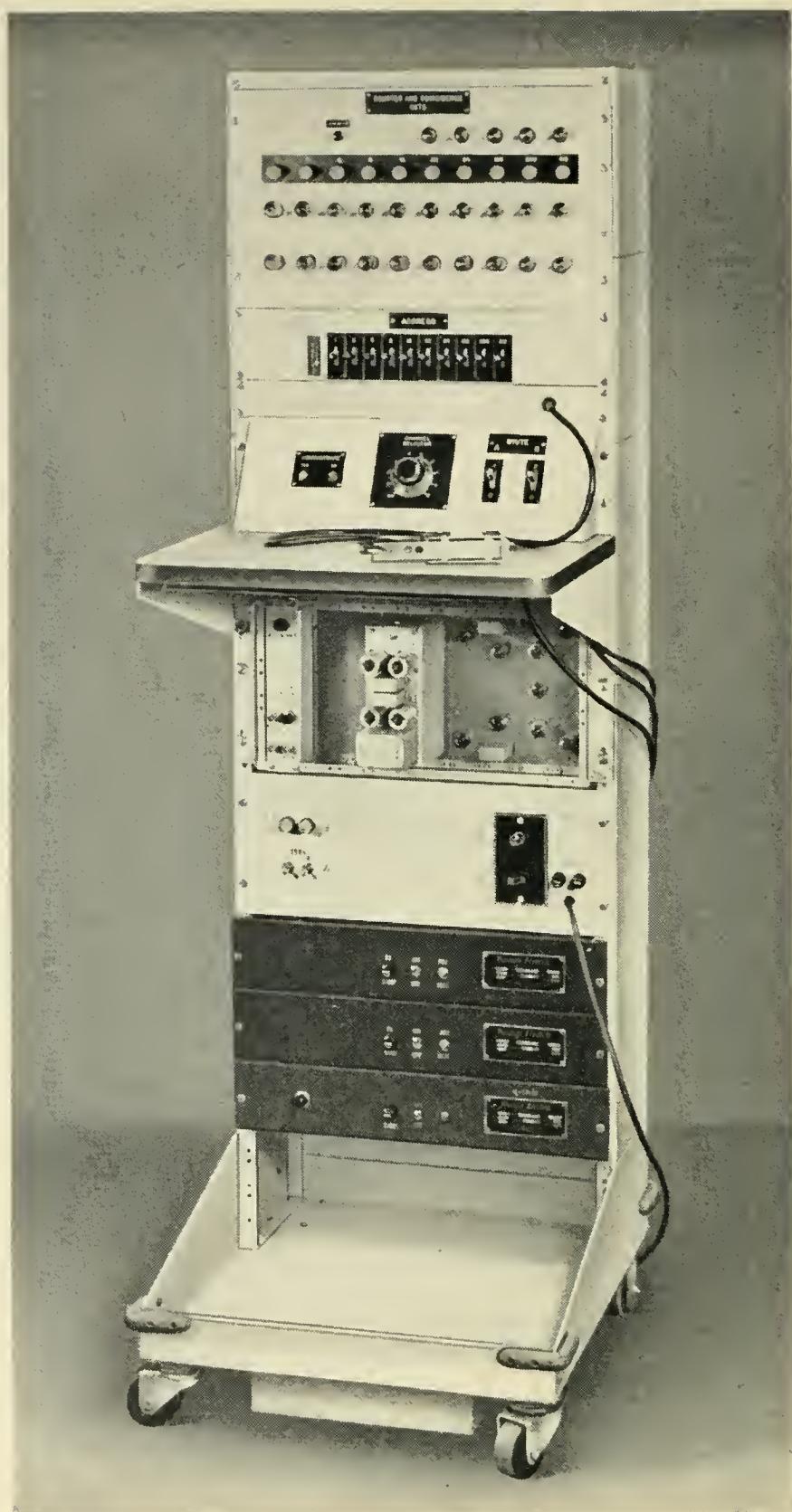


Fig. 14 — Administration unit. Three co-ax leads entering under shelf bring A, B and F pulses from translator. Cable leading to plug with bail-handle resting on shelf serves to connect writing amplifier output to magnetic heads in translator. Bottom cable connects to 60-cycle source which supplies all power.

former designed for the audio frequency range, and a cold-cathode gas tube. The starter-anode of the gas tube has a dc bias of about +24 volts with respect to its cathode to reduce the value of pulse voltage required to ionize it. When +130 volt battery is applied via the winding of the channel relay to the main anode of the gas tube, ionization established in the starter gap by the pulse stretcher signal will transfer to the main gap and cause the relay to operate. Closure of one of the relay make-contacts serves to divert the winding current from the gas tube directly to ground, thereby extinguishing the tube and prolonging its life. Other contacts, not shown, make the registered information available.

Components

A full complement of the electronic apparatus described in the last few sections utilizes plug-in components in the following quantities:

Twin-triode electron tubes	186
Cold-cathode gas tubes	121
Germanium varistors	552
Point-contact transistors	120

Only one type of each of these components is used in the translator; this uniformity greatly simplifies the maintenance problem and imposed little if any handicap on the circuit designs.

ADMINISTRATION EQUIPMENT

Whenever it is desired to add, or to change, a translation item on the drum, the auxiliary administration unit pictured in Fig. 14 is connected to the translator by three shielded cables, shown leaving the rack just under the shelf, and a ten-conductor cable, shown with its plug resting on the shelf. The shielded cables convey the A and B sync pulses and the once-per-drum-revolution fiducial F pulse to the administrator. The ten-conductor cable, with plug, is used to establish paths extending directly to magnetic heads on the drum. During the recording of any one complete translation item on the drum, this plug is successively shifted to each of nine multi-connector jacks located in the amplifier compartment of the translator.

The manual controls are located just above the shelf. At the right are the two keys for ordering a writing operation, one for the A slot and another for the B slot of the chosen pair. If either key is lifted, it will order the entry of a magnetic mark (write "1"). If depressed, the key will order the removal of a mark (write "0"). It is obvious that the translation is

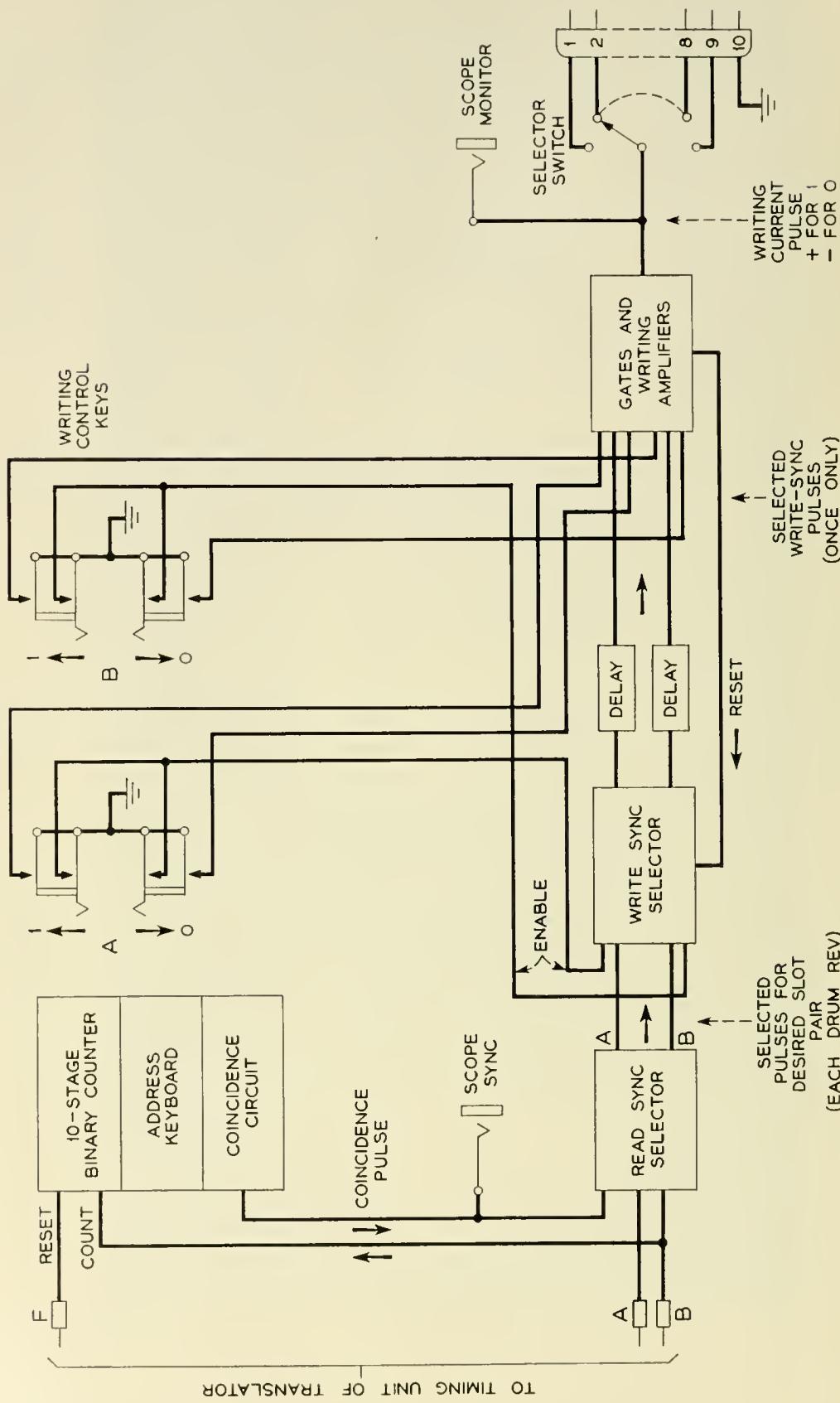


Fig. 15.—Administration unit block diagram.

inserted piecemeal by working in each track successively. The manual switching operation of connecting a single pair of writing amplifiers to each of eighty magnetic heads, in turn, is accomplished partly by setting the nine-position switch shown at the center of the panel, and partly by shifting the plug of the ten-conductor cable. At the left are two signal lights which serve as alarms to warn the operator of possible incorrect functioning of the equipment.

The operation of the administration unit can best be traced with the aid of the schematic block-diagram of Fig. 15. A ten-stage binary counter is supplied with B sync pulses from the translator; the 1,024 possible states of the counter are traversed in the course of exactly one revolution of the translator drum. The F pulse from the translator will, midway between two B pulses, set all counter stages to zero, once per revolution. After the first such reset, however, if the counter is working properly, it will always have returned to the zero condition just before the occurrence of the F pulse, by having counted 1,024 B pulses; under these conditions the F pulse, though still initiating reset action, does not change the state of the counter. The basis for the alarm signals mentioned above is a circuit arranged to detect if a change of state is occasioned by the F pulse.

Associated with the counter is a coincidence circuit with a keyboard on which may be set up any "address" between 0 and 1,023. When the count of B pulses equals the address set up on the keyboard, the coincidence circuit delivers a pulse which persists until the next B pulse alters the count; this coincidence pulse spans the time of occurrence of an A pulse, and is used in the read synce selector to gate-out a "selected" A pulse uniquely assigned to the address set up on the keyboard. A slot-spanning pulser, triggered by the selected A pulse, gates-out the associated "selected" B pulse.

These selected pulses, which occur once per revolution of the drum, are passed through gates under control of bistable electron-tube pairs which can be set by the manual writing keys and are re-set by the writing action itself. This insures that the desired action takes place only once per key operation, instead of repeating, once per drum-revolution, as long as the keys are held operated. The manually-gated unique selected A or selected B sync pulse is then slightly delayed in time to become a selected write-synce pulse. It is passed on through further gates under direct control of the writing keys, and is employed as an input to a writing amplifier.

A pair of writing amplifiers is provided, one to write "1" and the other to write "0"; the circuits are identical quiescent blocking-oscillators sharing a common output transformer, and one or the other is triggered into

action by the write-sync pulses. The output transformer supplies the writing current pulses, under control of the selector switch, to the chosen magnetic head. Arrangements are provided for synchronizing an oscilloscope to display the writing current pulses or the voltage outputs from the head at the selected address, as required.

When a new translation item is to be entered, or an existing one altered, the address corresponding to the desired slot-pair is determined from a card-index, or ledger, listing all items on the drum. The address keyboard is then set to the assigned number, thereby singling-out the desired slot pair so that the writing operation can proceed as described above. During this procedure, the monitoring oscilloscope may be used for verifying the new entry, two cells at a time. Over-all verification is accomplished by exercising the translator through facilities already available in the toll switching office. There is nothing about this procedure which precludes the use of automatic facilities for performing the administration. There is also no fundamental need to take the translator out of routine service during the administration operation, since each writing operation disables the equipment for only a few microseconds and would rarely delay a translation by as much as one drum revolution.

CONCLUSION

After short preliminary tests, the equipment described and pictured was installed in the switching systems laboratory at Bell Laboratories. A rapid-transfer arrangement permitted direct interchangeability with a card translator in a skeletonized model of a toll switching office.

A testing program was then begun entailing continuous 24-hours-per-day operation of the magnetic drum translator for approximately one year. After an initial shakedown period during which wiring faults and other minor troubles were recognized and cleared, many millions of translations were handled with only a small proportion of failures. The accumulated data on failure rate and cause was significant, being one of the primary objectives of the experiment. An analysis of the data indicated the desirability of certain simple design changes in the existing circuitry and established a basis for the selection of future designs.

If, in the future, consideration is given to the design of equipment of this type for some specific application, new electronic developments must also be taken into account. Many more types of transistors are now available than when the present design was undertaken, and some of the newer types have capabilities which make them obvious candidates for many of the jobs now done in the translator with electron tubes. Such a substitution would not only increase reliability and decrease power con-

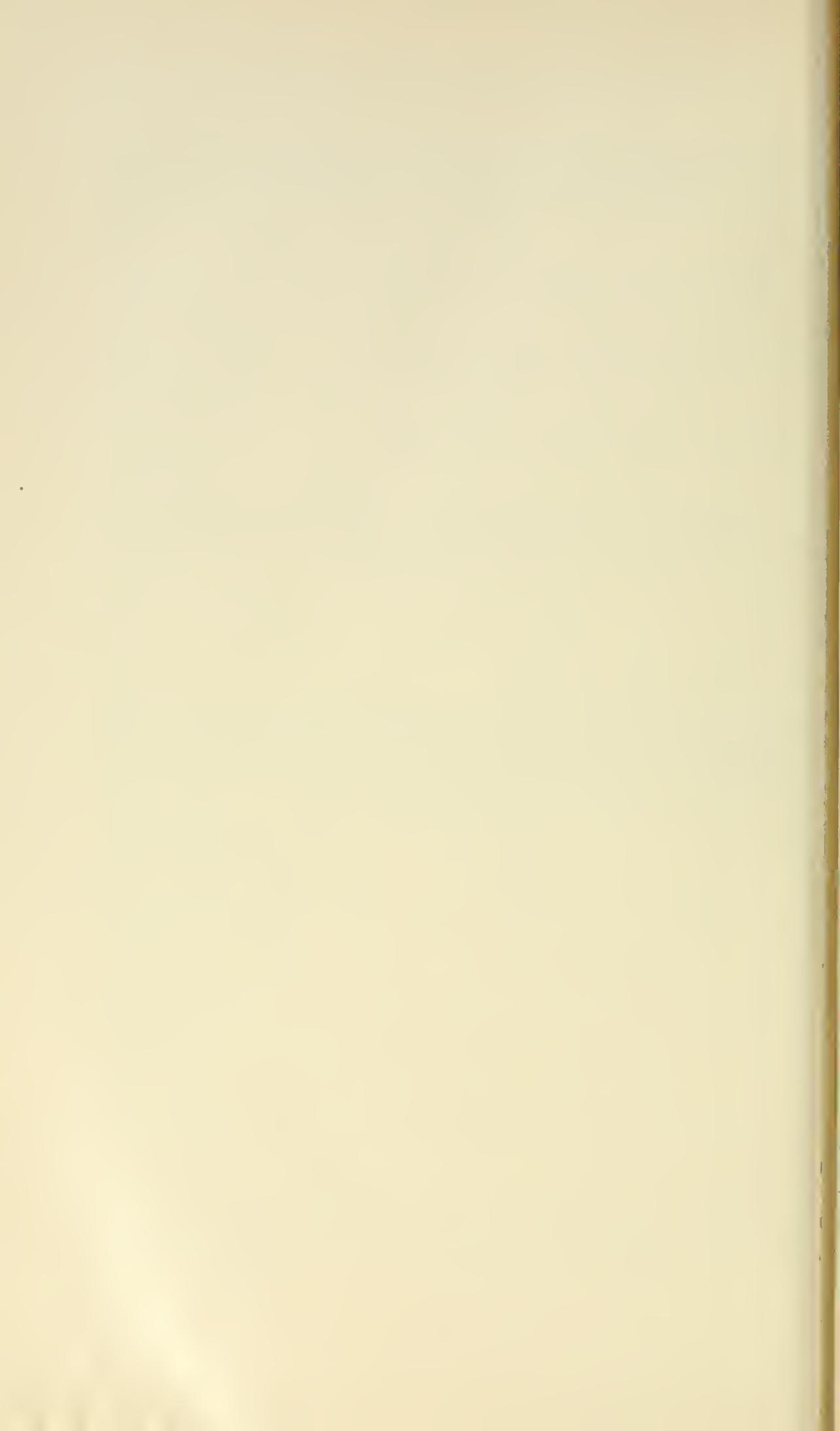
sumption, but since transistors are essentially current-operated devices they would seem to be particularly suitable for working with microsecond pulses in the environment of existing relay-equipped offices where the majority of interference transients are capacitively-propagated voltage-disturbances.

Evaluation of the magnetic drum reveals it to be a safe and very reliable means of storing several hundred thousand bits of information. During the course of these tests, the drum functioned perfectly, and the translations that were recorded at the beginning of the test were retained until near the end, when they were deliberately altered. During this interval of nearly continuous operation there was no detectable deterioration, or change in the signals obtained from the drum.

The results obtained from the tests of this particular drum translator indicate that the associated circuitry, working with microsecond pulses, can be designed to measure up to the exacting standards demanded for telephone office apparatus, whether the application be that of a magnetic drum translator or some other type of equipment.

REFERENCES

1. W. D. Lewis, Electronic Computers and Telephone Switching, Proc. I.R.E., **41**, pp. 1242-1244; Oct., 1953.
2. W. A. Malthaner and H. E. Vaughan, An Automatic Telephone System Employing Magnetic Drum Memory, Proc. I.R.E., **41**, pp. 1341-1347; Oct., 1953.
3. J. H. McGuigan, Combined Reading and Writing on a Magnetic Drum, Proc. I.R.E., **41**, pp. 1438-1444; Oct., 1953.
4. L. N. Hampton and J. B. Newsom, The Card Translator for Nationwide Dialing, B. S. T. J., **32**, pp. 1037-1098; Sept., 1953.



Tables of Phase of a Semi-Infinite Unit Attenuation Slope

By D. E. THOMAS

(Manuscript received February 24, 1956)

Five and seven place tables of the integral

$$B(x_c) = \frac{1}{\pi} \int_{x=0}^{x=x_c} \log \left| \frac{1+x}{1-x} \right| \frac{dx}{x}$$

which gives the phase associated with a semi-infinite unit slope of attenuation, are now available in monograph form. The usefulness of this integral and its tabulation are discussed.

H. W. Bode¹ has shown that on the imaginary axis, the values of the imaginary part of certain functions of a complex variable may be obtained from the corresponding values of the real part, and vice versa. This theorem was immediately recognized as a powerful tool in the communications and network fields. The most generally useful function which was given by Bode for use in applying this theorem to the solution of communications problems, is the phase associated with a semi-infinite unit slope of attenuation. This is given by the integral²

$$B(x_c) = \frac{1}{\pi} \int_{x=0}^{x=x_c} \log \left| \frac{1+x}{1-x} \right| \frac{dx}{x} \quad (1)$$

where: $B(x_c)$ is the phase in radians at frequency f_c ,

$$x = \frac{f}{f_0}, \quad x_c = \frac{f_c}{f_0} < 1.0$$

and f_0 = the frequency at which the semi-infinite unit slope begins

The usefulness of Integral (1) is illustrated by some of the communication problems which stimulated its accurate tabulation.

¹ Bode, H. W., Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co., Inc., New York, 1945, Chap. XIV.

² Ibid: Chap. XV, pp. 342-343.

When the development program on deep sea repeatered submarine telephone cable systems was reactivated at the close of World War II, one of the first problems to present itself was the determination of the delay distortion of a transatlantic repeatered cable system. The only means then known of obtaining an answer to this problem was by computing the minimum phase of the system from its predictable attenuation characteristic, using Bode's straight line approximation method,³ and then determining the delay distortion from the non-linear portion of this minimum phase. However, the non-linear phase is such a small part of the total phase, that a five figure accuracy tabulation of Integral (1) was needed for a satisfactory determination of the non-linearity. The necessary table was therefore compiled. A numerical computation was used to evaluate the integral because of the simplicity of its integrand. The minimum phase of the projected transatlantic repeatered telephone cables was then computed using this table and the anticipated delay distortion was determined from the non-linear portion of this minimum phase.

About this time the delay equalization of coaxial cable systems for television transmission became a pressing problem. Bode's technique proved to be the simplest means for determining the delay to be equalized and so the existing phase table was immediately put to use in the coaxial cable delay equalization program.

The increasing use of the tables led to a decision to publish them in THE BELL SYSTEM TECHNICAL JOURNAL.⁴ In order to make the tables more generally useful, the published paper included a tabulation of the phase in radians as well as in degrees. The radian tables can, for example, be used to determine the reactance characteristic associated with a given resistance characteristic of a minimum reactance impedance function.

Because of the demand for higher accuracy which occasionally arose after the publication of the five place tables, it was decided to undertake the computation of seven-place tables. These tables were also computed numerically using intervals selected to give at least ± 1 accuracy in the final figure. The complete tables require forty-nine pages for tabulation. Since it is probable that only a fraction of the JOURNAL readers would need these tables, it did not seem desirable to publish the actual tables in the JOURNAL. They are therefore being published in original monograph form as Bell System Monograph 2550⁵ entitled "Tables of Phase of a Semi-Infinite Unit Attenuation Slope." The phase is tabulated in the

³ Ibid: Chap. XV.

⁴ Thomas, D. E., Tables of Phase Associated with a Semi-Infinite Unit Slope of Attenuation, B.S.T.J., 26, pp. 870-899, Oct., 1947.

⁵ This Monograph will be available about June 15, 1956.

monograph both in degrees and radians for values of f greater than f_0 as well as for f less than f_0 . The tabular intervals are 0(0.001) 0.600 (0.0005) 0.9000 (0.0001) 0.9940 (0.00005) 0.99800 (0.00001) 1.00000. These intervals were selected to permit linear interpolation for intermediate values of the phase to an accuracy of the same order as the accuracy of the tabulated values, i.e., ± 1 in the last place. The original JOURNAL article discussed the construction of the tables and the errors involved in the numerical evaluation of Integral (1), described and illustrated the use of the tables, and gave five-place tabulations of the integral. This entire article is therefore included in Monograph 2550 for completeness along with the newer seven-place tables.

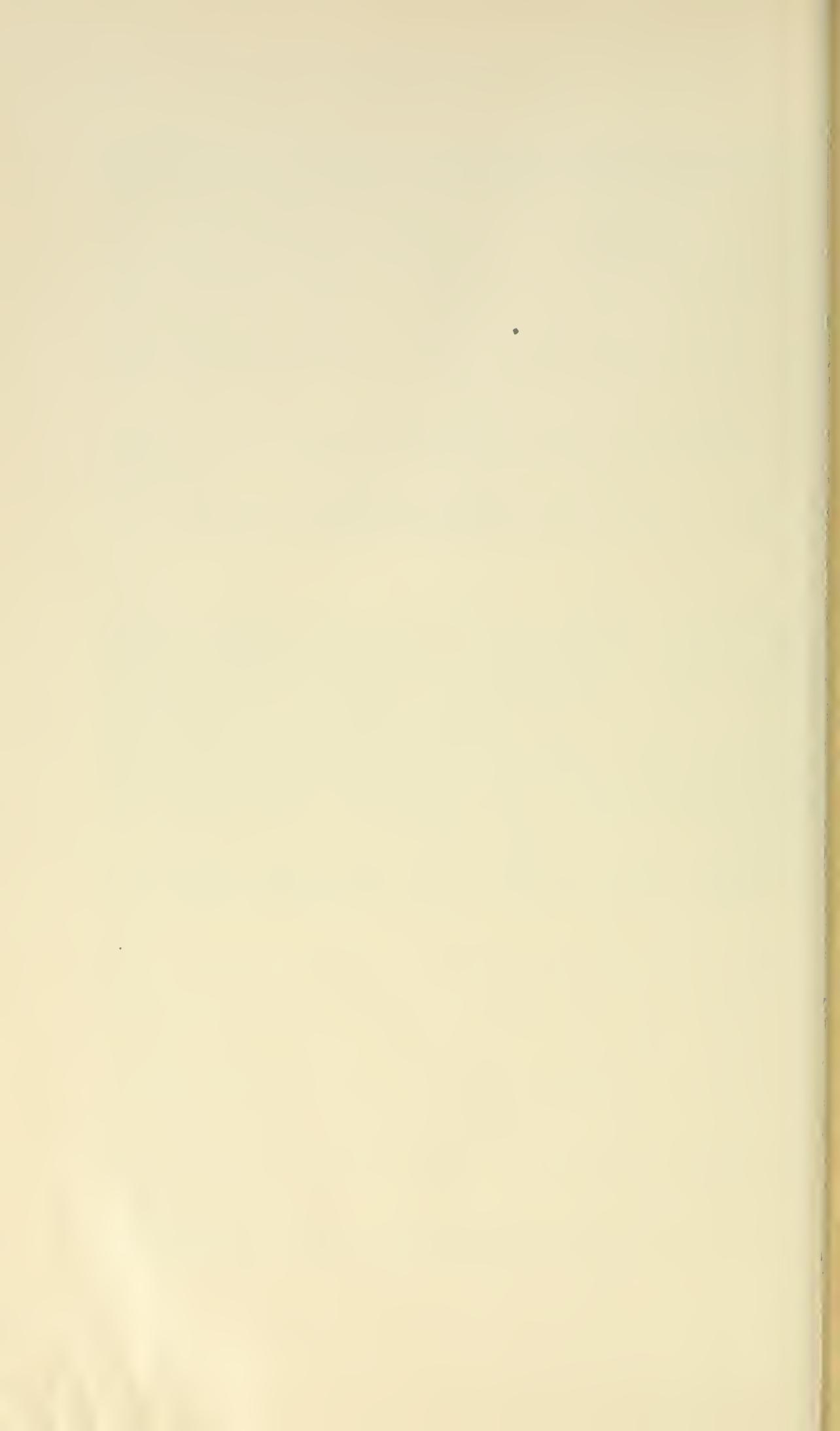
B. A. Kingsbury⁶ has pointed out that the Integral (1) which is tabulated in the phase tables in question is useful in other than the communications and network fields. A bibliography covering other possible fields of interest is given in an article by Murakami and Corrington.⁷

ACKNOWLEDGMENT

The author is indebted to R. W. Hamming of the Mathematical Research Department who supervised the computation of the seven place tables, to Miss R. A. Weiss who planned, programmed, ran, and checked the IBM computations of the tables and to Miss J. D. Goeltz who computed the ten-figure accuracy check points required for the construction of the tables. He also wishes to acknowledge the support and encouragement given to the project by R. L. Dietzold and P. H. Richardson, and the continued interest and helpful comments of B. A. Kingsbury.

⁶ Kingsbury, B. A., private communication.

⁷ Murakami, T., and Corrington, M. S., Relation Between Amplitude and Phase in Electrical Networks, R.C.A. Review, 9, pp. 602-631, Dec., 1948.



Bell System Technical Papers Not Published in This Journal

ANDERSON, P. W.,¹ and SUHL, H.¹

Instability in the Motion of Ferromagnets at High Microwave Power Levels, Phys. Rev., Letter to the Editor, **100**, pp. 1788–1789, Dec. 15, 1955.

ANDRUS, J., see Bond, W. L.

BEACHELL, H. C., see Veloric, H. S.

BECK, A. C.,¹ and MANDEVILLE, G. D.¹

Microwave Traveling Wave Tube Millimicrosecond Pulse Generators, I.R.E. Trans., **MTT-3**, pp. 48–51, Dec., 1955.

BENEDICT, T. S.¹

Single-Crystal Automatic Diffractometer — Part II, Acta Cryst., **8**, pp. 747–752, Dec. 10, 1955.

BENNETT, W. R.¹

Application of the Fourier Integral in Circuit Theory and Circuit Problems, I.R.E. Trans., **CT-2**, **3**, pp. 237–243, Sept., 1955.

BIONDI, F. J.¹

Corrosion-Proofing Electronic Parts Against Ozone, Ceramic Age, **66**, p. 39, Oct., 1955.

BOND, W. L.¹

Single-Crystal Automatic Diffractometer — Part I, Acta Cryst., **8**, pp. 741–746, Dec. 10, 1955.

¹ Bell Telephone Laboratories, Inc.

BOND, W. L.¹ AND ANDRUS, J.¹

Photographs of the Stress Field Around Edge Dislocations, Phys. Rev., Letter to the Editor, **101**, p. 1211, Feb. 1, 1956.

BOYLE, W. S., See Germer, L. H.

BOYLE, W. S.,¹ and HAWORTH, F. E.¹

Glow-to-Arc Transitions, Phys. Rev., **101**, pp. 935-938, Feb. 1, 1956.

BOZORTH, R. M.¹

The Physics of Magnetic Materials, Elec. Engg., **75**, pp. 134-140, Feb. 1956.

BRIDGERS, H. E.¹

A Modern Semiconductor — Single Crystal-Germanium, Chem. and Engg. News, **34**, p. 220, Jan., 1956.

BURRUS, C. A.,¹ and GORDY, W.⁵

Millimeter and Submillimeter Wave Spectroscopy, Phys. Rev., **101**, pp. 599-603, Jan. 15, 1956.

CHYNOWETH, A. G.¹

Dynamic Method for Measuring the Pyroelectric Effect with Special Reference to Barium Titanate, J. Appl. Phys., **27**, pp. 78-84, Jan., 1956.

CUTLER, C. C.¹

Spurious Modulation of Electron Beams, Proc. I.R.E., **44**, pp. 61-64, Jan., 1956.

DAVIS, H. M., see Wernick, J. H.

DUNCAN, R. A.,¹ and STONE, J. A., JR.¹

A Survey of the Application of Ferrites to Inductor Design, Proc. I.R.E., **44**, pp. 4-13, Jan., 1956.

¹ Bell Telephone Laboratories, Inc.

⁵ Duke University.

FEHER, G.,¹ FLETCHER, R. C.,¹ and GERE, E. A.¹

Exchange Effects in Spin Resonance of Impurity Atoms in Silicon,
Phys. Rev., Letter to the Editor, **100**, pp. 1784–1785, Dec. 15, 1955.

FELDMANN, W. L., see Pearson, G. L.

FEWER, D. R.¹

Design Principles for Junction Transistor Audio Power Amplifiers,
I.R.E. Trans., **AU-3**, pp. 183–201, Nov.–Dec., 1955.

FLASCHEN, S. S.,¹ and VAN UITERT, L. G.¹

New Low Contact Resistance Electrode, J. Appl. Phys., Letter to the
Editor, **27**, p. 190, Feb., 1956.

FLETCHER, R. C., see Feher, G.

FRY, T. C.¹

Mathematics as a Profession Today in Industry, Am. Math. Monthly,
63, pp. 71–80, Feb., 1956.

FULLER, C. S., see Reiss, H.

GEBALLE, T. H., see Hrotowski, H. J.

GERE, E. A., see Feher, G.

GERMER, L. H.,¹ and BOYLE, W. S.¹

Short Arcs, Nature, Letter to the Editor, **176**, p. 1019, Nov. 26, 1955.

GERMER, L. H.,¹ and BOYLE, W. S.¹

Two Distinct Types of Short Arcs, J. Appl. Phys., **27**, pp. 32–39, Jan.,
1956.

GIANOLA, U. F.¹

Photovoltaic Noise in Silicon Broad Area p-n Junctions, J. Appl.
Phys., **27**, pp. 51–53, Jan., 1956.

GORDY, W., see Burrus, C. A.

¹ Bell Telephone Laboratories, Inc.

HAGELBARGER, D. W., see Pfann, W. G.; Shannon, C. E.; and Williams, H. J.

HAGSTRUM, H. D.¹

Electron Ejection from Metals by Positive Ions, Appl. Sci. Res. B5, Nos. 1-4, pp. 16-17, 1955.

HAWORTH, F. E., see Boyle, W. S.

HERRING, C.¹ and VOGT, E.¹

Transport and Deformation Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering, Phys. Rev., 101, pp. 944-961, Feb. 1, 1956.

HERRMANN, D. B., see Williams, J. C.

HOLDEN, A. N.,¹ MERZ, W. J.,¹ REMEIKA, J. P.,¹ and MATTHIAS, B. T.¹

Properties of Guanidine Aluminum Sulfate Hexahydrate and Some of its Isomorphs, Phys. Rev., 101, pp. 962-967, Feb. 1, 1956.

HOROTOWSKI, H. J.,¹ MORIN, F. J.,¹ GEBALLE, T. H.,¹ and WHEATLEY, G. H.¹

Hall Effect and Conductivity of InSb, Phys. Rev., 100, pp. 1672-1677, Dec. 15, 1955.

INGRAM, S. B.¹

The Graduate Engineer His Training and Utilization in Industry, Elec. Engg., 75, pp. 167-170, Feb., 1956.

KAPLAN, E. L.¹

Transformation of Stationary Random Sequences, Mathematica Scandinavica, 3, FASCI, pp. 127-149, June, 1955.

LEWIS, H. W.¹

Superconductivity and Electronic Specific Heat, Phys. Rev., 101, pp. 939-940, Feb. 1, 1956.

MANDEVILLE, G. D., see Beck, A. C.

¹ Bell Telephone Laboratories, Inc.

MATTHIAS, B. T., see Holden, A. N.

MERZ, W. J., see Holden, A. N.

MILLER, L. E.¹

Negative Resistance Regions in the Collector Characteristics of the Point-Contact Transistor, Proc. I.R.E., **44**, pp. 65-72, Jan., 1956.

MOLL, J. L.,¹ and ROSS, I. M.¹

The Dependence of Transistor Parameters on the Distribution of Base Layer Resistivity, Proc. I.R.E., **44**, pp. 72-78, Jan., 1956.

MONTGOMERY, H. C., See Pearson, G. L.

MORIN, F. J., see Hrotowski, H. J.

MUMFORD, W. W.,¹ and SCHAFERMAN, R. L.¹

Data on the Temperature Dependence of X-Band Fluorescent Lamp Noise Sources, I.R.E. Trans., MTT-3, pp. 12-16, Dec., 1955.

NESBITT, E. A., see Williams, H. J.

OLMSTEAD, P. S.¹

QC Concepts Useful in OR, Ind. Qual. Cont., **12**, pp. 11, 14-17, Oct., 1955.

OWENS, C. D.¹

Stability Characteristics of Molybdenum Permalloy Powder Cores, Elec. Engg., **74**, pp. 252-256, Feb., 1956.

PEARSON, G. L.,¹ MONTGOMERY, H. C.,¹ and FELDMANN, W. L.¹

Noise in Silicon p-n Junction Photocells, J. Appl. Phys., **27**, pp. 91-92, Jan., 1956.

PFANN, W. G.,¹ and HAGELBARGER, D. W.¹

Electromagnetic Suspension of a Molten Zone, J. Appl. Phys., **27**, pp. 12-17, Jan., 1956.

¹ Bell Telephone Laboratories, Inc.

QUINLAN, A. L.³

Roll-Welding Precious Metals for Telephone Contacts, Elec. Engg., **75**, pp. 154-157, Feb., 1956.

REISS, H.,¹ and FULLER, C. S.¹

The Influence of Holes and Electrons on the Solubility of Lithium in Boron-Doped Silicon, J. of Metals, **12**, p. 276, Feb., 1956.

REMEIKA, J. P., see Holden, A. N.

ROSS, I. M., see Moll, J. L.

SCHAFFERMAN, R. L., see Mumford, W. W.

SCHAWLOW, A. L.¹

Structure of the Intermediate State in Superconductors, Phys. Rev., **101**, pp. 573-580, Jan. 15, 1956.

SCHAWLOW, A. L.,¹ and TOWNES, C. H.⁴

Effect on X-Ray Fine Structure of Deviations from a Coulomb Field near the Nucleus, Phys. Rev., **100**, pp. 1273-1280, Dec. 1, 1955.

SHANNON, C. E.,¹ and HAGELBARGER, D. W.¹

Concavity of Resistance Functions, J. Appl. Phys., **27**, pp. 42-43, Jan., 1956.

SIMKINS, Q. W.,¹ and WOGELSONG, J. H.¹

Transistor Amplifiers for Use in a Digital Computer, Proc. I.R.E., **44**, pp. 43-54, Jan., 1956.

SNOKE, L. R.¹

Specific Studies on the Soil-Block Procedure for Bioassay of Wood Preservatives, Appl. Microbiology, **4**, pp. 21-31, Jan., 1956.

SOUTHWORTH, G. C.¹

Early History of Radio Astronomy, Sci. Mo., **82**, pp. 55-66, Feb., 1956.

¹ Bell Telephone Laboratories, Inc.

³ Western Electric Company.

⁴ Columbia University.

STONE, H. A., see Duncan, R. A.

SUHL, H., see Anderson, P. W.

THOMAS, E. E.¹

Tin Whisker Studies — Observation of some Hollow Whiskers and Some Sharply Irregular External Forms, Letter to the Editor, *Acta Met.*, **4**, p. 94, Jan., 1956.

TOWNES, C. H., see Schawlow, A. L.

TOWNSEND, M. A.¹

A Hollow Cathode Glow Discharge with Negative Resistance, *Appl. Sci. Research, Sec. B*, **5**, pp. 75–78, 1955.

VALDES, L. B.¹

Frequency Response of Bipolar Transistors with Drift Fields, *Proc. I.R.E.*, **44**, pp. 178–184, Feb., 1956.

VAN UITERT, L. G., see Flaschen, S. S.

VELORIC, H. S.,¹ and BEACHELL, H. C.⁶

Absorption Isotherms, Isobars and Isoteres of Diborane on Palladium on Charcoal and Boron Nitride, *J. Phys. Chem.*, **60**, p. 102, Jan., 1956.

VOGELSONG, J. H., see Simkins, Q. W.

VOGT, E., see Herring, C.

WEIBEL, E. S.¹

Strains and the Energy in Thin Elastic Shells of Arbitrary Shape for Arbitrary Deformation, *Zeitchrift f. Mathematik and Physik*, **6**, pp. 153–189, May 25, 1955.

WERNICK, J. H.,¹ and DAVIS, H. M.⁷

Preparation and Inspection of High-Purity Copper Single Crystals, *J. Appl. Phys.*, **27**, pp. 144–153, Feb., 1956.

¹ Bell Telephone Laboratories, Inc.

⁶ University of Delaware.

⁷ Penn State University.

WHEATLEY, G. H., see Hrotowski, H. J.

WILLIAMS, H. J.,¹ HEIDENREICH, R. D.,¹ and NESBITT, E. A.¹

Mechanism by which Cobalt Ferrite Heat Treats in a Magnetic Field,
J. Appl. Phys., **27**, pp. 85-89, Jan., 1956.

WILLIAMS, J. C.,¹ and HERRMANN, D. B.¹

Surface Resistivity of Non-Porous Ceramic and Organic Insulating Materials at High Humidity with Observations of Associated Silver Migration, I.R.E. Trans., PGRQC-6, pp. 11-20, Feb., 1956.

WOOD, MRS. E. A.¹

A Heated Sample-Holder for X-Ray Diffractometer Work, Rev. Sci. Instr., **27**, p. 60, Jan., 1956.

¹ Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ANDERSON, P. W., and HASEGAWA, H.

Considerations on Double Exchange, Monograph 2532.

BAKER, W. O., see Winslow, F. H.

BARSTOW, J. M.

The ABC's of Color Television, Monograph 2529.

BEMSKI, G.

Lifetime of Electrons in p-type Silicon, Monograph 2534.

BENNETT, W. R.

Application of the Fourier Integral in Circuit Theory, Monograph 2533.

BRATTAIN, W. H., see Pearson, G. L.

BROWN, W. L.

Surface Potential and Surface Charge Distribution from Semiconductor Field Effect Measurements, Monograph 2501.

BULLINGTON, K.

Characteristics of Beyond-the-Horizon Radio Transmission, Monograph 2494.

BULLINGTON, K., INKSTER, W. J., and DURKEE, A. L.

Propagation Tests at 505 mc and 4,090 mc on Beyond-Horizon Paths, Monograph 2503.

* 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.

DURKEE, A. L., see Bullington, K.

FREYNIK, H. S., see Gohn, G. R.

GELLER, S., and THURMOND, C. D.

On the Question of the Existence of a Crystalline SiO, Monograph 2536.

GOHN, G. R., GUERRARD, J. P., and FREYNIK, H. S.

The Mechanical Properties of Wrought Phosphor Bronze Alloys, Monograph 2531.

GUERRARD, J. P., see Gohn, G. R.

HASEGAWA, H., see Anderson, P. W.

HAYNES, J. R., see Hornbeck, J. A.

HORNBECK, J. A., and HAYNES, J. R.

Trapping of Minority Carriers in Silicon, Monograph 2368.

INKSTER, W. J., see Bullington, K.

LEWIS, H. W.

Search for the Hall Effect in a Superconductor. II. Theory, Monograph 2523.

LINVILL, J. G., and MATTSON, R. H.

Junction Transistor Blocking Oscillators, Monograph 2487.

LOGAN, R. A.

Precipitation of Copper in Germanium, Monograph 2524.

LOGAN, R. A., and SCHWARTZ, M.

Restoration of Resistivity and Lifetime in Heat-Treated Germanium, Monograph 2525.

MATTSON, R. H., see Linvill, J. G.

MAYS, J. M., see Shulman, R. G.

McCALL, D. W., see Shulman, R. G.

MOLL, J. L.

Junction Transistor Electronics, Monograph 2537.

PEARSON, G. L., and BRATTAIN, W. H.

History of Semiconductor Research, Monograph 2538.

SANDSMARK, P. I.

Ellipticity on Dominant-Mode Axial Ratio in Nominally Circular Waveguides, Monograph 2539.

SCHWARTZ, M., see Logan, R. A.

SHULMAN, R. G., MAYS, J. M., and McCALL, D. W.

Nuclear Magnetic Resonance in Semiconductors. I, Monograph 2528.

THURMOND, C. D., see Geller, S.

VAN UITERT, L. G.

Low Magnetic Saturation Ferrites for Microwave Applications, Monograph 2504.

VAN UITERT, L. G.

Dc Resistivity in the Nickel and Nickel Zinc Ferrite System, Monograph 2540.

WEIBLE, E. S.

Vowel Synthesis by Means of Resonant Circuits, Monograph 2541.

WINSLOW, F. H., BAKER, W. O., and YAGER, W. A.

Odd Electrons in Polymer Molecules, Monograph 2486.

YAGER, W. A., see Winslow, F. H.

Contributors to This Issue

DONALD C. BENNETT, B.S. 1949 and M.S. 1951, Rensselaer Polytechnic Institute; Battelle Memorial Institute, 1951-1952; Bell Telephone Laboratories, 1952-. Mr. Bennett has been engaged in the development of processes for producing single crystals suitable for use in transistors. He is a member of the American Institute of Mining and Metallurgical Engineers.

F. G. BUHRENDORF, B.S.M.E. and M.E., Cooper Union Inst. Tech. 1925. Bell Telephone Laboratories 1925-. Mr. Buhrendorf's early Laboratories work included the design of switchboard apparatus and sound recording and reproducing equipment; among the latter were the Mirrophone and the stereophonic equipment demonstrated at the New York World's Fair. During World War II he was concerned with the design of mechanical components of a number of radar systems, particularly antenna drives and range units. After the war he resumed his work on high-quality sound reproduction and more recently has devoted his efforts to the design of magnetic drum units for digital data storage and special machinery for the purification and production of single-crystal semiconductors. He is a New York State Professional Engineer.

CALVIN S. FULLER, B.S. 1926 and Ph.D. 1929, University of Chicago. Bell Telephone Laboratories, 1930-. His early work was on organic insulating material, after which he made studies of plastics and synthetic rubber including investigations of the molecular structure of polymers and the development of plastics and rubbers. Since 1948 Dr. Fuller has concentrated on semiconductor research and the development of semiconductor devices. His work led to a technique of diffusing impurities into the surface of a silicon wafer, a preparation basic to the Bell Solar Battery and other silicon devices. He is a member of the A.C.S., an associate member of the A.P.S. and a member of the A.A.A.S.

H. A. HENNING, B.S. in Electrochemical Engineering, Pennsylvania State College 1926; Columbia University 1930-33. Bell Telephone

Laboratories, 1926-. Mr. Henning's early Laboratories work was connected with the development of high-quality sound recording and reproducing equipment and techniques. During this interval he developed the 9A disc phonograph reproducer. Other pre-war experience included development of telephone voice recorders, noise reduction studies of the dynamics of teletype equipment, and design of coin collector slug rejectors and coin disposal relays. During World War II he was concerned with improvements to the sound power telephone, and later with development of specialized magnetic sound recording-reproducing systems. After the war he resumed his work on high quality sound recording equipment and supervised the design of the 2A lateral disc feedback recorder. More recently he has been concerned with the principles and design of magnetic drum digital data storage and apparatus. He is currently engaged in investigating the application of square hysteresis loop magnetic cores to digital computer systems.

DAVID, A. KLEINMAN, S.B. in Chemical Engineering, 1946, S.M. in Mathematics, 1947, Massachusetts Institute of Technology; Ph.D. in physics, Brown University, 1952. Dr. Kleinman joined Bell Telephone Laboratories at Murray Hill in July, 1953. Since then he has studied theory of transistor devices and has been engaged in research in the band theory of solids in the Solid State Electronics Research Department. He is a member of the American Physical Society.

F. J. MORIN, B.S. and M.S., University of New Hampshire, 1939 and 1940; University of Wisconsin, 1940-1941; Bell Telephone Laboratories, 1941-. During World War II, Mr. Morin was involved in research on elemental and oxide semiconductors and the development of thermistor materials. Since that time he has worked on fundamental investigations into the mechanism of conduction in silicon, germanium and oxide semiconductors. Mr. Morin is a member of the American Chemical Society and the American Physical Society.

O. J. MURPHY, B.S. in Electrical Engineering, University of Texas, 1927; Columbia University, 1928-31. Bell Telephone Laboratories, 1927-. Mr. Murphy's early Laboratories projects included studies of voice-operated switching devices, effects of transmission delay on two-way telephone conversation, and voice-frequency signaling systems. During World War II he was concerned with design and development of the M-9 electrical gun director and related projects. After the war he resumed his research work on signaling systems and more recently has

concentrated on the design of magnetic drum digital data storage apparatus and circuits. He is a member of the A.I.E.E., a senior member of the I.R.E., and is a licensed professional engineer.

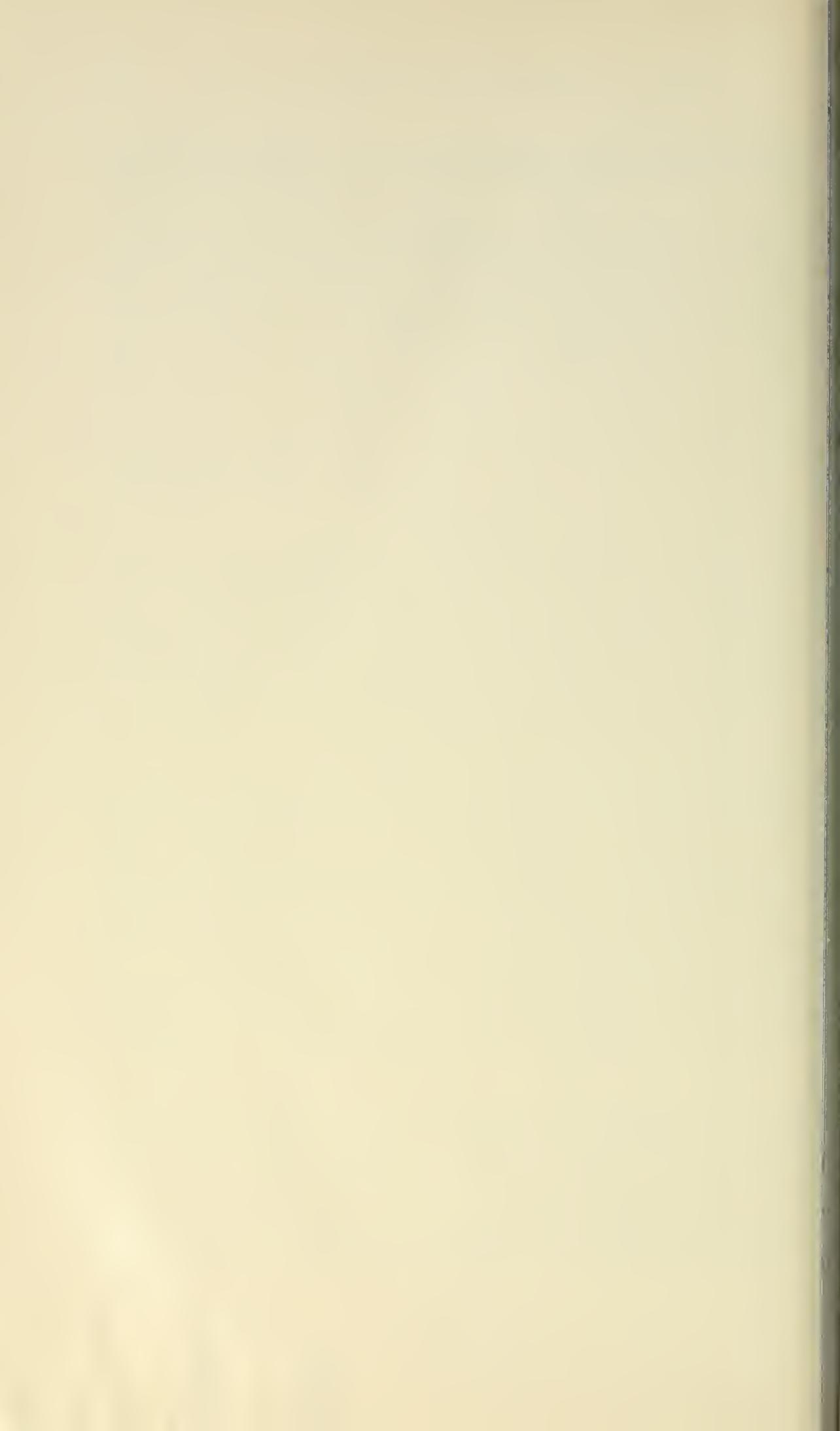
M. B. PRINCE, A.B., Temple University, 1947; Ph.D., Massachusetts Institute of Technology, 1951; Bell Telephone Laboratories, 1951-1956; National Semiconductor Products, 1956-. Between 1949-51 he was a research assistant at the Research Laboratory of Electronics at M.I.T. where he was concerned with eryogenic research. At Bell Telephone Laboratories, Dr. Prince was concerned with the physical properties of semiconductors and semiconductor devices and was associated with the development of silicon devices, including the Bell Solar Battery and the silicon power rectifier. Dr. Prince is a member of the I.R.E., the American Physical Society, and Sigma Xi.

HOWARD REISS, B.A., New York University, 1943; Ph.D., Columbia University, 1949; Instructor and Assistant Professor in Chemistry, Boston University, 1949-51; Head of the Fundamental Research Section, Celanese Corporation, 1951-52; Bell Telephone Laboratories, 1952-. Dr. Reiss is engaged in the theoretical chemistry of defects in semiconductors. He is a member of the American Chemical Society, the American Physical Society, Sigma Xi and Phi Lambda Upsilon.

BALDWIN SAWYER, B.E., Yale University, 1943; D.Sc., Carnegie Institute of Technology, 1952; Manhattan Project, University of Chicago, 1943-1946; Instructor and Research Associate in Physics, Carnegie Institute of Technology, 1948-1951; Bell Telephone Laboratories, 1951-. Dr. Sawyer's first work at the Laboratories was on the development of semiconductor devices, especially the silicon alloy junction diode. Since 1953 he has been in charge of a group at Allentown concerned with the growth, measurement and characterization of germanium and silicon crystals for use in semiconductor devices. He is a member of the American Physical Society, the American Institute of Mining and Metallurgical Engineers, Tau Beta Pi, Sigma Xi, and an associate of the I.R.E.

DONALD E. THOMAS, B.S. in E.E., Pennsylvania State University, 1929; M.A., Columbia University, 1932; Bell Telephone Laboratories, 1929-. Mr. Thomas specialized in the development of repeatered submarine cable systems until 1940 when he became engaged in the development of sea and airborne radar. In 1942 he entered military service where he was active in electronic countermeasures research and development.

Following the war he took part in the development and installation of the first deep-sea repeatered submarine telephone cable system between Key West and Havana. During this period he also served as a civilian member of the Department of Defense's Research and Development Board Panel on Electronic Countermeasures. At present Mr. Thomas is engaged in characterization and feasibility evaluation of research models of semiconductor devices. He is a senior member of the I.R.E. and a member of Tau Beta Pi and Phi Kappa Phi.



THE BELL SYSTEM *Technical Journal*

DEDICATED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXV

JULY 1956

AUG 2 NUMBER 4

- The Effect of Surface Treatments on Point-Contact Transistors
J. H. FORSTER AND L. E. MILLER 767

- The Design of Tetrode Transistor Amplifiers
J. G. LINVILL AND L. G. SCHIMPF 813

- The Nature of Power Saturation in Traveling Wave Tubes
C. C. CUTLER 841

- The Field Displacement Isolator S. WEISBAUM AND H. SEIDEL 877

- Transmission Loss Due to Resonance of Loosely-Coupled Modes in a
Multi-Mode System A. P. KING AND E. A. MARCATILI 899

- Measurement of Atmospheric Attenuation at Millimeter Wave-
lengths A. B. CRAWFORD AND D. C. HOGG 907

- A New Interpretation of Information Rate J. L. KELLY, JR. 917

- Automatic Testing of Transmission and Operational Functions
of Intertoll Trunks

H. H. FELDER, A. J. PASCARELLA AND H. F. SHOFFSTALL 927

- Intertoll Trunk Net Loss Maintenance Under Operator Distance
and Direct Distance Dialing

H. H. FELDER AND E. N. LITTLE 955

-
- Bell System Technical Papers Not Published in This Journal 973

- Recent Bell System Monographs 979

- Contributors to This Issue 985

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

B. McMILLAN, <i>Chairman</i>	R. K. HONAMAN
A. J. BUSCH	H. R. HUNTLEY
A. C. DICKIESON	F. R. LACK
R. L. DIETZOLD	J. R. PIERCE
K. E. GOULD	H. V. SCHMIDT
E. I. GREEN	G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

JULY 1956

NUMBER 4

Copyright 1956, American Telephone and Telegraph Company

The Effect of Surface Treatments on Point-Contact Transistor Characteristics

By J. H. FORSTER and L. E. MILLER

(Manuscript received January 23, 1956)

A description is given of the electrical properties of formed point contacts on germanium. A useful technique for observation of the equipotentials surrounding such contacts is described. The contrasting properties of donor-free and donor-doped contacts, used as diodes or transistor collectors are emphasized.

It is shown that unformed point contacts (which have electrical properties largely determined by a surface barrier layer), may exhibit analogous differences. Such changes are produced by chemical treatments calculated to influence properties of a soluble germanium oxide film on the surface.

The above information is applied to a study of transistor forming as it is done in present point-contact transistor processing. It is shown that high yields from the forming process can be expected on oxidized surfaces, and that chemical washes which remove soluble germanium oxide drastically lower forming yields. These and other effects are evaluated as sources of variability in forming yield.

TABLE OF CONTENTS

1. Introduction.....	768
2. Properties of Formed Point Contacts.....	770
2.1 Effects of Electrical Forming on Point Contacts.....	770
2.2 Donor-Free and Donor-Doped Contacts.....	774

2.2.1 Potential Probes.....	774
2.2.2 Use of the Copper Plating Technique.....	776
2.3 Under-Formed and Over-Formed Contacts.....	781
3. Properties of Unformed Point Contacts.....	783
3.1 Physical Properties of Metal-Semiconductor Contacts.....	783
3.2 Experimental Procedures.....	785
3.3 Experimental Results.....	786
3.3.1 Unformed Transistors on Superoxol-Etched Surfaces.....	786
3.3.2 Unformed Transistors on CP ₄ -Etched Surfaces.....	789
3.3.3 Diode Characteristics on Electro-Etched Surfaces.....	789
3.3.4 Output Characteristic Anomalies.....	789
3.3.5 Floating Potential Measurements.....	790
3.3.6 Contamination of Collector Points and Surfaces.....	792
3.4 Discussion of Experimental Results.....	794
3.4.1 Effects of the Chemical Treatment on the Superoxol-Etched Surfaces.....	794
3.4.2 CP ₄ -Etched Surfaces.....	795
4. Relation of Germanium Surface Properties to Transistor Forming.....	796
4.1 Pilot Production Problems.....	796
4.2 Experimental Results.....	797
4.2.1 Pilot Process Forming Yields.....	797
4.2.2 Relation of Unformed Diode Characteristics to Transistor "Formability".....	801
4.2.3 Controlled Ambient Experiments.....	804
4.2.4 A Statistical Survey Experiment on Transistor Forming.....	805
4.2.5 Effect of Contamination Before Etching.....	806
4.3 Conclusions.....	807
5. General Concluding Remarks.....	808
5.1 Point-Contact Transistors with High Current Gain.....	809
5.2 Current Multiplication in Unformed Transistors.....	809
5.3 Surface Properties and Transistor Forming.....	810

1. INTRODUCTION

The point-contact transistor, on the basis of several years use in the field in Bell System applications, has proved itself to be rugged and dependable. For certain military applications, a lasting demand exists for high-speed point-contact transistors. The adaptation of cartridge type units to a hermetically sealed structure has been completed, with further benefits to reliability. To date, the point-contact transistor is one of the few transistors to successfully pass all military specifications for shock, vibration, and high acceleration. Thus, although there are at present limitations to the electrical characteristics that can be built into a point-contact transistor which make it unsuitable for use in some switching circuits, there are many applications in which this type of transistor can give consistent and reliable performance. In fact, applications exist wherein the specific requirements are uniquely satisfied by the point-contact transistor.

However, the basic operational principles of this kind of device are not as well understood as would be desirable for facilitating developmental studies for manufacture. Although considerable effort has been

expended towards the analysis and understanding of the physical mechanisms of the point-contact transistor since its announcement in 1948, a complete design theory for these transistors is not available. This lack probably results partially from a more general interest in the readily designable junction transistor types, and partially from the relative complexity of the device itself. Actually the physical mechanisms which account for the operation of this device have their counterparts in at least three basically unique devices: the point diode, the junction transistor, and the filamentary transistor.

Thus, although the empirical knowledge of point-contact transistor design and operation is large enough to allow a reasonable degree of designability, and manufacture of these transistors in large quantities is possible, there are, from time to time, manufacturing problems which are often difficult to solve without sound theoretical understanding of the physical mechanisms which make the device work.

This article is concerned with describing the results of a general study of the physical properties of a few specific kinds of point contacts. The kinds of contact studied have been those of specific interest to those concerned with manufacture and processing of point-contact transistors. This investigation was conducted in parallel with the final development for manufacture of the hermetically sealed point-contact transistor. The study of these properties has led to practical solutions of several problems encountered during manufacture of point-contact transistors, and has provided experimental data which is of interest in consideration of the basic physical mechanisms involved in the operation of the point-contact transistor.

The work to be described, primarily experimental in nature, follows in Sections 2, 3 and 4. In section 2, the properties of formed, or electrically pulsed point contacts, and their relation to the source of output characteristic anomalies often responsible for lowering forming yields in point-contact transistor production is discussed. The properties of point contacts which have received no electrical forming in the conventional sense are considered in section 3. The electrical properties of these contacts, used as diodes or transistor collectors, are shown to be dependent on chemical history of the etched germanium surface. Thus "chemical forming" of point contacts is possible. Section 4 deals with application of these results to forming problems which arise during manufacture of point-contact transistors. The important relation between the chemical history of the surface and the forming on that surface is considered.

2. PROPERTIES OF FORMED POINT CONTACTS

2.1 Effects of Electrical Forming on Point Contacts

The simplest form of point-contact transistor collector is a metal to semiconductor contact which has not been subjected to excessive power dissipation either in short high energy pulses, or in the form of more prolonged aging at lower power levels. Such contacts will be referred to as unformed contacts, and their properties will be discussed in detail in Section 3. Unformed point-contact transistors sometimes exhibit power gain, but in general they are not suitable for use as active devices because the gain, although it may be highly variable from unit to unit, is usually low. The electrical characteristics of such contacts depend on a metal-semiconductor contact at the semiconductor surface, and control of these properties requires exacting control of surface preparation, surrounding ambient, and mechanical stability of the point.

In early experiments, Brattain¹ used electrical forming to improve both the power gain and stability of the transistor. For present purposes, the process of electrical forming will be defined as the passage of a short pulse of reverse current through a point contact which produces permanent changes in the electrical properties of the contact. This is usually accomplished by charging a condenser to several hundred volts, and subsequently discharging it through a resistor in series with the transistor collector. Bardeen and Pfann,² investigating electrical forming of phosphor bronze points on etched germanium surfaces, indicate, as a possible explanation of their data, that the forming pulse changes the height of the potential barrier at the germanium surface. This would, in absence of large surface conductivity, increase the reverse current through the point and increase the efficiency of hole collection by the point.³ Thus, the formed point may, according to theory,⁴ act as a collector with a current multiplication (α) greater than unity. Thermal and potential probing of an *n*-germanium surface under a formed phosphor bronze point indicates, according to Valdes,⁵ that an appreciable volume of germanium is converted to *p*-type conduction. Thus, the reverse current through a formed point probably depends on the characteristics of a *p-n* junction a small distance from the point, rather than on a potential barrier at the germanium surface.

A characteristic of the point-contact transistor is that the current gain can be substantially greater than unity. The current gain, α , is usually defined as the current multiplication at constant voltage, that is:

$$\alpha = -\frac{\partial I_c}{\partial I_e} \Big|_{V_c} \quad (1)$$

where I_c and I_e are the collector and emitter currents. The α can be considered as the product of three terms, that is:

$$\alpha = \alpha_i \beta \gamma \quad (2)$$

where γ and β represent the injection efficiency and transport factor respectively for minority carriers. The term α_i is the "intrinsic" current multiplication of the collector itself. As mentioned above, there are theoretical reasons to account for an α_i as large as $(1 + b)$, where b is the ratio μ_n/μ_p of the mobilities of electrons and holes, and thus the term α_i may be roughly as large as 3.1. The average current gain, $\bar{\alpha}$, taken over a large interval of emitter current, is seldom found to be greater than this value, and is usually about 2.5. However, the small signal α at low emitter current usually is found to be considerably larger than 3.1.

Several mechanisms have been proposed to account for this excess current gain at low emitter bias in formed transistors. The most generally known of these are the p-n hook hypothesis and the trapping model.^{6, 7}

The experiments to be described in this section will be concerned primarily with the characteristics of formed points as transistor collectors, and thus with the transport factor β . The subject of the origin of the intrinsic α_i will be discussed further in a later section.

The experiment of Valdes indicates that the properties of a formed point contact depend on the physical properties of a small region of germanium near the point, produced by impurity diffusion from the point or imperfections introduced during the forming pulse. A highly idealized representation of the physical situation is shown in Fig. 1. This is a radial model of a formed point contact on a semi-infinite block of n -germanium (respectively ρ), with a hemispherical p -layer (radius $\simeq r_0$). The electron and hole concentrations in the formed layer near the junction are designated as n_p and p . If a reverse bias V_c is applied to the point, a potential difference $V(r_1) - V(r_2) = V_J$ results from the resistance of the junction at r_0 . For $r \geq r_2$, at distances well outside r_0 , the potential $V(r)$ and the magnitude of the field $E(r)$ are given by

$$V = \frac{\rho I}{2\pi r} \quad (3)$$

$$E = \frac{\rho I}{2\pi r^2} \quad (4)$$

where I is the total current through the point. For

$$|V_c - V(r_1)| \ll |V_J|, \quad V(r_2) \simeq V_c - V_J,$$

and the junction resistance limits the magnitude of the drift field that can be set up near the point. For example, if the lifetime τ_n of electrons in the p -layer⁸ is substantially lower than τ_p , that of holes in the germanium bulk, the reverse current density across the junction can be increased by an increase in n_p , and junction resistance lowered.

Pfann⁹ reports a substantial increase in the reverse current of formed point contacts with donor concentration of the point wire. The increase in n_p will depend on the distribution of donors in the p -layer after the forming pulse. A high donor concentration near the collector point may

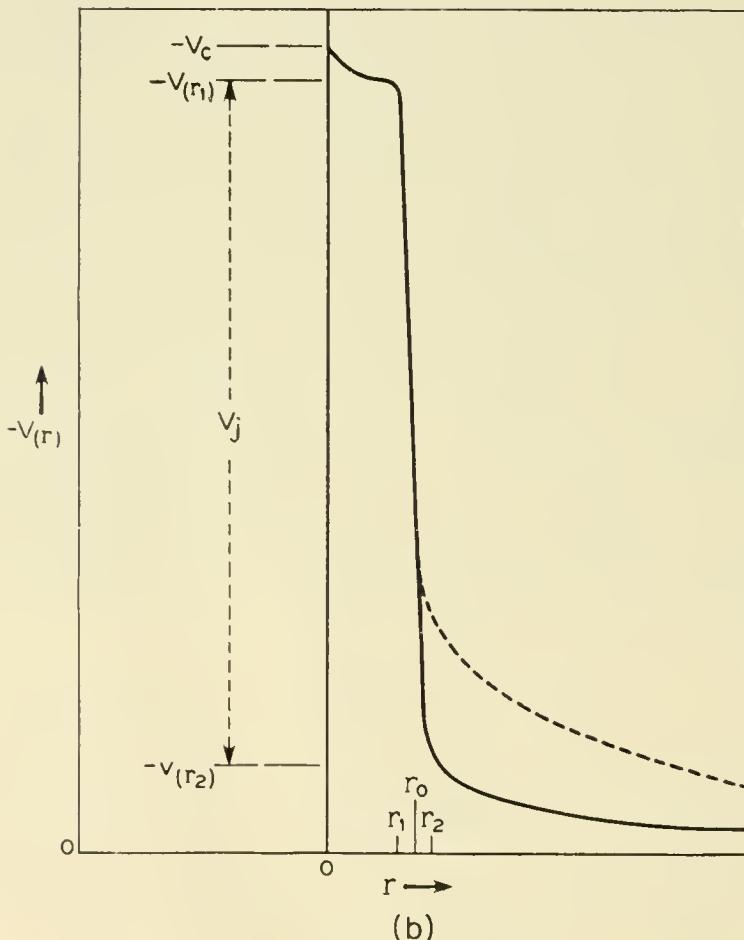
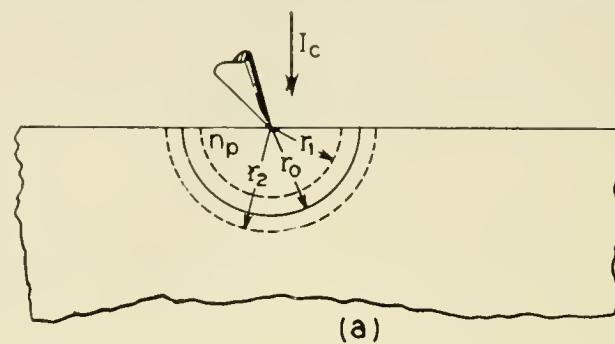


Fig. 1 — Formed point contact under reverse bias — schematic representation.

form an *n*-type inversion layer under the point (p-n hook) which, when the point is under reverse bias, acts as an electron emitter. Such a situation might arise as a result of diffusion of impurities from the collector point at the high temperature reached during the forming pulse. An acceptor element, such as copper, with a high diffusion coefficient¹⁰ might penetrate substantially farther into the germanium than donor elements such as phosphorous or antimony¹¹ with lower diffusion constants. Thus, the donor concentration near the point might be substantially higher than the acceptor concentration if the solubility of the acceptor element is low.

On the other hand, an appreciable number of donor atoms may penetrate the germanium as far as do the acceptors. Thus, the equilibrium value of n_p may be increased simply by decreasing the effective concentration of acceptors in the *p*-layer. Such a case might arise when a collector point such as copper is doped with a suitable amount of a donor element with a large diffusion coefficient and limited solubility.

The observation of regions of melted germanium¹² under heavily formed points gives evidence for a somewhat different interpretation of the forming process. It has been suggested¹³ that forming is essentially a remelt process. For example, forming of a phosphor-bronze point may produce a copper-germanium eutectic, allowing the introduction of a sizeable phosphorus concentration in the remelt region which is maintained after freezing. Thus the depth of penetration of the donor element depends upon the size of the remelt region, and the penetration of the acceptor element depends upon its solid state diffusion coefficient. This mechanism can lead either to the formation of a p-n hook, or at least to a layer of *p*-germanium with a high equilibrium electron concentration.

Whatever the reason for the decrease in resistance of the collector barrier, if it is sufficient, the magnitude of $E(r)$ for $r > r_2$ can be increased by forming to sufficient value to ensure efficient collection of holes and a transport factor β close to unity.

It would then be expected that for a formed donor-free point, such as the beryllium-copper alloy points often used as unformed emitters, the formed *p*-region would have a high acceptor concentration, n_p would be small, and under reverse bias, the magnitude of V_J would be large, with $|I_{co}|$, $|V(r_2)|$, and average α small, [solid curve, Fig. 1(b)]. On the other hand, a formed phosphor bronze point of the kind conventionally used to make transistor collectors, should exhibit under reverse bias, a lesser magnitude of V_J , with $|I_{co}|$, $|V(r_2)|$, and α as much as an order of magnitude larger, (dashed line in Fig. 1(b)].

2.2 Donor-Free and Donor-Doped Contacts

The qualitative picture of the conventional formed contact given above has been substantially supported by the work of Valdes, who observed a large increase in floating potential near the reverse biased collector after the forming pulse and a substantial *p*-region in the bulk of the germanium after forming.

Experiments have been directed to a comparison of the properties as diodes and collectors, between two kinds of points. Phosphor bronze points of the type used as transistor collectors, and beryllium copper points, normally used as emitters, were investigated. Thus a direct comparison can be made between donor-doped and donor-free points which have been given similar forming pulses. The forming pulses were of the capacitor discharge type, with voltage and RC values similar to those used in conventional transistor forming. The points used were of the cantilever variety, and the *n*-germanium was zone-leveled material in the 3 to 4 ohm-cm range. Two points were supported in a double-ended micro-manipulator which allowed freedom of movement in 3 dimensions for each point.

2.2.1 Potential Probes

Conventionally, point-contact transistors are made on a superoxol-etched wafer. This etch leaves a rough surface which is unsuitable for accurate potential probing. Some measurements were made of the floating potentials on this kind of surface, but accurate results were difficult to obtain. In a later section it is shown that the kind of etch used in surface preparation can have profound effects on the degree of forming obtained. However, it is shown that forming characteristics of an "aged" CP₄-etched surface are quite similar to the superoxol surface. Thus this kind of surface was used, since its topographical uniformity allows very reproducible results in the measurement of floating potentials.

Fig. 2 is a comparison of the floating potentials for the two kinds of transistor points examined. The log-log plot shows the magnitude of the floating potential, V_p , near the reverse biased collector as a function of r , the distance of the probe from the collector measured between centers of the two points. The bars represent the uncertainty in measurement of the linear distance. Three curves are shown. The lowest Curve I represents the potential near a Be-Cu point formed with a conventional forming pulse. Curve II is a plot of the potential near a similarly formed phosphor bronze point, while Curve III represents data obtained using such a point more heavily formed. In all cases the magnitude of the floating potential decreases inversely as the distance from the point, and is given

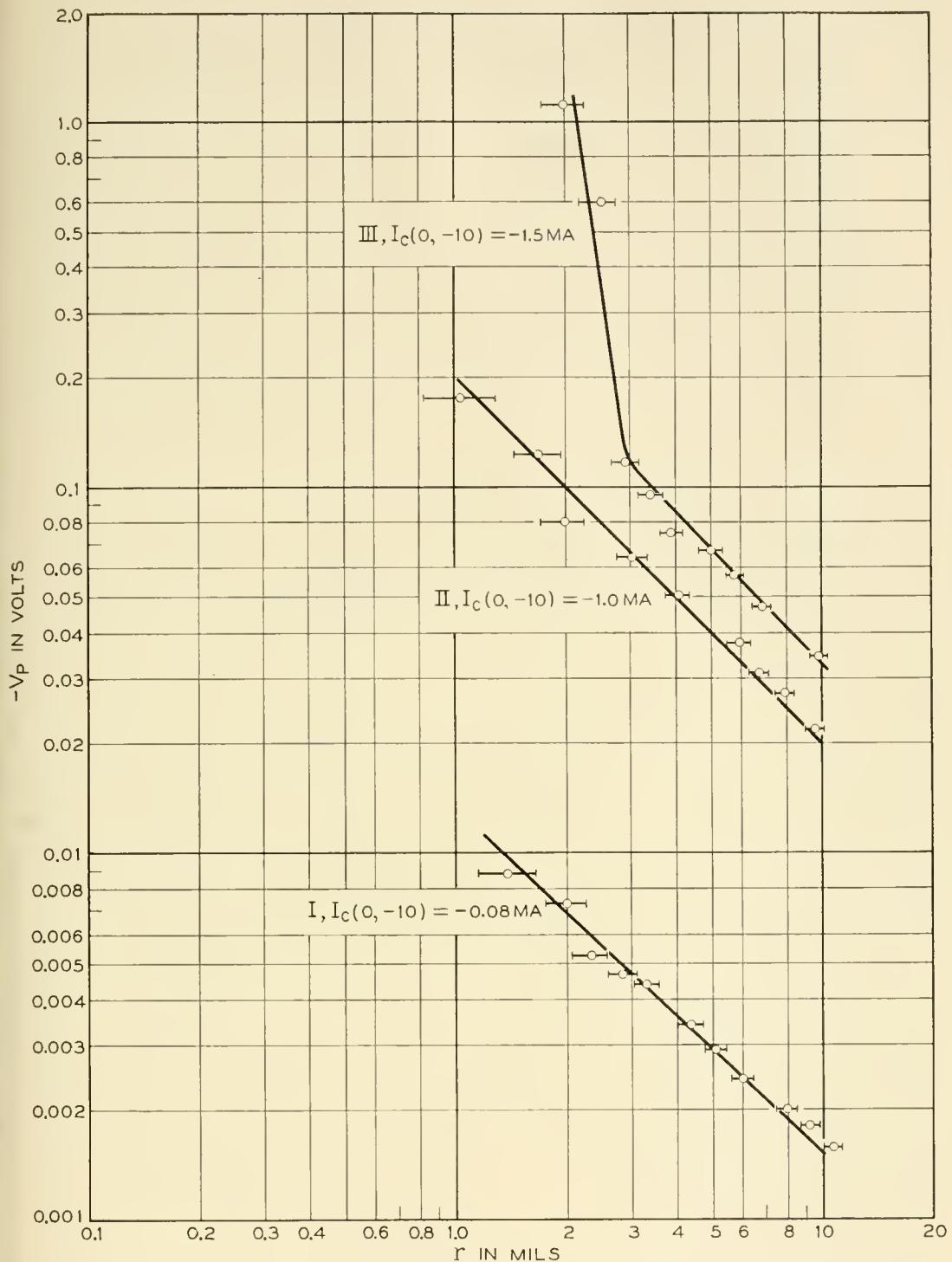


Fig. 2 — Comparison of floating potentials near formed points.

by $\rho I / 2\pi r$ where ρ is well within the range of the measured resistivity (3-4 ohm-cm).

Thus the effect of adding the donor to the point wire is to increase the reverse current and increase the floating potential near the point by an order of magnitude. One would therefore expect an accompanying

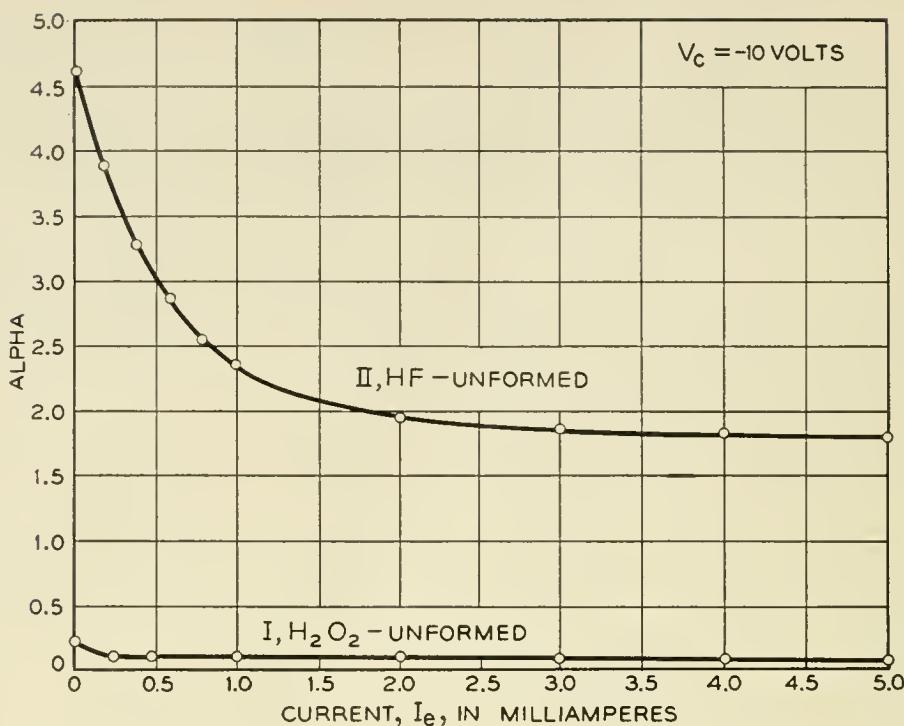


Fig. 3 — Comparison of alpha-emitter-current characteristics of formed points.

increase in the drift field near the point and a corresponding increase in α . Fig. 3 indicates that such is the case. The small signal α is plotted as a function of emitter current in Curves I and II. The point spacing in this case is 2.5 mils. It is interesting to note that the peak at low emitter currents is present in both cases, in spite of the fact that presence of a p-n hook is not likely when the Be-Cu point is formed.

It is thus apparent that the forming the Be-Cu point produces a structure which more closely resembles a p-n junction. The effect of adding the donor is to reduce the resistance of the junction. Further contrast between these two kinds of contacts is demonstrated by comparing forward currents through the contacts and their capacities. In Table I, a summary of all the contrasting properties is given. All values quoted are representative values.

2.2.2 Use of the Copper Plating Technique

During the investigation of these contact properties, an interesting way of illustrating their physical properties was developed. This technique, borrowed from junction transistor technology, can be used to identify visually the boundary between the formed region and the bulk germanium in a metallographic section of a point-contact transistor. It further appears that modifications of the technique will enable determination of

TABLE I

Contact	Formed Be Cu	Formed Phosphor Bronze
$I_c(0, -10)$ ma.....	-0.01 ma	- 1.0 ma
$I_c(6, -5)$ ma.....	-1.0 ma	-14.0 ma
$I_c(0, +0.5)$ ma.....	2.8 ma	0.8
Peak value of α	0.25	4.5
$\alpha (5.0, -10)$	0.1	1.7
Capacity ($V_c = -5V$).....	3.0 $\mu\mu f$	< 0.1 $\mu\mu f$

the equipotentials surrounding a collector or emitter point under bias, and visualization of current flow patterns in point contact transistors under bias operating conditions.

Use of this technique in identification of formed transistor properties is quite simple. A transistor container (including only the completed header, wafer, and point-contact structure) is filled with araldite plastic, which is allowed to harden. The collector point is then electrically formed. The plastic is necessary to ensure that the collector point does not subsequently move from the formed area. The can itself is then embedded in a plastic block, which is lapped down to expose a cross section of the unit, Fig. 4(a) and (b). Both the collector point and the base electrode are well masked, Fig. 5. A droplet of $CuSO_4$ solution of fairly low concentration is placed on the germanium, so that it is in physical contact only with the germanium and the masking plastic. In order to identify the formed region, a reverse bias of 20 volts or so is applied between the collector point and the base contact for a time usually of 0.1 second or less. Actually, best results have been obtained by applying the reverse

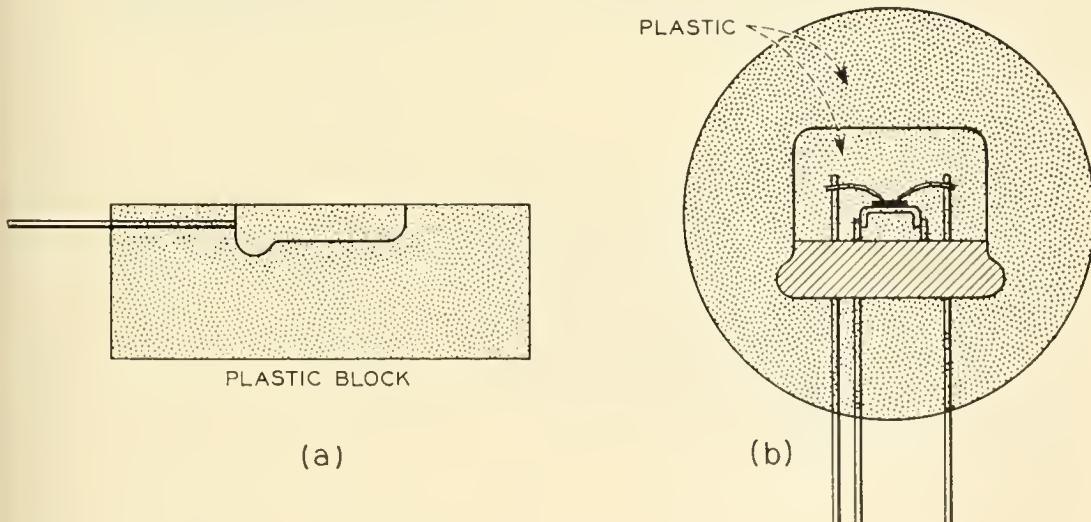


Fig. 4 — Preparation of a transistor for copper plating.

bias in the form of a condenser discharge pulse. Care must be taken to avoid changes in contact characteristics resulting from the plating pulse. The deposit of copper does not appear instantly after pulse application, but may require several seconds before becoming visible. At the instant the deposit becomes visible, the plating solution is washed off.

Fig. 6(a) and 6(b) show the results of the plating operation on a formed collector point and a formed emitter point. Both pulses were similar to, though somewhat "heavier" than those usually used to form transistors. These units were plated under the conditions illustrated in Fig. 5(a). The floating potential in the vicinity of the reversed bias point can be measured as a function of the distance, r , from its center, using an auxiliary tungsten point. Qualitatively this potential is shown as a function of the distance, r , in Fig. 5(b). In this case most of the drop in magnitude of the potential appears within a radius, r , less than 0.002 inches, provided surface conductivity is small. The conductivity of the plating solution is kept small to ensure that the potential distribution in the germanium is not altered by presence of the solution. Under these conditions, it is assumed that, although copper ions in solution are attracted towards the highly negative regions of the germanium, the main current flow is through the germanium, except for regions of high potential gradients. In these regions some of the current will be carried by ions in the solution, by-passing the region. If the formed region boundary is a sharp p-n junction, one would expect a plating pattern as observed in Fig. 6(b) and 6(d), as is observed with the donor-free emitter point. For the more complicated structure produced by forming the

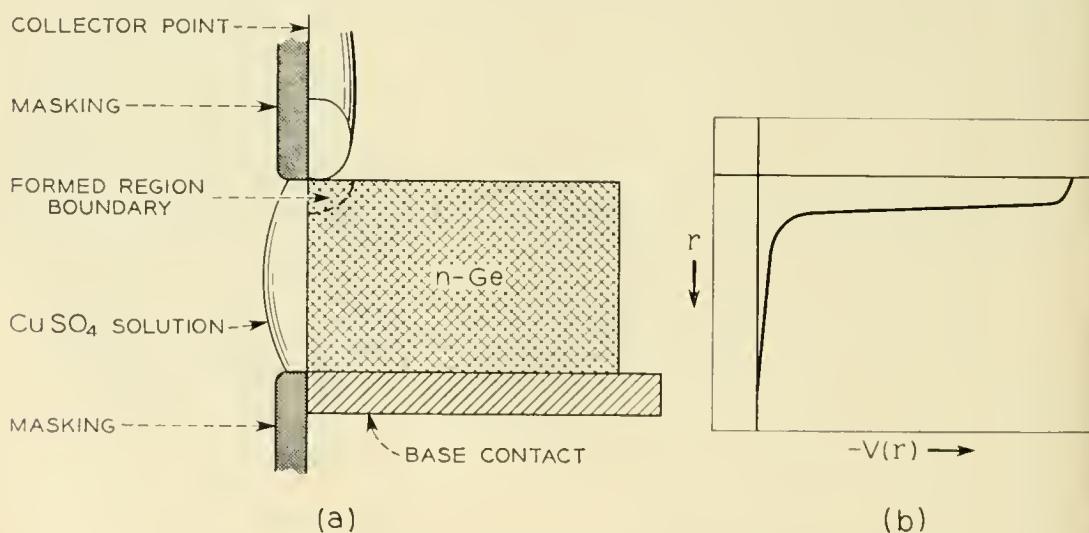


Fig. 5 — Experimental conditions for copper plating.

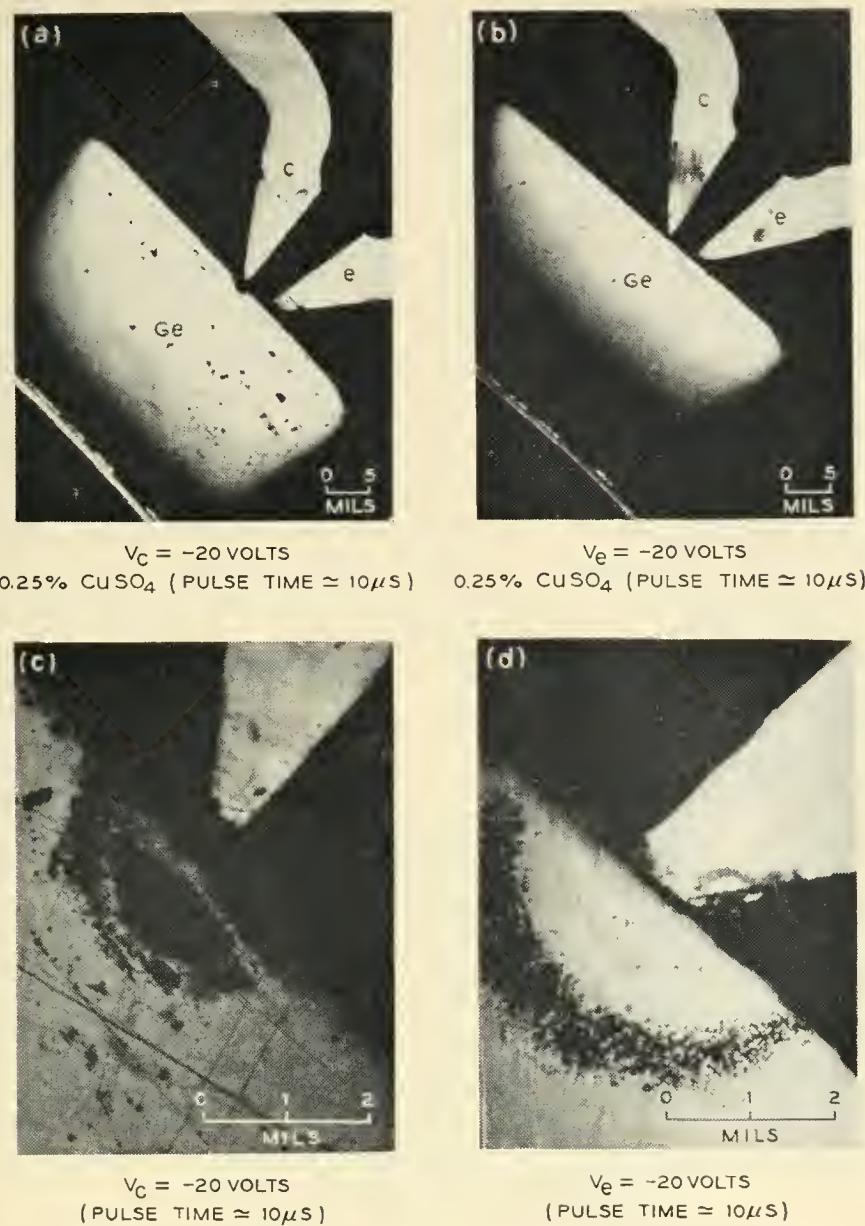


Fig. 6 — Copper plated formed layers in point-contact transistors.

collector, the pattern obtained is more difficult to interpret, Fig. 6(a) and (c). However, in both cases the disturbed areas are roughly comparable in shape and size.

Differences in the forward characteristics of the collector and emitter points may also be graphically observed by means of the plating technique. In Figs. 7(a) and 7(b) are sketches of patterns obtained by applying forward bias to contacts for plating. In this case a more concentrated solution is used, and the plating time is longer. In Fig. 7(a) is shown the pattern obtained when an unformed collector point is biased for-

ward during the plating pulse. The copper deposits to within the order of a diffusion length from the emitter point. Fig. 7(b) shows the pattern obtained by plating the region near a forward biased formed collector. Here again the copper has deposited over practically all of the base wafer surface, except for a much smaller hemispherical region near the collector point.

By adjustment of the plating time and solution concentration, the almost radial field in the bulk germanium under a reverse-biased collector point can be detected. Under similar conditions, an emitter point biased to the same voltage shows a plating pattern similar to that of Fig. 6(b), with little evidence of the radial field. This would be expected from the potential plots shown in Fig. 2.

These techniques serve merely to illustrate graphically the differences in the two types of contact. Although both points when formed give rise to a formed region in the bulk germanium of similar size and shape, the diode characteristics of the junction under the donor-doped point are degraded.

The plating technique may also be adjusted to allow sensitivity to the current flow pattern in a transistor with both points biased to operating values. The example shown in Fig. 8 demonstrates visually the bulk nature of the current flow in the point contact transistor. Here the copper plates out on the negative regions of the crystal and is noticeably absent from the regions of high hole density under the emitter point. In the region to the left of the collector indicated by the arrow, the plating is partially obscured by masking. The size of the copper-free region under the emitter point may be reduced to substantially zero for the same I_e , by increasing the bias applied to the collector.

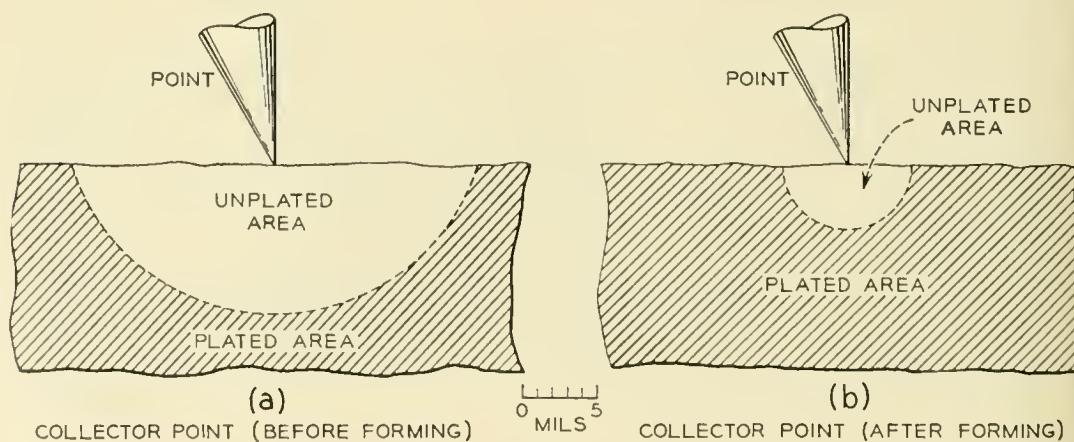


Fig. 7 — The effect of forming and current flow in point-contact collectors.

2.3 Under-Formed and Over-Formed Contacts

One of the problems encountered in the large-scale manufacture of point-contact transistors is the variation in the forming yield. Thus, forming to a specified criterion of transistor performance does not always result in a uniform product. Although considerable care may be taken to ensure uniformity of all bulk properties and forming technique, a large variation may be encountered in the output characteristics of the transistors. In Section 4, a prime factor in determining the efficiency of forming is shown to be the chemical history of the germanium surface. Uncontrollable variations in surface conditions may therefore often account for much of the variations in results of a specific forming technique.

Such variations often manifest themselves merely as differences in degree, but may show up as differences in kind, taking the form of anomalous output characteristics. These have been classified by L. E. Miller¹⁴ into three qualitatively different phenomena. The first of these, referred

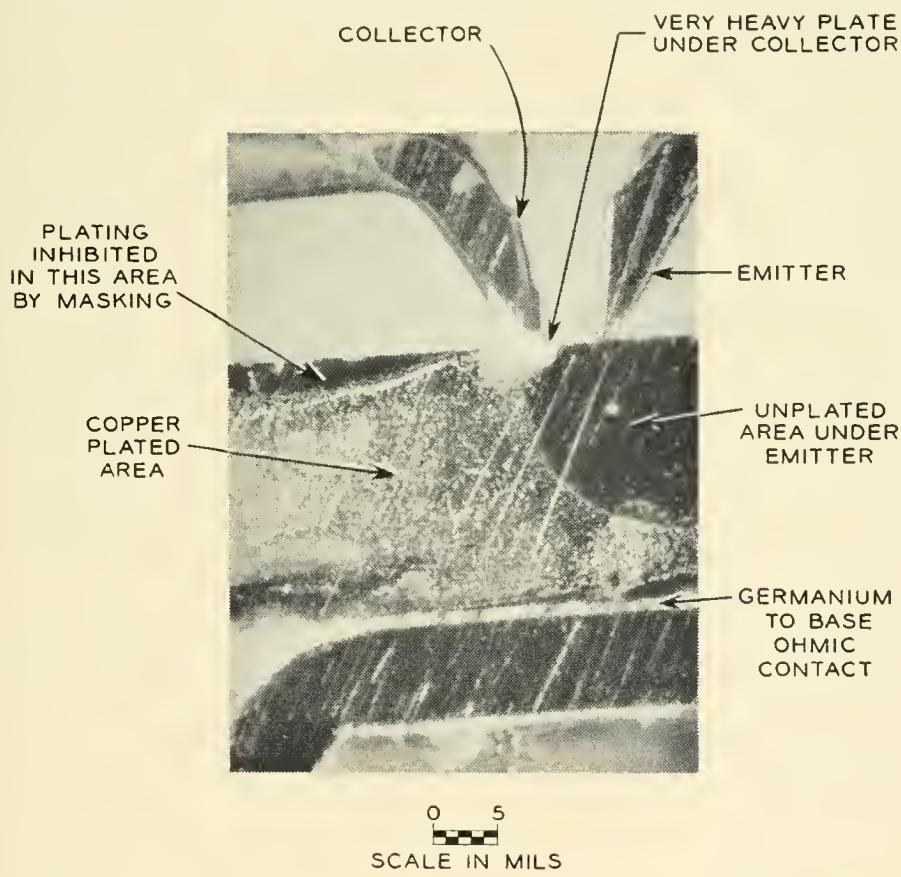


Fig. 8 — Flow geometry for a low alpha point-contact transistor.

to as the type (1) anomaly, is of interest here since it represents a collector contact whose physical properties are between the extremes listed in Section 2.2. Miller has shown that the source of this kind of output characteristic can be identified as the formed area under the collector point.

Essentially this anomaly consists of an abrupt rise in the current gain as the collector voltage V_c is increased at constant emitter current. Beyond the critical value of V_c , the characteristic of the unit resembles that of a well formed transistor. One is led to consider that such a contact is under-formed, in the sense that at low V_c , collection of holes is inadequate. Further support is lent to such a definition by the data of Miller, which shows a definite increase in the occurrence of anomalous units with a decrease in the I_{co} of the contact. Such an increase occurs regardless of whether the I_{co} decrease is obtained by decreasing the donor concentration of the point wire, or by increasing the time constant of the forming pulse. In Table II are compared collector capacity and I_{co} measurements made in units with and without output characteristic anomalies. The capacity of these anomalous collectors also appears to range between the two extremes listed in Table I. Thus there is evidence that these collectors are intermediate between the extremes cited in Table I in the sense that at low reverse biases the drift field is low, and the properties of the formed barrier resemble those of a formed donor-free point.

The results of detailed investigation of the properties of such anomalous characteristics now being conducted will be published at a later date. The present experimental results indicate that the instability occurs when the extra current to the collector, ΔI_c , reaches a critical value. In this respect, increasing the transport factor β , by increasing V_c , or increasing the emitter current are equivalent. At a roughly critical ΔI_c , the transition between a low α and a higher value of α occurs. After the transition, the unit behaves like a conventional point contact transistor, with a current multiplication on the order of $(1 + b)$ at higher values of I_e . Thus the origin of this kind of anomaly may lie in the lowering of the formed barrier by the space charge of the holes, a mechanism suggested by Bardeen.¹⁵

TABLE II

	$I_c(I_e = 0, V_c = -10 \text{ volts})$	$C_c(I_e = 0, V_c = -10 \text{ volts})$
Typical Transistor	1.0 ma	$0.1 \mu\mu\text{f}$
Typical Anomalous Transistor	0.2 ma	$0.5 \mu\mu\text{f}$

The other anomalous collector characteristics considered by Miller have their origin in the relation between the transport factor and the properties of the emitter at various operating conditions. In view of the relations existing between the occurrence of these anomalies and the I_{co} of the collector contact, there is some justification for classification of these contacts as "over-formed."

3. PROPERTIES OF UNFORMED POINT CONTACTS

3.1 *Physical Properties of Metal-Semiconductor Contacts*

The classical ideas on the nature of the rectifying metal-semiconductor contact have undergone substantial revision since the consideration by Bardeen¹⁶ of the importance of surface states and the work on the point contact transistor by Bardeen and Brattain. According to Bardeen's model, the nature of the space charge layer at such a contact is to be considered largely independent of the metal used for contact, and is primarily dependent on the charge residing in localized states at the germanium surface. Thus the rectifying properties of the metal semiconductor contact in air are expected to be largely independent of the work function of the contact metal.

The question of the exact nature of the surface charges is not yet readily answerable. Charges may arise which consist of electrons and holes residing in surface states of the type proposed by Tamm.¹⁷ On the other hand, other surface charges may arise as a result of adsorbed impurity ions, or from adsorbed atoms or molecules having electrical dipole moments. Brattain and Bardeen¹⁸ have shown that the space charge layer is dependent on the surrounding ambient and have indicated that charge may reside on the outer surface of a film (presumably an oxide layer) at the germanium surface as well as in surface states of the type mentioned above, which are presumably those responsible for surface recombination processes.

Thus, it is the surface charge on the semiconductor, rather than the nature of the metal, which primarily determines the nature of the potential barrier which exists at a metal semiconductor junction.

A schematic electron energy diagram for the contact between a metal and an *n*-type semiconductor is shown in Fig. 9. The potential barrier φ_0 , and the nature of the space charge layer in the semiconductor are determined by the surface charge system and the bulk properties of the semiconductor. In turn, the surface charge system is dependent upon such factors as the ambient at the germanium surface and the chemical history of the surface.

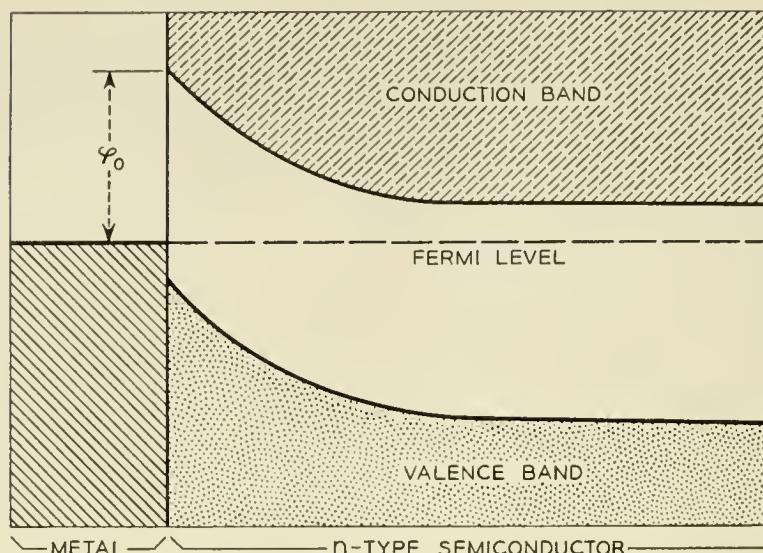


Fig. 9 — Electron energy diagram for a metal-semiconductor contact.

The experiments of Brown¹⁹ indicate that the presence of charge on the surface of *p*-germanium can alter the space charge in the crystal near its surface and, in some cases, produces an inversion layer of *n*-germanium at the surface. Garrett and Brattain²⁰ have shown that a change of ambient from sparked oxygen to dry oxygen to wet oxygen can increase I_{co} and floating potential on n-p-n junction transistors, and the process is reversible. Their interpretation is that sparked oxygen builds up a film, presumably germanium oxide. Oxygen atoms on the surface, negatively charged, can give rise to a p-type inversion layer on n-germanium. Moisture apparently counteracts this negative charge, and humid oxygen can cause an n-type inversion layer on p-germanium, which can be removed with a dry oxygen ambient.

Thus, the electrical resistance of an unformed metal-germanium contact on an etched germanium surface can be expected to be extremely sensitive to any chemical treatment which tends to affect the constitution of the oxide layer present on the surface, regardless of the metal used for contact in air. Bardeen and Brattain,¹ in early transistor experiments, have shown that such is the case. They have used transistor collector points on germanium surfaces which, after etching, were subjected to an oxidation treatment (heating in air).

In this section are described experiments which seem to indicate that the reverse resistance of unformed diodes on etched n-germanium surfaces can be decreased by chemical surface treatment, and the magnitude of the floating potential near such contacts is increased to sufficient extent that the point can serve as a multiplying collector. Average α for

these points approaches values found in electrically formed collectors. Subsequent parts of this section will be concerned with description of the experiments involved and comparison of the electrical characteristics of these points with those of conventionally formed points.

The effects of electrical forming on donor-doped and donor-free point contacts have been described in earlier sections. It has been stressed that the addition of the donor element to the point results in a contact with degraded diode characteristics, but which serves as an excellent collector.

The possibility of an analogous situation in an *unformed* point collector exists, with the electrical forming of the donor-doped point being replaced by a suitable chemical treatment of the surface. The experiments described below indicate that such is the case.

3.2 Experimental Procedures

The germanium used in these experiments was zone-leveled material. The n-germanium was in the 3 to 4 $\Omega\text{-cm}$ range. Originally, experiments were run using slices, about 0.025" in thickness, soldered on flat brass blocks, with the brass well masked with polystyrene. Germanium dice, already mounted on standard base-header assemblies used in a hermetic-seal transistor process pilot line, were also used.

The ground surface of a slice was given a three-minute chemical etch (CP₄ or superoxol), washed in pure water (conductivity <0.1 micromho), and blown dry in a nitrogen stream. This surface could then be exposed for several minutes to 24 per cent HF, hot zinc chloride-ammonium chloride solder flux, or other chemical treatments as the experiment might require. These solutions were applied to the slice or die in the form of large droplets, so the solution did not come in contact even with the masking. Later, in order to make doubly sure that contamination from the base or base contact was not involved, all experiments were repeated using a two-inch length of a zone leveled bar with a base contact soldered on one end, and the other end, freshly ground between treatments, used as the surface under examination. The etching was done by lowering one end of the bar about one-half inch into the etch, leaving the contact end a good distance from the etch. The etched surface could subsequently be exposed to any desired chemical treatment. After the chemical treatment, the sample surface was again washed in low conductivity water for several minutes and blown dry with nitrogen.

The sample, after chemical treatment, was placed on a double ended manipulator base, used to control the position and pressure of two anti-

lever points on the treated surface. The electrical characteristics of a beryllium copper point, operating as transistor collector on the treated surface, could then be investigated. An auxiliary etched tungsten point doubled as a potential probe and as an emitter. A switching arrangement allowed oscilloscope presentation of the I_c-V_c collector family and the alpha-emitter current sweep, measurement of the emitter floating potential on a high impedance VTVM, and determination of other transistor parameters for any desired position of the emitter point.

Phosphor bronze collector points were not used since it was found that, on certain chemically etched surfaces the mere application of a negative bias of 15–40 volts for a few seconds sometimes is sufficient to cause electrical forming of the point in the sense that I_{co} and average α are increased by an appreciable amount.

The beryllium copper points were carefully cleaned to prevent contamination by donor elements. Their cleanliness was then tested by other methods described in Section 3.3.6.

With this arrangement, most of the electrical properties of a given manipulator unit could be inspected during the time the unit "survived." These electrical measurements were made in room air (R. H. between 20 and 30 per cent), although provision was made for directing a continuous stream of dry nitrogen at the points and surrounding surface.

3.3 Experimental Results

3.3.1 Unformed Transistors on Superoxol Etched* Surfaces

A striking difference was observed in the electrical characteristics of unformed collector points on the various n-germanium surfaces examined. In particular, surprisingly large values of $I_c(0, -10)$ and $I_c(6, -5)$, (the latter taken as a measure of average α), were encountered on the superoxol etched surface subsequently "soaked" for about 10 minutes with 24 per cent HF. At these locations the unformed transistor action was quite similar to that observed with a conventional phosphor bronze point formed on a freshly etched surface.

These large values were found only in specific locations on the treated surface, there being a random fluctuation of $I_c(0 - 10)$ and $I_c(6, -5)$ with location of the points on the surface. However, no such large values of these parameters were found (together) on surfaces freshly etched in superoxol. The α as a function of emitter current for the unformed points (2.5 mil spacing) on a superoxol etched surface, before (Curve I) and after (Curve II) HF treatment is shown in Fig. 10. Comparison with

* One part 30 per cent H_2O_2 , one part 48 per cent HF and four parts water.

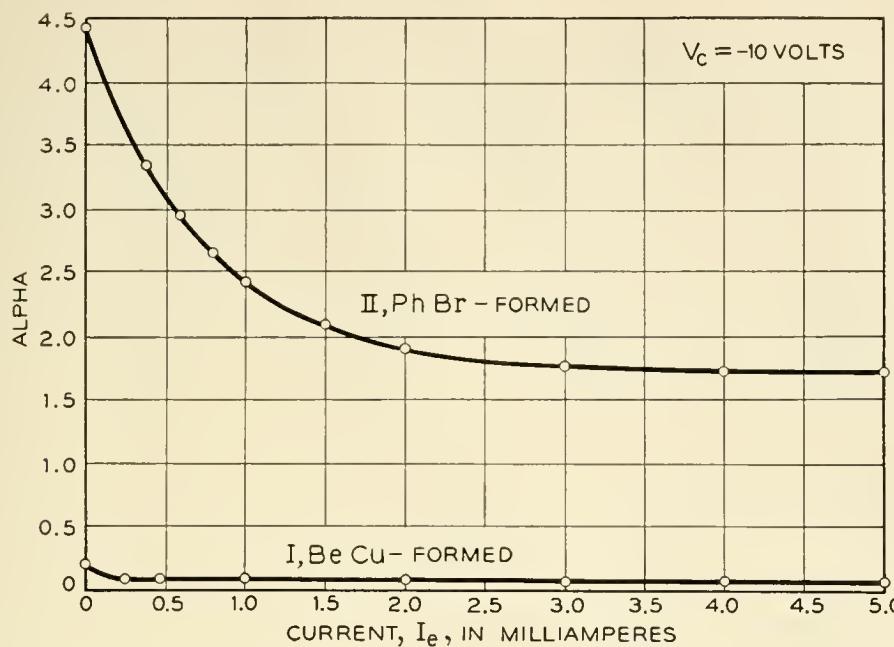


Fig. 10 — Comparison of alpha-emitter-current characteristics for unformed collectors.

Curve II, Fig. 3, indicates that the $\alpha(I_e)$, obtained after the HF treatment, is comparable to that of a phosphor bronze collector formed conventionally on the same etched surface before treatment. (It turns out that conventional electrical forming on the etched surface after the HF treatment is more difficult, and in cases as referred to above, where the α is not initially high, requires an excessive number of pulses to bring the α to a normal value.)

In Table III are listed the maximum and minimum values of some transistor parameters found on the same superoxol-etched surface before and after the HF treatment (point spacing about 2 mils).

It is seen that the effect of the subsequent HF treatment after the superoxol etch is at least in some locations on the treated surface to increase the $I_c(0, -10)$ and the average α , in some cases to values approaching those encountered in conventionally formed point-contact transistors. There is also a lowering of the forward current of the unformed collector point after the HF treatment. It is not to be implied from this table that the I_{co} is always found to be low on fresh superoxol-etched surfaces. Actually high values of $I_c(0, -10)$ have been occasionally found on surfaces freshly etched in superoxol. However, these collectors seldom have high values of average α , and it is suspected that here the higher reverse current is associated with excessive surface conductivity. Treatment of such a surface with HF always serves to increase the average α , and decrease the forward emitter current, with

TABLE III

Parameter	After 3 Min. Superoxol Etch		After Subsequent 10 Min. "Soak" in 24% HF	
	Max. Value Observed	Min. Value Observed	Max. Value Observed	Min. Value Observed
$I_c(0, -10)$ ma.....	-0.16	-0.06	-0.98	-0.20
$I_c(6, -5)$ ma.....	-7.0	-0.95	-13.5	-7.0
$I_c(0, +0.5)$ ma.....	2.8	2.0	2.3	1.3
Peak value of α	3.0	0.15	6.0	2.0
α (5.0, -10).....	0.50	0.09	2.0	0.5

no significant changes in the extreme values of $I_c(0, -10)$ encountered initially. Some of the unformed units have collector families quite similar to those of an electrically formed point-contact transistor. However, the resemblance ends when stability of operation is considered. When the unformed units are operated in room ambient, hysteresis loops are occasionally observed, either in the I_c-V_c output characteristic sweep, or the α -emitter current sweep. This hysteresis can be eliminated by directing a stream of dry nitrogen across the germanium surface in the vicinity of the points. It is not known whether the hysteresis is thermal or electrolytic in nature. The operation of these unformed units, even in the absence of hysteresis, is extremely erratic and unstable. Operating a unit at a high power level will cause loss of α and I_{co} , and mechanical shock delivered to the collector point while the unit operates under bias may cause loss or gain of α and I_{co} . In cases where I_{co} (and α) are low when the collector point is initially set down on the treated surface, an increase in I_{co} and α may be brought about by mechanical motion of the point, (such as "tapping" the manipulator base, or dragging the point across the surface). In other cases the high α and I_{co} are found immediately after the point is set down on the freshly treated surface, without any such procedure. None of these effects is observed to an appreciable degree on a freshly etched surface without further treatment.

The effect of zinc chloride-ammonium chloride solder flux on fresh superoxol-etched surfaces was also investigated. In this case, after the etch, the surface was immersed in almost boiling solder flux for about ten minutes. The effect of this surface treatment on the performance of the unformed transistors was entirely similar to the results quoted in connection with the HF treatment. The treatment increased the reverse collector current and average α , and decreased the forward collector current, on the average. Magnitudes of $I_c(6, -5)$ as high as 14 ma were observed on surfaces treated in this way.

3.3.2 Unformed Transistors on CP₄-Etched Surfaces

With reference to unformed point contact properties, the CP₄-etched surface is not at all similar to the superoxol etched surface. If two beryllium-copper points are put down on a ground surface freshly etched in CP₄, and operated as a transistor, high values of $I_c(0, -10)$ and $I_c(6, -5)$ are often encountered. However, after an hour or so in room air, both these parameters decrease and after an overnight exposure to room air, the properties of the surface with regard to the transistor action resemble those of a surface freshly etched in superoxol. At this point, a treatment in 24 per cent HF will return $I_c(0, -10)$ and $I_c(6, -5)$ to their originally high values. These effects are summarized in Table IV.

3.3.3 Diode Characteristics on Electro-Etched Surfaces

It has been found that the rectification properties of unformed point diodes may also be changed conveniently by changing the conditions during an electrolytic etch in KOH solution. These results are summarized in Table V which represents typical variation in reverse current, I_r , with surface variation attainable by adjusting the current density and etching time. In each case the measurements represent data taken on germanium cut from adjacent sections of the same ingot and given the surface treatment noted in the table. In general the electro-etched and chemically etched results agree; that is, any treatment which appears most likely to leave an oxide film (such as the use of a high current density during electro-etching) will yield a diode with improved rectification characteristics.

3.3.4 Output Characteristic Anomalies

In the process of examining these chemically treated surfaces, some of the superoxol-etched n-germanium surfaces were given additional

TABLE IV

Parameter	Value after 3 Min. CP ₄ Etch		Value after 16 Hrs. in Room Air		Value after 10 Min. in 24% HF	
	Max. Value Observed	Min. Value Observed	Max. Value Observed	Min. Value Observed	Max.	Min.
$I_c(0, -10)$ ma...	-1.7	-0.30	-0.10	-0.04	-1.0	-0.06
$I_c(6, -5)$ ma...	-13.3	-11.0	-7.0	-2.0	-17.5	-8.0
Peak value of α	4.5	2.5	2.0	0.75	9.0	3.0
$\alpha(5.0, -10)$	1.8	1.0	1.0	0.25	2.0	0.75

TABLE V

Etch Treatment in 0.1% KOH	$I_r (-10 \text{ volts})$
10 ma for 30 sec.....	-0.16 ma
5 ma for 30 sec.....	-0.37
2.5 ma for 30 sec	-0.55
5 ma for 1 min.....	-0.04
2.5 ma for 1 min.....	-0.18
1.75 ma for 1 min.....	-0.74

treatments in H_2O_2 (superoxol strength). In general, no great differences were observed in the unformed alpha and $I_c(0, -10)$ after the treatment. However, in isolated cases, unformed units made on etched p-germanium treated in this way exhibit output characteristic anomalies of the type characterized by Miller as type (1). It was later found that the same surface treatment can produce a similar result on etched n-germanium surfaces, again only in isolated locations on the surface. An output characteristic of this form is shown in Fig. 11. This unformed unit was made on a superoxol-etched n-germanium surface with a subsequent three-minute soak in H_2O_2 . This characteristic was extremely sensitive to variation in point pressure.

Miller has also referred to output anomalies of types (2) and (3), which are usually associated with close point spacing in conventional point-contact transistors. Such types of anomaly have been observed in unformed units (with high average alpha) made on HF treated surfaces.

3.3.5 Floating Potential Measurements

In all cases where the $I_c(0, -10)$ and average alpha on etched surfaces are increased by the HF or solder flux treatment, these increases are accompanied by an increase in the magnitude of the floating potential near the reverse-biased collector. In Fig. 12 the magnitude of the floating potential V_p of a sharp tungsten probe near the reverse-biased collector is shown as a function of r , the distance of the probe from the collector (r is approximately the distance between the center of the two point contacts). The surface used in this experiment was prepared by chemical polish for three minutes in CP_4 and subsequent storing in room air for sixteen hours. This provided a smooth surface which resembled, at least with regard to electrical characteristics, a freshly etched superoxol surface.

Curve I represents the potential-distance plot for an unformed BeCu point on the aged superoxol-etched surface. Curve II represents a similar plot for an unformed BeCu point taken after the surface was given a ten-minute soak in 24 per cent HF.

The measured resistivity of the germanium used in this experiment was 3.3 to 3.6 $\Omega\text{-cm}$. It can be seen from Curves I and II that increase in the magnitude of the floating potential near the unformed point on the etched surface after the HF treatment is, to a rough extent, proportional to the increase in $I_c(0, -10)$ produced by the treatment. Values of $2\pi V_p r/I$ taken from lines of slope (-1) drawn for best fit through points on the individual curves give reasonable agreement with the measured resistivity. For curve I, $2\pi V_p r/I = 3.3$ ohm-cm, and for Curve II, $2\pi V_p r/I = 3.5$ ohm-cm.

By comparing Curves I and II of Fig. 2 with Curves I and II of Fig. 12, it can be seen that the effect of treating the surface under the unformed point with HF is analogous to adding donor to the formed point

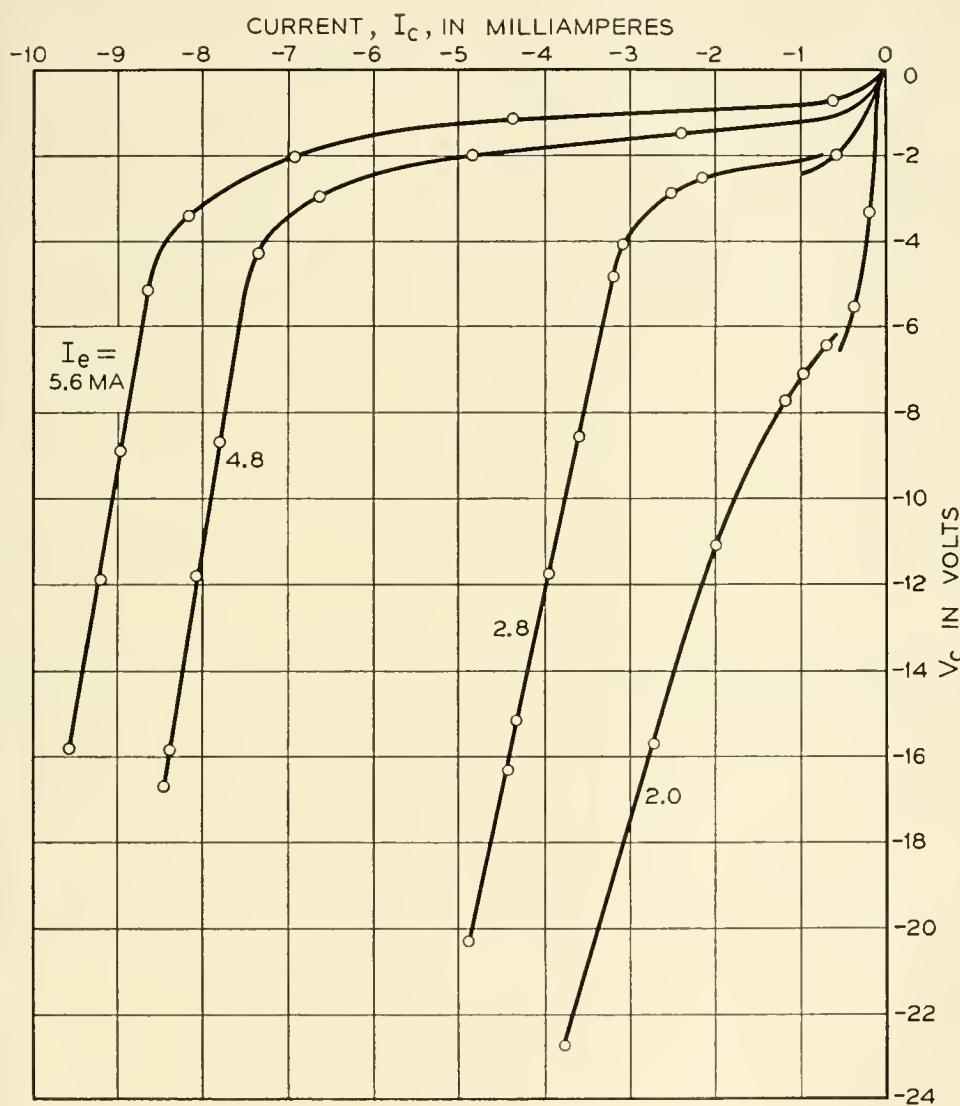


Fig. 11 — Type (1) collector anomaly observed in unformed unit (n-type germanium).

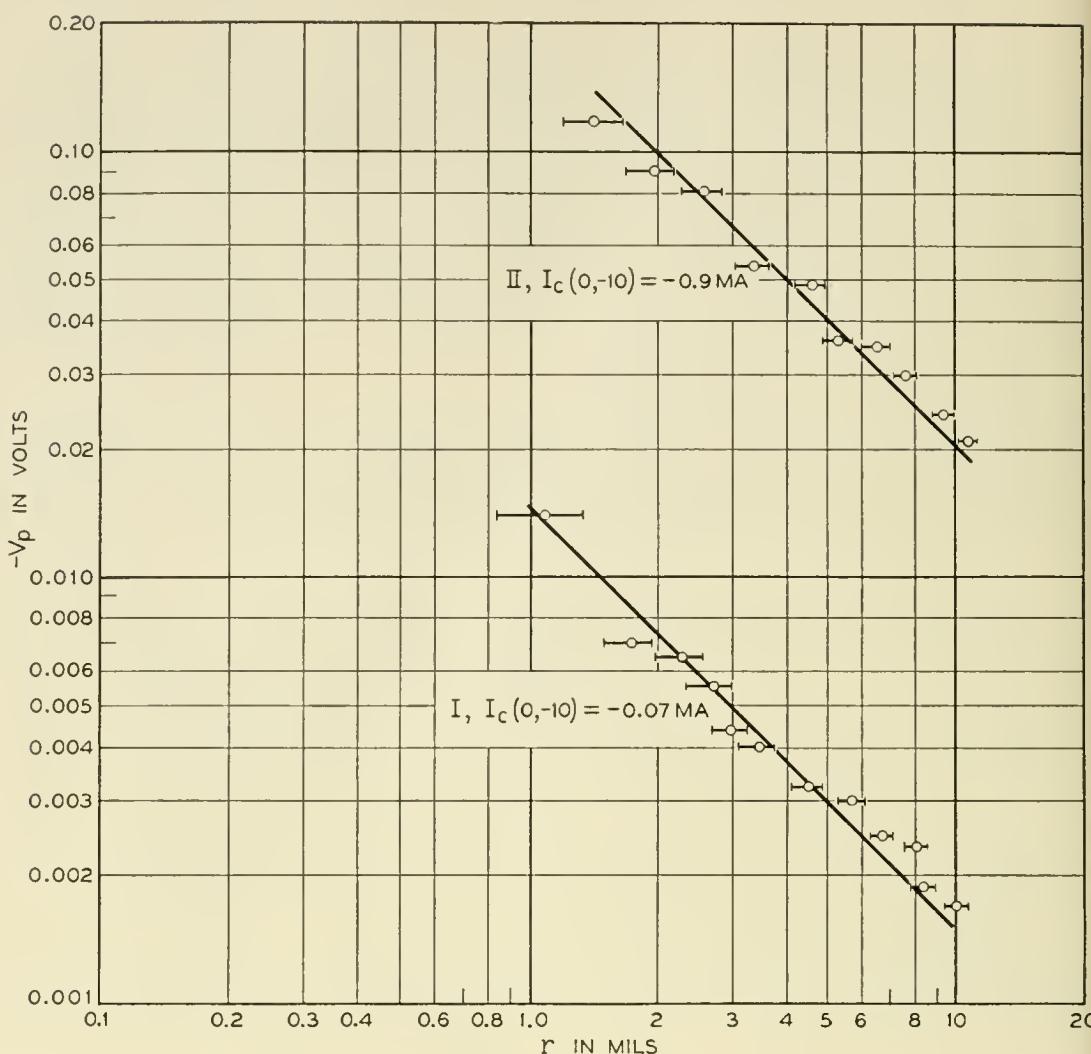


Fig. 12 — Comparison of floating potentials for unformed point-contact collectors.

on the etched surface. It seems reasonable to ascribe the increased negative floating potential after the HF treatment to an increase in current density through the surface under the point, rather than to any increase in surface conductivity. It is worth noting that on superoxol-etched surfaces, the negative floating potential near an unformed collector point can often be increased by an order of magnitude by blowing a stream of dry nitrogen near the point. This effect may possibly be a result of excess surface conductivity, but in these cases is *not* accompanied by any appreciable changes in $I_c(0, -10)$ or average alpha.

3.3.6 Contamination of Collector Points and Surfaces

Past experience with use of point-contacts as transistor collectors indicates that experiments may often be confused or confounded by unsus-

pected contamination of the points used. For this reason, particular attention was given to chemical processing of the beryllium copper points used in the preceding experiments. These points were chemically cleaned to remove oxides and unwanted contaminants, and carefully washed before use. Several lots were processed at different times, and all experiments repeated on the different lots, with no contradictory results.

It is particularly important that the point be free from donor elements, since it has been observed that phosphor bronze points or "poisoned" beryllium copper points washed with a lithium chloride solution often exhibit on superoxol etched surfaces a kind of "forming" after the application of reverse bias. The symptoms of this are a sudden increase in I_{co} which take place as the reverse bias is increased above 15–20 volts. The alpha emitter current sweep shows evidence of excessive noise in such a case, and it is not until the collector is given a conventional forming pulse that this excessive noise is eliminated, and the unit becomes stable in operation.

A donorless point can be reasonably identified by the fact that electrical pulsing, heavy or light, will not increase the initially low average alpha on a superoxol-etched surface to values much above 1.0, although I_{co} may be increased or decreased depending on the type of condenser discharge used. The beryllium copper points used were tested on superoxol-etched surface to make sure they showed no tendency to form electrically.

If high values of alpha can be found when these points are used as unformed collectors on the surfaces treated in HF or solder flux, the question arises whether such values may be attributable to presence of a donor element left on the surface in some mysterious way by the chemical treatment. If such is the case, the donor might, at high enough reverse bias, be responsible for an increased alpha in a manner similar to that observed in connection with the forming in under bias of phosphor bronze collectors on etched surfaces. Two precautions were taken in this connection. No reverse bias greater than 10 volts was ever applied intentionally to these collectors during experiments (with exception of the unit in Figure 11), and secondly, forming characteristics of both phosphor bronze points and the beryllium copper points on this type of surface were investigated.

It was found that on a superoxol surface treated with HF or the solder flux, a phosphor bronze point would form to a high average α , but this invariably required more forming pulses than on a superoxol etched surface. "One-shot" forming is common for a superoxol etched surface, whereas after the HF or solder-flux treatment, forming to high average

TABLE VI

	Point			
	Beryllium Copper Collector		Phosphor Bronze Collector	
	Occasion			
	Before Forming	After Four Forming Pulses	Before Forming	After Four Forming Pulses
$I_e(0, -10)$	-0.75	-0.50	-0.80	-2.5
$\alpha(5.0, -10)$	1.5	0	1.5	2.0
Noisy	Yes	Yes	Yes	No

alpha invariably requires at least three and sometimes many more "shots", although it can be done. This type of formed unit does not exhibit excessive noise in the α - I_e sweeping gear. However, pulsing of the beryllium copper points on the latter kind of surface, in similar fashion, invariably results in loss of alpha and never eliminates the excessive noise. Initially, the pulsing decreases the I_{co} magnitude, but continued pulsing will eventually cause large increases in this case. These results provide circumstantial evidence, at least, that the treated surface and the point are operationally free of any donor element and that the transistor collector barrier involved is at the germanium surface. For example, in Table VI are given some typical data obtained during pulsing of points on a superoxol-etched surface after treatment with near boiling zinc chloride-ammonium chloride solder-flux. A tungsten emitter was used.*

3.4 Discussion of Experimental Results

3.4.1 Effects of the Chemical Treatment on the Superoxol-Etched Surfaces

It might be presumed that an inversion layer and a relatively high surface conductivity is responsible for the increase in negative floating potential and reverse current observed on the superoxol-etched n-germanium surface after the HF treatment. On the other hand, if it be assumed that at the etched surface, in room air, an inversion layer exists which does not introduce excessive surface conductivity, one can say that the effect of the HF treatment is merely to raise the surface potential, (i.e., to reduce the barrier height for electrons). This might

* Alpha values are usually lower in any given situation when the conventional chisel-type beryllium copper emitter point is replaced by an etched tungsten point.

account for the increase in reverse current density* and a proportional increase in the magnitude of the floating potential near the point. In this case the geometry of current flow across the contact should remain relatively unchanged as indicated by the floating potential measurements. In this way the effect of the HF treatment is somewhat analogous to the addition of a small donor concentration near the surface to counteract the inversion layer. Since soluble oxide layers²¹ have been identified on etched germanium surfaces, it is not unlikely that HF (known to dissolve germanium oxide)²² might act to reduce the effective thickness of an oxide layer. Such a hypothesis is in agreement with the results of other experimenters,²³ who have attributed a surface inversion layer under the point of an n-germanium rectifier to the presence of germanium oxide. They have presumed the oxide is essential to the formation of a good point contact rectifier. The fact that, for a given ambient, the surface potential is determined by the oxide layer thickness has been postulated by Kingston.²⁴

3.4.2 CP₄-Etched Surfaces

Sullivan,²⁵ in connection with an experimental investigation of humidity stability of electrolytically-etched and chemically-etched p-n grown junction diodes, shows that CP₄ chemically-etched surfaces become more stable with respect to humidity variation after humidity exposure and cycling at room temperature. Referring to the fact that electron diffraction studies fail to reveal a crystalline oxide film on CP₄ chemically-polished surfaces and to the results of Law,²⁶ which indicate that oxide films may be formed slowly at room temperature on exposure to water vapors, he attributes the changes of stability on the CP₄ polished surface to the building up of an oxide film. If such a change can take place on the CP₄ chemically-polished surface on exposure to humid room air, then the results of Section 3.3 can be understood under the assumption that the action of the HF treatment is to remove the oxide film.

After the chemical polish, values of $I_c(0, -10)$ and average alpha for the unformed units are high, as might be expected if the polishing operation leaves the germanium surface with no appreciable oxide film. As the oxide film builds up on continued exposure to room air, both of these parameters are reduced. The subsequent application of HF tends to restore these parameters to their original values by removal of some of this oxide film. Thus, the results of this section are in accord with the

* Evidence for an increase in surface recombination velocity on HF treated surfaces is given in Section 4.2.3.

hypothesis discussed in the previous section to account for the effect of HF on the unformed transistors.

Such evidence, however, is at best only indirect evidence for the build-up of an oxide layer on prolonged exposure to room air. In experiments with grown p-n junction diodes, the authors have found great variations in the length of time required for the electrical properties of the diodes to recover after short wash periods in low conductivity water. Thus the slow changes mentioned above may at this point result from simply a longer time required for the surface to "dry out" after the washing treatment. However, a substantial difference in the physical properties of the oxide layer left by the two etches concerned is still implied. In this connection it is also worth noting that hysteresis effects appear primarily in unformed units made on HF treated surfaces.

The results of these experiments have important implications in the technology of point contact transistors. The results of an application of these results to transistor forming procedures are given in the following section.

4. RELATION OF GERMANIUM SURFACE PROPERTIES TO TRANSISTOR FORMING

4.1 *Pilot Production Problems*

The pilot production and early manufacturing stages of cartridge-type point-contact transistors has generally been characterized by periods during which the forming yields have been very high and similar periods of very low yield. Often these alternate periods occurred during the use of germanium taken from the same rod-grown or zone-leveled crystal. Considerable effort has been expended in attempting to correlate these variations in yield to variations, from crystal to crystal, or in different portions of the same crystal, or such bulk properties as resistivity or minority carrier lifetime. Although these properties of germanium do have some effect on device parameters such as average alpha, reverse emitter current, and I_{co} , there has not been any positive indication that variations in yield are attributable to the amount of variation of bulk properties normally found in the germanium which meets the specifications of the particular device concerned.

This problem was compounded during the early stages of the development of the process for hermetically sealing the point-contact transistor. It was found that although reasonable yields were obtained in the cartridge process, equivalent transistors in the hermetically sealed structure were made only with greatly reduced yield. Further, although micro-

manipulator units could be made with no difficulty, the same material fabricated into a completed structure showed completely different characteristics. In the course of investigation of this problem, it was found that the nature of the germanium surface treatment and specifically treatments calculated to produce or react with germanium oxide can profoundly affect the "formability" of the germanium surface as well as a number of other transistor parameters in the fabricated units.

It is the purpose of this section to emphasize the importance of considering the surface properties of germanium in attempting to solve such specific problems of development encountered in devices of this type. In particular, the striking variability of transistor forming on etched germanium surfaces subjected to varying chemical treatments and ambients will be described, as well as the effects of such pre-forming treatments on the parameters of the finished units. The experiments discussed in the previous section indicate how changes in the double layer at the germanium surface can influence the characteristics of an unformed point diode. In turn, the experiments below indicate how the characteristics of the unformed diode are related to the device properties of the transistor collector produced by forming the diode.

4.2 Experimental Results

4.2.1 Pilot Process Forming Yields

The forming yield of a point-contact transistor is determined by the values of the acceptance criteria and the allowable limits for each of these. Often, different criteria as well as different forming techniques are used for different transistors, so that direct comparison of results is quite complex. There are, however, certain common requirements placed on all point-contact transistors:

(a) The unit is formed so that the average alpha is roughly two or more. The collector current at a relatively high emitter current and low collector voltage is usually an approximate measure of this value, $I_c(6, -5)$ for example.

(b) The collector current with no emitter current flowing should be as low as is commensurate with the first objective.

The other transistor parameters are either directly or indirectly related to these. The number of pulses required to achieve the minimum forming objective, therefore, is one direct measure of the formability of a particular transistor; the average alpha obtained after pulsing is another. However, one must consider both average alpha and I_{co} , since while forming to a given average alpha, the I_{co} may increase prohibi-

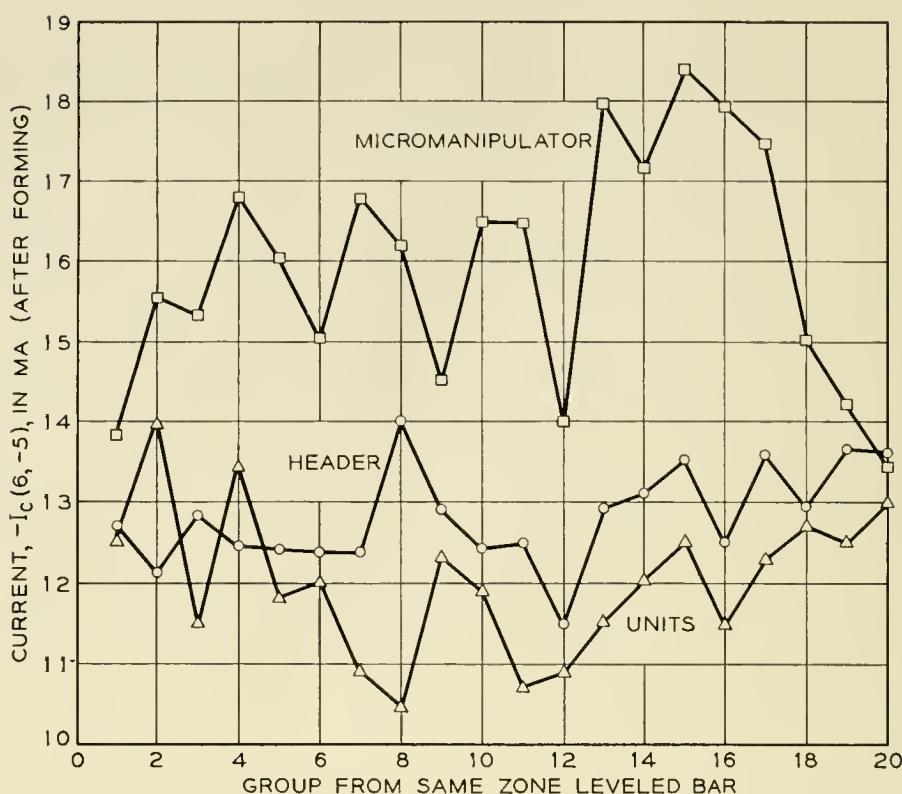


Fig. 13 — A pilot production process control chart.

tively. In later sections the authors have adopted the ratio $I_c(6, -5)/I_c(0, -20)$ as a measure of the success of the forming.

The 2N21 transistor is a hermetic seal version of a point-contact medium-speed switching transistor. During the early stages of the development of this device, it became evident that although similar germanium and point wire are used for both structures, the electrical parameters by which the devices are characterized belong to different universes. However, if the geometry of the 2N21 unit is duplicated in a manipulator transistor, the resulting device parameters do resemble those of the earlier unsealed unit. It is therefore likely that an unknown variable in the 2N21 process is responsible for the different universes mentioned above. The effect of such a variable is shown in Fig. 13, which shows a chart of a continuous process control. Each point represents the average of four different units sampled at the particular point in the process denoted in the legend. The micromanipulator data represents measurements taken on wafers which have been processed up to but not including point-wire attachment. The curve denoted "header" represents data taken immediately after the point-wire attachment. This is one additional process step beyond the point at which the manipulator data was found. It is evident from this curve that a severe degradation

TABLE VII

Treatment	Ave. No. of Pulses to Form	Average $I_c(6, -5)$	Average $I_c(0, -20)$	Fig. of Merit $I_c(6, -5)/I_c(0, -20)$
None.....	2	-13.8 ma	-1.7 ma	8.1
ZnCl ₂ -NH ₄ Cl Flux.....	3	-13.5	-1.8	7.5
Flux and heat.....	7	-10.4	-6.0	1.7

in the attainable average alpha has occurred even though the forming objective was the same. Finally, the curve denoted "unit" represents data on the first four completed units out of the same group from which the manipulator and header samples were taken. A slight decrease in average alpha is observed at this point. However, previous experience has indicated that this is an expected effect caused by the addition of the impregnant. This chart suggests that the point soldering operation in the process is causing a significant degradation in the formability of transistors passing through this step.*

This process step consists of placing the germanium wafer, which has already been etched and mounted on the header, in a point alignment tool. The point spacing and force is adjusted and the points are then soldered to the header point-wire support. In the early stages of this process a corrosive zinc chloride-ammonium chloride solder flux was necessary to obtain efficient soldering. The effect of this solder flux on the formability of micromanipulator transistors made on such surfaces is shown in Table VII. These units were formed to the acceptance criterion of $V_c(3, -5.5) \leq 2.0$ volts. Each figure represents the average of ten units treated in the same way.

The value of the use of a figure of merit such as suggested earlier is illustrated in this table. Since the average alpha (denoted here by $I_c(6, -5)$) is related to the forming objective, one might presumably keep forming until the average alpha was the same as for an easily formed transistor. In this case I_{co} tends to increase. Under these conditions, if one examined only average alpha, the data might easily be misleading. From an examination of the figures of merit in Table VII one concludes that the corrosive flux plus a heating cycle tends to degrade the germanium surface to such an extent that transistors are formed only with great difficulty.

The function of a flux during the soldering process is to remove any

* Curves of this nature have also been obtained by N. P. Burcham in investigation of soldering flux effects in hermetically sealed point contact transistor processes.

TABLE VIII

Treatment	No. of Pulses to Form	Average $I_c(6, -5)$	Average $I_c(0, -20)$	Figure of Merit $I_c(6, -5)/I_c(0, -20)$
3 min. in normal superoxol etch.....	2	-15.5 ma	-0.69 ma	22.4
1 min. in 48% HF.....	4	-10.2	-3.2	3.2
1 min. in 30% H_2O_2	1	-17.7	-1.9	9.3

oxides which are present so that a good solder joint may be made. Since the oxide on chemically-etched germanium is likely of the soluble form, one might assume that the results of Table VII imply that the action of the flux and heat tends to dissolve or remove this layer. Also implied by the data is that the presence of such an oxide layer is essential to efficient forming.

The experiments summarized in Table VIII further substantiate this hypothesis. These data represent manipulator transistors made on the same germanium wafers which had been treated in succession to a normal superoxol etch, a treatment in 48 per cent hydrofluoric acid, and a treatment in hydrogen peroxide, superoxol strength. Since the soluble form of germanium dioxide is known to react with hydrofluoric acid,²² it is presumed that the action of the HF is to partially or wholly remove any oxide left by the etch. The H_2O_2 tends to restore the original surface conditions left by the etch. Each figure represents the average of five transistors formed to the 2N21 acceptance criterion, ($V_c(3, -5.5) \leq 2.0$ volts).

In this case the hydrogen peroxide treated units have an extremely high average alpha, but the I_{co} is also higher than for normally etched units. In terms of the device properties, a unit with a more or less typical average alpha with a low I_{co} is more desirable than the one with an extremely high average alpha but accompanying high I_{co} . It has not been determined whether the I_{co} would be lower for the superoxol treated units if it had been possible to form to the same average alpha as the normally etched units. This is an important piece of device design information which is currently under investigation.

It is clear from these experiments that the nature of the germanium surface, and most probably the nature of the germanium oxide layer on it, to a large extent, determines the properties of the transistor formed on this surface. Direct application of this knowledge to the fabrication process of the hermetically sealed point contact transistor has been carried out by N. P. Burcham.

4.2.2 Relation of Unformed Diode Characteristics to Transistor "Formability"

From the results of the previous sections, it appears that superoxo-etched germanium surfaces treated with reagents in which germanium dioxide is soluble provide point contact diode characteristics unsuited to electrical pulse forming. Part of this difficulty, manifested in the inability to reach a specified value of average α without a prohibitive increase in I_{co} , probably results from a lower injection efficiency, γ , for the emitter on such a surface. This seems reasonable in view of the lower forward and higher reverse currents indicated in Table III produced by an HF soak. In Section 4.2.3 evidence will be shown that surface recombination is greater on n-type germanium surfaces treated with HF. This effect can also lead to difficulty in forming to high α without increase in I_{co} , since, for the same drift field, one would expect more minority carriers to die at the surface during their transit to the collector.

On the other hand, there is evidence for believing that the nature of the forming process itself may be quite different on an HF treated surface. Fig. 14(a) shows the time dependence of the collector voltage during a typical condenser discharge forming pulse.

The envelope of the voltage pulse follows roughly an exponential decay of a condenser-resistor series combination. However, inspection shows that during the discharge time, the resistance of the combination

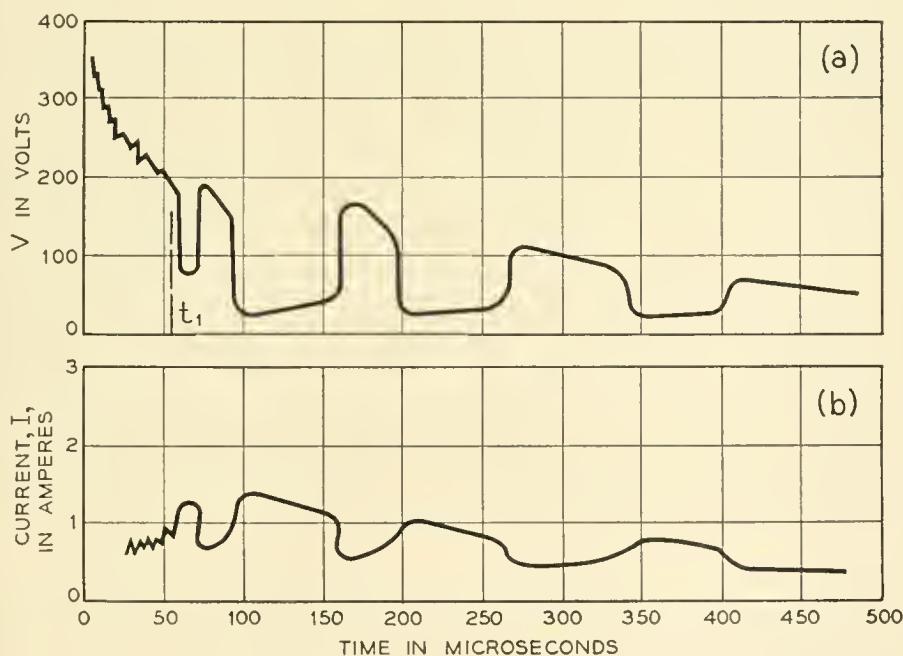


Fig. 14 — Collector current and voltage versus time for a condenser discharge forming pulse.

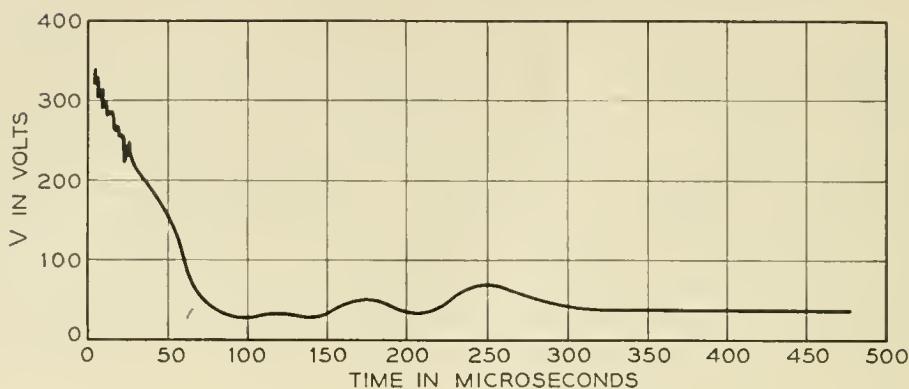


Fig. 15 — Forming voltage pulse for HF treated surface.

undergoes a succession of breakdown and recovery intervals. In Fig. 14(b) is the accompanying plot of current against time. Comparison of these two plots shows that following the application of the voltage, the resistance of the contact decreases until a rather sudden more rapid decrease in resistance occurs, taking place at time t_1 . In view of this time scale, the first decrease can be attributed to a heating of the contact, a form of thermal breakdown at the metal-semiconductor surface.²⁷ Any reason invoked to account for the second more rapid decrease in resistance must account for the short time (a few μs) in which this change occurs. In any event, shortly after the second "breakdown," a quenching results, with the collector resistance returning to a value nearer to its original value. This sequence of events is roughly repeated until the condenser is discharged.

The properties of the contact at nominal reverse voltage and currents are usually changed as soon as one such condenser discharge pulse has occurred, and often one such pulse is sufficient to reach the forming objective. A typical forming pulse obtained under similar conditions to those for Fig. 14 is shown in Fig. 15, with the exception that the surface has been treated in HF for a few minutes. On this case it is apparent that the second, rapid breakdown is entirely absent. The well-defined forming pulse of Fig. 14 is usually obtained on surfaces with good pre-forming diode characteristics, and results in production of a usable transistor.

From results of the previous sections it is well established that etched surfaces treated with reagents in which germanium dioxide is soluble provide point contact diode characteristics unsuited to electrical pulse forming.

It is often assumed, on the basis of the results of Valdes,⁵ that forming effects result from the diffusion of impurities from the point into the semiconductor during the forming pulse. Since the high temperature required for such diffusion results from the power dissipated at the metal

to semiconductor contact, more efficient forming probably results on surfaces which display very low initial saturation currents. On surfaces which produce a poor initial rectifying diode, the local energy of the forming pulse may be dissipated too far out into the bulk of the semiconductor. This situation would result in inefficient forming.

Since the low-voltage diode characteristics and the forming are probably related, one should be able to predict the "formability" of any particular surface. Fig. 16 shows that this can be done qualitatively. In the graph each point represents the average of at least five units formed on electro-etched surfaces to the forming objective, $V_c(3, -5.5) \leq 2.0$ volts. Fig. 16(a) represents the reverse emitter current before forming plotted on a log scale versus the percentage of units taking more than five pulses to form. The reverse emitter current rather than the reverse collector current is a desirable preforming parameter to use since this pre-

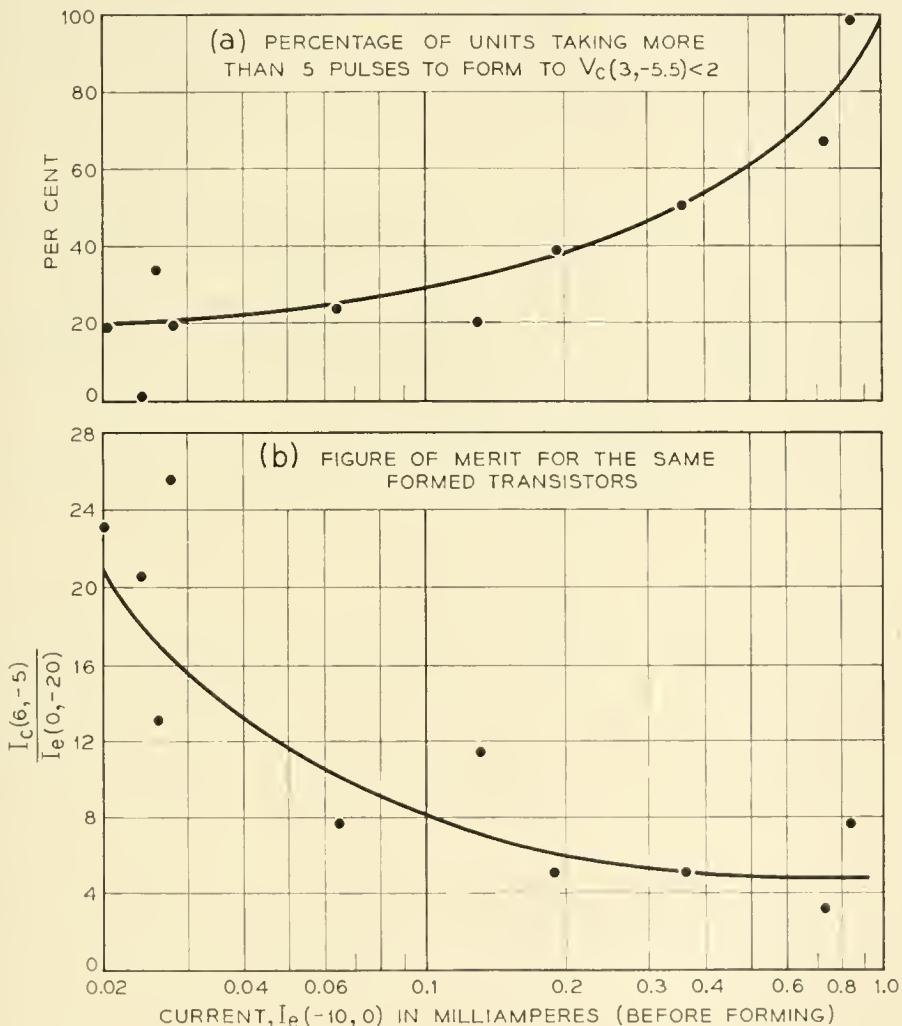


Fig. 16 — Relation of forming to pre-forming characteristics: electro-etched surfaces.

cludes any premature forming which could occur. This curve shows that a low reverse emitter current (high back impedance) is associated with easy forming and that a high reverse emitter current is associated with hard forming. Fig. 16(b) represents data on the same group of units with $I_c(6, -5)/I_c(0, -20)$ plotted versus the reverse emitter current on a log abscissa. It is significant to note that the figure of merit is consistently high for units with low reverse emitter current and low for units with high reverse emitter currents. It was possible to achieve this wide range in reverse currents on the same material by adjusting the current density in the manner summarized by Table V. In each case a high current density results in the low reverse currents.

Some other oxidizing agents may be used interchangeably with the materials just discussed. A dilute nitric acid solution produces a surface on which excellent diode properties are observed and good forming results on these surfaces. It has also been found that a treatment in potassium cyanide results in a surface which appears to be well oxidized. There are, however, some indications that certain chemical treatments tend, more than others, to passivate the germanium surface to any subsequent treatment.

Although it has been shown that variations in the surface oxide layer markedly affect the transistor made on that particular surface, variations in forming yield such as illustrated by the manipulator line in Fig. 13 are still unaccounted for. The etching procedure in the fabrication of the point contact transistor has always been one of the most carefully controlled steps. It therefore becomes necessary to examine the process for some subtle interaction between the germanium surface and the ambient to which the surface is subjected during processing.

4.2.3 *Controlled Ambient Experiments*

The experiment summarized by Fig. 17 represents a "dry box" experiment designed to investigate the effect of ambient on the forming yield. Ten germanium wafers were mounted on hermetic seal headers, they were electro-etched, and then five treated for one minute in HF. The wafers were rinsed in deionized water, dried for three minutes in a stream of nitrogen, and placed in a nitrogen dry box where the relative humidity was maintained at less than 1 per cent. One micromanipulator transistor was formed on each wafer immediately and then at subsequent intervals of one day, always in widely different locations on the wafer. These manipulations were carried out inside the dry box using rubber gloves so that at no time was the RH greater than 1 per cent. After two days the box was opened to room air and the experiment continued.

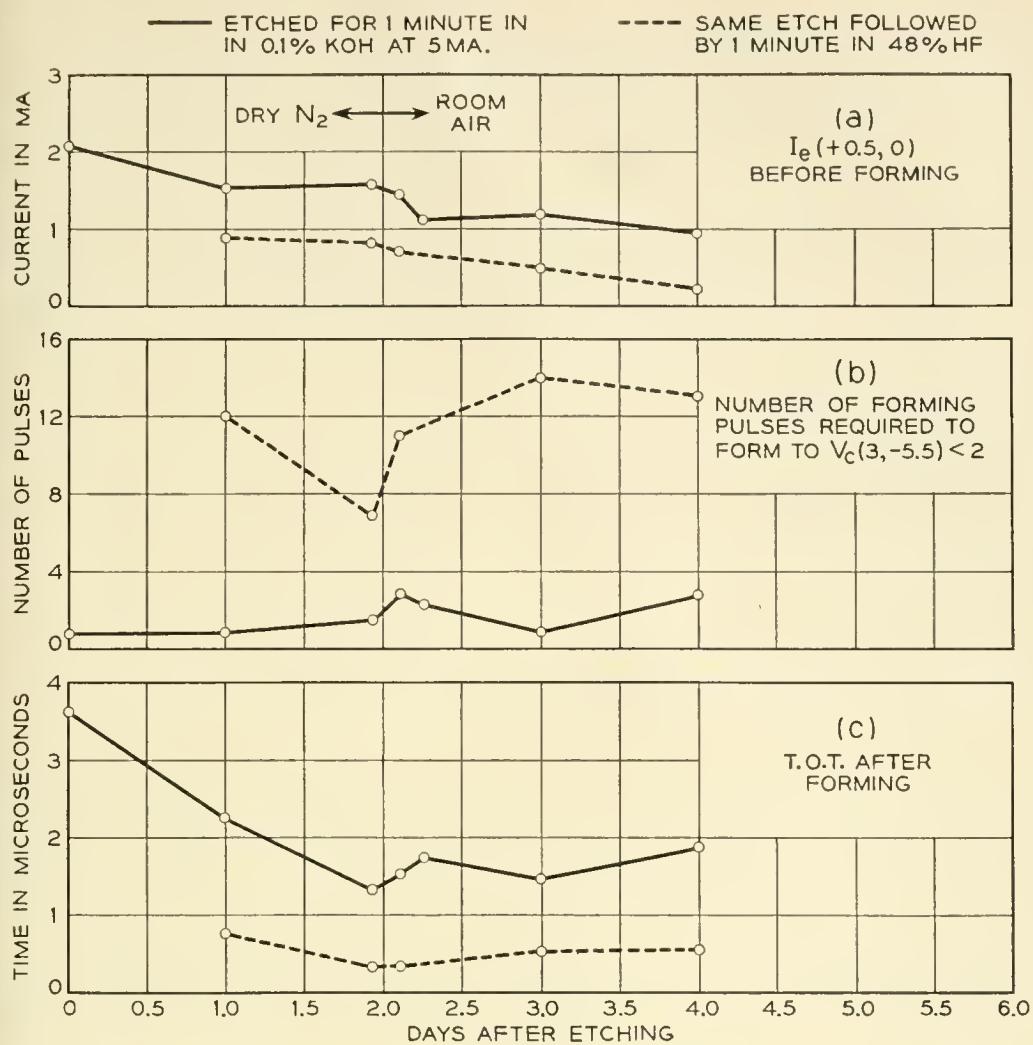


Fig. 17 — Effect of storage ambient on transistor characteristics — electro-etched surfaces.

Each point on Fig. 17 represents the average of five units on five different wafers.

The difference in the electrical properties of the two surfaces in air already noted in previous sections is observed. In addition an increase in surface recombination is indicated on the HF treated surface by a decrease in the turn-off-time measurement (TOT).* Finally, any influence of ambient on the electrical properties of the two surfaces used is apparently small.

4.2.4 A Statistical Survey Experiment on Transistor Forming

The experiment described here was designed to check some of the effects noted in earlier sections as well as to investigate possible interactions between the germanium surface and various ambients experienced during the processing of point contact units. The experimental design

* TOT is a nonparametric measurement indicative of the switching speed when used in a specific circuit.

TABLE IX—EXPERIMENTAL DESIGN OF RANDOMIZED BLOCK EXPERIMENT

	Surface Treatments					
	A Normal Superoxol Etch	B Electro Etch 5 ma. for 1 min. in 0.1% KOH	C 1 min. H ₂ O ₂ after Normal Etch	D 1 min. H ₂ O ₂ after Normal Etch and Shelf in Ambient	E 1 min. HF after Normal Etch	F 1 min. HF after Normal Shelf in Ambient
Ambient Shelf Conditions						
Formed immediately after treatment	x	x	x	x	x	x
Formed after shelf in room ambient	x	x	x	x	x	x
Formed after shelf over drierite	x	x	x	x	x	x
Formed after shelf in dry N ₂	x	x	x	x	x	x
Formed after shelf at 76.5% RH	x	x	x	x	x	x

Note: Shelf represents storage for 24 hours.

used is a 5×6 randomized block experiment with multiple subgroups.²⁸ Table IX shows the general plan of the experiment. The six columns represent different etch treatments, and the five rows represent some possible variations in storage conditions. Each subgroup represents five transistors, and the experiment represents a total of 150 transistors made on germanium from the same zone-leveled slice, given 30 different treatments. Although nine measurements were made for each transistor, the figure of merit appeared to be most significantly dependent on the treatments.

As expected from the results already quoted, the major variability was found in units formed on surfaces freshly treated with HF, with considerable improvement in formability during storage. However, the looked for influence of storage ambients does not appear when the column F has been removed from consideration. One concludes that the variation between treatments is small, and the effect of ambient is even less than the effect of the treatments. Thus when surface treatment does not vary to extremes, the effect of storage ambient is relatively minor. Thus variations found in such experiments as exemplified on the manipulator line in Fig. 13 must be attributed to a still unknown factor.

4.2.5 Effect of Contamination Before Etching

Since etching removes the damaged surface and is usually done with highly corrosive materials, it seems unlikely that any contamination

before etching could affect the efficiency of etch. There have, however, been some indications that this does occur. Certain chemical treatments appear to passivate the surface to any subsequent treatment, for example, the results in Sections 4.2.3 and 4.2.4. The electro-etched surface followed by an HF treatment does not change rapidly with time in room air, while the superoxol-etched surface followed by an HF treatment changes quite rapidly. Surfaces which have been etched in CP₄ and subsequently treated in HF appear to be as stable as electro-etched surfaces. Subsequent treatments in superoxol do not appear to result in significant changes in the surface characteristics. Experiments on unetched germanium wafers indicate that none of the components of CP₄ alone will prevent normal etching, but if an unetched wafer is treated with a combination of 50 per cent nitric acid plus 48 per cent HF for a few moments, the surface will be stabilized as to retard the formation of the normal pyramidal etch pattern when the surface is etched in superoxol etch. Taken together these observations may imply that certain types of oxide surfaces are more stable than others and perhaps may even be passivated to subsequent environmental conditions.

With this background of information it becomes more believable that chemical treatments before etching could affect the surface of the germanium resulting from the subsequent etching. It is not unreasonable to believe that any variation in surface potential resulting from pre-etch treatment might influence the reaction between the etchant and the germanium. An experiment was performed using gold-bonded bases to isolate the contribution of the solder flux normally used in the base-wafer attachment. Twenty wafers from the same slice were divided into four subgroups of five. The groups were treated in such a way that any effects of HF or solder flux soaking before superoxol etching could be detected.

The results of this experiment do indicate that presence of flux before etching significantly affects the collector currents and turn-off time of transistors made on such surfaces. Although there was no apparent difference in forming yield between sub-groups, it is felt that this variation would show up as a difference in forming yield in a process where the forming efficiency is decreased somewhat by the impregnant.

4.3 *Conclusions*

Treatment of an etched surface with germanium dioxide solvents such as HF or KOH degrades the surface to such an extent that transistor forming efficiency is decreased. A similar effect is produced by corrosive flux and heat. Thus, pre-forming measurements may be used to predict

the formability of a particular germanium surface. It is shown that poor diode characteristics are usually associated with poor forming yields. One convenient way of controlling the diode characteristics to ensure successful forming is to etch electrolytically. High current density results in the most desirable surface characteristics. Electro-etched germanium which has been subsequently treated in hydrofluoric acid shows little tendency to oxidize either in room air or dry nitrogen ambient, while superoxol-etched germanium, given the same HF treatment, changes quite rapidly in room air presumably due to oxidation of surface. Sullivan²⁵ has also observed differences in the stability of electro-etched and chemically-treated surfaces.

Different surfaces can be prepared chemically which show more than the amount of variation normally found in pilot and manufacturing process lines. However, extreme variations in storage ambients have relatively little significant effects on any of these surfaces. It is therefore concluded that although certain chemical treatments may affect forming, the variations in process yields are not attributable to interaction between the germanium surface and storage ambients.

The results of Sections 4.2.2 and 4.2.3 suggest the possibility of passivation of the germanium surface. An electro-etched surface followed by an HF treatment exhibits a higher degree of stability to ambient than does a superoxol-etched surface treated in the same way. Treatment of a lapped germanium surface with two components of CP₄ (HF + HNO₃) will inhibit subsequent etching in superoxol.

The possibility that contamination *before* etching may affect the characteristics of the germanium surface after etching is considered. Experiments show that contamination of the germanium with corrosive zinc chloride-ammonium chloride flux before etching significantly affects the rectification properties of the germanium surface obtained after etching. The surface recombination velocity (in so far as it is determinative of the turn-off time of the transistor) is also significantly affected. However, on the basis of the results quoted here, it is not possible to conclude that such contamination can account for an appreciable amount of the unassignable variability in forming yields experienced in pilot and manufacturing process lines involving soldered base-wafer connections.

5. GENERAL CONCLUDING REMARKS

The experiments which have been described have implications which are important in both design and processing of point-contact transistors. These are summarized below:

5.1 Point-Contact Transistors with High Current Gain

In most switching applications the combination of high current gain and low reverse current is desirable. The measurements of current gain, taken together with the potential probe measurements in Section 2.2.1, indicate that, for the structures used here, the reverse collector current at operating voltage must be large enough to set up a substantial drift field before efficient collection of holes can occur. If this condition is not met, either the unit has low gain at all values of emitter current (unformed), or develops a bistability of the kind described in Section 2.3 (partially formed). For a given structure, the drift field can be increased by increasing resistivity of the germanium at the expense of increased base resistance. Here thermal stability of the contact also provides a limit. A more likely expedient, in the case of germanium, is to decrease the area of the formed collector junction by using sharper points and modified forming technique. The limits here are produced by reliability requirements for mechanical stability of the point structure.

5.2 Current Multiplication in Unformed Transistors

Many experiments have reported on junction transistors with high current gains which are attributable to the p-n hook mechanism. The high values of current gain observed with conventionally formed point contact transistors have been attributed to various mechanisms, among which is the hypothesis of a p-n hook structure,⁶ primarily in the bulk of the germanium, introduced by the pulsing of the donor-doped point. In particular, at small emitter currents small signal α -values in conventionally formed collectors may reach values as high as ten, and values of α as large as 100 are encountered in formed collectors exhibiting anomalous output characteristics. However, the average α over a 6-ma emitter current range is usually near the value of 3.1 which would be expected from the mobility ratio of holes and electrons with the Type-A transistor geometry. The increase in reverse current of a formed collector by addition of donor to the point wire may result from the production of a hook structure. However, information is needed concerning the importance of the hook structure in accounting for the high values of α encountered at low emitter currents, or in connection with collector characteristic anomalies in conventionally formed point-contact transistors.

The unformed transistors discussed in this article differ from electrically formed units in that the collector barrier is the one at the metal-semiconductor surface. It has been found that certain chemical treatments can produce a collector barrier which allows an increased reverse

current flow and a substantial drift field near the emitter. Some of these units show an α value at all emitter currents quite comparable in magnitude to that of conventionally formed collectors, and surface treatment alone can also introduce in these unformed collector characteristics anomalies similar to those found in some formed units. It is difficult to visualize a p-n hook structure arising at the germanium surface as a result of the chemical treatments discussed. If such a possibility is precluded, the p-n hook mechanism does not seem necessary to the attaining of high α values at low emitter currents, or an α emitter current dependence of the kind normally observed in anomaly-free units. To account for values of α obtained with unformed collectors at low emitter currents, other mechanisms, such as the suggestion of Shockley, involving hole trapping in the germanium under the collector point^{6, 7} or the suggestion of Van Roosbroeck,²⁹ involving conductivity modulation, might in this case be more suitable.

Further, unformed transistors made by appropriate chemical treatments can duplicate qualitatively the electrical characteristics of conventionally formed units, including alpha-emitter current dependence and output characteristic anomalies of types (1), (2) and (3). These phenomena can thus occur under circumstances where a well-defined hook structure is improbable.

5.3 Surface Properties and Transistor Forming

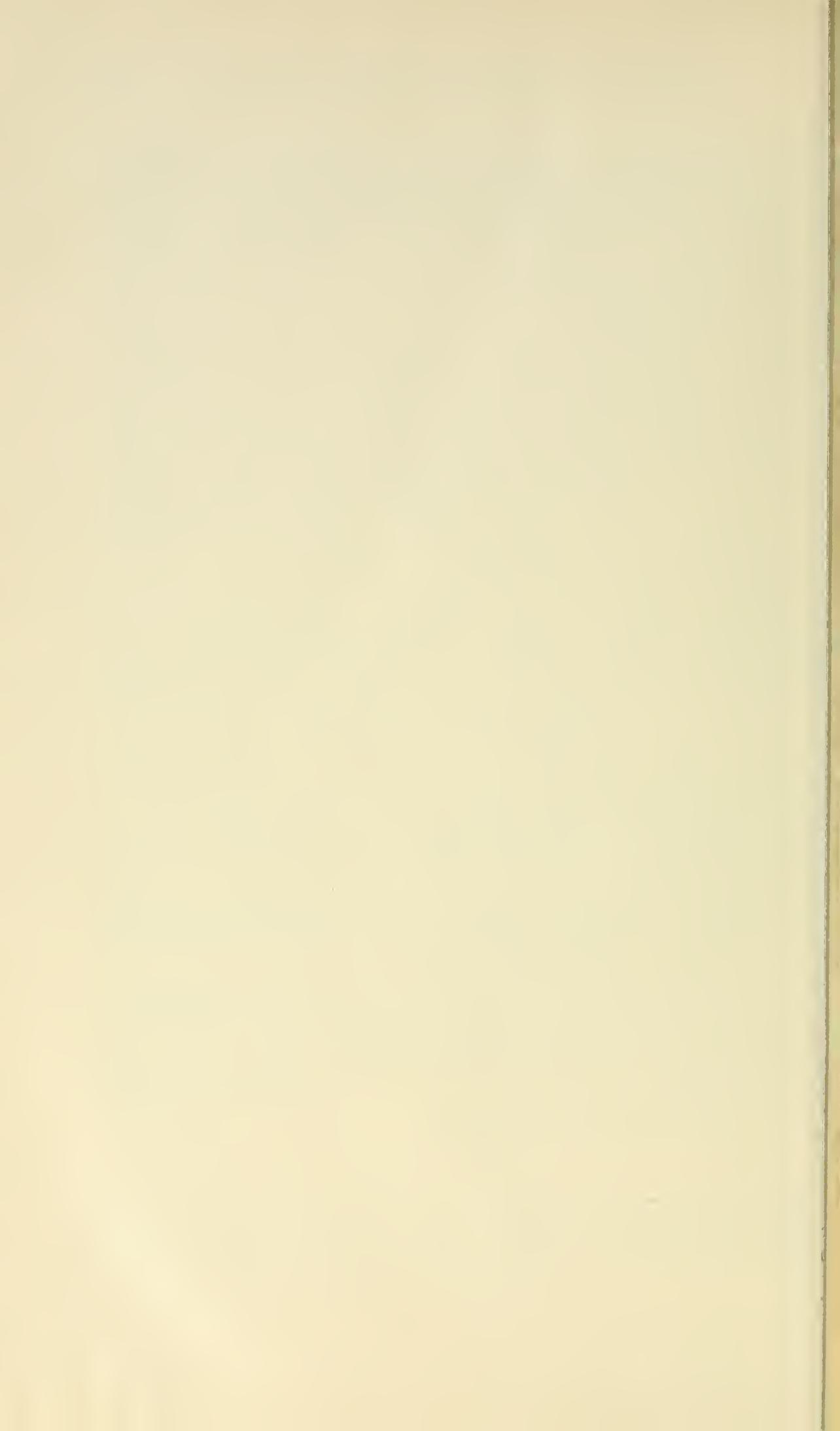
It has been found that a major factor in determining the forming yield of point-contact transistors is the chemical history of the surface. Thus in processing of point-contact transistors, major attention should be paid to ensuring chemical control of the base wafer surface if the forming yield is to be kept high. On the other hand, considerable variation may apparently be tolerated in storage ambients. Of course it has not been shown that such variations in storage conditions do not have an effect on subsequent reliability of the product. Processes which permit exposure of surfaces to solder fumes either before or after etching are to be regarded with suspicion. Monitoring of the reverse emitter diode characteristics should prove useful as a means of securing proper control of the pre-forming surface.

ACKNOWLEDGEMENT

The authors wish to acknowledge the help of M. S. Jones, who carried out many of the experiments mentioned here, and N. Carthage who did the electroetching work. The continued support and encouragement of N. J. Herbert has been greatly appreciated.

REFERENCES

1. J. Bardeen and W. H. Brattain, Physical Principles Involved in Transistor action, *Phys. Rev.* **75**, p. 1213, April 15, 1949.
2. J. Bardeen and W. G. Pfann, Effects of Electrical Forming on the Rectifying Barriers of n- and p-Germanium Transistors, *Phys. Rev.* **77**, p. 401-402, Feb. 1, 1950.
3. W. Shockley, Electrons and Holes in Semiconductors, D. VanNostrand Company, New York, N. Y., p. 110.
4. Reference 3, p. 111.
5. L. B. Valdes, Transistor Forming Effects in n-Type Germanium, *Proc. I.R.E.* **40**, p. 446, April, 1952.
6. W. Shockley, Theories of High Values of Alpha for Collector Contacts on Germanium, *Phys. Rev.* **78**, p. 294-295, May 1, 1950.
7. W. R. Sittner, Current Multiplication in the Type A Transistor, *Proc. I.R.E.*, **40**, pp. 448-454, April, 1952.
8. Valdes (Reference 5) reports large concentrations of copper present in the p-germanium under heavily formed phosphor-bronze points.
9. W. G. Pfann, Significance of Composition of Contact Point in Rectifying Junctions on Germanium, *Phys. Rev.* **81**, p. 882, March 1, 1951.
10. C. S. Fuller and J. D. Struthers, Copper as an Acceptor Element in Germanium, *Phys. Rev.* **87**, p. 526, Aug. 1, 1952.
11. C. S. Fuller, Diffusion of Acceptor and Donor Elements into Germanium, *Phys. Rev.* **86**, p. 136, April 1, 1952.
12. Reference 5, p. 448.
13. Personal communication, H. E. Corey, Jr.
14. L. E. Miller, Negative Resistance Regions in the Collector Characteristics of Point Contact Transistors, *Proc. I.R.E.*, **40**, p. 65-72, Jan. 1, 1956.
15. Reference 1, p. 1225.
16. John Bardeen, Surface States and Rectification at a Metal Semiconductor Contact, *Phys. Rev.*, **71**, p. 717-727, May, 15, 1947.
17. I. Tamm, über eine Mögliche Art der Elektronenbindung an Kristalloberflächen, *Physik, Zeits. Sowjetunion*, **1**, 1932, p. 733.
18. W. H. Brattain and J. Bardeen, Surface Properties of Germanium, *B. S. T. J.*, **32**, pp. 1-41, Jan., 1953.
19. W. L. Brown, n-Type Surface Conductivity on p-Type Germanium, *Phys. Rev.* **91**, pp. 518-527, Aug. 1, 1953.
20. W. H. Brattain and C. G. B. Garrett, private communication.
21. R. D. Heidenreich, private communication.
22. O. H. Johnson, Germanium and its Inorganic Compounds, *Chem. Rev.* **51**, pp. 431-469, 1952.
23. M. Kikurehi and T. Onishi, A Thermo-Electrical Study of the Electrical Forming of Germanium Rectifiers, *J. App. Phys.*, **24**, pp. 162-166, Feb., 1953.
24. R. H. Kingston, Water-Vapor Induced n-Type Surface Conductivity on p-Type Germanium, *Phys. Rev.*, **98**, 1766-1775, June 15, 1955.
25. M. V. Sullivan, personal communication.
26. J. T. Law, A Mechanism for Water Induced Excess Reverse Current on Grown Germanium n-p Junctions, *Proc. I. R. E.*, **42**, pp. 1367-1370, Sept., 1954.
27. E. Billig, Effect of Minority Carriers on the Breakdown of Point Contact Rectifiers, *Phys. Rev.* **87**, p. 1060, Sept. 15, 1952.
28. G. W. Snedecor, Statistical Methods, The Iowa State College Press, Ames, Iowa, 1946.
29. W. VanRoosbroeck, Design of Transistors with Large Current Amplification, *J. App. Phys.*, **23**, p. 1411, Dec., 1952.



The Design of Tetrode Transistor Amplifiers

By J. G. LINVILL and L. G. SCHIMPF

(Manuscript received March 7, 1956)

The design of tetrode transistor amplifiers encounters problems of the type that occurs with other transistor uses. Desired frequency characteristics, limitations of parasitic elements, and other practical considerations impose constraints on the range of terminations that can be employed. With many transistors, one can terminate a transistor so that it will oscillate without external feedback; this oscillation or other exceedingly sensitive terminations must be avoided.

The two-port parameters of the transistor in any orientation in which it is to be used constitute the fixed or given information which is the starting point of the amplifier design. Using this starting point, methods are developed by which one can select, on simple bases, the kinds of terminations that will be suitable. To facilitate the design of amplifiers, a set of charts has been developed from which one can read power gain and input impedance as functions of the load termination.

Illustrative tetrode amplifiers are described. These include a common base 20-mc video amplifier, a common-emitter 10-mc video amplifier, an IF amplifier centered at 30 mc, and an IF amplifier centered at 70 mc. Predicted and measured gains are compared.

INTRODUCTION

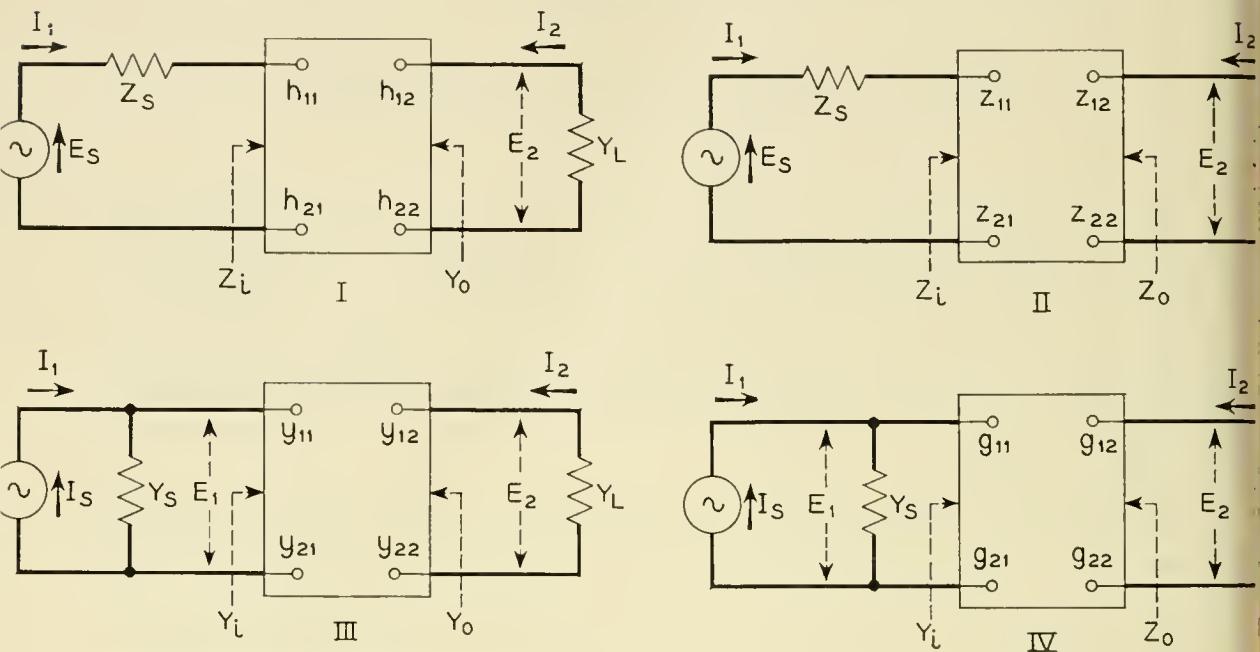
Junction tetrode transistors¹ of the type currently produced for research purposes at Bell Telephone Laboratories are suitable for high-frequency applications. They are being studied for use in video amplifiers, as IF amplifiers where the center frequency is below 100 mc, for oscillators up to 1,000 mc and for very fast pulse circuits.

Their application in amplifiers brings up design considerations similar to those encountered for other transistors but with differences resulting

¹ R. L. Wallace, L. G. Schimpf and E. Dickten, A Junction Transistor Tetrode for High-Frequency Use, Proc. I.R.E., 40, pp. 1,395-1,400, Nov. 1952.

from different parameter values and variation. The analysis presented in this paper regarding amplifier design was motivated by the study of tetrodes, but the results are equally applicable for other types.

The design of an amplifier begins with a characterization of the transistor which is suitable for the study of its performance as an amplifier. From this characterization, or functional representation, one



CORRESPONDING QUANTITIES

I	II	III	IV
h_{11}	Z_{11}	y_{11}	g_{11}
h_{12}	Z_{12}	y_{12}	g_{12}
h_{21}	Z_{21}	y_{21}	g_{21}
h_{22}	Z_{22}	y_{22}	g_{22}
I_1	I_1	E_1	E_1
E_1	E_1	I_1	I_1
E_2	I_2	E_2	I_2
I_2	E_2	I_2	E_2
E_s	E_s	I_s	I_s
Z_s	Z_s	Y_s	Y_s
Y_L	Z_L	Y_L	Z_L
Z_i	Z_i	Y_i	Y_i
Y_o	Z_o	Y_o	Z_o

RELATIONSHIPS BELOW ARE BETWEEN QUANTITIES IN COLUMN I. CORRESPONDING RELATIONSHIPS WRITTEN DIRECTLY FOR CORRESPONDING QUANTITIES IN ANY OTHER COLUMN.

$$(1) \quad E_1 = I_1 h_{11} + E_2 h_{12}$$

$$(2) \quad I_2 = I_1 h_{21} + E_2 h_{22}$$

$$(3) \quad Z_i = h_{11} - \frac{h_{12} h_{21}}{Y_L + h_{22}}$$

$$(4) \quad Y_o = h_{22} - \frac{h_{12} h_{21}}{Z_s + h_{11}}$$

$$(5) \quad I_2 = \frac{h_{21} E_s Y_L}{(h_{11} + Z_s)(h_{22} + Y_L) - h_{12} h_{21}}$$

Fig. 1 — Two-port parameters with summary of relationships.

determines the potentialities of amplifiers employing the transistor and designs a suitable amplifier circuit. This step involves answering two questions: What performance, maximum power gain for instance, is it possible to obtain? What source and load impedances should the transistor be associated with?

Two-Port Parameters of Transistors

For circuit applications, the two-port parameters are the most convenient for characterization of the transistor. These parameters implicitly but completely characterize the device from the performance standpoint.

Four sets of two-port parameters are illustrated in Fig. 1. Any set can be calculated from any other set, and the choice of the set to employ is determined only by convenience in the use of available measuring equipment and the preference of the designer. The relationships between parameters, input and output impedances, voltage and current ratios are summarized on Fig. 1. The same expressions given there for h 's can be used for any parameter set so long as one uses the corresponding quantities applicable to the desired parameter set.

Though the transistor can be operated as an amplifier with the base, emitter or collector common between the input and output terminal pairs, the two-port parameters for any of the connections can be used to calculate the parameters for any other connection.

For determination of the two-port parameters of tetrode transistors, R. L. Wallace suggested the use of two-terminal impedance measurements with subsequent calculation of the two-port parameters of interest from these. The impedances indicated in Fig. 2 have proved simple to measure at typical operating points with conventional high-frequency bridges. These impedances have been measured at a set of frequencies extending to 30 me. Because of the number of transistors measured it has been economical to program a digital computer to calculate two-port parameters and other quantities of interest from the measured two-terminal impedances.

THE RELATIONSHIPS OF TRANSISTOR PARAMETERS TO AMPLIFIER PERFORMANCE

Any of the sets of two-port parameters implicitly characterize all of the linear properties of the transistor for the range of frequencies for which the parameters have been measured. As mentioned before, it is necessary to translate the parameters into answers to the following questions. How much amplification can the transistor give at a particular

frequency? What impedance should it be supplied from? What impedance should it feed? What gain will be obtained using a pair of impedances different from the optimum ones? The answering of these and related questions amounts to establishing a convenient means of translating the parameter values into the quantities of interest applying to the amplifier. Such a convenient translating means for solving these problems is described in this section.

Earlier explicit solutions to special cases of the problem are well known. Wallace and Pietenpol² have given simple expressions in terms of the transistor parameters for matching input and output impedances and the maximum available gain when the transistor has purely real parameters. An implicit solution for optimum source and load impedances for maximum gain in the complex case has been known for a long time. It is simply that the transistor be terminated at the input and output by conjugate matching impedances. The implicit nature of this solution arises from the fact that the input impedance is a function of the load impedance, and the output impedance is a function of the source impedance for transistors with internal feedback. The solution for optimum source and load impedance from this approach amounts to the solution of simultaneous quadratic equations with complex unknowns and becomes involved.

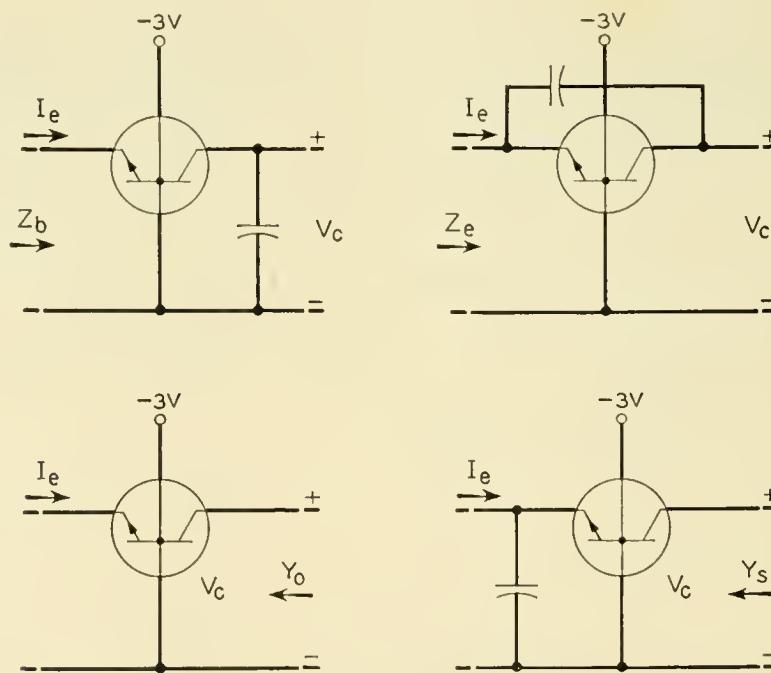


Fig. 2 — Two terminal impedance measurements for determination of two-port parameters.

² R. L. Wallace and W. J. Pietenpol, Some Circuit Properties of n-p-n Transistors, Proc. I.R.E., 39, pp. 753-67, July, 1951.

From the approach to the problem taken in this paper, one solves first for the maximum power gain and subsequently determines the optimum terminations. It turns out that the solutions leads to explicit relationships for optimum performance and terminations and also leads to charts from which power gains and input impedance can be read for any terminations.

In all expressions to be developed, the h parameters are used. Precisely the same expressions can be obtained for z 's, y 's, or g 's provided that one uses the corresponding quantities in the table of Fig. 1.

The maximum power gain is a quantity of primary interest in transistors since the transistor ordinarily has a resistive component in its driving-point impedance. Thus voltage or current amplification is constrained by the limited power gain attainable. In some cases, however, because of the inherent feedback internal to the device, instability can result simply from proper passive terminations without application of any additional feedback. Such cases are distinct because of this property. Transistors exhibiting this possibility are said to be potentially unstable at the frequency in question.

A quantity of interest presented here and derived later is a particular power gain defined for h -parameters as

$$\frac{\text{power out}}{\text{power in}} = \frac{P_{00}}{P_{i0}} = \frac{|h_{21}|^2}{4h_{11r}h_{22r} - 2\text{Re}(h_{12}h_{21})} \quad (1)$$

where h_{11r} and h_{22r} mean the real part of h_{11} and of h_{22} . $\text{Re}(h_{12}h_{21})$ means the real part of the product of h_{12} and h_{21} . Unless the amplifier is potentially unstable, the quantity P_{00}/P_{i0} is within 3 db of the maximum available gain for the transistor.

The matter of potential instability of the transistor is of great interest. Certainly the transistor is potentially unstable if P_{00}/P_{i0} is negative. Otherwise potential instability is indicated by greater than unity values of the criticalness factor

$$C = 2 \frac{P_{00}}{P_{i0}} \left| \frac{h_{12}}{h_{21}} \right| \quad (2)$$

If the transistor is not potentially unstable the maximum available gain is $K_G(P_{00}/P_{i0})$ where

$$K_G = \frac{2(1 - \sqrt{1 - C^2})}{C^2} \quad (3)$$

For $0 \leq C \leq 1$, $1 \leq K_G \leq 2$. A plot of K_G as a function of C is shown in Fig. 3. The function is seen to be exceedingly flat near $K_G = 1$ for

C between zero and 0.6. Thus the value P_{00}/P_{i0} in the majority of cases where the transistor is not potentially unstable is a close approximation to the maximum available gain.

The optimum source and load impedances can be expressed in terms of the transistor parameters and other quantities given in terms of them by the following relationships where the transistor is not potentially unstable.

$$G = 1 \left| \operatorname{Arg} (-\overline{h_{12}h_{21}}) \right| = e^{j\theta} \quad (4)^3$$

$$Z_s \text{ opt} = \bar{Z}_{in} = h_{11} - \frac{h_{12}h_{21}}{2h_{22r}} \left(1 - \frac{CK_g G}{2} \right) \quad (5)$$

$$Y_L \text{ opt} = -h_{22} + \frac{\frac{2h_{22r}}{1 - \frac{CK_g G}{2}}}{1 - \frac{CK_g G}{2}} \quad (6)$$

Though explicit relationships for ideal terminations and for the maximum power gain which one can achieve with a transistor are of interest, such terminations limit the band width of the amplifiers. Therefore, it is important to have convenient means for evaluating power gain and input impedance for other than ideal terminations in order to realize a desired bandwidth. A chart which facilitates computation of these quantities is now developed from an analysis which leads to the other results quoted above.

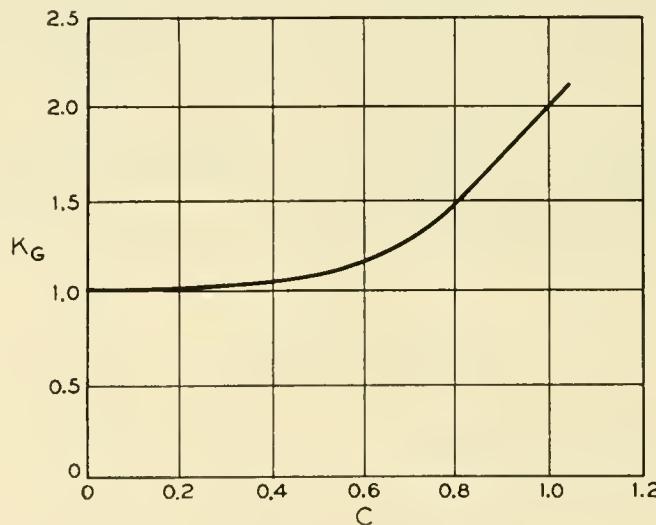


FIG. 3 — K_G plotted as a function of C .

³ If $-\overline{h_{12}h_{21}}$ is $c + jd$, then $\theta = \tan^{-1}(d/c)$; $G = e^{j\theta}$ and $\overline{-h_{12}h_{21}}$ is the conjugate of $-\overline{h_{12}h_{21}}$.

Power Flow in a Two-Port Device

A convenient point of departure in the analysis of power amplification in a transistor or other linear two-port device is the arrangement shown in Fig. 4. The two-port is supplied by a unit current at the frequency of interest and at reference phase at the input terminal pair. The output of the two-port is connected to a voltage source of the same frequency. The input-current and output-voltage time functions are

$$i_1 = Re\sqrt{2}e^{j\omega t} = Re\sqrt{2}I_1 e^{j\omega t} \quad (7)$$

and

$$\begin{aligned} e_2 &= Re\sqrt{2}(a + jb)e^{j\omega t} = Re\sqrt{2}(L + jM) \left(\frac{-h_{21}}{2h_{22}r} \right) e^{j\omega t} \\ &= Re\sqrt{2}E_2 e^{j\omega t} \end{aligned} \quad (8)$$

In (8), L and M are introduced for simplicity in some later relationships.

The whole analysis is essentially a study of power flow in the circuit shown in Fig. 4 as L and M of (8) are varied. All possible terminations and excitations can be simulated simply by varying L and M . Under some conditions the voltage source will absorb power; under others it will supply power to the two-port. Ordinarily the current source supplies power to the two-port, but for appropriate ranges of L and M if the two-port is potentially unstable, the transistor may supply power both to the current source and the voltage source. The problem of evaluating maximum power gain is simply finding the values of L and M corresponding to the greatest ratio of power out to power in. The load impedance to which this situation corresponds is $E_2/-I_2$. The input impedance for this condition is simply E_1/I_1 , and the optimum source impedance is the complex conjugate of the latter quantity.

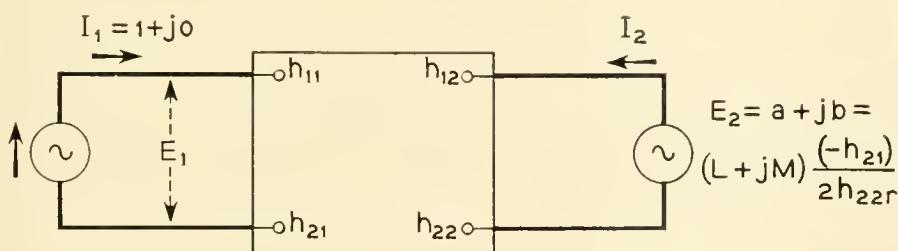


Fig. 4 — A two-port device supplied by a current source and feeding into a voltage source.

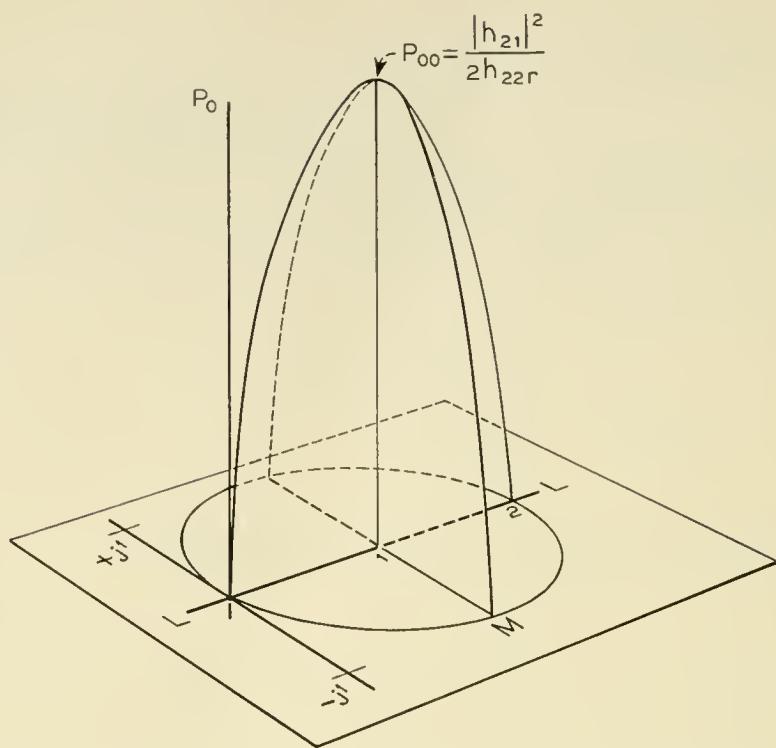


Fig. 5 — Sketch of power output as a function of L and M .

PROJECTION IN L - M PLANE OF GRADIENT

LINE IS G OR $\text{Arg}-h_{12}h_{21}$

$$\text{SLOPE OF PLANE ALONG } G \text{ IS } \left| \frac{h_{21} h_{12}}{2h_{22r}} \right|$$

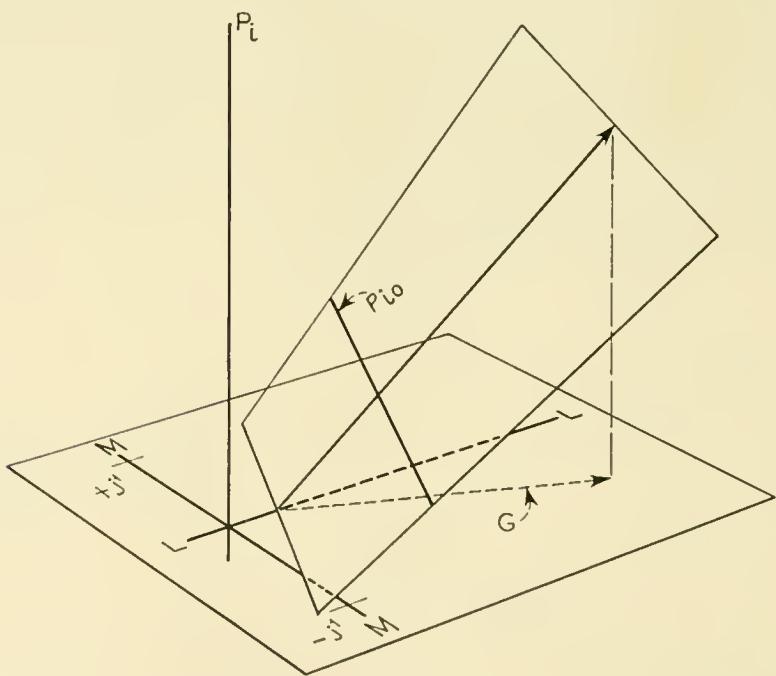


Fig. 6 — Sketch of power input as a function of L and M .

The output power can be readily evaluated in terms of L and M .

$$I_2 = I_1 h_{21} + E_2 h_{22} \quad (9)$$

$$I_2 = (1 + j0)h_{21} + (L + jM)h_{22} \frac{(-h_{21})}{2h_{22r}} \quad (10)$$

$$\text{Power out} = P_0 = Re(-\bar{E}_2 I_2) \quad (11)$$

$$P_0 = Re \left[\frac{(L - jM)\bar{h}_{21}}{2h_{22r}} h_{21} - (L^2 + M^2) \frac{|h_{21}|^2}{4h_{22r}} \right] \quad (12)$$

$$= L \frac{|h_{21}|^2}{2h_{22r}} - (L^2 + M^2) \frac{|h_{21}|^2}{2h_{22r}} \quad (13)$$

On the basis of (13) the power output plotted as a function of L and M is a paraboloid as shown in Fig. 5, having the pertinent dimensions indicated there. Only within the circle centered at $L = 1$, $M = 0$ and passing through the origin does one obtain positive power output. The apex of the paraboloid corresponds to

$$P_0 = P_{00} = \frac{|h_{21}|^2}{4h_{22r}} \quad (14)$$

The input power can similarly be evaluated in terms of L and M .

$$E_1 = I_1 h_{11} + E_2 h_{12} \quad (15)$$

$$= (1 + j0)h_{11} + (L + jM) \frac{(-h_{21})}{2h_{22r}} h_{12} \quad (16)$$

$$\text{Power in} = P_i = Re[E_1 I_1] \quad (17)$$

$$P_i = Re \left[h_{11} + (L + jM) \frac{(-h_{21})h_{12}}{2h_{22r}} \right] \quad (18)$$

$$= h_{11r} - LRe \frac{(h_{12}h_{21})}{2h_{22r}} + MIm \frac{(h_{12}h_{21})}{2h_{22r}} \quad (19)$$

where $Im[(h_{12}h_{21})/2h_{22r}]$ means the imaginary part of the expression in parenthesis.

On the basis of Eq. 19 the input power plotted as a function of L and M is simply an inclined plane having the properties indicated on Figure 6.

Since Figures 5 and 6 turn out to be such simple geometrical figures the problem of finding the point of maximum ratio of P_0 to P_i is very simple and other interpretations are easy to make. First, a negative value of P_{i0} (P_i at 1, 0) certainly indicates potential instability for both input and output terminations receive power from the two-port. Even if the plane of P_i intersects the L - M plane within the unit circle centered at

1, 0, then the two-port is potentially unstable since on one side of the intersection both input and output terminations receive power from the two-port. The change in P_i from the minimum value found on the unit circle centered at 1, 0 to P_{i0} divided by P_{i0} is the criticalness factor, C . Values of C greater than unity indicate potential instability.

The power input at 1, 0 is

$$P_{i0} = \frac{2h_{11r}h_{22r} - Re(h_{12}h_{21})}{2h_{22r}} \quad (20)$$

Using (14) and (20), one obtains

$$\frac{P_{00}}{P_{i0}} = \frac{|h_{21}|^2}{4h_{11r}h_{22r} - 2Re(h_{12}h_{21})} \quad (21)$$

$$C = \frac{\frac{|h_{12}h_{21}|}{2h_{22r}}}{\frac{2h_{11r}h_{22r} - Re(h_{12}h_{21})}{2h_{22r}}} = 2 \frac{P_{00}}{P_{i0}} \left| \frac{h_{12}}{h_{21}} \right| \quad (22)$$

Now if the plane of power input, Fig. 6, is parallel to the L - M plane and above it, certainly the point of maximum power gain is the apex of the paraboloid, 1, 0 in Fig. 5. If the plane is inclined but always above the unit circle centered at 1, 0 certainly the point of maximum power gain is downward along the gradient line which lies above the point 1, 0. This must be so since for any contour of equal power out (a circle of fixed elevation around the paraboloid) the minimum power input (or greatest gain) lies along the line of steepest descent from 1, 0 in Fig. 6. Thus the problem of evaluation of the maximum available gain reduces to the simple problem of finding the abscissa of Fig. 7 where the ratio of ordinates of the parabola and straight line is a maximum. The parabola

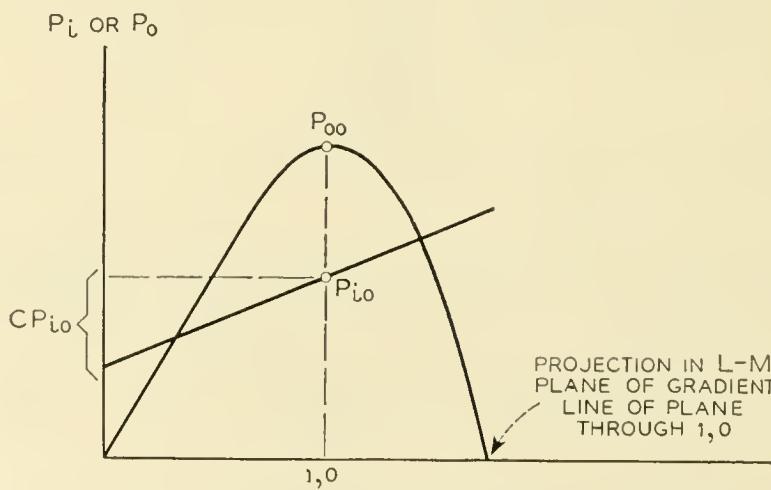


Fig. 7 — Section of paraboloid and inclined plane of Figs. 5 and 6.

and straight line are sections of the paraboloid and plane through the gradient line of the plane over 1, 0.

A straightforward analysis indicates that the point in the $L-M$ plane where the maximum of P_0/P_i occurs is at

$$L + jM = 1 - \frac{CK_g G}{2} \quad (23)$$

where these quantities are defined as in (2), (3), and (4). The power gain at this optimum point is K_g times that obtained at 1, 0. One finds that the maximum gain is only two times P_{00}/P_{i0} even if C approaches unity which corresponds to the marginal case of potential instability.

The analysis just described leads to the maximum values of power gain and to the best terminating impedances. For many design problems these answers are a guide but one may prefer to use other than optimum values for other compelling reasons. For such a case charts from which one can get the pertinent quantities are very helpful.

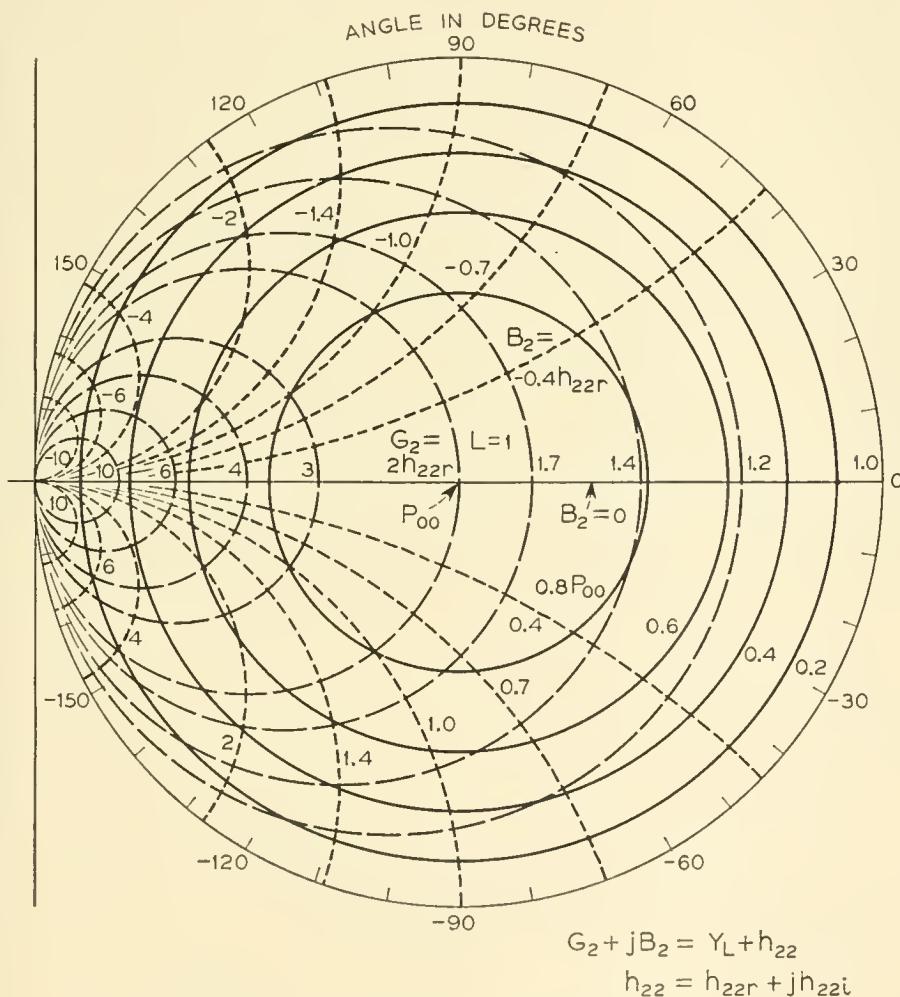
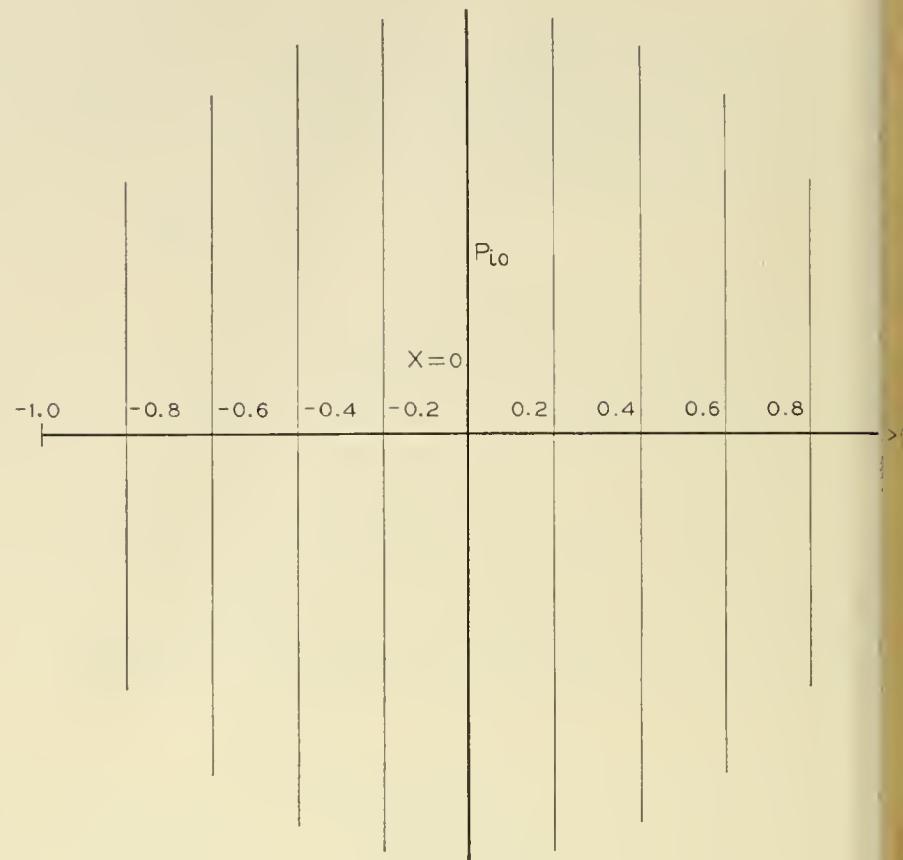


Fig. 8 — Gain and impedance chart.

Fig. 9(a) — Input power as a function of X .*Development of Transmission and Impedance Charts*

The same point of departure employed in the evaluation of optimum cases leads to a convenient set of charts. Equation 12 shows that a set of concentric circles centered at 1, 0 are loci in the L - M plane of constant power output for a unit current source at the input. It is convenient to plot these as is done on Figure 8, showing P_0 as a fraction of P_{00} .

$$\frac{P_0}{|h_{21}|^2} = \frac{P_0}{P_{00}} = 1 - (L - 1)^2 - M^2 \quad (24)$$

Since Y_L , the load admittance, is $-I_2/E_2$, using (10) one obtains

$$\frac{-I_2}{E_2} = Y_L = -h_{22} + \frac{2h_{22r}}{L + jM} \quad (25)$$

Now it is clear that if one defines G_2 and B_2 by

$$Y_2 = G_2 + jB_2 = Y_L + h_{22} = \frac{2h_{22r}}{L + jM} \quad (26)$$

loci of constant real and imaginary parts of Y_2 become the mutually orthogonal circles shown in Fig. 8. Thus the value of $L + jM$ is determined by the load admittance and two-port parameters.

Contours representing constant input power, with equal increments of power between successive contours, are always parallel equally-spaced lines in the $L-M$ plane. However, as may be seen from (19) and Fig. 6 different cases have different directions for the line normal to the contours, (the gradient line) and also different power increments for a given spacing of equal-power-input contours. It is convenient to define a new variable X which is the component along the gradient line of the vector starting at $L = 1, M = 0$ and going to L, M . Thus

$$P_i = P_{i0}(1 + CX) \quad (27)$$

Equation 27 suggests Fig. 9(a) which shows loci of constant power input plotted as a function of X . If Fig. 9(a) is shown on a transparent ma-

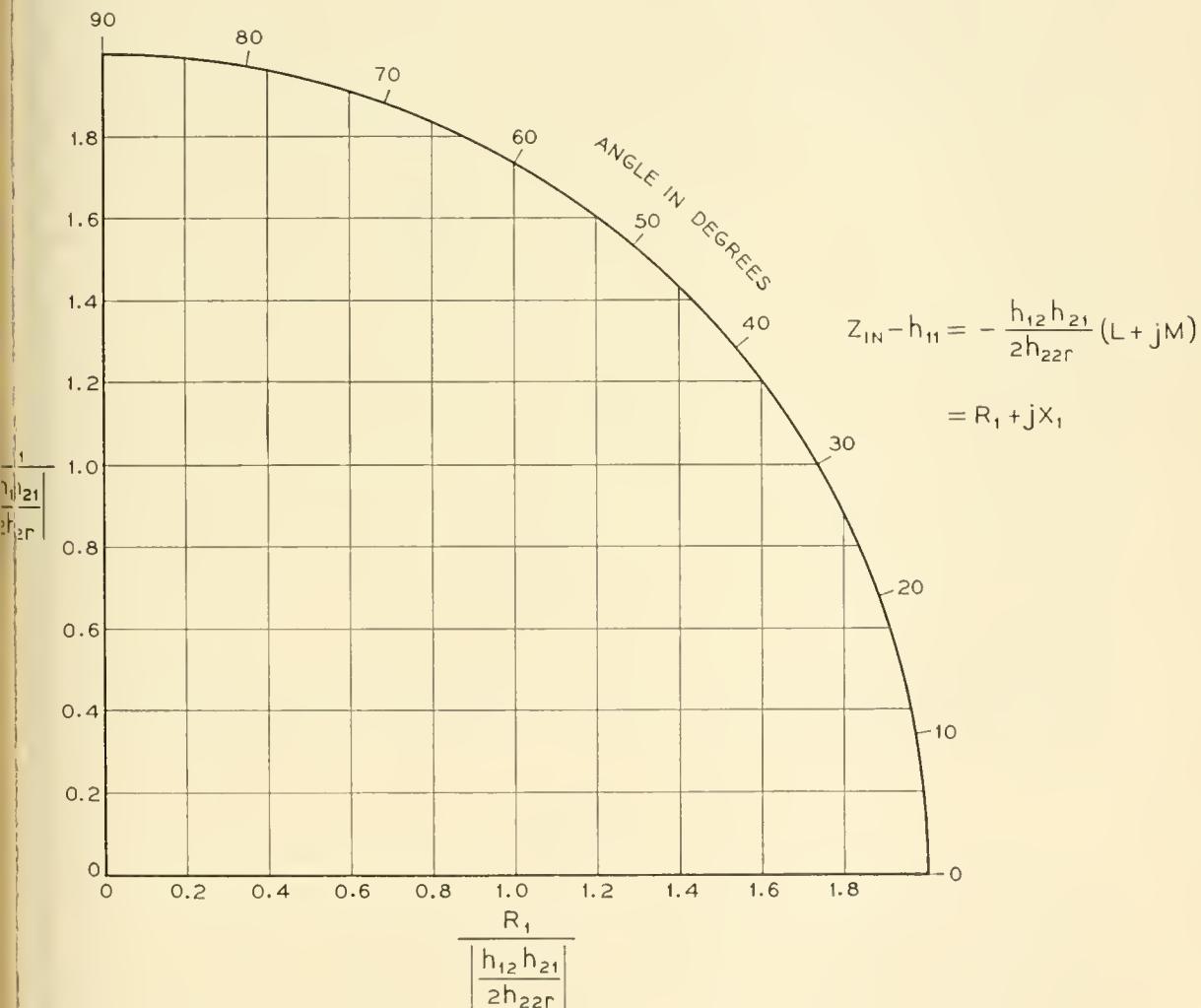


Fig. 9(b) — Input impedance as a function of L and M .

terial, its center (at $X = 0$ along the gradient line) can be superposed with the point $L = 1, M = 0$ of Fig. 8. With the gradient line of Fig. 9(a) oriented at the argument of $-\overline{h_{12}h_{21}}$, or θ in the L - M plane, one can easily determine graphically the power gain at any point in the L - M plane compared to the power gain at 1, 0. With Fig. 9(a) superposed on Fig. 8 as just described the viewer gets a bird's-eye-impression of the paraboloid of power output and the inclined plane of power input simultaneously. With such a bird's-eye view, it is easy to assess possibilities for power gain with all possible angles of load termination.

The evaluation in input impedance is done through use of (16) from which is obtained

$$\frac{E_1}{I_1} = Z_{in} = h_{11} + (L + jM) \frac{(-h_{12}h_{21})}{2h_{22r}}, \quad \text{or} \quad (28)$$

$$Z_{in} = h_{11} + (L + jM)(e^{-j\theta}) \left| \frac{h_{12}h_{21}}{2h_{22r}} \right| \quad (29)$$

For evaluating the second component of (29), it is convenient to have a second transparent overlay, Fig. 9(b), consisting of a rectangular grid to the same scale as the L - M plane, Fig. 8, with coordinates marked as

$$Re \frac{(Z_{in} - h_{11})}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|} = \frac{R_1}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|}$$

and

$$Im \frac{(Z_{in} - h_{11})}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|} = \frac{X_1}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|}$$

This overlay is placed over the L - M plane with the

$$\frac{R_1}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|}$$

axis making the angle θ with respect to the L axis. Thus on the rectangular overlay for any point in the L - M plane, one reads

$$\frac{Z_{in} - h_{11}}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|}$$

PARTICULAR DESIGNS OF TETRODE TRANSISTOR AMPLIFIERS

The charts and optimum relationships developed in the preceding section are convenient starting points in the design of amplifiers. They

do not ordinarily constitute a finished solution, however, since practical constraints frequently modify the design used. Moreover, all of the relationships are expressed on a single frequency basis, and many times the amplifier must operate over a range of frequencies broad enough that parameters change significantly over the range.

Four amplifier designs are described in this section: a single stage, common-base, 20-mc video amplifier; a common-emitter, 10-mc video amplifier; an IF amplifier at 30 me and a 60 to 80-mc IF amplifier. Parameter measurements made with bridges support the first three designs.

Parameter values and associated constants of a typical tetrode transistor are given in Table I. The quantities shown there reveal some interesting facts about the typical tetrode transistor represented. First, in the common-base connection the tetrode is potentially unstable at 30 mc but not at the lower frequencies. The common-emitter amplifier is potentially unstable at 1 and 3 mc. Second, the power gains of common-emitter and common-base stages are about the same at 30 me, the common-emitter connection giving more gain at low frequencies.

The matter of potential instability requires further consideration from a practical point of view. Potential instability at a frequency neither implies that a stable amplifier cannot be built at that particular frequency, nor does it imply that one can obtain an unlimited amount of stable amplification at that frequency. It does mean that by simultaneously tuning output and input one can adjust for oscillation. The region of potential instability corresponds to a region in which the input resistance may be negative for appropriate loads. Instability is avoided in the physical amplifier if one supplies the amplifier from a sufficiently high impedance that the input loop impedance always has a positive real part. To operate the amplifier with such a load that it presents a negative resistance to the source is attended by the difficulty that the amplification is more sensitive to changes in the source impedance than it is when the input resistance is positive. Hence the possible higher gain with internal positive feedback goes along with a greater sensitivity to changing termination impedance.

A Common-Base 20-Mc Video Amplifier

The data presented in Table 1 gives a quite comprehensive picture of possibilities for amplifier designs. To it must be added a practical fact. It is difficult to connect the load impedance without adding about 2 μuf of capacitance. This means that any termination considered must include about this amount of capacitance. By a theorem regarding

TABLE I — PARAMETERS AND ASSOCIATED CONSTANTS OF TETRODE NO. 668

Freq. Mc	0.6	1.0	3.0	10.0	30.0
Common Base					
h_{11}	$49 + j0.0$	$49 + j2.0$	$49 + j2.0$	$50 + j^2.0$	$51 + j^4.0$
h_{12}	$(1.1 + j0.24) \cdot 10^{-3}$	$(1.1 + j0.40) \cdot 10^{-3}$	$(1.2 + j1.3) \cdot 10^{-3}$	$(1.3 + j4.6) \cdot 10^{-3}$	$(2.1 + j11) \cdot 10^{-3}$
h_{21}	$-0.93 + j0.015$	$-0.93 + j0.023$	$-0.92 + j0.055$	$-0.88 + j0.13$	$-0.82 + j0.24$
h_{22}	$(3.3 + j5.3) \cdot 10^{-6}$	$(3.7 + j8.9) \cdot 10^{-6}$	$(6.8 + j25) \cdot 10^{-6}$	$(28 + j57) \cdot 10^{-6}$	$(35 + j140) \cdot 10^{-6}$
P_{00}/P_{i0}	310	310	230	87	40
C	0.78	0.79	0.89	0.94	1.3
$-\theta$	11°	20°	45°	65°	64°
$\left \frac{h_{12}h_{21}}{2h_{22r}} \right $	160	150	120	77	150
Common Emitter					
h_{11}	$720 - j160$	$660 - j210$	$440 - j270$	$220 - j210$	$130 - j140$
h_{12}	$(2.0 + j2.8) \cdot 10^{-3}$	$(3.0 + j4.3) \cdot 10^{-3}$	$(8.0 + j6.9) \cdot 10^{-3}$	$(1.5 + j11.0) \cdot 10^{-3}$	$(18 - j2.2) \cdot 10^{-3}$
h_{21}	$13 - j3.3$	$12 - j4.3$	$7.2 + j5.6$	$2.8 - j4.1$	$0.98 - j2.7$
h_{22}	$(6.4 + j6.4) \cdot 10^{-6}$	$(8.6 + j9.8) \cdot 10^{-6}$	$(1.9 + j17) \cdot 10^{-6}$	$(34 + j18) \cdot 10^{-6}$	$(40 + j17) \cdot 10^{-6}$
P_{00}/P_{i0}	1570	1330	567	121	44.4
C	0.80	1.1	1.3	0.75	0.57
$-\theta$	-139°	-145°	-177°	128°	103°
$\left \frac{h_{12}h_{21}}{2h_{22r}} \right $	359	381	250	111	65

passive impedances⁴ this puts an upper limit on the level of impedance presented by the load over a band of frequencies. The greatest possible constant level of load impedance over 20 mc is

$$|Z| = \frac{2}{C\omega} = \frac{2}{2.10^{-12} \cdot 2.10^7 \cdot 2\pi} = 7,960\Omega \quad (30)$$

Thus number, though not strictly applicable to this case, nonetheless gives a measure of the sort of value which one can expect. Hence one observes that for a broad-band video amplifier the load impedance is certainly going to be considerably less than $1/|h_{22}|$ which up to 10 mc is not less than 15,000 ohms. Moreover, the gain, if it is to be uniform, will certainly be limited by the gain obtainable at 20 mc.

Recognition that the load admittance will be a number of times h_{22r} ,

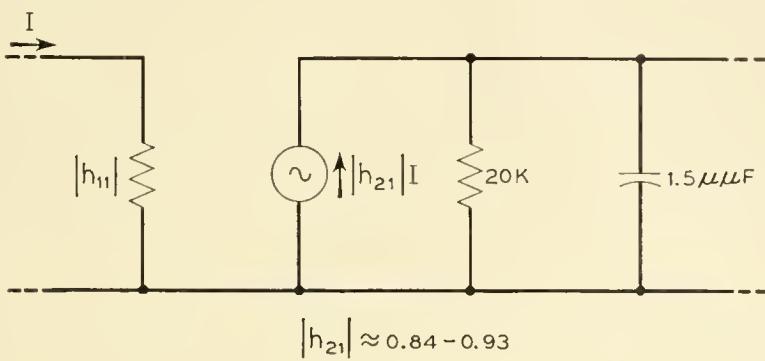


Fig. 10 — A rough approximant for the common base video amplifier. The variation of $|h_{21}|$ is a function of frequency and not variation between units.

five to ten, means that in Fig. 8 one will be operating near the origin where $(L + jM)$ is much less than one. Thus Z_{in} in (28) will be approximately h_{11} . Moreover, by superposing Fig. 9(a) on Fig. 8 at the correct angle for a frequency of 30 mc ($\theta = -64^\circ$) one observes that negative power input occurs only in the small section of circle cut-off by a chord running from the 80° to the 155° points on the periphery. This region is quite a way from the likely point of operation. Thus, this points out that the low impedance termination precludes instability due to internal feedback.

If the amplifier is supplied by a 75-ohm source, its output admittance at 30 mc (Equation 4, Figure 1) is $(7.0 + j20) \cdot 10^{-5}$ mho. At 10 mc the output admittance is $(4.2 + j8.8) \cdot 10^{-5}$ mhos.

These computations reveal that the amplifier in the common-base connection appears quite like the model shown in Fig. 10. Clearly, the

⁴ H. W. Bode, Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co., New York, 1945.

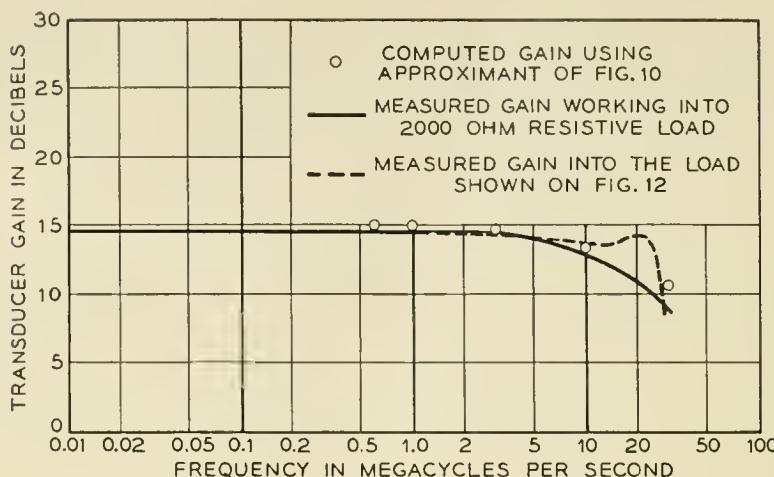


Fig. 11 — Measured and computed gain of a common base video amplifier.

amplification is obtained through the ratio of impedances of load to source.

Since it is impossible to match the output impedance for maximum gain due to reasons outlined above, Equation 5 of Fig. 1 can be used to compute the gain once the load impedance is determined. If we use a load impedance of 2,000 ohms and a source impedance of 75 ohms, the difference between the computed gain using the approximate of Fig. 10 and the exact expression (Equation 5 of Fig. 1) amounts to less than 1 db at frequencies up to 10 mc. At 30 mc, the exact expression results in a computed gain 1.5 db lower than that obtained from the approximation. A comparison of measured and computed gain for a common base video amplifier is shown on Fig. 11. Using a resistive load of 2,000 ohms a gain of 14.5 db is obtained at low frequencies and the response is down 3 db at 17 mc. To equalize the decreasing $|h_{21}|$ with frequency and the increasing effect of the capacitance, a load consisting of an inductor and a resistor is used. The circuit is shown on Fig. 12 and it will be noted from the response on Fig. 11, that the low frequency gain is 14.5 db with the 3 db point occurring at about 26 mc.

Common base stages can, of course, be cascaded to advantage only if impedance transformation is provided in the interstage coupling. Practical transformers or coupling networks may introduce undesirable band limitation. In the next section we will consider common-emitter stages which can be cascaded without impedance transformation.

Common Emitter 10-Mc Video Amplifier

To get a first idea of feasible impedance levels for a common-emitter video amplifier, one recognizes from Table I that the input impedance

will be within an order of magnitude of h_{11} , perhaps in the vicinity of 500 ohms. One sees that in this case if the termination impedances are equal, $Y_L \gg h_{22}$. Again, with reference to Fig. 8, the point of operation will be close to the origin of the L - M plane. Again the input impedance approximates h_{11} . The output admittance is given by

$$Y_0 = h_{22} - \frac{h_{12}h_{21}}{h_{11} + Z_s} \quad (31)$$

and if $Z_s = 500 \Omega$, Y_0 is $(3.3 + j1.0)10^{-4}$ mhos. A rough approximant to the common-emitter transistor is shown in Fig. 13. On an order of magnitude basis, one expects an iterative power gain of $|h_{21}|^2$ per stage. Final choice of elements amounts to computation using the approximant of Fig. 13 and experimental adjustment.

A video amplifier circuit employing the common emitter connection is shown in Fig. 14. If it is assumed that R_1 is zero and no compensating network is used in the output circuit, the gain characteristic can be computed from the approximant of Fig. 13, or if desired, using the exact expression (Equation (5) of Fig. 1). The two methods agree to within

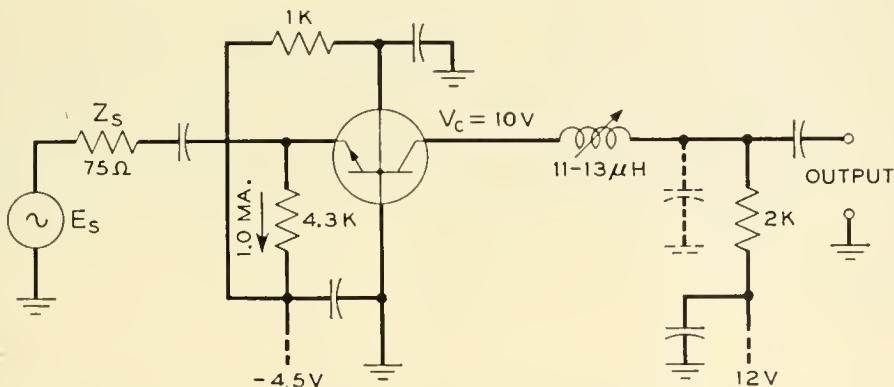


Fig. 12 — Circuit of a common base video amplifier. The series coil compensates for the decrease of $|h_{21}|$ with increasing frequency.

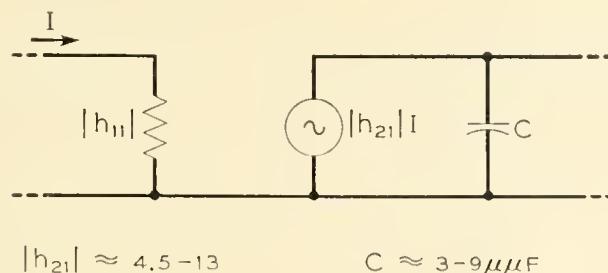


Fig. 13 — An approximant for a common-emitter stage video amplifier when terminated in a few hundred ohms. The variation of $|h_{21}|$ and C are a function of frequency.

about 1 db at frequencies up to 10 mc. Comparison of the measured and computed values is shown on Fig. 15 for a load of 500 ohms with no high-frequency compensation. The low-frequency gain is higher than for the common base connection but the response is down 3 db at 7 mc. By using the combination of R_1 in parallel with $800 \mu\mu F$ in the emitter circuit, negative feedback is introduced at low frequencies which results in the reduction of low frequency gain tending to make the response more uniform. In addition the $L-C$ network has been added in the output to compensate for the drop of $|h_{21}|$ with increasing frequency and the increasing effect of the output capacitance.

This results in the response shown as the dotted curve on Fig. 15. The low-frequency gain has been reduced to 17.5 db, but the response is now flat to within ± 0.3 db up to 13 mc and is 3 db down at 18 mc.

Although the data given on video amplifiers shows the results obtained using one transistor, similar response curves were obtained from some 6 or 8 units.

An I-F Amplifier Centered at 30 Mc.

The design of an IF amplifier at 30 mc is distinct from the preceding two cases in that one can use matching techniques over the narrow band.

Reference to Table 1 reveals that the common-base connection provides more potential gain at 30 mc than the common emitter connection; in fact, the common-base connection can be made to oscillate with certain terminations. The common-base connection is chosen for the 30-mc amplifier.

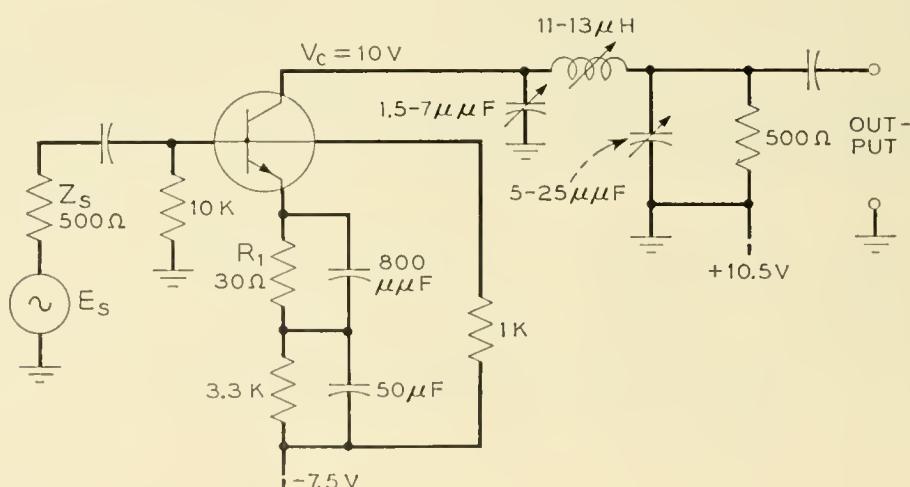


Fig. 14 — Circuit of a common emitter video amplifier. R_1 in parallel with $800 \mu\mu F$ and the $L-C$ network in the output circuit peak the response at 10 to 12 mc.

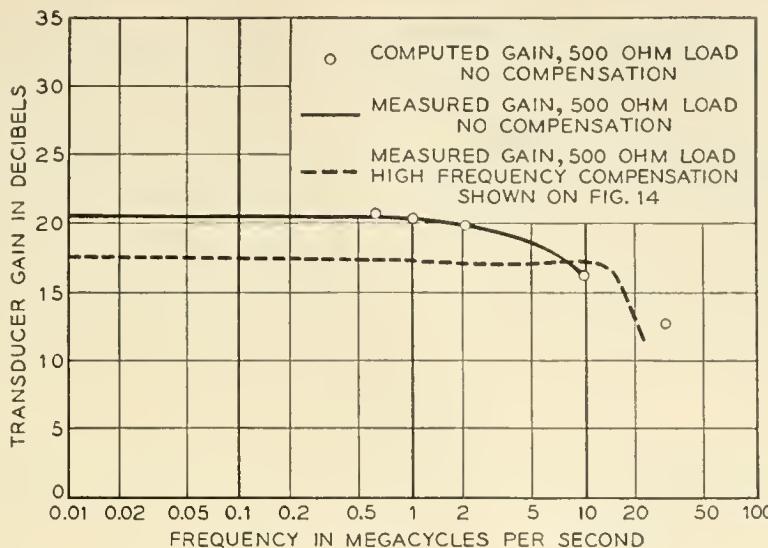


Fig. 15 — Computed and measured response of a common emitter amplifier.

In the design of the IF amplifier one is interested in a moderate range of frequencies. It will generally be true that the most frequency dependent parameters are the output and load admittances, since the load is to be tuned. One can take as a suitable load a parallel combination of a fixed conductance with a frequency dependent susceptance, the sort of termination typical of tuned circuits. Thus on Fig. 8, the locus of $G_2 + jB_2$ is one of the $G_2 = \text{Const.}$ circles.

Superposition of Fig. 9(b) on Fig. 8 with the

$$\frac{R_1}{\left| \frac{h_{12}h_{21}}{2h_{22r}} \right|}$$

axis making an angle of -64° with the L axis reveals that $Z_{in} - h_{11}$ has a negative real part on the upper left edges of all of the contours of constant G_2 . On the $G_2 = 2h_{22r}$ contour, $\text{Re}(Z_{in})$ reaches a minimum of 22.5 ohms. We select a load with $G_L = 2h_{22r}(G_2 = 3h_{22r})$ to avoid low values of input resistance resulting from the internal feedback.

Superposition of Fig. 9(a) on Fig. 8 with the gradient line making an angle of -64° through the point $L, M = 1, 0$ reveals that the maximum value of P_0/P_i on the $G_2 = 3h_{22r}$ circle is $1.87 P_{00}/P_{i0}$ and it occurs for $B_2 = -2h_{22r}$. The input impedance at this point is $36 + j87$ ohms.

For an amplifier one is primarily interested in

$$\frac{P_0}{\text{Power Available from Source}}$$

(which is called transducer gain) rather than P_0/P_i , the quantities just

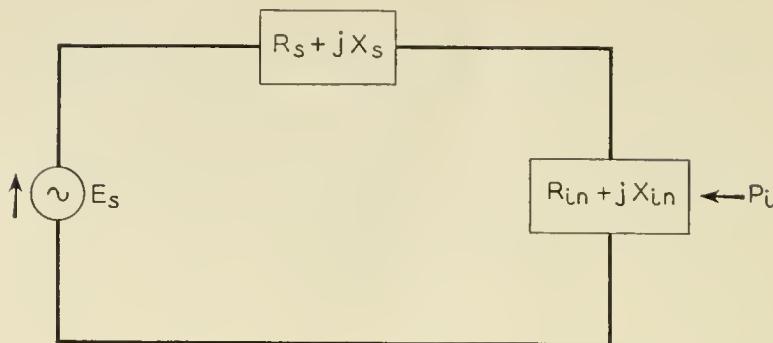


Fig. 16 — Typical input circuit.

read from the charts. From the source-load arrangement shown in Fig. 16, one readily computes

$$\frac{P_i}{\text{Power Available from Source}} = \frac{\frac{|E_s|^2 R_{in}}{(R_s + R_{in})^2 + (X_s + X_{in})^2}}{\frac{|E_s|^2}{4R_s}}$$

$$= \frac{4R_s R_{in}}{(R_s + R_{in})^2 + (X_s + X_{in})^2} \quad (32)$$

The source impedance selected for the amplifier is $75 - j87$ ohms at 30 mc. The 75 ohms is selected to reduce the effect of variations in input impedance when it is reduced further by the internal feedback. The 87 ohms of capacitive reactance is selected to tune the input reactance at the peak of response. Using Fig. 8 with the overlay of Fig. 9(a) along with (32) under the assumption that X_s varies insignificantly over the frequencies involved one obtains Table II. This table shows the variation of transducer gain as the value of B_2 is changed as well as indicating the value of B_2 required for the maximum gain. Thus, if the total output capacitance is known, the load admittance required to give the maximum gain at the desired frequency can be computed. As will be shown

TABLE II — EVALUATION OF TRANSDUCER GAIN OF I-F AMPLIFIER

B_2	$-5h_{22r}$	$-4h_{22r}$	$-3h_{22r}$	$-2h_{22r}$	$-h_{22r}$	$0h_{22r}$	h_{22r}
P_0/P_i	50	64	73	75	59	50	38
Z_{in}	$22 + j48$	$23 + j58$	$28 + j73$	$36 + j87$	$67 + j99$	$96 + j94$	$120 + j67$
$P_i/\text{Power available from source}$	0.61	0.66	0.78	0.87	0.98	0.98	0.94
Transducer gain	30	42	57	66	58	49	36
Gain, db	14.8	16.2	17.5	18.2	17.6	16.9	15.5

below, the bandwidth at which the response is down a given number of db can also be computed.

From Table II one observes that this design provides a gain of about 18 db with half power frequencies where the susceptance B_2 has changed by $\pm 3h_{22r}$ mhos from its value of $-2h_{22r}$ at the center of the pass band. The value of h_{22i} corresponds to approximately $1 \mu\text{uf}$ of capacitance and if the stray capacitance amounts to $3.5 \mu\text{uf}$, then the bandwidth is $\Delta B_2/2C$ since the slope of the susceptance of a tuned circuit is

$$2C \frac{\text{mhos}}{\text{rad/sec}}$$

Thus the bandwidth is approximately

$$\frac{6 \cdot 3.5 \cdot 10^{-5}}{2 \cdot 2.5 \cdot 10^{-12} \cdot 6.28}$$

or 3.7 mc. This is the actual value of load capacitance measured on an experimental amplifier with a vacuum tube voltmeter connected to the output. The measured response of this amplifier with a load of $Y_L = (68 - j215) \cdot 10^{-6}$ at 30 mc ($G_L = 2h_{22r}$) shows a peak gain of 18.3 db and half power points separated by 3.8 mc. For a given value of G_L , the bandwidth of the amplifier will vary inversely with the total capacitance in the output circuit. The same gain as obtained in the sample given above, can be obtained over a narrower band by increasing the load capacitance. Since the minimum capacitance is fixed, if one wishes to increase the width of the pass band, a higher value of G_2 must be used. In the same manner as is used to arrive at the data shown on Table II, Table III is computed for a value of $G_2 = 6h_{22r}$ ($G_L = 5h_{22r}$).

In this case, the maximum value of P_0/P_i occurs when $B_2 = -3h_{22r}$. The source impedance is selected to be $75 - j45$ ohms at 30 mc and the remainder of the table is computed. The maximum computed gain is approximately 16 db with half power frequencies where the susceptance

TABLE III — EVALUATION OF TRANSDUCER GAIN OF I.F.
AMPLIFIER

B_2	$-8h_{22r}$	$-5h_{22r}$	$-3h_{22r}$	$-2h_{22r}$	$-h_{22r}$	$+h_{22r}$	$+4h_{22r}$
P_0/P_i	25	33	43	40	39	31	21
Z_{in}	$35 + j22$	$35 + j35$	$50 + j45$	$56 + j46$	$65 + j48$	$77 + j39$	$83 + j20$
$P_{in}/\text{Power available from source}$	0.83	0.86	0.96	0.97	0.99	0.99	0.97
Transducer gain	21	28	41	39	39	31	20
Gain db.....	13.2	14.5	16.1	15.9	15.9	14.9	13.0

B_2 has changed by $\pm 6h_{22r}$ mhos from its value at the center of the band. Using the same value of circuit capacitance as above, the indicated bandwidth is about 7.4 mc. The measured response on an amplifier with this value of load impedance indicates a gain of 16.1 db at 30 mc, with the frequencies at the half power point separated by 7.6 mc.

Often it is desirable to build tuned amplifiers to work between like impedances in which case at least the output network must perform both the function of selectivity and impedance transformation. An example of a simple network to perform these functions is shown on Fig. 17. The impedance transforming properties of such a circuit are well known. With a given value of load resistance, the load admittance presented to the transistor can be made to have a given value at a certain frequency. However, since the circuit performs both the function of impedance transformation and selectivity the bandwidth is determined by the output impedance selected. This circuit does not present a fixed value of conductance as a function of frequency but for frequencies near the maximum gain it is a fair approximation to assume it constant. The output circuit of Fig. 17 was designed to present a load admittance such that $G_2 + jB_2 = 3h_{22r} - j2h_{22r}$ at 30 mc. This is the same condition as computed in Table II so one would expect the same value of maximum gain. However, in order to present the proper value of load impedance, a total load capacitance of about $4 \mu\text{f}$ must be used. This indicates a bandwidth of 3.3 mc between the half power points. The measured response of this amplifier is shown on Fig. 18 as the solid line. The points indicate the computed maximum gain and the frequencies at which the gain is down 3 db.

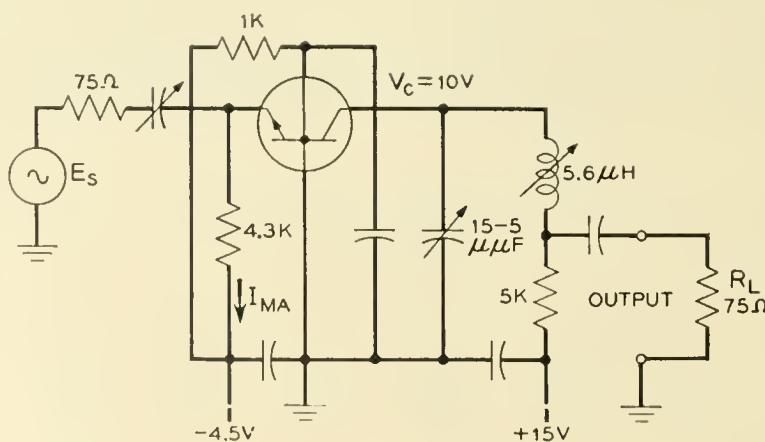


Fig. 17 — Simple tuned amplifier. The output circuit performs both the functions of impedance transformation and selectivity.

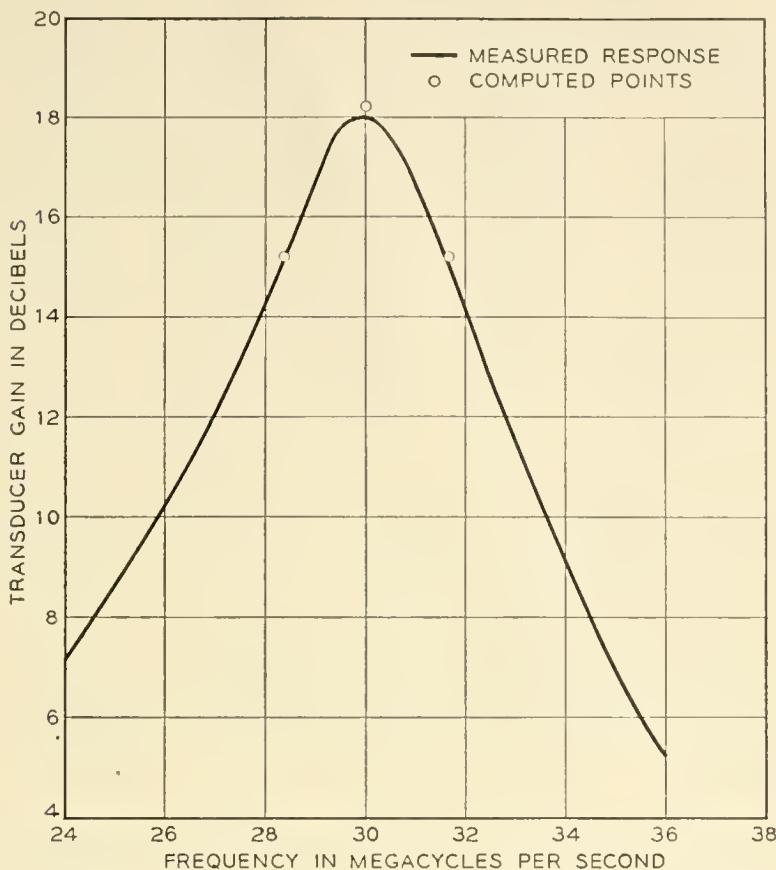


Fig. 18 — Measured and computed response of the stage shown on Fig. 17.

An IF Amplifier Centered at 70 Mc.

Although we do not have complete data on the parameter values of tetrode transistors in this frequency range, amplifiers with a center frequency of 70 mc have been built and their performance measured. The amplifier was designed to provide a flat gain characteristic over the frequency range from 60 to 80 mc. The stage was designed with the equivalent of a double tuned transformer, interstage circuit with the transformer being replaced by the equivalent tee section. The selective circuit is terminated at its output into the load resistance in the case of the last stage or by the input impedance of the following transistor when it is used as an interstage network. The impedance transformation of the network is approximately 75 ohms to 1,500 ohms so it is essentially unterminated at the collector. By using a sweeping oscillator, such a stage can be adjusted to result in a fairly flat frequency response. A typical stage is shown on Fig. 19. The output terminals are connected to either the load or the next emitter. The response obtained from a 3-stage amplifier is shown on Fig. 20. In order to determine the variation of gain

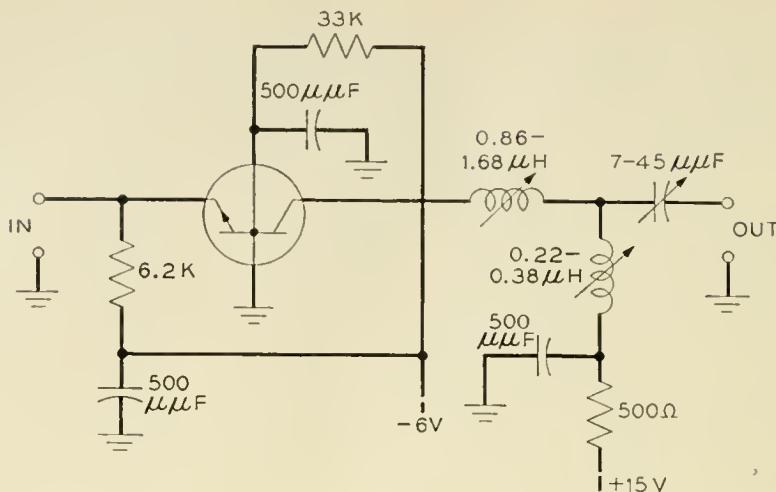


Fig. 19 — Circuit of a 60 to 80-mc band pass amplifier stage.

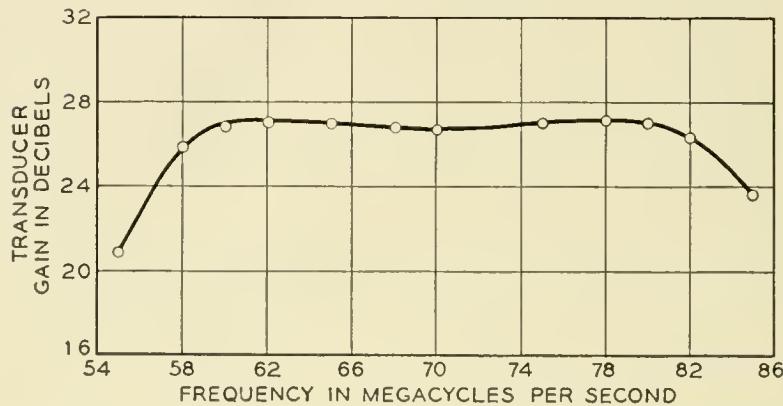


Fig. 20 — Gain of a 3-stage band pass amplifier working between 75-ohm impedances. Each stage uses the circuit shown on Fig. 19.

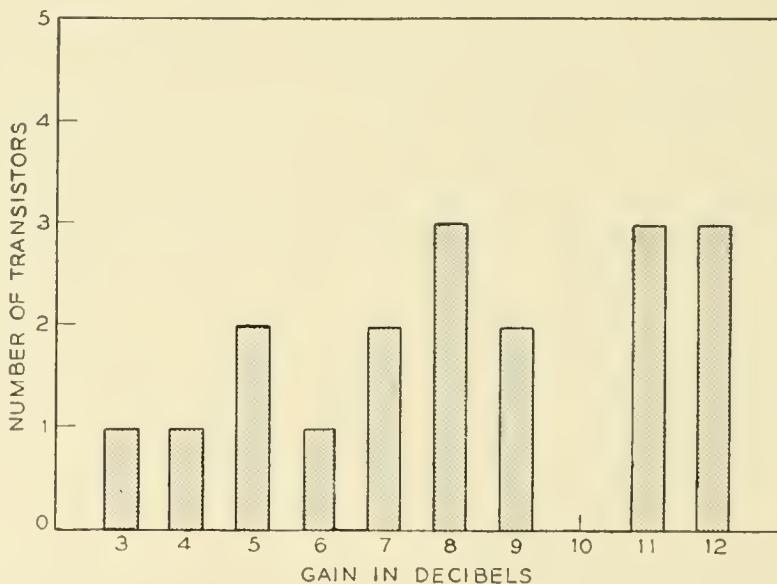


Fig. 21 — Variation of gain for a group of transistors used in the circuit of Fig. 19.

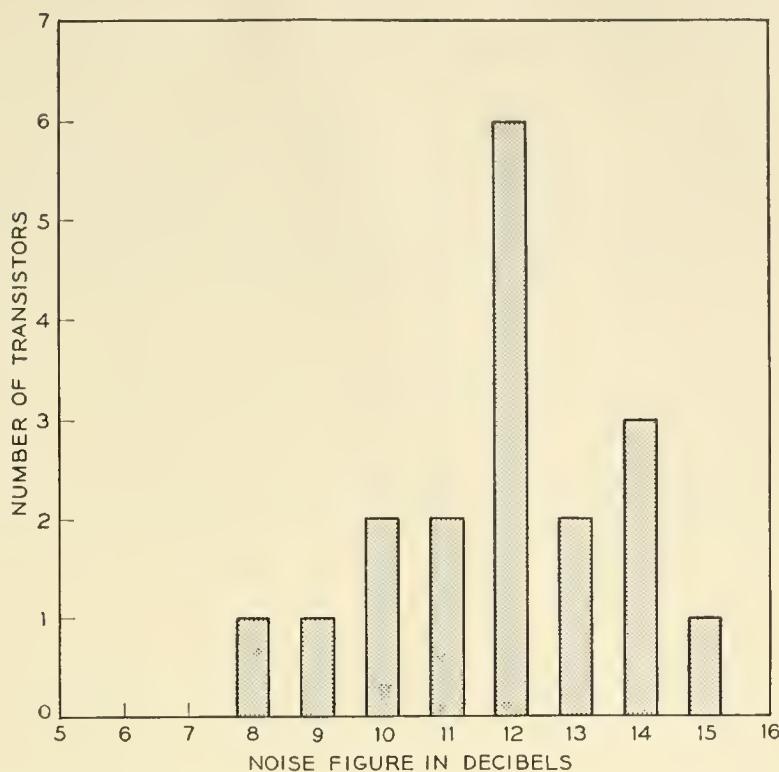


FIG. 22—Noise figure for a group of transistors used in the circuit of Fig. 19.

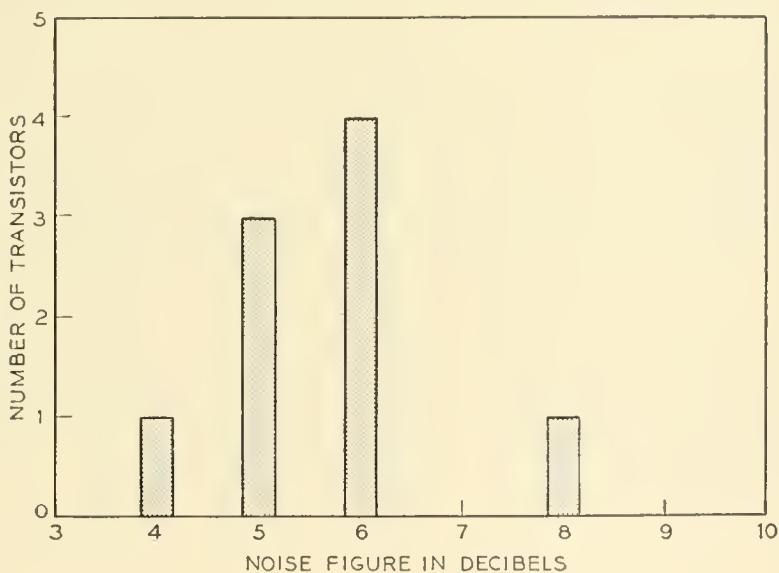


Fig. 23—Noise figure for a group of transistors used in a 10-mc bandpass amplifier.

between various transistors, 18 tetrodes were measured in the first stage of the amplifier. If the measured gain of each transistor is rounded off to the nearest db and the number of transistors having this gain plotted as the abscissa, the results shown on Fig. 21 are obtained. Of the 18 transistors measured, 11 have a gain of 8 db or greater. Similar data has been obtained on the noise figure of the same 18 transistors, the results being shown on Fig. 22. In general, the transistors having the highest gain also have the lowest noise figure. The noise figure depends to some extent on the source impedance but a 75-ohm source results in a noise figure which is within a few tenths of a db of the minimum. The value of the noise figure does not vary a great deal as the collector voltage and emitter current are changed except that if the collector voltage is lowered below 6 or 8 volts the gain decreases and in general the noise figure increases.

Noise Figure at 10 Mc.

Although not described here, bandpass amplifiers centered at 10 mc with a 200-ke pass band have been constructed using tetrode transistors. A gain of slightly over 20 db per stage can be realized at this frequency. The noise figure of transistors tried in this circuit is shown on Fig. 23, the data being shown in the same manner as described above. At 10 mc the noise figures are lower than at 70 mc. The remarks made above concerning variation of noise figure with operating conditions also apply to this case.

ACKNOWLEDGMENTS

We are happy to acknowledge the advice and encouragement given us by R. L. Wallace, Jr., and others in the Laboratories. We also wish to express our thanks to E. Dickten who fabricated the transistors used to obtain the experimental data presented. W. F. Wolfertz made the transistor parameter measurements used in the computations. R. H. Bosworth and C. E. Scheideler were responsible for construction of the circuits and some of the gain measurements. We also wish to thank W. R. Bennett for his aid in preparing the manuscript.

The Nature of Power Saturation in Traveling Wave Tubes

By C. C. CUTLER

(Manuscript received February 2, 1956)

The non-linear operating characteristics of a traveling wave tube have been studied using a tube scaled to low frequency and large size. Measurements of electron beam velocity and current as a function of RF phase and amplitude show the mechanism of power saturation.

The most important conclusions are:

I. There is an optimum set of parameters ($QC = 0.2$ and $\gamma r_0 = 0.5$) giving the greatest efficiency.

II. There is a best value of the gain parameter "C" which leads to a best efficiency of about 38 per cent.

III. A picture of the actual spent beam modulation is now available which shows the factors contributing to traveling wave tube power saturation.

INTRODUCTION

The highest possible efficiency of the traveling wave tube has been estimated from many different points of view. In his first paper on the subject¹ J. R. Pierce showed that according to small signal theory, when the dc beam current reaches 100 per cent modulation an efficiency of

$$\eta = \frac{C}{2} \quad (1)$$

is indicated,* and thus the actual efficiency might be limited to something like this value. Upon later consideration² he concluded that the ac convection current could be twice the dc current and that one might expect an efficiency of

$$\eta = 2C \quad (2)$$

He also considered the effects of space charge, and concluded on the

* Symbols are consistent with Reference 2 and are listed at the end of this paper.

same basis that under high space charge and elevated voltage conditions, efficiencies might be as high as

$$\eta = 8C \quad (3)$$

J. C. Slater³ on the other hand considered the motion of electrons in a traveling wave and concluded that the maximum possible reduction in beam velocity would also indicate a limiting efficiency of $2C$. Taking a more realistic account of the electron velocity, Pierce² showed that these considerations lead to a value of

$$\eta = -4y_1C \quad (4)$$

which, since y_1 ranges between $-\frac{1}{2}$ and -2 , leads to the same range of values as the other predictions.

None of these papers purport to give a physical picture of the overloading phenomenon, but only specify clear limitations to the linear theory. L. Brillouin⁴ on the other hand found a stable solution for the flow of electrons bunched in the troughs of a traveling wave. This he supposed to represent the limiting high level condition of traveling wave tube operation. His results give an efficiency of

$$\eta = 2bC \quad (5)$$

In the first numerical computations of the actual electron motion in a traveling wave tube in the nonlinear region of operation, Nordsieck⁵ predicted efficiencies ranging between 2.5 and 7 times C and showed that there would be a considerable reduction in efficiency for large diameter beams, due to the non-uniformity of circuit field across the beam diameter. He also gave some indication of the electron dynamics involved. Improving on this line of attack, Poulter⁶ calculated some cases including the effect of space charge and large values of C .

Tien, Walker and Wolontis⁷ carried computations still further for small values of C by including the effect of small beam radii upon the space charge terms, and showed that space charge and finite (small) beam radii result in much smaller efficiencies than were previously predicted. J. E. Rowe⁸ got similar results and gave more information on the effects of finite values of C . Computations for large values of C by Tien⁹ showed that a serious departure from the small C conditions takes place above values of $C = 0.1$ if space charge is small (i.e., below $QC = 0.1$) and above $C = 0.05$ for larger values of space charge. They indicated that a maximum value of efficiency as high as 40 per cent should be possible using $C = 0.15$, $QC = 0.1$ and elevated beam voltages.

These five papers give some insight into the electron dynamics of power

saturation, but still involve questionable approximations which make it desirable to compare predictions with the actual situation.

Theoretical considerations of the effects of attenuation upon efficiency have not led to conclusions coming even close to the observed results. Measured characteristics^{10, 11} show that the effect of attenuation is very large, but that attenuation may be appropriately distributed to attain stability and isolation between input and output of the tube without degrading the output power.

There are also several papers in the French and German periodicals which deal with the question of traveling wave tube efficiency. Some of these are listed in References 12 through 20.

This paper describes measurements of efficiency and of beam modulation made on a traveling wave tube scaled to large size,* and low frequencies. The construction of the tube, shown in Fig. 1, and the measurement of its parameters were much more accurate than is usual in the design of such tubes. The results are believed to be generally applicable to tubes having similar values of the normalized parameters.

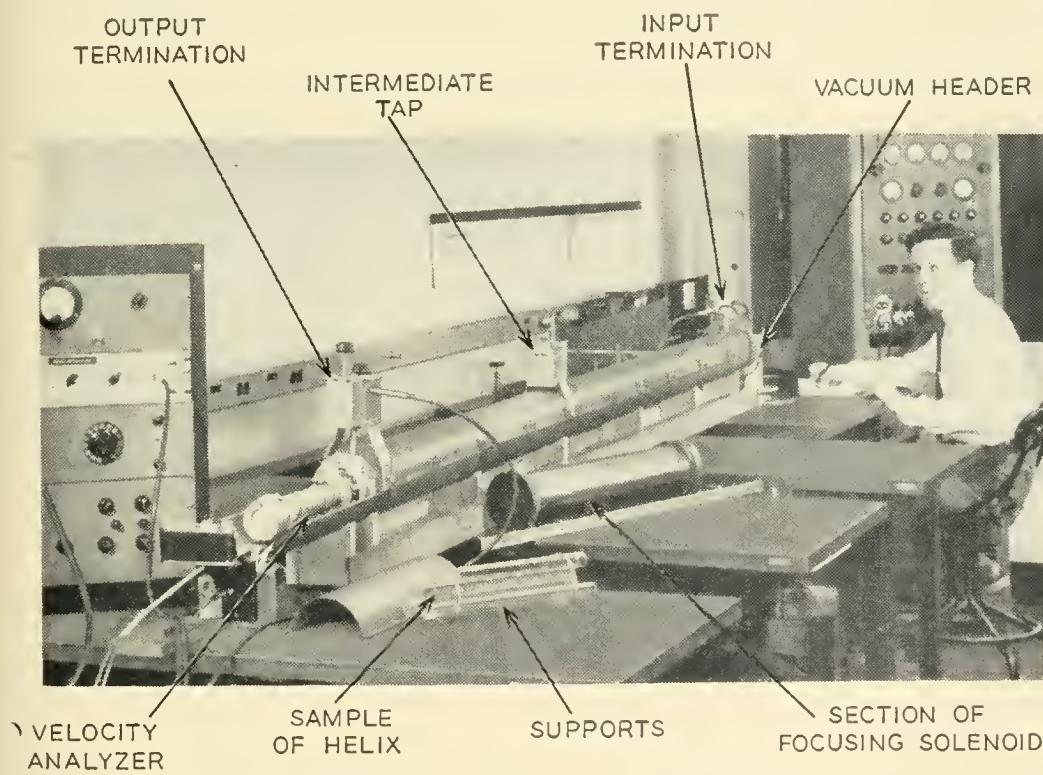


Fig. 1 — The scale model traveling wave tube. The tube is 10 feet long with a copper helix supported by notched glass tubing from an aluminum cylinder overwound with a focusing solenoid. It is continuously pumped and readily demountable.

* See Appendix.

Two kinds of measurements are described. First, the efficiency and power output are determined for various conditions of operation, and second the spent beam ac velocity and current are measured. The principal results are shown in Figs. 2 to 4 which give the obtainable efficiencies, and in Figs. 7 to 10 which show some of the factors which contribute to power saturation. These figures are discussed in detail later. The most significant phenomenon is the early formation of an out-of-phase bunch of electrons which have been violently thrown back from the initial bunch, absorbing energy from the circuit wave, and inhibiting its growth. The final velocity of most of the electrons is near to that of the circuit wave which would lead to a value of

$$\text{limiting efficiency } \eta = -2y_1 C \quad (6)$$

if the wave velocity maintained its small signal value. Actually the wave slows down, under the most favorable conditions giving rise to a somewhat higher efficiency. For other conditions, space charge, excess electron velocity, or nonuniformity of the circuit field enter in various ways to prevent the desired grouping of electrons and result in lower efficiencies.

The observed efficiencies are a rather complicated function of QC , γr_0 and C . To compare with efficiencies obtained from practical tubes one must account for circuit attenuation and be sure that some uncontrolled factor such as helix non-uniformity and secondary emission is not seriously affecting the tubes' performance. Measured efficiencies of several carefully designed tubes have been assembled and are compared with the results of this paper in Table I.

The results of these measurements compare favorably with the computations of Tien, Walker and Wolontis⁷, and of Tien⁹. There are, however some important differences which are discussed in a later section.

TRAVELING WAVE TUBE EFFICIENCY MEASUREMENTS

Reasoning from low level theory, efficiency should be a function of the gain parameter, "C," the space charge parameter "QC," the circuit attenuation, and (for large beam sizes), the relative beam radius " γr_0 ." It was soon found that efficiency is a much more complicated function of γr_0 than expected. The initial objective was to determine the effect of QC , C , and γr_0 separately on efficiency, but it was necessary to give a much more general coverage of these parameters, not assuming any of them to be small.

Most of the measurements have been made with small values of loss

TABLE I

Laboratory	Freq. mc.	QC	γr_0	C	η meas- ured	η (from Fig. 3)	η (From Fig. 3 with allowance for circuit attenuation ¹⁰)
McDowell*	4,000	0.27	0.62	0.078	19.5	26	21.6
	6,000	0.29	0.8	0.058	13.2	16.2	12.5
Brangaccio and Cutler†	4,000	0.61	0.87	0.041	11	6	6
Danielson and Watson*	11,000	0.35	1.2	0.05	6.6	7	4.8
R. R. Warnecke ^{16, 17, 18}	870	0.32	0.3	.125	27	33	33
W. Kleen and W. Friz ¹⁵	4,000	0.5	0.43	0.05	7.8	11.5	5.7
W. Kleen‡	4,000	0.2	0.94	0.1	20	26	22
L. Brück§	3,500	0.19	0.6	0.065	15	23	18.5
Hughes Aircraft Co.	3,240	0.19	0.94	0.12	39	31	29
	9,000	0.15	1.3	0.11	25	15.5	12.7

* At Bell Telephone Laboratories.

† Reference 10 (a slight beam misalignment could account for most of this difference).

‡ Siemens & Halske, Munich, Germany.

§ Telefunken, Ulm, Germany.

and of the gain parameter, where efficiency is proportional to C , as expected from small-signal small- C predictions. This reduces the problem to a determination of η/C versus QC and γr_0 .

Many measurements of this kind have been made, and the data are summarized in Figs. 2 and 3, with efficiency shown as a function of QC and γr_0 . In Fig. 2 we have the efficiency when the beam voltage is that which gives maximum low-level gain. Fig. 3 shows the efficiency obtained when the beam potential is raised to optimize the power output, and contours of constant efficiency have been sketched in. There is significantly higher efficiency than before in the region of maximum efficiency, but not much more elsewhere.

Fig. 4 shows how efficiency varies with C for a small value of QC , a representative value of γr_0 , and with beam voltage increased to maximize the output. This indicates a maximum of about 38 per cent at $C = 0.14$.

Some of the computed results of Tien, Walker and Wolontis,⁷ and of Tien⁹ are also indicated in the figures. Their results generally indicate somewhat greater efficiencies than were observed, but in the most significant region the comparison is not too bad as will be seen in a later section.

The measurements are for conditions having negligible circuit loss near the tube output. There are no new data on the effect of loss, but earlier results¹⁰ have been verified by measurements at Stanford University¹¹ and are still believed to be a satisfactory guide in tube design.

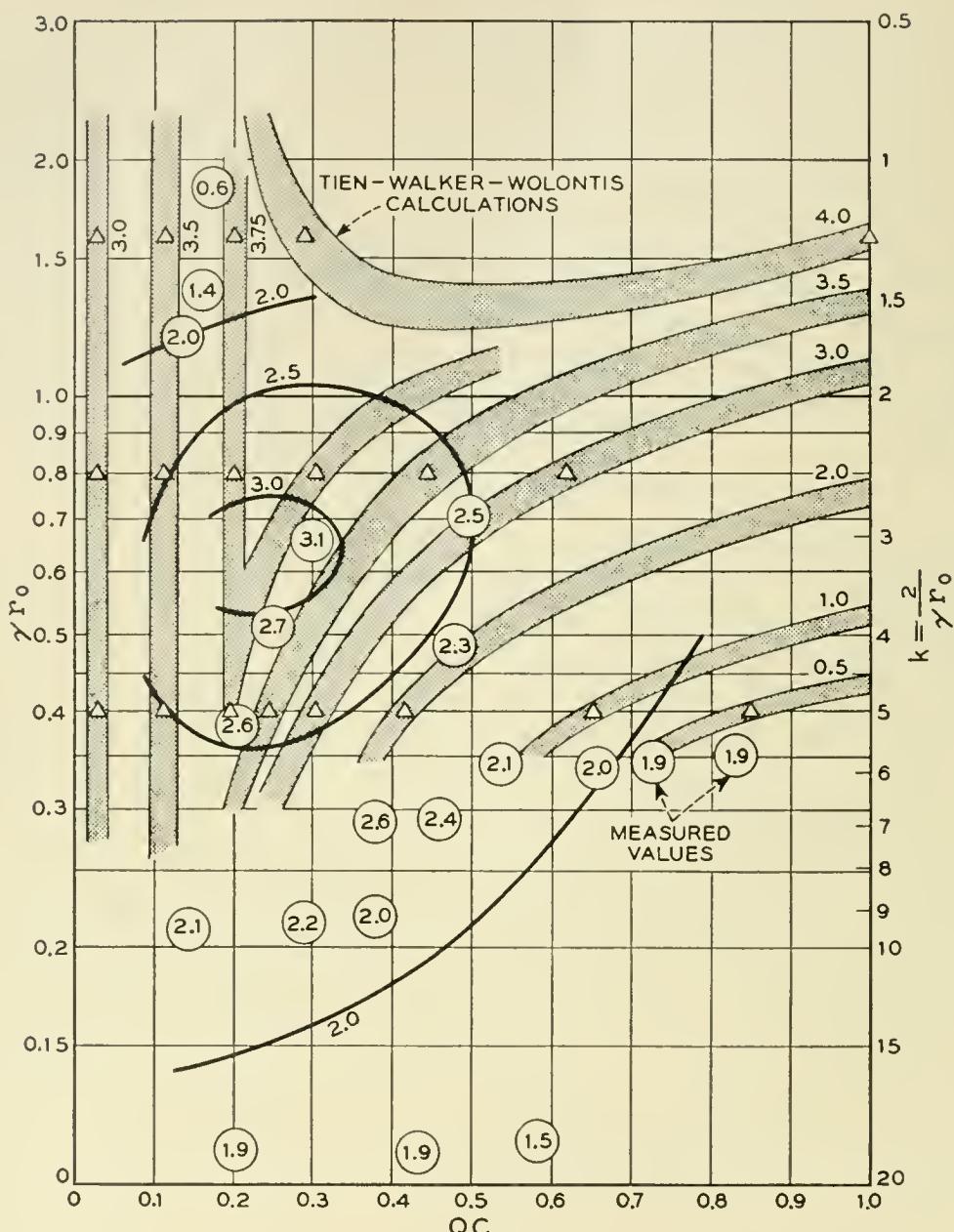


Fig. 2 — Values of efficiency/ C as a function of QC and γr_0 at the voltage giving maximum gain per unit length. The shaded contours and triangular points are from the computations⁷ of Tien, Walker and Wolontis. The circled points are from the measurements and the line contours are estimated lines of constant efficiency. The most significant difference is for large beam radii, where the RF field varies over the beam radius in a way not accounted for in the computations.

SPENT BEAM CHARACTERISTICS

The scale model traveling wave tube was followed by a velocity analyzer as sketched in Fig. 5 and described in the Appendix. A sample of the beam at the output end of the helix is passed through a sweep circuit to separate electrons according to phase, and crossed electric and

magnetic fields to sort them according to velocity. The resulting beam draws a pattern on a fluorescent screen as shown in Fig. 6 from which charge density and velocity can be measured as a function of signal phase. The velocity coordinate is determined by photographing the ellipse with several different beam potentials, as in Fig. 6(a), and the phase coordinate is measured along the ellipse. From pictures like this a complete determination of electron behavior is obtained from the linear region up to and above the saturation level.

The results of such a run are plotted in Fig. 7. The upper lefthand

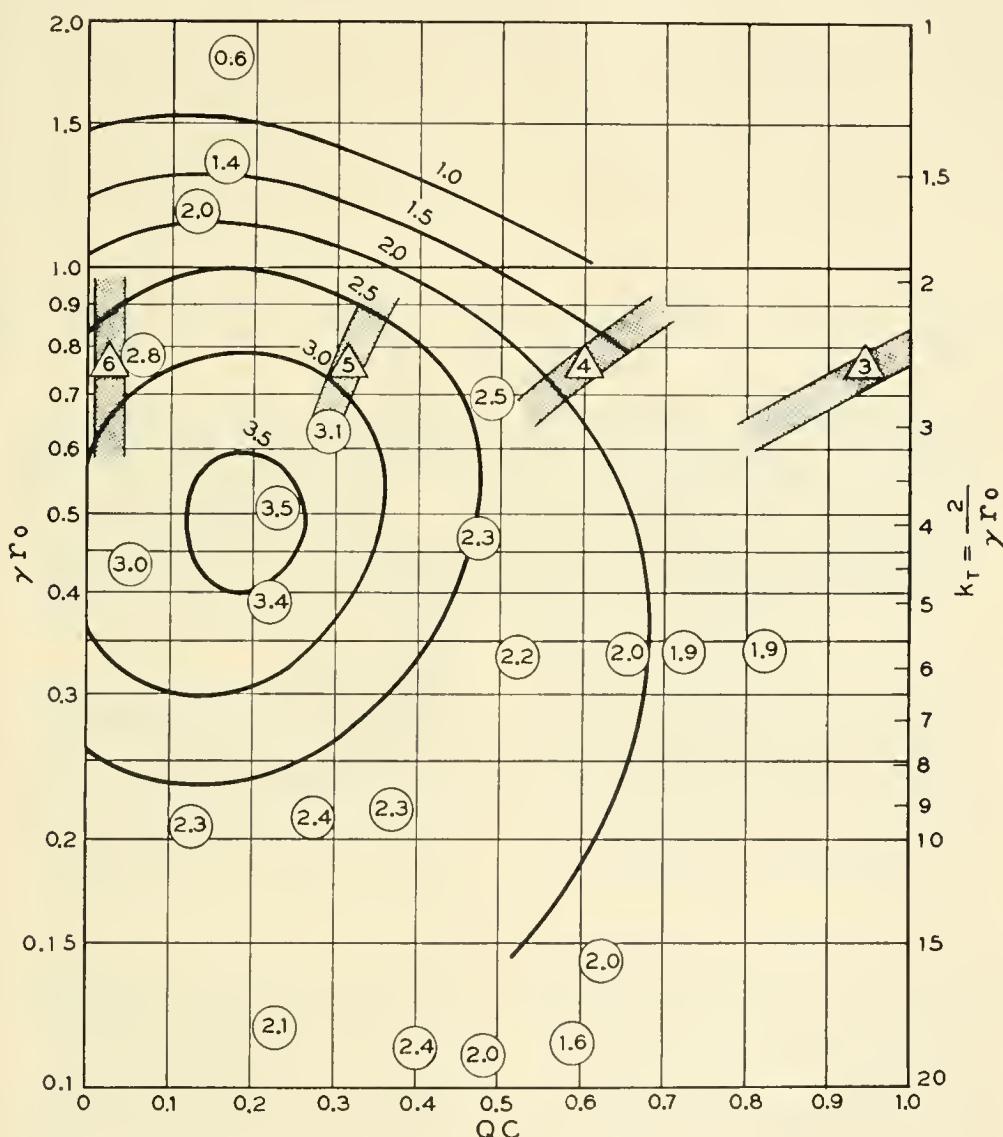


Fig. 3 — Values of efficiency/ C as a function of QC and γr_0 at elevated beam voltage. Raising the beam voltage has little effect at large QC and small γr_0 , and less than expected anywhere. Again the triangular points are from Tien, Walker and Wolontis,⁷ and the line contours are estimated from the measured data.

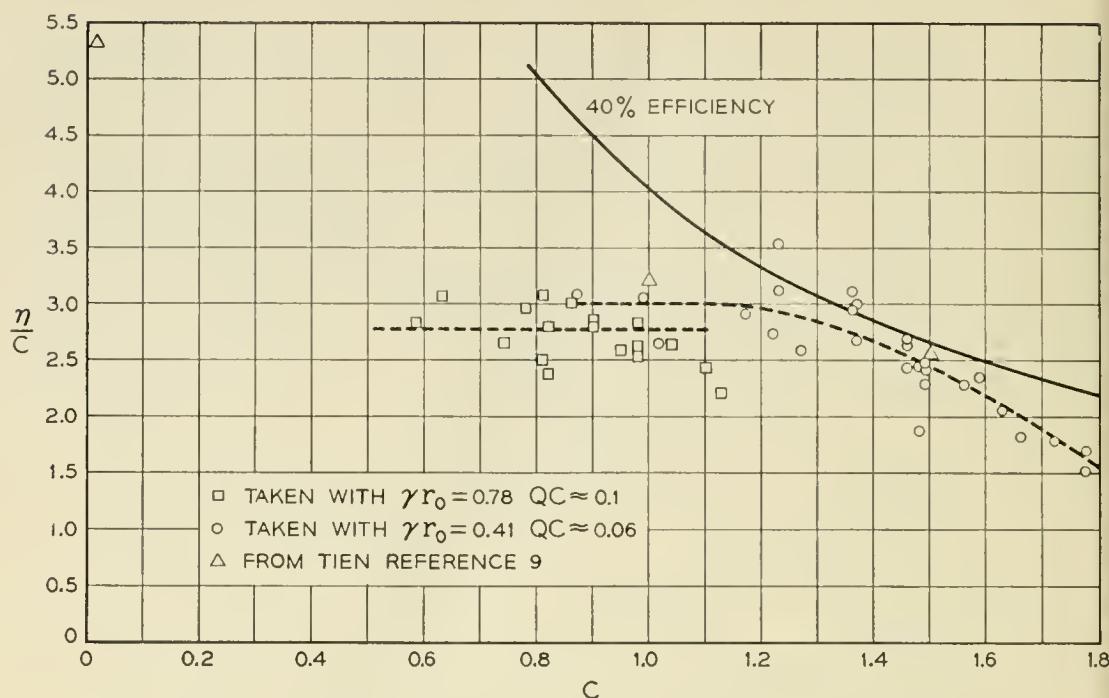


Fig. 4 — Efficiency/ C for large values of C and with elevated beam voltage. Efficiency seriously departs from proportionality to C at $C = 0.14$, where a maximum efficiency of about 38 per cent is measured.

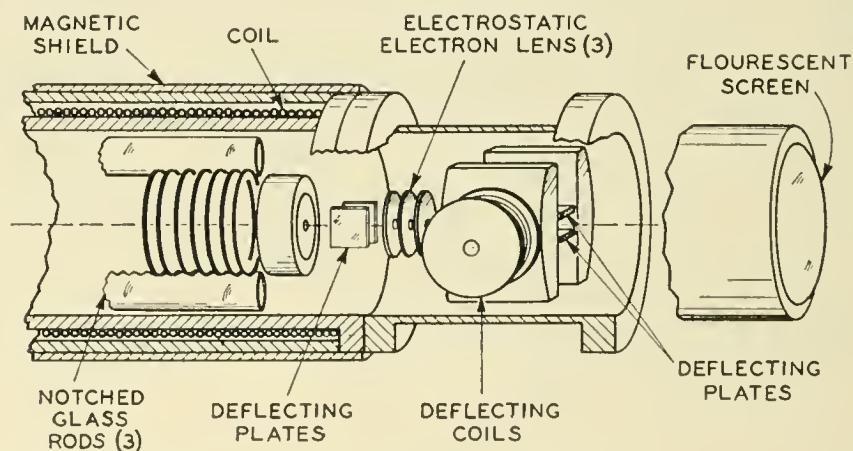


Fig. 5 — The velocity analyzer. A sample of the spent electron beam is accelerated to a high potential, swept transversely with a synchronous voltage, sorted with crossed electric and magnetic fields, and focused onto a fluorescent screen.

pattern, Fig. 7(a), is representative of the low level (linear) conditions (22 db below the drive for saturation output). The dashed curve represents the voltage on the circuit, inverted so that electrons can be visualized as rolling down hill on the curve. The phase of this voltage relative to the electron ac velocity is computed from small signal theory, but

everything else in Fig. 7, including subsequent variations of phase, are measured. The solid line patterns represent the ac velocity, and the shaded area, the charge density corresponding to that velocity. Thus in each pattern we have a complete story of (fundamental) circuit voltage, electron velocity and current density as a function of phase, for a particular signal input level. The velocity and current modulations at small signal levels check calculated values well, and it is not difficult to visualize the dynamics giving this pattern.

Consider first the situation in the tube at small signal amplitudes. At the input an unmodulated electron beam enters the field of an electromagnetic wave moving with approximately the same velocity as the electrons. The electrons are accelerated or decelerated depending upon their phase relative to the wave, and soon are modulated in velocity. The velocity modulation causes a bunching of the electrons near the potential maxima (i.e., the valleys in the inverted potential wave shown) and these bunches in turn induce a new electromagnetic wave component onto the circuit roughly in quadrature following the initial wave. The addition of this component gives a net field somewhat retarded from the initial wave and larger in amplitude. Continuation of this process

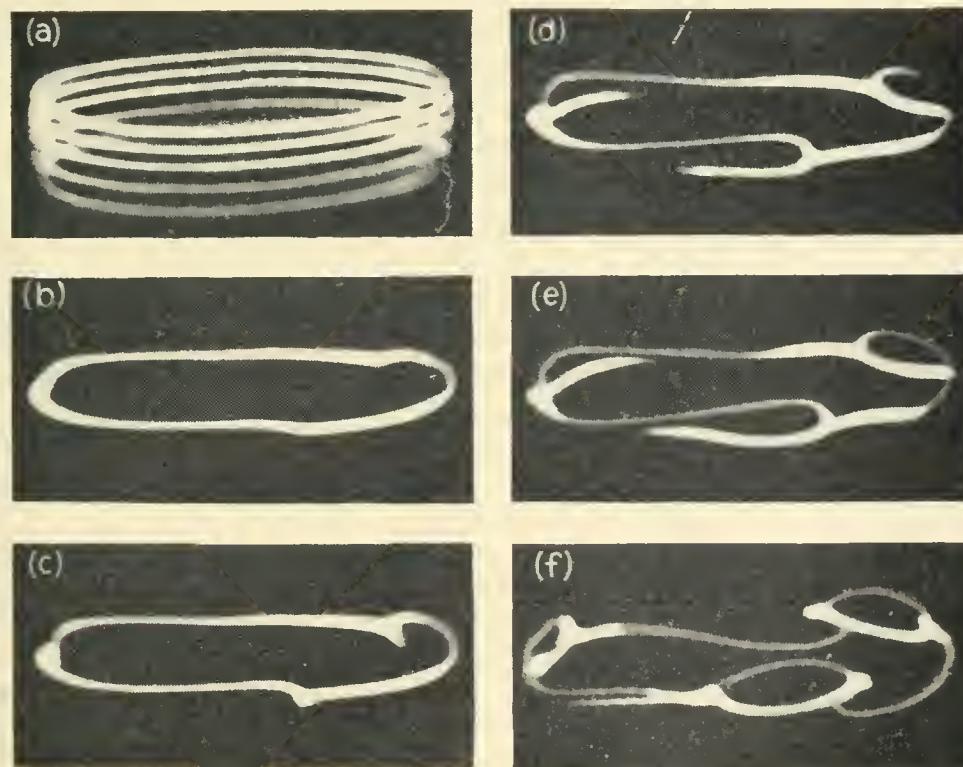


Fig. 6 — Velocity analyzer patterns. The beam sample is made to traverse an ellipse at $\frac{1}{3}$ the signal frequency. Current density modulation appears as intensity variation, and velocity variation as vertical deflection from the ellipse.

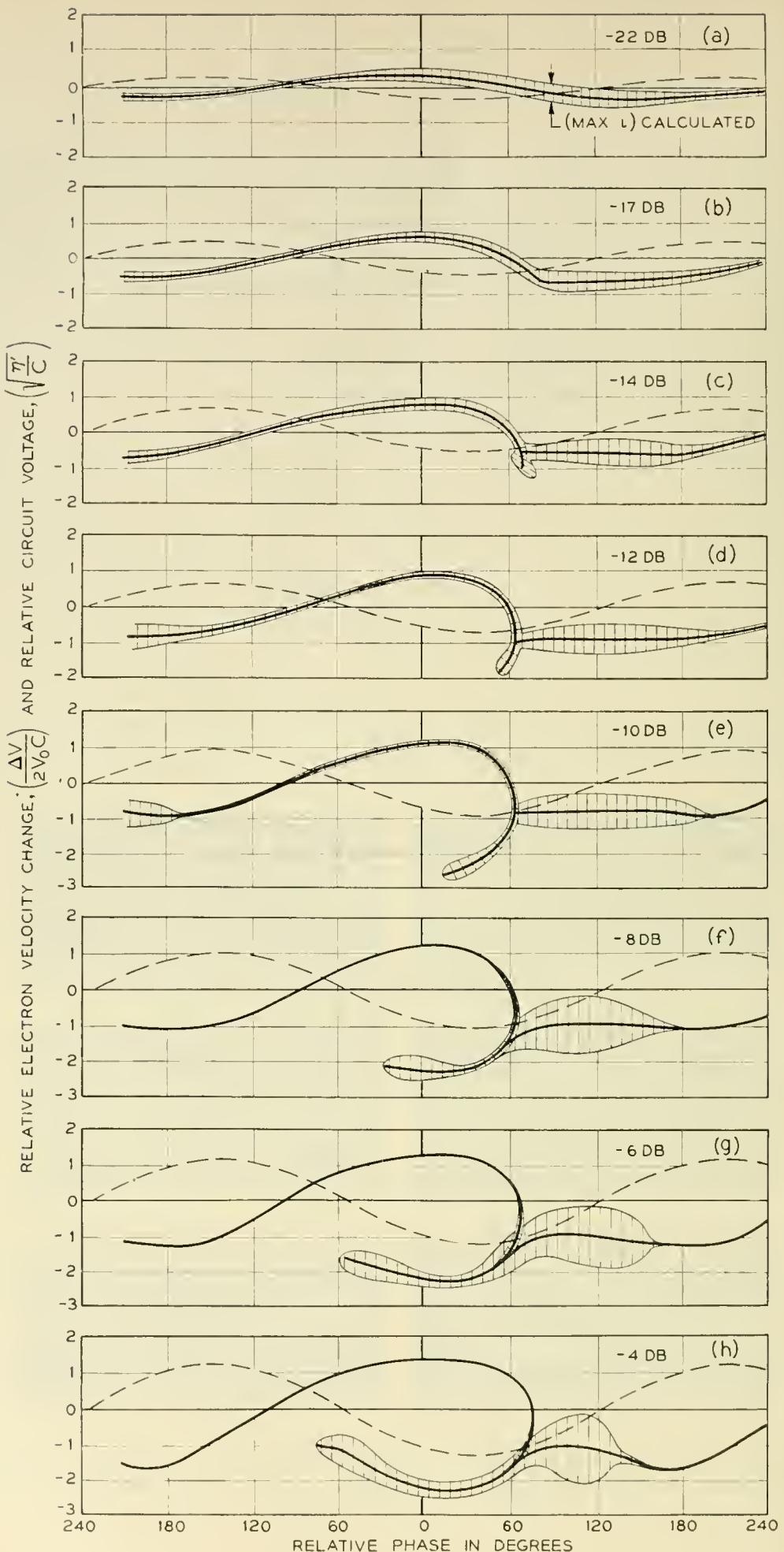
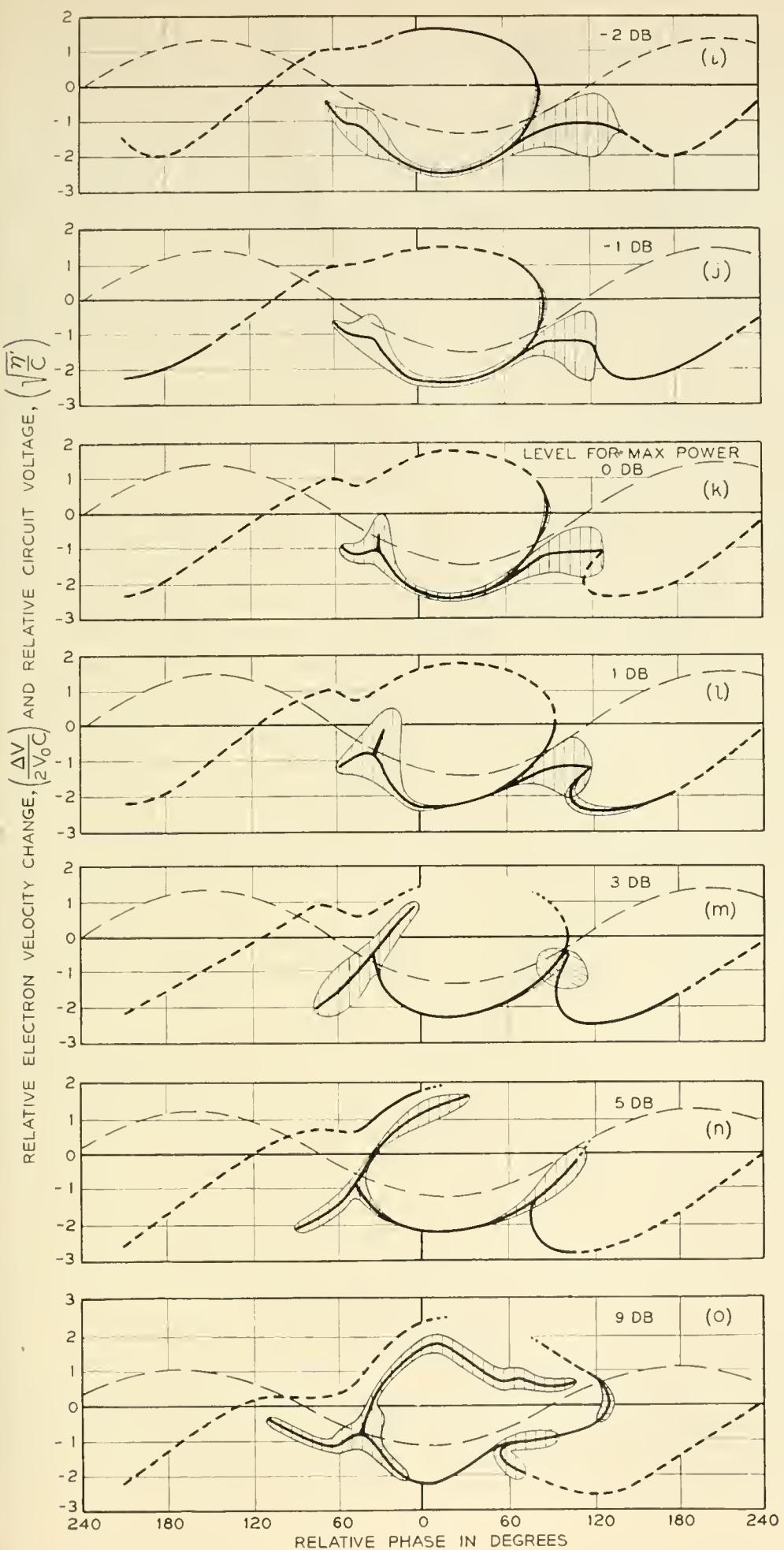


Fig. 7 — Curves of current and velocity as a function of phase for various input levels. The velocity becomes multivalued at a very low level, a tail forming a nucleus for a second electron bunch which eventually caused saturation in the output. For this run $C = 0.1$, $Q/C = 0.06$, $\gamma r_0 = 0.4$ and $b = 0.26$.



may be seen to give a resultant increasing wave traveling somewhat slower than the initial wave, and thus slower than the electron velocity. Returning to Fig. 7 we see that electrons in the decelerating field [from +30 to +210° in Fig. 7(a)] have been slowed down, and because of their initial velocity being faster than the wave velocity, have moved forward in the wave giving a region of minimum velocity somewhat in advance of the point of maximum retarding field (greatest negative slope in the wave potential). Also, bunching due to acceleration and deceleration of electrons has produced a maximum of electron current density which, because of the initial excess electron velocity, is somewhat to the right of the potential maximum (downward).

As the level is increased the modulation increases and at 17 db below saturation drive, Fig. 7(b), some nonlinearity is evident. The velocity and current are no longer sinusoidal, but show the beginnings of a cusp in the velocity curve and a definite non-sinusoidal bunching of the electrons in the retarding field region (between +30 and 210°).

In the next pattern, Fig. 7(c), at 14 db below saturation a definite cusp has formed with a very sharp concentration of electrons extending significantly below the velocities of the other electrons. We already have a wide range of velocities in the vicinity of the cusp, and at this level the single valued velocity picture of the traveling wave tube breaks down. Although it cannot be distinctly resolved, the study of many such pictures leaves little doubt that the cusp and its later development is really a folding of the velocity line.

The next pattern at 12 db below saturation drive, Fig. 7(d), shows a greater development of the spur and a somewhat greater consolidation of current in the main bunch between +60° and +180°. It is interesting that the velocity in this region has not changed significantly. In order for this to be true the space charge field must just compensate for the circuit field. In the vicinity of the 60° point the space charge field obviously must reverse, accounting for the very sharp deceleration evident in the very rapid development of the low velocity spur. The decelerating field must be far from that of the wave, inasmuch as the electrons just behind the cusp are much more sharply decelerated than those preceding the cusp. We conclude that there are very sharply defined space charge fields much stronger than the helix field. At this relatively low drive, the velocity spread has already achieved its maximum peak value.

The succeeding three patterns show a continuing growth of the spur; a continued bleeding of electrons from the higher velocity regions, and a consolidation of the main bunch just in advance of the spur. Presumably the increased concentration of space charge in the bunch has kept

pace with the increasing helix field, so that the net decelerating field still balances to nearly zero. At 4 db below the saturation drive, Fig. 7(h), the spur has moved well into the accelerating region, and has been speeded up. The main bunch of electrons is still to the right of the spur, and has been consolidated into a 60° interval. The few electrons in advance of this region evidently no longer find the space charge field sufficient to balance the circuit field, and are being decelerated into a second low velocity loop.

The next three patterns show a continued growth of this second low velocity loop, further consolidation of the 'main bunch', and the rapid formation of a second bunch in the accelerating field at the end of the spur. It is interesting that at saturation drive, Fig. 7(k) the two bunches are very nearly equal, and in equal and opposite circuit fields, nearly 180° apart. The reason for the saturation is that while the main bunch is still giving up energy to the wave, the new one is absorbing energy at an equal rate. The fundamental component of electron current is evidently small, and is in quadrature with the circuit field. The current density in the dashed regions is less than 1 per cent of that in the bunches, and probably more than 95 per cent of the electrons are in the two bunches. Two new effects are observable at this level. The second electron bunch has begun to come apart, presumably because of strong localized space charge forces. These forces are also evident in the kink in the velocity pattern drawn by the fast electrons at the same phase as the second bunch.

Since the majority of the current is in the two bunches at a reduced velocity of

$$\frac{\Delta V}{2V_0 C} = -1.1$$

one would expect an output efficiency of

$$\frac{\Delta V}{V_0} = 2.2C$$

The actual measured efficiency

$$\frac{\text{RF power output}}{\text{DC power input}}$$

was 2.0 C. Under the conditions described, (6) would give 1.4 C.

At still higher drive levels the pattern continues to develop, electrons from the first bunch falling back into the second, which in turn continues to divide, one part accelerating ahead into a new spur, and the other

slowing down and falling further back in phase. At 9 db above saturation, Fig. 7(o), the pattern is quite complex, and at still higher levels it is utterly indescribable.

It is interesting that the velocity gives a line pattern, even though a multivalued one. It is reasonable to suppose that the development of the spur is really a folding of the velocity line so that the spur is really a double line. Thus, at the 9 db level, and at 0° phase, for instance, there must be electrons originating from five different parts of the initial distribution. In an attempt to verify this the resolution of the velocity analyzer was adjusted so that a difference in velocity of 2 per cent of the overall spread could be observed, but there was no positive indication of more than one velocity associated with any line shown.

There has been a long-standing debate as to whether or not electrons are trapped in the circuit field, or continue to override the wave at large amplitudes. The observations indicate that with low values of space charge and near synchronous voltage the electrons are effectively trapped in the wave until well above saturation amplitude. In other circumstances this is not the case, as we shall see.

SPACE CHARGE EFFECTS

The data of Fig. 7 were taken with a very small value of the space charge parameter QC , so small in fact as to be almost negligible as far as low level operation is concerned. Yet the space charge forces evidently played a very strong role in the development of the velocity and current patterns. It is doubtful that space charge would ever be negligible in this respect, because if the space charge parameter were smaller, the bunching would be more complete, the electron density in the bunch would be greater limited only by the balance of space charge field and circuit field in the bunch. The effect of decreasing QC further therefore is a greater localization of the space charge forces, rather than a reduction of their magnitude, at least until the bunch becomes short compared to the beam radius.

Increasing the value of the space charge parameter has quite the opposite effect. In Fig. 8 are shown three velocity-current distributions at the saturation level, for different values of QC . It can be seen that a result of increased space charge is a greater spread of velocities, and a wider phase distribution of current.

With the introduction of space charge, the velocity difference between the electrons and the circuit wave at low levels is increased. Consequently electrons spend a longer time in the decelerating field before being thrown back in the low velocity spur, and thus lose more energy. The

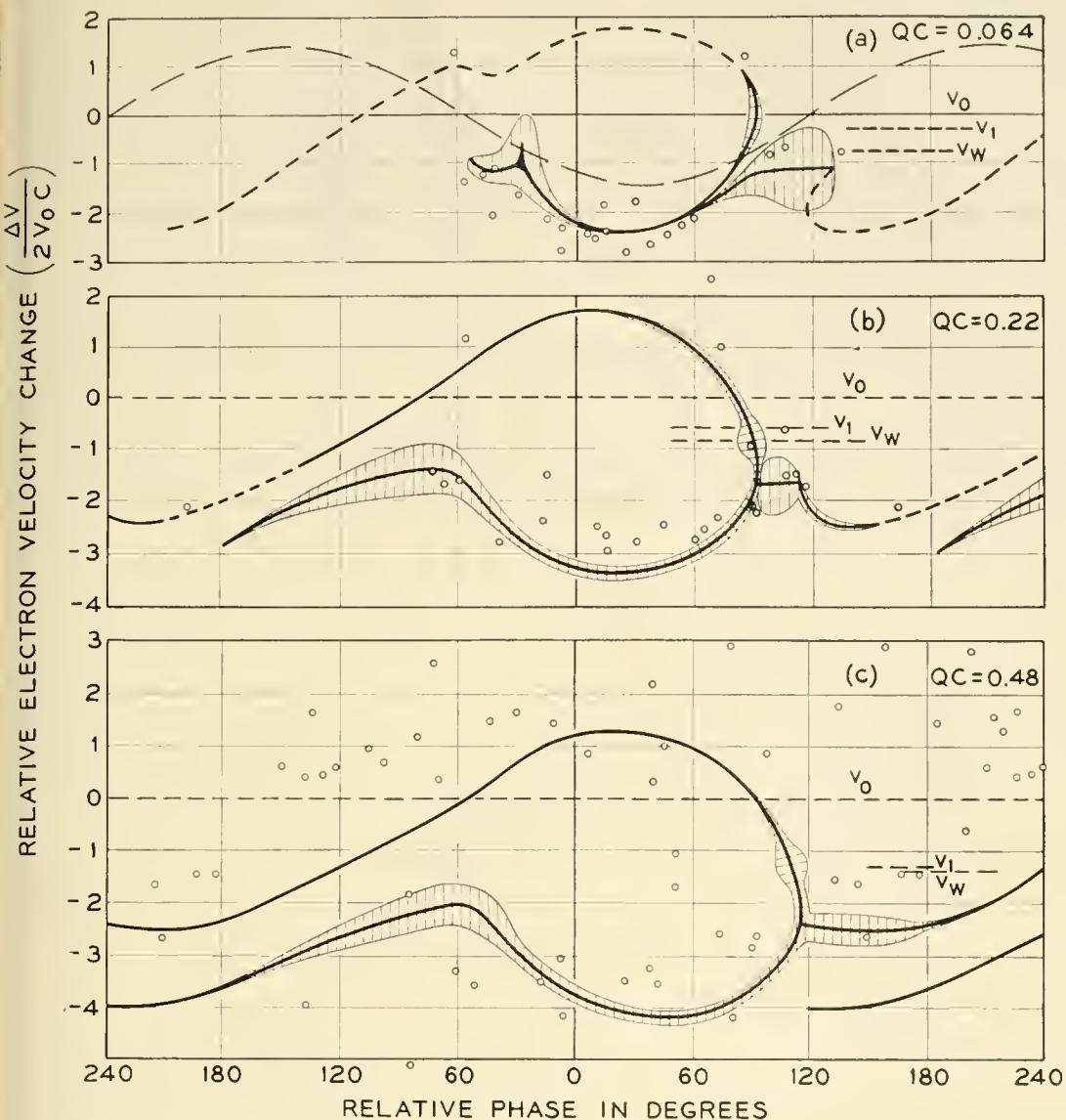


Fig. 8 — A comparison showing the effect of the space charge parameter QC on the velocity and current at overload. The points represent the disc electrons of the computations⁷ of Tien, Walker and Wolontis. For this run $\gamma r_0 = 0.4$ and b is chosen for maximum x_1 .

greater reduction of velocity results in a faster and farther retarding of the current in the spur before the retarded electrons recover velocity in the accelerating region. Also the larger space charge forces prevent as tight bunching of the electrons anywhere, so that at overload they are spread over a much wider phase interval (about 360° for $QC = 0.5$). Space charge also prevents electrons from the forward part of the bunch from being trapped so that more electrons escape ahead of the decelerating field and more current is found in the upper half of the velocity curve. This very likely is the reason that efficiency decreases when QC is increased above about 0.3.

EFFECT OF BEAM SIZE

In small signal operation, decreasing the beam radius below that which assures a constant circuit field throughout the beam has no effect except that accounted for by its effect on QC . Fig. 9 shows that for large signals, however, it has a pronounced effect. When the beam is made smaller (with QC maintained by changing frequency and beam current), the slowed up tail is formed at a much lower signal level (not shown), by a very few electrons which begin to collect in the accelerating region before the beam is strongly modulated. As the level is increased, the current is redistributed, more going into the tail without much alteration in the shape of the velocity pattern, and with no strong bunching at any part of the curve. This result is exaggerated in Fig. 9(c) by measuring with a

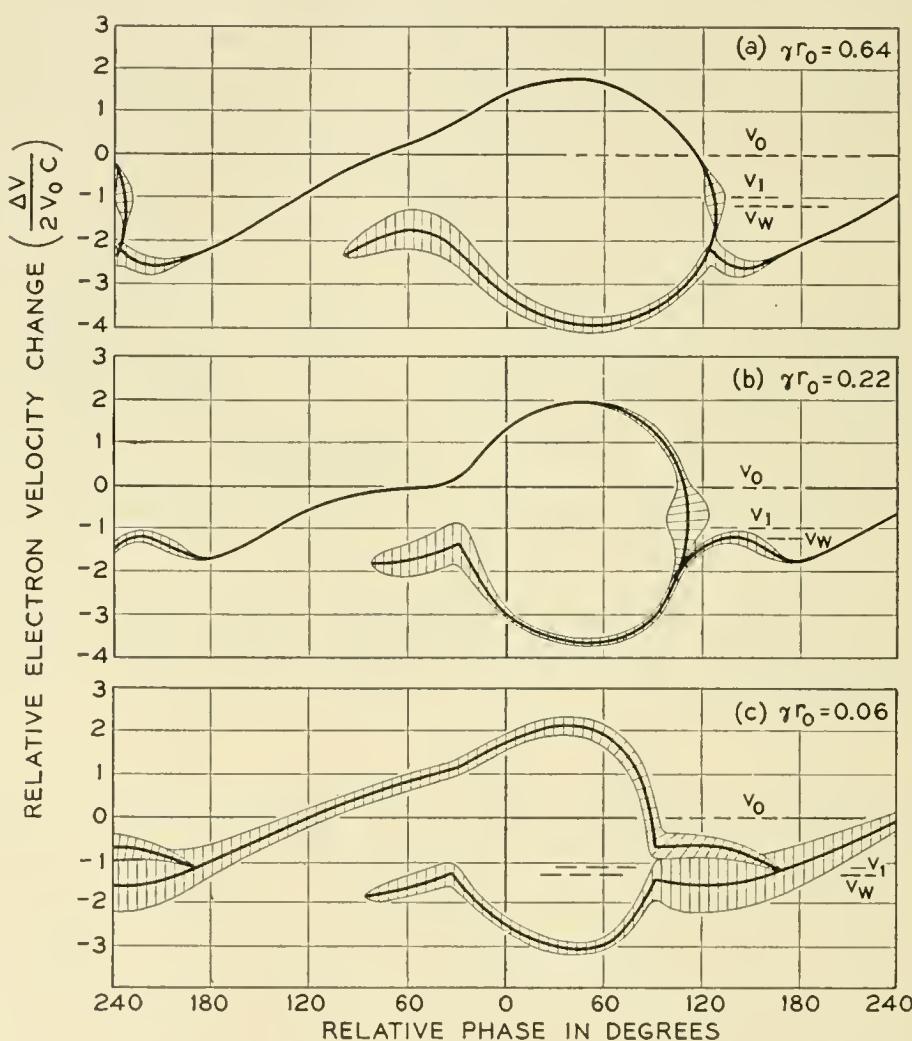


Fig. 9 — Curves of current and velocity as influenced by γr_0 . Space charge becomes a very potent factor near overload, especially when the beam is small. For this run $QC = 0.34$ and $b = 1.0$.

ridiculously small beam. By comparison with curves taken for larger beams, the tail is diminutive, electrons are much more uniformly distributed over all velocities and phases, and a peculiar splitting of velocities in the main bunch is found. The latter indicates that electrons entering from the higher velocity region move forward in the bunch, and the rest gradually retard. The smaller reduction in velocities, and the spread of electrons into the higher velocity regions is consistent with the lower efficiency measured (Fig. 2).

To explain the observed difference in high level performance of tubes with different size beams we must consider the character of the ac longitudinal space charge field. The coulomb field from an elemental length of an electron beam is inversely proportional to the square of the distance from the element

$$E = \text{Const} \frac{q\Delta z}{(z - z_1)^2} \quad (7)$$

provided $(z - z_1) \gg r_0$ and $(z - z_1) \ll a$.

For $(z - z_1)$ not small compared to a , (i.e., circuit radius not awfully large) the field would drop even faster with $(z - z_1)$ due to the shielding effect of the circuit. On the other hand, very near to the beam element ($z - z_1 \ll r_0$), the field is approximately that of a disc, which is nearly independent of z , i.e.,

$$E = \text{Const} \cdot \frac{q\Delta z}{\pi r_0} \quad (8)$$

independent of z for $z \ll r_0$.

Thus to a fair approximation the space charge field may be considered to be uniform for an axial distance of the order of a half a beam radius, and to drop rapidly at greater distances. For a given current element, a small diameter beam has an intense field extending only a short distance, while an equal charge element in a larger beam has a weaker longitudinal field extending to a greater distance.

At low amplitudes the extent of the forces makes no difference in operation, for a sinusoidal current gives a sinusoidal space charge field in either case. However, at large amplitudes, a sharp change in current density has a very high short range space charge field if the beam is small, or a much smaller smoothed out long range field if the beam is large. For $\gamma r_0 = 0.5$ which appears to be an optimum compromise between the effects of space charge and field non-uniformity, the space charge field could scarcely be confined closer than about $\pm 30^\circ$ in phase. On the other hand, a sharp bunching of electrons in a beam having

$\gamma r_0 = .05$ would have 100 times the space charge field, extending however only one tenth as far from the current discontinuity.

Returning to Fig. 9 we can see how these considerations enter into the development of the beam modulation. In the case of the small beam, Fig. 9(c), at the very beginning of the formation of a cusp, the strong highly concentrated space charge force causes a rapid deceleration of nearby electrons, resulting in the relatively early formation of a diminutive tail. The very high localized space charge force also prevents as tight bunching of electrons, forcing some to move forward and continuously repopulate the accelerating part of the wave. The relatively early falling apart of the initial bunch and the greater acceleration of the overriding electrons evidently give the latter enough velocity to penetrate the main bunch of electrons and form the second class of electrons in the main bunch, 90° – 150° in Fig. 9(c). Thus the net result of reducing the beam size is a severe aggravation of space charge debunching effects, with a consequent reduction in efficiency. To get high efficiency, we conclude, the beam should not be small. It should not be larger than $\gamma r_0 = 0.7$ however, for then the circuit field is not uniform enough over the beam cross-section to excite it properly, resulting in a loss in efficiency as is evident in Figs. 2 and 3.

EFFECT OF INCREASED BEAM VOLTAGE

It is common practice in the operation of traveling wave tubes to elevate the beam voltage, taking a sacrifice in gain in order to obtain increased power output. The effects on the beam modulation are shown in Fig. 10. In Fig. 10(a), the voltage is somewhat below that giving maximum gain. The curve is characteristic of what we have already seen but the bunching is less pronounced and the velocities are less reduced. In Fig. 10(b) the voltage is somewhat above that giving maximum gain and the curve is much like that of Fig. 8 except that the decelerated electrons are slowed by a greater amount, consistent with the increased separation of electron and wave velocity, and also with the measured increase in power output.

Increasing the beam voltage still further gives only a slight increase in efficiency. Fig. 10(c) shows that even though electrons are slowed to still lower velocities, and the velocity spread is increased, many more electrons override the circuit wave and are accelerated, thereby offsetting the greater contribution of the slower electrons. This is much like what was seen with increasing space charge (QC) and indeed the effects are almost equivalent. As one would expect therefore, little is gained by elevating the beam voltage if the space charge is large, the

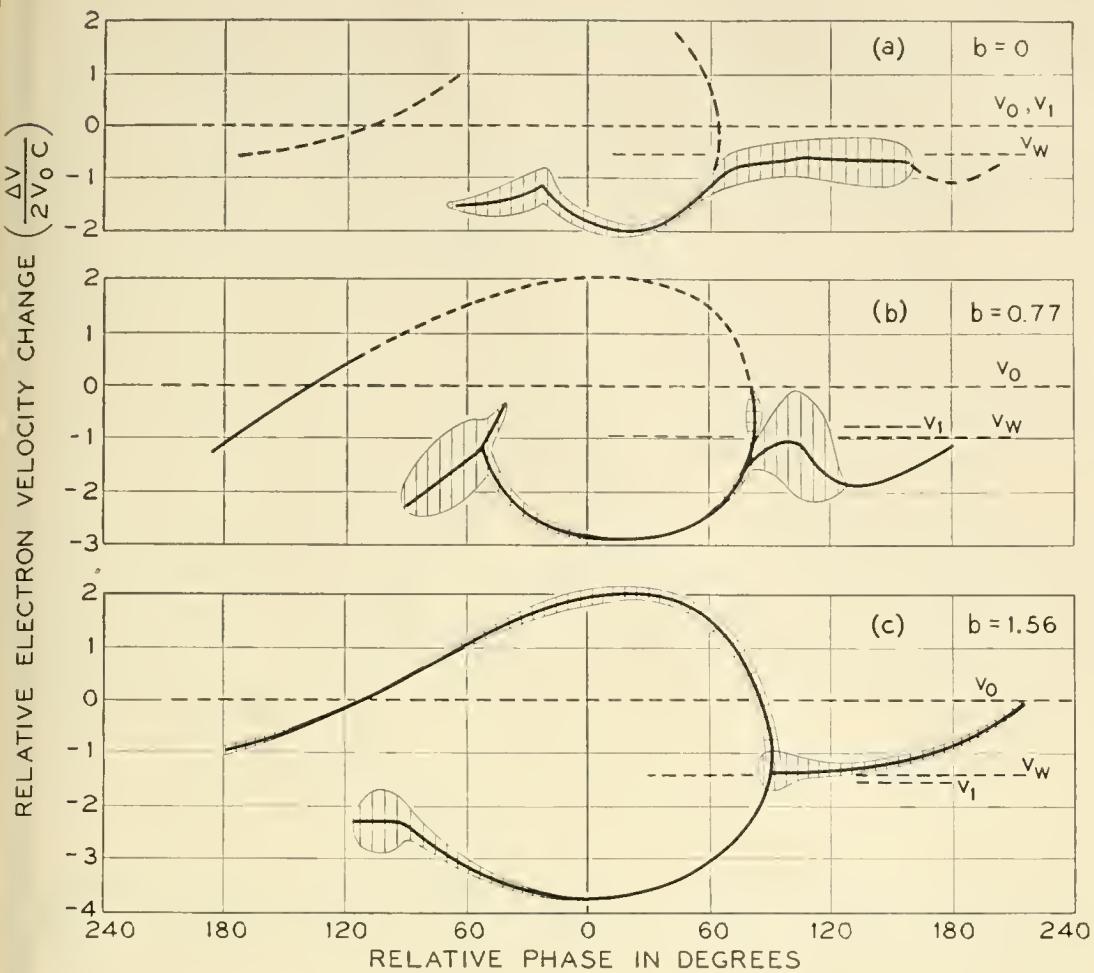


Fig. 10 — The influence of beam velocity on ae velocity and current. When the velocity is raised too high, the electrons are not effectively trapped by the wave, and override into the accelerating field. With large QC and/or small γr_0 the electrons override in any case, and little is gained by increasing b . For this case $QC = 0.13$ and $\gamma r_0 = 0.21$.

main effect being to push more electrons forward into the accelerating region.

ELECTRIC FIELD IN THE BEAM

Besides telling a clear story of the non-linear dynamics of the traveling wave tube, the foregoing curves contain a lot of information about average current and velocity distributions. From the current or velocity curves we can in turn deduce the distribution of longitudinal electric field in the beam. Figs. 11(a) and (b) show the instantaneous current as a function of phase, taken from the curves of Figs. 8(a) and (b). The infinite differential in the velocity curve necessarily gives a pole in the charge density (at about 88°). The total charge in the vicinity of the

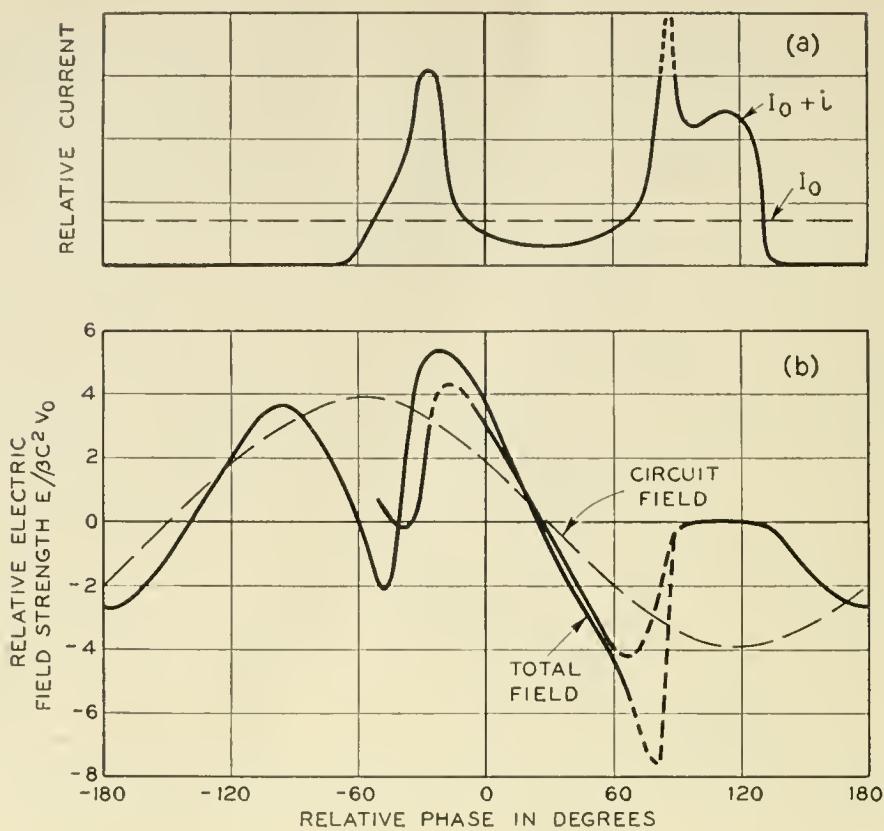


FIG. 11 — AC current and electric field in the beam. The upper curve comes directly from Fig. 8(a). The lower curve is deduced by an approximate method from the velocity curve of Fig. 8(a). The double value below 90° is partly due to inconsistency between the two parts of the velocity curve, and partly due to the nature of the approximation.

pole, and the range of the space charge force (dependent upon QC and γr_0) determines its effect upon the electron dynamics.

Most of the current is incorporated in the two bunches nearly 180° apart, as we have seen, each bunch having a current density many times the average.

We might obtain the space charge fields from the current density, but this would require a rather definite knowledge of the characteristic space charge field versus distance as influenced by beam diameter. It would also be pushing the accuracy of charge density measurement, which is crude at best. A better way is to compute the electron acceleration from the velocity curves. This may be done by taking two velocity patterns at slightly different signal levels, and tracing electrons from one to the next, using the measured velocity to determine the relative phase shift of any electron.

In the appendix it is shown that a close approximation to this is

$$E_\Phi = 2\beta C^2 V_0 \left[\frac{(V_0 - V_w) + \Delta V}{2V_0 C} \right] \frac{d}{d\Phi} \left(\frac{\Delta V}{2V_0 C} \right) \quad (10)$$

where the parameters are all obtained from a single velocity curve, and

E_Φ is field strength in volts meter at phase Φ

$\frac{\Delta V}{2V_0C}$ is the value of the ordinate of the velocity characteristic of interest (Figures 7 to 10) and

$\left(\frac{V_0 - V_w}{2V_0C}\right)$ is the value of the ordinate corresponding to the wave velocity. (To be precise, the wave velocity at the associated output level, but to a reasonable approximation, that of the wave velocity at low levels. (This value is indicated by V_w in the velocity curves.)

The total electric field has been computed for the case of Figs. 8(a) and (b) and is given in Figs. 11(b) and 12(b) together with the circuit field calculated for the associated power level and plotted with an arbitrarily chosen phase. In each case it is seen that the space charge field is comparable in magnitude to the circuit field, is far from sinusoidal, and

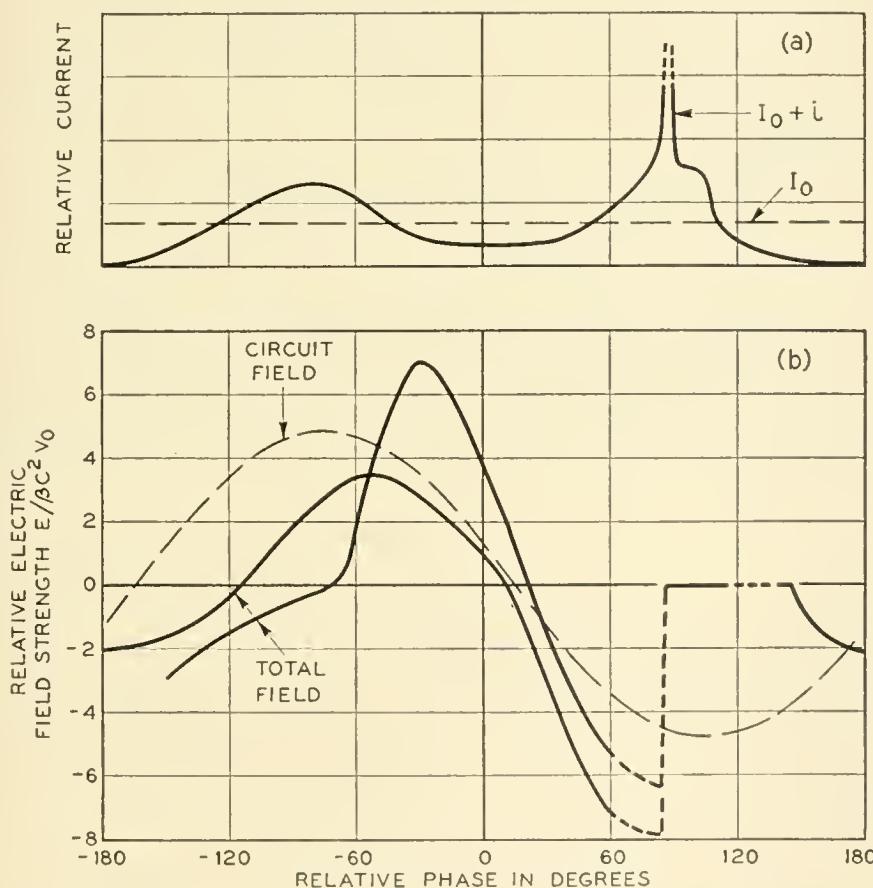


Fig. 12 — AC current and electric field in the beam deduced from Fig. 8(b). The greater space charge results in a less defined bunch, and smoother space charge field than in Fig. 11.

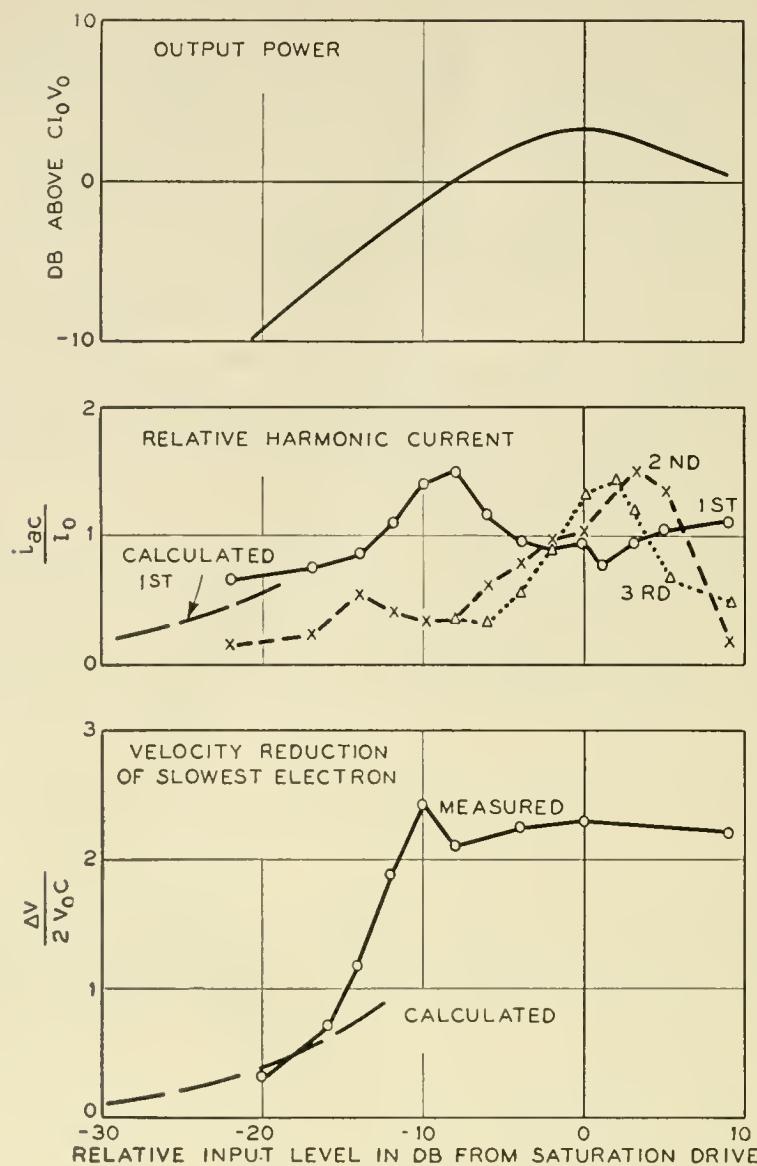


Fig. 13.—Curves of output level, fourier component amplitudes of beam current, and peak velocity as a function of input level for low space charge. These curves were deduced from Fig. 8(a).

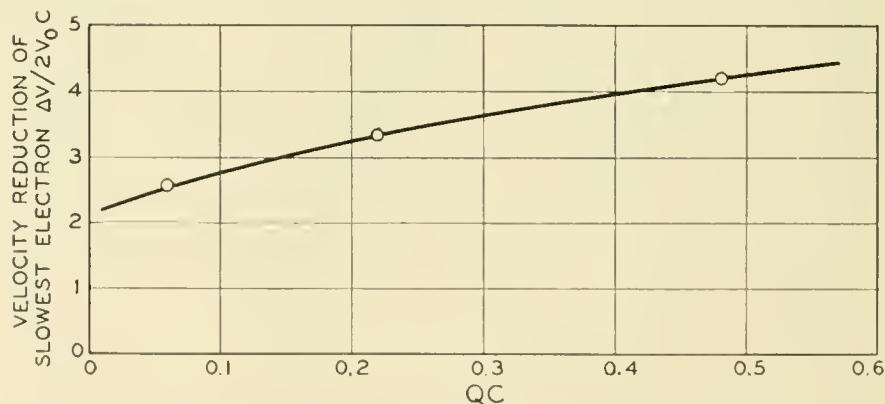


Fig. 14.—Maximum velocity reduction as a function of space charge (from Fig. 8). The velocity reduction is about $3.5 y_1$.

agrees qualitatively with what would be expected from the associated curve of beam current.

To determine the curves of Figs. 11 and 12 is rather stretching the accuracy of the measurements as can be seen by the large discrepancy in the field calculated from the two parts of the velocity curve which of course should be identical. The figures do give an interesting qualitative picture of traveling wave tube behavior however, and are included here for that reason.

OVERALL VELOCITY SPREAD

Of more practical importance is the overall velocity spread in the spent beam. It is often desirable to reduce the power dissipation in a traveling wave tube by operating the collector at a potential below that of the electron beam, and it is interesting to see how far one might go. Fig. 13 shows how the velocity reduction of the slowest electron, together with the output level and fourier current components of beam current vary with input level. For small amplitudes, the low level theory accurately predicts the velocity, but near overload, as we have seen, the minimum velocity drops sharply to a value several times lower than that projected from small signal theory.

The maximum velocity spread dependence upon the space charge parameter QC is shown in Fig. 14. Similar data for values of the other parameters may be obtained from the velocity diagrams.

From the foregoing data, one can deduce the amount of reduction of collector potential that should be theoretically possible without turning back any electrons. An idealized unipotential anode could collect all the current at a potential ΔV (in the foregoing figures) above the cathode, decreasing the dissipated power by a factor of $\Delta V/V_0$ below the dc beam power.

STOPPING POTENTIAL MEASUREMENTS

Information on spent beam velocity has also been obtained by a stopping-potential measurement at the collector of a more conventional 4,000-mc traveling wave tube.* Two fine mesh grids were closely spaced to a flat collecting plate, and collector current was measured as a function of the potential of second grid. The first grid was very dense, to prevent reflected electrons from returning into the helix. One curve taken with this arrangement is shown in Fig. 15 and for comparison we have

* Similar measurements have been reported by Atsumi Kondo, Improvement of the Efficiency of the Traveling Wave Tube, at the I.R.E. Annual Conference on Electron Tube Research, Stanford University, June 18, 1953.

plotted the distribution predicted from Fig. 9(b). The *RF* losses in the 4,000-mc tube were not negligible, and probably account for slightly smaller power output and greater proportion of higher velocity electrons.

COMPARISON WITH COMPUTED CURVES

Non-linear calculations of traveling wave tube behavior have been made by Tien, Walker and Wolontis⁷ and by Tien⁹ covering the same region of parameter values as is reported here. In Figs. 2, 3, 4 and 9 are shown some of their data on our coordinates. The similarity of the results over much of the range is rather reassuring. It is interesting that in order to make the computations it was necessary to assume two space charge factors, just as was found experimentally. There are, however, some significant differences:

1. In general, the computed values give a higher value of efficiency than is measured, by about 25 per cent. Thus, the computations indicate

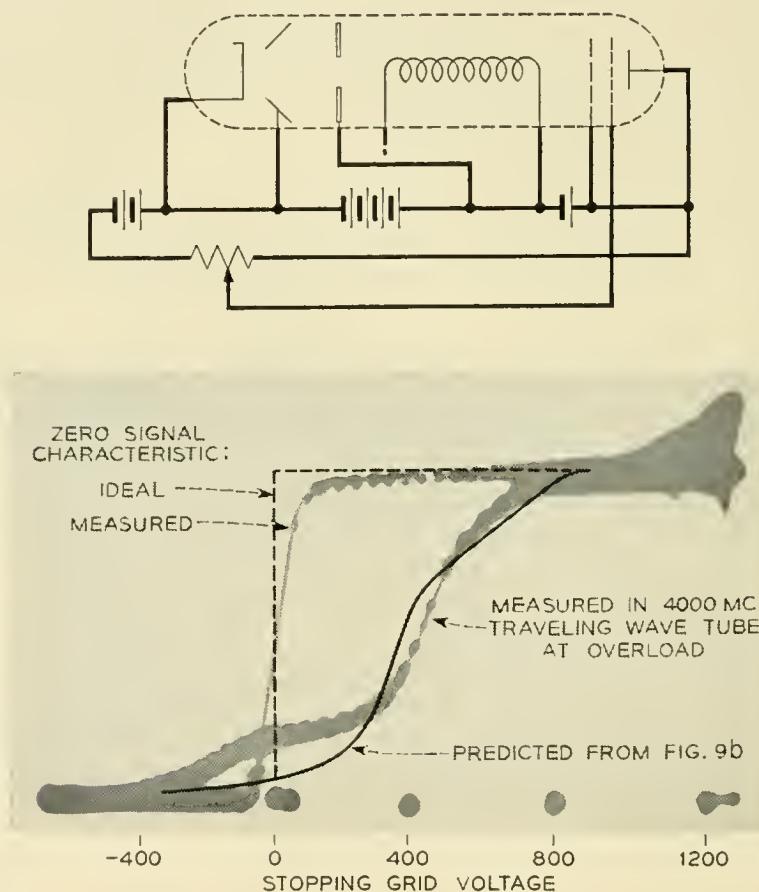


Fig. 15 — Collected current versus stopping potential. The oscilloscope curve is for a 4,000-mc tube, and the other that predicted from the scale model measurements. By integrating current as a function of velocity for Figs. 7-10 stopping potential distributions can be deduced for other conditions.

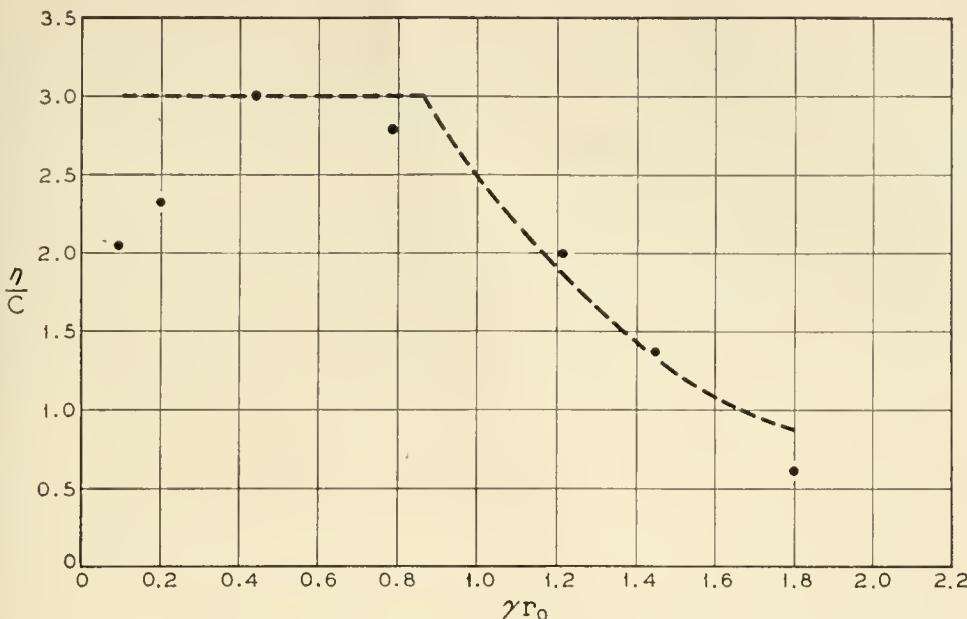


Fig. 16 — Efficiency versus γr_0 for small QC . The dashed curve is proportional to the amount of beam current in the circuit field strength having at least 85 per cent of the intensity at the edge of the beam. This illustrates the fact that for large beams only the edge of the beam is effective.

that with the reasonable values of $QC = .25$ and $\gamma r_0 = 0.8$ ($k_T = 2.5$), the efficiency would be about $3.8C$, whereas the measured value is $3.1C$.

2. The largest discrepancy in the measured and computed value of η/C is for large values of γr_0 (small k_T), where the computations show a steady increase in efficiency instead of a sharp decrease. This arises because the computational model assumed the electric field to be uniform across the beam, whereas in the actual tube it varies as $I_0(\gamma r)$, and for large values of γr_0 the field is weak near the beam axis. This effect is shown in Fig. 16 where η/C is plotted versus γr_0 for small values of QC , on the same scale with a curve proportional to the square of the fraction of the beam within a cylindrical shell such that

$$1 - \frac{I_0(\gamma r_1)}{I_0(\gamma r_0)} = 0.85 \quad (11)$$

where r_1 is the inside radius, and r_0 the outside beam radius (i.e., the fraction of the beam in a field greater than 85 per cent of that at the beam edge).

No serious studies of velocity were made for large beams, but on cursory examination it was evident that the beam modulation varied considerably over the cross section when the beam was very large, and scarcely at all when it was smaller than around $\gamma r_0 < 0.8$.

3. The observed effect of small beam radius upon efficiency is not as pronounced as was found in the computations. The reason is not known but may be due to modulation of the beam diameter at large signal levels. This effect would be negligible with the larger γr_0 's, due to the focusing fields being relatively much larger.

4. The computations, and also those of Nordsieck,⁵ Poulter⁶ and Rowe⁸ indicate a much higher efficiency than has been observed at elevated beam voltages and small C and QC . The reason for this may be that the limited number of "electrons" used in the computational models fail to adequately account for the very sharp space charge cusp that forms under low QC conditions, or that interpolation between their points should not be linear, as assumed in making the comparison. On the other hand it would be difficult to be sure that nonuniformities in electron emission were not influencing the measurements in the case of the large beams by giving a larger QC than calculated.

5. The increase in efficiency to be had by elevation of beam voltage is much smaller than is indicated by the computations. This may be a real difference, or it may be that at elevated voltages, the measurements are beginning to feel the influence of overloading in the attenuator. The margin of safety on attenuator overloading is not as great as one would like at the higher frequencies.

6. The velocity curves, Fig. 8, compare the computed and measured data on three runs. For small QC , Fig. 8(a), the agreement is remarkably good considering the fact that in the computation only 24 "electrons" were used to describe a rather complicated function. The effect of the lumping of space charge in the artificial 'disc' electrons causes a scatter-of points which is different from that in an actual tube as is especially apparent in Figure 8c. In spite of this the computational results indicate a velocity spread and current distribution not greatly different from that observed.

CONCLUSIONS

The large scale model traveling wave tube is a means for the determination of non-linear behavior, and has been valuable in determining relationships and limitations important to efficient operation of such tubes. It has shown that there is a broad optimum in tube parameters around $C = 0.14$ $QC = 0.2$ and $\gamma r_0 = 0.5$ for which values it is possible to obtain efficiencies well above 30 per cent. The measured ac beam velocity and current near overload show that it is unlikely that significant increase in efficiency can be obtained by any simple expedients such

as operations on the helix pitch alone, or the use of an auxiliary output circuit.

The results being in normalized form, are believed to be generally applicable to conventional traveling wave tube design. With determination of an equivalence in beams, they should even be a useful guide in the design of tubes using hollow beams or other configurations.

The work described could not have been done without the able assistance of G. J. Stiles and L. J. Heilos and the helpful council of many of my colleagues at Bell Telephone Laboratories.

APPENDIX

SCALE MODEL TUBE DESIGN

There were a larger number of factors to be accounted for in the design of this tube. Its proportions should be such as to make it representative of the usual design of traveling wave tube. Its size should be such as to make it easy to define the electron beam boundary, and to dissect the beam. The size should also be such that the electron beam velocity analysis could be done before the beam character would be changed either by space charge, or its velocity spread. The voltage should be low so that further acceleration in the velocity analyzer would not lead to an inconveniently high voltage. Finally, the availability of suitable measuring gear over a 3-1 frequency range, and the size of the laboratory must be considered. All of these factors led to low frequency operation, limited principally by the laboratory size and the mechanics of construction.

A moderate perveance of around 0.2×10^{-6} was taken, with a γa of 1.2 and γr_0 of 0.8 in a representative helix with small impedance reduction due to dielectric and space harmonic loading. This is representative of practical tube design in the microwave range and is centered on the parameter values of most general interest. At a frequency of 100 mc and a beam potential of 400 volts this resulted in a helix 10 feet long and $1\frac{1}{2}$ inches in diameter, with an electron beam 1 inch in diameter. The choice of frequency was finally determined by the availability of measuring equipment, and the voltage was selected to give a convenient size for dissection of the electron beam.

By changing frequency, beam current, and beam diameter it was possible to cover a reasonable range of γr_0 , and QC , and to make some observations into the region of large C operation.

In all of the measurements described, a very strong uniform magnetic field was used to confine the beam, and therefore sealing of the magnetic

focusing field need not be considered. The electron beam was produced in a gridded gun and is thus near to the ideal confined flow, which is the only focusing arrangement which is known to determine a reasonably uniform boundary to the beam. The beam size and straightness was checked using a fluorescent screen at the collector end.

NORMALIZING FACTORS

The measurements described are expressed relative to the linear theory, in Pierce's² notation, which are generally used in the design of traveling wave tubes. Thus, instead of being presented in the terms of measurement or simply normalized to efficiency, perveance, impedance, etc., they are expressed in terms of C , QC , γr_0 , etc., with normalized fields, currents and velocities. In this way the results become adjuncts to the linear theory and are more easily applied to tube design. Electron velocity is plotted on the same scale as the relative velocity parameters b and y_1 used in low level theory, (i.e., normalized to $\Delta V/2V_0C$). Efficiency is normalized as η/C , which for C less than 0.1 is relatively independent of C . Field strength in the linear region is proportional to

$$\sqrt{\frac{\eta'}{C}}$$

(η' being efficiency measured at the appropriate signal level). Solving the equation for C^3 ,

$$C^3 = \frac{E^2}{2\beta^2 P} \frac{I_0}{2V_0} \quad (12)$$

gives us

$$\sqrt{\frac{\eta'}{C}} \approx \frac{E}{\beta C^2 V_0} \quad (13)$$

which we use as the normalizing parameter for electric field. Circuit potential is the integral of circuit field over a quarter period, giving a normalized parameter V/V_0C^2 . For convenience in the use of common coordinates, circuit potential was plotted as $-V/2V_0C^2$ in Figure 7.

The other curves are plotted as values relative to dc quantities or to saturation level.

Strictly speaking, the results hold only for tubes having the same proportions as the model. Practically, however, as long as the helix impedance and radius (ka or γa) are not different by orders of magnitude from the values used, and as long as the perveance is low (below 2×10^{-6} for

instance), the results are believed to be significant for tubes having the indicated values of γr_0 and QC .

HELIX IMPEDANCE

It is important to the measurements to have an accurate evaluation of the helix impedance. Several methods of measurement have been discussed in the literature.^{21, 22} That described by R. Kompfner was selected, wherein the circuit impedance is correlated with the beam current and voltage which gives a null in the output signal. When the beam voltage and current are adjusted to give zero transmission for a lossless section of helix (neglecting space charge) $CN = 0.314$ and $\delta V/V_0 \cong 1/N$. Using the measured length of the helix, and measuring the voltage and current giving the null in signal transmission, we can compute C , and thus the impedance and velocity (synchronous voltage) of the helix.

The impedance was calculated by P. K. Tien,²³ and the results are compared in Fig. 17. The measured impedance at the high frequency end was much too low until space charge in the beam was accounted for in interpreting the measurements. Fortunately, in the absence of attenu-

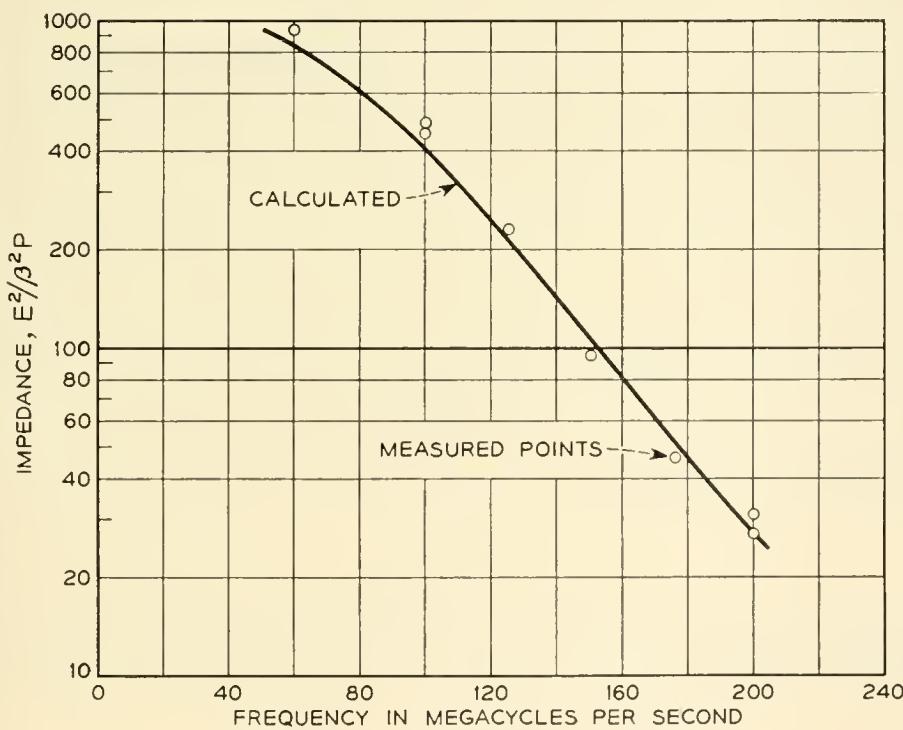


Fig. 17 — Helix impedance as a function of frequency. The impedance was calculated taking into account dielectric loading and wire size. It was measured using the Kompfner dip method, taking account of space charge.

ation, the conditions for start of oscillation in a backward wave oscillator are the same as for the output null in a traveling wave tube. Space charge was first accounted for using the results of H. Heffner^{24, 25} giving an excellent check between predicted and computed helix impedance. Later C. F. Quate²⁶ showed that the same measurement could be used to determine the space charge parameter QC as well as the helix impedance. Since thermal velocity effects and the uncertainty of some of the assumptions used in evaluating the small signal effects of space charge cast some doubt on the proper evaluation of this term, further measurements were made on this factor, and a satisfactory correlation between the observed value of QC and that computed from the Fletcher²⁷ curves was obtained.

TOTAL ACCELERATING FIELDS

From the velocity characteristics shown in Figs. 7 through 10, we can deduce the electron accelerations, and thus the electric fields at any point. While the curves are actually diagrams of velocity as a function of phase, they closely correspond to the velocity-time or distance distribution of the electrons in the traveling wave tube. Knowing these characteristics we can deduce the motion of any element of charge, and thus the force under which it moves. It is observed that over most of the curve the shape of the velocity pattern does not change nearly so rapidly as the redistribution of electrons within the pattern. Thus, we can approximate the situation at any amplitude by assuming the velocity pattern to be constant, and that electrons move within the pattern according to simple particle dynamics. This is a good approximation except where the acceleration is high (i.e., vertical crossings of the wave velocity line).

Consider then an element of the velocity pattern at phase Φ_1 and velocity $(u_0 + \Delta u)$. In an interval dt this element will move a distance

$$(u_0 + \Delta u) dt \quad (14)$$

and will change velocity by

$$du = E \frac{e}{m} dt \quad (15)$$

At the same time the wave will have moved a distance $v dt$, resulting in a relative change in phase between wave and current element of

$$d\Phi = \beta(u_0 - v + \Delta u) dt \quad (16)$$

In terms of equivalent differences the term in brackets can be written

$$(u_0 - v + \Delta u) = \sqrt{2 \frac{e}{m} V_0} C \left(\frac{V_0 - V_w + \Delta V}{2V_0 C} \right) \quad (17)$$

from (16) and (17) we can write:

$$\begin{aligned} \frac{du}{dt} &= \frac{du}{d\Phi} \frac{d\Phi}{dt} \\ &= \frac{d}{d\Phi} \left(\frac{\Delta V}{2V_0C} \sqrt{2 \frac{e}{m} V_0} C \right) \left[\beta \sqrt{\frac{e}{m} V_0} C \left(\frac{V_0 - V_w + \Delta V}{2V_0C} \right) \right] \end{aligned} \quad (18)$$

giving (from 15)

$$\frac{E}{\beta V_0 C^2} = 2 \left[\left(\frac{V_0 - V_w}{2V_0C} \right) + \left(\frac{\Delta V}{2V_0C} \right) \right] \frac{d}{d\Phi} \left(\frac{\Delta V}{2V_0C} \right) \quad (19)$$

β , V_0 and C are constants of the tube, the first inner parenthesis may be calculated from the tube constants and is shown in the curves. $\Delta V/V_0C$ and its differential are the value and the slope of the velocity curve in question.

The important approximations here are that the velocity-phase curves are representative of velocity-distance characteristics, which is true for small values of C , and that the electrons move roughly tangent to the given velocity pattern. By comparing several patterns at different signal levels it is observed that this is true to a fair accuracy over most of the curve. Also it is assumed that the wave velocity at large amplitudes is the same as that for small signals, which is not quite true. The resulting curves give at least a qualitative picture of the field distribution within a traveling wave tube, and serve to emphasize the importance of space charge fields in determining the non-linear characteristics.

ELECTRIC FIELD OF THE HELIX WAVE

In order to see what part of the field is due to space charge we must evaluate the corresponding helix fields. A value for this can be derived from the basic traveling wave tube equations assuming the helix fields to be sinusoidal and not seriously affected in impedance by the beam (small C again). By definition

$$\frac{E^2}{2\beta^2 P} \frac{I_0}{4V_0} = C^3 \quad (18)$$

and

$$\frac{\eta'}{C} = \frac{P}{I_0 V_0 C} \quad (19)$$

where η' is normalized power level, i.e., efficiency corresponding to the signal level E of interest. From this we deduce for the normalized circuit

field

$$\frac{E}{\beta V_0 C^2} = 2\sqrt{2} \sqrt{\frac{\eta'}{C}} \quad (20)$$

which integrates to give a normalized ac circuit voltage

$$\frac{V}{2\sqrt{2}V_0 C^2} = \sqrt{\frac{\eta'}{C}} \quad (21)$$

RELATIVE PHASE BETWEEN WAVE, VELOCITY AND CURRENT

The velocity analyzer provides no convenient measure of relative phase between the helix wave and the beam modulation. Therefore we compute the relation of helix field and beam modulation for a small signal, and for large amplitudes measure the phase of each relative to that at small amplitudes.

$$\text{Pierce gives the relationship}^2 \quad v = \frac{-\eta \Gamma V}{u_0(j\beta_e - \Gamma)} \quad (22)$$

$$\text{which using (9) and the fact that } \beta_e C \delta = j\beta_e - \Gamma \quad (23)$$

gives for the small signal beam modulation

$$\frac{\Delta V}{2V_0 C} = -j \frac{\sqrt{2}}{\delta} \sqrt{\frac{\eta}{C}} = \frac{\sqrt{2}}{|\delta|} \sqrt{\frac{\eta'}{C}} \left[\left(\tan^{-1} \frac{y_1}{x_1} \right) - \frac{\pi}{2} \right] \quad (24)$$

Similarly we have for the small signal current modulation

$$i = I_0 \sqrt{2 \frac{\eta}{C} \cdot \delta^2} = I_0 \sqrt{2 \frac{\eta}{C} |\delta^2|} \left[2 \tan^{-1} \frac{x_1}{y_1} \right] \quad (25)$$

The value of $\delta (= x_1 + jy_1)$ is given in Fig. 18, drawn from data supplied by P. K. Tien, from Pierce² and from Birdsall and Brewer.²⁸ This figure was also used as a basis for determining the values of y_1 and b used in several of the curves.

MEASUREMENT OF POWER

The output power, and relative output phase was measured using a micro-oscilloscope.²⁹ The subharmonic of the signal was used for a sweep voltage, and phase was measured from the shape of the observed lissajou figures. The oscilloscope deflection was compared with the de deflection from a battery standard, and checked on occasion with a bolometer power meter at the operating frequency.

THE VELOCITY ANALYZER

There are many ways in which one may separate velocities in an electron stream. Crossed electric and magnetic fields were used in this ex-

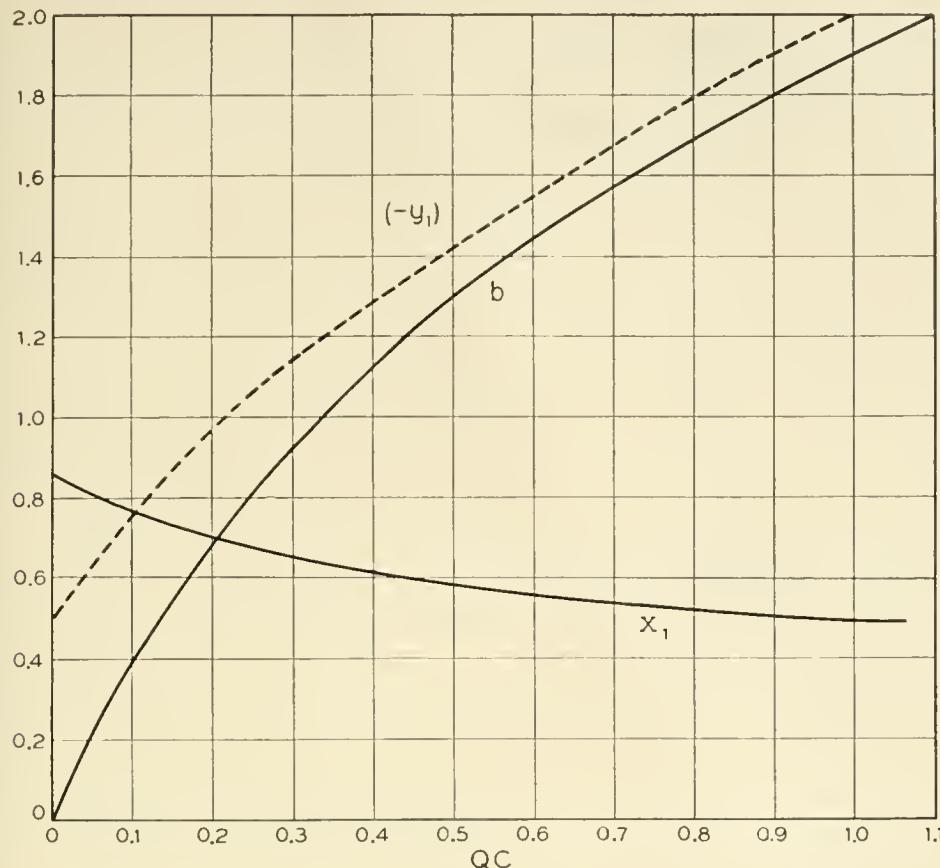


Fig. 18 — Increasing wave propagation factors used in interpreting the measurements. These are the maximum value of x_1 and the corresponding value of b and y_1 for given values of QC .

periment because a simple control of sensitivity was important in order to study velocity differences ranging from 1 per cent up to as much as 100 per cent of the dc beam velocity.

The velocity analyzer is sketched in Fig. 5. It consists of an aperture which transmits only a few microamperes of the electron stream; a magnetic pole piece (not shown) terminating the focusing field; a pair of horizontal deflection plates; an electrostatic lens system; pole pieces and deflection plate to provide a region with crossed electric and magnetic fields; and finally a drift tube, a post deflection acceleration electrode and fluorescent screen. The whole assembly is raised 1,000 volts above the helix potential and the 0.001" aperture is very close to the end of the helix, so that the electrons are very quickly accelerated to a high voltage. By this means, the region of debunching outside of the helix field is kept below 1.4 radians transit angle and the velocity spread within the analyzer is reduced by a factor of four. Space charge within the analyzer is entirely negligible because of the small current transmitted.

In order to diseriminate in phase before the electrons are scrambled

due to their spread in velocity, the horizontal sweeping plates are mounted just as close to the aperture as is deemed practical. The observed velocity spreads in the beam were such as to give less than 0.2 radians error in phase under the worst conditions.

The horizontal deflecting plates were driven synchronously with a sub-harmonic of the RF input to the helix, and the resulting deflection served to separate electrons according to phase in the final display.

Placing the focusing lens after the deflection plates results in a considerable reduction in deflection sensitivity. However, undesirable magnification of the pinhole aperture dictated that the lens could not be close to it, and it was important to initiate the deflection as early as possible. The lens consists of three discs, the center one being biased to about 800 volts above the mean voltage of the rest of the system.

Immediately after the lens there are two iron pole pieces and two insulated electric deflection plates which extend parallel to the beam for $1\frac{1}{4}$ inches. The pole pieces provide a dc magnetic field up to about 20 gauss induced by small coils outside of the envelope, and the electric deflection plates are biased with up to a corresponding 50 volts dc polarized to oppose the magnetic deflection of the beam. The electric and magnetic fields are adjusted so that the normal unmodulated electron beam traverses the region with no deflection and strikes the center of the fluorescent screen. In the crossed field region

$$\frac{E}{B} = \sqrt{2 \frac{e}{m} V_0}. \quad (26)$$

Electrons having greater or lesser velocity are deflected parallel to the electric field, and give a corresponding deflection from the center of the fluorescent screen.

To get a display in which the various elements are not hopelessly entangled, it was necessary to sweep the trace in an initial ellipse at a subharmonic rate. The sweep voltage was applied to the horizontal deflection plates, with just a little applied to the vertical plates through a phase shifter. The relative phase of any part of the trace was measured from the ellipse, and the velocity sensitivity was calibrated by observing the ellipse deflection as a function of the de beam potential, as shown in Fig. 6(a). There is a small error due to the sensitivity of deflection to velocity, and due to distortion of the ellipse by fringing fields.

In order to measure velocity and current density in the displayed pattern, the fluorescent screen was photographed, and the negative projected in a microcomparator. It was assumed that with the small currents used, the light intensity was proportional to current, and the film linearity was calibrated by making exposures of several different durations. The trace density was measured with a densitometer, sweeping

over the trace width to account for variations in focus for different parts of the pattern. Admittedly, the process is not very accurate, but it does give a rough measure of current density and helps considerably in interpreting the observed velocity patterns.

NOMENCLATURE

<i>a</i>	Circuit radius
<i>b</i>	Parameter relating electron velocity to that of the cold circuit wave $u_0 - v_1/u_0 C = \Delta V/2V_0 C$
<i>B</i>	Magnetic field
β	the axial phase constant ω/v_1
<i>C</i>	The gain parameter $= (E^2/2\beta^2 P) (I_0/4V_0)$
γ	Radial phase constant $\cong \beta = \omega/v_1$
δ_1	Complex propagation constant for the increasing wave
<i>E</i>	Electric field
E_Φ	Electric field at phase Φ
e/m	Charge to mass ratio of the electron
I_0	Beam current in amperes
$I_0()$	Modified Bessel function
k_T	Tien's constant $k_T = 2/\gamma r_0$
ka	Circuit circumference measured in (air) wavelengths
<i>N</i>	Number of wavelengths
η	Maximum efficiency
η'	Efficiency at an intermediate power level
<i>P</i>	RF power obtainable from the circuit
<i>QC</i>	Space charge parameter
<i>q</i>	Charge per unit length in the electron beam
<i>r</i>	Radial distance from the axis
r_0	Beam radius
<i>t</i>	Time variable
<i>u</i>	Electron velocity
u_0	DC beam velocity
<i>v</i>	AC velocity of the electron beam
v_1	Wave velocity
V_0	DC beam voltage
T_w	Voltage corresponding to the wave velocity
ΔV	Voltage difference corresponding to the difference in velocity of an electron and the dc beam velocity
δV	Difference between synchronous voltage and that giving the Kompfner dip
Φ	Relative phase
<i>z</i>	Distance measured along the beam

REFERENCES

1. Pierce, J. R., Theory of the Beam Type Traveling Wave Tube, Proc. I.R.E., **35**, pp. 111-123, Feb., 1947.
2. Pierce, J. R., Traveling Wave Tubes, D. VanNostrand Co., Chapter XII.
3. Slater, J. C., Microwave Electronics, D. VanNostrand Co., 1950, pp. 298.
4. Brillouin, L., The Traveling Wave Tube (Discussion of Waves of Large Amplitudes), *J. Appl. Phys.*, **20**, p. 1197, Dec., 1949.
5. Nordsieck, A., Theory of the Large Signal Behavior of Traveling Wave Amplifier, Proc. I.R.E., **41**, pp. 630-647, May, 1953.
6. Poulter, H. C., Large Signal Theory of the Traveling Wave Tube, Tech. Report No. 73 Electronics Research Laboratory, Stanford University, Stanford, California, Jan., 1954.
7. Tien, P. K., Walker, L. R., and Wolontis, V. M., A Large Signal Theory of Traveling Wave Amplifiers, Proc. I.R.E., **43**, pp. 260-277, Mar. 1955.
8. Rowe, J. E., A Large Signal Analysis of the Traveling Wave Amplifier, Technical Report No. 19, Electron Tube Laboratory, University of Michigan.
9. Tien, P. K., A Large Signal Theory of Traveling Wave Amplifiers Including the Effects of Space Charge and Finite C, *B.S.T.J.*, **34**, Mar., 1956.
10. Brangaccio, D. J., and Cutler, C. C., Factors Affecting Traveling Wave Tube Power Capacity, *Trans. I.R.E. Professional Group of Electron Devices, PGED* **3**, June, 1953.
11. Crumly, C. B., Quarterly Status Progress Report No. 26, Electronics Research Laboratory, Stanford University, Stanford, California, pp. 10-12.
12. Doehler, O., et Kleen, W., Phénomènes non Linéaires dans les Tubes à Propagation D'onde" *Annales de Radioélectricité* (Paris), **3**, pp. 124-143, 1948.
13. Doehler, O., et Kleen, W., Sur le Rendement du Tube à Propagation D'onde," *Annales de Radioélectricité*, Tome IV No. 17 Juillet, 1949 pp. 216-221.
14. Berterotière, R., et Convert, G., Sur Certains Effets de la Charge D'espace dans les Tubes à Propagation D'onde, *Annales de Radioélectricité*, Tome V, No. 21, Juillet, 1950.
15. Klein, W., und Friz, W., Beitrag zum Verhalten von Wanderfeldröhren bei Hohen Engangspegeln, *F.T.Z.*, pp. 349-357, July, 1954.
16. Warnecke, R. R., L'évolution des Principes des Tubes Électroniques Modernes pour Micro-ondes, *Convegno di Eletronica e Televisione*, Milano, p. 12-17, Aprile, 1954.
17. Warnecke, R. R., Sur Quelques Résultats Récemment Obtenus dans le Domaine des Tubes Electroniques pour Hyperfréquences, *Annales de Radioélectricité*, Tome IX, No. 36, Avril, 1954.
18. Warnecke, R., Guenard, P., and Doehler, O., Phénomènes fondamentaux dans les Tubes à onde Progressive, *Onde Electrique*, France, **34**, No. 325, p. 323-338, 1954.
19. Brück, L., und Lauer, R., Die Telefunken Wanderfeldröhre TL6, *Die Telefunken-Röhre Heft* 32, pp. 1-21, Februar, 1955.
20. Brück, L., Vergleich der Verschiedenen Formeln für den Wirkungsgrad einer Wanderfeldröhre, *Die Telefunken-Röhre Heft* 32, pp. 23-37, Februar, 1955.
21. Cutler, C. C., Experimental Determination of Helical Wave Properties, Proc. I.R.E., **36**, pp. 230-233, Feb., 1948.
22. Kompfner, R., On the Operation of the Traveling Wave Tube at Low Level, *Journal British I.R.E.*, **10**, p. 283, Aug.-Sept., 1950.
23. Tien, P. K., Traveling-Wave Tube Helix Impedance, Proc. I.R.E., **41**, pp. 1617-1623, Nov., 1953.
24. Heffner, H., Analysis of the Backward-Wave Traveling-Wave Tube, Proc. I.R.E., **42**, pp. 930-937, June, 1954.
25. Johnson, H. R., Kompfner Dip Conditions, Proc. I.R.E., **43**, p. 874, July, 1955.
26. Quate, C. F., Power Series Solution and Measurement of Effective QC in Traveling-Wave Tubes, Oral presentation at Conference on Electron Tube Research, University of Maine, June, 1954.
27. Fletcher, R. C., Helix Parameters in Traveling Wave Tube Theory, Proc. I.R.E., **38**, pp. 413-417, Apr., 1950.
28. Birdsall, C. K., and Brewer, G. R., Traveling Wave Tube Characteristics for Finite Values of C, *Trans. I.R.E., PGED-1*, pp. 1-11, Aug., 1954.
29. Pierce, J. R., Traveling Wave Oscilloscope, *Electronics*, **22**, Nov., 1949.

The Field Displacement Isolator

By S. WEISBAUM and H. SEIDEL

(Manuscript received February 7, 1956)

A nonreciprocal ferrite device (*field displacement isolator*) has been constructed with reverse to forward loss ratios of about 150 in the region from 5,925 to 6,425 mc/sec. The forward loss is of the order of 0.2 db while the reverse loss is 30 db. These results are obtained by using a single ferrite element, spaced from the sidewall of the guide. The low forward loss suggests the existence of an electric field null at the location of a resistance strip on one face of the ferrite. We discuss the various conditions, derived theoretically, under which the electric field null may be obtained and utilized. Furthermore, a method of scaling is demonstrated which permits ready design to other frequencies.

I. INTRODUCTION

The need for passive nonreciprocal structures has long been recognized.¹ In the microwave field, Hogan's gyrator² paved the way for an increasingly important class of such devices. The isolator, in particular, has emerged as one of the more useful ferrite components. It performs the function, as its name implies, of isolating the generator from spurious mismatch effects of the load. Unlike lossy pads, which consume generator power, the isolator provides a unidirectionally low loss transmission path.

A. G. Fox, S. E. Miller and M. T. Weiss³ have pointed out that nonreciprocal ferrite devices may exploit any of the following waveguide effects:

1. Faraday rotation
2. Gyromagnetic resonance
3. Field displacement
4. Nonreciprocal phase shift

In the present paper we shall discuss an isolator, based upon the field displacement effect, which was developed to meet the following requirements for a proposed microwave relay system (5,925–6,425 mc/sec):

1. Forward loss 0.2 db

2. Reverse loss 20 db
3. Return loss 30 db

The field displacement isolator employs an ordinary rectangular waveguide and requires no specialized adaptation to the rest of the guide system. It is relatively compact and does not require excessive magnetic fields. In contrast to the field displacement structure of Reference 3, in which a symmetrically disposed pair of ferrite slabs is used, the present unit (see Fig. 1) contains only a single slab. Other differences of a more substantial nature may be noted — in the present case the slab is displaced from the guide wall, it occupies a partial height of the waveguide, and it employs a novel disposition of the absorption material on one face. These features result in a broadband device.

In the analysis presented in this paper the isolator field characteristics for a full height slab are determined by exact solution of Maxwell's equations, as opposed to the "point-field" perturbation approximation used in Reference 3. An exact solution of the partial height geometry of the experimental device would be exceedingly difficult to obtain. However, such a solution did not appear to be essential for this investigation since good correspondence has been obtained between the experimental results and the idealized full height slab calculations.

The following performance of the isolator was obtained from 5,925-6,425 mc/sec:

1. Forward loss ~ 0.2 db

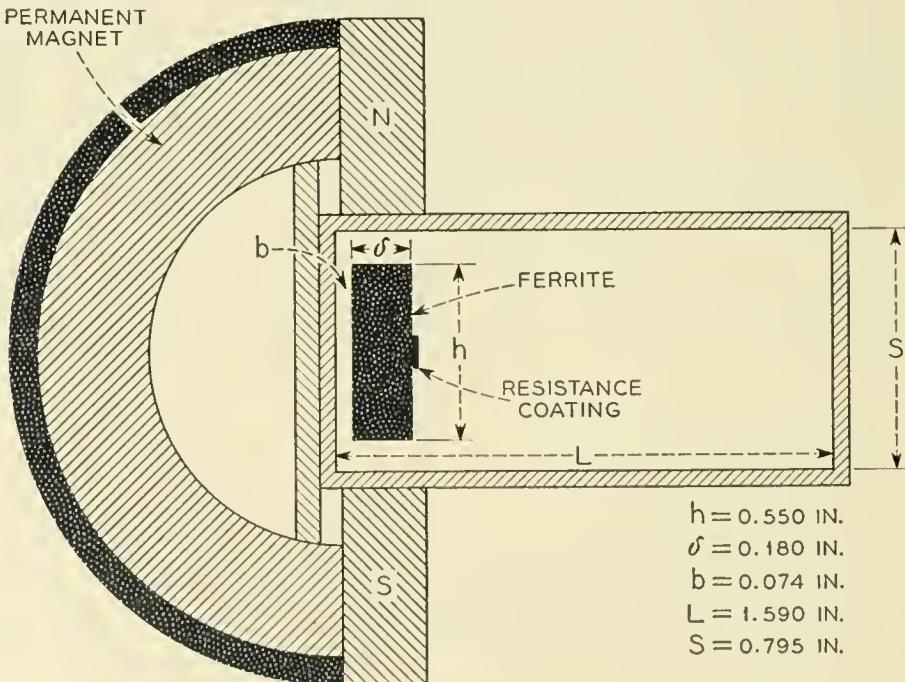


Fig. 1 — Field displacement isolator.

2. Reverse loss ~ 30 db
3. Return loss ~ 30 db

The extremely low forward loss strongly suggested the existence of an electric field null in the plane of the resistance material. Consequently, a theoretical investigation of the null condition was made and a set of criteria established for the existence and utilization of the null. (E. H. Turner⁴ independently developed the same null conditions.) An extension of the analysis leads directly to a set of scaling laws which permits the ready design of isolators of comparable performance at other frequency bands.

II. DESCRIPTION OF OPERATION

In Section IIA we will show how the "point-field" approach³ is used to predict the qualitative behavior of the structure and in Section IIB we will apply a more rigorous analysis to the determination of the optimum design parameters.

A. Qualitative

Prior to introducing the actual isolator configuration, we shall review some elementary properties both of the ferrite medium and of an unloaded rectangular waveguide. It is in terms of these properties that we can understand, in a qualitative sense, the interaction of an *rf* wave with a ferrite in such a waveguide. Since the behavior of a ferrite medium in the presence of a static magnetic field and a small *rf* field has been discussed in the literature⁵ the following resumé is not intended to be detailed. It is presented, however, to maintain continuity.

If a static magnetic field is applied to a ferrite medium the unpaired electron spins, on the average, will line up with the field. If now an *rf* magnetic field, transverse to the dc field, excites the spin system these electrons will precess, in a preferential sense, about the static field. The precession gives rise to components of transverse permeability at right angles to the *rf* magnetic field, leading to a tensor characterization of the medium. This tensor has been given by Polder⁵ and may be diagonalized in terms of circularly polarized wave components. Corresponding to the appropriate sense of polarization we use the designation + and -. When the polarization is in the same sense as the natural precessional motion of the spin system, gyromagnetic resonance occurs for an appropriate value of the static magnetic field. The scalar permeabilities μ_- and μ_+ are shown in Fig. 2 as functions of the internal static magnetic field as would be observed at an arbitrary frequency.

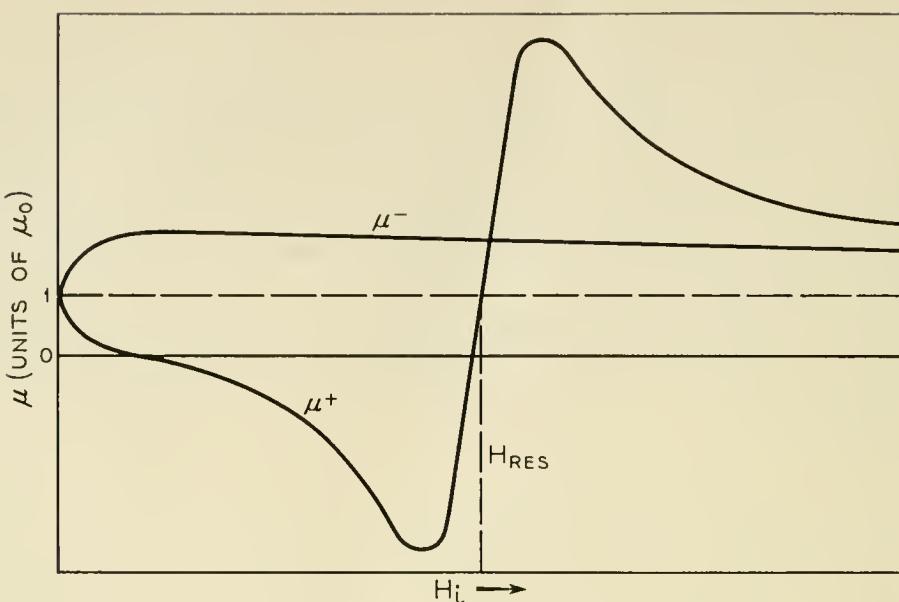


Fig. 2 — Permeability versus magnetic field.

Clearly, in employing a ferrite medium, we intend to use the basic difference between the scalar permeabilities μ_+ and μ_- . To this end we may exploit the fact that the magnetic field configuration at any given point in a rectangular waveguide is, in general, elliptically polarized. Traveling loops of magnetic intensity appear in Fig. 3 for the fundamental (TE_{10}) mode. At point P an observer sees a counterclockwise elliptically polarized magnetic intensity if the wave is traveling in the ($+y$) direction.* The propagating wave may be decomposed into two oppositely rotating circularly polarized waves of different amplitudes:

$$\text{Diagram of a large circle with a clockwise arrow} = \text{Diagram of a large circle with a counter-clockwise arrow} + \text{Diagram of a small circle with a clockwise arrow}$$

For propagation in the ($-y$) direction the *rf* polarization is reversed:

$$\text{Diagram of a large circle with a counter-clockwise arrow} = \text{Diagram of a large circle with a clockwise arrow} + \text{Diagram of a small circle with a counter-clockwise arrow}$$

Let us now consider the actual experimental configuration shown in Fig. 4 (the partial height geometry was chosen on an experimental basis, in that it gave VSWR considerably less than that for a full height ferrite slab). The precession of the spin magnetic moments is counterclockwise

* It is evident that a point converse to P exists symmetrically to the right of center. This is utilized in a double slab isolator which has been investigated by S. Weisbaum and H. Boyet, I.R.E., 44, p. 554, April, 1956.

looking along the direction of the dc magnetic field shown in Fig. 4. Since the major component of circular polarization for ($+y$) propagation is also counterclockwise the permeability will be less than unity for this direction of propagation. This occurs provided we are using small static fields, as is readily verified from Fig. 2. The permeability will be greater than unity for ($-y$) propagation. Physically, this is equivalent to energy being crowded out of the ferrite for ($+y$) propagation and to energy being crowded in, in the reverse direction. The electric field will thus be distorted as shown in a qualitative way in Fig. 5. The vertical dimension in this figure serves both to identify the guide configuration and to provide an ordinate for the electric field intensity.

The fields as shown in Fig. 5 merely represent a qualitative picture of the distributions in the guide and are not intended to be exact. There is no question, however, that the electric fields at the ferrite face are dif-

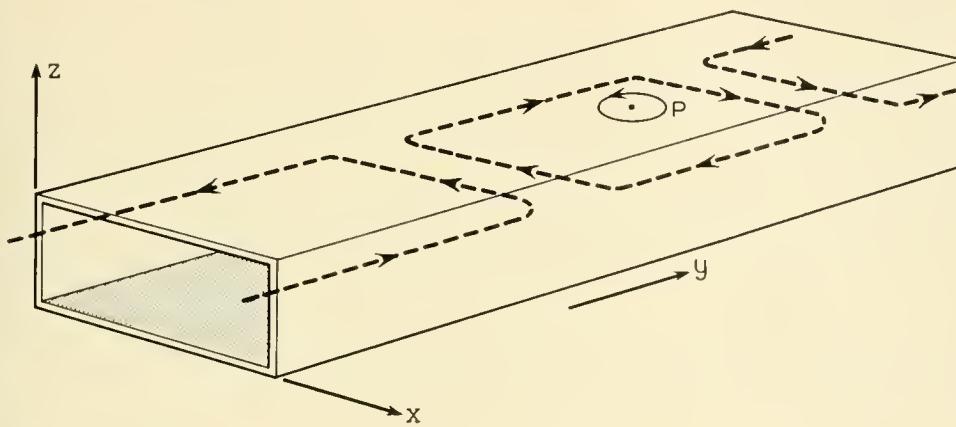


Fig. 3 — Magnetic field configuration — Dominant TE_{10} mode.

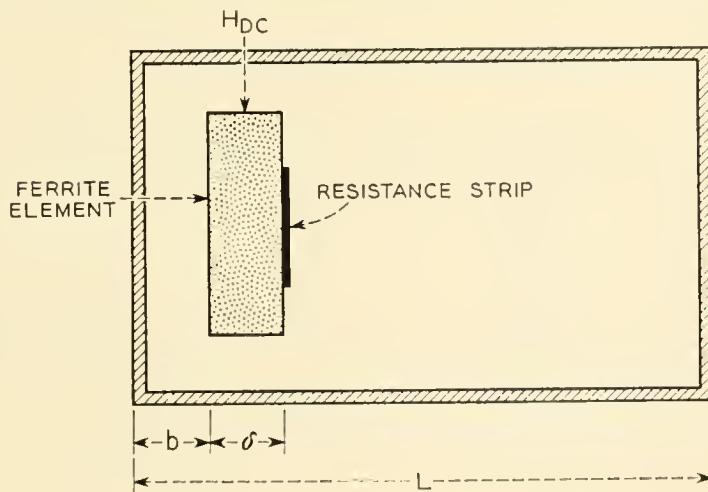


Fig. 4 — Experimental configuration.

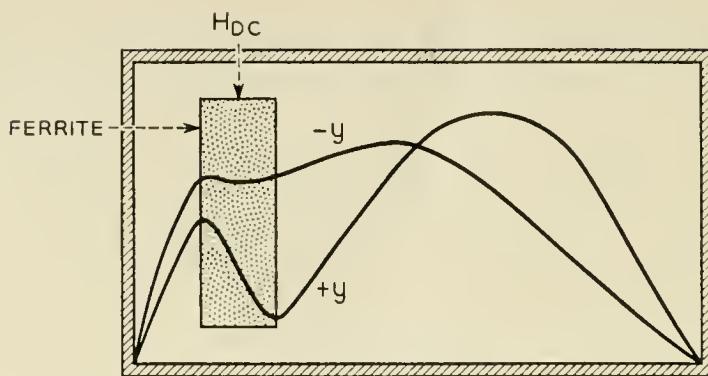


Fig. 5 — Electric field distortion.

ferent in magnitude corresponding to the two directions of propagation. Hence, if resistance material is placed at the interior face of the ferrite (see Fig. 1) we may expect to absorb more energy in one direction of propagation.

B. Analysis of Electric Field Null:-Full Height Ferrite

The description we have given in Section II A is based on a perturbation approach and does not take into account the higher order interaction effects of the ferrite and the propagating wave. In this section we consider an analysis of the idealized case, namely that of a full height ferrite slab, and impose the condition of an electric field null at the face of the ferrite for the forward direction of propagation. While this too does not represent the true experimental situation, we believe it to be a better approximation than the "point-field" perturbation viewpoint.

The fields of the various regions shown in Fig. 6 are described as follows:

$$E_z^{(1)} = \sin \alpha_1 x$$

$$E_z^{(2)} = Ae^{-i\alpha_2 x'} + Be^{i\alpha_2 x'} \quad \text{where } x' = x - a \quad (\text{II} - 1)$$

$$E_z^{(3)} = V \sin \alpha_1 x'' \quad \text{where } x'' = x - L$$

where

α_j = transverse wave number in the j^{th} region

a = transverse dimension from narrow wall to ferrite face

L = broad waveguide dimension

x = variable dimension along broad face

z = height variable

A, B, V = constants

Setting up the wave equation, there results

$$\alpha_2^2 = K^2 \left[\frac{\varepsilon_r}{\mu_r} (\mu_r^2 - k_r^2) - 1 \right] + \alpha_1^2 \quad (\text{II} - 2)$$

where μ_r and k_r are the relative diagonal and off-diagonal terms of the Polder tensor, respectively, K is the free space wave number and ε_r is the relative dielectric constant.

$$\mu_r = 1 + \frac{4\pi M_s \gamma \omega_0}{\omega_0^2 - \omega^2}$$

$$k_r = \pm \frac{4\pi M_s \gamma \omega}{\omega_0^2 - \omega^2}$$

$$\gamma = 2.8 \times 10^6 \text{ cycles/sec/oersted}$$

$4\pi M_s$ = saturation magnetization in gauss

H_0 = static magnetization in oersteds

$$\omega_0 = \gamma H_0$$

$$K = \frac{2\pi}{\lambda}$$

The following transcendental equation results from satisfying the boundary conditions on E and H :⁶

$$\frac{\tan \alpha_1 a [\mu_r \alpha_2 + k_r \beta \tan \alpha_2 \delta] + (\mu_r^2 - k_r^2) \alpha_1 \tan \alpha_2 \delta}{(\beta^2 - K^2 \mu_r \varepsilon_r) \tan \alpha_1 a \tan \alpha_2 \delta + \alpha_1 (\mu_r \alpha_2 - k_r \beta \tan \alpha_2 \delta)} + \frac{\tan \alpha_1 b}{\alpha_1} = 0 \quad (\text{II} - 3)$$

where β is the propagation constant.

The minimum nontrivial value of α_1 causing a null to appear at the ferrite face is $\alpha_1 = \pi/a$. Placing this value in (II - 3) produces the following transcendental equation for the null:

$$\frac{\frac{\pi}{a} (\mu_r^2 - k_r^2) \tan \alpha_2 \delta}{\mu_r \alpha_2 - k_r \beta \tan \alpha_2 \delta} + \tan \alpha_1 b = 0 \quad (\text{II} - 4)$$

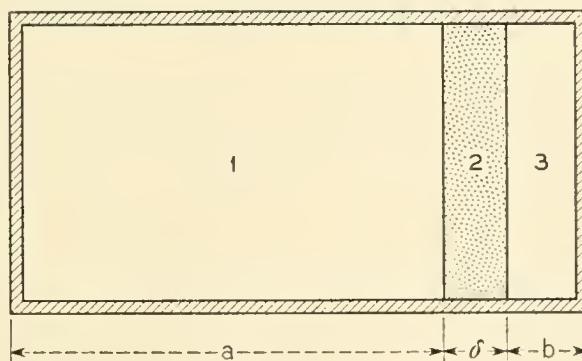


Fig. 6 — Full height geometry.

where $b = L-a-\delta$. Equation (II — 4) demonstrates that the null condition is nonreciprocal since, in general, the solutions differ for k_r positive and k_r negative. The quantity k_r has the same sign as the direction of the dc magnetic field; reversing the sign of k_r is equivalent to reversing the direction of propagation.

A numerical analysis of equation (II — 4) has led to the conclusion that the null condition is most broadband when $|\mu_r| < |k_r|$.* We use the criterion $|\mu_r| = |k_r|$ to determine a critical magnetic field:

$$H_c = \frac{\omega}{\gamma} - 4\pi M_s \quad (\text{II} - 5)$$

Clearly we require $\omega/\gamma > 4\pi M_s$ for physically realizable solutions. The saturation magnetization ($4\pi M_s$) is subject to the following:

1. A choice of too large a $4\pi M_s$ might create a mode problem and in addition will not satisfy the limit on $4\pi M_s$ implied in (II — 5).
2. $4\pi M_s$ must be sufficiently large so that the field needed to make $|\mu_r| < |k_r|$ not be excessive.
3. $\gamma\sqrt{H(H + 4\pi M)}$ (this being the slab resonance frequency for small slab thickness⁷) must be sufficiently far from the operating frequency to avoid loss due to resonance absorption. In addition, this condition improves the frequency insensitivity of the null.

Further analytic considerations are presented in Section IV.

III. EXPERIMENTAL DESIGN CONSIDERATIONS

Aside from the partial height nature of the slab, there are two other basic factors in the experimental situation which are not present in the analysis of Section IIB (see also Section IV). First, the ferrite has both finite dielectric and magnetic loss. Second, higher order modes may be present. These deviations from the simplified analysis are by no means trivial and it would not be surprising if one found a considerable modification of the analytic results. As it turns out, there are broad areas of general agreement between the theoretical and experimental results and in no case examined here does one find a basic inconsistency. In considering the various parameters which must be adjusted to optimize the broadband performance of the isolator we will point out, where possible, how the theoretical results are modified by the factors mentioned above. The parameters of interest are:

* This is partially evident from equation II — 4. The quantity $|\mu_r|$ must be less than $|k_r|$ if the angle $(\alpha_1 b)$ is to be small and in the first quadrant. Second quadrant solutions cause the guide cross section to be excessively large, with attendant higher mode complication.

- A. The saturation magnetization ($4\pi M_s$) and the applied magnetic field (H_{DC}).
- B. The ferrite height.
- C. The thickness (δ) of the ferrite and its distance (b) from the nearest sidewall.
- D. The placement of the resistance material and its resistivity (ρ).
- E. The length of the ferrite (ℓ).

A. $4\pi M_s$ and H_{DC}

Theoretically, minimum forward loss occurs with a true null at the face of the loss film and has been given in the condition $|\mu_r| < |k_r|$. Although this inequality is required in the full height slab analysis, experiment (Fig. 7) indicates the low loss region to be so broad as to extend well into the low field, or $|\mu_r| > |k_r|$ region.

There is inherent loss in the ferrite so that a more accurate statement of the bandwidth of operation is that in which the losses in the film are of equal order to the ferrite losses at the band edges. Even discounting ferrite losses, it will be shown in Section IV that we have a good analytic basis for the observed broadness of the low loss region. In general, therefore, we need not be as restrictive as the null analysis of Section II B would imply. It is not surprising then that optimum operation actually occurs in the region $|\mu_r| > |k_r|$. There are several reasons why this may be so:

1. Shift of operation occurs due to the partial height nature of the ferrite slab.
2. Reverse loss has a peak in the low field region, requiring a compromise of low forward loss and high reverse loss for best isolation ratios (see Fig. 8).
3. Optimum compromise between low ferrite loss and low film loss must be made.

The internal magnetic field, determining $|\mu_r|$ and $|k_r|$, differs from the applied field by the demagnetization of the ferrite slab. Although not ellipsoidal, it may nonetheless be considered to have an average demagnetization which has been computed, for this case, to be 460 oersteds. A further complication in knowledge of the internal field is the proximity effect of the pole pieces. This latter correction was obtained experimentally and, all in all, it was determined that the internal field for optimum operation was of the order of 300 oersteds. For the given ferrite and the range of frequency of operation, this internal field corresponds to the condition that $|\mu_r| > |k_r|$, as stated above.

Taking all effects into account, it was found that optimum permanent magnet design occurred for an air gap field of 660 oersteds.

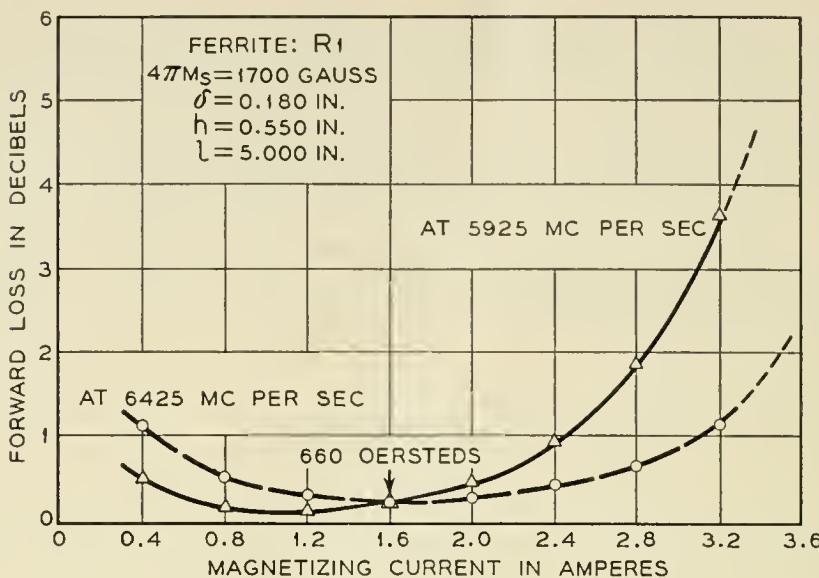


Fig. 7 — Forward loss versus magnetizing current.

Using the experimental values $4\pi M_s = 1,700$ gauss and internal magnetic field = 300 oersteds, the frequency at which ferromagnetic resonance occurs was estimated to be about 2200 mc/sec. This value is sufficiently far from our operating range (5,925–6,425 mc/sec) that we would expect a negligible loss contribution due to resonance absorption. This is confirmed by the low forward loss actually observed.

B. Ferrite Height

We have already pointed out that when the ferrite height is reduced from full height a more reasonable VSWR is obtained. This is due to the fact that we have relieved the stringent boundary requirements at the

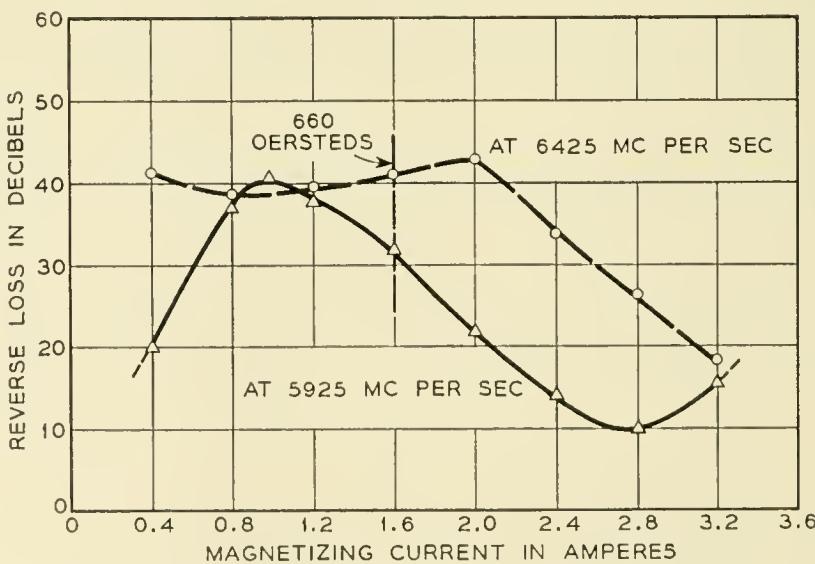


Fig. 8 — Reverse loss versus magnetizing current.

top and bottom faces of the ferrite and approach, in a sense, a less critical rod type geometry. A ferrite height of 0.550" gave a VSWR ~ 1.05 over the band. With full height slabs (0.795"), VSWR values as high as 10:1 have been observed for typical geometries.

C. δ and b

Experimentally, we have examined various ferrite thicknesses at different distances from the sidewall until optimum broadband performance was obtained. Table I shows the ferrite distance from the wall which gave the best experimental results (highest broadband ratios, low forward loss, high reverse loss) for each thickness δ of one of the BTL materials. It is interesting to note that the empirical quantity $\delta + b/2 -$

TABLE I

δ (mils)	b (mils)	$\delta + \frac{b}{2}$ (mils)	t (mils)	$\delta + \frac{b}{2} - 2t$ (mils)
201	11	206.5	3	200.5
189	35	206.5	3	200.5
186	42	207.0	3	201.0
176	65	208.5	3	202.5
189	42	210.0	6	198.0

$2t$, where t is the thickness of the resistive coating, is very nearly constant (within a few mils) for the stated range of δ and for this type of design.*

In Section IV a theoretical calculation using the null condition at 6175 mc/sec for a full height ferrite gives

$$\delta = 180 \text{ mils}$$

$$b = 38.7 \text{ mils}$$

so that $\delta + b/2 = 199.3$ mils. In the theoretical case t is assumed to be very small. It will be noted that the theoretical result for $\delta + b/2$ (with small t) agrees quite well with the experimental $\delta + b/2 - 2t$. The question of the possible physical significance of this quantity is being investigated.

D. Placement of Resistance Material and Choice of Resistivity

The propagating mode with a full height ferrite slab is of a TE_0 variety, the zero subscript indicating that no variation occurs with re-

* In one design of the isolator we used a General Ceramics magnesium manganese ferrite with $\delta = 0.180"$, $b = 0.074"$ and $t = 0.009"$ so that $\delta + b/2 - 2t = 199$ mils, in good agreement with Table I.

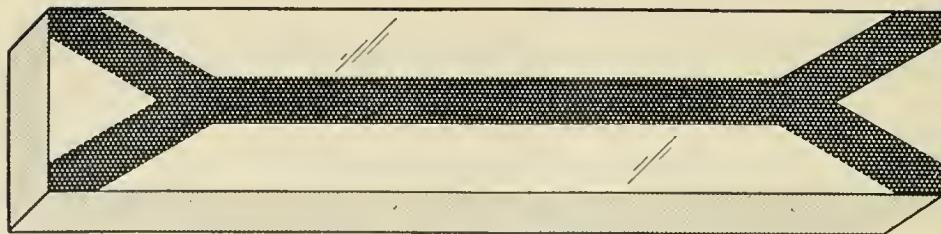


Fig. 9 — Distribution of small tangential electric fields at interior ferrite face.

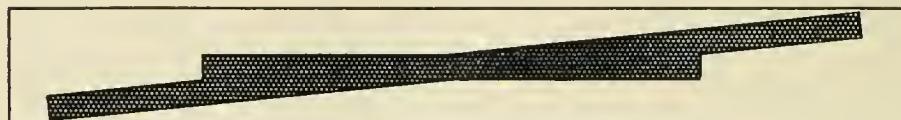


Fig. 10 — Resistance configuration.

spect to height. A field null in this construction therefore extends across the entire face of the full height ferrite and all of this face is then "active" in the construction of an isolator. This field situation no longer accurately applies to the partial height slab. The departure of the ferrite from the top wall creates large fringing fields extending from the ferrite edges, and large electric fields may exist tangential to the ferrite face close to these edges. We would therefore expect the null condition to persist only in a small region about the vertical center of the ferrite face. We may, however, also expect longitudinally fringing modes (TM-like) to be scattered at the input edge of the ferrite slab so that a longitudinal field maximum will exist at the central region of the ferrite. However, this is a higher mode, so that this maximum decays rapidly past the leading edge.

Considering all the effects, the distribution of small tangential electric fields at the ferrite face may be expected to appear as shown in Fig. 9. Experimentally, we have utilized this low loss region and have avoided the decay region of the higher TM-like modes by using the resistance configuration shown in Fig. 10. The resistivity is uniform and about 75

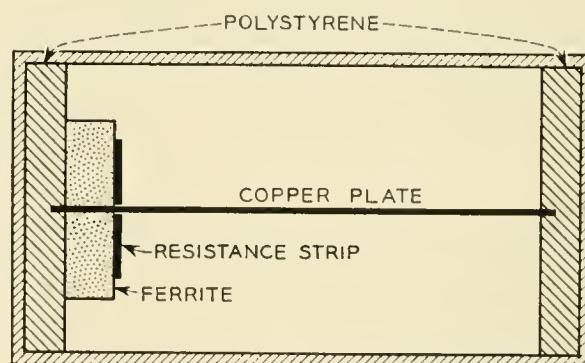


Fig. 11 — Elimination of longitudinal components.

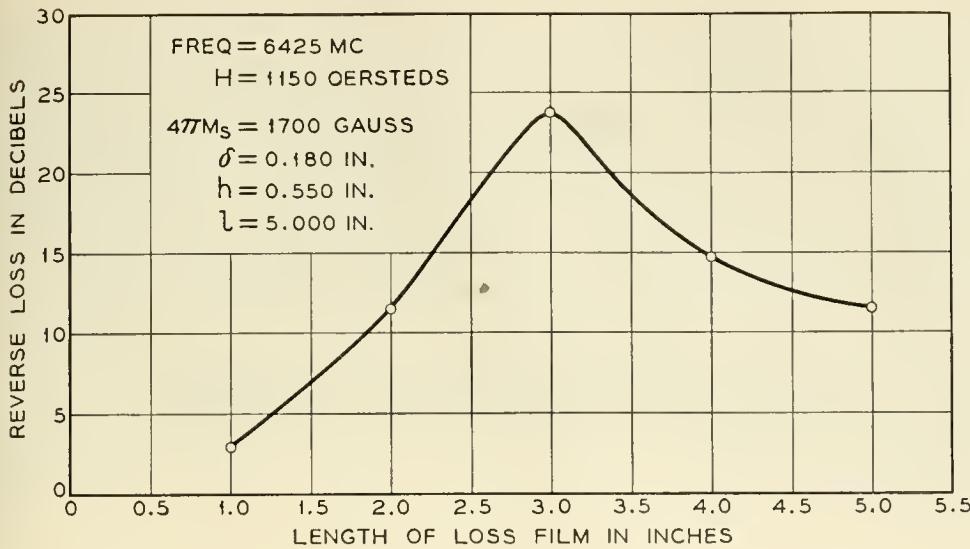


Fig. 12 — Attenuation versus length of resistance strip.

ohms/square. Variations of about ± 30 ohms/square about this value result in little deterioration in performance.

Some further discussion of the perturbed dominant mode is of interest. We may think of the height reduction as primarily a dielectric discontinuity where we have effectively added a negative electric dipole density to a full height slab. Since this addition is smaller for the forward case (where there was initially a small electric field) than for the reverse case, we may expect the longitudinal components to be smaller for the forward propagating mode. The other type of longitudinal electric field,

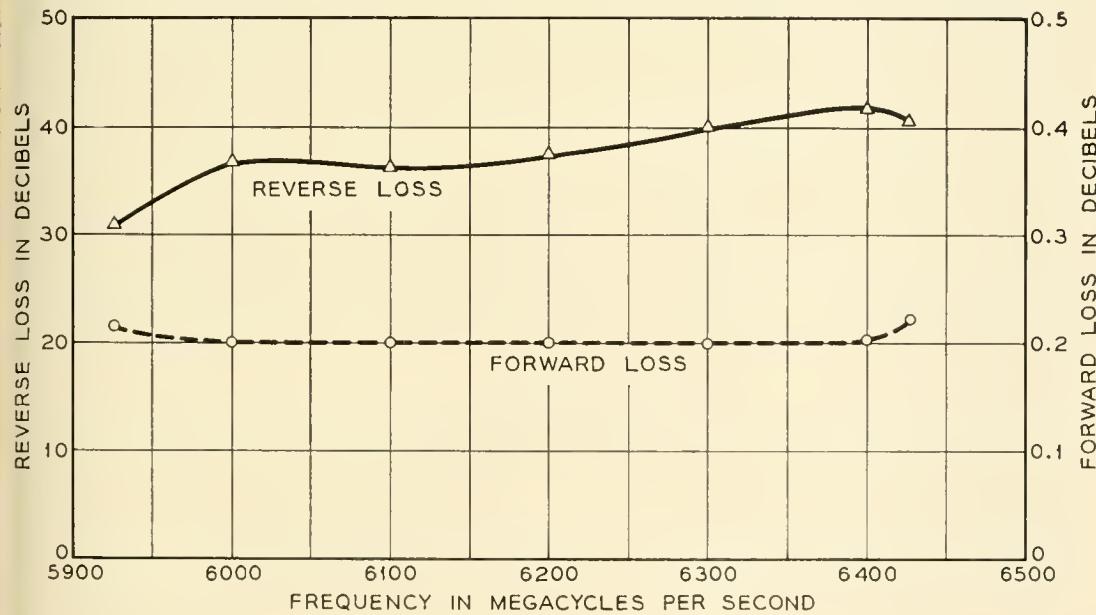


Fig. 13 — Loss versus frequency.

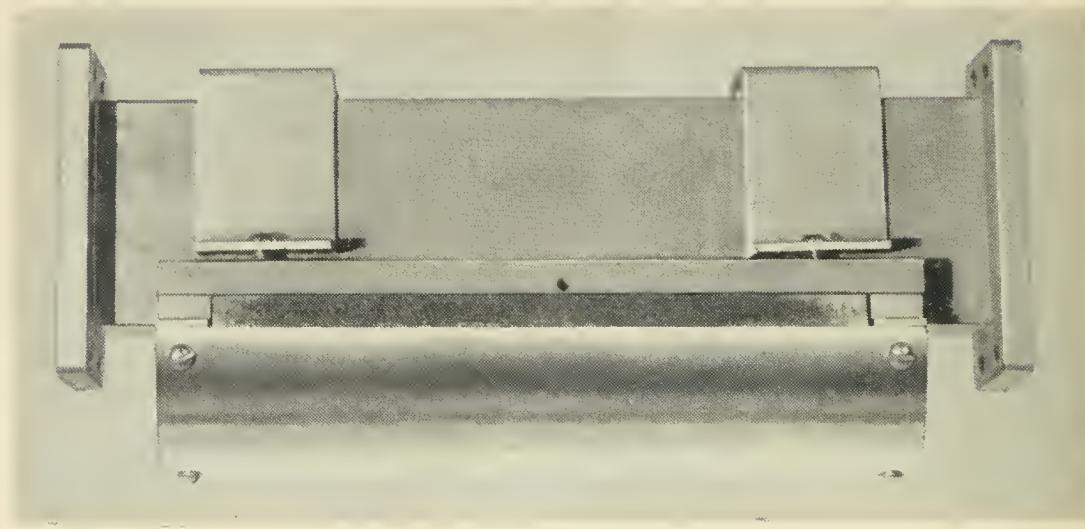


Fig. 14 — Isolator model.

which occurs due to the scattering of the TM-like longitudinal modes. decays rapidly and is not of consequence in an experiment now to be described. This experiment was designed to demonstrate the nonreciprocal nature of the longitudinal electric fields associated with the distorted dominant mode. It also shows that the existence of these components is significant as a loss mechanism for the reverse direction of propagation in the isolator. The geometry employed is shown in Fig. 11.

The copper plate was inserted to minimize longitudinal electric field components, and we may therefore expect to obtain less reverse loss than in the condition of its absence. The result of this experiment was that the reverse loss decreased from about 25 db* without the plate to 18 db with the plate. The forward loss was unaffected.

E. Determination of Length

Given a dominant mode distribution in a waveguide, attenuation will be a linear function of length, once this mode has been established. Consequently, one would expect that doubling the loss film length would double the isolator reverse loss. The isolator does not exhibit this behavior, however, as is illustrated in Fig. 12.

This occurrence might be explained by the appearance of still another longitudinal mode, peculiar in form to gyromagnetic media alone, which propagates simultaneously with the transverse electric mode, and is essentially uncoupled to the loss material. The maximum reverse loss

* This experiment was conducted with a different ferrite than that employed in the eventual design.

thus obtainable is limited by the scattering into this mode. The character of these singular modes will be discussed in a subsequent paper.

Results

The performance of the isolator as a function of frequency is shown in Fig. 13. Fig. 14 shows a completed model of the isolator.

IV. FURTHER ANALYSIS

While an exact characteristic equation is obtainable for the overall geometry of the full height isolator, including the lossy film, the expressions which result are sufficiently complex to be all but impossible to handle. However, if the resistance film is chosen to have small conductivity we may utilize a simple perturbation approach in which the field at the ferrite face is assumed to be unaffected by the presence of the loss film. A quantity η may then be defined* so that

$$\eta = \frac{|E_R|^2}{P} \quad (\text{IV-1})$$

For small conductance values η is proportional to attenuation to first order in either direction of propagation. E_R , in equation (IV-1), is the electric field adjacent to the film and P is the power flowing across the guide cross section. The loss in the ferrite material is not taken into account in this approximation, but it would naturally have a deteriorating effect on the isolator characteristics.

The ratio of the values of η corresponding to backward and forward direction of propagation defines the isolation ratio, given in db/db, for the limit of very small conductivity.

Fig. 15 shows a calculated curve of the forward value of η and Fig. 16 shows the backward case. The isolation ratio shown in Fig. 17 demonstrates surprisingly large bandwidth for values of the order of 200 db/db. Fig. 18 portrays propagation characteristics for both forward and backward power flows and provides the interesting observation, in conjunction with Fig. 16, that peak reverse loss occurs in the neighborhood of $\lambda = \lambda_g$.

Fig. 19 is a plot of α_1 , the transverse wave number, over the frequency range. The flatness of the forward wave number means that the position of null moves very little with frequency across the band. Hence the lossless transmission in the forward direction is broadband. Since the forward and backward wave numbers have such radically different

* See Appendix

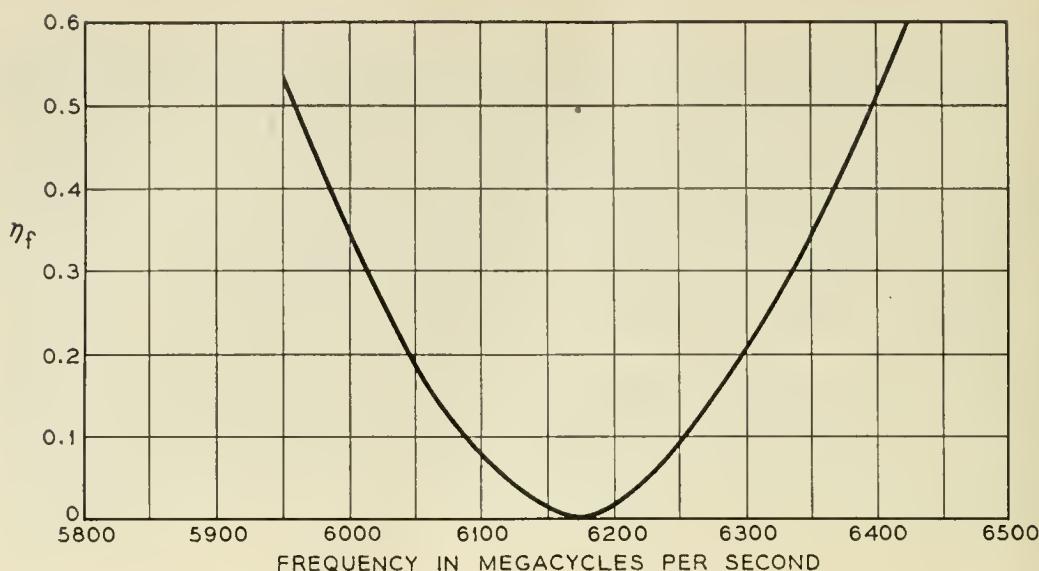


Fig. 15 — Relative attenuation — forward direction

rates of variation, a simple adjustment of parameters may be made to cause the forward null and maximum reverse attenuation to appear at the same frequency, resulting in an optimum performance.

The occurrence of the reverse maximum loss in the region of $\lambda = \lambda_g$ may roughly be explained as follows. As the transverse air wave number decreases, the admittance of the guide, defined on a power flow basis, increases. The electric field magnitude distribution must therefore generally decrease in such a fashion as to cause the overall power flow to re-

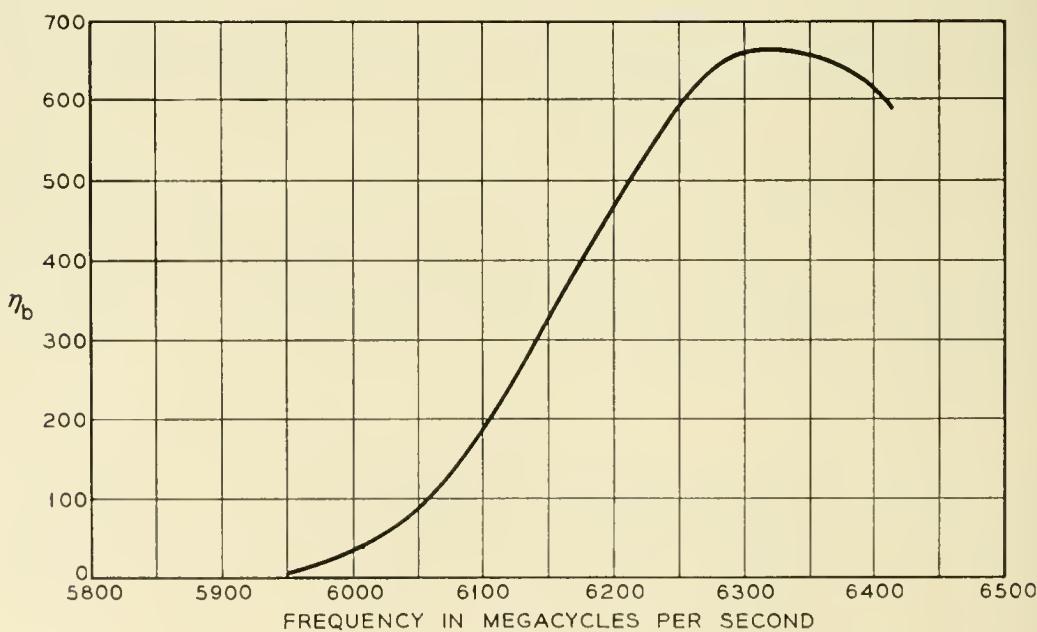


Fig. 16 — Relative attenuation — backward direction

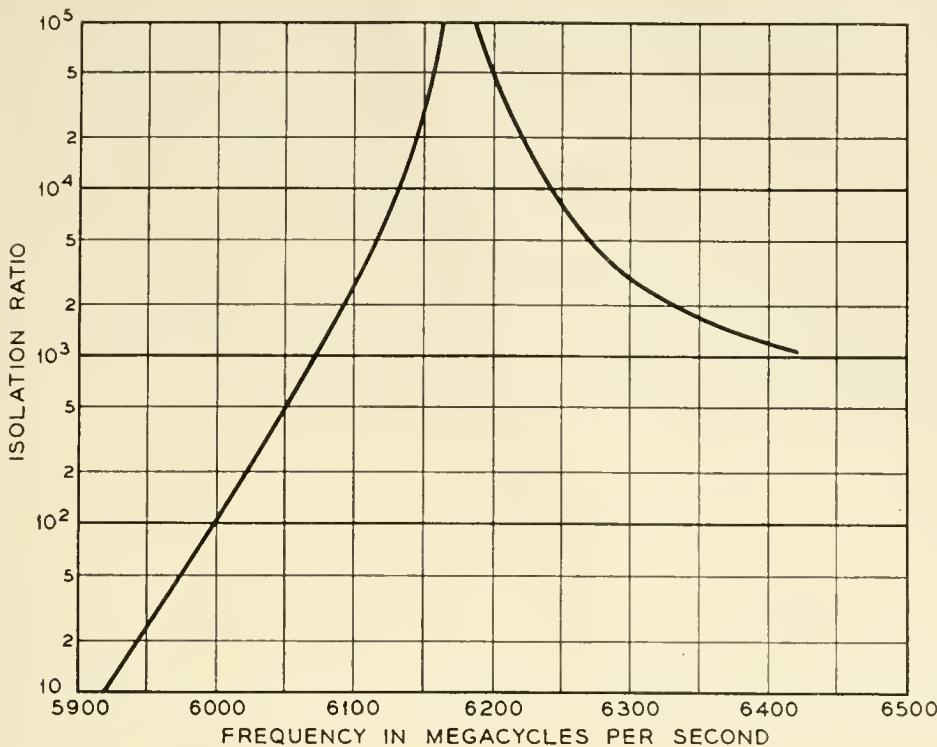


Fig. 17 — Ideal isolation characteristics.

main constant. On the other hand as the transverse air wave number decreases through real values, the electric field adjacent to the ferrite becomes relatively large. At $\lambda = \lambda_g$, the distribution is linear with relatively large dissipation at the ferrite face. As the transverse air wave number increases through imaginary values the distribution becomes exponential such that the field adjacent to the ferrite is always the maximum for the air region and the growth of the field at the face of the ferrite would not seem to be so great as formerly. One would therefore expect a maximum reverse loss somewhere in the region $\lambda = \lambda_g$.

The above considerations plus the transcendental equation for the null show consistency with the experimental design values which were:

$$\delta = 0.180"$$

$$L = 1.59$$

$$4\pi M_s = 1,700 \text{ gauss}$$

Using $H_{DC} = 600$ oersteds in the calculation we obtain the spacing from the guide wall $b = 0.0387"$. The fact that we used 600 oersteds for the full height slab calculation as opposed to the internal field of 300 oersteds found experimentally for the partial height slab should not be a source of confusion. It has been indicated earlier that the peak reverse loss shifts

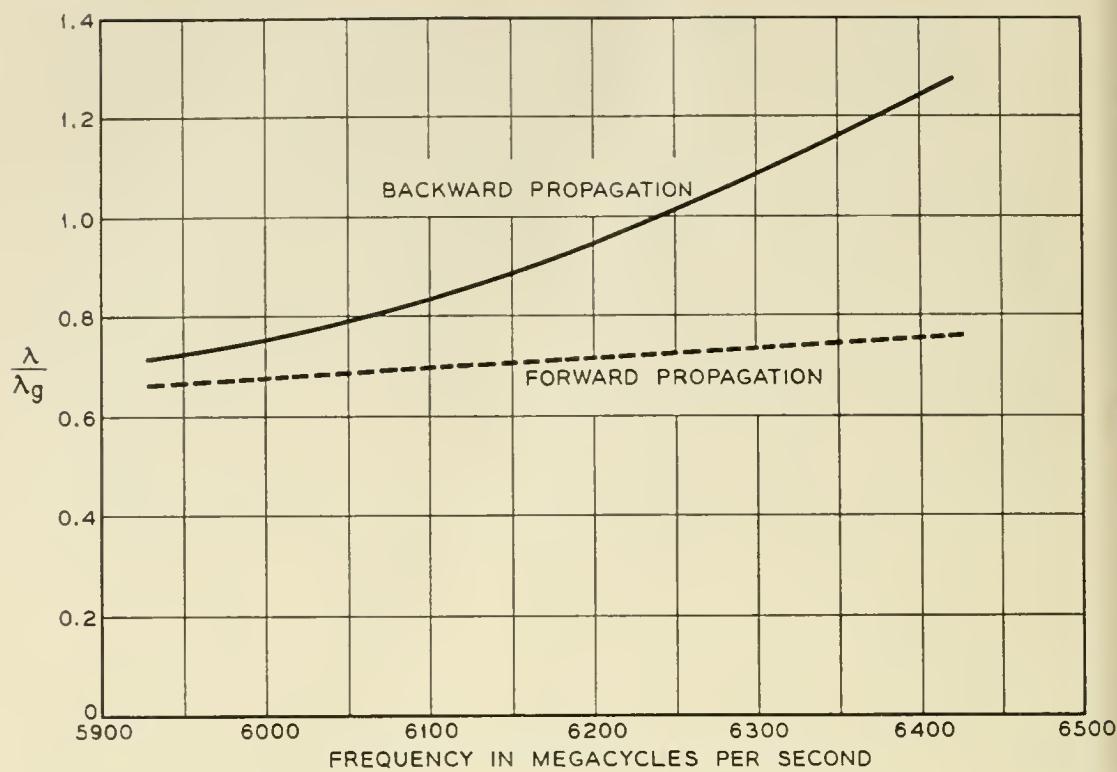


Fig. 18 — Ferrite isolator characteristics.

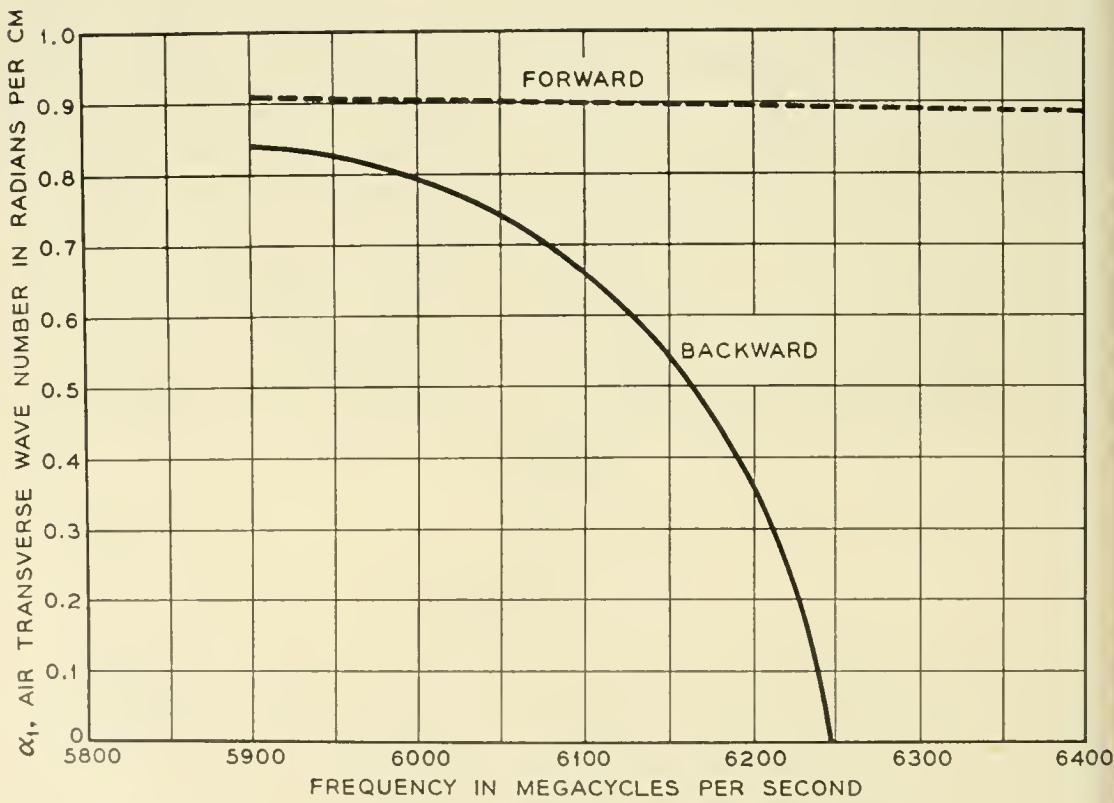


Fig. 19 — Transverse characteristics of a ferrite isolator

with ferrite height reduction. It is not inconsistent therefore to choose 600 oersteds for the full height analysis in contrast to the value determined from the experiment.

V. SCALING

Once the optimum set of parameters has been decided upon for a given frequency range (e.g., 5,925–6,425 me/sec, $\delta = 0.180''$, $b = 0.074''$, $\ell = 5''$, $h = 0.550''$, $4\pi M_s = 1,700$ gauss, $H_{DC} = 660$ oersteds) it is a simple matter to scale these parameters to other frequency ranges. From Maxwell's equations:

$$\text{Curl } H = i\omega \epsilon E + gE$$

$$\text{Curl } E = -i\omega T \cdot H$$

where T is the permeability tensor, and g is the conductivity in mhos/meter. The first of Maxwell's equations suggest that frequency scaling may be accomplished by permitting both the curl and the conductance to grow linearly with respect to frequency. The curl, which is a spatial derivative operator, may be made to increase appropriately by shrinking all dimensions by a $1/\omega$ factor, which will keep the field configuration the same in the new scale.

Having imposed this condition on the first equation we must satisfy the second of Maxwell's equations by causing T to remain unchanged with frequency. T is a tensor given as follows for a cartesian coordinate system:

$$T = \begin{pmatrix} \mu_r & ik_r & 0 \\ -ik_r & \mu_r & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (\text{V}-1)$$

for a magnetizing field in the z direction. The components may be expanded in the following fashion:

$$\mu_r = \frac{1 + 4\pi \left(\frac{\gamma M_s}{\omega}\right)\left(\frac{\gamma H}{\omega}\right)}{\left(\frac{\gamma H}{\omega}\right)^2 - 1} \quad (\text{V}-2)$$

$$k_r = \frac{4\pi \left(\frac{\gamma M_s}{\omega}\right)}{\left(\frac{\gamma H}{\omega}\right)^2 - 1}$$

where $4\pi M_s$ is the saturation magnetization in gauss and γ is the magnetomechanical ratio. The Polder tensor evidently remains unchanged if M_s and H are both scaled directly with frequency.

Since the field distributions are assumed unchanged relative to the scale shift, normal and tangential E and H field components continue to satisfy the appropriate boundary equalities at interfaces. Then, invoking the uniqueness theorem, the guide characteristics are only as presumed and the model has been properly scaled as a function of frequency.

The scaling equations are:

$$\begin{aligned} d_1 &= \frac{\omega_2}{\omega_1} d_2 \\ g_1 &= \frac{\omega_1}{\omega_2} g_2 \\ M_{s_1} &= \frac{\omega_1}{\omega_2} M_{s_2} \\ (H_0)_1 &= \frac{\omega_1}{\omega_2} (H_0)_2 \end{aligned} \quad (\text{V}-3)$$

where d is any linear dimension.

CONCLUSION

An isolator with low forward loss and high reverse loss can be constructed by a proper choice of parameters. Once a suitable design has been reached the scaling technique can be used to reach a suitable design for other frequencies.

As yet, a theoretical analysis of this problem has been carried out only for a full height ferrite.

ACKNOWLEDGMENT

We would like to thank F. J. Sansalone for his assistance in developing the field displacement isolator. We would also like to thank Miss M. J. Brannen for her competent handling of the numerical computations.

APPENDIX

It is desirable to establish an isolator figure of merit. A simple quantity characterizing the isolator action is the normalized rate of power

loss in the resistive strip, for an idealized ferrite, in the low conductive limit of such a strip. Let

$$\eta = \frac{|\mathbf{E}_r|^2}{P}$$

where η is the appropriate quantity, E_r is the field at the resistance, and P is the total power flow across the guide cross-section. This figure of merit is related to the rate of change of the attenuation constant (A) with respect to strip conductance in the following manner:

$$\frac{dA}{dg} = 0.04343\eta h \text{ (db)(ohms)/cm}$$

where h is the fractional height of the loss strip, and g is the reciprocal of the surface resistivity in ohms/square.

The total power flow may be divided into integrations of the Poynting vector over the three regions of the guide cross-section. The following results are obtained normalized to $E_r = \sin \alpha_1 a$:

Region 1: $0 \leq x \leq a$

$$P_y^{(1)} = \frac{\beta}{2\omega\mu_0} \left(a - \frac{\sin 2\alpha_1 a}{2\alpha_1} \right)$$

Region 2: $a \leq x \leq a + \delta$

$$\begin{aligned} P_y^{(2)} = & \frac{\beta}{2\omega\mu_0} \left(\frac{\mu_r \delta}{\mu_r^2 - k_r^2} (d_1^2 + d_2^2) + \frac{\sin 2\alpha_2 \delta}{2\alpha_2 (\mu_r^2 - k_r^2)} \right. \\ & \cdot \left[\mu_r (d_1^2 - d_2^2) - \frac{k_r}{\beta} \alpha_2 (2d_1 d_2) \right] + \frac{1 - \cos 2\alpha_2 \delta}{2\alpha_2 (\mu_r^2 - k_r^2)} \\ & \cdot \left. \left[\mu_r (2d_1 d_2) + \frac{k_r \alpha_2}{\beta} (d_1^2 - d_2^2) \right] \right) \end{aligned}$$

Region 3: $a + \delta \leq x \leq L$

$$P_y^{(3)} = \frac{\beta}{2\omega\mu_0} \left(b - \frac{\sin 2\alpha_1 b}{2\alpha_1} \right) \left(\frac{d_1 \cos \alpha_2 \delta + d_2 \sin \alpha_2 \delta^2}{\sin \alpha_1 b} \right)$$

where

$$d_1 = \sin \alpha_1 a$$

and

$$d_2 = \frac{1}{\mu_r \alpha_2} [(\mu_r^2 - k_r^2) \alpha_1 \cos \alpha_1 a + k_r \beta \sin \alpha_1 a]$$

BIBLIOGRAPHY

1. Tellegen, B. D. H., Philips Res. Rep., **3**, 1948.
2. Hogan, C. L., B.S.T.J. **31**, 1952.
3. Fox, A. G., Miller, S. E., and Weiss, M. T., B.S.T.J. **34**, p. 5., Jan. 1955.
4. Turner, E. H., URSI Michigan Symposium on Electromagnetic Theory, June, 1955.
5. Polder, D., Phil. Mag., **40**, 1949.
6. Lax, B., Button, K. J., Roth, L. M., Tech Memo No. 49, M.I.T. Lincoln Laboratory, Nov. 2, 1953.
7. Kittel, C., Phys. Rev., **73**, 1948.

Transmission Loss Due to Resonance of Loosely-Coupled Modes in a Multi-Mode System

By A. P. KING and E. A. MARCATILI

(Manuscript received January 17, 1956)

In a multi-mode transmission system the presence of spurious modes which resonate in a closed environment can produce an appreciable loss to the principal mode. The theory for the evaluation and control of this effect under certain conditions has been derived and checked experimentally in the particularly interesting case of a TE₀₁ transmission system, where mode conversion to TE₀₂, TE₀₃ . . . is produced by tapered junctons between two sizes of waveguide.

INTRODUCTION

In a transmission system, the presence of a region which supports one or more spurious modes can introduce a large change in the transmission loss of the principal mode when the region becomes resonant for one of the spurious modes. This phenomenon can occur even when the mode conversion is low and the waveguide increases in cross section smoothly to a region which supports more than one mode. In general, the conditions required to resonate the various spurious modes are not fulfilled simultaneously and, in consequence, interaction takes place between the principal mode and only one of the spurious modes for each resonating frequency. Under these conditions the resonating environment can be visualized as made of only two coupled transmission lines, one carrying the desirable mode and the other the spurious one. This simplification makes it possible to calculate the transmission loss as a function of (1) the coefficient of conversion between the two modes and (2) the attenuation of the modes in the resonating environment. The theory has shown good agreement with the measurement of transmission loss of the TE₀₁ mode in a pipe wherein a portion was tapered to a larger diameter which can support the TE₀₂ mode.

TRANSMISSION LOSS OF A WAVEGUIDE WITH A SPURIOUS MODE RESONATING REGION

Let us consider a single-mode waveguide connected to another of different cross-section that admits two modes. Since these two modes are orthogonal, the junctions may be considered as made of three single-mode lines connected together, provided we define the elements of the scattering matrix properly. The three modes, or lines in which they travel, are indicated by the subscripts 0, 1, and 2, as shown in Fig. 1. If a_0, a_1, a_2 and b_0, b_1, b_2 are the complex amplitudes of the electric field of the incident and reflected waves respectively, then

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = [S] \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

where

$$[S] = \begin{bmatrix} \Gamma_{00} & \Gamma_{01} & \Gamma_{02} \\ \Gamma_{01} & \Gamma_{11} & \Gamma_{12} \\ \Gamma_{02} & \Gamma_{12} & \Gamma_{22} \end{bmatrix} \quad (1)$$

is the scattering matrix.¹

This specific type of change of cross section may be treated as a three-port junction.

Now, if a length ℓ of a two mode waveguide is terminated symmetrically at both ends with a single mode waveguide (Fig. 2), each joint is described by the same matrix (1), and the connecting two mode wave guide has the following scattering matrix:

$$\begin{bmatrix} 0 & e^{-j\theta_1} & 0 & 0 \\ e^{-j\theta_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{-j\theta_2} \\ 0 & 0 & e^{-j\theta_2} & 0 \end{bmatrix} \quad (1')$$

in which

$$j\theta_1 = \gamma_1 \ell = (\alpha_1 + j\beta_1) \ell$$

$$j\theta_2 = \gamma_2 \ell = (\alpha_2 + j\beta_2) \ell,$$

γ_1 and γ_2 are the propagation constants of modes 1 and 2.

¹ N. Marcuvitz, Waveguide Handbook, 10, M.I.T., Rad. Lab. Series, McGraw-Hill, New York, 1951, pp. 107-8.

Matrices 1 and 1' describe the system completely and from them, the transmission coefficient results,

$$\Gamma = \frac{b_0'}{a_0} = \Gamma_{01}^2 e^{-j\theta_1} \frac{A - \Gamma_{22}^2 e^{-j2\theta_2} A^* \left(1 + \left| \frac{\Gamma_{12}}{\Gamma_{22}} \right|^2 + \frac{\Gamma_{00}}{\Gamma_{01}} \frac{\Gamma_{02}^*}{|\Gamma_{22}|^2} \Gamma_{12} \right)^2}{[1 - (\Gamma_{12}^2 - \Gamma_{11}\Gamma_{22})e^{-j(\theta_1+\theta_2)}]^2 - (\Gamma_{11}e^{-j\theta_1} + \Gamma_{22}e^{-j\theta_2})^2} \quad (2)$$

where

$$A = 1 + \left(\frac{\Gamma_{02}}{\Gamma_{01}} \right)^2 e^{-j(\theta_2-\theta_1)}$$

A^* is the complex conjugate of A

Γ_{01}^* is the complex conjugate of Γ_{01}

Γ_{02}^* is the complex conjugate of Γ_{02}

Furthermore, let us make the following simplifying assumptions

$$\Gamma_{00} = 0 \quad (3)$$

$$|\Gamma_{02}| \ll 1 \quad (4)$$

$$\sum_{\beta=0}^{\beta=2} \Gamma_{\beta n} \Gamma_{\beta m}^* = \begin{cases} 1, & \text{if } m = n = 0, 1, 2 \\ 0, & \text{if } m \neq n \end{cases} \quad (5)$$

Equation (3) indicates that if in Fig. 1, lines 1 and 2 were matched, line 0 would also be matched looking toward the junction. Equation (4) states that almost all the transmission is made from 0 to 1, or that there is small mode conversion to the spurious mode 2. Equation (5) assumes that the transition is nondissipative. The first two conditions are fulfilled when the transition is made smoothly. The last is probably the most stringent one, especially if the transition is a long tapered waveguide section, but it is always possible to imagine the transition as lossless and attribute its dissipation to the waveguides.

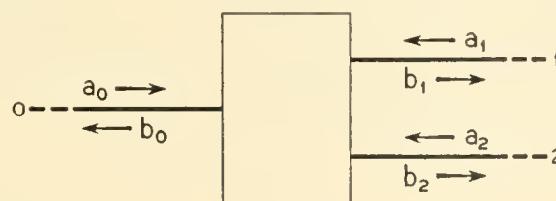


Fig. 1 — Schematic of a three-port junction.

From (2), (3), (4) and (5)

$$\Gamma = \frac{\Gamma_{01}^2 e^{-j\theta_1}}{|\Gamma_{22}|^2} \frac{1}{1 + \left| \frac{\Gamma_{11}}{\Gamma_{22}} \right| e^{-j\varphi}} \left\{ 1 - \frac{2 |\Gamma_{12}|^2 (1 - \cos \varphi)}{1 - [\Gamma_{22} e^{-j\theta_2} + \Gamma_{11} e^{-j\theta_1}]^2} \right\} \quad (6)$$

where

$$\Gamma_{11} = |\Gamma_{11}| e^{j\varphi_{11}}$$

$$\Gamma_{22} = |\Gamma_{22}| e^{j\varphi_{22}}$$

$$\varphi = \theta_1 - \theta_2 - \varphi_{11} + \varphi_{22}$$

In order to understand this expression physically, let us suppose first that there is no attenuation. The transmission coefficient Γ becomes 0 when the following equations are fulfilled simultaneously

$$\beta_2 \ell - \varphi_{22} = p\pi \quad p = 0, 1, 2, 3, \dots \quad (7)$$

and

$$\varphi = (2q + 1)\pi \quad q = 0, 1, 2, 3, \dots \quad (8)$$

The first of these equations states that the line carrying the feebly coupled mode must be at resonance, since this condition is satisfied when the electric length of this line is modified by a multiple of π radians. The second condition, (8), implies that both paths, in lines 1 and 2, must differ in such a way that electromagnetic waves coming through them must arrive in opposite phase at the end of the two-mode waveguide. This is quite clear if we think that, in order to get complete reflection, signals coming through lines 1 and 2 must recombine again with the same intensity and opposite phase. In order to get both modes with the same intensity, the converted mode must be built up through resonance; the opposite phase is obtained by an appropriate electric length adjustment. When attenuation is present, Γ will not be 0, and conditions (7) and (8) for minimum transmission are modified only slightly if the

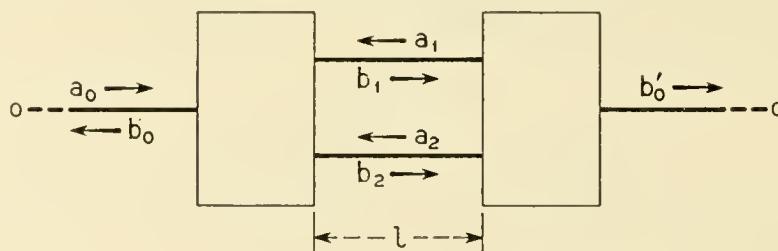


Fig. 2 — Schematic of a two-mode waveguide terminated symmetrically on each side with a single-mode waveguide.

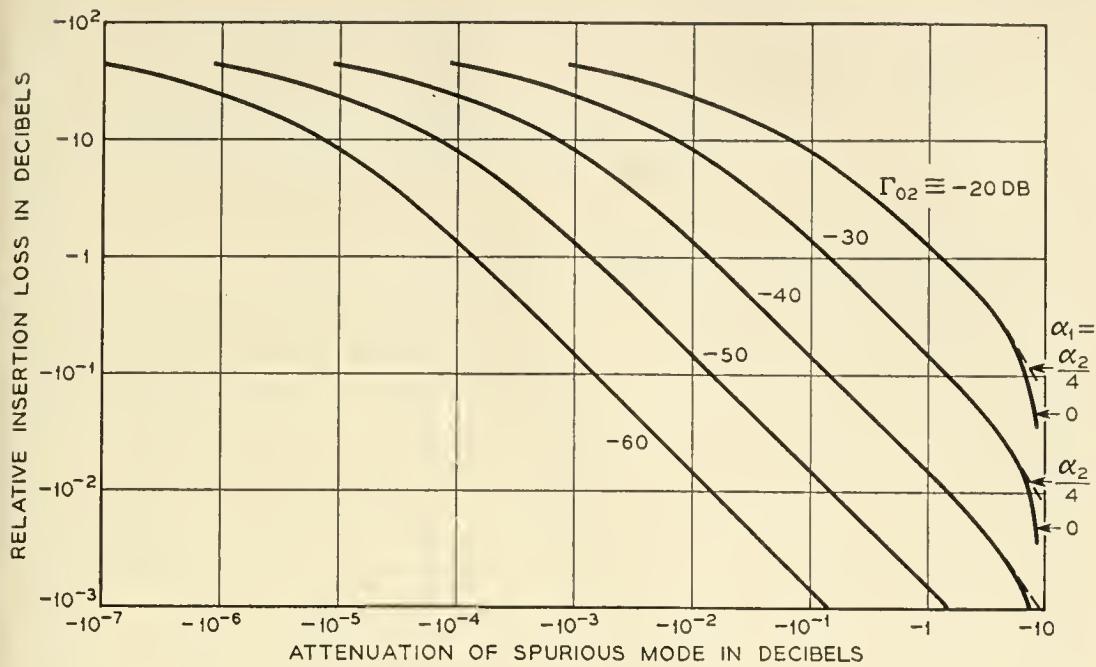


Fig. 3 — Relative insertion loss as a function of the spurious mode attenuation and mode conversion level.

attenuation is low, but the general interpretation of the phenomenon is still the one given above.

From (6) we can calculate the extreme values of $|\Gamma|$ differentiating with respect to ℓ , and we define the relative insertion loss I in db, as the ratio between the minimum and maximum transmitted power expressed in db.

$$I = 20 \log_{10} \left| \frac{\Gamma_{\min}}{\Gamma_{\max}} \right| = 20 \log_{10} \frac{B}{C} \frac{1 - \frac{2 |\Gamma_{12}|^2 (1 + \cosh \alpha \ell)}{1 - C^2 |\Gamma_{22}|^2 e^{-2\alpha_2 \ell}}}{1 - \frac{2 |\Gamma_{12}|^2 (1 - \cosh \alpha \ell)}{1 + B^2 |\Gamma_{22}|^2 e^{-2\alpha_2 \ell}}} \quad (9)$$

where

$$B = 1 + \left| \frac{\Gamma_{12}}{\Gamma_{22}} \right|^2 e^{-\alpha \ell} \quad C = 1 - \left| \frac{\Gamma_{12}}{\Gamma_{22}} \right|^2 e^{-\alpha \ell} \quad \alpha = \alpha_1 - \alpha_2$$

For the most important practical case, that is, when the maximum value attainable by $\cosh \alpha \ell$ is of the order of 1, and knowing from (3), (4) and (5) that

$$|\Gamma_{12}|^2 = |\Gamma_{02}|^2 (1 - |\Gamma_{02}|^2)$$

$$|\Gamma_{22}|^2 \cong 1 - 2 |\Gamma_{02}|^2$$

$$I \cong 20 \log_{10} (1 + 2 |\Gamma_{02}|^2 e^{-\alpha_2 \ell}) \quad (10)$$

$$\left\{ 1 - \frac{2 |\Gamma_{02}|^2 (1 + \cosh \alpha \ell)}{1 - e^{-2\alpha_2 \ell} + 2 |\Gamma_{02}|^2 (1 + e^{-\alpha_2 \ell}) e^{-2\alpha_2 \ell}} \right\}$$

From this expression we deduce

(a), I is strongly reduced when $\alpha_2 \ell \gg \Gamma_{02}$.

(b), Attenuation in line 1 is not an important factor until $\alpha_1 \ell$ and $|\alpha_1 - \alpha_2| \ell$ are of the order of 1. In other words, for low attenuation in both lines, $\alpha_2 \ell$ assumes a major importance in the determination of I because it influences the conditions of resonance. That the effect of $\alpha_1 \ell$ is small is shown in Fig. 3 (dotted line for the particular case $\alpha_1 = \alpha_2/4$).

In order to handle the general problem, (10) has been plotted in Fig. 3. We can enter with any two and obtain the third following quantities: I_1 , relative insertion loss in db; $10 \log_{10} e^{-2\alpha_2 \ell}$, attenuation in db of the spurious mode in the resonating environment; and $20 \log_{10} \Gamma_{02}$ conversion level at the junction, in db, of power in the spurious mode relative to that in the first line.

APPLICATION OF THESE RESULTS TO A TE_{01} TRANSMITTING SYSTEM

The results of the preceding section have been checked experimentally by measuring the relative insertion loss of different lengths of $\frac{7}{8}$ " diameter round waveguide tapered at both ends to round waveguides of $\frac{7}{16}$ " diameter. This waveguide is shown in Fig. 4 with a schematic diagram of the measuring set. In the round transmission line $A-B$, section A will propagate only TE_{01} . Section B, which has been expanded by means of the conical taper T_1 , can support TE_{02} and TE_{03} in addition to the principal TE_{01} mode. This section is a closed region to the spurious modes (TE_{02} , TE_{03}) whose length can be adjusted to resonate each one of these modes. A sliding piston provides a means for varying the length, ℓ , of section B.

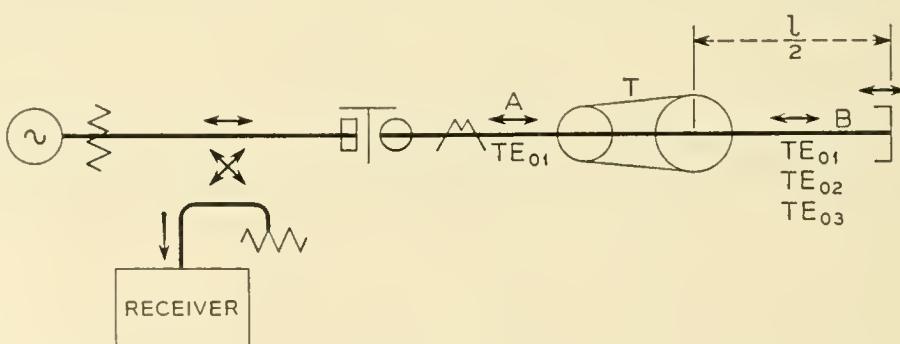


Fig. 4 — Circuit used to measure TE_{01} insertion loss due to resonance of the TE_{02} and TE_{03} modes.

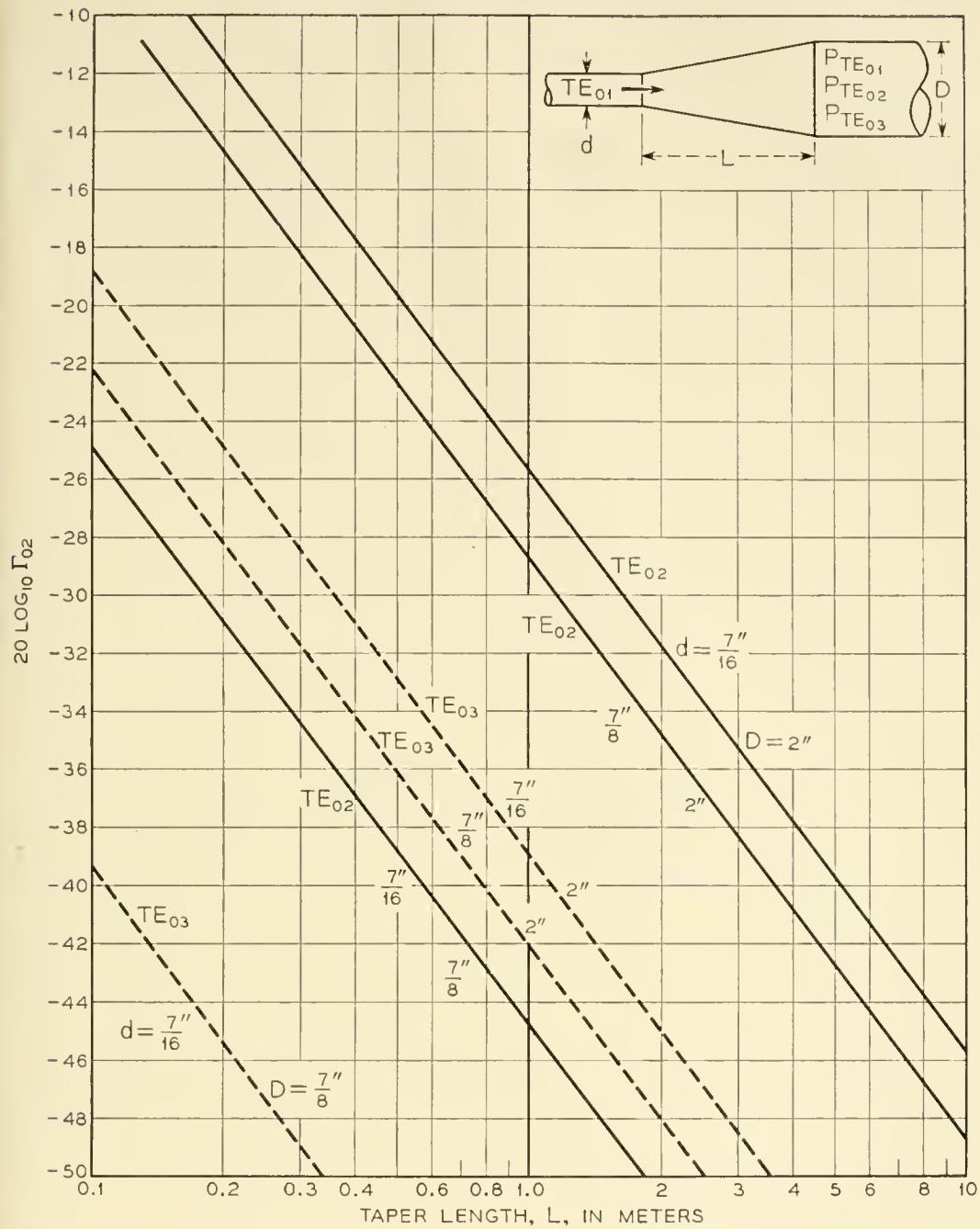


Fig. 5 — Mode conversion of TE_{02} and TE_{03} relative to TE_{01} generated by a conical taper.

The relative levels of TE_{02} and TE_{03} conversions, which have been calculated from unpublished work of S. P. Morgan are shown plotted in Fig. 5 for the waveguide sizes employed in the millimeter wavelength band. The conversions, $20 \log_{10} \Gamma_{02}$, are plotted in terms of the TE_{02} and TE_{03} powers relative to the TE_{01} mode power and are expressed in db as a function of the taper length L , in meters.

Fig. 6 shows the theoretical and experimental values obtained for TE_{01} relative insertion loss. Since the minimum length of pipe tested is

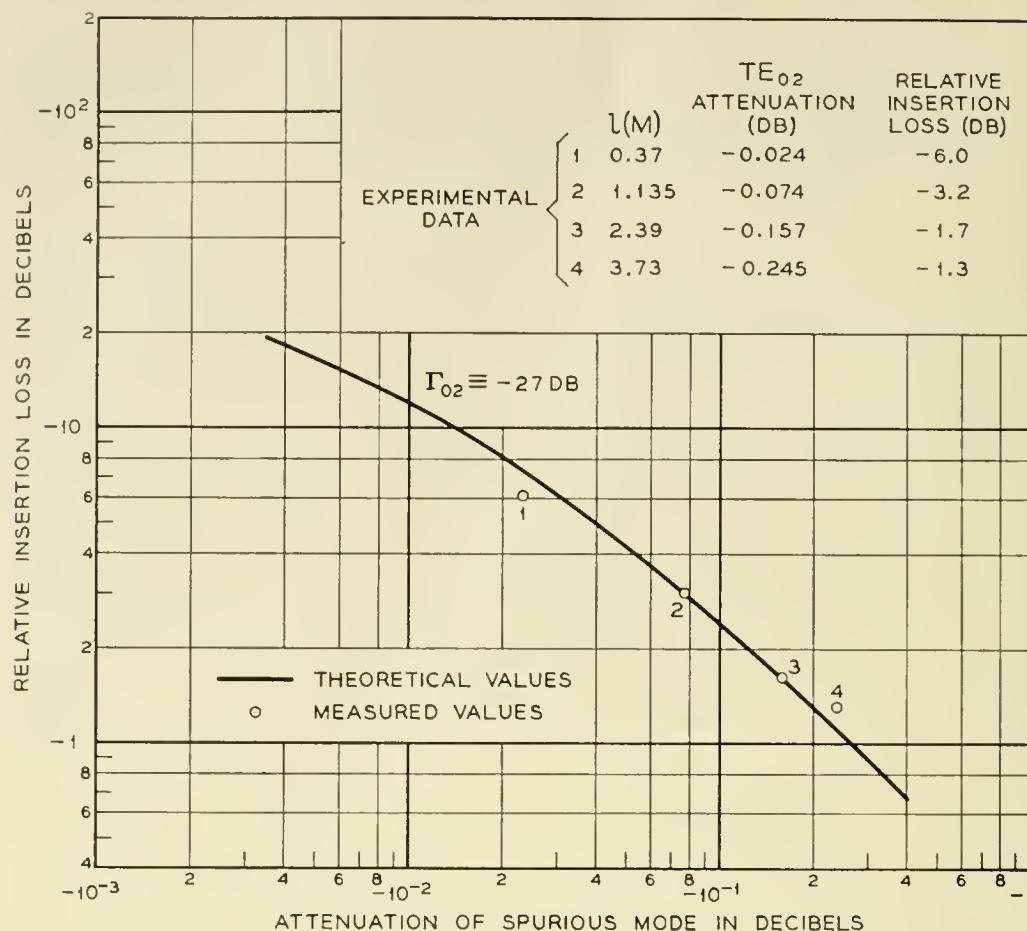


Fig. 6 Theoretical and measured relative insertion loss in the TE_{01} transmission system of Fig. 4.

several times the length of the tapers, the losses in the transitions are fairly small compared to the losses in the multimode guide and this justifies assumption (5). The resonance due to the other modes is too small to be appreciable. This is understandable since, according to (10), the value of the mode conversion for the TE_{03} (Fig. 5) and the attenuation for the shortest length of pipe tested, the calculated relative insertion loss is less than -0.1 db.

CONCLUSIONS

The resonance of spurious modes in a closed environment can produce a large insertion loss of the transmitting mode. In a fairly narrow band device it is possible to avoid this problem by selecting a proper wave-guide size for the closed environment. In a broad-band system the losses can be minimized by providing a high attenuation and a low mode conversion for the spurious mode. For example, it may be noted, by referring to Fig. 3, that mode conversion as high as -20 db with a spurious mode loss of -8 db results in only an -0.1 db insertion loss for the transmitting mode.

Measurement of Atmospheric Attenuation at Millimeter Wavelengths

By A. B. CRAWFORD and D. C. HOGG

(Manuscript received September 20, 1955)

A frequency-modulation radar technique especially suited to measurement of atmospheric attenuation at millimeter wavelengths is described. This two-way transmission method employs a single klystron, a single antenna and a set of spaced corner reflectors whose relative reflecting properties are known. Since the method does not depend on measurements of absolute antenna gains and power levels, absorption data can be obtained more readily and with greater accuracy than by the usual one-way transmission methods.

Application of the method is demonstrated by measurements in the 5-mm to 6-mm wave band. The results have made it possible to assign an accurate value for the line-breadth constant of oxygen at atmospheric pressure; the constant appropriate to the measurements lies between 600 and 800 MCS per atmosphere.

INTRODUCTION

It is well known that certain bands in the microwave region are attenuated considerably due to absorption by water vapour and oxygen in the atmosphere. A theory of absorption for both gases was given by Van Vleck.¹ Numerous measurements have been made on the gases when confined to waveguides or cavities² and several when unconfined in the free atmosphere.³ Nevertheless, there is some uncertainty regarding the line-breadth constants which should be used in calculating water vapour and oxygen absorption. In particular, at atmospheric pressure there is doubt as to the amount of absorption on the skirts of the bands where the absorption is small. The present work was undertaken to test a new method of measurement and to improve the accuracy of experimental data measured in the free atmosphere.

The method of measurement is one of comparison of reflections from

spaced corner reflectors whose relative reflecting properties are known. The free-space attenuation is readily calculated and any measured attenuation in excess of this represents absorption by the atmospheric gases.

A description of the method and the apparatus is followed by a discussion of data taken in the wavelength range 5.1 to 6.1 mm (which includes the long wavelength skirt of the oxygen absorption band centered at 5 mm). These data, when compared with the theory,¹ indicate that the line-broadening constant of oxygen at atmospheric pressure is of the order of 600 me. Some rain and fog attenuation measurements at a wavelength of 6.0 mm are included.

METHOD

The experimental setup is shown in Fig. 1. It consists of a high-gain antenna for both transmitting and receiving and a pair of spaced corner reflectors. Corner reflectors can be built to have good mechanical and electrical stability, and their reflecting properties are relatively insensitive to slight misalignments. The reflectors are mounted well above the ground to ensure free-space propagation conditions.

At the outset, the relative reflecting properties of the corner reflectors are measured by placing them side by side at a convenient distance (d_1 for example) from the antenna. By alternately covering one and the other with absorbent non-reflecting material and measuring the reflected signals, the relative effective areas are determined. The reflectors are then separated as shown and consecutive measurements are made of the signals returned from each reflector. From these measurements, knowing the distances d_1 and d_2 and the calibration of the reflectors, one determines the attenuation over the path d_2-d_1 in excess of the free-space attenuation.* This excess, in the absence of condensed water in the air, represents absorption by the atmosphere.

* The power received from the reflector at distance d_1 is,

$$P_1 = P_T \frac{A^2 A_1^2}{\lambda^4 d_1^4} Q(\lambda, d_1)$$

where A and A_1 are the effective areas of the antenna and corner-reflector respectively, and P_T is the transmitted power; $Q(\lambda, d_1)$ is a loss factor which accounts for atmospheric absorption. A similar relation holds for the power received from the reflector at distance d_2 . The ratio of the received powers is then,

$$\frac{P_1}{P_2} = \left(\frac{A_1}{A_2} \right)^2 \left(\frac{d_2}{d_1} \right)^4 Q[\lambda, (d_2 - d_1)]$$

The accuracy of the measurements will be affected, of course, by spurious reflections in the neighborhood of the corner-reflectors. The sites for the experiment were chosen to minimize such reflections and checks were made by observing the decrease in the return signals when the corner-reflectors were covered by absorbent material. In all cases, the background reflections were at least 30 db below the signal from the corner-reflector.

The method of measuring the reflected signals is illustrated in Fig. 2. The transmitted signal is frequency modulated in a saw tooth manner with a small total frequency excursion, F . The signal reflected from the near corner-reflector is delayed with respect to the transmitted signal by a time, τ_1 , equal to twice the distance to the reflector divided by the velocity of light. During a portion, $T_1 - \tau_1$, of the sawtooth cycle, there is a constant frequency difference, f , between the transmitted and received signals, ($f/F = \tau_1/T_1$). Power at this frequency is produced by mixing the initial source signal with the delayed received signal and amplifying the difference frequency in a narrow-band amplifier centered at frequency f . The output of this amplifier is, therefore, a pulse at frequency f , of length $T_1 - \tau_1$ and repetition rate $1/T_1$.

To measure the signal returned from the far corner-reflector it is necessary merely to increase the period of the sawtooth modulation proportionate to the increase in distance. The frequency excursion, F , remains the same; hence the average power output of the transmitter is unchanged. As may be seen in Fig. 2, the frequency difference, f , between the transmitted and received signals is unchanged; thus the same amplifier and output meter can be used for the two cases. Another advantage in changing only the sawtooth repetition rate is that the delay is the same fraction of a period in both cases; therefore the duty cycle is unchanged and the intermediate frequency pulses can be detected by either an average or a peak measuring device.

Since the beat frequency, f , is not affected by slow changes in the fre-

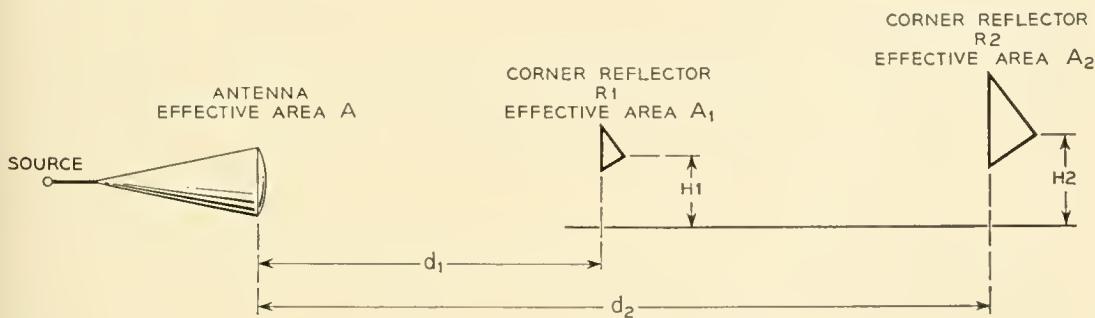


Fig. 1 — Siting arrangement for the atmospheric absorption measurements.

quency of the transmitter, the bandwidth of the intermediate frequency amplifier need be only wide enough to take care of non-linearity in the sawtooth modulation. A signal-to-noise advantage is obtained by the use of the narrow-band amplifier.

Table I gives the distances, heights and effective areas of the reflectors as well as the sawtooth repetition rates that were used in the experiment. The frequency excursion of the sawtooth modulation was 5.8 me.

It will be noted that three reflectors were used; this was done to provide a long path (comparison of reflections from R_1 and R_3) for wavelengths at which the absorption was relatively low, and a short path (comparison of R_1 and R_2) for wavelengths at which the absorption was high. The small reflector, R_1 , was one foot on a side; the large reflectors, R_2 and R_3 , were about 5.6 feet on a side. Fig. 3 is a set of side-by-side measurements showing the reflecting properties of the large reflectors relative to the small one for the wavelengths at which they were used.

APPARATUS

A schematic diagram of the waveguide and electronic apparatus is shown in Fig. 4; Fig. 5 is a photograph of the waveguide equipment so mounted that it moves as a unit with the horn antenna. The antenna is adjusted in azimuth and elevation by means of the milling vise at the bottom of the photograph. The box at the left contains the transmitting tube, a low voltage reflex klystron* which has an average power output of about 12 milliwatts over its 5.1- to 6.1-mm tuning range. About 2 milliwatts of the klystron output is fed through a 6-db directional coupler to a balanced converter that contains two wafer-type millimeter rectifier units.† The remainder of the power proceeds into a 3-db coupler which

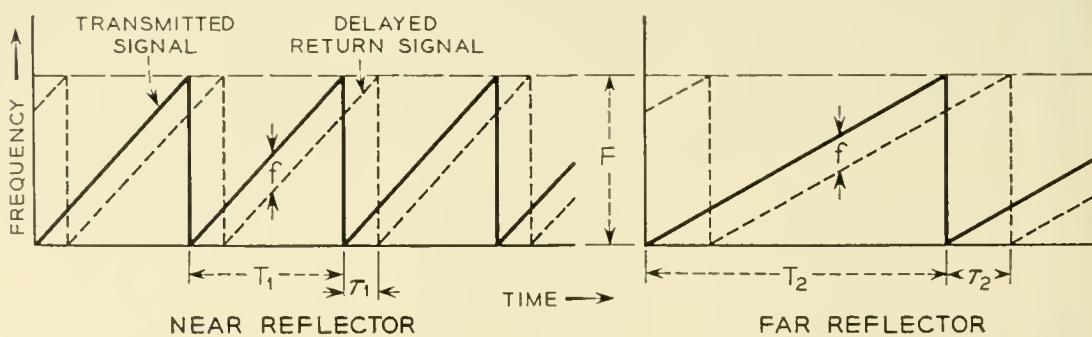


Fig. 2 — Transmitted and reflected frequency-modulated signals.

* This klystron was developed by E. D. Reed, Electron Tube Development Department, Murray Hill Laboratory.

† These millimeter-wave rectifiers were developed by W. M. Sharpless, Radio Research Department, at the Holmdel Laboratory.

TABLE I

Reflector	Distance	Height	Effective Area (Average)	Sawtooth Rep. Rate	Intermediate Frequency-f
	km	m	m^2	kc	kc
R1	$d_1 = 0.59$	6.7	0.05	33	750
R2	$d_2 = 1.36$	21.5	0.67	14.4	750
R3	$d_3 = 2.87$	75	0.79	6.8	750

has the antenna on one arm and an impedance composed of an adjustable attenuator and shorting plunger on another arm. This impedance is adjusted to balance out reflections from the antenna so that a negligible amount of the power flowing toward the antenna enters the converter which is on the remaining arm of the coupler. The delayed energy that re-enters the antenna after reflection from a corner reflector passes through the 3-db coupler to the converter.

The intermediate frequency amplifier shown in Fig. 4 operates with a bandwidth of 300 kc centered at $f=750$ kc. The output of the amplifier is fed to a square law detector and meter for accurate measurement and to an oscilloscope for checking operation of the equipment. Oscillograms of the pulses obtained from the three corner reflectors are shown in Fig. 6; these are all on the same time scale. The gap between the pulses is the delay, τ , shown schematically in Fig. 2.

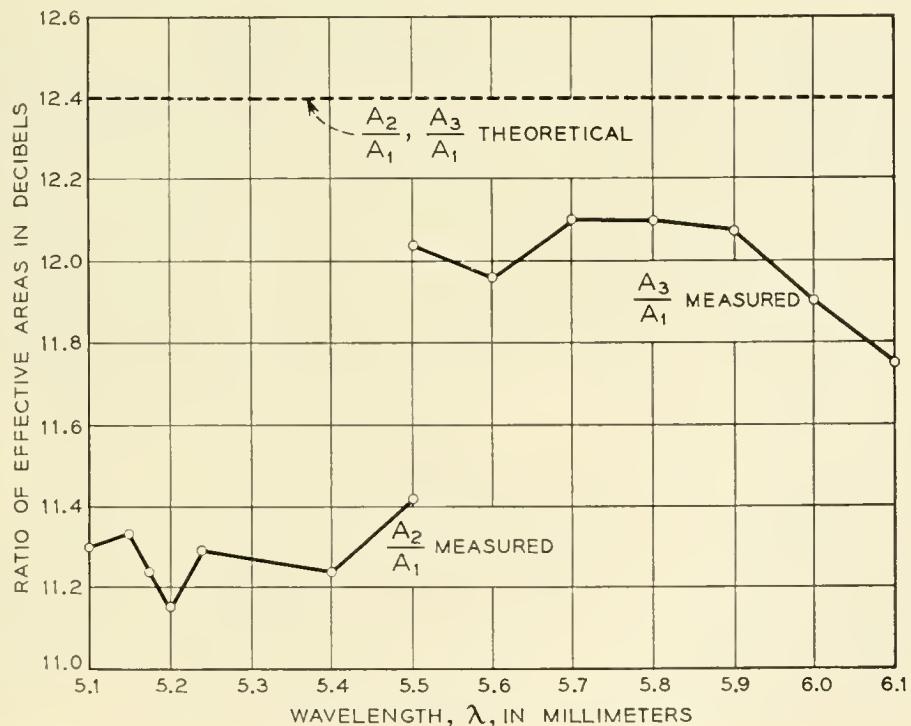


Fig. 3 — Calibration of corner-reflectors R2 and R3 using R1 as a standard.

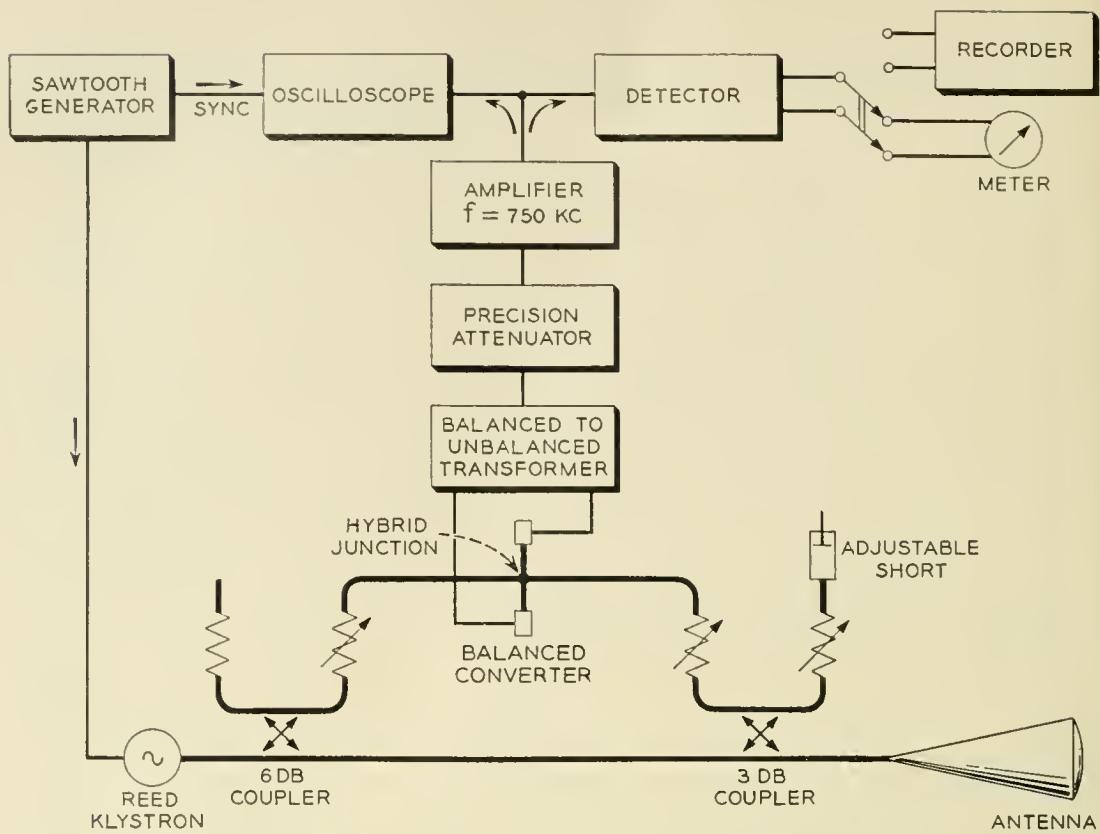


Fig. 4 — Schematic diagram of frequency-modulation radar.

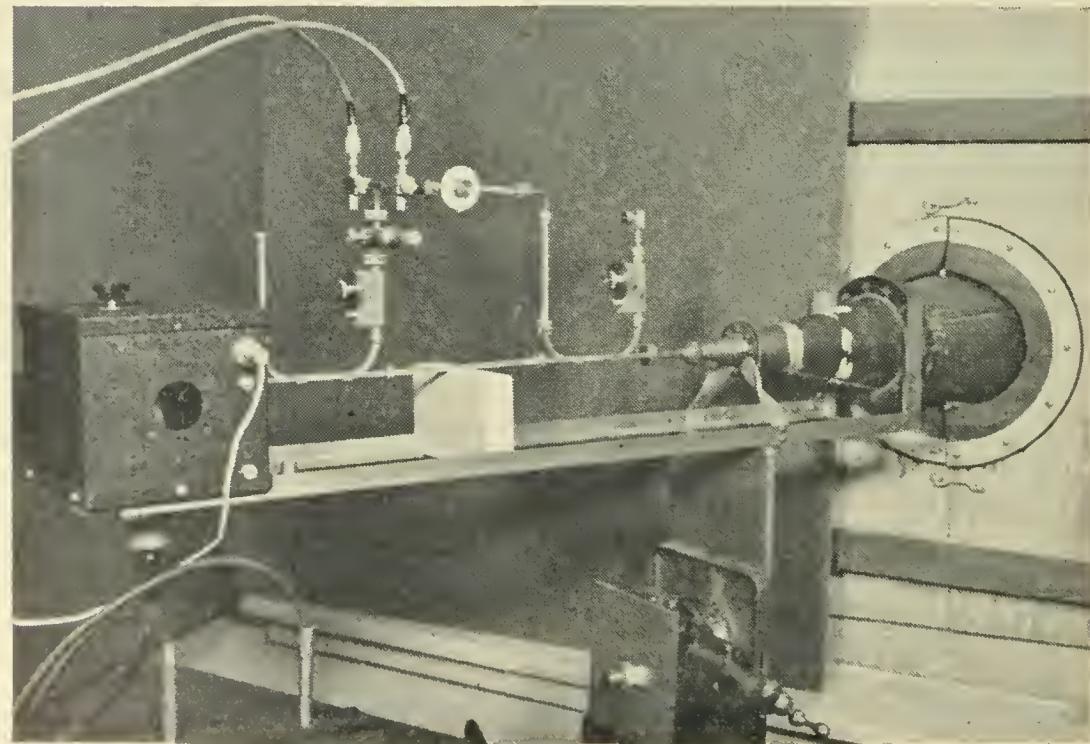


Fig. 5 — Waveguide apparatus and antenna.

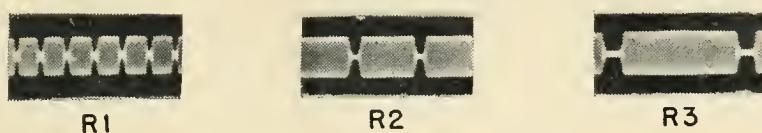


Fig. 6 — 750-kc pulses corresponding to the data in Table I.

Fig. 7 shows the conical horn-lens antenna supported by two bearings to allow adjustment of azimuth and elevation angles. The aperture of the antenna is fitted with a polyethylene lens 30 inches in diameter. The antenna has a gain of about 51 db and a beam width of about 0.5 degrees in the middle of the 5- to 6-mm wave band. This narrow beam, together with well-elevated reflectors, essentially eliminated ground reflections from the measurements.

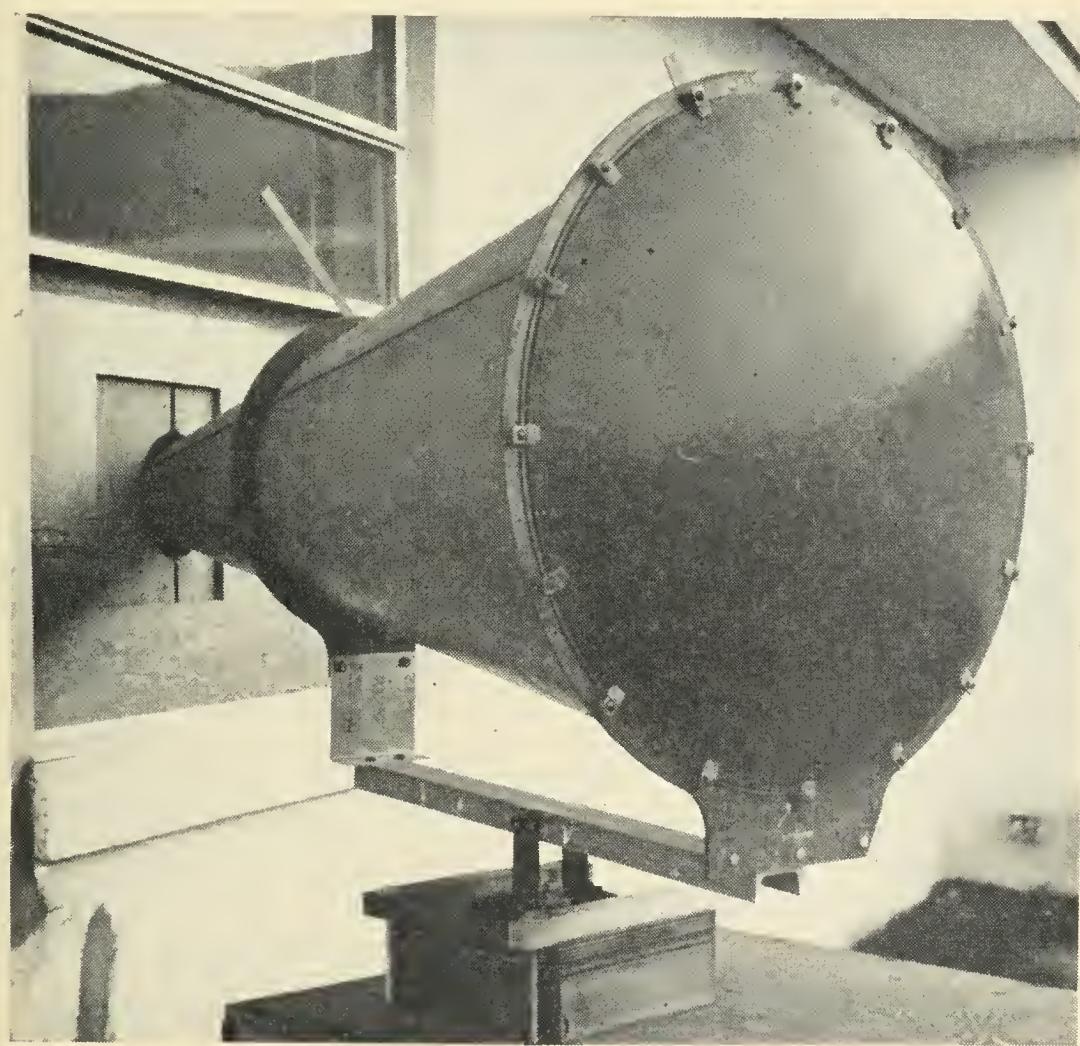


Fig. 7 — Conical horn-lens antenna and mount.

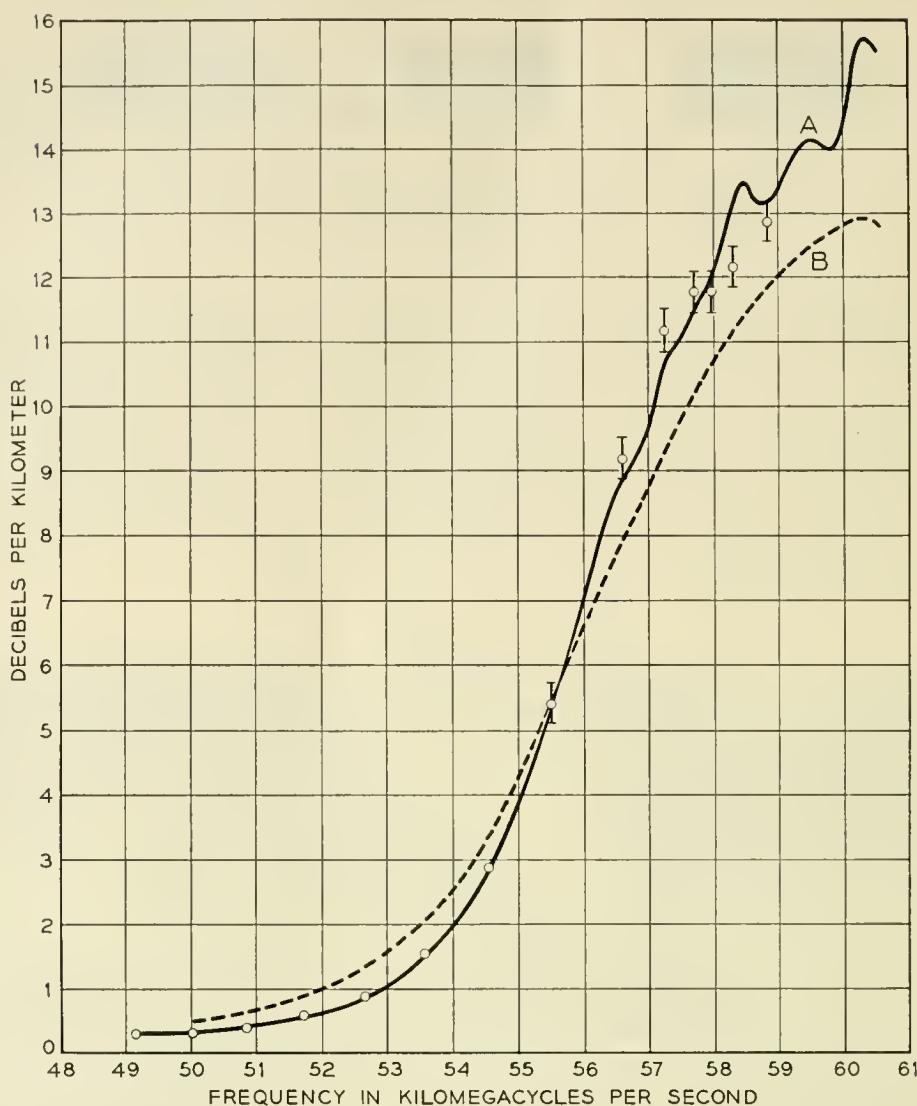


Fig. 8—Calculated and measured absorption by air at sea level. The dots represent the experimental data; the vertical lines indicate the spread in the measured values. Curves A and B are calculated curves of oxygen absorption using line-breadth constants of 600 and 1200 mc, respectively, and a temperature of 293° K. (Courtesy of T. F. Rogers, Air Force Cambridge Research Center.)

RESULTS

The data to be discussed are shown in Fig. 8; they were taken at Holmdel, N. J., during the months of December, 1954, and January, 1955, on days when the temperature was between 25 and 40 degrees Fahrenheit; the absolute humidity was less than 5 grams/meter³ during the measurements. It is believed, therefore, that the resonance of the oxygen molecule is the main contributor to the absorption.

The spread in the measurements is indicated by vertical lines through the average values. Each point represents an average of six or more measurements taken on different days. In the range 49 to 54.5 kmc, (5.5 to

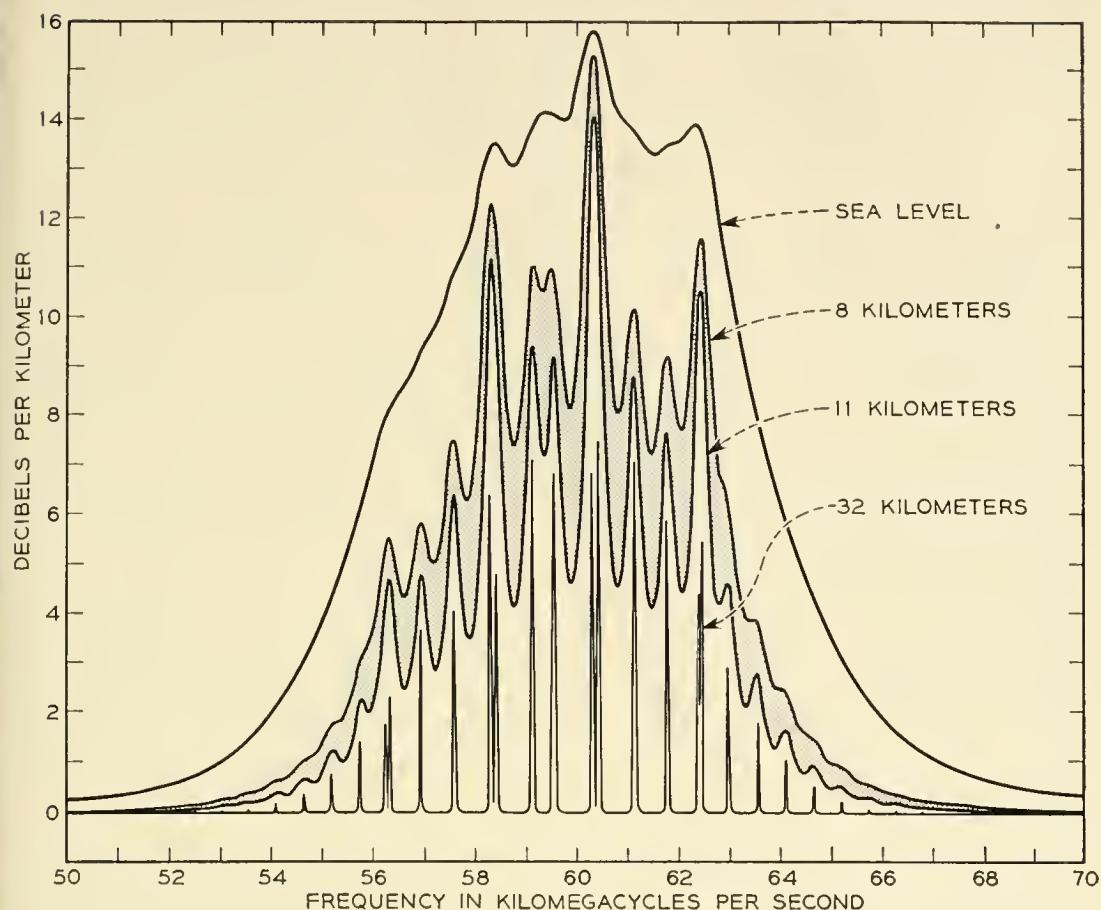


Fig. 9—Calculated curves of oxygen absorption at various altitudes for a line-breadth constant of 600 megacycles and a temperature of 293° K. (Courtesy of T. F. Rogers, Air Force Cambridge Research Center.)

6.1 mm) the measurements were highly consistent, due mainly to the longer path that was used. Errors in the absolute values of the absorption are estimated not to exceed ± 0.05 db/km in the 49 to 54.5 kmc region, ± 0.25 db/km in the 55.5 to 59 kmc region. The errors in absolute absorption are governed mainly by the structural and thermo-mechanical stability of the corner reflectors.

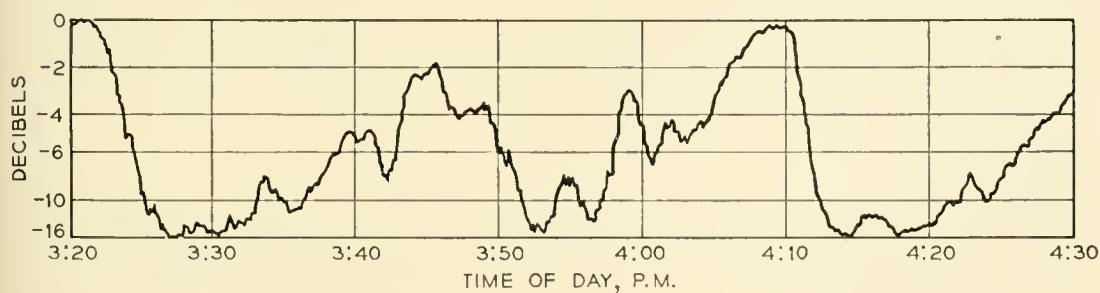


Fig. 10—Attenuation of 6.0-mm radiation caused by a light rain.

Round-trip path length = 2.72 kilometers

Average rainfall rate = 5 millimeters per hour

TABLE II

Approximate Optical Visibility (miles)	Attenuation due to Land Fog DB/KM
1½	0.06
½	0.13
¼	0.22

In Fig. 8, measured values are compared with the theory of Van Vleck as calculated by T. F. Rogers using line-breadth constants of 600 mc and 1200 mc per atmosphere. The fit with the 600-mc curve is good from 49 to 55.5 kmc, but discrepancies are evident between 56.5 and 59 kmc. For completeness, Rogers' calculations for the absorption at higher altitudes are reproduced in Fig. 9.

A few continuous recordings of rain attenuation have been made at a wavelength of 6.0 mm; a record taken during a light rain is shown in Fig. 10. The median value of the signal is -6.7 db which corresponds to an attenuation of 2.5 db/km for this 5 mm per hour rainfall. During more intensive rainfalls, short-term attenuations in excess of 25 db/km have been observed.

On one occasion, it was possible to measure attenuation by land fog. The measurements given in Table II were made at a wavelength of 6.0 mm. No information regarding water content or drop size was available for this fog.

CONCLUSION

A frequency-modulation, two-way transmission technique has proven reliable for measurement of atmospheric attenuation at millimeter wavelengths. Prerequisite to the success of the method are corner reflectors with good mechanical, thermal and electrical stability.

The frequency-modulation method has been demonstrated by absorption measurements in the free atmosphere in the 5.1- to 6.1-mm band. The data thus obtained are in good agreement with Van Vleck's theory of oxygen absorption; the line-breadth constant appropriate to the measurements lies between 600 and 800 mc per atmosphere.

REFERENCES

1. J. H. Van Vleck, Phys. Rev., **71**, pp. 413 ff, 1947.
2. R. Beringer, Phys. Rev., **70**, p. 53, 1946. R. S. Anderson, W. V. Smith and W. Gordy, Phys. Rev. **87**, p. 561, 1952. J. O. Artman and J. P. Gordon, Phys. Rev., **96**, p. 1237, 1954.
3. R. H. Dicke, R. Beringer, R. L. Kyhl, A. B. Vane, Phys. Rev., **70**, p. 340, 1946. G. E. Mueller, Proc. I.R.E., **34**, p. 181, 1946. H. R. Lamont, Phys. Rev., **74**, p. 353, 1948.

A New Interpretation of Information Rate

By J. L. KELLY, JR.

(Manuscript received March 21, 1956)

If the input symbols to a communication channel represent the outcomes of a chance event on which bets are available at odds consistent with their probabilities (i.e., "fair" odds), a gambler can use the knowledge given him by the received symbols to cause his money to grow exponentially. The maximum exponential rate of growth of the gambler's capital is equal to the rate of transmission of information over the channel. This result is generalized to include the case of arbitrary odds.

Thus we find a situation in which the transmission rate is significant even though no coding is contemplated. Previously this quantity was given significance only by a theorem of Shannon's which asserted that, with suitable encoding, binary digits could be transmitted over the channel at this rate with an arbitrarily small probability of error.

INTRODUCTION

Shannon defines the rate of transmission over a noisy communication channel in terms of various probabilities.¹ This definition is given significance by a theorem which asserts that binary digits may be encoded and transmitted over the channel at this rate with arbitrarily small probability of error. Many workers in the field of communication theory have felt a desire to attach significance to the rate of transmission in cases where no coding was contemplated. Some have even proceeded on the assumption that such a significance did, in fact, exist. For example, in systems where no coding was desirable or even possible (such as radar), detectors have been designed by the criterion of maximum transmission rate or, what is the same thing, minimum equivocation. Without further analysis such a procedure is unjustified.

The problem then remains of attaching a value measure to a communi-

¹ C. E. Shannon, A Mathematical Theory of Communication, B.S.T.J., 27, pp. 379-423, 623-656, Oct., 1948.

cation system in which errors are being made at a non-negligible rate, i.e., where optimum coding is not being used. In its most general formulation this problem seems to have but one solution. A cost function must be defined on pairs of symbols which tell how bad it is to receive a certain symbol when a specified signal is transmitted. Furthermore, this cost function must be such that its expected value has significance, i.e., a system must be preferable to another if its average cost is less. The utility theory of Von Neumann² shows us one way to obtain such a cost function. Generally this cost function would depend on things external to the system and not on the probabilities which describe the system, so that its average value could not be identified with the rate as defined by Shannon.

The cost function approach is, of course, not limited to studies of communication systems, but can actually be used to analyze nearly any branch of human endeavor. The author believes that it is too general to shed any light on the specific problems of communication theory. The distinguishing feature of a communication system is that the ultimate receiver (thought of here as a person) is in a position to profit from any knowledge of the input symbols or even from a better estimate of their probabilities. A cost function, if it is supposed to apply to a communication system, must somehow reflect this feature. The point here is that an arbitrary combination of a statistical transducer (i.e., a channel) and a cost function does not necessarily constitute a communication system. In fact (not knowing the exact definition of a communication system on which the above statements are tacitly based) the author would not know how to test such an arbitrary combination to see if it were a communication system.

What can be done, however, is to take some real-life situation which seems to possess the essential features of a communication problem, and to analyze it without the introduction of an arbitrary cost function. The situation which will be chosen here is one in which a gambler uses knowledge of the received symbols of a communication channel in order to make profitable bets on the transmitted symbols.

THE GAMBLER WITH A PRIVATE WIRE

Let us consider a communication channel which is used to transmit the results of a chance situation before those results become common knowledge, so that a gambler may still place bets at the original odds. Consider first the case of a noiseless binary channel, which might be

² Von Neumann and Morgenstern, *Theory of Games and Economic Behavior*, Princeton Univ. Press, 2nd Edition, 1947.

used, for example, to transmit the results of a series of baseball games between two equally matched teams. The gambler could obtain even money bets even though he already knew the result of each game. The amount of money he could make would depend only on how much he chose to bet. How much would he bet? Probably all he had since he would win with certainty. In this case his capital would grow exponentially and after N bets he would have 2^N times his original bankroll. This exponential growth of capital is not uncommon in economies. In fact, if the binary digits in the above channel were arriving at the rate of one per week, the sequence of bets would have the value of an investment paying 100 per cent interest per week compounded weekly. We will make use of a quantity G called the exponential rate of growth of the gambler's capital, where

$$G = \lim_{N \rightarrow \infty} \frac{1}{N} \log \frac{V_N}{V_0}$$

where V_N is the gambler's capital after N bets, V_0 is his starting capital, and the logarithm is to the base two. In the above example $G = 1$.

Consider the case now of a noisy binary channel, where each transmitted symbol has probability, p , or error and q of correct transmission. Now the gambler could still bet his entire capital each time, and, in fact, this would maximize the expected value of his capital, $\langle V_N \rangle$, which in this case would be given by

$$\langle V_N \rangle = (2q)^N V_0$$

This would be little comfort, however, since when N was large he would probably be broke and, in fact, would be broke with probability one if he continued indefinitely. Let us, instead, assume that he bets a fraction, ℓ , of his capital each time. Then

$$V_N = (1 + \ell)^W (1 - \ell)^L V_0$$

where W and L are the number of wins and losses in the N bets. Then

$$\begin{aligned} G &= \lim_{N \rightarrow \infty} \left[\frac{W}{N} \log (1 + \ell) + \frac{L}{N} \log (1 - \ell) \right] \\ &= q \log (1 + \ell) + p \log (1 - \ell) \text{ with probability one} \end{aligned}$$

Let us maximize G with respect to ℓ . The maximum value with respect to the Y_i of a quantity of the form $Z = \sum X_i \log Y_i$, subject to the constraint $\sum Y_i = Y$, is obtained by putting

$$Y_i = \frac{Y}{X} X_i,$$

where $X = \sum X_i$. This may be shown directly from the convexity of the logarithm.

Thus we put

$$(1 + \ell) = 2q$$

$$(1 - \ell) = 2p$$

and

$$\begin{aligned} G_{\max} &= 1 + p \log p + q \log q \\ &= R \end{aligned}$$

which is the rate of transmission as defined by Shannon.

One might still argue that the gambler should bet all his money (make $\ell = 1$) in order to maximize his expected win after N times. It is surely true that if the game were to be stopped after N bets the answer to this question would depend on the relative values (to the gambler) of being broke or possessing a fortune. If we compare the fates of two gamblers, however, playing a nonterminating game, the one which uses the value ℓ found above will, with probability one, eventually get ahead and stay ahead of one using any other ℓ . At any rate, we will assume that the gambler will always bet so as to maximize G .

THE GENERAL CASE

Let us now consider the case in which the channel has several input symbols, not necessarily equally likely, which represent the outcome of chance events. We will use the following notation:

- $p(s)$ the probability that the transmitted symbol is the s 'th one.
- $p(r/s)$ the conditional probability that the received symbol is the r 'th on the hypothesis that the transmitted symbol is the s 'th one.
- $p(s, r)$ the joint probability of the s 'th transmitted and r 'th received symbol.
- $q(r)$ received symbol probability.
- $q(s/r)$ conditional probability of transmitted symbol on hypothesis of received symbol.
- α_s the odds paid on the occurrence of the s 'th transmitted symbol, i.e., α_s is the number of dollars returned for a one-dollar bet (including that one dollar).
- $a(s/r)$ the fraction of the gambler's capital that he decides to bet on the occurrence of the s 'th transmitted symbol *after* observing the r 'th received symbol

Only the case of independent transmitted symbols and noise will be considered. We will consider first the ease of "fair" odds, i.e.,

$$\alpha_s = \frac{1}{p(s)}$$

In any sort of parimutuel betting there is a tendency for the odds to be fair (ignoring the "track take"). To see this first note that if there is no "track take"

$$\sum \frac{1}{\alpha_s} = 1$$

since all the money collected is paid out to the winner. Next note that if

$$\alpha_s > \frac{1}{p(s)}$$

for some s a bettor could insure a profit by making repeated bets on the s^{th} outcome. The extra betting which would result would lower α_s . The same feedback mechanism probably takes place in more complicated betting situations, such as stock market speculation.

There is no loss in generality in assuming that

$$\sum_s a(s/r) = 1$$

i.e., the gambler bets his total capital regardless of the received symbol. Since

$$\sum \frac{1}{\alpha_s} = 1$$

he can effectively hold back money by placing canceling bets. Now

$$V_N = \prod_{r,s} [a(s/r)\alpha_s]^{W_{sr}} V_0$$

where W_{sr} is the number of times that the transmitted symbol is s and the received symbol is r .

$$\begin{aligned} \text{Log } \frac{V_N}{V_0} &= \sum_{rs} W_{sr} \log \alpha_s a(s/r) \\ G &= \lim_{N \rightarrow \infty} \frac{1}{N} \log \frac{V_N}{V_0} = \sum_{rs} p(s, r) \log \alpha_s a(s/r) \end{aligned} \tag{1}$$

with probability one. Since

$$\alpha_s = \frac{1}{p(s)}$$

here

$$\begin{aligned} G &= \sum_{rs} p(s, r) \log \frac{a(s/r)}{p(s)} \\ &= \sum_{rs} p(s, r) \log a(s/r) + H(X) \end{aligned}$$

where $H(X)$ is the source rate as defined by Shannon. The first term is maximized by putting

$$a(s/r) = \frac{p(s, r)}{\Sigma_k p(k, r)} = \frac{p(s, r)}{q(r)} = q(s/r)$$

Then $G_{\max} = H(X) - H(X/Y)$, which is the rate of transmission defined by Shannon.

WHEN THE ODDS ARE NOT FAIR

Consider the case where there is no track take, i.e.,

$$\sum \frac{1}{\alpha_s} = 1$$

but where α_s is not necessarily

$$\frac{1}{p(s)}$$

It is still permissible to set $\sum_s a(s/r) = 1$ since the gambler can effectively hold back any amount of money by betting it in proportion to the $1/\alpha_s$. Equation (1) now can be written

$$G = \sum_{rs} p(s, r) \log a(s/r) + \sum_s p(s) \log \alpha_s .$$

G is still maximized by placing $a(s/r) = q(s/r)$ and

$$\begin{aligned} G_{\max} &= -H(X/Y) + \sum_s p(s) \log \alpha_s \\ &= H(\alpha) - H(X/Y) \end{aligned}$$

where

$$H(\alpha) = \sum_s p(s) \log \alpha_s$$

Several interesting facts emerge here

(a) In this case G is maximized as before by putting $a(s/r) = q(s/r)$. That is, *the gambler ignores the posted odds in placing his bets!*

(b) Since the minimum value of $H(\alpha)$ subject to

$$\sum_s \frac{1}{\alpha_s} = 1$$

obtains when

$$\alpha_s = \frac{1}{p(s)}$$

and $H(X) = H(\alpha)$, any deviation from fair odds helps the gambler.

(c) Since the gambler's exponential gain would be $H(\alpha) - H(X)$ if he had no inside information, we can interpret $R = H(X) - H(X/Y)$ as the increase of G_{\max} due to the communication channel. When there is no channel, i.e., $H(X/Y) = H(X)$, G_{\max} is *minimized* (at zero) by setting

$$\alpha_s = \frac{1}{p_s}$$

This gives further meaning to the concept "fair odds."

WHEN THERE IS A "TRACK TAKE"

In the case there is a "track take" the situation is more complicated. It can no longer be assumed that $\sum_s a(s/r) = 1$. The gambler cannot make canceling bets since he loses a percentage to the track. Let $b_r = 1 - \sum_s a(s/r)$, i.e., the fraction not bet when the received symbol is the r^{th} one. Then the quantity to be maximized is

$$G = \sum_{rs} p(s, r) \log [b_r + \alpha_s a(s/r)], \quad (2)$$

subject to the constraints

$$b_r + \sum_s a(s/r) = 1.$$

In maximizing (2) it is sufficient to maximize the terms involving a particular value of r and to do this separately for each value of r since both in (2) and in the associated constraints, terms involving different r 's are independent. That is, we must maximize terms of the type

$$G_r = q(r) \sum_s q(s/r) \log [b_r + \alpha_s a(s/r)]$$

subject to the constraint

$$b_r + \sum_s a(s/r) = 1$$

Actually, each of these terms is the same form as that of the gambler's exponential gain where there is no channel

$$G = \sum_s p(s) \log [b + \alpha_s a(s)]. \quad (3)$$

We will maximize (3) and interpret the results either as a typical term in the general problem or as the total exponential gain in the case of no communication channel. Let us designate by λ the set of indices, s , for which $a(s) > 0$, and by λ' the set for which $a(s) = 0$. Now at the desired maximum

$$\frac{\partial G}{\partial a(s)} = \frac{p(s)\alpha_s}{b + a(s)\alpha_s} \log e = k \quad \text{for } s \in \lambda$$

$$\frac{\partial G}{\partial b} = \sum_s \frac{p(s)}{b + a(s)\alpha_s} \log e = k$$

$$\frac{\partial G}{\partial a(s)} = \frac{p(s)\alpha_s}{b} \log e \leq k \quad \text{for } s \in \lambda'$$

where k is a constant. The equations yield

$$k = \log e, \quad b = \frac{1 - p}{1 - \sigma}$$

$$a(s) = p(s) - \frac{b}{\alpha_s} \quad \text{for } s \in \lambda$$

where $p = \sum_{\lambda} p(s)$, $\sigma = \sum_{\lambda} (1/\alpha_s)$, and the inequalities yield

$$p(s)\alpha_s \leq b = \frac{1 - p}{1 - \sigma} \quad \text{for } s \in \lambda'$$

We will see that the conditions

$$\sigma < 1$$

$$p(s)\alpha_s > \frac{1 - p}{1 - \sigma} s \in \lambda$$

$$p(s)\alpha_s \leq \frac{1 - p}{1 - \sigma} \quad \text{for } s \in \lambda'$$

completely determine λ .

If we permute indices so that

$$p(s)\alpha_s \geq p(s+1)\alpha_{s+1}$$

then λ must consist of all $s \leq t$ where t is a positive integer or zero. Consider how the fraction

$$F_t = \frac{1 - p_t}{1 - \sigma_t}$$

varies with t , where

$$p_t = \sum_1^t p(s), \quad \sigma_t = \sum_1^t \frac{1}{\alpha_s}; \quad F_0 = 1$$

Now if $p(1)\alpha_r < 1$, F_t increases with t until $\sigma_t \geq 1$. In this case $t = 0$ satisfies the desired conditions and λ is empty. If $p(1)\alpha_1 > 1$ F_t decreases with t until $p(t+1)\alpha_{t+1} < F_t$ or $\sigma_t \geq 1$. If the former occurs, i.e., $p(t+1)\alpha_{t+1} < F_t$, then $F_{t+1} > F_t$ and the fraction increases until $\sigma_t \geq 1$. In any case the desired value of t is the one which gives F_t its minimum positive value, or if there is more than one such value of t , the smallest. The maximizing process may be summed up as follows:

- (a) Permute indices so that $p(s)\alpha_s \geq p(s+1)\alpha_{s+1}$
- (b) Set b equal to the minimum positive value of

$$\frac{1 - p_t}{1 - \sigma_t} \quad \text{where} \quad p_t = \sum_1^t p(s), \quad \sigma_t = \sum_1^t \frac{1}{\alpha_s}$$

- (c) Set $a(s) = p(s) - b/\alpha_s$ or zero, whichever is larger. (The $a(s)$ will sum to $1 - b$.)

The desired maximum G will then be

$$G_{\max} = \sum_1^t p(s) \log p(s)\alpha_s + (1 - p_t) \log \frac{1 - p_t}{1 - \sigma_t}$$

where t is the smallest index which gives

$$\frac{1 - p_t}{1 - \sigma_t}$$

its minimum positive value.

It should be noted that if $p(s)\alpha_s < 1$ for all s no bets are placed, but if the largest $p(s)\alpha_s > 1$ some bets might be made for which $p(s)\alpha_s < 1$, i.e., the expected gain is negative. This violates the criterion of the classical gambler who never bets on such an event.

CONCLUSION

The gambler introduced here follows an essentially different criterion from the classical gambler. At every bet he maximizes the expected value of the logarithm of his capital. The reason has nothing to do with

the value function which he attached to his money, but merely with the fact that it is the logarithm which is additive in repeated bets and to which the law of large numbers applies. Suppose the situation were different; for example, suppose the gambler's wife allowed him to bet one dollar each week but not to reinvest his winnings. He should then maximize his expectation (expected value of capital) on each bet. He would bet all his available capital (one dollar) on the event yielding the highest expectation. With probability one he would get ahead of anyone dividing his money differently.

It should be noted that we have only shown that our gambler's capital will surpass, with probability one, that of any gambler apportioning his money differently from ours but still in a fixed way for each received symbol, independent of time or past events. Theorems remain to be proved showing in what sense, if any, our strategy is superior to others involving $a(s/r)$ which are not constant.

Although the model adopted here is drawn from the real-life situation of gambling it is possible that it could apply to certain other economic situations. The essential requirements for the validity of the theory are the possibility of reinvestment of profits and the ability to control or vary the amount of money invested or bet in different categories. The "channel" of the theory might correspond to a real communication channel or simply to the totality of inside information available to the investor.

Let us summarize briefly the results of this paper. If a gambler places bets on the input symbol to a communication channel and bets his money in the same proportion each time a particular symbol is received his capital will grow (or shrink) exponentially. If the odds are consistent with the probabilities of occurrence of the transmitted symbols (i.e., equal to their reciprocals), the maximum value of this exponential rate of growth will be equal to the rate of transmission of information. If the odds are not fair, i.e., not consistent with the transmitted symbol probabilities but consistent with some other set of probabilities, the maximum exponential rate of growth will be larger than it would have been with no channel by an amount equal to the rate of transmission of information. In case there is a "track take" similar results are obtained, but the formulae involved are more complex and have less direct information theoretic interpretations.

ACKNOWLEDGMENTS

I am indebted to R. E. Graham and C. E. Shannon for their assistance in the preparation of this paper.

Automatic Testing of Transmission and Operational Functions of Intertoll Trunks

By H. H. FELDER, A. J. PASCARELLA and
H. F. SHOFFSTALL

(Manuscript received October 19, 1955)

Conditions brought about by nationwide dialing increase intertoll trunk maintenance problems substantially. Under this switching plan with full automatic alternate routing there is a considerable increase in the amount of multiswitched business, and as many as eight intertoll trunks in tandem are permissible. In addition, operator checks of transmission on the connections are lost on most calls. These factors impose more severe limitations on transmission loss variations in the individual trunks and throw on the maintenance forces additional burdens of detecting defects in the distance dialing network.

New methods of analyzing transmission performance to locate the points where maintenance effort will be most effective continue to be studied. The automatic testing arrangements described in this paper enable the maintenance forces to collect over-all transmission loss data quickly and with a minimum of effort. They also facilitate the collection of such data on groups of trunks in a form to make statistical analyses easier. The use of these testing arrangements will permit the maintenance forces to keep a closer watch on intertoll trunk performance and will assist in disclosing trouble patterns.

INTRODUCTION

The advent of nationwide dialing, especially with full automatic alternate routing, has presented additional problems in the maintenance of intertoll trunks. Transmission requirements are more rigorous, the intertoll trunk connections are more complex, and certain irregularities in the performance of the distance dialing network are difficult to detect. Automatic test equipment has been provided to aid and increase the efficiency of over-all testing. This equipment is capable of automatically

testing the operational (signaling and supervisory) functions of dial-type intertoll trunks, and of making two-way transmission loss measurements and a noise check at each end. The test results may be recorded at the originating end by means of a Teletypewriter.

Automatic trunk testing has been used for many years in the local plant for checking the signaling and supervisory features of interoffice trunks. The automatic intertoll trunk testing equipment serves a similar function with respect to these operational features of the intertoll trunks. Because published material is available on automatic operational testing,* these features will not be discussed in detail in this paper; more emphasis is given to the transmission testing features which are new.

MAINTENANCE ARRANGEMENTS FOR INTERTOLL TRUNKS

Except in the very small offices, intertoll trunks usually have a test jack appearance in the toll testboard for maintenance purposes. Cord ended testing equipment in the toll testboard positions enables the attendants to perform various operational tests and to make transmission loss, balance, noise or crosstalk measurements. Facilities are provided for communication with distant offices and with intermediate points where carrier or repeater equipment may be located. Testing of carrier or repeater equipment as individual components or systems is an important aspect of the trunk maintenance problem but is beyond the scope of the present paper.

The maintenance of intertoll trunk net losses close to their specified values is currently a most important transmission problem. Various aspects of the problem are discussed in a companion paper.†

Although the manual testing equipment mentioned above is vital to trunk net loss maintenance, the need for reduction in time and effort required to make measurements has led to the provision of semi-automatic testing arrangements. These arrangements permit a testboard attendant to check transmission in the incoming direction by dialing code 102 over a trunk. The trunk is connected to a source of one milliwatt test power at the far end and a measurement of the received power indicates the net loss. The equivalent of a semi-automatic two-way test may be obtained by making a code 102 test in each direction. If complete information on the test results is desired by one testboard attendant, the attendant at the other end of the trunk must report back his results.

* R. C. Nance, Automatic Intertoll Trunk Testing, Bell Labs. Record, Dec., 1954.

† H. H. Felder and E. N. Little, Intertoll Trunk Net Loss Maintenance Under Operator Distance and Direct Distance Dialing, page 955 of this issue.

In both the manual and semi-automatic methods of measurement, the results must be recorded manually. For statistical analysis of trunk transmission performance in terms of "bias" and "distribution grade", as discussed in the companion paper,* deviations of the measured losses from the respective specified losses must be computed and summarized manually.

The automatic testing equipment described in this paper has been developed as an additional maintenance tool. It will not supplant existing arrangements discussed above but rather is intended to increase the capabilities of plant personnel to do an effective maintenance job. The following features of the equipment contribute particularly to this end:

1. Large numbers of trunks can be tested and the results recorded without the continuous attention of a testboard attendant.
2. The attendant is informed by an alarm whenever the loss of a trunk deviates excessively from the specified value.
3. Computation and summarizing of net loss deviations into class intervals are done automatically, thus facilitating statistical analysis of trunk performance.
4. Data can be collected quickly in large volume for indicating the performance of groups of trunks. Confusion occurring with manual measurements because of changing conditions with time is reduced.
5. Stability of an individual trunk may be checked by a series of repetitive tests.
6. Semi-automatic two-way trunk tests can be made by one attendant when required.

To do an equivalent job entirely by manual methods would require an appreciable increase in the amount of manual test equipment and in the number of test personnel. A comparison of the times required for operational and transmission tests by manual, semi-automatic and automatic methods is shown in Fig. 1. The time shown for the code 102 test does not include coordination time required if information on test results in both directions is required at one end.

GENERAL DESCRIPTION OF AUTOMATIC TESTING EQUIPMENT

Automatic intertoll trunk testing requires automatic equipment at both ends of the trunk. At the originating or control end, an automatic test circuit sets up the test call and controls the various test features. In the distant offices, test lines reached through the switching train provide appropriate automatic test terminations. The automatic equipment for

* H. H. Felder and E. N. Little, Intertoll Trunk Net Loss Maintenance Under Operator Distance and Direct Distance Dialing, page 955 of this issue.

use at the control end, adapted for transmission testing, is presently available only for No. 4 type toll switching offices.

Fig. 2 is a block schematic of the arrangement for automatic intertoll trunk testing, including transmission tests. In the originating No. 4 toll crossbar office an automatic outgoing intertoll trunk test circuit is used which consists of an automatic outgoing intertoll trunk test frame and one or more associated test connector frames. These frames have been provided in all No. 4 type offices and perform the functions of setting up

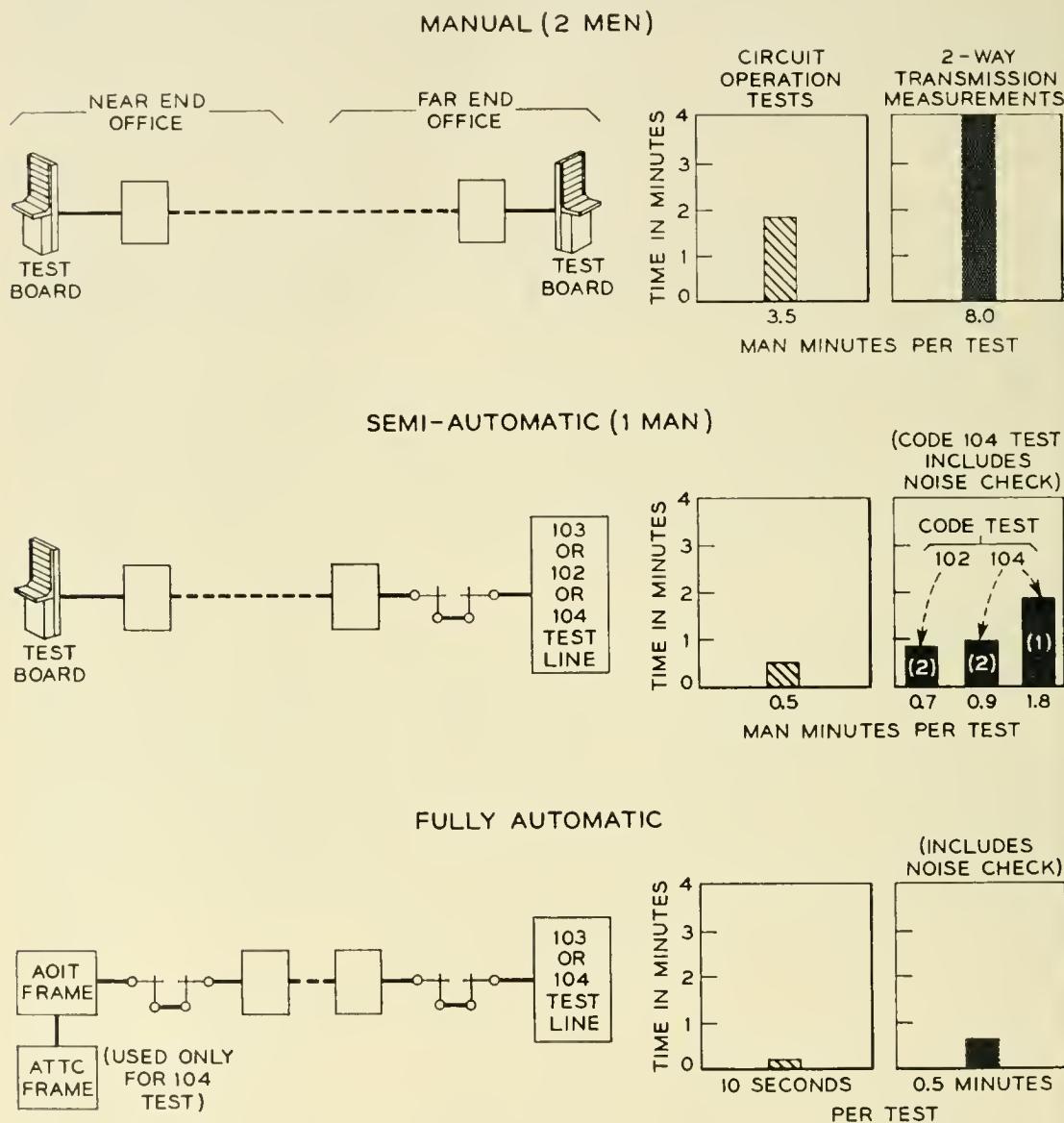


Fig. 1 Time required for manual tests versus semi-automatic and fully automatic tests. (1) Average time per test for 52,000 field test measurements in 20 offices under normal operating test conditions (includes test preparation, time waiting to be served, testing time, and recording of results). (2) Average time per test for test measurements made in rapid sequence during light load period. The semi-automatic code 104 test includes a noise check at the far end only.

the test call and making operational tests on the intertoll trunks. For automatic transmission tests an automatic transmission test and control circuit, provided in a separate frame as an adjunct to the test frame, is brought into play. A Teletypewriter, mounted in the transmission test frame, is adapted for use with the equipment in the originating office for recording test results. The Teletypewriter is used to make a record of trunks having some defect in their operational features, busy trunks passed over without test and, during transmission tests, to record the results of the transmission measurements. The test frame and associated transmission test and control circuit and Teletypewriter are used principally by the toll test board forces and, therefore, are usually located near the toll test board. Figure 3 shows such an installation.

The intertoll trunk test connector frames in the originating office, not shown in Fig. 3, are frames of crossbar switches and there may be several such frames in a large office. Each crosspoint on the switches of the test connector frames represents an individual intertoll trunk. When a trunk is to be tested, the test frame closes the crosspoint of the test connector switches which serves that particular trunk. This extends the selecting leads (trunk sleeve and select magnet leads) of the trunk to the test frame for use in setting up the call. A class contact on the test connector crosspoint also operates one of several class relays in the test frame when the crosspoint is closed. The function of the class relay is discussed later.

The test frame has an appearance on the incoming link frame of the office switching train. The intertoll trunks to be tested appear on the outgoing link frames of the office switching train. When a trunk is to be tested, the test frame engages the office common control equipment (decoder and marker), through a connector, and requests a path between the test frame appearance on the incoming link frame and the particular intertoll trunk which is to be tested. The common control equipment is able to set up this path since the test frame has closed a test connector cross-point to bring the selecting leads of the trunk to be tested into the test frame. The common control equipment uses the select magnet lead to identify the trunk to be tested and thus is able to set up the path to that particular trunk. The test frame uses the trunk sleeve lead for busy test purposes and for controlling the test call.

In the distant offices separate groups of test lines provide automatic test terminations for operational and transmission tests, respectively. These are reached through the switching train as indicated in Fig. 2. The three digit service code 103 is reserved in toll switching offices for reaching the operational test lines and code 104 is reserved for reaching the transmission test lines. A transmission measuring and noise checking

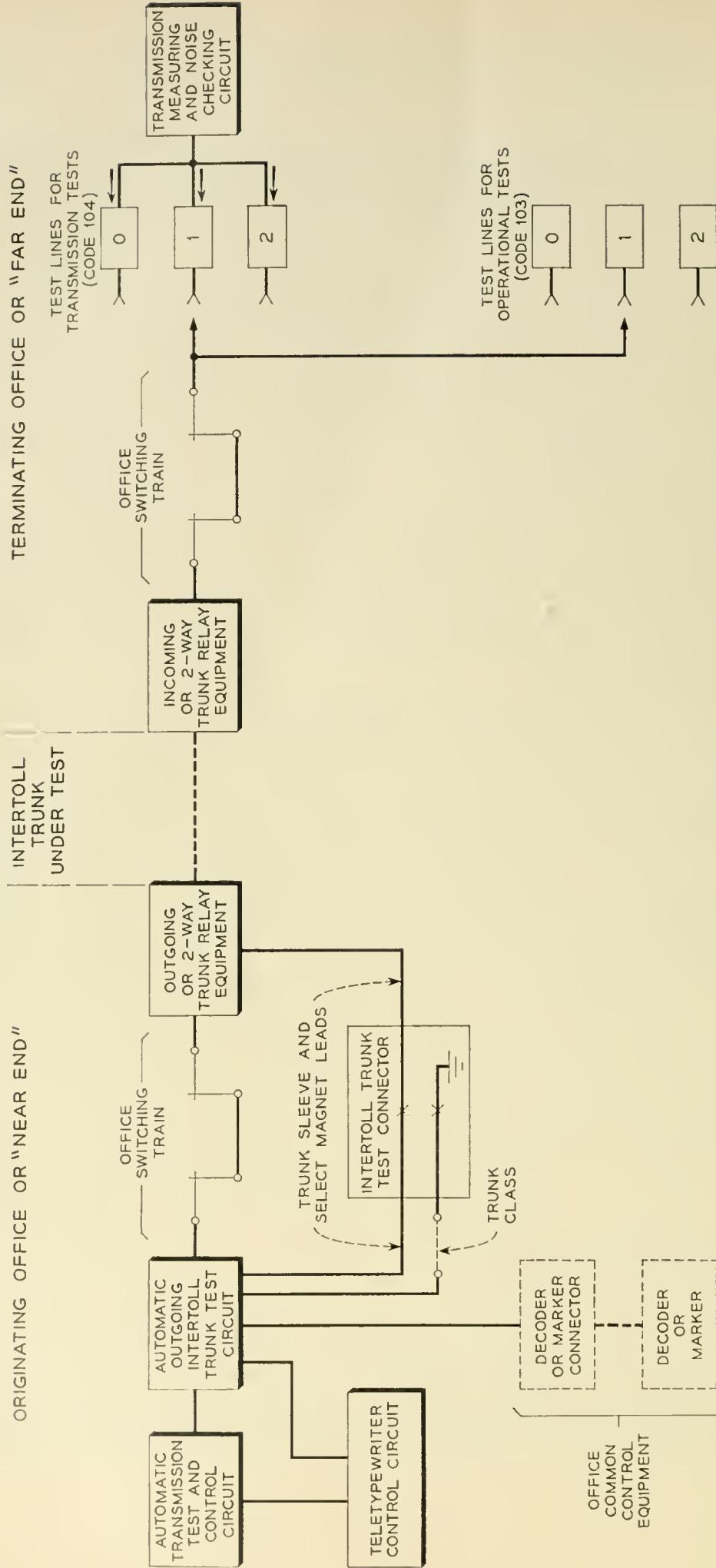


Fig. 2 — Arrangement for automatic tests.

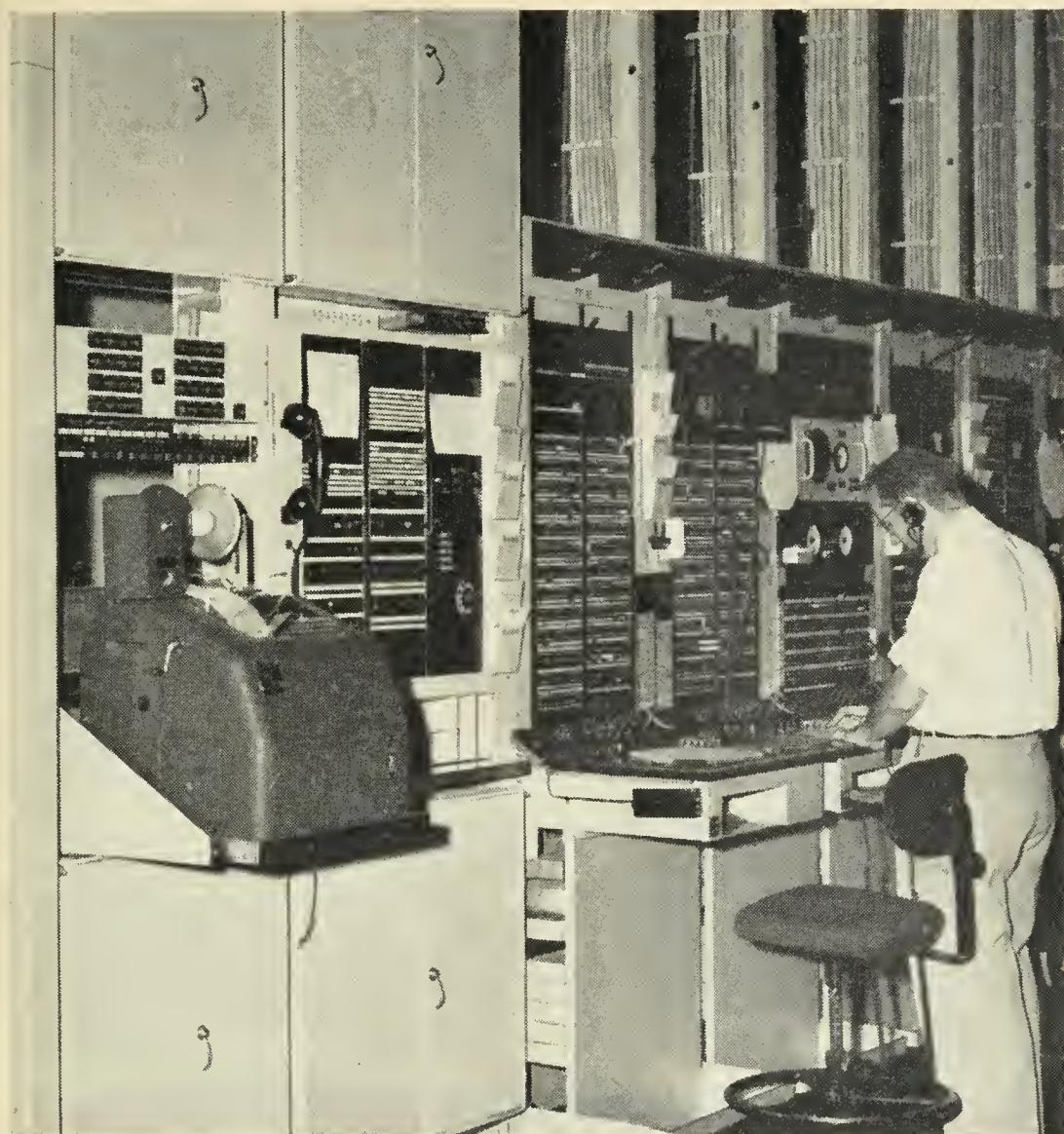


Fig. 3 — Automatic intertoll trunk testing equipment.

circuit, sometimes referred to as "far end equipment," is associated with the group of code 104 test lines for performing the transmission measurements. When simultaneous transmission test calls arrive in the distant office from different originating offices, the calls wait on the code 104 test lines and are served by the transmission measuring and noise checking circuit, one at a time, in their proper turn.

After the trunk test frame has obtained a path through the switching train in the originating office to the trunk to be tested as explained above, it pulses forward over the trunk the desired test line code, either code 103 or code 104. In response to the code, the switching equipment in the distant office sets up a path through the switching train to one of

the code 103 or code 104 test lines. A steady off-hook signal is returned over the established connection to the test frame at the originating office to indicate that the test may proceed.

The process of setting up a test call, as described above, simulates very closely the procedures followed in setting up a normal call from an incoming trunk in the originating office to a desired number in the distant office. Therefore, any irregularities in the operational features while setting up the test call will be detected by the test frame and will result in appropriate trouble indication at the test frame at the originating end.

The automatic testing arrangements will operate with offices where terminating calls are switched either on a terminal net loss (TNL) or on a via net loss (VNL) basis. This is done by including appropriate pads for use at VNL offices.

FEATURES OF INTERTOLL TRUNK TEST FRAME

Control of Test Connector

When the trunk test frame is put into operation, it closes the first crosspoint of the first test connector crossbar switch on the test connector frames to prepare for testing the first trunk in the test sequence. When the trunk test is completed, the test frame advances to the next crosspoint for testing the next trunk. This progression continues through all the test connector frames until all intertoll trunks in the office have been tested. The test frame stops when the test cycle is completed. Particular circuit selection keys are provided on the test frame so that the test connector can be directed manually to any point for testing an individual trunk or for starting a test cycle at some intermediate point in the test sequence, rather than with the first trunk. As the test frame progresses through a test cycle, it also displays on lamps a 4-digit "trunk identification number" corresponding to the test connector crosspoint which is closed. When a trouble is encountered, the attendant uses this 4-digit number to identify the trunk being tested as a particular trunk to a particular destination. The Teletypewriter prints the trunk identification number as a part of each trunk test record.

Busy Test

Before starting a test, the test frame tests the trunk sleeve lead for busy. If the trunk is busy, the test frame waits for the trunk to become idle. A "pass busy" key is provided which, when operated, cancels the waiting period and causes the test frame to immediately pass over busy

trunks to save time. The circuits are arranged so that, if desired, the Teletypewriter may print a record of busy trunks passed over without test. By means of a timing key this record can be delayed two minutes or four minutes to wait for the trunk to become idle. This is used when it is preferable to wait a reasonable time for trunks to become idle to secure tests on a larger proportion of the trunks.

Trunk Classes

A class relay, operated by a contact on the test connector crosspoint as previously mentioned, indicates to the test frame the type of trunk being tested so that it can properly handle the test call. There are 33 of these relays. A flexible cross-connection in the path of the class contact on each test connector crosspoint permits each crosspoint to be assigned to the particular one of the 33 class relays which represent the characteristics of the intertoll trunk associated with that crosspoint. Twenty-eight of the class relays are used in connection with trunks on which automatic transmission tests are made and indicate, among other things, the specified loss of the trunk being tested. These relays are provided in such a manner that, for any trunk, a class can be chosen which agrees with the specified loss of the trunk to within ± 0.1 db over the range 3.8 db to 12.1 db. The specified loss is used by the automatic transmission test and control circuit when computing the deviation of the measured loss from the specified value, as covered later.

Test Cycles

The test frame may be set up by means of control keys to perform various kinds of test cycles. Some of these test cycles are described briefly below.

Code 103 Tests. A complete check is made of all the circuit operating features including the ringing and the supervisory features while the connection is established. If this test is passed successfully, one may assume that the intertoll trunk circuit and the associated signaling channel will properly handle normal calls although this does not prove that the transmission performance is satisfactory.

Signaling Channel Tests. This is an abbreviated code 103 test to verify the integrity of the trunk and its signaling channel. It can be made quite rapidly and is useful for checking the correctness of patching after day and night circuit layout changes.

Pass Idle Test. This is a test cycle which may be run occasionally during a very light load period to detect trunks which may be falsely busy.

Repeat-2 Tests. This consists essentially of two code 103 tests on the same trunk in rapid succession. It is used to insure that a connection through the switching train in the distant office will be properly broken down when a call is completed.

Repeat Test. A "repeat test" key on the test frame cancels the advance of the intertoll trunk test connector. Since the test frame cannot then advance to the next test connector crosspoint, it tests the same trunk repeatedly. This is useful for verifying a trouble condition and for detecting an intermittent trouble, or for obtaining data on stability versus time.

Manual Tests. When the test frame is set up for making manual tests, it engages the office common control equipment to set up a path through the switching train to the trunk to be tested but does not pulse forward code 103 or 104. Instead the attendant can pulse forward the proper code to reach the toll test board at the distant end. This permits dial type trunks to distant offices, not equipped with test lines, to be tested manually.

Code 104 Tests. A 3-position key controls transmission testing. When the key is normal, the test frame makes operational tests. When the key is operated to the transmission and noise position, a two-way transmission loss measurement is made and is followed by a noise check at each end of the trunk. When the key is operated to the transmission only position the noise checks are omitted. The latter position is used when it is permissible to omit the noise checks to save time. When making code 104 tests, the test frame sets up and breaks down the test connection in the same way as when making code 103 tests and, while the connection is established, it receives supervisory signals from the far end. Thus most of the trunk operating features, except for ringing, are also checked as an incidental part of the transmission test. Irregularities in the circuit operating features can result in a trouble indication in the same way as when making operational tests.

Trouble Indications

During a test, progress lamps display the progress of the test call and, when trouble is detected, one of a number of trouble indicating lamps may also be lighted. The progress and trouble indicating lamps indicate the general nature of the trouble.

When the Teletypewriter is not in operation, a trouble indication causes the test frame to stop, to hold the trunk busy and to sound an alarm while awaiting the attention of the attendant. It is the usual practice for the attendant to make a repeat test on the same trunk to

verify the trouble condition. He notes the nature of the trouble from the progress and trouble indicating lamps and then causes the test frame to resume testing by advancing it to the next trunk with a manual advance key.

When the associated Teletypewriter is provided and operating, it is not always necessary to sound an alarm and thus interrupt the regular work of the attendant. Instead the Teletypewriter may print a trouble record. For this purpose troubles are grouped into 18 categories. When a trouble is detected, the Teletypewriter prints a record of the trunk identification number together with a letter in a separate column indicating one of these categories. The test frame then usually makes a repeat test on the same trunk to verify the trouble except when the connection must be held, as discussed later. If the second test is satisfactory, the trouble was of a transient nature and the test frame resumes testing, leaving a single line trouble record on the Teletype tape. If the trouble is still present on the second trial, a second record is printed on the next line for the same trunk.

If the nature of the trouble, as indicated by its category, is such as to render the trunk unfit for service, the test frame will stop after the second trial, hold the trunk busy, and sound an alarm to attract the immediate attention of the attendant. If, however, the trouble is of a minor nature that can be tolerated temporarily, the test frame advances automatically to the next trunk after the second record is printed, and resumes testing without sounding an alarm. By periodic inspection of the Teletype record the attendant can note those trunks needing maintenance attention by means of the double line trouble records. A test cycle can thus be completed with the minimum of supervision on the part of the attendant.

When the nature of a trouble is such that its identity is likely to be lost if the original connection is broken down, e.g., failure of a holding ground, the test frame will not attempt a second trial but stop, hold the trunk busy, and sound an alarm. Failure to complete a transmission test satisfactorily is included in this class because such failures can be due to the testing equipment itself.

AUTOMATIC TRANSMISSION TESTS

Basic Scheme of Measurement

An automatic transmission loss measurement consists essentially of adjusting the loss of a pad at the receiving end of the trunk to bring the test power level at the pad output to a fixed value. A functional diagram of the arrangement is shown in Fig. 4.

The standard one milliwatt source of test power is used at the sending end. The receiving end includes an amplifier, a set of adjustable resistance pads which are relay controlled and an amplifier-rectifier with a measuring relay (M) in its output circuit. Relay (M) is a polarized relay of a type widely used in the telephone plant.

The amplifier has a fixed gain of 19.9 db and it includes considerable negative feedback so that its gain is constant. The pad components are precision resistors to insure accuracy.

The amplifier-rectifier consists of a two-stage amplifier followed by a rectifier tube and a detector tube for controlling relay (M). This circuit is designed so that the margin between the input power which will hold relay (M) operated and the input power which will insure that relay (M) will release is less than 0.1 db. The gain is adjusted, by means of a potentiometer, so that relay (M) will operate when the test power level at the output from the receiving pads in Fig. 4 is one milliwatt or higher and so that it will release when the power level at this point is 0.1 db or more below one milliwatt. This close margin between operate and release permits relay (M) to be used as an accurate measuring device with a precision comparable with that of manual transmission measuring equipment using direct reading meters. Negative feedback, built into the amplifier portion of the amplifier-rectifier, insures gain stability and the amplifier-rectifier will maintain its gain adjustment over a long period.

When making a transmission loss measurement, the power from the sending end operates relay (M) in the amplifier-rectifier. The loss in the receiving pads is then increased, by means of control circuitry, until the power level at their output is reduced to one milliwatt. In making this adjustment, relay (M) is used as the power level indicating device. When this adjustment is finished the trunk loss will be

$$\text{Intertoll Trunk Loss} = 19.9 \text{ db} - \text{Receiving Pad Loss.}$$

Adjustment of Receiving Pads

The receiving pads, shown in Fig. 4 consist of 9 individual pads having losses of 10, 5, 4, 2, 1, 0.5, 0.4, 0.2 and 0.1 db. Each pad is inserted into the input circuit to the amplifier-rectifier by the operation of a corresponding pad control relay. Adjustment of the pad loss takes place in steps.

When relay (M) operates on arrival of the test power, the control circuit operates relay 10 to insert the 10 db pad. If this reduces the test power level at the output from receiving pads to a value below one milli-

watt, relay (M) in the amplifier-rectifier will release. The control circuit then releases relay 10 also to remove the 10 db pad before it proceeds to the next step. If the test power level remains one milliwatt or higher after the 10 db pad is inserted, relay (M) remains operated. The control circuit then locks relay 10 in its operated position to retain the 10 db pad before it proceeds to the next step. In the next step, pad control relay 5 is operated to insert the 5 db receiving pad. The 5 db receiving pad will then be rejected or retained, as described above, depending upon which position relay (M) takes after the 5 db pad is inserted. This process continues until all 9 individual receiving pads have been tried in descending order ending with the 0.1 db pad. When this process is completed, the combination of the 9 pad control relays which remain locked in the operated position determines, additively, the receiving pad loss and consequently, this combination is related directly to the trunk loss. At the originating or control end this combination of operated relays will be translated to the measured loss of the intertoll trunk being tested, when the results of the measurement are recorded. The method of transmitting the measured loss from the far end to the originating end is discussed later.

The transmitting and check pads shown in Fig. 4 are a separate set of pads also controlled by the pad control relays. At the start of the test the total loss in these pads is 19.9 db. Whenever a pad control relay operates to insert a receiving pad, it removes an equal loss from the transmitting pads. Therefore, when the receiving pad adjustment is finished, the loss remaining in the transmitting pads will be equal to the loss of the trunk. Also, the sum of the losses in the two sets of pads is always 19.9 db regardless of the trunk loss being measured, provided all pad components and all pad control relay contacts are in perfect order. This condition permits a precise accuracy check to be made, as discussed later.

Whenever the control circuit leaves pad control relay 4 or 0.4 in its operated position to retain the 4 db or 0.4 db pad, the subsequent 2 db and 1 db or 0.2 db and 0.1 db pad control relays are disabled. There will then be no action as the control circuit passes through the 2 db and 1 db or the 0.2 db and 0.1 db steps. This limits the maximum receiving pad loss to 19.9 db, which is the maximum range of the automatic measurement. This range amply covers the range of losses of intertoll trunks in a usable condition. Loss measurements attempted outside the range of 0 to 19.9 db will cause failure of the built-in checks, mentioned later, and will result in an alarm at the control end of the trunk.

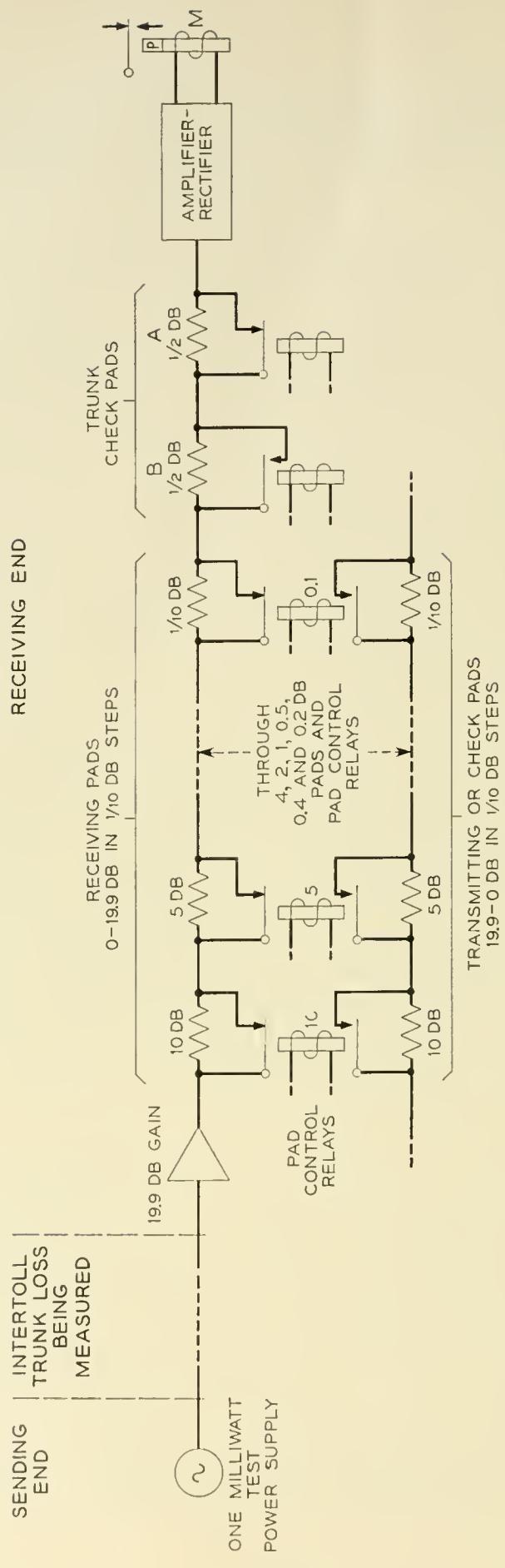


Fig. 4 — Schematic of transmission measurement.

Accuracy Checks

If the receiving pad adjustment has been successful, the power level at the pad output will be very close to one milliwatt and the measuring relay (M) will be just on the verge of moving from its front to its back contact, or vice versa. Errors may creep in, however, to prevent these things from being true. Some such sources of error are:

(1) One or more of the pad control relays might fail to lock in the operated position when they should, or fail to release when they should. It would then be impossible to adjust the total receiving pad loss to the correct value.

(2) The trunk loss might change suddenly while the pad adjustment is in progress and make it impossible, with the pads remaining to be tried, to bring the power level at the pad output to one milliwatt.

(3) The amplifier or amplifier-rectifier gains might increase or decrease due to a defective component.

(4) The milliwatt test power supply might deviate from the standard value.

(5) Defective components or faulty control relay contacts might cause the individual pad losses to be incorrect.

To detect errors of the type in items (1) and (2) a "trunk check" is made immediately after the pad adjustment is finished. Referring to Fig. 4, two 0.5 db pads, A and B, are provided in the input circuit to the amplifier-rectifier, pad A being normally out. Before the sending end removes the test power, pad A is inserted, momentarily. The resulting decrease in input power to the amplifier-rectifier should cause relay (M) to release. Both pads A and B are then cut out. The resulting increase in input power should cause relay (M) to operate. If relay (M) fails to pass either of these checks the receiving pad loss is in error by 0.5 db or more and another trial is needed to secure a more accurate adjustment. Premature removal of the test power at the sending end would, of course, cause relay (M) to fail on the second check and result in another trial.

Immediately after the trunk check and while pad B is still cut out, the receiving end rearranges its circuit locally as shown in Fig. 5 for a "loop check" to guard against errors of the types mentioned in items (3), (4) and (5) above. This rearrangement inserts a 0.3 db pad in place of the 0.5 db pad B, which is cut out. The local milliwatt supply then applies power to the amplifier-rectifier at a level about 0.2 db higher than necessary to operate relay (M). Relay (M) will fail to operate and pass this check if the combined effect of any decrease in the value of the milliwatt test power supply, any decrease in the amplifier and the amplifier-rectifier gains and cumulative errors in the receiving pads and check pads adds more than 0.2 db loss. After the above check, a 0.5 db loss is

added in the looped circuit. This reduces the input power to the amplifier-rectifier about 0.2 db below that which causes relay (M) to release. Relay (M) will remain operated and fail to pass this check if the combined effect of increases in the milliwatt test power supply or amplifier and amplifier-rectifier gains and cumulative errors in the pads exceeds 0.2 db gain. By means of the loop check the maintenance forces will be notified whenever the measuring equipment drifts more than ± 0.2 db from the initially calibrated setting.

After the loop check the receiving end restores its circuit to the original connections shown in Fig. 4 and by means of relays not shown, cuts out all of the receiving pad loss. Relay (M) then reoperates. The circuit rests in this condition to await the removal of the test power at the sending end.

When the sending end removes the test power, relay (M) releases. If all accuracy checks have been passed successfully, the receiving end then prepares for the next phase of the test. If, however, the accuracy checks failed in any respect, the receiving end restores its circuit to the original condition at the start of the measurement and returns a signal to the sending end to request reconnection of the test power for another trial.

Intertoll Trunk Loss Measurement and Noise Check

An intertoll trunk loss measurement consists of two successive one-way measurements, as described above, one for each direction of transmission. The transmission test call is set up to one of the code 104 test lines in the distant office. If the transmission measuring and noise checking circuit at the far end is already engaged because another call arrived

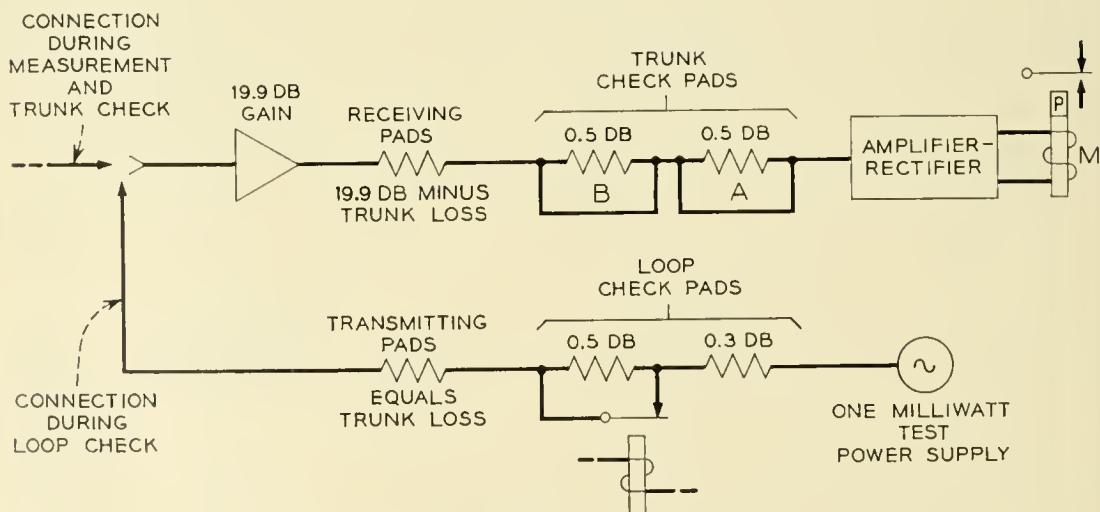


Fig. 5 — Arrangement for loop check.

just previously from some other originating office, this call waits on the test line. When the transmission measuring and noise checking circuit is ready to serve this call, it connects to the test line on which this call is waiting and then returns a steady off-hook signal to the originating end. This notifies the originating end that the transmission test may begin.

The philosophy of a two-way transmission measurement is as follows. The near end sends test power over the trunk and the far end measures the loss as previously described. In this process the loss of the transmitting pad at the far end is adjusted to a value equal to the trunk loss in the near to far direction. The far end then returns test power, first directly over the trunk and next, through the transmitting pad. The power levels received at the near end are a measure of first, the trunk loss in the far to near direction and next, the sum of the losses in the two directions. Measurement of these levels provides data for recording the loss in each direction at the near end.

The two-way transmission measurement takes place in four steps as shown in Fig. 6. These steps are described below.

Step 1

The near end sends one milliwatt and the far end adjusts its pads and checks the measurement. After about 3 seconds the near end removes the test power and then pauses for a short interval to wait for a signal denoting whether or not the accuracy tests were successful.

If they were unsuccessful, the far end will restore itself to the condition prevailing at the start of Step 1 and will also return a short (about $\frac{1}{2}$ second) on-hook signal to the near end. The near end then reconnects the test power for three seconds for another trial. The test frame at the near end stops and sounds an alarm after a third unsuccessful trial.

If the far end is successful in any one of the first three trials, an on-hook signal will not be returned to the near end when the test power is removed. The near end, after the short pause, then sends a short spurt of test power which reoperates the measuring relay at the far end. This signal at the far end, after a successful Step 1, indicates to the far end that this is a full automatic test.

Step 2

For Step 2 the near end connects a far-near amplifier, a set of far-near receiving pads and an amplifier-rectifier. The far end disconnects its receiving equipment and returns one milliwatt over the trunk. The far-near receiving pads at the near end are now inserted in the proper com-

bination to reduce the power level at their output to one milliwatt. The combination of the nine far-near pad control relays remaining operated after the adjustment is finished will be translated later to measured loss of the intertoll trunk in the far to near direction. After about 3 seconds the far end removes the test power and pauses for a short interval. This completes Step 2. Each end then prepares for Step 3.

Step 3

For Step 3 the near end retains the far-near amplifier and the setting of the far-near receiving pads and adds a near-far amplifier and a set of near-far receiving pads in tandem in the input circuit to the amplifier-rectifier. The far end, after the short pause, again sends one milliwatt but this time it sends through the transmitting pad which was adjusted in Step 1 to represent the near-to-far trunk loss. The near-far receiving pads at the near end are now automatically arranged to reduce the power level at their output to one milliwatt.

In the adjustment of Step 2 the over-all loss, including the trunk in the far-to-near direction, the far-near amplifier and the far-near receiving pads, was made 0 db. Consequently, the net loss being measured in Step 3 is simply that of the transmitting pad at the far end, which is the same as the trunk loss in the near-to-far direction. Therefore the combination of the 9 near-far pad control relays remaining operated after Step 3 is finished can be translated to measured loss of the intertoll trunk in the near-to-far direction. After about 3 seconds the far end will remove the test power to complete Step 3 and it will then pause for a short interval before proceeding with Step 4.

At the near end there are also two sets of check pads, not shown, which are associated with the far-near and near-far receiving pads, respectively, as indicated in Fig. 4. During Step 2 and Step 3 the near end makes the trunk check previously described to verify the accuracy of the pad loss settings and, in addition, in Step 3, rearranges its circuit in the manner shown in Fig. 5 for the loop check. Thus at the near end the two sets of check pads, the far-near and near-far amplifiers, and the two sets of receiving pads are all connected in tandem for the loop check.

During the short pause following Step 2 and Step 3 the far end reconnects its amplifier and amplifier-rectifier as shown for Step 1 in Fig. 6. If the near end is unsuccessful in the trunk check in Step 2 or in either the trunk check or loop check in Step 3, it will restore the circuit to the original condition at the beginning of Step 2 and will also send a short spurt of test power to the far end as shown for Step 1 in Fig. 6. This reoperates the measuring relay (M) at the far end momentar-

ily. The far-end then repeats Steps 2 and 3 for another trial. The test frame at the near end will stop and sound an alarm after a third unsuccessful attempt.

If the trunk loss in the near-to-far direction exceeds 10 db, the loss in the transmitting pad at the far end will exceed 10 db. Under this condition the far end will, prior to Step 3, remove 10 db loss from the transmitting pad to increase the test power level on the trunk. This is done to improve the test power level-to-noise ratio and to reduce the error when measuring losses of intertoll trunks having apparatus whose loss is dependent on signal amplitude. The far end will also return to the near end a short on-hook signal. This on-hook signal at the near end, just prior to Step 3, is an "add 10" signal and causes the near end to add 10 db to its loss measurement in Step 3, to compensate for the loss which was removed at the far end.

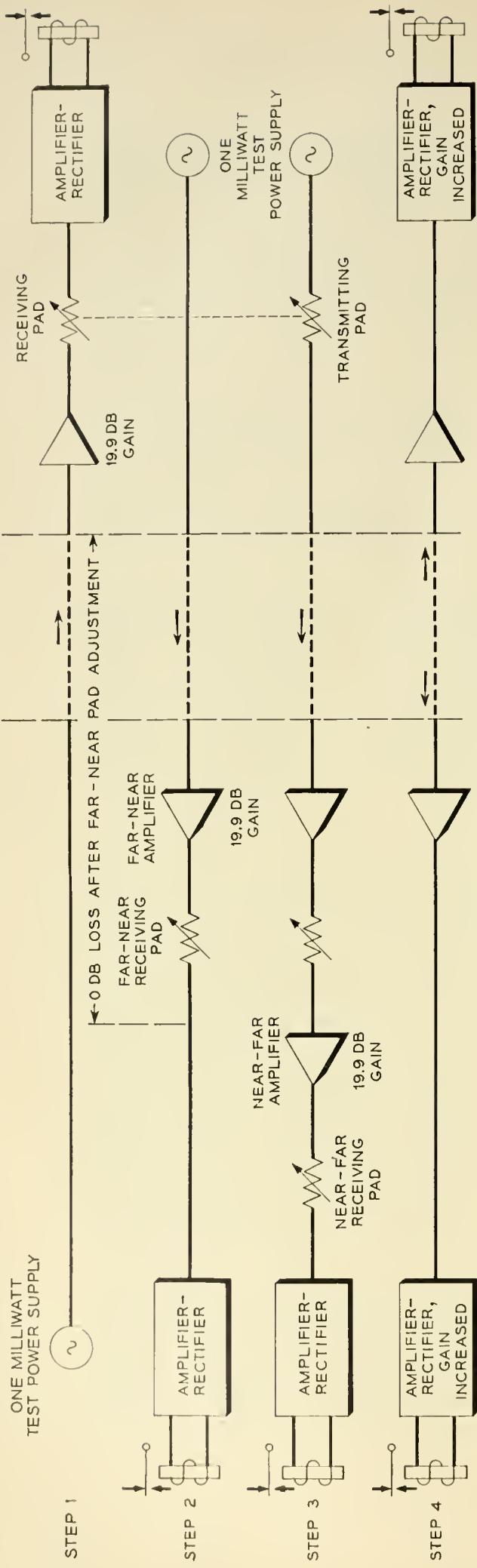
Immediately after Step 3, if the transmission test control key on the test frame is in the transmission only position, the test frame will cause the teletypewriter to record the results of the measurements and will then break down the connection and advance to the next trunk. If the transmission test control key is in the transmission and noise position, the test frame will wait after Step 3 for each end to complete a noise check in Step 4.

Step 4

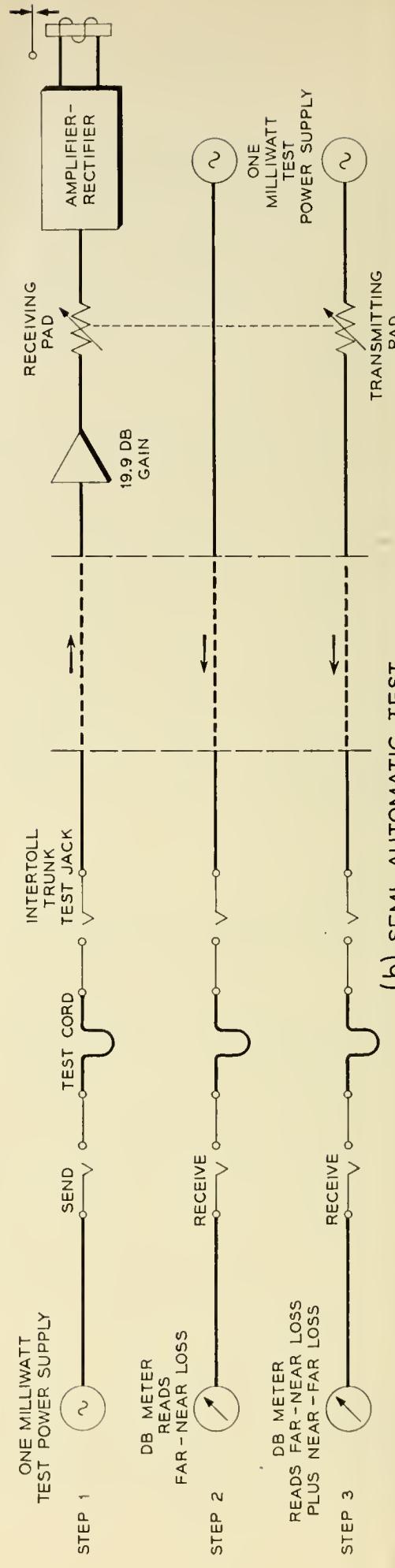
For Step 4 the near-end removes its near-far amplifier and the near-far and far-near receiving pads and increases the gain of the amplifier-rectifier for a noise check at the near-end. Likewise, the far-end removes the receiving pads and increases the gain of the amplifier-rectifier for a noise check at the far end. Each end rests in this condition while the amplifier-rectifier at each end integrates the noise voltage over a 5-second interval. If the integrated value of noise voltage at either end exceeds a pre-determined value, the amplifier-rectifier at that end will operate measuring relay (M) in its output which causes a high noise condition to be registered at that end. If neither end registers a high noise condition, the test call proceeds to completion without a noise indication being recorded at the near end.

When the transmission measuring and noise checking circuit at the far end completes the noise check, it releases itself from the test line and is then free to serve a new call while the test line returns an on-hook signal to notify the originating end that the test is completed. This will be either a steady on-hook signal if the far end has not registered a high

UNDER TEST



(a) FULLY AUTOMATIC TEST



(b) SEMI-AUTOMATIC TEST

noise condition, or a 120 IPM flashing signal if the far-end has registered a high noise condition. The near end is thus advised of the results of the noise check at the far end. The test frame, on receipt of this signal, causes the Teletypewriter to complete the record and then breaks down the connection and advances to the next trunk.

The amplifier, which precedes the amplifier-rectifier includes a network which provides F1A noise weighting during the noise check. The amplifier-rectifier is adjusted in the noise checking condition (that is, when its gain is increased) so that a noise indication will be given when the noise exceeds about 35 or 40 or 45 dba. Since this test is intended only as a rough check to detect any abnormal noise condition, the noise rejection limit used in any given office will be governed by the types of intertoll trunk facilities terminating in that office. No correction is made for the measured loss of the trunk at the time of the noise check, hence the noise is checked at the receiving switchboard level. For the usual types of noise the results of the noise check agree roughly with those which would be obtained by an average observer using a 2-B Noise Measuring Set for a similar "go-no go" type of check.

As is evident from the previous description, each end is expected to complete the various steps of its functions within allotted time intervals. Timing intervals at the far end are controlled by a multivibrator circuit. Timing at the near end is controlled by a similar multivibrator in the intertoll trunk test frame. To insure that the test circuits always perform as they should and that the timing circuits are functioning properly, checks are built into the circuits so that anything which prevents the successful completion of a 2-way measurement on schedule causes the automatic outgoing intertoll trunk test frame at the near end to stop, hold the trunk busy and sound an alarm while awaiting attention of the attendant. The transmission measuring and noise checking circuit at the far end will, however, release itself from the test line so that it will be free to handle other calls.

Semi-Automatic Test

One-milliwatt test power supply outlets have been provided in toll offices for some time for making a one-way transmission measurement frequently referred to as a code 102 test. A test board attendant can reach the one milliwatt test power supply by pulsing forward code 102 or by requesting an operator at the distant end of a manual trunk for a connection. The test power is applied at the distant end for about 10 seconds during which time the attendant measures the loss in the receiving (far-to-near) direction. This is a fairly fast semi-automatic test but, of

course, has the disadvantage that it is a one-way test and cannot be used for all purposes.

In order to provide a semi-automatic two-way test, the far-end equipment is arranged so that a test board attendant can make a code 104 measurement unassisted. This measurement is carried out in 3 steps as shown in the lower portion of Fig. 6.

Step 1

The attendant connects a test cord to the test jack of the intertoll trunk and pulses forward code 104 using his test position dial or key set. When the far end is ready, it returns an off-hook signal which retires the test cord supervisory lamp. He then connects the other end of the cord to the one milliwatt test power supply. The far end then adjusts the receiving and transmitting pads in the same way as for a full automatic test. After about 3 seconds the attendant disconnects the test power and at that time observes the cord supervisory lamp; a single flash indicates that the far end was unsuccessful and is requesting a second trial. If the supervisory lamp remains steadily dark he connects the cord to the receive jack of his transmission measuring circuit to prepare for Step 2.

Step 2

The far end will pause about 2 seconds after the attendant removes the test power to give him time to prepare for Step 2. During this pause the far end will not receive a short spurt of test power as in the case of a full automatic test. Consequently, after the 2 second interval the far end will return one milliwatt for 10 seconds on a semiautomatic test to give the attendant time to complete a measurement. The received power is read directly on the meter of the transmission measuring circuit and is the loss in the far-to-near direction. When the far end removes the test power, the meter reading drops back to the position of no current (infinite loss) and at that time the attendant observes the cord lamp. A single flash at this time is an "add 10" signal and indicates that 10 db should be added to the next measurement. A steady dark lamp indicates that the next measurement should be recorded without correction.

Step 3

After about 2 seconds delay to give the attendant time to record the first measurement, the far end again returns 1 milliwatt, this time through the transmitting pad set up in Step 1. The meter now reads the

loss of the trunk plus the loss of the transmitting pad at the far end. Since the transmitting pad loss equals the trunk loss in the near-to-far direction, the difference between the measurements in Step 3 and Step 2 is the trunk loss in the near-to-far direction. After about 10 seconds the far end removes the test power and starts the noise check in the same way as if this were a full automatic test.

When the far end removes the test power after Step 3, the attendant leaves the connection intact until the cord supervisory lamp lights to indicate completion of the noise check at the far end. A flashing lamp indicates that the noise at the far end exceeds the prescribed limit and a steadily lighted lamp indicates the noise at the far end is below this value. A noise test at the near end may be made by the attendant if he judges, after a listening test, that a noise test is desirable. For this test he uses the standard noise measuring equipment.

PRESENTATION OF TEST RESULTS

When making operational tests and a Teletypewriter is not being used, troubles are registered by means of an audible alarm and accompanying display lamps. When making transmission loss measurements, however, a complete record of the measurements on all trunks tested, both good and bad is frequently needed. A Teletypewriter then becomes a practical necessity; otherwise the attendant would be required to supervise the automatic equipment continuously and to record, from a lamp display or similar indication, the results of each measurement as it was made. Having provided the Teletypewriter for transmission testing, its ability to print letters to represent trouble indications is utilized to avoid halting the progress of the tests when operational troubles are experienced, except when completely inoperative conditions are encountered.

Computer Circuit

As mentioned earlier intertoll trunk transmission performance is rated in terms of bias and distribution grade which are calculated from the deviations of the measured losses of the intertoll trunks from their specified values. For such calculations the maintenance forces are, therefore, more interested in the deviations than they are in actual measured losses. Accordingly, the automatic transmission test and control circuit at the near-end has a computer built into it which will compute the deviation for each measurement so that the deviation can be recorded by the Teletypewriter.

The computer is a bi-quinary relay type adder similar to those used

for other purposes in the telephone plant, for example, in the computer of the automatic message accounting system. It obtains the specified net loss of the trunk being tested from the class relay which remains operated throughout the test. When a computation is to be made of the deviation in the far-to-near direction, for example, the control circuit extends to the adder a number of leads from the contacts of the far-near pad control relays. Some combination of the 9 far-near pad control relays remains operated after the far-near pad adjustment is finished and therefore some combination of the leads extended to the adder will be closed. These leads furnish to the adder the measured loss of the intertoll trunk in the far-to-near direction. The adder then subtracts the specified loss from the measured loss and presents the answer together with the proper sign, + or -, to the teletypewriter for a printed record. The deviation in the near-to-far direction is computed in the same manner by extending corresponding leads from the near-far pad control relays to the adder at the proper time.

Deviation Registers

In determining bias and distribution grade by the method discussed in the companion article,* the deviations from specified net loss are calculated for each measurement. These deviations are grouped together in 0.5 db increments from +8 db to -8 db, all deviations exceeding +7.8 db or -7.8 db being considered as +8.0 db and -8.0 db respectively. For example, all deviations of +0.3 db to +0.7 db, inclusive are considered to be +0.5 db and are so tallied on the data, or stroke, sheet.

To assist in this work the automatic test equipment includes thirty-three manually resettable counters corresponding to the 0.5 db increments from +8.0 db to -8.0 db inclusive. Just prior to a transmission test cycle all these counters are reset to zero. At the time a deviation computation is made, the computer also causes the proper counter to register one count. After the test run on a group of trunks, the counter readings can be transcribed directly as the final tally on the stroke sheet and may be used to determine the bias and distribution grade. A "total tests" counter keeps a tally of all the computations. At the end of the test run the total count serves as a check of the total count of the other 33 counters.

Check for Excessive Deviations

In addition to obtaining data for the calculation of bias and distribution grade, the maintenance forces would also like to know promptly

* H. H. Felder and E. N. Little, Intertoll Net Loss Maintenance Under Operator Distance and Direct Distance Dialing, page 955 of this issue.

when the loss of an intertoll trunk deviates an abnormal amount from its specified value. The maintenance practices currently require that, whenever an intertoll trunk is found to have a deviation of ± 5 db or more in either direction, the trunk should be removed from service immediately and the cause of the abnormal deviation corrected. Accordingly, the computer circuit includes an alarm feature which sounds an alarm to attract the immediate attention of the attendant whenever the computed deviation is ± 5.0 db or greater.

The maintenance forces may also like to know promptly about trunks with wide deviations but which are not so bad as to require immediate removal from service. For this purpose the computer also includes a limit checking feature. This can be set, by means of optional wiring, to detect deviations in excess of ± 3.0 db, ± 4.0 db or ± 5.0 db. Whenever a deviation exceeds the limit for which the computer is wired, this feature performs as follows:

- (1) When the Teletypewriter is not in operation the test frame stops and sounds an alarm.
- (2) When the Teletypewriter is recording all measurements, the letter U is added in a separate column at the end of the test record. The letter stands out on the record to permit quick spotting of trunks with abnormal deviations.
- (3) By means of a control key, a transmission test record can be printed only for those trunks whose deviation exceeds the computer checking limit or which are "noisy" at either end.

Teletypewriter Record

The Teletypewriter is put into operation by means of a key on the test frame. When this key is normal, no records are printed. Under this condition a trouble causes the test frame to stop and sound an alarm. When the Teletypewriter is operating it prints various records and a minor operational trouble may result only in a record, without an alarm. Each record occupies a separate line on the tape. Each line starts with the four-digit trunk identification number in the first column. Fig. 7 shows a short specimen of the Teletypewriter record.

When the pass busy key on the test frame is in its nonoperated position, the Teletypewriter will print the trunk identification number, followed by the letter B, for each trunk passed over without test because it was busy. This is done on both operational and transmission test cycles. When the pass busy key is operated no record is made of busy trunks passed without test.

During operational tests no record is printed for trunks which are

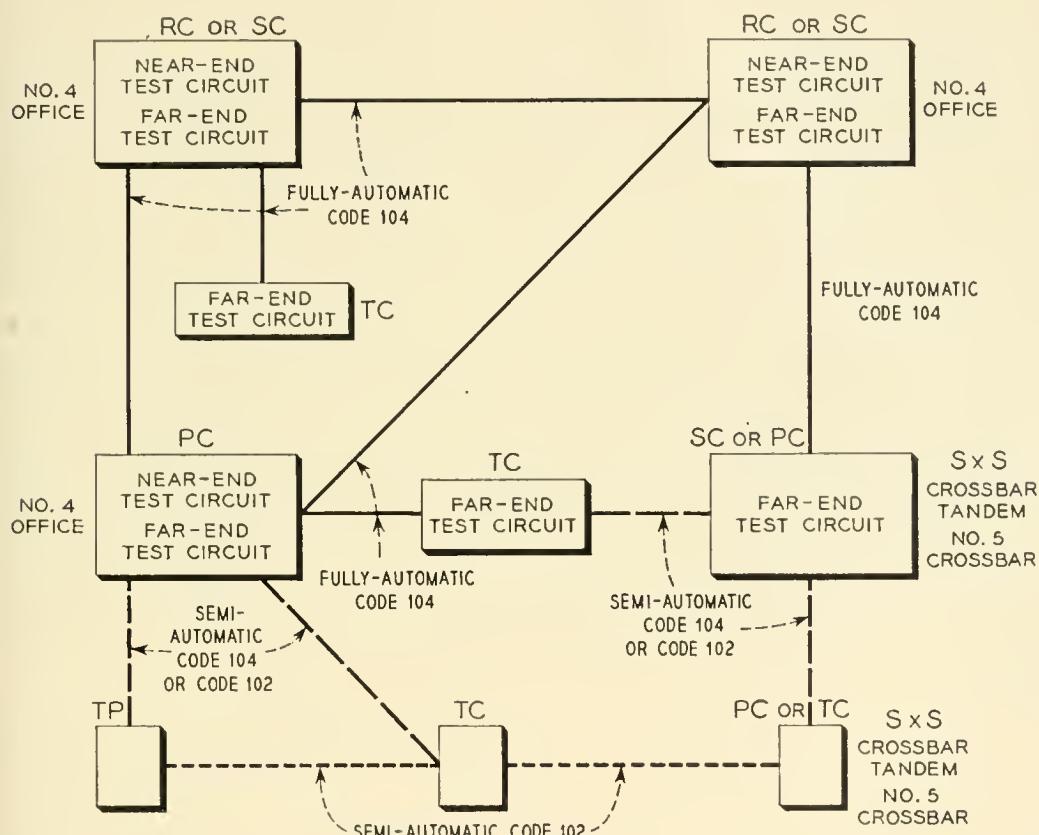
satisfactory. Troubles during either operational or transmission test cycles result in a record of the trunk identification number followed by a cue letter in a separate column denoting the nature of the trouble. This may be a single line record or a double line record for a repeat test on the same trunk as previously discussed. For example, in Fig. 7, the letter Y in the second line indicates that on trunk 1267 the far end was unable to complete its transmission measurement successfully. The letter A in lines 3 and 4 indicates that the test frame was unable to establish a connection over trunk 1293 on either its first or second attempt. The record of transmission tests is printed in several columns. Reading from left to right (see Fig. 7) these are (1) trunk identification number, (2) specified net loss, (3) deviation in the far-to-near direction together with the sign, and (4) deviation in the near-to-far direction together with the sign. In columns 2, 3, and 4 the decimal points are omitted and the ten's digits are omitted when they are zero (0). Column 5 will contain an N if the far end is "noisy" or the letter U if the deviation in the far-to-near direction exceeds the computer check limits, preference being given to N if both conditions occur on the same trunk. Likewise column 6 contains an N if the near end is "noisy" or a U if the deviation in the near-to-far direction exceeds the computer check limits. Transmission test cycles will, of course, include a trouble record whenever an operational trouble is encountered or whenever the transmission test cannot be completed successfully.

TRUNK IDENTIFICATION NUMBER	BUSY OR TROUBLE CUE	SPECIFIED LOSS	DEVIATION, FAR TO NEAR DIRECTION	DEVIATION, NEAR TO FAR DIRECTION	ABNORMAL DEVIATION OR NOISE CUE
1234	B				
1267	Y				
1293	A				
1293	A				
1376		72 + 08 - 04			
1377		75 - 11 + 07		N	
1378		75 - 52 + 07		U	
1379		75 + 07 - 51			U

Fig. 7 — Teletypewriter test record.

APPLICATION

Maximum benefits can be derived from the automatic testing equipment by locating it at points having a considerable number of intertoll trunks. This suggests that the near-end installations be placed in offices in larger cities and far-end installations be placed at points with enough trunks available to near-end equipment to justify the far-end equipment. As has been indicated the far-end equipment can operate with either an automatic transmission test and control circuit or with a test board attendant at the opposite end. Therefore, once an office has been supplied with far-end test equipment, all incoming and two-way dial type intertoll trunks from offices provided with near end equipment can be tested on a fully automatic basis and all incoming and two-way dial type intertoll



NOTE:

IT IS ASSUMED THAT CODE 102 MW SUPPLY CIRCUITS WILL BE AVAILABLE AT ALL OFFICES AND CAN BE USED, OR THAT TEST BOARD TO TEST BOARD MEASUREMENTS CAN BE MADE WHEN DESIRED

LEGEND

RC = REGIONAL CENTER
SC = SECTIONAL CENTER
PC = PRIMARY CENTER
TC = TOLL CENTER
TP = TOLL POINT

Fig. 8 — Typical layout for automatic testing.

trunks from other toll offices can be tested on a semi-automatic basis from the toll test board in the distant office.

Fig. 8 shows a possible application of automatic test circuits. In such an application, all No. 4 type toll crossbar offices would have both near-end and far-end equipments. Other offices would have far-end equipment only when they have a sufficient number of direct trunks to No. 4 type offices to justify its use. The several types of tests which would be possible are indicated in the illustration.

It can be seen that a well distributed number of near-end and far-end test circuits will make it possible to test automatically a large percentage of the intertoll trunks throughout the country. This is particularly true in the more populous sections, where the concentration of trunks results in the probability of toll centers having trunks to more than one office furnished with near-end equipment.

ACKNOWLEDGMENTS

Automatic intertoll trunk testing arrangements, including transmission tests, are the result of the ideas, efforts and experiences of many people concerned with intertoll switching and maintenance problems throughout the Bell System. Mr. L. L. Glezen and Mr. L. F. Howard deserve particular mention in this regard. Specific credit should also be given to Mr. B. McKim and Mr. T. H. Neely for the basic scheme of two-way transmission measurements and accuracy checks and to Mr. C. C. Fleming for the design of the amplifier and amplifier-rectifier. Appreciation is given to various departments of the American Telephone and Telegraph Company for their assistance during the development and trial of this equipment. Mention should also be made of the hearty cooperation and aid given by the A.T. & T. and Associated Company plant forces during the field trial of automatic transmission testing.

Intertoll Trunk Net Loss Maintenance under Operator Distance and Direct Distance Dialing

By H. H. FELDER and E. N. LITTLE

(Manuscript received March 15, 1956)

Nearly all of the components of an intertoll trunk contribute in some degree to its variations in transmission loss. Automatic transmission regulating devices in carrier systems and in many voice-frequency systems control inherent variations in the intertoll trunk plant. These variations in transmission come mainly from unavoidable causes such as temperature changes. The success of these devices depends on how precisely the trunk is lined up and the manner in which the maintenance adjustments are made. When the nationwide dialing plan with automatic alternate routing is in full swing, maintenance requirements will be more severe because of the material increase in switched business and the number of possible links in tandem, and because operator checks will not be obtained on most calls. Therefore, the maintenance forces will have to keep closer watch on intertoll trunk transmission performance and insure that the necessary adjustments are made in the right places. This article discusses some of the maintenance techniques now used and suggests fields for further study.

TABLE OF CONTENTS

	Page
Introduction	956
The Problem of Net Loss Maintenance	956
Effect of Switching Plans	957
Manual Operation	957
Dial Operation	958
Effect of Carrier Operation	960
Table I	960
Quantitative Aspects of the Problem	962
Table II	963
Use of Transmission Loss Data	964
Procedure for Analyzing Measurements	965
Effectiveness of Over-all Trunk Test and Analysis	969
Simple Layouts	969
Complex Layouts	970
Need for Education	971
Summary and Conclusions	972

INTRODUCTION

Currently there are over 230,000,000 long distance calls made in the Bell System per month. They range from relatively simple connections involving a single intercity trunk to complex connections involving several intercity trunks in tandem, perhaps totaling 4,000 miles in length. In each case there is a toll connecting trunk at each end. Almost half of this traffic involves distances over 30 miles. The transmission engineer's problem is how to provide uniformly good and dependable transmission so that every one of these calls will be satisfactory to the customers involved. To accomplish this requires among other things that:

1. The design loss of every trunk must be the lowest permissible from the standpoint of echo, singing, crosstalk and noise.
2. The actual loss of every trunk must be kept close to the design loss at all times.

Meeting the first requirement is a matter of system design and circuit layout engineering. The factors involved have been covered in a previous article.¹ Meeting the second requirement is an important function of the maintenance forces and is discussed in this article.

THE PROBLEM OF NET LOSS MAINTENANCE

The transition from manual operation under the "general toll switching plan"² to dial operation under the "nationwide dialing plan"^{3, 4} is requiring material changes in intertoll trunk design and also in techniques for maintaining these trunks. While precise maintenance is becoming increasingly necessary, it is also becoming more difficult to achieve. There are three important reasons for this.

First, the nationwide dialing plan increases both the possible number of trunks used in tandem for a given call and the variety of the connections in which any particular trunk may be used. This increases the chances of impairment due to deviations from assigned loss in individual trunks since these deviations may combine unfavorably in multi-switched connections. To minimize this, the transmission stability of the individual trunk links must be better than under the old plan.

Second, more and more of the trunks are being put on carrier because

¹ H. R. Huntley, Transmission Design of Intertoll Telephone Trunks, B.S.T.J., Sept. 1953.

² H. S. Osborne, A General Switching Plan for Telephone Toll Service, B.S.T.J., July, 1930.

³ A. B. Clark and H. S. Osborne, Automatic Switching for Nationwide Telephone Service, B.S.T.J., Sept., 1952.

⁴ J. J. Pilliod, Fundamental Plans for Toll Telephone Plant, B.S.T.J., Sept. 1952.

it is the best solution to the transmission and economic problems. However, carrier involves many more variable elements and requires higher precision of adjustment than voice-frequency systems need. These increase the difficulty of maintaining trunk losses close to design values on a day-by-day basis.

Third, as operator distance and direct distance dialing grow, there is constantly diminishing opportunity for operators to detect and change unsatisfactory connections or to report unsatisfactory transmission conditions to the appropriate testboards for action.

Thus the maintenance problem is in two parts:

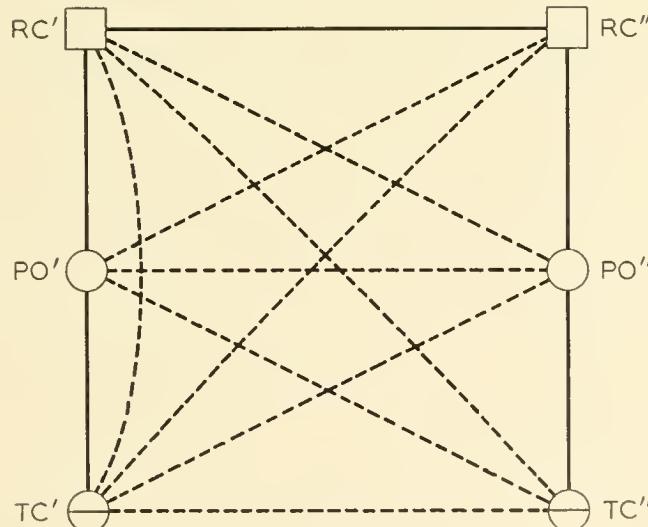
1. How can we reduce departures from design standards even in the face of increasing complexity of plant?
2. What substitute can we find for operator detection of troubles, and can we find even better means of detection?

The ways in which switching plans and the use of carrier reflect upon the problem of trunk net loss maintenance is discussed in more detail in the following sections.

EFFECT OF SWITCHING PLANS

Manual Operation

For many years long distance traffic has been handled on a manual basis under the "general toll switching plan" illustrated in Fig. 1. Between two points indicated by toll centers, TC' and TC'' , it was theoretically possible to get as many as five trunks in tandem. This rarely occurred be-



TC = Toll Center PO = Primary Outlet RC = Regional Center

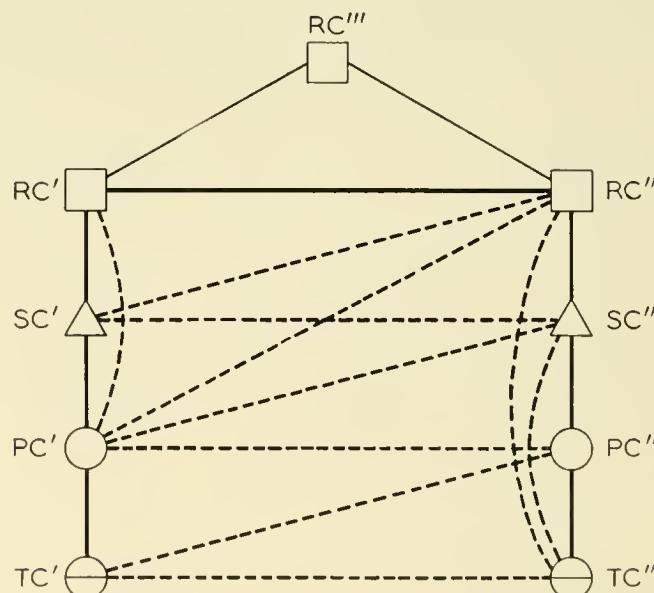
Fig. 1 — General toll switching plan — manual operation.

cause handling switched connections manually was so complicated and expensive that direct trunks were provided wherever they were economical and alternate routes were assigned and used sparingly. The result was that the manual switching plan was characterized by a minimum of switching.

Under manual operation, operators passed information over every trunk in the connection, as well as over the completed connection, before it was turned over to the customers. If anything was radically wrong with a trunk, the operators recognized it and substituted another trunk. When this was necessary, they could report the defective trunk to the appropriate testboard for action. Under these conditions, if trunk losses wandered appreciably from their specified values, the consequences were seldom serious.

Dial Operation

With dial operation, not only is the plan more complex (as shown on Fig. 2), with an abundance of alternate routes, but intertoll trunk switching is so fast and reliable that the number of switching points has little effect on speed of service. Thus the dial operating plan can take full advantage of alternate routing and the use of trunks in tandem will occur much more frequently than with manual operation.



TC = Toll Center

PC = Primary Center

SC = Sectional Center

RC = Regional Center

Fig. 2 — Nationwide dialing plan — dial operation.

Here again the alternate routing follows a definite plan.⁵ As shown in Fig. 2, a call from toll center, TC', to toll center, TC'', will follow the direct route, if it is available and not busy. A second choice will be via a higher ranking office in the chain from TC'' to the regional center, RC''. A third choice may be available to a still higher ranking office. Thus, if the originating office cannot use its direct route, the call will be advanced over the alternate routes according to a predetermined pattern. If all other alternatives fail, the call will follow the heavy solid line 7-link route shown on Fig. 2, or, in special cases, the 8-link route via RC'''.

These attempts involve many operations but the automatic equipment completes them quickly. This makes it feasible to provide small, high usage, direct trunk groups between two points, with the realization that in busy periods alternate routes can handle the overflow traffic with negligible time delay. Thus over a good part of the day, the direct trunks or first choice trunks will handle the traffic. In the busy periods, use of alternate routes with a number of links in tandem will be a frequent occurrence. Therefore, it is important to have losses on alternate routes not greatly different from those on direct routes so customers will not experience noticeable contrasts.

Operators will seldom talk to each other over the complete connection, and even less over the individual trunks. Only on person-to-person or collect calls will they talk even to the called party. On station-to-station calls they merely dial or key up the desired number and rely on supervisory signals to disclose the progress of the call.

On operator dialed calls, the operator may sometimes pick up the intertoll trunk in her switchboard multiple, but in many cases she will reach it over a tandem trunk. In the former case she can identify the intertoll trunk forming the first link in the connection but assistance would be needed at intermediate testboards to identify succeeding trunks. In the latter case, testman assistance would be required at the originating office in order to identify even the first trunk of the connection. In either case the need for holding the customer's line during identification, to avoid breaking down the connections makes such means of identification impracticable with presently available techniques.

On direct distance dialed calls there are no operators involved and present means of identification of trunks in trouble after the connection has been established are even more impracticable. This is because the calling party must release the connection before he can report a trouble, thus destroying any possibility of trunk identification.

⁵ R. I. Wilkinson, Theory for Toll Traffic Engineering in the U. S. A., B.S.T.J., March, 1956.

Thus under dial operation there is a need for better trunk stability. Therefore, a greater burden is placed on the plant forces to locate unsatisfactory trunks so that proper maintenance action can be taken before customers experience difficulty.

EFFECT OF CARRIER OPERATION

Carrier is the principal transmission instrumentality which makes it possible to go ahead with nationwide dialing with assurance that people can talk satisfactorily over the complex connections set up by the switching systems. But it brings with it formidable problems of maintenance. The high attenuation per mile of the line conductors at carrier frequencies increases the number of variable elements as well as the precision with which they must be adjusted. The interrelation between the elements adds to the complication.

Table I illustrates this by giving some figures comparing 100 miles of a voice-frequency cable trunk with 100 miles of a typical trunk on K carrier, which is widely used on cable facilities. The figures apply in both cases to one direction of transmission.

The ten-to-one ratio in the number of electron tubes represents a greater chance of trouble developing in the carrier trunk due to aging or failure of electron tubes. In the carrier trunks there are more automatic adjustable features. For instance, in a typical K2 carrier system there are five flat gain regulators and one twist regulator in one twist section of approximately 100 miles, against a single regulator in a voice-frequency trunk 100 miles long. These regulators are depended upon to keep the loss variations to tolerable amounts. Any malfunction can have a serious effect on trunk loss. Furthermore, they must be adjusted to the desired regulating range and therefore they are points at which maladjustments may be made.

The channels of any one carrier system or of a 12-channel group are commonly routed by the circuit layout engineers to a number of terminal

TABLE I

	V-f Trunk	K2 Carrier Trunk
Total Conductor Loss -db.....	35	378
Gain Required to Reduce to Via Net Loss -db.....	31	377.4
Percentage of Line Loss Represented by a 2 db Variation.....	5.7	0.53
Number of Electron Tubes	3	28
Number of Amplifiers.....	3	7
Number of Automatic Regulators.....	1	6

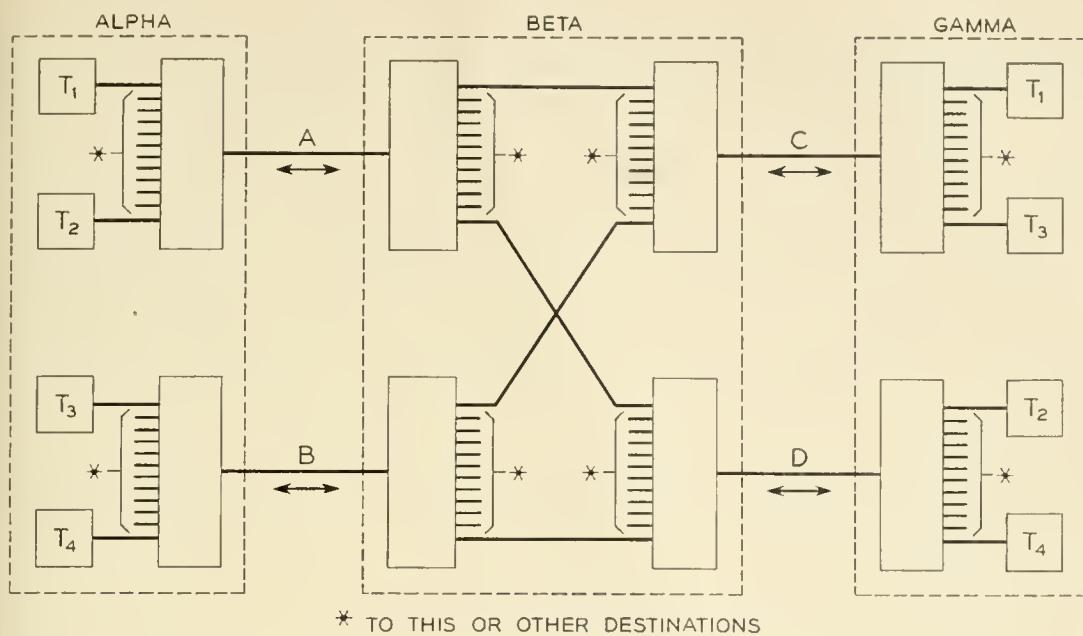


Fig. 3 — Typical carrier channel assignments.

points even though circuit requirements to a given point are sufficient to utilize 12 or more channels. This is done to minimize the chances that all of the trunks between two points will be interrupted by a system failure. A simple case is illustrated by Fig. 3 which shows trunks between Alpha and Gamma connected at an intermediate point, Beta, in such a manner that a failure in any one of the systems A, B, C, or D will affect only half the trunks.

This routing problem, however, complicates the maintenance problem. For example, if trunk T1 were found to have excess loss in the Alpha-Gamma direction it could be corrected by raising the channel gain at Gamma. On the other hand, a correct diagnosis might have disclosed that the trouble was due to a repeater in system A. If this were the case, merely compensating for the excess loss in T1 by changing the channel gain would still leave all other trunks associated with system A in trouble. Later on, if the repeater difficulty were corrected, and no further action were taken, the net loss of T1 would then be too low.

Thus, the flexibility which is so desirable to minimize interruptions of whole circuit groups leads to a difficult problem in the administration of trunk loss adjustment and maintenance. Furthermore, because of the larger numbers and greater dispersion of trunks and terminal points, the situation in the actual telephone plant is much more complex than in the above example. Also, the diagnosis of trouble conditions is made more difficult by the normal variations of channel losses in the carrier systems and consequently of the trunk losses about their design values. This can

be better appreciated when some quantitative aspects of the problem are considered.

QUANTITATIVE ASPECTS OF THE PROBLEM

When the nationwide dial switching plan began to take shape some 8 or 10 years ago, intensive study of the transmission maintenance problem was undertaken. The existing situation was examined to determine whether or not the plant would continue to be satisfactory under the changed conditions. This was done by analyzing the results of many thousands of transmission measurements which had been made on a routine basis in toll test rooms all over the Bell System. Both the measured and the assigned losses were available so the differences between them could be derived and analyzed statistically.

Although the distribution of differences expressed in db for an office did not necessarily follow precisely a normal probability law, the distributions were close enough to normal law so that they could be treated as normal. The results were similar throughout the System. The differences within an office were random as also were the means of the differences from office to office. However, the means tended to be biased in the direction of excess loss. The performance of trunks in multi-link connections which would be set up by the switching machines could therefore be estimated with reasonable accuracy. In the statistical analysis of measurements on the group of trunks, the performance was expressed in terms of "distribution grade" and "bias." In telephone transmission maintenance terminology, bias is the algebraic average of the measured transmission departures in db from individual specified net losses for the group of trunks. The distribution grade is the standard deviation of the differences between measured and specified trunk losses about this bias value. The distribution grades found in these studies were about as follows:

For trunks under 500 miles — about 1.8 db.

For longer trunks — about 2.5 db.

Table II illustrates the effects of the distribution grades on connections involving various combinations of these trunk links, assuming that bias can be neglected.

The design loss objective for a 4-link connection, say 1,000 miles long, is about 7 or 8 db (including 2 db of connecting trunk or pad loss at each end). Table II shows that, in an appreciable percentage of the 4-link connections involving the above type of plant, the variations can be expected to exceed the design loss. Variations of this magnitude can result in transmission impairment due to echo, hollowness, singing, crosstalk,

TABLE II

Number of Intertoll Trunks in the Connection	2	4*	6*	8*
Distribution Grade in db	2.5	4.4	5.0	5.6
Per cent of Connections Departing from Average				
± 2 db or more	42	65	69	73
± 4 db or more	11	36	42	47
± 8 db or more	0.2	7	11	15

* Includes two trunks over 500 miles long.

noise or low volume. Furthermore, undesirable contrast may be encountered on successive calls between the same two telephones.

The results of the study as well as experience with the beginning of automatic alternate routing show that the performance of the existing trunk plant must be improved. Three immediate objectives have been set:

1. Reduction of distribution grades to about $\frac{1}{2}$ of the values mentioned above, i.e., about 1.0 db.
2. Maintenance of office bias within ± 0.25 db.
3. Removal from service of individual trunks differing widely from their design losses (in the order of 4 or 5 db).

To achieve these objectives requires effort along four lines. First, systems should be designed to have sufficient stability once they are adjusted. This involves the inclusion of stable circuit elements and the provision of automatic regulating devices to compensate for unavoidable transmission variations arising from natural causes. These features have been applied to existing systems within limits imposed by economic considerations and the state of the art. Further extension of these features will be required in the future in order to meet the above objectives.

Second, before a trunk is placed in service, each of its component parts and the over-all trunk should be adjusted to give the correct loss. From the transmission maintenance point of view, it is extremely important for each trunk to start out with all of its adjustments correctly made.

Third, existing and incipient troubles, and deterioration or maladjustment of components, must be detected and corrected by routine maintenance of individual systems used in making up trunks. Such activity must make up for the inability to design systems to have the desired stability.

Fourth, significant departures from trunk design losses must be detected by over-all transmission measurements, and must be corrected be-

fore service reactions occur. Such measurements will also be of aid in determining the effectiveness of efforts along the first and third lines.

As discussed earlier, the presence of the operator on every call was of material assistance in the detection of unsatisfactory trunks. On operator or direct distance dialed calls, there will be little or no operator conversation over the intertoll trunk connection. As a substitute, the maintenance forces may need to make more frequent checks of the transmission performance of the trunks unless the stability of individual systems and components of systems can be improved. Manual methods have been used by the maintenance forces in the past to measure trunk losses. Semi-automatic measuring methods have been developed to reduce the time and effort required. In many cases the necessary number of measurements will be economical only when made by automatic devices. One form of such gear is described in a companion paper.⁶

The ability to measure over-all trunk losses simply and frequently is of direct aid in detecting when loss deviations exceed maximum tolerances. Such measurements in themselves, however, are insufficient to detect incipient troubles or to indicate the component part responsible for unsatisfactory transmission. An attempt has been made to achieve these objectives by using statistical analysis of the measured data as an aid to diagnosis. The following sections discuss the application of such analysis.

Use of Transmission Loss Data

It has been shown that considerable variation can be expected in trunk losses even in the absence of trouble conditions. For any given group of trunks selected for analysis, the performance is described by the distribution grade and the bias. If a group of trunks is found to have bias, it is usually an indication of some assignable cause. One such cause might be a change in gain of an amplifier common to the group. Another cause might be improper gain adjustment for channel units of a carrier terminal associated with the group.

If a group of trunks is found to have a greater distribution grade than the distribution grade for all the trunks in the office, this may indicate excessive instability in a component part common to the trunks in the group. If analysis of all the trunks terminating in an office shows a higher distribution grade than is usually found in similar offices, the fault may be due to maintenance routines being inadequately or improperly applied.

⁶ H. H. Felder, A. J. Pascarella and H. F. Shoffstall, Automatic Testing of Transmission and Operational Functions of Intertoll Trunks, page 927 of this issue.

Statistical analyses must thus be made of data for small groups as well as for large groups of trunks. Furthermore, the groups which are studied must have elements or factors in common in order for the statistics to have significance. Analyses of periodic measurements of losses for the same trunk or groups of similar trunks can likewise indicate significant changes in performance.

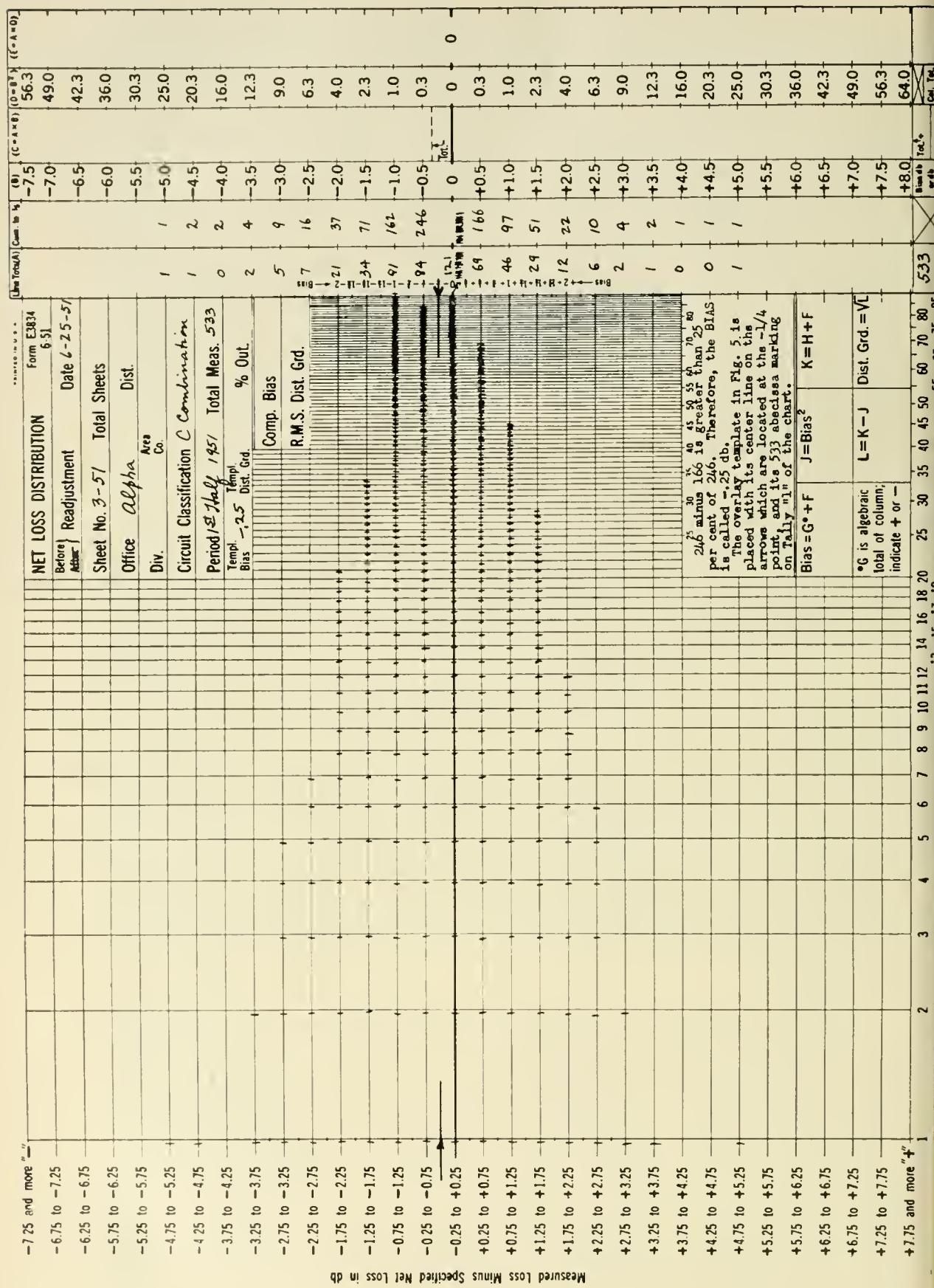
As yet, the problems of properly selecting the trunks to be analyzed and of correlating the results of the analyses with particular system elements needing maintenance attention have been solved only partially. In addition to the need for proper procedures, there is the need for thorough training of maintenance personnel. The complexity of the telephone plant today is increasing the importance of all maintenance personnel having a thorough knowledge of how individual systems function and how the performance of the various system elements reacts upon overall trunk performance.

Procedure for Analyzing Measurements

In an effort to facilitate the application of statistical analysis of trunk performance by plant personnel, a special data sheet and associated templates have been devised. These are shown in Figs. 4, 5, and 6. The method of analysis gives only approximate results but has been found to be sufficiently accurate for reasonably large amounts of data. It is simple, rapid and easily comprehended by the plant personnel. The procedure to be followed consists first of subtracting the specified loss from the measured loss for each of the trunks under study. A stroke is placed on the chart for each of the resulting deviations at the intersection of the appropriate classification and tally lines. For example, the first deviation between -3.25 db and -3.75 db would be stroked on the horizontal line for that band, just to the left of the vertical line for tally 1 (See Fig. 4). The second deviation in that band would be stroked just to the left of the tally 2 line. This is continued until all the deviations have been recorded.

The last stroke in each $\frac{1}{2}$ db band indicates the number of deviations found having values within that band. As shown on Fig. 4, for the analysis by the template method this value is written in the first column, marked "Line Tots. (A)." These values are added and should equal the total number of measurements in the study (533 in the example).

Next, the column "Cum. to $\frac{1}{2}$ " is filled out. Beginning at the top line, totals are accumulated to the point where adding the next line total will result in a value exceeding $\frac{1}{2}$ the grand total of measurements (266 in the example). Similarly a value is obtained accumulating the totals from the bottom. In Fig. 4 these values are 246 and 166, respectively.



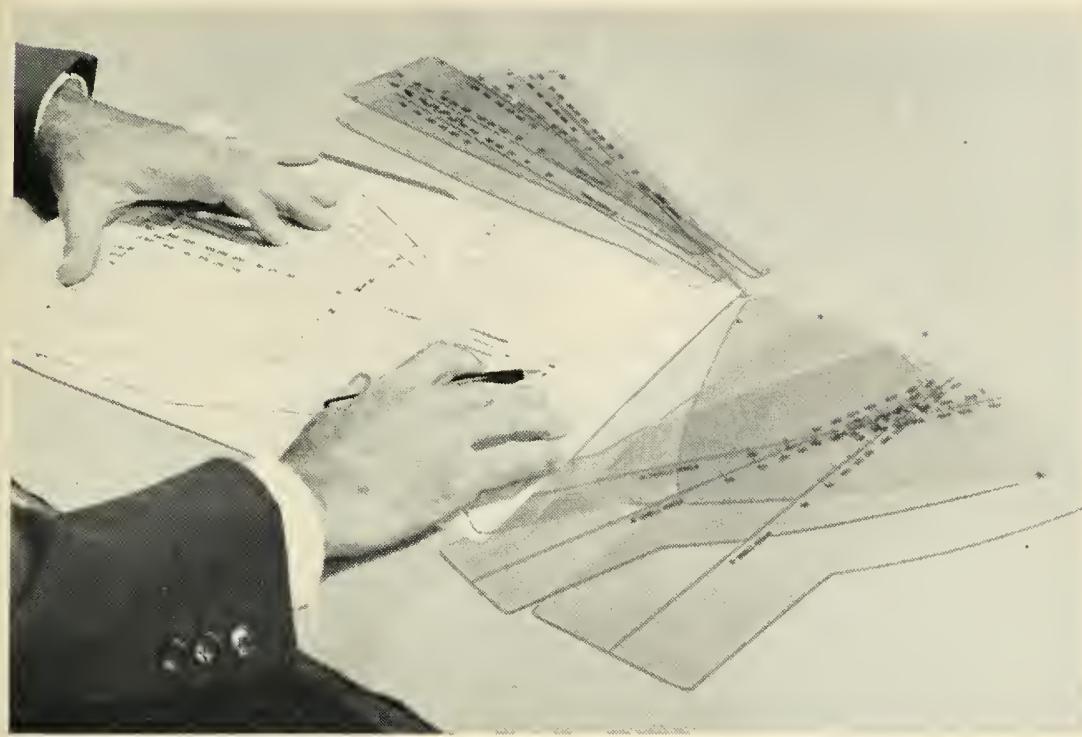
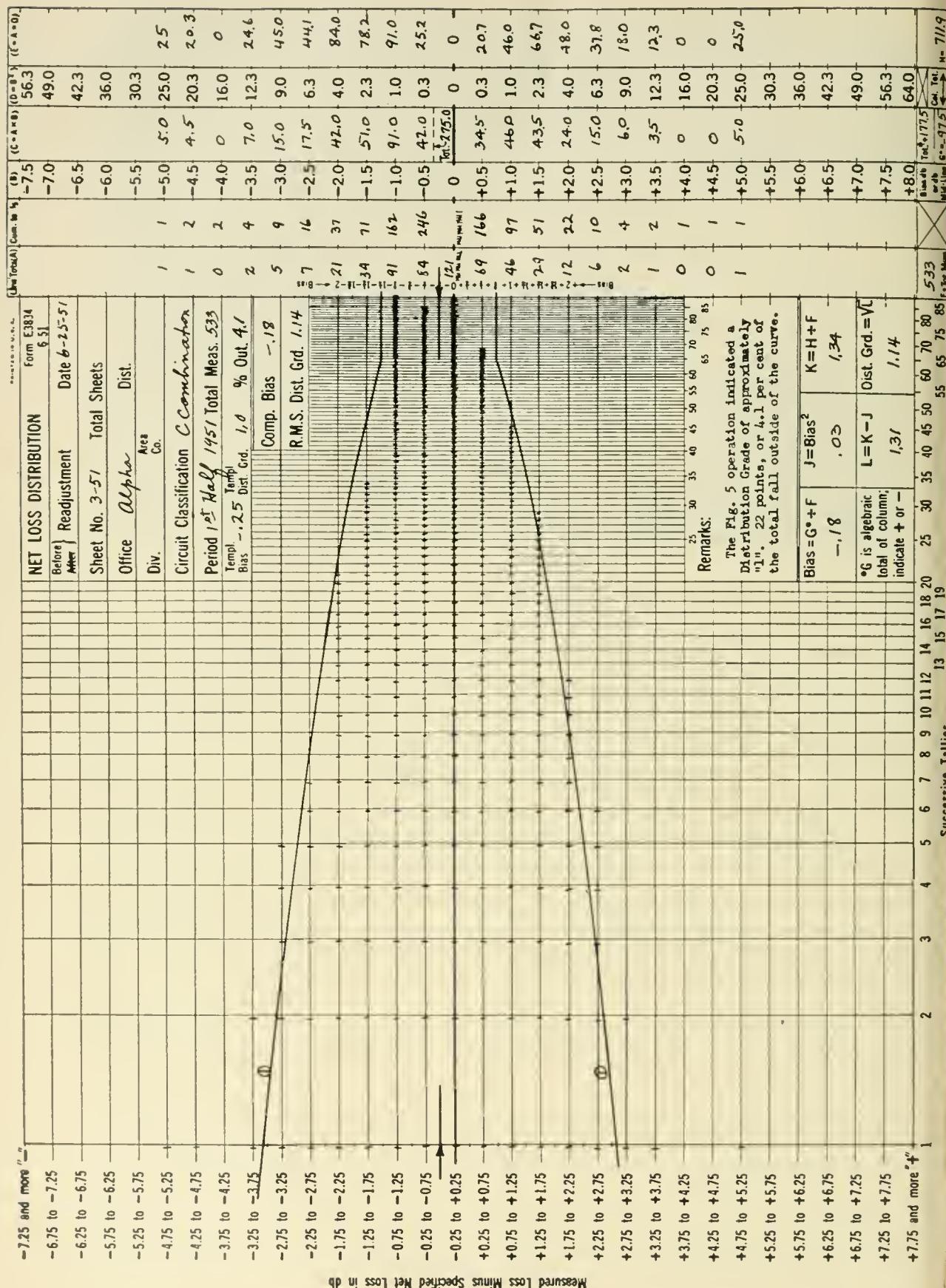


Fig. 5 — Combined template on stroke chart.

By use of this information the approximate bias is determined. The scale of the bias values on the stroke sheet is shown in $\frac{1}{4}$ db steps along the left-hand edge of the "Line Tots. (A)" column, and bias is determined to the nearest $\frac{1}{4}$ db. If the two cumulative totals differ from each other by less than 25 per cent of the larger value, an arrow indicating the bias is placed midway between the two class lines representing these cumulative totals. Its value is read on the bias scale. If the two cumulative totals differ from each other by 25 per cent or more of the larger value, an arrow indicating the bias is placed $\frac{3}{4}$ of the distance between the two class lines representing these cumulative totals and nearer the larger value. In Fig. 4, since 246 minus 166 (80) is greater than 25 per cent of 246 (61.5), the arrow is placed $\frac{3}{4}$ of the way from the line representing 166 toward the line representing 246; i.e., at the $-\frac{1}{4}$ db point on the bias scale. A second arrow is placed at the corresponding point on the tally 1 line.

As shown in Fig. 5, a combined template is then placed over the chart so that the center line of the template coincides with the two arrows. Along the center line of the template there is a scale indicating numbers of measurements from 50 through 700. The template is moved horizontally so that the point on the scale corresponding to the grand total of measurements (533 in the example) is placed on the 1 tally line. Envelope



curves for distribution grades from 0.38 db to 3 db are shown on the template. The smallest envelope having not over 8 per cent of the grand total of measurements outside of the envelope represents the approximate distribution grade. In the example, this is the 1 db curve, for which 22 points or 4.1 per cent of the total fall outside of the curve. Using a cut-out template corresponding to the distribution grade, a trace is placed on the stroke sheet, as shown on Fig. 6.

In cases where the small number of measurements or the character of the dispersion makes it difficult to fit the data with any of the envelope curves of the template, RMS methods of determining the distribution grade and bias afford a better estimate. In the example on Fig. 6, the bias is thus found to be -0.18 db and the distribution grade is found to be 1.14 db.

When the automatic transmission test and control circuit described in the companion paper is used for measuring net losses, the bias and distribution grade can be determined more quickly and easily. This circuit measures the transmission in terms of deviations from the specified loss and records these by a teletypewriter. In addition, registers indicate the total number of measurements and the number of deviations falling in the $\frac{1}{2}$ db bands shown on the stroke chart. The final strokes for each band can thus be placed on the chart directly without the need for stroking each measurement. From this point on, the analysis and the final tracing of the envelope curve which is selected are the same as in the case illustrated by Fig. 6.

EFFECTIVENESS OF OVER-ALL TRUNK TESTS AND ANALYSES

Simple Layouts

With simple trunk layouts particularly those involving one voice-frequency or carrier link, plant forces have been able to use over-all trunk measurements and analyses as a direct aid in maintenance. Early field trials of the stroke chart method were made at two operating telephone company offices. The testers made up stroke sheets from their routine measurements and interpreted the results to find clues as to what to investigate. Stroke sheets made at successive routine testing periods also showed them what improvements they were obtaining in the operation of the trunks.

Both offices started with distribution grades of about 1.8 db and with biases of about $\frac{3}{4}$ db. The trunk plant was then given a thorough cleanup and realignment more rigorous than that called for in the maintenance practices at the time. Similar rigorous circuit order tests were followed

as circuit order changes were made. Many small troubles were found and cleared. As the result of such rigorous circuit order work and the use of statistical analyses, the distribution grades at the end of the trial were reduced to about 1 db and the biases were brought close to zero.

The maintenance activities were conducted by the regular test forces during normally available maintenance time. Although the initial work involved in cleaning up the trunks necessitated some slippage in the periodic maintenance tests, troubles requiring realignment were eventually reduced to the point where it became possible to carry on the periodic testing work concurrently with the more rigorous circuit order work.

Complex Layouts

During the field trial of the automatic transmission test and control circuit discussed in a companion paper, there was an opportunity for studying transmission data taken on intertoll trunks of greater length and complexity of layout. These trunks were composed of two or more carrier links and connected Washington, D.C. to several outlying points; namely, Atlanta, Georgia; Boston, Massachusetts; Hempstead, New York; New York, New York; Oakland, California; and Richmond, Virginia. A total of 231 trunks were in the groups. When the trial began, without preliminary rigorous circuit order work on the trunks involved, the distribution grade was 2.26 db and the bias was +0.35 db. Maintenance investigation was initiated only when trunks were found to have departed more than a prescribed amount from their specified net losses. Initially this value was 4 db and later it was reduced to 3.5 db.

As many of these wide deviations were investigated and corrected as available manpower permitted. The layouts were so complex, however, that it was found impracticable to give prompt attention to all of them; and in many cases it was impossible to check carrier systems that were suspected of being the source of some of the deviations. At the end of the trial the distribution grade had been reduced from the original 2.26 db to a range of about 1.8 to 2 db. The bias had not been changed significantly from the original +0.35 db.

These results indicated very little improvement from the limited readjustments found practicable during the tests. Analysis of the test results has shown that transmission maintenance methods must be improved in some respects. An example of this was a case where the data indicated several trunks to be affected by excessive variation from some common cause. This was traced to a group pilot being out of limits. If routine maintenance methods had indicated this difficulty earlier, the amount of time in which service could have been affected by these trunks would have been reduced. This is important because of the difficulty of finding evidence of common trouble sources, with complex layouts.

The scope of the trial was then limited to a smaller group of intertoll trunks which could be given close attention. The 42 trunk group between Washington, D.C., and Atlanta, Ga., was selected and these trunks were put through rigorous circuit order tests and adjustments approaching the completeness of initial line-up tests. A test cycle composed of transmission loss measurements made on the 42 trunks in both directions was performed four times daily for a period of about five months. During the period covered by this phase of the trial, adjustments were made only as indicated by carrier pilot variations, by deviations from specified net loss large enough to operate the limit feature of the automatic transmission test and control circuit, or with other trouble clearance.

The tests for each day were analyzed as a group. On the first day the distribution grade of the deviations from specified net loss for the group was 0.8 db and the bias was +0.5 db. On the last day the distribution grade was 1.2 db and the bias was -0.25 db. For the entire group of measurements (584 test cycles), the distribution grade of the deviations was 1.26 db and the bias was -0.08 db. This represented a substantial improvement over the results obtained in the first phase of the trial. It showed that a great deal can be accomplished by improving the circuit order procedures and increasing the thoroughness with which they are carried out.

It was found that combination carrier trunks composed of permanently connected links, thus not having the benefit of control by terminal-to-terminal pilots, have more variability than individual trunks having over-all pilots. Adjustment of such combination trunks requires coordinated action at the various pilot terminals through which the trunk passes, in order that readjustment of the over-all trunk loss can be made at the point in the system responsible for the deviation. In the case of many route junctions, the complexity of the layout makes it difficult to coordinate the necessary measurements at several points so that the proper point for adjustment can be determined.

NEED FOR EDUCATION

The complexity of carrier system layout as indicated above, has imposed a difficult task on the plant transmission maintenance forces. Although our present transmission maintenance practices seem to be adequate for systems in simple layouts, some expansion appears needed for the more complex layouts. This will require further study.

It is important to keep in mind, however, that the provision of good practices and training of personnel in following the detailed steps therein are not in themselves sufficient to assure good transmission maintenance. There is an additional need for *education* of plant personnel in fundamental considerations affecting operation of carrier systems. This must in-

clude over-all objectives, inherent capabilities and limitations, and the interrelation of functions of the many basic blocks comprising carrier systems. Personnel so educated can approach the problems of transmission maintenance with understanding and avoid the maladjustments and troubles due to "man-failure" which are potential hazards in any complex systems.

SUMMARY AND CONCLUSIONS

In summary, the problem of maintaining satisfactory transmission over trunks under distance dialing involves, primarily:

1. Improving the over-all trunk net loss stability so that the distribution grade does not exceed 1 db, as an initial objective.
2. Reducing trunk loss bias for individual offices to less than ± 0.25 db.
3. Removing from operation those trunks having excessive loss deviations before unfavorable service reactions occur.

To do these things in the face of the increasing complexity of our plant and the absence of operator surveillance will require that:

1. Individual systems have adequate short term stability to keep day-to-day variations small.
2. Routine tests and adjustments be made on individual systems and components to correct for long-term deterioration.
3. Frequent over-all trunk tests be made to locate trunks whose performance is beyond acceptable limits and, as a quality control measure, to monitor the performance of the trunk plant.
4. Trunk trouble-shooting be performed on a well coordinated basis to locate and correct the source of trouble. (Compensating maladjustments must be avoided.)

Although facilities are available and methods are known for doing some of these things, considerable effort is required as follows:

1. Study of performance of individual systems to determine capabilities of present design and major sources contributing to over-all trunk instability.
2. Study of transmission maintenance procedures, both routine and trouble-shooting, to determine the proper test intervals and how best the procedures can be carried out on a coordinated basis.
3. Development of improvements in systems and test facilities as indicated by the above studies. Convenience is an important factor in test arrangements.
4. Thorough education of personnel in the over-all make-up, function and interrelation of systems within the trunk plant, and in the significance of transmission maintenance in providing uniformly good and dependable transmission.

Bell System Technical Papers Not Published in This Journal

ABBOTT, L. E.,¹ and POMEROY, A. F.¹

How To Get More Range From An Air Gage, Am. Machinist, **100**, pp. 113-115, Feb. 27, 1956.

AHEARN, A. J.,¹ and LAW, J. T.¹

Russell Effect in Silicon and Germanium, J. Chem. Phys., Letter to the Editor, **24**, pp. 633-634, Mar., 1956.

ANDERSON, P. W., see Clogston, A. M.

ARLT, H. G.¹

Standardization of Materials, Standards Engineering, **8**, pp. 6-7, Mar., 1956.

BASHKOW, T. R.¹

DC Graphical Analysis of Junction Transistor Flip-Flops, A.I.E.E. Commun. and Electronics., **23**, pp. 1-6, Mar., 1956.

BECKER, J. A.,¹ and BRANDES, R. G.¹

A Favorable Condition for Seeing Simple Molecules in a Field Emission Microscope, J. Appl. Phys., **27**, pp. 221-223, Mar., 1956.

BENNETT, W. R.¹

Characteristics and Origins of Noise — Part I., Electronics, **29**, pp. 154-160, Mar., 1956.

BENNETT, W. R.¹

Electrical Noise — Part II: Noise Generating Equipment, Electronics, **29**, pp. 134-137, Apr., 1956.

BENNETT, W. R.¹

Synthesis of Active Networks, Proc. Symp. Modern Network Synthesis, MRI Symposia Series, **5**, pp. 45-61, 1956.

¹ Bell Telephone Laboratories, Inc.

BLECHER, F. H.¹**A Junction Transistor Integrator**, Proc. National Electronics Conference, **11**, pp. 415-430, Mar. 1, 1956.BOMMEL, H. E.,¹ MASON, W. P.,¹ and WAINER, A. W.¹**Dislocations, Relaxations and Anelasticity of Crystal Quartz**, Phys. Rev., **102**, pp. 64-71, Apr. 1, 1956.BOZORTH, R. M.¹**Quelques Propriétés Magnétiques, Électriques Et Optiques Des Films Obtenus Par Électrolyse Et Par Évaporation Thermique**, Le J. De Physique Et Le Radium, **17**, pp. 256-262, Mar., 1956.

BOYET, H., see Weisbaum, S.

BRADY, G. W.¹**X-Ray Study of Tellurium Oxide Gas**, J. Chem. Phys., Letter to the Editor, **24**, p. 477, Feb., 1956.

BRANDES, R. G., see Becker, J. A.

BRATTAIN, W. H., see Garrett, C. G. B.

BRAUN, F. A.¹**Mounting Scheme for Large Cathodes**, Rev. Sci. Instr., Lab. and Shop Notes Section, **27**, p. 113, Feb., 1956.CLOGSTON, A. M.,¹ SUHL, H.,¹ WALKER, L. R.,¹ and ANDERSON, P. W.¹**Possible Source of Line Width in Ferromagnetic Resonance**, Phys. Rev., Letter to the Editor, **101**, pp. 903-905, Jan. 15, 1956.DE LEEUW, K.,¹ MOORE, E. F.,¹ SHANNON, C. E.,¹ and SHAPIRO, N.¹
Computability by Probabilistic Machines, Automata Studies, (Princeton Univ. Press), pp. 183-212, Apr., 1956.

EIGLER, J. H., see Sullivan, M. V.

FOX, A. G.¹**Wave Coupling by Warped Normal Modes**, I.R.E. Trans., PGM'TT, **3**, pp. 2-6, Dec., 1955.

¹ Bell Telephone Laboratories, Inc.

FRANCOIS, E. E., see Law, J. T.

GARDNER, M. B.¹

Speech We May See, Volta Review, **58**, pp. 149-155, Apr., 1956.

GARRETT, C. G. B.,¹ and BRATTAIN, W. H.¹

Some Experiments on, and a Theory of, Surface Breakdown, J. Appl. Phys., **27**, pp. 299-306, Mar., 1956.

HAYNES, J. R.,¹ and WESTPHAL, W. C.¹

Radiation Resulting from Recombination of Holes and Electrons in Silicon, Phys. Rev., **101**, pp. 1676-1678, Mar. 15, 1956.

HERRMANN, D. B., see Williams, J. C.

KELLY, M. J.¹

Contributions of Research to Telephony—A Look at Past and Glance into Future, Franklin Inst. J., **261**, pp. 189-200, Feb., 1956.

KLEIMACK, J. J., see Wahl, A. J.

LAW, J. T.,¹ and FRANCOIS, E. E.¹

Adsorption of Gases on Silicon Surface, J. Chem. Phys., **60**, pp. 353-358, Mar., 1956.

LAW, J. T., see Ahearn, A. J.

LLOYD, S. P.,¹ and McMILLAN, B.¹

Linear Least Squares Filtering and Prediction of Sampled Signals, Proc. Symp., PIB, V, pp. 221-247, Apr., 1955.

LOGAN, R. A.¹

Thermally Induced Acceptors in Germanium, Phys. Rev., **101**, pp. 1455-1459, Mar. 1, 1956.

MASON, W. P.¹

Comments on Weertman's Dislocation Relaxation Mechanism, Phys. Rev., Letter to the Editor, **101**, p. 1430, Feb., 15 1956.

MASON, W. P., see Bommel, H. E.

McMILLAN, B., see Lloyd, S. P.

¹ Bell Telephone Laboratories, Inc.

MENDEL, J. T.¹

Microwave Detector, Proc. I.R.E., **44**, pp. 503-508, Apr., 1956.

MERZ, W. J., see Remeika, J. P.

MOORE, E. F.¹

Gedanken-Experiments on Sequential Machines, Automata Studies (Princeton Univ. Press), pp. 129-153, Apr., 1956.

MOORE, E. F., see de Leeuw, K.

POMEROY, A. F., see ABBOTT, L. E.

REMEIKA, J. P.,¹ and MERZ, W. J.¹

Guanidine Vanadium Sulfate Hexahydrate: A New Ferroelectric Material, Phys. Rev., Letter to the Editor, **102**, p. 295, Apr. 1, 1956.

ROBERTSON, S. D.¹

Ultra-Bandwidth Finline Coupler, I.R.E. Trans., PGM TT, **3**, pp. 45-48, Dec., 1955.

ROSE, D. J.¹

On the Magnification and Resolution of the Field Emission Electron Microscope, J. Appl. Phys., **27**, pp. 215-220, Mar., 1956.

SHANNON, C. E., see de Leeuw, K.

SHAPIRO, N., see de Leeuw, K.

SUHL, H.¹

Subsidiary Absorption Peaks in Ferromagnetic Resonance at High Signal Levels, Phys. Rev., Letter to the Editor, **101**, pp. 1437-8, Feb. 15, 1956.

SUHL, H., see Clogston, A. M.

SULLIVAN, M. V.,¹ and EIGLER, J. H.¹

Five Metal Hydrides as Alloying Agents on Silicon, J. Electrochem. Soc., **103**, pp. 218-220, Apr., 1956.

SULLIVAN, M. V.,¹ and EIGLER, J. H.¹

Electrolytic Stream Etching of Germanium, J. Electrochem. Soc., **103**, pp. 132-134, Feb., 1956.

¹ Bell Telephone Laboratories, Inc.

TRENT, R. L.¹

Design Principles of Junction Transistor Audio Amplifiers, I.R.E. Trans., **PQA**, **3**, pp. 143–161, Sept.–Oct., 1955.

TURNER, D. R.¹

The Anode Behavior of Germanium in Aqueous Solutions, J. Electro-chem. Soc., **103**, pp. 252–256, Apr., 1956.

UHLIR, A., JR.¹

High-Frequency Shot Noise in PN Junctions, Proc. I.R.E., Correspondence, **44**, pp. 557–558, Apr., 1956.

VAN HASTE, W.¹

Statistical Techniques for a Transmission System, A.I.E.E. Commun. and Electronics, **23**, pp. 50–54, Mar., 1956.

VAN HASTE, W.¹

Component Reliability in a Transmission System, Elec. Engg., **75**, p. 413, May, 1956.

VAN ROOSBROECK, W.¹

Theory of the Photomagnetoelectric Effect in Semiconductors, Phys. Rev., **101**, pp. 1713–1724, Mar. 15, 1956.

VAN UITERT, L. G.¹

High Resistivity Nickel Ferrites — The Effects of Minor Additions of Manganese or Cobalt, J. Chem. Phys., **24**, p. 306, Feb., 1956.

WAHL, A. J.,¹ and KLEIMACK, J. J.¹

Factors Affecting Reliability of Alloy Junction Transistors, Proc. I.R.E., **44**, pp. 494–502, Apr., 1956.

WAINER, A. W., see Bommel, H. E.

WALKER, L. R., see Clogston, A. M.

WEISBAUM, S.,¹ and BOYET, H.¹

A Double-Slab Ferrite Field Displacement Isolator at 11 KMC, Proc. I.R.E., **44**, pp. 554–555, Apr., 1956.

WESTPHAL, W. C., see Haynes, J. R.

¹ Bell Telephone Laboratories, Inc.

WILLIAMS, J. C.,¹ and HERRMANN, D. B.¹

Surface Resistivity of Nonporous Ceramic and Organic Insulating Materials at High Humidity with Observations of Associated Silver Migration, I.R.E. Trans., PGRQC, 6, pp. 11-20, Feb., 1956.

WOLONTIS, V. M.¹

A Complete Floating-Decimal Interpretive System for the IBM 650 Magnetic Drum Calculator, IBM Technical Newsletter, 11, Mar., 1956.

¹ Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ARNOLD, W. O., and HOEFLER, R. R.

A System Plan for Air Traffic Control, Monograph 2483.

BABCOCK, W. C., RENTROP, E., and THAELER, C. S.

Crosstalk on Open-Wire Lines, Monograph 2520.

BECK, A. C., and MANDEVILLE, G. D.

Microwave Traveling-Wave Tube Millimicrosecond Pulse Generators, Monograph 2551.

BOZORTH, R. M., WILLIAMS, H. J., and WALSH, DOROTHY E.

Magnetic Properties of Some Orthoferrites and Cyanides at Low Temperatures, Monograph 2591.

BRIDGERS, H. E.

A Modern Semiconductor — Single-Crystal Germanium, Monograph 2552.

CETLIN, B. B., see Galt, J. K.

CHYNOWETH, A. G.

Measuring the Pyroelectric Effect with Special Reference to Barium Titanate, Monograph 2545.

CORENZWIT, E., see Matthias, B. T.

CUTLER, C. C.

Spurious Modulation of Electron Beams, Monograph 2543.

DAIL, H. W., JR., see Galt, J. K.

* 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.

DAVIS, H. M., see Wernick, J. H.

DESOER, C. A.

Iterative Solution of Networks of Resistors and Ideal Diodes, Monograph 2583.

DUNCAN, R. S., and STONE, H. A., JR.

A Survey of the Application of Ferrites to Inductor Design, Monograph 2579.

FELDMAN, W. L., see Pearson, G. L.

FEWER, D. R., see Kircher, R. J.

FRY, THORNTON C.

Mathematics as a Profession Today in Industry, Monograph 2585.

GALT, J. K., YAGER, W. A., MERRITT, F. R., CETLIN, B. B., and DAIL, H. W., JR.

Cyclotron Resonance in Metals: Bismuth, Monograph 2535.

GEBALLE, T. H., see Hrostowski, H. J.

GIANOLA, U. F.

Photovoltaic Noise in Silicon Broad Area p-n Junctions, Monograph 2546.

GOSS, A. J., see Hassion, F. X.

GYORGY, E. M., see Heinz, O.

HAGELBARGER, D. W., see Pfann, W. G.; also Shannon, C. E.

HARKER, K. J.

Periodic Focusing of Beams from Partially Shielded Cathodes, Monograph 2553.

HASSION, F. X., THURMOND, D. C., TRUMBORE, F. A., and GOSS, A. J.
Germanium: on the Melting Point; on the Silicon Phase Diagram,
Monograph 2489.

HEIDENREICH, R. D., see Williams, H. J.

HEINZ, O., GYORGY, E. M., and OHL, R. S.

Solid-State Detector for Low-Energy Ions, Monograph 2568.

HERRMANN, D. B., see Williams, J. C.

HOEFLE, R. R., see Arnold, W. O.

HROSTOWSKI, H. J., MORIN, F. J., GEBALLE, T. H., and WHEATLEY, G. H.

Hall Effect and Conductivity of InSb, Monograph 2586.

INGRAM, S. B.

The Graduate Engineer — His Training and Utilization in Industry, Monograph 2554.

KELLY, M. J.

Contributions of Research to Telephony, Monograph 2590.

KETCHLEDGE, R. W.

Distortion in Feedback Amplifiers, Monograph 2488.

KIRCHER, R. J., TRENT, R. L., and FEWER, D. R.

Audio Amplifier Applications of Junction Transistors, Monograph 2484.

KUH, E. S.

Special Synthesis Techniques for Driving Point Impedance Functions, Monograph 2581.

LEE, C. Y.

Similarity Principle with Boundary Conditions for Pseudo-Analytic Functions, Monograph 2587.

MANDEVILLE, G. D., see Beck, A. C.

MATTHIAS, B. T., and CORENZWIT, E.

Superconductivity of Zirconium Alloys, Monograph 2526.

MERRITT, F. R., see Galt, J. K.

MILLER, L. E.

Negative Resistance Regions in Collector Characteristics of Point-Contact Transistor, Monograph 2574.

MOLL, J. L., and ROSS, I. M.

Dependence of Transistor Parameters on Distribution of Base Layer Resistivity, Monograph 2575.

MONTGOMERY, H. C., see Pearson, G. L.

MORIN, F. J., see Hrostowski, H. J.

NESBITT, E. A., see Williams, H. J.

OHL, R. S., see Heinz, O.

OWENS, C. D.

Stability Characteristics of Molybdenum Permalloy Powder Cores, Monograph 2576.

PEARSON, G. L., MONTGOMERY, H. C., and FELDMANN, W. L.

Noise in Silicon p-n Junction Photocells, Monograph 2555.

PEDERSEN, L.

Aluminum Die Castings for Carrier Telephone Systems, Monograph 2593.

PEDERSON, D. O.

Regeneration Analysis of Junction Transistor Multivibrators, Monograph 2452.

PFANN, W. G., and HAGELBARGER, D. W.

Electromagnetic Suspension of a Molten Zone, Monograph 2556.

PRINCE, M. B.

High-Frequency Silicon-Aluminum Alloy Junction Diodes, Monograph 2557.

RENTROP, E., see Babcock, W. C.

ROSS, I. M., see Moll, J. L.

SCHAWLOW, A. L.

Structure of the Intermediate State in Superconductors, Monograph 2569.

SHANNON, C. E., and HAGELBARGER, D. W.

Concavity of Resistance Functions, Monograph 2547.

SIMKINS, Q. W., and VOGELSONG, J. H.

Transistor Amplifiers for Use in a Digital Computer, Monograph 2548.

SNOKE, L. R.

Specific Studies on Soil-Block Procedure for Bioassay of Wood Preservatives, Monograph 2577.

SOUTHWORTH, G. C.

Early History of Radio Astronomy, Monograph 2544.

STONE, H. A., JR., see Duncan, R. S.

TANNER, T. L.

Current and Voltage-Metering Magnetic Amplifiers, Monograph 2582.

THAELER, C. S., see Babcock, W. C.

THURMOND, D. C., see Hassion, F. X.

TRENT, R. L., see Kircher, R. J.

TRUMBORE, F. A., see Hassion, F. X.

ULRICH, W., see Yokelson, B. J.

VOGELSONG, J. H., see Simkins, Q. W.

WALSH, DOROTHY E., see Bozorth, R. M.

WERNICK, J. H., and DAVIS, H. M.

Preparation and Inspection of High-Purity Copper Single Crystals, Monograph 2571.

WHEATLEY, G. H., see Hrostowski, H. J.

WILLIAMS, H. J., see Bozorth, R. M.

WILLIAMS, H. J., HEIDENREICH, R. D., and NESBITT, E. A.

How Cobalt Ferrite Heat Treats in a Magnetic Field, Monograph 2558.

WILLIAMS, J. C., and HERRMANN, D. B.

Surface Resistivity of Non-Porous Ceramic and Organic Insulating Materials, Monograph 2560.

YAGER, W. A., see Galt, J. K.

YOKELSON, B. J., and ULRICH, W.

Engineering Multistage Diode Logic Circuits, Monograph 2592.

Contributors to This Issue

ARTHUR B. CRAWFORD, B.S.E.E. 1928, Ohio State University; Bell Telephone Laboratories 1928-. Mr. Crawford has been engaged in radio research since he joined the Laboratories. He has worked on ultra short wave apparatus, measuring techniques and propagation; microwave apparatus, measuring techniques and radar, and microwave propagation studies and microwave antenna research. He is author or co-author of articles which appeared in The Bell System Technical Journal, Proceedings of the I.R.E., Nature, and the Bulletin of the American Meteorological Society. He is a Fellow of the I.R.E. and a member of Sigma Xi, Tau Beta Pi, Eta Kappa Nu, and Pi Mu Epsilon.

C. CHAPIN CUTLER, B.S. 1937, Worcester Polytechnic Institute. Bell Telephone Laboratories 1937-. Mr. Cutler's early work was in research related to the problems of the short wave multiplex radio transmitter. During World War II he was engaged in research on the proximity fuse and microwave antennas for radar use. Since the war he has been concerned with research on the microwave amplifier and the traveling wave tube. Mr. Cutler is a member of the I.R.E. and Sigma Xi.

HARRY H. FELDER, B.S. in Electrical and Mechanical Engineering, Clemson A. and M., 1918. After some months in the U. S. Signal Corps he joined the Engineering Department of the American Telephone and Telegraph Company in 1919. He joined the Laboratories in 1934. He has been engaged in general transmission problems in connection with telephone repeater development and toll circuit layout and switching. During World War II, Mr. Felder assisted in the development of a method of laying telephone wires from airplanes. Since that time he has continued to work on the transmission aspects of intertoll trunk design, switching, maintenance and loading. He was also associated with adapting of cable carrier circuits for radio broadcast networks. Mr. Felder is a member of Tau Beta Pi.

J. H. FORSTER, B.A. 1944, M.A. 1946, University of British Columbia; Ph.D. 1953, Purdue University; Bell Laboratories 1953-. Since joining the Laboratories, Dr. Forster has been engaged in research on semi-

conductor devices including point-contact transistor development, transistor reliability studies and the development of low-noise alloy transistors. He also served as instructor of semiconductor electronics in the Laboratories Communications Development Training program. At present he is engaged in surface studies and semiconductor device reliability. Member of Sigma Pi Sigma and Sigma Xi.

DAVID C. HOGG, B.S.C., University of Western Ontario, 1949; M.Sc. and Ph.D., McGill University, 1950 and 1953. Dr. Hogg joined Bell Telephone Laboratories in July 1953 and has worked at the Holmdel Laboratory. He has been engaged in studies of artificial dielectrics for microwaves, antenna problems, and over-the-horizon and millimeter wave propagation as a member of the Radio Research Department. During World War II Dr. Hogg was in the Canadian Army and spent five years in Europe. From 1950 to 1951 he was engaged in research for the Defense Research Board of Canada. He is a member of Sigma Xi.

JOHN L. KELLY, JR., B.A. in 1950, M.A. in 1952, and Ph.D. in 1953, all in Physics at the University of Texas. Dr. Kelly joined Bell Telephone Laboratories in 1953 as a member of the Television Research Department at the Murray Hill Laboratory. He has been engaged in experimental work on the nature of television pictures as well as theoretical investigations pertaining to applications of the Information Theory to television. In 1944 he was commissioned a Navy pilot and served three years.

ARCHIE P. KING, B.S. California Institute of Technology, 1927. After three years with the Seismological Laboratory of the Carnegie Institution of Washington, Mr. King joined Bell Telephone Laboratories in 1930. Since then he has been engaged in ultra-high-frequency radio research at the Holmdel Laboratory, particularly with waveguides. For the last ten years Mr. King has concentrated his efforts on waveguide transmission and waveguide transducers and components for low-loss circular electric wave transmission. He holds at least a score of patents in the waveguide field. Mr. King was cited by the Navy for his World War II radar contributions. He is a Senior Member of the I.R.E. and is a Member of the American Physical Society.

J. G. LINVILL, A.B., William Jewell College, 1941; S.B. in 1943, S.M. in 1945 and Sc.D. in 1949, all in electrical engineering at Massachusetts Institute of Technology. Dr. Linvill served at M. I. T. as assistant pro-

fessor in electrical engineering from 1949 to 1951 and was a consultant to Sylvania Electrical Products. He joined Bell Telephone Laboratories in 1951 and worked on active network problems involving applications of transistors as the active element. In March, 1955, he became Associate Professor of Electrical Engineering at Stanford University. He is a member of the American Institute of Electrical Engineers, Institute of Radio Engineers, Sigma Xi, and Eta Kappa Nu.

EDWARD N. LITTLE, A.B., Yale, 1916; S.B., Massachusetts Institute of Technology, 1919; Signal Corps and Air Service Radio Officer training, World War I. Joined Long Lines Department of A. T. & T. in 1919 to work on transmission studies. Transferred to Transmission Section of O. & E. Department in 1922 in work dealing with telephone repeaters. Nine years later joined the group working on transmission maintenance, and since then has worked principally on various phases of voice-frequency toll transmission maintenance. For the last eight years he has been working on the problems of intertoll trunk transmission maintenance posed by the advent of nationwide intertoll dialing with full automatic alternate routing. One angle of this work has been the development and application of statistical analyses as tools for helping to attain the required reduction in net loss variations.

ENRIQUE A. J. MARCATILI, University of Cordoba, Argentina. Mr. Marcatili was awarded the Argentine title of Aeronautical Engineer in 1947 and the title of Electrical Engineer in 1948. He received a Gold Medal from the University of Cordoba for the highest scholastic record. He joined Bell Telephone Laboratories in 1954 after studies of Cherenkov radiation in Cordoba, and has been engaged in waveguide research at Holmdel. Specifically, Mr. Marcatili has been concerned with the theory and design of filters in the millimeter region to separate channels in waveguides. He has published technical articles in Argentina and belongs to the A. F. A. (Physical Association of Argentina).

LEWIS E. MILLER, B.S. in Engineering Physics, Lafayette College, 1949; General Aniline and Film Corp., 1949-1952; Bell Telephone Laboratories, 1952-. Since joining the Laboratories Mr. Miller has specialized in the development of transistors. His early work was on the development for manufacture of the point-contact transistor. From 1954 to May 1956 he was concerned with surface problems and the development of germanium alloy transistors. At present he is concentrating on diffused silicon transistors. Mr. Miller is a member of the American Physical Society.

A. J. PASCARELLA, E.E., Columbia University, 1916. After his graduation he entered the student course of the General Electric Company at Schenectady. Shortly after our entrance into World War I, Mr. Pascarella joined the U. S. Navy and was put in charge of the electrical laboratory of the Gas Engine School at Columbia. In 1921 he joined the Western Electric Company and in 1925 the Technical Staff of the Laboratories. Here with the Systems Department he was concerned with the development of toll testboards, toll signaling, telegraph, carrier and miscellaneous testing equipment. Later his work consisted of formulating maintenance requirements for the over-all testing of toll lines and the detecting and location of faults on toll cables. During World War II he was concerned with developing high level auditory systems for use in psychological warfare. He also acted as editor of repair manuals used by the Armed Services. At the present time he is working on military projects. Licensed Professional Engineer, New York State.

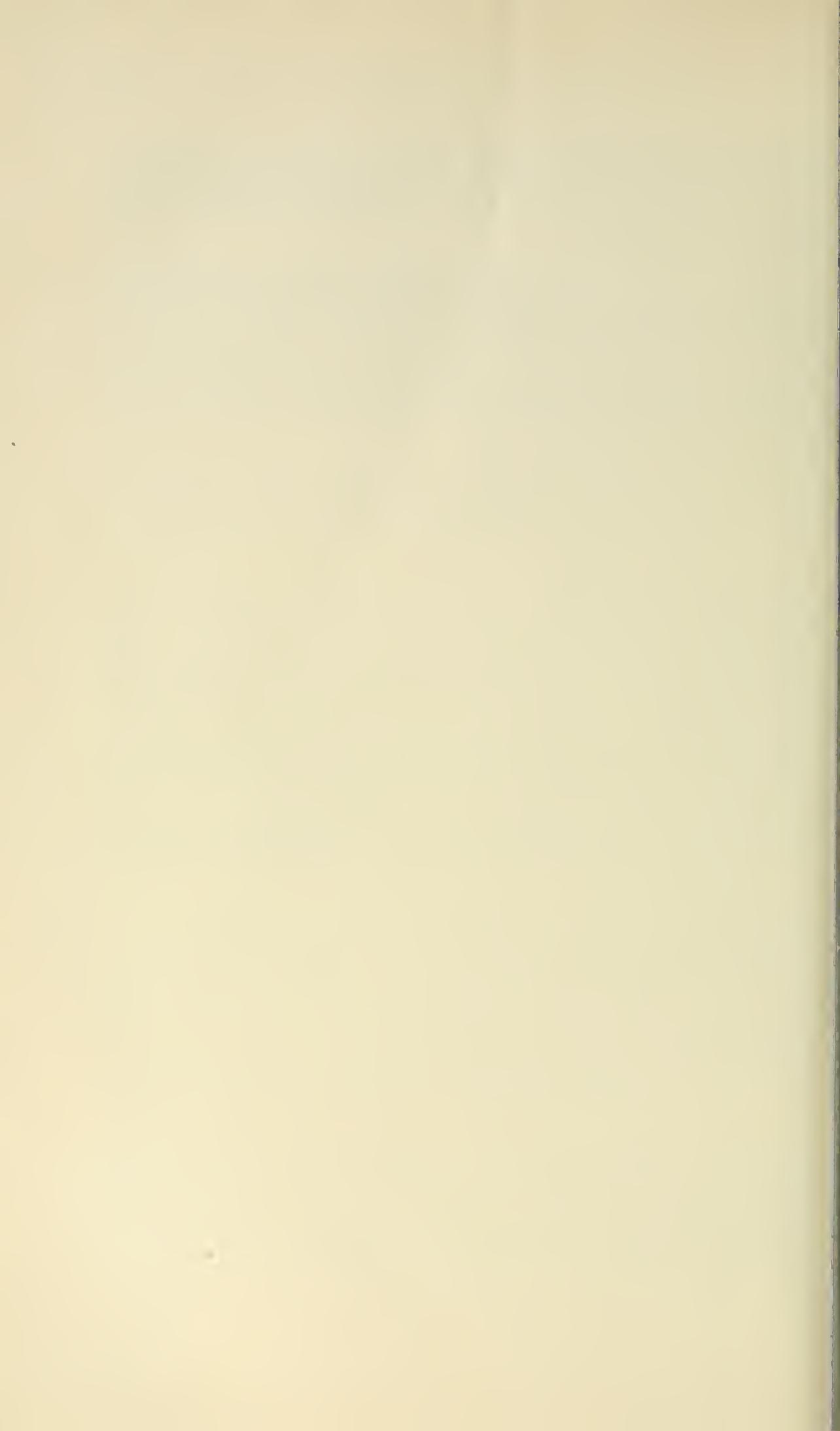
L. G. SCHIMPF, B.E.E., Ohio State University, 1937; Bell Telephone Laboratories, 1937-. From 1937 to 1940 Mr. Schimpf was engaged in research on the application of electronic devices to switching functions, with particular emphasis on cold cathode tubes. With the outbreak of World War II, he turned his attention to research and development work on military projects. For six years after the war he specialized in transmission research studies of local subscriber station circuits and acoustics. Since 1952 he has been engaged in transistor circuit research. In this field he has concentrated particularly on the high frequency operation of transistors in transmission circuits. Senior Member of I.R.E., member of Acoustical Society of America, Eta Kappa Nu, and Tau Beta Pi.

H. F. SHOFFSTALL, B.E.E., Ohio State University, 1916; American Telephone and Telegraph Company, 1916-35; Bell Telephone Laboratories, 1935-. Mr. Shoffstall worked on the development of telephone repeaters and on toll equipment for central offices until he came to the Laboratories in 1935. Since then he has been associated with the switching development group engaged in the design of toll-switching circuits. Member of the American Institute of Electrical Engineers.

HAROLD SEIDEL, B.E.E., College of the City of New York, 1943; M.E.E., D.E.E., Polytechnic Institute of Brooklyn, 1947 and 1954. Dr. Seidel joined Bell Telephone Laboratories in 1953 after employment with the Microwave Research Institute of the Polytechnic Institute of Brooklyn, the Arma Corporation and the Federal Telecommunications Labora-

tories. His work at the Laboratories has been concerned with general electromagnetic problems, especially regarding waveguide applications, and with analysis of microwave ferrite devices. Dr. Seidel is a member of Sigma Xi and the I.R.E.

S. WEISBAUM, B.A., M.S. and Ph.D., New York University, 1947, 1948 and 1953; instructor in physics, New York University, 1950-53; Bell Telephone Laboratories, 1953-. Since joining the Laboratories, Dr. Weisbaum has specialized in the development of microwave ferrite devices, such as isolators and circulators. He is a member of the American Physical Society and Sigma Xi.



H E B E L L S Y S T E M

Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

Volume XXXV

SEPTEMBER 1956

NUMBER 5

Electronics in Telephone Switching Systems	A. E. JOEL	991
Combined Measurements of Field Effect, Surface Photo-Voltage and Photo-Conductivity	W. H. BRATTAIN AND C. G. B. GARRETT	1019
Distribution and Cross-Sections of Fast States on Germanium Surfaces	C. G. B. GARRETT AND W. H. BRATTAIN	1041
Transistorized Binary Pulse Regenerator	L. R. WRATHALL	1059
Transistor Pulse Regenerative Amplifiers	F. H. TENDICK, JR.	1085
Observed 5-6 mm Attenuation for the Circular Electric Wave in Small and Medium-Sized Pipes	A. P. KING	1115
Automatic Testing in Telephone Manufacture	D. T. ROBB	1129
Automatic Manufacturing Testing of Relay Switching Circuits	L. D. HANSEN	1155
Automatic Machine for Testing Capacitors and Resistance-Capaci- tance Networks	C. C. COLE AND H. R. SHILLINGTON	1179
A 60-Foot Diameter Parabolic Antenna for Propagation Studies	A. B. CRAWFORD, H. T. FRIIS AND W. C. JAKES, JR.	1199
The Use of an Interference Microscope for Measurement of Ex- tremely Thin Surface Layers	W. L. BOND AND F. M. SMITS	1209
<hr/>		
Bell System Technical Papers Not Published in This Journal		1223
Recent Bell System Monographs		1230
Contributors to This Issue		1233

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

F. R. KAPPEL, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

B. McMILLAN, <i>Chairman</i>	R. K. HONAMAN
A. J. BUSCH	H. R. HUNTLEY
A. C. DICKIESON	F. R. LACK
R. L. DIETZOLD	J. R. PIERCE
K. E. GOULD	H. V. SCHMIDT
E. I. GREEN	G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

SEPTEMBER 1956

NUMBER 5

Copyright 1956, American Telephone and Telegraph Company

Electronics in Telephone Switching Systems

By A. E. JOEL

(Manuscript received March 18, 1956)

In recent years a number of fundamentals has been discovered through research which place new tools at the disposal of the circuit and system designers. Examples of this "new art" are concepts such as information theory, dealing with the quantization and transmission of information, and solid state principles from which have developed the transistor and other devices. This paper surveys certain new art principles, techniques and devices as they apply to the design of new telephone switching systems.

Over the past forty years a great background and fund of knowledge has developed in the field of telephone switching. Constant improvement in available devices has resulted in increasing the scope of their application. The field has almost reached a point of perfection as an art and is now rapidly entering a more scientific era.

The tools of the present day telephone system design engineer are well known and some are illustrated in Figure 1. These are the relay and the various forms of electromechanical switching apparatus. But over the years, while the art employing these tools was developing, the field of electronics has also been developing. Its applications were most needed when dealing with its characteristics of sensitivity rather

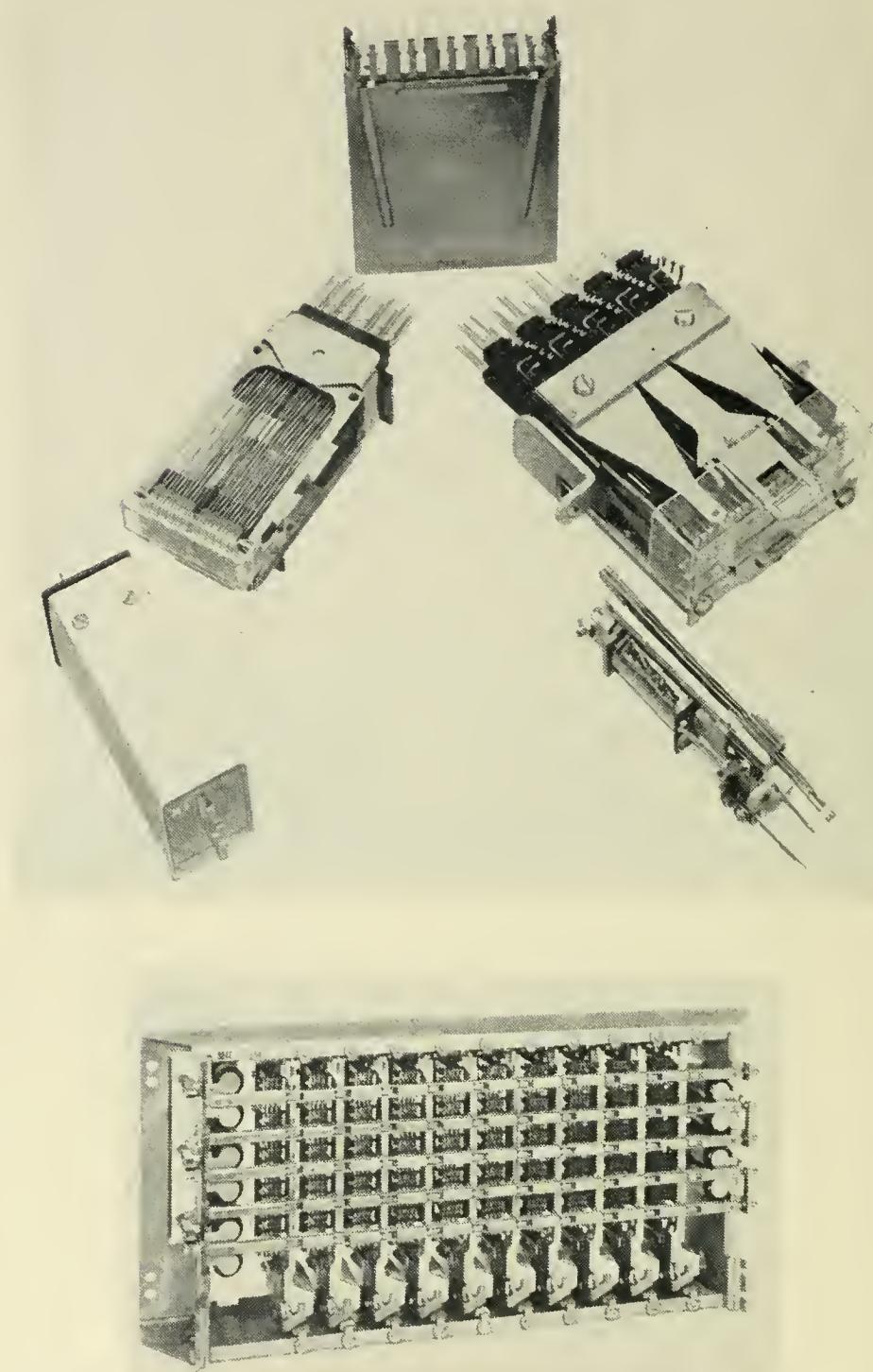


Fig. 1 — Typical telephone relays and switches.

than speed. Even in the telephone switching field, this property of electronics has made its inroads to provide us with better signaling and more accurate timing.

It was not, however, until World War II that the speed advantages of electronics were exploited. This exploitation came primarily in the quantizing of information, both in transmission and information processing equipment. In the latter field new digital computers made their appearance. These machines brought forth the development of new forms of electronic devices, most important of which are those classified as "bulk memory" devices.¹ Later in this paper the characteristics of many of these devices will be discussed in more detail.

In the post-war period the exploitation of another phase of electronics developed from research in semiconductor devices. The transistor is perhaps the best known invention to emerge from these investigations. The impact of the application of semiconductor devices is yet to be felt in the electronics industry and it will most likely find greatest application in the information processing field and in communications generally.

Before one may understand and appreciate the impact electronics will have on the design of new telephone switching systems it is necessary to consider the question: "What is a Telephone Switching System?" By evolution it is now generally recognized that the central office portion of a telephone switching system consists of two principal parts and certain physical and operational characteristics of these parts. These parts, as illustrated in Figure 2, are the interconnecting network, or conversation channel, and its control.

In some switching systems, particularly those of the progressive

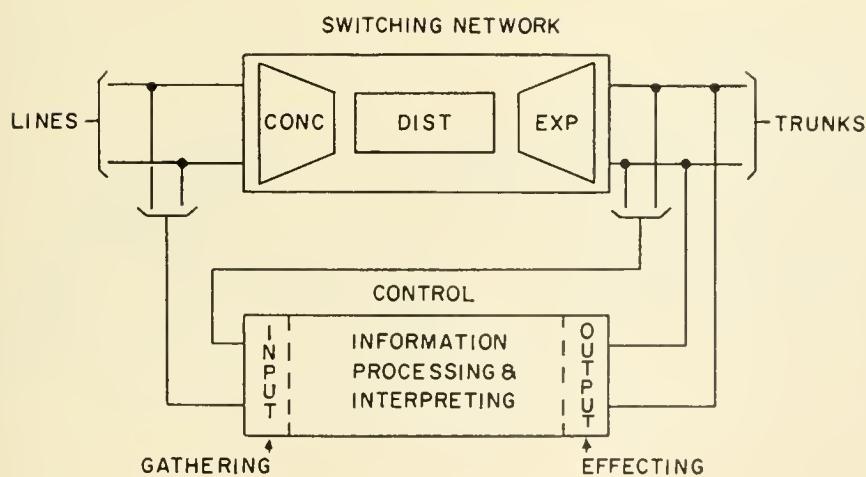


Fig. 2 — Principal parts of common control telephone switching system.

direct control type, such as the step-by-step system, these parts are inexorably integrated. But in the modern systems they have largely been separated. For purposes of the following discussion this type of system, viz., common control, will be assumed. The bulk nature of the electronic memory devices makes them more readily adaptable to systems of the common control type, where the control functions consisting of the receipt, interpretation, and processing of input signals and the effecting of output signals may be concentrated.

INTERCONNECTING NETWORK

In electromechanical switching systems the interconnecting network is composed of crossbar switches or other electromechanical devices. Each connection through the network is physically separated in space from the others and hence the type of network can be called generically a "Space Division" type of network. Such networks are subdivided functionally. First there is the concentration stage where active lines are separated from those not being called or served at a particular time. Next there is distribution stage where interconnection of active lines and trunks is accomplished. Finally there may be an expansion stage where active call paths are connected to selected destinations.

In electronic switching systems three classes of switching networks have been described.² These are:

a. "Space Division" similar to the space division for electromechanical apparatus except that electronic devices such as gas tubes are employed in place of mechanical contacts as the crosspoint element.⁷

b. "Time Division" where calls are sampled in time, each one being given a "time slot" on a single channel.^{3,4}

c. "Frequency Division" such as employed in carrier systems where each call is modulated to a different frequency level on a single transmission medium.^{5,6}

Thus in electronic switching the interconnecting networks derive their basic characteristics from the known methods of telephone transmission. Since transmission techniques are used it is generally not feasible to pass direct current signals through such networks. Also certain ac signals such as 20-cycle current now used for ringing are of such a high power level that they would overload the electronic switching devices employed. For this reason it appears that to accomplish switching with an electronic interconnecting network a change is required in the customer's apparatus to make it capable of responding to a lower level ac for the call signal. Telephone sets with transistor amplifiers and an acoustical horn are being developed. (See Fig. 3.) Interrupted

tones in the voice frequency range can be used effectively to call the user to the telephone.⁸

As in most electromechanical switching networks, the concepts of connecting successive stages of switching devices (stages to perform the functions of concentration, distribution and expansion) to form the network also apply. Since there is more than one method of interconnection, the successive stages of a network may employ different switching techniques — electronic, electromechanical, or both. In electromechanical switching, different devices may also be used in different stages.

In electromechanical space division networks certain types of crosspoints are more adapted to common control operation than others. Systems with electromechanical selector switches most generally are set progressively. In systems with relays or relay-like crosspoints all crosspoints involved in a connection may be actuated simultaneously. In either case the switching device, or the circuit in which it is used, has a form of memory. This memory, shown as a square labeled M in Fig. 4, may be the ability of a selector to remain mechanically held in a particular path connecting position or in a locking or holding circuit associated with a crosspoint relay or crossbar switch magnet.

To minimize the time consumed by the common control elements, simultaneous operation of relay or relay-like crosspoints is most desirable. However, this type of network requires a grid of link testing and control leads such as shown in Fig. 5 for a typical stage of a crossbar switching network. In a network of this type the calling rate capacity is limited by the slow actuating speed of the electromechanical relay or switch. Efficient network configurations can be devised for

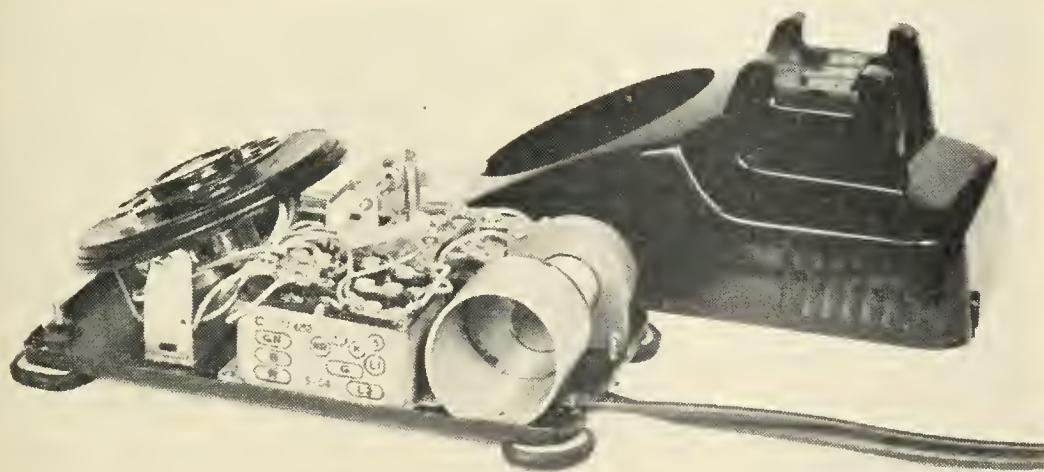


Fig. 3 — Tone ringer telephone set.

large capacity. To set up connections at a high rate in such a network requires a plurality of controls each capable of operating on all or part of the network. In any case, the controls function in parallel on the network because of the speed considerations.

With electronics applied to space division switching networks, two improvements over the operation of relay type space division networks may be achieved. First, the speed of operation of the crosspoint elements may be made high enough so that only one control is needed to operate on networks of the size now requiring a plurality of controls. Second, the properties of proposed electronic crosspoint elements are such that the principle of "end-marking" may be employed.

In contrast to the grid of testing and actuating wires required in electromechanical versions of space division networks, the electronic space division switching network requires only the selectors at each end of a desired network connection to apply the marking potentials. (This is what is meant by "end-marking"; see Fig. 6). The electronic cross-

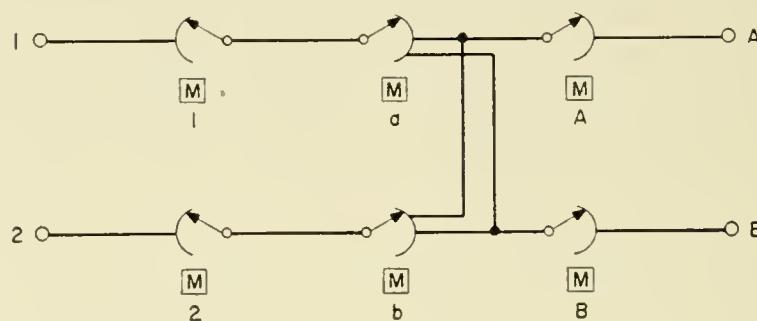


Fig. 4 — Space division switching.

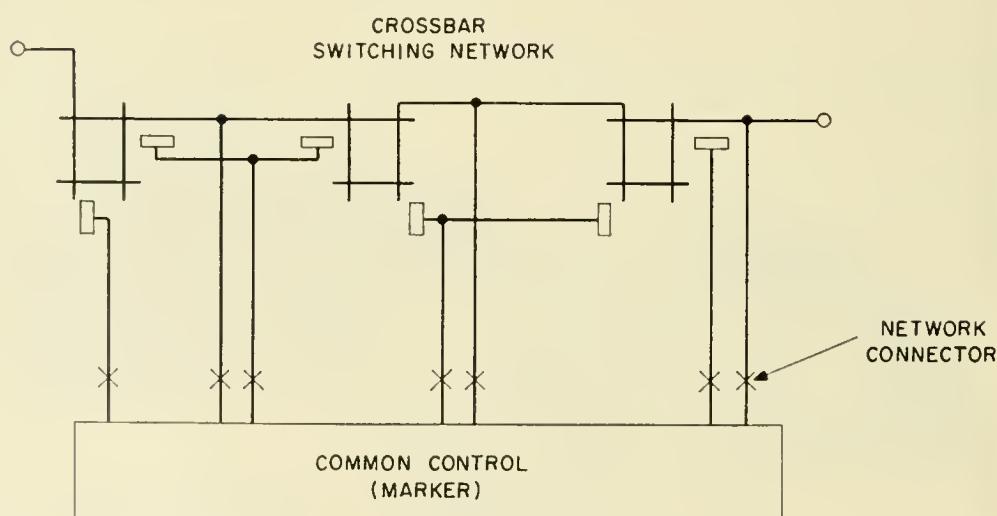


Fig. 5 — Typical common control of a crossbar switching network.

point element will be actuated if the link to which it connects is idle. Eventually all available paths between input and output will be marked. Means must be provided for sustaining only one of the possible idle paths. Here the memory property of the crosspoint device takes over to hold the path until it is released by release marks or removal of the sustaining voltages. So it may be seen that in space division networks the memory requirements must be satisfied the same as in electromechanical networks.

Multiplexing and carrier transmission systems⁹ employ time and frequency division but the physical terminals at both ends of a channel for which the facilities are derived have a one-to-one correspondence which can only be changed manually. In a switching system means must be provided to change automatically the input-output relations as required for each call. Here the need arises for a changeable memory for associating a given time or frequency slot to a particular call at any given time. At some other time these points in time or frequency must be capable of being assigned automatically to different inputs and outputs. For the period that they are assigned, some form of memory must record this assignment and this memory is consulted continuously or periodically for the duration of the call.

With time division switching this new concept in the use of memory in a switching network appears most clearly, see Fig. 7(a). To associate an input with an output during a time slot the memory must be consulted which associates the particular input with the particular output. To effect the connection during a time slot the input and output must be selected. A memory is consulted to operate simultaneously high speed

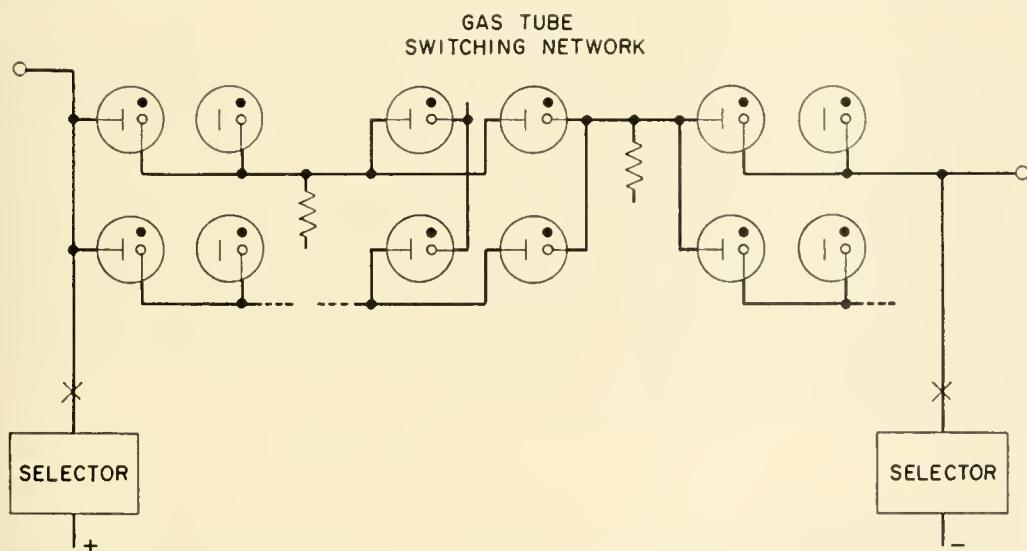


Fig. 6 — Typical "End Marking" control of a gas tube switching network.

selectors for both the input and output. Each selector receives information from a memory which actuates crosspoints to associate the input or output with the common transmission medium. The information from the memory which controls the selection process is known as an "address". The crosspoint is non-locking since it must open when the selector receives its next address. The individual memory of crosspoints for space division networks has thus been changed by time division to changeable memory, usually in the form of a coded address associated with each time slot. Furthermore since the successive addresses actuate the same selectors and hence may be held in a common high speed device, electronic bulk memory is ideally suited for this task. The memory must be changeable to allow for different associations of input to output at different times.

In frequency division the control characteristics of the interconnecting network require a modulation frequency to be assigned each simultaneous conversation to be applied within the bandwidth of the common medium. As shown in Fig. 7(b) the application of the modulation frequencies requires a separate selector for each input and output. These

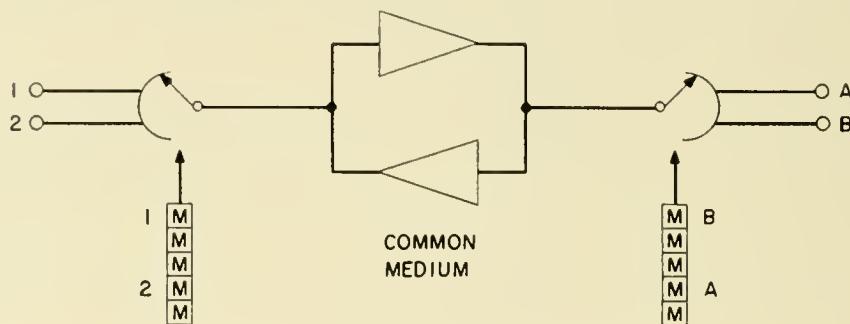


FIG. 7(a) — Time division switching.

selectors are nothing more than space division switching networks and therefore require memory in the switching devices whether they are electromechanical or electronic.

In addition to memory for associations within the switching network, selecting means are also needed to activate a terminal to be chosen in space division (e.g., Fig. 6), to place address information in the proper time slot in time division switching or to set the frequency applying switching network in frequency division.

CONTROL

The control of the switching system provides the facilities for receiving, interpreting and acting on the information placed into it. In par-

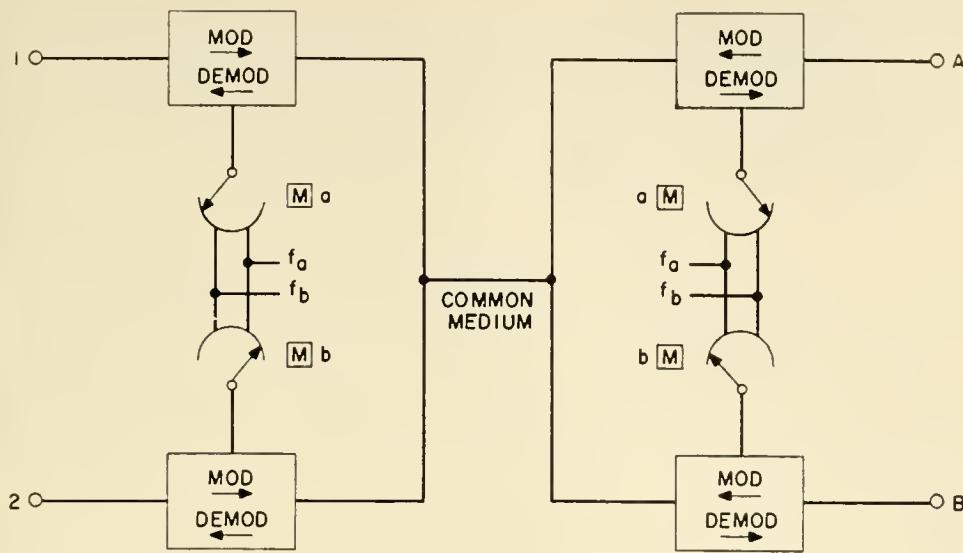


Fig. 7(b) — Frequency division switching.

ticular this is the address of the output desired. A service request detector (SR-D) is provided for each line or trunk.

In electromechanical systems these logic and information gathering functions are performed by relays or electromechanical switches. In order to keep up with the flow of information from a large number of customers, a number of register circuits must be provided to perform the same function simultaneously on different calls. Here information is being gathered on a "space division" basis and therefore a control switching network may be visualized as depicted in Fig. 8. The registers designated R-M constitute the memory used to store the input information as it is being received in a sequential manner from lines and trunks. As in the case of the conversation switching network, a space division control switching network has been used in electromechanical systems because the speed of these devices is not adequate to accommodate the rate at which information flows into the system. It is interesting to note in passing that in the step-by-step system the control and conversation switching networks are coincident. In the No. 5 crossbar system¹⁰ the same network is used for both control and conversation on call originations but when so used the functions are not coincident, that is, the network is used for either control or conversation. In other common control systems, separate control networks known as "register or sender links" are employed.

When using relays to receive the information pulsed into the office by customers or operators a plurality of register circuits are needed. The number of the registers required is determined by the time required to actuate the calling device and for it to pulse in the information. The

registering function has two parts, one to detect or receive the information and the second to store it until a sufficient amount has been received for processing. The processing function is usually allotted to other circuits such as the markers in Crossbar systems.

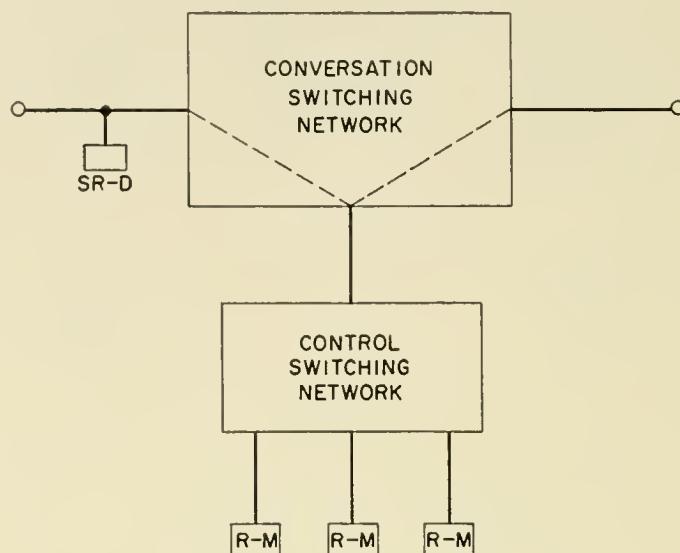


Fig. 8 — Control access.

Since the input of information to a switching system is usually limited to two conductors, a serial form of signaling is used. It would seem only natural that if a detector were fast enough it could function to receive the serial information in several simultaneously active inputs. Relays are not fast enough to do this, but high speed time sharing electronic devices have been designed to perform this information gathering function. Since it is a time sharing arrangement it is analogous to the time division switching. A time division control access as shown in Fig. 8 and 9 requires memory to control the time division switching function. Time sharing when applied to the gathering of information in telephone switching systems has been called "scanning". The individual register memories are still in parallel form because of the relatively long time required for sufficient information to be received before processing may start. Higher speed means for placing information into switching systems such as preset keysets is one way of reducing, if not eliminating, this need for parallel register storage in the switching system prior to processing. However, with this type of device one merely transfers the location of the storage from the central office to the customer's telephone set. The fundamental limitation is the rate at which a human being is able to transfer information from his brain into some physical representation.

Lower cost memory is a practical means for improving this portion

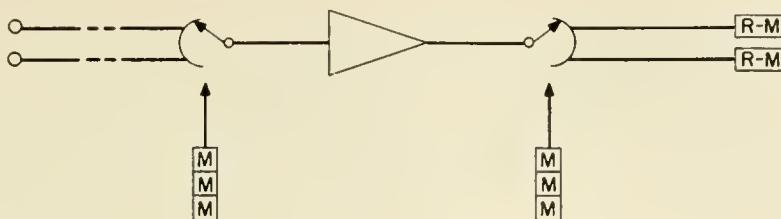


Fig. 9 — Time division control access with separate functional memory.

of the switching system. Many small low cost relay registers have been designed and placed into service.¹⁰ Electronics, however, offers memory at one tenth, or less, of the cost per bit if used in large quantities with a common memory access control. New low cost bulk electronic memories are now available to be used in this manner. As shown in Fig. 10 the memory for the control of the time division control access network and the register memory may be combined in the bulk memory.

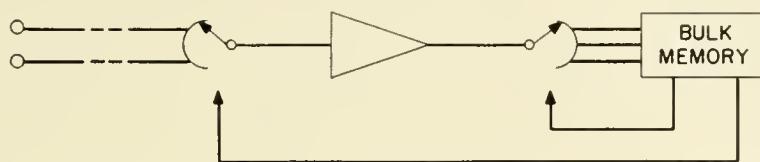


Fig. 10 — Time division control access with bulk memory.

Memory appears in the control portion of a switching system in many ways. Some are obvious and others are more subtle. Fig. 11 shows a typical electromechanical switching system, much like No. 5 crossbar and attempts to indicate various memory functions. First there is active memory designated A such as the call information storage A_2 whether in a register, sender or marker during processing. There is also certain pertinent call information storage associated with trunk circuits such as a "no charge class" on outgoing calls or the ringing code used on incoming calls. Another type of active memory A_1 has been mentioned in connection with switching networks to remember the input-output associations. In most electromechanical systems active memory has been implemented with relays or switches.

Another form of memory is also employed in all telephone switching systems and much effort has been devoted to devising improved means for effecting this memory. This memory is of the type that is not changed with each call but is of a more permanent nature. Examples of this type of memory, which may be called passive memory, designated P, (Fig. 11), are the translations required in common control systems to obtain certain flexibility between the assignment of lines to the switching network and their directory listing. These translations between

equipment numbers (network location) and directory numbers are required to direct incoming calls to the proper terminals (such as the number group frame in No. 5 crossbar, Fig. 12) and to provide on originating calls information for charging purposes (such as the AMA "Dimond" ring translator,¹² Fig. 13). Each of these translators for a 10,000 line office represent about 10^6 bits of information. Another use for passive memory is to translate central office codes into routing information. In local central offices this is also done by cross-connections as shown in Fig. 14.

Another form of passive memory is the punched card or tape. These have been used widely in telephone accounting systems. A step toward electronic memory is the card translator which provides routing information in the crossbar toll switching system¹³ (see Fig. 15). Here the cards represent passive memory and are selected and read by a combination of electromechanical action and light beam sensing with phototransistor detectors. One such device equipped with 1,000 cards represents the storage of approximately 10^5 bits of information.

In all of the above types of passive memory limitations in the speed are involved in the choice of devices used within the memory or the access to it. This is one of the reasons these translators are subdivided so that the various portions may be used in parallel in order to satisfy the total information processing needs of the office.

A discussion of passive memory would not be complete without one further illustration, Fig. 16. This is a wiring side view of a typical relay circuit in the information processing portion of a switching system. It

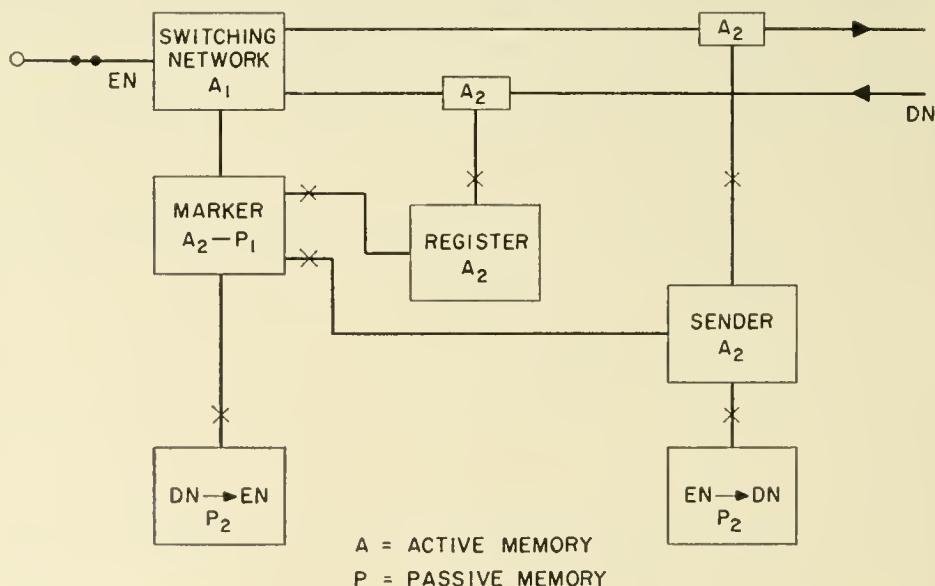


Fig. 11 — Memory in typical electromechanical switching system.

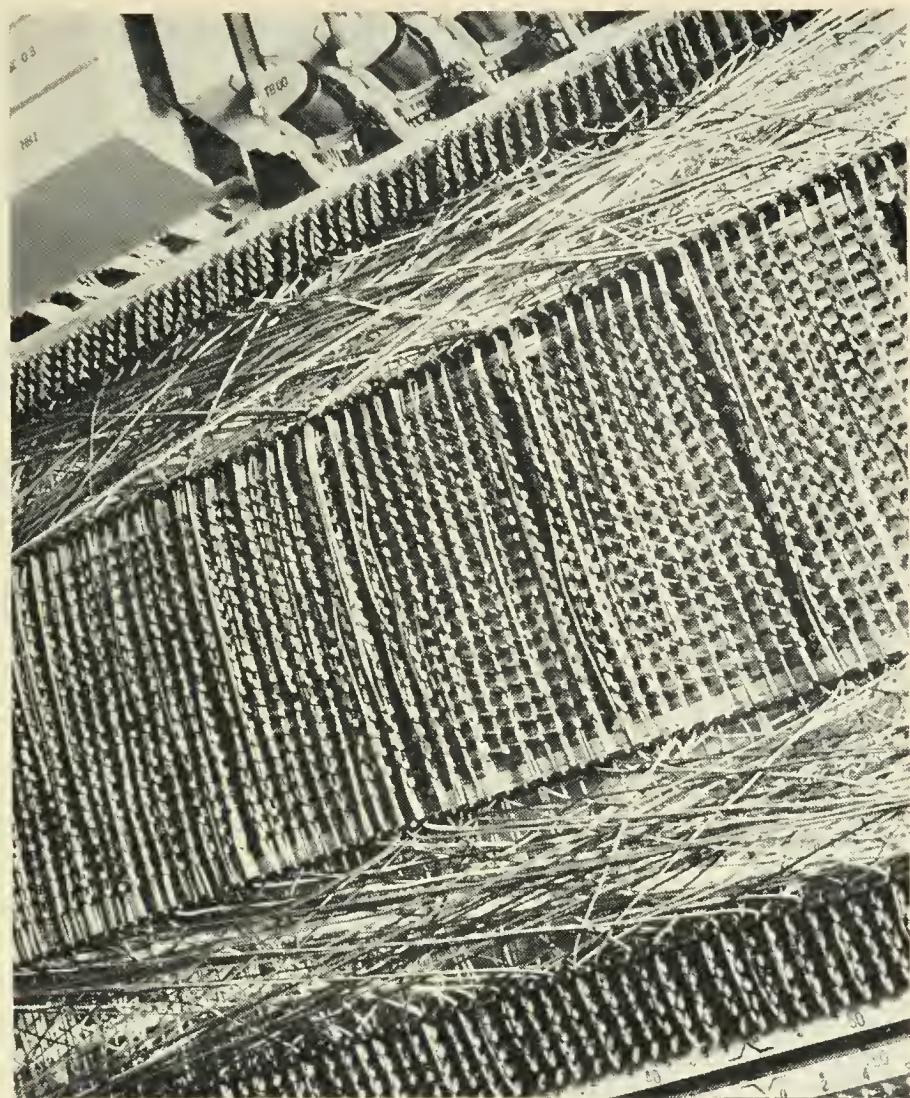


Fig. 12 — No. 5 number group.

could be any other unit, for example, a trunk circuit. The principal point is that each wire on such a unit is remembering some passive relationship between the active portions of the circuit, such as relays. This is the memory of the contact and coil interrelationships as conceived by the designer and based on the requirements of what the circuit is required to accomplish. It is the program of what the central office must do at each step of every type of call. Modern digital computers have been built with the ability to store programs in bulk memories for the solutions of the various types of problems put to them. It is conceivable that the program of a telephone control office may also be stored in bulk memories to eliminate the need for much of the fixed wiring such as appears in relay call processing circuits.

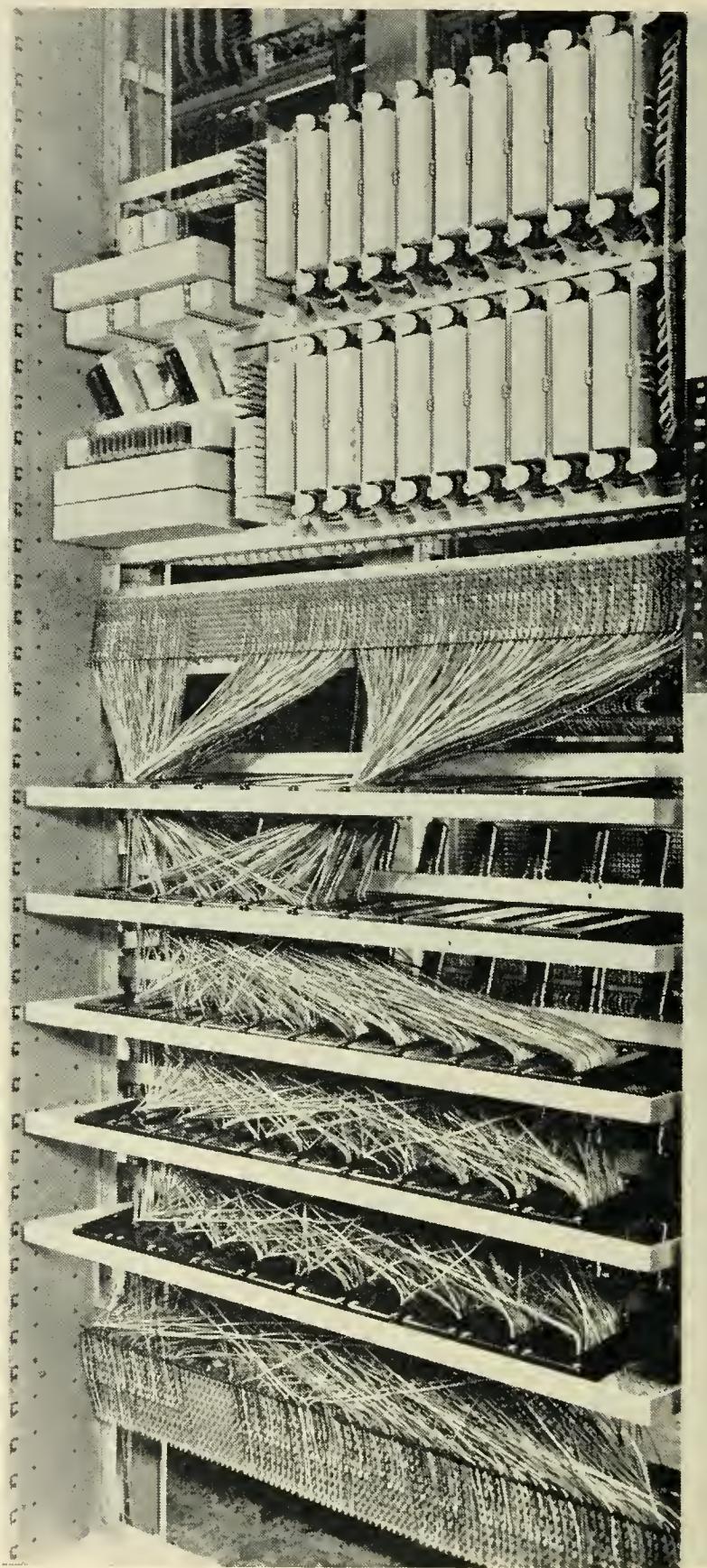


Fig. 13 — AMA translator.

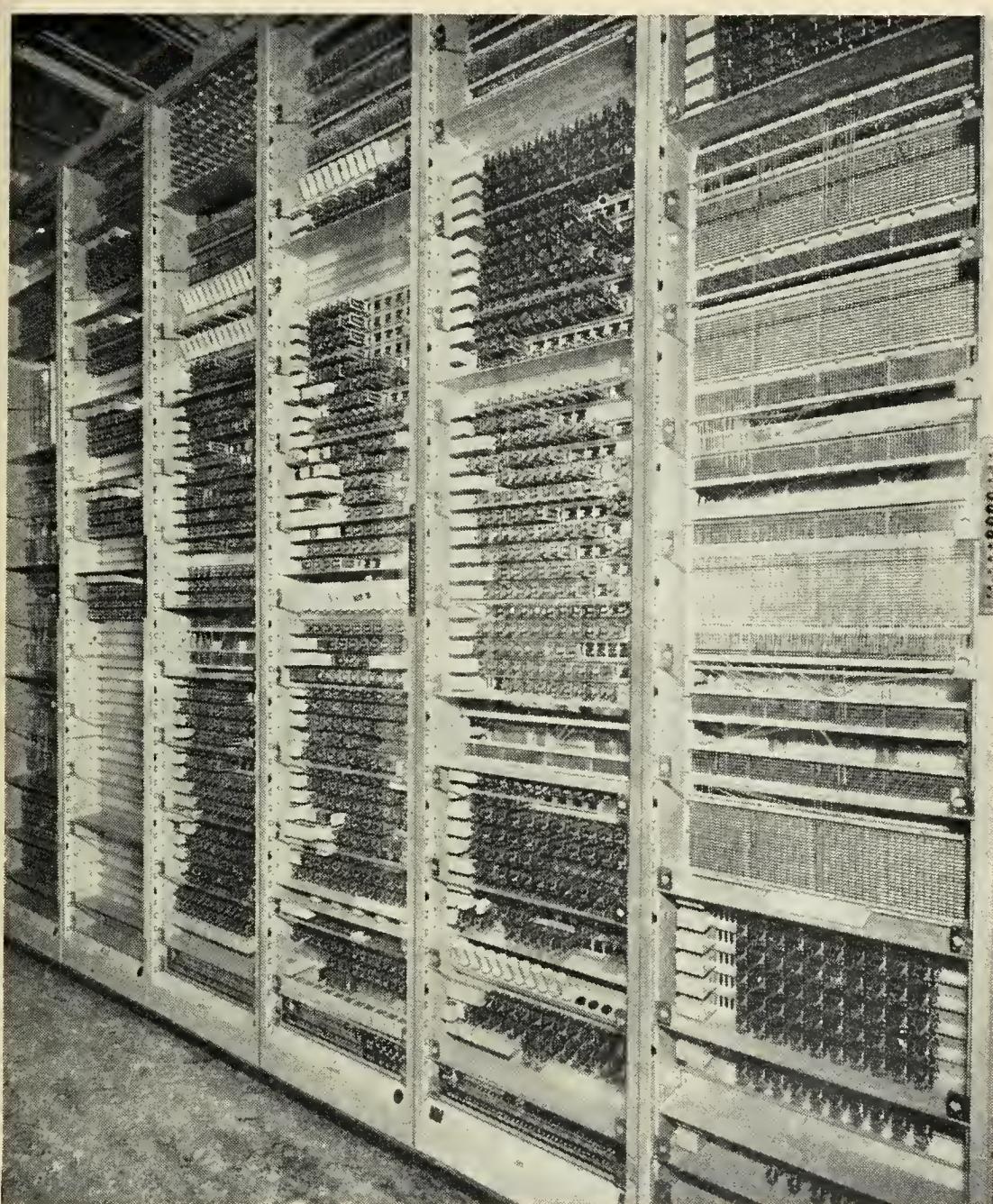


Fig. 14 — No. 5 route relay frame.

The form of memory available in electronics is considerably different from that which has been previously available. Electronic memory has been characterized as "common medium" or "bulk" memory. A single device is used capable of storing more than a single bit of information which is the limit of most relays or other devices capable of operating in a bistable manner. A number of different types of electronic bulk memories have been devised for digital processing. They differ appre-



Fig. 15 — No. 4A card translator.

ciably in physical form, each taking advantage of the phenomenon of some different area of the physical sciences — electrostatic, electromagnetic, optic. Magnetic tapes¹⁴ and drums¹⁵ (Fig. 17), cores¹⁸ (Fig. 18), electrostatic storage in tubes^{16, 17} (Fig. 19) and ferroelectrics^{19, 20} (Fig. 20) and photographic storage²¹ (Fig. 21) are available.

Several properties of these memory devices are of interest. Being electronic, the speed with which stored information may be read is of primary interest. This is known as "access speed". Another property of these common medium memory systems or devices is the ability to change what has been written. If the changes can be made rapidly enough they may be used in electronic systems in much the same manner as relays are used in electromechanical systems to process information. If the change must be made relatively infrequently, such as changing photographic plates, they may be used as substitutes for the type of memory in these systems which are provided by cross connections and wiring. The required fixed or semipermanent electronic memory may be characterized primarily by a high reading speed, large capacity, and the ability to hold stored information even during pro-

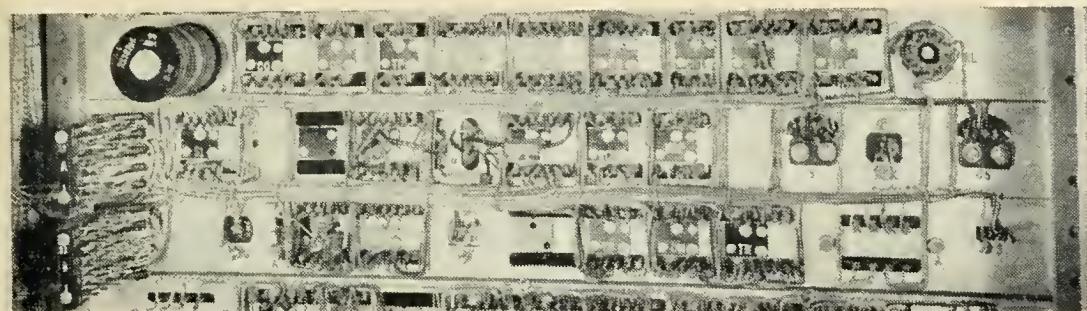


Fig. 16 — Wiring side of relay unit.

longed intervals of loss of power. The amount of memory is measured in terms of binary digits or "bits". The number of bits equivalent to single cross connection can be rather large. Therefore, electronic memory replacing fixed memory such as in the card translator in modern electromechanical systems should be high in bit capacity, from 10^5 to 10^7 bits for 10,000 lines.

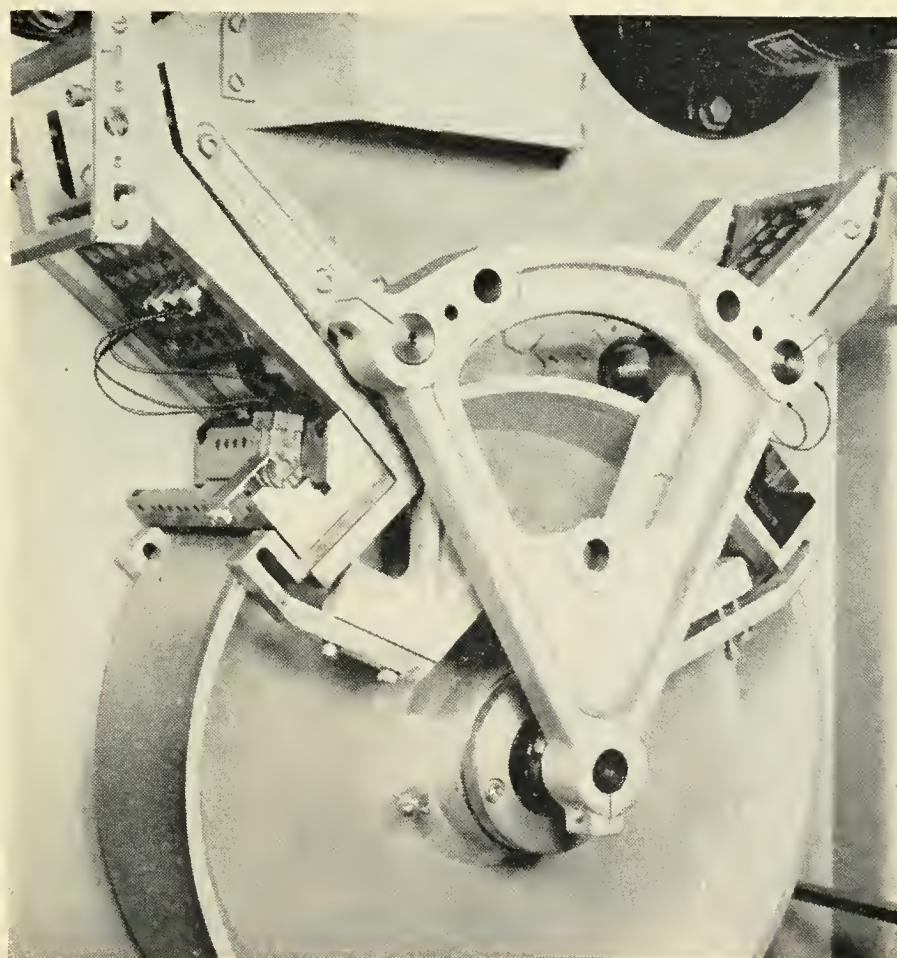


Fig. 17 — Magnetic drum.

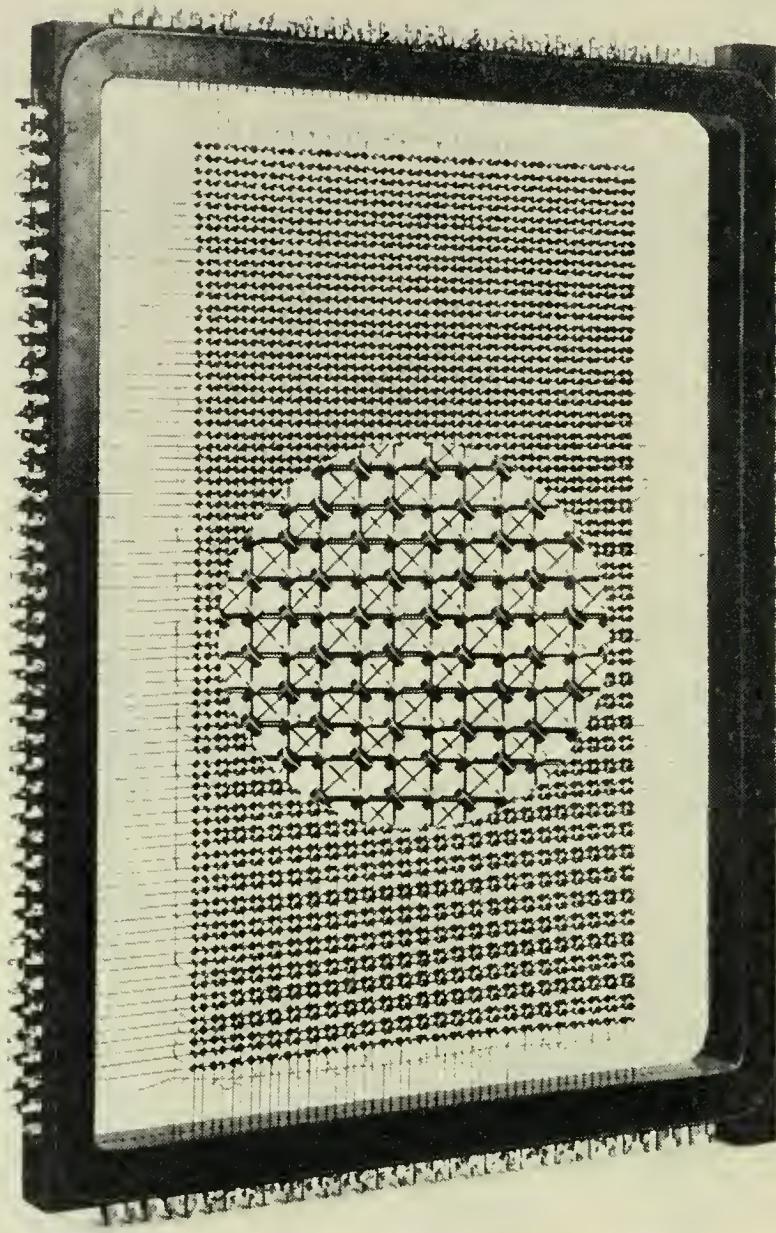


Fig. 18 — Magnetic core array (Courtesy of IBM).

One way in which electronic memory for various system applications may be evaluated is given by the chart of Fig. 22. This chart attempts to show, for the various forms of storage, the relation between the capacity in bits and cycle time, which includes access, reading and, if necessary, the regeneration time of the stored information. For sake of simplicity, ferroelectric and magnetic core memories have been combined as coordinate access arrays. Single bit electronic memory will be described in more detail later.

In the control portion of a switching system it is not only necessary to gather and store information but it must be interpreted and appropriate action taken. This function is called "processing". Processing circuits control the information gathering and storage functions and perform logical functions to produce the necessary flow of information. In the logic circuits of electronic systems, to keep pace with the time sharing nature of the information gathering function, the devices used must be several orders of magnitude faster than their counterparts, the relays, of the electromechanical system. The scanning and bulk memory

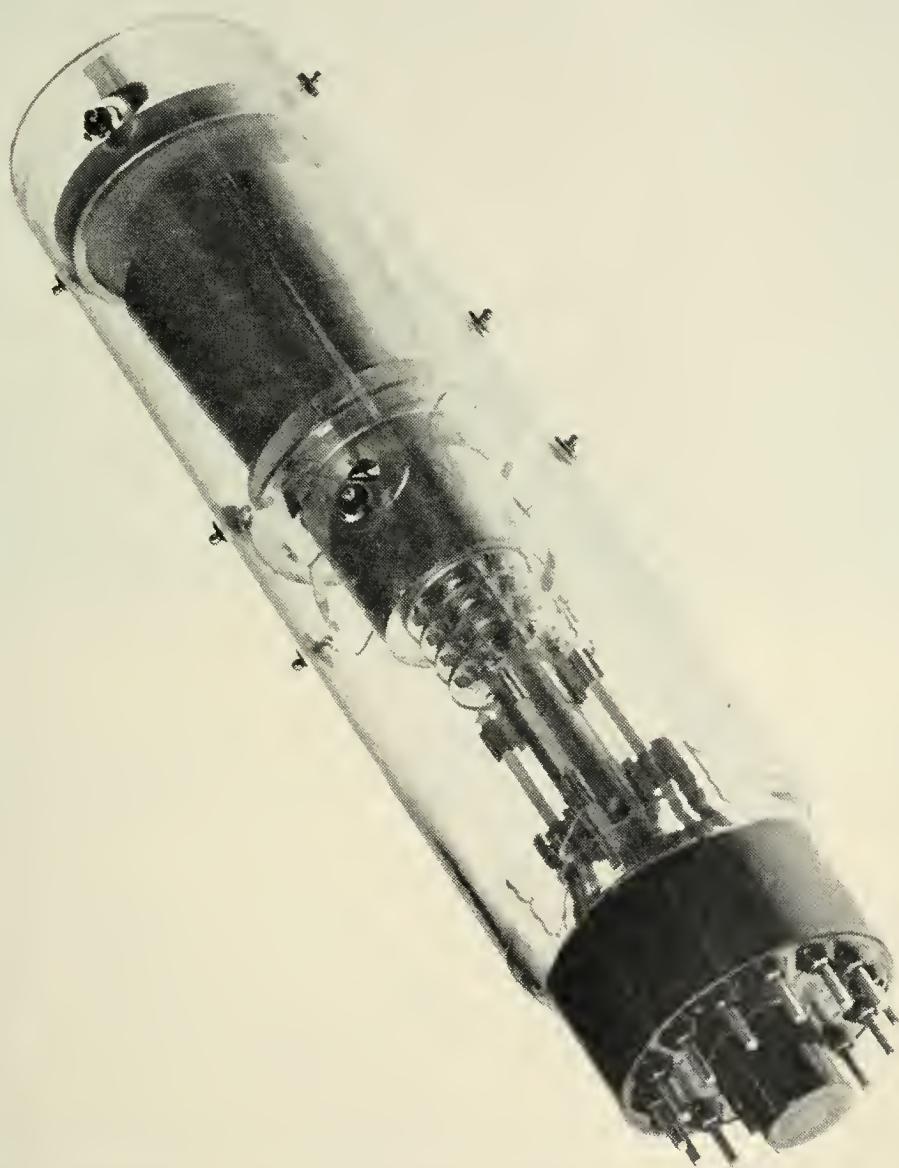


Fig. 19 — Electrostatic storage tube.

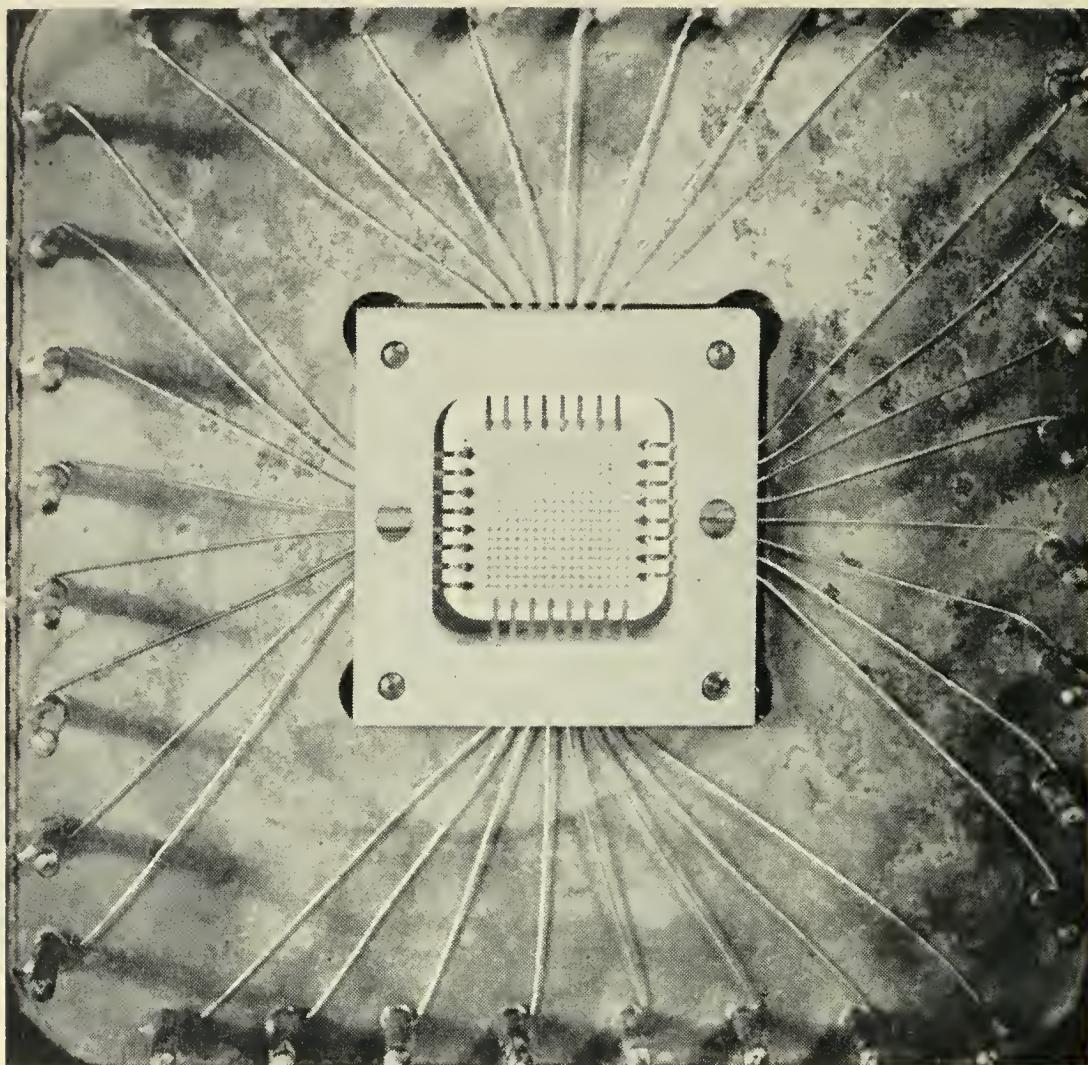


Fig. 20 — Ferroelectric array.

access speeds must be comparable in speed if they are not to become the speed bottleneck. All portions of the system must be in balance time-wise.

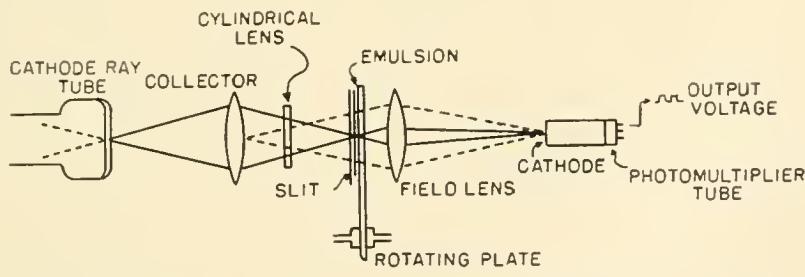
Devices and techniques for use in the design of high speed logic circuits are available.²² With such devices information processing previously carried out by complex relay circuitry may be carried out in microseconds instead of milliseconds. Devices such as semiconductor diodes and transistors seem to be pointing the way to the future in performing these functions.²³ Previously, hot cathode tubes with high power consumption were needed to achieve the same functions at similar high speeds and for a long time this has been one of the greatest deterrents to electronic switching.

Semiconductor diode gate circuits are now quite familiar²³ and take

the place of the conventional make and break contacts in the electro-mechanical switching art (see Fig. 23 for the "AND" function). Magnetic core circuitry is also being exploited to perform high speed switching functions²⁴ (Fig. 24).

There are a number of differences between the circuit configuration used for relay contacts and diode or magnetic core gates for switching logic. When interconnecting such gates to realize complex logic functions other gates are required when circuit elements are placed in series or parallel, whereas in the wiring of relay contacts in series or in parallel no additional circuit elements are required (Fig. 25). Pulse signals passing through diode gate circuits are usually attenuated since the electronic device is not a perfect switcher (infinite impedance open circuit to zero impedance closed circuit). Some minute currents flow when open and some resistance is encountered when closed. Therefore, some amplification is needed at various places in logic circuits and this can be provided by transistor amplifiers. The use of transistors as the gating element eliminates this shortcoming by providing amplification in each gate (see Fig. 24). Transistors have also been successfully used in a new form of logic to provide relay contact like logic thus eliminating the need for gate elements to represent the series of paralleling functions²⁵ (see Fig. 26).

The processing of information usually requires a sequence of logic actions. To provide such sequences, momentary elements similar to locking relays but with microsecond action times are required. When this condition obtains a bistable or "flip-flop" circuit using transistors may be employed. Several forms of transistor circuits have been devised using either the Eccles-Jordan principle,²⁶ negative resistance properties,²⁷ such as achieved with a gas tube, or a regenerative approach.²⁸ Some suggestions have been made on the use of semiconductor diodes in special energy storing circuits to amplify pulses instead of the more conventional transistor amplifiers.²⁹



FLYING SPOT SCANNING A ROTATING DISC

Fig. 21 — Photographic storage (from Proc. I.R.E., Oct. 1953).

EQUIPMENT CONCEPTS

In what has been said, consideration was given only to the concepts and circuitry of electronic telephone switching systems, but the things which the manufacturer and user come in contact with are the physical or equipment realizations of these concepts. One thing that is outstanding about the physical aspects of an electronic system is the large number of small components which are required. Fortunately, most of these components such as resistors, diodes, transistors, condensers, etc., are all of the same physical or similar mechanical design. From the manufacturer's point of view the problem then is to find the most economical way in which these many devices may be manufactured, assembled and tested, because of the large numbers required in a system. The basic solution appears to be: automatic production. This has led to the concept of small packages of components. These packages are the building blocks of a system and contain basic circuits which may be used repetitively. The trend in making such packages appears to be the use of printed wiring with automatic means of placing the components on the printed wiring boards.³¹

Despite the fact that there are large numbers of these small com-

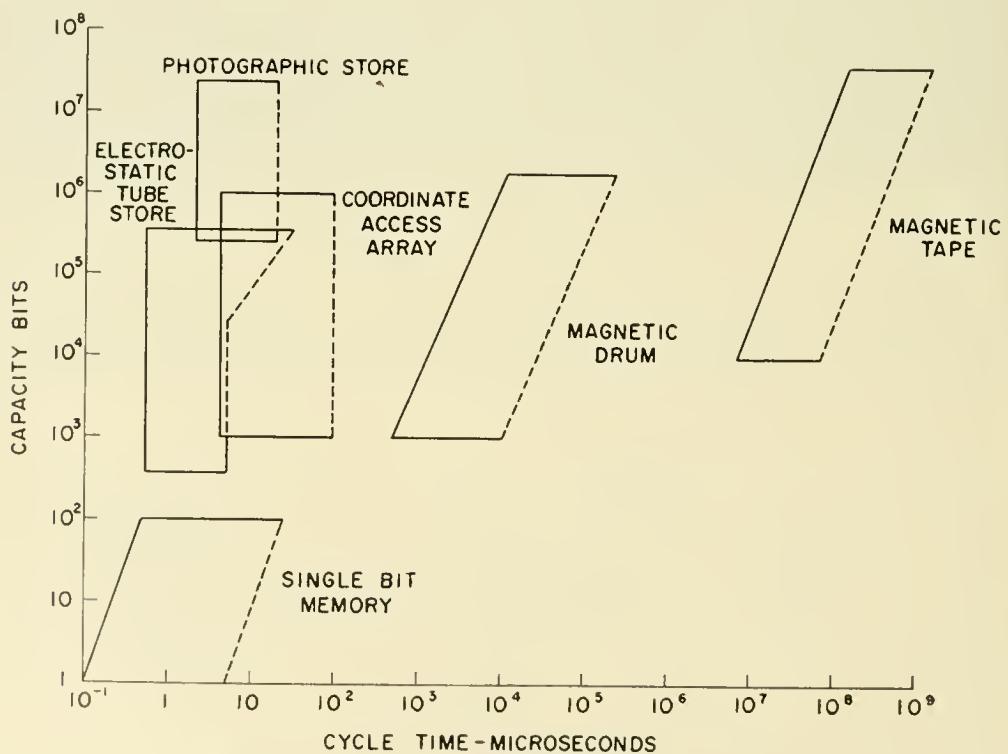


Fig. 22 — Memory system capabilities.

ponents required in electronic telephone switching they are small and when equipment using a multiplicity of printed wiring boards is assembled it takes on the aspect of a three-dimensional arrangement of components, with components mounted in depth as well as on the surface. This is in contrast to electromechanical systems where all components are generally mounted on a vertical surface. By using only one or two common control circuits of a given type (due to high speed) and

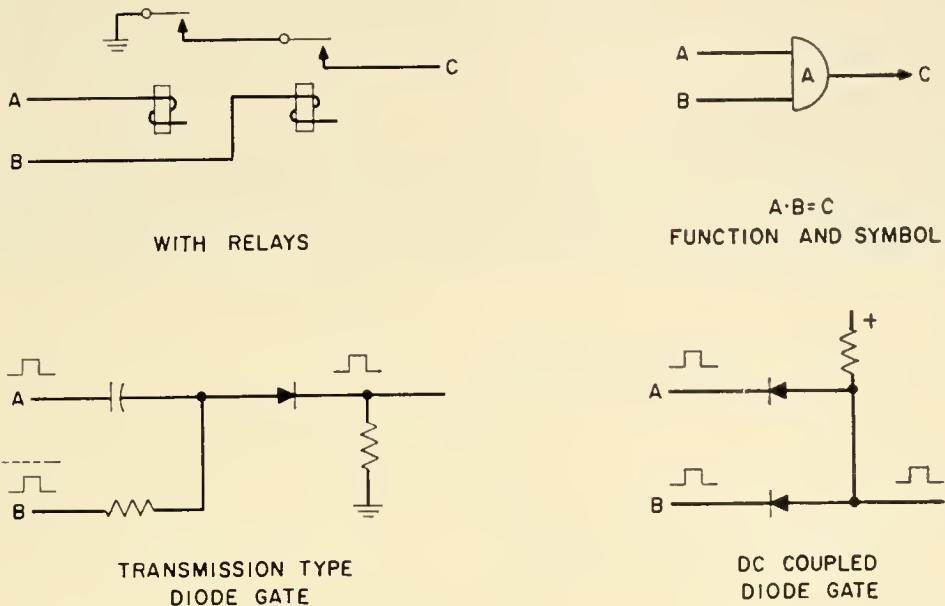


Fig. 23 — The "And" function.

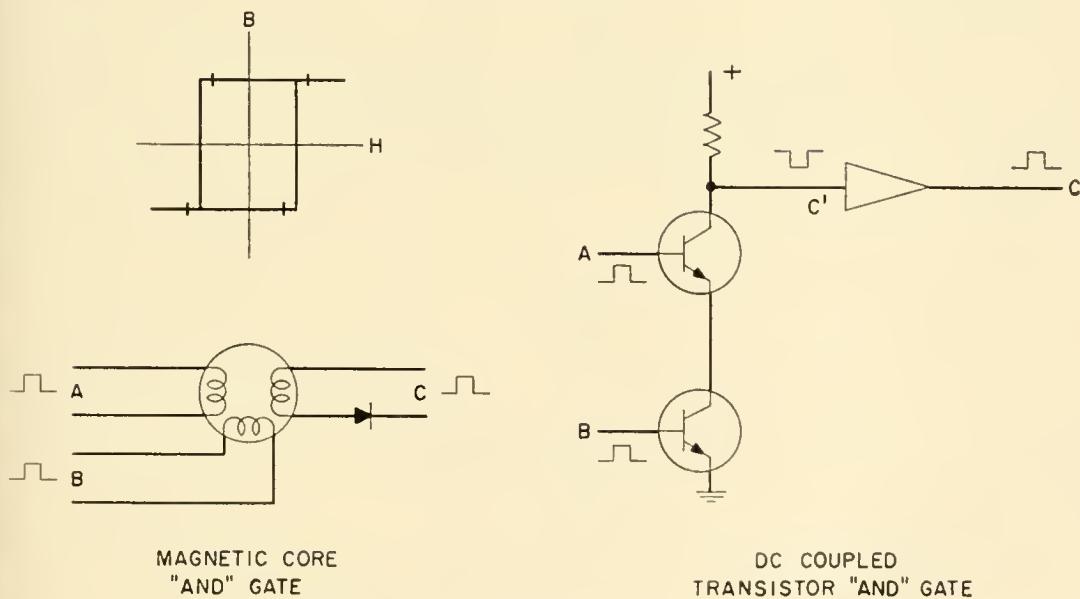


Fig. 24 — Other "And" circuits.

common medium bulk memory, fewer system elements are required which in the overall result in material space saving.

Another phase of the equipment aspects of electronic switching is that the devices require closer environmental control. Air conditioning appears necessary in early systems because of temperature limitations and other characteristics of some of the devices presently available. Also, vacuum tubes and other high power devices may develop objectionable hot spots in the equipment which make it advisable to exhaust hot air.

MAINTENANCE CONCEPTS

There is insufficient experience at this time to say what the maintenance problems of electronic telephone systems will be. Much has been written about the problems encountered in maintaining electronic

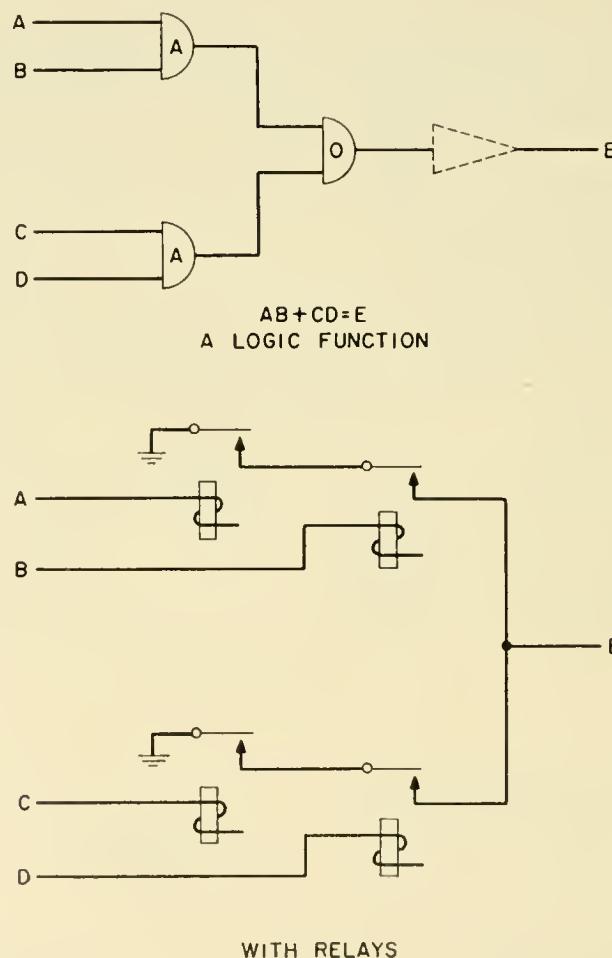


Fig. 25 — A logic function with relays.

computers; however, in designing a telephone system an entirely different philosophy must be pursued since it should not be necessary to have engineering caliber maintenance forces. At no time should the system be incapable of accepting and completing calls. This does not mean that portions of the system may not be worked on for routine or trouble maintenance.

A promising approach appears to be the use of marginal condition routine tests for detecting in advance components which are about to fail.³⁴ Automatic trouble locating arrangements may be devised for giving information as to the specific location of a package in trouble when it occurs.³⁵ This automatic trouble locator combined with the equipment concept of plug-in units means that service may be maintained without long interruptions. By designing devices which are reliable, employing them in a manner to give maximum service life and

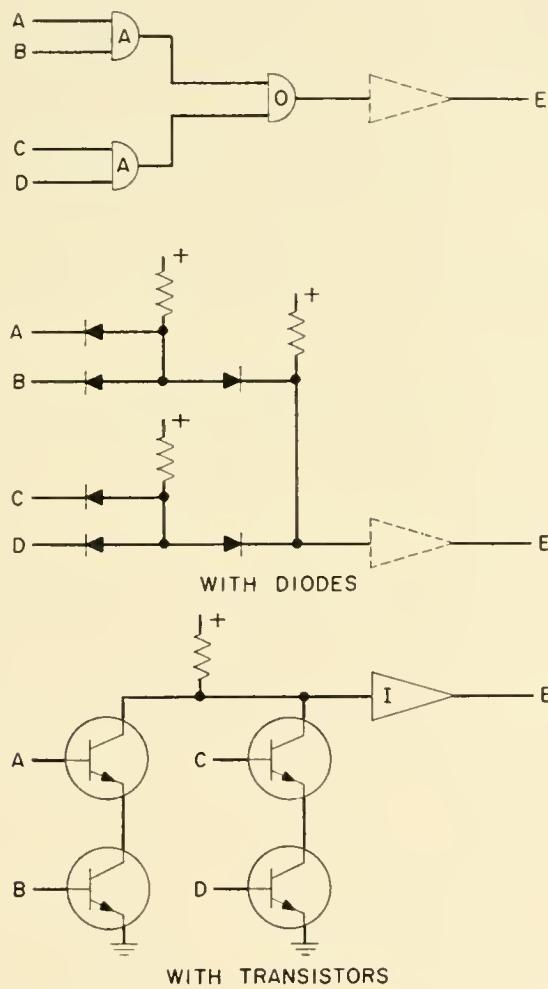


Fig. 26 — A logic function with diodes or transistors.

by judiciously introducing redundancy into the equipment, the chance of simultaneous failures of any two identical parts should be extremely improbable.³² With automatic trouble locating, the maintenance forces will not be required to have a thorough understanding of the device characteristics and the circuitry used. Centralized repair of defective units as in modern telephone transmission systems³³ and perhaps even expendability of defective units are a distinct possibility.

As a result of some of these maintenance considerations it is quite likely that equipment in the future, besides being smaller and more compact, will appear more generally in enclosed low cabinets rather than exposed frames. The administrative control may be from consoles rather than vertical panels. More attention will be paid to appearance. The appurtenances, such as ladders required for high frames in electro-mechanical systems, may be eliminated.

Another change in concept which may come with electronics in telephone switching is the form of the power supply. Present day telephone systems use a centralized single voltage dc distribution system with reserve battery. The wide variety of devices and associated voltages, and the need for close regulation in some portions of electronic systems make a reliable ac distribution system with individual power rectifiers at the point of use appear quite attractive. To insure reliability of service the ac distribution must be continuous and not dependent directly upon the commercial sources.

There is no question that reliability is imperative if electronic switching systems are to survive among electromechanical systems which have achieved a high degree of reliability over a long period of years. The device reliability of the first electronic system may not be comparable since some of the components of the electronic switching systems will not in their initial applications be as reliable as the least reliable component in our present day systems. Reliability will be earned and this will probably require considerable effort. Even if initially some devices employed in electronic systems do not measure up to the present high standard which has been set, continuity of high quality service is a must. It is, therefore, necessary to design a system which will mask the shortcomings of any individual electronic component.³² As their reliability is proven an optimum balance will be sought between system redundancy and component quality. Telephone engineers familiar only with the high degree of reliability of present day apparatus will have to accommodate themselves to the characteristics of new electronic devices.

REFERENCES

1. J. R. Eckert, A. Survey of Digital Computer Memory Systems, I.R.E. Proceedings, **41**, pp. 1393-1406, Oct., 1953.
2. T. H. Flowers, Electronic Telephone Exchanges, Proceedings I.E.E., **99**, Part I, pp. 181-201, 1952.
3. U. S. Patent 2,387,018.
4. U. S. Patent 2,490,833.
5. U. S. Patent 2,408,462.
6. U. S. Patent 2,379,221.
7. W. A. Depp, M. A. Townsend, Cold Cathode Tubes for Audio Frequency Signaling, B.S.T.J., **32**, pp. 1371-1391, Nov., 1953.
8. Tone Ringer May Replace Telephone Bell, Bell Laboratories Record, pp. 116-117, March, 1956.
9. W. R. Bennett, Time Division Multiplex Systems. B.S.T.J., **20**, p. 199, 1941.
10. F. A. Korn, J. G. Ferguson, No. 5 Crossbar Dial Telephone Switching System, *Elec. Eng.*, **69**, pp. 679-684, Aug., 1950.
11. J. W. Dehn, R. E. Hersey, Recent New Features of the No. 5 Crossbar Switching System, A.I.E.E. *Paper No. 55-580*.
12. T. L. Dimond, No. 5 Crossbar AMA Translator, Bell Laboratories Record, p. 62, Feb., 1951.
13. L. N. Hampton, J. B. Newsom, The Card Translator for Nationwide Dialing, B.S.T.J., **32**, pp. 1037-1098, Sept., 1953.
14. Review of Input and Output Equipment Used in Computing Systems. A.I.E.E. *Special Publication S53*.
15. Cohen, A. A., Magnetic Drum for Digital Information Processing Systems, Mathematical Aids to Computation, **4**, pp. 31-39, Jan., 1950.
16. M. E. Hines, M. Chrunev, J. A. McCarthy, Digital Memory in Barrier Grid Storage Tubes, B.S.T.J., **43**, p. 1241, Nov., 1955.
17. M. Knoll, B. Kazan, Storage Tubes and Their Basic Principles, John Wiley & Sons, 1952.
18. M. K. Haynes, Multidimensional Magnetic Memory Selection System. Transactions of the I.R.E., Professional Group on Electronic Computers, pp. 25-29, Dec., 1952.
19. D. A. Buck, Ferroelectrics for Digital Information Storage and Switching, Report R212, M.I.T. Digital Computer Laboratories, June, 1952.
20. J. R. Anderson, Ferroelectric Materials as Storage Elements for Digital Computers and Switching Systems, Communications and Electronics, pp. 395-401, Jan., 1953.
21. G. W. King, G. W. Brown, L. N. Ridenour, Photographic Techniques for Information Storage, Proc. I.R.E., pp. 1421-1428, Oct., 1953.
22. Staff of Harvard Computation Laboratory, Synthesis of Electronic Computing and Control Circuits, Vol. 27 of *Annals of Harvard Computation Laboratory*, 1951.
23. B. J. Yokelson, W. Ulrich, Engineering Multistage Diode Logic Circuits, *Communications and Electronics*, pp. 466-474, Sept., 1955.
24. M. Karnaugh, Pulse Switching Circuits Using Magnetic Cores, Proc. I.R.E., **43**, pp. 576-584, May, 1955.
25. R. H. Beter, W. E. Bradley, R. B. Brown, M. Rubinoff, Surface Barrier Transistor Switching Circuits, I.R.E. Convention Record, Part 4, pp. 139-145, 1955.
26. R. L. Trent, Two Transistor Binary Counter, Electronics, **25**, pp. 100-101, July, 1952.
27. A. E. Anderson, Transistors in Switching Circuits, Proc. I.R.E., **40**, pp. 1541-1558, Nov., 1952.
28. J. H. Felker, Regenerative Amplifier for Digital Computer Applications, Proc. I.R.E., **40**, pp. 1584-1596, Nov., 1952.
29. J. H. Felker, Typical Block Diagrams for a Transistor Digital Computer, Communications and Electronics, pp. 175-182, July, 1952.

30. Promising Electronic Components — Diode Amplifiers, *Radio Electronics*, p. 45, Nov., 1954.
31. A. A. Lawson, Mass Production of Electronic Subassemblies, *Electrical Manufacturing*, **54**, p. 134, Oct., 1954.
32. C. J. Crevelens, Increasing Reliability by the Use of Redundant Circuits, *Proc. I.R.E.*, pp. 509-515, April, 1956.
33. A. L. Bonner, Servicing Center for Short-Haul Carrier System, *Communications and Electronics*, pp. 388-396, Sept., 1954.
34. N. L. Daggett, E. S. Rich, Diagnostic Programs and Marginal Checking in Whirlwind I Computer, *I.R.E. Convention Record*, Part 7, pp. 48-54, 1953.
35. MAID Service for Computer Circuits, *Automatic Control*, p. 23, Aug., 1955.

Combined Measurements of Field Effect, Surface Photo-Voltage and Photoconductivity

By W. H. BRATTAIN and C. G. B. GARRETT

(Manuscript received May 10, 1956)

Combined measurements have been made of surface recombination velocity, surface photo-voltage, and the modulation of surface conductance and surface recombination velocity by an external field, on etched germanium surfaces. Two samples, cut from an n-type and a p-type crystal of known body properties, were used, the samples being exposed to the Brattain-Bardeen cycle of gaseous ambients. The results are interpreted in terms of the properties of the surface space-charge region and of the fast surface states. It is found that the surface barrier height, measured with respect to the Fermi level, varies from -0.13 to +0.13 volts, and that the surface recombination velocity varies over about a factor of ten in this range. From the measurements, values are found for the dependence of charge trapped in fast surface states on barrier height and on the steady-state carrier concentration within the semiconductor.

I. INTRODUCTION

This and the succeeding paper are concerned with studies of the properties of fast surface states on etched germanium surfaces. The experiments involve simultaneous measurement of a number of different physical surface properties. The theory, which will be presented in the second paper, interprets the results in terms of a distribution of fast surface states in the energy gap. The distribution function, and the cross-sections for transitions from the states into the conduction and valence bands, may then be deduced from the experimental results.

Early experiments¹ on contact potential of germanium, and on the change of contact potential with light, indicated that there are two kinds of surface charge associated with a germanium surface, over and above the holes and electrons that are distributed through the surface space-charge region. One kind of surface charge, usually called "charge

in fast traps" can follow a change in the space-charge region very fast in comparison with the light-chopping time used in that work ($\frac{1}{100}$ sec); the other kind, imagined to be more closely connected with adsorbed chemical material, can only change rather slowly. In a previous paper by the authors² it was pointed out that the Brattain-Bardeen experiments, taken by themselves, do not furnish unambiguous information concerning the distribution of these "fast" traps, but that such information might be obtained by performing, simultaneously, other measurements on the germanium surface. More recently Brown and Montgomery^{3, 4} have provided a valuable tool in their studies of large-signal field effect; they point out that if, under given chemical conditions, it is possible to apply a field, normal to the surface, large enough to force the surface potential to the minimum in surface conductivity; then it becomes possible to determine the initial surface potential absolutely (provided certain considerations as to the mobility⁵ of the carriers near the surface are valid).

This paper concerns studies of a number of physical properties that depend on the distribution and other characteristics of the surface traps or "fast" states. Measurements are reported of (i) the change of conductivity of a sample with field; (ii) the photoconductivity; (iii) the change of photoconductivity with field; (iv) the filament lifetime; and (v) the surface photo-voltage. Measurements were made in a series of gaseous ambients, first described by Brattain and Bardeen.¹ Evidence is presented to the effect that the variation in gas ambient changes only the "slow" states, leaving the distribution and other properties of the traps substantially unaffected. From measurements (i) to (iii) it is possible to construct the whole field-effect curve (conductance versus surface charge), even though the fields used were in general not large enough to reach the minimum in conductance.

Using the field effect data, values for the surface potential Y in units of kT/e could be obtained at each point, and also of the quantity $(\partial\Sigma_s/\partial Y)_{\delta=0}$, where Σ_s is the charge in surface traps, and the suffix $\delta = 0$ implies zero illumination. From measurements (ii) and (iv), the surface recombination velocity s could be deduced. (A more detailed study of photoconductivity in relation to surface recombination velocity will be reported at a later date.) Combined with the field effect data, this enables one to deduce the relation between s and Y .

Measurements of the surface photo-voltage may be presented in terms of the quantity $dY/d\delta$, where δ is equal to $\Delta p/n_i$, Δp being the density of added carrier-pairs in the body of the material, and n_i the intrinsic carrier density. The quantity $dY/d\delta$ is closely related to the ratio of the

change in surface potential produced by illumination of the surface to the change in the quasi-Fermi level for minority carriers. By measuring $dY/d\delta$ rather than dY/dL , discussed in Reference 2, the surface recombination velocity is eliminated from the surface photo-voltage data: the limiting values of $dY/d\delta$, after correction for the Dember effect, ought to be (p_0/n_i) and $-(n_i/p_0)$, no matter what the surface recombination velocity may be.

By combining this information with the field-effect data, one can deduce the quantity $(\partial\Sigma_s/\partial\delta)_r$. This and the previous differential, deduced directly from the field-effect data, completely define the dependence of charge in surface traps on the two independent parameters Y and δ — that is, the dependence on chemical environment and on the bulk non-equilibrium carrier level.

The further interpretation of the quantities $(\partial\Sigma_s/\partial Y)_{\delta=0}$, $(\partial\Sigma_s/\partial\delta)_r$ and s in terms of the distribution of surface traps is postponed to the succeeding paper. Here it is sufficient to say that the results are consistent with the assumption that the traps responsible for surface recombination are also those pertinent to the field effect and surface photovoltage experiments. Then the quantity $(\partial\Sigma_s/\partial Y)_{\delta=0}$ depends only on an integral over the distribution in energy of traps; $(\partial\Sigma_s/\partial\delta)_r$ depends also on the ratios of cross-sections for transitions to the valence and conduction bands; and s depends in addition on the geometric mean cross-sections.

II. OUTLINE OF THE EXPERIMENT

The experiment is carried out with a slice of germanium, 0.025 cm thick, which is supported in such a way that there is a gap 0.025 cm wide between the slice and a metal plate. Substantially ohmic contacts are attached to the ends of the slice. Three kinds of experiment are now carried out:

- (i) The conductance of the slice is modulated by illuminating it with a short flash of light; the subsequent decay of photoconductivity with time is studied, and the time-constant of the exponential tail measured.
- (ii) A sinusoidally varying potential difference of about 500 volts peak-to-peak is applied between the metal plate and the germanium. Facilities are available for measuring the changes in conductance produced by the field. The sample is also illuminated with light chopped at a frequency different from that of the applied field. One measures: (a) the magnitude of the peak-to-peak conductance change in the dark; (b) the same in the presence of the light; and (c) the change in con-

ductance, at zero field, produced by the light. The applied field is sufficiently small for the dark field effect and the apparent field effect in the presence of light to be substantially linear.

(iii) The metal plate, disconnected from the high voltage supply, is connected to a high-impedance detector; chopped light is shone on the germanium, and the change in contact potential produced at the surface opposite the metal plate by illumination of the sample measured, and compared with the photoconductivity.

The interpretation of the field effect data has been given by Brown and Montgomery^{3, 4} and by the authors.² The surface conductance ΔG is equal to $e\mu_p(\Gamma_p + b\Gamma_n)$,² where Γ_p and Γ_n are surface excesses of holes and electrons, and are, in equilibrium (i.e., in the absence of light) functions of the surface potential Y and of the body type and resistivity. The minimum in the surface conductance curve occurs at a particular value of Y , so that, if a field effect experiment allows passage through this minimum, values of Y may be obtained.*

In our experiments, measurements were made in a series of different chemical environments, and the minimum in surface conductance did not, in general, occur within the range of field employed. However, it was found to be possible to piece together the complete surface conductance curve (ΔG versus surface charge) by making use of simultaneous measurements of the photoconductance and the change in photoconductance with field. (See Section VI.) From the surface conductance curve, one may deduce the fraction of the surface charge (whether induced electrically, by application of a field, or chemically, by changing the environment) which goes into the fast surface states or traps.^{3, 4} There is, indeed, an assumption here, to the effect that the distribution of traps is unaffected by a change in the chemical environment. The justification for this is the observation of Brown and Montgomery⁴ that it was possible to superpose overlapping large signal field effect curves obtained in different environments. There is also evidence for the validity of this assumption from the self-consistency of the procedure used (see Section V and Fig. 4).

The photoconductivity measurements have been interpreted on the following basis. Illumination of the sample will do two things: it will change the surface excesses Γ_p and Γ_n ,⁶ and it will also change the

* The question of the mobility of carriers near the surface should be mentioned here. For extreme values of Y , the mobility of the carriers that are constrained to move in the narrow surface well is reduced. Values for this reduction in mobility have been calculated by Schrieffer.⁵ However, for values of Y near zero the Schrieffer correction is small, and at somewhat larger (positive or negative) values ΔG is increasing so fast that the error in Y introduced by ignoring the Schrieffer correction is small.

steady-state carrier density deep inside the sample. If the sample is thin in comparison with the body diffusion length and with (D/s) , as was the case in our experiments, the added carrier density Δp will be almost uniform throughout the thickness t of the sample, and one can easily convince oneself that the photoconductance arising from this cause is of the order of (t/\mathfrak{L}) times larger than that arising from the changes in the surface excesses, where \mathfrak{L} is a Debye length for the material. This being the case, the photoconductivity may be considered to be a bulk rather than a surface effect, the surface entering only through the surface recombination velocity s . Under the conditions of the present work the magnitude of the photoconductivity was in fact inversely proportional to s , as was verified in a separate set of experiments. Surface recombination is of interest in that this also calls for "fast" trapping centers on the surface; in fact any trap contributing to the field effect experiment may be a recombination centre, if the cross-sections are right. The questions as to whether the recombination centres and the "fast states" affecting the field effect are the same, or not, is taken up in the succeeding paper.

The surface photo-voltage, like the field effect, is affected both by changes in the surface excesses and by changes in Σ_s , the charge in surface traps. In the experiments, the change in contact potential in a certain light (usually chosen so that the change is small in comparison with kT/e) is compared with the change in conductance produced by the same light. From the latter one may calculate δ (defined as $\Delta p/n_i$) directly. The change in contact potential, measured in units of kT/e , is taken to be equal to ΔY . Thus the surface photo-voltage experiment measures the quantity $(dY/d\delta)$, the differential being taken at constant surface charge. By a slight generalization of the argument previously given by the authors,² one can show that:

$$\frac{dY}{d\delta} = - \frac{(\partial/\partial\delta)_Y(\Gamma_p - \Gamma_n) + (\partial\Sigma_s/\partial\delta)_Y}{(\partial/\partial Y)_\delta(\Gamma_p - \Gamma_n) + (\partial\Sigma_s/\partial Y)_\delta} \quad (1)$$

Now the first terms in the numerator and denominator on the right-hand side are determinate functions of Y , and so are known; the quantity $(\partial\Sigma_s/\partial Y)_\delta$ may be deduced from the field-effect measurements, so that the only remaining quantity, $(\partial\Sigma_s/\partial\delta)_Y$, may be deduced from the measurements of surface photo-voltage.

In concluding this section, a word as to the meaning to be attached to $(\partial\Sigma_s/\partial\delta)_Y$ is in order. The sign of this quantity depends, roughly speaking, on whether the traps in question (i.e., those near the Fermi level under the conditions of the experiment) are in better contact with the

conduction or the valence band. This in turn depends both on the surface potential and on the ratio of cross-sections for transitions to the two bands. For $Y \ll -1$, one expects $(\partial\Sigma_s/\partial\delta)_Y/(\partial\Sigma_s/\partial Y)_\delta$ to have the value $-\lambda^{-1}$; for $Y \gg +1$, the value $+\lambda$. These limiting values may be deduced by a somewhat general argument.⁷

At some intermediate value of surface potential, the above ratio must change sign. If the distribution of surface states in energy is known from the field effect measurements, then the value of Y at which the above ratio changes sign determines the ratio of cross-sections for those traps which are close to the Fermi level for that value of Y . By repeating the experiment for samples of differing bulk resistivity, it is then possible to determine whether the same ratio holds for the states at some different position in the energy gap.

III. EXPERIMENTAL DETAILS

Fig. 1 shows the experimental arrangements. The sample of germanium, of dimensions shown, was prepared by cutting, sandblasting, etching in CP4¹ and washing in distilled water. The exposed faces were approximately (100). The end contacts were made by sandblasting and soldering. The slots A , A' in the ceramic were incorporated in order to

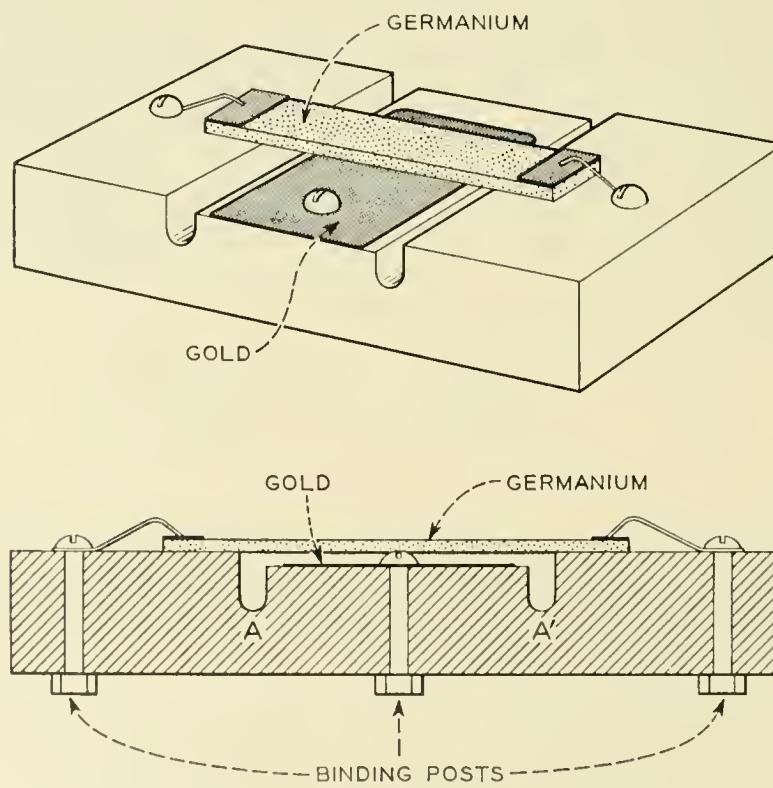


Fig. 1 — Experimental arrangement.

reduce the high field that would otherwise be present near the edges of the ceramic. The gold electrode was deposited by evaporation through a mask. Connections from the gold and from the ends of the sample were made to binding posts passing through the ceramic block.

The ceramic block was set into a metal box, divided into two compartments. In the upper compartment, which contained the sample, there were inlet and outlet tubes, to allow the gas to be changed. The lower compartment contained electrical components, which were thereby protected to a large extent from the changes in gas in the upper compartment. Facilities were available for the type of cycle of gas environment described by Brattain and Bardeen,¹ which cycle was found by them to produce reversible cyclic changes in surface potential. In the top of the box was a window, through which light could be shone onto the germanium either from a chopped or a flash source.

The electrical circuit is shown in Fig. 2. The condenser C_1 is that formed between the germanium and the gold, and has a capacity of about $2 \mu\text{F}$. Impedances Z_1 and Z_2 form a Wagner ground, which has to be balanced first. Then, by adjusting resistance R_1 and condenser C_2 , one may obtain a balance in the case that there is no current flowing through the sample. A current (determined by the battery B and the

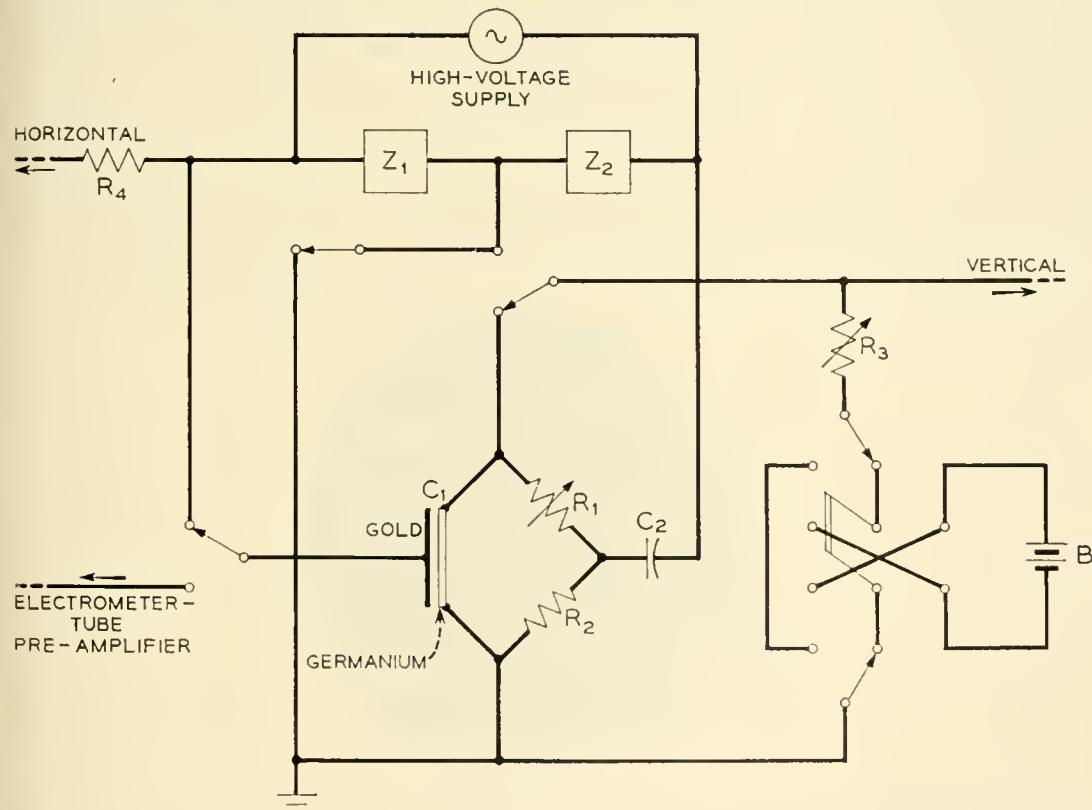


Fig. 2 — Electrical circuit.

resistance R_3) is now switched in, and the resulting off-balance (representing the field modulation of conductivity) presented on the vertical plates of an oscilloscope. The supply voltage is connected, via the high bleeder resistance R_4 , to the horizontal plates. The frequency of the oscillator was chosen to be 25 cyc/sec, a value sufficiently low to obviate lifetime difficulties; the peak-to-peak swing was generally 500 volts.

During a field-effect measurement, the sample was also illuminated with light chopped at 90 cyc/sec. This had the result of causing to be presented on the oscilloscope screen a pattern such as that shown in Fig. 3. The lower tilted line represents the (dark) field effect curve; the vertical separation represents the photoconductivity, as modified by the applied field. Measurements were made of the mean vertical separation, and of the slopes of the upper and lower lines (by reading gain settings).

During a surface photo-voltage measurement, the gold electrode was disconnected from the high-voltage supply, and connected to a high-impedance detector, similar to that used in the work of Brattain and Bardeen.¹ A value for the chopped light intensity was chosen to give a contact potential change that was generally not more than 5 mV. A simultaneous measurement of the photoconductivity was also made.

The gas cycle was similar to that described by Brattain and Bardeen.¹ Some variations were made in it to try to spread out the rate of change with time so that the data could be obtained without large gaps. The cycle used was: (i) sparked oxygen 1 min, (ii) dry O_2 , (iii) mixture of dry and wet O_2 , (iv) wet O_2 , (v) wet N_2 , (vi) a mixture of dry and wet N_2 , (vii) dry O_2 , (viii) dry O_2 , triple flow, and (ix) ozone normal flow. The normal rate of gas flow was about 2 liters per minute; the wet gas was obtained by bubbling through water (probably about 90 per

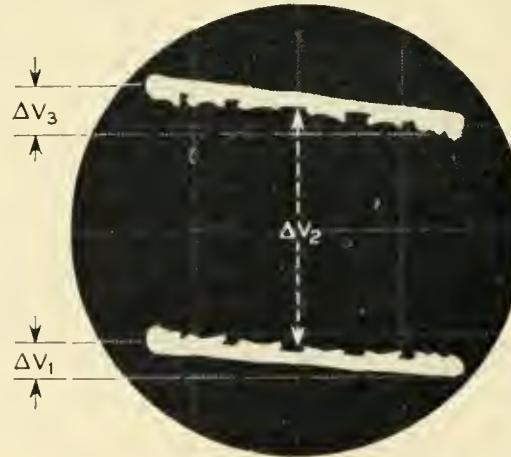


Fig. 3 — Picture of field effect-photoconductivity pattern, as observed on oscilloscope. Dark curve at the bottom.

cent r.h.) and the mixture of dry and wet was obtained by letting approximately one-half the gas flow bubble through H₂O. In carrying out the experiment, it was found convenient to carry out alternately a complete cycle of field effect and surface photo-voltage measurements. The values of the photoconductivity at equivalent points in successive cycles could be compared, in order to check that no systematic error was introduced by this procedure.

In addition to the foregoing, the following measurements were made:

1. All dimensions were determined.
2. The resistivity of the sample was found, and also the body lifetime, on another specimen cut from the same crystal.
3. The amplitude of the voltage swing was measured.
4. The amplifiers in the field effect circuit were calibrated.
5. The capacity of the germanium-gold condenser was determined (by a substitutional method). The value obtained was larger than that calculated from the parallel-plate formula, because of the edge effects.
6. A standard square-wave voltage was introduced into the surface photo-voltage circuit, in order to calibrate the high-impedance detector.
7. At several points in the cycle, the fundamental mode lifetime of the sample was determined by the photoconductivity decay method. This calibrated the 90 cyc/sec photoconductivity measurements, without the necessity for a knowledge of the light intensity.

IV. RESULTS

Measurements were made on two samples: one n-type, 22.6 ohm cm ($\lambda = 0.345$), the other p-type, 8.1 ohm cm ($\lambda = 17.7$). The body lifetime for both samples was greater than 10⁻³ sec, so that for slices of the thickness used (0.025 cm. or less), and for values of *s* in the range encountered, body recombination may be ignored.

Results of typical field-effect runs for the two samples are indicated in Tables I and II. The first column in each table gives the time in minutes from the beginning of the cycle at which the measurements were made. The second column shows the "effective mobility," $d\Delta G/d\Sigma$, obtained from the observed (dark) field effect signal voltage ΔV_1 (see Fig. 3) by use of the formula: $\mu_{\text{eff}} = w^2 t^2 \Delta V_1 / I \rho_0^2 C V_{\text{app}}$, where *w* is the width of the slice, *t* the thickness, *I* the dc flowing through it, ρ_0 the resistivity, *C* the capacity of the germanium-gold condenser, and V_{app} the voltage applied across it. The third column shows the mean value of $\delta (= \Delta p/n_i)$, obtained from the mean photoconductivity signal voltage

TABLE I — 22.6 OHM CM *n*-TYPE CYCLE 12.
RELATIVE LIGHT INTENSITY 0.082

Time min.	μ_{eff} $\frac{cm^2}{volt \ sec}$	δ	μ_{eff}^* $\frac{cm^2}{volt \ sec}$	$s \frac{cm}{sec}$
0	Sparked O ₂			
1	Changed to dry O ₂			
1.5	334	7.85×10^{-2}	520	90
2.5	344	6.9×10^{-2}	585	103
5.5	344	6.2×10^{-2}	595	114
6.5	344	6.07×10^{-2}	595	117
7.0	Changed to mixture of dry & wet O ₂			
7.5	136	4.4×10^{-2}	520	161
8.5	84	4.02×10^{-2}	440	177
9.0	52	3.60×10^{-2}	270	196
10.0	Changed to full wet O ₂			
11.0	-660	2.26×10^{-2}	-440	314
11.5	-890	1.74×10^{-2}	-780	408
12.0	-960	1.67×10^{-2}	-910	425
13.0	Changed to full wet N ₂			
18.0	Changed to mixture of dry & wet N ₂			
19.5	-1150	1.67×10^{-2}	-1060	425
20.5	-1050	1.74×10^{-2}	-960	408
22.5	-990	1.83×10^{-2}	-890	390
23.0	Changed to dry O ₂			
23.5	-430	2.98×10^{-2}	-220	238
23.8	-290	3.2×10^{-2}	0	222
24.0	-84	3.81×10^{-2}	240	186
24.5	31	4.3×10^{-2}	310	165
26.5	146	4.7×10^{-2}	410	151
27.0	Tripled flow of dry O ₂			
27.5	220	5.1×10^{-2}	450	139
29.5	260	5.7×10^{-2}	510	124
31.5	280	6.2×10^{-2}	510	114
35.0	Changed to ozone			
35.5	310	6.8×10^{-2}	510	104
37.5	320	8.2×10^{-2}	490	87

ΔV_2 by use of the formula: $\delta = w t \rho_i \Delta V_2 / I \ell_{ill} \rho_0^2$, where ρ_i is the intrinsic resistivity and ℓ_{ill} the length of the illuminated part of the slice. The fourth column shows the apparent effective mobility in the presence of light,* obtained from the field effect signal voltage ΔV_3 in the presence of light, using the same formula as that giving μ_{eff} . The last column shows the surface recombination velocity, which is proportional to δ^{-1} for fixed light intensity, the constant of proportionality being determined by comparison with measurements of the fundamental mode lifetime.

The results of typical surface photo-voltage runs are shown in Tables

* One must be careful to avoid thinking of μ_{eff}^* as a true field effect mobility, since it is really a sum of two quite different components: the true field effect mobility μ_{eff} , and a term, proportional to thickness of the slice, arising from the photoconductivity.

TABLE II—8.1 OHM CM *p*-TYPE CYCLE 5.
RELATIVE LIGHT INTENSITY 0.25

Time min.	$\mu_{\text{eff}} \frac{\text{cm}^2}{\text{volt sec}}$	δ	$\mu_{\text{eff}}^* \frac{\text{cm}^2}{\text{volt sec}}$	$s \frac{\text{cm}}{\text{sec}}$
0	Started sparked O ₂			
1.0	Changed to dry O ₂			
1.5	307	4.1×10^{-2}	490	503
3.5	318	3.2×10^{-2}	490	660
6.0	Changed to mixture of wet & dry O ₂			
7.5	273	1.4×10^{-2}	376	1480
9.5	239	1.3×10^{-2}	320	1580
11.0	Changed to wet O ₂			
11.5	94	1.2×10^{-2}	-194	1690
12.5	-200	1.1×10^{-2}	-230	1820
15.5	-216	1.2×10^{-2}	-285	1690
17.0	Changed to wet N ₂			
18.5	-352	1.6×10^{-2}	-570	1310
22.0	Changed to mixture of wet & dry O ₂			
25.0	-80	1.1×10^{-2}	-137	1820
26.5	0	1.2×10^{-2}	31	1690
27.5	3.3	1.3×10^{-2}	58	1630
28.0	Changed to dry O ₂			
29.0	193	1.9×10^{-2}	330	1070
29.5	239	2.2×10^{-2}	400	1000
30.5	250	2.4×10^{-2}	420	873
33.0	Tripled flow of dry O ₂			
33.5	296	3.2×10^{-2}	500	645
34.5	296	3.7×10^{-2}	525	560
36.5	296	4.2×10^{-2}	570	490
37.5	296	4.6×10^{-2}	570	455
38.0	Changed to ozone			
42.5	330	6.4×10^{-2}	535	323

III and IV. Values of δ were obtained from the photoconductivity signal, as before, taking the actual illuminated length as the length of the sample. In making use of the standard square-wave calibration for the surface photo-voltage measurement (Section III), it is necessary to allow for the fact that the measured capacity involves the whole length of the sample, plus end and side fringing effects, whereas the surface photo-voltage measurements involves only the illuminated length, plus the fringe effect at the sides.

The penultimate column in Tables III and IV shows the ratio of the change in contact potential, measured in units of (kT/e) , to the added-carrier parameter δ , which was deduced from the photoconductivity. This is not yet, however, the true surface photo-voltage function $(dY/d\delta)$, since the observed change in contact potential includes also the Dember potential $\Delta V_i^{(10)}$ which occurs between the illuminated and non-illuminated parts of the body of the semiconductor. The last column in Tables III and IV shows the true values of $(dY/d\delta)$, obtained by sub-

TABLE III—22.6 OHM CM *n*-TYPE CYCLE 7

Time mins.	Relative Light Intensity	δ	ΔCP volts	$\frac{\beta \Delta CP}{\delta}$	$\frac{dY}{d\delta}$
Starting condition wet N_2					
6.5	2.25	0.36	6.5×10^{-3}	0.7	-0.10
11.5	2.25	0.34	1.0	0.115	-0.045
12.0	Changed to mixture wet and dry N_2				
12.5	2.25	0.32	2.2	0.27	0.10
13.0	2.25	0.34	3.5	0.40	0.23
13.5	2.25	0.35	4.6	0.51	0.34
14.5	2.25	0.38	6.8	0.70	0.53
15.5	0.56	0.10	3.1	1.2	1.03
17.5	0.56	0.11	3.7	1.33	1.16
18.0	Changed to dry O_2				
18.5	0.56	0.16	5.7	1.4	1.2
19.5	0.14	0.06	3.4	2.2	2.0
22.0	Changed to dry O_2 triple flow				
24.5	0.14	0.082	6.1	2.9	2.7

TABLE IV—8.1 OHM CM *p*-TYPE CYCLE 8

Time mins.	Relative Light Intensity	δ	ΔCP volts	$\frac{\beta \Delta CP}{\delta}$	$\frac{dY}{d\delta}$
Starting condition wet N_2					
0	Changed to mixture wet and dry N_2				
0.5	0.14	0.011	-3.5×10^{-3}	-12.5	-12.6
1.0	0.14	0.0088	-2.1	-9.6	-9.7
5.0	Changed to mixture wet and dry O_2				
5.5	0.56	0.0275	-1.8	-2.6	-2.7
6.0	0.56	0.03	-1.45	-1.9	-2.0
7.5	0.56	0.0325	-1.16	-1.4	-1.5
9.5	0.56	0.035	-1.08	-1.2	-1.3
10.0	Changed to dry O_2				
10.5	0.56	0.044	-0.71	-0.63	-0.69
11.5	0.56	0.055	-0.49	-0.35	-0.41
14.0	Changed to dry O_2 triple flow				
14.5	0.56	0.0625	-0.32	-0.20	-0.26
16.5	2.25	0.28	-0.72	-0.10	-0.16
20.5	2.25	0.33	-0.42	-0.05	-0.11
30.0	Changed to ozone				
31.5	2.25	0.47	+0.47	+0.039	-0.023
32.5	2.25	0.53	+1.4	+0.103	+0.041

tracting from $(\beta \Delta e.p./\delta)$ a Dember potential correction, given by $(b - 1)/(\lambda + b\lambda^{-1})$. (The boundaries of the illuminated region were sufficiently distant from the contacts for this formula to apply.)

Tables III and IV include only data from the second half of the cycle (wet $N_2 \rightarrow$ ozone), since the rate of change of Δ e.p. during that part of the first half in which dry oxygen was replaced by wet oxygen was too fast to follow.

The reproducibility of all the data from cycle to cycle was good. One surprising result is that the surface recombination velocity assumed its maximum value close to the "wet nitrogen" extreme for both p-type and n-type. This behavior is quite different from that reported by Brattain and Bardeen,¹ who found s to be constant within 20 per cent throughout the range and Stephenson and Keyes,⁸ who found a maximum value sometimes at one end, sometimes at the other, and sometimes in the middle. There is quite good agreement on the other hand, with the results of Many et al.⁽¹²⁾, who report a maximum in s near the wet end of the cycle. The result of Brattain and Bardeen is not understood at the present time, and is probably wrong. The differences between the present work and that of Stephenson and Keyes may be associated with differences in surface preparation.

V. ANALYSIS OF THE RESULTS

From now onwards we shall express all experimental and calculated quantities in terms of the following dimensionless ratios:

$$\begin{aligned}\bar{\Gamma}_p &= \Gamma_p/n_i\mathfrak{L} & \bar{\Gamma}_n &= \Gamma_n/n_i\mathfrak{L} \\ \bar{\Sigma}_s &= \Sigma_s/en_i\mathfrak{L} & \bar{\Sigma} &= \bar{\Sigma}_s + \bar{\Gamma}_p - \bar{\Gamma}_n \\ \bar{\Delta G} &= \Delta G/en_i\mu_p\mathfrak{L}, & \bar{\mu}_{\text{eff}} &= \mu_{\text{eff}}/\mu_p, & \bar{\mu}_{\text{eff}}^* &= \mu_{\text{eff}}^*/\mu_p\end{aligned}\quad (2)$$

where ΔG is the surface conductance, \mathfrak{L} the Debye length for intrinsic germanium (1.4×10^{-4} cm), and μ_p is the mobility for holes ($1800 \text{ cm}^2\text{V}^{-1}\text{sec}^{-1}$). Tables V and VI show values of the quantities we shall need, as functions of the surface potential Y , calculated from the theoretical considerations of Garrett and Brattain.² The surface conductance, and the differentials in the fifth and sixth columns, are evaluated for $\delta = 0$.

TABLE V — 22.6 OHM CM *n*-TYPE

Y	$Y - \ln \lambda$	$\bar{\Gamma}_p - \bar{\Gamma}_n$	$\bar{\Delta G}$	$\left(\frac{\partial(\bar{\Gamma}_p - \bar{\Gamma}_n)}{\partial Y}\right)_\delta$	$\left(\frac{\partial(\bar{\Gamma}_p - \bar{\Gamma}_n)}{\partial \delta}\right)_Y$
3	4.1	-10.3	17.5	-4.1	-1.3
2	3.1	-7.0	10.6	-2.6	-0.8
1	2.1	-4.9	6.2	-1.8	-0.4
0	1.1	-3.4	3.3	-1.3	0.0
-1	0.1	-2.3	1.45	-1.1	0.5
-2	-0.9	-1.2	0.36	-1.1	1.3
-3	-1.9	0.0	0.0	-1.4	2.7
-4	-2.9	1.7	0.65	-2.1	5.2
-5	-3.9	4.4	2.65	-3.5	9.4
-6	-4.9	8.9	6.8	-5.8	16.3
-7	-5.9	16.4	14.4	-9.5	27.7

TABLE VI — 8.1 OHM CM *p*-TYPE

<i>Y</i>	<i>Y</i> — $\ln \lambda$	$\bar{\Gamma}_p - \bar{\Gamma}_n$	$\Delta \bar{G}$	$\left(\frac{\partial(\bar{\Gamma}_p - \bar{\Gamma}_n)}{\partial Y} \right)_{\delta}$	$\left(\frac{\partial(\bar{\Gamma}_p - \bar{\Gamma}_n)}{\partial \delta} \right)_Y$
8	5.1	-8	9.8	-5.5	-87
7	4.1	-4	3.4	-3.1	-42
6	3.1	-2	0.8	-1.9	-19
5	2.1	0	0.0	-1.45	-8.2
4	1.1	1	0.25	-1.3	-3.4
3	0.1	2	1.3	-1.45	-1.5
2	-0.9	4	2.8	-1.75	-0.62
1	-1.9	6	4.8	-2.4	-0.21
0	-2.9	9	7.4	-3.4	0.0
-1	-3.9	12	10.9	-4.3	0.15
-2	-4.9	18	16.4	-6.4	0.31
-3	-5.9	26	25.0	-10.0	0.53

The first problem is the constructing, from the experimental results, of the curve relating $\Delta \bar{G}$ and $\bar{\Sigma}$. The experiments provide a series of pictures like Fig. 3, each one corresponding to a different chemical environment, and so to a different *Y*. At each of two succeeding pictures of this sort one knows (i) the vertical displacement (photoconductivity) between the dark and light field effect curves; and (ii) the mean difference in the dark and light slopes, and hence the rate of change of photoconductivity with applied field, and therefore with $\bar{\Sigma}$. The problem is to deduce the horizontal displacement (in $\bar{\Sigma}$) between the two pictures.

A correction must first be made for the fact that the ambient changes $\bar{\Sigma}$ uniformly on both surfaces, whereas the applied field induces charge only on the lower surface, plus fringing effects.* The correction is applied by taking the difference in slopes ($\mu_{\text{eff}}^* - \mu_{\text{eff}}$), and multiplying this by (2/1.27), where the number 1.27 is deduced for the given geometry from the standard edge-effect formula.⁹ This having been done, it is now possible to take the revised pictures and piece them together to form two smooth curves (Fig. 4). The process of assembling such a diagram determines the horizontal and vertical distances, and therefore the change of $\bar{\Sigma}$ and $\Delta \bar{G}$, between successive experiments.

This argument may be given analytically as follows. First notice that the photoconductivity voltage in the absence of field (ΔV_2 in Fig. 3) is proportional to $(1/s)$. The application of a voltage between the gold and the germanium induces some charge density Σ at each point on the germanium surface, Σ being (due to fringing effects) a complicated function of position. At each point $(1/s)$ is changed by an amount $\Sigma[d(1/s)/d\Sigma]$. This causes the photoconductivity in the presence of field

* We are indebted to W. L. Brown for bringing this to our attention.

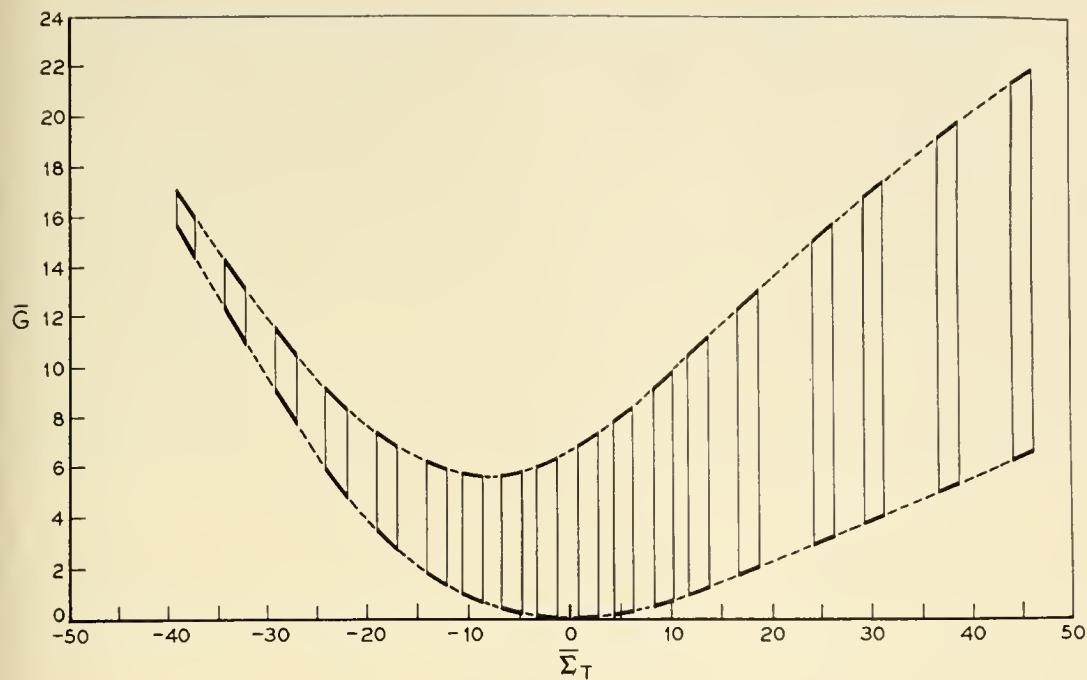


Fig. 4.—Construction of the curve relating \bar{G} (surface conductivity, in units of $e\mu_p n_i \mathfrak{L}$) and $\bar{\Sigma}$ (surface charge, in units of $en_i \mathfrak{L}$).

to differ from that in zero field, and gives rise to the voltage difference $(\Delta V_3 - \Delta V_1)$ shown in Fig. 3. Expressing this difference in terms of the difference $(\mu_{\text{eff}}^* - \mu_{\text{eff}})$ between the apparent and true effective mobilities in the presence of light (see Section IV), one finds:

$$(\mu_{\text{eff}}^* - \mu_{\text{eff}}) = \left(\frac{w^2 t^2}{I \rho_0^2 C} \right) K \frac{d(1/s)}{d\Sigma} \left(\frac{C_{\text{unit}}}{2w} \right) \quad (3)$$

where K is the constant of proportionality between $(1/s)$ and the photoconductivity signal ΔV_2 , and C_{unit} is the capacity per unit of the germanium-gold condenser in the illuminated region, which is 1.27 times the parallel-plate formula. From a series of measurements of $(\mu_{\text{eff}}^* - \mu_{\text{eff}})$ and ΔV_2 it is now possible to obtain $\bar{\Sigma}$ by graphical integration:

$$\Sigma = \left(\frac{1}{en_i \mathfrak{L}} \right) \left(\frac{w^2 t^2}{I \rho_0^2 C} \right) \left(\frac{C_{\text{unit}}}{2w} \right) \int \frac{d\Delta V_2}{\mu_{\text{eff}}^* - \mu_{\text{eff}}} \quad (4)$$

This and the graphical method are of course equivalent. It is worthwhile emphasizing again that either technique depends for its validity on the fact that the distribution of fast states is unaffected by the gas changes in the Brattain-Bardeen cycle, as shown in the experiments of Brown and Montgomery.⁴ If, however, the assumption were too far from the truth, the fitting of both slopes in Fig. 4 would be impossible. The only

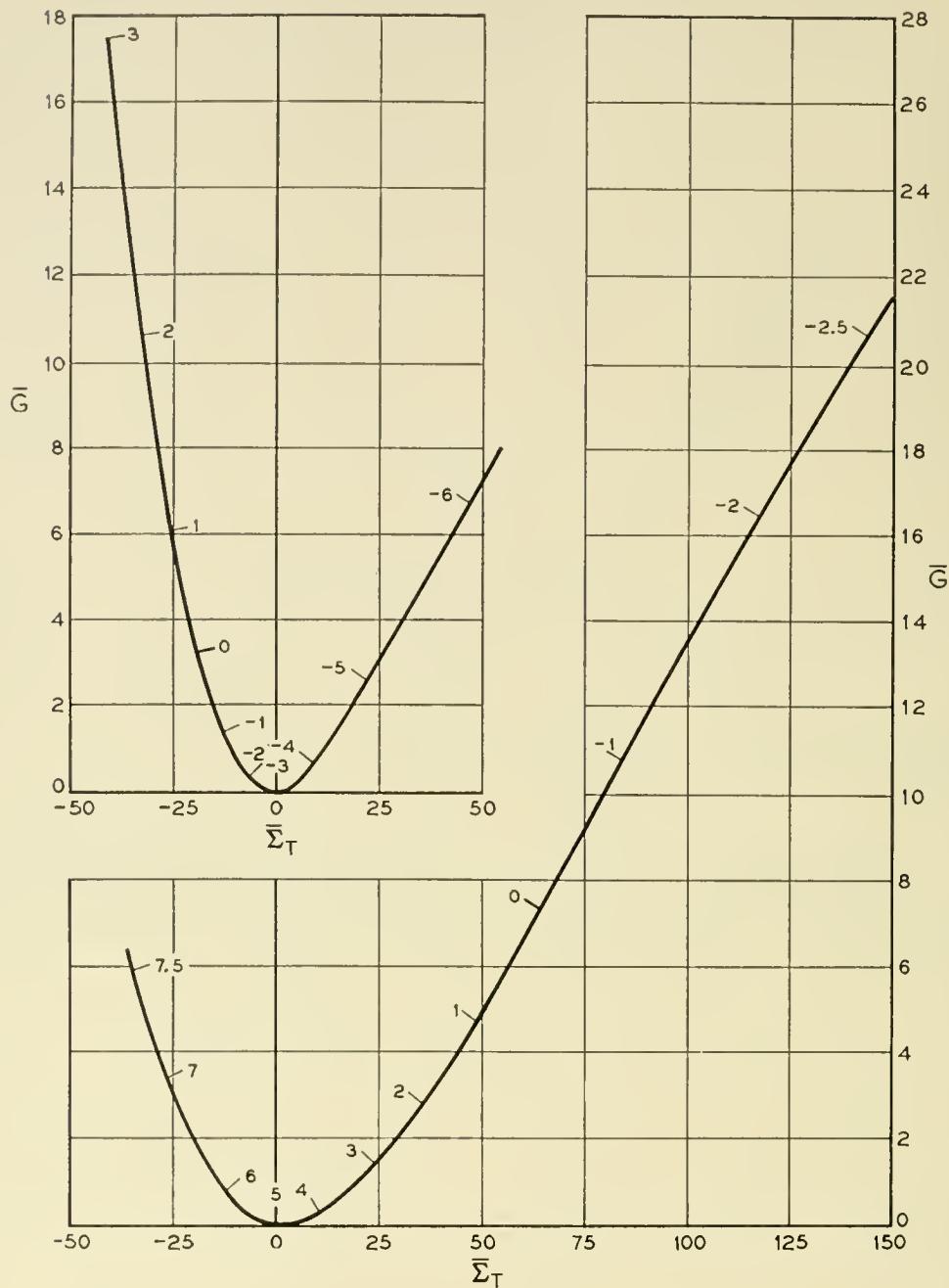


Fig. 5 — Curves showing $\Delta\bar{G}$ (surface conductivity, in units of $e\mu_p n_i \mathcal{L}$) and Σ surface charge, in units of $en_i \mathcal{L}$) for the 22.6 ohm-cm sample (upper curve) and for the 8.1 ohm-cm sample (lower curve). Values of Y , deduced from the surface conductivity, are indicated on the curves.

place at which fitting was at all difficult was at the extreme wet end. For most of the range, therefore, the method is at least internally consistent.

Fig. 5 shows the result of carrying out this procedure for the n and p -type samples. The data were averaged over a number of runs. The numbers appearing on the curves represent values of Y , obtained by reference to Tables V and VI.

From Fig. 5 one may now calculate⁴ the changes occurring in $\bar{\Sigma}_s$, the (reduced) charge in fast states, since $\bar{\Gamma}_p - \bar{\Gamma}_n$ may be read from Tables V and VI, and $\bar{\Sigma}_s = \bar{\Sigma} - (\bar{\Gamma}_p - \bar{\Gamma}_n)$. Fig. 6 shows $(\partial\bar{\Sigma}_s/\partial Y)_\delta$ as a function of $Y - \ln \lambda$, calculated from the experimental results in this way. [The reason for plotting against $Y - \ln \lambda$ instead of Y is that this quantity represents the difference, in units of (kT/e) , between the electrostatic potential at the surface and the Fermi level. In this way the effects of difference from sample to sample in the position of the Fermi level in the interior are eliminated.] Notice that the measurements of $(\partial\bar{\Sigma}_s/\partial Y)_\delta$ for the two samples have the same general shape, and that the turning points of the two curves occur at about the same value of

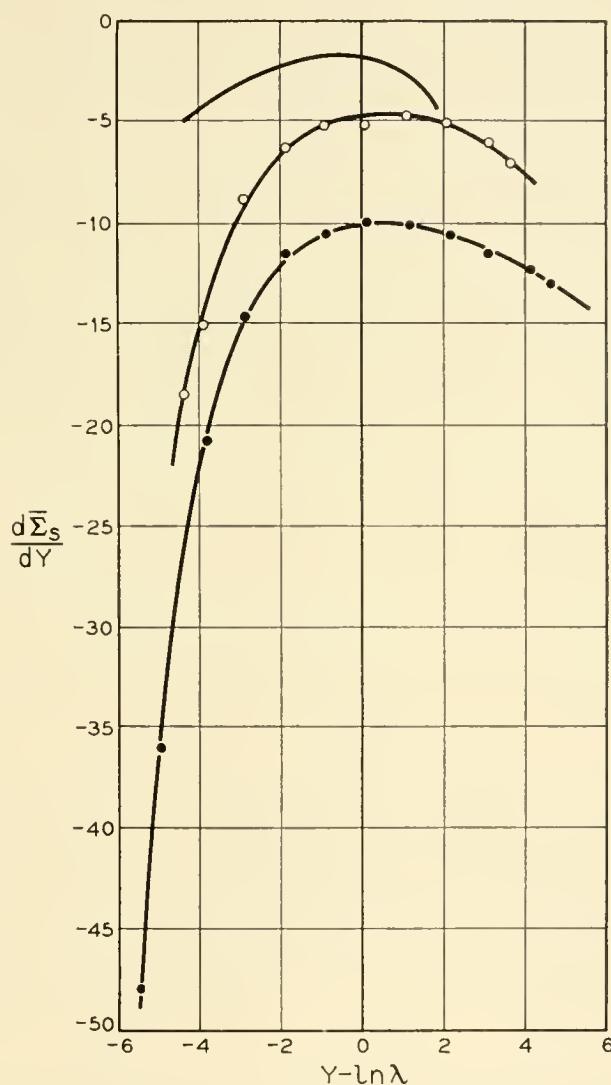


Fig. 6 — Differential charge in fast states versus surface potential. The graphs show $(\partial\bar{\Sigma}_s/\partial Y)$ plotted against $Y - \ln \lambda$. Dots: p-type; circles: n-type. A typical result of Brown and Montgomery, using 28 ohm-cm p-type germanium, is also shown.

$(Y - \ln \lambda)$. Fig. 7 shows the variation of surface recombination velocity with $Y - \ln \lambda$, using the experimental photoconductivity data and values of Y read from Fig. 5. The values of s have been divided by $(\lambda + \lambda^{-1})$, as indicated, since $s/(\lambda + \lambda^{-1})$ is expected to be the same, at a given value of $(Y - \ln \lambda)$, for all samples, so long as the distribution of fast states is the same. The agreement shown in Fig. 7 is probably closer than would be expected in the light of the experimental accuracy.

Fig. 8 shows the observed dependence of $dY/d\delta$ on $(Y - \ln \lambda)$ for both samples, using the data of Tables III and IV, and using the photoconductivity to determine, from Fig. 7, the value of Y at each point. On the figure the expected limiting values ($-\lambda$ and λ^{-1}) are shown for both samples. Of the four asymptotes, the higher limit of $(dY/d\delta)$ for the n -type sample is satisfactorily reached for large negative values of Y ; the experimental values for the p -type sample appear to be approaching the expected limit for large positive values of Y , while the information regarding the approach to the two lower limits is too fragmentary to do more than show that the order of magnitude is as expected. Now taking the data shown in Fig. 8, making use of (1) and the calculations given in Tables III and IV, one calculates $(\partial \bar{\Sigma}_s / \partial \delta)_Y / (\partial \bar{\Sigma}_s / \partial Y)_\delta$. The values so found are plotted against Y in Fig. 9. Fig. 6, 7 and 9, showing the observed variation of $(\partial \bar{\Sigma}_s / \partial Y)_\delta$, s and $(\partial \bar{\Sigma}_s / \partial \delta)_Y / (\partial \bar{\Sigma}_s / \partial Y)_\delta$ with Y , furnish a complete description of the properties of the fast states at the

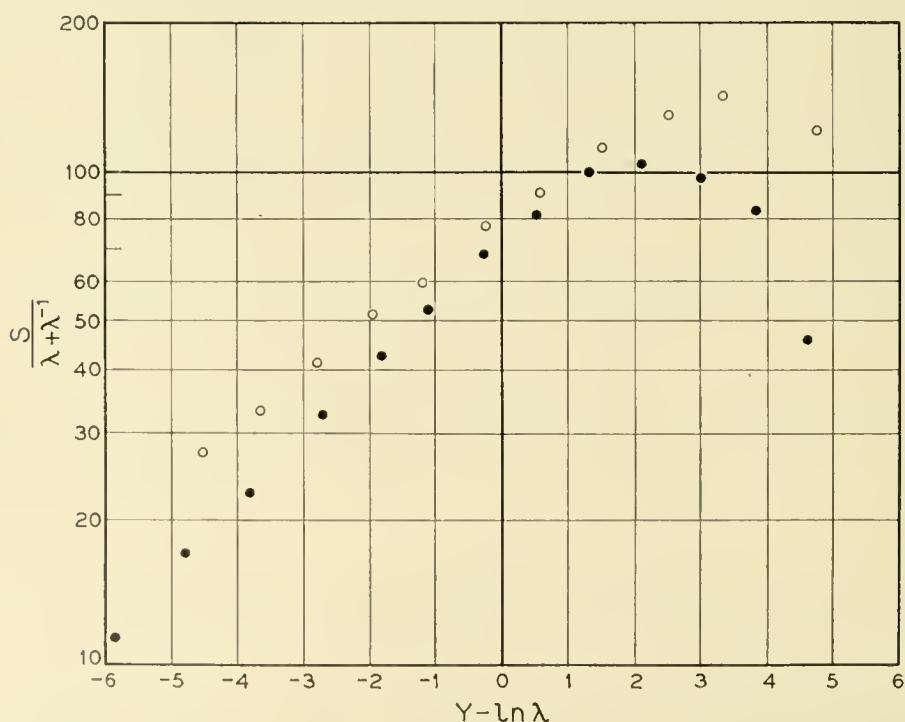


Fig. 7 -- Surface recombination velocity versus surface potential. The curves show $s/(\lambda + \lambda^{-1})$ plotted against $Y - \ln \lambda$. Dots: p -type; circles: n -type.

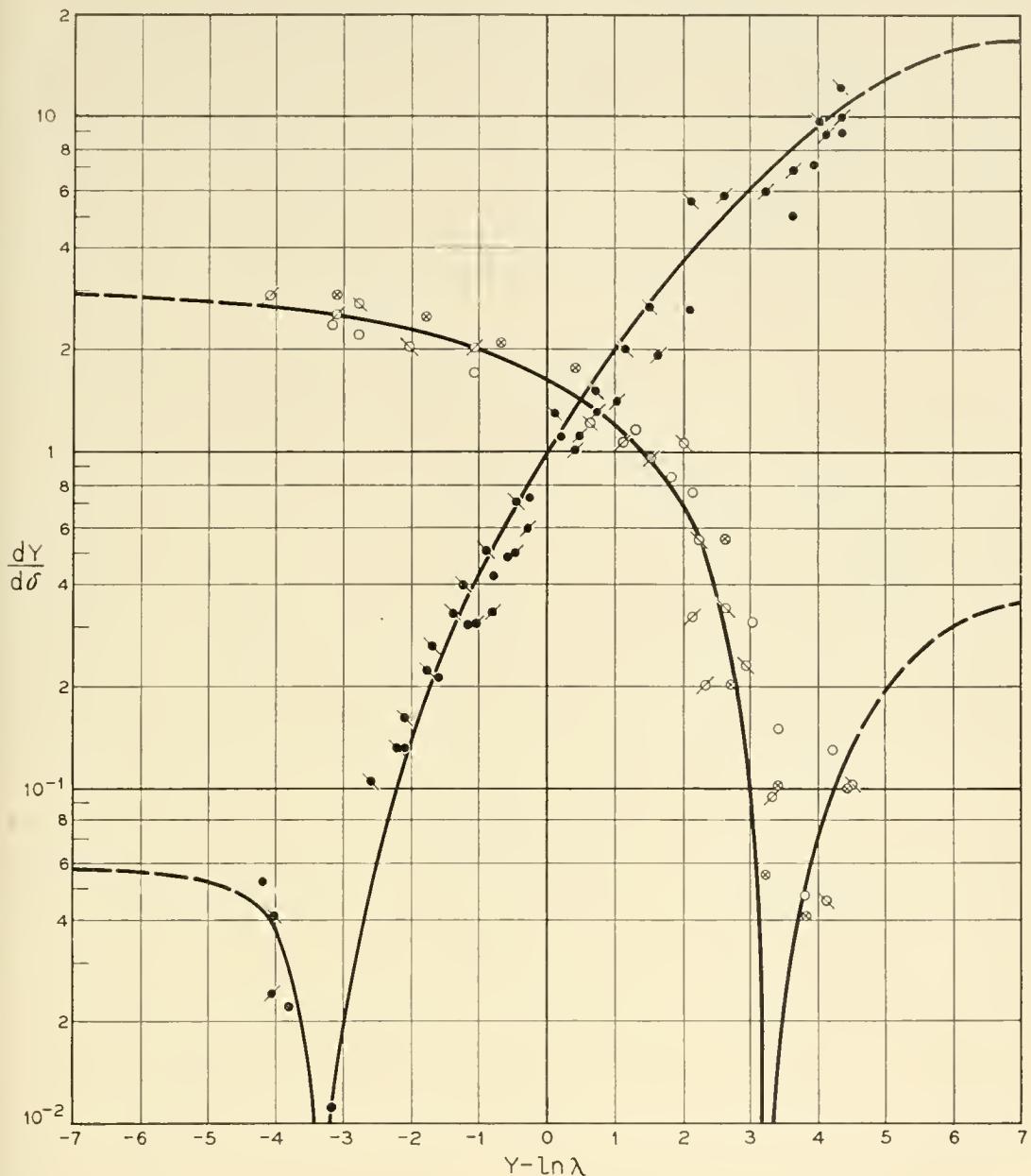


Fig. 8—Surface photo-voltage (change in contact potential in relation to added carrier concentration). $dY/d\delta$ is shown plotted against $Y - \ln \lambda$. Dots: p -type; circles: n -type. Data from different runs are distinguished by modifications to these symbols. The left-hand branches denote absolute magnitudes, since the ratio is negative there. At the extreme left hand of the diagram, the fast states near to the Fermi level are in good contact with the valence band: at the extreme right hand, to the conduction band. The theoretical asymptotes (λ^{-1} to the left and λ to the right) are also indicated.

temperature studied. This is the basic information which any theoretical treatment must explain. In the succeeding paper this matter is discussed from the point of view of the statistics of a distribution of fast states, and information on the cross sections, as well as on the distribution itself, is derived from the data just presented.

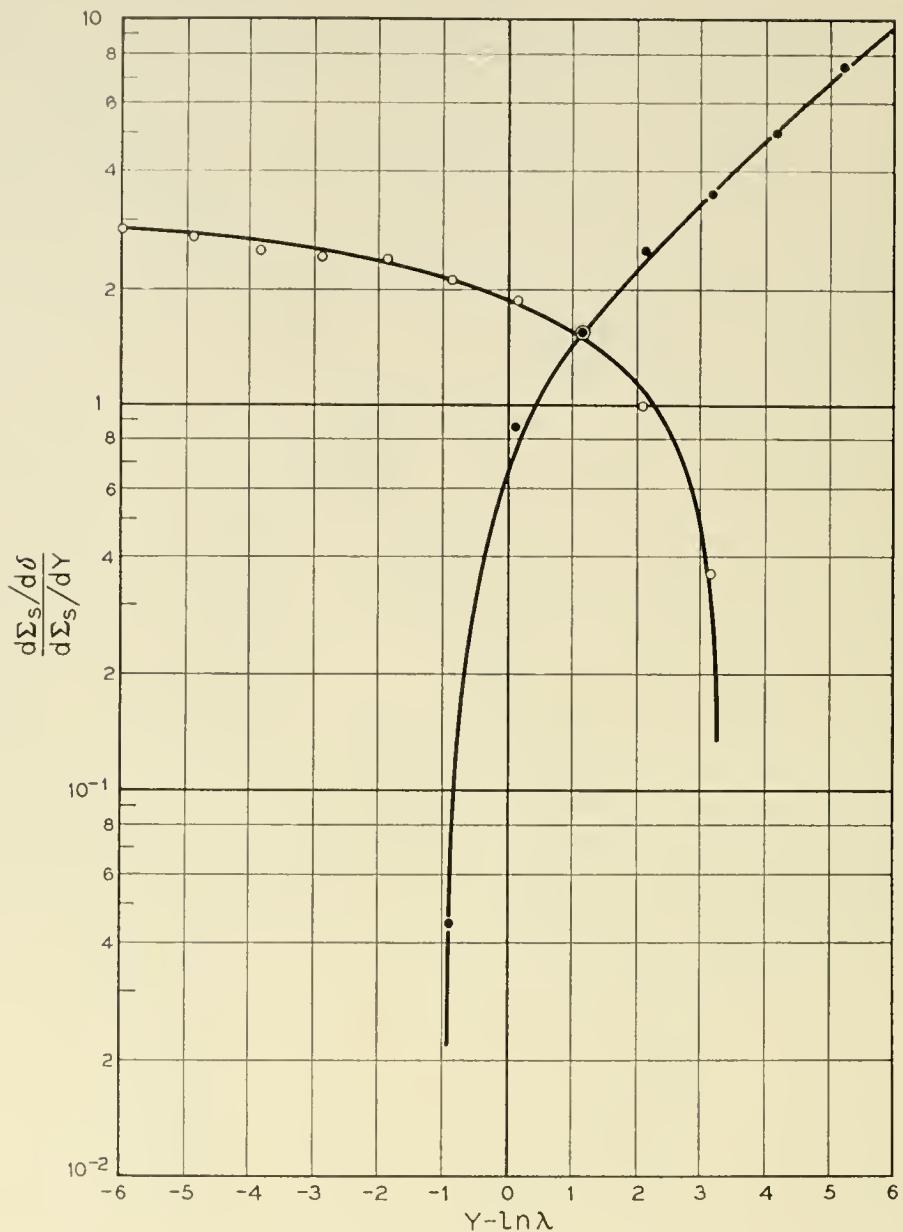


FIG. 9—The function $(\partial \Sigma_s / \partial \delta) Y / (\partial \Sigma_s / \partial Y)_\delta$ plotted against $Y - \ln \lambda$. Dots: *p*-type; circles: *n*-type.

VI. FURTHER COMMENTS

The development given in the previous section has concerned particularly the properties of the fast states. As to the slow states, the experiments are much less informative. The variations of Y with gas are generally consistent with the variations of contact potential previously reported,¹ although the total range in Y (± 0.13 volt) is smaller by about a factor of 2 than that in contact potential found in the previous work. One must say that roughly half the change of contact potential is in V_B , (i.e., βY) and half in V_D , the potential drop across the ion layer.

It may be seen from the figures that it is the quantity ($Y - \ln \lambda$), rather than Y , which appears to be characteristic of the point in the cycle reached. This property of a semiconductor surface, and possible reasons therefore, have often been discussed in the literature.¹¹ The total range of surface potential is illustrated in Fig. 10, which is drawn to scale, and also shows sundry other points of interest found in the present research. The potential diagrams for n-type and p-type are drawn with the Fermi levels aligned, to show the relation between the property ($Y - \ln \lambda$) = const. and the frequently observed smallness of the contact potential difference between n and p-type germanium.

As to the reproducibility and accuracy of the work presented here, the following points may be of interest: (i) The measurements were repeated on another n-type sample of nearly the same resistivity as the one reported here, but cut from a different crystal. The results on this sample were indistinguishable, within the experimental error, from those found on the first n-type sample. (ii) If the sample was re-etched in precisely the same way as before, and the experiments repeated, the results were in good agreement with those obtained before. However, variations in the etching procedure sometimes gave quite different re-

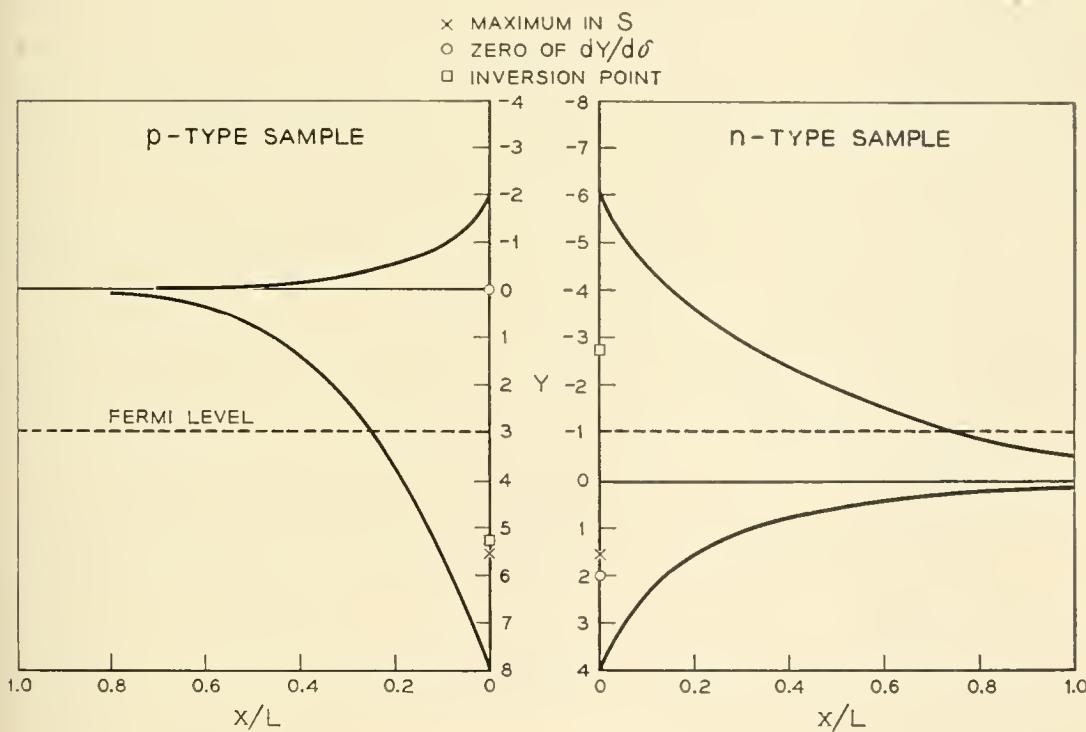


Fig. 10 — The shapes of the surface space-charge regions for the *p*-type and *n*-type samples in the extremes of gaseous environment. The two surfaces are to the center of the figure. The solid curves show the center of the gap (intrinsic Fermi level) plotted against distance, in units of an intrinsic Debye length. Also shown are the positions of the zeros of $(dY/d\delta)$, the maxima of s , and the minima of surface conductivity.

sults. We hope to discuss this at a future date. (iii) The accuracy of the measurements is not high. Some of the more directly-derivable quantities, such as s , should be known to 5 per cent, but a quantity like $(\partial \bar{\Sigma}_s / \partial \delta) / (\partial \bar{\Sigma}_s / \partial Y)$, which is only obtained after a long and elaborate calculation involving a number of corrections, is perhaps uncertain to 30 per cent.

VII. CONCLUSIONS

This paper has presented results of combined measurements of field effect, photoconductivity, change of photoconductivity with field, filament lifetime and surface photo-voltage, on slices of germanium. From the measurements, the surface potential Y has been found at each point, and the variations of the quantities $(\partial \bar{\Sigma}_s / \partial Y)$, s and $(\partial \bar{\Sigma}_s / \partial \delta) / (\partial \bar{\Sigma}_s / \partial Y)$ with Y determined.

It is a pleasure to record our thanks to W. L. Brown, for comments on field effect techniques and many stimulating discussions, to H. R. Moore, who constructed the high-voltage power supply, and to A. A. Studna, who assisted in the experiments. We are also grateful to C. Herring for comments on the text.

BIBLIOGRAPHY

1. W. H. Brattain and J. Bardeen, Surface Properties of Germanium, B.S.T.J., **32**, pp. 1-41, Jan. 1953.
2. C. G. B. Garrett and W. H. Brattain, Physical Theory of Semiconductor Surfaces, Phys. Rev., **99**, pp. 376-387, July 15, 1955.
3. W. L. Brown, Surface Potential and Surface Charge Distribution from Semiconductor Field Effect Measurements, Phys. Rev., **98**, p. 1565, June 1, 1955.
4. H. C. Montgomery and W. L. Brown, Field-Induced Conductivity Changes in Germanium, Phys. Rev., **103**, Aug. 15, 1956.
5. J. R. Schrieffer, Effective Carrier Mobility in Surface Charge Layers, Phys. Rev., **97**, pp. 641-646, Feb. 1, 1955.
6. C. G. B. Garrett and W. H. Brattain, Interfacial Photo-Effects in Germanium at Room Temperature, Proc. of the Conference on Photo Conductivity, Nov., 1954, Wiley, in press.
7. W. H. Brattain and C. G. B. Garrett, Surface Properties of Germanium and Silicon, Ann. N. Y. Acad. of Science, **58**, pp. 951-958, Sept., 1954.
8. D. T. Stevenson and R. J. Keyes, Measurements of Surface Recombination Velocity at Germanium Surfaces, Physica, **20**, pp. 1041-1046, Nov., 1954.
9. J. Clerk Maxwell, Electricity and Magnetism, 3rd Edition, **1**, p. 310, Clarendon Press, 1904.
10. W. van Roosbroeck, Theory of Photomagnetoelectric Effect in Semiconductors, Phys. Rev., **101**, pp. 1713-1725, March 15, 1956.
11. J. Bardeen and S. R. Morrison, Surface Barriers and Surface Conduction, Physica, **20**, p. 873, 1954.
12. E. Harnik, A. Many, Y. Margoninski and E. Alexander, Correlation Between Surface Recombination Velocity and Surface Conductivity in Germanium, Phys. Rev., **101**, pp. 1434-1435, Feb. 15, 1956.

Distribution and Cross-Sections of Fast States on Germanium Surfaces

By C. G. B. GARRETT and W. H. BRATTAIN

(Manuscript received May 10, 1956)

A theoretical treatment of the field effect, surface photo-voltage and surface recombination phenomena has been carried out, starting with the Hall-Shockley-Read model and generalizing to the case of a continuous trap distribution. The theory is applied to the experimental results given in the previous paper. One concludes that the distribution of fast surface states is such that the density is lowest near the centre of the gap, increasing sharply as the accessible limits of surface potential are approached. From the surface photo-voltage measurements one obtains an estimate of 150 for the ratio (σ_p/σ_n) of the cross-sections for transitions into a state from the valence and conduction bands, showing that the fast states are largely acceptor-type. On the assumption that surface recombination takes place through the fast states, the cross-sections are found to be: $\sigma_p \sim 6 \times 10^{-15} \text{ cm}^2$ and $\sigma_n \sim 4 \times 10^{-17} \text{ cm}^2$.

1. INTRODUCTION

The existence of traps, or "fast" states, on a semiconductor surface, becomes apparent from three physical experiments: measurements of field effect,¹ of surface photovoltage,² and of surface recombination velocity s . Results of combined measurements of these three quantities on etched surfaces of *p*- and *n*-type germanium have been presented in the preceding paper.³ The present paper is concerned with the conclusions which may be drawn from these experiments as to the distribution in energy of these surface traps, and the distribution of cross-sections for transitions between the traps and the conduction and valence bands.

The statistics of trapping at a surface level has been developed by Brattain and Bardeen² and by Stevenson and Keyes,⁴ following the work on body trapping centers of Hall⁵ and of Shockley and Read.⁶

It is known that surface traps are numerous on a mechanically damaged surface⁷ or on a surface that has been bombarded but not annealed;⁸

and that on an etched surface their density is comparatively low. It is also known that the available results cannot be accounted for by a single level, or even two levels, so that one is evidently dealing either with a large number of discrete states or a continuous spectrum. A given trapping centre is completely described by specifying: (i) whether it is donor-like (either neutral or positive) or acceptor-like (neutral or negative); (ii) its position in energy; and (iii) the values for the constants C_p and C_n (related to cross-sections) occurring in the Shockley-Read theory. In this paper we shall deduce what we can about these quantities, using the experimental results previously presented.

At the outset it must be admitted that it is by no means certain that the same set of surface states appear in the field-effect experiment and give rise to surface recombination. However, (i) it is found that such surface treatments as increase s also reduce the effective mobility in the field-effect experiment; (ii) any surface trap must be able to act as a recombination centre, unless one of the quantities C_p and C_n is zero;⁹ and (iii) the capture cross-sections obtained by assuming that the field-effect traps are in fact recombination centres are, as we shall see below, eminently reasonable.

As to the nature of the surface traps, not too much can be said at the moment. The lack of sensitivity to the cycle of chemical environment used argues against their being associated with easily desorbable surface atoms; the intrinsically short time constants (Section 5) suggest that they are on or very close to the germanium surface. The possibility that the surface traps are Tamm levels¹⁰ remains; or they could be corners or dislocations. However, the reproducibility with which a given value of s may be obtained by a given chemical treatment of a given sample, followed by exposure to a given ambient, suggests that there is nothing accidental about their occurrence.

II. STATISTICS OF A DISTRIBUTION OF SURFACE TRAPS

We start by quoting results from the work of Shockley and Read⁶ and Stevenson and Keyes⁴ on the occupancy factor f_t and the flow U of minority carriers (per unit area) into a set of traps having a single energy level and statistical weight unity:

$$f_t = (C_n n_s + C_p p_1) / [C_n(n_s + n_1) + C_p(p_s + p_1)] \quad (1)$$

$$U = C_n C_p (p_s n_s - n_i^2) / [C_n(n_s + n_1) + C_p(p_s + p_1)] \quad (2)$$

where the symbols have the following meanings:

n_s , p_s — densities of electrons and holes present at the surface

n_1, p_1 — values which the equilibrium electron and hole densities at the surface would have if the Fermi level coincided with the trapping level

$C_n = N_t v_{Tn} \sigma_n$; $C_p = N_t v_{Tp} \sigma_p$, where N_t stands for density of traps per unit area, v_{Tn} is the thermal speed for electrons and v_{Tp} that for holes, and σ_n and σ_p are the cross-sections for transitions between the traps and the conduction and valence bands respectively.

If we introduce the surface potential Y and the quantity δ , defined as $(\Delta p/n_i)$, where Δp is the added carrier density in the body of the semiconductor, we may write:

$$\begin{aligned} n_s &= \lambda^{-1} n_i e^Y (1 + \lambda \delta) \\ p_s &= \lambda n_i e^{-Y} (1 + \lambda^{-1} \delta) \end{aligned} \quad (3)$$

where $\lambda = p_0/n_i$, p_0 being the equilibrium hole concentration in the body of the semiconductor. We further introduce the notation:

$$\begin{aligned} n_1 &= n_i e^{-\nu} & p_1 &= n_i e^\nu \\ (C_p/C_n)^{\frac{1}{2}} &= \chi \end{aligned} \quad (4)$$

The quantity ν thus represents the energy difference, measured in units of (kT/e) , between the trapping level and the centre of the gap;* and is positive for states below, negative for those above, this level. The parameter χ will be most directly associated with whether the state is donor-like or acceptor-like. If it is donor-like (neutral or positive), a transition involving an electron in the conduction band will be aided by Coulomb attraction whereas one involving a hole will not; so one would expect $\chi \ll 1$. For an acceptor-like trap, (neutral or negative) the contrary holds, and one expects $\chi \gg 1$.

Using (4), the occupancy factor (1) becomes

$$\begin{aligned} f_t &= \frac{\chi^{-1} \lambda^{-1} e^Y (1 + \lambda \delta) + \chi e^\nu}{\chi^{-1} \lambda^{-1} e^Y (1 + \lambda \delta) + \chi^{-1} e^{-\nu} + \chi \lambda e^{-Y} (1 + \lambda^{-1} \delta) + \chi e^\nu} \\ &= \frac{1}{2} \lambda^{-\frac{1}{2}} e^{-\frac{1}{2} Y} e^{\frac{1}{2} \nu} \operatorname{sech} [\frac{1}{2} (Y + \nu) - \frac{1}{2} \ell \ln \lambda] \quad \text{for } \delta = 0 \end{aligned} \quad (5)$$

Note that, in thermodynamic equilibrium, the occupancy factor does not depend in any way on the cross-sections, whereas for $\delta \neq 0$ it does, through the ratio χ .

* Strictly speaking, one should say "position of the Fermi level for intrinsic semiconductor" instead of "centre of the gap." These will fail to coincide if the effective masses of holes and electrons are unequal, as they certainly are in germanium.

Similarly, the flow of carrier-pairs to the surfacee (2) becomes:

$$U =$$

$$N_t(v_{Tn}v_{Tp})^{1/2}(\sigma_n\sigma_p)^{1/2}n_i \frac{(\lambda + \lambda^{-1})\delta + \delta^2}{\chi^{-1}\lambda^{-1}e^r(1 + \lambda\delta)\chi^{-1}e^{-r} + \chi\lambda e^{-r}(1 + \lambda^{-1}\delta)\chi e^r} \quad (6)$$

which, for $\delta \rightarrow 0$, tends to the linear law $U = sn_i\delta$, where s , the surface recombination velocity, is given by:

$$s/(v_{Tn}v_{Tp})^{1/2} = N_t S_t$$

where

$$S_t = (\lambda + \lambda^{-1})(\sigma_n\sigma_p)^{1/2}/2[ch(\nu + \ell n \chi) + ch(Y - \ell n \lambda - \ell n \chi)] \quad (7)$$

The surface density Σ_s of trapped charge is given by:

$$\Sigma_s = N_t f_t \quad (8)$$

where f_t is the occupancy factor, given by (5).

Now let us turn to the question of a distribution of surfacee traps through the energy ν . Suppose that the density of states having ν lying between ν and $\nu + d\nu$ is $\bar{N}(\nu) d\nu$, expressed in units ($n_i\mathfrak{L}$). Then the total surface recombination velocity arising from all traps, and the total trapped surface charge density, are given by:

$$s/(v_{Tn}v_{Tp})^{1/2} = n_i \mathfrak{L} \int S_t(\nu) \bar{N}(\nu) d\nu \quad (9)$$

$$\bar{\Sigma}_s = \int f_t(\nu) \bar{N}(\nu) d\nu \quad (10)$$

where $S_t(\nu)$ and $f_t(\nu)$ are explicit functions of ν , given by (5) and (7). The limits of the integrals in (9) and (10) are the values of ν corresponding to the conduction and valence band edges; however, as we shall see, it is often possible to replace these limits by $\pm \infty$.

In summing up the contributions in the way represented by (9), we have implicitly ignored the possibility of inter-trap transitions, supposing that the population of each trap depends only on the rates of exchange of charge with the conduction and valence bands, and is independent of the population of any other trap of differing energy.

What kind of function do we expect $N(\nu)$ to be? Brattain and Bardeen² postulated that $N(\nu)$ was of the form of two delta-functions, corresponding to discrete trapping levels high and low in the band. This assumption is not consistent with the observed facts in regard to field effect, surfacee

photo-voltage, or surface recombination velocity. The general difficulty is that the observed quantities usually vary less rapidly with surface potential than one would expect. It is possible to fit the field-effect observations of Brown and Montgomery¹¹ with a larger number of discrete levels, but this would call for a "sharpening up" of the trapped charge distribution as the temperature is lowered, and this appears to be contrary to what is observed.* It is always possible that the surface is patchy, in which case almost *any* variation with mean surface potential could be explained. The simplest assumption, however, seems to be that $N(\nu)$ is a rather smoothly-varying function. All we need assume for the moment is that it is everywhere finite, continuous and differentiable. We may then differentiate equation (10) with respect to Y and δ under the integral sign, and get $(\partial \bar{\Sigma}_s / \partial Y)_\delta$ and $(\partial \bar{\Sigma}_s / \partial \delta)_Y$, the quantities for which experimental measurements were reported in the previous paper:³

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial Y} \right)_\delta = \int \frac{\bar{N}(\nu) d\nu}{4 ch^2[\frac{1}{2}(\nu + Y) - \frac{1}{2}\ln \lambda]} \quad (11)$$

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial \delta} \right)_Y = - \int \frac{\bar{N}(\nu) (\frac{1}{2}(\lambda^{-1} + \lambda)lh[\frac{1}{2}(\nu - Y) + \frac{1}{2}\ln \lambda] + \ln \chi) d\nu}{4ch^2[\frac{1}{2}(\nu + Y) - \frac{1}{2}\ln \lambda]} \quad (12)$$

Notice that the expression in brackets in the numerator of (12) generally has the value λ^{-1} or $-\lambda$, except near the point $\nu = Y - \ln \lambda - 2\ln \chi$. This is indicative of the fact that, whatever the exact form of $\bar{N}(\nu)$, the ratio of $-(\partial \Sigma_s / \partial \delta)_Y / (\partial \Sigma_s / \partial Y)_\delta$ tends to these limiting values (λ^{-1} and $-\lambda$) for sufficiently large negative and positive Y respectively.

It may be verified from (7), (11) and (12) that $(\partial \bar{\Sigma}_s / \partial Y)_\delta$, found from the field effect experiment, depends only on $N(\nu)$; $(\partial \bar{\Sigma}_s / \partial \delta)_Y$, found from the surface photo-voltage, depends on $\bar{N}(\nu)$ and χ ; while s , the surface recombination velocity, depends in addition on the geometric mean cross-section $(\sigma_n \sigma_p)^{1/2}$. Both χ and $(\sigma_n \sigma_p)^{1/2}$ might themselves, of course, be functions of ν . Thus relations (7), (11) and (12) are integral equations, from which the three unknown functions of ν may in principle be deduced from the experimental results. (Equation 11, in fact, may be solved explicitly. P. A. Wolff¹⁷ has shown, however, that, to determine $N(\nu)$ unambiguously, it is necessary to know $(\partial \bar{\Sigma}_s / \partial Y)_\delta$ for all values of Y in the range $\pm \infty$.)

The foregoing considerations apply to "small-signal" measurements.

* There are some changes with temperature, but not what one would expect if there were only discrete surface states.

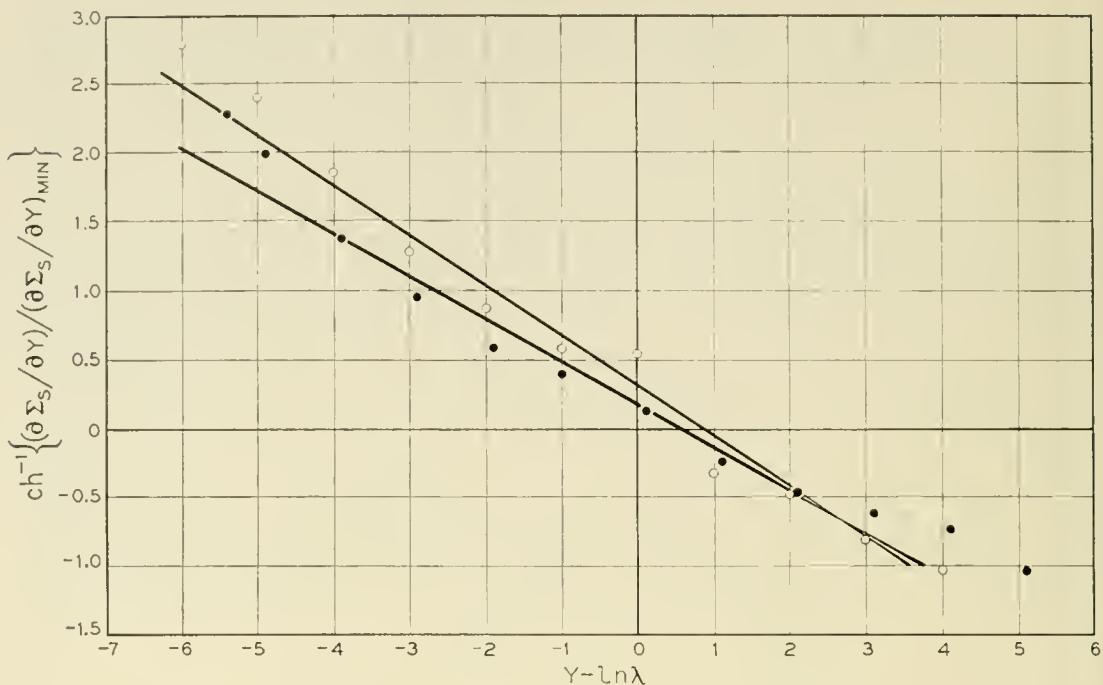


Fig. 1 — The fit between Equations (13) and (14) and the experimental data. The circles and dots give the experimental data for the *n* and *p*-type samples respectively and the solid straight lines represent Equations (13) and (14).

But it is also possible, once $N(\nu)$, χ and $(\sigma_n \sigma_p)^{1/2}$ are known, to calculate the expected behavior of the surface photo-voltage and surface recombination rate at high light intensities, and compare the answer with the experimental findings. We hope to discuss this matter in a later paper.

III. ANALYSIS OF THE EXPERIMENTAL DATA BY USE OF THE DELTA-FUNCTION APPROXIMATION

Let us first consider the interpretation of our field effect measurements by means of (11). We start by finding empirical expressions that describe the observed dependence of $(\partial \bar{\Sigma}_s / \partial Y)$ on Y (Fig. 6 of the preceding paper³). Except at values of $(Y - \ln \lambda)$ close to the extremes reached one may fit quite well by a hyperbolic cosine function. Fig. 1 shows the function whose hyperbolic cosine is $(\partial \bar{\Sigma}_s / \partial Y) / (\partial \bar{\Sigma}_s / \partial Y)_{\min}$ plotted against $Y - \ln \lambda$. From this figure we find:

22.6 ohm-cm *n*-type:

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial Y} \right)_s = 4.5ch[0.36(Y - \ln \lambda) - 0.8] \quad (13)$$

(for $(Y - \ln \lambda) > -4$

8.1 ohm-cm p-type:

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial Y} \right)_s = 9.7 ch[0.31(Y - \ln \lambda) - 0.5] \quad (14)$$

$$\text{for } 2 > (Y - \ln \lambda) > -4$$

For values of $(Y - \ln \lambda)$ less than -4 , it appears that Σ_s is changing more rapidly with Y than is indicated by (13) and (14). We shall comment on this point later. Excluding this region, we note that in both cases the variation with Y is everywhere slow in comparison with e^Y , and proceed on the assumption that $\bar{N}(\nu)$ is a function of ν that varies everywhere slowly in comparison with e^ν . Then (11) indicates that there is one fairly sharp maximum in the integrand in the range $\pm \infty$, occurring at that value of ν which coincides with the Fermi level:

$$\nu = -Y + \ln \lambda \quad (15)$$

The integral in (11) could be evaluated in series about this point (method of steepest descents). The zero-order approximation is got by replacing

$$\frac{1}{4} \operatorname{sech}^2 [\frac{1}{2}(\nu + Y) - \frac{1}{2}\ln \lambda] \quad \text{by} \quad \delta(\nu + Y - \ln \lambda).$$

Later we shall proceed to an exact solution, and we shall find that this delta-function approximation is not too bad. From (11) we now find:

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial Y} \right)_s \sim \int \bar{N}(\nu) \delta(\nu + Y - \ln \lambda) d\nu = \bar{N}(-Y + \ln \lambda) \quad (15)$$

This mathematical procedure will be seen to be equivalent to identifying $(\partial \bar{\Sigma}_s / \partial Y)_s$ with the density of states at the point in the gap which coincides with the Fermi-level at the surface. Using (13) and (14), one gets:

22.6 ohm-cm n-type:

$$\bar{N}(\nu) = 4.5 ch(0.36\nu + 0.8) \quad (16)$$

8.1 ohm-cm p-type:

$$\bar{N}(\nu) = 9.7 ch(0.31\nu + 0.5) \quad (17)$$

As we shall see in the next section, the exact solutions differ from (16) and (17) only in the coefficients preceding the hyperbolic cosines.

Turning to the surface photo-voltage measurements, we take (12) and again replace

$$\frac{1}{4} \operatorname{sech}^2 [\frac{1}{2}(\nu + Y) - \frac{1}{2}\ln \lambda] \quad \text{by} \quad \delta(\nu + Y - \ln \lambda)$$

Using (15), one gets:

$$\begin{aligned} & - \frac{(\bar{\partial}\Sigma_s/\partial\delta)_Y}{(\bar{\partial}\Sigma_s/\partial Y)_\delta} \\ & = \frac{1}{2}(\lambda^{-1} + \lambda) \operatorname{th}(-Y + \ln\lambda + \ln\chi) + \frac{1}{2}(\lambda^{-1} - \lambda) \end{aligned} \quad (18)$$

This procedure, inaccurate as it is, has the advantage that no particular assumption need be made concerning the functional dependence of χ on ν , it being understood that χ in (18) has the value holding for $\nu = -Y + \ln\lambda$. In particular, if Y_0 is that value of Y at which the ratio $-(\bar{\partial}\Sigma_s/\partial\delta)_Y/(\bar{\partial}\Sigma_s/\partial Y)_\delta$ changes sign,

$$\ln\chi_0 = Y_0 - \ln\lambda + \operatorname{th}^{-1}[(\lambda - \lambda^{-1})/(\lambda + \lambda^{-1})] \quad (19)$$

From the experimental data, one finds, for the n -type sample, $\ln\chi_0 \sim 2.4$ (at $\nu = -3.5$); for the p -type sample, $\ln\chi_0 \sim 1.0$ (at $\nu = 1.9$).

In view of the approximations made, these estimates would not be expected to be more precise than ± 1 to 2 units. Notice that both values are positive, and that the difference between them is small in comparison with the difference in ν . This suggests that we start afresh with the assumption that χ is independent of ν , and work out the surface photovoltage integral exactly. This is done in the next section.

IV. EXACT TREATMENT FOR THE CASE $\bar{N}(\nu) = A \operatorname{ch}(q\nu + B)$, WITH CONSTANT CROSS-SECTIONS

The results of the previous section suggest the procedure of assuming that $N(\nu)$ is of the functional form given by (16) and (17), and evaluating the integrals (9), (11) and (12) exactly. The integral for $(\bar{\partial}\Sigma_s/\partial Y)$, (11), depends only on the form of $N(\nu)$ and may be evaluated at once. To get $(\bar{\partial}\Sigma_s/\partial\delta)$, (12), one must know how χ depends on ν . On the basis of the work of the previous section, we shall suppose that χ is independent of ν . (Properly, we need only assume that χ varies with ν more slowly than e^ν . Since the function $\operatorname{th}[\frac{1}{2}(\nu - Y) + \frac{1}{2}\ln\lambda + \ln\chi]$ has one of the values ± 1 everywhere except close to $\nu = Y - \ln\lambda - 2\ln\chi$, and since the denominator of (12) has a sharp minimum at $\nu = -Y + \ln\lambda$, it follows that the region in which $(\bar{\partial}\Sigma_s/\partial\delta)_Y$ changes sign will be governed mainly by the value of χ at $\nu = -\ln\chi$.) To get s [(9)], using (7), one must also assume something about the geometric mean cross-section, $(\sigma_n\sigma_p)^{1/2}$. In the absence of any information on this score, we shall assume that $(\sigma_n\sigma_p)^{1/2}$ also is independent of ν , and see how the computed variation of s with Y compares with the experimental results.

We assume:

$$\bar{N}(\nu) = A \operatorname{ch}(q\nu + B) \quad (20)$$

and substitute in (11), (12) and (7). In view of the sharp maximum in the integrands of these expressions, it is permissible to set the limits which should correspond to the edges of the gap or of the state distribution equal to $\pm \infty$. The integrals are conveniently evaluated by the contour method (see Appendix 1) and yield the following results:

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial Y} \right)_\delta = A \pi q \operatorname{cosec} \pi q \operatorname{ch}[B - q(Y - \ln \lambda)] \quad (21)$$

$$\left(\frac{\partial \bar{\Sigma}_s}{\partial \delta} \right)_Y = -A \pi q \operatorname{cosec} \pi q \operatorname{ch}[B - q(Y - \ln \lambda)] \times$$

$$\left[\frac{1}{2}(\lambda^{-1} + \lambda) \left(-\operatorname{coth} \mathcal{Y} + \frac{\operatorname{sh} q\mathcal{Y} \operatorname{ch} \mathfrak{B}}{q \operatorname{sh}^2 \mathcal{Y} \operatorname{ch}(q\mathcal{Y} - \mathfrak{B})} \right) + \frac{1}{2}(\lambda^{-1} - \lambda) \right] \quad (22)$$

where

$$\begin{aligned} \mathcal{Y} &= Y - \ln \lambda - \ln \chi \\ \mathfrak{B} &= B - q \ln \chi \end{aligned} \quad \left. \right\} \quad (23)$$

$$\begin{aligned} \frac{s}{(v_{Tn} v_{Tp})^{1/2}} \\ = \frac{1}{2}(\lambda + \lambda^{-1})(\sigma_n \sigma_p)^{1/2} n_i \mathcal{L} 2\pi A \operatorname{sh} q\mathcal{Y} \operatorname{ch} \mathfrak{B} \operatorname{cosec} \pi q \operatorname{cosech} \mathcal{Y} \end{aligned} \quad (24)$$

Comparing (21) with (15), we see that the delta-function approximation is in error to the extent that it replaces $\pi q \operatorname{cosec} \pi q$ by 1. With the value of q found experimentally, this is not too bad; we can now, however, by fitting the right-hand side of (21) to the experimental facts, (13) and (14), obtain exact solutions for $N(\nu)$:

22.6 ohm-cm n-type

$$N(\nu) = 3.6 \operatorname{ch}(0.36\nu + 0.8) \quad (\text{for } \nu < 4) \quad (25)$$

8.1 ohm-cm p-type

$$N(\nu) = 8.3 \operatorname{ch}(0.31\nu + 0.5) \quad (\text{for } \nu < 4) \quad (25)$$

The question arises as to whether this solution for the distribution is unique. We have already pointed out that the mathematical methods fail if the distribution is discontinuous. It seems that (25) represents the only solution that is slowly-varying, in the sense used in the previous section; its correctness could presumably be checked by carrying out experiments at different temperatures. For $\nu > 4$, the above expressions

do not fit the observed facts, because, for $Y - \ln \lambda < -4$, the charge in fast states is found to change more rapidly than is given by the empirical expressions in (13) and (14). The behaviour in this region is perhaps indicative of the existence of a discrete trapping level just beyond the range of ν which can be explored by our techniques. The observations (see Fig. 6 of preceding paper³) can be described by postulating, in addition to the continuous distribution of states given above, a level of density about 10^{11} cm^{-2} , situated at $\nu = 6$, or a higher density still further from the center of the gap. Statz et al.,¹³ using the "channel" techniques, which are valuable for exploring the more remote parts of the gap, have proposed a level of density $\sim 10^{11} \text{ cm}^{-2}$, situated at about 0.14 volts below the center of the gap ($\nu = 5.5$): this is not in disagreement with the foregoing.

In order to compare (22) with the experimental data derived from the surface photo-voltage, it is necessary to choose a value for χ . Fig. 2 shows the comparison with the results presented in the preceding paper. On the vertical axis, the values of $(\partial \Sigma_s / \partial \delta) / (\partial \Sigma_s / \partial Y)$ plotted have been divided by $(\lambda + \lambda^{-1})$, in order to show the n and p-type results on the same scale. (Note that the limiting values of this quantity should be $\lambda / (\lambda + \lambda^{-1})$ and $-\lambda^{-1} / (\lambda + \lambda^{-1})$, so that the vertical distance between the limiting values should be 1, independent of λ). The theoretical curves have been drawn with the value $\ln \chi = 2.5$, in order to give best fit between theory and experiment at the points at which the ordinate changes sign. (It may be seen from the form of (22) that, with the actual value of the other parameters, the main effect of adopting a different value of $\ln \chi$ would be to shift the theoretical curve horizontally, while a change of λ shifts it *vertically* without in either case greatly modifying its shape). The fit between theory and experiment is not quite as good as could be expected, even taking into account the rather low accuracy of the measurements. The variation of $(\partial \Sigma_s / \partial \delta) / (\partial \Sigma_s / \partial Y)$ with Y found experimentally seems to be rather slower than the theory would lead one to expect. The main points to make are: (i) the difference in Y between the zeros for the two samples (5.4 ± 1) is about what it should be (4.8) on the assumption that $\ln \chi$ is the same for both samples and of the order of unity; and (ii) paying attention mainly to the zeros, the estimate $\ln \chi = 2.5$ is likely to be good to ± 1 .

Now let us consider the surface recombination velocity. Here we are on somewhat shakier ground, in that, in deriving (24), we have had to assume not only that χ is independent of ν , but $(\sigma_n \sigma_p)^{1/2}$ also. First we note from (24) that the maximum value of s should occur at $Y - \ln \lambda = \ln \chi$. Comparing with the experimental results given in the preceding paper,

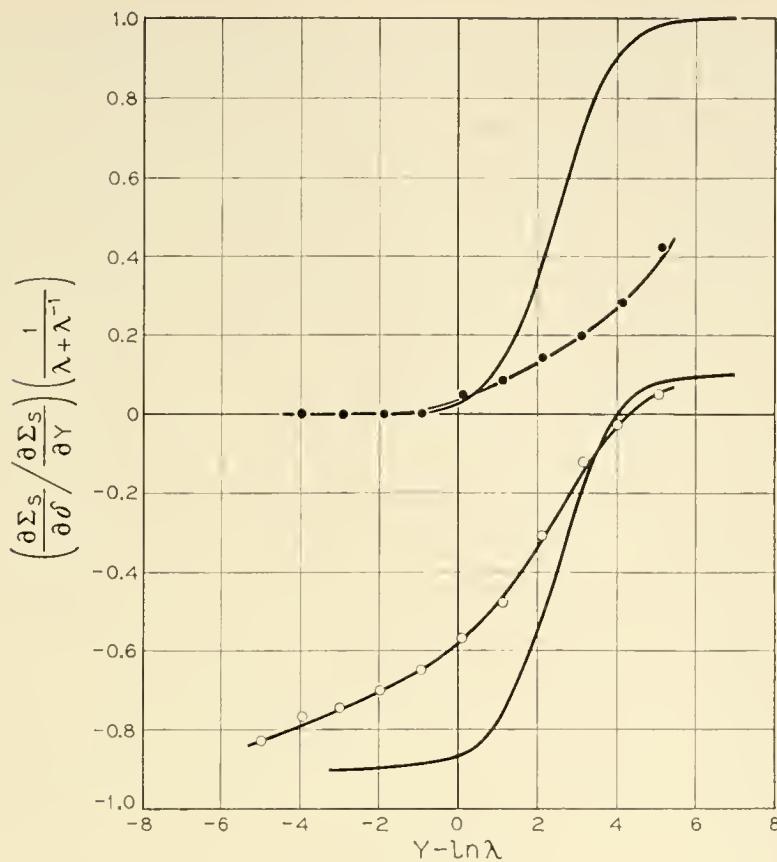


Fig. 2 — Experiment and theory for

$$\left[\left(\frac{\partial \Sigma_s}{\partial \delta} \right) / \left(\frac{\partial \Sigma_s}{\partial Y} \right) \right] / \frac{1}{\lambda + \lambda^{-1}}$$

Solid lines theory; circles and dots, with smooth curves through the points, represent experimental results for *n* and *p*-type samples, respectively.

we see maxima at $Y - \ln \lambda = 2.0$ for the *p*-type sample, and 3.5 for the *n*-type sample. Both these values are within the limits to $\ln \chi$ given in the previous paragraph, thus confirming the estimate made there. Fig. 3 shows a comparison between the experimental results and (24). The graph has been fitted horizontally, by setting $\ln \chi = 2.5$, as found above; vertically, to agree with the mean value at that point. The agreement with experiment is reasonable, although again, just as in Fig. 2, the experimental variation of *s* with $(Y - \ln \lambda)$ is rather slower than one would expect.

The fact that the experimental values, both of surface photo-voltage and of surface recombination velocity, vary more slowly than expected, is susceptible of a number of interpretations: (i) The deduced distribution of fast states might be wrong. However, the most likely alternative distributions — isolated levels, or a completely uniform distribution —

give (in at least some ranges of Y) a more rapid instead of a smoother variation of these quantities so long as the surface is homogeneous. (ii) The estimates of the changes in Y might be too large. It is unlikely that our calibration is sufficiently in error, and other workers have obtained results comparable to ours. The only possibility would be that the mobility of carriers near the surface is larger (instead of smaller, as found by Schrieffler) than inside — which seems quite out of the question. (iii) The ratio of capture cross-sections varies with ν . This, however, would only be in the right direction if one were to assume that the ratio χ increases with the height of the level in the gap — i.e., that the high states behave like acceptors, and the low ones like donors. While not quite impossible, this is an unlikely result. (iv) The surface is patchy. It is probable that a range of variation of two to four times (kT/e) in surface potential would be sufficient to account for the observed slow variation of surface photo-voltage and recombination velocity with mean surface potential. We have refrained from detailed calculations of patch effects, on the grounds that, without detailed knowledge of the magnitude and distribution of the patches, it would be possible to construct a model that could indeed fit the facts, but one would have little confidence in the result. The possibility of patches warns us to view with caution the exact distribution function deduced for the fast states. It would still be conceivable, for example, that one has but two discrete states, as originally proposed by Brattain and Bardeen,² and that the apparent existence of a band of states in the middle of the gap arises from the fact that there are always some parts of the surface at which the Fermi level is close to one or other of these states. Fortunately the conclusions as to the cross-sections are not too sensitive to the exact distribution function assumed.

Using the mean of the two coefficients in (25), substituting $n_i = 2.5 \times 10^{13} \text{ cm}^{-3}$, $\mathfrak{L} = 1.4 \times 10^{-4} \text{ cm}$, $(v_{Tn}v_{Tp})^{1/2} = 1.0 \times 10^7 \text{ cm/sec}$, in (24), and using the experimental result (see Fig. 3) that $s_{\max}/(\lambda + \lambda^{-1}) = 1.2 \times 10^2 \text{ cm/sec}$, one obtains $(\sigma_p\sigma_n)^{1/2} = 5 \times 10^{-16} \text{ cm}^2$. Now setting $(\sigma_p/\sigma_n) = \chi^2 \sim e^5 \sim 150$, one gets for the separate cross-sections:

$$\sigma_p = 6 \times 10^{-15} \text{ cm}^2$$

$$\sigma_n = 4 \times 10^{-17} \text{ cm}^2$$

These values appear to be eminently reasonable. Burton et al.¹² who studied recombination through body centres associated with nickel and copper in germanium, found $\sigma_p > 4 \times 10^{-15} \text{ cm}^2$, $\sigma_n = 8 \times 10^{-17} \text{ cm}^2$ for nickel, and $\sigma_p = 1 \times 10^{-16}$, $\sigma_n = 1 \times 10^{-17}$ for copper. The fact that

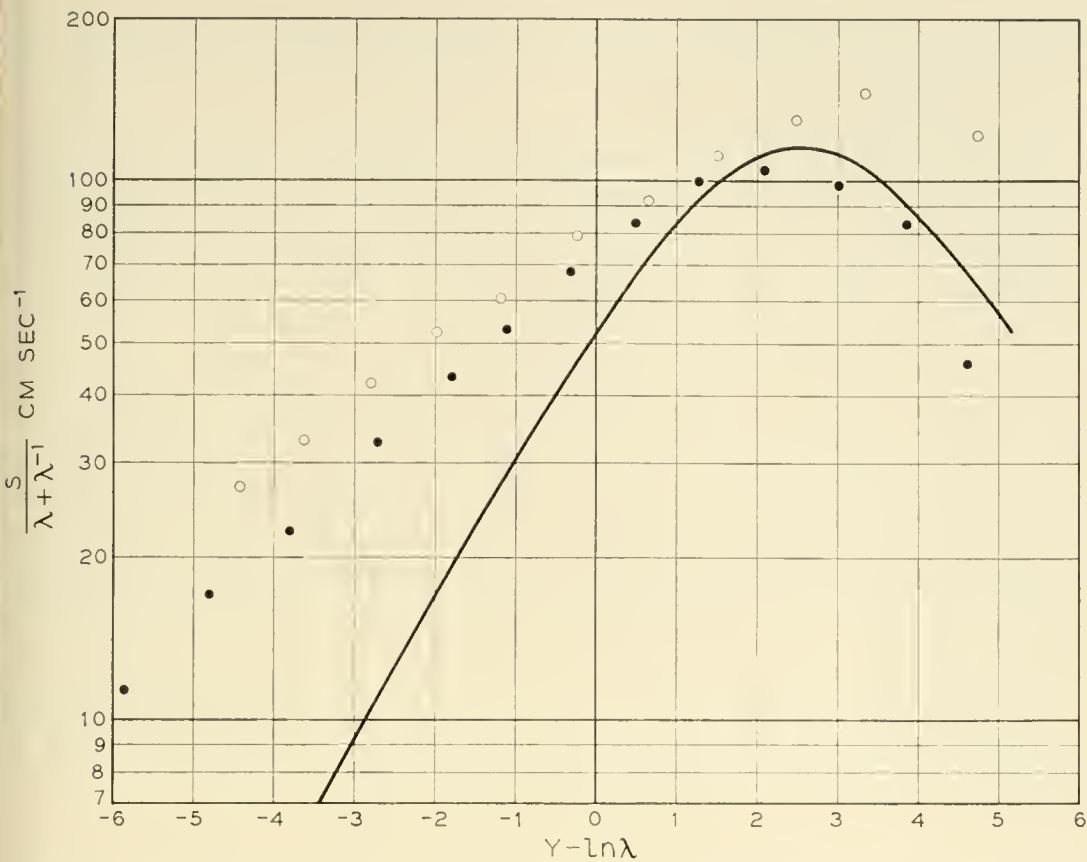


Fig. 3—Experiment and theory for surface recombination. Solid curve theory circles and dots for *n* and *p*-type samples, respectively.

our estimates for σ_p and σ_n appear to be of the expected order of magnitudes lends strong support to the view that identifies the traps appearing in the field-effect and surface photo-voltage experiments with those responsible for surface recombination.

The result that $(\sigma_p/\sigma_n) = 150$ is good evidence that the fast states are acceptor-like. This statement must be restricted to the range $|\nu| < 4$; the states that are outside this range might be of either type. Also one might allow a rather small fraction of the states near the middle to be donor-type, without serious trouble; but the experimental results compel one to believe that most of the fast states within 0.1 volts or so of the centre of the gap are acceptor-like.

V. TRAPPING KINETICS

The foregoing considerations have concerned the steady-state solution to the surface trapping problem. If the experimental constraints are changed sufficiently rapidly, however, there may be effects arising from the finite time required for the charge in surface states to adapt itself

to the new conditions.¹⁴ This section will concern itself with the trapping time constants (which are not directly related to the rate of recombination of minority carriers).

One case of trapping kinetics has been discussed by Haynes and Hornbeck.⁹ A general treatment of surface trapping kinetics is necessarily quite involved, and will be taken up in a future paper. Here we shall restrict ourselves to giving an elementary argument relating to the high-frequency field effect experiment of Montgomery.¹⁵ To simplify the discussion, we assume that the surface in question is of the "super" type; i.e., the surface excess of the bulk majority carrier is large and positive. At time $t = 0$, a large field is suddenly applied normal to the surface; the induced charge appears initially as a change in the surface excess of the bulk majority carrier; as time elapses, charge transfer between the space-charge region and the fast states takes place, until equilibrium with the fast states has been re-established. What time constant characterizes this process?

Take electrons as the majority carrier. Then the flow of electrons into the fast states must equal the rate of decrease of the surface excess of electrons. For a single level one may write:

$$\begin{aligned} U_n &= N_t v_{Tn} \sigma_n [(1 - f_t) n_s - f_t n_1] \\ &= -\dot{\Gamma}_n \end{aligned} \quad (26)$$

For a continuous distribution of levels, one can say that only those levels within a few times (kT/e) of the Fermi level at the surface will be effective, so that one may regard the distribution as being equivalent to a single state with $n_1 = n_i \exp(Y - \ln \lambda)$, which will be about half full. We assume further that the density of fast states is sufficient for the changes in Γ_n to be large in comparison with those in f_t , as is reasonable, having regard to the relative magnitudes of the measured values of $(\partial \Sigma_s / \partial Y)_\delta$ found in the present research, and of $(\partial \Gamma_p / \partial Y)_\delta$ and $(\partial \Gamma_n / \partial Y)_\delta$. Thus f_t may be treated as a constant in equation (26). Further, we may set $n_s = 4\Gamma_n^2 / n_i \mathcal{E}^2$, as may be proved from considerations on the space-charge region.¹⁶ Solving (26) with these conditions, one finds, for the transient change in Γ_n between the initial and the quasi-equilibrium state:

$$\Delta \Gamma_n \propto \left(1 - th \frac{t}{\tau} \right) \quad (27)$$

where

$$\tau = \lambda e^{-Y} \mathcal{E} / [N_t v_{Tn} \sigma_n \sqrt{2} \sqrt{f_t(1 - f_t)}]$$

To clarify the order of magnitude of time constant involved, let us substitute $\mathfrak{L} \sim 10^{-4}$ cm, $N_t \sim 10^{11}$ cm⁻², $v_{Tn} \sim 10^7$ cm/sec., $\sigma_n \sim 10^{-15}$ cm², $f_t \sim 0.5$, $\lambda e^{-Y} \sim 1$. This gives $\tau \sim 10^{-7}$ sec, which suggests that one would be unlikely to run into trapping time effects in the field-effect experiment at frequencies less than 10 Mcyc/sec. This conclusion is consonant with the findings of Montgomery.¹⁵

APPENDIX 1

EVALUATION OF THE INTEGRALS IN SECTION 4

The integrals occurring in Section 4, giving the experimentally accessible quantities $(\partial \bar{\Sigma}_s / \partial Y)$, $(\partial \bar{\Sigma}_s / \partial \delta)$ and s in terms of the surface trap distribution and cross-sections, are conveniently evaluated by contour integration. In view of the general applicability of this method in dealing with integrals of the sort that arise from such a distribution of traps, we include here a short note on the procedure used. The integrals needed are:

$$\begin{aligned} I_1 &= \int_{-\infty}^{+\infty} ch(cx + g) \operatorname{sech}^2 x dx \\ I_2 &= \int_{-\infty}^{+\infty} th(x + b) ch(cx + g) \operatorname{sech}^2 x dx \\ I_3 &= \int_{-\infty}^{+\infty} \frac{ch(cx + g)}{chx + chk} dx \end{aligned}$$

To evaluate I_1 , we evaluate $\int ch(cz + g) \operatorname{sech}^2 z dz$ around the contour shown in Fig. 4. The contributions from the parts $z = \pm R$ vanish in the limit $R \rightarrow \infty$, so that the integral has the value:

$$(1 - \cos c\pi) \int_{-\infty}^{+\infty} ch(cx + g) \operatorname{sech}^2 x dx - i \sin c\pi \int_{-\infty}^{+\infty} sh(cx + g) \operatorname{sech}^2 x dx$$

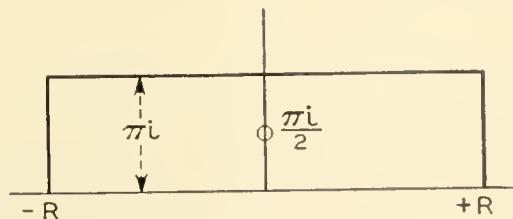


Fig. 4 — Evaluation of I_1 .

The integrand has one pole within the contour, at $x = \frac{1}{2}i\pi$, at which the residue is $-c(\cos \frac{1}{2}c\pi \operatorname{sh} g + i \sin \frac{1}{2}c\pi \operatorname{ch} g)$. Multiplying by $2\pi i$ and equating the real part to that in the above expression, one obtains:

$$I_1 = \pi c \operatorname{cosec} \frac{1}{2}c\pi \operatorname{ch} g$$

The same contour is used in evaluating I_2 ; there are now poles at $z = \frac{1}{2}i\pi$ and at $z = \frac{1}{2}i\pi - b$, and one obtains:

$$I_2 = \pi c \coth b \operatorname{ch} g \operatorname{cosec} \frac{1}{2}c\pi$$

$$- 2\pi \operatorname{cosec} \frac{1}{2}c\pi \operatorname{cosech}^2 b \operatorname{sh} \frac{1}{2}bc \operatorname{ch} (\frac{1}{2}bc - g)$$

To evaluate I_3 , one integrates $\int [ch(cz + g)/(chz + chk)] dz$ around the contour shown in Fig. 5. There are poles at $i\pi \pm k$. Proceeding as before, one finds:

$$I_3 = 2\pi \operatorname{sh} ck \operatorname{ch} g \operatorname{cosec} \pi c \operatorname{cosech} k$$

APPENDIX 2

LIMITATION OF SURFACE RECOMBINATION ARISING FROM THE SPACE-CHARGE BARRIER

The question of the resistance to flow of carriers to the surface arising from the change in potential across the space-charge layer has been discussed by Brattain and Bardeen.² Here we shall recalculate this effect by a better method, which again shows that, within the range of surface potential studied, the effect of this resistance on the surface recombination velocity is for etched surfaces quite negligible.

Let I_p and I_n be the hole and electron (particle) currents towards the surface, and let x be the distance in a direction perpendicular to the surface, measuring x positive outwards. Then the gradient of the quasi-Fermi levels φ_p and φ_n at any point is given by:

$$\nabla \varphi_{\frac{n}{p}} = \mp (I_p/\mu_p)/\binom{p}{n} \quad (1)$$

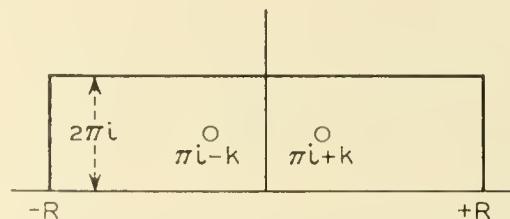


Fig. 5 — Evaluation of I_3 .

Then the total *additional* change in φ_p and φ_n across the space-charge region, arising from the departure in uniformity in the carrier densities p and n , is:

$$\begin{aligned}\Delta\varphi_p &= -\frac{I_p}{\mu_p} \int \left(\frac{1}{p} - \frac{1}{p_0} \right) dx \\ \Delta\varphi_n &= \frac{I_n}{\mu_n} \int \left(\frac{1}{n} - \frac{1}{n_0} \right) dx\end{aligned}\quad (2)$$

Suppose now that the true surface recombination rate is infinite, so that the quasi-Fermi levels must coincide at the surface, and:

$$\varphi_p + \Delta\varphi_p = \varphi_n + \Delta\varphi_n \quad (3)$$

These equations, together with the known space-charge equations,¹⁶ complete the problem. Notice first, from (2), that $\Delta\varphi_p$ will be large only if there is a region in which p is small ($Y \gg 1$), while $\Delta\varphi_n$ is large only when, in some region, n is small ($Y \ll -1$). Introducing the quantity δ , approximating for δ small, equating I_p and I_n and setting the result equal to $s n_i \delta$, one finds:

$$Y \ll -1$$

$$s \rightarrow (D_n/\mathcal{L})(\lambda^{1/2} + \lambda^{-3/2})e^{\frac{1}{2}Y} \quad (4)$$

$$Y \gg 1$$

$$s \rightarrow (D_p/\mathcal{L})(\lambda^{-1/2} + \lambda^{3/2})e^{-\frac{1}{2}Y}$$

The coefficients (D_n/\mathcal{L}) and (D_p/\mathcal{L}) are of the order of 4×10^5 cm/sec. The most extreme case encountered in our work is that occurring at the ozone extreme for the n-type sample ($\lambda = 0.34$, $Y = -6$), for which the surface recombination velocity, if limited by space-charge resistance alone, would be about one-quarter of this (10^5 cm/sec). The fact that the observed surface recombination velocity is lower than that by more than two orders of magnitude shows that space-charge resistance is not a limiting factor in the present experiments. Equations 4 might well hold on a sand-blasted surface, however, where the trap density is much higher.

REFERENCES

1. W. L. Brown, Surface Potential and Surface Charge Distribution from Semiconductor Field Effect Measurements, Phys. Rev. **98**, p. 1565, June 1, 1955.
2. W. H. Brattain and J. Bardeen, Surface Properties of Germanium, B.S.T.J., **32**, pp. 1-41, Jan., 1953.
3. W. H. Brattain and C. G. B. Garrett, page 1019 of this issue.

4. D. T. Stevenson and R. J. Keyes, Measurements of Surface Recombination Velocity at Germanium Surfaces, *Physica*, **20**, pp. 1041-1046, Nov. 1954.
5. R. N. Hall, Electron-Hole Recombination in Germanium, *Phys. Rev.*, **87**, p. 387, July 15, 1952.
6. W. Shockley and W. T. Read, Jr., Statistics of the Recombination of Holes and Electrons, *Phys. Rev.*, **87**, pp. 835-842, Sept. 1, 1952.
7. T. M. Buck and F. S. McKim, Depth of Surface Damage Due to Abrasion on Germanium, *J. Elec. Chem. Soc.*, in press.
8. H. H. Madden and H. E. Farnsworth, Effects of Ion Bombardment Cleaning and of Oxygen Adsorption on Life Time in Germanium, *Bull. Am. Phys. Soc.*, II, **1**, p. 53, Jan., 1956.
9. J. A. Hornbeck and J. R. Haynes, Trapping of Minority Carriers in Silicon. I. P-Type Silicon. II. N-Type Silicon, *Phys. Rev.*, **97**, pp. 311-321, Jan. 15, 1955, and **100**, pp. 606-615, Oct. 15, 1955.
10. Ig. Tamm, Über eine mögliche Art der Elektronenbindung an Kristalloberflächen, *Phy. Zeits. Sowj.*, **1**, pp. 733-746, June, 1932.
11. H. C. Montgomery and W. L. Brown, Field-Induced Conductivity Changes in Germanium, *Phys. Rev.*, **103**, Aug. 15, 1956.
12. J. A. Burton, G. W. Hull, F. J. Morin and J. C. Severiens, Effects of Nickel and Copper Impurities on the Recombination of Holes and Electrons in Germanium, *J. Phys. Chem.*, **57**, pp. 853-859, Nov. 1953.
13. H. Statz, G. A. deMars, L. Davis, Jr., and H. Adams, Jr., Surface States on Silicon and Germanium Surfaces, *Phys. Rev.*, **101**, pp. 1272-1281, Feb. 15, 1956.
14. C. G. B. Garrett, The Present Status of Fundamental Studies of Semiconductor Surfaces in Relation to Semiconductor Devices, *Proc. West Coast Electronics Components Conf.* Los Angeles, pp. 49-51, June, 1955.
15. H. C. Montgomery and B. A. McLeod, Field Effect in Germanium at High Frequencies, *Bull. Am. Phys. Soc.*, II, **1**, p. 53, Jan., 1956.
16. C. G. B. Garrett and W. H. Brattain, Physical Theory of Semiconductor Surfaces, *Phys. Rev.*, **99**, pp. 376-387, July 15, 1955.
17. P. A. Wolff, Private communication.

Transistorized Binary Pulse Regenerator

By L. R. WRATHALL

(Manuscript received March 14, 1956)

A simple transistorized device has been constructed for amplifying and regenerating binary code signals as they are transmitted over substantial lengths of transmission line. By the use of simple circuitry, means are provided whereby the distortion in the output of one repeater due to low frequency cutoff is compensated in the next repeater. Furthermore, the repeater is effectively and simply timed from its own regenerated output. A brief discussion of the theory of the circuit is presented along with measured results and oscillograms showing its performance. The effects of extraneous interference on the production of errors in such a repeater are reported. These results are in substantial agreement with theory.

1. INTRODUCTION

Long distance communication using digital transmission is not new but was used by man in his earliest communication system. In fact, his first successful electrical system, the telegraph, made use of binary pulse codes. It was not until the invention of the telephone that the emphasis was shifted from the digital to carrier and voice systems. During recent years the development of new electronic devices and techniques have brought digital transmission into the picture again, and it now seems possible to use it not only for telephony but for television as well. Future systems will probably make use of the binary code, this choice being dictated by circuit simplicity and performance.

The fundamental requirement for perfect binary transmission is to be able to detect the presence or absence of a pulse in each of a regular set of discrete time intervals. From this requirement the principal advantages of such a system may be tabulated. First, a pulse can be recognized in the presence of large amounts of interference. Second, when a pulse is recognized it can be faithfully regenerated, suppressing the effect of the interfering noise to any desired degree. Third, simple high-efficiency non-linear devices such as multivibrators or blocking oscillators can be used to regenerate the pulses. The great disadvantage,

common to all pulse systems is the large bandwidth required for transmission.

On wire lines this large transmission band will create a number of problems. The phase-loss variations, crosstalk and temperature effects will be greatly increased over the transmission band as compared to that of the more conventional systems. It can be shown however that if the repeater spans are made sufficiently short these problems will largely disappear. Only rough equalization will be needed, crosstalk and temperature effects become negligible. Furthermore the repeater power requirements will be small and the circuitry comparatively simple, since only partial regeneration will be required. The problem remains to build a regenerative repeater so simple that it will be economically sound to use on short spans of line. The development of the transistor with its small size and low power requirements has made such a repeater feasible.

1.1 Pulse Distortion Caused by Low Frequency Cutoff

Since the frequency spectrum of a binary pulse train will extend down to and include dc, the ideal repeater should be able to handle the complete frequency band to avoid signal distortion. This would preclude the use of coupling transformers and condensers which attenuate the low frequencies and remove the dc. Practical considerations however dictate the use of these elements which means that the repeater will have a low frequency cutoff. The distortion of a binary pulse train produced by low frequency cutoff presents one of the most vexing problems the designer of a regenerative repeater must cope with. It produces what is probably the most potent source of intersymbol interference found in an average binary pulse communication system. This interference consists of a transient response whose effect may be appreciable far beyond the end of the pulse itself.

When a train of ideal flat top pulses with infinitely steep sides is applied to a load through a condenser or a transformer, the transient response persisting beyond the end of the pulse is an exponential and may be expressed as

$$T = kP_0 e^{-bt} \quad (1)$$

The time t , is measured from the end of the pulse and the damping coefficient b is a function of the low frequency cutoff.* P_0 is the amplitude

* The value of b may be approximated by

$$b = 2\pi f_0$$

where f_0 is the frequency in cycles/sec at which the low frequency loss characteristic of the transformer is 6 db above that of the pass band.

of the pulse and k is given as

$$k = 1 - e^{-bt_p}$$

where t_p is the pulse duration. The sum of the transients of a sequence of pulses will shift the zero potential from the base of the pulse toward its average value as shown on Fig. 1(b). This phenomenon has been re-

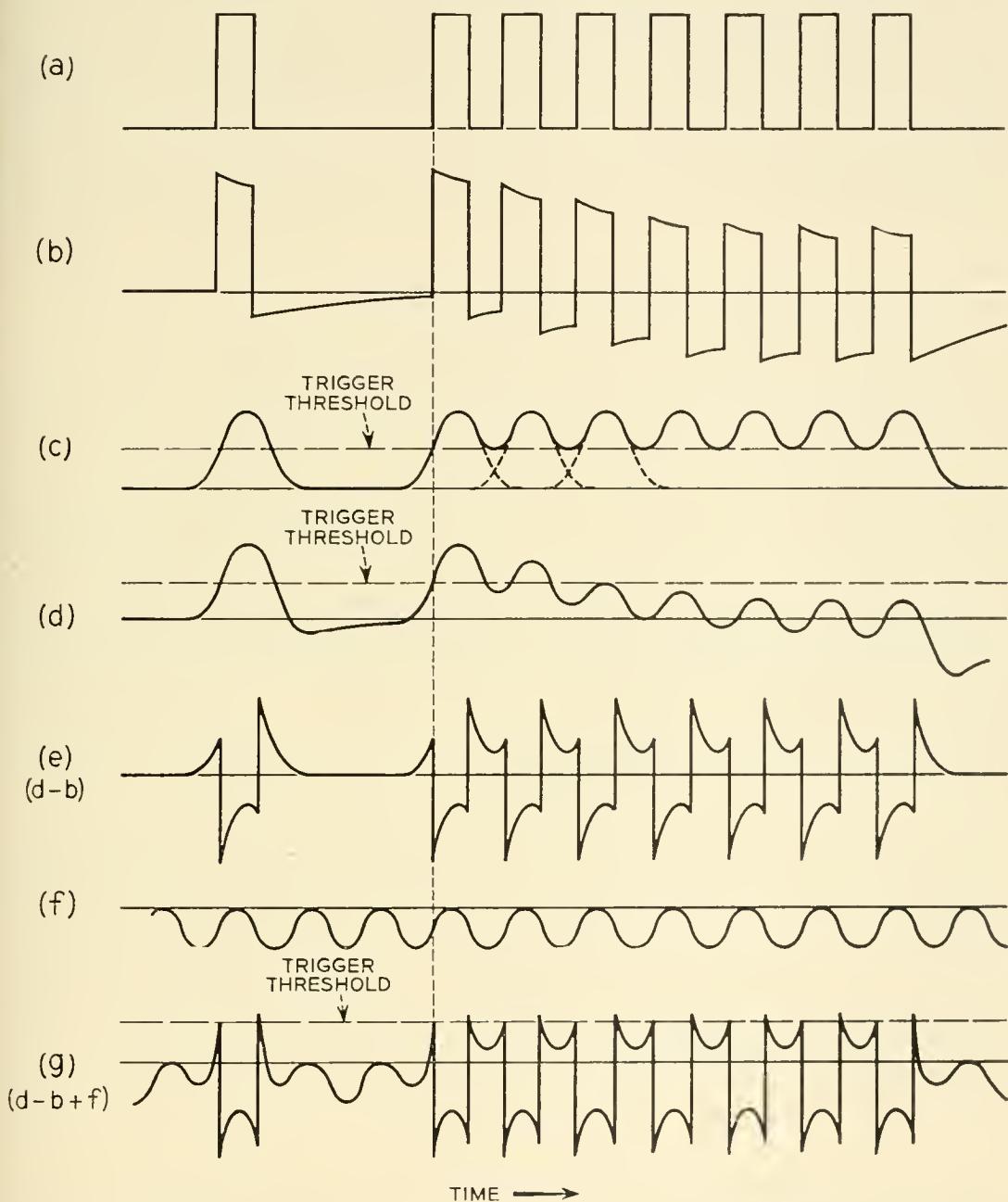


Fig. 1 — (a), a perfectly regenerated pulse train; (b) showing the effect of low-frequency cutoff; (c), showing (a) after passing over equalized line; (d), showing (b) after passing over equalized line; (e), effect of (d) minus (b); (f), inverted pedestal timing wave; (g), composite wave at input to repeater, namely, (d) minus (b) plus (f).

ferred to as "zero wander." In a regenerative repeater the trigger potential is tied to the zero level by a constant bias. Zero wander then will produce a changing bias which reduces the signal to noise margins of the repeater, or in some cases even prevents regeneration. Suppose, for example, a transmission line is equalized so the ideal pulse train shown on Fig. 1(a) will appear as Fig. 1(c) after being transmitted over the line. The individual pulses have widened until the envelope of a sequence of consecutive pulses shows as a ripple with a much smaller amplitude than the individual pulse. If the pulse train distorted by low frequency cutoff shown on Fig. 1(b) is transmitted over this line its output will appear similar to that shown on Fig. 1(d). The portion of the signal where the peak amplitude lies below the trigger threshold will not be regenerated.

1.2 Compensation for Low-Frequency Distortion

In the past many circuits have been devised to prevent zero wander, but none have been completely satisfactory. The repeater described in this paper effectively eliminates zero wander in a string of consecutive repeaters by means of a new and simple method. This may be better understood by referring to Fig. 2. Here are represented two successive repeaters of a transmission system. These repeaters have what appears as a conventional negative feedback loop consisting of a pair of resistors, R . The function performed by this feedback loop bears little if any resemblance to the negative feedback of linear amplifiers and is referred to as "Quantized feedback" in this paper.*

Suppose an isolated pulse of amplitude P_m is regenerated in repeater M and is applied to the line through its output transformer. The low frequency cutoff of this transformer will produce a transient response to the regenerated pulse as given in (1). A spectrum analysis of the transient tail shows that most of its energy occurs in the lower portion of the pass band of the equalized line. Consequently, it will be transmitted over the line to the next repeater with little if any frequency or phase distortion, but will be attenuated by a factor α . This transient at the input of the following repeater may be expressed as

$$T_M = \alpha k_M P_M e^{-bt} \quad (2)$$

where t is again measured from the end of the pulse. Suppose the regeneration of the pulse at the output of repeater N is delayed by time t_1

* A paper by Rajko Tomovich entitled "Quantized Feedback" was published in the I.R.E. Transactions on Circuit Theory. There are some fundamental differences in the meaning of the term, quantized feedback, as used in these papers.

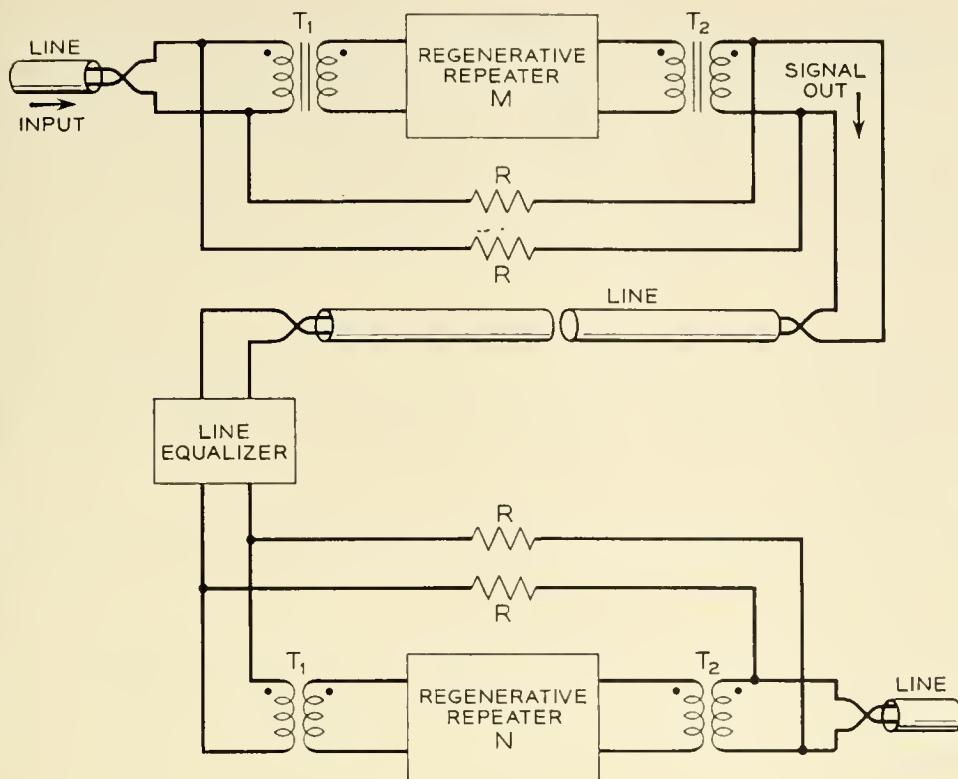


Fig. 2 — Block diagram of a section of equalized line and its terminating repeaters.

compared to the pulse at the input of the repeater. The transient response of the regenerated pulse after passing through its output transformer† will be

$$T_N = k_N P_N e^{-b(t-t_1)} \quad (3)$$

$$= k_N P_N e^{bt_1} e^{-bt} \quad (4)$$

If the transient (4) is attenuated by factor β and added in opposite phase to T_M through the feedback loop at the input of the repeater, their sum is

$$T_M - \beta T_N = \alpha k_M P_M e^{-bt} - \beta k_N P_N e^{bt_1} e^{-bt} \quad (5)$$

$$= e^{-bt} (\alpha k_M P_M - \beta k_N P_N e^{bt_1}) \quad (6)$$

This can be made equal to zero if

$$\alpha k_M P_M = \beta k_N P_N e^{bt_1} \quad (7)$$

which is accomplished by adjusting the value of β which represents the feedback attenuation introduced by resistances R . If the regenerated

† It is assumed that the electrical characteristics of the output transformers of all the repeaters are identical. In this case the damping coefficients will be identical for all the regenerated outputs.

output pulses of M and N are identical, then $P_M = P_N$ and $k_M = k_N$ and eq. (6) becomes

$$T_M - \beta T_N = e^{-bt} k_M P_M (\alpha - \beta e^{bt_1}) \quad (8)$$

This expression can be made equal to zero if

$$\beta = \alpha e^{-bt_1} \quad (9)$$

By this means zero wander produced in one repeater can be eliminated at the input of the next repeater. The low frequency distortion of one repeater corrects for the corresponding distortion produced in the previous repeater.

If the electrical characteristics of all the repeater output transformers are identical it is possible to completely remove the effects of the transient tails due to low frequency cutoff.* It is important however that t_1 should not be so large that the feedback pulse occupies the next timing interval. W. R. Bennett has shown that a similar cancellation of transients can be accomplished for more complicated types of low frequency cutoff characteristics. In this case the transient tails will be the sum of a number of exponentials having different amplitudes and damping coefficients. Here the quantized feedback must be provided by multiple loops, of greater complexity.

It may be disturbing at first to observe the resultant sum of the incoming signal and feedback as shown on Fig. 1(e). It should be noted however that the signal is not changed in any way until the repeater has triggered the regenerated pulse, and at the next time slot the tails have been cancelled, so that when the next pulse arrives it too will begin at the zero axis. Tails may also be produced by high frequency phase-loss characteristics. These however, may be removed by proper equalization.

1.3 Timing In a Regenerative Repeater

The binary regenerative repeater must not only regenerate the shape and amplitude of each individual pulse but it must also keep them in proper time sequence with other signal pulses. To accomplish this a suitable timing wave must be provided. This timing wave may be transmitted over separate pairs of wires or it may be derived from the signal. In the past it has been common to obtain a sine wave of the repetition

* It can be shown that, with reasonable differences in damping coefficients, quantized feedback will greatly reduce intersymbol interference even when considering a single pulse. If the contributions from all the transients of an infinite train of random pulses are summed, the resultant interference is further reduced and can be considered negligible.

frequency by exciting a high Q filter circuit from the received pulse train. Short timing pips generated from this wave are used to time the regenerated output pulses precisely. This procedure is far too involved to be used in a simple repeater. If less precision in timing is acceptable it may be accomplished with a minimum of circuitry by use of a sinusoidal wave derived from the repeater output. This is referred to in this paper as "self timing."

Self timing prohibits the use of short timing pips derived from the regenerator output. In this case most of the timing control would be exercised by the filter circuit and little, if any, by the input signal. The direct use of the sinusoidal output of this filter provides sufficient control by the input signal with only a small penalty due to less precise timing.* Self timing also sets certain requirements on the regenerator. If the timing wave is derived from an independent source it can be added to the signal in such a way as to act as a pedestal, lifting the signal above the trigger level. In such a circuit neither the signal nor the timing wave alone can trigger the regenerator. If the timing wave is derived from the output it is obvious that the signal alone must be able to trigger the regenerator, since the generation of a timing wave depends upon the signal triggering the regenerator. A timing wave derived by filtering the output of a random pattern of binary pulses will also have a varying amplitude which could cause variations in repeater noise margins. It is apparent then that self timing output cannot be used as a pedestal in a regenerator. All these objections can be overcome by the use of "inverted pedestal" timing.

Inverted pedestal timing is produced by tying the peaks of the timing wave having the same polarity as the signal pulses to a fixed level by means of a diode. This is illustrated on Fig. 1(f). The timing wave is added to the signal at the input so the sum of the signal, feedback and timing looks somewhat like the wave on Fig. 1(g). The effect of the inverted pedestal timing is to inhibit triggering except in the time interval near the peaks of the timing wave. This permits the signal to trigger the regenerator without a timing wave, yet allows timing control to be exercised as the amplitude of the timing wave builds up. With sinusoidal timing, noise often causes the regenerator to trigger either early or late, introducing a phase shift in the regenerated output which will be reflected in the timing wave. Since the timing wave is derived from the code pattern by a relatively high Q tuned circuit, the phase distortion of the timing wave from a shift of a single pulse will be small. With a random dis-

* E. D. Sunde, Self-timing Regenerative Repeaters (paper being prepared for publication).

tribution of noise the resultant phase shift of the timing wave will be negligible. If the interference has low frequency components, the phase shift of the timing wave may be appreciable but these are slow and consequently will not seriously effect the performance of the regenerator.

2.0 DESCRIPTION OF REPEATER CIRCUIT

The circuit diagram shown on Fig. 3 will aid in understanding the operation of the repeater. The incoming signal after being transmitted over the equalized line is applied through the input transformer T_1 to the emitter of transistor (1). The function of this transistor is to provide gain to the incoming signal. This amplified signal is applied to the emitter of transistor (2) through the blocking condenser C_2 . The second transistor functions in a single shot blocking oscillator circuit being biased in the "off" condition through the resistance R_2 . When the positive signal exceeds the trigger threshold, a pulse is regenerated by the blocking oscillator. During the pulse period a large emitter current flows through D_1 in the conducting direction. T_2 is the output transformer while transformer T_3 provides the essential positive feedback for the blocking oscillator.

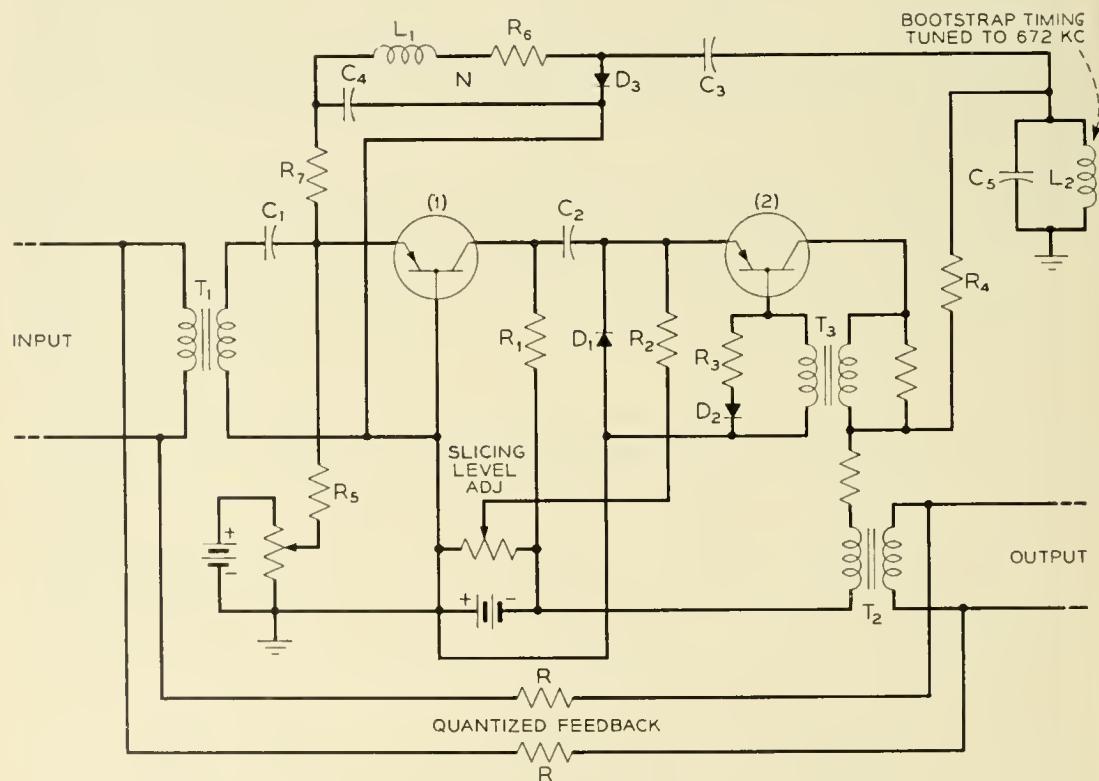


Fig. 3 — Circuit diagram of the regenerative repeater.

2.1 Inhibiting in Blocking Oscillator

The secondary of T_3 is connected between the transistor base and ground with the diode D_2 and resistor R_3 in series across it. The combination of diode and resistance across T_3 serves a very important function, the inhibiting of multiple triggering on a single input pulse. During the interval in which the pulse is regenerated a negative potential is applied between the base and ground. A current I_0 flows through the base of the transistor, the diode D_2 being poled to restrict the flow of current in R_3 . At the end of the pulse the current I_0 in T_3 drops suddenly to a low value. This current change in the inductive winding of T_3 induces a relatively large potential across the base of the blocking oscillator. The impedance of D_2 becomes low and current flows in R_3 and T_3 . The potential across T_3 decays exponentially and with proper circuit values will take the form of a damped cosine wave.

$$E = E_0 e^{-\alpha t} \cos \omega_0 t \quad (10)$$

where t is the time measured from the peak of the pulse. The values of α and ω_0 can be adjusted by varying the inductance the transformer and the capacity and resistance connected across it. E should become substantially zero at or near the next timing interval. The damping coefficient α should be sufficiently large to prevent an appreciable negative excursion of E since this will reduce the effective bias on the repeater and consequently its noise margins. This will be further discussed in the section on the measurements of errors.

2.2 Quantized Feedback

The quantized feedback is provided by coupling the input and output transformers by means of resistances R . The fed back pulse must be in the opposite phase compared to the input signal.

2.3 Timing Wave Circuit

The timing wave is derived by means of the parallel resonant tank circuit L_2C_5 which is tuned to the signal repetition frequency. The regenerated pulses are applied to this network through the relatively large resistance R_4 . The amount of energy added to the network by each pulse as well as the amount dissipated in it is a function of Q . The higher the Q the smaller will be the variations of timing wave amplitude as the average pulse density of the signal train changes. This does not mean that the highest Q will be the most desirable for increased Q means larger,

more expensive coils. Higher Q 's also produce greater variations in impedance and phase with small changes of resonant frequency which require much closer control of inductance and capacity with temperature. In the circuit described here the Q has a value of about 100 and its operation is quite satisfactory. The tank circuit is coupled through the small condenser C_3 to the diode D_3 . This diode ties the positive peaks of the timing wave to ground as is required for inverted pedestal timing. The network N provides the timing delay needed for optimum repeater performance.

2.4 DC Compensation in Timing Wave

The timing wave amplitude from the tank circuit is insufficient to allow it to be applied directly to the emitter of the blocking oscillator. Consequently in the interest of circuit simplicity the signal amplifier is used for the timing wave as well. To avoid the complications introduced by dc coupled circuits when close bias tolerances must be maintained, the amplifier was coupled to the blocking oscillator by condenser C_2 . This presents a problem as to how to neutralize the charge the dc component of the timing wave builds up on C_2 . The means by which this is accomplished can be more easily understood by referring to Fig. 4.

In this figure the time constant of the feedback loop $R_0C_1R_1$, is made large so that substantially equal charges are added to C_1 by each regenerated pulse. In the timing loop this is also nearly true even though noise

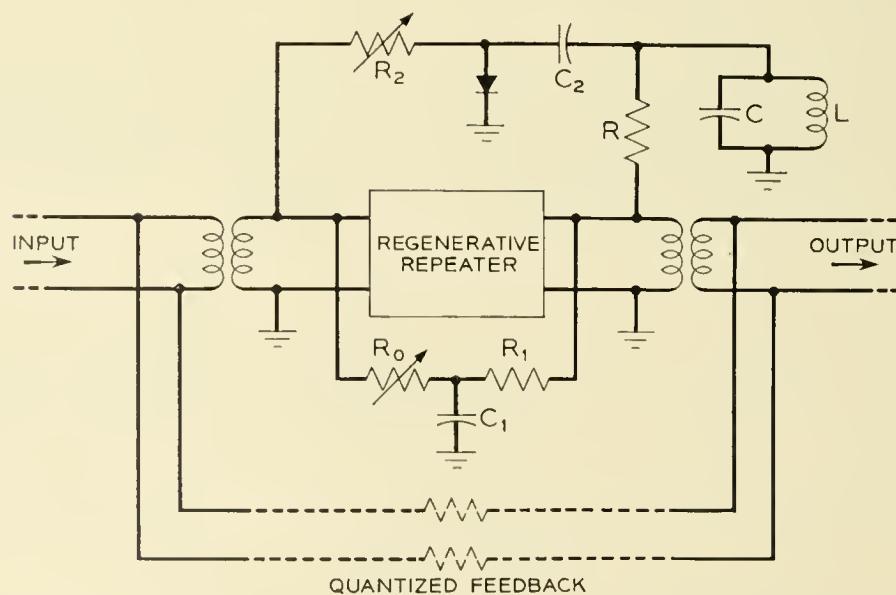


Fig. 4 — Method for maintaining the dc values of timing wave.

may change the phase of individual pulses. The change of amplitude of the sinusoidal timing wave in one pulse period will be

$$\Delta A_T = A_T [1 - e^{-\pi t_m/Q}] \quad (11)$$

where $Q = \omega L/R$ and t_m is the timing interval. In a similar manner the variation of the amplitude of the voltage across C_1 will be

$$\Delta A_c = A_c [1 - e^{-t_m/R_1 C_1}] \quad (12)$$

If now R_1 and C_1 are adjusted until

$$\frac{\pi}{Q} = \frac{1}{R_1 C_1} \quad (13)$$

and R_0 varied until the amplitude A_c is equal to the average value of A_T , the charge on the interstage coupling condenser should be effectively neutralized at all times. Since both loops are made up of passive elements with common inputs and outputs a single adjustment should suffice even though the pulse amplitude, width, or signal pulse density may vary.

In the repeater circuit shown on Fig. 3 this neutralizing principle is used but is more difficult to see. When a pulse is regenerated, a large emitter current flows in D_1 , which produces a sharp negative voltage spike. This voltage adds a charge to C_2 which tends to neutralize the one the timing wave adds to it. The time constant of C_2 and its associated circuit may be made to equal the decrement of the tank circuit and the two amplitudes made equal by adjusting the level of the timing wave. By this means effective dc transmission of the timing wave is achieved through capacity coupling.

2.5 Line Equalization

The line equalizer is not essentially a part of the repeater itself. It is however so intimately connected with the repeater it is logical that they be considered together. One of the important equalizer requirements is simplicity, another, that the impedance seen from the repeater input shall be substantially constant over a relatively large frequency range. This latter requirement comes from the need of transmitting the feed-back pulse around the feedback loop to the emitter of the first transistor without too much distortion. The equalizer is not used to equalize the low frequency losses of transformers but only the frequency characteristic of the line. The equalization must be such that the individual pulses are allowed to widen but not enough to cause inter-symbol interference.

A gaussian shaped pulse at the output of the line is one of the most economical to use and can have a maximum span of one timing interval at its base. However, in this case the envelope of a long consecutive sequence of such pulses will show substantially no ripple. It can be readily seen that in such a sequence the only timing control exercised by the input upon the timing wave comes from the first pulse. In the interest of better timing and consequently better repeater performance one should be content with narrower pulses at the repeater input. The resulting ripple of the envelope of a consecutive pulse sequence allows each incoming pulse some control over the repeater timing.

3.0 REPEATER PERFORMANCE

To check the performance of the regenerative repeaters a binary code generator was built having a nominal pulse repetition rate of 672 kc producing an eight digit code. Any code combination from the possible 256 can be selected or the code automatically changed at periodic intervals reproducing all possible codes in orderly sequence. Random codes may also be generated by making the absence or presence of a pulse

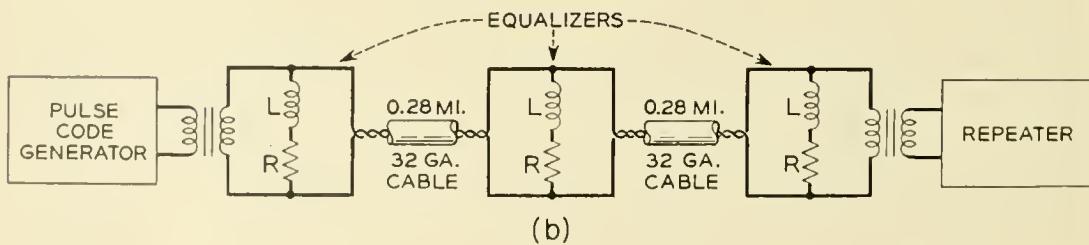
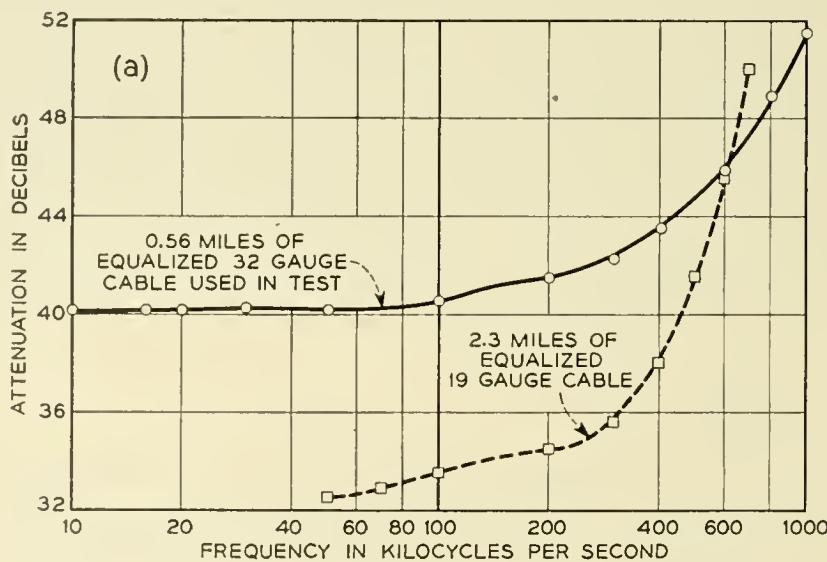


Fig. 5(a) — Equalized characteristics of 19 and 32 gauge line.
Fig. 5(b) — Block diagram of equalizer for 32 gauge line.

in any time slot dependent on the polarity of random noise. The output of the code generator was made substantially the same as the outputs of the repeaters both in shape and amplitude. Two types of transmission line were used, a line from a 51 pair 19 gauge exchange cable and a pair from a 32 gauge experimental cable. The nominal lengths of cable between repeaters was 2.3 miles for the 19 gauge and 0.56 miles for the 32 gauge cable. Fig. 5(a) shows the equalized characteristics for both these lines. The important differences between the two is a greater flat loss with a better high frequency characteristic for the 32 gauge cable. This was advantageous in the study of error production and consequently, the error measurements were all made with this cable. The 19 gauge characteristic represents about the maximum high frequency loss that can be tolerated by these regenerative repeaters.

The performance of the regenerative repeater circuit can best be shown by photographs taken from a cathode ray oscilloscope representation. Plate I shows the effect of the 19 gauge line equalizer. The output pulse (1) transmitted over the unequalized line has become very broad, extending over several timing intervals, which are indicated by small pips along the trace. The addition of the equalizer reduces the width of the received pulse (2) until it is somewhat narrower than the normal pulse interval of the code. Plate II shows a series of photographs taken of the input and output of a repeater with or without interference added at the repeater input.* A signal code at the input of the repeater is shown on (a) and its regenerated output on (b). A sinusoidal interference having a frequency of about 100-kc pictured on (c) is added to the signal as represented on (d). The regenerated output of input (d) is shown on (e). From these it can be seen that while interference does not change the pulse shape or size, it does produce a phase modulation.

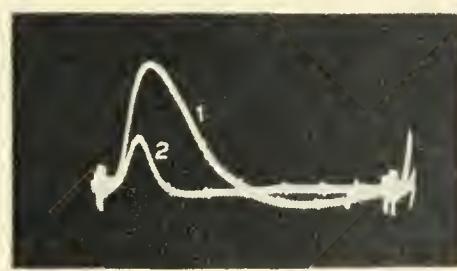


Plate I — Single pulse at output of 2.3 miles of 19 gauge cable. 1 — Unequalized. 2 — equalized.

* The input signal of this and some of the following photographs was taken with the repeater in an inoperative condition. This was done in order to avoid the resulting complexity that results when both the quantized feedback and timing wave are added to the combinations of incoming signal and interference.

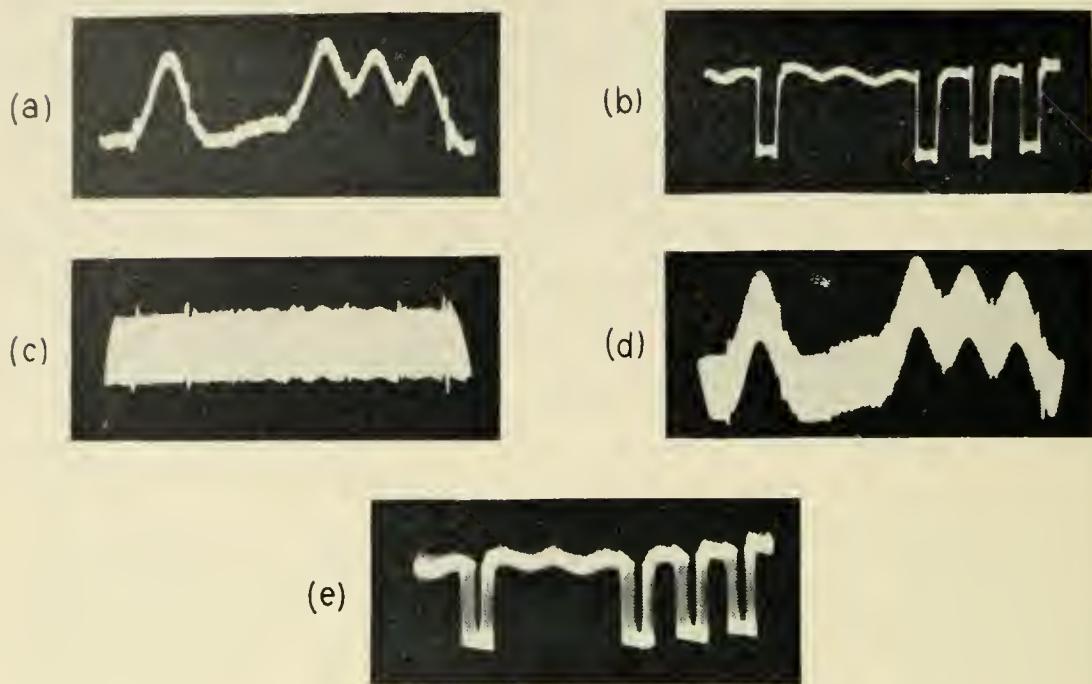


Plate II — (a), repeater input, no interference; (b), regenerated output with input (a); (c), sinusoidal interference; (d), repeater input, signal (a) plus sinusoidal interference (c); (e), regenerated output of (d).

3.1 Performance of Repeaters in Tandem

Plate III shows the results when certain phase modulated codes are transmitted through a series of repeaters in tandem. The regenerated signal from each successive repeater is transmitted over 2.3 miles of equalized 19 gauge line. One code which has two out of a possible eight pulses present has most of the phase jitter removed after passing through the three additional repeaters. The other fixed code shown contains four out of a possible eight pulses. The jitter is removed much more rapidly with this code, after passing through two repeaters it is regenerated almost perfectly. The reason for the difference in the regeneration of the two codes is variations in the amplitude of the timing wave. In any period of time the energy delivered to the tank circuit is proportional to the number of regenerated pulses in that interval. The amplitude of the timing wave for a fixed code with two pulses of the eight will be half the one produced by the code having four pulses out of eight present. The average number of pulses in a normal PCM signal will be half the maximum possible pulses. The timing wave should then average the same as that produced by the fixed code having four out of a possible eight pulses present. The phase jitter of the random code should be removed as quickly as it was with this fixed code. This is confirmed by

regenerating a noise-dictated random code having the same pulse density expected of a normal PCM signal. The results are shown on Plate III(c). After passing through two repeaters the jitter has been substantially removed as shown by the sharp vertical lines marking the pulses. The thickening of the horizontal lines are produced by transients produced by low frequency cut off distortion. In all these photographs the oscillograph synchronization was obtained from the code generator.

3.2 Possible Effects of Line Temperature Variations

The gain and phase characteristics of a particular wire transmission line is a function not only of its length but of temperature as well. To the first order approximation the effect of an increase in temperature may be considered as caused by an increase in the length of the line. In order to better understand the effect of temperature change on repeater performance the following steps were taken; The repeater was adjusted for optimum performance with 2.3 miles of line between it and the preceding repeater and then the length of the connecting transmission line was decreased by about 25 per cent. It was found that for the same interference on the input of the repeater no difference in the performance of the repeater was observed. Plate IV shows a fixed code signal after it has traversed 2.3 miles of equalized cable. Superimposed

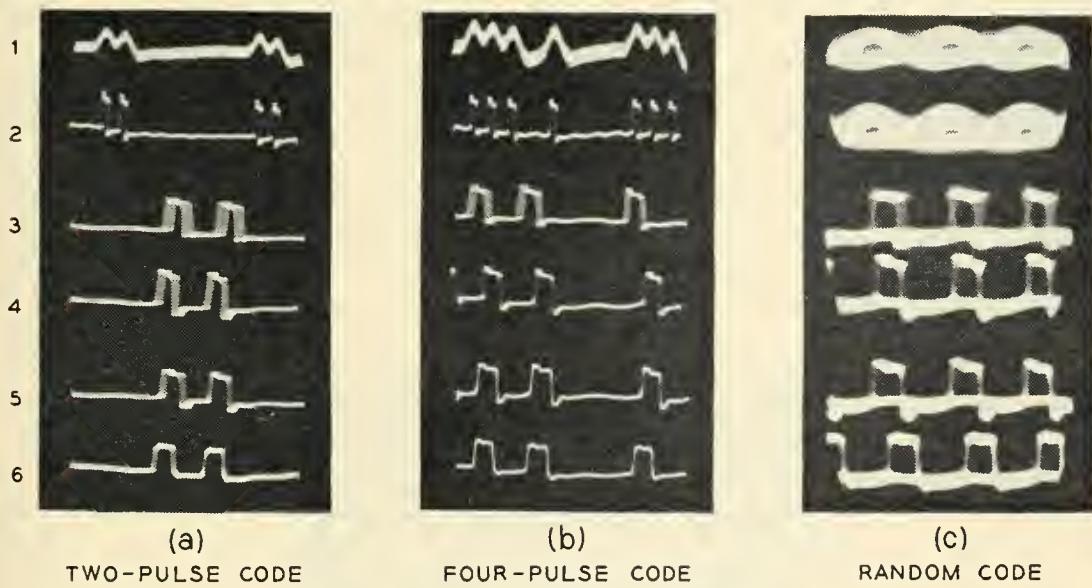


Plate III — (a), set code having 2 pulses out of possible 8; (b), set code having 4 pulses out of possible 8; (c), random code having an average of 4 pulses out of a possible 8.1 (a and b), input signal plus interference; 2 (a and b), regenerated output of 1; 3, expanded section of 2; 4, output of 2nd repeater; 5, output of 3rd repeater; 6, output of 4th repeater. 1(c), input signal alone; 2(c), input signal plus interference.

on this is the same signal after traversing a 1.75 mile length of line and the same equalizer. Shortening the line results in the transmitted pulses having higher peak amplitudes and narrower widths. Faulty high frequency equalization of the shorter lengths produces the short tail following the pulse. It is interesting to observe that the transient tail due to the low frequency cut off has not changed appreciably as the line was shortened. This is to be expected since it can be shown that the energy of the low frequency cut off transient is concentrated in low frequency end of the transmission spectrum. In this region changes in the length of the line, or changes in the primary constants will result in inconsequential changes in attenuation and phase as is shown on Fig. 6. If the quantized feedback is adjusted for the worst condition, i.e., the highest temperature likely to be encountered, it will not need to be changed with lower temperatures.

4.0 ERROR PRODUCTION BY EXTRANEous INTERFERENCE

A knowledge of the performance of a regenerative repeater with various types and amounts of interference added to the input signal is important. Consequently a study of such errors produced in one of these repeaters was undertaken. Two general types of extraneous interference was used in this study. The first is impulse noise, the type which is produced by telephone dials, switches, lightning surges and crosstalk from other pulse systems. The second is sinusoidal noise, the type which come from power line or carrier crosstalk. This interference may affect the regenerated output in a number of ways. It may produce a phase shift or "jitter" in the output; cause a pulse to be omitted; or cause a spurious pulse to be inserted in the signal code. The phase jitter will be largely removed by timing regeneration in subsequent repeaters, but omission and most insertion errors will be carried through the remaining repeaters, causing distortion in the decoded signal.

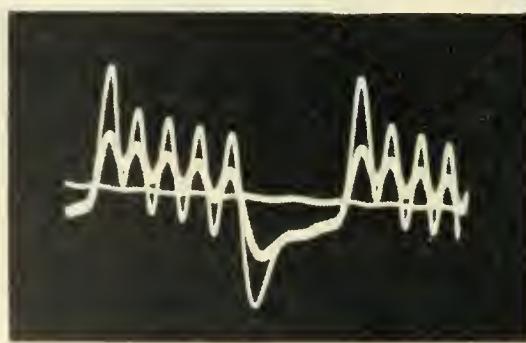


Plate IV — Superimposed picture of the outputs of 2.3 and 1.75 miles of 19 gauge cable with identical inputs.

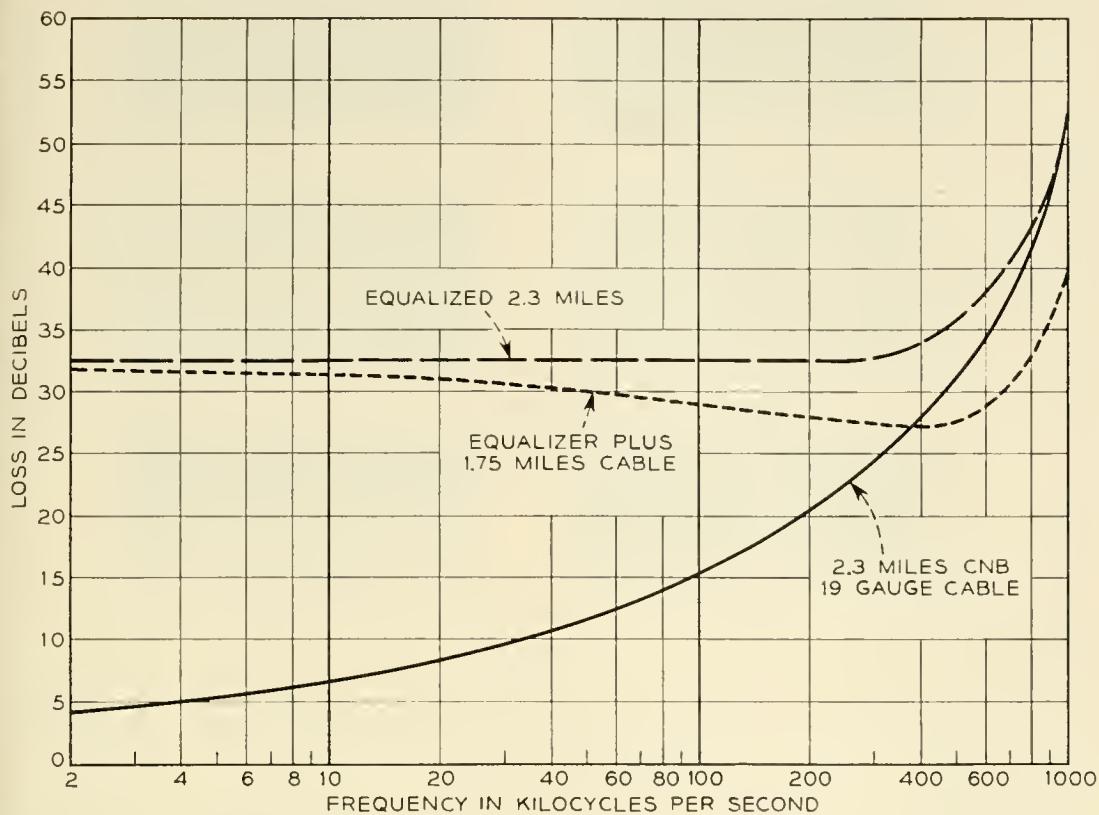


Fig. 6 — Effect of changing the length of 19 gauge line with fixed equalization.

4.1 Description of Error Detecting and Counting Circuit

An error detecting and counting circuit was built to count insertion and omission errors. This circuit (block diagram, Fig. 7) is a coincidence detector in which each pulse or space of the repeater input signal is compared to its corresponding regenerated output. As long as the two sources are the same, i.e., having corresponding pulses or spaces, there is no output from the detector. If the two differ the detector produces an output pulse which may be caused to actuate the counting circuit. The code generator as has already been described produces a number of different types of signal codes.

The output of the code generator is transmitted over 0.56 miles of equalized 32 gauge cable to the regenerative repeater under test. Interference is introduced at the repeater input when desired. A portion of the code generator output is differentiated and passed over a delay cable whose delay is substantially that of the section of 32 gauge line over which the signal is transmitted. This delayed signal is regenerated without error by the single shot blocking oscillator A. The width of the blocking oscillator pulses are adjusted to be about half of the total timing interval. The width of the pulses from the regenerative repeater

are likewise widened to a corresponding width by blocking oscillator B. Unfortunately a variable phase shift is introduced in the repeater output by interference and by variations in the timing wave amplitude and phase. This variable phase shift prevents perfect coincidence between the outputs of blocking oscillators A and B. An example of phase "jitter" caused by interference is shown on Plate V(a). To overcome this a sharp sampling pip; as shown on the same plate, is provided to enable the detection of the narrow region of coincidence between the two signals. These pips are generated from the repeater timing wave, hence they follow the timing wave phase variations. The regenerated signal pulses also follow the timing wave phase. If the sampling pulse is positioned to fall in the center of the regenerated pulses, it will tend to maintain that position as the timing wave changes.

The gates require a signal pulse and sampling pip to be present simultaneously before there can be an output. This output, then, will have substantially the same shape and position as the sampling pip. When a signal pulse is simultaneously applied to each gate the two outputs can be made to cancel when added in opposite phase as is done in T_1 . If however there is a pulse on one gate and a blank on the other, an output pulse will be produced. The polarity of this pulse will depend upon which gate contains the signal pulse. Since the decade counter is

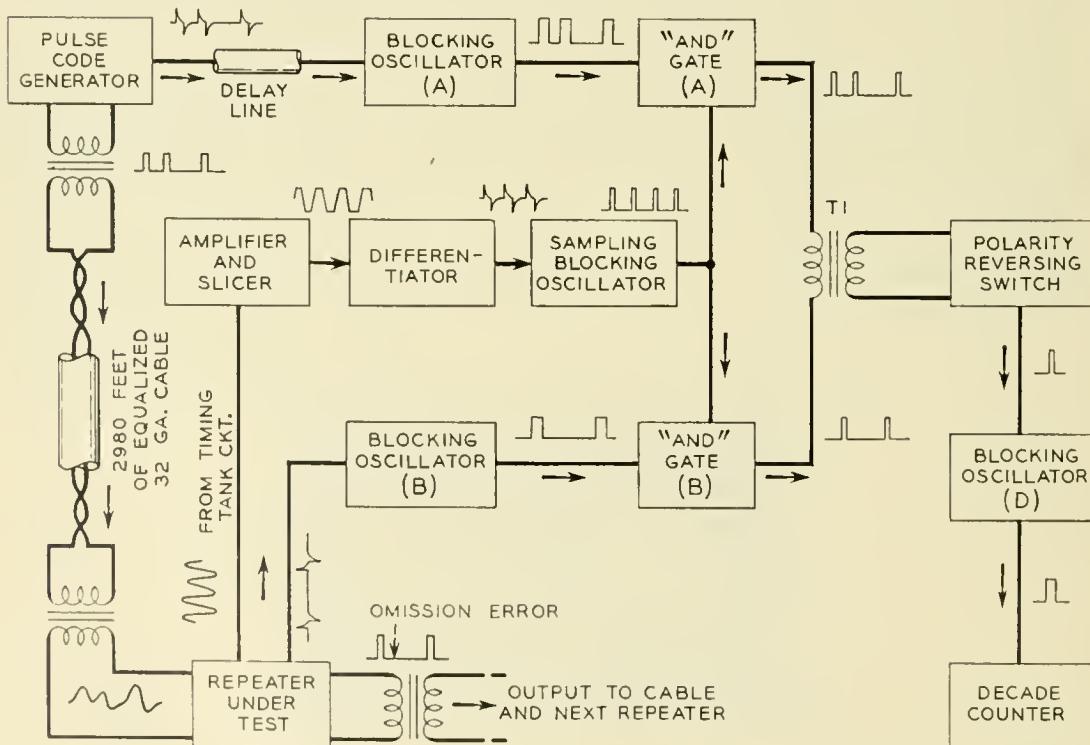


Fig. 7 — Block diagram of error detecting circuit.

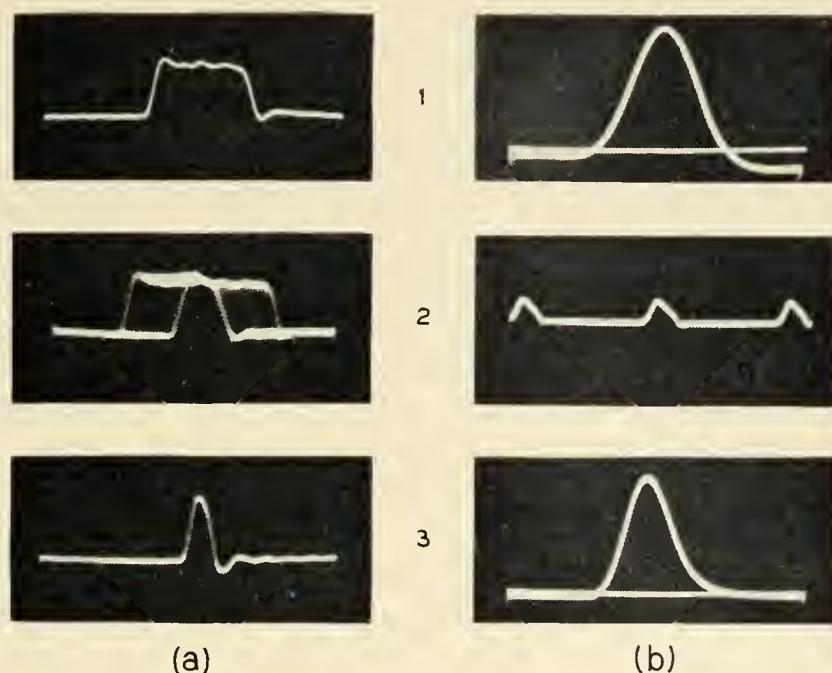


Plate V — (a) with 1, repeater output; 2, jitter on output pulse; 3, sampling pulse. (b) with 1, signal pulse at repeater input; 2, 672-ke timing pips; 3, interference input.

triggered by pulses of one polarity, the reversing switch permits the independent measuring of different types of errors. The counter used in this study has 9 decades capable of counting and recording ($10^9 - 1$) errors at 10^6 counts per second.

4.2 Discussion of Impulse Noise Generator

A study of the noise in cable pairs leading from a central office indicate that impulse noise will cause much of the expected interference on pulse systems. In order to simulate the effect of this type of interference, a generator was built which produces uniformly shaped pulses over a wide range of rates. The polarity of these pulses can be reversed and their amplitude varied continuously from zero to a value exceeding the peaks of the signal pulses. These impulses were introduced into the center of a transmission cable through a high impedance. Plate V(b) shows photographs comparing the impulse with a signal pulse. The repetition rate for the impulse interference used in this investigation was $10^4/\text{sec}$, which is low compared to the nominal pulse repetition rate of the signal ($6.72 \times 10^5/\text{sec}$). With the relatively large separation between interfering impulses, there is no measurable interaction between errors produced in the repeater. At the same time the impulse rate is high enough to get an excellent statistical distribution in the 10 second interval used in these measurements.

4.3 Production of Impulse Errors—Nomenclature and Discussion

To expedite the discussion of impulse errors, the following system of nomenclature is used. Any impulse having the same polarity as the signal pulse is designated as "plus." Those having the opposite polarity are "minus." Two types of errors are produced. First, a spurious pulse may be added to the regenerated signal; this is called an "insertion" error. Second, a signal pulse may be removed, which is called an "omission" error. A "plus insertion" error is a spurious pulse introduced by an impulse having the same polarity as the signal. A "plus omission" error on the other hand is pulse omitted because of a pulse of same polarity as the signal. A "minus omission" error is a pulse omitted because of an impulse having a polarity opposite to that of the signal.

A positive pulse, if large enough, can produce a spurious pulse at any instant of time not already occupied by a pulse. The only requirement for the production of such a pulse is that the sum of the impulse and timing wave exceed the trigger level.* On the other hand, a negative impulse cannot produce a spurious pulse but can only cause a signal pulse to be omitted. If a pulse is to be omitted the sum of its amplitude, the timing wave and the impulse must not exceed the trigger level. It would be expected that the number of plus insertion errors will exceed the minus omission errors. This follows from the fact that a spurious pulse may be produced at any point not already occupied by a pulse. On the other hand if a signal pulse is to be omitted the negative impulse must occur in the time interval occupied by the signal pulse. A positive impulse is indirectly responsible for the positive omission error. When a spurious pulse is produced a short interval of time ahead of a signal pulse, the latter may be removed by the inhibiting reaction of the spurious pulse. There is no apparent way in which a minus insertion error can be produced. This is confirmed by the fact that no error of this type was observed in this investigation. Thus we have three types of errors produced: plus insertion, minus omission and plus omission.

4.4 Results of Impulse Interference Measurements

Preliminary measurements of errors as functions of impulse amplitude were made using random code. These measured values, shown on Fig. 8 exhibit many of the expected characteristics. For example the insertion errors are more numerous than the omission and the threshold of the plus omission errors is considerably higher than those of the other two.

* The trigger level is normally considered to be the negative dc bias applied to the emitter of the blocking oscillator. There are however other components of the bias that will be discussed later.

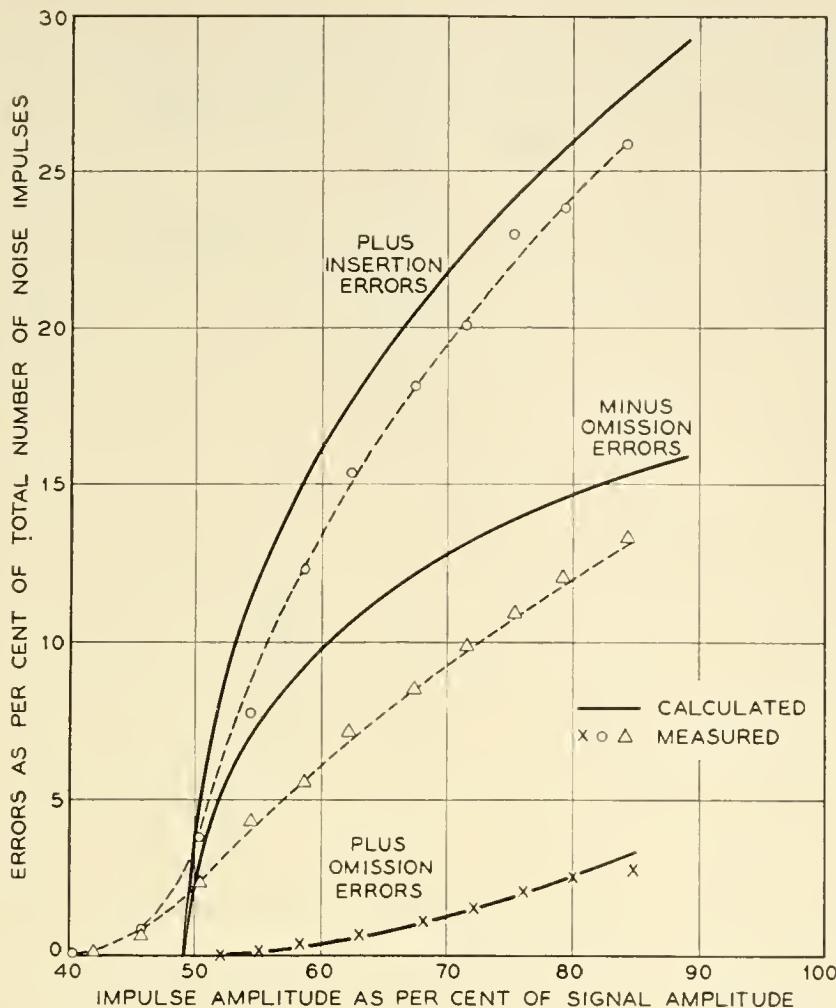


Fig. 8 — Repeater errors as a function of interference amplitude.

On the other hand there are some deviations from the simple theory of a perfect regenerator such as the low common threshold value of the plus insertion and minus omission errors. Some of the differences can be attributed to the extremely sensitive method of measuring errors. Here the maladjustments of timing tank circuit, quantized feedback amplitude as well as other factors which cannot be readily detected by other means are reflected as sources of error. However with care these errors can be made small and the measured values should follow the theoretical values reasonably well.

Most variations from theoretical values are due to changes in the effective bias caused by intersymbol crosstalk. This can be demonstrated by measurements made using set codes. In all these codes the number of pulses equaled the number of blanks but combinations varied from one to another. On Fig. 9 the omission errors are plotted for a fixed impulse amplitude as a function of the number of pulses which

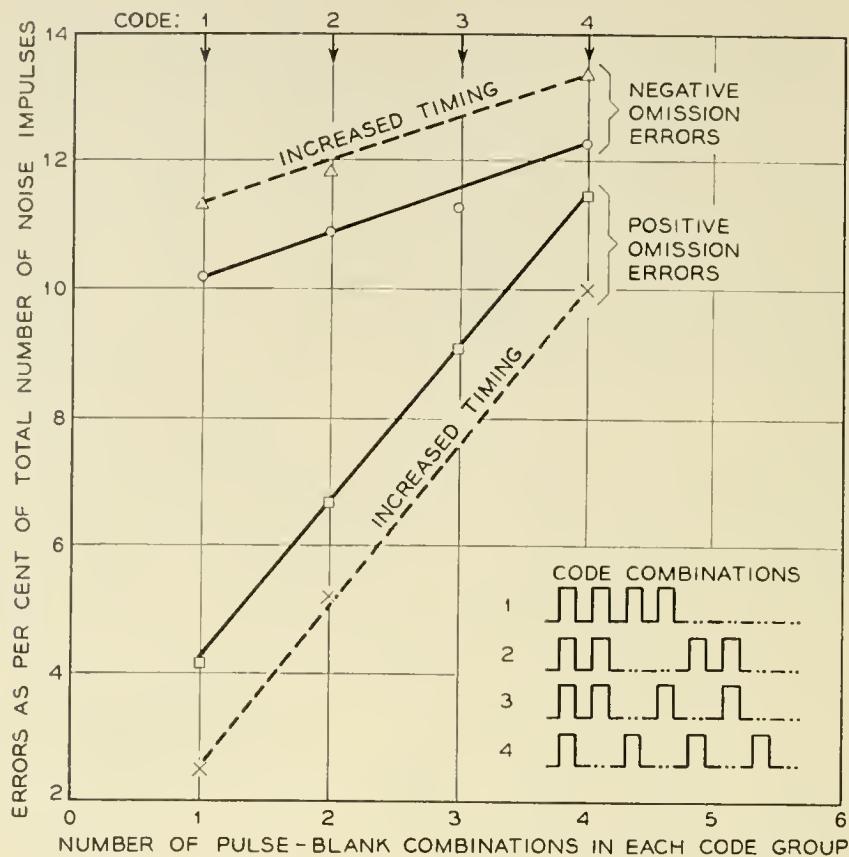


Fig. 9 — Repeater errors as a function of pulse distribution in code.

are followed by a space in the particular code. The codes used for various points on the abscissa are shown on the graph. The omission error curves plotted in this manner are linear. These data demonstrate that the presence of a pulse modifies the trigger level in the next timing interval. This is largely due to the negative excursion of the damped cosine voltage from base to ground in the blocking oscillator. On Fig. 10(a) is shown the circuit of the single shot blocking oscillator used in the repeater. With no timing an incoming signal must overcome bias V_{DC} to trigger the repeater. The solid curve on Fig. 10(b) shows the dc bias with the timing wave added at the blocking oscillator emitter. Fig. 10(c) shows the base voltage when a pulse is produced in the first timing interval. The pulse begins at t_0 and ends at t_1 . As previously mentioned the sudden rise of the base and collector impedance coupled with the fall of the current in the transformer windings, produces an inductive voltage surge across transformer T_3 at t_1 . The decay of this voltage surge can be controlled by the inductance of the transformer and the damping resistor R_b . This positive decay voltage across the base will inhibit the blocking oscillator from triggering. It is essential that this decay be adjusted so it will inhibit triggering until the following time slot. If

the decay transient is a damped oscillation and the base voltage passes through zero at the next normal triggering time, sufficient damping must be provided so the negative excursion is negligible. The dashed line shows how the effective bias at the emitter is modified by this voltage across the base.

Fig. 11 shows the measured values of plus insertion and minus omission errors for two set codes. These are plotted as functions of impulse amplitude. The first code has alternate pulses and blanks while the second consists of pairs of pulses separated by pairs of blanks. With these two curves the error threshold values may be determined from

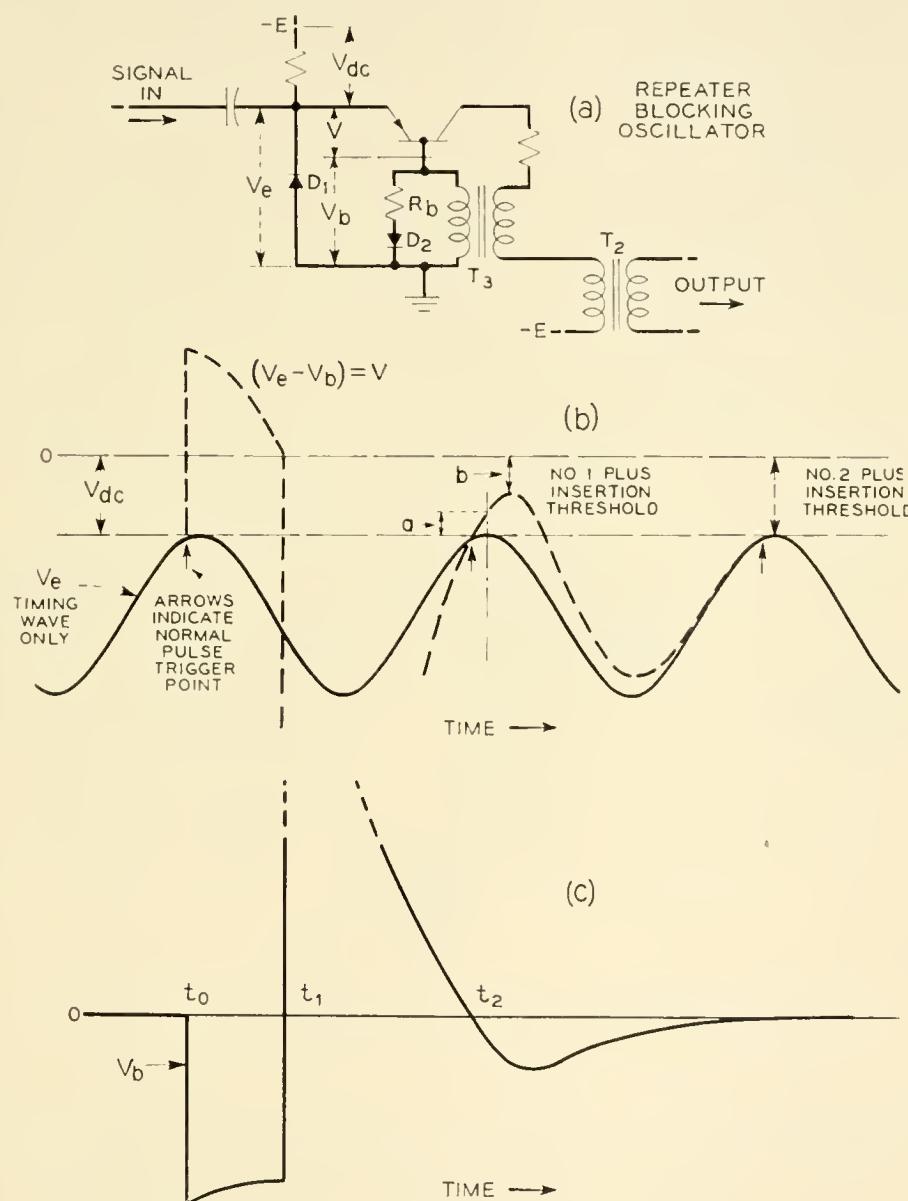


Fig. 10 — (a) Circuit diagram of blocking oscillator showing various components of the effective bias. (b) The effective bias as a function of time. (c) Inhibiting voltage V_b produced by a regenerated pulse.

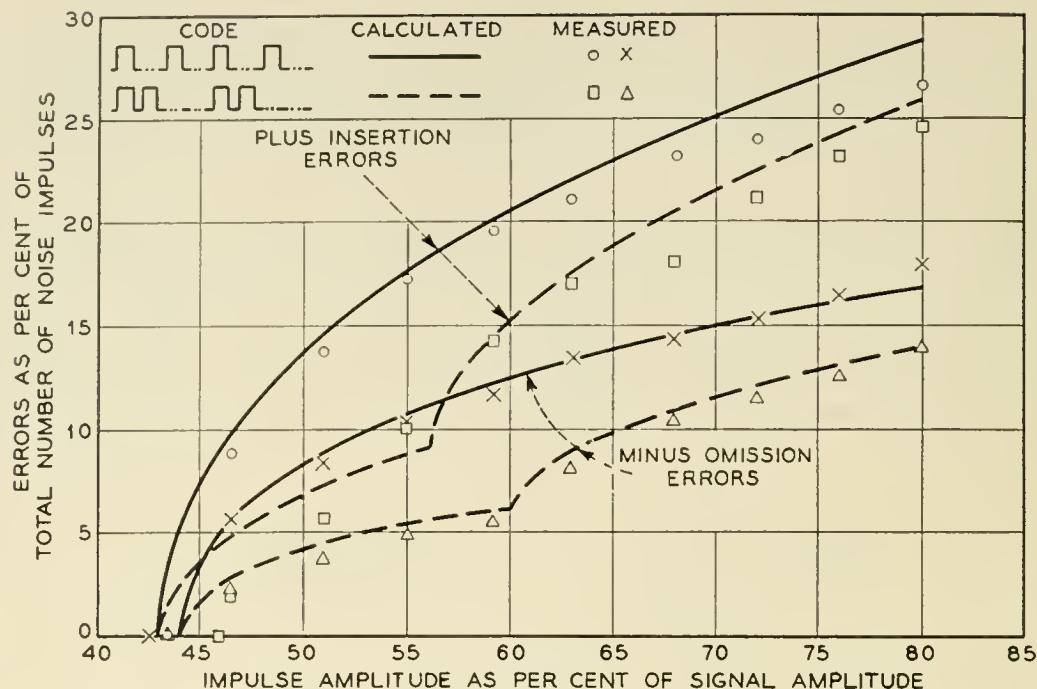


Fig. 11 — Calculated and measured repeater errors for two set codes.

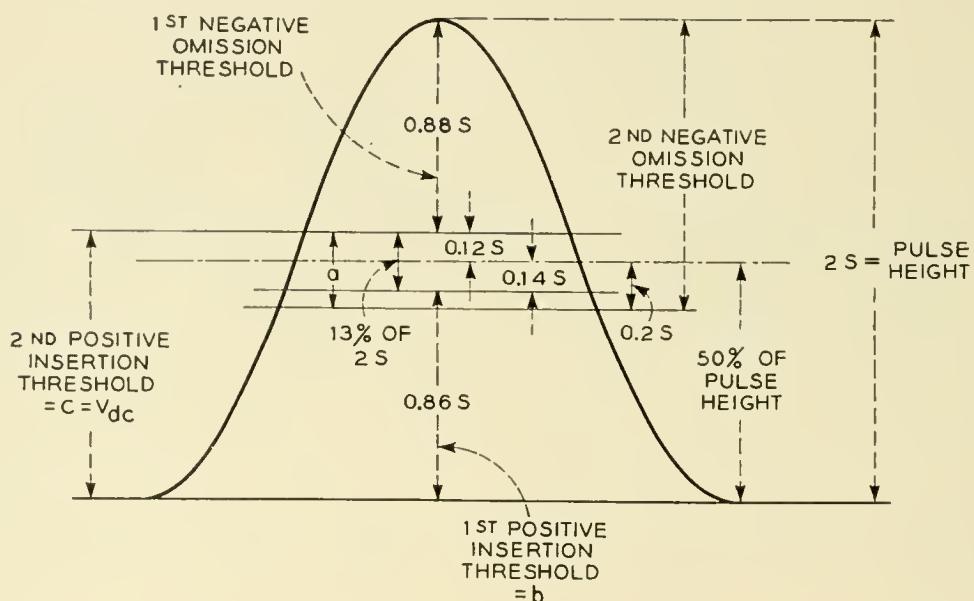


Fig. 12 — Bias levels used in calculating repeater errors.

the points of discontinuity. Fig. 12 illustrates these various error thresholds with reference to a signal pulse. Theoretical curves were plotted using these values and the observed values of timing and signal amplitudes as shown on Fig. 11. It can be seen that very good agreement exists between the measured and computed values.

The separate lower thresholds for insertion and omission errors may

be explained from Fig. 10(b). These are caused by the phase shift introduced by the inhibiting voltage to the effective bias compared to that of the timing wave. The omission thresholds are determined chiefly by the maximum signal amplitude. On the other hand the insertion thresholds are determined by the point of maximum trigger bias. There exists then two separate threshold values for a timing interval which follows a regenerated pulse. These values can be measured from points "a" and "b" on Fig. 10(b).

4.5 Result of Sinusoidal Interference Measurements

On Fig. 13 are shown the errors produced by sinusoidal interference. Here a 110-kc sine wave is added to the signal and the various types

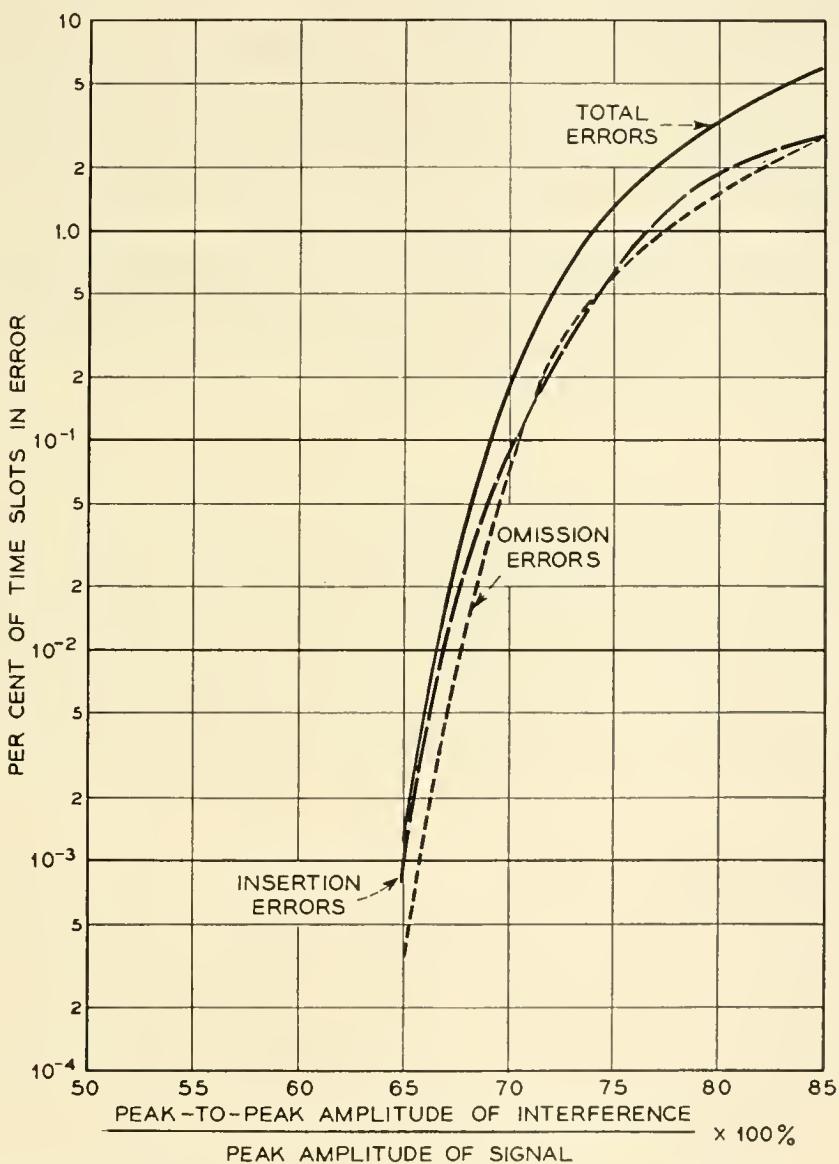


Fig. 13 — Repeater errors as a function of interferences level for sinusoidal interference.

of errors counted. Random code was used in this case and the repeater bias was adjusted to provide equal omission and insertion thresholds. The threshold for this particular case occurred when the peak to peak sinusoidal interference was 63 per cent of the signal amplitude. This is lower than the theoretical maximum which with a constant bias centered at the half amplitude point, would be 100 per cent of the peak to peak signal amplitude. For the bias conditions illustrated on Fig. 12, this percentage would be 86 per cent for the positive insertion threshold and 88 per cent for the minus omission. This becomes apparent when the negative and positive excursions of the interfering sine wave are considered as minus and positive impulses respectively. The remaining loss in the interference margins can easily be due to maladjustments of timing, quantized feedback or inhibiting.

When the frequency of the sinusoidal interference is varied, the number of errors for a constant interference voltage at the blocking oscillator emitter does not change appreciably. However, the input transformer and condenser coupling introduce a substantial frequency characteristic. This reduces considerably the errors caused by power line crosstalk. One of the striking things about the sinusoidal interference errors is the rate at which they increase above the threshold. For example, a change of 1 per cent of the interference amplitude can triple or quadruple the total number of errors.

5.0 SUMMARY

New techniques and devices now make it possible to build practical regenerative repeaters for use in digital transmission. Such a repeater which is suitable for a 12-channel, 7-digit PCM system, is discussed. Simple, inexpensive devices are used to eliminate the effects of distortion due to low frequency cutoff and to provide self timing for the circuit. Experimental evidence is presented which shows the repeater to function as expected.

ACKNOWLEDGEMENTS

I am deeply indebted to J. V. Scattaglia for his aid in this project and to the pioneering work of A. J. Rack on quantized feedback which was of great help in the development of this regenerative repeater. I also wish to thank W. R. Bennett, C. B. Feldman and Gordon Raisbeck for their aid and many valuable suggestions.

Transistor Pulse Regenerative Amplifiers

By F. H. TENDICK, JR.

(Manuscript received April 5, 1956)

A pulse regenerative amplifier is a bistate circuit which introduces gain and pulse reshaping in a pulse transmission or digital data processing system. Frequently it is used also to retime the pulses which constitute the flow of information in such systems. The small size, reliability, and low power consumption of the transistor have led naturally to the use of the transistor as the active element in the amplifier. It is the purpose of this paper to describe some of the techniques that are pertinent to the design of synchronized regenerative amplifiers operating at a pulse repetition rate of the order of one megacycle per second. An illustrative design of an amplifier for use in a specific digital computer is presented.

1. INTRODUCTION

A basic building block of many modern digital data processing or transmission systems is a pulse regenerative amplifier. The particular high speed transistor regenerative amplifiers to be discussed in this paper are intended for use in systems where the logic operations on the digit pulses are performed by passive circuits and the amplifiers are inserted at appropriate intervals to amplify, reshape, and retime the pulses. The design of these amplifiers for any specified system involves a knowledge of the environment of the amplifier in the system, a study of possible functional circuits which are combined to form an amplifier circuit, and the selection of a combination of these functional circuits to achieve the desired amplifier performance. Although a study of the functional circuits constitutes the major portion of this paper, the design of an amplifier for a particular digital computer is presented to illustrate the general design procedure.

One important way in which these amplifiers differ from many pulse amplifiers is that they must function properly under adverse conditions. That is, instead of merely expecting superior performance most of the

time under relatively special operating conditions, consistently good performance is demanded at all times, even with wide variations of circuit parameters and operating conditions (as, for example, a twenty-to-one variation in the required output current). Therefore, various circuit possibilities will be examined from the standpoint of reliable performance.

When the switching and mathematical operations of a digital data processing system are accomplished by a network of passive logic circuits with amplifiers interspersed to overcome circuit losses,^{1, 2} the environment of an amplifier is generally as indicated in Fig. 1. The signal information that passes from one logic network to another is represented in a code by a group of discrete pulses. Due to the nature of this digital information, utmost reliability of each amplifier is an important requirement that greatly influences the amplifier design. Since the position of a pulse in time or place determines its significance to the system, it is necessary that each pulse be identically amplified and that noise or extraneous disturbances do not cause false output pulses from an amplifier. The effect of an error or a failure in operation is different for different systems and in a given system depends upon the time or place of the failure. In some computers a single mistake will invalidate an entire computation cycle, while a permanent failure of even a single amplifier will cause complete system failure in almost any digital machine. Experience with the type of amplifier under discussion indicates that failure rates of less than a tenth of one percent per thousand hours are attainable.

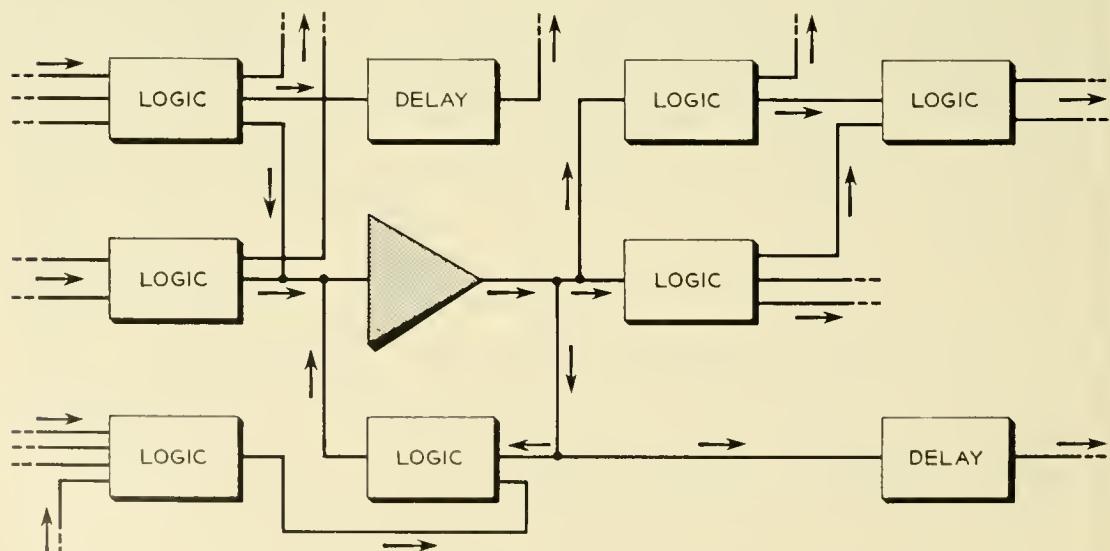


Fig. 1 — Typical environment of an amplifier.

This goal of reliable circuit operation can be realized if the amplifiers have:

- a. Simple circuitry with a minimum number of parts.
- b. The ability to operate with wide variations of signal level.
- c. Ample margins against crosstalk and noise.
- d. Low sensitiveness to changes in component values.
- e. Low power dissipation to realize long component life.
- f. Sufficient gain margins with system variations.

Although these features are desirable in any circuit, they are often subordinated in order to obtain special performance, usually at the expense of reliability. In the amplifiers under discussion these features represent the primary design goal.

As is so often true, some compromises usually must be made to obtain a suitable balance of these features in a particular design. It is sometimes possible to accept an increase in power consumption for other desired performance. However, because of the large number of amplifiers employed, low power operation is desirable in order to reduce the physical size and weight of a system. In this paper considerable emphasis is placed on efficient low power circuits which do not require critical components.

A convenient way to study regenerative amplifiers is to consider an amplifier as a small system. The following functional breakdown has been found useful:

- a. Transistor properties.
- b. Feedback circuits.
- c. Input trigger circuits.
- d. Output coupling circuits.
- e. Synchronizing circuits.

The block diagram of an amplifier then might take the form shown in

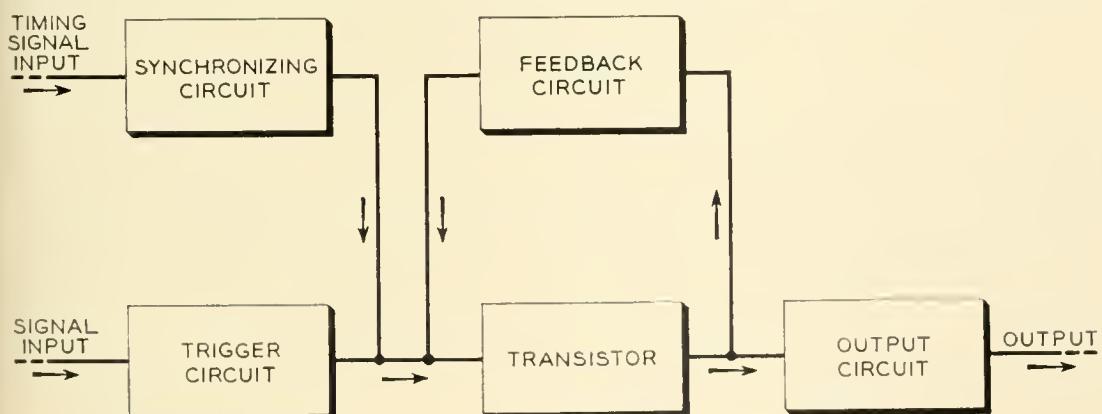


Fig. 2 — Regenerative amplifier block diagram.

Fig. 2. In the following sections the relation between each of the above functional features and amplifier performance is discussed, various circuit configurations to achieve each function are investigated, and the interactions between the functional circuits are examined. The design of any particular amplifier then consists of a suitable selection of a transistor and functional circuits to achieve the desired amplifier performance.

2. TRANSISTOR PROPERTIES

In a regenerative amplifier the transistor operates as a switch with power gain. The "on" and "off" state usually are characterized, respectively, by high and low collector current levels, and changes of state are initiated by applied control signals. The performance items of interest are the power dissipation in the two states, the speed with which the transistor changes state, the amount of power gain available, and the attainable margins against false operation. The transistor parameters related to these items, as discussed below, are listed in Table I with typical values for several classes of transistors. Desirable and satisfactory values have been indicated in italics.

The power dissipated in a transistor in the "off" state is proportional to I_{co} , the collector current with the emitter open circuited, and to the collector supply voltage. This is wasted power and, since the minimum collector supply voltage usually is dictated by other considerations, a low I_{co} current is desirable to reduce standby power. Point contact units are relatively poor in this respect. In junction units the I_{co} power is almost negligible compared to other circuit standby power.

The power dissipated in a transistor in the "on" state is proportional to the saturation voltage between the collector and the common terminal.

TABLE I — TRANSISTOR SWITCHING PROPERTIES

Switching Features	Point Contact Transistors (Low Resistivity Ge)	Junction Triode Transistors		
		Ge Grown	Ge Alloy	Si Grown
I_{co} at $V_c = 10v$	1500 μa	5 μa	5 μa	0.01 μa
Collector to emitter saturation voltage at $I_e = 10$ ma..	0.8 V	0.5 V	0.05 V	4 V
α cut-off.....	15 mc	2 mc	4 mc	4 mc
Base resistance.....	50 ohms	500 ohms	100 ohms	500 ohms
Collector capacitance at $V_c = 10V$	0.5 UUF	10 UUF	20 UUF	10 UUF
Collector breakdown voltage.....	40 V	100 V	35 V	100 V
Punch through voltage.....	no punch through	100 V	35 V	100 V
Emitter breakdown voltage.....	40 V	5 V	35 V	1 V
Ratio of alpha at $I_e = 10 \mu a$ to alpha at $I_e = 1$ ma....	3	0.8	0.8	0.6

Again, this represents wasted power, but also important is the fact that it places an upper limit on the output power available from the transistor. Hence, it is desirable to have as low a saturation voltage as possible. Alloy junction transistors are especially good in this respect.

The speed with which a transistor changes state is principally a function of the alpha cut-off frequency (which should be high), base resistance, and collector capacitance (both of which should be low).^{3, 4} Both the rise and fall times of the transistor response are greatly influenced by the associated circuitry; generally a blocking oscillator circuit yields the fastest response.

The amount of effective power gain available from a regenerative amplifier is influenced by two transistor properties. One property is the breakdown voltage, which may be the collector to base breakdown voltage or the collector to emitter punch through voltage (whichever is lower). This limits the output power by limiting the collector supply voltage. The other factor is the variation of alpha with emitter current, especially at low emitter currents. The minimum average emitter current required to initiate self-sustaining positive feedback determines the minimum input power. Point contact units are especially good in this respect in that alpha may approach ten at emitter currents as low as five microamperes. Junction units are poor since alpha generally decreases rapidly at emitter currents below one hundred microamperes.

Even though the attainable margins against false operation are largely a matter of circuit design, two transistor properties occasionally become important. In point contact units trouble with lock up in the "on" state may occur due to internal base resistance. Although this property of base resistance is exploited in negative resistance feedback circuits, it is undesirable in circuits where the feedback is obtained by external coupling. In grown junction units the emitter to base reverse breakdown voltage may limit the voltage margin against false triggering caused by noise or crosstalk. Normally it is desirable to have a one or two volt margin.

From the above discussion it can be seen that no one type of transistor is outstanding in all features. The choice of which unit to use in a specific amplifier depends upon the repetition rate, gain, and power requirements desired of the amplifier. Although the point contact type has the best overall performance of the types shown in Table I, it is quite possible that new types (such as PNIP or diffused triodes¹³) and improved designs of the present types will change the picture.

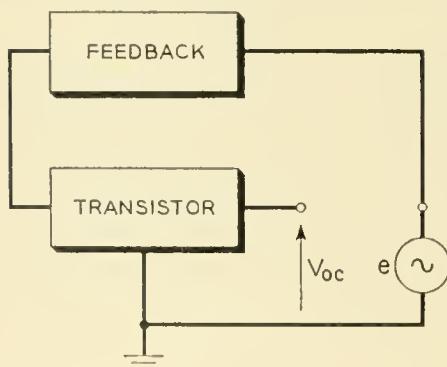
3. FEEDBACK CIRCUITS

The use of positive feedback in an amplifier results in high gain and short rise time. If the input circuit is isolated from the feedback loop by

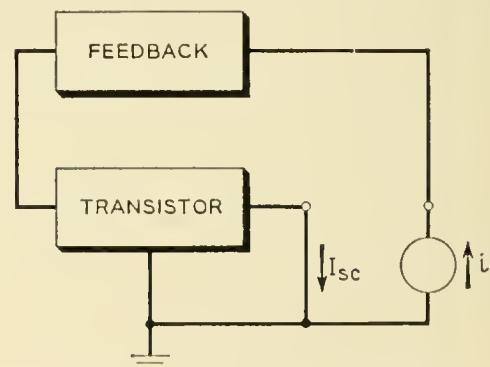
a diode or large resistor, these effects are enhanced and the shape, duration, and amplitude of the output signal become independent of the input signal. These results are possible because once the circuit has been triggered and the feedback loop gain is greater than unity, the response proceeds independently of input conditions and is determined solely by the transistor and circuit parameters.

By definition a regenerative amplifier must have positive feedback sufficient to cause instability during the transition period between the "off" and "on" states. When investigating various circuits, it is necessary to eliminate circuits which are never unstable when a pulse is applied to the input circuit. If the circuit is unstable under either of the conditions shown in Fig. 3, sufficient instability is possible. However, if the circuit is stable, linear, and either the small signal open circuit voltage gain or the short circuit current gain is less than unity or negative at all frequencies, it is impossible to have instability. These latter conditions for instability often can be easily checked by inspection without tedious computation or experimentation.

This use of positive feedback requires that attention be given to its control. To be useful, the amplifier must be stable in one state and at least quasi-stable in the other state. The change from instability in the transition period to stability in the end states is accomplished by a non-linear change in the gain or impedance of some element in the feedback loop. Usually the "off" state is made stable by causing the voltage and current conditions in the input circuit to reverse bias the transistor input. The "on" state may be made stable (or quasi-stable when there are reactive coupling elements in the loop) in several ways. For example, the transistor may be permitted to saturate when the desired pulse voltage is reached; a "catching" diode may be used to clip the pulse voltage at an appropriate level; or a current switch may be used to



(a) OPEN-CIRCUIT LOOP VOLTAGE



(b) SHORT-CIRCUIT LOOP CURRENT

Fig. 3 — A check for instability.

introduce an impedance in the feedback loop at a predetermined current level.

The degree of stability of the amplifier in the "on" state may be thought of as the amount of power required to initiate the transition to the "off" state. During the early portion of the output pulse duration the degree of stability should be large, but near the end of the pulse duration it should be relatively small to make turn-off easier. Also, the degree of stability should not change over the range of output loading expected for the amplifier and should be effected without excessive wastage of pulse or supply power. These conditions are difficult to fulfill when the range of output load current may be as large as 20 to 1.

Three methods of obtaining positive feedback in transistor circuits will now be considered: (a) negative resistance feedback; (b) capacitor coupled feedback; and (c) transformer coupled feedback. Of these, transformer coupled feedback appears to be the best for most applications. It will be assumed that the type of feedback under discussion is the dominant or only type present; circuits employing more than one feedback mechanism generally violate the premise of simple circuitry and will not be discussed.

3.1 Negative Resistance Feedback

With the advent of point contact transistors a novel form of negative resistance was offered to circuit designers for use in positive feedback applications.⁵ This negative resistance property occurs when the current gain of a transistor is greater than unity and the emitter and base small signal currents are in phase.* At first sight this property appears to lead to attractively simple regenerative amplifiers. However, as systems become more complex and, consequently, amplifier requirements more severe, the original simplicity often is lost due to the additional circuitry required to control the negative resistance. An example, shown in Fig. 4, is similar to a regenerative amplifier described by J. H. Felker.² The functional circuits are indicated by dashed outlines.

This amplifier operates at a one megacycle pulse repetition rate with one-half microsecond, three volt pulses. It is capable of driving from one to six similar amplifiers. The output pulse rise time is 0.05 microsecond, the average dc standby power is 33 milliwatts, only a few components operate at as much as half of maximum ratings, and the supply voltage margins† are greater than ± 15 per cent. Seven hundred of these ampli-

* Although point contact transistors are noted for this property, certain types of junction transistors also exhibit it. For example, see Reference 7.

† Supply voltage margins, the amount by which the supply voltage may be

fiers operated in a system for over 17,000 hours with a failure rate of slightly less than 0.07 per cent per thousand hours.

These features, however, are obtained at the expense of relative complex circuitry. This negative resistance type of high speed regenerative amplifier has the following inherent limitations.

1. The degree of stability in the "on" state depends critically on the collector current. In the example a dummy load must be strapped in when the amplifier drives less than four logic circuits.
2. A steering diode (D3) and a timing circuit diode (D1) have critical reverse recovery time⁶ specifications.*
3. The requirements on transistor parameters (primarily the dynamic alpha versus emitter current and base resistance characteristics) are relatively critical.
4. A relatively large amount of synchronizing power is required.
5. With transformer output coupling (as discussed in Section 5.1) a large amount of the total standby power is absorbed by a circuit required to protect the transistor in case the timing voltage fails (In the example 21 milliwatts, or 64 per cent of the standby power, is absorbed by R3.)

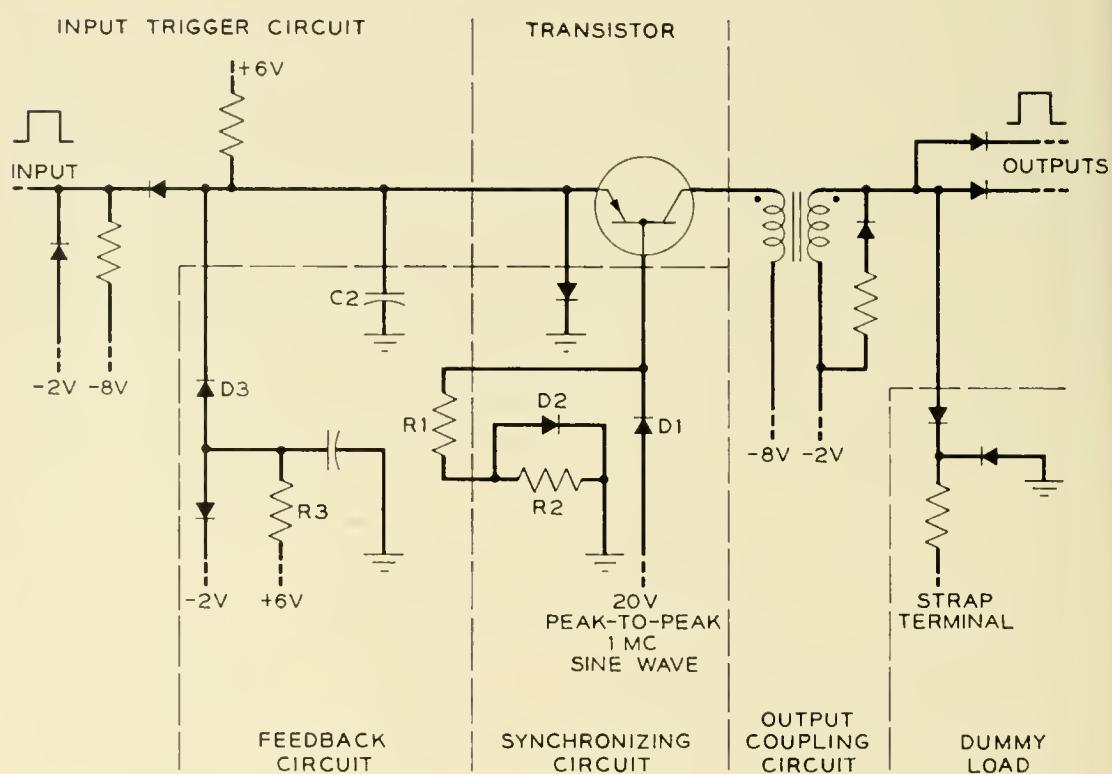


Fig. 4 — Negative resistance feedback amplifier.

varied without causing an operational failure, are an indication of the sensitivity of the amplifier to changes in component values.

* At lower pulse repetition rates this property may not be critical.

The use of an inductor, instead of a resistance, in the base lead does not appear to mitigate the limitations.

3.2 Capacitor Coupled Feedback

A second method of obtaining positive feedback is by external coupling through a capacitor or capacitor-resistor network. This method is seldom used for the principal feedback for reasons to be mentioned. Occasionally, in conjunction with some other type of feedback, it may be used to provide additional feedback during the rise time of an amplifier.

Since the voltage and current gain of a capacitor can not exceed unity, the open circuit voltage gain and the short circuit current gain of the rest of the loop (Fig. 3) must be greater than unity for instability. This criterion indicates that capacitor feedback is limited to point contact, or other transistors with an alpha greater than unity, or to a junction transistor in the common emitter configuration.*

A circuit with capacitor feedback around a short-circuit stable point contact transistor might take the form shown in Fig. 5. Although this type of circuit has the merit of simplicity, it has the following limitations:

1. The initial feedback current is highly dependent upon the incremental output load impedance. This may result in a failure to trigger when the load approximates a short circuit, as in the case of diode gates or a large stray capacitance.

2. The degree of stability in the "on" state is critically dependent on the load current and the collector supply voltage. Variations in either may cause a foreshortened output pulse or require an excessive timing signal current for turn-off.

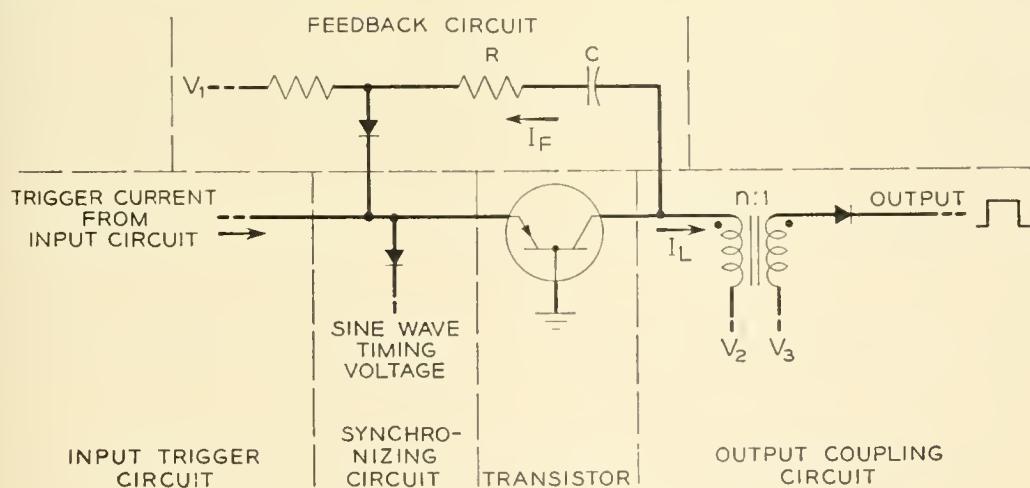


Fig. 5 — RC feedback amplifier.

* An inverting transformer is necessary with the junction transistor.

3. The necessity of a feedback circuit time constant equal to or shorter than the output pulse length results in a relatively low output power efficiency.

Due to the above considerations, capacitor feedback appears to be the least attractive type of feedback.

3.3 Transformer Coupled Feedback

A transformer appears to be the most convenient and versatile component for feedback coupling in a regenerative amplifier. The pertinent features* of a transformer are:

1. Current or voltage gain (impedance matching.) This feature permits full use of the power gain of the transistor, even if such gain be in the form of voltage or current gain only.
2. Bias isolation between circuit parts and the possibility of supplying dc voltage bias without the use of additional elements.
3. Phase inversion, if desired.

All of these features, conveniently combined in a transformer, provide great design freedom to meet specified circuit objectives. Since positive feedback is possible with any type of transistor (with power gain, of course), the choices of transistor and connection are determined by other circuit requirements.

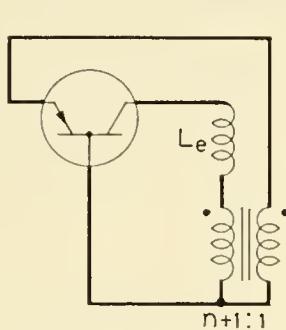
The use of transformer coupled feedback yields the familiar blocking oscillator circuit. An important feature of this circuit is the fast rise time that is obtainable. Linvill and Mattson⁴ have shown that a junction transistor with an alpha cutoff frequency of two megacycles may exhibit a rise time of 0.1 microsecond in an unloaded blocking oscillator with collector to emitter coupling, Fig. 6 (a). It can be shown that the same response may be expected with collector to base or base to emitter coupling, provided that the transformer turns ratio is modified, Figs. 6 (b) and 6 (c). When the circuit is providing useful output power into a load, a slightly different turns ratio would be used for optimum rise time, which may be appreciably slower than in the unloaded case. However, it should be noted that the foregoing gives no information about the initial response of the circuit from the time that the input trigger is applied until the output reaches ten per cent of its final value. In some instances this initial time, which is a complicated function of the transistor non-linearities, may be comparable to the output rise time.

In a blocking oscillator circuit with a fixed output load, the degree of stability in the "on" state decreases with time. The reason is that the

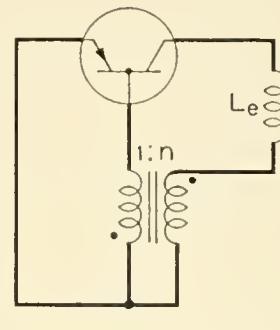
* The operation of a transformer over the non-linear portion of its magnetization characteristic is outside the scope of this paper.

voltage across the coupling transformer, which is approximately constant during the pulse duration, causes an increasing magnetizing current to be subtracted from the initial feedback. When the feedback current can no longer support the required output current, the circuit turns off. In a synchronized amplifier the value of the feedback transformer mutual inductance may be specified to give the desired degree of stability at the end of the predetermined pulse length. Thus, the least stable condition occurs at the end of the pulse duration and is under the circuit designer's control. At other times during the pulse duration the circuit is more stable, which reduces the possibility of premature turn-off.

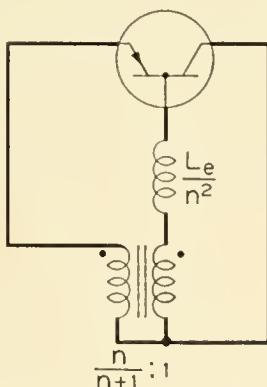
Other considerations, such as stability variations with output current, power dissipation, and output voltage regulation, depend upon whether the output load is in series or in shunt with the feedback loop. Therefore, these considerations are discussed in connection with output coupling in



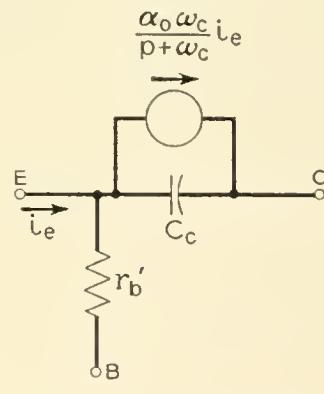
(a) COMMON BASE



(b) COMMON EMMITTER



(c) COMMON COLLECTOR



(d) ASSUMED TRANSISTOR EQUIVALENT CIRCUIT

L_e = LEAKAGE INDUCTANCE OF TRANSFORMER

n = TURNS RATIO FOR COMMON EMMITTER CONNECTION

α_0 = LOW FREQUENCY VALUE OF COMMON BASE
SHORT CIRCUIT CURRENT GAIN

ω_c = CUTOFF RADIAN FREQUENCY OF α

Fig. 6 — Transformer coupled blocking oscillator circuits.

Section 5.2. For constant voltage, variable current loads, transformer coupled feedback with the output load in series with the feedback loop results in low power dissipation, relatively small degree of stability variations versus output current variations, and non-critical components. The possible limitations are that transformers generally are more expensive than other passive components and are not as readily available in a variety of stock values.

4. INPUT TRIGGER CIRCUITS

The primary function of the input trigger circuit is to initiate the transition from the "off" to the "on" state when there is an input signal. At all other times the input circuit must provide a threshold or margin against false triggering due to noise or spurious disturbances.

Although the input circuit must supply sufficient energy to establish regeneration, it is unnecessary and undesirable that any additional energy be supplied. To do so reduces the gain of the amplifier, since gain may be defined as the ratio of the output power to the input power during one cycle of operation. Because regeneration makes the input and output power independent of each other, any reduction in input power results in greater amplifier gain.

In an amplifier with external feedback coupling it is possible, but not always practical, to have the input circuit trigger the transistor at the collector, base, or emitter terminal. The collector terminal seldom is selected because then the input circuit must supply energy to the output load as well as to the transistor. Also, the base is usually not used (except occasionally with negative resistance feedback) because extra components are required to steer the triggering energy into the transistor and it is difficult to apply a timing signal.* However, the following discussion and the dc input characteristic of Fig. 7 (a) are equally valid for triggering at the base or emitter terminal of junction or point contact transistors which are short-circuit stable.

One of the simplest types of triggering circuits is shown in Fig. 7 (b). The voltage and current increments assumed necessary to initiate regeneration are designated V_t and I_t . Therefore, the required input signal voltage V_s and current I_s are:

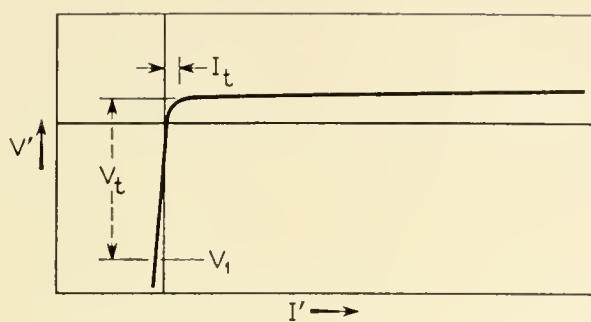
$$V_s \geq V_t + I_t R_1 \quad (1)$$

$$I_s \geq I_t \left(1 + \frac{R_1}{R_2}\right) + \frac{V_t}{R_2} + \frac{V_1 - V_2}{R_2} \quad (2)$$

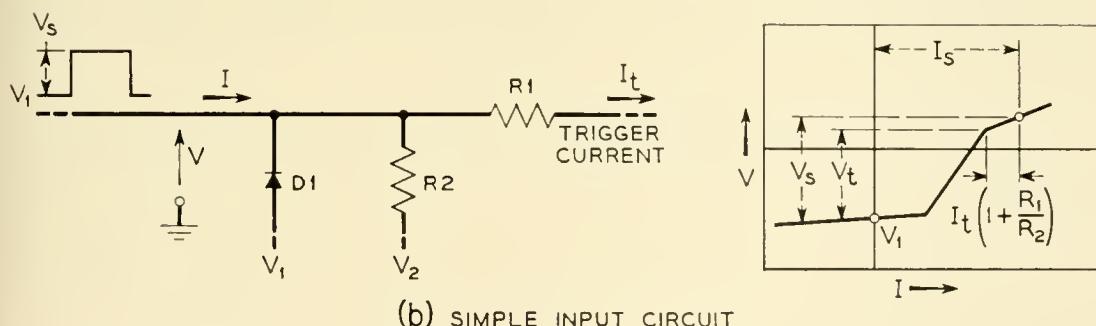
* Also, for junction transistors, about twice as much energy is required to trigger at the base as at the emitter.⁴

The purpose of diode D_1 is to provide a low impedance current threshold, the amount of current given by the last term of (2). This type of threshold is especially effective for preventing false operation from electrostatically induced crosstalk. Also, it allows a faster rate of discharge of stray capacity on the input terminal at the end of the input pulse period.

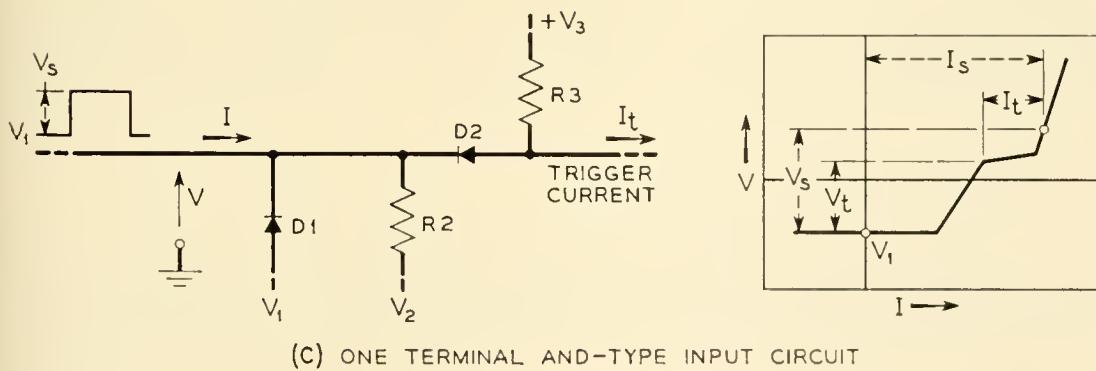
Although the circuit of Fig. 7 (b) is attractively simple, it is undesirably sensitive to variations in signal voltage. An increase in the input pulse voltage causes excessive triggering current and a decrease may easily result in failure to trigger. Since the circuit must be designed to operate reliably with the smallest expected input pulse, it is wasteful of input power with the average amplitude of input pulse.



(a) TRANSISTOR INPUT CHARACTERISTIC



(b) SIMPLE INPUT CIRCUIT



(c) ONE TERMINAL AND-TYPE INPUT CIRCUIT

Fig. 7 — Input trigger circuits.

The single terminal AND-type circuit^{9, 10} Fig. 7 (c) has the desirable characteristics of the previous circuit, and is relatively insensitive to input signal variations. In this circuit the input pulse switches the current through R3 into the transistor input and then encounters the relatively high resistance R2, as compared to the parallel resistance of R2 and R1 in Fig. 7 (b). The blocking action of D2 thus reduces variations in the input signal current. However, R2, R3, V_3 and V_2 cannot be increased without limit to reduce the variations; the dc power dissipated in R2 and R3 would become excessive.

Another advantage of the AND-type circuit is that several inputs may be paralleled with a common R3 to provide an AND logic function as well as an input trigger function. This feature, when desired, saves components and does not reduce the gain of the amplifier.

When both the input circuit and the feedback circuit terminate at the same transistor input terminal, as is usually the case, some additional components are generally required to prevent one circuit from shunting the other circuit. To steer the trigger current into the transistor, a diode may be placed in the feedback path so that the diode is reverse biased except when there is feedback current. Similarly, a diode or a resistor may be placed in the input circuit so as to prevent the feedback current from flowing into the input circuit.*

Although the discussion has assumed positive polarity input pulses, the remarks apply equally well to negative pulses if the polarity of the diodes and the supply voltages are reversed.

It is recognized that the preceding remarks assume that the minimum triggering energy is known and that a step function of current or voltage is the optimum form of the triggering energy. Actually, until a study is made of the circuit and transistor parameters (including the non-linear aspects) that affect the initial triggering before the feedback is established, the design of an optimum input trigger circuit will remain an experimental art. Experience with the AND-type input circuit has indicated that appreciably more current is required to trigger junction units than point contact units.

5. OUTPUT COUPLING CIRCUITS

In addition to the obvious function of efficient power transfer from the amplifier to a load, the output coupling circuit is a convenient point at which to perform other functions, as for example, dc level restoration

* This precaution is not necessary if the transistor input exhibits appreciable negative resistance.

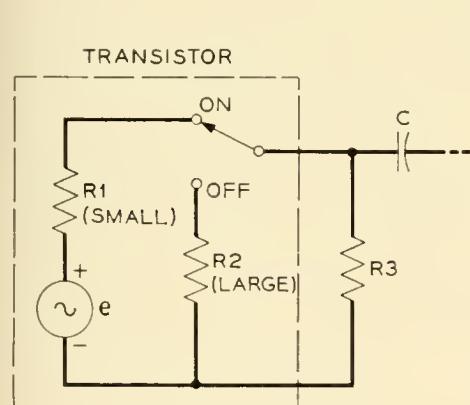
and pulse inversion. In a system of logic circuits interspersed with amplifiers at regular intervals, it is apparent that the dc level at similar points, such as the outputs of the amplifiers, must be identical if the amplifiers are to be interchangeable. Without some circuit or element to restore the dc level, the levels along the transmission path will monotonically decrease* due to the dc voltage loss through the logic circuits and across the transistor in the amplifier. The output circuit is one point where restoration of the dc level may be readily combined with other functions.†

In the following two sections three methods of output coupling are discussed and the interaction between the output and feedback circuits is considered.

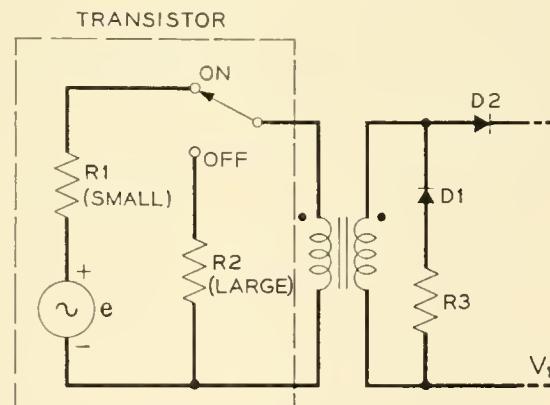
5.1 Output Coupling Elements

Three types of coupling circuits are RC, transformer, and diode coupling. Each of these methods permits the dc level of the signal pulses to be corrected to a predetermined level. However, the restoration,‡ efficiency, and versatility characteristics of each circuit are quite different.

Although RC coupling is common in linear amplifiers, it is seldom used in transistor pulse amplifiers that operate at duty cycles near 50 per cent. The reason is that the time constants encountered do not permit both proper restoration of the capacitor and high efficiency of the output circuit. As indicated in Fig. 8 (a), the transistor is a low impedance in



(a) RC COUPLING



(b) TRANSFORMER COUPLING

Fig. 8 — Reactive output coupling circuits.

* Decrease for positive pulses; increase for negative pulses.

† An exception, to be discussed, is diode output coupling where it is occasionally more convenient to correct the dc level in the input of the logic circuits or the amplifier.

‡ This refers to restoration of a reactive element (i.e., the return to a quiescent state) and is not to be confused with restoration of the dc level of a circuit.

the "on" state and a high impedance in the "off" state. Since C must be relatively large to make the voltage drop across it small during the pulse duration, R3 must be equal to or smaller than R1 for satisfactory restoration (50% duty cycle assumed). But then the current transmission efficiency of the coupling network is less than 50 per cent because generally R1 is smaller than the input resistance of the driven circuits during the pulse duration. Unless the pulse length is only a small fraction of the pulse repetition period, it is seldom possible to effect a suitable compromise. Also, it might be noted that variations of I_{co} current, which flows through R3, cause variations in the output pulse amplitude. Finally it is not possible to obtain pulse inversion.

A transformer coupled circuit, Fig. 8 (b), works efficiently with a transistor. Diode D2 isolates the transformer from the load and interlead stray capacitance during the interdigit period* so that the restoration time of the transformer is controlled by the value of R3. The restoration time is approximately proportional to the mutual inductance divided by the total shunting resistance. Diode D1 prevents R3 from shunting down the output during the pulse duration, thus permitting high output efficiency and proper restoration.†

As noted in Section 2, the maximum output power from the transistor is determined by the maximum collector voltage (as set by breakdown or punch-through) and the maximum collector current consistent with the permissible dissipation in the transistor. Usually this maximum voltage exceeds the desired amplifier output voltage and, occasionally, the maximum collector current is insufficient; in such instances a voltage step down is desirable. When the transistor is not required to operate at maximum power dissipation, it often is advantageous to balance the "off" and "on" power dissipation. An increase in the collector supply voltage increases the "off" power and decreases the "on" power (by decreasing the required collector current for the same output power). Thus the collector voltage may be adjusted to give the lowest total power dissipation consistent with the average duty cycle of the amplifier. The transformer turns ratio is specified to match the optimum collector voltage to the desired output voltage. Furthermore, I_{co} variations have negligible effect on the output voltage amplitude and pulse inversion (if desired, for example, for inhibition) is possible. For these reasons transformer coupling appears to give optimum output coupling performance.

* The minimum time interval between the end of one pulse and the beginning of a succeeding pulse; for a 50 per cent duty cycle the interdigit period is equal to the pulse duration.

† Occasionally it is possible to specify the collector impedance, the transformer losses, and the reverse impedance of D2 so that D1 and R3 are not necessary.

A third method of coupling, which is attractive for systems using only AND- and OR-type logic, utilizes the reverse characteristic of a breakdown diode, Fig. 9 (a). The interesting feature of this diode is the sharp transition between the high and low incremental resistance regions of the reverse characteristic. With this diode it is possible to shift dc levels by an amount equal to the reverse breakdown voltage of the diode, as indicated in Fig. 9 (b). In the quiescent state D2 operates in the breakdown region and D1 serves to clamp the collector voltage at $-V_3$; during the pulse duration D2 operates in the high resistance portion of its reverse characteristic. If the driven circuit has a voltage threshold, like the transistor threshold in Fig. 7 (a), less than $-V_3 + V_B + V_s$ and $V_s < |V_B|$, the circuit operates like a normal AND-type circuit except for the dc level change. For this reason it is convenient with AND-OR logic circuits to include only D1 and R2 in the output circuit of the amplifier and use D2 and R1 as the AND input elements in the logic circuits.

The principal advantages of diode coupling are simplicity and the lack of an energy storage element. The limitations are that there is no opportunity to match transistor and output conditions, variations in

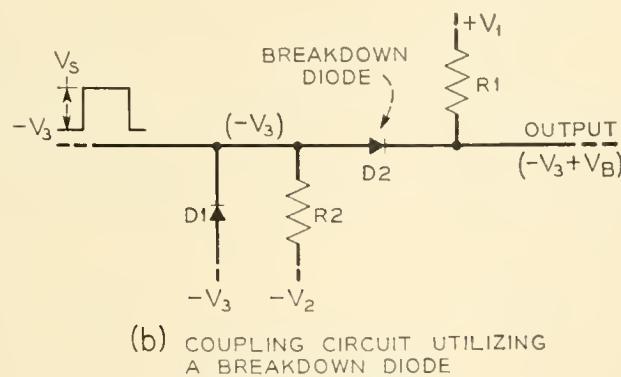
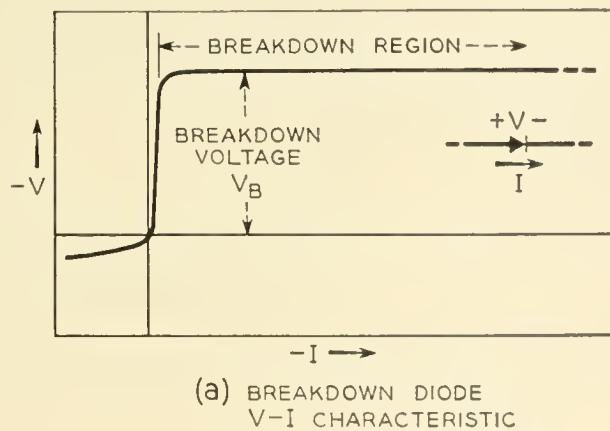
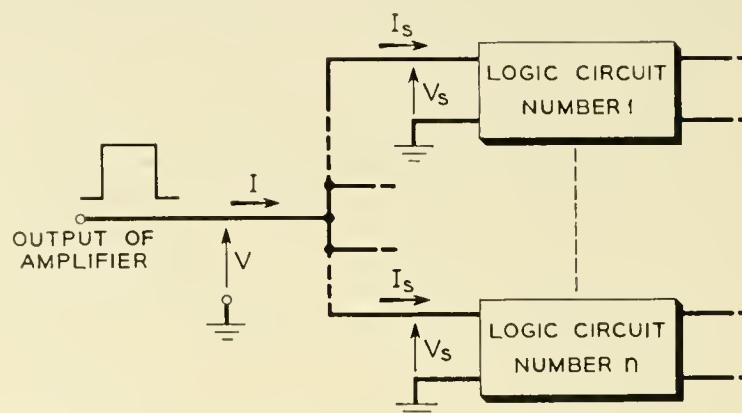


Fig. 9 — Direct output coupling circuit.

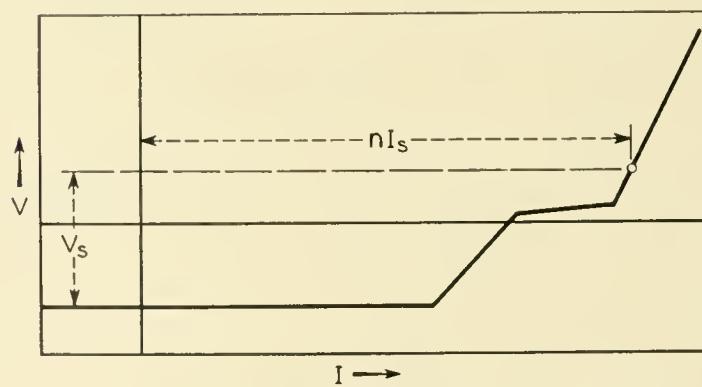
diode breakdown voltage reduce amplifier margins, and pulse inversion is not possible. For these reasons diode coupling has limited utility, but is attractive for some applications.

5.2 Connection of Output and Feedback Circuits

The performance of the amplifier is greatly affected by the method used to connect the output circuit and the feedback circuit together at the output of the transistor. Should these two circuits be connected in a shunt or a series fashion? Performance features, such as rise time, sufficient output voltage, degree of stability versus load current variations, and power dissipation directly depend upon this choice. With transformer output coupling, the choice always exists; with other types of output coupling the choice may or may not exist, depending upon the type of feedback coupling. The following discussion is in terms of transformer coupled output and feedback circuits and the general conclusions may be extended to other cases.



(a) CONNECTION OF AMPLIFIER LOAD



(b) V-I CHARACTERISTIC OF AMPLIFIER LOAD

Fig. 10 — Output load characteristic.

The principal factor that influences the choice of the output-feedback connection is the nature of the output load of the amplifier. In the majority of computer and switching systems the amplifier must drive a multiplicity of paralleled load circuits, as indicated in Fig. 10 (a). The input characteristic of each load circuit is assumed to be of the threshold type, like the AND-type input characteristic of Fig. 7 (c), which results in the amplifier load characteristic of Fig. 10 (b). During the initial portion of the rise time of the output pulse the incremental impedance is almost zero and during the remainder of the pulse duration it is relatively large. Due to the voltage threshold nature of the load, the amplifier load variations are current variations at a constant voltage. The minimum current is encountered in the system position where the amplifier drives the smallest number of logic circuits, often a single logic circuit; the maximum current is limited by the maximum output power of the amplifier. Although a desirable ratio of maximum to minimum current may be as high as 20:1, the amplifier is expected to exhibit optimum performance at any load current within this range.

The shunt connection of the output and feedback circuits is illustrated in Fig. 11.¹² Windings $1:n_1$ constitute the feedback coupling and $1:n_2$ the output coupling. The two circuits shunt each other in the sense that the ratio of the feedback to the output current is determined by the ratio of the impedance of these circuits as modified by the turns ratio of the transformer.

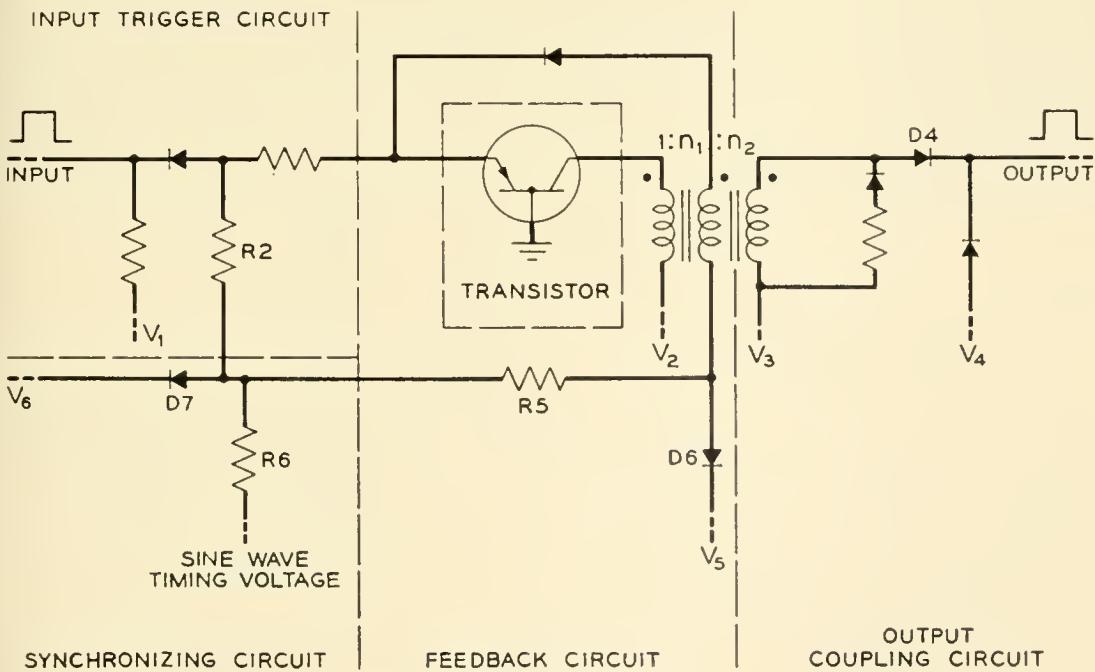


Fig. 11 — Shunt connection of output and feedback.

There are two limitations associated with this output-feedback connection. In the first place there is the possibility of insufficient output voltage, slow rise time, or complete failure of regeneration. This is caused by the shunt effect of the output load which places an almost zero initial incremental impedance across the feedback path. In order to overcome this limitation a current switch (R_5 and D_6 in Fig. 11) is used to obtain a low initial feedback impedance and the output diode (D_4) is reverse biased so that the initial load impedance is large. The price paid is the undesirable power dissipation in the current switch. Moreover, stray capacity across the output terminal or a load current that exceeds the design value may still result in a long rise time, low output voltage, or regeneration failure.

The series connection of the output and feedback circuits is shown in Fig. 12. In this connection the output load is in series with the feedback loop. Thus, the transistor output current, feedback current, and output load current are all proportional to each other. This situation assures regeneration regardless of output load current variations.

The regeneration cycle of the series type amplifier is as follows. In the quiescent state diode D_2 is reverse biased by V_1 to prevent false triggering. After the arrival of an input signal, the timing signal voltage goes positive and steers the trigger current into the transistor. No

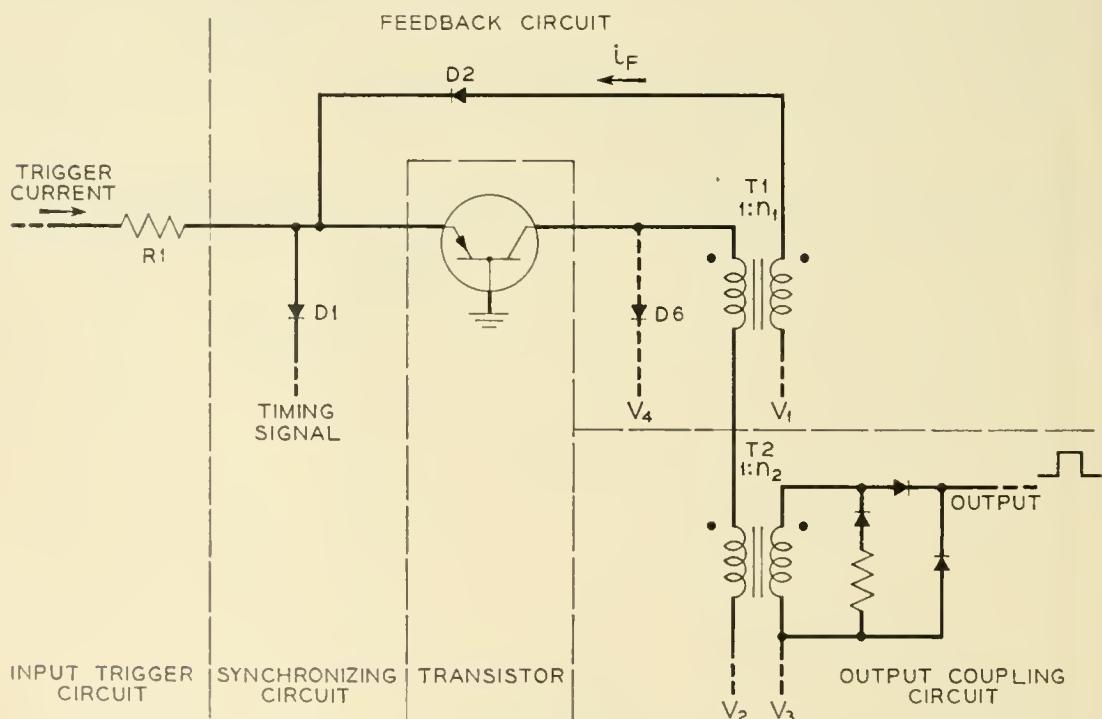


Fig. 12 — Series connection of output and feedback.

appreciable output current flows until the voltage across transformer T1 is sufficient to forward bias diode D2. Then both the feedback and output current build up simultaneously and rapidly since the turns ratio $1:n_1$ of T1 is selected to give a feedback loop gain greater than unity. When the sum of the voltages across the primaries of the feedback and output transformers almost equals the collector supply voltage, the transistor saturates and stabilizes the feedback loop. At the end of the pulse duration the timing signal voltage goes negative and robs current from the feedback loop, thus forcing the transistor out of saturation and causing the amplifier to turn off.

Because the feedback current is proportional to the output current during the rise time, the amplifier can deliver any value of load current up to the current corresponding to the maximum allowable collector current. Also, assuming that the leakage inductances of the transformers are small, a large stray capacitance across the output terminal does not appreciably degrade the rise time. Since a current switch is not necessary, the standby power dissipation in the feedback loop is negligible. These are the outstanding features of the series connection.

Two important performance considerations of the series type amplifier are the change in the degree of stability versus load current variations and the action of the amplifier when the timing signal fails. Both of these items may be controlled by the selection of suitable values for the turns ratio and the primary inductance of the feedback and output transformers.* In order to prevent burnout of the transistor in the event that the timing signal fails, the amount of excess feedback current must decrease during the pulse duration. Due to the low impedance of the feedback loop, this condition may be approximately† stated in terms of the primary inductances as:

$$\left| \frac{V_1}{n_1 L_1} \right| > \left| \frac{V_2 - \frac{V_1}{n_1} - V_{sat}}{\frac{L_2}{1 - \frac{n_1}{\alpha}}} \right| \quad (3)$$

where V_{sat} is the collector saturation voltage and L_1 and L_2 are the primary inductances of T_1 and T_2 respectively.

The degree of stability in the series amplifier at the end of the pulse duration is proportional to the output load current. This situation may be seen more clearly if a "catching" diode (D6 in Fig. 12) is added to the

* If the transistor is not short circuit stable, it is also usually necessary to use a small resistance in series with the emitter.

† The principal approximation is that alpha is constant versus collector current. The value of alpha at the end of the pulse duration is a conservative value.

circuit to prevent saturation in the transistor.* Because the feedback loop gain, as determined by the alpha of the transistor and the turns ratio n_1 , must be greater than unity for regeneration reasons, there will be current flow through D6 during the pulse duration. This current is proportional to the degree of stability. An increase Δi_{out} in the output current causes an increase of

$$\Delta i_c = \frac{\alpha n_2 \Delta i_{out}}{n_1} \quad (1)$$

in the collector current. Therefore, the current in D6 increases by an amount equal to

$$\Delta i_{D6} = \left[\frac{\alpha}{n_1} - 1 \right] n_2 \Delta i_{out} \quad (5)$$

This variation in the degree of stability may be reduced by selecting α/n_1 close to unity and reducing n_2 . However, since it is desirable to have α/n_1 much larger than unity for short rise time and since any reduction in n_2 increases the I_{co} standby power,† a compromise is necessary.

6. SYNCHRONIZING CIRCUITS

The majority of modern digital data processing systems employ coincidence gate circuits to perform the logical functions. In order to insure that digit pulses will coincide at the inputs to the logic circuits, it is convenient to synchronize the amplifiers. Usually a master oscillator, or "clock," produces the timing signals that are distributed to the amplifiers. The function of the synchronizing circuit in the amplifier is to turn on and to turn off the amplifier at predetermined time intervals in response to the clock signal.

In a regenerative amplifier there is always a small delay from the time triggering commences until the full output pulse is developed. Then there are variations in the transmission time to other amplifiers. For these reasons the clock signal must lag the input signal to the amplifier in order to maintain control of turn-on and to obtain a uniform pulse length from all amplifiers. Generally the time lag is one-fourth of the

* In an actual amplifier D6 is not required if the transistor saturation voltage is relatively constant versus collector current and the pulse fall time is not adversely affected by minority carrier storage in the transistor. Often the inductive "kick" of the transformers and the regenerative feedback are sufficient to make the minority carrier storage effect negligible. If D6 is used, its reverse recovery time may adversely affect the pulse fall time, thus nullifying its usefulness.

† The I_{co} standby power is proportional to V_2 , which, for a given output voltage, is inversely proportional to n_2 .

repetition period and, in such a case, the clock signal is made available in four phases.

Although the clock signal may have any one of a number of forms, a sine or a square wave are the most common forms. Usually a sine wave is preferred because it is simpler to distribute to a large number of amplifiers. Exceptions occur in cases where exceptionally precise timing is necessary, or the use of a square wave requires considerably less clock power. In the following discussion of where to synchronize, a square wave will do as well or better than the assumed sine wave. With either signal it is desirable to keep the clock power to a minimum.

If the synchronizing circuit is to be effective, the clock signal must be capable of accomplishing the following actions:

- a. It must be able to hold the transistor in the "off" state in the presence of trigger current in order to control turn-on.
- b. At the turn-on time it must rapidly inject the trigger current into the transistor.
- c. At the turn-off time it must alter the conditions in the feedback loop in such a manner that the transistor turns off promptly.

In other words the synchronizing circuit must act like an inhibit logic circuit with the clock signal appearing as the inhibit signal during the interdigital period.

It is recognized that there are many amplifier configurations and several ways to synchronize each configuration. Generally it is preferable to synchronize at only one input terminal of the transistor or at only one point in the feedback circuit. A relatively complete discussion can be given with the aid of the following four examples.

A circuit that employs negative resistance feedback, such as in Fig. 4, requires a relatively large amount of clock power for synchronization. Because a capacitor (C_2) is required on the emitter for regeneration,² the clock signal must be applied to the base of the transistor to control turn-on accurately. As far as turn-off is concerned, another clock signal might be applied to the emitter or to the current gate in the feedback circuit. However, this would result in additional components, a second clock signal 180° out of phase with the base clock signal, and approximately the same required clock power as if the base clock signal alone were used. Turn-off at the emitter is impractical due to the negative resistance characteristic. The power that the clock signal on the base must furnish is made up of two parts. One part is the average standby power that is absorbed every time the clock voltage is positive. It is composed of the I_{ce} power supplied to the transistor plus the power dissipated in R_1 and R_2 . R_2 and D_2 serve to reduce the clock current in R_1 and

R₂, but the maximum value of R₂ is limited by stray capacitance from the transistor base to ground.* The average clock standby power for this circuit (with a 10 volt peak clock voltage) is approximately 13 milliwatts. The second part of the clock power occurs at turn-off when the clock must supply approximately the full "on" state collector current. In this design the clock supplies about 20 milliamperes of current for 0.1 microsecond at voltages up to about 6 volts peak before the transistor turns off. Therefore, a negative resistance feedback circuit usually requires a relatively large amount of standby clock power continuously and a high peak clock power at turn-off. Also it should be noted that diode D₁ must have a short reverse recovery time in order to prevent false triggering during the negative portion of the clock cycle.

A second example of synchronization is shown in Fig. 11. Here the clock signal is introduced in the feedback circuit to control turn-off. It is also applied to R₂ in the input circuit so as to control turn-on. In this circuit most of the clock power is dissipated in R₅ and R₆ when the clock voltage is positive during the output pulse time slot (whether or not an output pulse is produced). Necessarily, this power is relatively large because the clock must supply the full amount of feedback current. Also, it is necessary to clip the positive peak of the clock voltage in order to prevent false triggering via R₂ when there is no input pulse. A square wave clock signal would eliminate the need for R₆ and D₇, but would not change the power in R₅. The average clock power in a typical circuit of this type is approximately 20 milliwatts, which is relatively large. The principal advantage of this method is that diode reverse recovery time is not a problem.

A third method of synchronization is to apply a square wave clock signal (a sine wave is not suitable in this case) between the base of the transistor and ground (for example, assume in Fig. 11 that R₂ and R₅ are returned directly to V₆ and that the base of the transistor is the clock terminal instead of ground). Before turn-on the clock voltage must be more positive than the trigger voltage on the emitter. At turn-on the clock voltage drops rapidly to ground potential and triggering takes place. During the pulse duration the base current of the transistor is supplied by the clock source. At turn-off the clock voltage must rise rapidly several volts until D₆ conducts and robs current from the feedback loop. The clock power required by this method is relatively large (order of 20 milliwatts) for point contact transistors because the base current of such units is large. In a junction transistor with alpha close

* The capacitance causes the base voltage to lag the clock voltage at turn-on if R₂ is large, which degrades the timing.

to unity the base current is small and the required clock power may be as low as 3 milliwatts. However, it should be noted that this method of synchronization applies only to amplifiers with a gated feedback circuit (such as R5 and D6 in Fig. 11). In other circuits (Fig. 12, for example), a clock voltage applied to the base terminal of the transistor may never be able to turn off the transistor (the feedback current may actually increase instead of decrease). Thus, this method of synchronization is limited and is a low power method only when used with junction transistors.

A fourth synchronization method, which avoids the limitations cited in the previous examples, is illustrated in Fig. 12. The timing circuit is simply diode D1. The operation of the circuit, which is like an inhibit logic circuit, is as follows. When trigger current commences, the clock voltage is negative and D1 conducts the trigger current away from the emitter terminal. As the clock voltage rises positiveward, the emitter voltage follows until it reaches the threshold voltage of the transistor, usually ground potential. Then the trigger current flows into the transistor which turns on. As the clock voltage continues positiveward the emitter conduction clamps the emitter voltage so that D1 opens and the clock does not shunt the feedback path during the pulse duration. At the end of the pulse duration the clock voltage goes negativeward through ground potential and D1 becomes conducting. This action robs current from the feedback loop, thus causing the transistor to turn off. If no input pulse is present, D1 is always non-conducting and any small reverse leakage current is drained off through R1 (which is returned to voltage V1).

Because diode D1 is always non-conducting when no input pulse is present, the standby clock power is essentially zero. During a pulsing cycle the clock conducts only a small current before turn-on and only instantaneously at a low voltage at turn-off. Hence, the required clock power is usually less than two milliwatts.

It is important to note that the amplitude of the negative peak of the clock voltage usually should not be more negative than the quiescent bias voltage on the emitter. If it should be, D1 will conduct and, due to minority carrier storage, may cause false triggering when the clock voltage goes positive. The current through D1 at turn-off might have the same effect in the succeeding cycle except that the flyback voltage of the transformers during the interdigit period removes the minority carriers from both D1 and D2. Since D2 carries a larger current for a longer period than D1, the carriers are cleared from D1 first. It is then reverse-biased for almost one-half the repetition period before there is any chance

of false triggering. Hence, diode reverse recovery time is not a problem. However, D1 should have a short forward recovery time in order that turn-off will occur rapidly.

One possible limitation of this synchronization method is that a low impedance clock source is necessary. This is usually not difficult to obtain with a resonant circuit in the output of the clock signal source. Offsetting this point are the advantages of low clock power, essentially zero standby clock power, only one additional component, and no critical component tolerances.

7. ILLUSTRATIVE DESIGN

In the preceding sections the features of various configurations for the functional circuits of an amplifier have been described. The following discussion illustrates the application of these ideas to an amplifier design for use in a digital computer system. It is intended that the description of the design philosophy be sufficient to permit its application to other systems.

In the computer under consideration the amplifier is to be combined with a single level, diode logic circuit to form a logic network. The logic networks, together with delay lines, will be connected in appropriate arrays to perform the logic functions of the system, such as addition, multiplication, etc. Digital information is to be represented by one-half microsecond pulses and the amplifiers are to be synchronized at a one megacycle pulse repetition rate by a four phase sine wave master oscillator. Other system requirements are mentioned in connection with the selection of the corresponding functional circuit.

Since the amplifier is considered as a small system of functional circuits, it is necessary, as in most system designs, to re-examine, and possibly change, circuit choices as the design progresses. However, for the sake of clarity, the following discussion omits the re-examination and frequently refers to the final schematic shown in Fig. 13.

The first step in the design is to select the feedback configuration most suitable to the computer requirements. For this computer the dc and clock power are to be minimized and the amplifier should be able to drive from 1 to 12 logic networks. Miniaturization of the computer implies that there may be an appreciable amount of stray capacity across the amplifier output. These considerations suggest transformer coupled feedback connected in series with an output circuit. Since both positive and negative output pulses are to be required (one polarity for AND and OR logic and the other polarity for inhibition), transformer output coupling is indicated.

The next basic selection is the choice of an appropriate transistor. In this computer it is expected that pulses will occur in only about one third or less of the pulse time slots due to the nature of the digital information. In order to minimize the dc standby power an alloy junction transistor is a logical choice for this application because of the low I_{co} current. However, even with a junction unit possessing an alpha cut-off frequency of eight megacycles, it is difficult if not impossible to obtain acceptable gain and rise time with the desired output load current at a one megacycle repetition rate. If the rise time is improved by increasing the trigger current, the gain is decreased. The principal cause of the poor "gain-bandwidth" appears to be the depletion layer capacitance.¹¹ The difficulty can be overcome by selecting a point contact transistor. A particular germanium transistor coded GA-52996* appears to be suitable and has the following pertinent characteristics:

- Collector capacitance less than 0.5 uuf.
- Alpha cut-off frequency in excess of 80 mc.
- Base resistance less than 100 ohms.

Since the alpha of this unit is greater than 2 at collector currents of the order of 10 ma, the common base connection will yield the greatest current gain. The disadvantage of a point contact unit, of course, is the I_{co} current. For this reason the amplifier will have to be designed to use the smallest possible collector supply voltage.

The point contact transistor, due to its high cut-off frequency relative to the amplifier pulse repetition rate and its high alpha at small emitter

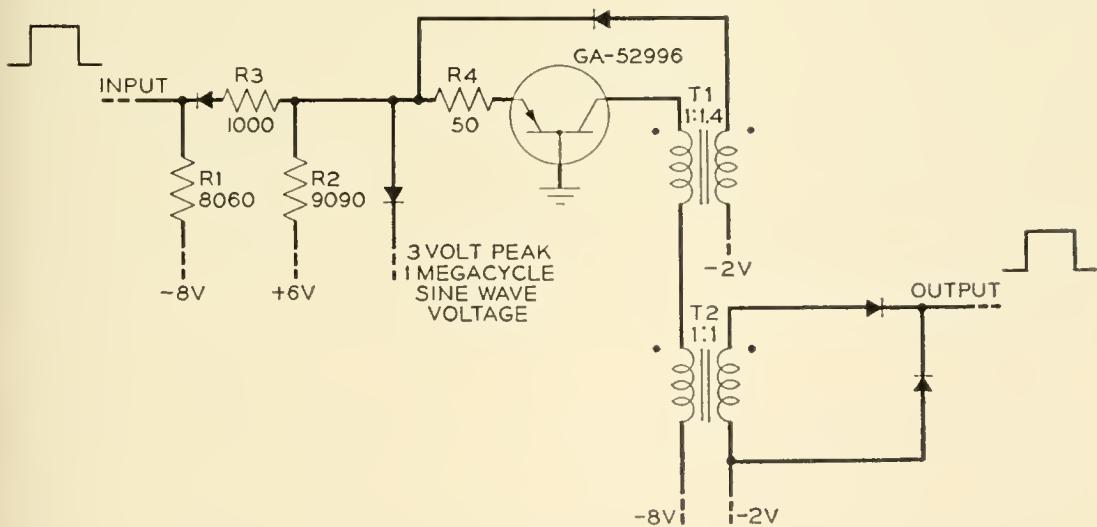


Fig. 13 — Illustrative design

* This is a relatively special unit especially suited for high speed switching applications.

currents,* permits the use of a simple input circuit. The AND type input circuit is suitable and desirable for another reason. When AND type logic is added to the amplifier, it may be paralleled with the basic input circuit and the input sensitivity of the complete network will be the same as for the amplifier alone. Other logic circuits will be added to an amplifier in a manner similar to that described by Felker² so that the input sensitivity will be reduced at most by the voltage drop across one series diode (approximately 0.3 volts).

The input pulse voltage and current requirements depend upon the voltage threshold necessary to prevent false operation and the minimum trigger current for reliable regeneration. A test of several sample transistors indicates that approximately 0.3-ma emitter current is required to trigger the transistor with an estimated collector supply voltage of 10 volts. The emitter breakpoint† voltage is found to vary between -0.25 and +0.25 volts. To allow for aging variations of the transistor and of R2, it seems reasonable to use a 6-volt source and R2 equal to 9090 ohms, which results in a trigger current a little more than twice the required minimum. Previous experience with computers of this type indicates that a 2-volt threshold will be sufficient to prevent false triggering. Thus, the secondary winding of the feedback transformer is returned to -2 volts and R1 is chosen to give a quiescent emitter voltage of -2 volts. With these considerations and an estimated voltage drop across R3, the input pulse amplitude is calculated to be 2.3 volts and 0.9 ma. Allowing 0.3 volts for a series logic diode, the minimum output voltage and current of the amplifier are 2.6 volts and 0.9 ma per driven network.

The selection of the collector supply voltage and the turns ratio of T2 depends upon the dc power dissipation due to I_{ce} current and output voltage regulation versus collector current. For this transistor a unity turns ratio appears to represent a reasonable compromise. Then, by estimating the voltage drops across T1, T2, and the transistor, it is found that a collector supply voltage of -8 volts is sufficient to produce an output pulse voltage about 0.5 volt greater than the required minimum.

The next step is the selection of the turns ratio of T1 and the primary inductances of both T1 and T2. The two considerations involved are sufficient feedback with the minimum output current (the worst case with respect to feedback) and the maximum collector dissipation in the event that the clock fails. By means of the formulas and assumptions indicated in section 5, primary inductance values of 0.4 mh for T1 and

* Usually $\alpha > 4$ for $i_e = 0.5$ ma.

† The transition point of the emitter diode from cut-off to conduction.

0.2 mh for T2 together with a turns ratio of 1.4 for T1 are selected. Since the GA-52996 transistor is not quite short circuit staple, a 50-ohm resistor is added in series with the emitter. The excess emitter current at the end of the pulse duration is greater than 2 ma, thus assuring sufficient stability, and, if the clock fails, the amplifier will turn off by itself in approximately 7 μ sec, at which time the instantaneous collector dissipation will be approximately 240 mw (considered to be a safe instantaneous dissipation for this transistor).

For low clock power and circuit simplicity the single diode synchronizing circuit is chosen. Although a peak clock voltage of 2 volts would normally be used (this value corresponds to the quiescent emitter bias voltage) it is found that the clock may be varied between 1 volt and 6 volts peak without a failure occurring. Therefore, the nominal clock voltage is set at a centered value of 3 volts peak. The dc level of the clock voltage is 0 volts, which approximately corresponds to the emitter break point voltage of the transistor. This concludes the basic selections in the design procedure.

The power dissipated in the amplifier is quite modest. In the quiescent state the amplifier absorbs only 0.2 mw average clock power and 30 mw dc power (this would be only 10 mw if the I_{ce} power were negligible). When the amplifier is pulsing every microsecond the dc power is 50 mw and the average clock power is 2 mw. Since the amplifier is so conservative of power, it is possible to use 4,000 networks in a computer and require less than 200 watts dc power.

One indication of the component sensitivity of a pulse amplifier is the magnitude of the supply voltage margins. In this amplifier the supply voltages may be varied, one at a time, over ± 12 per cent of the nominal values before a failure occurs. Generally margins of this magnitude under the worst conditions are considered sufficient to guarantee against failures caused by aging, or to insure that such failures will be indicated by routine checks before they occur. It is interesting to note that in a temperature test the amplifier continued to operate properly over a temperature range from -20 to $+80^\circ\text{C}$. Even at $+75^\circ\text{C}$ the supply voltage margins were 10 per cent or better.

8. SUMMARY

A method of analysis and design procedure have been presented in which a transistor regenerative amplifier is considered as an interconnected system of functional circuits. Each functional circuit may be evaluated or chosen in terms of the requirements of the complete digital system in which the amplifier is to be used. In general no particular cir-

cuit or collection of circuits can result in an amplifier suitable for use in every type of digital system. The use of an AND type input circuit, transformer coupled output and feedback circuits, and an inhibit type synchronizing circuit appear to be an optimum set of functional circuits to make up an amplifier for use in a synchronous digital computer system employing passive logic circuits. An illustrative design is presented for such an amplifier which operates at a pulse repetition rate of 1 mc, uses 12 components (none of which are especially critical), requires an average of 40-mw dc power and 1-mw clock power, is capable of driving from 1 to 12 similar amplifiers, and has voltage margins in excess of 12 per cent. Although the design philosophy was developed for this type of amplifier, it is believed that much of the philosophy is applicable to regenerative amplifiers for use in other digital data processing systems.

9. ACKNOWLEDGEMENT

The final design and the performance data of the illustrative amplifier are due to L. C. Thomas and H. E. Coonce. The author also wishes to express his appreciation for the many helpful and stimulating discussions with other colleagues, especially A. J. Grossman, T. R. Finch, J. H. Felker, and J. R. Harris.

REFERENCES

1. S. Greenwald, et al., SEAC, Proc. I.R.E., Oct., 1953.
2. J. H. Felker, Regenerative Amplifier for Digital Computer Applications, Proc. I.R.E., Nov., 1952.
3. J. L. Moll, Large-Signal Transient Response of Junction Transistors, Proc. I.R.E., Dec., 1954.
4. J. G. Linvill and R. H. Mattson, Junction Transistor Blocking Oscillators, Proc. I.R.E., Nov., 1955.
5. A. E. Anderson, Transistors in Switching Circuits, B.S.T.J., Nov., 1952.
6. T. E. Firle, et. al., Recovery Time Measurements on Point-Contact Germanium Diodes, Proc. I.R.E., May, 1955.
7. S. L. Miller and J. J. Ebers, Alloyed Junction Avalanche Transistors, B.S.T.J., Sept., 1955.
8. J. J. Ebers and S. L. Miller, Design of Alloyed Junction Germanium Transistors for High Speed Switching, B.S.T.J., July, 1955.
9. T. C. Chen, Diode Coincidence and Mixing Circuits in Digital Computation, Proc. I.R.E., May, 1950.
10. L. W. Hussey, Semiconductor Diode Gates, B.S.T.J., Sept., 1953.
11. J. M. Early, Design Theory of Junction Transistors, B.S.T.J., Nov., 1953.
12. Q. W. Simkins and J. H. Vogelsong, Transistor Amplifiers for Use in a Digital Computer, Proc. I.R.E., Jan., 1956.
13. M. Tanenbaum and D. E. Thomas, Diffused Emitter and Base Silicon Transistors, B.S.T.J., Jan., 1956.

Observed 5–6 mm Attenuation for the Circular Electric Wave in Small and Medium-Sized Pipes

By A. P. KING

(Manuscript received March 20, 1956)

At frequencies in the 50–60 kmc region the use of circular electric wave transmission can provide lower transmission losses than the dominant mode, even in relatively small pipes.

The performance of two sizes of waveguide was investigated. In the small size ($\frac{1}{16}$ " I.D. $\times \frac{1}{16}$ " wall) the measured TE₀₁ attenuation was approximately 5 db/100 ft and is appreciably less than that of the dominant mode. The measured attenuation for the medium sized ($\frac{7}{8}$ " I.D. $\times \frac{1}{8}$ " wall) waveguide was 0.5 db/100 ft which is about one-fourth that for the dominant mode.

This paper also considers briefly some of the spurious mode conversion-reconversion effects over the transmission band and their reduction when spurious mode filters are distributed along the line. Allowance has been made for the added losses due to oxygen absorption when air is present.

INTRODUCTION

Since 5.4-mm dominant-mode rectangular waveguide has attenuations of the order of 60 db/100 ft, another transmission technique is required in applications which involve appreciable line lengths. Losses may be reduced by the use of oversize waveguide; some earlier work with dominant mode transmission in slightly oversize round waveguide (two or three propagating modes) has been reported.¹ The possibility of still lower losses exists with circular electric wave transmission in an oversize round waveguide. Miller and Beck² have computed the theoretical relative transmission losses of the TE₀₁ and TE₁₁ modes as functions of

¹ A. P. King, Dominant Wave Transmission Characteristics of a Multimode Round Waveguide, Proc. I.R.E., **40**, pp 966–969, Aug., 1952.

² S. E. Miller and A. C. Beck, Low Loss Waveguide Transmission, Proc. I.R.E., **41**, pp 348–358, March, 1953.

guide size and frequency. At 5.4 mm, a $\frac{7}{16}$ " I.D. waveguide has an appreciably lower attenuation with the circular electric mode than with the dominant mode. A $\frac{7}{8}$ " I.D. guide has a circular electric attenuation approximately one-fourth that of the dominant mode in the same pipe.

It is the purpose of this paper to present some experimental results which have been observed with circular electric wave transmission in the 5-6 mm wavelength region. The attenuation for three different lines and the transmission variations due to moding effects are reported. Allowance for the loss due to oxygen absorption has been included.

DESCRIPTION OF THE TEST LINES

The TE_{01} mode attenuation measurements were made on approximately straight runs of line ranging from about 100 to 200 feet in length. The copper pipe comprising these lines is believed to conform to the best tolerances and internal smoothness which are current manufacturing practice for waveguide tubing. The relative tolerances and their effect upon transmission are considered in a later section. Three kinds of copper line were measured: a waveguide of oxygen-free copper, one line of low phosphorous-deoxidized copper and one line of steel with a 20-mil low phosphorous-deoxidized copper inner lining. The oxygen-free high-conductivity-copper with its higher conductivity and somewhat greater ductility was chosen to provide comparative performance data with the low phosphorous-deoxidized copper which is commonly used in waveguide manufacture. A waveguide whose outer wall is constructed of steel to provide the necessary strength and wall thickness to support a very thin copper inner wall has the advantage that such waveguide requires less copper. This composite wall tubing was obtained to ascertain whether the tolerances and the nature of the inner surface would yield transmission data comparable to solid copper waveguide.

The lines were supported on brackets which were accurately aligned and spaced at 6-ft intervals. Although the brackets provided for an accurately straight line, the manufactured pipe was not perfectly straight but, in some samples, varied as much as $\frac{3}{8}$ " in a 12-ft length. Installing the pipe on the brackets tended to straighten the line and reduce these variations to about half this amount. A general view of the lines is shown in the photograph of Fig. 1.

The sections of waveguide were joined together with a more or less conventional threaded coupling, but with one very important difference. The threads, which are cut at the ends of each section, are cut relative to center of the inside diameter and not the outside diameter. This is achieved by employing a precision pilot to provide a center for the cut-

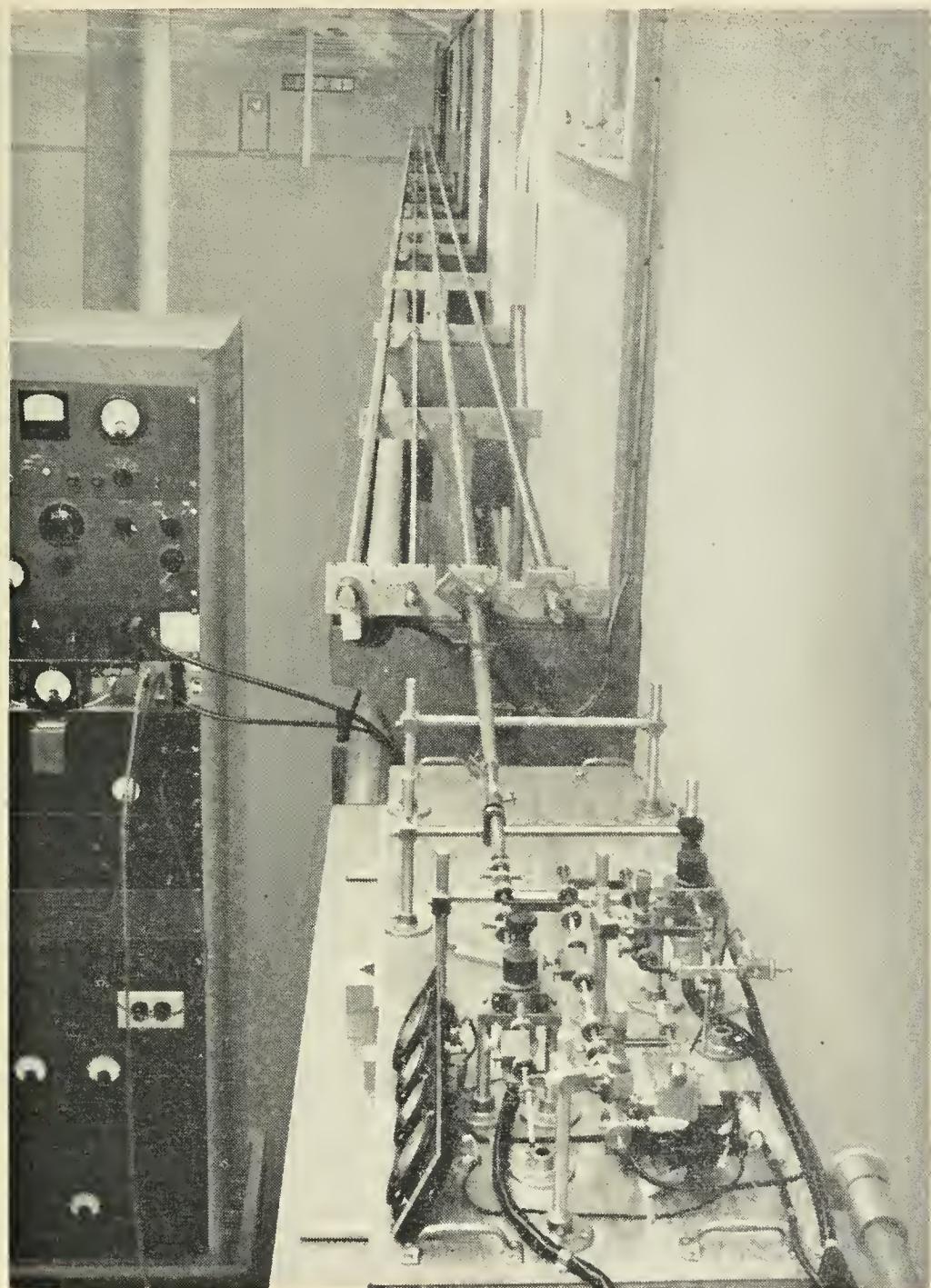


Fig. 1 — General view of the circular waveguide lines and the millimeter wave measuring equipment.

ting die. Since the internal diameter is made as precise as possible, the variations of outside diameter become a function of the tolerances of both the internal diameter and wall thickness and cannot be as precise as the inside of the pipe. Any thread cut relative to the outside diameter as in regular plumbing practice, will not, in general, be concentric to the

inside wall. To avoid an offset at the joint it is therefore important that the thread be centered relative to the inside diameter. After a section was threaded the ends were faced off to make the ends square and thus avoid any tilt between sections when the ends are butted together.

Of the two sizes tested the smaller diameter ($\frac{7}{16}$ " I.D. $\times \frac{1}{16}$ " wall) was chosen to provide a moderate line loss, while limiting the number of propagating modes. In the band of interest (5.2–5.7 mm) the theoretical TE_{01} wave attenuation is about 4 db/100 ft. The number of modes which can be supported at $\lambda = 5.2$ mm is limited to 12 modes and to only one of the circular electric modes. The higher order TE_{on} modes are beyond cut-off. These features limit the number of spurious modes and simplify the mode filtering problem. Furthermore, in this smaller sized waveguide, the associated components which may set up TE_{cn} waves, for example conical tapers, need not be as long proportionately as in larger waveguides. The $\frac{7}{16}$ " I.D. guide has the advantage of smaller size, lower cost and greater ease of transmitting TE_{01} through specially constructed bends. The attenuation of this smaller diameter guide is large enough that system requirements will usually restrict its usage to lengths of line of a hundred feet or so.

The larger size ($\frac{7}{8}$ " I.D. $\times \frac{1}{8}$ " wall) is exactly twice the diameter of the small size discussed in the preceding paragraph but has only one-tenth the attenuation, or about 0.4 db/100 ft. The low loss of this larger size becomes more attractive for runs as long as several hundred feet. This diameter guide will, of course, support more modes, 50 at $\lambda = 5.2$ mm; four of which are circular electric modes — TE_{01} , TE_{02} , TE_{03} and TE_{04} . Some of the disadvantages which accompany the increased diameter are: (1) greater care must be taken as to line straightness, (2) longer conical tapers are required when converting from one guide diameter to another, and (3) longer mode filters are required since the desired mode-filtering attenuations vary inversely with the filter diameter at a given frequency. Flexible spaced-disk lines employed as uniform bends for TE_{01} transmission require much greater bending radii than bends in the smaller diameter guide if the bend loss is to be kept proportionately low. This problem is considered in some detail in another paper.³ With reasonable care the accumulative effect of these foregoing factors can be held to a reasonably low value. Expressed in terms of the ratio of measured to theoretical attenuation the values are, on the average, about 10 per cent higher in the $\frac{7}{8}$ " I.D. waveguide than in the $\frac{7}{16}$ " I.D. waveguide.

³ A. P. King, forthcoming paper on bends.

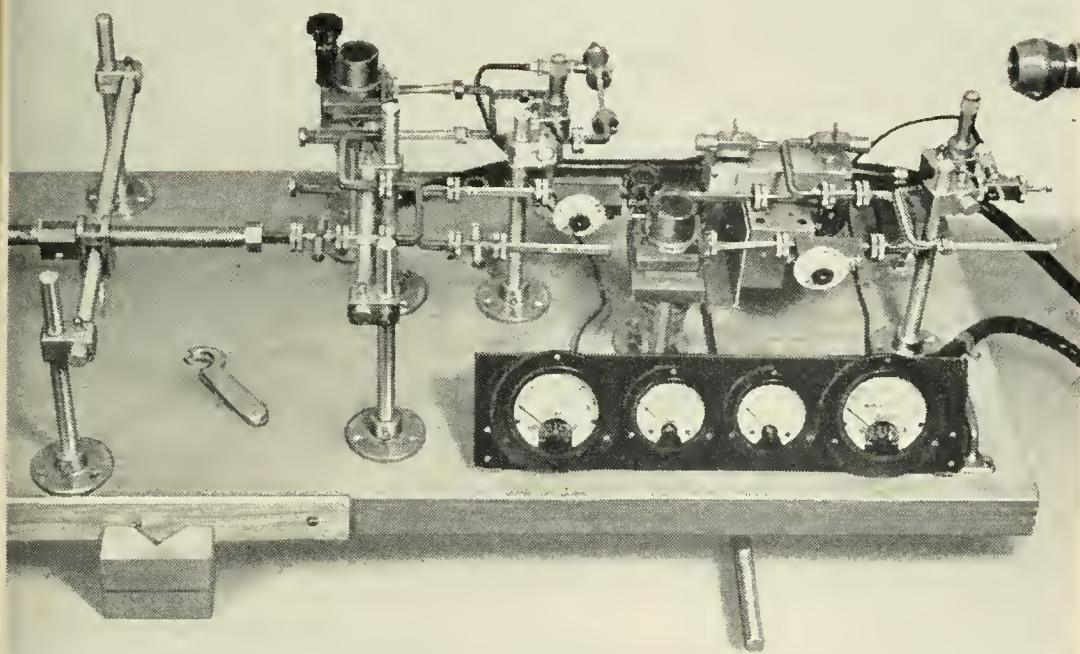


Fig. 2 — Waveguide portion of millimeter wave measuring set.

MEASURING PROCEDURE

With straight runs of round, TE_{01} waveguide lines whose length lies in the 100–200 ft range, it is convenient to make attenuation measurements on a round trip basis. This method has the advantage of convenience in that the attenuation can be measured directly by using a waveguide switch but has the disadvantage of requiring a careful impedance match of the measuring equipment to the line. Fig. 1 shows an overall view of the lines; Fig. 2 shows the arrangement of the 5–6 mm measuring set, and Fig. 3 shows a block diagram of the set-up employed.

This measuring set makes use of two klystrons developed by these laboratories.⁴ The double detection receiver features a separate beating oscillator klystron which is frequency modulated and a narrow band (1.7 mc at 60 mc) IF amplifier. The resulting IF pulses are detected with a peak detector and then amplified to provide the usual meter indication. This method with its circuitry has been developed by W. C. Jakes and D. H. Ring,⁵ and provides a greater amplitude stability than is possible with a cw beating oscillator.

In the waveguide schematic of Fig. 3 about a tenth of the power is

⁴ E. D. Reed, A Tunable, Low Voltage Reflex Klystron for Operation in the 50-60 Kmc Band, B.S.T.J., **34**, p. 563, May 1955.

⁵ W. C. Jakes and D. H. Ring, unpublished work.

taken from the signal oscillator to provide monitoring and wavemeter indication. The remaining power, after suitable padding, is fed into a 3-db directional coupler or hybrid junction 2. This junction is employed as a waveguide bridge so that, when arms A and B are properly terminated, no power flows in receiving arm C. Any reflection in line A will,

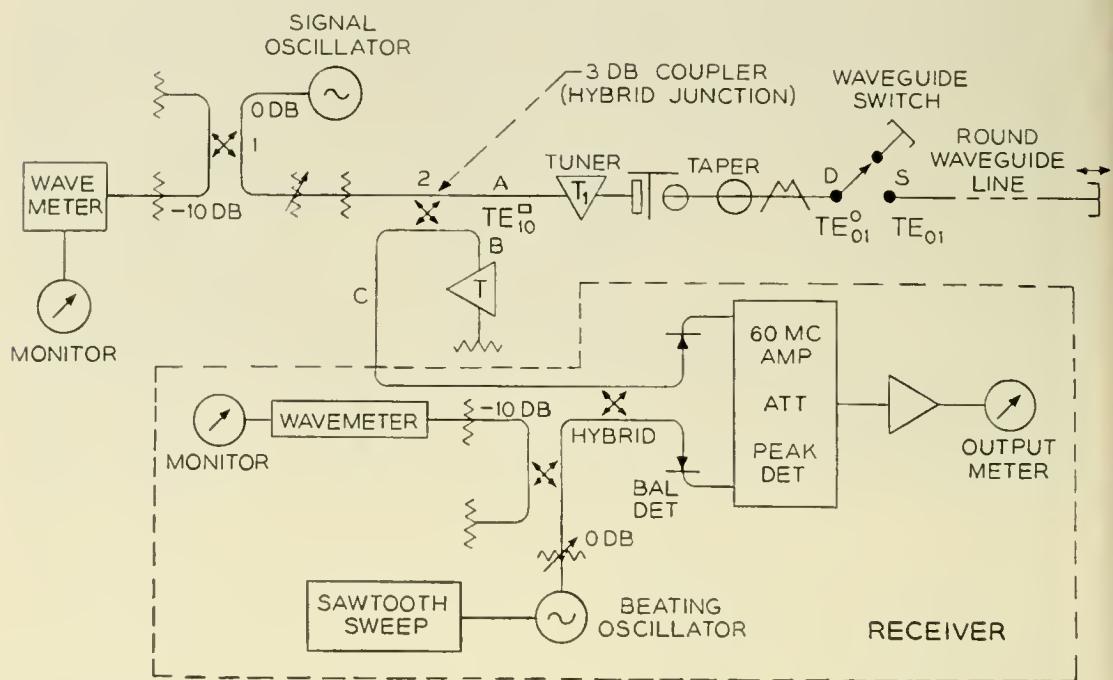


Fig. 3 — Schematic of measuring equipment.

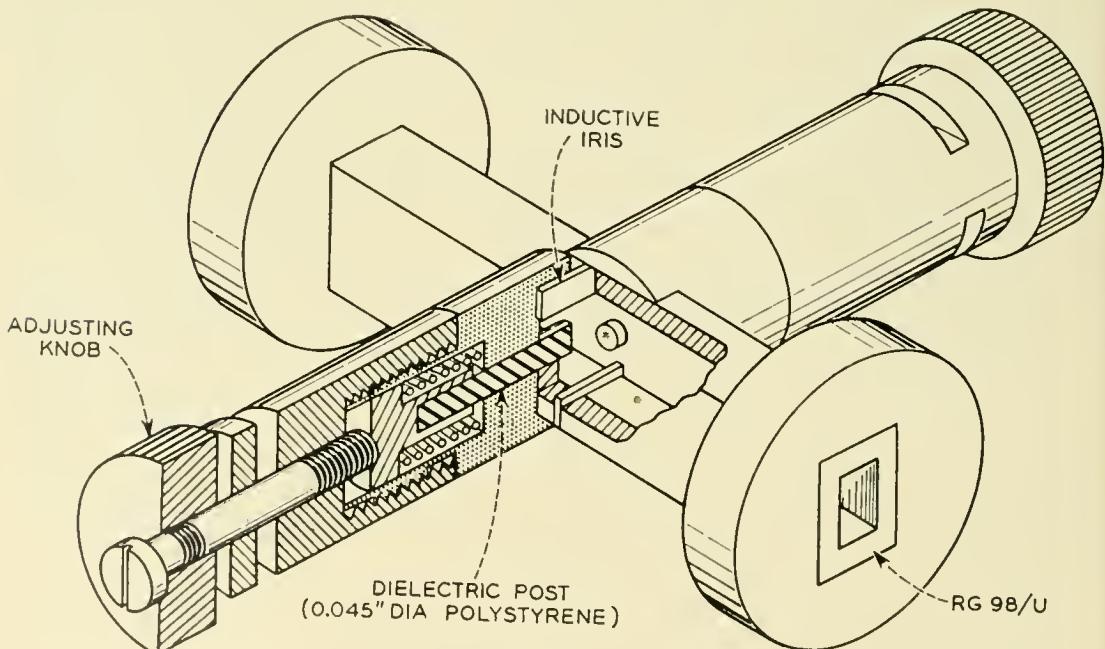


Fig. 4 — Structure of impedance matching tuner.

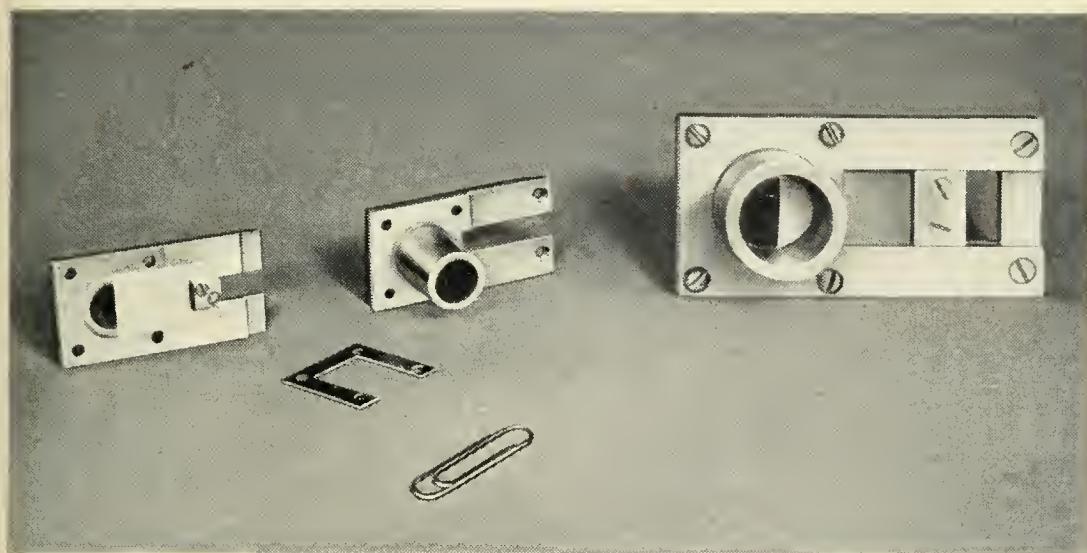


Fig. 5 — Structure of waveguide switch.

however, produce a power flow in the arm C to the balanced converter of the receiver and an indication in the output meter. So far this set is similar to a setup for measuring the round trip loss in a terminated waveguide system. The impedance of the $\text{TE}_{10}^{\square} \Leftrightarrow \text{TE}_{01}^{\circ}$ wave transducer,⁶ taper section and mode filter connected as shown in the section A-D of Fig. 3 can be matched to the rectangular waveguide at A by an appropriate adjustment of the dielectric post tuner⁷ T_1 whose structure is shown in Fig. 4. Under these conditions a conical taper termination placed in the round waveguide at D will again produce a balance and again no power will flow in arm C. A waveguide switch whose structure is shown in Fig. 5 is connected between the point D and the line under test. A movable short at the far end of the line completes the set-up.

With the impedances matched as described above, the only reflection which reaches the receiver will be from the far end of the line when the switch S is open or, when shorted, from the switch itself. The round-trip attenuation is the difference in attenuation measured for the two positions of the switch. By means of a movable short at the far end of the line, the line length can be varied to produce mode conversion and mode reconversion effects, and the resultant variation in TE_{01} mode transmission can be observed. This phenomena is described in some detail elsewhere.⁸

⁶ Reference 2, page 354, Fig. 14.

⁷ C. F. Edwards, U.S. Patent 2,563,591, Aug. 7, 1951. The millimeter tuner employs an adjustable dielectric post in place of a metallic tuning screw described in the patent.

⁸ Reference 2, pp 356, 357.

LOSSES DUE TO OXYGEN ABSORPTION

In addition to the losses which result from imperfect conductivity, surface effects, and mode conversions, there is a very appreciable loss due to oxygen absorption when the guide is open to the atmosphere. In a waveguide the loss due to O_2 absorption is:

$$\frac{A}{\sqrt{1 - \nu^2}} \quad (1)$$

where

A is the absorption due to oxygen in the atmosphere

$\nu = \lambda/\lambda_c$

λ = free space wavelength

$\lambda_c = \frac{\pi d}{k} = \frac{\pi d}{3.83}$ = cut-off wavelength

d = internal diameter of waveguide

k = Bessel root for TE_{01} mode = 3.832

The loss due to absorption of oxygen which is present in the atmosphere (at approximately sea level) was obtained from the experimental data of D. C. Hogg.⁹ The added loss produced by the presence

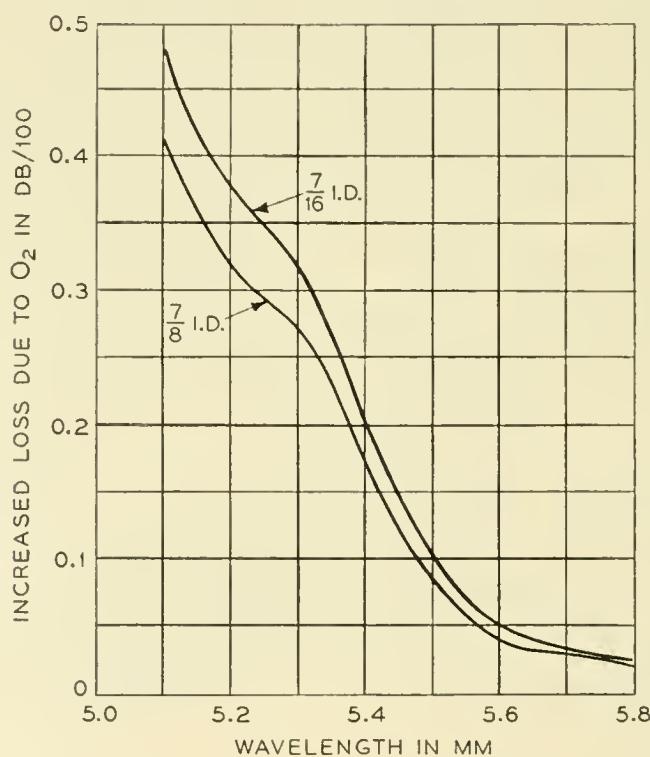


Fig. 6 — TE_{01} transmission loss in waveguides due to oxygen absorption.

⁹ A. B. Crawford and D. C. Hogg, Measurement of Atmospheric Attenuation at Millimeter Wavelengths, B.S.T.J., 35, pp. 907-917, July, 1956.

of oxygen in the waveguide in terms of (1) is plotted in Fig. 6. It will be noted that this loss becomes very appreciable at the short wavelength end of the band. At $\lambda = 5.2$ mm this loss is in the 0.3–0.4 db/100 ft range. For the larger size waveguide line ($7/8"$ I.D.) the loss due to O_2 is approximately equal to the theoretical wall losses; for the smaller size lines this amounts to about a tenth the wall loss. At the other end of the millimeter band the O_2 losses are very small, being in the 0.02 – 0.03 db/100 ft range at $\lambda = 5.7$ mm.

The relative effects of theoretical wall and expected oxygen absorption losses are shown plotted in Fig. 7. For the two sizes of waveguide the upper dashed curve represents the combined effect of these two factors and the lower solid line curve is the theoretical attenuation of the TE_{01} mode in empty pipe. The shaded area indicates the increase which is the result of oxygen absorption.

In order to minimize the transmission losses in any practical system it becomes desirable to exclude the presence of oxygen from the line, for example, by introducing an atmosphere of dry nitrogen. Since the ex-

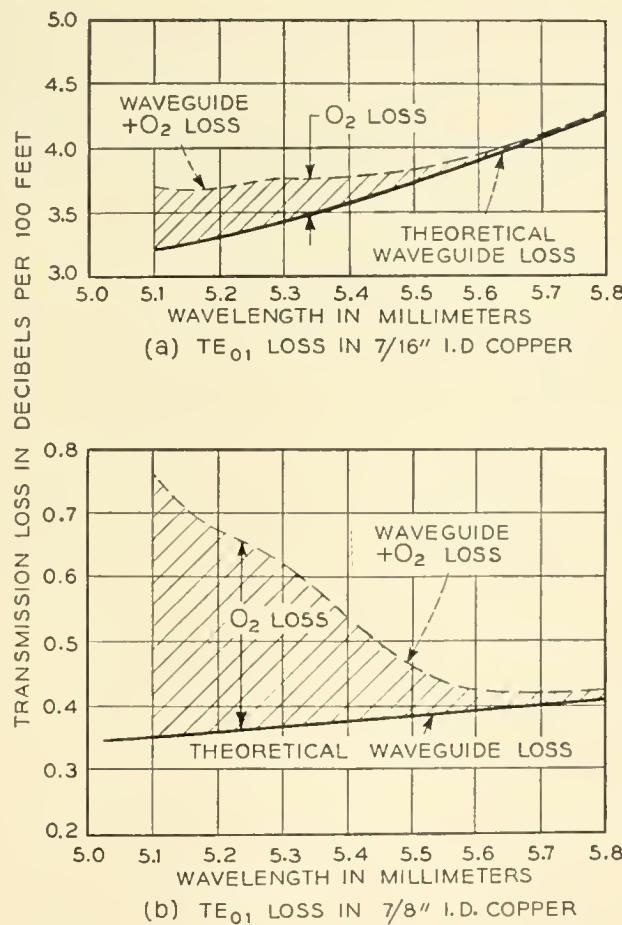


Fig. 7 — TE_{01} transmission losses.

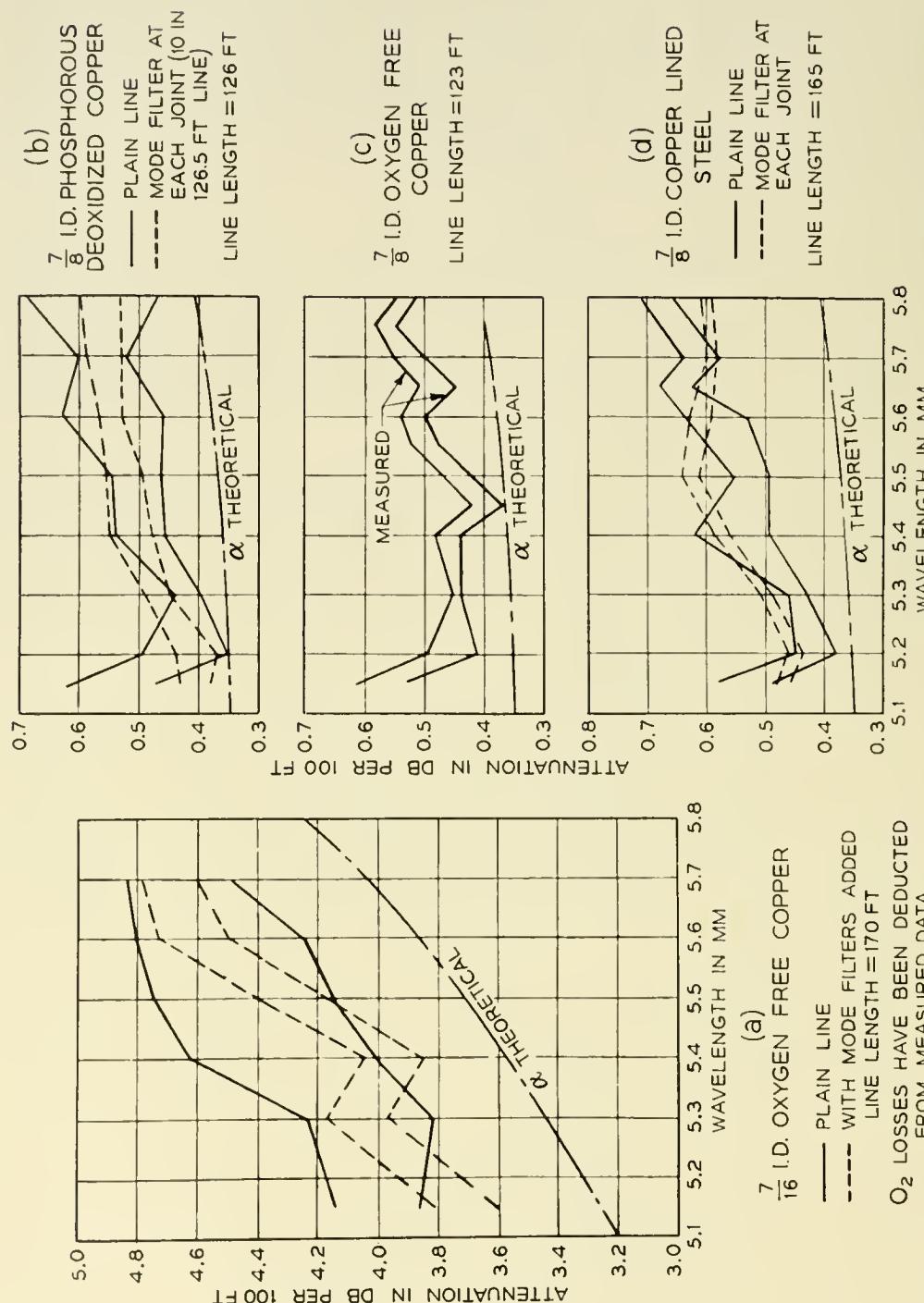


Fig. 8 — Measured and theoretical loss of the four lines studied.

clusion of oxygen was not very feasible in the experimental TE₀₁ lines, the effects due to oxygen absorption were included in the measurements. However, in order to simplify the presentation of the attenuation data these absorption losses, as indicated in Figs. 6 and 7, have been subtracted from the measured data.

The measured attenuation of the four lines are shown in Fig. 8 as a function of wavelength (5.1–5.8 mm). In each case the dash-dot-dash lines represent the theoretical attenuation for copper. Each plot shows two solid lines which indicate the range of values measured over the mm band. The same range was observed either by varying the length of the line by means of a sliding piston at the far end of the line or by imposing a sweep voltage on the repeller of the signal klystron to produce a small frequency modulation. These variations in attenuation correspond to piston movements which are greater than a half wavelength and are due to the mode interference effects produced by spurious modes generated in the line. The resultant signal fluctuations which are due to mode conversion and reconversion effects have been described in considerable detail by Miller.¹⁰

Referring again to Fig. 8, the measured data shown by the solid lines, which are for a plain line without mode filters, indicates that the oxygen-free high conductivity copper line gave the lowest measured average attenuation as well as the least variation. The low phosphorous deox-

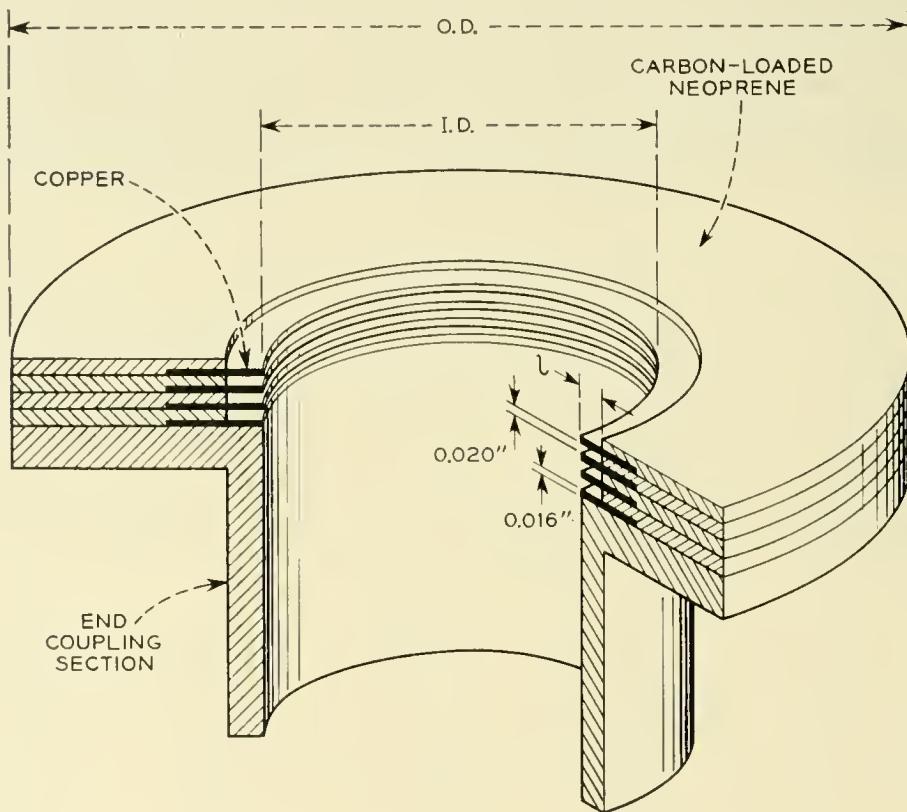
TABLE I

	1/16" I.D.	1/8" I.D.	1/8" I.D.	1/8" I.D.
	OFHC Copper	OFHC Copper	Low Phos. Deoxidized Copper	Copper Lined Steel
Wall Thickness.....	1/16"	1/8"	1/8"	1/8"
α meas. (db/100 ft) ..	4.33 \pm 0.24	0.47 \pm 0.02	0.49 \pm 0.05	0.52 \pm 0.04
α meas.....	1.17	1.29	1.34	1.42
α calc.....				
Average ovality				
A.....	1/1100	1/1100	1/1200	1/585
B.....	0.0004"	0.0008"	0.00075"	0.0015"
Maximum ovality				
A.....	1/730	1/875	1/875	1/290
B.....	0.0006"	0.001"	0.001"	0.003"
Maximum tolerance				
A.....	1/310	1/730	1/430	1/290
B.....	0.0014"	0.0012"	0.002"	0.003"

¹⁰ S. E. Miller, Waveguide as a Communication Medium, B.S.T.J., **33**, pp. 1229–1247, Nov. 1954.

idized copper was next best while the steel line with a 20-mil inner copper lining was the poorest.

In the $\frac{7}{16}$ " I.D. oxygen-free high conductivity copper line the measured attenuation was 17 per cent higher than the calculated value (see $\alpha_{\text{meas}}/\alpha_{\text{calc}}$ in Table I). This higher loss is attributed to spurious mode conversion and to surface conductivity effects. In the $\frac{7}{8}$ " line of the same material the $\alpha_{\text{meas}}/\alpha_{\text{calc}} = 1.29$ which is an increase of 12 per cent relative to the smaller waveguide. Since the $\frac{7}{8}$ " diameter line supports about four times the number of modes of the $\frac{7}{16}$ " diameter line, this increase in loss is attributed to mode conversion. In the other two $\frac{7}{8}$ " diameter guides the added losses are believed to be increased mode conversion which results from the poorer dimensional tolerances. These data are listed in Table I together with dimensional tolerances. In this table α_{meas} is the measured attenuation averaged over the 5.2-5.7 mm band together with the variations shown in Fig. 8; α_{calc} is the average theoretical attenuation for standard (IACS) copper. The I.D. tolerances are listed in two sets of rows A and B; row A gives the fractional variation



TYPE	O. D.	I. D.	<u>l</u>
MEDIUM	$2"$	$\frac{7}{8}"$	$0.080"$
SMALL	$1\frac{1}{2}"$	$\frac{7}{16}"$	$0.081"$

Fig. 9 — Structure of spaced-disk mode filter.

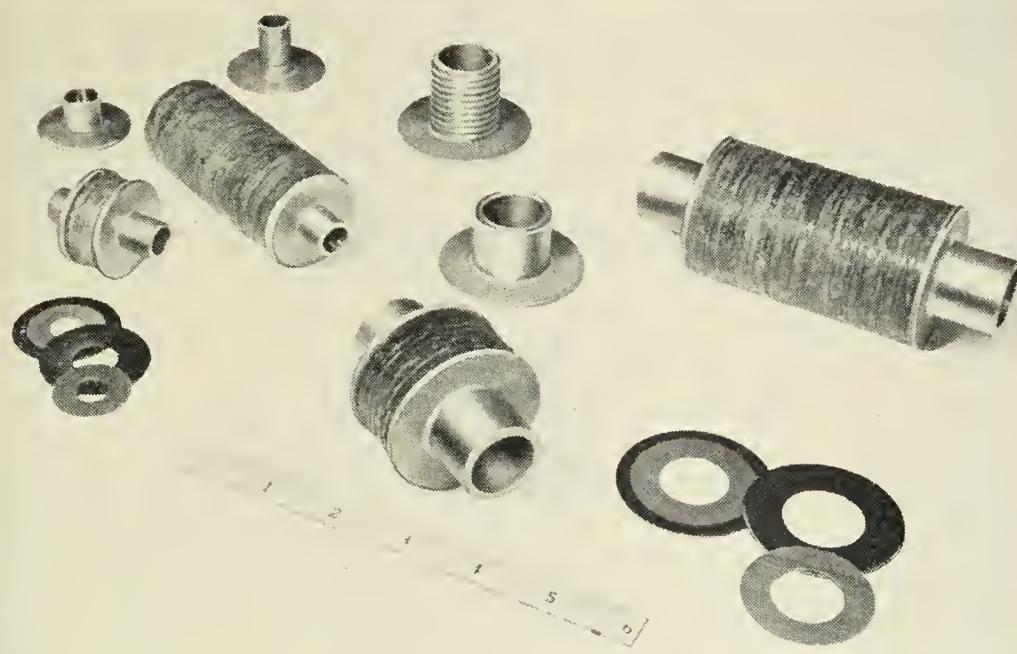


Fig. 10 — Mode filters.

relative to the average diameter and the rows marked B indicate the corresponding variations in inches. The average ovality gives the average difference between maximum and minimum diameters, maximum ovality the maximum difference in diameter and the maximum tolerance gives the maximum difference between diameter and ovality. These measurements have been limited to measuring at the two ends of each section of pipe. In spite of this small sampling the TE_{01} loss measurement appears to follow the I.D. tolerances quite well; the OFHC line shows both the lowest attenuation and the best tolerances.

Mode interference effects can be reduced considerably by increasing the loss to the undesired modes. This effect can be accomplished by modifying the structure so that the spurious modes are highly attenuated while the TE_{01} losses are increased only slightly. One way is to construct

TABLE II — AVERAGE PERFORMANCE OF TE_{01} WAVEGUIDES WITH MODE FILTERS

	$\frac{3}{16}$ " I.D. OFHC copper	$\frac{3}{8}$ " I.D. low phos. deoxidized copper	$\frac{7}{8}$ " I.D. copper lined steel
α measured (average db/100 ft.) . . .	4.24 ± 0.1	0.51 ± 0.025	0.56 ± 0.012
α measured	1.16	1.39	1.52
α calculated			

the waveguide wall with a series of disks which are closely spaced as shown in Fig. 9 and the photograph of Fig. 10. The spacers serve a dual purpose; to hold the disks in alignment and to provide loss for the spurious modes. The circular disks provide the necessary continuity to support the TE₀₁ and TE₀₂ modes and the gaps introduce high resistivity to the longitudinal currents of the other modes. The spaced-disk filters, which were arbitrarily designed to provide a 10 db loss to the TM₁₁ wave, were 1 $\frac{5}{8}$ " and 3 $\frac{1}{4}$ " long for the $\frac{7}{16}$ " and $\frac{7}{8}$ " waveguide sizes, respectively. In the experiments to be described, a mode filter was inserted at each joint of the line, at approximately 12-ft. intervals.

The measured attenuation data with mode filters at each joint of the various lines are indicated by the dashed lines of Fig. 8. As shown the effect of the mode filters is to reduce the TE₀₁ loss variation by a factor of at least two.

The average attenuation is, however, generally somewhat higher than for the unfiltered lines. This higher loss is partly due to spurious mode power which is absorbed by the mode filter and is not reconverted to TE₀₁ power and to a slight degree to the increased TE₀₁ loss introduced by the mode filters. These results are shown in tabular form in Table II, where the nomenclature is the same as in Table I. Because of the excellent performance of the $\frac{7}{8}$ " I.D. line (OFHC copper) by itself no measurements with mode filters were performed on this line.

CONCLUSIONS

The measured data presented above indicate the feasibility of realizing transmission losses as low as 0.5 db/100 ft. with the TE₀₁ mode over distances up to several hundred feet. The transmission variations which occur over the frequency band are a function of the circularity or tolerances of the waveguide. In a particular line the variations can be reduced considerably by adding mode filters along the line. It is reasonable to expect that these variations can be reduced further by adding longer mode filters at the joints or adding more mode filters at shorter intervals along the line. Oxygen must be excluded from the line if the losses are to be a minimum.

ACKNOWLEDGMENT

The author wishes to thank J. W. Bell and W. E. Whitaere for their help in the measurements.

This study was carried out at Holmdel and was sponsored in part by a Joint Service Contract administered by the Office of Naval Research, Contract Nonr-687(00).

Automatic Testing in Telephone Manufacture

By D. T. ROBB

(Manuscript received May 8, 1956)

A general discussion is given on the philosophy behind the development of automatic test facilities and the relationship of this activity to product design and manufacturing engineering. A brief historical discussion of early automatic test machines used by the Western Electric Company leads to a summary of design considerations. These considerations are then illustrated by descriptions of the specific techniques used in three automatic facilities of considerable diversity.

INTRODUCTION

Many of the parts used in the telephone plant are made in such numbers that automatic shop testing of them is desirable. The cost of manual testing by suitable personnel is high, and its nature so repetitive and dull that accuracy suffers. Fortunately, in many cases the complexity of the test requirements has matched the state of the art and the business picture well enough to warrant the development of machine methods. It is our purpose in these articles to review the art as it has evolved in the Manufacturing Division of the Western Electric Company, and to describe some of the techniques. This is done with the hope that improvements or extensions to other testing or manufacturing problems may be suggested.

It should be emphasized that the developments treated here and in the other papers³⁻⁹ have required cooperation among testing and manufacturing engineers in the Western and product design engineers in the Bell Telephone Laboratories. Modifications of design for Western's convenience, changed methods for translating basic requirements into manufacturing test requirements, informal Laboratories suggestions of approaches to manufacturing and testing problems, all are commonplace. The boundaries of the specialists' domains are readily crossed.

Testing is a process for proving something such as quality of a prod-

uct or accuracy of a computation. In one form or another, testing is essential in manufacture. It insures against further investment of effort in product found bad. More importantly, it provides information for the manual or automatic correction of earlier processes, to prevent manufacture of additional faulty product. Also, its techniques and devices are used in many applications where testing is not the object. Table I gives a listing of functions, with examples of some of our automatic means, that illustrates this. Of these, 1a, 4, and 5a are testing functions. The remainder are manufacturing processes.

TABLE I

<i>Function</i>	<i>Example</i>
1. Sorting, either	
a. sorting good from bad or	A network testing machine at Indianapolis. ¹
b. sorting into cells for selective assembly	A relay coil test set at Kearny. ²
2. Adjusting:	A capacitor test machine at Hawthorne. ³
3. Calibrating:	An adjusting machine for flat type resistors at Haverhill. ⁴
4. Plotting data:	A calibrating machine for oscillator film scales at Kearny. ⁵
5. Operation of wired equipment,	Continuous thickness test systems for alpeth and stalpeth cable sheath at Hawthorne and Kearny. ^{6, 7}
a. to verify accuracy of wiring or fulfillment of purpose, and	Cardomatic and tape-o-matic test sets for key telephone equipments and wired relay units at Hawthorne and Kearny. ^{8, 9}
b. to enable prompt location and correction of faults.	

GENERAL

The fundamental steps necessary to any testing operation are:

1. Putting the item to be tested in location;
2. Subjecting the item to a specified set of conditions;
3. Observing the results or the reaction of the item to the conditions;
4. Comparing the observed results to required results;
5. Deciding on the basis of the comparison what disposition to make of the item;
6. Indicating the disposition;
7. Making the disposition. (This may mean transportation, repair or adjustment.)

In purely manual testing all of these steps would be initiated by human

operators. In many cases it is feasible for all steps to be taken automatically. The bulk of our accomplishment in automatic testing, however, has been in steps 2 through 6. We do not ordinarily use "automatic" to describe rudimentary automaticity in combinations among steps 3, 4, and 5.

The present models of many of our machines have evolved from earlier models, either because of changed product or test requirements or through improved designs worked out for plant expansion or cost reduction. The names of engineers associated with the various developments mentioned are included in the references. About 1927 there were put in use at Hawthorne two machines, one for gaging a number of critical dimensions and performing a breakdown test on carbon protector blocks,¹⁰ and the other for heat coils.¹¹ In the protector block machine the blocks follow a linear course drawn by an indexing chain conveyor through a number of positions where the various checks are performed. Failure of any block at a position causes a jet of air to blow the block into the opening of a chute which conducts it to a reject pan. Good blocks are delivered into a pan at the end of the run. The heat coil machine has an indexing turret over a ring of ports which open selectively to permit good or rejected coils to fall into chutes. The test parameters are three gaged dimensions and dc resistance.

In 1929 a machine with an indexing turret was put in use, testing paper capacitors for dielectric strength and leakage resistance,¹² and sorting them into 13 cells for capacitance grouped around a nominal 1 mf. The 13 cells correspond to 13 segments in a commutator disposed along the scale of a microfarad meter. For a given test capacitor, when the meter needle reaches its deflection a bow depresses it against the nearest segment, establishing a circuit through a relay. A system of relays then locks up and serves as a memory to operate a solenoid later when the turret has brought the capacitor to the point of disposition. Action of the proper solenoid causes the capacitor to be deposited in its cell. The cells are arranged as parallel files in a horizontal plane and, starting with the cells empty, the machine will in effect produce a stovepipe distribution curve. Capacitors from the middle cell and its upper neighbors may be used as 1 mf capacitors, and those from more remote cells combined, large with small, to make 2-mf capacitors.

Also in 1929 a turret type machine was first used for sorting mica laminations.¹² The sorting parameter was ac dielectric strength, the criterion being failure at 1760 volts r.m.s. The individual laminations were carried from position to position by vacuum fingers mounted on a turret. Again locking relays were used, in this case to operate a solenoid controlled

valve in the vacuum line at the right time in the turret indexing cycle to drop the laminations as class "A" or "B" mica.

Experience with these machines and with others that followed brought into being a more or less orderly body of knowledge as to what features are desirable and what constitutes good design in an automatic test machine.

If the machine is to have speed, reliability and long life, attention should be paid to the following matters:

1. *Reduction of the test process time to as low a figure as the capabilities and use of the product will permit.* Thus, if one of the requirements of a capacitor is a maximum limit on its leakage current measured after a charge time of 60 seconds, and if the materials and manufacturing process are such that a unit is surely good or bad after a 25-second charge, then the machine may be designed to charge for, say, 30 seconds. Frequently the only limitation is the speed of the machine itself. When this is true, it must be worked out so as to satisfy the needed production rate. Obviously the machine should satisfy the rate of the line it serves, or more than one machine should be provided.

2. *Rationalization of the number of test positions in the machine with the production rate and the total test process time.* This requires breaking the test time down into bits equal to the desired output cycle. In the example above, if the output needed is a capacitor every 5 seconds then the 30-second charge will have to extend over 6 positions.

3. *Ruggedness.* This must be stressed, even at the expense of space, power consumption, and dollars of first cost. If a project is large enough to justify automatic test facilities, then any down time associated with it will be expensive. A good mechanical design is essential.

4. *Provision of self-stopping and alarm features to serve in the event of certain types of failure.* A limited torque clutch in the main drive will prevent jamming and damage caused by parts getting into the wrong places, or in certain applications overload cutouts will suffice. Gong and lamp alarms are desirable to attract attention. The point is that allowance must be made for mishaps which, without precautions, could result in shutdowns of the equipment.

5. *Provisions of adequate checking for accuracy.* Accessible check points and suitable easy-to-use standards are essential. Checking intervals are determined by experience, but schedules should be laid out to cause as little interference with use as possible. Where practicable there may be means for self-checking in the regular operation of the machine. In this case, periodic checking of the checking devices themselves is necessary.

6. *Incorporation of features in the product and in the handling methods*

that will facilitate feed automatic testing. This requires the cooperation of the product design and product manufacturing interests. It is almost axiomatic that automation in manufacture requires special consideration in product design. Automatic testing imposes the same requirement. A notch or a lug may be needed for proper use of automatic feed devices, or terminals may have to be properly chosen. Again, the method of transport from the previous operation needs to be studied, rationalized, and fully agreed upon. If continuous conveyor transportation can be justified, so much the better. In the consideration of conveyor feed, the need for time flexibility must not be overlooked. It is important that provision be made for easy storage of product whenever the test machine is inoperative, lest a breakdown of this machine shut down the entire line.

7. *Arrangement of the events in the operating cycle in such a way that their sequence is reliably self determined.* This is comparatively straightforward when the programming is done by gear driven cams or other mechanical means. It requires care when switching logic is used. Switching engineers are familiar with the phenomena known as "relay races" and "sneak circuits." These have psychophysical analogies wherever humans and machines work together. The prevention of both the switching errors and their analogs is essential in automatic test set design. Interlocks must be provided against any conceivable mishap.

8. *Enough margin and design flexibility in electrical and mechanical parameters to cope with reasonable variations in product design.* Improvements are constantly being made in telephone apparatus and equipment, and these occasionally result in major redesigns or in entirely new systems. Also the need for adding new features to a historical complex of existing telephone plant causes the generation of an endless variety of special equipments. The product designer needs as much freedom as we can afford. There has to be enough flexibility in the costly automatic test sets to permit adaptation as new designs of product come along.

These considerations are in addition to the fundamental matters of personnel safety and comfort, motion economy, quietness and appearance.

While dealing with general considerations we must recognize one important difference between the product design and the facilities design problems. In product design there is a premium on optimization of parameters, or striving toward perfection. There is generally also opportunity for winning this premium on later tries even though the rush for first production may have denied it to us in the original design. In facilities design there is no such premium and frequently no such opportunity. While careful design is very important, the real premium here is on a

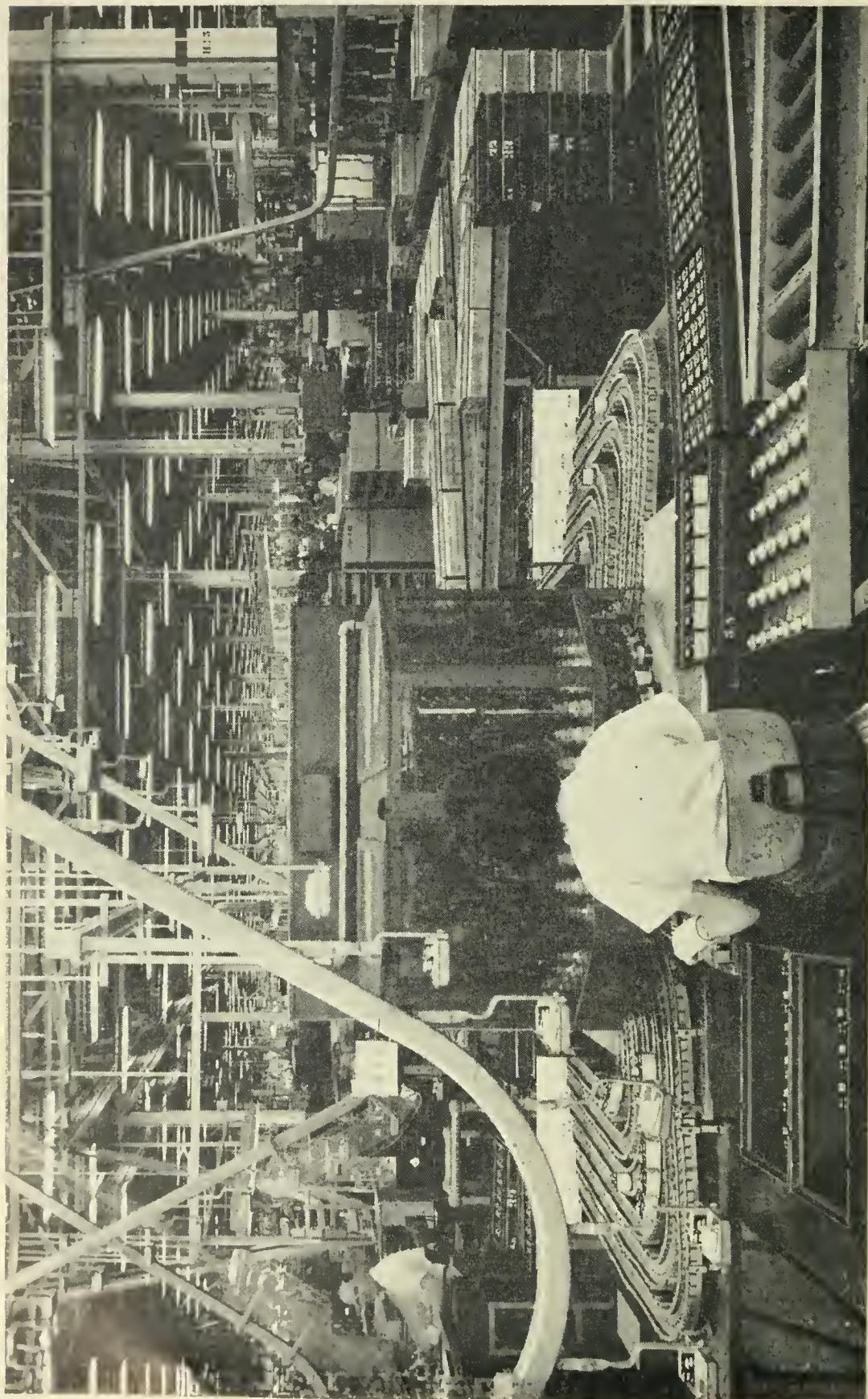


Fig. 1 — Network test position.

device that will do the required job and that can be put in use in time for early production. Once the facility is in use it may be starting on a productive life that will run thirty years or longer. The designer may think of countless ways to improve it or to redesign it completely. If his improvements or redesign can be proved in on a business basis, they may be undertaken. Sometimes they cannot be proved in. The evolution that has taken place in test set designs has been possible mainly because the customers have wanted newer products, or products delivered at a greater rate. Advancement has been attained under a compulsion to take each step quickly and surely. This has represented a real and continuing challenge to the test engineering force.

With these general considerations in mind the author has chosen three automatic testing devices of diverse character to discuss in some detail. The associated papers^{3, 9} cover additional machines. The machines described illustrate in various ways the principles discussed above.

THE NETWORK TESTING MACHINE AT INDIANAPOLIS¹

The 425B network¹³ is used in the 500 series telephone sets to furnish the transmission link between the handset and the line. Its shop testing requires three tests for transmission, three for capacitance tolerance, three for leakage current, two for ac dielectric strength, one for dc dielectric strength and four for continuity. The rotating turret type test machine (Figs. 1 and 2) performs all these tests, applies a conditioning "burnout" voltage and counts and date stamps the good networks. Rejects from each test position are segregated in roller conveyors. In the rotation of the turret an empty test fixture is presented to the operator every $3\frac{3}{4}$ seconds moving from left to right. She must load each position, taking networks from the pans at her right; good networks, ejected automatically in a roller chute at the left, are hand loaded into the carriage fixtures of the overhead storage type conveyor, which pass within easy reach of the operator's left hand. The pans at the left are used to store good networks when the accessible fixtures of the overhead conveyor are full. The twelve roller conveyors for rejected networks are arranged along the sides of the machine, six on each side.

The turret contains forty test fixtures (Fig. 3 and 4) and the machine forty positions. The turret rotates continuously, causing eleven contact brushes associated with each fixture to pass against fixed commutator segments and a ground ring associated with the test positions. As each fixture advances past one test position a gear connected cam shaft rotates through a complete cycle. Seventeen switches are operated by the cams

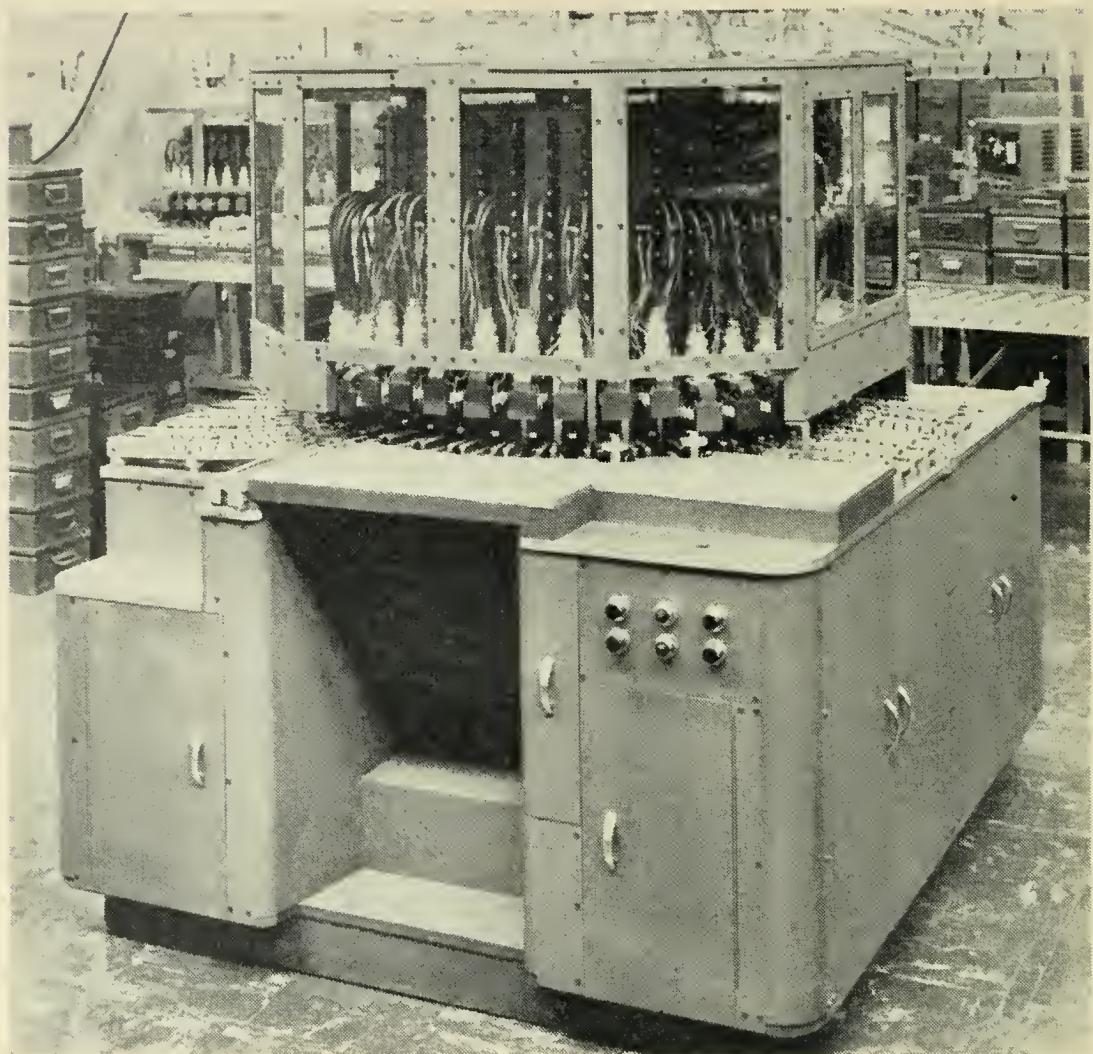


Fig. 2 — The network test machine.

to assure the proper sequence and timing of the conditioning and testing events occurring at the various positions. Table II shows the order of the positions and the approximate timing, with respect to cam rotation. The result of the test at each test position is remembered by a self-locking relay until the fixture comes just opposite the entrance to the corresponding rejection chute. At that instant a cam switch closes and causes rejection if the test result was a failure. Unloading into the rejection chutes is effected by compressed air operated cylinders as explained below.

The clamping movement of each fixture as it leaves the loading area (entering position 7) is driven by a helical spring which lowers the contact fixture over the terminals of the network, bringing spring loaded plungers into contact with the terminals. (See Fig. 3) At a rejection location a plunger rises, driven by an air cylinder under the control of a solenoid operated valve. The rising of the plunger first forces the fixture to

unclamp against the compression of the helical spring, and then operates an ejection arm which drives the network horizontally out of the fixture. The top rollers of several of these ejection arms can be seen in the fixtures at the front of the machine in Fig. 2.

The measuring circuits associated with the various test positions are straightforward. If there is a dielectric failure in one of the breakdown tests at position 8 or 9, the current through a relay coil in series with the

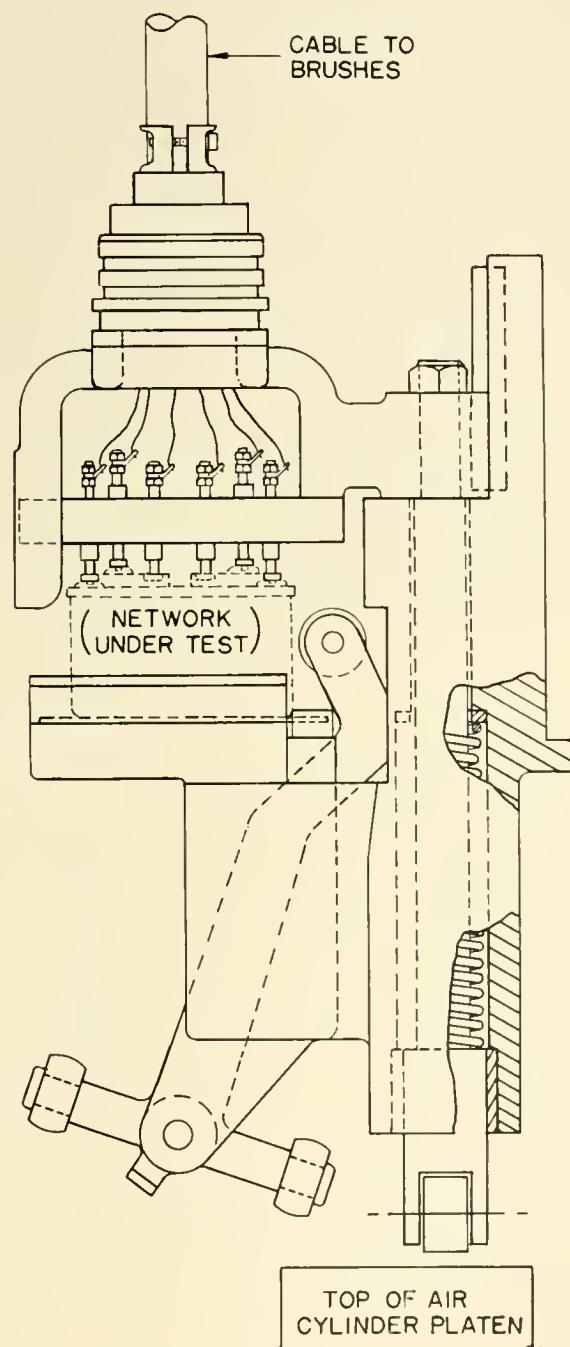


Fig. 3 — Test fixture, loaded.

test exceeds a predetermined value. This causes another relay to lock up and remember the failure until the network reaches the reject location.

In a typical transmission test position (Position 10, 35 or 36) a fixed-voltage, swept-frequency signal, 300 to 3,500 c.p.s., is impressed across two terminals of the network. The three tests are for transmission and short and long line sidetone with suitable terminations connected as in actual use. In each case the signal from two output terminals should be less than or greater than a specified value. This signal is amplified and fed

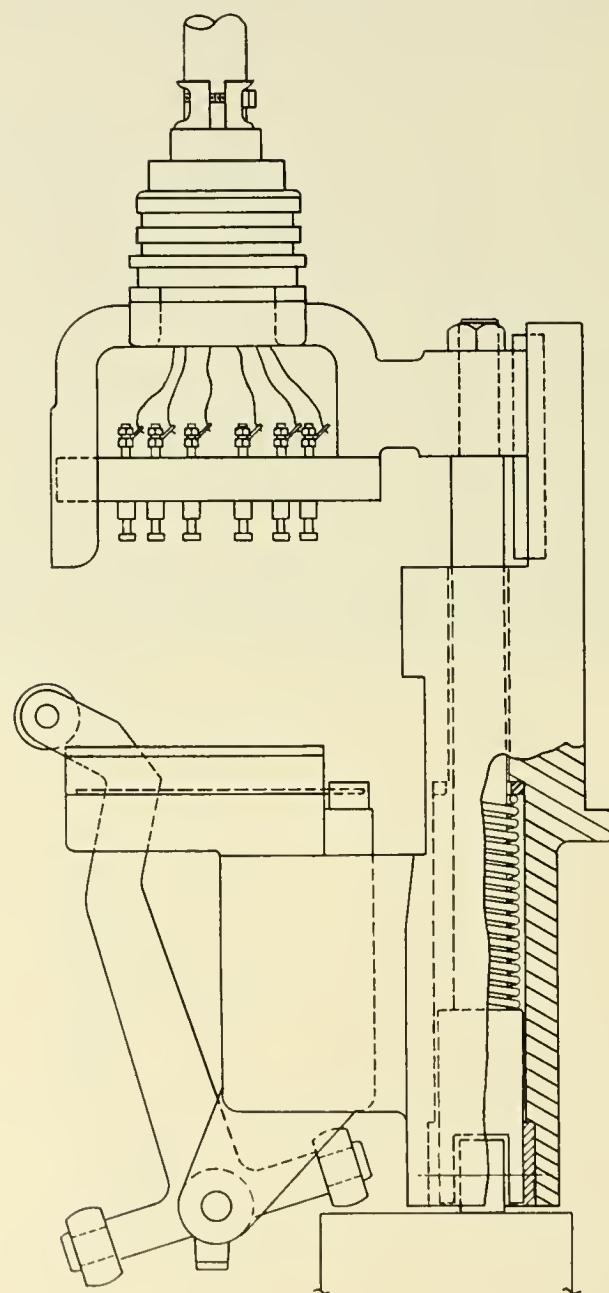


Fig. 4 — Test fixture, unloaded.

to a sensitrol relay, which is mechanically biased in amount and sense to correspond to the limit. If the sensitrol operates it prevents rejection.

In the three capacitance test positions 12, 13, and 14, capacitors in the networks are connected into a 60 c.p.s. comparison bridge. The output signal from the bridge is amplified and rectified, and impressed on a balanced dc amplifier which drives a sensitrol relay. If the bridge is out of balance (that is if the capacitance is greater or less than nominal) current flows in the relay, but always in the same sense. If the current in the relay exceeds an amount corresponding to either capacitance limit, rejection occurs. Determination of which capacitance limit was violated is done manually in a separate analysis of defects. It may be observed also that any rejection at the capacitance positions could have been caused by a loss unbalance of the bridge. If the conductance of the test capacitor were such as to cause this it would so appear in the separate analysis, mentioned above. The effect of any ordinary conductance deviation at 60 c.p.s. is negligible. Quality is protected by the fact that a conductance deviation could not cause an out-of-limit capacitor to be accepted.

Considerable pains are taken at each capacitance test position to prevent damage to the equipment from various kinds of mishaps. The sensitive winding of the sensitrol is short circuited at all times except for about 0.2 second when the actual test is performed. This prevents damage and erroneous rejections that would otherwise be caused by switching

TABLE II — SEQUENCE OF EVENTS IN NETWORK TEST MACHINE

		CAM ROTATION →								
		(TWELFTHS OF A POSITION)								
POSITION	PROCESS	0	2	4	6	8	10	12		
1 TO 6	LOAD									
7	BURNOUT			CHARGE		DISCHG				
8	AC BKDN.			TEST						
9	DC BKDN.			TEST						
10	TRANSMISSION 1		TEST							
11	DISCHARGE									
12	CAPACITANCE 1	CIRCUIT SETUP	TEST		MEMORY					
13	" 2	"	"		"					
14	" 3	"	"		"					
15	BURNOUT	CHARGE			DISCHG					
16 TO 31 (1 MINUTE)	CHARGE FOR LEAKAGE TEST	CHARGE	CHARGE							
32	LEAKAGE 1		TEST							
33	" 2				"					
34	" 3				"					
35	TRANSMISSION 2				"					
36	" 3				"					
37	CONTINUITY	TEST 4 CIRCUITS AT ONCE								
38	UNLOAD						UNLOAD			
39	RESET									
40	LOAD									
1	"									

EJECTION AT POSITIONS WHERE FAILURES OCCUR

transients from this and other circuits. During the short interval of actual test no other switching takes place in the machine.

A fixture that has no network because of rejection at an earlier test position or because of operator failure to load it, would cause open circuit in one bridge arm on capacitance test. Without intervention this would cause a violent unbalancing of the bridge, overloading of the detector system and possible damage to the sensitrol. Ordinary methods of limiting the overload signal would be only partially effective and would detract from the sensitivity. To forestall this trouble from empty fixtures, each capacitance test position is equipped with a microswitch which is operated by a dog at the bottom end of the ejection arm of any empty fixture (Fig. 3). When the microswitch operates it causes the bridge to be disconnected from the test leads and connected to a capacitor that is just out of limits, several tenths of a second before the removal of the short circuit from the sensitrol. Then when the test is made it results in a rejection.

There is also an interlock circuit which will stop the machine if a failure of the bridge and detector system causes an empty fixture not to show rejection. This serves as a random occasional check on the functioning of the circuit.

The conditioning of the three capacitors for the leakage current tests begins at position 16. Because of charging and absorption currents obscuring the effect of pure leakage, the test for leakage is made to an arbitrary current limit specified at one minute of charge. To insure that good units pass the test, it is desirable to use the whole minute. But if the leakage current reading is taken after more than a minute of charge, quality is jeopardized. Accordingly it is necessary to make sure that the charge is for a minute and no longer on each capacitor. Therefore, at position 16 the first unit is put on charge, at 17 the second, and at 18 the third. Then at position 32 the first unit is tested while the other two remain on charge. At 33 the first unit is discharged, the second tested, and so on.

The leakage test itself is made by measuring the voltage across a large resistor in series with the test capacitor and a dc voltage source. The energy in this signal is small and must be amplified before there is enough to operate a sensitrol. A dc amplifier with high input impedance is used for this purpose. In addition the mechanical bias of the sensitrol is kept small to increase sensitivity, and a carefully controlled dc biasing source is used to insure accuracy and stability.

At position 37 three capacitors and a coil winding are given a final check for continuity. The test of the winding is made by connecting it in series with a relay coil (say No. 1) and battery. If current passes, relay

No. 1 operates. The three capacitors are tested simultaneously by connecting each of them in series with an 8,000 c.p.s. source and detectors. The detectors consist of bridge type rectifiers and relays. If all of these three relays operate, a series connection through their closed contacts causes another relay to operate and lock up. Finally this relay when operated has open contacts in parallel with open contacts on relay No. 1, so that when the reject cam closes it finds an open circuit and rejection does not occur.

The reader may question the necessity for continuity tests on capacitors that have already been tested for capacitance. Perhaps the most convincing answer is that there is an occasional failure on the continuity test. Telephone apparatus is always exposed to more severe conditions in test than it will encounter in ordinary use. The leakage resistance charge and test operations and the transmission tests can on rare occasions cause the metallized connections at the ends of the capacitors to open. As the cost of making the final continuity test is vanishingly small, the additional insurance is economical.

The detail list of checking standards for this machine contains some twenty items. Most of them are modified 425B networks, specially arranged in one way or another to check certain functions of the machine. These are used right in the individual fixtures.

It is interesting to reflect on the labor saving virtues of this machine. The operator in one eight hour shift handles over five tons of networks. She does it easily and without fatigue. The testing would not be even attempted on a manual basis, because over and above multiple handlings, the added human effort of closing fixtures, operating switches and the like could not be tolerated.

In contrast to the multiposition set described above, it is instructive to consider two single position sets of diverse character. They are a relay coil test set and a film scale calibrating set.

THE RELAY COIL TEST SET AT KEARNY²

Coil assemblies for the U, Y and UA types of relays¹⁴ are tested for dc resistance, direction of winding and breakdown before assembly into complete relays. Many thousands of the relays are used in any crossbar office. Minimum and maximum tolerance limits are placed on their winding resistances, to control cumulative current requirements and to insure a proper margin of relay operation. Each coil assembly, as presented to the test position, consists of a magnetic core, a solenoidal winding assembly and a terminal assembly. A winding assembly may have one, two or



Fig. 5 — Relay coil test set control panel.

three windings (called primary, secondary and tertiary). The primary and secondary are wired to corresponding pairs of terminals on the terminal assembly, while the tertiary leads at this stage are not on terminals and must be connected to the test contact fixture by hand.

Direction of winding is important in the multiwinding coils because of external fields and the fact that the relays are required to respond to currents in more than one winding and the proper direction of flow in each, relative to the other, must be known. In some relays one or two of the windings may be noninductively wound, to serve merely as resistors. Also, many windings are wound part copper and part resistance wire to obtain the desired resistance without unnecessary increase in copper, inductance and response time. In such cases the percentages of copper and resistance wire are known. This is important because of the effect of temperature on the resistivity of copper. Resistance tolerances on the test windings are specified at 68°F, but shop testing is done at any value of room temperature. The effect of the difference on copper is serious enough to cause errors larger than some of the tolerances, and the effect on resistance wire may be neglected. Therefore, it is necessary to have the test set compensated for temperature in such a way as to allow for the proportions of copper and resistance wire.

The coil test set (Fig. 5) tests all windings for resistance and direction of winding and for breakdown to each other and the core. The maximum total test time for three-winding coils is less than 3 seconds under normal conditions. A borderline winding resistance will cause some delay. There are lamps to indicate the type of failure on a rejection. Other lamps indicate satisfaction of the requirements. At the completion of test on a good coil an "OK" lamp lights on the test fixture, so that the operator need look at the set itself only when there is a rejection.

Requirements data are stored in the set before a given code of coil is tested. The codes come to the set in batches, so that one setup will serve for a large number of coils. Three six-decade resistance standards are set to the nominal values for the respective windings. If there are fewer than three windings, a key is operated to disable bridges and furnish substitute continuity paths. The percentage tolerances for the windings are set on selector switches: ± 1 , 2, 5, 10 and 15 per cent tolerances are available. Also, the known percentages of resistance wire in the windings are set on selector switches in steps of 5 per cent from 0 to 100. Keys are operated to warn the set of noninductive windings and bypass the direction of winding circuits as needed.

Once a coil is placed and connected in the test fixture and the fixture closed by operation of a pedal, the test is automatic up to the point where

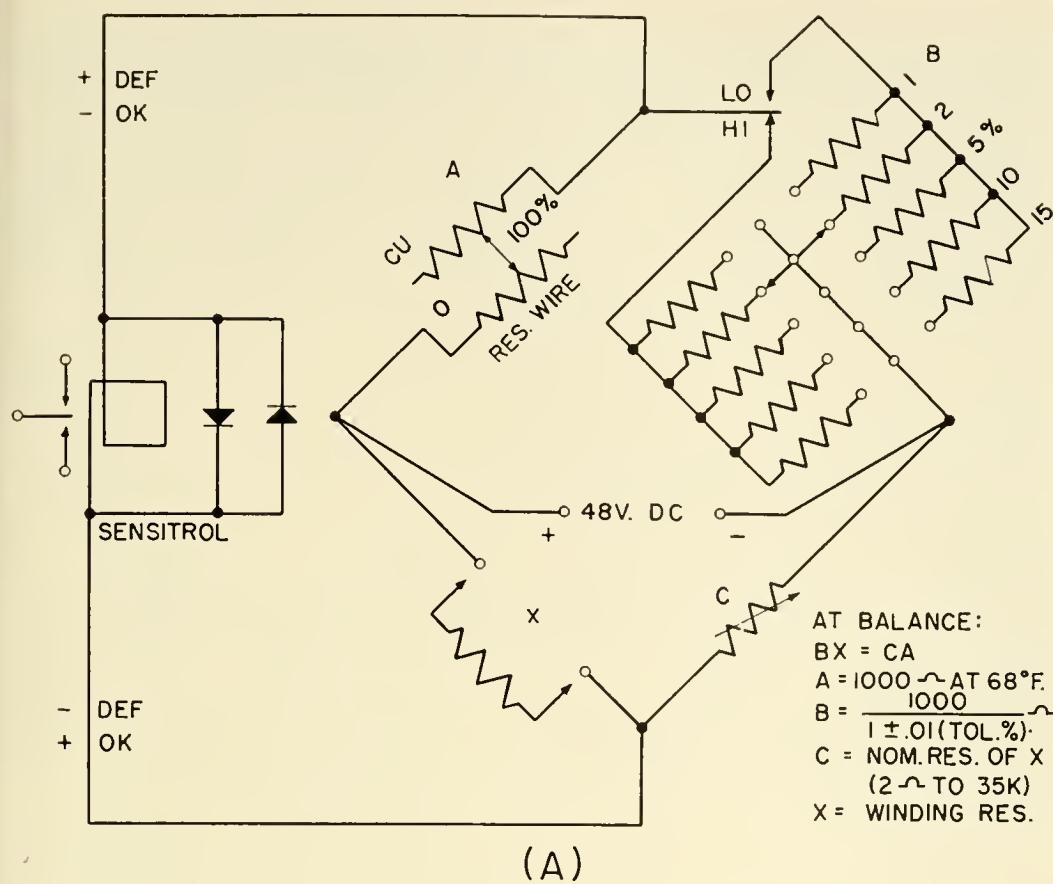
the operator must make disposition. The sequence of events within the set is controlled by a switching circuit containing thirty telephone relays, a sensitrol relay and two electron tubes. The sensitrol is used in succession to detect the existence and sense of unbalance of six dc bridge circuits (high and low limit for each of three windings). The operation sequence for primary windings is shown in Table III.

Fig. 6(a), shows schematically a typical bridge arrangement for testing a winding at one tolerance limit. A and B correspond to the ratio arms of an ordinary Wheatstone bridge, and are nominally 1,000 ohms each. The temperature compensation referred to above is obtained by including the same resistance percentage (within 2.5 per cent) of copper in the A arm of the bridge as there is known to be in the winding. Inspection of the bridge balance equation in Fig. 6(a) will show that an error in X could be compensated by a proportional error in either A or C. A is chosen as the compensating arm because of its simplicity. It has available twenty resistors of copper and twenty of low temperature coefficient resistance wire. Each resistor is 50 ohms, measured at 68°F. The selector switch is arranged so that the arm always has twenty resistors, the indicated percentage being resistance wire.

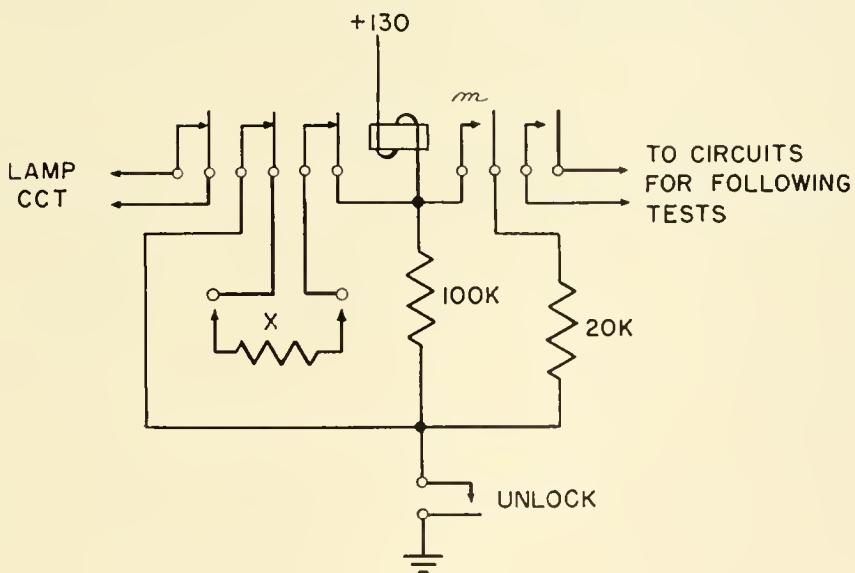
For proper compensation it is necessary that the A arm be as near ambient temperature and the temperature of the coils as possible. The di-

TABLE III — SEQUENCE OF EVENTS IN TEST OF PRIMARY WINDING FOR HIGH LIMIT RESISTANCE

"OK" LAMP	STEP	DEFECT LAMP
	POWER SWITCH CLOSED	
	SENSITROL RESETS AND HOLDS	
	OPERATOR CLOSES FIXTURE	
	FIXTURE START SWITCH CLOSES	
	CONTINUITY TEST - ALL WINDINGS	"P OPEN" ETC.
	"HIGH" B ARM CONNECTED TO PRI. BRIDGE	
	SENSITROL RESET RELEASED	
	SENSITROL OPERATES	"HIGH"
	"LOW" B ARM CONNECTED TO PRI. BRIDGE	
	BREAKDOWN TEST ON PRIMARY	"BREAKDOWN"
	SENSITROL RESETS AND HOLDS	
	SENSITROL RESET RELEASED	
"P RES. GOOD"	SENSITROL OPERATES	"LOW"
	DIRECTION OF WINDING DETECTOR ENABLED	
	D.C. POWER DISCONNECTED FROM PRIMARY BRIDGE	
	INDUCED VOLTAGE IN PICKUP COIL	
	SENSITROL RESETS AND HOLDS	"P DIR. OF WDG. DEFECT"
(OK)	"HIGH" B ARM CONNECTED TO SEC. BRIDGE	
	(SECONDARY TEST PROCEEDS; SIMILAR TO PRIMARY)	
	(FOR SINGLE-WINDING COIL) (LAMP ON FIXTURE LIGHTS)	



(A)



(B)

Fig. 6 — Circuits used in Relay Coil Test Set. (a), resistance bridge, simplified schematic; (b) continuity, simplified schematic.

vision into twenty resistors helps in this by maintaining high effectiveness of dissipation. In addition, the automatic switching circuits are arranged to keep the duty cycle of current in the bridge arms low.

The B arm of the bridge is selected by the setting of the percentage tolerance switch. Each resistor is used alone and consists of low temperature coefficient resistance wire as in standard bridge practise. The value of each resistor in ohms is 1,000 divided by one plus or minus the corresponding tolerance fraction. Thus, for ± 1 per cent tolerances the resistors are $1,000/1.01 (=990.0)$ and $1,000/0.99 (=1010.1)$, respectively. One setting of the switch indicates zero tolerance and is equipped with 1,000-ohm resistors to permit easy checking of the C arm precision.

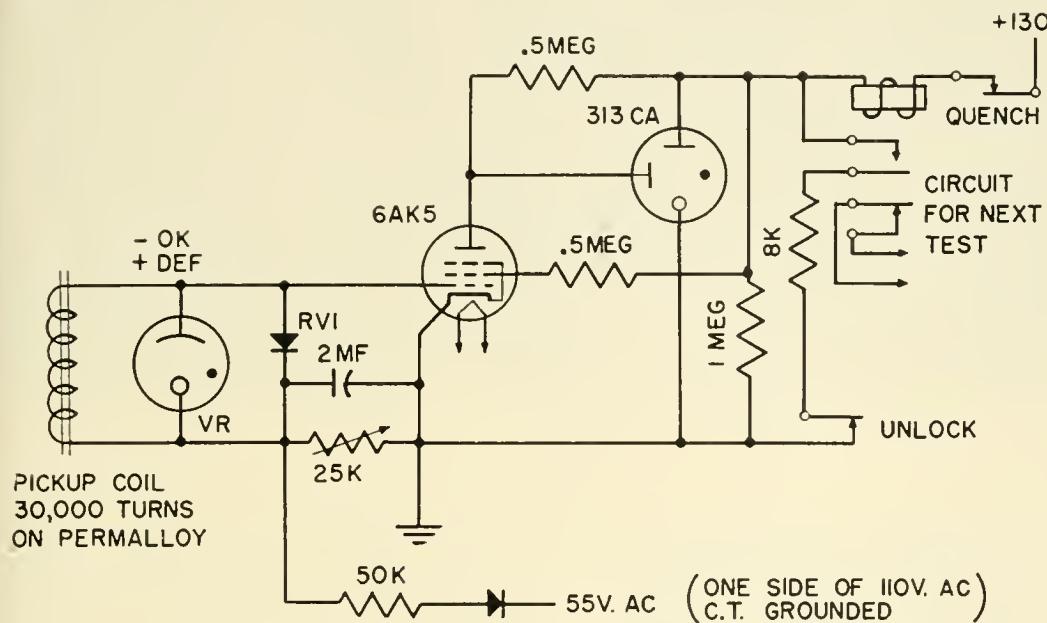
The six-decade standard resistor in the C arm, which is set to the nominal value of the test winding, is of a high quality commercial type with a range of 0 to 40,000 ohms in steps of 0.1 ohm. Because the C and X arms may contain values as low as 2 ohms, no relay contacts are used in them. Relay switching is done in the A and B arms where the resistances are always of the order of 1,000 ohms and small variations in contact resistance are negligible. The more stable wiping contacts of selector switches do appear in the X arm. These switches permit any contact in the test fixture to be connected to any bridge terminal, to enhance flexibility.

A continuity test on all windings, before resistance test, is desirable for two reasons. The effect on the sensitrol of the severe bridge unbalance caused by an open winding would be life-shortening and is to be avoided if possible. Also, the result of the resistance test would only show high resistance, and separate analysis would be needed to reveal that a winding was open. The continuity test circuit in Fig. 6(b) was devised to prove continuity for windings having resistance values as high as 35,000 ohms. A relay (UA-104) was chosen which is sensitive enough to close a pair of "preliminary make" contacts (m) on 0.005 ampere, and which provides the number of other contacts needed to satisfy circuit requirements. When the test winding is connected at X, the currents through it and the 100,000 ohms combine to equal 0.005 ampere or more. This closes m, connecting the 20,000-ohm resistor in parallel with the 100,000 ohms, thus locking the relay and assuring that all the other contacts operate. In the act of proving continuity, the relay disconnects itself from the test winding and remains locked. The make contacts shown at the right end of the relay symbol are in series with similar contacts on the continuity relays for the other two test windings, and when all are closed they pass operating current to a relay which initiates the first resistance test (for primary high limit).

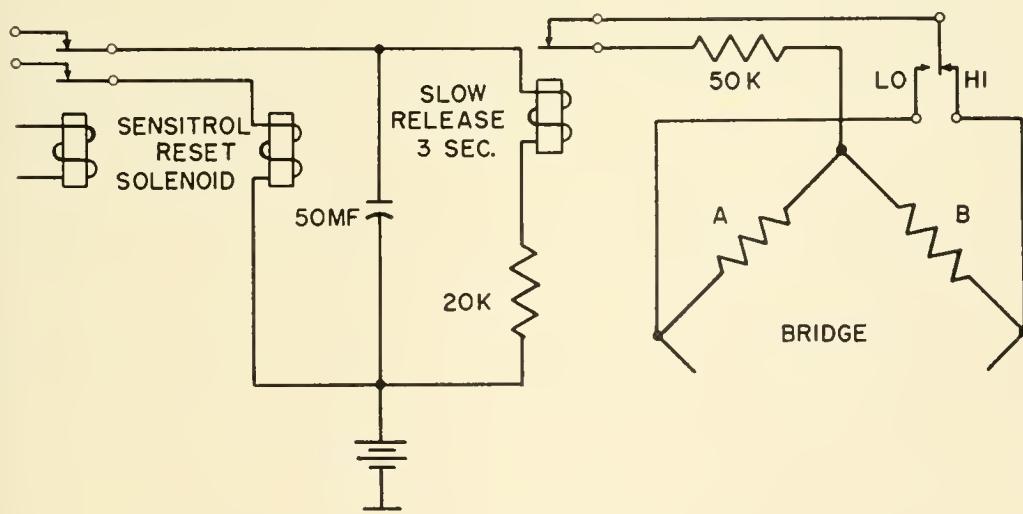
In the direction of winding circuit, Fig. 7(a), it is necessary to have a negative pulse from the pickup coil, in the test fixture, cause the 313CA

gas tube to fire and the relay to operate. The circuit is designed to handle a wide range of pulse amplitudes. The VR tube limits negative pulses to 90 volts to protect the 6AK5. The varistor dissipates positive pulses and prevents any false acceptance that might be caused by damped oscillations following a positive pulse. The 6AK5 furnishes the needed sensitivity for small pulses.

Occasionally a winding will have a value of resistance just equal to



(A)



(B)

Fig. 7 — Circuits used in relay coil test set. (a), direction of winding, simplified schematic; (b) anti-stall, simplified schematic.

its upper or lower tolerance limit. On the corresponding resistance test, the sensitrol will balance and not operate either way. Without an anti-stall device the test cycle would then be stalled until the balance failed. Current flowing through the A arm would eventually heat it up and vitiate the temperature compensation feature. The anti-stall circuit in Fig. 7(b) is essentially a slow release device to which external energy is interrupted at the same time as the sensitrol reset is released. Energy stored in the 50-mf capacitor prevents release of the relay for about 3 seconds, long after the bridge test is ordinarily finished. If at release the bridge is still balanced, a 50,000-ohm resistor is thrown in parallel with that ratio arm which will make the sensitrol accept the test winding.

A prominent and hitherto valuable feature of this test set is its adaptability to a large variety of coil assemblies. Some hundreds of distinct designs of product are presently accommodated. In the Kearny relay coil shop there are four sets of the design described here and four sets of earlier designs. It is possible that future development, if justifiable, will be directed toward greater automaticity for some of the simpler and more numerous product codes, with less emphasis on universal application.

THE CALIBRATING MACHINE FOR 56-A OSCILLATOR FILM SCALES⁵

Photographic films are used for the frequency scales of some oscillators to afford scale length and enhance readability. There have been several successive designs of film scale calibrators built and put in use at the Bell Telephone Laboratories and at Kearny. Some have been described in the literature.^{16, 17, 18} One very early design is still in use on production at the Marion Shops in Jersey City. In its use, a calibrating run requires about an hour, and the possibility of frequency drift due to temperature variations makes the use of an air conditioned room essential. All of those used at Western, prior to the one described here, depended for accuracy on the film scale of a standard prototype of the oscillator to be calibrated. Using a frequency controlled servo linkage, the scale of the standard was reproduced photographically on the film of the product. Some of the prior art appears in the design of the new machine. In order to describe the principle clearly, it seems necessary to discuss some features which were previously covered, but which now are used in new ways.

The 56A is a heterodyne oscillator designed for use in the field testing of L3 installations.¹⁵ It has a usable range of 50 kc to 10 mc. One component oscillator is fixed at or near 90 mc and the other may be varied between 80 and 90 mc by means of a tunable cavity. The calibrated portion of the 35-mm film scale geared to the cavity tuner is about 17

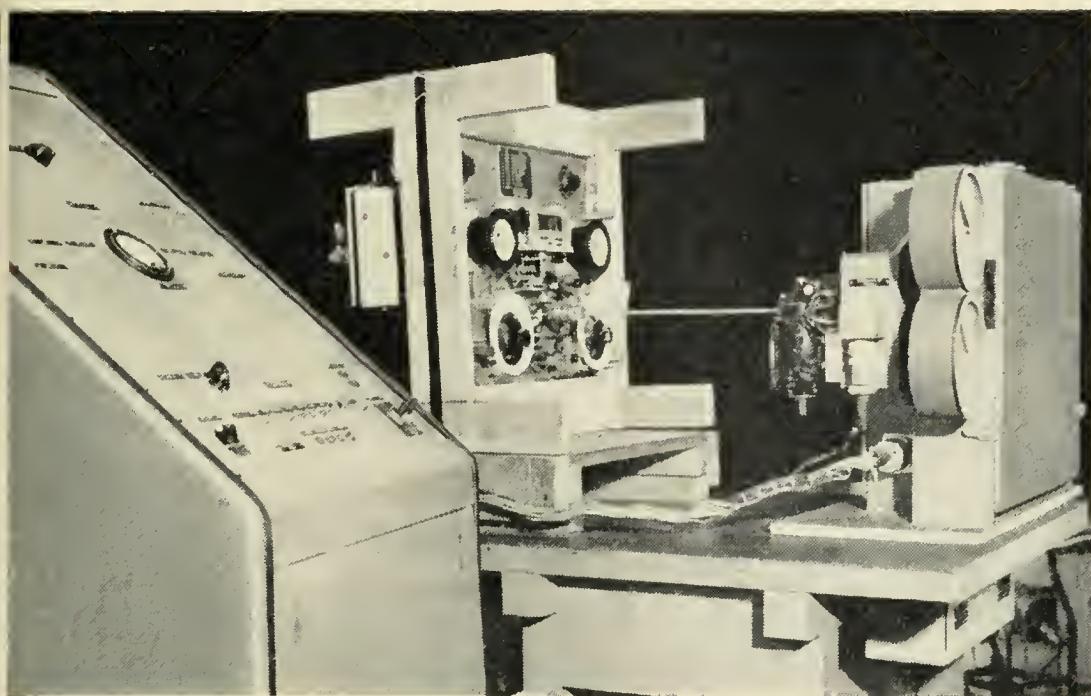


Fig. 8 — Film scale calibrator.

feet long. It has sprocket holes and is moved by a standard movie sprocket. The required precision of each calibration mark is ± 2 ke. Two resonant devices are included in the circuit to permit checking and adjusting two widely separated points on the scale, 100 ke and 7,266 ke. Considering the output frequency as a function of scale setting, one of the two adjustments controls the lateral displacement of the curve and the other its average slope. By design the curve approaches linearity but not closely enough to permit less than a unique calibration for each oscillator manufactured.

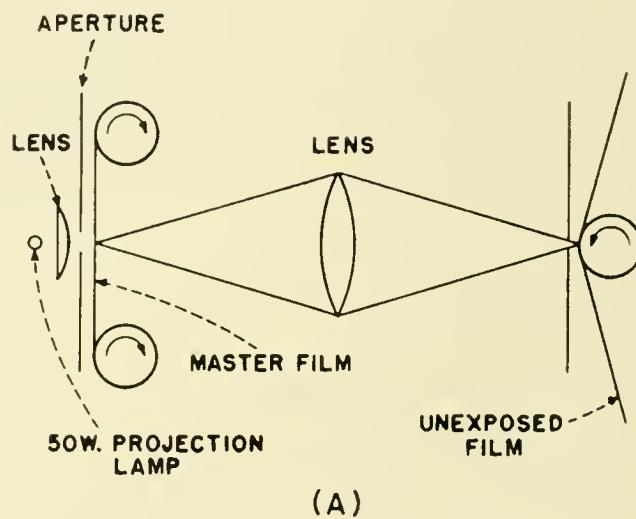
Fig. 8 shows the machine which performs the calibration, with an oscillator connected, and the control cabinet. The oscillator is shown in its shipping frame. An unexposed photographic film to be calibrated is mounted in a camera so that it can be driven by a sprocket. The sprocket is connected by gears to a drive motor which also drives the take-up reel and, through a flexible shaft, the cavity tuner and sprocket in the oscillator itself. The gear arrangement is such that the peripheral speeds of the two sprockets are the same.

A positive master film is provided which has a scale similar to the one to be made for the product except that it is very precisely linear. A portion of the master is shown in Fig. 9(b). The master film passes over a sprocket which is driven by a servo motor. A lamp illuminates and shines through that portion of the master which is in front of an aperture at

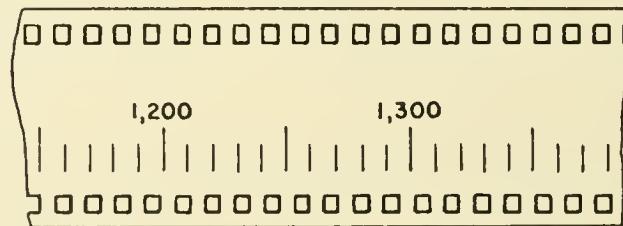
any instant. An optical system, Fig. 9(a), produces on the unexposed film an image of the illuminated portion of the master. As the oscillator, its film, and the master advance, the markings on the master can be reproduced on the new film.

The problem in control is to cause each mark on the master film to pass the slit just as the oscillator goes thru the corresponding value of frequency. To do this we drive the oscillator and its scale together at a constant linear speed. The oscillator frequency increases steadily but not at a constant rate. Its rate of increase varies according to the law of its particular cavity. So our problem reduces to causing the master film to move according to that same law.

The method is to time the passage of known points in the oscillator frequency spectrum, and then to pace the movement of the master film to maintain precise correspondence. The pacing is done by detecting small differences in times of arrival at corresponding points and correcting the speed of the master film to keep successive differences small. Fig. 10 is a block schematic of the automatic control system. The varying oscillator output passes through multiples of 10 kc at a rate near five multiples per second. When it is compared in a balanced modulator with



(A)



(B)

Fig. 9 — Film scale calibrator. (a), optical system schematic; (b) section of master film.

the fixed harmonics of a standard 10-kc signal, the first order difference frequency in the modulator output varies back and forth between 0 and 5 kc. It passes through the 2500 c.p.s. point twice per period of variation, or twice per 10-kc interval of the oscillator frequency.

The output of the modulator is sent through a narrow band amplifier which peaks at 2500 c.p.s. A burst of signal, therefore, leaves this amplifier twice per 10-kc interval. The bursts are further amplified and rectified and become pulses which time the progress of the oscillator through its spectrum. The pulses are impressed across the winding of a high speed relay, causing its contacts to close momentarily twice per 10-kc interval. During the instant when the contacts are closed they connect a particular value from a sawtooth voltage wave to a 0.1-mf capacitor.

The voltage of the capacitor biases the grid of a cathode follower tube, and the output voltage from this tube is fed to a servo system and controls the speed of its motor. Thus the motor runs at a speed determined by the voltage of the sawtooth at the instant when the relay contacts close. As the sawtooth itself is timed by the rotation of the servo motor, its voltage-time relationship is the device for pacing the master film. The sawtooth wave originates in the alternate shorting and charging of a 1-mf capacitor. Each tooth begins when a pair of shorting contacts is closed momentarily by a cam geared to the servo motor. After a discharge, the voltage on the 1-mf capacitor increases negatively as a practically linear function of time, with charging current flowing through a one megohm resistor. Thus the value of voltage transmitted to the 0.1-mf capacitor at the instant of closure of the relay contacts depends on the time elapsed since the most recent shorting of the 1-mf capacitor. Twenty volts at the input to the servo system corresponds to midvoltage of the sawtooth and to 3,600 rpm of the motor, which is the same as the constant speed of the motor driving the oscillator and undeveloped film.

If the characteristic of the oscillator causes a given 2,500-cycle point to occur early, the contacts of the relay will close at a higher positive voltage point on the corresponding sawtooth. The servo motor will start to speed up to make subsequent sawteeth start earlier than they otherwise would have. The motor will slow down if the 2,500-cycle points fall later and lower on the teeth.

Several design features in the system are of interest. The servo system was supplied by Industrial Control Company (SL-1035). It has a tachometer feedback in inverse sense to enhance system stability. The cam used to operate the shorting contactor and start the sawtooth is a small permanent magnet mounted on a wheel. The moving field causes the contactor to operate very briefly as the magnet swings past. The contactor itself is a Western Electric 222-A mercury switch, which has a

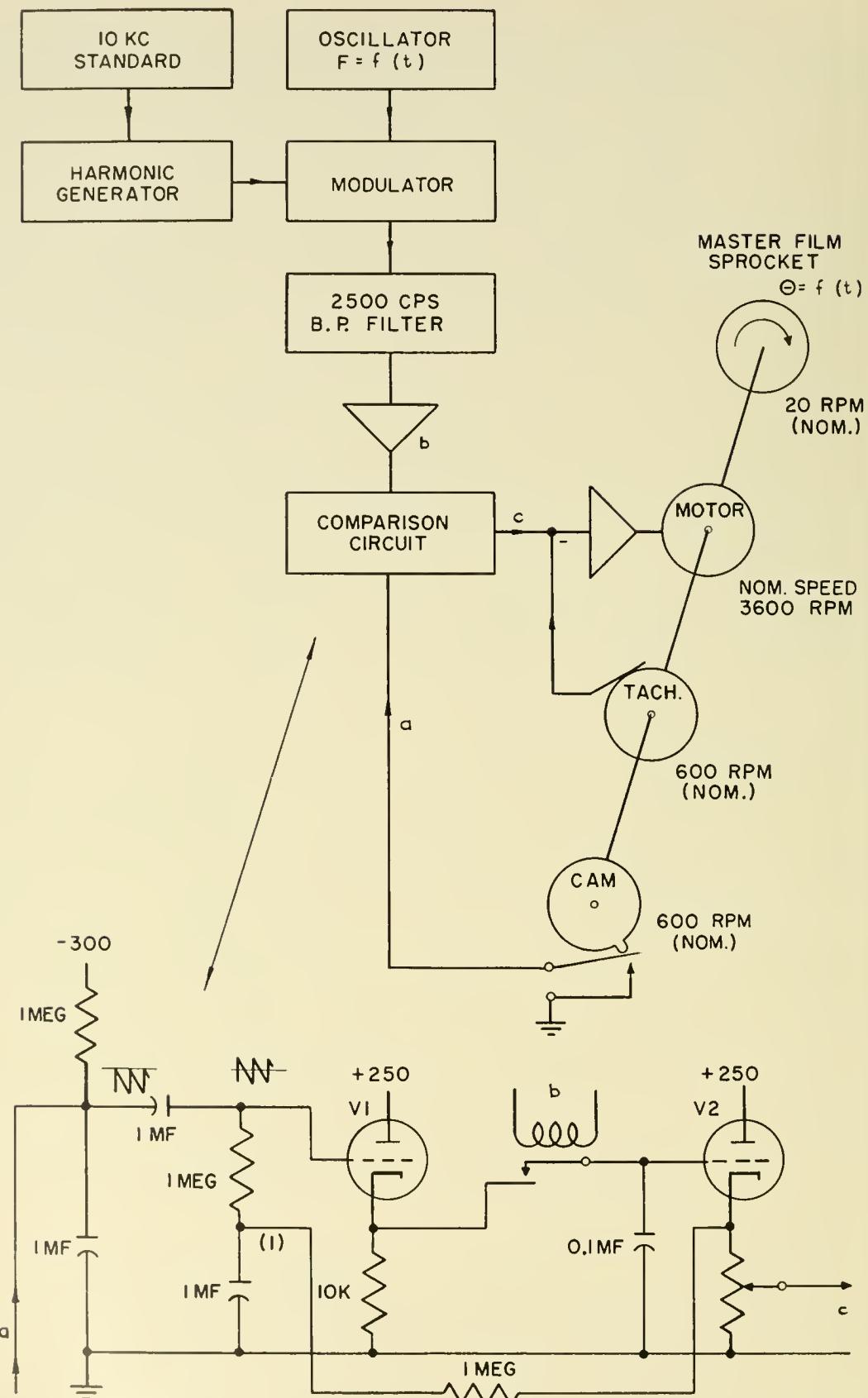


Fig. 10 — Block diagram of film scale calibrator with schematic of comparison circuit.

hydrogen atmosphere, high speed capability and high current capacity. The magnetic arrangement reduces shock torque loads on the servo motor, which might result from mechanical operation. The high speed relay which operates at the 2,500-cycle points is a Western Electric 275-B, chosen because of the speed required (about 10 operations per second).

The time comparison circuit has a small amount of long time constant positive feedback (shown at 1 in Fig. 10) to raise or lower the midvoltage of the sawtooth wave in cases of extreme correction and prevent the control point from slipping one or more teeth. In effect this supplies extra acceleration to the master film when needed.

There is also incorporated in the design an arrangement which permits an important variation in the method of use. A magnetic tape is driven by a sprocket which is geared to the main drive motor and moves with the oscillator drive. The magnetic head for recording on the tape receives its signal in the form of 2,500 c.p.s. bursts through an amplifier. These are the same bursts that time the progress of the oscillator through its spectrum. Thus it is possible to separate the function of calibration from that of printing the film scale. The calibration data on the oscillator is stored on the tape and may be checked for absence of abrupt departures from linearity before it is used to drive the servo and master film in an actual printing run. This eliminates some wastage of raw film. Also a recording (or calibrating) run is made without the servo linkage and can be made at twice the speed of a printing run. A 56A oscillator can be driven through its spectrum, 50 to 10,000 kc, in 100 seconds, allowing very little opportunity for temperature effects to change the check points. In fact no particular effort need be made to control the temperature beyond an ordinary warm up interval.

The control portion of the machine contains various circuits for convenience in setting up and starting the runs. For example one relay circuit under the control of a start button brings a fixed dc voltage into the servo loop, and automatically disconnects after a period long enough for the motor to reach approximately the right speed. A gear shift lever permits changing the ratios between the speeds for the recording run and the printing run.

It is doubtful that a calibration of the 56A oscillator could be performed by manual means. It has been estimated that even if possible, such a task would require more than a week of the most painstaking effort, under very carefully controlled conditions. By comparison, the calibrator requires one minute forty seconds to obtain the data, and three minutes twenty seconds to reproduce it. Development and checking of the exposed film takes about a day. Accuracy of the scales has always been well within the ± 2 -kc limit.

CONCLUSION

In this and the accompanying articles we have given a partial picture of the facilities for automatic testing in the Western Electric Company. At this writing several new machines are under development, and modifications are in progress extending the application of some of the present machines. There is a continuing search for new fields in which to apply these techniques. A staff portion of the manufacturing engineering force now devotes its full attention to automation techniques in general, keeps abreast of the field, bulletinizes important additions to the literature, lends assistance in the solution of problems, and develops specific applications. It is likely that the near future will see important extensions in the use of automatic test equipment.

ACKNOWLEDGMENTS

The author is indebted to the people cited in the references for information used in this article, and particularly to A. L. Bennett, J. Lamont and F. W. Schramm who furnished valuable comments on the early drafts.

REFERENCES

1. Developed by A. L. Bennett and C. R. Rasmussen.
2. The original design of automatic relay coil test set was developed at Hawthorne by R. W. Brown. The set discussed here is a Kearny modification developed by J. Lamont.
3. C. C. Cole and H. R. Shillington, page 1179 of this issue.
4. Developed by G. H. Harmon and A. E. Rockwood.
5. Developed by F. W. Schramm based on suggestions by T. Sloneczewski, Bell Telephone Laboratories.
6. B. M. Wojciechowski, Continuous Incremental Thickness Measurements of Non-Conductive Cable Sheath, B.S.T.J., p. 353, 1954.
7. W. T. Eppler, Thickness Measurement and Control in the Manufacture of Polyethylene Sheath, B.S.T.J., p. 599, 1954.
8. A. N. Hanson, Automatic Testing of Wired Relay Circuits, A.I.E.E. Technical Paper 53-407, Sept., 1953.
9. L. D. Hansen, Tape Control, Automation, p. 26, May, 1956. Also see page 1155 of this issue.
10. Developed by C. F. Dreyer and A. W. Schoof.
11. Developed by L. H. Brown and N. K. Engst.
12. Information was supplied by C. A. Purdy.
13. A. F. Bennett, An Improved Circuit for the Telephone Set, B.S.T.J., p. 611, 1953.
14. Improved U, UA, and Y Type Relays, Bell Lab. Record, p. 466, 1951.
15. J. O. Israel, Broadband Test Oscillator for the L-3 Coaxial Carrier System, Bell Lab. Record, p. 271, July, 1955.
16. W. J. Means and T. Sloneczewski, Automatic Calibration of Oscillator Scales, A.I.E.E. Miscellaneous Paper 50-80, Dec., 1949.
17. T. Sloneczewski, A Servo System for Heterodyne Oscillators, A.I.E.E. Technical Paper 51-218, May, 1951.
18. F. W. Schramm, Calibrating Strip Type Dials, Electronics, pp. 102-3, May, 1950.

Automatic Manufacturing Testing of Relay Switching Circuits

By L. D. HANSEN

(Manuscript received May 18, 1956)

The large variety and quantity of shop-wired relay switching equipments produced by the Western Electric Company lead to the use of comprehensive and flexible manufacturing testing facilities to insure quality of product and to reduce costs. An older manual type test set is briefly described and used to illustrate the functions and operation of two automatic test sets designated as Card-O-Matic and Tape-O-Matic respectively.

INTRODUCTION

Early telephone central office installations were of the manual switch-board type which were relatively simple and required few relay circuits other than those located in switchboards themselves. Installation effort, in addition to actual erection of the switchboards, equipment frames, fuse boards and the like consisted largely of running and terminating the central office cabling. As the telephone art grew, both with the introduction of the dial telephones, and carrier and repeater equipments for long distance calls and the consequent need for interconnection of these various types of systems, a considerable variety of relay switching circuits was required.

To reduce the installation time and effort the practice of doing as much circuit wiring in the factory as possible was introduced. Relay switching units are now completely assembled, wired to terminal strips and tested in the shop. Since these are in effect working circuits the installation testing effort, after the connection of office cabling, consists largely of overall tests required to insure the proper functioning of the entire office.

Due to the wide variety and complexity of these units, many of which have optional circuit conditions that can be supplied on order and few of which have sufficient demand to justify specially designed high pro-

duction test sets for their exclusive use, adaptable manually operated test sets were first used. These sets required a high degree of flexibility in interconnecting the terminals of the circuit under test to those of the test set and in applying the proper potentials in sequence that would insure putting the circuit through its paces and checking that the switching functions are properly performed.

It should be stated here that since all apparatus components of these circuits such as relays, transformers, capacitors, inductors and resistors are tested and inspected for their respective electrical and mechanical requirements when manufactured, except in the case of some types of relays which require adjustment to meet their particular circuit requirements, the testing of switching circuits is largely confined to verification of the circuit wiring with normal voltages. Although marginal component tests are not normally applied, operation tests will, of course, detect defective apparatus components which cause malfunctioning of the circuit.

MANUAL TEST SET

Fig. 1 shows a representative manual type test set that was extensively used for wired relay unit testing before the introduction of the automatic test sets to be described later. On the left side is a pin jack field into which the numbered wires of the connecting cable can be individually plugged in order to connect the test set terminals to the proper terminals of the relay unit under test. The other end (not shown) of the cable is equipped with a contact fixture arranged to give quick electrical connections to the terminals of the wired relay unit. The plugging of the pins into the proper pin jacks is a feature needed to provide flexibility in a test set arranged to test many types of circuits and is a part of the setup operation for any one circuit. It is a slow and time-consuming operation since each lead has to be identified and plugged into the proper pin jack. The pin plug setup must be taken down and rearranged in order to test any other type of relay circuit.

The test set is equipped with signal lamps for visual response indications and manually operated keys for the use of the tester in performing the test operations. Separate power cords are plugged into power distribution jacks which supply the various potentials commonly used in telephone central offices.

After the initial setup the tester operates the numbered keys and observes the lamp signal responses in accordance with the chart clipped to the front of the test set. Failure to get a particular lamp indication



Fig. 1 — Manual wired unit test set.

requires that he analyze the circuit conditions and locate the cause of the trouble. Usually a circuit fault must be corrected before testing can proceed.

Fig. 2 shows a small portion of a simplified circuit test arrangement for such a manual test set. In this illustration a single key, when operated, supplies battery and ground potentials to the winding of a relay in the circuit under test. Assumption is made that the three relay contact terminals are wired directly to the relay unit terminal strip so that

they can be connected to ground and to battery through lamps for circuit closure indications. The switching functions of the relay can then be checked by operating the test key and observing that signal lamp (1) extinguishes and that (2) lights.

While such an arrangement can adequately test most switching circuits of any complexity by further extension of the basic scheme, when supplemented by internal circuit connections where necessary, the

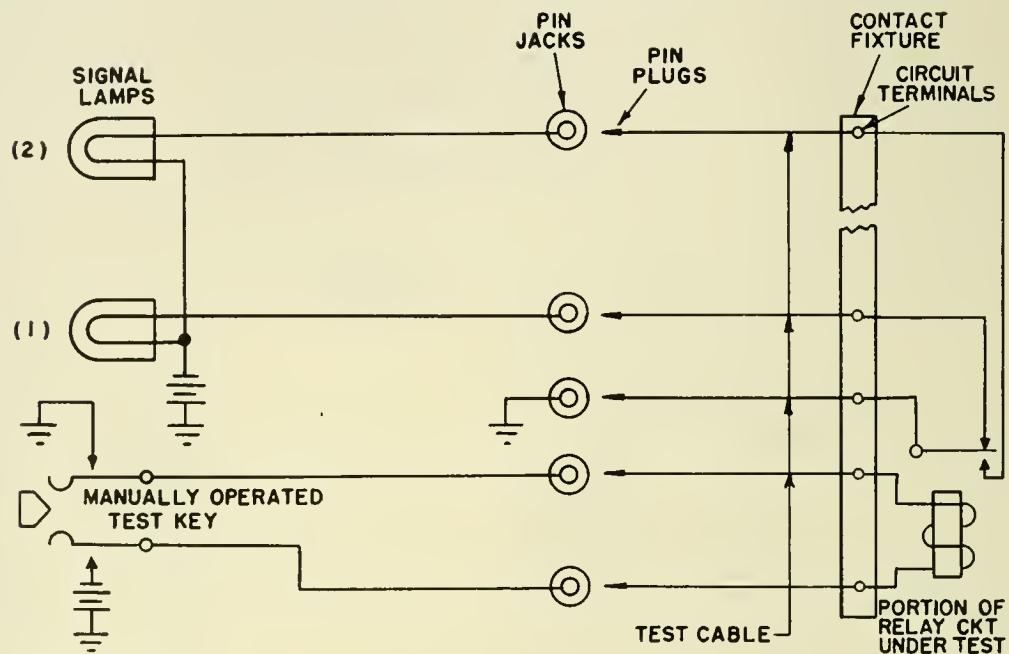


Fig. 2 — Simplified circuit sketch for manual test operation.

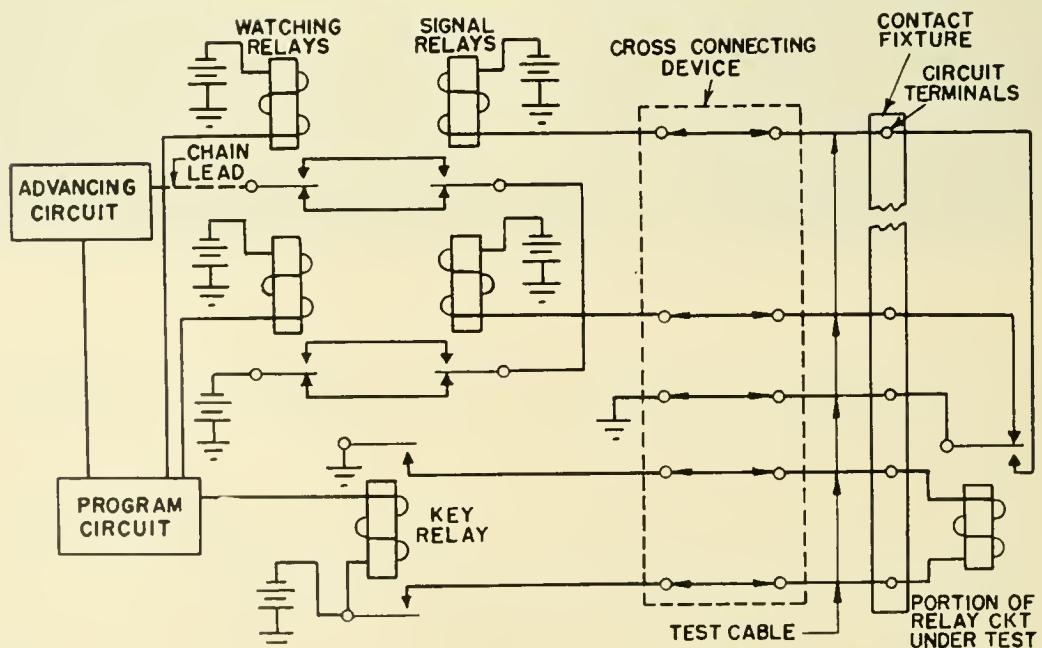


Fig. 3 — Simplified circuit sketch for automatic test operation.

system is at best a slow and laborious one which is subject to human error. Wages for testers are determined not primarily on their ability to operate keys and check the indications of lamps but on their skill in analyzing and clearing trouble conditions. If some quick and automatic means could be devised to make the initial cross connection setup, apply the potentials in the proper sequence under control of some programing device and check the circuit responses at each step a real advance in speeding up tests and reducing human errors would be accomplished. Such an automatic set ideally should have improved response indications to aid the tester in locating circuit troubles when the test set stops on the failure of meeting any test requirement.

THE AUTOMATIC TEST SET

The key and visual lamp indicating functions of the manual test set can be replaced by relays in an automatic test set which perform these operations if they are under control of suitable programing and advancing circuits as shown in Fig. 3. Here the "signal" relays operate through the contacts of the relay under test and their operating positions are checked by the "watching" relays whose contact closures must match those of the signal relays. The series path through the contacts of all signal and watching relays is called a chain lead. The program circuit establishes the positions of the watching relays to meet the expected conditions prior to operating the key relay and then any lack of continuity through the chain lead caused by failure to satisfy test conditions halts the progress of the tests under control of the advancing circuit. At this point additional contacts (not shown) on the signal and watching relays may be used to light signal lamps to convey information to the tester as to which portion of the circuit failed to operate properly.

For quick setup a pre-wired multi-contact adapter plug may be used as a cross-connection device to permit establishing the proper test connections to the unit under test. One will be required for each type of relay circuit to be tested. These, together with some means whereby the sequential operation of the programing circuit can be controlled, constitute the essential features of an elementary automatic relay switching circuit test set. How these basic features can be extended into practical embodiments will be explored further below.

THE CARD-O-MATIC TEST SET

Key equipment relay units are small switching circuits used as circuit building blocks to provide the desired optional features in conjunction with the key boxes or key-in-base telephones often seen in small

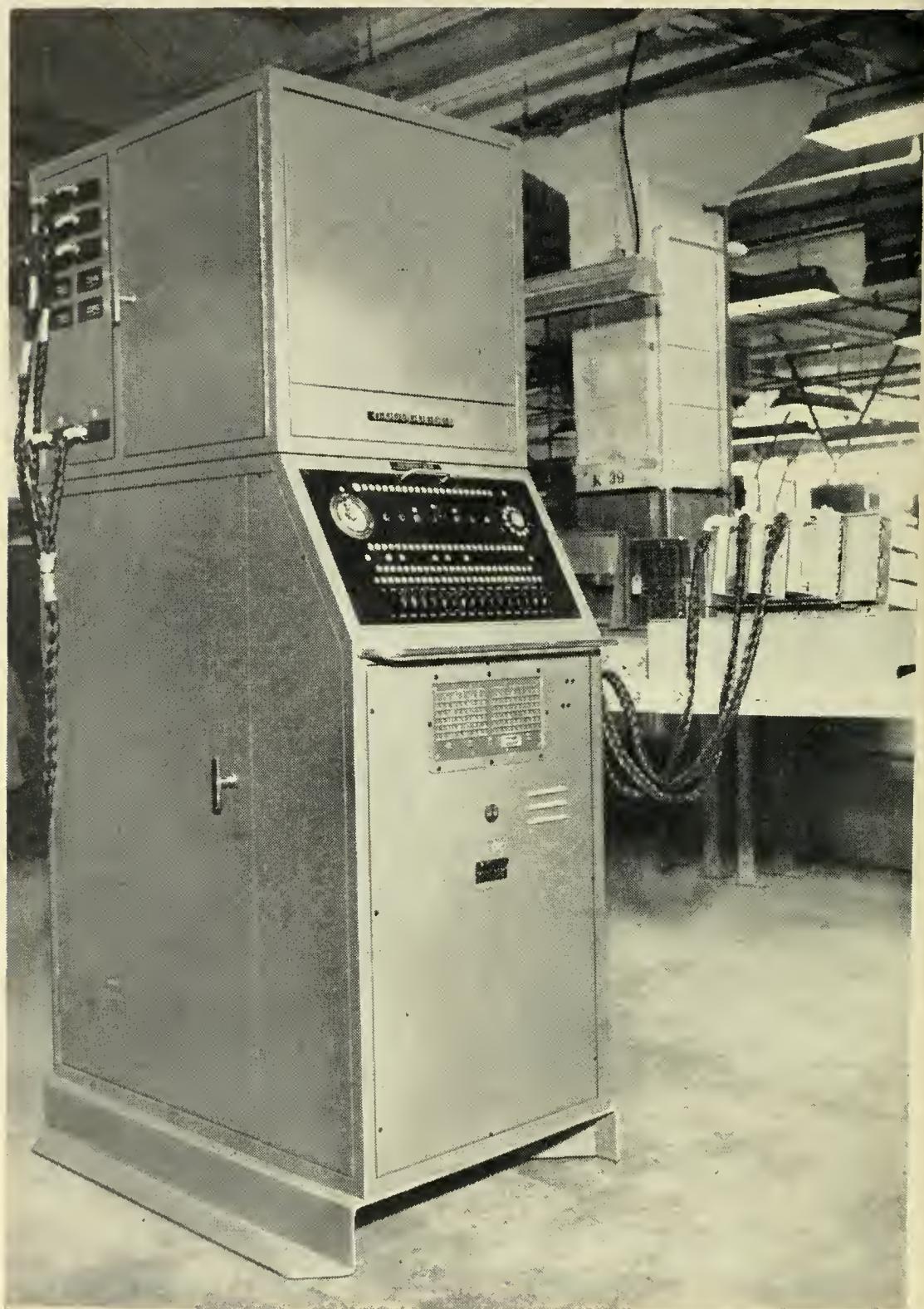


Fig. 4 — Card-O-Matic test set.

business offices to furnish the flexibility needed in answering and transferring calls. These systems are used where the number of telephones served does not warrant the use of a regular PBX switchboard.

These circuits are relatively simple but their large scale production warrants the use of high speed automatic test sets to perform the test functions and to indicate circuit trouble.

Fig. 4 shows the operating position of the Card-O-Matic* test set which was developed to test such unit assemblies. The keys shown are used to initiate and control the automatic operation of the test set and in trouble shooting. They are not to be confused with those that perform the actual testing functions described previously for the manual test set. The lamps provide indications of the progress of the tests and of the positions of the watching relays which are also needed to aid in determining the point of circuit failure. The meter type relay in the upper left corner of the operating panel provides a sensitive checking device for audio frequency tests through the voice transmission circuits. The telephone dial affords a simple means of generating any required number of pulses for operating stepping selectors on some types of units. The terminal field in the lower front of the cabinet gives the tester access to the circuit terminals of both the unit under test and the test set for his use in analyzing and locating faults. The upper cabinet was a later addition and contains the multi-contact relays needed to permit testing units with more than one circuit. The row of push buttons are used to select the circuit to be tested.

Fig. 5 is a rear view of the set that shows the perforated insulating card from which the set derives its name. The coded card controls the sequence of test operations and is hung on pins over the field of 1,000 spring plungers (20×50) as a part of the setup operation for a particular relay unit. Closing the door and screwing up the hand wheel, which is necessary to provide the force required to depress the plungers, will ground those which coincide with holes in that particular card.

Cross-connection setup of the test leads is achieved by the use of a plug-board such as is commonly used for quick change over on perforated card type business machines. Fig. 6 shows the plug board being inserted into the transport mechanism. The relatively large number of terminals are required because each of 60 test leads must be capable of being patched in to an equivalent number of terminals on a maximum of ten different circuits. Not all of our test sets are equipped with the upper cabinet since most key units have only one circuit and on these a simpler

* Patent No. 2,329,491.

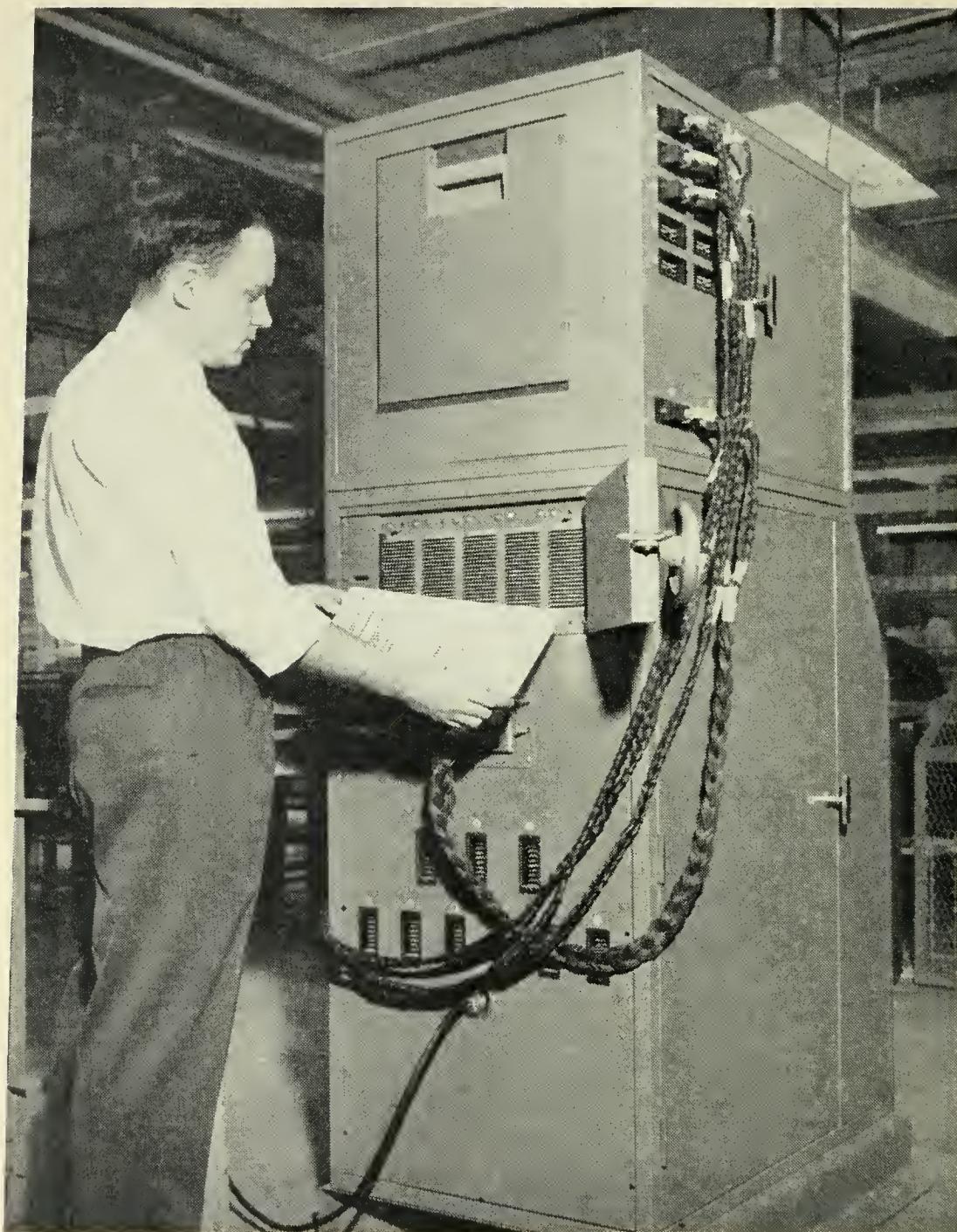


Fig. 5 — Rear view of Card-O-Matic test set showing insertion of perforated card.

cross connection fixture is plugged into the location where the lower end of the cable joining the two cabinets is shown terminated in Fig. 5.

A side view of the test set is shown in Fig. 7 to give an indication of the amount of switching equipment and wiring necessary for an automatic test set of this sort. The set is powered from a 120-volt 60-cycle

source from which are derived the 24-volt dc, 90-volt 20-cycle ringing current and 600-cycle audio tone supplies that are required. The test circuit features include tone transmission checking, dial pulsing, 90-volt 20-cycle ringing and ground and battery supplied either directly or under relay control. Other battery and ground relays are available for checking the response of the circuit under test.

These test features have been sufficient to perform operation tests on most relay units associated with key telephone systems. The test cycle is fast and the twenty test steps can be performed in approximately ten seconds. The lamp indications given when the test is interrupted by an open-circuited chain lead, convey information to the tester as to which test step is involved and when any pairs of signal and watch relays fail.

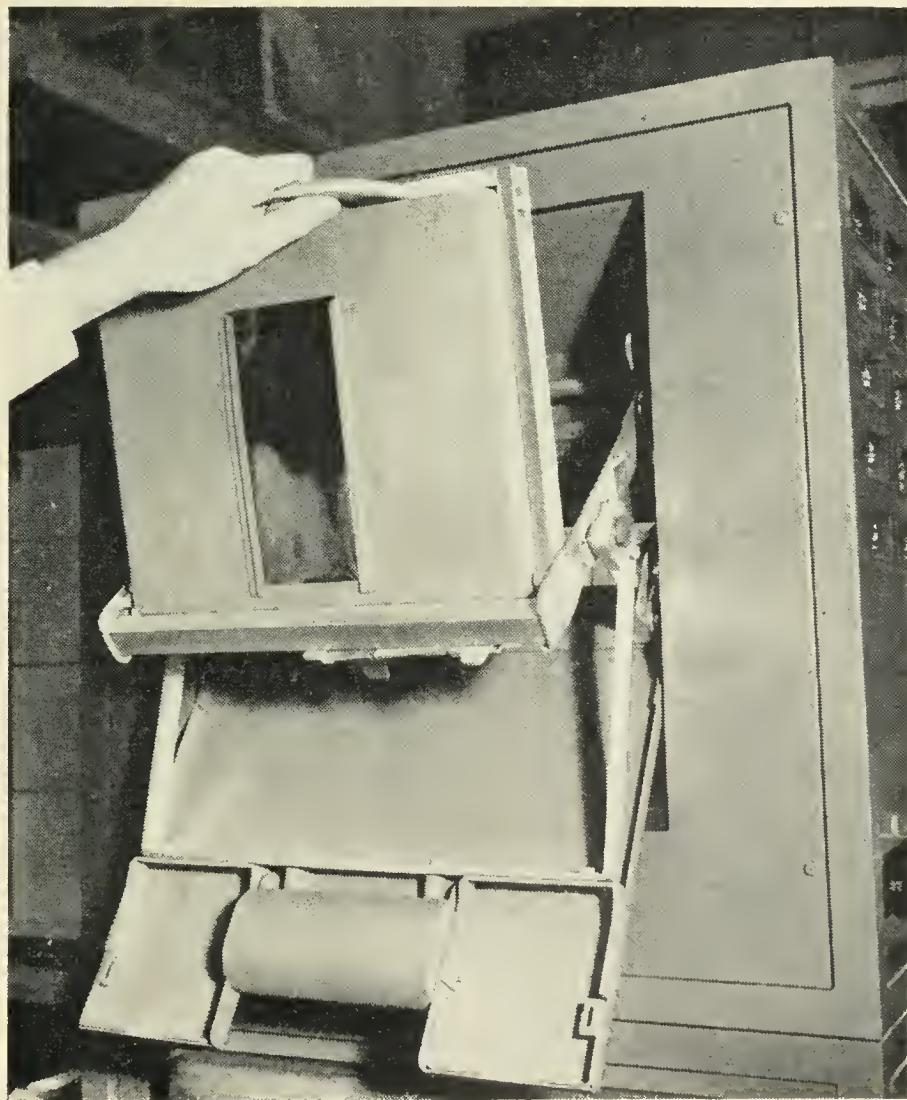


Fig. 6 — Insertion of cross-connection plug board into Card-O-Matic test set.

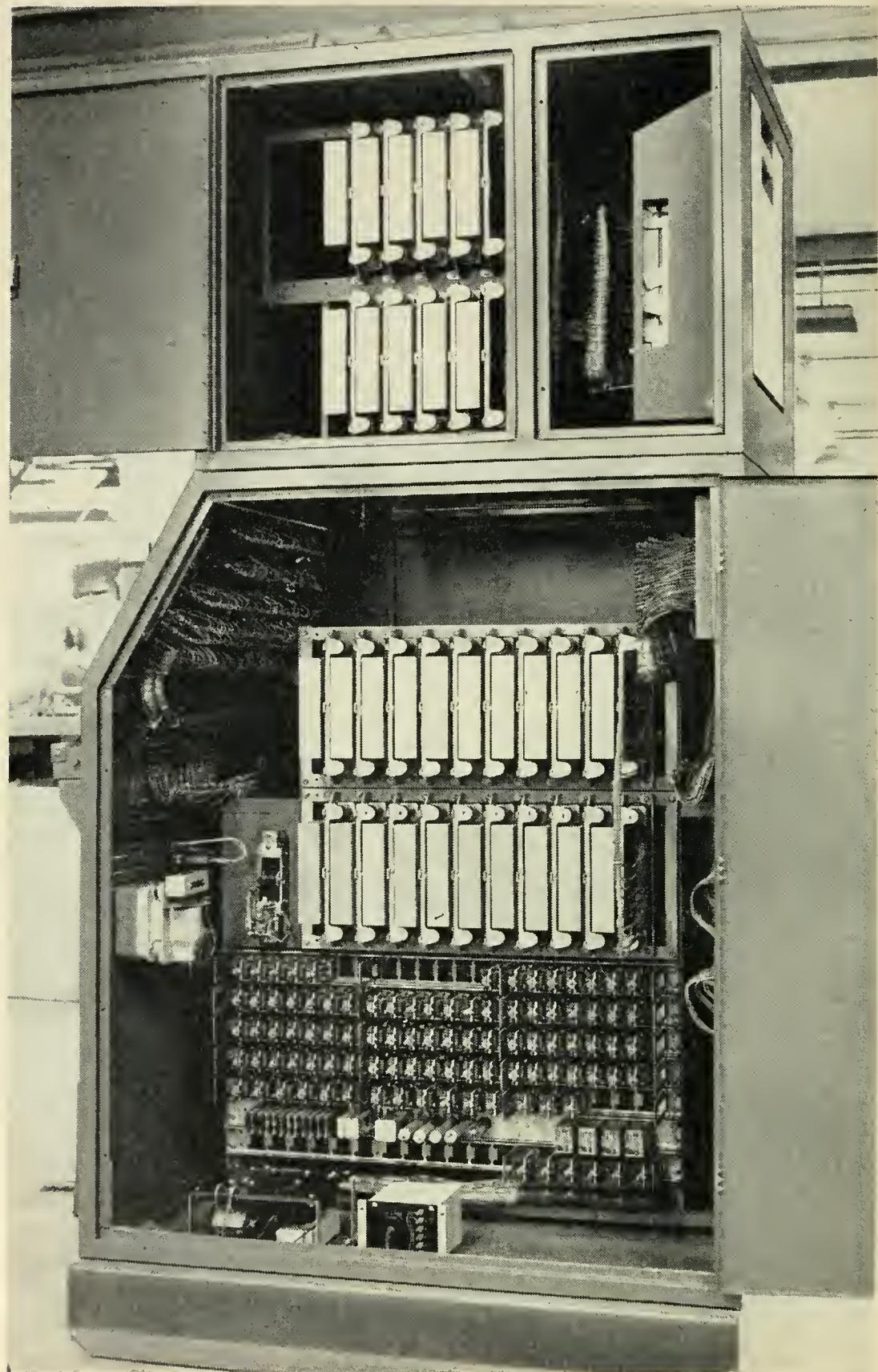


Fig. 7 — Interior of Card-O-Matic test set.

to match each other. Simplified circuit sketches which show the interconnection of test set and wired unit circuits are provided to enable the tester to determine quickly the cause of the failure.

The Card-O-Matic test set, while performing admirably on the relatively simple relay circuits within its range and capabilities, falls down on the more complicated relay switching circuits used in telephone central offices for several reasons. The most important of these are:

1. A fixed cycle within a maximum of twenty steps with any one coded card.
2. No provision for alternate or optional circuit conditions on a card.
3. The only power supplies provided to operate relays are negative 24-volt dc and 90-volt 20-cycle ringing whereas telephone office units frequently also require negative 48-volt and positive 130-volt dc as well as positive or negative biased ringing currents for party line ringing.
4. The increase of either test steps or features would increase the size of the perforated card beyond a practical size.

THE TAPE-O-MATIC TEST SET

The experience gained in the design and successful operation of the Card-O-Matic test set led naturally to the exploration of ways and means whereby a more versatile and comprehensive set could be devised. The five hole coded perforated teletype tape was selected as a cheap and flexible programing device. It afforded a means of providing a test cycle of any required length and, since the perforating and reading mechanisms were already available, it appeared to be nearly ideal for its purpose.

Consideration was given to the following desirable features all of which were incorporated in the design of the new set:

1. Provision for cross-connecting (under control of the coded tape) any test set circuit to any terminal of the circuit under test for as long as necessary and then disconnecting for reuse in later testing steps if required. This would greatly extend the range and capabilities of the set.
2. Provision for several power voltage sources which could be selected as required to meet the normal telephone office voltage requirements of the unit under test.
3. Provision for alternate or optional tests to be coded into the tape to meet the various circuit arrangements that may be wired into the unit as required by the Telephone Company who is our customer. Such optional test arrangements could be applied by the test set under the control of keys to be operated by the tester as part of the setup at the start of the tests.

4. Provision for stopping the test cycle to enable the tester to perform manual operations such as inserting a test plug in a jack on the unit or insulating relay contacts in order to isolate portions of the circuit for test simplification and to obtain a more detailed test.
5. Provision of improved lamp indications to aid the tester in clearing wiring faults or in locating defective apparatus. These would include the necessary information as to which test set circuits are connected to which unit terminals as well as which relays of the wired unit should be operated at that stage of the test cycle.
6. Provision for connecting several terminals of the unit under test together as a means of providing circuit continuity where required.
7. Provision for measuring resistance values of circuit components.
8. Provision for insertion of various resistors in battery or ground leads to control currents to desired values.
9. Provision for checking voice transmission paths through non-metallie circuits such as transformers or capacitors.
10. Provision for measuring circuit operating times in steps of approximately 100 milliseconds.
11. Provision for sending and receiving dial pulses.
12. Provision for a single code for releasing all test connections and conditions previously established by the coded tape as a means of quick disconnect. This is in addition to the release of individual connections mentioned in (1) above.
13. Provision for audible and visual indications of completion of a successful test cycle.

Through the use of two letters (each of which has its own combination of the five holes) for each signal it was possible to obtain the over 500 codes required to control all test and switching functions even though the teletype keyboard has only 32 keys. The only Teletype transmitter (tape reader) available when the test set was first designed operated at a speed of 368 operations per minute and was arranged for sequential read out on two wires by means of a commutator. Conversion to five wire operation and removing the commutator permitted reading each row of holes simultaneously. The gearing was also changed to permit 600 operations per minute but even so the hole reading contact dwell time was increased from approximately 20 milliseconds to 70 milliseconds for more reliable operation with ordinary telephone relays.

The machine which was designated as the Tape-O-Matic* test set, is shown in Fig. 8 in operation on a typical wired relay unit mounted in its shipping frame. The contact fixture is attached to the unit terminal

* Patent No. 2,328,750.

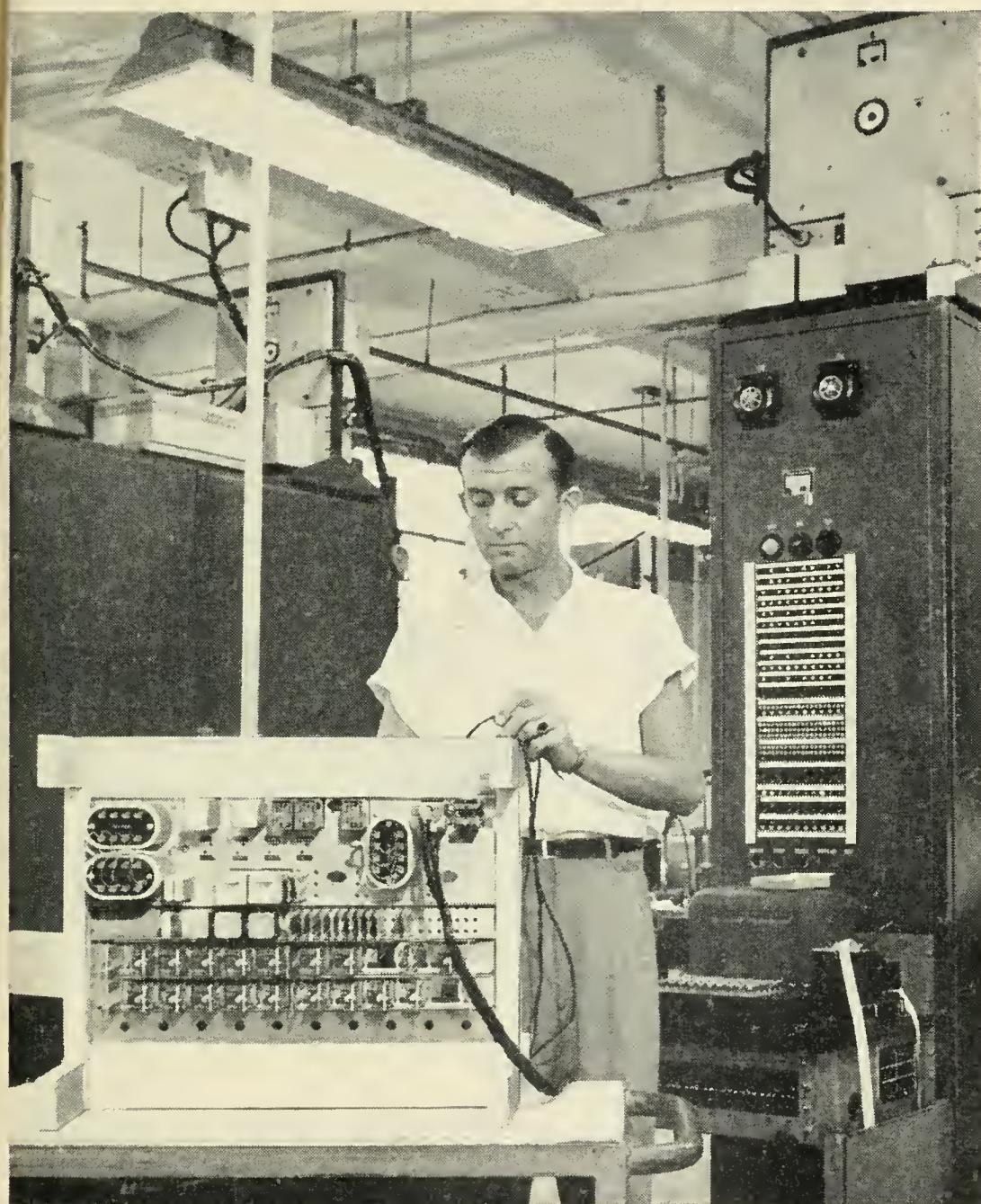


Fig. 8 — Tape-O-Matic test set in operation:

strip and cabled to a gang plug which in turn is plugged into a receptacle behind the operator. These leads are extended through a duct to the metal enclosure at the base of the set for entry to the test set proper. The coded tape is dropped into the receptacle at the side of the key shelf to which it returns after its traverse through the reader. A row of circuit breakers on the end of the key shelf control the application of

and provide protection for the various power supplies. Two of these supplies are mounted on the top of the set.

The rows of vertical push button keys on the key shelf afford the tester a means of determining (for trouble shooting purposes) the association (through lamp display signals) of the wired unit circuit terminals with those of the test set and the corresponding test voltages which are connected at that particular stage of the test. The lamp display panel also indicates which test set circuits are in use and through fast or slow (0.5 or 1 second) flashes whether the fault thus indicated is the result of a failure to meet either an expected condition or the occurrence of an unexpected condition. This feature is illustrated in Fig. 9 which shows one link of the chain leads which extend through all pairs of signal and watching relays for the check of satisfaction of all test conditions and the application of steady or interrupted ground to the associated test feature lamp. The operating condition of all test set key relays as previously established by the tape is also indicated by the display lamps. Another type of information obtained from the lamp display panel which is valuable to the tester in trouble clearing is the indication of the particular unit relays which should be operated at that part of the test cycle. By checking the lamps against the operated or non-operated position of the relays he can frequently localize the fault in a minimum of time.

As mentioned above an important part of the test set flexibility is the ability of the tester to set up the test set to test only those optional circuit arrangements which are provided in any particular unit ordered

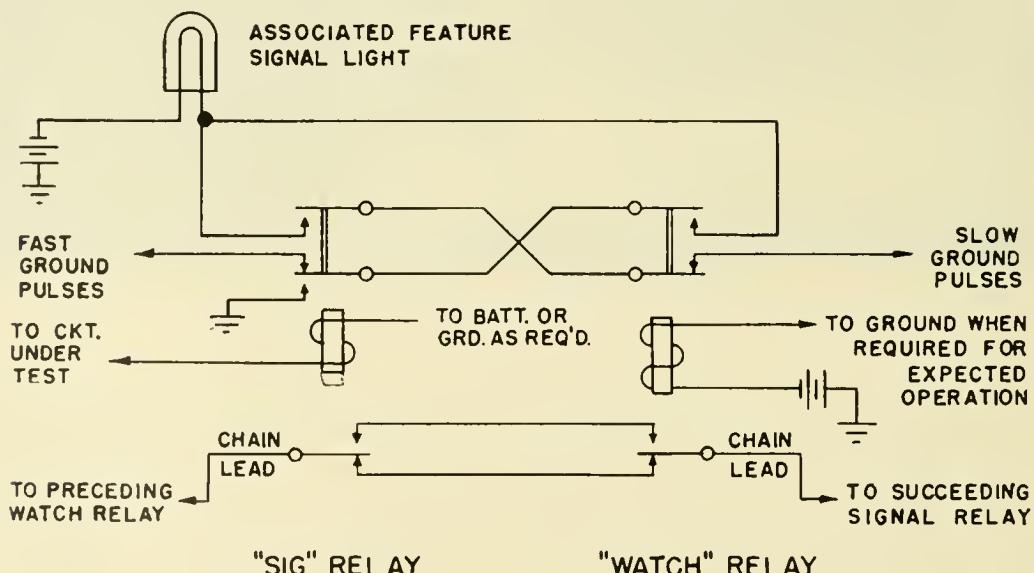


Fig. 9 — Chain circuit showing watching relay function.

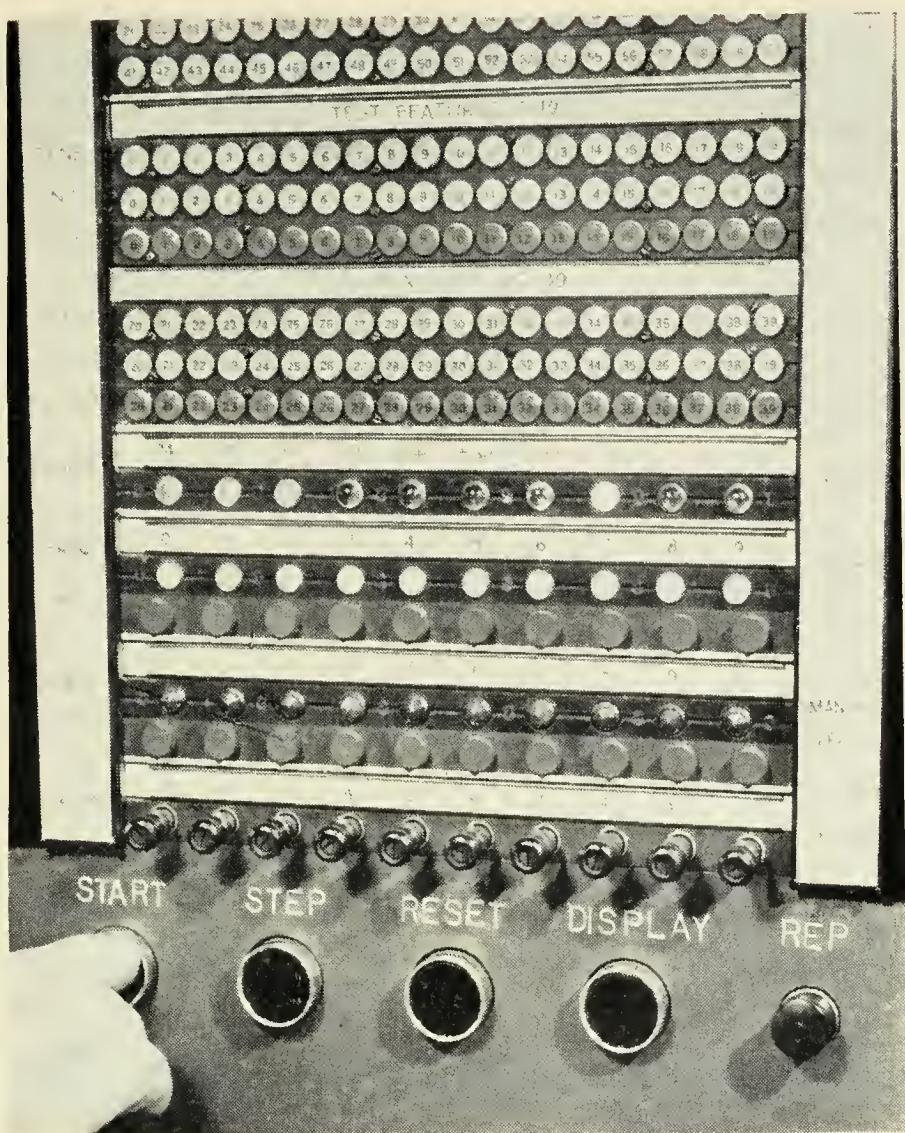


Fig. 10 — Lower portion of lamp display panel.

by the customer. Failure to provide this would result in fixed test cycles and many more tapes, which might be similar but varying only in regard to the options, would have to be prepared. Figure 10 shows the lower portion of the lamp display panel with the push-pull option keys on the bottom row. Directly above are the manual operation keys with their associated lamps which the tester must operate to cause the test set to resume the testing cycle after it has stopped for him to perform a manual operation.

A side view of the interior of the set is shown in Fig. 11. Two bays each facing the opposite direction from the other are housed within the cabinet and are used for mounting the crossbar switches and telephone type relays which are the principal circuit components. Two doors on

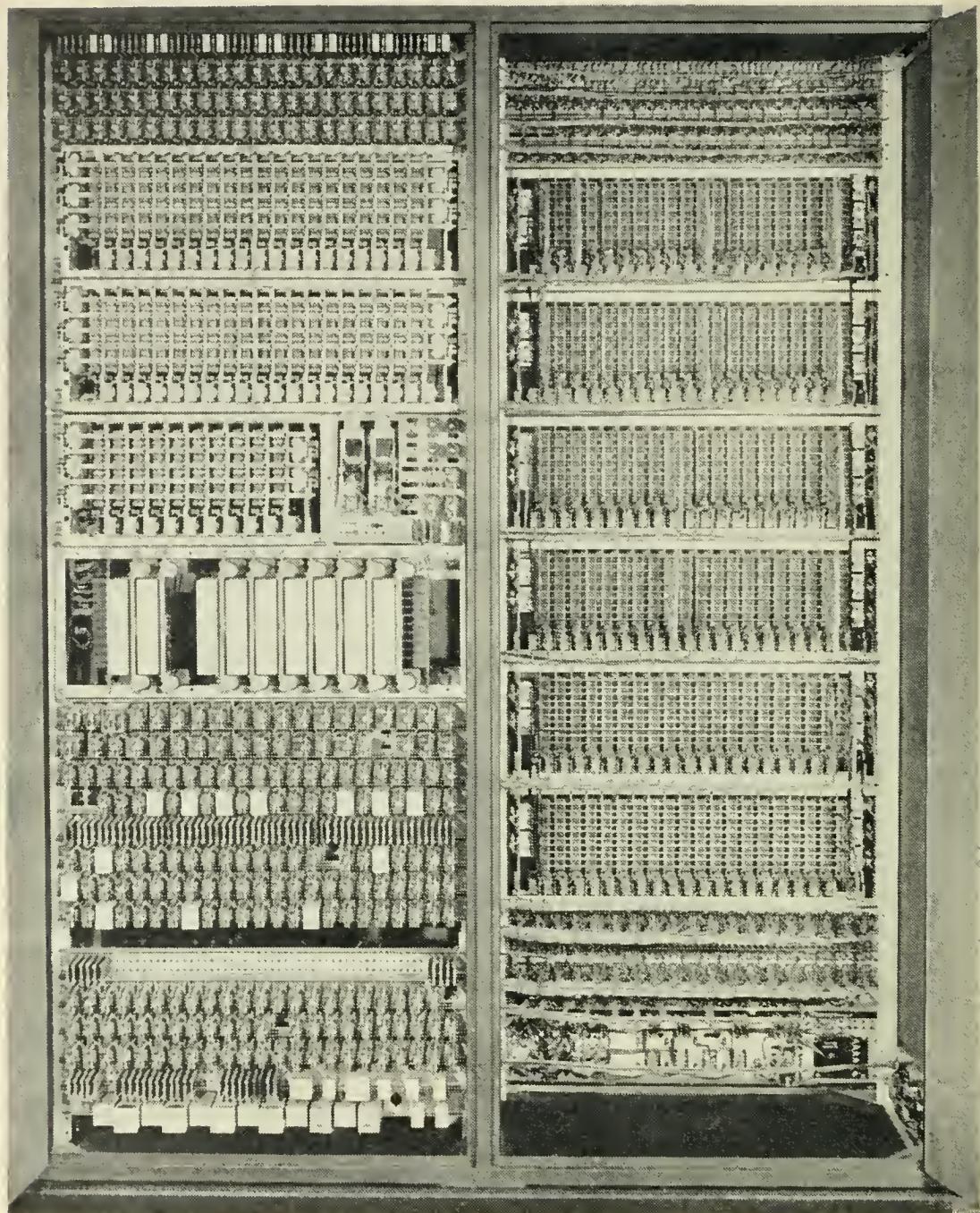


Fig. 11 — Interior of Tape-O-Matic test set.

each side give convenient access to all wiring and apparatus for maintenance purposes.

A fairly large portion of the mounting space is occupied by the cross-bar switches which perform the functions of interconnecting the circuit terminals of the unit under test and those of the test set. They also connect the proper voltages to these circuits. The switching plan Fig. 12

shows in abbreviated diagrammatic form that the unit terminals 0-99 appear on the horizontal inputs of the two 10×20 and one 10×10 switches that comprise the primary group. The horizontal multiple of these switches are split so that each section runs through five verticals to afford connection to each of the hundred unit terminals.

The vertical outputs of the primary switches are connected to the horizontal inputs of the two 10×20 secondary switches. The horizontal multiple of these switches are split so that each section runs through eight verticals. The verticals of the secondary switches are linked to the horizontals of the two 10×20 tertiary switches which have their

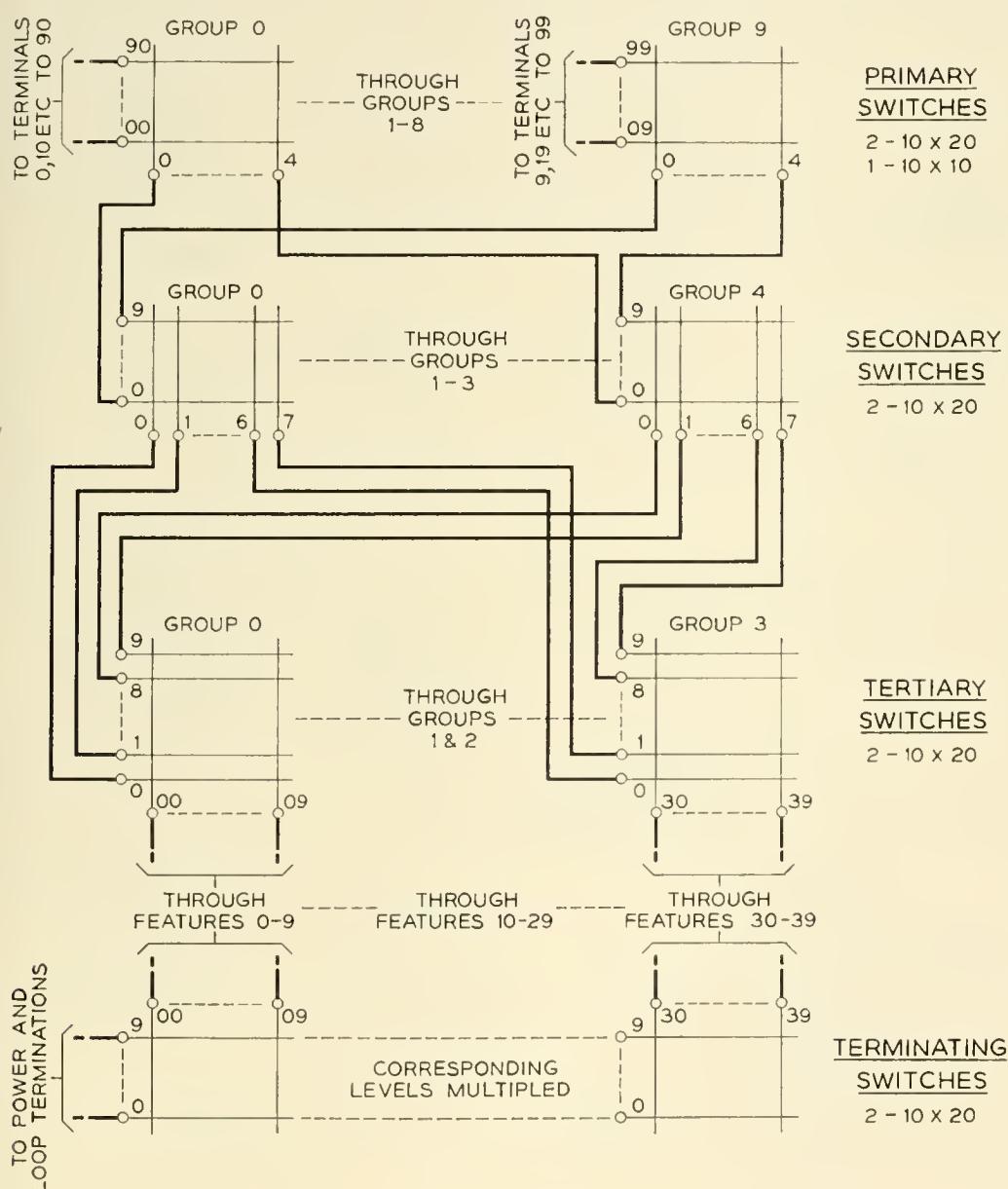


Fig. 12 — Switching plan.

multiple split into groups of ten. The 40 verticals of the latter are connected directly to the 40 test set feature circuits designated 0-39.

Two additional 10×20 crossbar switches perform the function of connecting any of the five power or five multiple terminations to any of the forty test set features. These terminations are comprised of 5 loops and one each of ground, negative 24 volts, negative 48 volts, 90-volt 20-cycle ringing current and positive 130 volts.

Thus it can be seen that, through proper operation of the primary, secondary, tertiary and terminating crossbar switch cross points, a path can be established from any circuit terminal to any test set feature and supplied with any of the available power or loop terminations. It is

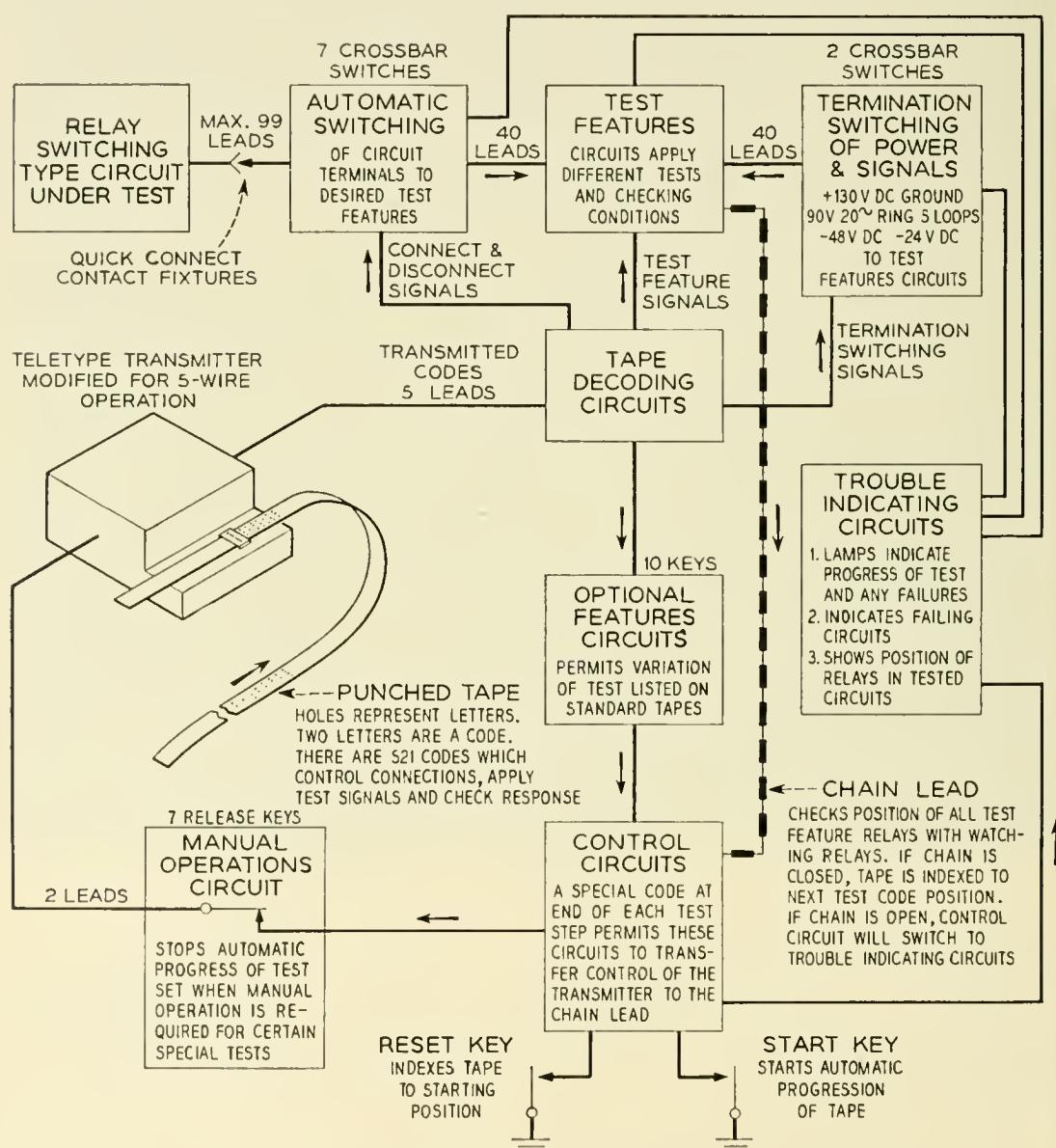


Fig. 13 — Block schematic.

also apparent that several paths can be found that will satisfy any one switched connection. Paths are assigned in sequence by a series relay loop circuit. The entry point on this circuit is changed periodically to distribute wear on the relays and switch cross points.

Although only one lead for the switched circuit is shown for each cross point in Fig. 12 there are actually four leads through corresponding pairs of contacts through each cross point. The remaining leads are associated with the holding and signalling functions of the switch.

The block schematic (Fig. 13) shows the principal functions which must be included in an automatic test set of this sort. A somewhat more detailed schematic is presented in Fig. 14 in order to show the functions of the forty test features 0-39. These are tabulated in Table I.

The coding of the two letter combinations in the tape must follow a definite sequence in order that the machine may recognize and act on the information it receives. This sequence is as follows:

1. Code FW to stop the tape at the end of the reset cycle after which tests will proceed when the start button is pressed. This is the first code on all tapes.

2. Codes to set up crossbar switches to connect each circuit terminal to its proper test set terminal and the proper termination. Knock down or release codes may also be sent.

3. Codes to operate or release "Key" relays. These relays are shown without windings in Fig. 14.

4. Codes to operate or release the watching relays associated with the "Signal" relays which are shown with windings in Fig. 14.

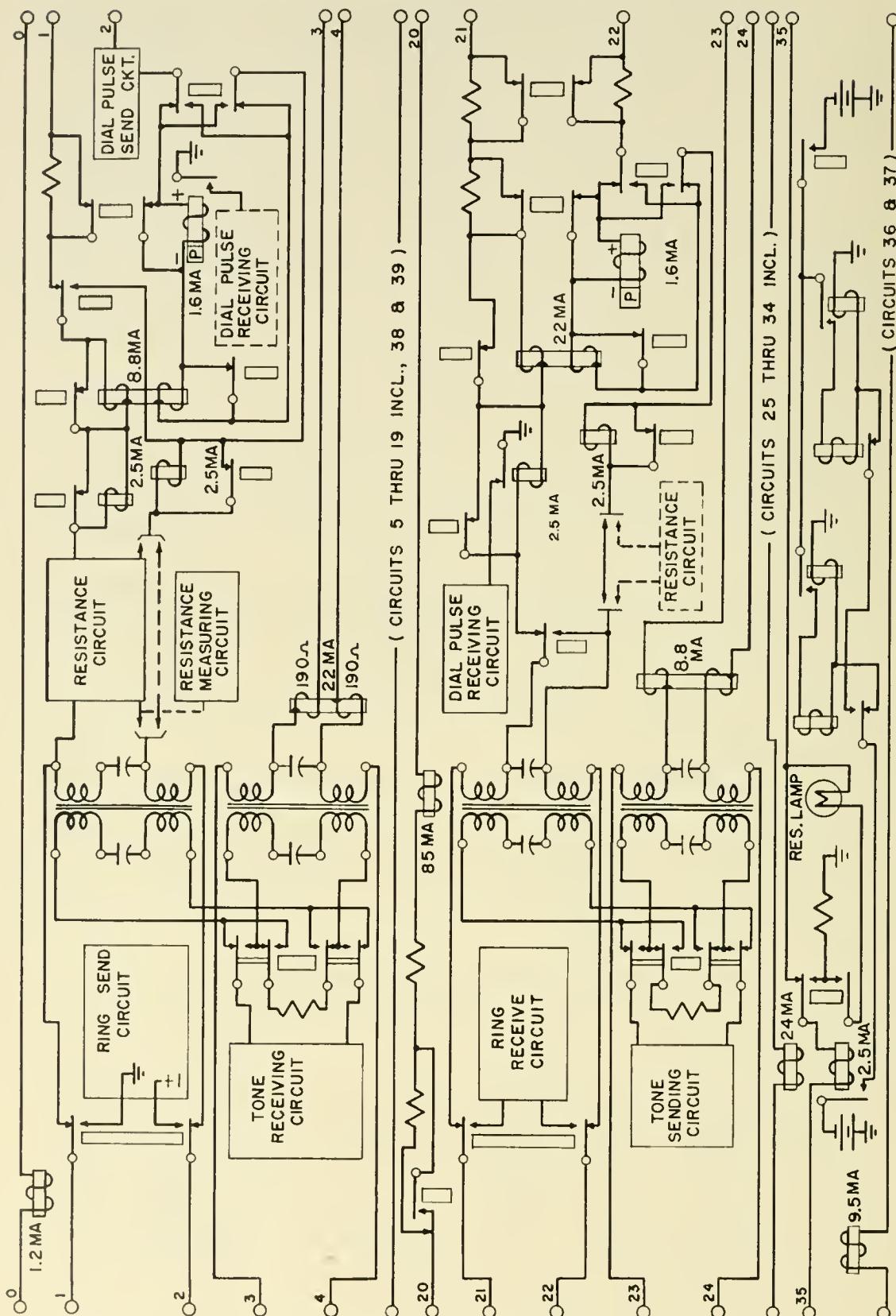
5. Codes to operate or release relays controlling the lamps associated with relays in the circuit under test to aid in trouble shooting.

6. Codes to delay the timing out interval up to a maximum of ten seconds.

7. Code FJ which checks the matching of all signal and watching relays through the chain circuit for satisfaction of all test conditions being applied.

In addition to the above, additional codes can be inserted after each FJ test signal to stop progress of the test to permit the tester to perform some required manual operation. After completion of this step he presses a button associated with that operation and the test proceeds. Option codes can also be inserted at the beginning and end of each testing step to permit bypassing of that part of the tape if the corresponding option keys are operated at the beginning of the test. A common knock down code FR can be inserted at any time to release all connections and relays for a quick disconnect and make all test set features available for

THRU CROSSEBAR SWITCHES TO TERMINATIONS AS REQUIRED



THRU PRI., SEC. AND TERTIARY CROSSEBAR SWITCHES TO RELAY UNIT TERM. 0-99 AS REQ'D

reuse. A final code SC must be put in every tape to operate the OK lamp and gong if a successful test cycle has been performed or conversely to indicate that the tape should be re-run if trouble has been found and cleared during the test cycle to be certain that no new faults have been introduced.

Preparation for testing a particular wired relay unit requires only the selection of a test cable one end of which is equipped with a suitable contact fixture for attachment to the unit terminal strip and the other with a gang plug for connection to the set. The proper tape is selected from a nearby file cabinet and inserted in the gate of the tape reader as shown in Fig. 15. The tape is stored in a cardboard carton $3\frac{3}{4} \times 4$

TABLE I

Feature Numbers	Description of Functions
0	High sensitivity relay circuit. Simulates 1,800-ohm sleeve circuit for busy test and general continuity through high resistance circuits.
1 and 2	Simulates the distant tip and ring terminations of a subscriber or exchange trunk. Provides for ringing, tone receival, dial pulse sending, line resistance, high-low or reverse battery supervision, pad control, continuity, and resistance verification.
3 and 4	Auxiliary tip and ring circuit for holding, checking continuity, receival of tone on four wire or hybrid coil circuits. Loss range of less than 0.5 db, 0.5 to 1.5 db, 1.5 to 6 db and 6 to 15 db can be checked.
5 through 19	Direct connections for supplying any of the ten terminating conditions.
20	Simulates low or medium resistance sleeve circuits for marginal tests.
21 and 22	Simulates the local tip and ring terminations of a switchboard or trunk circuit. Provides for ringing and dialing receival, high-low reverse battery supervision, transmission pad control, tone transmission, continuity and resistance check by balancee.
23 and 24	An auxiliary tip and ring circuit for holding, checking continuity, tone transmission on four-wire hybrid coil circuits.
25 through 34	Low sensitivity relay circuits for general continuity checking.
35	A circuit for checking balance on the (M) lead of composite or simplex signalling circuits and for checking receival of none, one or two pulses.
36 and 37	Medium sensitivity relay circuits for continuity checking.
38 and 39	Direct connections for supplying any of the ten terminations.

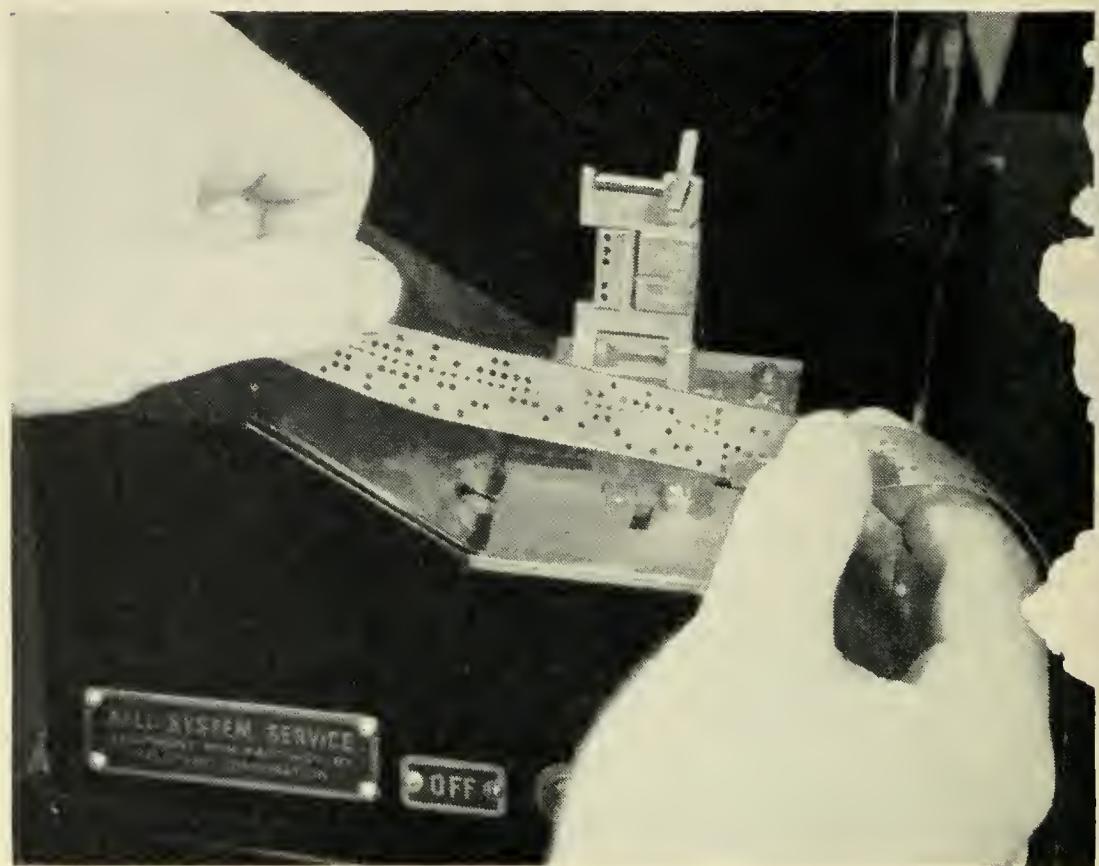


Fig. 15 — Perforated tape being inserted in reader.

inches in size, the label of which carries all pertinent information required for setup of the option keys, preliminary tests and manual operations during test. A separate 12 conductor cable equipped with individual test clips permits connection to internal parts of the circuit if needed for adequate tests. No other information than that on the box label, the circuit schematic and the lamp panel display is needed by the tester to operate the test set and to analyze and locate circuit faults when they occur.

With the tape inserted, the test connections established and any preliminary operations performed the tester has only to push the RESET button to index the tape to the initial perforation on the tape and the START button to initiate the test cycle. The set will continue to operate until either a circuit trouble is encountered or a manual operation must be performed. After a defect has been repaired, the automatic progression of the tape is again started by the momentary depression of the STEP button. When a manual operation is performed the tape is restarted by the momentary depression of the red button associated with the lighted manual operation lamp signal.

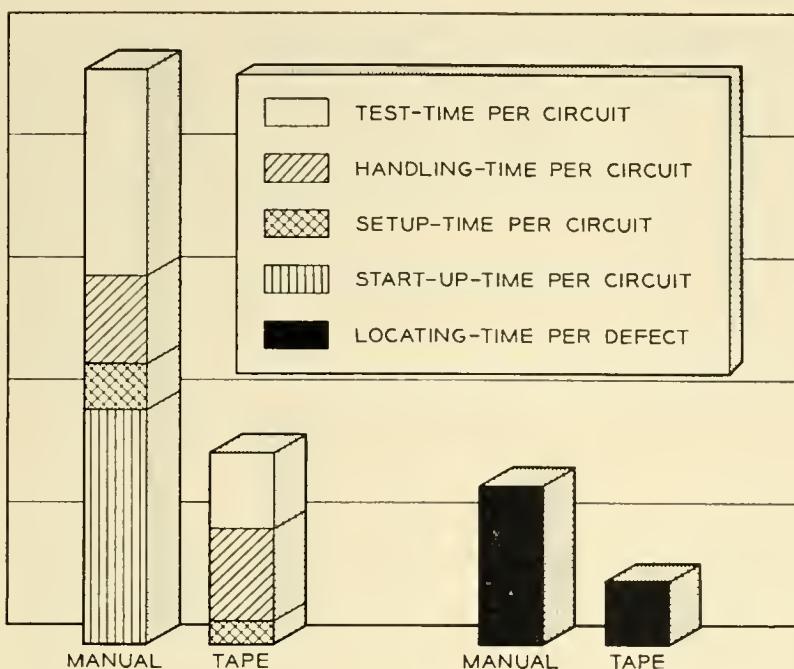


Fig. 16 — Comparison of manual and Tape-O-Matic test operation times.

As might be expected the easy setup, automatic testing and superior trouble indicating features of the Tape-O-Matic test set have materially improved the quality and reduced the testing time and effort required for wired relay units as compared to the older manually operated sets. The average time per circuit for six representative units are shown graphically in Fig. 16. One time consuming operation on manual testing is the start up time allowance for reading and understanding the written test instructions which has no counterpart in the Tape-O-Matic tests and this alone represents a sizeable gain. The handling time of the unit itself is the only operation which is not reduced in automatic testing.

HISTORY

The initial Card-O-Matic test set was installed in 1938 in the Western Electric, Kearny, New Jersey plant. Post war and subsequent expansions of production levels have necessitated construction of six more sets of improved design of the type described earlier in this article.

The first three Tape-O-Matic test sets were built in 1942 for the Wired Relay Unit Shop and additional sets have since been constructed to bring the number to twenty-six including six that are used in testing trunk units in the Toll Crossbar Shop. They have performed admirably with few changes from the initial design. They have been used to test well over a million wired units with a minimum of maintenance. This

may be accounted for, in part, by the fact that most of the component parts are telephone type apparatus designed for heavy duty use.

A maintenance feature is the use of 18 specially coded tapes which, together with a properly strapped input plug, permit the maintenance technician to obtain indications on the lamp display panel of the performance of the set.

Nearly three thousand tapes have been coded to date. Of these approximately two thousand are in active use on the many types of wired relay units made at the Kearny plant. More tapes are being added weekly as the Bell System telephone plant grows in size and complexity.

CONCLUSION

Automatic testing of wired relay switching circuits has been successfully applied to the manufacture of these equipments at the Kearny, New Jersey, plant of the Western Electric Company for a number of years. Even though the total production is large, manufacture is essentially of a job lot nature due to large number of types made and is further compounded by the optional circuit arrangements that may be ordered. The solution to the problem was found through provision of flexibility in programing and cross connection leading to quick setup, rapid testing and improved transmittal of essential information to the tester to aid him in clearing circuit faults.

Automatic Machine for Testing Capacitors and Resistance-Capacitance Networks

By C. C. COLE and H. R. SHILLINGTON

(Manuscript received May 8, 1956)

The modern telephone system consists of a variety of electrical components connected as a complex network. Each year, millions of relays, capacitors, resistors, fuses, protectors, and other forms of apparatus are made for use in telephone equipment for the Bell System. Each piece of apparatus must meet its design requirements, if the system is to function properly. This article describes an automatic machine developed by the Western Electric Company for testing paper capacitors and resistance-capacitance networks used in central office switching equipment.

INTRODUCTION

The capacitors discussed in this article are the ordinary broad limit units made with windings of paper and metal foil, packaged in a metal case. They include both single and double units in a package, connected to two, three, or four terminals. The networks consist of a capacitor of this same type connected in series with a resistor.

The testing requirements for capacitors include dielectric strength, capacitance, and insulation resistance. These same tests plus impedance measurements are specified for networks. In general, requirements of the kind involved here could be adequately verified by statistical sampling inspection. However, in equipment as complex as automatic telephone switching frames, even the minor number of dielectric failures that would elude a properly designed sampling inspection would result in an intolerable expense in the assembly and wiring operations. While engineering considerations thus called for a detailed inspection for dielectric breakdown, it was recognized that detailed inspection of the other electrical requirements could be obtained at no additional expense for labor with automatic testing machines.

DESIGN CONSIDERATIONS

In the development of this machine, the designer was faced with the same problems that obtain in the conception and design of any unit of complex equipment. These included the economic feasibility, reliability, simplicity, and versatility of such a machine.

Economic Feasibility

This can be determined by comparing the cost of performing the operations to be made by the proposed machine with the cost by alternative methods. Estimates indicated that the cost of the machines could be recovered within two years by the saving in labor that would be effected.

Reliability

Reliability has two connotations, (1) freedom from interruptions of production because of mechanical or electrical failure and (2) consistent reproducible performance. A rugged mechanical design combined with the use of the most reliable electrical components available is necessary. In addition, safeguards are required to protect the equipment from mechanical or electrical damage. To achieve consistent reproducible performance, it is important that testing circuits of adequate stability be used. Besides, it was recognized that each circuit should be so arranged that in case of a circuit failure, there would be immediate and positive action by the machine to prevent acceptance of defective product. All circuits are designed to provide positive acceptance. This means that the machine must take action to accept each item of product at each test position. In the case of the dielectric strength tests, a self-checking feature is included.

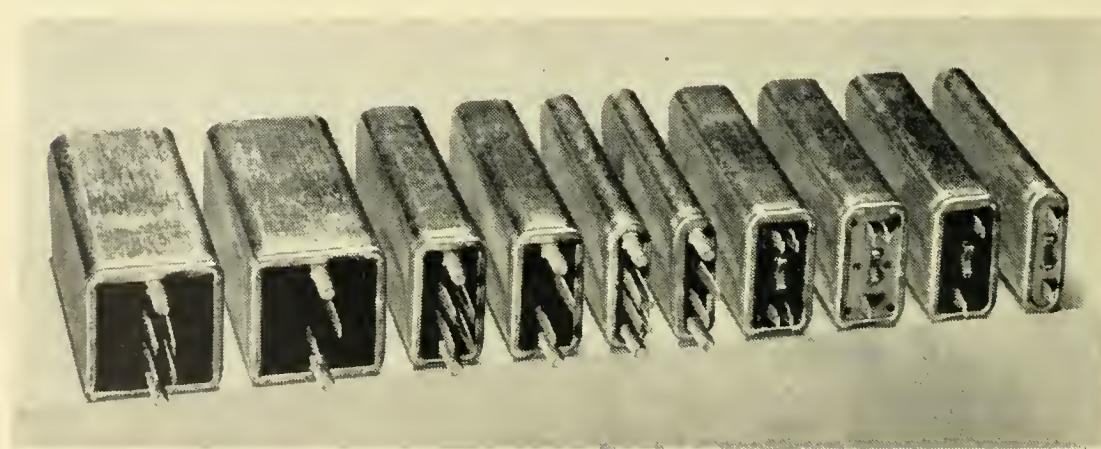


Fig. 1 — Types of capacitors and networks tested.

Simplicity

This type of equipment is operated by non-technical personnel. To minimize the possibility of improper operation of the equipment, it is important that adjustments and judgment decisions by the operator be minimized. From a production standpoint, it is important that the machine be designed to permit quick changes to handle the variety of product to be tested. All "set-ups" are made by the operator and the switching of circuits and changing of contact fixtures are simply and easily done.

Versatility

The product tested by this machine includes a variety of physical sizes and terminal arrangements with a wide range of electrical test requirements (Fig. 1).

a. *Physical Sizes.* The aluminum containers for this type of capacitors and R.C. Networks all have the same nominal length and width but are made in three different thicknesses.

b. *Terminals.* The product is made with terminals of two different lengths, two different spacings, and four different patterns connected in eight combinations. It is necessary to provide contact fixtures and switching facilities to handle all of these combinations.

c. *Electrical Tests*

(1) Dielectric strength tests are made between terminals, and between terminals and can, on single unit packages. Two-unit packages require an additional test between units.

(2) Capacitance: The capacitance of the product to be tested ranges from 0.02 mf to 5.0 mf or any combination within this range in one- or two-unit packages with no series resistance in the case of capacitors, but with a series resistor from 100 ohms to 1,000 ohms in the case of networks. This problem is discussed in more detail in the description of the capacitance test circuit.

(3) Insulation Resistance: The minimum requirements vary from 375 megohms to 3,000 megohms.

(4) Impedance: The RC networks have impedance requirements at 15 kc that range from 100 ohms to 1,000 ohms.

MECHANICAL ASPECTS OF TESTING MACHINE

Packaging of the product precludes a magazine type of feed because the variety of terminal combinations associated with two-unit packages necessitates orientation in the contact fixtures that can not be done by



1. HANDWHEEL FOR POSITIONING TEST FIXTURES.
2. ROTARY FEED MECHANISM.
3. PRODUCT PASSING ALL TESTS EJECTED FROM FIXTURE.
4. INSULATION RESISTANCE TEST PANEL AND TERMINAL COMBINATION "SETUP" SWITCHES.
5. CABINET HOUSING TEST CIRCUITS.
6. CONTAINERS FOR REJECTED PRODUCT.

Fig. 2 — Testing machine in operation.

mechanical means. A turret type construction is used to permit one operator to perform both the loading and unloading operations.

Fig. 2 shows this machine in operation. The networks or capacitors are fed into the fixtures by an operator and as the turret carries the fixtures past the feed mechanism, rollers on the feed mechanism are synchronized with the fixtures and the roller forces the unit under test into the contact fixture against a spring loaded plunger to make contact with the fixture contact springs. Also, synchronized with the feed mechanism is the closing of the gripper hook on the bottom end of the can containing the unit under test.

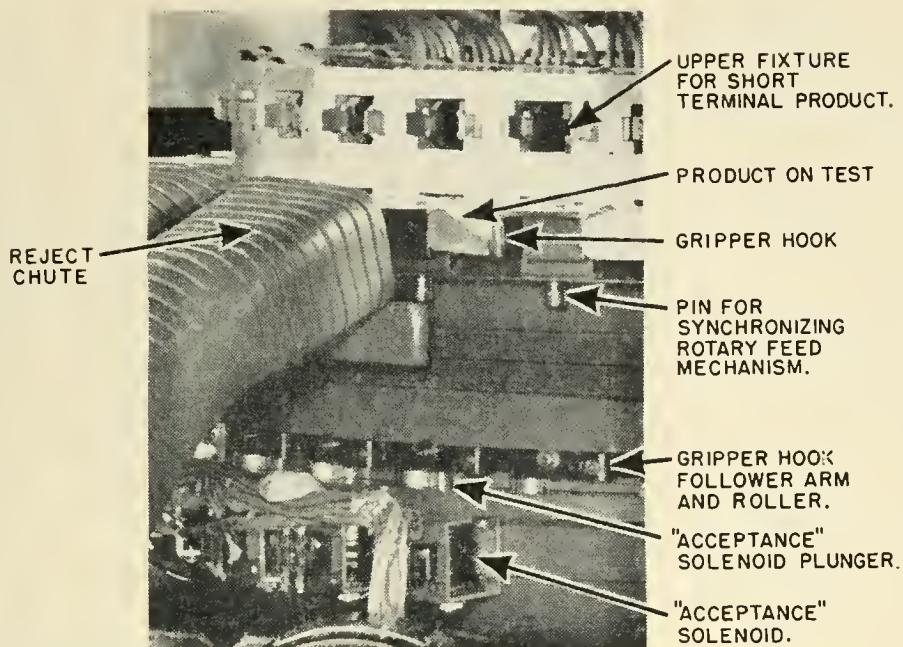


Fig. 3 — View of rejection and acceptance mechanisms.

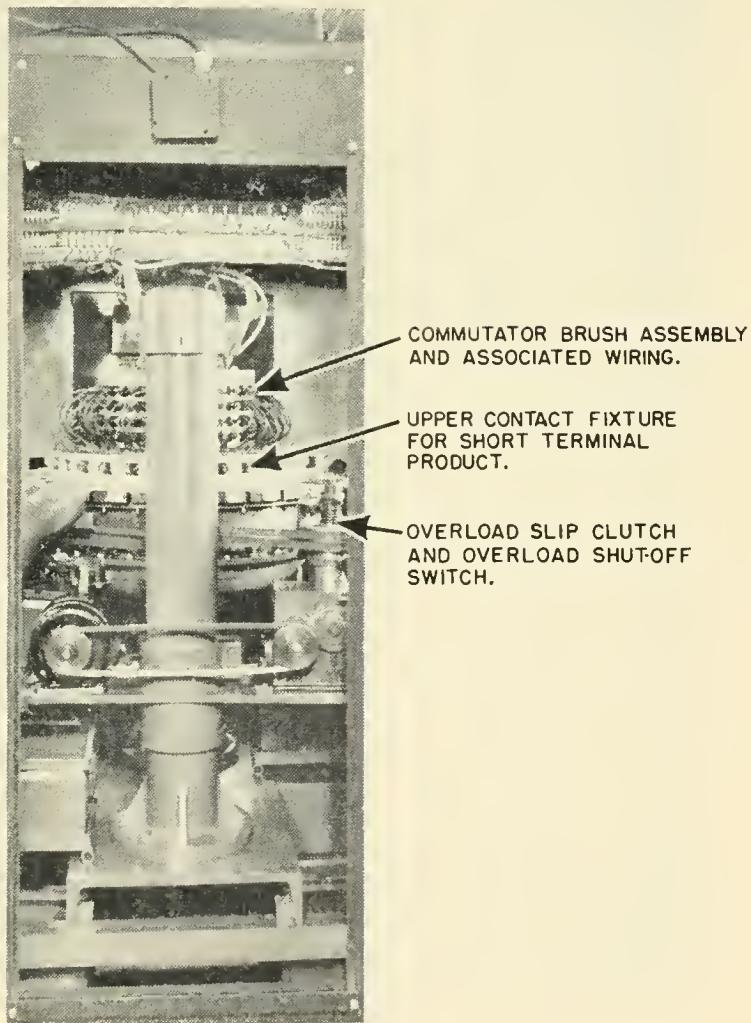


Fig. 4 — View of turret.

The acceptance or rejection of a unit under test at any one of the six test positions depends on whether the test on the unit energizes the "acceptance" solenoid associated with that test position. The gripper hook, which locks the unit under test in the contact fixture, is connected to a release shaft, follower arm, and roller (see Fig. 3). The roller rides in a track in which the plunger of each "acceptance" solenoid lies unless removed by energizing the solenoid from its associated test circuit. In the case of a defective unit, the acceptance solenoid is not energized and the roller in passing over the plunger of the "acceptance" solenoid trips the gripper hook and the spring loaded plunger in the contact fixture ejects the defective unit. Units that pass all tests are ejected on a turntable to the left of the operator from which they are stacked in handling trays by the operator.

The turret assembly includes the test fixtures, the gripper hooks and associated release shaft, follower arm and roller, and the brush assembly

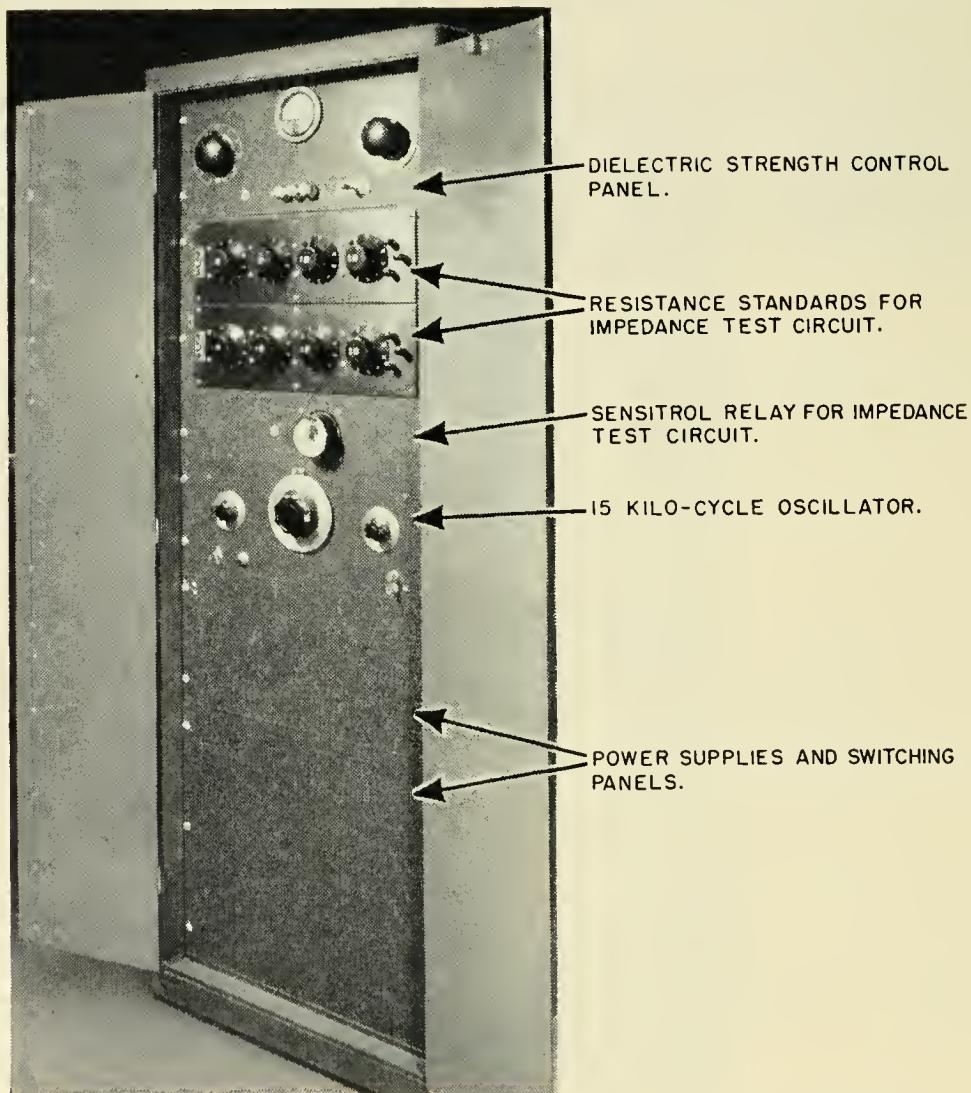


Fig. 5 — Control panels for dielectric strength and impedance tests.

connected to the test fixtures. The commutator is stationary and its segments are connected to the test circuit through permanent wiring. Fig. 4 shows the turret. Each fixture has two sections, one above the other, with the contacts wired in parallel. The lower section is designed for making contact to stud mounted units with long terminals and the upper section for strap mounted units with short terminals. To change the machine "set-up" from one fixture to the other, the turret assembly is raised or lowered by means of the hand wheel, shown on Fig. 2, located at the right of the operator. This feature was incorporated in this machine to facilitate rapid "set-up" which is essential for testing small lots. An overload clutch is incorporated in the driving mechanism to prevent mechanical damage to the machine in case of a "jam".

Fig. 5 shows the control panels for dielectric strength and impedance and Fig. 6 shows the control panels for the capacitance circuits.

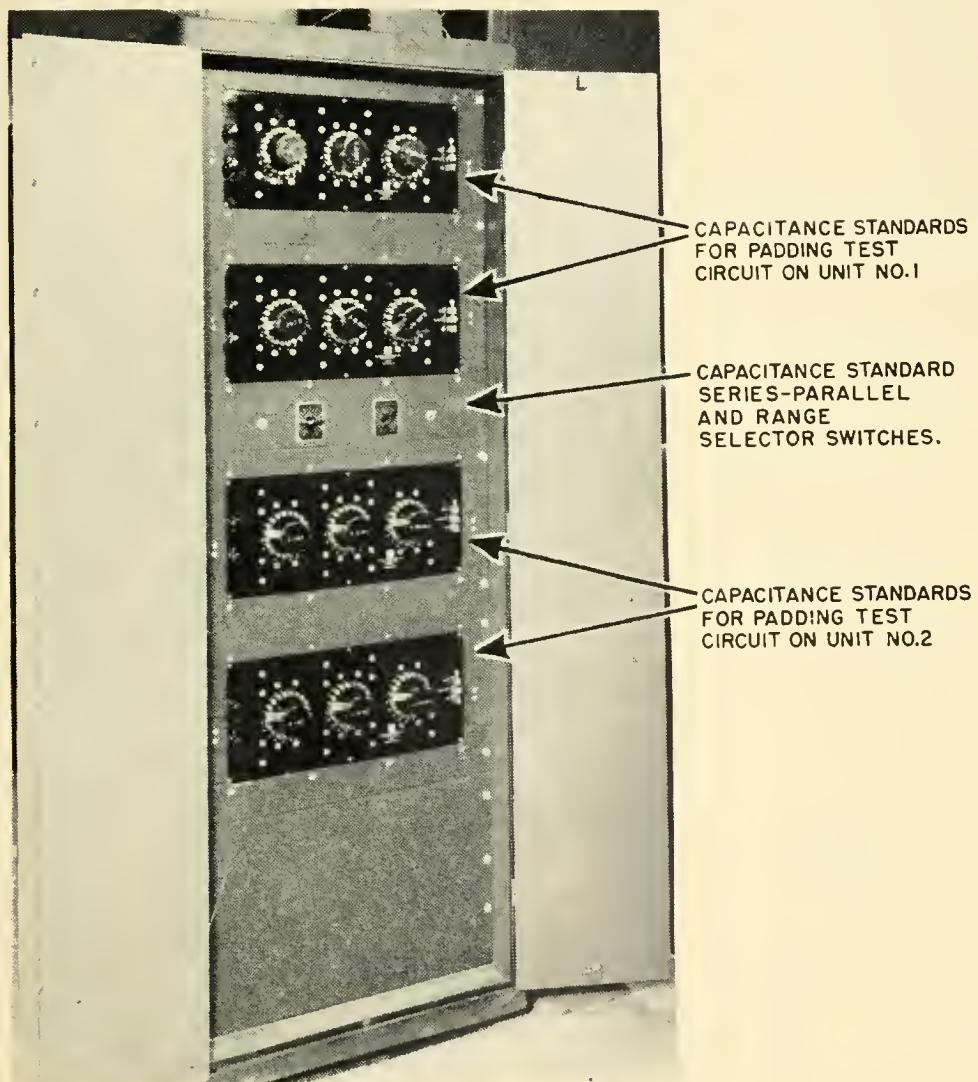


Fig. 6 — Control panels for capacitance circuits.

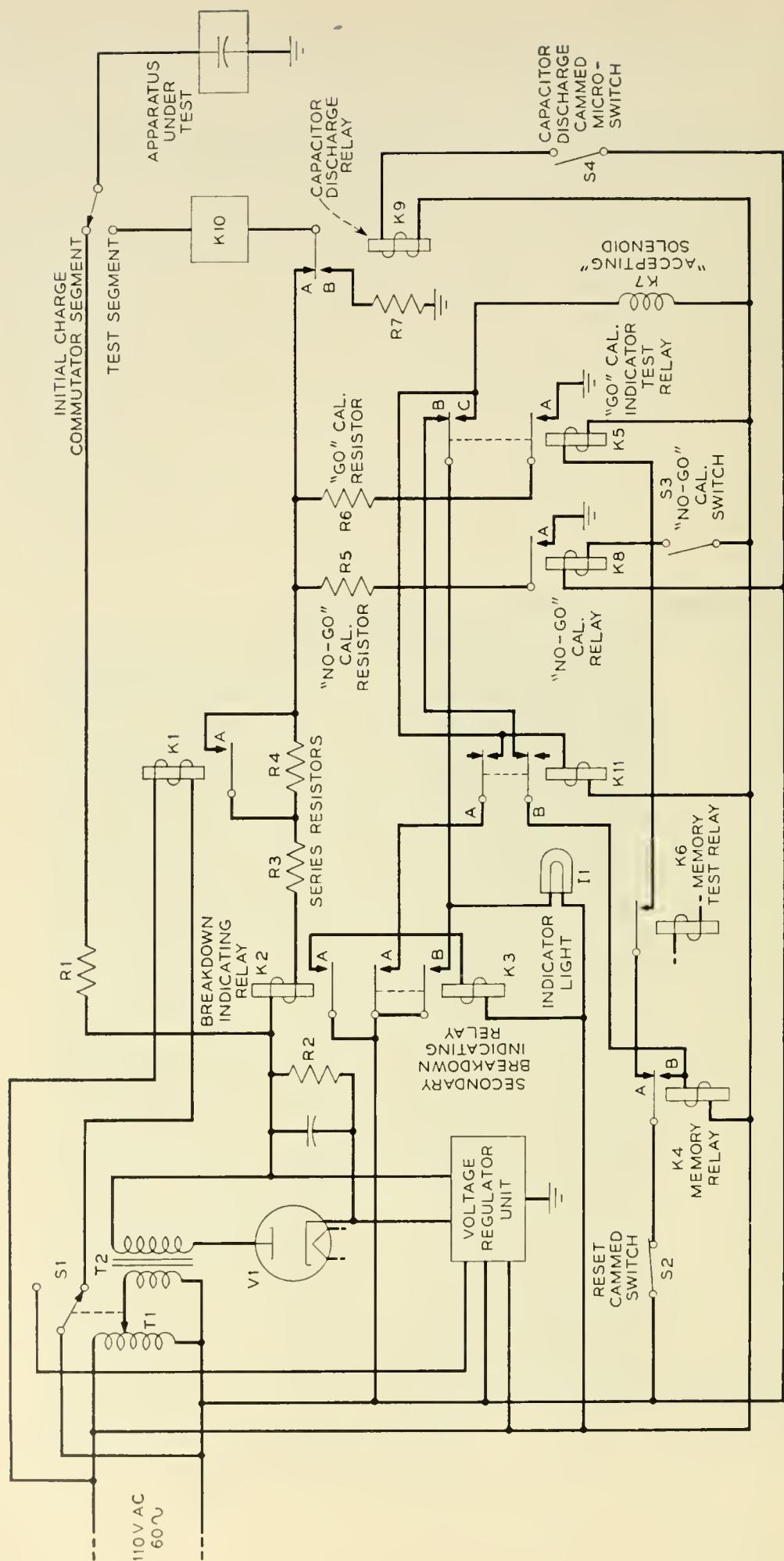


Fig. 7 — Simplified schematic of dielectric strength test circuits.

ELECTRICAL ASPECTS OF TESTING MACHINES

Tests are applied to the product in sequence during one revolution of the turret.

1. Dielectric strength test between terminals and can, and between terminals and studs.
2. Dielectric strength test between units in the same can when the can contains two units.
3. Dielectric strength test between terminals of each unit.
4. Impedance test.
5. Capacitance test.
6. Insulation resistance test.

Dielectric Strength Test Circuit Operation

Since the three dielectric strength tests are made on similar circuits, the operation of one of these circuits is described using the nomenclature and circuit designations shown in Fig. 7. A graphic interpretation of the circuit operations shown in Fig. 7 is given in Fig. 8.

The "heart" of each circuit is a calibrated current sensitive relay K2 that operates on minute values of current resulting when a defective unit under test attempts to charge on the "test" commutator position.

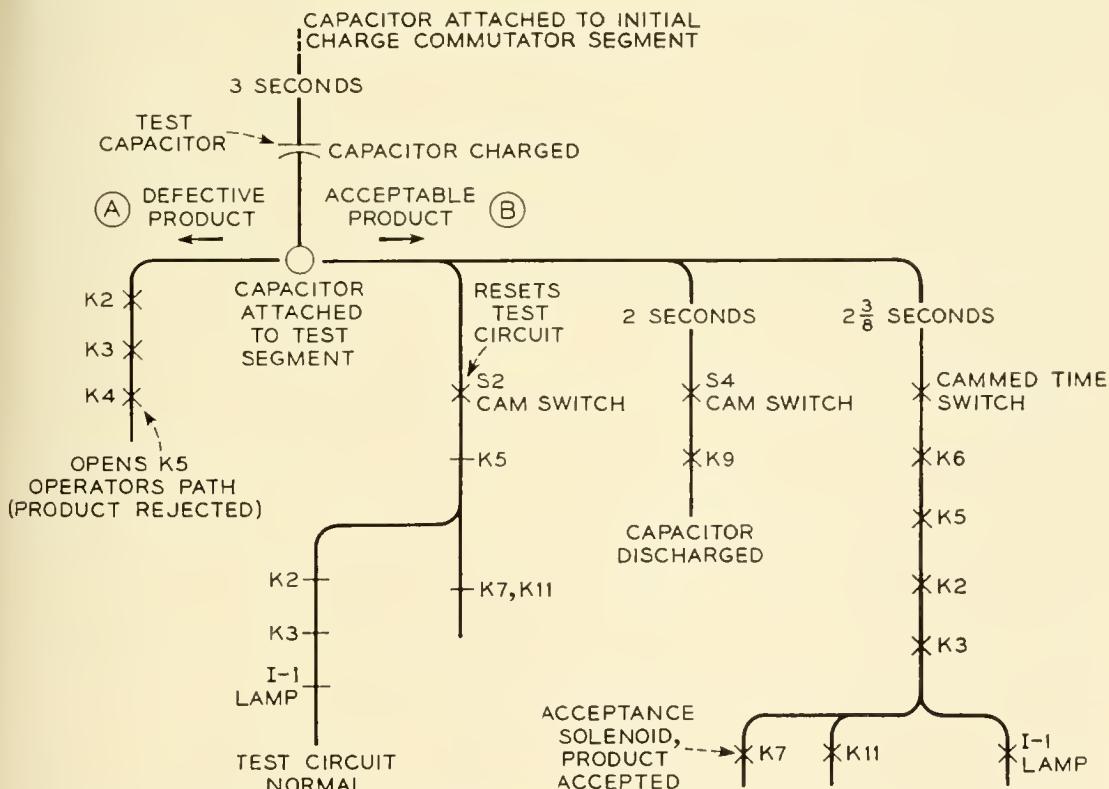


Fig. 8 — Sequence chart for dielectric strength test circuit operation.

Two commutator segments are required to make a dielectric strength test. These segments are known as "initial charge" and "test". After the unit under test has been charged at the test voltage for three seconds on the "initial charge" segment, it passes to the "test" segment in which the unit is again connected to the test voltage through relay K2, current limiting and calibrating resistors R3 and R4 and the contacts on the preset terminal selecting relays K10.

One of the two conditions (under heading A and B below) may be encountered in making this test and the circuit operation for each will be discussed separately.

A. Circuit Operation for Acceptable Product. An acceptable product retains the charge received on the "initial charge" segments and when this unit reaches the "test" segment, no further charging current of a magnitude great enough to operate relay K2 will flow through the unit. Two seconds after the unit under test has been connected to the test segment, a cammed timing switch S4 closes to operate discharge relay K9 to discharge the unit under test to ground through R7. The "self-checking" feature mentioned earlier in this article under "Design Considerations" functions as follows: After the unit under test has been on the "test" segment for approximately $2\frac{3}{8}$ seconds, a cammed timing switch (not shown) closes the memory test relay K6 which in turn closes the "go" calibration indicator relay K5 and the "A" contacts on this relay grounds the high voltage test circuit through resistor R6. This resistor is of such a value as to permit sufficient current to operate relay K2. The contacts on relay K2 are not adequate to carry much current, so an auxiliary relay K3 is closed through contacts "A" on relay K2. Contacts "B" on relay K3 closes the indicator light circuit I1 and operates relay K11 and the acceptance solenoid K7. Contacts "A" on the same relay lock relay K11. The circuit is reset for the next unit to be tested by momentarily opening the reset cammed switch S2. Relay K11 was added to the circuit to eliminate a "sneak circuit" that occurred occasionally following the reset when relay K5 opened faster than relay K3. This would result in relay K4 operating to reject the next unit tested. Relay K1 is controlled by switch S1 operated by the manual control T1 on the test voltage power supply. The function of this relay is to add calibrating resistor R4 to the test circuit for voltages above 1,000 volts. Resistor R5, relay K8, and switch S3 control the manual calibrating "No Go" circuit for breakdown indicating relay K2.

B. Circuit Operation for Defective Product. Defective product will not retain the charge it received on the "initial charge" segment and when it reaches the "test" segment, current will flow through the breakdown

indicating relay K2 in an attempt to charge the defective unit, but this current will close relay K2 which in turn closes relay K3. This completes the circuit through the "B" contacts of relays K3, K5, and K11 to close memory relay K4. The closure of relay K4 prevents the memory test relay K6 from closing the "go" calibration indicator relay K5, thereby leaving contact "C" open on relay K5 and no power is applied to the "acceptance solenoid" K7 circuit, which rejects the unit under test.

IMPEDANCE — TEST CIRCUIT OPERATION

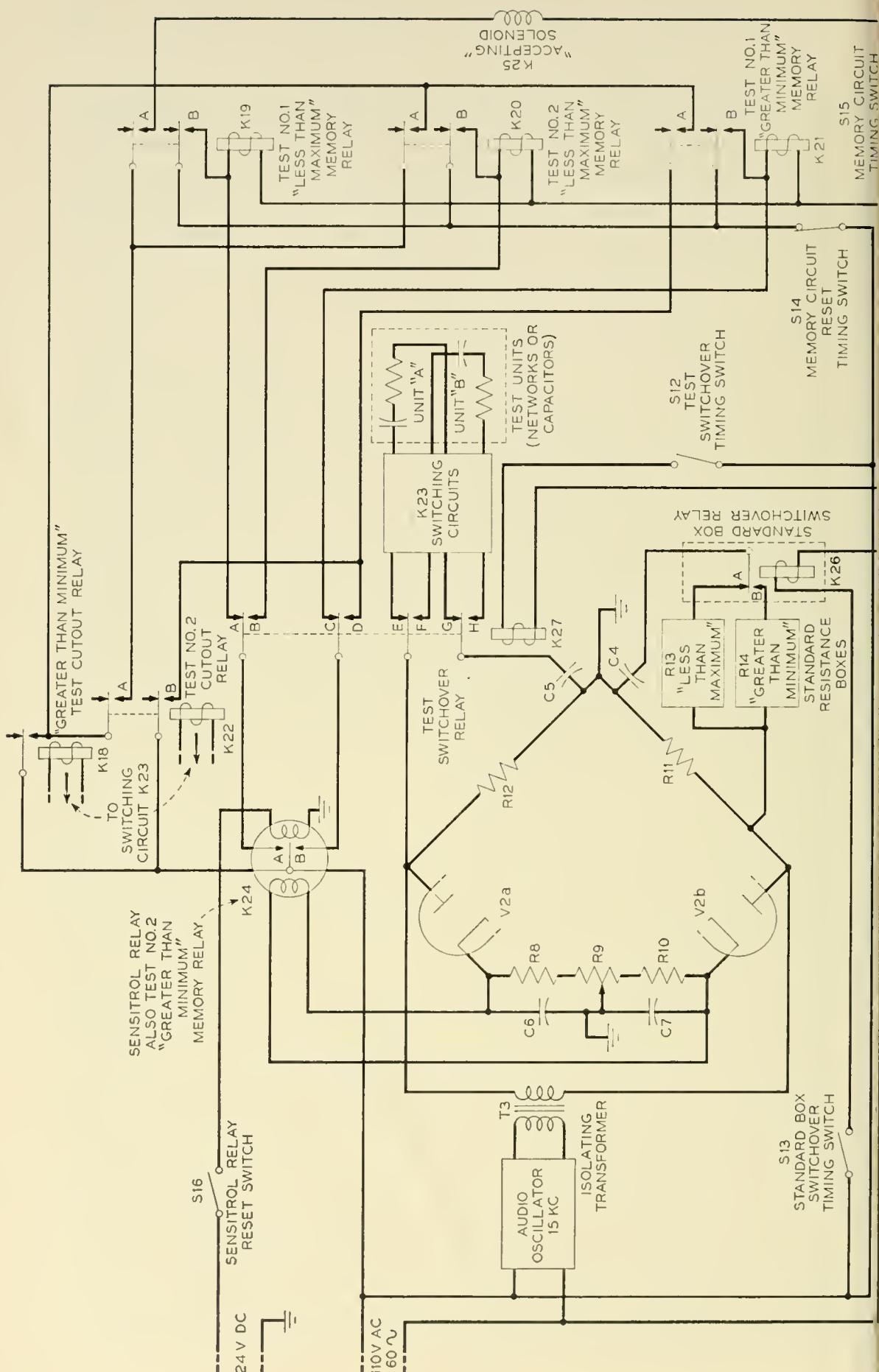
The impedance test is made with a 15-ke circuit (see Fig. 9). One arm of the circuit, composed of resistor R12 paralleled by capacitor C5 and the unit under test, is compared with another arm, composed of resistor R11, paralleled by capacitor C4 and either one of two resistance boxes, R13 and R14 respectively, representing maximum and minimum impedance limits. The detector consists of a balanced diode V2 with a 1-0-1 microampere sensitrol relay K24 connected between the diode cathodes. If the impedance of the unit under test falls within the limits for which the resistance boxes were set, the acceptance solenoid will be energized to accept the unit under test. A product outside the preset limits is rejected because the acceptance solenoid is not energized.

The circuit operation is discussed for the following four conditions under A, B, C, and D.

A. Impedance Test on Dual Unit Capacitors

This test is made on capacitors to prevent shipment of resistance-capacitance networks mislabeled as capacitors. Fig. 9 shows dual unit networks connected to the test terminals. Capacitors to be tested are connected to these same terminals. The greater than minimum test cutout relay K18 is preset closed by the switching circuit K23. The cammed memory reset timing switch S14 (normally closed) is opened momentarily to clear relay K19, K20, and K21 at the start of the test.

The sensitrol relay reset switch S16 is cammed shut momentarily to reset the contactor on the sensitrol relay K24. With relay K26 open, the "less than maximum" resistance box R13 is connected to the test circuit. If unit "A" of the dual unit capacitor under test is acceptable product, the contactor on sensitrol relay K24 will close on contact "A", which applies power to close and lock test No. 1 "less than maximum" memory relay K19. Cam operated switch S13 applies power to close relay K26 to connect the "greater than minimum" resistance box R14 into the test circuit. This resistance box is set on zero ohms when capaci-



tors are tested. Sensitrol relay reset switch S16 is cammed shut momentarily to reset sensitrol relay K24, after which the sensitrol relay contactor closes on its "B" contact, thereby applying power to close and lock test No. 1 "greater than minimum" memory relay K21.

Switch S13 is cammed open which opens relay K26 and connects the "less than maximum" resistance box R13 into the test circuit. At the same time switch S12 is cammed shut to close relay K27 which disconnects unit "A" from test and connects unit "B" to the test circuit. Switch S16 is cammed shut momentarily to reset the sensitrol relay contactor. If the unit "B" on test is an acceptable product, the sensitrol relay contactor will close on its "A" contacts and applies power to close and lock test No. 2 "less than maximum" memory relay K20 through contacts "B" of relay K27.

Switch S13 is cammed shut to close relay K26 and connect the "greater than minimum" resistance box R14 into the test circuit. Switch S16 is cammed shut momentarily to reset the sensitrol relay K24 after which its contactor closes on the "B" contact for acceptable product. Memory circuit timing switch S15 is cammed shut and power from one side of the 110 volt ac line flows through the acceptance solenoid, contacts "A" on relay K19, contacts "A" on relay K20, the closed contacts on relay K18 to the other side of the 110-volt ac line to close K25 and to accept the dual unit capacitor under test. The failure of either relay K19 or K20 to operate because of defective product tested opens the acceptance solenoid circuit and rejects the capacitor tested.

B. Impedance Test on Single Unit Capacitors

The impedance test on a single unit capacitor is identical with the testing of dual unit capacitors, except test No. 2 cutout relay K22 is preset closed and test No. 2 "less than maximum" memory relay K20 is not operated since only a single unit is tested.

C. Impedance Test on Dual Unit Networks

The impedance test on dual unit networks is identical with the test for dual unit capacitors, except the "greater than minimum" test cutout relay K18 is not preset closed and the resistance boxes R13 and R14 are set to represent maximum and minimum limits.

D. Impedance Test on Single Unit Networks

The impedance test on single unit networks is identical with the test of dual unit networks except test No. 2 cutout relay K22 is preset closed for the same reason given above for the test of single unit capacitors.

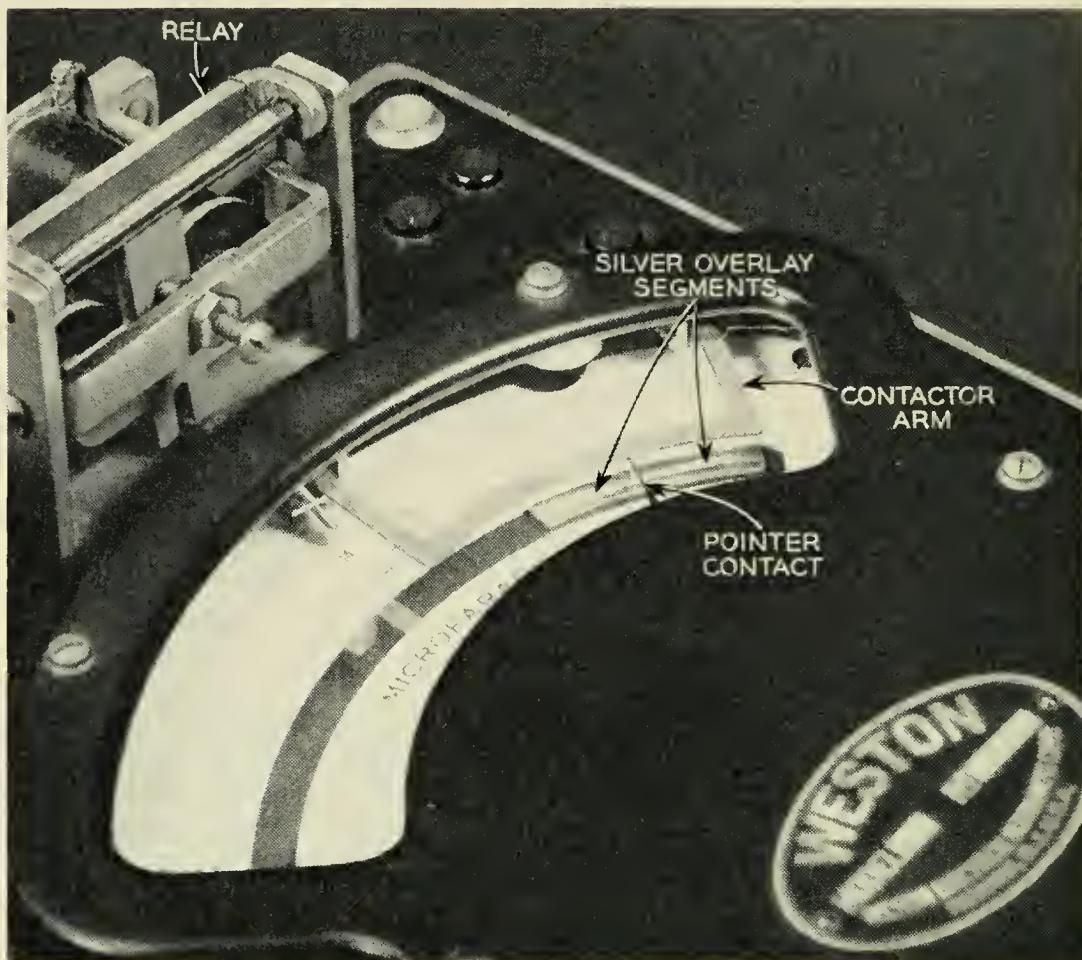


Fig. 10 — Details of modified microfarad meter.

CAPACITANCE TEST CIRCUIT OPERATION

The wide range of capacitance values to be measured, both with and without the series resistance in resistance-capacitance networks, and the one and two-unit construction of the product imposed limitations on the type of capacitance measuring circuits that could be used in this machine. The method selected consists of modified Weston Model 372 microfarad meters that automatically set up external circuits associated with the meters to accept or reject the product as determined by limits preset into the machine.

Two decade capacitance boxes, having a range from 0.001 to 1.0 mf in steps of 0.001 mf are connected in series or parallel with the capacitor on test to make the resultant capacitance fit the range of the meter and control the maximum and minimum limits. This procedure increases the number of capacitor codes that may be tested on a given meter. Capaci-

tors from 0.02 to 5 microfarads are tested on this machine to an accuracy of ± 2 per cent.

The modified microfarad meters are equipped with two brass segments, covered with an overlay of silver (Fig. 10). These segments are mounted end to end in a predetermined cutout portion of the meter scale, representing maximum and minimum capacitance conditions. The physical distance between the adjacent ends of these two segments is as small as possible without the two segments touching. A small silver contact is mounted on an insulated portion of the meter pointer, directly above but not touching the segments while the meter pointer traverses its arc of rotation. The armature of the relay, mounted on the meter, actuates a contactor arm which forces the silver contact on the pointer down against the silver overlay segment, thus closing external circuits connected to the segments and contactor.

The testing machine is equipped with three ranges of the special microfarad meters as follows:

1. Suppressed scale from 1.2 to 1.8 mf, with the dividing point between the two segments at 1.60 mf.
2. Suppressed scale from 0.25 to 0.75 mf, with the dividing point between the two segments at 0.63 mf.
3. Suppressed scale from 0.051 to 0.075 mf with the dividing point between the two segments at 0.062 mf.

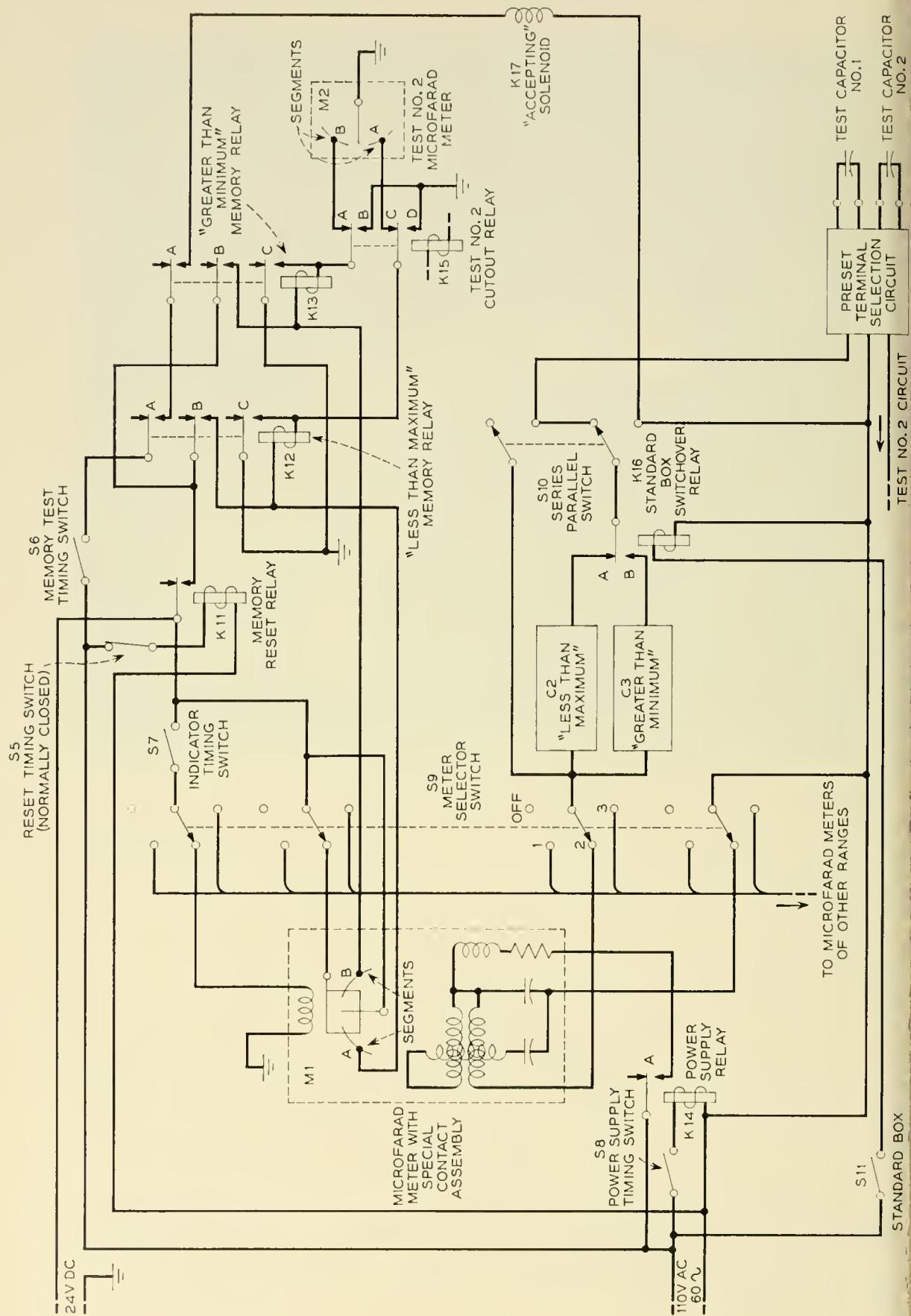
Two meters for each of the above ranges are necessary in each testing machine, one for each unit in a dual unit. Likewise, four capacitance boxes are necessary, two for each unit in a dual unit.

The discussion that follows, which is divided into two headings, A and B, is a detailed description of the capacitance test circuit. The circuit component designations are those shown in Fig. 11.

A. Capacitance Test on Dual Unit Capacitors or Networks

The cammed switch S5 is opened momentarily at the beginning of the test to restore the test circuit to normal; following this, the cammed switch S8 closes and operates relay K14, which applies power and closes the power supply circuit through the microfarad meters and the capacitor on test.

The capacitance decade box "less than maximum" C2 is shown in series with test capacitor No. 1 by the preset series-parallel switch S10, and in a like manner a capacitance box is connected in series with test capacitor No. 2.



Note: The capacitance decade box for Test No. 2 is not shown in Fig. 11. Also, only the segments for M2 Test No. 2 microfarad meter are shown.

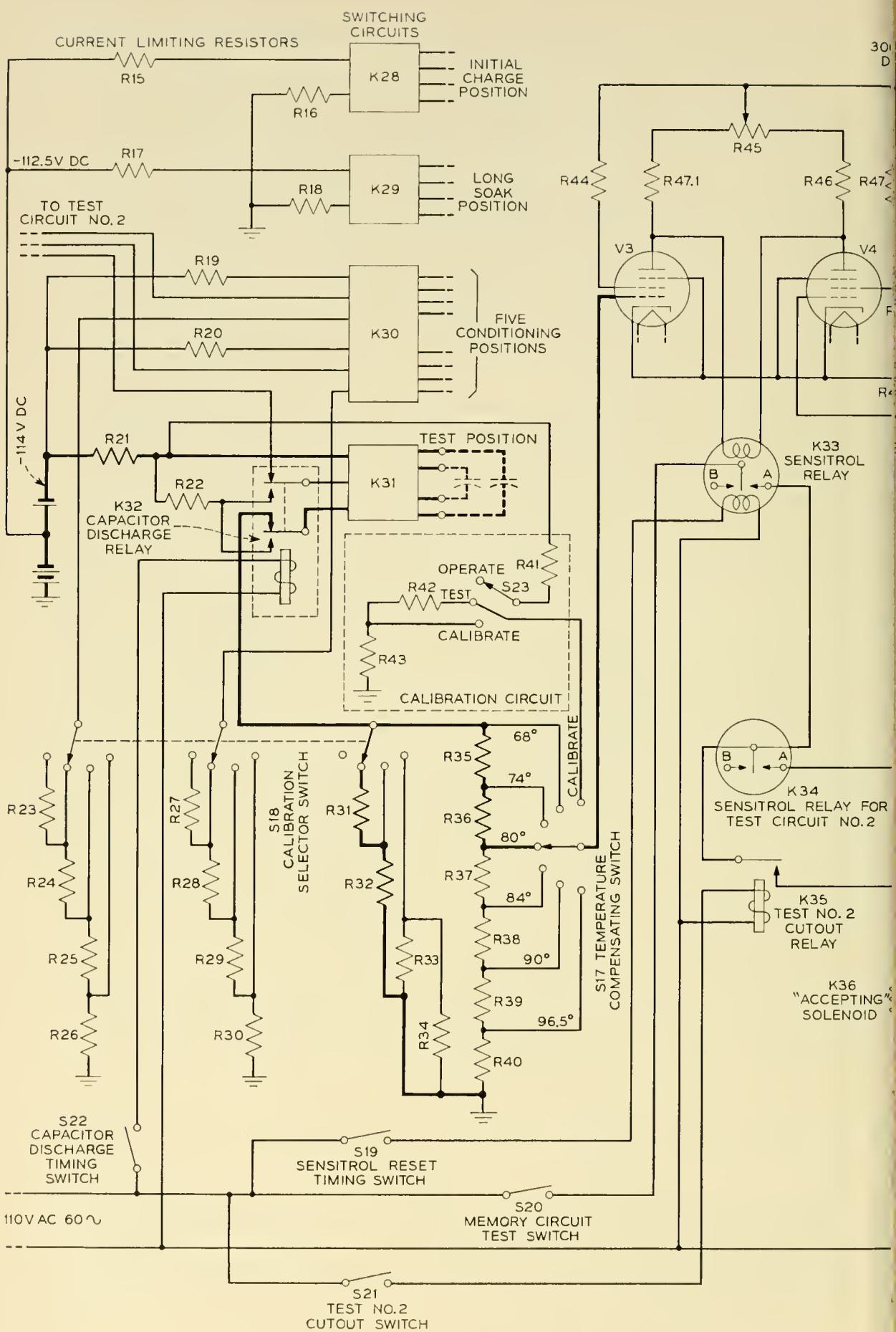
If capacitor units No. 1 and No. 2 under test are acceptable product, the pointer on the microfarad meters M1 and M2 will both swing to segment "A". The cammed switch S7 will close and energize the relay on the microfarad meters (not shown on meter M2) which will operate the meter contactor and close the circuit through segments "A" of meters M1 and M2 and apply 24 volts dc to close and lock the "less than maximum" memory relay K12.

The cammed switch S7 is opened to release the meter pointer from segments "A" on M1 and M2. Cammed switch S8 is opened momentarily to release relay K14 which removes the test voltage from the capacitors on test and from meter M1 and M2. During this interval cammed switch S11 is closed to energize relay K16 which connects the "greater than minimum" capacitance box C3 in series with capacitor unit No. 1 on test and meter M1. In a like manner a second "greater than minimum" capacitance decade box (not shown on Fig. 10) is connected in series with capacitor unit No. 2 and microfarad meter No. 2. If the capacitor units No. 1 and No. 2 under test are acceptable product, the microfarad meter pointers will swing to segments "B". The cammed switch S7 will close and energize the relay on the microfarad meters M1 and M2 which will operate the contactor that depresses the M1 and M2 meter pointers against segments B and closes and locks the "greater than minimum" memory relay K13. With relays K12 and K13 closed as described above, the cammed switch S6 is closed which operates the acceptance solenoid K17 through the "A" contacts on relays K12 and K13 to accept the dual unit capacitor under test.

It may be readily observed that in case either or both of the capacitor units on test are out of limits, the circuit will not close either or both relays K12 and K13, which would leave the acceptance solenoid K17 circuit open, and the product would be rejected.

B. Capacitance Test of Single Unit Capacitors or Networks

The capacitance test of single unit capacitors is the same as for dual unit capacitors, except test No. 2 circuit and test No. 2 microfarad meter M2 are not used. Test No. 2 cutout relay K15 is closed to apply ground to its contacts B and D.



INSULATION RESISTANCE TEST CIRCUIT OPERATION

In general, the insulation resistance test consists of a charging period and a test period. The charging of the unit under test requires 10 positions or 30 seconds time to insure that the unit is thoroughly charged before it reaches the test position. At the test position the capacitor or network on test is connected to form part of a voltage divider in the grid circuit of a sensitive balanced detector. This sets up relays to accept or reject the unit under test depending on whether the unit meets the limits for which the circuit was preset and calibrated. Two insulation resistance circuits are required, one for each unit in a dual unit capacitor or network. A calibrating circuit is provided by switch S23 and resistors R41, R42, and R43.

The discussion that follows is a detailed description of the sequence of operation of the insulation resistance test circuit. The component designations are those shown on Fig. 12. The discussion is divided into two headings A and B as follows:

A. Insulation Resistance Test on Dual Unit Capacitors or Networks

The capacitor or network on test is automatically connected in succession to the INITIAL CHARGE POSITION, the LONG SOAK POSITION and FIVE CONDITIONING POSITIONS which assures that the acceptable product is thoroughly charged before it reaches the test position. The switching circuits K28, K29, K30, K31 switch S18, and the temperature compensating switch S17 are manual preset switch circuits for the particular code on test.

For the sake of simplicity, the balanced detector and the reset solenoid for sensitrol relay K34 for test circuit No. 2 are not shown. If the insulation resistance of the units on test meets the limits for which the circuit was calibrated and preset, the contactor on K33 and K34 both close on the "A" contacts. Switch S20 is then cammed closed to apply power through the "A" contacts on the sensitrol relays to energize the acceptance solenoid K36 to accept the units on test. At the close of the test, capacitor discharge timing switch S22 is cammed closed, thereby closing the capacitor discharge relay K32 which discharges the units on test before they are ejected as acceptable product. It may be readily observed from the schematic that a unit or units defective for insulation resistance will fail to close either or both of the "A" contacts on the sensitrol relays K33 and K34, which leaves the acceptance solenoid circuit K36 open, thereby rejecting the units tested.

B. Insulation Resistance Test on Single Unit Capacitors or Networks

The insulation resistance test on single unit capacitors or networks is the same as for dual units, except the second test circuit is not required and test No. 2 cutout switch S21 is closed to operate test No. 2 cutout relay K35 which eliminates the second test circuit and its sensitrol relay K34.

CONCLUSION

This machine has been in successful operation on a multishift basis for several years and has proven itself economically. Inspection of the product tested shows that the machine's performance, quality wise, is highly satisfactory. Difficulties that have been encountered were largely those associated with product handling, contact fixtures, etc. Machines of this type that are planned for the future will make use of circuitry developed since this machine was built, but many of the features described will be incorporated.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions to the development of this machine of G. E. Weeks of the Western Electric Company S. V. Smith and S. E. Frisbee of the Electric Eye Company.

A 60-Foot Diameter Parabolic Antenna for Propagation Studies*

By A. B. CRAWFORD, H. T. FRIIS and W. C. JAKES, JR.

(Manuscript received February 2, 1956)

A solid-surface parabolic antenna, sixty feet in diameter and of aluminum construction, has been erected on a hilltop near Holmdel, New Jersey. This antenna can be steered in azimuth and elevation and was specially designed for studies of beyond-the-horizon radio propagation at frequencies of 460 mc and 4,000 mc.

The electrical properties of the antenna and the technique of measurement are described; construction and mechanical details are discussed briefly.

INTRODUCTION

Studies in recent years have demonstrated that transmission of useful amounts of microwave energy is possible at distances considerably farther than the horizon.¹ The exact mechanism responsible is not as yet completely understood, although scattering by atmospheric irregularities seems to play a significant part. A program to study the nature of these effects has been started at the Holmdel Laboratory. An important and necessary tool for this work is a steerable antenna having unusually high gain and narrow beam width. Such an antenna has been built, and it is the purpose of this paper to describe its design and the methods used to measure its radiation properties.

DESCRIPTION OF THE ANTENNA

The antenna is a 60-foot diameter paraboloid made up of forty-eight radial sectors, each constructed of sheet aluminum. Each sector is held to the correct doubly-curved surface by reinforcing ribs, and all are fastened to a central hub eight feet long and thirty inches in diameter. During assembly, the axis of the paraboloid was vertical; thus no scaf-

* This work was supported in part by Contract AF 18(600)-572 with the U.S. Air Force, Air Research and Development Command.

¹ Proc. I.R.E., October, 1955, contains many papers by workers in this field.

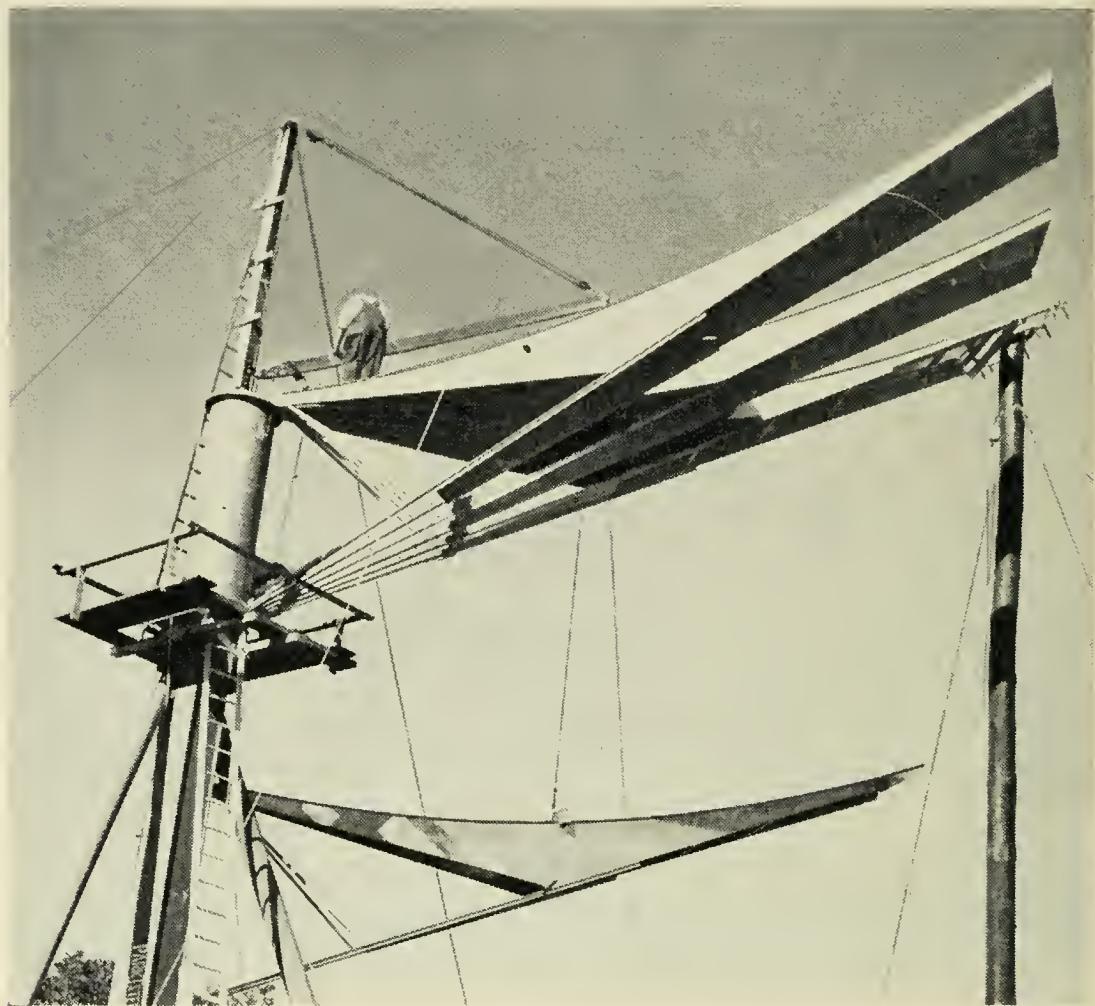


Fig. 1 — Fastening the radial sectors to the hub.

folding was required. Figs. 1 to 5 illustrate the paraboloid construction and support. The weight of the antenna is carried on a vertical column which is mounted on bearings to permit movement in azimuth. The column is held upright by a tripod structure. The central hub of the paraboloid is fastened to a steel girder which extends to the rear along the paraboloid axis and is pinned to a yoke carried by the vertical column, thus permitting movement in elevation. The antenna is scanned by two motors mounted on an A-frame and connected to the end of the axial girder by crank mechanisms. The total scanning range of the antenna is about 3° in both azimuth and elevation.

The antenna is designed for use at frequencies of 460 mc and 4,000 mc. The tolerance on the parabolic reflecting surface is set by the higher frequency and thus must be $\pm \frac{3}{16}$ inch to meet the usual $\pm \lambda/16$ criteria. The focal length is 25 feet, so that the total angle intercepted by the paraboloid as seen from the focal point is 124° . Design of a feed horn for

this angle so that the illumination is tapered to -10 db at the edge of the paraboloid is not difficult; the horn used is diagramed in Fig. 6, with dimensions given in wave-lengths. The feed horn is mounted in a tripod support extending out from the front surface of the paraboloid. It is made strong enough so that two 460 mc horns can be mounted side by side.

The paraboloid itself weighs approximately $5\frac{1}{2}$ tons; the frontal wind load for a 100 mph wind is about 40 tons. It is expected that winds of this force will be withstood.

The antenna is mounted atop Crawford Hill near Holmdel, New Jersey, at an altitude of 370 feet. It is aimed towards Pharsalia, New York, a distance of about 171 miles.

MEASUREMENT TECHNIQUE

The two important properties of the antenna which had to be determined before it could be put into use were its gain and radiation pattern at the operating frequencies of 460 mc and 4 kmc. It was also hoped to



Fig. 2 — Assembling the sectors.

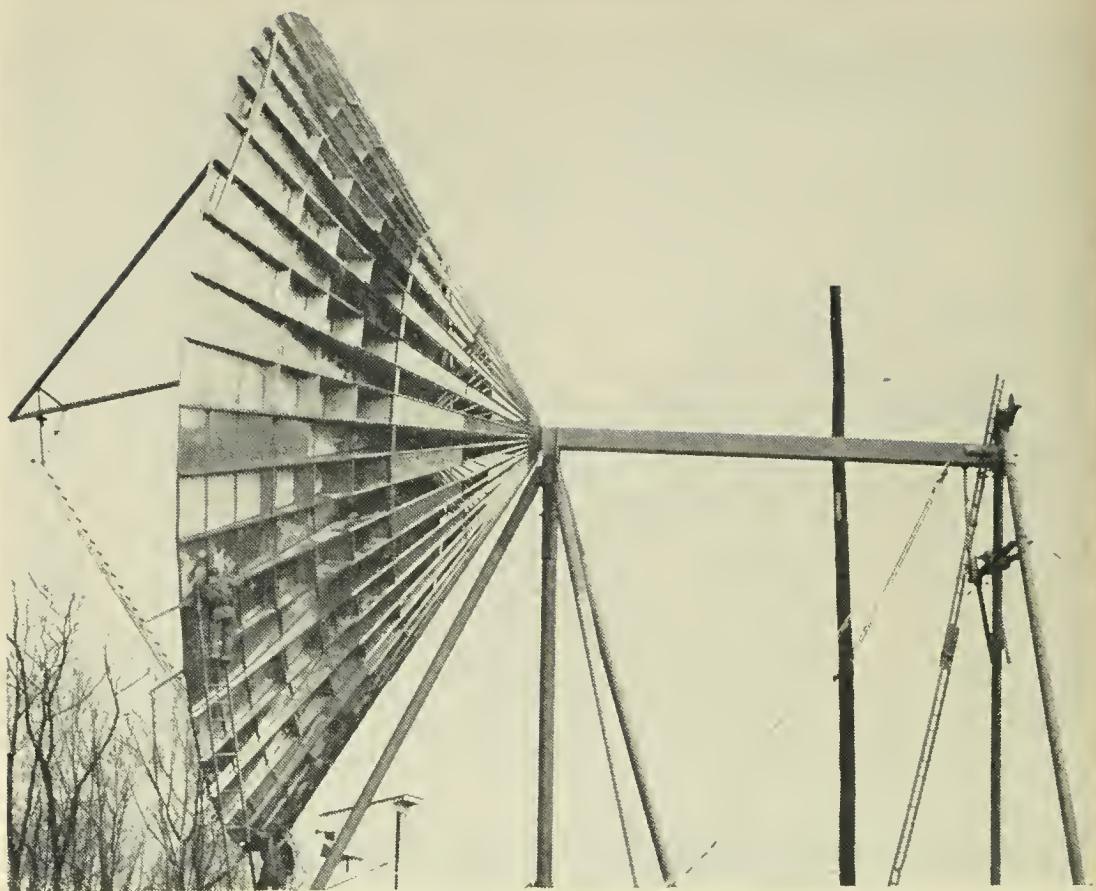


Fig. 3 — The completed antenna.

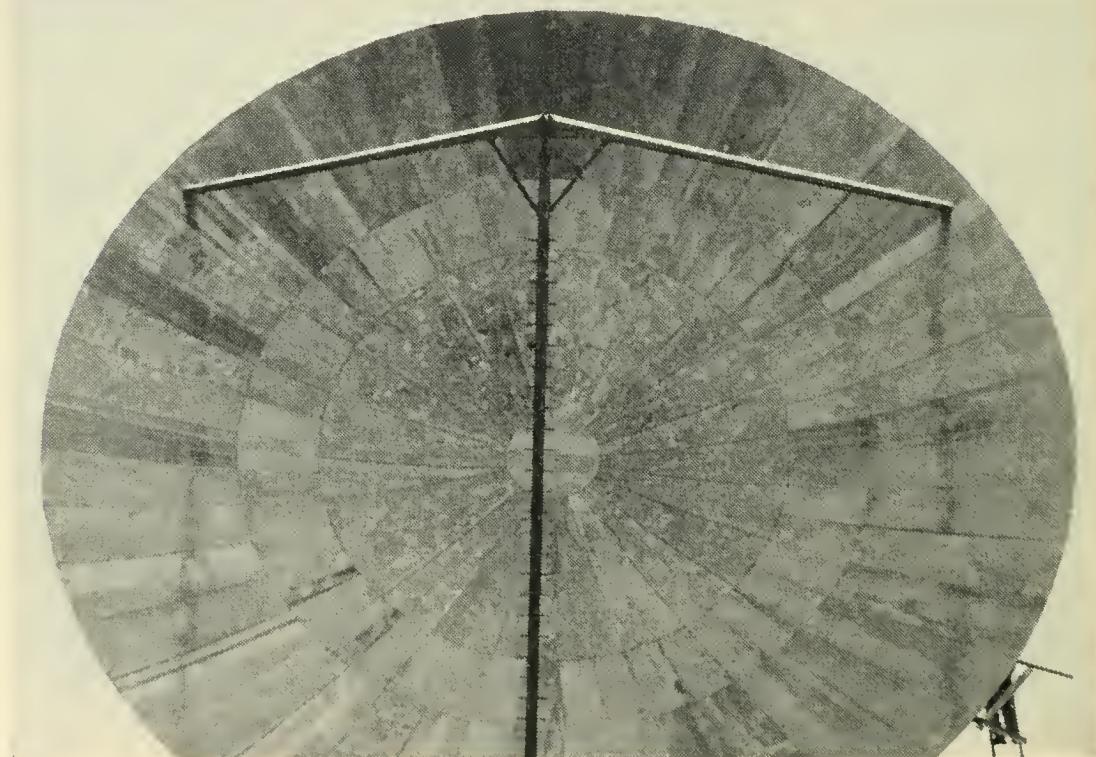


Fig. 4 — Front view of the paraboloid.



Fig. 5 — Antenna scanning motors.

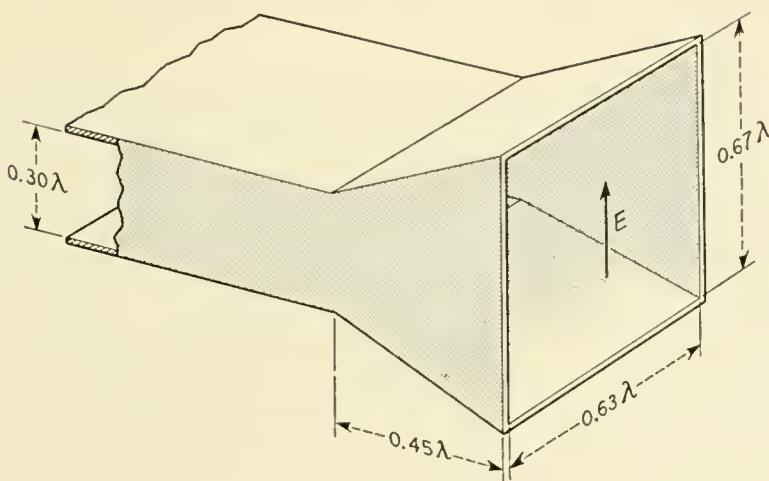


Fig. 6 — Feed horn dimensions.

measure these properties at 9.4 kmc to get some idea of how good the mechanical tolerances actually are.

The first requisite for making antenna measurements is a sufficiently uniform incident field. The source producing this field must be located at a distance of at least $2b^2/\lambda$, (b is the paraboloid diameter), which means a distance of about 0.6 mile at 460 mc, six miles at 4 kmc, and thirteen miles at 9.4 kmc. An obvious and convenient place for the sources was at Murray Hill, 22.8 miles away, which is on the transmission path to Pharsalia. Having located the sources at a suitable distance it was then necessary to test the incident field for uniformity. A 64-foot



Fig. 7 — Height run tower with the three standard horns during preliminary studies of the incident field.

tower was used for this purpose, and the variation of the incident field with height was measured before the antenna was erected. Figs. 7 and 8 show a typical set up. Height runs were taken at intervals of 15 feet along a line normal to the direction of transmission in the plane which would eventually contain the antenna aperture. The results of these tests showed that the Murray Hill location was satisfactory for the 4 and 9.4 kmc sources, with ground reflections giving rise to ± 1 db variations with height. In each case several complete cycles occurred in the 60-foot height run so that an average signal level could be established with an accuracy of a few tenths of db.

However, at 460 me the variation with height was about 5 db, and only a portion of one cycle was available, so that the average signal could not be determined. The solution was to bring the source to a location as close as possible to the effective ground reflecting surface. Such a location was found at the far edge of a large body of water lying in the path, and the source antenna was placed in a mobile truck 10 feet above the water and eight miles away. The resulting variation with height was now only about 1 db.

In all cases the variation of field at right angles to the direction of transmission was found to be no worse than ± 1 db; thus it was felt that suitable sources for test at all three frequencies were now ready.

The standard method of measuring the gain of a microwave antenna is to compare the signal received from the antenna to that from another antenna whose gain is accurately known. A pyramidal horn of about 20 db gain is usually used as the standard. Such horns are readily available at 4 kmc and 9.4 kmc, and, in principle, also at 460 mc. Under the present set up, however, the physical dimensions of the standard horn were limited by the necessity of raising the horn on a carriage attached to the 64-foot tower. The largest horn that could be so mounted had an aperture of 4 feet \times 4 feet, or 1.8λ on a side at 460 mc. Since the gain of a horn of this small aperture size cannot be accurately calculated by the usual formulas a scale model was made and tested at 4 kmc. The result of this test showed that the actual horn gain was 15.05 db, which is about 0.4 db more than the calculated gain.

A typical gain measurement on the 60-foot paraboloid was thus made as follows:

1. The feed position and antenna orientation were adjusted to obtain maximum received signal level.

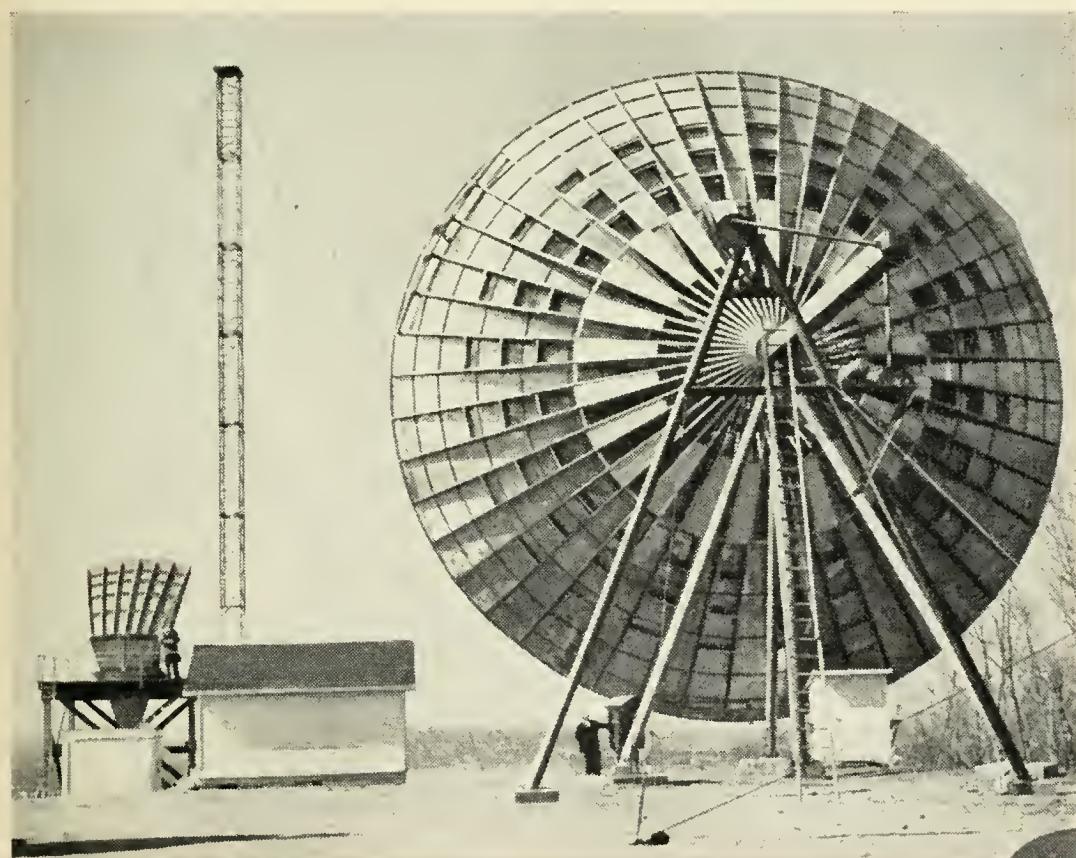


Fig. 8 — Position of height run tower during gain measurements.

2. The average incident field was determined by a height run with a standard horn.

3. The decibel gain of the antenna was then calculated by adding the db gain of the standard horn to the db difference in the signal levels determined in (1) and (2).

The problem of adjusting the 60-foot antenna for maximum received signal at 4 kmc and 9.4 kmc was complicated by the scintillations of the

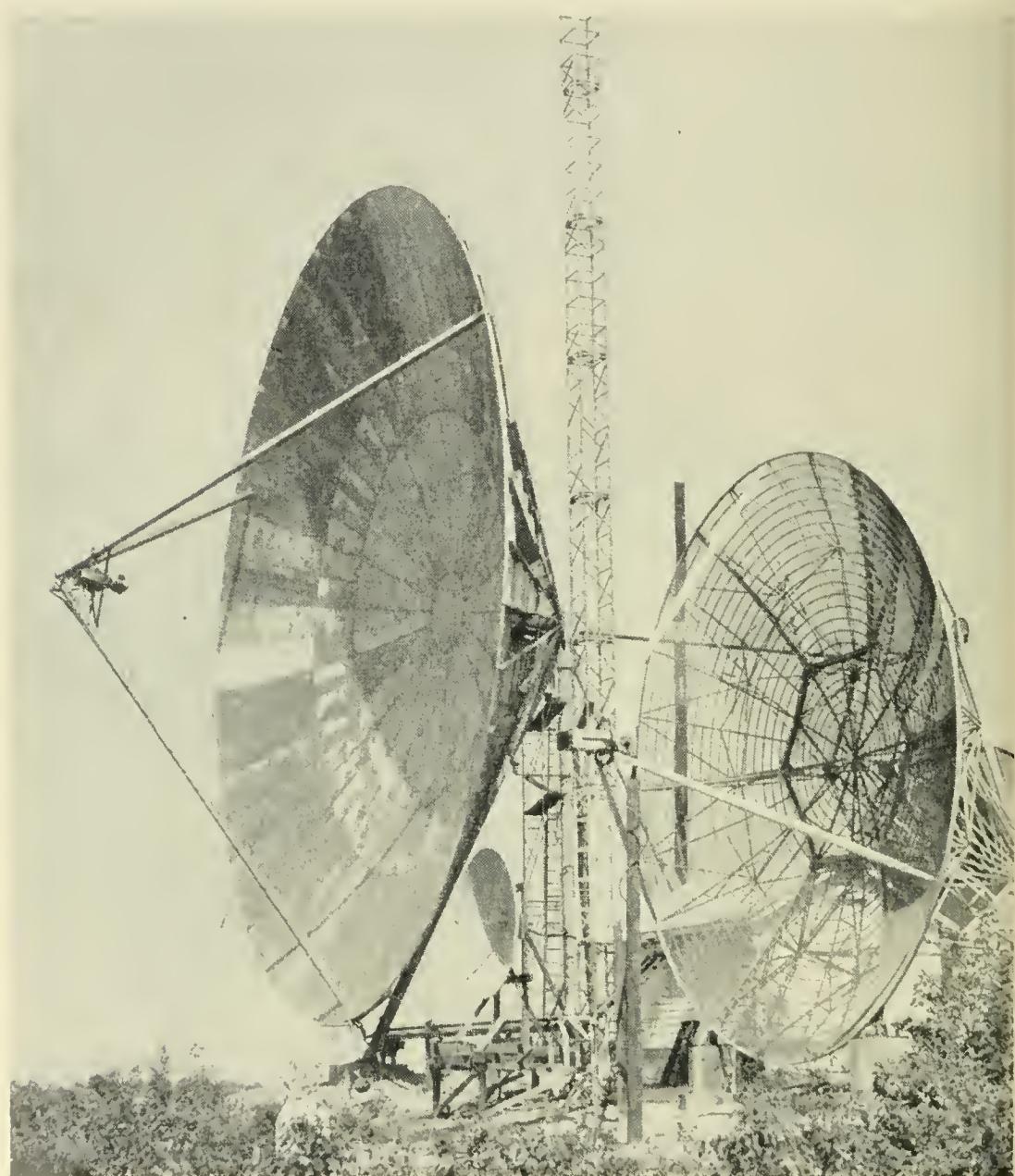


Fig. 9—A view of the antennas at Crawford Hill used for beyond-the-horizon propagation studies and showing the 60-foot, a 28-foot and an 8-foot paraboloid, the latter between the two larger ones.

TABLE I

Frequency	Area Gain, [*] db	Gain, db Meas.	Ratio of Effective Area to Actual Area	3 db beam width		1st Minima	1st Minor lobes
				Calc.	Meas.		
460 mc	38.90	37.0 \pm 0.1	0.65	2.35°	2.45°	—	—
3.89 kmc	57.44	54.6 \pm 0.2	0.52	0.28°	0.3°	-33db	-23db
9.40 kmc	65.12	61.1 \pm 0.5	0.40	0.12°	0.14°	-25db	-18db

* The area gain is defined as $10 \log \frac{4\pi}{\lambda^2}$, where A is the paraboloid projected area, 2,830 square feet.

incident field at these frequencies due to the remote location of the source. Accordingly, instead of adjusting the feed position for maximum signal level, it was adjusted to give vertical and horizontal radiation patterns having the best possible symmetry, deepest minima, and lowest minor lobes. It was then assumed that this was also the point of maximum gain. At 460 mc the scintillations were so small that the conventional technique of adjusting for maximum output was effective.

A double detection receiver was used for making all measurements. Signal level decibel differences were established by an attenuator in the intermediate frequency (65 mc) channel, and could be determined to an accuracy of ± 0.02 db.

RESULTS

Carrying out the measuring procedure described above the results given in Table I were obtained. At 460 mc the restricted scanning range did not permit inspection of the minor lobes.

CONCLUDING REMARKS

The overall performance of this antenna is considered to be excellent. In general the radiation patterns are clean with satisfactory minor lobe structure. The good performance at 9.4 kmc (61 db gain) is particularly gratifying, since the mechanical tolerance of $\pm \frac{3}{16}$ inch is equivalent to $\pm \lambda/7$ at this frequency.

As stated earlier, this antenna was designed to provide a research tool for propagation studies and thus has some features which are neither necessary nor desirable in an antenna intended primarily for communication use. A consideration of the problem of providing a sturdy 60-foot

antenna for fixed point-to-point service led to the square "bill-board" design* and antennas of this type are now in production.

ACKNOWLEDGEMENTS

The construction of the antenna described in this paper was carried out under the general direction of H. W. Anderson, Supervisor of the Holmdel Shops Department. The paraboloid was assembled in place by members of the Carpenter Shop supervised by C. P. Clausen. Daniel Beaton, of Lorimer and Rose, served in an advisory capacity on some features of the construction. Assistance in the design and testing of the antenna was given by many members of the technical staff.

* A picture and short description of this antenna appeared in Bell Laboratories Record, **34**, p. 37, Jan., 1956.

The Use of an Interference Microscope for Measurement of Extremely Thin Surface Layers

By. W. L. BOND and F. M. SMITS

(Manuscript received March 15, 1956)

A method is given for the thickness measurement of p-type or n-type surface layers on semiconductors. This method requires the use of samples with optically flat and reflecting surfaces. The surface is lapped at a small angle in order to expose the p-n junction. After detecting and marking the p-n junction, the thickness is measured by an interference microscope. Another application of the equipment is the measurement of steps in a surface. The thickness range measurable is from 5×10^{-6} cm to 10^{-3} cm.

INTRODUCTION

Extremely thin p-type or n-type surface layers can be obtained on semiconductors by recently developed diffusion techniques.^{1, 2} Layer thicknesses of the order of 10^{-4} cm are currently used for making diffused base transistors.^{3, 4} The thickness of the diffused layer is an important parameter for the evaluation of such transistors. Its measurement is facilitated by lapping a bevel on the original surface, thus exposing the p-n junction within the bevel where the thickness appears in an enlarged scale. With a sharp and well defined angle, one would obtain the thickness by the measurement of the angle and of the distance between the vertex and the p-n junction.

However, it is extremely difficult to obtain vertices sharp enough for measurements of thicknesses of the order of 10^{-4} cm. To avoid this difficulty, an interferometric method was developed in which the depth is measured directly by counting interference fringes of monochromatic light. The method can also be used for the measurement of small steps

¹ C. S. Fuller, Phys. Rev., **86**, p. 136, 1952.

² J. S. Saby and W. C. Dunlap, Jr., Phys. Rev., **90**, p. 630, 1953.

³ C. A. Lee, B.S.T.J., **35**, p. 23, 1956.

⁴ M. Tanenbaum and D. E. Thomas, B.S.T.J., **35**, p. 1, 1956.

and similar problems occurring, for example, in the evaluation of controlled etching and of evaporated layers.

PRINCIPLE⁵

A half-silvered mirror is brought into contact with a reflecting surface. If this combination is illuminated with monochromatic light, one observes interference fringes. Dark lines appear where the distance between mirror and surface is $n \times \lambda/2$, where n is an integer. Between two points on adjacent fringes the difference in this distance is therefore $\lambda/2$. Hence the fringes can be regarded as contour lines for the distance between the mirror and the surface under consideration. Since the mirror is optically flat, one can deduce the profile of the surface. Equidistant and parallel fringes, for example, prove the surface to be flat. By taking the profile across a bevel or a step, one is able to measure the depth of one part of the surface with respect to another optically flat part of the surface. The reflectivity of the crystal surface should be as high as possible, and that of the mirror should be of the same order. The fringes are then produced by the interference of several wave trains which make the fringes very sharp, and one can measure fractions of $\lambda/2$. With the equipment described here, one is able to interpolate to $1/10$ of $\lambda/2$ or less.

Since small linear dimensions are involved, this principle was adapted for use under a microscope. Hence, it is possible to measure small linear dimensions and the correlated depth simultaneously.

The measurement of small steps, or steps not too steep in an otherwise flat surface, can be done without altering the sample. For measurement on steep and high steps a bevel must be lapped on the sample.

For the measurement of p-type or n-type surface layers on semiconductors, it is essential to lap a bevel on the original surface. The p-n junction is thus exposed and can be found by an electrical method. After marking its position within the bevel, it is then possible to measure its depth with respect to the original surface by taking the profile across the bevel. The marking has to be such that it will be visible in the fringe pattern. By a proper adjustment of the optical flat, a fringe pattern can be produced in which the profile is easily interpreted and the depth measurement amounts to a counting of fringes.

PREPARATION OF THE SAMPLES

The method requires the use of samples with optically flat and highly reflecting surfaces with respect to which a depth can be measured. It is

⁵ S. Tolansky, *Multiple-Beam Interferometry of Surfaces and Films*, Oxford, at the Clarendon Press, 1948.

also advisable to use plane-parallel samples to facilitate the lapping of a bevel at a small angle.

For lapping, the sample is waxed with its back side to the face of a short steel cylinder. The face is cut at a small angle. Angles of 0.5° or 1.0° are practical. The cylinder is placed in a jig, in such a position that approximately half of the sample surface projects above the plane of the jig (Fig. 1). A short grind on a slightly rough glass plate using a fine abrasive with water gives usable bevels. For a shiny finish just the right degree of roughness of the glass is important. The use of a lapping machine with a vulcanized fiber plate and fine abrasive gives a better surface finish, but the ridge is not as sharp. A 0.5° bevel could be obtained only on a glass plate.

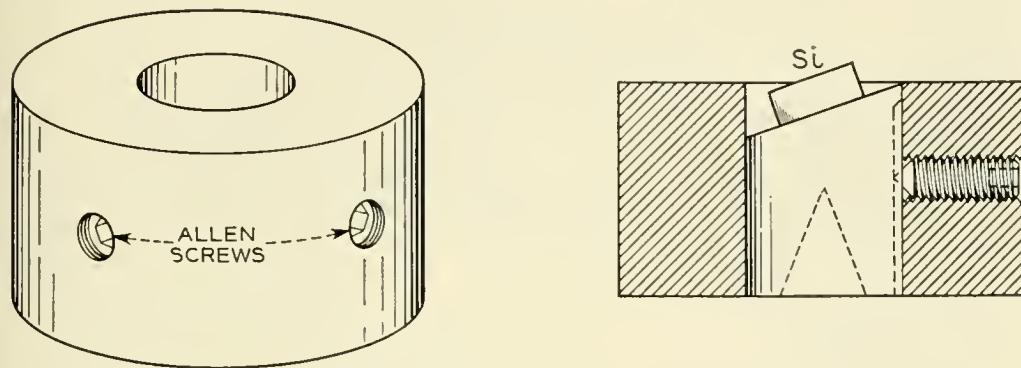


Fig. 1 — Jig for lapping a bevel.

MARKING OF p-TYPE OR n-TYPE SURFACE LAYERS

In a sample with a p-type or n-type surface layer the junction is exposed within the bevel. The next step is to detect and mark the junction.

The sample is fixed on a microscope stage which allows a micrometer controlled movement in two directions (Fig. 2 shows a Wilder micrometer cross slide). The sample is oriented in such a way that the ridge is parallel to one direction of movement (y-direction). One or two lines of aquadag are applied to the surface of the sample, perpendicular to the ridge. The aquadag should be diluted with water in such a proportion as to achieve a thin film which is non reflecting.

A needle is fixed to the base of the stage with a suitable linkage leaving a vertical degree of freedom. The needle is brought into contact with the surface of the sample outside the aquadag. Thus, the sample can be moved under the needle while the needle maintains contact. In a suitable electrical circuit, the needle serves as detector of the junction. The sample is moved in the direction perpendicular to the ridge (x-direction)

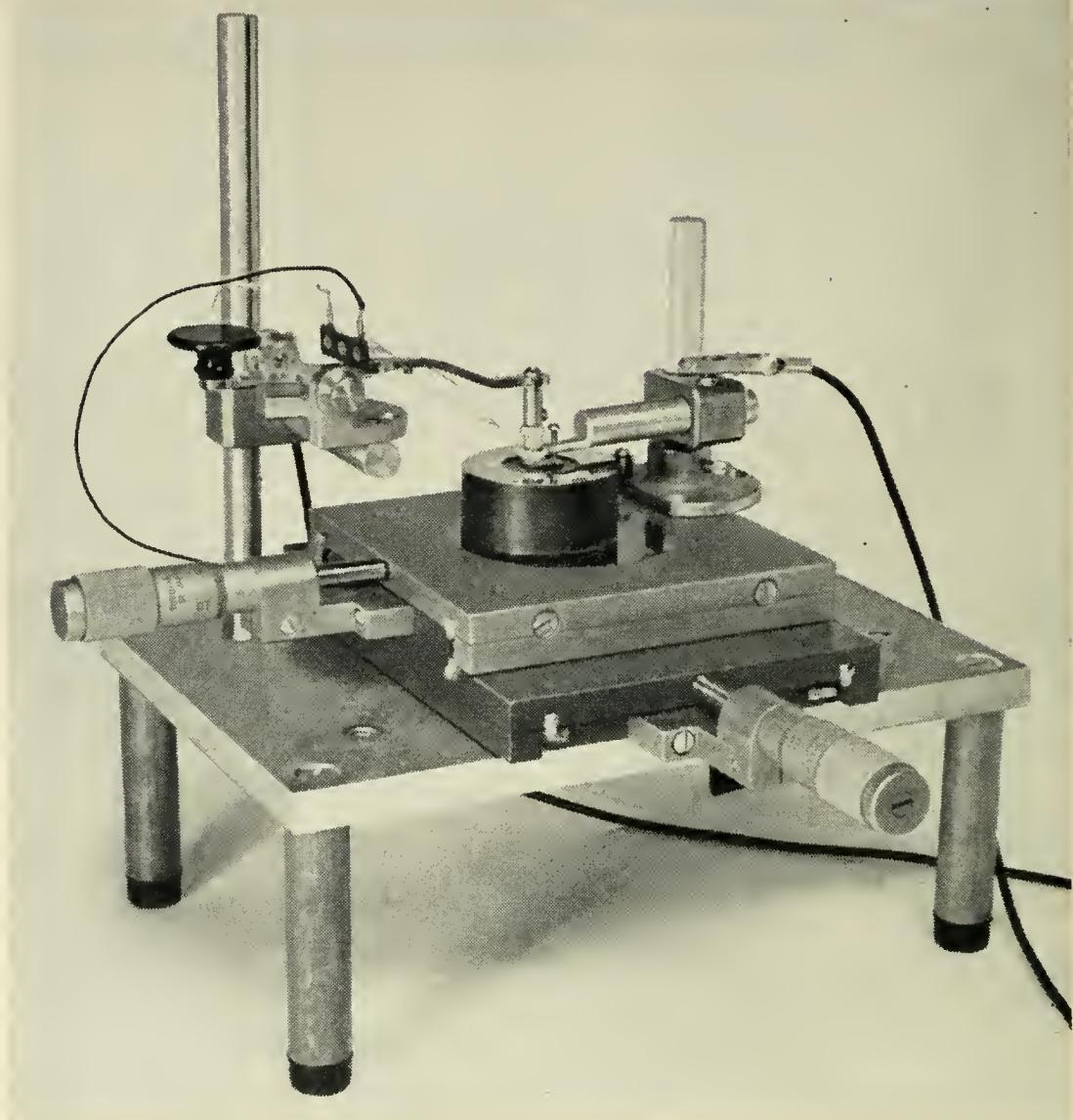


Fig. 2 — Apparatus for locating and marking *p-n* junctions.

until the needle rests on the *p-n* junction as seen by the electrical detector. By moving the sample in the *y*-direction the same needle scrapes a line through the aquadag. In this line, the reflecting sample surface is bared and thus, a reflecting line is produced within a non-reflecting surrounding and can be seen in the fringe pattern.

If the ridge of the sample is exactly lined up with the *y*-direction, the needle moves along the junction and the line in the aquadag indicates the position of the junction exactly. To minimize an error due to poor alignment, it is advisable to locate the junction close to the edge of the aquadag. By doing this on two different sides of the coating, the average

of both readings compensates the error. To obtain, however, the maximum of accuracy, one can locate the junction at any point. (See Fig. 3, Point A). Moving the sample in the y-direction scribes a line through the point at which the junction was found. A movement in the x-direction with the needle in the aquadag marks a point B on this line. The distance from this point to the junction can be obtained from the readings on the micrometer. Thus, the exact point at which the junction was located can be reproduced under the microscope.

DETECTION OF THE p-n JUNCTION

1. Thermoelectric Probe

The thermoelectric voltage⁶ occurring between a hot and a cold contact to the sample, changes sign by crossing the junction with the hot contact. The advantage of this probe is that it does not depend upon the rectification properties of a p-n junction. The thermoelectric probe is most suitable for germanium since lapping across a p-n junction normally produces a "short" between the two regions. However, it is likely to give a p-reading on lightly doped n-material. It is therefore only usable on heavily doped layers, where the nearly compensated zone is very small. In the case of silicon, the junction normally maintains rectifying properties after lapping; thus, a photocurrent is present. This current is superimposed upon the thermocurrent. Therefore, the thermoelectric probe is only usable in the dark. The photoelectric method (see below) is more convenient for these cases.

The thermoelectric probe used, consisted of a commercial phonograph needle, which had a good hemispherical point and was surrounded by a piece of ceramic tubing carrying a heating coil. Between needle and sam-

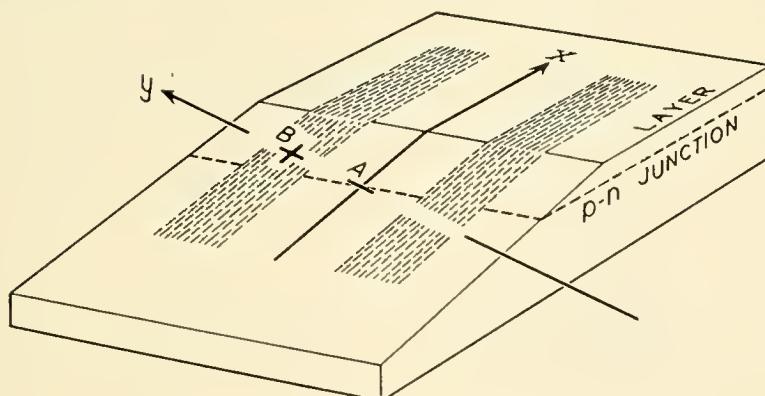


Fig. 3 — Schematic view of a scribed p-n junction.

⁶ V. A. Johnson and K. Lark-Horowitz, Phys. Rev., **69**, p. 259, 1946.

ple a sensitive galvanometer is connected. It is unimportant whether the contact is made to the p-type or to the n-type side of the sample. The best results are obtained on freshly lapped and clean surfaces. It is, therefore, advisable to keep the sample on the steel cylinder while applying the thermoelectric probe.

When applying the probe, the sample is moved in the x-direction to the point of zero deflection on the galvanometer, whereby the point rests on the junction.

2. Point Rectification on the Surface

This test is also usable on p-n junctions which are not rectifying. With one fixed ohmic contact to the sample, the point rectification of the movable needle can be displayed on an oscilloscope. By crossing the junction with the needle, the characteristic changes from p-n to n-p. Thus, the needle again can be placed on the junction.

This test was applied on lightly doped Ge-layers. The oscilloscope pattern is not very definite, since on a lapped surface the point rectification is poor. However, with some experience the junction can be located. It is advisable to repeat the measurements several times. Boiling the sample in water before applying the probe improves the surface.

3. Photoelectric Probe

This method requires that the junction exhibit rectifying properties. It is most successfully applied to silicon. Between the needle and a contact to either the p-type side or the n-type side of the sample, a high impedance voltmeter is connected. While the sample is strongly illuminated, it is moved in the x-direction. When the needle crosses the junc-

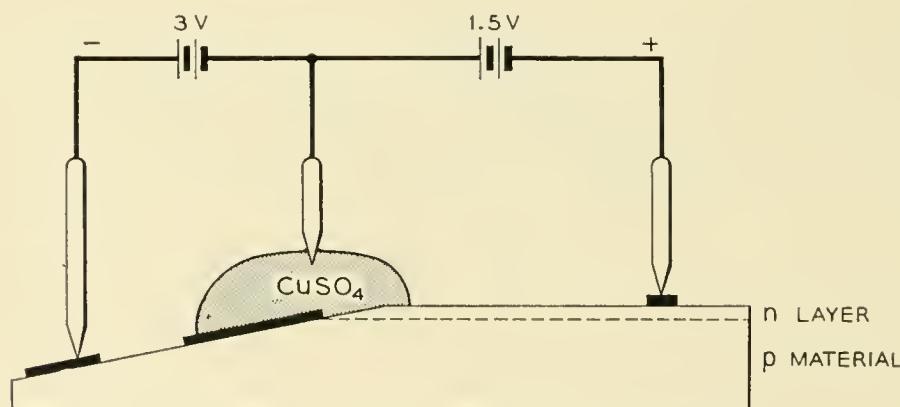


Fig. 4—Arrangement for Cu-plating the p-type side of a p-n junction.

tion, a change in the photoelectric voltage occurs. For more careful measurements one might plot the photovoltage versus the x-coordinate in units of the micrometer. Such a plot allows an accurate location of the junction in these units. If the micrometer is set for this reading, the needle will rest on the p-n junction.

4. Potential Probe

This is another method for locating the junction where the junction is at least slightly rectifying. One needs two contacts to the sample, one on the p-type side and the other on the n-type side. When a current is passed through the sample in the reverse direction, the voltage between the needle and either contact shows a discontinuity at the junction. The voltage can be plotted in a similar way as described for the previous method, and thus the needle can be set on the p-n junction.

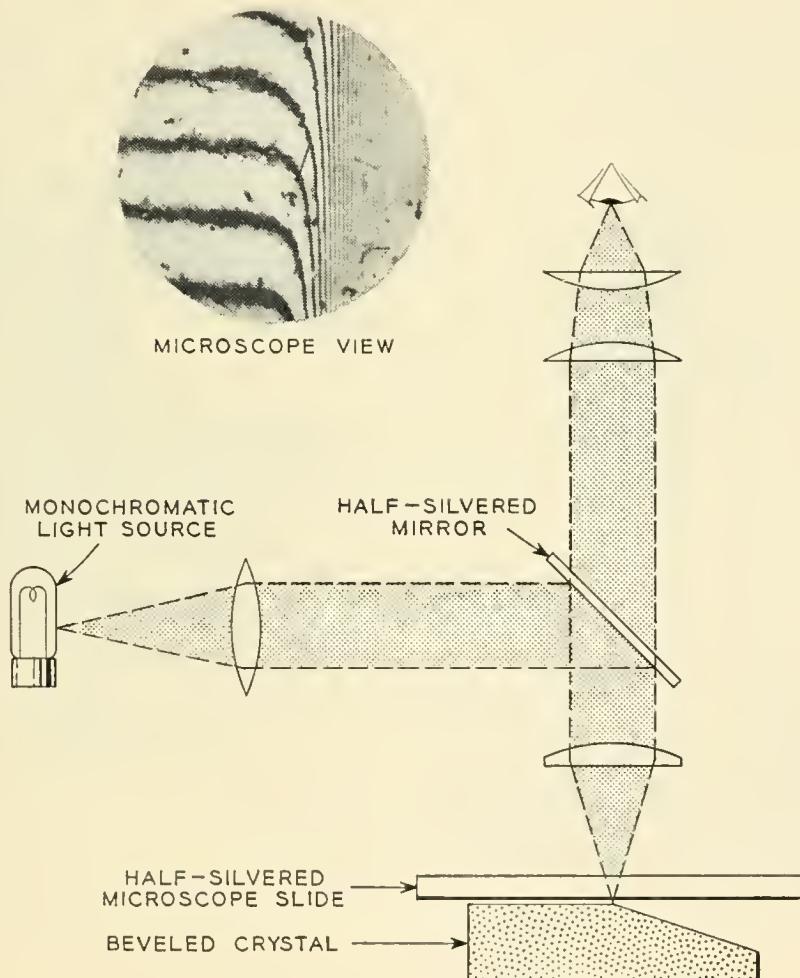


Fig. 5 — Diagrammatic view of the light path in the interferometer.

5. Plating the p-side of the p-n Junction*

This method detects and marks the junction in one process without using the micrometer arrangement. Voltages are applied in such a way that only the p-type side is plated (See Fig. 4). Since the plating projects up and is not optically flat, it can be recognized under the interferometer. It has the advantage of showing the junction as a line. The disadvantage is that it is only convenient on rectifying p-n junctions (silicon), with n-type layers since the plating ought to take place on the body side of the p-n junction.

THE INTERFEROMETER

The main part of the interferometer is a microscope with illumination through the objective. As a source of monochromatic light, a sodium lamp for which $\lambda = 5.89 \times 10^{-5}$ cm is most convenient. The use of a

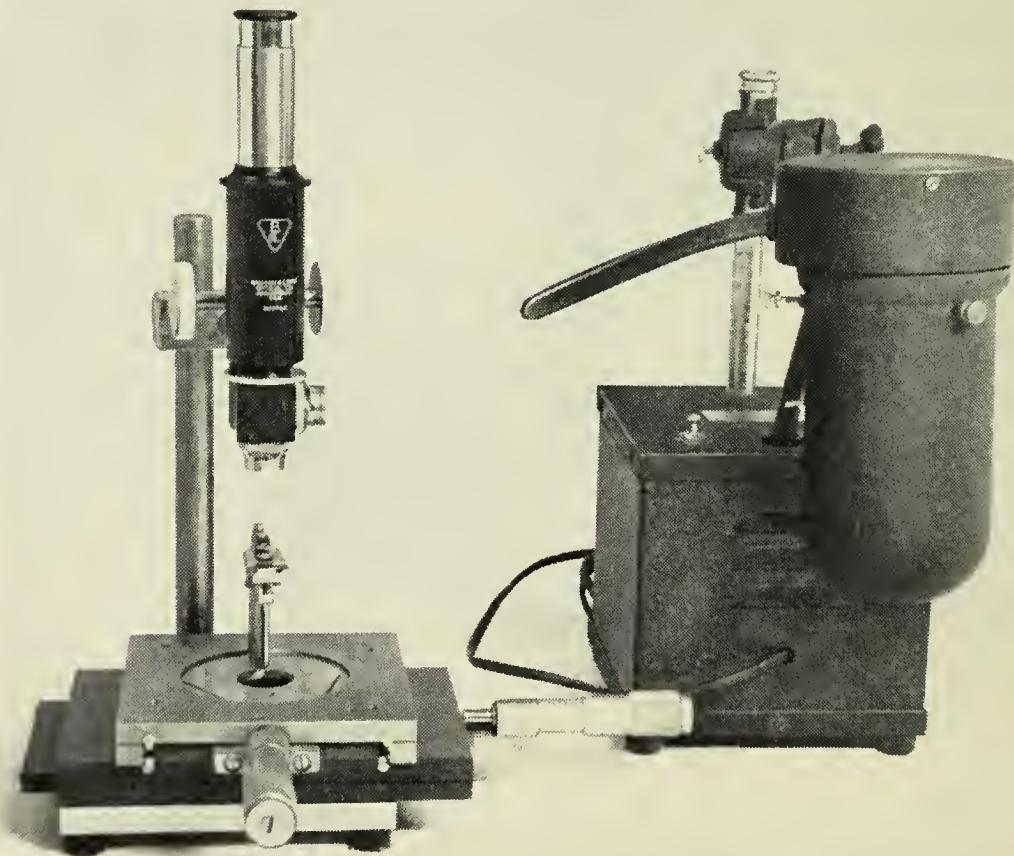


Fig. 6 — Interferometer with light source.

* This method was developed by N. Holonyak.

shorter λ would increase the resolution. However, a sodium lamp gives enough light that one can easily work in daylight.

The microscope is mounted above a micrometer cross-slide of the same kind as used in the procedure for marking the p-n junctions. The stage carries a special sample support. Fig. 5 gives a diagrammatic view of the light path in the interferometer. A normal microscope is used with an attachment carrying a semi-transparent mirror. Fig. 6 shows a photo of the complete arrangement, and Fig. 7 gives the details of the sample support.

The prepared sample is waxed to a microscope slide and covered by a half-silvered mirror. Both are placed on the adjustable lower jaw of the sample support. The lower jaw is raised so that the upper jaw presses against the mirror. In this position it is fixed by tightening the screw in the back. Thus the mirror and sample are in contact, and the fringes can be observed through the microscope. Three screws in the lower jaw make it possible to change the relative position of mirror and sample. Thus the fringe pattern can be adjusted to make it most suitable for the particular case.

THE MEASUREMENTS

The measurement of a layer thickness was chosen to demonstrate the principle of evaluating a fringe pattern. (See Fig. 8.) The first illustration

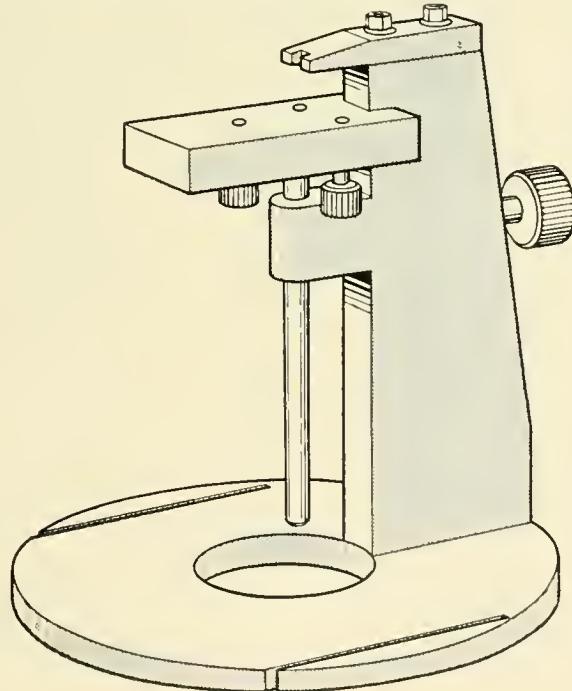


Fig. 7 — Sample support.

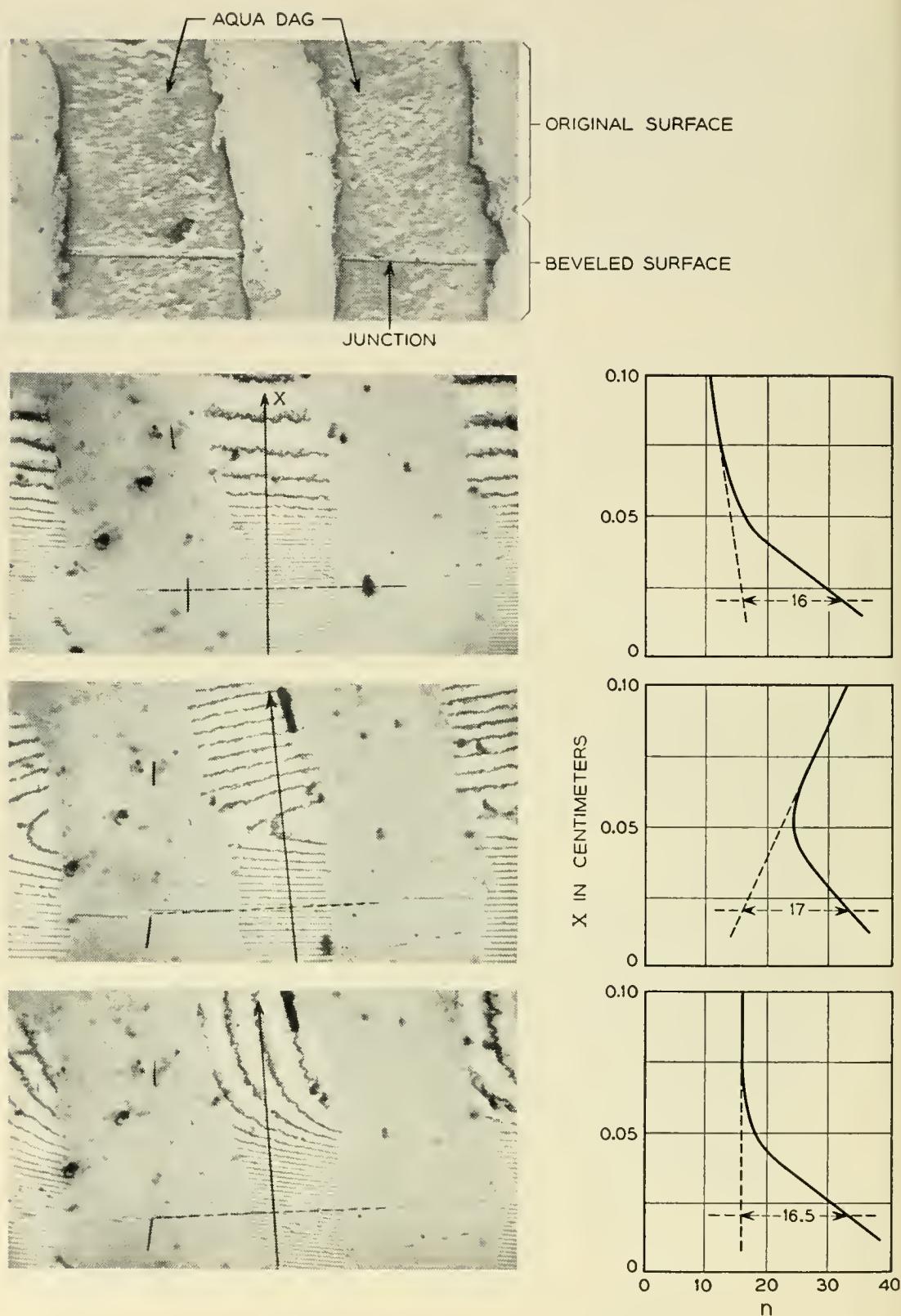


Fig. 8 — Evaluation of the interference fringe pattern on a scribed *p-n* junction.

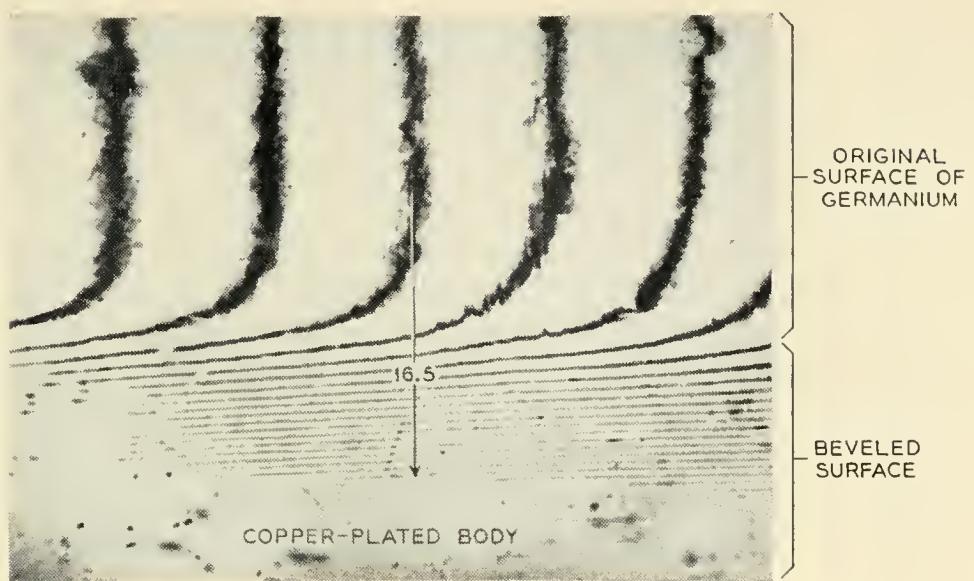


Fig. 9 — Evaluation of the interference fringe pattern on a Cu-plated *p-n* junction.

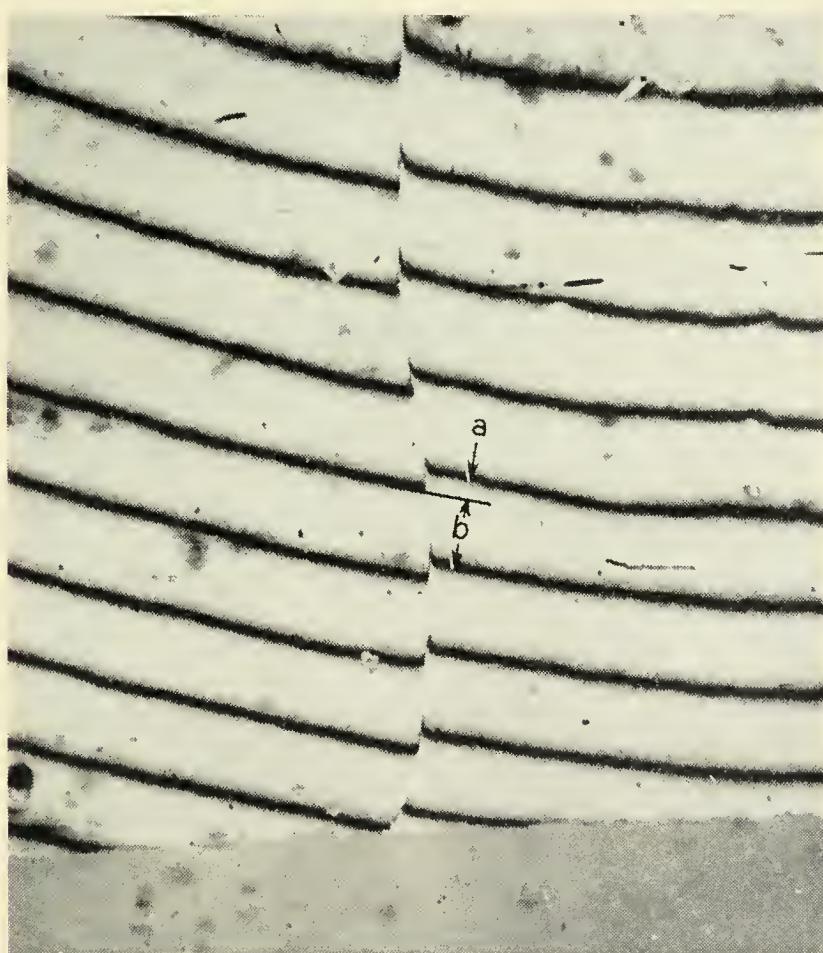


Fig. 10 — Evaluation of the interference fringe pattern of a shallow step in a surface.

shows a sample with aquadag coating and the lines marking the p-n junction. Under this, three typical fringe patterns obtained on this sample are presented. As pointed out in the beginning, one can regard the fringe pattern as contour lines for the distance between mirror and sample. The profile along any arbitrary straight line will show the structure of the sample surface. The profiles along the marked x-axes are shown to the right in each case. They were obtained by plotting the points of intersection between the n -th fringe and the x-axis against n . The original surface and the bevel (in this case 1°) are easily recognized. The dashed line is an extrapolation of the original surface. The vertical line marks the position of the p-n junction. The layer thickness is obtained as the difference in n at this point between the extrapolated original surface and the beveled surface. Note that the beveled surface need not necessarily be flat.

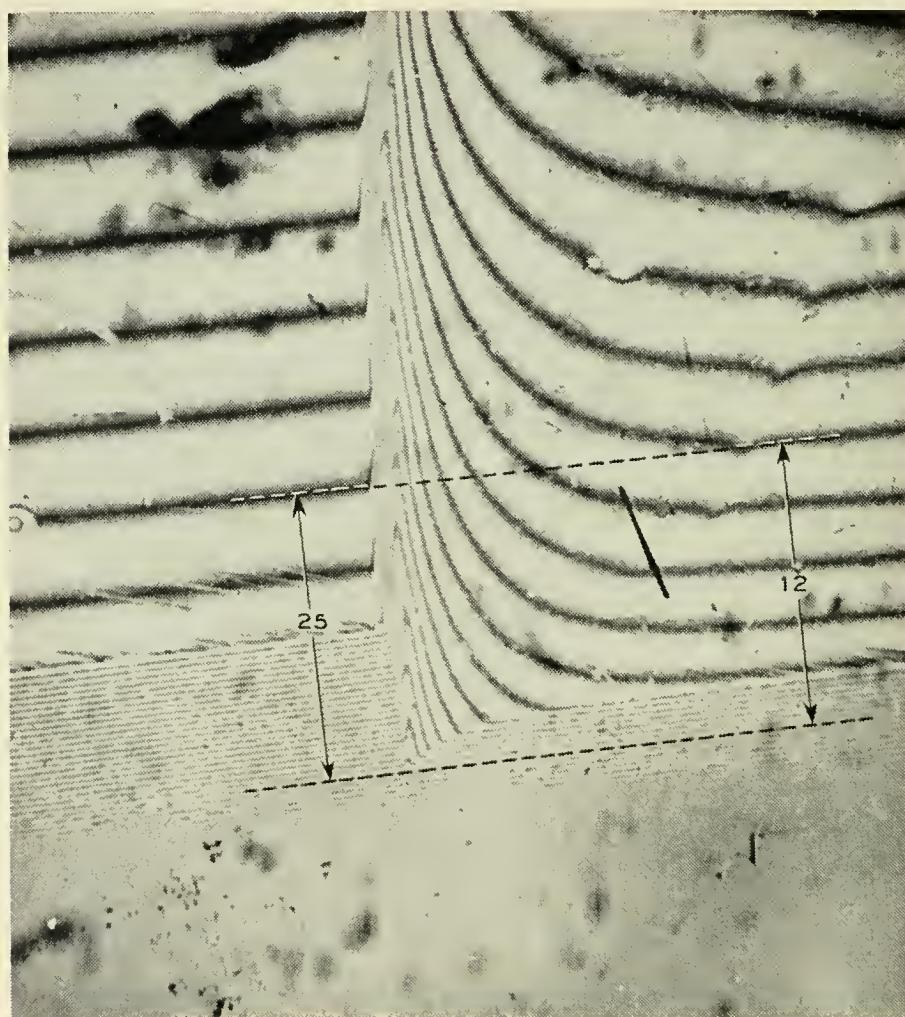


Fig. 11 — Evaluation of the interference fringe pattern of a steep and high step in a surface.

In the second case the fringes turn back. With a slightly different setting of the mirror the fringes could be almost parallel with the turning point outside the field of sight. The fringe pattern then resembles the first case, and therefore it might be easily misinterpreted.

The third case makes the plot unnecessary. The x-axis is chosen in such a way that it coincides with a fringe in the original surface. Thus, in the profile plot, the original surface is horizontal. Hence, the layer thickness can be obtained by counting the number of intersecting fringes between original surface and the p-n junction. This is, therefore, the most convenient setting of the mirror.

The noted number gives in each case the layer thickness in "fringes." All three cases are in essential agreement. The layer thickness in this case is

$$\begin{aligned}\Delta n \times \frac{\lambda}{2} &= (16.5 \pm 0.5) \times \frac{5.89}{2} \times 10^{-5} \text{ cm} \\ &= (4.85 \pm 0.15) \times 10^{-4} \text{ cm}\end{aligned}$$

Fig. 9 gives the fringe pattern obtained with a silicon p-n junction marked by the plating procedure.

The evaluation of steps in a surface is shown for two cases. The very shallow step in Fig. 10 is an example in which fractions of $\lambda/2$ are to be measured. The step here is

$$\frac{a}{a+b} \times \frac{\lambda}{2} = 0.195 \times \frac{5.89}{2} \times 10^{-5} \text{ cm} = 5.75 \times 10^{-6} \text{ cm}$$

In Fig. 11 the step is so high and steep that it is impossible to correlate the fringes crossing the step. But with the aid of the bevel, seen in the lower part of Figure 11, a correlation is possible. The height of the step along the drawn line is

$$(25 - 12) \times \frac{\lambda}{2} = 13 \times \frac{5.89}{2} \times 10^{-5} \text{ cm} = 3.8 \times 10^{-4} \text{ cm}$$

The accuracy of the method depends mainly on the quality of the optically flat mirror since it serves as a plane of reference. A thin mirror is likely to be slightly bent under the pressure of the clamp. Therefore, it is advisable not to work with too high a pressure. For the measurement of layer thicknesses the quality of the original surface is also important. An accuracy of 5 per cent is easily obtained using half-silvered microscope slides for the mirror. These slides are essentially flat over the small region covered by the microscope.

Bell System Technical Papers Not Published in This Journal

ALBRECHT, E. G.,⁵ DIETZ, A. E.,¹ CHRISTOFERSON, E. W.,⁶ and SLOTHOWER,, J. C.⁶

Co-ordinated Protection for Open-Wire Joint Use — Minneapolis Tests, A.I.E.E. Commun. and Electronics, **24**, pp. 217–223, May, 1956.

ANDERSON, P. W.¹

Note on Ordering and Antiferromagnetism in Ferrites, Phys. Rev., **102**, pp. 1008–1013, May 15, 1956.

ATALLA, M. M., see Preston, K., Jr.

BAKER, W. O., see Winslow, F. H.

BENSON, K. E., see Goss, A. J.

BENNETT, W. R.¹

Techniques for Measuring Noise. Part III, Electronics, **29**, pp. 162–165, May, 1956.

BENNETT, W. R.¹

Electrical Noise, Part IV. Design of Low Noise Equipment, Electronics, **29**, pp. 154–157, June, 1956.

BENNETT, W. R.¹

Electrical Noise. Part V. Noise Reduction in Communication Systems, Electronics, **29**, pp. 148–151, July, 1956.

BENNETT, W. R.¹

Methods of Solving Noise Problems, Proc. I.R.E., **44**, pp. 609–638, May, 1956.

¹ Bell Telephone Laboratories Inc.

⁵ Northwestern Bell Telephone Company.

⁶ Northern State Power Company, Minneapolis.

BOGERT, B. P.¹

The VOBANC — A Two-to-One Bandwidth Reduction System. *J. Acous. Soc.*, **28**, pp. 399–404, May, 1956.

BONNEVILLE, S., see Noyes, J. W.

BOYD, R. C.¹

Objectives and General Description of the Type-P1 Carrier System, *A.I.E.E. Commun. and Electronics*, **24**, pp. 188–191, May, 1956.

BOYET, H., see Weisbaum, S.

BULLARD, W. R.⁴ and WEPPLER, H. E.²

Co-ordinated Protection for Open-Wire Joint Use — Present Trends, *A.I.E.E. Commun. and Electronics*, **24**, pp. 215–216, May, 1956.

CHYNOWETH, A. G.¹

Surface Space Charge Layers in Barium Titanate, *Phys. Rev.*, **102**, pp. 705–714, May 1, 1956.

CHYNOWETH, A. G.¹

Spontaneous Polarization of Guanidine Aluminum Sulfate Hexahydrate at Low Temperatures, *Phys. Rev.*, **102**, pp. 1021–1023, May 15, 1956.

CHYNOWETH, A. G.¹ and MCKAY, K. G.¹

Photon Emission From Avalanche Breakdown in Silicon, *Phys. Rev.* **102**, pp. 369–376, Apr. 15, 1956.

DIETZ, A. E., see Albrecht, E. G.

DITZENBERGER, J. A., see Fuller, C. S.

DUDLEY, H. W.¹

Fundamentals of Speech Synthesis, *J. Audio Engg. Soc.*, **3**, pp. 170–185, Oct., 1955.

EBERHART, E. K.¹ HALLENBECK, F. J.¹ AND PERKINS, E. H.¹

Circuit and Equipment Descriptions of Type-P1 Carrier System, *A.I.E.E. Commun. and Electronics*, **24**, pp. 195–204, May, 1956.

¹ Bell Telephone Laboratories Inc.

² American Telephone and Telegraph Company, Inc.

⁴ Ebasco Services, Inc., New York.

ELLIS, H. M.,⁷ PHELPS, J. W.,¹ ROACH, C. L.,⁸ and TREEN, R. E.⁷

Co-ordinated Protection for Open-Wire Joint Use — Ontario Tests,
A.I.E.E. Commun. and Electronics, **24**, pp. 223–236, May, 1956.

FULLER, C. S.,¹ and DITZENBERGER, J. A.¹

Diffusion of Donor and Acceptor Elements in Silicon, J. Appl. Phys., **27**, pp. 544–553, May, 1956.

FULLER, C. S.¹

Some Analogies Between Semiconductors and Electrolyte Solutions,
Record of Chem. Progress, **17**, pp. 75–93, No. 2, 1956.

GARRETT, C. G. B., see Law, J. T.

GASTON, C. M.¹

Stop Playing Hide-and-Seek with Engineering Drawings, Iron Age Magazine, **177**, pp. 100–101, May 17, 1956.

GAUDET, S., see Noyes, J. W.

GELLER, S.¹

The Crystal Structure of Gadolinium Orthoferrite, GdFeO_3 , J. Chem. Phys., **24**, pp. 1236–1239, June, 1956.

GILLEO, M. A.¹

Magnetic Properties of a Gadolinium Orthoferrite, GdFeO_3 Crystal, J. Chem. Phys., **24**, pp. 1239–1243, June, 1956.

GILOTH, P. K.¹

A Simulator for Analysis of Sampled Data Control Systems, Proc. Natl. Simulation Conf., pp. 21.1–21.8, Jan., 1956.

GOSS, A. J.,¹ BENSON, K. E.,¹ and PFANN, W. G.¹

Dislocations at Compositional Fluctuations in Germanium-Silicon Alloys, Acta Met., Letter to the Editor, **4**, pp. 332–333, May, 1956.

HALLENBECK, F. J., see Eberhart, E. K.

HARROWER, G. A.¹

Auger Electron Emission in the Energy Spectra of Secondary Electrons from Mo and W., Phys. Rev., **102**, pp. 340–347, Apr. 15, 1956.

¹ Bell Telephone Laboratories Inc.

⁷ Hydro-Electric Power Commission of Ontario, Toronto, Ont., Canada.

⁸ Bell Telephone Company of Canada, Montreal, Que., Canada.

HARROWER, G. A.¹

Dependence of Electron Reflection on Contamination of the Reflecting Surface, Phys. Rev., 102, pp. 1288-1289, June 1, 1956.

HOWARD, J. D., JR.²

Application of the Type-P1 Carrier System to Rural Telephone Lines, A.I.E.E. Commun. and Electronics, 24, pp. 205-214, May, 1956.

HUTSON, A. R.¹

Effect of Water Vapor on Germanium Surface Potential, Phys. Rev., 102, pp. 381-385, Apr. 15, 1956.

KATZ, D.¹

A Magnetic Amplifier Switching Matrix, A.I.E.E. Commun. and Electronics, 24, pp. 236-241, May, 1956.

KOWALCHIK, M., see Trumbore, F. A.

LAW, J. T.¹ and GARRETT, C. G. B.¹

Measurements of Surface Electrical Properties of Bombardment-Cleaned Germanium, J. Appl. Phys., 27, p. 656, June, 1956.

LEWIS, H. W.¹

Two-Fluid Model of an "Energy-Gap" Superconductor, Phys. Rev., 102, pp. 1508-1511, June 15, 1956.

LOGAN, R. A., see Thurmond, C. D.

LOZIER, J. C.¹

A Study State Approach to the Theory of Saturable Servo Systems, I.R.E. Trans., PGAC, 1, pp. 19-39, 1956.

LUNDBERG, J. L.¹ and ZIMM, B. H.³

Sorption of Vapors by High Polymers, J. Phys. Chem., 60, pp. 425-428, Apr. 16, 1956.

MATTHIAS, B. T.¹ and REMEYKA, J. P.¹

Ferroelectricity in Ammonium Sulfate, Phys. Rev., Letter to the Editor, 103, p. 262, July 1, 1956.

¹ Bell Telephone Laboratories Inc.

² American Telephone and Telegraph Company, Inc.

³ General Electrical Research Laboratories.

MCKAY, K. G., see Chynoweth, A. G.

MCLEAN, D. A.¹

Tantalum Solid Electrolytic Capacitors, Proc. Natl. Conf. Aeronautical Electronics, pp. 289-294, May, 1956.

McSKIMIN, H. J.¹

Propagation of Longitudinal Waves and Shear Waves in Cylindrical Rods at High Frequencies, J. Acous. Soc., 28, pp. 484-494, May, 1956.

NOYES, J. W.,⁸ GAUDET, G.,⁸ and BONNEVILLE, S.⁸

Development of Communications in Canada, Elec. Engg., 75, p. 539, June, 1956.

O'BRIEN, J. A.¹

Cyclic Decimal Codes for Analog to Digital Converters, A.I.E.E. Commun. and Electronics, 24, pp. 120-122, May, 1956.

OWENS, C. D.¹

Stability Characteristics of Molybdenum Permalloy Powder Cores, Elec. Engg., 75, pp. 252-256, Mar., 1956.

PEARSON, G. L.¹

Electricity from the Sun, Proc. World Symp. Appl. Solar Energy, pp. 281-288, Book.

PERKINS, E. H., see Eberhart, E. K.

PFANN, W. G.¹

Zone Melting: A Fresh Outlook for Fractional Crystallization, Chem. & Engg. News, 34, pp. 1440-1443, Mar. 26, 1956.

PFANN, W. G., see Goss, A. J.

PHELPS, J. W.¹

Protection Problems in Telephone Distribution Systems, Wire and Wire Products, 31, pp. 555-596, May, 1956.

PHELPS, J. W., see Ellis, H. M.

¹ Bell Telephone Laboratories Inc.

⁸ Bell Telephone Company of Canada, Montreal, Que., Canada.

PIERCE, J. R.¹

Physical Sources of Noise, Pro. I.R.E., **44**, pp. 601-608, May, 1956.

POMEROY, A. F.¹ and SUAREZ, E. M.¹

Determining Attenuation of Waveguide From Electrical Measurements on Short Samples, I.R.E. Trans. **MTT-4**, pp. 122-129, Apr., 1956.

PONDY, P. R.¹

Dust-Lint Control in Tube Fabrication, Electronics, **29**, pp. 246-250, June, 1956.

PRESTON, K., JR.¹ and ATALLA, M. M.¹

Transient Temperature Rise in Semi-Infinite Solid Due to a Uniform Disc Source, J. Appl. Mechanics, **23**, p. 313, June, 1956.

PRINCE, E.¹

Neutron Diffraction Observation of Heat Treatment in Cobalt Ferrite, Phys. Rev., **102**, pp. 674-676, May 1, 1956.

REMEIKA, J. P., see Matthias, B. T.

REISS, H.¹

p-n Junction Theory by the Method of Functions, J. Appl. Phys., **27**, pp. 530-537, May, 1956.

RICE, S. O.¹

A First Look at Random Noise, A.I.E.E. Commun. and Electronics, **24**, pp. 128-131, May, 1956.

SMITH, D. H.¹

Power Supplies for the Type-P1 Carrier System, A.I.E.E. Commun. and Electronics, **24**, pp. 191-195, May, 1956.

SUAREZ, E. M., see Pomeroy, A. F.

THEUERER, H. C.¹

Purification of Germanium Tetrachloride by Extraction with Hydrochloric Acid and Chlorine, J. of Metals, **8**, pp. 688-690, May, 1956.

¹ Bell Telephone Laboratories Inc.

THURMOND, C. D.,¹ and LOGAN, R. A.¹

The Distribution of Copper Between Germanium and Ternary Melts Saturated with Germanium, J. Phys. Chem., **60**, pp. 591-594, May, 1956.

THURMOND, C. D., see Trumbore, F. A.

TRUMBORE, F. A.,¹ THURMOND, C. D.,¹ and KOWALCHIK, M.¹

The Germanium-Oxygen System, J. Chem. Phys., Letter to the Editor, **24**, p. 1112, May, 1956.

WEISBAUM, S.¹ and BOYET, H.¹

Broadband Non-Reciprocal Phase Shifts — Analysis of Two Ferrite Slabs in Rectangular Guide, J. Appl. Phys., **27**, pp. 519-524, May, 1956.

WEPPLER, H. E., see Bullard, W. R.

WINSLOW, F. H.,¹ BAKER, W. O.,¹ and YAGER, W. A.¹

The Structure and Properties of Some Pyrolyzed Polymers, Proc. Conf. on Carbon, pp. 93-102, 1956.

WOOD, E. A. MRS. ¹

The Question of a Phase Transition in Silicon, J. Phys. Chem., **60**, pp. 508-509, Apr., 1956.

YAGER, W. A., see Winslow, F. H.

¹ Bell Telephone Laboratories Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BASIKOW, T. R.

DC Graphical Analysis of Junction Transistor Flip-Flops, Monograph 2615.

BECKER, J. A., see Rose, D. J.

BITTRICH, G., see Compton, K. G.

BOYET, H., see Weisbaum, S.

BRANDES, R. G., see Rose, D. J.

BRATTAIN, W. H., see Garrett, C. G. B.

COMPTON, K. G., EHRHARDT, R. A., and BITTRICH, G.

Brass Plating, Monograph 2467.

EGERTON, L., and KOONCE, S. E.

Structure and Properties of Barium Titanate Ceramics, Monograph 2517.

EHRHARDT, R. A., see Compton, K. G.

EIGLER, J. H., see Sullivan, M. V.

FRANCOIS, E. E., see Law, J. T.

FULLER, C. S., see Reiss, H.

GARRETT, C. G. B., and BRATTAIN, W. H.

Some Experiments on, and a Theory of, Surface Breakdown, Monograph 2589.

HAGELBARGER, D. W.

SEER, A Sequence Extrapolating Robot, Monograph 2599.

* 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.

IAYNES, J. R., and WESTPHAL, W. C.

Radiation Resulting from Recombination of Holes and Electrons in Silicon, Monograph 2622.

IERRING, C., and VOGT, E.

Transport and Deformation-Potential Theory for Many-Valley Semiconductors, Monograph 2596.

IERRING, C., see Vogel, F. L., Jr.

KLEIMACK, J. J., see Wahl, A. J.

KONCE, S. E., see Egerton, L.

LAW, J. T., and FRANCOIS, E. E.

Adsorption of Gases on a Silicon Surface, Monograph 2600.

LEWIS, H. W.

Superconductivity and Electronic Specific Heat, Monograph 2597.

LOGAN, R. A.

Thermally Induced Acceptors in Germanium, Monograph 2601.

LUNDBERG, J. L., see Zimm, B. H.

MAY, J. E., JR.

Low-Loss 1000-Microsecond Ultrasonic Delay Lines, Monograph 2584.

MENDEL, J. T.

Microwave Detector, Monograph 2602.

PATERSON, E. G. D.

An Over-all Quality Assurance Plan, Monograph 2630.

POMEROY, A. F., and SUAREZ, E. M.

Attenuation of Waveguide from Electrical Measurements on Short Samples, Monograph 2625.

PRESS, H., and TUKEY, J. W.

Power Spectral Methods of Analysis and Application in Airplane Dynamics, Monograph 2606.

READ, W. T., JR., see Vogel, F. L., Jr.

REISS, H., and FULLER, C. S.

Influence of Holes and Electrons on Solubility of Lithium in Silicon,
Monograph 2603.

ROSE, D. J., BECKER, J. A., and BRANDES, R. G.

On the Field Emission Electron Microscope, Monograph 2588.

SUAREZ, E. M., see Pomeroy, A. F.

SULLIVAN, M. V., and EIGLER, J. H.

Electrolytic Stream Etching of Germanium, Monograph 2595.

THOMAS, D. E.

Tables of Phase of a Semi-Infinite Unit Attenuation Slope, Monograph 2550.

TUKEY, J. W., see Press, H.

VAN UITERT, L. G.

High-Resistivity Nickel Ferrites-Minor Additions of Manganese or Cobalt, Monograph 2594.

VOGT, E., see Herring, C.

VOGEL, F. L., JR., HERRING, C., AND READ, W. T., JR.

Dislocations in Plastic Deformation, Monograph 2616.

WAHL, A. J., and KLEIMACK, J. J.

Factors Affecting Reliability of Alloy Junction Transistors, Monograph 2604.

WESTPHAL, W. C., see Haynes, J. R.

WEISBAUM, S., AND BOYET, H.

A Double-Slab Ferrite Field Displacement Isolator at 11 kmc, Monograph 2605.

ZIMM, B. H., and LUNDBERG, J. L.

Sorption of Vapors by High Polymers, Monograph 2573.

Contributors to This Issue

W. L. BOND, B.S. 1927 and M.S. 1928, Washington State College; Bell Telephone Laboratories, 1928-. Mr. Bond has conducted investigations in the mineral field including studies of the piezoelectric effect in minerals and similar studies of synthetic crystals. He has designed optical, X-ray, and mechanical tools and instruments for the orientation, cutting and processing of crystals. Mr. Bond also served as consultant on quartz crystals with the War Production Board. He is a member of the American Physical Society, and of the American Crystallographic Association.

WALTER H. BRATTAIN, B.S., Whitman College, 1924; M.A., University of Oregon, 1926; Ph.D., University of Minnesota, 1929. Honorary D.Sc. Portland University, 1952, Whitman College and Union College, 1955. Radio section, Bureau of Standards, 1928-29. Bell Telephone Laboratories, 1929-. Co-inventor with Dr. John Bardeen of point contact transistor. Primary activity at Laboratories in semi-conductors. Research in field of thermionics, particularly electronic emission from hot surfaces. Frequency standards, magnetometers and infra-red phenomena. Studied magnetic detection of submarines for National Defense Research Committee at Columbia University, 1942-43. Visiting lecturer at Harvard University, 1952-53. Author of numerous technical articles. Recipient of John Scott Medal, 1955, and Stuart Ballantine Medal of Franklin Institute, 1952. Fellow of American Physical Society, American Academy of Arts and Sciences and American Association for the Advancement of Science. Member of Franklin Institute, Phi Beta Kappa and Sigma Xi.

C. C. COLE, B.S. in E.E., State College of Washington, 1923; U. S. Navy 1917-1919; Western Electric Company 1923-. His first assignment was in manufacturing development on paper and mica capacitors. Other assignments include manufacturing development on loading coils, quality control, and inspection development laboratory. During World War II he handled the design and construction of testing facilities for various defense projects. Since World War II he has been engaged in

inspection methods development and in the development and design of testing facilities for telephone apparatus and cable. Member of Sigma Tau and A.I.E.E.

ARTHUR B. CRAWFORD, B.S.E.E. 1928, Ohio State University; Bell Telephone Laboratories, 1928-. Mr. Crawford has been engaged in radio research since he joined the Laboratories. He has worked on ultra short wave apparatus, measuring techniques and propagation; microwave apparatus, measuring techniques and radar, and microwave propagation studies and microwave antenna research. He is author or co-author of articles which appeared in *The Bell System Technical Journal*, *Proceedings of the I.R.E.*, *Nature*, and the *Bulletin of the American Meteorological Society*. He is a Fellow of the I.R.E. and a member of Sigma Xi, Tau Beta Pi, Eta Kappa Nu, and Pi Mu Epsilon.

HARALD T. FRIIS, E.E., 1916, D.Sc., 1938, Royal Technical College (Copenhagen); Engineering Department of the Western Electric Company, 1919-1924. Bell Telephone Laboratories, 1925-. Dr. Friis, Director of Research in High Frequency and Electronics, has made important contributions on ship-to-shore radio reception, short-wave studies, radio transmission (including methods of measuring signals and noise), a receiving system for reducing selective fading and noise interference, microwave receivers and measuring equipment, and radar equipment. He has published numerous technical papers and is co-author of a book on the theory and practice of antennas. The I.R.E.'s Morris Liebmann Memorial Prize, 1939, and Medal of Honor, 1954. Valdemar Poulsen Gold Medal by Danish Academy of Technical Sciences, 1954. Danish "Knight of the Order of Dannebrog," 1954. Fellow of I.R.E. and A.I.E.E. Member of American Association for the Advancement of Science, Danish Engineering Society and Danish Academy of Technical Sciences. Served on Panel for Basic Research of Research and Development Board, 1947-49, and Scientific Advisory Board of Army Air Force, 1946-47.

C. G. B. GARRETT, B.A., Cambridge University (Trinity College), 1946; M.A., Cambridge, 1950; Ph.D., Cambridge, 1950. Instructor in Physics, Harvard University, 1950-52. Bell Telephone Laboratories, 1952-. Before coming to the Laboratories, Dr. Garrett's principal research was in the field of low-temperature physics. At the Laboratories he has been engaged in research and exploratory development on semiconductor surfaces and, for the past year, has supervised a group working in this field. He is the author of "Magnetic Cooling" (Harvard

University Press, 1954). Senior Scholar of Trinity College, Cambridge, 1945. Twisden Student of Trinity College, 1949. Fellow of Physical Society (London). Member of American Physical Society.

L. D. HANSEN, B.S., Montana State College, 1924; Western Electric Company, 1924-. Mr. Hansen joined the Equipment Engineering Organization at the Hawthorne Plant of The Western Electric Company in Chicago in 1924 where he was engaged in preparation of telephone central office specifications. He transferred to the Kearny, N. J., Plant in 1928 where he was promoted to section chief in 1929. He transferred to the Engineer of Manufacture Organization in 1930 and worked on carrier and repeater test development and methods until 1941 when he was promoted to Department Chief in charge of wired switching apparatus and equipment test set development and methods.

WILLIAM C. JAKES, JR., B.S.E.E., Northwestern University, 1944; M.S., Northwestern, 1947; Ph.D., Northwestern, 1948. Bell Telephone Laboratories, 1949-. Dr. Jakes is engaged in microwave antenna and propagation studies and holds a patent in microwave antennas. He is the author of chapter in antenna engineering handbook (McGraw-Hill). Member of Sigma Xi, Pi Mu Epsilon, Eta Kappa Nu, I.R.E. and Phi Delta Theta.

AMOS E. JOEL, JR., B.S., Massachusetts Institute of Technology, 1940; M.S., M.I.T., 1942; Bell Telephone Laboratories, 1940-. Mr. Joel is Switching Systems Development Engineer responsible for coordinating the exploratory development of a trial electronic switching system. Prior to his present position he worked on relay engineering, crossbar test laboratory, fundamental development studies, circuits for relay computers, preparation of a text and teaching switching design, designing AMA computer circuits and making fundamental engineering studies on new switching systems. He holds some forty patents. Member of A.I.E.E., I.R.E., Sigma Xi and Association for Computing Machinery.

ARCHIE P. KING, B.S., California Institute of Technology, 1927. After three years with the Seismological Laboratory of the Carnegie Institution of Washington, Mr. King joined Bell Telephone Laboratories in 1930. Since then he has been engaged in ultra-high-frequency radio research at the Holmdel Laboratory, particularly with waveguides. For the last ten years Mr. King has concentrated his efforts on waveguide transmission and waveguide transducers and components for low-loss circular

electric wave transmission. He holds at least a score of patents in the waveguide field. Mr. King was cited by the Navy for his World War II radar contributions. He is a Senior Member of the I.R.E. and is a Member of the American Physical Society.

D. T. ROBB, B.S., University of Chicago, 1927; Western Electric Company, 1927-. Mr. Robb has been concerned with measurement and testing problems throughout his career. In the electrical laboratory at Hawthorne Works, Chicago, he specialized in ac standardization. Later he worked on the development of shop test methods and test sets. In 1944 he transferred to take charge of radar test engineering at the Eleventh Avenue Plant of Western Electric in New York City. In 1946 he supervised the engineering of the standards laboratory at Chatham Road Plant in Winston Salem, N. C. Currently, he has charge of transmission test set development and test set design at Kearny Works, N. J.

HARRY R. SHILLINGTON, B.S. in E.E. Iowa State College, 1937; Long Lines Department of the American Telephone and Telegraph Company, 1928-1932; Western Electric Company, 1937-. Mr. Shillington's first assignment was that of product engineering on panel dial equipment. During World War II and the Korean War he was engaged in test engineering on various defense projects. He is presently concerned with the development of special test facilities for telephone apparatus. Member of Eta Kappa Nu and Tau Beta Pi.

FRIEDOLF M. SMITS, Dipl.Phys. and Dr.Rer.Nat., University of Freiburg, Germany, 1950; research assistant, Physikalisches Institut, University of Freiburg, 1950-54; Bell Telephone Laboratories, 1954-. As a member of the Solid State Electronics Research Department of the Laboratories, Dr. Smits has been concerned with diffusion studies of germanium and silicon for semiconductor device applications. He is a member of the American Physical Society and the German Physical Society.

FRANK H. TENDICK, Jr., B.S.E.E., 1951, University of Michigan; Bell Telephone Laboratories, 1951-. Mr. Tendick was first engaged in work pertaining to the synthesis of networks employed in the L3 coaxial cable system. Later he engaged in the design of transistor networks for digital computers. More recently, he has been associated with exploratory studies of submarine cable systems. He is a member of the I.R.E. Mr.

Tendick also belongs to four honor societies, Tau Beta Pi, Eta Kappa Nu, Sigma Xi and Phi Kappa Phi.

LEISHMAN R. WRATHALL, B.S., 1927, University of Utah. Mr. Wrathall did another year of graduate work at the University of Utah and joined Bell Telephone Laboratories in 1929. For many years he was primarily concerned with studies of the characteristics of non-linear coils and capacitors. During World War II non-linear coils were used extensively in radar systems, and his work in this field was intensified. Later he was occupied with general circuit research. He is now engaged in studies of conductor problems, particularly digital repeaters, as a member of the Transmission Research Department at Murray Hill.



H E B E L L S Y S T E M

Technical Journal

VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXV

NOVEMBER 1956

NUMBER 6

Nobel Prize in Physics Awarded to Transistor Inventors

DEC 14 1956

i

Theory of the Swept Intrinsic Structure

W. T. READ, JR. 1239

A Medium Power Traveling-Wave Tube for 6,000-Mc Radio Relay

J. P. LAICO, H. L. McDOWELL AND C. R. MOSTER 1285

Helix Waveguide

S. P. MORGAN AND J. A. YOUNG 1347

Wafer-Type Millimeter Wave Rectifiers

W. M. SHARPLESS 1385

Frequency Conversion by Means of a Nonlinear Admittance

C. F. EDWARDS 1403

Minimization of Boolean Functions

E. J. MCCLUSKEY, JR. 1417

Detection of Group Invariance or Total Symmetry of a Boolean
Function

E. J. MCCLUSKEY, JR. 1445

Bell System Technical Papers Not Published in This Journal 1454

Recent Bell System Monographs 1461

Contributors to This Issue 1465

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

A. B. GOETZE, *President, Western Electric Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

B. McMILLAN, *Chairman*

S. E. BRILLHART

E. I. GREEN

A. J. BUSCH

R. K. HONAMAN

L. R. COOK

H. R. HUNTLEY

A. C. DICKIESON

F. R. LACK

R. L. DIETZOLD

J. R. PIERCE

K. E. GOULD

G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

Nobel Prize in Physics Awarded to Transistor Inventors

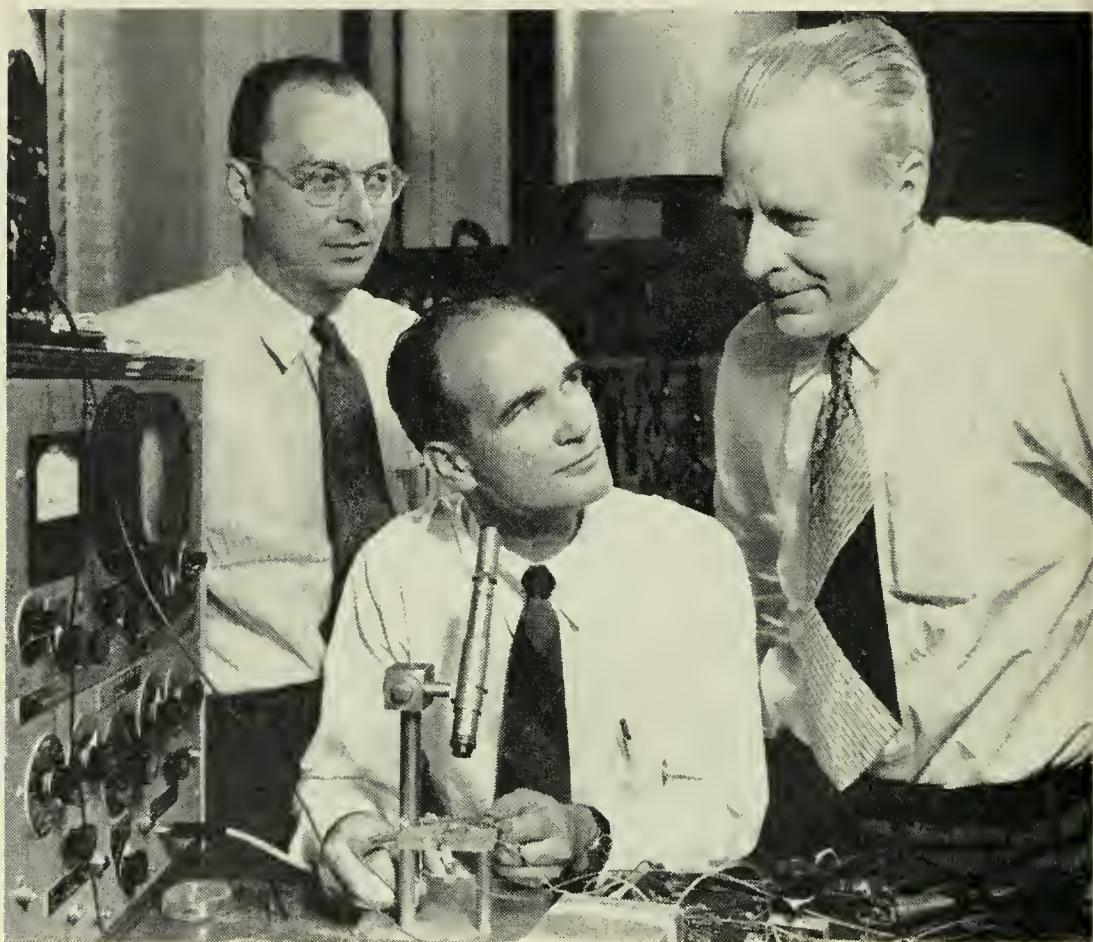
The Swedish Royal Academy of Sciences announced on November 1 that a Nobel Prize in Physics, most highly coveted award in the world of physics, had been awarded jointly to Dr. Walter H. Brattain of the Laboratories Physical Research Department, with Dr. John Bardeen and Dr. William Shockley, both former members of the Laboratories. The prize was awarded for "investigations on semiconductors and the discovery of the transistor effect."

This marks the second time that work done at the Laboratories has been recognized by a Nobel Prize. The previous recipient was Dr. C. J. Davisson who shared in the 1937 prize for his discovery of electron diffraction as a result of experiments carried out with Dr. L. H. Germer, also of the Laboratories.

Each of the three winners of this year's prize will receive a gold medal, a diploma and a share of the \$38,633 prize money. When he was notified that he was one of these winners, Dr. Brattain said, "I certainly appreciate the honor. It is a great satisfaction to have done something in life and to have been recognized for it in this way. However, much of my good fortune comes from being in the right place, at the right time, and having the right sort of people to work with."

The principle of transistor action was discovered as a result of fundamental research directed toward gaining a better understanding of the surface properties of semiconductors. Following World War II, intensive programs on the properties of germanium and silicon were undertaken at the Laboratories under the direction of William Shockley and S. O. Morgan. One group in this program engaged in a study of the body properties of semi-conductors, and another on the surface properties. Dr. John Bardeen served as theoretical physicist and R. B. Gibney as chemist for both groups. These investigations, which resulted in the invention of the transistor, made extensive use of knowledge and techniques developed by scientists here and elsewhere, particularly by members of the Laboratories—R. S. Ohl, J. H. Seaff and H. C. Theuerer.

Since the transistor was announced, little more than eight years ago, it has become increasingly important in what has been called the "new



The Nobel Prize winners in an historic photograph taken in 1948 when the announcement of the invention of the transistor was made. Left to right, John Bardeen, William Shockley and Walter H. Brattain.

electronics age." As new transistors and related semiconductor devices are developed and improved, the possible fields of application for these devices increase to such an extent that they may truly be said to have "revolutionized the electronics art."

The invention of the transistor, basis for the Nobel Prize award, represents an outstanding example of the combination of research teamwork and individual achievement in the Bell System that has meant so much to the rapid development of modern communications systems.

Dr. Brattain received a B.S. degree from Whitman College in 1924, an M.A. degree from the University of Oregon in 1926, and a Ph.D. degree from the University of Minnesota in 1928. He joined Bell Telephone Laboratories in 1929, and his early work was in the field of thermionics, particularly the study of electron emission from hot surfaces. He also studied frequency standards, magnetometers and infra-red phenomena.

Subsequently, Mr. Brattain engaged in the study of electrical conductivity and rectification phenomena in semiconductors. During World War II, he was associated with the National Defense Research Committee at Columbia University where he worked on magnetic detection of submarines.

Mr. Brattain has received honorary Doctor of Science degrees from Whitman College, Union College and Portland University. His many awards include the John Scott Medal and the Stuart Ballantine Medal, both of which he received jointly with John Bardeen. Mr. Brattain is a Fellow of the American Academy of Arts and Sciences.

Dr. Bardeen received the B.S. in E.E. and M.S. in E.E. degrees from the University of Wisconsin in 1928 and 1929 respectively, and his Ph.D. degree in Mathematics and Physics from Princeton University in 1936. After serving as an Assistant Professor of Physics at the University of Minnesota from 1938 to 1941, he worked with the Naval Ordnance Laboratory as a physicist during World War II. In 1945 he joined the Laboratories as a research physicist, and was primarily concerned

Clinton J. Davisson Previous Laboratories Nobel Laureate

In December, 1937, Dr. Clinton J. Davisson of the Laboratories was awarded the Nobel Prize in Physics for his discovery of electron diffraction and the wave properties of electrons.

He shared the prize with Professor G. P. Thompson of London, who worked in the same field, though there was little in common between their techniques. Dr. Davisson's work on electron diffraction started as an attempt to understand the characteristics of secondary emission in multi-grid electron tubes. In this work he discovered patterns of emission from the surface of single crystals of nickel. By studying these patterns, Dr. Davisson, with Dr. L. H. Germer and their associates, proved that reflected electrons have the properties of trains of waves.

Dr. Davisson was awarded the B.S. degree in physics from the University of Chicago in 1908 and the Ph.D. degree from Princeton in 1911. From September, 1911, until June, 1917, he was an instructor in physics at the Carnegie Institute of Technology, coming to the Laboratories on a wartime leave of absence. He found the climate of the Laboratories conducive to basic research, however, and remained until his retirement in 1946. Besides his work on electron diffraction, Dr. Davisson did much significant work in a variety of fields, particularly electron optics, magnetrons, and crystal physics.

with theoretical problems in solid state physics, including studies of semiconductor materials.

Mr. Bardeen, whose honors include an honorary Doctor of Science degree from Union College, the Stuart Ballantine Medal, the John Scott Medal, and the Buckley Prize, is a member of the National Academy of Sciences. He joined the University of Illinois in 1951.

Dr. Shockley received a B.Sc. degree from the California Institute of Technology in 1932, and a Ph.D. degree from the Massachusetts Institute of Technology in 1936. He joined the staff of Bell Telephone Laboratories in 1936. In addition to his many contributions to solid state physics and semiconductors, Mr. Shockley has worked on electron tube and electron multiplier design, studies of various physical phenomena in alloys, radar development and magnetism.

His many awards include an honorary degree from the University of Pennsylvania, the Morris Liebmann Memorial Prize, the Buckley Prize, the Comstock Prize and membership in the National Academy of Sciences. Dr. Shockley left the Laboratories to form the Shockley Semiconductor Laboratory at Beckman Instruments, Inc., in 1955.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXV

NOVEMBER 1956

NUMBER 6

Copyright 1956, American Telephone and Telegraph Company

Theory of the Swept Intrinsic Structure

By. W. T. READ, JR.

(Manuscript received March 4, 1956)

The electric field and the hole and electron concentrations are found for reverse biased junctions in which one side is either intrinsic (I) or so weakly doped that the space charge of the carriers cannot be neglected. The analysis takes account of space charge, drift, diffusion and non linear recombination. A number of figures illustrate the penetration of the electric field into a PIN structure with increasing bias for various lengths of the I region. For the junction between a highly doped and a weakly doped region, the reverse current increases as the square root of the voltage at high voltages; and the space charge in the weakly doped region approaches a constant value that depends on the fixed charge and the intrinsic carrier concentration.

The mathematics is greatly simplified by expressing the equations in terms of the electric field and the sum of the hole and electron densities.

I. INTRODUCTION

Applications have been suggested for semiconductor structures having both extrinsic and intrinsic regions. Examples are the "swept intrinsic" structure, in which a region of high resistivity is set up by an electric field that sweeps out the mobile carriers, and the analogue transistors, where the intrinsic region is analogous to the vacuum in a vacuum tube. However, the junction between an intrinsic region and an N or P region

is considerably less well understood than the simple *NP* junction. Most of the assumptions that make the *NP* case relatively simple to deal with do not apply to junctions where one side is intrinsic. Specifically, the space charge is that of the mobile carriers; thus the flow and electrostatic problems cannot be separated as they can in *PN* junction under reverse bias. The following sections analyze the *N*-intrinsic-*P* structure under reverse bias.

For a given material with fairly highly doped extrinsic regions, the problem is defined by the length of the intrinsic region and the applied voltage. Taking the intrinsic region infinitely long gives the solution for a simple *N*-intrinsic or *P*-intrinsic structure. The results are given and plotted in terms of the electric field distribution. From this the potential, space charge and carrier concentrations can be found; so also can the current-voltage curve. The final section considers the case where the middle layer contains some fixed charge but where the carrier charge cannot be neglected.

Qualitative Discussion of an N-intrinsic-P Structure

Consider an *N*-intrinsic-*P* structure where the intrinsic, or *I*, region is considerably wider than the zero bias, or built-in, space charge regions at the junctions, so that there is normal intrinsic material between the junctions. The field distribution at zero bias can be found exactly from the zero-current analysis of Prim.¹ Throughout the intrinsic region, hole and electron pairs are always being thermally generated and recombining at a rate determined by the density and properties of the traps, or recombination centers. Under zero bias the rates of generation and recombination are everywhere equal. Suppose now a reverse bias is applied causing holes to flow to the right and electrons to the left. Some of the carriers generated in the intrinsic region will be swept out before recombining. This depletes the carrier concentration in the intrinsic region and hence raises the resistivity. It also produces a space charge extending into the intrinsic layer. The electrons are displaced to the left and the holes, to the right. Thus the space charge opposes the penetration of the field into the intrinsic region; that is, the negative charge of the electrons on the left and positive charge of the holes on the right gives a field distribution with a minimum somewhere in the interior of the intrinsic region and maxima at the *NI* and *IP* junctions. If holes and electrons had equal mobilities, the field distribution would be symmetrical with a minimum in the center of the intrinsic region. Likewise, the total carrier

¹ R. C. Prim, B. S. T. J., **32**, p. 665, May, 1953.

concentration (holes plus electrons) would be symmetrical with a maximum in the center. As the applied bias is increased the hole and electron distributions are further displaced relative to one another and the space charge increases. Finally, at high enough biases, so many of the carriers are swept out immediately after being generated that few carriers are left in the intrinsic region. Now the space charge decreases with increasing bias until there is negligible space charge, and a relatively large and constant electric field extends through the intrinsic region from junction to junction. This may happen at biases that are still much too low to appreciably affect the high fields right at the junction or in the extrinsic layers, which remain approximately as they were for zero bias.

The current will increase with voltage until the total number of carriers in the intrinsic region becomes small compared to its normal value. After that, there is negligible further increase of current with voltage. All the carriers generated in the intrinsic region are being swept out before recombining. In general, the current will saturate while the minimum field in the intrinsic region is still small compared to the average field.

Comparison with the NP Structure

The analysis is more difficult than in a simple reverse-biased *NP* structure. In the *NP* case there is a well defined space charge region in which carrier concentration is negligible compared to the fixed charge of the chemical impurities; so the field and potential distributions are easily found from the known distribution of fixed charge. Outside of the space charge region are the diffusion regions in which the minority carrier concentration rises from a low value at the edge of the space charge region to its normal value deep in the extrinsic region. However, there is no space charge in this region because the majority carrier concentration, by a very small percentage variation, can compensate for the large percentage variation in minority carrier density. The minority carriers flow by diffusion. Since the disturbance in carrier density is small compared to the majority density, the recombination follows a simple linear law (being directly proportional to the excess of minority carriers). Thus the minority carrier distribution is found by solving the simple diffusion equation with linear recombination.

None of these simplifications extend to the *NIP* or *NI* or *IP* structure. There is, in the intrinsic region, no fixed charge; hence the space charge is that of the carriers. There is no majority carrier concentration to maintain electrical neutrality outside of a limited space charge region.

It is necessary to take account of (1) space charge, (2) carrier drift, (3) carrier diffusion and (4) recombination according to a nonlinear bimolecular law. Of these four, only space charge and recombination are never simultaneously important in practical cases. Nevertheless certain simplifications can be made if the problem is formulated so as to take advantage of them. The field and carrier distributions in the intrinsic region are found by joining two solutions: one solution is for charge neutrality; the other, which we shall call the no-recombination solution is for the case where the recombination rate is negligible compared to the rate of thermal generation of hole-electron pairs. We shall show that in practical cases the ranges of validity of the two solutions overlap; that is, wherever recombination is important, we have charge neutrality.

Prim's Zero-Current Approximation

Prim* derived the field distribution in a reverse biased *NIP* structure on the assumption that the hole and electron currents are negligibly small differences between their drift and diffusion terms, as in the zero-bias case. He showed that the average diffusion current is large compared to the average current. However, as it turns out, this is misleading. Throughout almost all of the intrinsic region (where the voltage drop occurs in practical cases) the diffusion current is comparable to or smaller than the total current. The larger average diffusion current comes from the extremely large diffusion current in the small regions of high space charge at the junctions. Prim's analysis, in effect, neglects the space charge of the carriers generated in the intrinsic region. These may be neglected in calculating the field distribution if the intrinsic region is sufficiently narrow or the reverse bias sufficiently high. In the appendix we derive the limits within which Prim's calculation of the field and potential will be valid. The range will increase with both the Debye length and the diffusion length in the intrinsic material. However, in cases of practical interest the zero-current approximation may lead to serious errors in the field distribution and give a misleading idea of the penetration of the field into the intrinsic region. The present, more general analysis, reduces to Prim's near the junctions where the zero-current assumption remains valid. The zero current approximation was, of course, not intended to give the hole and electron distributions in the intrinsic region or to evaluate the effects of interacting drift, diffusion and recombination.

* Ibid.

Outline of the Following Sections

Sections II through V deal with the ideal case of equal hole and electron mobilities. Here the problem is somewhat simplified and the physics easier to visualize because of the resulting symmetry. In Section VI, the general case of arbitrary mobilities is solved by an extension of the methods developed for solving the ideal case. The technique is to deal not with the hole and electron flow densities but with two linear combinations of hole and electron flow densities that have a simple form.

Section II deals with the basic relations and in particular the formula for recombination in an intrinsic region for large disturbances in carrier density. The nature and range of validity of the various approximations are discussed. Section III derives the field distribution in regions where recombination is small compared to pair generation. Section IV treats the recombination region and the smooth joining of the recombination and no-recombination solutions. Section V considers the role of diffusion in current flow and the situation at the junctions where the field and carrier concentration abruptly become large. The change in form of the solution near the junetions is shown to be represented by a basic instability in the governing differential equation. Section VI extends the results to the general case of unequal mobilities. Section VII deals with the still more general case where there is some fixed charge in the "intrinsic" region. If the density of excess chemical impurities is small compared to the intrinsic carrier density, the solution remains unchanged in the range where recombination is important. In the no-recombination region the solution is given by a simple first order differentiatial equation which can be solved in closed form in the range where the carrier flow is by drift. The fixed charge may have a dominant effect on the space charge even when the excess density of chemical impurities is small compared to the density n_i of electrons in intrinsic material. Consider, for example, a junction between an extrinsic *P* region and a weakly doped *n* region having an excess density $N = N_d - N_a$ of donors. In the limit, as the reverse bias is increased and the space charge penetrates many diffusion lengths into the *n* region, the field distribution becomes linear, corresponding to a constant charge density equal to

$$\frac{1}{2}[N + \sqrt{N^2 + 8 n_i^2 \mathfrak{L}^2 / L_i^2}]$$

where L_i is the diffusion length in the weakly doped *n* type region and \mathfrak{L} is the Debye length for intrinsic material. For germanium at room temperature \mathfrak{L}/L_i is the order of 10^{-3} . Thus, in this example, a donor density as low as 10^{11} cm^{-3} will have an appreciable effect on the space charge.

II. BASIC RELATIONS

The problem can be stated in terms of the hole density p , the electron density n , and the electric field E and their derivatives. Let the distance x be measured in the direction from N to P . The field will be taken as positive when a hole tends to drift in the $+x$ direction. The field increases in going in the $+x$ direction when the space charge is positive. Poisson's equation for intrinsic material is

$$\frac{dE}{dx} = a(p - n) \quad (2.1)$$

where the constant a has the dimensions of volt cm and is given in terms of the electronic charge q and the dielectric constant κ by

$$a = \frac{4\pi q}{\kappa}$$

For germanium $a = 1.17 \times 10^{-7}$ volt em.

The hole and electron flow densities J_p and J_n are²

$$\begin{aligned} J_p &= \mu E p - D \frac{dp}{dx} = \mu p \left[E - \frac{kT}{q} \frac{d}{dx} \ln p \right] \\ J_n &= -b \left(\mu E n + D \frac{dn}{dx} \right) = -b \mu n \left[E + \frac{kT}{q} \frac{d}{dx} \ln n \right] \end{aligned} \quad (2.2)$$

where μ and $D = \mu kT/q$ are the hole mobility and diffusion constant respectively, k is Boltzmann's constant (8.63×10^{-5} ev per °C) and T is the absolute temperature. The ratio b of electron mobility to hole mobility we take to be unity. This makes the problem symmetrical in n and p and consequently easier to understand. Section VI will extend the results to the general case of arbitrary b .

Charge and Particle Flow

For some purposes it helps to express the flow not in terms of J_p and J_n but rather in terms of the current density I and the flow density $J = J_p + J_n$ of particles, or carriers. The current density $I = q(J_p - J_n)$. Each carrier, hole or electron, gives a positive contribution to J if it goes in the $+x$ direction and a negative contribution if it goes in the $-x$ direction. In other words, J is the net flow of carriers regardless of their charge sign. The current I is constant throughout the intrinsic

² See, for example, Electrons and Holes in Semiconductors, by W. Shockley. D. Van Nostrand Co., New York, 1950.

region. Particle flow is away from the center of the intrinsic region. Carriers are generated in the intrinsic region and flow out at the two ends, the electrons going out on the *N* side and holes on the *P* side. Thus J is positive near the *IP* junction and negative near the *NI* junction.

From the definitions of I and J and equations (2.2)

$$\begin{aligned}\frac{I}{q} &= \mu E(p + n) - D \frac{d}{dx}(p - n) \\ J &= \mu E(p - n) - D \frac{d}{dx}(p + n)\end{aligned}\quad (2.3)$$

It is convenient to express the equations in terms of E and a dimensionless variable

$$s = \frac{n + p}{2n_i} \quad (2.4)$$

which measures how "swept" the region is. In normal intrinsic material $s = 1$. In a completely swept region $s = 0$; at the junctions with highly extrinsic material $s \gg 1$. Using Poisson's equation to express $p - n$ in terms of E , equations (2.3) become

$$\begin{aligned}I &= \sigma s E - \frac{qD}{a} \frac{d^2 E}{dx^2} \\ J &= \frac{d}{dx} \left[\frac{\mu E^2}{2a} - 2n_i D s \right]\end{aligned}\quad (2.5)$$

where $\sigma = 2\mu n_i q$ is the conductivity of intrinsic material. The particle flow J is thus seen to be the gradient of a flow potential that depends only on E and s .

Equations (2.5) can be written in the form

$$I = \sigma \left[sE - \mathcal{L}^2 \frac{d^2 E}{dx^2} \right] \quad (2.6)$$

$$J = D2n_i \frac{d}{dx} \left[\frac{E^2}{E_1^2} - s \right] \quad (2.7)$$

where $\mathcal{L} = \sqrt{kT/2an_i q}$ is the Debye length in intrinsic material and

$$E_1 = 2 \sqrt{\frac{an_i kT}{q}} = \frac{\sqrt{2}kT}{q\mathcal{L}} \quad (2.8)$$

is a field characteristic of the material and temperature. Specifically E_1 is $\sqrt{2}$ times the field required to give a voltage drop kT/q in a Debye

length. For germanium at room temperature $\mathcal{L} = 6.8 \cdot 10^{-5}$ cm and $E_1 = 383$ volts per cm.

Both I and J are the sum of a drift term and a diffusion term. For charge neutrality, where $p - n$ is small compared to $p + n$, both charge diffusion and particle drift can be neglected. We shall see later that, except right at the junctions, charge diffusion is negligible.

The Equations of Continuity

The two equations of continuity are

$$\frac{dJ_p}{dx} = \frac{dJ_n}{dx} = g - r \quad (2.9)$$

where g is the rate of pair generation and r the rate of recombination. In terms of I and J , these become

$$\frac{dI}{dx} = 0 \quad (2.10)$$

or $I = \text{constant}$ and

$$\frac{dJ}{dx} = 2(g - r) \quad (2.11)$$

which says that the gradient of particle flow is equal to the net rate of particle generation, that is, twice the net rate of pair generation.

To complete the statement of the problem it remains to express g and r in terms of n and p .

Generation and Recombination

The direct generation and recombination of holes and electrons follows the mass action law, in which $g - r$ is proportional to $n_i^2 - np$. The constant of proportionality can be defined in terms of a lifetime τ as follows: Let $\delta p = \delta n \ll n_i$ be a small disturbance in carrier density. Then defining $\tau(g - r) = -\delta n$, we see that the proportionality constant in the mass action law is $(2n_i\tau)^{-1}$. So

$$g - r = \frac{n_i^2 - np}{2n_i\tau} \quad (2.12)$$

and the generation rate

$$g = \frac{n_i}{2\tau} \quad (2.13)$$

is independent of carrier concentration.

In actual semiconducting materials, recombination is not direct. Rather it occurs through a trap, or recombination center. The statistics of indirect recombination has been treated by Shockley and Read³ for a recombination center having an arbitrary energy level ε_t somewhere in the energy gap. At any temperature the trap level can be expressed by the values n_1 and p_1 which n and p would have if, at that temperature, the Fermi level were at the trap level. Shockley and Read showed that, at a given temperature, the lifetime for small disturbances in carrier density is a maximum in intrinsic material. It drops to limiting values τ_{n0} and τ_{p0} in highly extrinsic n and p material, respectively. The formula for $g - r$ in terms of n and p is

$$g - r = \frac{n_i^2 - np}{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)} \quad (2.14)$$

For our purposes it is more convenient to define the lifetime τ not by $\tau(g - r) = -\delta n \ll n_i$, but rather as the proportionality factor in the mass action law. Then τ is not necessarily constant independent of carrier density. From (2.12) and (2.14)

$$\tau = \frac{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)}{2n_i} \quad (2.15)$$

We shall be interested in the lifetime in the region where n and p are equal to or less than n_i . τ decreases as n and p decrease; that is, τ is less in a swept region than in normal intrinsic material. Let $\tau = \tau_i$ for $n = p = n_i$ and $\tau = \tau_0$ for $n = p = 0$. The total range of variation of τ is by a factor of

$$\frac{\tau_i}{\tau_0} = 1 + \frac{n_i(\tau_{p0} + \tau_{n0})}{p_1\tau_{n0} + n_1\tau_{p0}} \quad (2.16)$$

Let the energy levels be measured relative to the intrinsic level, and define a level ε_0 by

$$\varepsilon_0 = kT \ln \sqrt{\frac{\tau_{n0}}{\tau_{p0}}}$$

Then if $\varepsilon_t = \varepsilon_0$, $n_1\tau_{p0} = p_1\tau_{n0}$. Now eq. (2.16) becomes

$$\frac{\tau_i}{\tau_0} = 1 + \frac{1}{2} \left(\sqrt{\frac{\tau_{n0}}{\tau_{p0}}} + \sqrt{\frac{\tau_{p0}}{\tau_{n0}}} \right) \operatorname{sech} \left(\frac{\varepsilon_t - \varepsilon_0}{kT} \right) \quad (2.17)$$

Thus the variation in τ increases as the ratio of τ_{n0} to τ_{p0} deviates from unity and as the trap level moves away from the level ε_0 .

³ W. Shockley and W. T. Read, Jr., Phys. Rev., **87**, p. 835, 1952.

The data of Burton, Hull, Morin, and Severien⁴ shows that a typical value of the ratio of τ_{p0} and τ_{n0} is about 10. This means that the variation in τ with carrier concentration will be less than 10 per cent provided ε_t is about $4kT$ from ε_0 . In what follows we shall assume that this is so. Then we have the mass action law (2.12) with τ a constant, which could be measured by one of the standard techniques involving small disturbances in carrier density. The general case of variable τ is considered briefly at the end of Section IV.

Outline of the Solution

To conclude this section, we discuss briefly the form of the equations and the solution in different parts of the intrinsic region. First consider (2.6) for the current in the ideal case of equal mobilities. In Sections III and V we shall show that throughout almost all of the intrinsic region the current flows mainly by pure drift so we can take $I = \sigma sE$. The reason for this is as follows. The quantity \mathfrak{L}^2 is so small that the diffusion term remains negligible unless the second derivative of E becomes large — so large in fact that the E versus x curve bends sharply upward and both the drift and diffusion terms become large compared to the current I . This is the situation at the junction where I is the small difference between large drift and diffusion terms. Thus (2.6) has two limiting forms:

(1) Except at the junctions the current is almost pure drift so $I = \sigma sE$ is a good approximation. In Section III we derive an upper limit for the error introduced by this approximation and show how the approximate solution can be corrected to take account of the diffusion term.

(2) At the junction, the drift term becomes important and the current rapidly becomes a small difference between its drift and diffusion terms and the solution approaches the zero current solution, for which $sE = \mathfrak{L}^2 d^2 E / dx^2$. In Section V we derive an approximate solution that joins onto the $I = \sigma sE$ solution near the junction and then turns continuously and rapidly into the zero current solution. We shall call this the *junction* solution.

The abrupt change in the solution from (1) to (2) near the junction is shown to be related to a basic instability in the differential equation. This makes it impractical to solve the equations on a machine.

When the applied bias is large compared to the built-in voltage drop, the junction region will be of relatively little interest so the $I = \sigma sE$ solution can be used throughout.

In the region where $I = \sigma sE$ there are two overlapping regions in which the equations assume a simple form. These are the following:

⁴ Burton, Hull, Morin and Severiens, *J. Phys. Chem.*, **57**, p. 853, 1953.

The No-Recombination Solution

Here recombination is small compared to generation, $r \ll g$. This will be so in at least part of the intrinsic region for reverse biases of more than a few kT/q . The E versus x curve turns out to be given by a simple, cubie algebraic equation.

The Recombination, or Charge Neutrality, Solution

Here $n - p$ is small compared to $n + p$, so the particle flow is by diffusion. We shall find that the s versus x curve is given by a third degree elliptic integral. As we move away from the center of the intrinsic region and toward the junctions, recombination becomes small compared to generation and the recombination solution goes into the no-recombination solution. In the region where both solutions hold, the solution has the simple form $s = I/\sigma E = A - x^2$ where A is a constant that must be less than $\frac{2}{3}$ and the unit of length is twice the diffusion length.

As the bias on an NIP structure is increased and the space charge penetrates through the intrinsic region, the region where the recombination is important will shrink and eventually disappear.

Fig. 1 is a schematic plot of the field distribution for the case where the applied bias is large compared to the built-in potential drop but not large enough to sweep all the carriers out of the intrinsic region. As the voltage is increased, the drop in field in the intrinsic region will become less and finally the field distribution will be almost flat from junction to junction. Only half of the intrinsic region is shown in Fig. 1. For equal mobilities the field distribution will be symmetrical about the center x_i of the intrinsic region.

The illustration shows the recombination solution (1), which holds near the center of the intrinsic region and overlaps (2), the no-recombination solution. The junction solution (3) joins continuously onto the no-recombination solution at the point x_0 and rapidly breaks away and approaches the zero-current solution at the junction. The figure is schematic and has not been drawn to scale. In most cases of interest, the low fields in the recombination region will be much lower and the junction solution will hold over a smaller fraction of the intrinsic region.

It is convenient to take $x = 0$ not at the center x_i of the intrinsic region but at the minimum on the no-recombination solution. As the applied bias increases, x_i approaches zero.

Unequal Mobilities

In the general case of unequal mobilities, it is no longer so that I is pure drift except at the junctions. However we can define a linear com-

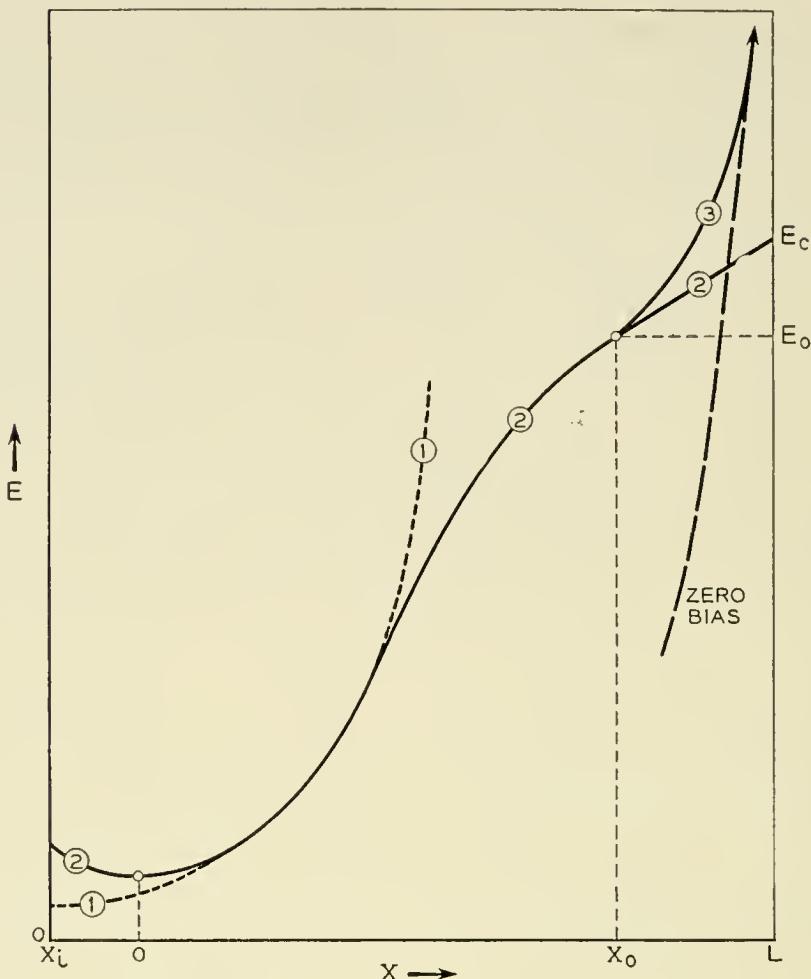


Fig. 1 — Schematic of the field distribution and the three overlapping solutions.

bination of J_p and J_n which has the same form as I in (2.6) and in which the diffusion term is negligible except near the junction. As we show in section VI, the effect of unequal mobilities is (1) to introduce some asymmetry into the curve in the region where the curvature is upward and (2) to displace the curve toward the NI junction (for the case where the electrons have the higher mobility). Thus the field is higher on the side where the carrier mobility is lower, as would be expected.

III. THE NO-RECOMBINATION CASE

This section deals with the case where recombination can be neglected in comparison with generation. This will be so where np is small compared to n_i^2 .

The continuity equation for J now becomes

$$\frac{dJ}{dx} = 2g = \frac{n_i}{\tau} \quad (3.1)$$

Combining this with (2.7) gives

$$\frac{d^2}{dx^2} \left(\frac{E^2}{E_1^2} - s \right) = \frac{1}{2L_i^2} \quad (3.2)$$

where $L_i^2 = D\tau$ is the diffusion length in intrinsic material.

Equation (3.2) can be immediately integrated. There are two constants of integration, one of which can be made to vanish by choosing $x = 0$ at the center of the intrinsic region, where the first derivatives of E and s vanish. (E is a minimum here and s a maximum). The solution obtained by two integrations is

$$\left(\frac{E}{E_1} \right)^2 - s = \left(\frac{x}{2L_i} \right)^2 - A \quad (3.3)$$

As we shall see later, the constant A is determined by the voltage drop across the unit.

The exact procedure now would be to substitute s from (3.3) into (2.6). The resulting second order differential equation could, in principle, then be solved for E versus x . The exact solution, however, would be quite difficult. We shall discuss it in Section V. Here we make the assumption that throughout the intrinsic region the charge flow is mainly by drift, so that we can neglect the diffusion term in (2.6) and take $I = \sigma s E$, as discussed in Section II. Later in this section we find an upper limit on the error due to this assumption and show how the cubic can be corrected to take account of the diffusion term.

Putting $s = I/\sigma E$ in (3.3) gives a cubic equation

$$\begin{aligned} \left(\frac{E}{E_1} \right)^2 - \frac{I}{\sigma E} &= \left(\frac{x}{2L_i} \right)^2 - A \\ \left(\frac{E}{E_1} \right)^2 - \frac{I}{\sigma E_1} \left(\frac{E_1}{E} \right) &= \left(\frac{x}{2L_i} \right)^2 - A \end{aligned} \quad (3.4)$$

for E/E_1 as a function of $x/2L_i$. This equation contains two parameters I and A . A determines the voltage and I is determined by the length $2L$ of the intrinsic region. The relation is as follows: Let the applied voltage drop across each junction be at least a few kT/q . Then the minority carrier currents from the extrinsic regions will have reached their saturation values. Call I_s the sum of the hole current from the N region and the electron current from the P region. I_s comes from pairs generated in the extrinsic regions near the junctions. I_s can be made arbitrarily small by making the N and P regions sufficiently highly doped (provided the diffusion length in the extrinsic material does not decrease with doping faster than the majority carrier concentration in-

creases). The current generated in the intrinsic region is qg per unit volume. So the density of current from pairs generated in the intrinsic layer is $2Lqg = qn_i L/\tau$. Hence

$$I = I_s + \frac{qn_i L}{\tau}$$

In what follows we shall assume that I_s is negligibly small compared to I . Then

$$I = \left(\frac{qn_i}{\tau} \right) L = \left(\frac{qn_i D}{L_i} \right) \frac{L}{L_i} = \left(\frac{\sigma}{2L_i} \frac{kT}{q} \right) \frac{L}{L_i}$$

Thus I is L/L_i times a characteristic current equal to (1) the diffusion current produced by a gradient n_i/L_i or (2) the drift current produced by a field that gives the voltage drop kT/q in two diffusion lengths in normal intrinsic material. In germanium this characteristic current is about 5 milliamperes per cm^2 .

That the current I is proportional to L and independent of voltage follows from the neglect of recombination. When recombination is small compared to generation, then the current has reached its maximum, or saturation, value. All the carriers generated in the intrinsic region are swept out before recombining. It will sometimes be convenient to take σE_1 as the unit of current. From the above and (2.8)

$$\frac{I}{\sigma E_1} = \frac{\sqrt{2} \mathcal{E} L}{(2L_i)^2} \quad (3.5)$$

In germanium σE_1 is about 7 amperes per cm^2 . In general we will be dealing with currents that are small compared to this. For example, if L_i is 1 mm, we would have to sweep out an intrinsic region 3 meters long in order to get a current this large. If we take E_1 as the unit field, σE_1 as the unit current and $2L_i$ as the unit length then the cubic becomes $E^2 - I/E = x^2 - A$.

For a given structure and temperature the field versus x curves form a one parameter family. A determines both the field distribution and the voltage. The voltage increases as A decreases. Fig. 2 is a plot of E/E_1 versus $x/2L_i$ for $L/2L_i = 0.1$ and several different values of A . Fig. 3 is for $L = 2L_i$ and Fig. 4 for $L/2L_i = 3$.

There is an upper limit on A but not lower limit. The reason is as follows: As A increases, the minimum value of E (at $x = 0$) decreases and the maximum value of s increases. So if A is too large, the maximum s will be so large that we cannot neglect recombination, which becomes important when np approaches n_i^2 , or s approaches 1. Fre-

quently recombination can be neglected over parts of the intrinsic region but not near the center, where the field is a minimum and the carrier concentration a maximum. Then (3.4) will represent the field distribution over that part of the region where recombination is unimportant. The correct solution will join onto the cubic as we move away from the center of the intrinsic region, which will no longer be at the $x = 0$ point on the cubic. In Section IV we solve the equations for the recombination region and show how the solution approaches the cubic. We will show that, as A increases, the distance from the center of the intrinsic region to the $x = 0$ point on the cubic also increases. The value $A = \frac{2}{3}$ corresponds to an infinitely long intrinsic region. For a larger A there exists no exact solution that could join onto the cubic as recombination becomes negligible. In Figs. 3 and 4 the $A = \frac{2}{3}$ curves join onto recombination solutions at values of E which are too low to show.

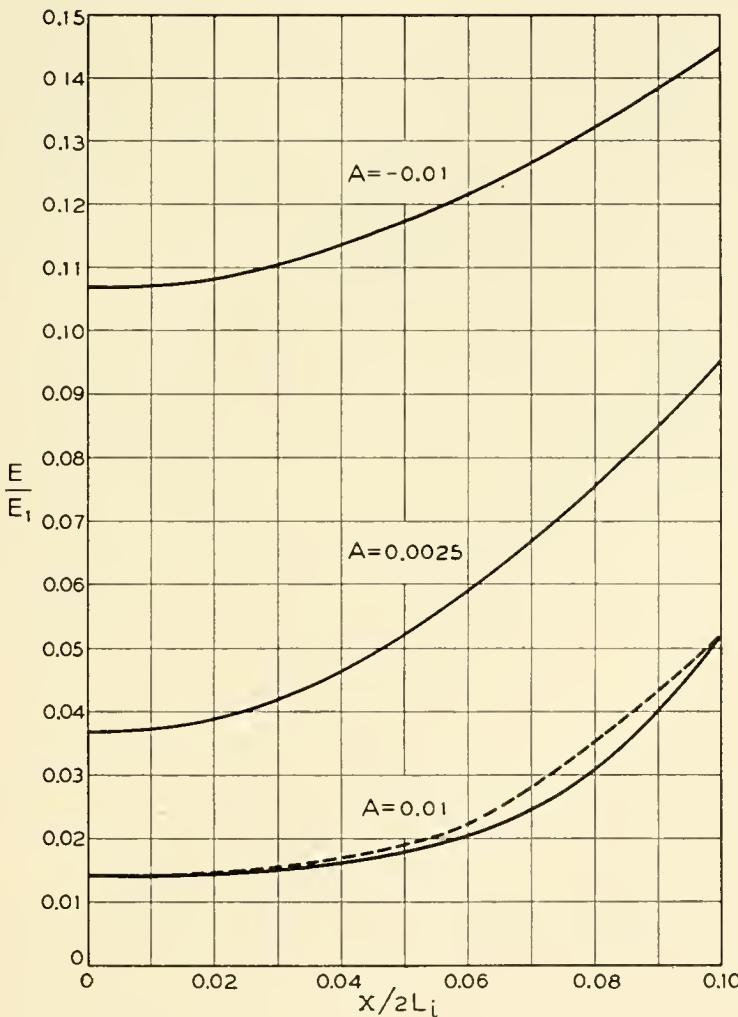


Fig. 2 — Field Distributions for $L = 0.2L_i$.

As A decreases and becomes negative the cubic approaches the form

$$E^2 = E_0^2 + E_1^2 \left(\frac{x}{2L_i} \right)^2 \quad (3.6)$$

where $E_0^2 = -AE_1^2$ is the minimum value of E^2 . This form of the solution will be valid when the minimum E is large compared to (IE_1^2/σ) . As E_0 increases, the voltage increases and the curve becomes flatter. This is because the increasing field sweeps the carriers out and reduces the space charge; so the drop in field decreases.

If (3.4) for E/E_1 versus $x/2L_i$ is extended to indefinitely large values of $x/2L_i$, it approaches the straight line of slope 1 going through the origin. Since E is always positive the curve is above this straight line at $x = 0$. If A is negative the curve is always above the straight line and always concave upward. If A is positive, the curve crosses the straight

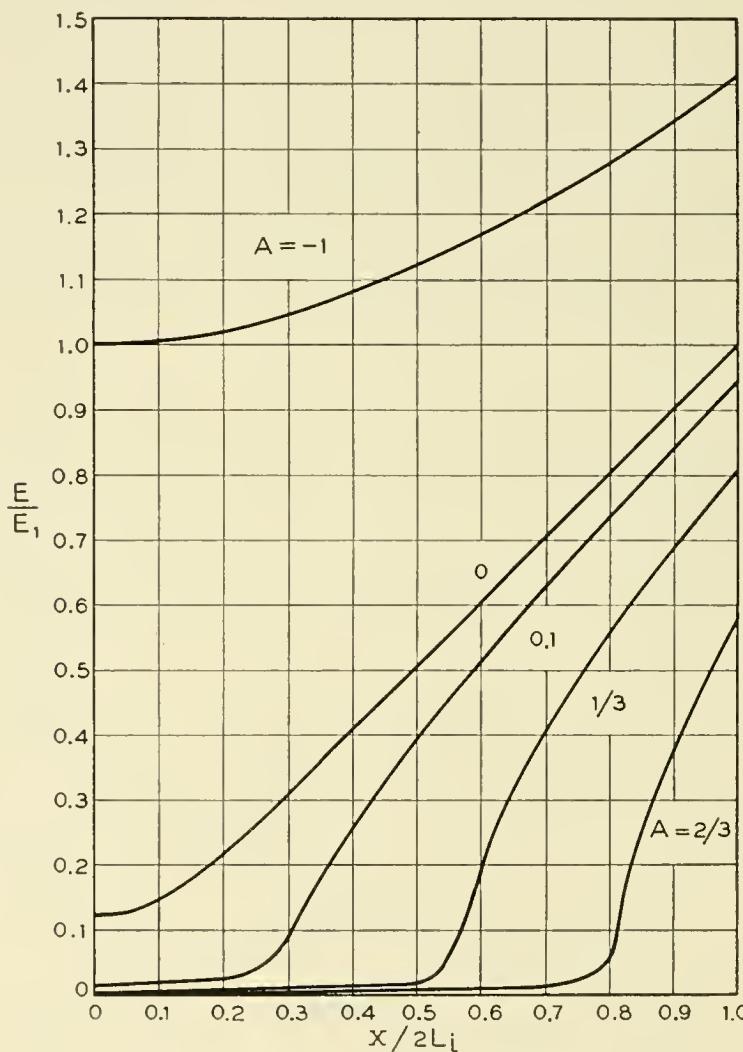


Fig. 3 — Field Distributions for $L = 2L_i$.

line at $E/E_1 = I/\sigma E_1 A$ and thereafter remains under it approaching it from below. For positive A the curvature, which is upward near the origin, changes to downward at about $x/2L_i = \sqrt{A}$.

The carrier concentrations n and p can be found from the E versus x curves with the aid of Poisson's equation $p - n = 1/a dE/dx$ and the definition $s = (n + p)/2n_i$ with $s = I/\sigma E$. These relations and (3.4) give

$$\frac{p - n}{p + n} = \frac{x}{L} \frac{1}{\left(1 + \frac{IE_1^2}{2\sigma E^3}\right)} \quad (3.7)$$

From (3.4) and (3.7) we may distinguish the following two regions on the cubic:

(1) When E^3/E_1^3 is smaller than $I/\sigma E_1$ (which as we have seen is usually smaller than unity), the E versus x curve is concave upward, the hole and electron concentrations are almost equal (charge neutrality) and the particle flow is by diffusion.

(2) When E^3/E_1^3 is greater than $I/\sigma E_1$, in general there is space charge and the particle flow, like the charge flow, is by drift. The curve is concave downward for positive A .

Figure 6, which we will discuss in Section IV, shows the field and carrier distributions for $L = 2L_i$ and $A = 0.665$ plotted on a logarithmic

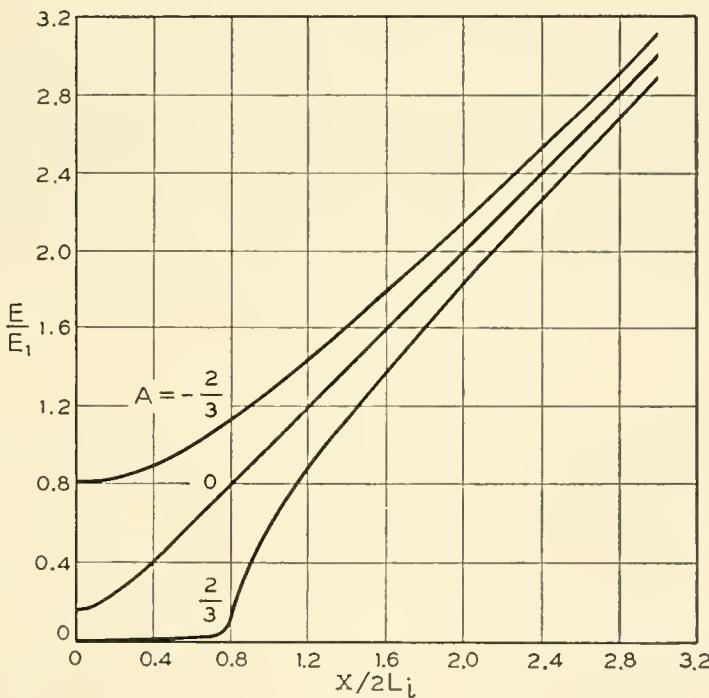


Fig. 4 — Field Distributions for $L = 6L_i$.

scale to show the behavior at low values of field and carrier density. In the region of no-recombination the field distribution is indistinguishable from that for $A = \frac{2}{3}$, which is plotted in Fig. 3 on a linear scale. In the region where recombination is important the solution is found from the assumption of charge neutrality as will be discussed in Section IV. The cubic and charge neutrality solutions are each shown dashed outside of their respective ranges of validity. For $A = 0.665$ the half length of the intrinsic region is $2.098 \times 2L_i$. Thus the length of the intrinsic region is more than twice the effective length $2L$ in which current is generated. The effective length will be discussed in more detail in Section IV and it will be shown that the effective length $2L$ of current generation is equal to the twice the distance from the $1P$ junction to the minimum on the cubic. As explained earlier, it is convenient to take $x = 0$ at the minimum on the cubic.

Intrinsic-Extrinsic Junction Under Large Bias

Consider the limiting case of an intrinsic-extrinsic junction as the bias is increased and the space charge penetrates many diffusion lengths into the intrinsic material. Then the field distribution approaches the straight line $E/E_1 = x/2L_i$. This, by Poisson's equation, means that there is a constant charge density of N_i where

$$N_i = \frac{E_1}{2aL_i} = \frac{\sqrt{2}\mathcal{L}}{L_i} n_i$$

Thus in the limit, the field in the intrinsic region approaches that in a completely swept extrinsic region having a fixed charge density of N_i . In germanium at room temperature N_i is about $4.10^{10} \text{ cm}^{-3}$. As the field approaches the limiting form, the voltage V approaches $E_1 L^2 / 4L_i$. Thus the limiting form of the current voltage curve is

$$\frac{I}{\sigma E_1} = \frac{\mathcal{L}}{L_i} \sqrt{\frac{V}{2E_1 L}}$$

So in the limit the current varies as the square root of the voltage. Typical values for germanium at room temperature are $\sigma E_1 = 7 \text{ amps cm}^{-2}$, $\mathcal{L}/L_i = 10^{-3}$ and $2E_1 L_i = 50 \text{ volts}$.

Equivalent Generation Length for an Intrinsic-Extrinsic Junction

It should be noted that for an IP structure the current is the same as for an NIP structure with an infinite I region, or at least an I region that is long compared to the distance of penetration of the space charge.

Thus the equivalent length of current generation is $2L$ even though the current is actually being generated in an effective length L . The reason is that for an *NIP* structure the holes entering the right hand half of the *I* region were generated in the left hand side. For an *IP* structure the holes entering the space charge regions from the left were injected at the external left hand contact to the *I* region.

Applied Voltage

In all cases the voltage can be found from the area under the E versus x curve. In Figs. 2 to 4 the area under the curves gives the voltage accurately; recombination becomes important only where the field is so low as to have a negligible effect on the total voltage drop.

Correction of the Cubic

To conclude this section we consider the error introduced by using the assumption $I = \sigma sE$. For simplicity take E_1 as the unit field, $2L_i$ as the unit length and σE_1 as the unit current. Then the cubic becomes $E^2 - I/E = x^2 - A$. The corresponding exact solution is $E^2 - s = x^2 - A$ where the relation between s and E is given by equation (2.6) which in dimensionless form is

$$\mathcal{L}^2 \frac{d^2 E}{dx^2} = sE - I \quad (3.8)$$

where \mathcal{L}^2 is of the order of 10^{-6} .

Let δE and δs represent the difference between the cubic and the correct solution at a given x . Assume that δE and its second derivative are small compared to E and its second derivative respectively. Then $\delta s = 2E\delta E$ and on the correct solution $sE - I = (2E^2 + I/E)\delta E$. So (3.8) becomes

$$\frac{\delta E}{E} = \left(\frac{\mathcal{L}^2}{2E^3 + I} \right) \frac{d^2 E}{dx^2} \quad (3.9)$$

To obtain a first approximation to $\delta E/E$ we use the cubic to evaluate $d^2 E/dx^2$. It is convenient to express the results in terms of a dimensionless variable $z = E/I^{1/3}$, or if E and I are measured in conventional units $z = E(\sigma/E_1^2 I)^{1/3}$. Then (3.9) becomes

$$\frac{\delta E}{E} = \frac{1}{2} \left(\frac{L_i \mathcal{L}}{L^2} \right)^{2/3} \left(\frac{z}{z^3 + \frac{1}{2}} \right)^2 + \left(\frac{x}{2L} \right)^2 \frac{z^3(1-z^3)}{(z^3 + \frac{1}{2})^4} \quad (3.10)$$

when the lengths are in conventional units.

The first term has a maximum value of $0.35 (L_i \mathcal{L}/L^2)^{2/3}$ at $z = 0.6$ and the second term a maximum value of 0.18 at $z = 0.5$ and $x = L$.

The dashed curve in Fig. 2 for $A = .01$ is the corrected cubic. For the other curves in Fig. 2, the correction is smaller. For the curves in Figs. 3 and 4 the correction is too small to show.

Limits on the Solution

We now show that δE as derived above is not only a first approximation but also upper limit on the correction necessary to take account of charge diffusion. That is, an exact solution to (3.8) lies between the cubic and the corrected cubic.

Consider the region where the second derivative of E is positive so that the perturbed curve lies above the cubic as in Fig. 2. On the cubic we have $sE - I = 0$. As we move upward from the cubic and toward the dashed curve, $sE - I$ increases. The value of $sE - I$ on the dashed curve just equals the value of $\mathcal{L}^2 d^2 E / dx^2$ on the cubic. However, the dashed curve has a smaller second derivative than the cubic. Thus in moving upward from the cubic toward the dashed curve $sE - I$ increases from zero and $\mathcal{L}^2 d^2 E / dx^2$, which is positive, decreases; on the dashed curve $sE - I$ is actually greater. Therefore there is a curve lying just under the dashed curve where the two sides of (3.8) are equal. The same argument applied to the region where the second derivative is negative shows that the equation is satisfied by a curve lying just above the first perturbation of the cubic. Where the curvature changes sign, the cubic is correct.

It should be emphasized again that the neglect of the diffusion term in the current is justified only for the ideal case of equal hole and electron mobilities. For unequal mobilities both drift and diffusion will be important in current flow. However, as we will discuss in section 5, we can simplify the problem of unequal mobilities by defining a fictitious current that has the same form as I in (2.6) and (3.8).

IV. RECOMBINATION

As discussed in Section III, when the voltage for a given current is reduced, s increases and near $x = 0$ becomes comparable to unity. Then recombination becomes important and the cubic solution breaks down, or rather joins onto a solution that takes account of recombination. When recombination is important the center x_i of the intrinsic region is no longer at the $x = 0$ point on the cubic but to the left of it. That is, if we want the same current with continually decreasing voltage, we even-

tually get to the point where a longer intrinsic region is required. Finally for a given current we reach a minimum voltage which corresponds to an infinite length of intrinsic region. Another way of saying this is that, when recombination becomes important, the length L defined in terms of the current by $I = qg2L = qn_i/\tau L$ is no longer the half length of the intrinsic region.

Equivalent Generation Length

We shall continue to define L by $I = qn_i/\tau L$. Thus L is an equivalent, or effective, half length of current generation and not the half length of the intrinsic region. By definition L is the length such that the amount of generation alone in the length L is equal to the net amount of generation (generation minus recombination) in the total half length of the intrinsic region. Hence

$$gL = \int_{x_i}^{x_p} (g - r) dx \quad (4.1)$$

where x_i is at the center of the intrinsic region and x_p at the IP junction. We shall for the most part deal with reverse biases of at least a few kT/q , in which case recombination is negligible at the junctions. Then the exact solution becomes the no-recombination solution before reaching the junctions. We shall continue to take $x = 0$ at the point $dE/dx = ds/dx = 0$ on the no-recombination solution which the exact solution approaches as recombination becomes negligible.

Simplifying Assumptions

The general differential equation with recombination will be the same as for no-recombination except that $g - r$ replaces g . From (3.1) and (3.2)

$$\frac{d^2}{dx^2} \left(\frac{E^2}{E_1^2} - s \right) = \frac{1}{2L_i^2} \left(1 - \frac{r}{g} \right) \quad (4.2)$$

From (2.12) and (2.13) and Poisson's equation

$$\frac{r}{g} = \frac{np}{n_i^2} = \left(\frac{n + p}{2n_i} \right)^2 - \frac{(n - p)^2}{(2n_i)} = s^2 - 2 \left(\frac{\mathcal{L}}{E_1} \frac{dE}{dx} \right)^2 \quad (4.3)$$

The following analysis will be based on the assumption of charge neutrality. That is we neglect terms in $p - n$ in comparison with those in $p + n$. In particular this means:

- (1) The charge flows by drift so $I = \sigma s E$. This is the same assumption

made in the no-recombination case. It will be an even better approximation in the recombination region, where the second derivative of E is less.

(2) The particle flow is by diffusion. That is, E^2/E_1^2 can be neglected in comparison with s .

(3) The ratio of recombination rate r to generation rate g is proportional to $g - r$; that is $g - r = g(1 - s^2)$.

All of these simplifying assumptions can be justified by substituting the resulting solution into the original expressions and showing that the neglected terms are small when recombination is important. If the solution is substituted into (4.3) and (2.6) the neglected terms will turn out to be negligible—and therefore assumptions (1) and (3), justified—when s^2 is large compared to \mathfrak{L}/L_i . Assumption (2) follows from (1) and the fact that $I/\sigma E_1$ is small compared to unity.

Assumptions (2) and (3) may also be justified by the discussion following (3.7) in the following way: Where recombination is important s must be near unity. So the cubic will begin to break down when $s = I/\sigma E$ becomes near to unity, or when E approaches I/σ . However, if E is approximately I/σ then $\sigma E^3/I E_1^2$ is approximately $(I/\sigma E_1)^2$, which, as we saw in the Section III, is small compared to unity in practical cases. Thus recombination becomes important and the solution joins onto the cubic in the range where E^3/E_1^3 is small compared to $I/\sigma E_1$. In this range the particle flow is by diffusion and $p - n$ is small compared to $p + n$. As we move toward the center of the intrinsic region s increases and E and dE/dx decrease. Therefore, since assumptions (2) and (3) are good where the solution joins onto the cubic, they are good throughout the region where recombination is important.

The Recombination Solution

The differential equation (4.2) now takes the form

$$\frac{d^2s}{dx^2} = -\frac{(1 - s^2)}{2L_i^2} \quad (4.4)$$

The solution for s in the recombination range is seen to be the same for all values of the current. When s has been found E is found from $E = I/\sigma s$.

For small disturbances in normal carrier concentration, s is only slightly different from unity and (4.4) takes the familiar form

$$\frac{d^2}{dx^2}(1 - s) = \frac{1 - s}{L_i^2}$$

which says that the disturbance in carrier concentration varies exponentially as x/L_i .

Equation (4.4) can be integrated once to give

$$\left(\frac{ds}{dx}\right)^2 = \frac{1}{L_i^2} \left(s_0 - s - \frac{s_0^3 - s^3}{3} \right) \quad (4.5)$$

where s_0 is the value of s at the center of the intrinsic region where s is a maximum.

As recombination becomes unimportant, s^2 becomes small compared to unity and (4.5) approaches the form

$$\left(\frac{ds}{dx}\right)^2 = \frac{1}{L_i^2} \left[s_0 \left(1 - \frac{s_0^2}{3} \right) - s \right] \quad (4.6)$$

and the solution joins onto the no-recombination solution.

Joining onto the Cubic.

We have seen that the solution joins onto the no recombination solution, in the region where particle flow is by diffusion so that the no recombination solution has the form $s = A - (x/2L_i)^2$. This is readily transformed to the form (4.6) with

$$A = s_0 \left(1 - \frac{s_0^2}{3} \right) \quad (4.7)$$

Thus the one arbitrary parameter s_0 in the recombination solution determines the parameter A in the cubic that the recombination solution approaches. Since the maximum value of s_0 under reverse bias is unity, the maximum value of A is $\frac{2}{3}$. Negative values of A correspond to solutions where recombination is always negligible.

The s versus x Curve

To find s versus x we integrate (4.5). There is one constant of integration, which is fixed by the choice of $x = 0$. We have taken $x = 0$ at the point where $dE/dx = ds/dx = 0$ on the cubic. To make the recombination solution join the cubic we choose the constant of integration so that the recombination solution extrapolates to $s = 0$ at $(x/2L_i)^2 = A$. Then

$$\frac{x}{2L_i} = \sqrt{A} - \frac{\sqrt{3}}{2} \int_0^s \frac{ds}{\sqrt{3(s_0 - s) - (s_0^3 - s^3)}} \quad (4.8)$$

which can be expressed in terms of elliptic integrals.

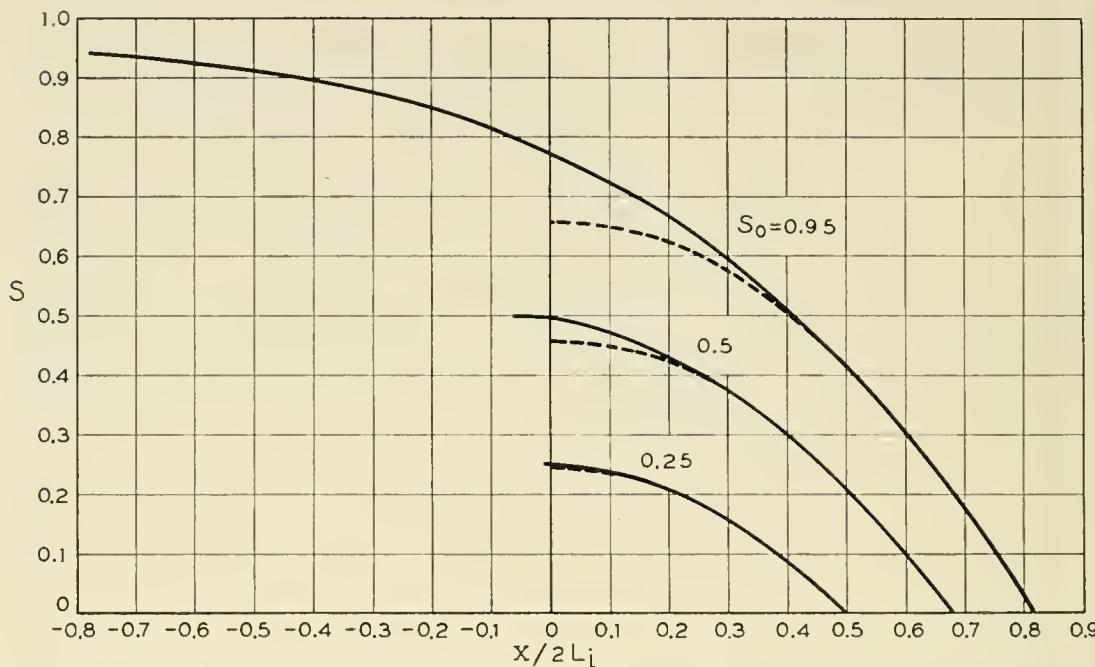


Fig. 5 — Variation of $s = p/n_i = n/n_i$ in the range where recombination is important.

Deep in an infinitely long intrinsic region the carrier densities approach their normal values $n = p = n_i$, or $s = 1$. Putting $s_0 = 1$ in (4.8), we find that as s approaches $s_0 = 1$, x becomes infinite. This will be the solution for a simple intrinsic-extrinsic junction. Fig. 5 is a plot of s versus x for various values of s_0 . The dashed curves represent the corresponding no-recombination solution $s = A - (x/2L_i)^2$.

The IP Junction

It remains to find the position of the *IP* boundary. We now show that if recombination is unimportant at the junction, so that the solution joins onto a no-recombination solution, then the position of the junction is at $x = L$ where L is the effective length of current generation and $x = 0$ is the point where $dE/dx = ds/dx = 0$ on the no-recombination solution (which of course will not be valid at $x = 0$). The proof is as follows: From the definition (4.1) of L and (4.2)

$$\begin{aligned} L &= \int_{x_i}^{x_p} (1 - r/g) dx = 2L_i^2 \int_{x_i}^{x_p} \frac{d}{dx} \left(\frac{E^2}{E_1^2} - s \right) dx \\ &= 2L_i^2 \left[\frac{d}{dx} \left(\frac{E^2}{E_1^2} - s \right) \right]_{x=x_p} \end{aligned} \quad (4.9)$$

If the boundary comes where recombination is negligible so that $(E/E_1)^2 - s = (x/2L_i)^2 - A$, then (4.9) gives $x_p = L$. Physically

this means that the amount of recombination in the interval from $x = 0$ to $x = L$ is just equal to the excess amount of generation in the interval from the center of the intrinsic region to $x = 0$.

If the applied reverse bias is less than a few kT/q then recombination is important even at the junction and there is no joining onto a no-recombination solution. In this case (4.9) says that for a given choice of current (and hence of L) the boundary comes where

$$\frac{ds}{dx} = - \frac{L}{2L_i^2} \quad (4.10)$$

Example. Fig. 6, which we discussed briefly in Section III, is a plot of the field and carrier distributions for $L = 2L_i$ and $s_0 = 0.95$, for which $A = 0.665$. The hole and electron densities were found from (3.7) and $p + n = 2n_is$ where s is found from Fig. 5. When s approaches s_0 (4.8) for x versus s takes the simple form

$$\frac{x - x_i}{2L_i} = \frac{s_0 - s}{1 - s_0^2} \quad (4.11)$$

This will be accurate when $s_0 - s$ is small compared to $1/s_0 - s_0$. We have used (4.11) to evaluate the s versus x curve beyond the range of the $s_0 = 0.95$ curve in Fig. 5.

It is seen that the recombination solution in Fig. 6 joins the cubic in the range where n and p are still almost equal.

Variable Lifetime.

Finally consider the general case where the variation in τ with carrier density cannot be neglected. Then, with $n = p = n_is$, (2.15) becomes $\tau = \tau_0 + (\tau_i - \tau_0)s$ and L_i^2 in (4.4) is replaced by $D\tau[1 + (\tau_i/\tau_0 - 1)s]$ where τ_i/τ_0 is given by (2.17). The more general form of (4.4) can be solved graphically after one integration. The solution will join onto a cubic if $(\tau_i/\tau_0 - 1)s$ becomes small compared to unity before space charge becomes important. This will be so if $(\tau_i/\tau_0 - 1)^{3/2}I/\sigma E_1$ is small compared to unity.

V. THE JUNCTION SOLUTION

In this section we consider the solution near the junctions, where the assumption $I = \sigma s E$ breaks down. We shall deal with reverse biases of at least a few kT/q so that recombination is negligible at the junctions. The junction solution will therefore join onto the no-recombination solution. We shall use the cubic solution in the no recombination region.

Again it is convenient to use dimensionless variables with E_1 as the

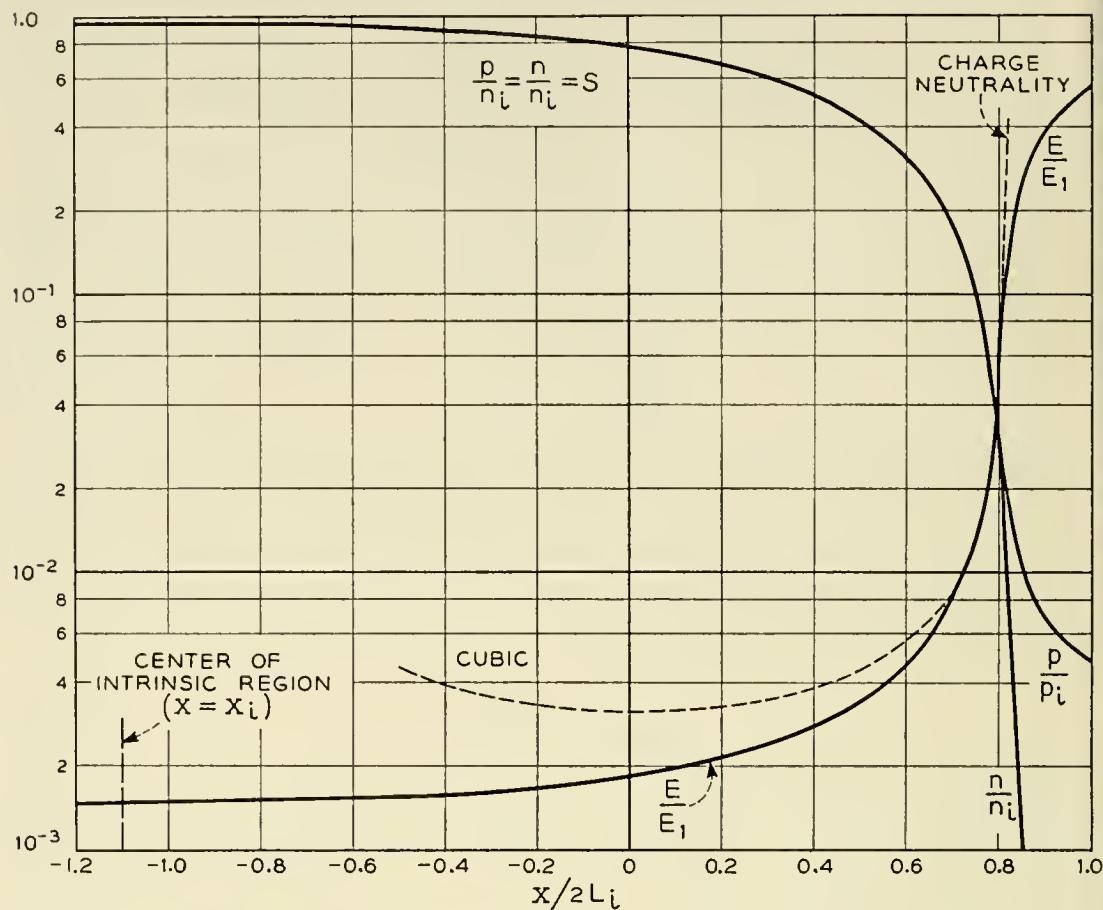


Fig. 6 — Field and carrier distributions for $L = 2L_i$ and $A = 0.665$ ($s_0 = 0.95$).

unit field, $2L_i$ as unit length and σE_1 as unit current. Then on the cubic $s = I/E$, and $E^2 - I/E = x^2 - A$. The current is related to L by $I = \sqrt{2\mathfrak{L}L}$ where the dimensionless \mathfrak{L} is of the order of 10^{-3} for germanium at room temperature. Substituting the exact no-recombination solution $E^2 - s = x^2 - A$ into the solution (2.6), or (3.8), for the current gives the second order differential equation

$$\frac{d^2E}{dx^2} = \frac{1}{\mathfrak{L}^2} \left[E^3 - E(x^2 - A) - I \right] \quad (5.1)$$

for E as a function of x . The two boundary conditions are as follows: At $x = 0$, $dE/dx = 0$ by symmetry. At the IP junction the carrier concentration must rise and approach that in the normal P material. For a strongly extrinsic P region the normal hole concentration P is large compared to both n_i and the electron concentration. Thus s must increase and approach $P/2n_i \gg 1$ as we approach the P region. Clearly the cubic cannot satisfy this requirement. On the cubic the maximum value of s comes at $x = 0$ and is less than unity. As we approach the junc-

tion E increases so $s = I/E$ must decrease. Thus the correct solution must break away from the cubic near the junction.

Instability of the Solution

Equation (5.1) has two limiting forms and makes a rather abrupt transition between them. Over most of the intrinsic region, the quantity in brackets $[Es - I] = [E(E^2 - x^2 + A) - I]$ almost vanishes. It differs from zero just enough that when multiplied by the very large factor $\mathfrak{L}^2 \approx 10^6$ it gives the correct second derivative of E . In Section III we derived an upper limit on the small deviation δE from the cubic required to satisfy the differential equation. If E deviates from the cubic by more than this small amount, then the second derivative of E becomes too large. This increases the deviation from the cubic, which further increases the second derivative and so on. E and s rapidly approach infinity in a short distance. This, of course, is the required behavior at the junction. The rapid increase in s makes it possible for s to approach $P/2n_i$.

In Section III we showed that there is a solution to the differential equation that lies within a small interval δE from the cubic. Suppose we try to solve (5.1) graphically or on a machine starting at $x = 0$. There are two boundary conditions: By symmetry $dE/dx = 0$ at $x = 0$. We choose for $E(0)$ a value somewhere in the interval $\delta E(0)$. The resulting solution will not long remain in the interval $\delta E(x)$. In fact there is only one choice of $E(0)$ for which the solution remains close to the cubic from $x = 0$ to $x = \infty$. For any other $E(0)$ the solution would remain close to the cubic for a certain distance and then abruptly become unstable and both E and s approach infinity. $E(0)$ must be so chosen that the solution becomes unstable and E and s become large at the junction. However it is impractical to set $E(0)$ on a machine with sufficient accuracy to insure that the solution will remain stable for a reasonable distance. A more practical procedure is to find a solution which holds near the junction and joins the cubic to a solution in the adjacent extrinsic region.

Zero Bias

It may be helpful to approach the junction solution by reviewing the simple case of an IP junction under zero bias. Both charge and particle flow vanish. The vanishing of particle flow means that in the intrinsic region $E^2 - s$ is constant, (2.7). Since $E = 0$ and $s = 1$ in the normal intrinsic material, it follows that $E^2 - s = 1$. With $I = 0$ the equation

for current becomes

$$\frac{d^2E}{dx^2} = \frac{sE}{\mathcal{L}^2} = \frac{E^3 + E}{\mathcal{L}^2} \quad (5.2)$$

This can be integrated at once. The boundary conditions are $dE/dx = 0$ when $E = 0$ and $E = E_j$ at $x = L$; the field E_j at the *IP* junction will be determined by joining the solutions for the *I* and *P* regions. The solution can best be expressed by parametric equations giving x and the potential V as functions of E .

$$L - x = \mathcal{L} \int_E^{E_j} \frac{dE}{E \sqrt{1 + E^2/2}} = \mathcal{L} \left[\operatorname{esch}^{-1} \frac{E}{\sqrt{2}} - \operatorname{esch}^{-1} \frac{E_j}{\sqrt{2}} \right] \quad (5.3)$$

$$V_j - V = \mathcal{L} \int_0^E \frac{dE}{\sqrt{1 + E^2/2}} = \frac{2kT}{q} \left[\sinh^{-1} \frac{E_j}{\sqrt{2}} - \sinh^{-1} \frac{E}{\sqrt{2}} \right] \quad (5.4)$$

where we have used the relation between dimensionless quantities $\mathcal{L} = \sqrt{2kT/q}$, which follows from (2.8) with $E_1 = 1$. It will be more convenient to express voltages in terms of kT/q rather than in terms of the unit voltage $2E_1L_i$; then the ratio qV/kT is independent of the units. For convenience we take the voltage as increasing in going toward the *IP* junction with $V = 0$ in the normal *P* material. The voltage V_j at the junction is found by joining solutions.

On the *P* side, let the excess acceptor density be *P*. Adding the term $-aP$ to the right hand side of Poisson's (2.1), and proceeding as before we have, instead of (2.5)

$$\frac{d}{dx} \left(\frac{E^2}{E_1^2} - s - s_p \frac{qV}{kT} \right) = J = 0$$

where $s_p = P/2n_i$. We shall assume that the *P* region is strongly extrinsic so that $n \ll p$. Then $s = s_p$ in the normal *p* material, where $E = V = 0$. Hence

$$E^2 - s = s_p \left(\frac{qV}{kT} - 1 \right) \quad (5.5)$$

In the intrinsic material the corresponding solution is $E^2 - s = -1$. Since both E and s are continuous at the junction, $qV_j/kT = 1 - 1/s_p$ where $1/s_p$ can be neglected. Thus $E_j^2 = s_j = s_p \exp [-(qV_j/kT)] = s_p/e$ where $e = 2.72$ is the base of the natural logarithms.

Knowing E_j we can find the field and potential distributions in the intrinsic material from (5.3) and (5.4).

Reverse Bias

Now in the intrinsic region, $E^2 - s = x^2 - A$. Let E_c be the value of E at the junction as given by the cubic, and let $s_c = I/E_c$ be the corresponding value of s . Then at the junction $x^2 - A = E_c^2 - s_c$. In the P material equation (5.5) will still be a good approximation near the junction, where the additional terms arising from the flow will be negligible. Joining the solutions for the I and P regions and neglecting s_c in comparison with s_p gives

$$\frac{qV_j}{kT} = 1 + \frac{E_c^2}{s_p}$$

Again using $s_j = s_p \exp [-(qV_j/kT)]$ we have

$$E_j^2 = E_c^2 + s_p \exp [-(1 + E_c^2/s_p)] \quad (5.6)$$

In most practical cases E_c^2 will be small compared to $s_p = P/2n_i$ so E_j will be the same as for zero bias.

Junction Solution

We now consider an approximate solution that joins smoothly onto the cubic and has the required behavior at the junction. Let $x = x_0$ be the point where this solution is to join the cubic. Then in (5.1) x^2 must lie between x_0^2 and L^2 . We can obtain two limiting forms of the solution by giving x the two constant values, x_0 and L respectively. It will be best to take $x = x_0$ since in practical cases the x^2 term is not important except near the point where the junction solution joins the cubic. In all cases the uncertainty due to taking $x^2 = \text{constant}$ can be estimated by comparing the solutions for $x = x_0$ and $x = L$.

With x^2 constant, (5.1) can easily be integrated. The two boundary conditions are (a) $E = E_j$ at $x = L$, where E_j is given by (5.6), and (b) to insure a smooth joining, the slope at $x = x_0$ must be the same as that of the cubic, namely

$$\left(\frac{dE}{dx}\right)_0 = \frac{2x_0}{2E_0 + I/E_0^2} \quad (5.7)$$

The first integration of (5.1) with $x = x_0$ gives

$$\left(\frac{dE}{dx}\right)^2 = \left(\frac{dE}{dx}\right)_0^2 + \frac{2}{L^2} \left[\frac{E^4}{4} - \frac{E^2}{2} (E_0^2 - I/E_0) - IE \right]_{E_0}^E \quad (5.8)$$

where (dE/dx) is given by (5.7) and $E_0^2 - I/E_0 = x_0^2 - A$. The E versus

x curve can now be found from (5.8) and

$$\begin{aligned}x - x_0 &= \int_{E_0}^E \left(\frac{dE}{dx} \right)^{-1} dE \\L - x &= \int_E^{E_j} \left(\frac{dE}{dx} \right)^{-1} dE\end{aligned}\quad (5.9)$$

In general we will be interested in cases where the junction solution holds over a length $L - x_0$ that is small compared to L , so we can take $x_0 = L$ in (5.7). It will also be valid to let E_0 in (5.7) and (5.8) be the value E_c on the cubic at $x = L$. Putting $E_c = E_0$ in equation (5.6) then gives E_j in terms of E_0 and $s_p = P/2n_i$, where P is the majority carrier concentration in the extrinsic region. In what follows we shall use these approximations. It will be convenient to express $x_0 = L$ in (5.7) in terms of I using $I = \sqrt{2}L\mathcal{L}$. We continue to use dimensionless quantities with E_1 , $2L_i$ and σE_1 as the units of field, length and current respectively, and $2L_i E_1$ as the unit of voltage. In general however we can express voltages in terms of kT/q .

When E_0^3 is either large or small compared to I , the junction solution takes a simple form and the field and potential distributions can be found analytically. We next consider two approximations that hold in those two cases respectively. Relatively good agreement between the two solutions at $E_0^3 = I$ indicates that each solution may be used up to this point.

Case of E_0^3 Large Compared to I

From (5.7) to (5.9)

$$x - x_0 = \sqrt{2}\mathcal{L} \int_{E_0}^E \left[\left(\frac{I}{E_0} \right)^2 + (E^2 - E_0^3)^2 \right]^{-1/2} dE \quad (5.10)$$

This can be solved in the following two overlapping ranges where the integrand has a simple form:

Range 1. Here $E - E_0$ is small compared to $2E_0$, so (5.10) becomes

$$x - x_0 = \frac{\mathcal{L}}{\sqrt{2E_0}} \sinh^{-1} \left[(E - E_0) \frac{2E_0^3}{I} \right] \quad (5.11)$$

Since E and E_0 are almost equal, we have for the voltage drop in this range

$$V - V_0 = E_0(x - x_0) \quad (5.12)$$

Range 2. Here $E^2 - E_0^2$ is large compared to $2(\mathcal{L}L/E_0)^2$, so (5.10) gives

$$L - x = \sqrt{2}\mathcal{L} \int_E^{E_j} \frac{dE}{E^2 - E_0^2} = \frac{\sqrt{2}\mathcal{L}}{E_0} \left(\operatorname{ctnh}^{-1} \frac{E}{E_0} - \operatorname{ctnh}^{-1} \frac{E_j}{E_0} \right) \quad (5.13)$$

From $E_0^3 \gg I$ it follows that Ranges 1 and 2 overlap. By joining the two solutions in the overlap region, the solution in Range 2 can be written as

$$x - x_0 = \frac{\mathcal{L}}{\sqrt{2} E_0} \ln \left[\frac{8E_0^3}{I} \frac{E - E_0}{E + E_0} \right] \quad (5.14)$$

Putting $E = E_j$ in (5.14) gives the distance over which the junction solution holds. In general we will be interested in cases where E_j is large compared to E_0 so (5.14) becomes

$$\frac{L - x_0}{l} = \frac{3}{2} \frac{\ln 2z_0}{z_0} \quad (5.15)$$

where $l = \sqrt{2}\mathcal{L}/I^{1/3}$ and as before $z_0 = E_0/I^{1/3}$. In conventional units

$$l = 2L \left(\frac{\mathcal{L}L_i}{L^2} \right)^{2/3} \quad (5.16)$$

Fig. 7 is a plot of $(L - x_0)/l$ versus z_0 . In germanium at room temperature $\mathcal{L}L_i$ will be around 10^{-5} cm. Thus the junction solution will hold over a region that is small compared to L if L is large compared to 3×10^{-3} cm.

Again it is convenient to use the relation $\mathcal{L} = \sqrt{2}kT/q$ to express the voltage in terms of kT/q .

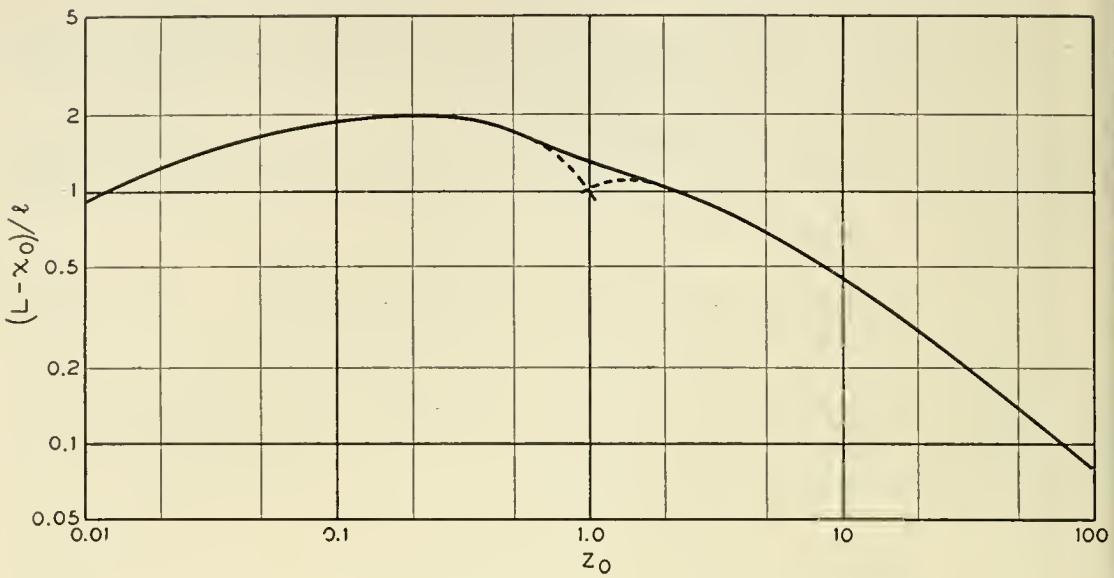
$$V_j - V = \int_E^{E_j} E \left(\frac{dE}{dx} \right)^{-1} dE = \frac{kT}{q} \ln \left(\frac{E_j^2 - E_0^2}{E^2 - E_0^2} \right) \quad (5.17)$$

By joining the two solutions in the overlap region, the voltage in Range 2 can be expressed as

$$V - V_0 = \frac{kT}{q} \ln \left(\frac{2E_0}{I} \right) (E^2 - E_0^2) \quad (5.18)$$

Setting $V = V_j$ and $E = E_j$ in (5.18) gives the total voltage drop in the region where the junction solution holds. Let ΔV be the difference between $V_j - V_0$ and the built in voltage drop at the junction. Then substituting (5.6) with $E_c = E_0$ into (5.18) and subtracting the built in drop we have for ΔV ,

$$\Delta V = \frac{kT}{q} \left[\ln \frac{E_0}{I} - \frac{E_0^2}{s_p} \right] \quad (5.19)$$

Fig. 7 — Variation of $(L - x_0)/\ell$ with z_0 .

I/E_0 is equal to the value of s on the cubic at $x = L$. For positive values of A the maximum value of E_0/I is $L/I = 1/\sqrt{2}\mathfrak{L}$ as can be seen from the cubic. In germanium at room temperature \mathfrak{L} is about 10^{-3} (for $2L_i = \text{unit length}$) so the reverse bias produces an additional voltage drop in the junction region equal to about $7kT/q$. For negative values of A the additional voltage drop near the junction would be higher.

Comparing (5.3) and (5.13) we see that the junction solution reduces to the zero bias solution when E^2 is large compared to $E_0^2 + 2$. In this case both solutions have the simple form

$$L - x = \sqrt{2}\mathfrak{L} \left(\frac{1}{E} - \frac{1}{E_j} \right) \quad (5.20)$$

and

$$V_j - V = \frac{2kT}{q} \ln \frac{E_j}{E} \quad (5.21)$$

Case of E_0^3 Small Compared to I

Now from (5.7) and (5.8) with $x_0 = L = I\sqrt{2}\mathfrak{L}$ we have

$$\begin{aligned} \left(\frac{dE}{dx} \right)_0^2 &= \left(\frac{2LE_0^2}{I} \right)^2 \\ \left(\frac{dE}{dx} \right)^2 &= \frac{1}{\mathfrak{L}^2} \left[2E_0^4 + (E - E_0)^2 \left(\frac{E^2}{2} + \frac{I}{E_0} \right) \right] \end{aligned} \quad (5.22)$$

Again there are two overlapping ranges where the solution has a simple form:

Range 1. Here E^2 is small compared to $2I/E_0$. This will be so even when E becomes large compared to E_0 . Setting $c_1^2 = 2E_0^5/I$ and $y = E - E_0$ in equation (5.22) and integrating gives

$$\begin{aligned} x - x_0 &= \mathfrak{L} \sqrt{\frac{E_0}{I}} \int_0^{E-E_0} \frac{dy}{\sqrt{c_1^2 + y^2}} \\ &= \mathfrak{L} \sqrt{\frac{E_0}{I}} \sinh^{-1} \left(\frac{E - E_0}{c_1} \right) \end{aligned} \quad (5.23)$$

and

$$V = V_0$$

$$= \frac{kT}{q} \sqrt{\frac{2E_0}{I}} (\sqrt{c_1^2 + (E - E_0)^2} - c_1) + 2E_0(x - x_0) \quad (5.24)$$

Range 2. Here E is large compared to E_0 . It follows from $E_0^3 \ll I$ that E is also large compared to c_1 . Setting $c_2^2 = 2I/E_0$ we have

$$\begin{aligned} L - x &= \sqrt{2}\mathfrak{L} \int_E^{E_j} \frac{dE}{E \sqrt{E^2 + c_2^2}} \\ &= \mathfrak{L} \sqrt{\frac{E_0}{I}} \left(\operatorname{csch}^{-1} \frac{E}{c_2} - \operatorname{csch}^{-1} \frac{E_j}{c_2} \right) \end{aligned} \quad (5.25)$$

Joining (5.21) and (5.23) where they overlap we have in range (2)

$$x - x_0 = \mathfrak{L} \sqrt{\frac{E_0}{I}} \ln \left(\frac{2I}{E_0^3} \right) \left[\frac{E}{c_2 + \sqrt{c_2^2 + E^2}} \right] \quad (5.26)$$

Putting $x = L$ and $E = E_j$ in (5.26) gives the length $L - x_0$ in which the junction solution holds. If E_j is large compared to c_2 , then

$$\frac{L - x_0}{l} = \sqrt{\frac{z_0}{2}} \ln \frac{4}{z_0^3} \quad (5.27)$$

where as before $z_0 = E_0/I^{1/3}$ and l is given by (5.16). Fig. 7 is a plot of $(L - x_0)/l$ versus z_0 . The two approximations (5.15) and (5.27) for $z_0^3 \ll 1$ and $z_0^3 \gg I$ respectively are shown dashed. Both become inaccurate as they are extended toward $z_0 = 1$. The point at $z_0 = 1$ was obtained graphically. Each approximation is in error by about 28 per cent here. The error will decrease as each approximation is extended away from $z_0 = 1$ toward its range of validity.

The voltage in Range 2 is given by

$$V_j - V = \frac{2kT}{q} \left[\sinh^{-1} \frac{E_j}{c_2} - \sinh^{-1} \frac{E}{c_2} \right] \quad (5.28)$$

or again joining (5.28) to the solutions in Range 1 we have in Range 2

$$V - V_0 = \frac{2kT}{q} \sinh^{-1} \frac{E}{c_2} + 2E_0(x - x_0) \quad (5.29)$$

The total voltage drop in the junction can be found by setting $V = V_j$ and $E = E_j$ in (5.29). The term $2E_0(L - x_0)$ will be negligible. When E^2 is large compared to $c_2^2 + 2$ the junction solution reduces to the zero current solution as can be seen by comparing (5.3) and (5.25). Then the solution has the simple form (5.20) and (5.21). E_j will always be large compared to c_2 . (E_j^2 is approximately s_p/e and $c_2^2 = 2s_0$ where s_0 is the value of s where the junction solution joins the cubic.) Thus the difference ΔV between $V_j - V_0$ and the built in voltage is

$$\Delta V = \frac{kT}{q} \ln \frac{E_0}{I} \quad (5.30)$$

Example. Fig. 8 shows the field distribution near the *IP* junction for the case $L = 2L_i$ and $A = \frac{2}{3}$, for which the intrinsic region is infinitely long. The field distribution near the junction, however, will be indistinguishable from that for $A = 0.665$, or $s_0 = 0.95$, for which the intrinsic region is about twice the effective length of current generation. We have taken $E_j = 30$, which corresponds to an excess acceptor density $P = 4.7 \times 10^3 n_i$ in the *P* region. Over the range where the junction solution holds the cubic gives an almost constant field $E = E_0 = E_c$. The junction solution goes from the cubic to the zero bias solution in a distance of the order of the Debye length. The sum of the built in voltage and the voltage derived from the cubic differ from the correct voltage by the order of kT/q or about kT/q . The total voltage is about $0.3 E_1 L_i$, which would be about 11 volts in germanium at room temperature.

VI. GENERAL CASE, UNEQUAL MOBILITIES

This Section deals with the general case where the ratio of the hole and electron mobilities is arbitrary. The procedure is similar to that used in the preceding Sections. Many of the results for $b = 1$ are useful in the present, more general, case. We shall deal first with the no-recombination case and again find that E is given by a cubic. However, the field distribution is no longer symmetrical and the coefficient of the I/E term in the cubic is a linear function of x instead of a constant. The differential equation for s in the recombination region remains un-

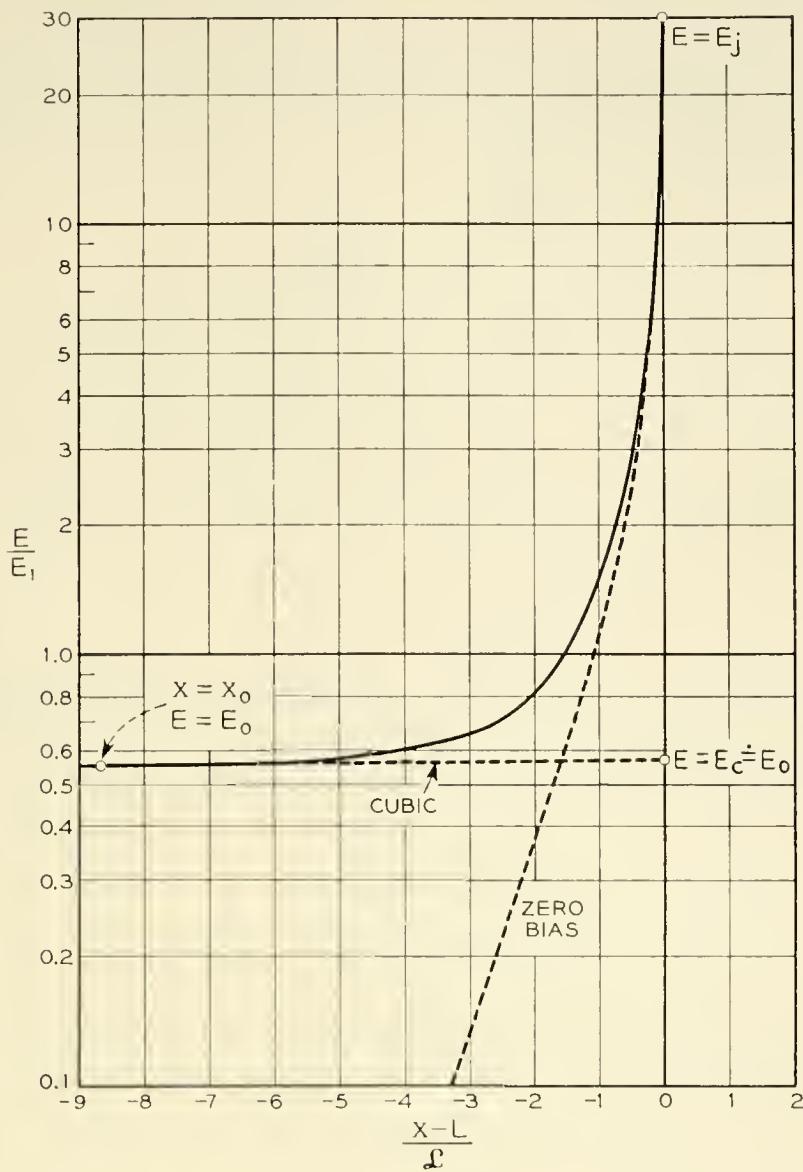


Fig. 8 — Field Distribution near the IP Junction for $L = 2L_i$ and $A = \frac{2}{3}$.

changed. It is no longer so that charge diffusion can be neglected except near the junctions. However, there is a linear combination of J_p and J_n in which the diffusion term is negligible except near the junctions.

Basic Relations

The equations are the two continuity (2.9) and Poisson's (2.1). The formulas for $g - r$ remain unchanged, since they involve only the statistics of recombination and are independent of mobility. The hole and electron currents are given by (2.2) with b arbitrary. Equation (2.2) for J_p in terms of E , p and n remains unchanged. Now J_n/b has the same

form as J_n had for the $b = 1$ case. It is therefore desirable to deal with the fictitious carrier flow $J_p + J_n/b$ and the fictitious current $q(J_p - J_n/b)$ since these have the same form in terms of E and $s = (n + p)/2n_i$ as J and I had for $b = 1$. Thus

$$J_p + \frac{J_n}{b} = 2n_i D \frac{d}{dx} \left(\frac{E^2}{E_1^2} - s \right) \quad (6.1)$$

$$q \left(J_p - \frac{J_n}{b} \right) = \frac{2}{1+b} \sigma \left[Es - \mathfrak{L}^2 \frac{d^2 E}{dx^2} \right] \quad (6.2)$$

where E_1 and \mathfrak{L} have the same meaning as before and the conductivity of intrinsic material is now $\sigma = qn_i\mu(1+b)$. As before D and μ are respectively the diffusion constant and mobility for holes. Equations (6.1) and (6.2) reduce respectively to (2.7) for J and (2.6) for $I = q(J_p - J_n)$ where $b = 1$.

When the flow is by pure diffusion, the holes and electrons diffuse "in parallel" so the effective diffusion constant is the reciprocal of the average of the reciprocal hole and electron diffusion constants. Hence the effective diffusion length is given by

$$L_i^2 = D\tau \frac{2b}{1+b} \quad (6.3)$$

We continue to let $2L = I/qg$ be the effective length of current generation; again it is the actual length for the no recombination case. Let x_n and x_p be the coordinates of the NI and IP junctions respectively. Since the problem is not symmetrical we will not take $x = 0$ in the center of the intrinsic region even for the no-recombination case.

No-Recombination Case

Setting $r = 0$ we can immediately integrate the continuity equat'

$$\frac{dJ_p}{dx} = \frac{dJ_n}{dx} = g$$

subject to the boundary conditions:

$$\begin{aligned} \text{at the } NI \text{ junetion, } x = x_n, \quad J_p &= 0, & J_n &= -I/q \\ \text{at the } IP \text{ junetion, } x = x_p, \quad J_p &= I/q, & J_n &= 0 \end{aligned} \quad (6.4)$$

The result is $J_p = g(x - x_n)$ and $J_n = g(x - x_p)$. This agrees with $I = q(J_p - J_n) = 2qgL$ since $2L = x_p - x_n$ is the length of the intrinsic region, which, for no-recombination, is also the effective length of cur-

rent generation. It will be convenient to choose $x = 0$ so that $x_n = -x_p/b$. Then the origin is nearer to the NI junction for $b > 1$. Now from this and the boundary conditions (6.4) and $I = 2qgL$ we have the positions of the junctions:

$$\frac{x_p}{L} = \frac{2b}{1+b}, \quad \frac{x_n}{L} = \frac{-2}{1+b} \quad (6.5)$$

As before, the junetions are at $x = \pm L$ for $b = 1$.

We can now find the fictitious carrier flow $J_p + J_n/b$ and the fictitious current $q(J_p - J_n/b)$ as functions of x .

$$J_p + \frac{J_n}{b} = \left(\frac{1+b}{b} \right) gx \quad (6.6)$$

$$q \left(J_p - \frac{J_n}{b} \right) = \frac{2I}{1+b} \left(1 + \frac{\beta x}{L} \right) \quad (6.7)$$

where the dimensionless parameter $\beta = (b^2 - 1)/4b$. Thus the fictitious current $q(J_p - J_n/b)$ is equal to the actual current times a linear function of x . This function is always positive and varies from a minimum of $1/b$ to a maximum of 1.

Combining (6.6) with (6.1) and integrating gives the equation

$$\frac{E^2}{E_1^2} - s = \left(\frac{x}{2L_i} \right)^2 - A \quad (6.8)$$

that we had before. Now, however, E is not a minimum at the same point where s is a maximum. As before, when recombination is negligible throughout all of the intrinsic region, A determines the voltage; and, when recombination is important over part of the region, A determines both the voltage and the length of the intrinsic region $x_p - x_n > 2L = r/a$.

Combining (6.7) with (6.2) gives

$$I \left(1 + \frac{\beta x}{L} \right) = \sigma \left[Es - \mathfrak{L}^2 \frac{d^2 E}{dx^2} \right] \quad (6.9)$$

which is similar to the previous (3.6) except that I is multiplied by the factor $1 + \beta x/L$, which varies from $1 + 1/b$ to $1 + b$. The same arguments used in Section V apply here and show that the second term in brackets (the diffusion term) can be neglected except near the junctions. In other words, although I is always part drift and part diffusion, $I(1 + \beta x/L)$ is approximately pure drift except at the junctions.

Eliminating s between (6.9) and (6.8) and neglecting the diffusion

term in (6.9) gives the cubic equation

$$\frac{E^2}{E_1^2} - \frac{I}{\sigma E_1} \left(1 + \frac{\beta x}{L}\right) = \left(\frac{x}{2L_i}\right)^2 - A \quad (6.10)$$

for the field distribution.

In germanium, where $b = 2.1$, $\beta = 0.406$, $x_p = 1.35L$ and $x_n = -0.65L$. The coefficient of $I/\sigma E_1$ therefore varies from 1.47 to 3.10, or by a factor of a little more than 2. This will introduce some asymmetry into the E versus x curve in the low field region where the fictitious carrier flow $J_p + J_n/b$ is by diffusion. It is evident that, as the voltage increases, the field versus x curve becomes increasingly symmetrical about the $x = 0$ point; so the effect of having $b \neq 1$ is simply to shift the field distribution along the x axis.

Recombination

The arguments of section 4 again apply. Where recombination is important, $n - p$ is small compared to $n + p$, so $g - r = g(1 - s^2)$. The diffusion term dominates in the fictitious particle flow $J_p + J_n/b$; that is, E^2/E_1^2 is small compared to s , so (6.1) becomes

$$J_p + \frac{J_n}{b} = -2n_i D \frac{ds}{dx}$$

The continuity equations give

$$\frac{d}{dx} \left(J_p + \frac{J_n}{b} \right) = \left(1 + \frac{1}{b}\right) (g - r) = \frac{n_i(1+b)}{2\tau b} (1 - s^2)$$

So again we have

$$\frac{d^2s}{dx^2} = -\frac{(1 - s^2)}{2L_i^2} \quad (6.11)$$

The solution joins the no recombination solution where $s = A - (x/2L_i)^2$. Therefore A is again related to s_0 , the maximum s , by $A = s_0(1 - s_0^2/3)$ and the s versus x curve is given by (4.8) and is symmetrical about the point where s is a maximum. When the recombination solution joins onto no-recombination solutions, there will be a different no-recombination solution on each side of the recombination region. The junctions will be at the points x_p and x_n on the respective no-recombination solutions. The length of the intrinsic region will not be $x_p - x_n = 2L$ since the $x = 0$ points are different on the two no-recombination solutions and are separated by a region of maximum recombination.

To find E when s is known we express the current $I = q(J_p - J_n)$ in terms of s and E . Since $n - p$ is small compared to $n + p$, we set $n = p = sn_i$ in (2.2) and obtain

$$I = \sigma \left[sE - \frac{1-b}{1+b} \frac{kT}{q} \frac{ds}{dx} \right] \quad (6.12)$$

Thus the current contains both a drift and a diffusion term. This is to be expected for unequal mobilities. When holes and electrons have the same concentration gradient, the electrons, which have the higher diffusion constant, diffuse faster than the holes; hence the diffusion gives a net current. It is seen that in the recombination region the total carrier concentration has a symmetrical distribution about the point where it is a maximum but the field remains unsymmetrical.

Junction Solution

When $(E_0/E_1)^3$ is large compared to $I/\sigma E_1$ the junction solution is independent of b ; so the solution obtained in Section V is valid. In all cases the junction solution can be found using the method of Section V. The effect of b will be small over most of the range where the junction solution holds because the concentration of one type of carrier will be negligible. To be exact, I in (5.8) should be multiplied by the factor $(1 + \beta x_0/L)$, which can be taken to be $(1 + b)/2b$ at the NI junction and $(1 + b)/2$ at the IP junction. Instead of equation (5.7) we have

$$\left[2E_0 + \frac{I}{E_0^2} \left(1 + \frac{\beta x_0}{L} \right) \right] \left(\frac{dE}{dx} \right)_0 = 2x + \frac{I}{E_0} \frac{\beta}{L} \quad (6.13)$$

as can be seen by differentiating (6.10) with $E_1 = 2L_i = \sigma = 1$.

VII. EFFECT OF FIXED CHARGE

This section will deal briefly with the case where there is some fixed charge but where the carrier charge cannot be neglected. For no recombination, the field distribution is given by a first order differential equation. Solutions in closed form are obtained for the case of pure drift flow. For recombination and charge neutrality the solution in Section IV is valid provided the fixed charge is small compared to n_i . We have seen that at large fields the E versus x curve becomes linear, corresponding to a fixed charge density of N_i where $N_i = \sqrt{2}n_i\mathcal{L}/L_i$. Thus the fixed charge may have a dominant effect on the space charge while having a negligible effect on the solution in the range where recombination is important.

Let the density of fixed charge be $N = N_d - N_a$ = excess density of donors over acceptors. N may be either positive or negative. In what follows we shall assume that N is positive. So the structure is $N\nu P$ where ν means weakly doped n-type. Equations (2.2) for the hole and electron currents remain unchanged. Poisson's equation becomes

$$\frac{dE}{dx} = a(p - n + N) \quad (7.1)$$

We shall deal with the general case of arbitrary mobilities. As in Section VI it is convenient to deal with a fictitious current $q(J_p - J_n/b)$ and a fictitious particle flow $J_p + J_n/b$. The extra term in (7.1) drops out by differentiation when (7.1) is substituted into the equation for $J_p - J_n/b$ so (6.2) remains unchanged. However, instead of (6.1) we have

$$J_p + \frac{J_n}{b} = 2n_i D \frac{d}{dx} \left(\frac{E^2}{E_1^2} - s \right) - \mu N E \quad (7.2)$$

So the fictitious particle flow is no longer the gradient of a potential involving only E and s .

No Recombination

As in Section VI the continuity equations can be immediately integrated to give (6.6) and (6.7). Again I is given by (6.9) where the diffusion term on the right can be neglected except at the junctions; so again we have $\sigma s E = I(1 + \beta x/L)$. Substituting this into (7.2) and combining (7.2) and (6.6) gives a first order differential equation for E versus x . It is convenient to again use dimensionless quantities with E_1 , $2L_i$ and σE_1 as the units of field, length and current respectively. Then the differential equation becomes

$$\frac{d}{dx} \left[E^2 - \frac{I}{E} \left(1 + \frac{\beta x}{L} \right) \right] = 2(x + \alpha E) \quad (7.3)$$

where

$$\alpha = \frac{N}{N_i}$$

and as before $N_i = \sqrt{2n_i \mathcal{L}}/L_i$, which is around 4×10^{10} in germanium at room temperature. The solution of (7.3) contains one arbitrary constant (which corresponds to A in the $N = 0$ case). The lower limit on the constant is determined by the necessity of joining onto a recombination solution where s approached unity. The positions of x_n and x_p of the $N\nu$ and νP junctions respectively are given by (6.5).

In the region of low fields where E^3 is comparable to or less than I , (7.3) would have to be solved graphically or on a machine. At higher fields the equation is easily integrated as discussed below.

Case of Pure Drift

When the flow is entirely by drift, $E^3 \gg I$ and (7.3) becomes

$$\frac{dE}{dx} = \frac{x}{E} + \alpha \quad (7.4)$$

which is made integrable by the substitution $E = yx$. A family of solutions for positive E throughout the ν region is

$$(E - a_1x)^{a_1}(E + a_2x)^{a_2} = E_0^{a_1+a_2} \quad (7.5)$$

where $2a_1 = \sqrt{4 + \alpha^2} + \alpha$ and $2a_2 = \sqrt{4 + \alpha^2} - \alpha$ and E_0 is the value of E at $x = 0$. For an intrinsic region $N = \alpha = 0$ and (7.5) reduces to $E^2 = E_0^2 + x^2$, which is the same as (3.9) for negative A . Fig. 9 shows several curves for various values of E_0 . These remain above, and at large distances approach, the asymptotic solutions $E = a_1x$ on the right of the origin and $E = -a_2x$ on the left. These curves differ from the corresponding curves for an intrinsic region in that the straight line asymptotes now have slopes of a_1 and $-a_2$ instead of ± 1 . Toward the P side the slope is greater because the positive charge qN of the excess donors is added to the charge of holes. Toward the N side of the ν region the slope is reduced because N compensates to some extent for the electron charge. As α increases and the ν region becomes more n type, the solution approaches that for a simple NP junction, where $E = \alpha x$ on the N side.

Another set of solutions of (7.4) are given by

$$(a_1x - E)^{a_1}(a_2x + E)^{a_2} = a_1^{a_1}a_2^{a_2}x_c^{-2} \quad (7.6)$$

Several of these are shown in Fig. 9. They remain below the linear asymptotes and go through zero field at $x = \pm x_c$. Actually these will join onto solutions of the more general equation (7.3) when E becomes small and the diffusion term becomes important.

Recombination. When the fixed charge density is small compared to the intrinsic hole and electron density the treatment of recombination in Section IV remains valid. The recombination solution joins onto a solution of (7.3) at small fields. When N is comparable to n_i the recombination solution is difficult even with the assumption of charge neutrality.

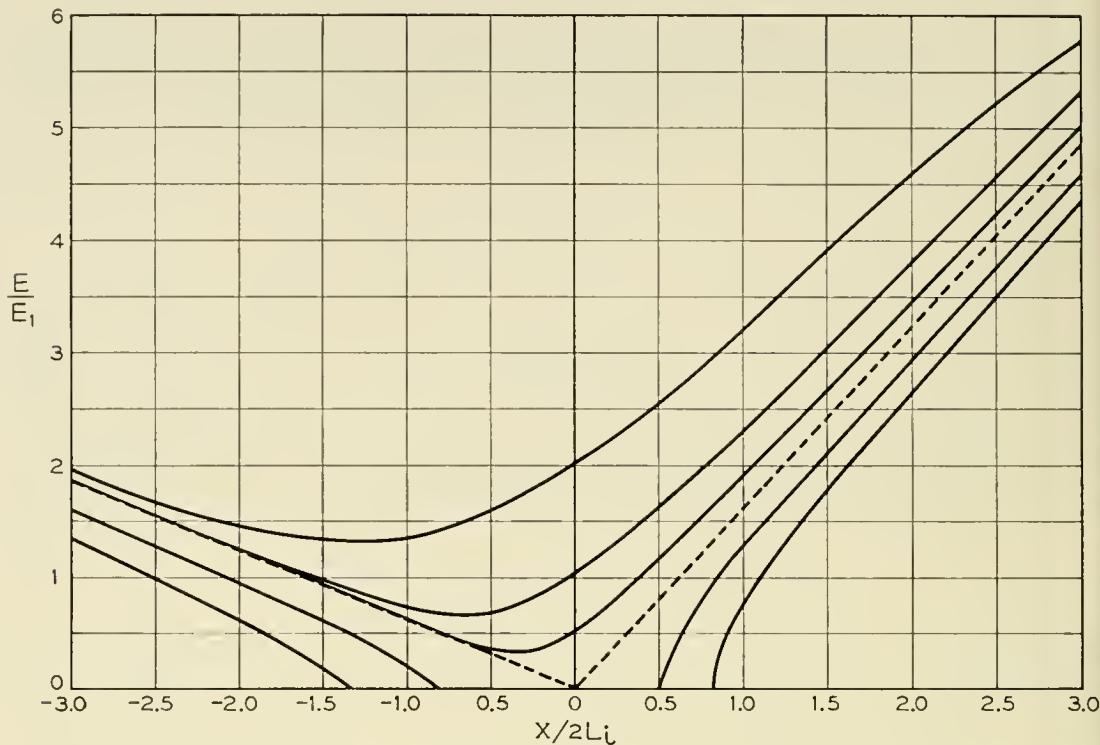


Fig. 9 — Field Distribution in the Range of Pure Drift for a fixed charge $N = N_i$, or $\alpha = 1$.

ACKNOWLEDGEMENTS

The author wishes to thank Miss M. M. Segrich for doing the extensive computations and plotting the curves, and Miss M. C. Gray for help with the calculations leading to Fig. 7.

APPENDIX A

Prim's Zero-Current Approximation

Prim's analysis is based on the assumption that the hole and electron currents are negligibly small differences between their drift and diffusion terms. Setting $J_p = J_n = 0$ then gives n and p as functions of the potential, which is found by substituting n and p into Poisson's equation and solving subject to the boundary conditions at the junctions. These conditions involve the applied bias and the majority carrier densities in the extrinsic regions. Since the current is assumed to vanish, the phenomena of carrier generation and recombination do not enter the problem and the results are independent of carrier mobility. The results will be exact when there is no applied voltage; the potential drop across the unit is then the built-in potential. In this appendix we use an internal consistency check to see for what values of applied bias the analysis

breaks down. First we find where the carrier concentration is in error by finding the bias at which the minimum drift current as calculated from $q\mu(n + p)E$ becomes equal to the total current, as found from the excess of generation over recombination in the intrinsic region. We then go on to find where the error in carrier concentration gives a sufficient error in space charge to affect the calculation of electric field. As we shall see, the zero-current approximation gives too low a carrier concentration in the interior of the intrinsic region. This will lead to serious errors in the field distribution only if the space charge of the carriers is important. When the bias is sufficiently high or the intrinsic region sufficiently narrow that the intrinsic region is swept so clean that the carrier space charge is, in fact, negligible, it will not matter that the calculated carrier density is too low, even by orders of magnitude. In such cases, the electric field is constant throughout most of the intrinsic region.

In the following we shall, for simplicity, take $b = 1$ and assume that the extrinsic regions are equally doped so that the problem is symmetrical.

Carrier Density

We now find where, on the zero current assumption, the drift current becomes equal to the total current. This involves knowing only the carrier concentrations and the field E_i in the center of the intrinsic region, where the drift current $q\mu(n + p)E_i$ is a minimum. By symmetry n and p are equal here and $n = p = n_i \exp(-qV_a/2kT)$ where V_a is the applied bias. The minimum field E_i is given by the total voltage drop V and the field penetration parameter η , which is the ratio of the minimum field to the average field. Thus $\eta = 2LE_i/V$ where $2L$ is the width of the intrinsic regions. The difference between V and V_a is the built-in voltage $(2kT/q)/\ln(N/n_i)$ where N is the majority carrier concentration in the extrinsic regions. We now have for the drift current in the center of the intrinsic region

$$q\mu(n + p)E = q\eta D \left(\frac{qV}{kT} \right) L n_i \exp \left(-\frac{qV_a}{2kT} \right) \quad (\text{A1})$$

We next find the total current from the excess of generation over recombination in the intrinsic region. From the zero current assumption, $np = n_i^2 \exp(-qV_a/kT)$ is constant throughout the intrinsic region. Hence $g - r$ is constant. So the current $I = q(g - r)2L = qL(n_i^2 - np)/\tau n_i$ is

$$I = \frac{qLn_i}{\tau} \left[1 - \exp \left(-\frac{qV_a}{kT} \right) \right] \quad (\text{A2})$$

Equating this to the drift current (A1) in the center of the intrinsic region gives

$$\left(\frac{L}{L_i}\right)^2 = \eta \frac{qV}{2kT} \operatorname{csch} \left(\frac{qV_a}{2kT} \right) \quad (\text{A3})$$

The error in carrier concentration is less for narrower intrinsic regions and lower biases. Thus (A3) gives a curve of L versus V_a such that the zero current solution gives a good approximation to carrier concentration for points in the $V_a L$ plane lying well below the curve. As expected, for zero bias, the solution is good for any value of L . However, for a bias of several kT/q , the solution for carrier concentration breaks down unless L is a very small fraction of a diffusion length.

Carrier Space Charge.

In Prim's analysis the carrier space charge is so low throughout most of the intrinsic region that the field remains approximately constant and equal to E_i . However there must be enough carriers present that the drift currents of holes and electrons can remove the carriers as fast as they are generated. In this section we ask where the space charge of the necessary carriers becomes large enough that its effect on the field can no longer be neglected. Let ΔE be the change in field due to the space charge in the intrinsic region (not counting the high field regions near the junctions). Unless ΔE is small compared to E_i the neglect of carrier space charge will not be justified. We shall find the ratio of ΔE to E_i .

If the field is to be approximately constant, then the hole and electron concentrations can easily be found from the hole and electrons currents. We shall deal with applied biases of at least a few kT/q , for which recombination is negligible and the total current $I = qg2L = qn_i L / \tau$. Since $g - r \doteq g$ is constant, the hole and electron currents are linear in x and, for constant field, are proportional to the hole and electron concentrations respectively. Thus the net space charge of the moving carriers $q(p - n)$ is proportional to x and varies from zero in the center of the intrinsic region to qp near the IP junction, where n is small compared to p and the current flows by hole drift, so $I = q\mu p E_i$. Thus the maximum charge is $I/\mu E_i$ and the total positive charge of the carriers on the P side of the center is $IL/2\mu E_i$. This gives a drop in field

$$\Delta E = \frac{aIL}{2q\mu E_i} = \frac{an_i}{2} \frac{kT}{qE_i} \frac{L^2}{L_i^2}$$

Dividing by $E_i = \eta V / 2L$ gives

$$\frac{\Delta E}{E_i} = \frac{L^4}{\mathfrak{L}^2 L_i^2} \left(\frac{kT}{\eta q V} \right)^2 \quad (\text{A4})$$

Setting ΔE equal to some fraction, say 20 per cent of E_i , gives a family of curves for V versus L with η as a parameter. Prim has plotted such curves in Fig. 11 of his paper. His curves will be good approximations when V for a given L and η lies above the V given by (A4).

Prim's results are expressed in terms of the parameters $U = qV/2kT$ and $\hat{L} = 2L/\mathfrak{L}_e$ where \mathfrak{L}_e is the Debye length in the extrinsic material. \mathfrak{L}_e is given by the same formula as \mathfrak{L} except that N replaces n_i . Substituting these into (A4) and setting $\Delta E = E_i/5$ gives

$$\hat{L} = 3.57 \frac{NL_i}{n_i \mathfrak{L}} \eta U \quad (\text{A5})$$

Prim's U versus \hat{L} curves will be accurate up to the point where they intersect the corresponding curves from (A5). For germanium a reasonable value of $NL_i/n_i \mathfrak{L}$ is about 10^6 . This says that Prim's curves go bad at about $\hat{L} = 10^4$, which would be about 2.1×10^{-2} cm in germanium at 300°C .

Branching of the V versus L Curves

An effect which does not emerge from the zero-current analysis is that V may have several values for the same L and η . In other words the V versus L curve for given η will have more than one branch. Specifically, there will be a single V versus L curve up to a certain L at which the curve splits into three branches that diverge as L increases. This may be seen as follows: Consider an intrinsic region that is long compared to the diffusion length. Suppose a bias is applied that is low enough not to appreciably affect the space charge and potential drop at the junctions. A current will flow and a proportional, ohmic voltage drop will be developed across the intrinsic region. If the intrinsic region is long enough, this ohmic voltage may become large compared to the built-in voltage before the voltage drop at the junctions has changed appreciably. In this range the field penetration parameter will be rising from zero to about unity as V increases from the built-in voltage and approaches the ohmic voltage. As the voltage continues to increase, the space charge begins to penetrate the intrinsic region and a majority of the voltage drop comes in the space charge regions. Let L be the effective length of current generation. When L is larger than a diffusion

length but small compared to the length of the intrinsic region, then the voltage drop at the ends of the intrinsic region will be proportional to L^2 while the current, and consequently the minimum field, will be proportional to L . Thus η will be proportional to $1/L$ and will decrease as V increases and the region becomes more swept. Finally the two space charge regions meet; then η rises again with V and approaches unity. Hence, for a given η and length of intrinsic region, there will be three different values of V . For lower L the dip in the η versus V curve will be less, and there will be only one V for some values of η . Since η starts from zero at the built-in voltage and approaches unity for infinite voltage, there must be either one or three values of V for every η . Thus when the V versus L curve (or in Prim's notation the U versus \hat{L} curve) branches, it branches at once into three curves. Prim's plot gives the upper branch in cases where all three are present.

A Medium Power Traveling-Wave Tube for 6,000-Mc Radio Relay

By J. P. LAICO, H. L. McDOWELL and C. R. MOSTER

(Manuscript received May 15, 1956)

This paper discusses a traveling-wave amplifier which gives 30 db of gain at 5 watts output in the 5,925- to 6,425-mc common carrier band. A description of the tube and detailed performance data are given.

TABLE OF CONTENTS

	Page
I. Introduction.....	1285
II. Design Considerations.....	1288
III. Description of the Tube.....	1291
3.1 General Description.....	1291
3.2 The Electron Gun and Electron Beam Focusing.....	1295
3.3 The Helix.....	1302
3.4 The Collector.....	1311
IV. Performance Characteristics.....	1314
4.1 Method of Approach.....	1314
4.2 Operation Under Nominal Conditions.....	1315
4.3 Operation Over an Extended Range.....	1325
4.4 Noise Performance.....	1333
4.5 Intermodulation.....	1336
V. Life Tests.....	1342
VI. Acknowledgements.....	1343

I. INTRODUCTION

During the past ten years traveling-wave tubes have received considerable attention in vacuum tube laboratories, both in this country and abroad. So far their use in operating systems has been somewhat limited, the most notable exceptions being in radio relay service in France, Great Britain, and Japan. However, it appears that sufficient progress in both tube and system design has been made so that traveling-wave tubes may see widespread application in the near future.

This paper describes an experimental helix type traveling-wave tube representative of a class which may see extensive use as a power amplifier in radio relay systems. The tube is designated as the Bell Laboratories type MI789. Stated briefly, the performance characteristics under nominal operating conditions are:

Frequency Range.....	5,925-6,425 mc
Power Output.....	5 watts
Gain at 5 watts output.....	31-35 db
Noise Figure.....	< 30 db

The tube is designed for use with waveguide input and output circuits. The input voltage standing wave ratio (VSWR) is less than 1.1 and the output VSWR is less than 1.4 over the 500-mc band when the tube is delivering 5 watts of output. Fig. 1 shows a photograph of an MI789 and of an experimental permanent-magnet focusing circuit.

In developing this tube we have endeavored to produce an amplifier which could be considered "practical" for use in a transcontinental radio relay system. Because such an application requires a high degree of reliability and refinement in performance, the tube was rather conservatively designed. This made it possible to obtain the desired gain and power output without difficulty. On the other hand, the contem-

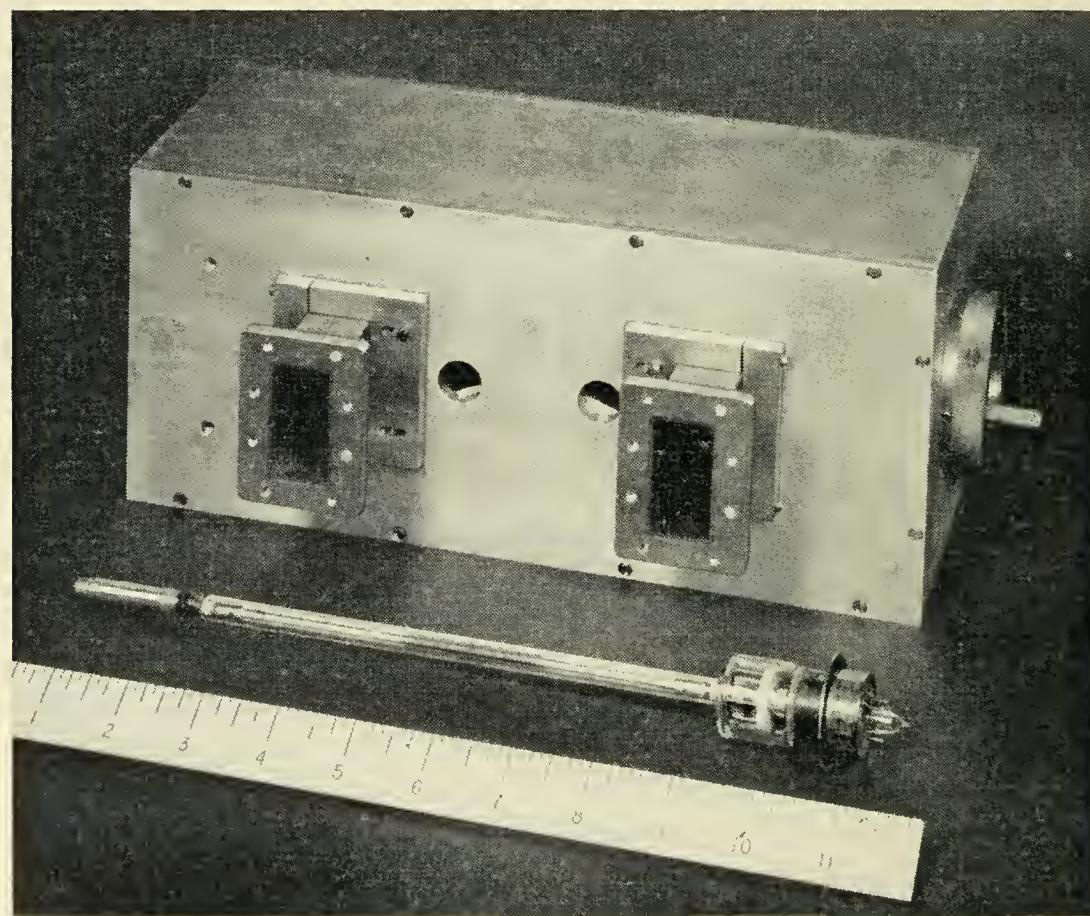


Fig. 1 — The MI789 traveling-wave tube and an experimental permanent magnetic circuit used to focus it. The circuit contains two specially shaped bar magnets between which the tube is mounted. The magnetic flux density obtained is 600 gauss, and the overall circuit weight is about 25 pounds.

plated system application made it necessary to investigate in detail the problems associated with band flatness, matching, noise output, certain signal distortions, reproducibility, and long life.

The solution of some of these problems required the development of a precisely constructed helix assembly in which the helix winding is bonded to ceramic support rods by glaze. Others required the initiation of a life test program. Early results indicate that life exceeding 10,000 hours can be obtained. This, in no small measure, is a result of a de potential profile which minimizes the ion bombardment of the cathode. Since power consumed by focusing solenoids seriously degrades the overall efficiency of a traveling-wave amplifier, permanent magnet focusing circuits such as the one shown in Fig. 1 have been designed. Finally, to further improve efficiency, a collector which can be operated at about half the helix voltage was developed.

The major difficulties encountered in the course of the MI789 development were: excessive noise output, ripples in the gain-frequency characteristic, and lack of reproducibility of gain. There is evidence that a growing noise current wave on the electron stream was the source of the high noise output. This phenomenon has been observed by a number of experimenters but is not yet fully explained. By allowing a small amount of the magnetic focusing flux to link the cathode, the growing noise wave was eliminated, and the noise reduced to a reasonable level for a power amplifier. Reflections caused by slight non-uniformities in the helix pitch were the source of the gain ripples. Precise helix winding techniques reduced these reflections so that the ripples are now less than ± 0.1 db. The lack of reproducibility in gain was caused by variations in helix attenuation. Here, too, careful construction techniques alleviated the problem so that in a recent group of tubes the range of gain variation at five watts output was ± 2 db.

We have divided this paper into four main parts. The next section discusses some of the factors affecting the design of the traveling-wave tube. (We will henceforth use the abbreviation TWT.) Section III describes the tube itself. Certain performance data are included there when closely related to a particular portion of the tube. Section IV considers the rf performance in detail. There comparisons are made between the performance predicted from TWT theory and that actually observed. Finally Section V summarizes our life test experience.

This paper is written primarily for workers in the vacuum tube field and assumes knowledge of TWT theory. However, we believe that readers interested in TWT's from an application standpoint may also benefit from the discussion of the rf performance in Section IV. Much of that section can be understood without detailed knowledge of TWT's.

II. DESIGN CONSIDERATIONS

While TWT theory served as a general guide in the development of the MI789, a number of important tube parameters had to be determined either by experimentation or by judgement based on past experience. The most important of these were:

Saturation power output.....	12 watts
Mean helix diameter.....	90 mils
γa	~ 1.6
Magnetic flux density.....	600 gauss
Cathode current density.....	$\sim 200 \text{ ma/cm}^2$

These quantities and the requirement of 30-db gain at five watts output largely determined the TWT design.

The saturation output of 12 watts was found necessary to obtain the desired linearity at five watts output and the γa value of 1.6 to obtain the flattest frequency response over the desired band.

The choice of helix diameter and magnetic flux density represented a compromise. For the highest gain per unit length, best efficiency, and lowest operating voltage, a small helix diameter was called for. On the other hand, a large helix diameter was desirable in order to ease the problem of beam focusing and to facilitate the design of a light-weight permanent magnet focusing circuit. In particular, the design of such a circuit can be greatly simplified if the field strength required is less than the coercive force of available magnetic materials. This allows the use of straight bar magnets instead of much heavier horseshoe magnets. Moreover, the size and weight of the magnetic circuit is minimized by employing a high energy product material. These considerations led us to choose a flux density of 600 gauss, thereby permitting us to design a magnetic circuit using Alnico bar magnets.

To obtain long tube life we felt it desirable to limit the helix interception to about one per cent of the beam current. On the basis of past results we estimated that this could be done with a magnetic flux density 2.6 times the Brillouin value for a beam entirely filling the helix. With this restriction, Fig. 2 shows how the TWT design is affected by varying the helix diameter. A choice of 600 gauss is seen to result in a mean helix diameter of 90 mils.

In the selection of cathode current density, a compromise between long life and ease of focusing had to be made. To obtain long life, the current density should be minimized. However, this calls for a highly convergent gun which in turn complicates the focusing problem. We decided to use a sprayed oxide cathode operating at about 200 ma/cm^2 . Experience with the Western Electric 416B microwave triode had shown

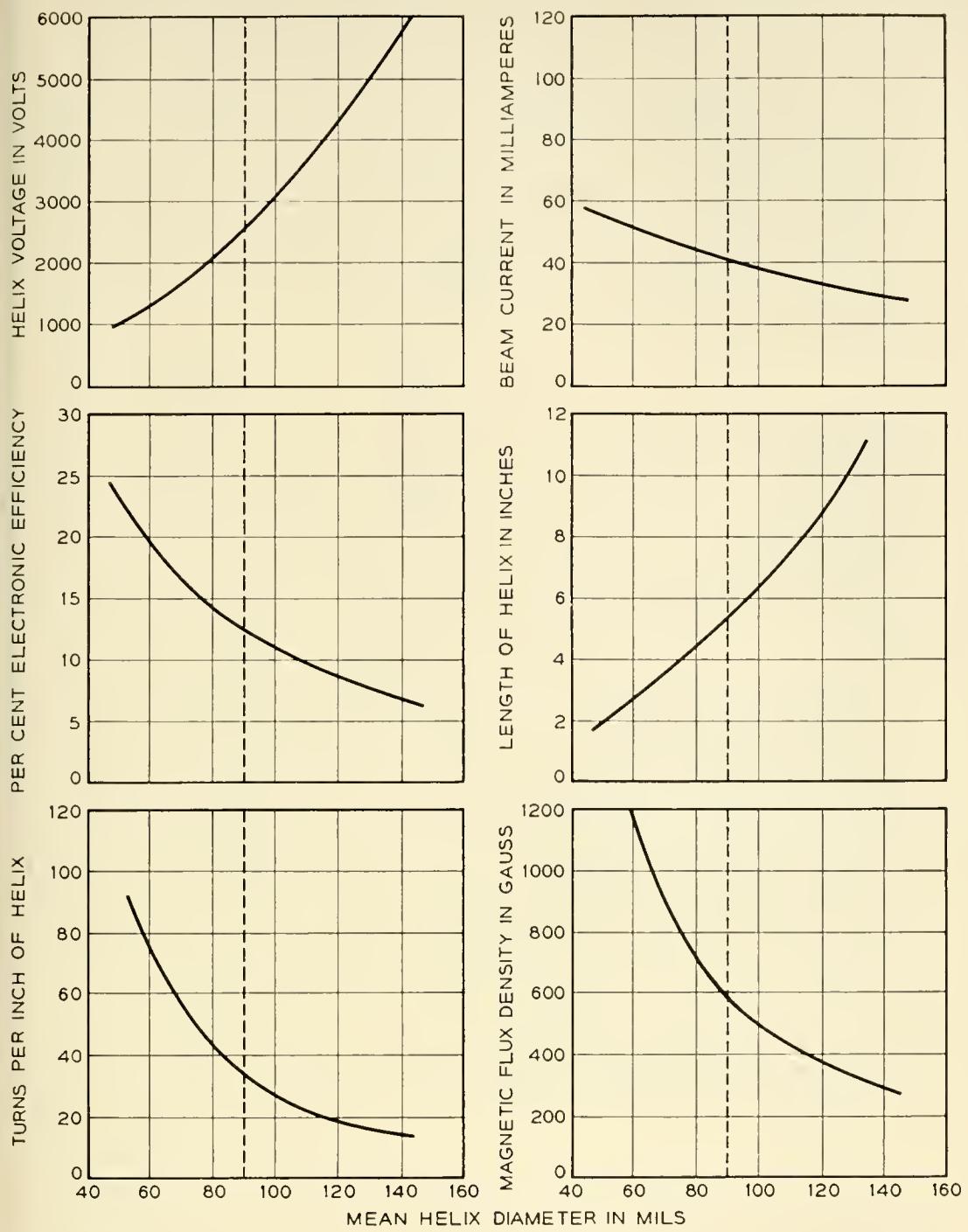


Fig. 2 — Alternate designs for the M1789. These curves are an estimate of how the TWT design would be affected by changing the helix diameter. They represent essentially a scaling of the M1789 design. In all cases the expected maximum power output is 12 watts and the low-level gain is 33 db. The line at 90 mils mean diameter in the curves represents the present M1789 design. In these calculations it was assumed that:

- $\gamma a = 1.6$
- power output = $2.1 C I_0 V_0 = 12$ watts
- the magnetic flux density is 2.6 times the Brillouin flux density for a beam entirely filling the helix.
- the ratio of wire diameter to pitch is 0.34.
- the dielectric loading factor is 0.79.
- the ratio of effective beam diameter to mean helix diameter is 0.5.

TABLE I — SUMMARY OF M1789 DESIGN

I. Helix Dimensions

Mean Diameter	90	mils
Inside Diameter	80	mils
Wire Diameter	10	mils
Turns per Inch	34	
Pitch	29.4	mils
Wire Diameter/Pitch	0.34	
Active Length	5½	inches

II. Voltages and Currents

Electrode	Voltage (volts)	Current (ma)
Cathode	0	40
Beam Forming Electrode	0	0
Accelerator	2600	< 0.1
Helix	2400	< 0.4
Collector	1200	> 39.5
Heater Power		6 watts

III. TWT Parameters at Midband (6175 mc)

γa	1.58
$k a$	0.148
C	0.058
$Q C$	0.29
N (number of λ 's on helix)	30
Dielectric Loading factor	0.79
Impedance Reduction factor	0.4

} As defined by Tien⁸

IV. Electron Gun

Gun type — Converging Pierce Gun
Cathode type — Sprayed oxide
Cathode Current Density 213 ma/cm ² (for $I_K = 40$ ma)
Cathode diameter — 192 mils
Convergence half angle 12° 40'
Cathode radius of curvature (r_c) 438 mils
Anode radius of curvature (r_a) 190 mils
r_c/r_a 2.3

Perveance 0.3×10^{-6} amps/volts^{3/2}

$$\sqrt{V_A/T_K} = 1.61 \text{ for } T_K = 720^\circ\text{C}$$

At the beam minimum in absence of magnetic field:

r_{min} (from Pierce ¹⁰)	11.5	mils
r_{95}/r_c	0.220	
r_{95}	20.5	mils
r_e/σ	3.50	
σ	4.80	mils
Brillouin flux density for 80 mil helix ID	240	gauss
Actual focusing flux density required	600	gauss
Beam transmission from cathode to collector at 5 watts output	99%	

V. RF Performance

Frequency range	5925–6425	mc
Saturation power output	12	watts
Nominal power output	5	watts
Gain at 5 watts	31–35	db
Noise figure	< 30	db
Input VSWR	< 1.1	
Output VSWR (at 5 watts)	< 1.4	} impedance match to WR 159 waveguide

For an explanation of symbols see page 1345.

that tube life in excess of 10,000 hours was possible with such a cathode. Moreover, an electron gun of the required convergence (about 13° half angle) could be designed using standard techniques.

The various dimensions, parameters, voltages and currents involved in the design of the MI789 are summarized in Table I. For the sake of completeness, some rf performance data are also included.

III. DESCRIPTION OF THE TUBE

3.1 General Description

This section describes the mechanical structure of the MI789 and presents some performance data closely associated with particular portions of the tube. The overall rf performance is reserved for consideration in the next section. In the MI789 we have tried to achieve a design which could be easily modified for experimental purposes and which would also be adapted to quantity production. To assist in obtaining low gas pressure, a rather "open" structure is used, thereby minimizing the pumping impedance. In addition, all parts are designed to withstand comparatively high temperatures during outgassing, both when the tube is pumped and, in the case of the helix and gun assemblies, during a vacuum firing treatment prior to final assembly. Fig. 3 shows an MI789 and its subassemblies. Fig. 4 shows a simplified drawing of the whole tube and Fig. 5 shows how the tube is mounted with respect to the permanent magnet circuit and to the waveguides. The permanent magnets are shown schematically in Fig. 5. In actual practice they are shaped so as to produce a uniform field between the pole pieces. The means of doing this was discussed by M. S. Glass at the Second Annual Meeting of the I.R.E. Professional Group on Electron Devices, Washington, D. C., October 26, 1956.

Control of Positive Ions

Our experience with previous TWT's has indicated that an improvement in life by as much as a factor of ten is obtained by arranging the dc potential profile so that positive ion bombardment of the cathode is minimized. This improvement has been observed even in tubes in which all reasonable steps have been taken toward minimizing the residual gas pressure. From Table I it is seen that the relative values of accelerator, helix, and collector voltage are arranged to drain positive ions formed in the helix region toward the collector. These ions are thereby kept from reaching the cathode. Spurious ion modulation which can result from accumulation of ions in the helix is also prevented.¹

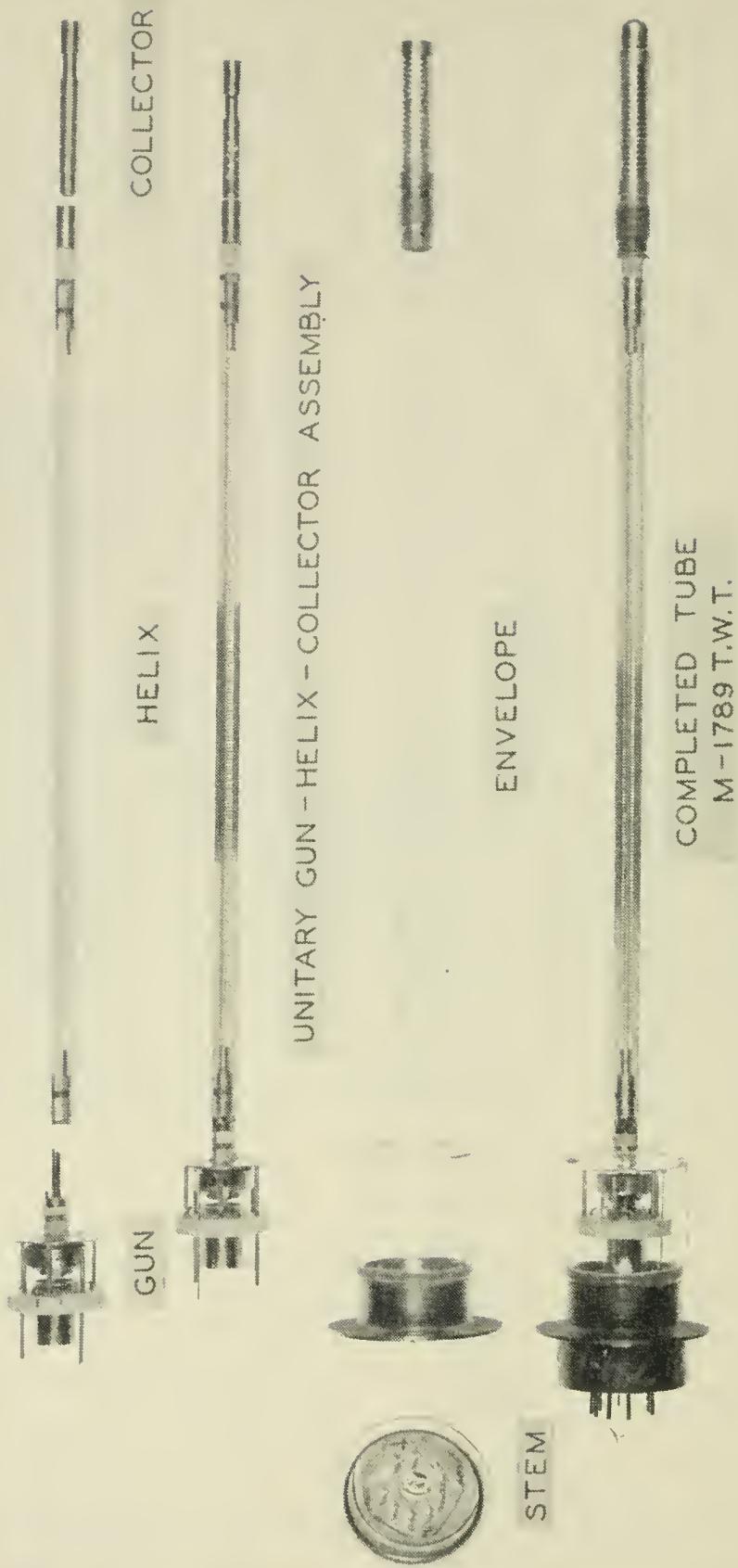


Fig. 3 — The M1789 TWT and its subassemblies. In assembling the tube, the gun is first connected to the helix-collector subassembly. This unit is then inserted into the envelope and an rf braze is made between the inner collector cylinder which is part of the helix assembly and the cooling fins which are part of the envelope. This braze extends for the entire length of the fins. Finally, the stem leads are connected to the gun and a second rf braze is made between the stem and the envelope.

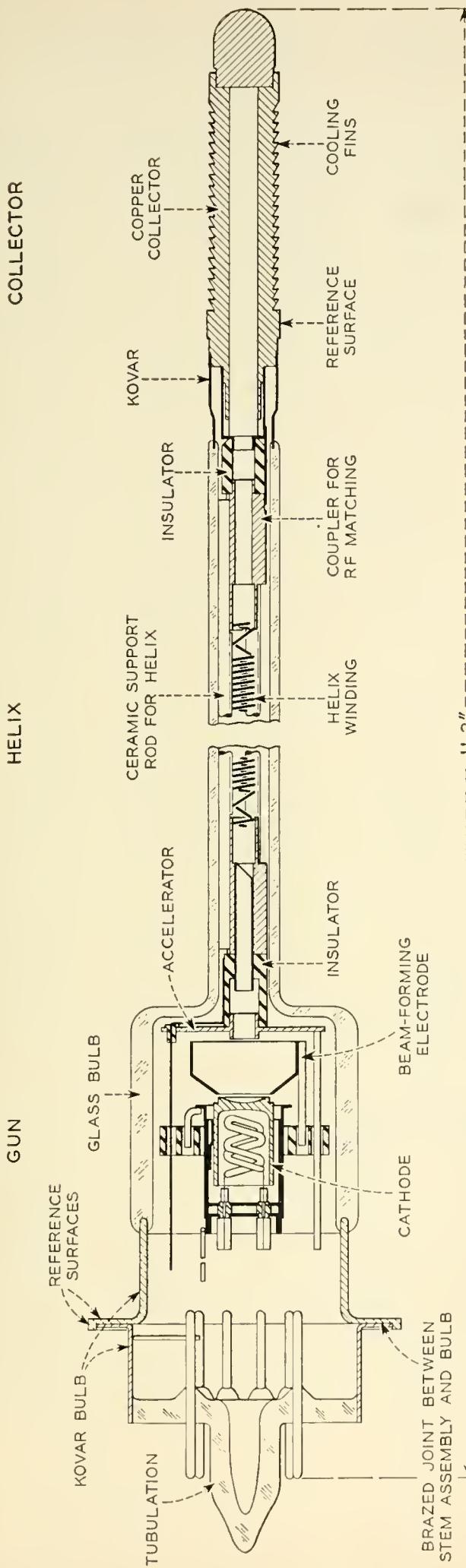


Fig. 4 — Simplified layout of the MI1789. Detailed drawings of the various parts of the tube are shown in Figs. 8, 14 and 21. The alignment surfaces are provided for mounting the TWT with respect to its associated magnetic circuit as is shown in Fig. 5. These surfaces are accurately concentric with the gun helix-axis. This is accomplished by shrinking the envelope in the helix region onto a precision mandrel and then grinding the surfaces concentric with this mandrel. The helix assembly is made a close fit inside of the glass envelope (less than two mils clearance), thereby making the helix axis concentric with the alignment surfaces. The gun is aligned with respect to the helix by telescoping cylinders which are held to a clearance of less than one mil. The ceramic insulators at the ends of the helix provide rf isolation of the helix from the gun and collector. These insulators have a larger inside diameter than do adjacent metal parts to prevent them from charging as a result of electron bombardment. We have not observed any effects in the MI1789 which could be attributed to charging of these ceramics.

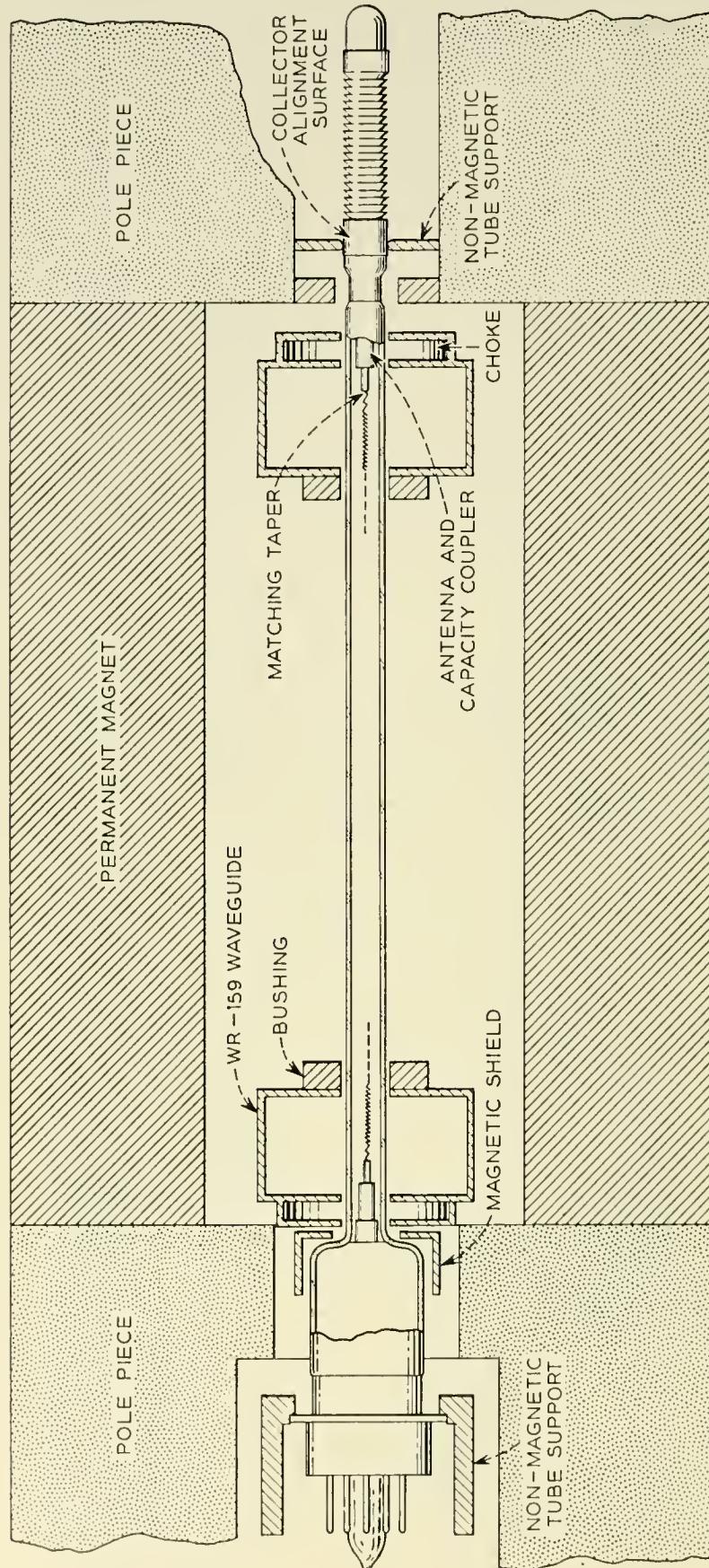


FIG. 5.—Schematic drawing of relationship of TWT to magnetic circuit and to waveguides. Mounting the gun and collector inside of holes through the pole pieces shields them from the magnetic focusing field. The helix passes through the center of the broad face of the waveguide. Energy is coupled between helix and waveguide by means of antennas and of matching tapers (see Fig. 14) at the ends of the helix. A shorting plunger is located in the waveguide about $\frac{1}{8}$ wavelength behind the TWT. The diameter of the cooling fins on the collector end is such that they can pass through the holes in the waveguide when the tube is inserted into the circuit. Forced air cooling of the collector is employed.

The effect that ions can have on cathode life was clearly demonstrated in a TWT which was in many aspects a prototype of the MI789. This tube operated with the accelerator, helix and collector at successively higher voltages, with consequent ion draining toward the cathode. Severe ion bombardment of the cathode brought about failure of most of these tubes in from 500 to 2,000 hours. In contrast to this the average life of the M1789 is in excess of 10,000 hours in spite of a cathode current density about twice that in the prototype tube. Moreover, failure of the M1789 comes about from exhaustion of coating material rather than as a result of ion bombardment. During the course of the work of the prototype tube, an experiment was performed to determine how much the ion bombardment would be affected by changing the potential difference between tube electrodes. In this experiment a small hole was drilled in the center of the cathode and an ion current monitoring electrode placed behind it. The ion monitor current was then investigated as a function of electrode voltages. Fig. 6 shows the results. We see that comparatively small potential differences are adequate to control the flow of positive ions.

3.2 *The Electron Gun and Electron Beam Focusing*

The electron gun used in the MI789 is a converging Pierce gun. The values of the gun parameters are summarized in Table I. Included are both the original parameters introduced by Pierce as well as those defined in a recent paper by Danielson, Rosenfeld and Saloom² in which the effects of thermal velocities are considered. Fig. 7 shows a drawing of the electrically significant contours of the MI789 gun. Fig. 8 shows the completed electron gun assembly. The method of constructing the gun is a modification of a procedure used in oscilloscope and television picture tubes. The electrodes are drawn parts made of molybdenum or, in the case of the cathode, of nickel. They are supported by rods which are in turn supported from a ceramic platform to which these rods are glazed. The whole gun structure is supported from the end of the helix by the helix connector detail. Since this part must operate at helix potential, it is insulated from the remainder of the gun by a ceramic cylinder which is glazed both to it and to the accelerator.

To obtain good focusing, the cathode must be accurately aligned with respect to the other electrodes. However, it must be omitted from the gun during the glazing process and during a subsequent vacuum outgassing because the cathode coating cannot withstand the temperatures involved. To insure proper placement of the cathode in the gun assembly

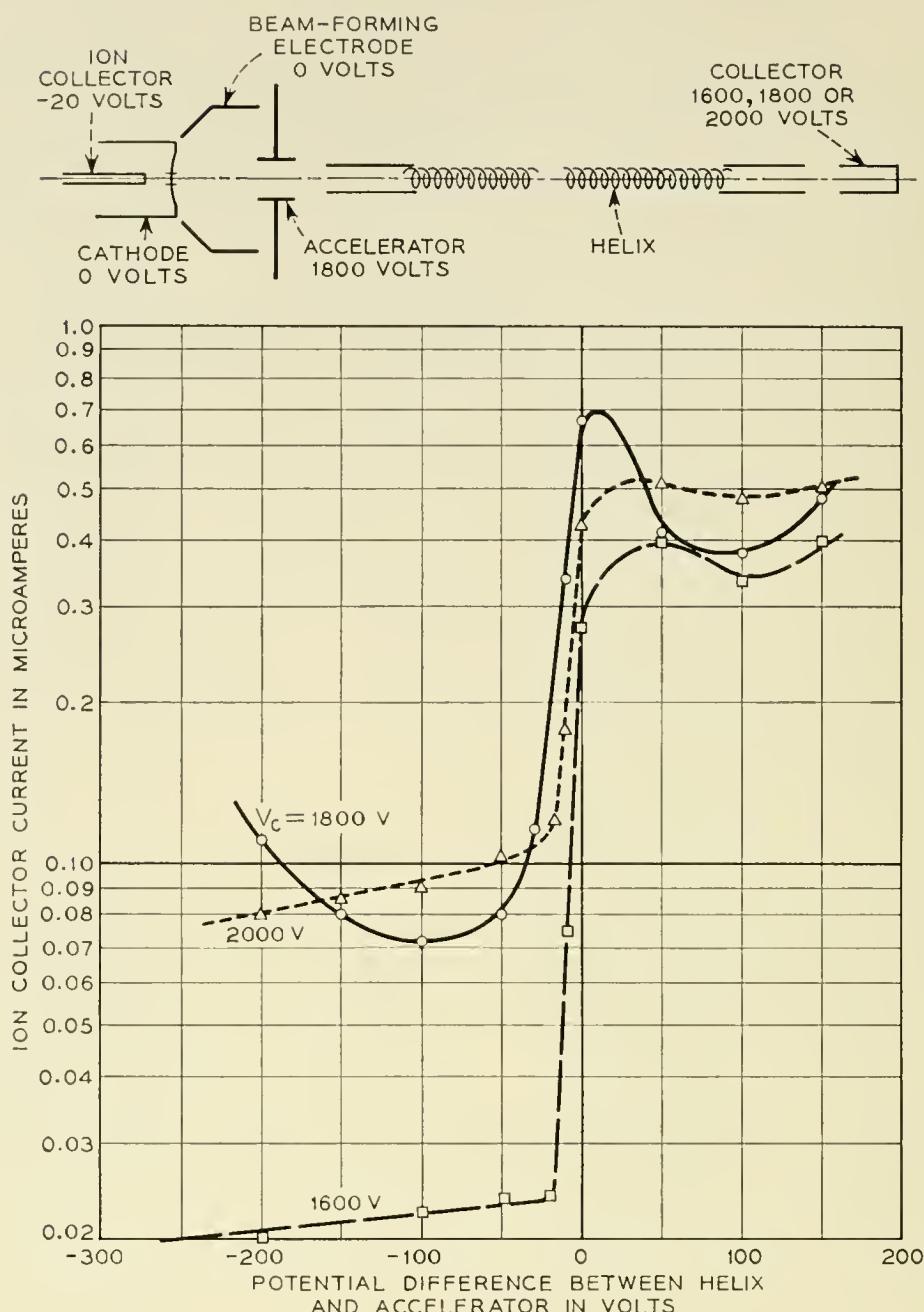


Fig. 6 — Effect of electrode voltages on ion bombardment of the cathode in a prototype of the M1789. In this experiment the helix voltage was varied while the positive ion current to a monitor electrode behind a hole in the cathode was measured. Curves are shown for the collector voltage greater than, equal to, and less than the accelerator voltage. During this experiment the accelerator voltage was held constant at 1800 volts with a resulting beam current of 40 ma. The experiment was performed on a continuously pumped system with the pressure maintained at 2×10^{-7} mm Hg. The helix ID was 80 mils, the cathode diameter 300 mils, and the cathode hole diameter 20 mils. These curves show that the ion bombardment of the cathode can be reduced by as much as a factor of 20 by properly arranging the voltage profile.

at a later stage, an alignment cylinder is included in the gun at the time of glazing (outer cathode alignment cylinder in Fig. 8). When the gun is ready to receive the cathode, the subassembly shown in Fig. 9 is slid into the outer alignment cylinder. The cathode to beam forming electrode spacing is set using a toolmakers microscope, and welds are made between the inner and outer alignment cylinders.

Initially, we thought that the cathode should be completely shielded from the magnetic field, and that the field should be introduced in the region between the accelerator and the point at which the beam would reach its minimum diameter in the absence of magnetic field. This ar-

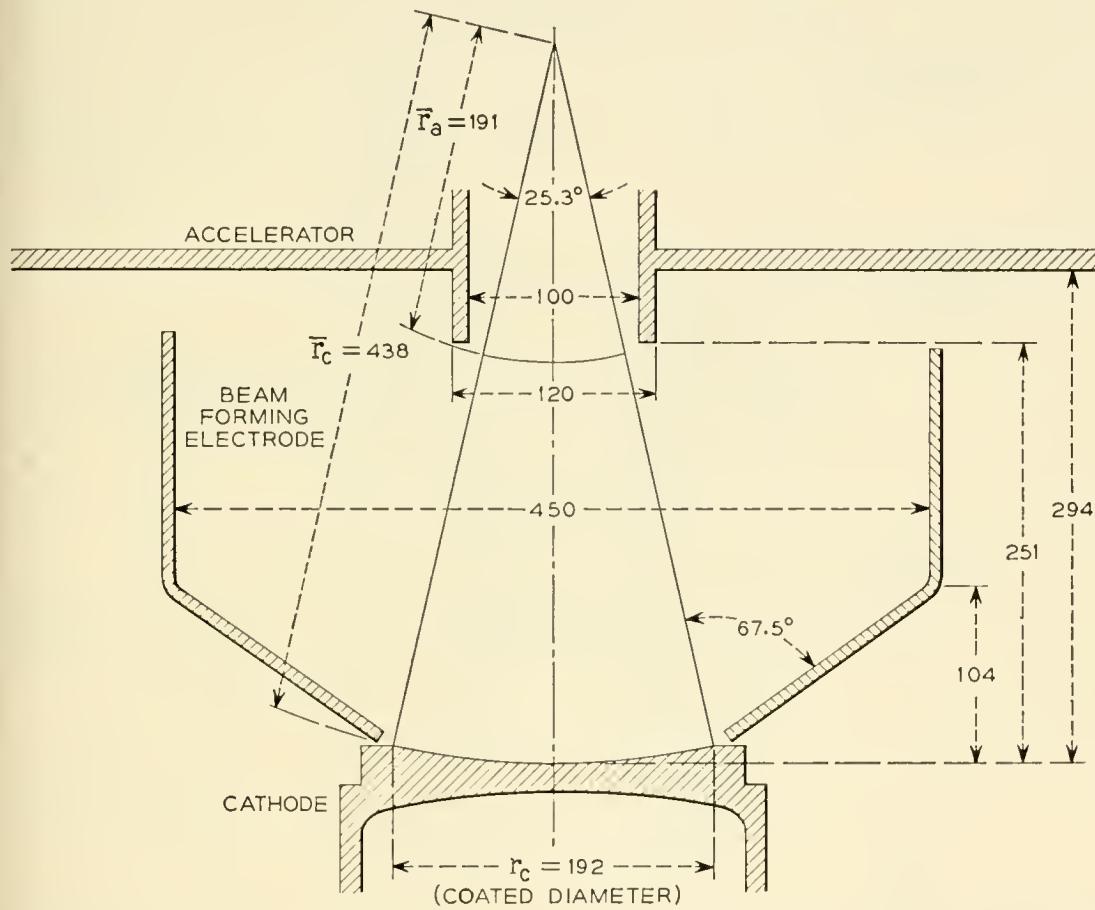


Fig. 7 — The electrically significant contours of the M1789 gun. All dimensions are in mils. These contours were determined using an electrolytic tank and following the procedure originated by Pierce. The measured potential at the beam boundary in the tank was made to match the calculated value within $\pm \frac{1}{4}$ per cent of the accelerator voltage to within 10 mils of the anode plane. The aperture in the accelerator was made sufficiently large so that substantially no beam current is intercepted on it. The significant parameters of this gun are:

P	$= 0.3 \times 10^{-6}$ amps/volts $^{3/2}$	$r_e/\sigma = 3.50$	At the beam minimum in absence of magnetic field
\bar{r}_c/r_a	$= 2.30$	$\sigma = 4.80$ mils	
θ	$= 12.67^\circ$	$r_{95} = 20.5$ mils	
$\sqrt{V_A/T_k}$	$= 1.61 (T_k = 720^\circ\text{C})$	$J = 213 \text{ ma/cm}^2$	

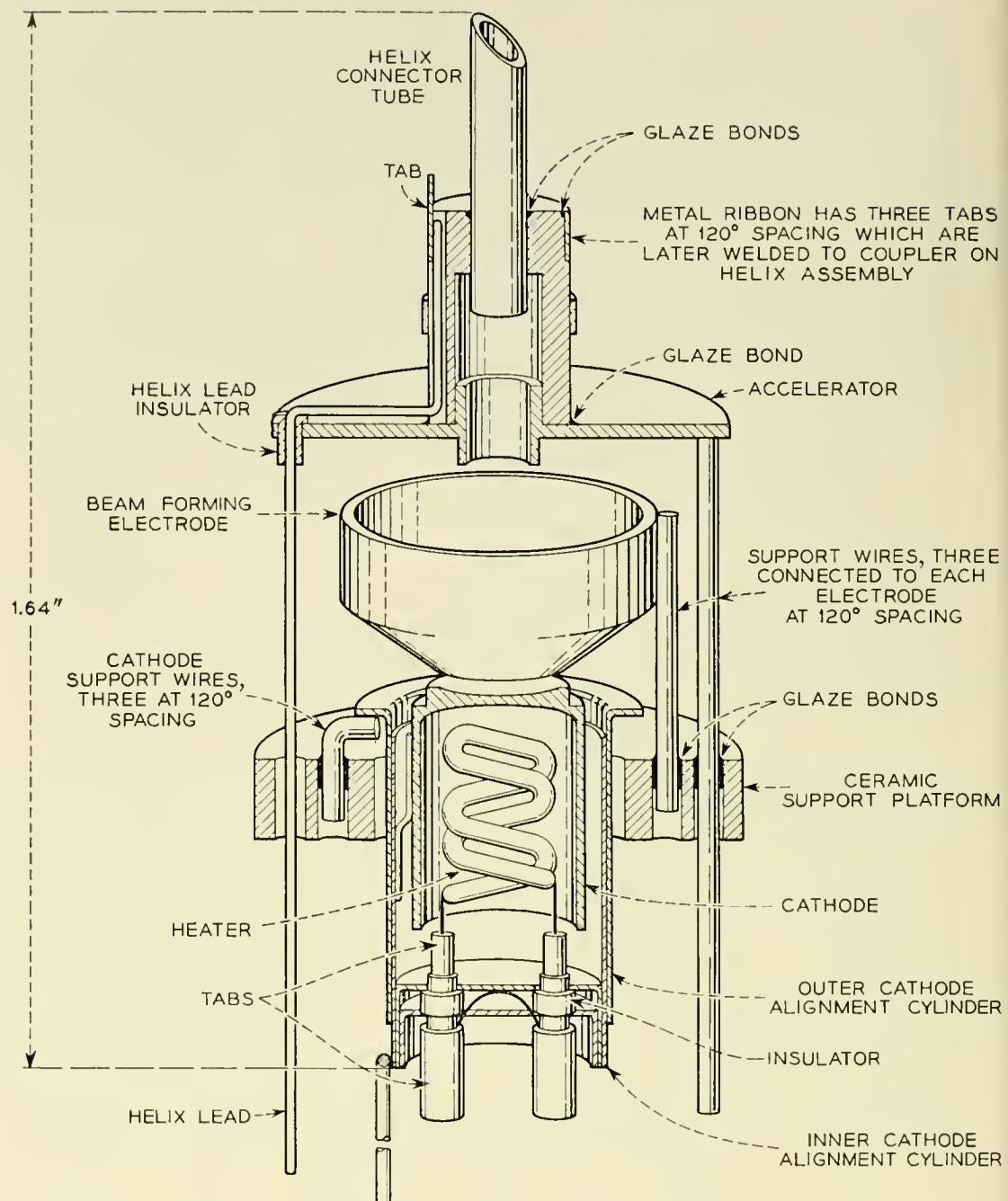


Fig. 8 — M1789 electron gun assembly. In constructing the gun, all the parts with the exception of the cathode, heater, and inner support cylinder are mounted on a mandrel which fixes their relative positions. Glass powder is applied to the areas where glazed joints are desired. The unit is then heated in forming gas (85% N₂, 15% H₂) to 1100°C to melt the glass and form the glazed bonds. With this technique the precision required for alignment and spacing of the electrodes resides entirely in the tools. The helix connector tube later slides into the coupler detail of Fig. 14 to align gun and helix assemblies. The inner and outer cathode alignment cylinders are welded together at two points at the end remote from the cathode. Optical comparator inspection shows that the significant dimensions of these guns are held to a tolerance of less than ± 2 mils.

angement did result in the best beam transmission to the collector. We later discovered, however, that the noise on the electron stream became extremely high when there was no magnetic flux at the cathode. This effect will be discussed further in Section IV. We found that by having a flux density of about 20 gauss at the cathode, the noise figure could be considerably reduced with the only penalty being a slight increase in interception on the helix. The penalty results from the fact that the flux linking the cathode causes a reduction in the angular velocity of the electrons in the helix region (from Busch's theorem), and this in turn diminishes the magnetic focusing force.

Fig. 10 shows the distribution of axial magnetic field in the gun region. The curve represents a compromise between that which gives best focusing (zero flux density at the cathode) and that which gives best noise performance (about 25 gauss flux density at the cathode). This flux density variation was arrived at by empirical methods.

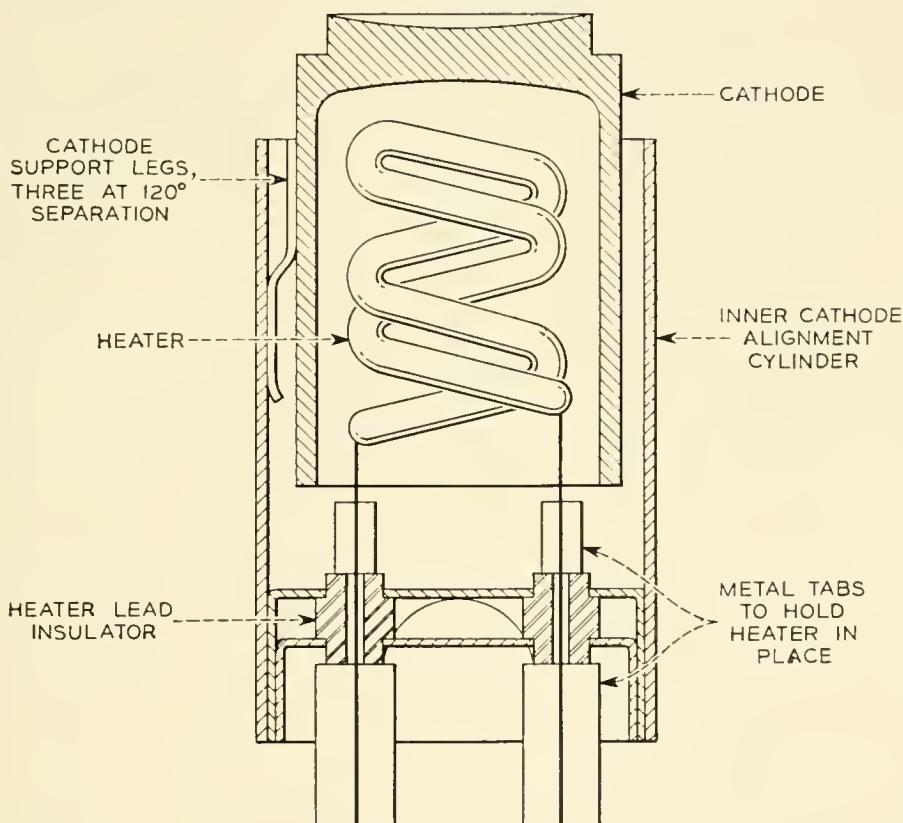


Fig. 9 — The cathode subassembly. In this unit the cathode is connected to the inner alignment cylinder by three legs. These legs are first welded to the cathode and then oven brazed to the alignment cylinder. During the brazing, a jig holds the cathode accurately concentric with this cylinder. The cathode is then coated and the unit is ready for assembly into the gun. The heater power required to raise the cathode to its operating temperature of 720°C is about six watts.

Measurements of beam interception as a function of magnetic flux density are shown for several beam currents in Fig. 11. These measurements were obtained without any rf input to the TWT. An interesting way of normalizing these data is shown in Fig. 12. Here the magnetic

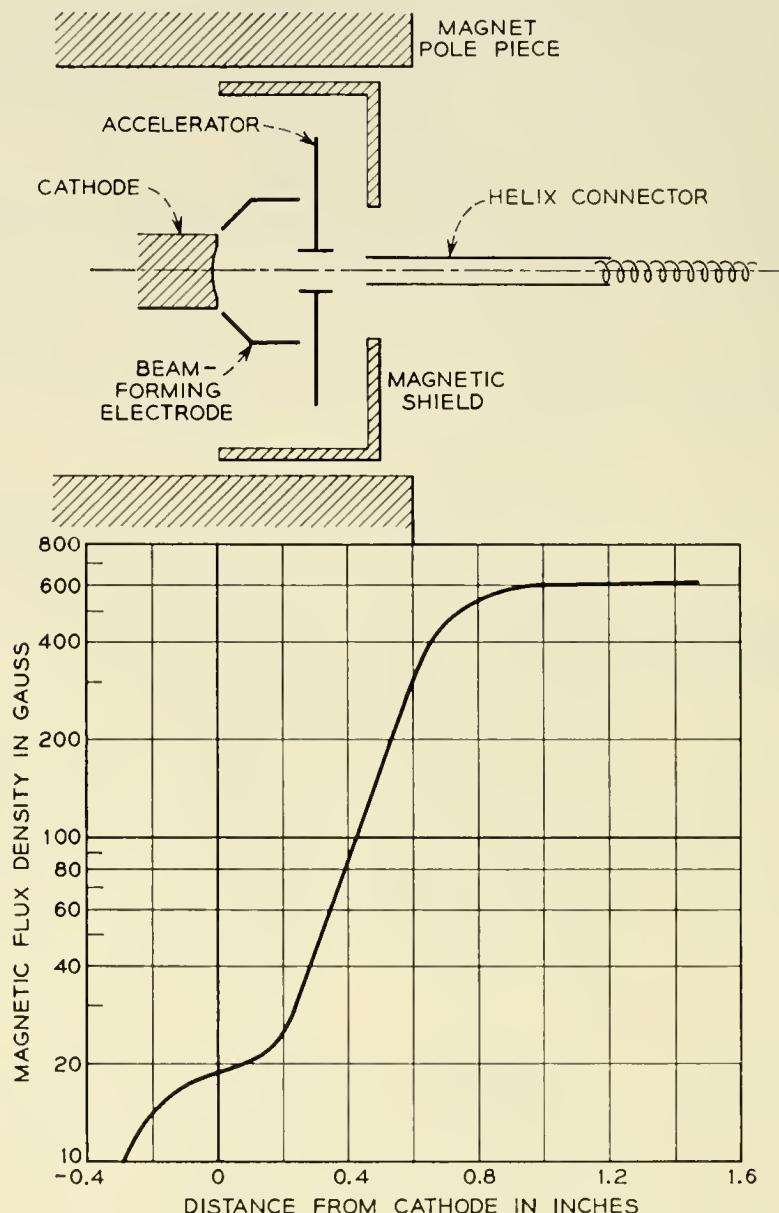


Fig. 10 — Variation in magnetic flux density as a function of distance from the cathode. A schematic representation of the gun electrodes and of the magnetic parts which have been used to control the flux is also shown. All the elements inside the tube are non-magnetic so that the flux density variation is determined entirely by magnetic parts external to the tube envelope. The flux density at the cathode is built up (i.e., the step is put into the curve) by having the magnetic shield end near the cathode. The flux which leaves the shield at this point increases the flux density at the cathode over what it would be if the shield extended well behind the cathode.

flux density has been divided by the Brillouin flux density for a beam entirely filling the helix. This quantity is the minimum flux density which could theoretically be used to focus the beam. This normalization tends to bring all of the curves together. Thus we see that, although the conditions in the MI789 are far from those of ideal Brillouin flow (because of transverse thermal velocities, aberrations in the gun, and magnetic field at the cathode), the concept of the Brillouin flux density still retains meaning, i.e., it appears that the flux density required maintains a fixed ratio to the Brillouin value.

Applying sufficient rf input to the MI789 to drive it into non-linear operation, results in defocusing caused by the high rf fields (both from the helix wave and from space charge) near its output end. Fig. 13 shows how the beam interception for different magnetic flux densities varies as a function of the power output of the TWT. From these curves we see that an output level of five watts can be maintained with about one per cent interception with a flux density of 600 gauss.

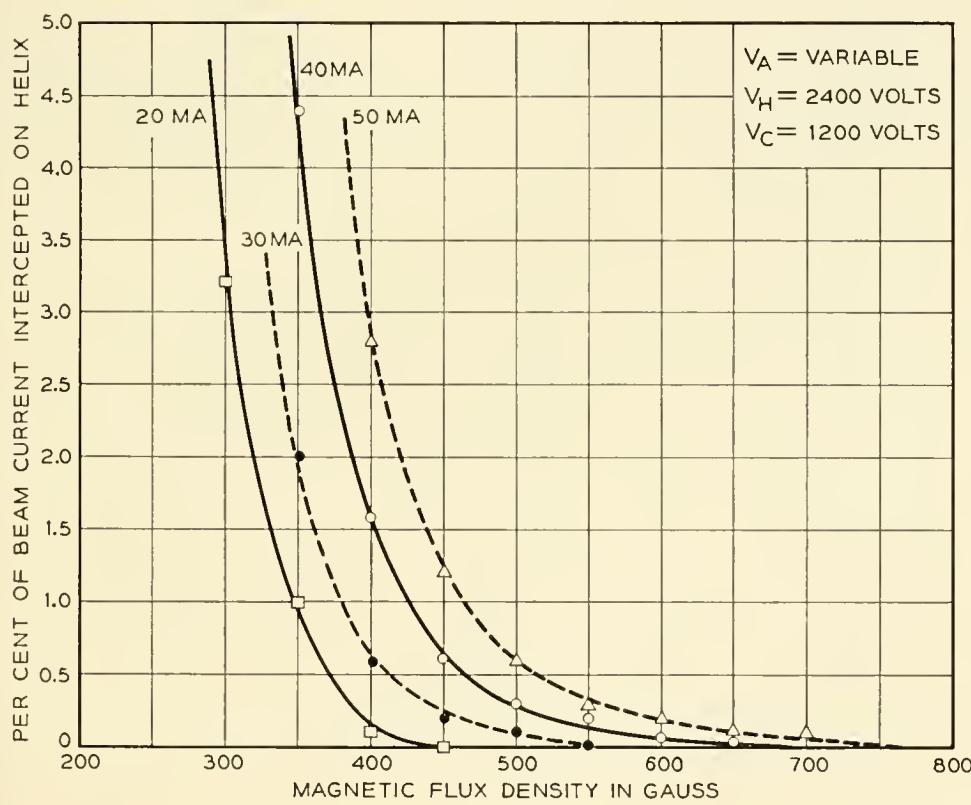


Fig. 11 — Per cent interception on the helix as a function of magnetic flux density. These measurements were taken using a precision solenoid to focus the TWT. The component of field perpendicular to the TWT axis was less than 0.1 per cent of the longitudinal field. During these measurements there was no rf input to the TWT and there was substantially no (<0.1 ma) interception on the accelerator electrode.

3.3 The Helix

The MI789 helix assembly is a rigid self-supporting structure composed of three ceramic support rods bonded with glaze to the helix winding. A drawing of the helix assembly is shown in Fig. 14. The support rods are made from Bell Laboratories F-66 steatite ceramic. This material was chosen because of its low rf losses and because these losses do not increase rapidly with temperature. Fig. 15 shows an enlarged photograph of the glaze bonds between the winding and one of the support rods. Attenuation is applied over a length of two inches starting $1\frac{1}{2}$ inches from the input end by spraying the helix assembly with aquadag (carbon in water suspension) and then baking it.

Supporting the winding by glazing it to ceramic support rods has the following advantages:

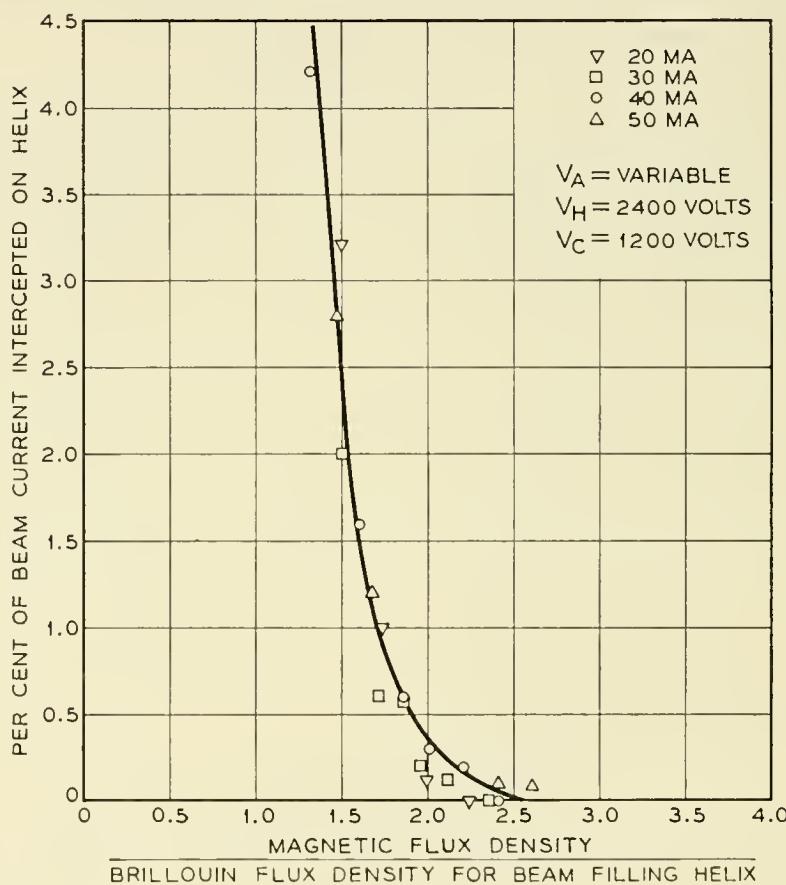


Fig. 12 — The measurements of Fig. 11 normalized in terms of the Brillouin flux density for a beam entirely filling the helix. The fact that the curves tend to come together indicates that the concept of the Brillouin flux density retains some meaning in the MI789. Because of the additional defocusing effects encountered when the MI789 is driven to high output levels, the tube is usually used with about 2.6 times the Brillouin flux density.

(1) The dielectric loading and intrinsic attenuation of the helix are comparatively low because the amount of supporting structure in the rf fields is small.

(2) High loss per unit length in the helix attenuator is made possible. The reason for this will be discussed further below.

(3) The heat dissipation capability of the helix is greatly increased because the glaze provides an intimate thermal contact between winding and support rods. This is illustrated by Fig. 16 which compares the heat dissipation properties of glazed and non-glazed helices.

(4) Mechanical rigidity is realized and therefore the helix can be handled without risk of disturbing the pitch or diameter of the winding.

On the other hand, use of the ceramic rods in the MI789 has a significant disadvantage in that it makes the outside radius of the vacuum envelope large compared to the helix radius, thus making coupled helix matching out of the question. However, since the MI789 is required to match over less than a 10 per cent band, this is not particularly serious.

To obtain reproducibility of performance in the MI789, the helix must be precisely constructed. Together, the pitch of the helix and the amount of dielectric loading determine the synchronous voltage. A pitch variation of ± 1 per cent results in a voltage variation of about ± 50 volts, and a loading variation of ± 1 per cent results in a variation

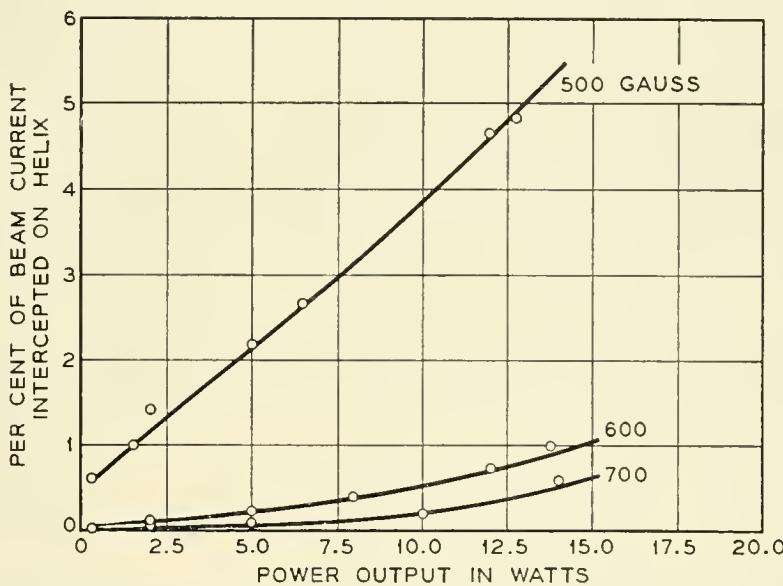


Fig. 13 — Per cent interception on the helix as a function of rf power output. These measurements were made using permanent magnet circuits charged to different field strengths. The magnetic field variation as a function of distance from the cathode was as shown in Fig. 10. The component of magnetic field perpendicular to the tube axis in these circuits was less than 0.2 per cent of the longitudinal field. All measurements were taken with a beam current of 40 ma and with the helix voltage adjusted to maximize the power output.

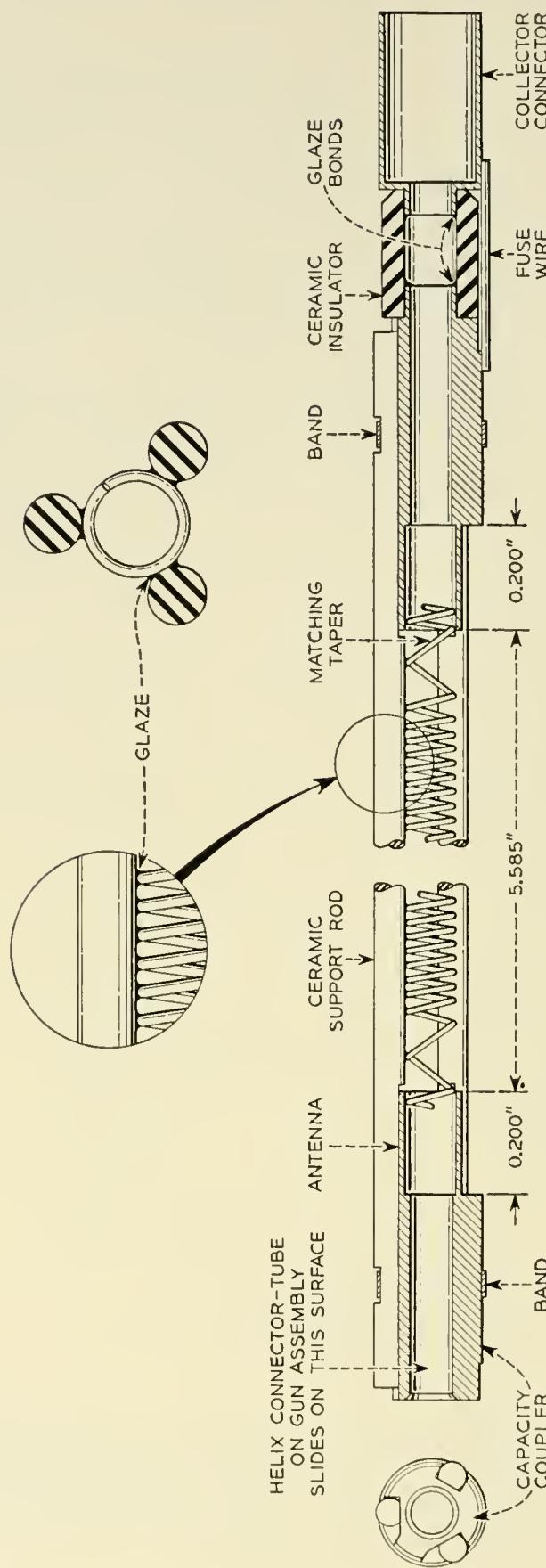


Fig. 14 — Helix assembly. The winding is supported from three F-66 ceramic rods to which it is glazed. On each end of the winding there are two turns with greater than nominal pitch to assist in transferring energy between helix and waveguides. The bands around the assembly are for the purpose of holding the support rods against the capacity couplers. There is no glaze in this region. The relationship of the antenna and capacity coupler to the waveguide circuit is shown in Fig. 5. The collector is later brazed inside of the collector connector cylinder. The fuse wire allows the helix to be heated on the pump station by passing current from the helix pin on the tube base to the collector. After outgassing the helix, the fuse is blown to isolate helix from collector. Before adding the helix attenuation, the rf loss of the helix is 3.6 ± 0.2 db. After adding the attenuation, it is 65 ± 80 db. The synchronous voltage is 2200 ± 50 volts.

of about ± 25 volts. It is not difficult to hold the average pitch variations to less than ± 1 per cent. The loading, however, is a more difficult problem for not only must the dielectric properties of the support rods and of the glaze material be closely controlled, but attention must also be paid to the size and density of the glaze fillets. The gain of the tube is affected by the amount of loss in the helix attenuator. For the particular loss distribution used in the MI789 a variation of ± 5 db out of a total attenuation of 70 db results in a gain variation of about ± 1 db. The helix attenuator depends to a large extent on a conducting "bridge" between helix turns and therefore the amount of attenuation is sensitive to the size and the surface condition of the glaze fillets. Thus, the glazing process must be in good control in order to minimize variations in both gain and operating voltage. With our present techniques, we are able to hold the voltage for maximum gain to within ± 50 volts of the nominal value. The gain is held to ± 2 db — about half of the spread we believe to be caused by variations in loss distribution and about half by differences in beam size.

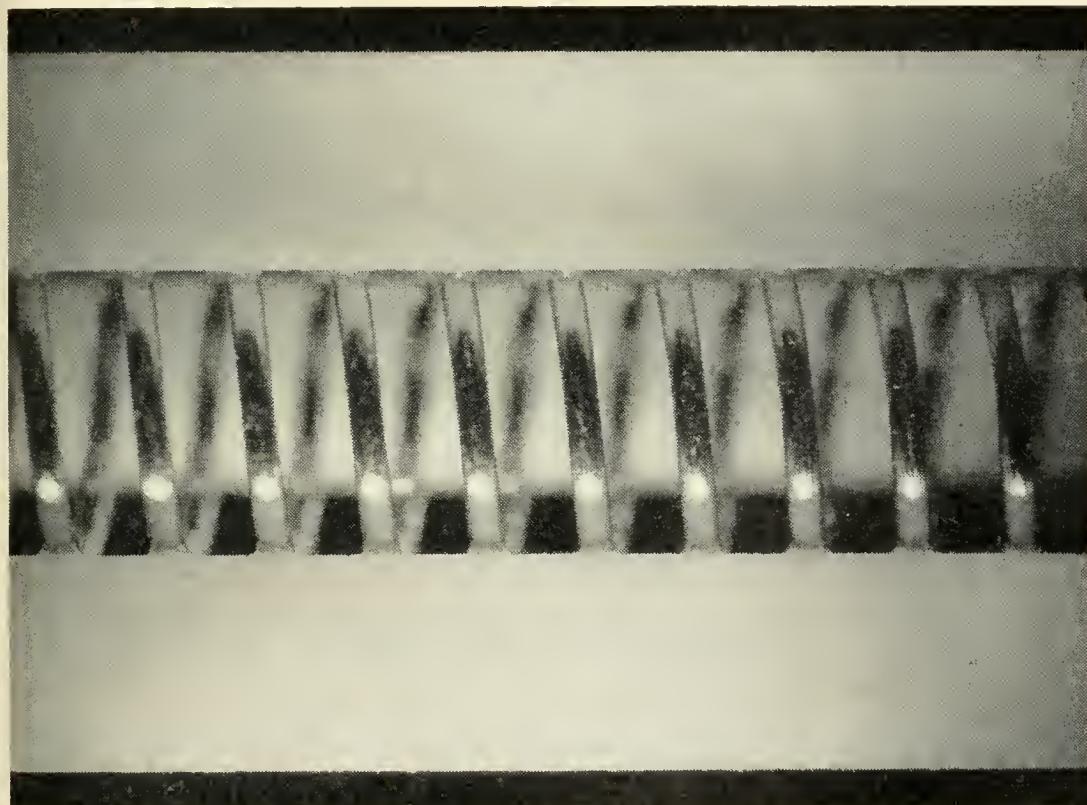


Fig. 15 — Enlarged photograph of part of an M1789 helix. Two of the ceramic support rods can be seen. The other is directly opposite the camera behind the helix and is out of focus. The fillets of glaze which bind the helix to the rods can be seen along the upper rod. This section of helix was free from applied loss.

Helix-to-Waveguide Matching

In the helix-to-waveguide transducer the helix passes through the center of the broad face of the waveguide and energy is coupled between helix and waveguide by an antenna and matching taper. A capacitive coupler on the helix and an rf choke on the waveguide place an effective ground plane at the waveguide end of the antenna. The rf choke also assists in minimizing leakage of rf power. Details of this transducer are shown in Figs. 5 and 14.

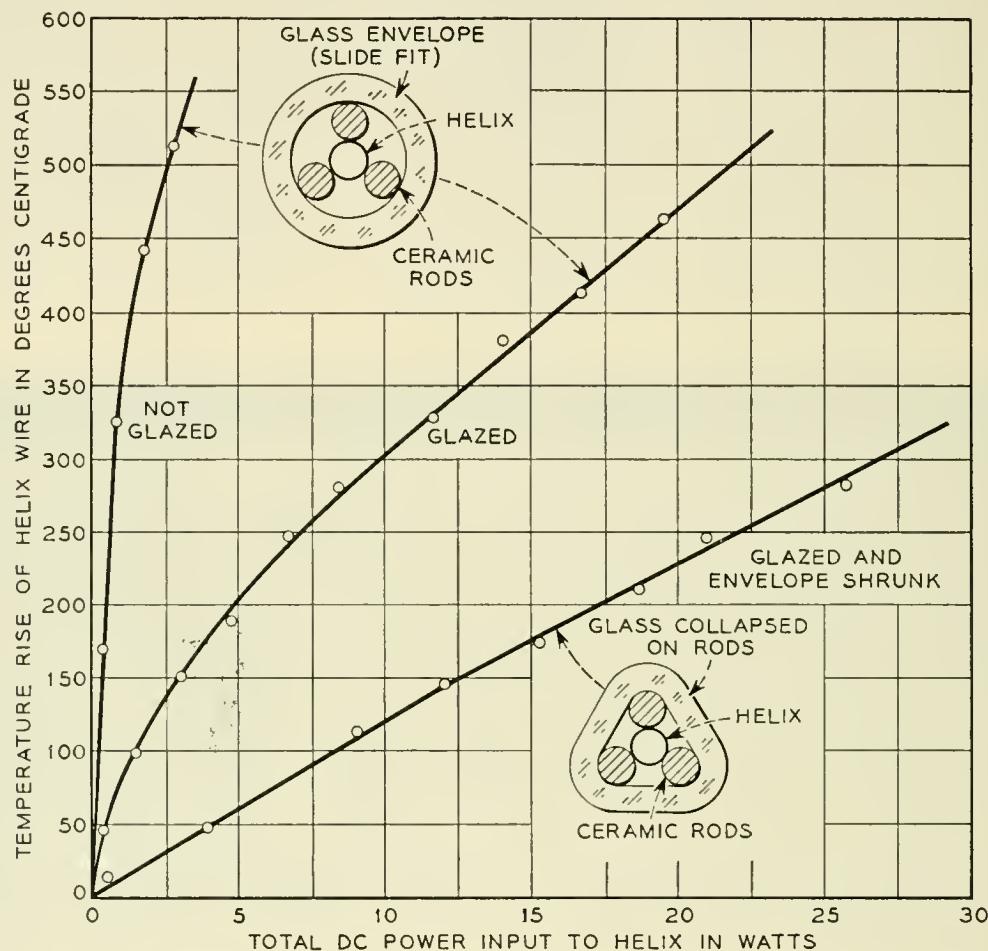


Fig. 16 — Comparison of heat dissipation properties of different helix structures. In this experiment, the helices were heated by passing dc current through them while they were mounted in a vacuum. The temperature was determined from the change in helix wire resistance.

Along with the results for glazed and non-glazed helices in a normal round envelope, this figure shows results on a structure consisting of a glazed helix in an envelope which has been shrunk around the helix support rods. This technique produces a structure which, by virtue of the good thermal contact between the support rods and the envelope, can dissipate more power than the conventional structure. The additional complication of shrinking the envelope is not necessary for the power levels used in the M1789. However, this method could be used if it were necessary to extend the tube's output range to higher power levels.

The dimensions of this transducer were determined empirically. It was found that the antenna length affects mainly the conductive component of the admittance referred to the plane of the helix. The length of the matching taper affects mainly the susceptive component, and the distance from helix to a shorting plunger, which closes off one end of the waveguide, affects both components. If for each tube, the position of the waveguides along the axis of the TWT and the position of the shorting plunger are optimized, the VSWR of the transducers will be less than 1.1 (~ 26 db return loss) over the entire 500-me frequency band. With these positions fixed at their best average value, the VSWR will be less than about 1.3 (~ 18 db return loss).

Internal Reflections

A problem that has required considerable effort has been that of "internal reflections." By this we mean reflections of the rf signal from various points along the helix as contrasted with reflections from helix-to-waveguide transducers. The principal sources of internal reflections are the edge of the helix attenuator and small variations in pitch along the helix. In the MI789 the pitch variations are the main source of difficulty.

The type of performance degradation caused by small internal reflections can be illustrated by the following. Consider a signal incident on the TWT output as a result of a reflection from a radio relay antenna. Except for a small reflection at the transducer, energy incident on the TWT output will be transferred to the helix, propagated back toward the input, and for the most part be absorbed in the helix attenuator. However, if there are reflection points along the helix, reflected signals will be returned to the output having been amplified in the process by the TWT interaction. Because of this amplification, even a small reflection of the backward traveling wave can result in a large reflected signal at the TWT output. In the MI789, these amplified internal reflections are considerably larger than the reflection from the output transducer. They limit the overall output VSWR to about 1.4, whereas the transducer alone has a VSWR of about 1.1.

If there is a long length of waveguide between the TWT and the antenna, the echo signal resulting from a reflection at the antenna and a second reflection at the TWT will vary in phase with respect to the primary signal as frequency is changed. This will cause ripples in both the gain and in the phase delay of the system as functions of frequency. Suppose the VSWR of the antenna is 1.2 and that of the TWT is 1.4 and the two are separated by 100 feet of waveguide. The amplitude of

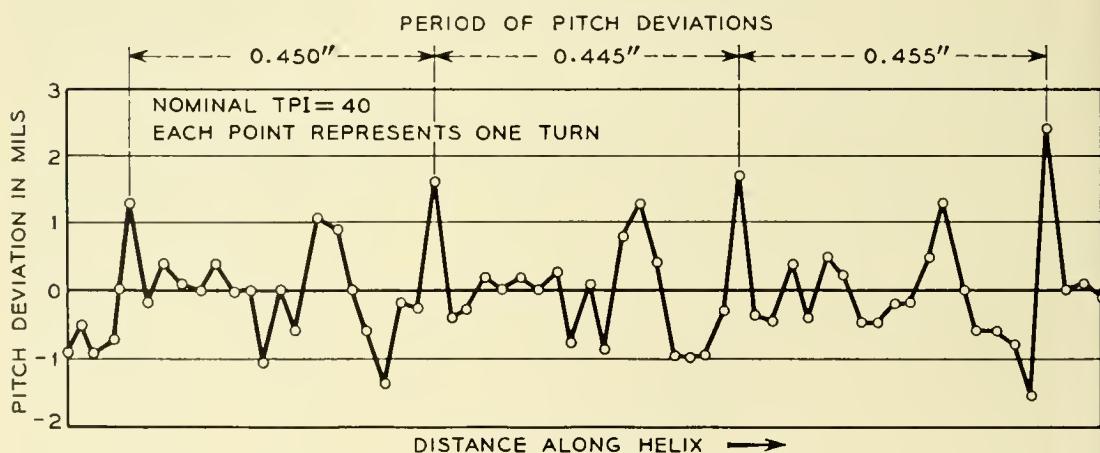
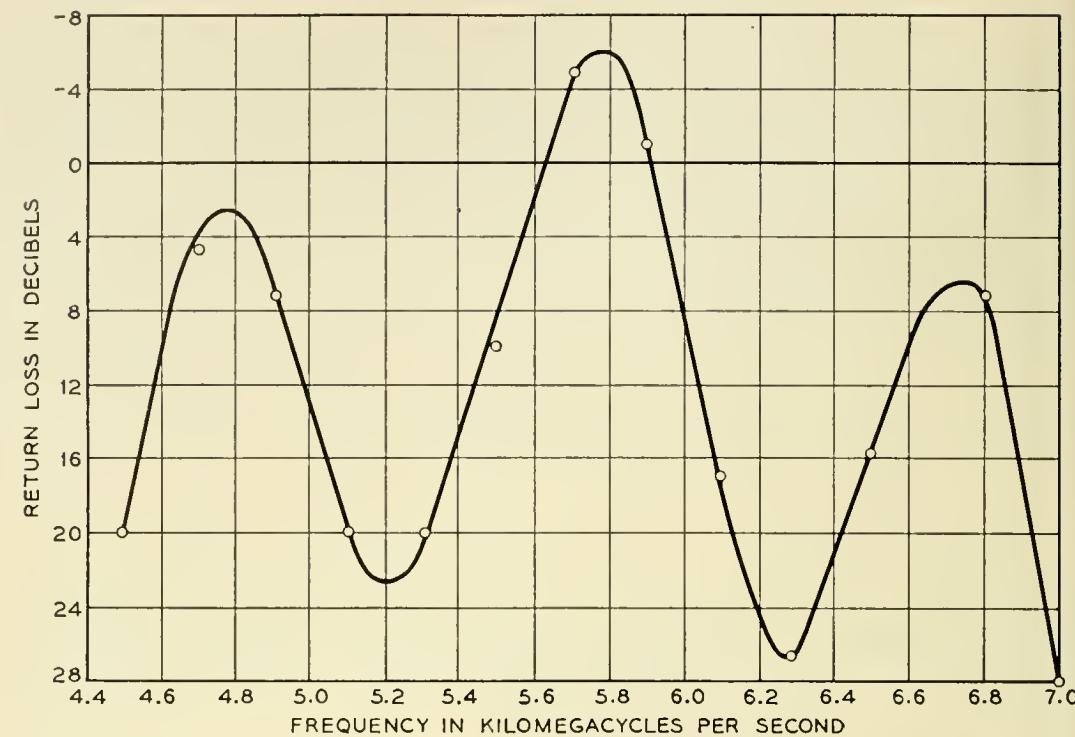


Fig. 17 — Pitch deviations and internal reflections in an early M1789 TWT. The ordinate of the pitch deviation curve is the difference between the measured spacing between helix turns and the nominal value, which for this particular helix was 25 mils. (The tube operated at 1,600 volts.) Each point represents a helix turn. It is seen that the pitch deviations are periodic in nature, repeating about every 0.450 inch.

The internal reflections were measured by matching the TWT with beam off at each individual frequency with a tuner to a VSWR of less than 1.01 (return loss greater than 40 db). The beam was then turned on and the resulting reflection taken as an approximate measure of the internal reflection. There appeared to be no appreciable change in the helix-to-waveguide transducer reflection as a result of turning the beam on. Evidence for this is the fact that when the beam was turned on with the helix voltage adjusted so that the TWT did not amplify, there was little change in the reflection.

The peaks of the internal reflection curve occur at five, six and seven half wavelengths per period of the helix pitch deviations, indicating that the reflections from each period are adding in phase at these frequencies. At the 5,800-mc peak the return loss is positive. This indicates a reflected signal larger than the incident signal. Shorting the TWT output caused the tube to oscillate at this frequency.

the gain fluctuations will be about 0.25 db, the amplitude of the phase fluctuations will be about 0.9 degree and the periodicity of the fluctuations will be about six mc. This effect may be eliminated by using an isolator between the TWT and the antenna to eliminate the echo signal.

In addition to echo signals that occur between the TWT and the antenna there are echoes which occur wholly within the TWT as a result of a reflection of the signal from the output transducer and a second reflection from some point along the helix. Thus even if a TWT is operating into a matched load it may have ripples in gain or phase characteristics. These ripples may be controlled by minimizing the internal reflections. In the MI789 they are less than ± 0.1 db in gain and one-half degree in phase. Their periodicity is about 100 mc.

In addition to causing transmission distortions, internal reflections can seriously reduce the margin of a TWT against oscillation. Outside of the frequency band of interest, the helix-to-waveguide transducer may be a poor match or the TWT may even be operating into a short circuit in the form of a reflection type bandpass filter. At such frequencies, the internal reflections must not be large enough so that an echo between transducer or filter and an internal reflection point will see any net gain, or else the TWT will oscillate.

With many types of helix winding equipment, variations in helix pitch are periodic in nature. This causes the helix to exhibit a filter-like behavior with respect to internal reflections. At frequencies at which the period of the pitch variations is an integral number of half-wavelengths, the resultant reflections from each individual period will add in phase, thereby causing the helix to be strongly reflecting at these frequencies. This effect can perhaps best be illustrated by considering some results obtained in an early stage of the MI789 development. Fig. 17 shows measurements of the spacing between turns of an early helix. Also shown is the return loss as a function of frequency that a signal incident on the output of an operating TWT would see as a result of internal reflections alone. Helix-to-waveguide transducer reflections were eliminated with waveguide tuners during this experiment. The deviations in helix pitch from nominal are rather large and are markedly periodic in nature. The resulting internal reflections show strong peaks at frequencies corresponding to five, six and seven half-wavelengths per period of the pitch deviations.

In the present M1789 this situation has been considerably improved by increased precision in helix winding and by insuring that the remaining periodicity does not produce a major reflection peak in the band. Fig. 18 shows pitch measurements and internal reflections for a recently constructed tube.

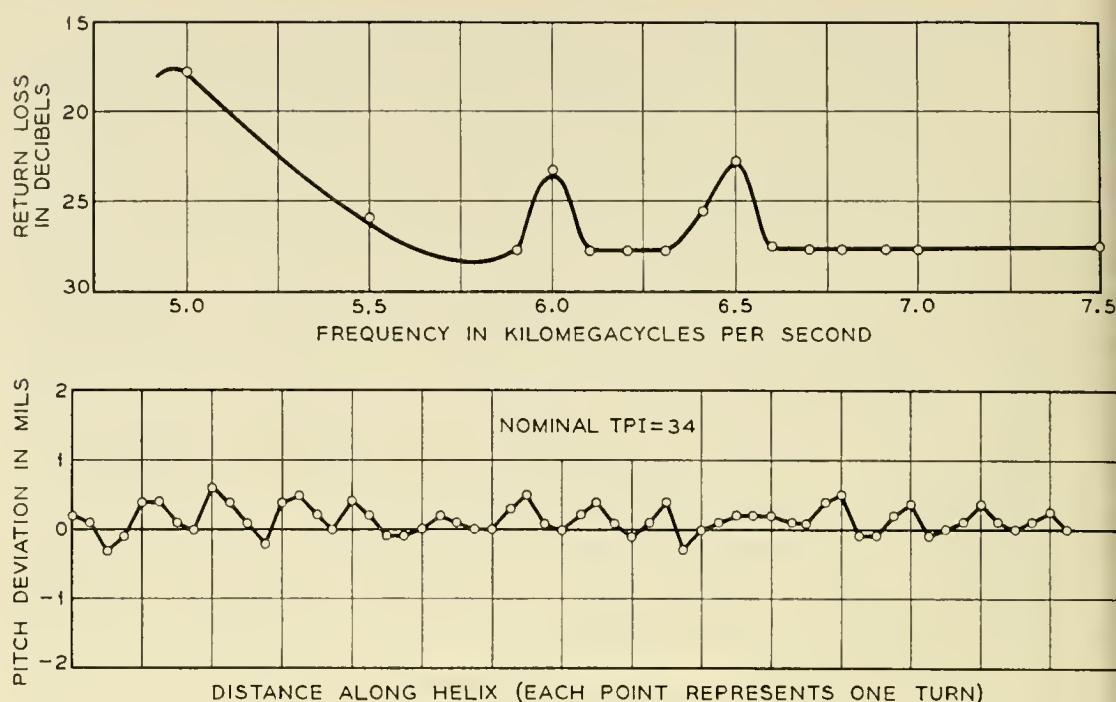


Fig. 18—Pitch deviations and internal reflections in a recent M1789 TWT. By precise helix winding techniques the pitch deviations have been reduced by a factor of about 10 over those occurring in early tubes. The resulting internal reflections have been improved by about 25 db although there is still a residual periodicity remaining.

For return losses greater than about 25 db, we begin to see internal reflections originating from the edge of the helix attenuator. At these values of return loss, the measurements also begin to be in appreciable error as a result of the residual transducer reflections.

Helix Attenuator

Attenuation is applied to the helix by spraying aquadag directly on the helix assembly and then baking it. The result is a deposit of carbon on the ceramic rods and on the glaze fillets. The attenuation is held between 65 and 80 db and is distributed as shown in Fig. 19. Evidently most of the loss is caused by a conducting bridge which is built up between helix turns. This was indicated by one experiment in which we cleaned the deposit off the rods of a helix by rubbing them with emery paper. Only the carbon directly between helix turns then remained. This decreased the total attenuation by less than 20 per cent. Having the helix glazed to the support rods is apparently necessary in order to get good contact between the winding and the carbon "bridge." We have been able to obtain about four times as much loss per unit length with glazed helices as with non-glazed ones. Using our method of applying attenuation we can add in excess of 80 db/inch to a glazed helix. The ability to obtain such high rates of attenuation allows us to concentrate the loss along the helix thereby minimizing the TWT length.

The machine used for spraying aquadag on the helix is shown in Fig.

20. A glass cylinder and photocell arrangement is used to monitor the amount of carbon deposited. In this manner the attenuation added is made independent of both the aquadag mixture and the nozzle setting of the spray gun. This machine has been checked alone by using it to spray glass slides which are then made into attenuator vanes. Over a two-year period we have found that a given light transmission through the monitor slide results in the same vane attenuation within ± 2 db out of 40 db.

After a helix has been sprayed, it is vacuum fired at 800°C for thirty minutes and then the loss is measured. About 60 per cent of the helices fall within the desired range of 65–80 db. The principal cause of the differences in attenuation is believed to be variation in the condition of the glaze fillets. Helices not meeting specifications are sprayed and fired a second time (after cleaning off excess aquadag if necessary). This second treatment, brings the attenuation of almost all helices to within the desired range.

3.4 The Collector

It is desirable to operate the collector at the lowest possible voltage in order to minimize the dc power input to the TWT. This increases the overall efficiency and simplifies the collector cooling problem. On the

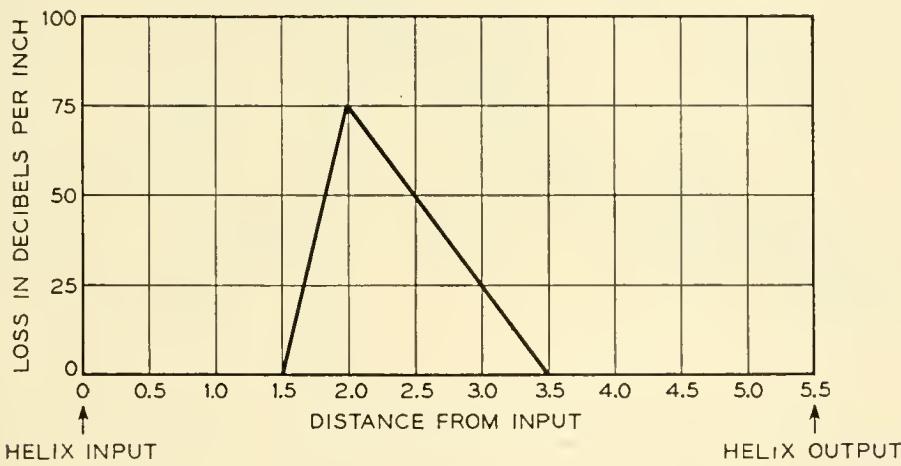


Fig. 19 — Distribution of helix attenuation. The attenuation pattern has a gradually slanting edge facing the output to provide a smooth transition into the loss for any signals traveling backwards toward the input. Reflections of these signals must be very small since the reflected signals will be amplified in the process of returning toward the output. Cold measurements (i.e., measurements on the helix without electron beam) made by moving a sliding termination inside the helix, indicate that the return loss from the attenuator output is better than 45 db, the limiting sensitivity of our measurement. The input side of the helix attenuator is also tapered to minimize reflections but this taper is much sharper than that on the output side because there is comparatively little gain between input and attenuator. Cold measurements with a sliding termination showed a return loss for this taper of about 40 db. (Surprisingly, even a sharp edge produces a reflection with a return loss of almost 30 db.)

other hand, if there is appreciable potential difference between helix and collector, we must insure that few secondary or reflected electrons return from the collector to bombard the helix and accelerator, or else we may overheat these electrodes. Fig. 21 shows a drawing of the collector used in the M1789. It takes the form of a long hollow cylinder shielded from the magnetic field. Inside of the collector the beam is allowed to gradually diverge and the electrons strike the walls at a grazing angle. This design reduces secondary electrons returned from the collector to almost negligible proportions.

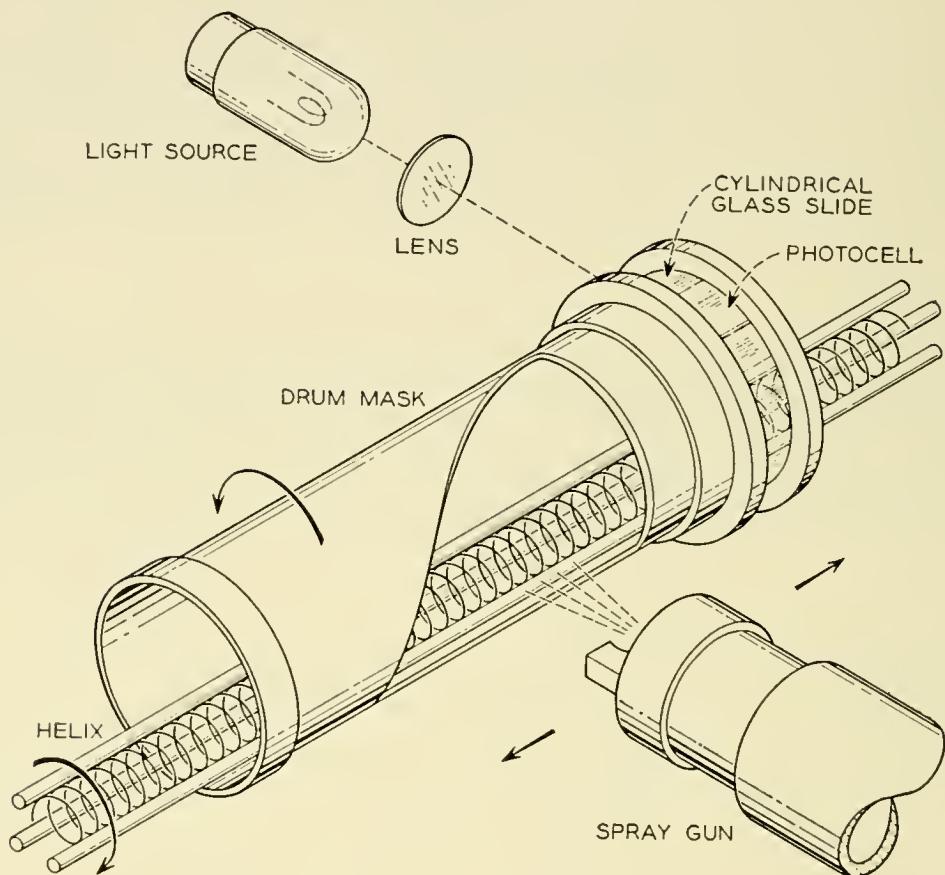


Fig. 20 — Schematic diagram of the machine used for spraying aquadag attenuation on the helix. In this machine the helix is rotated rapidly to insure uniform exposure to the spray. At the same time the masking drum rotates at a slower speed and the spray gun traverses back and forth along the masking drum. The drum therefore acts as a revolving shutter between the helix and the spray gun and its degree of opening serves to control the amount of aquadag reaching the helix. From a knowledge of the rate of attenuation increase as a function of the amount of carbon deposited (empirically determined) the shape of the drum opening can be calculated so as to give any desired attenuation pattern.

The spray gun also passes over a glass cylinder at one end of the masking drum so that it receives a sample of the aquadag spray. A photocell is used to monitor light transmitted through the cylinder. Before starting to spray, the glass is cleaned and the photocell reading is taken as 100 per cent light transmission. The helix is then sprayed until the light transmission has decreased to the proper value. The photoelectric monitoring technique makes the attenuation added insensitive to the aquadag composition and to the spray gun nozzle opening.

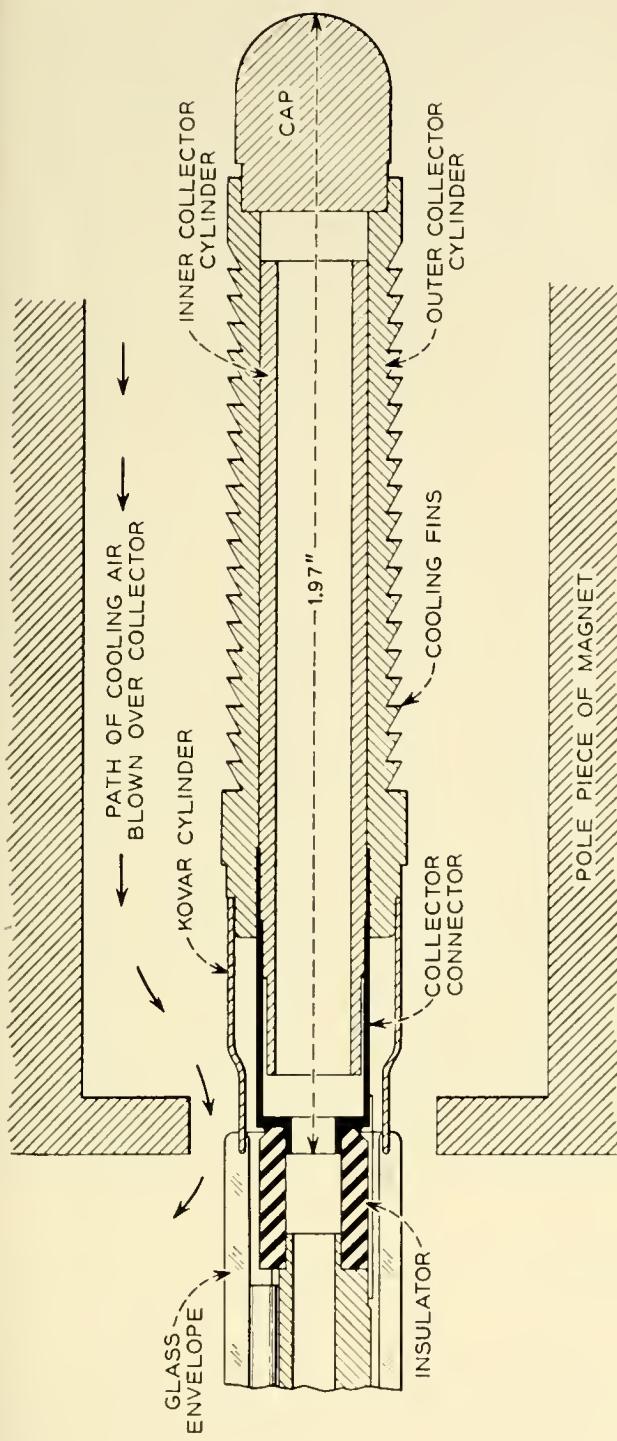


FIG. 21 — The collector consists of two cylinders. The inner cylinder is the electron collector proper and is part of the helix subassembly. The outer cylinder is part of the envelope subassembly. The two parts of the collector are brazed together in the final assembly of the tube. The collector is shown in its position with respect to the pole piece at the output end of the magnetic circuit. The magnetic field variation in the collector region is plotted to the same scale as the collector drawing. The electron beam diverges gradually inside of the collector and the electrons strike the walls at a grazing angle. With this design there are essentially no secondary electrons returned from the collector.

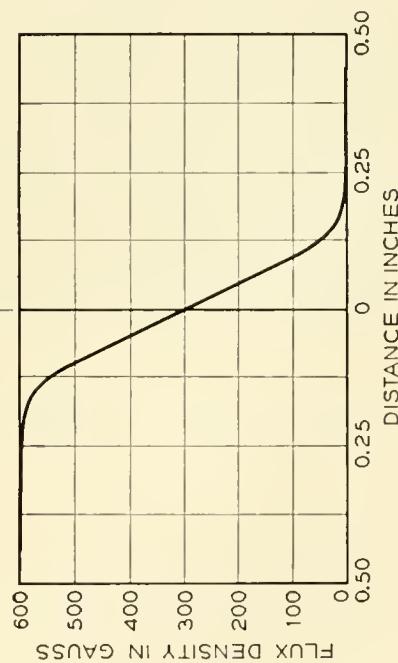


Fig. 22 shows the total accelerator and helix interception as functions of collector voltage at various output levels. When there is no rf drive, the intercepted current remains low to a collector voltage of about 200 volts at which point it suddenly increases to a high value. This appears to be caused by the phenomenon of space charge blocking. As the collector voltage is progressively lowered, the space charge density at the mouth of the collector increases because of the decrease in electron velocity at this point. Increasing the charge density causes the potential depression in the beam to increase until at some collector voltage the potential on the axis is reduced to cathode potential. At collector voltages lower than this, some of the beam is blocked, i.e., it is turned back by the space charge fields.

When the TWT is operated at appreciable rf output levels, the collector voltage must be increased to permit collection of all electrons which have been slowed down by the rf interaction. Unfortunately, some electrons are slowed far more than is the average, so that we must supply to the TWT several times more dc power than we can take from it in the form of rf power. However, as seen from Fig. 22, there is still an appreciable advantage to be gained by operating the collector at lower than helix potential. These curves should not be taken as an accurate measure of the velocity distribution because there are undoubtedly space charge blocking effects which even at higher collector voltages have some influence on the number of electrons returned from the collector. This arises from the fact that the rf interaction causes an axial bunching of the electrons, thereby causing the space charge density in an electron bunch to be much higher than it is in an unmodulated beam. Thus, as a bunch enters the collector, the local space charge density may be high enough to return some electrons.

IV. PERFORMANCE CHARACTERISTICS

4.1 *Method of Approach*

In this section we will consider the overall rf performance of the M1789 and make some comparisons between theory and observed results. The following TWT parameters can be varied: input level; helix voltage; beam current; frequency; and magnetic field. Our approach here will be to first consider the operation of the tube under what might be called nominal conditions. This will be followed by a discussion of the variations in low-level gain and in maximum output over an extended range of beam current, frequency, and magnetic field. By this procedure we are able to obtain a description of tube performance without presentation of a formidable number of curves. Two topics, noise and inter-

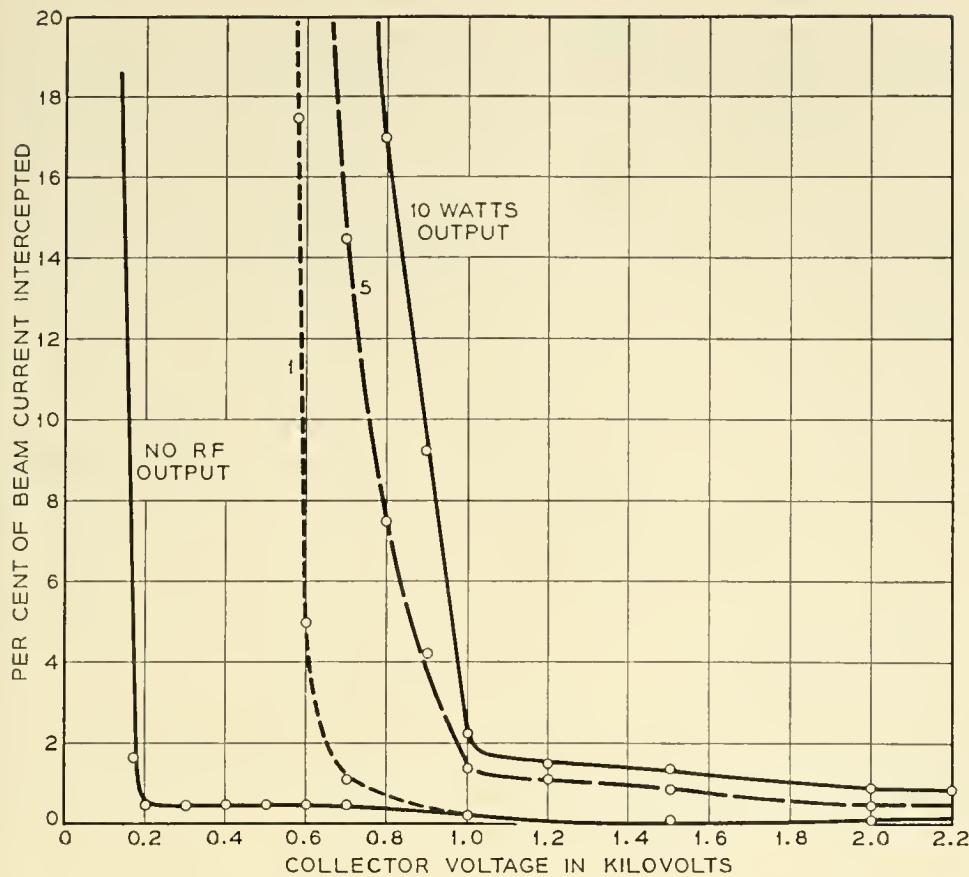


Fig. 22 — Intercepted current as a function of collector voltage with helix and accelerator voltages held constant at their nominal values. Below the knee of the curves about three quarters of the total intercepted current goes to the helix and about one quarter is focused all the way back to the accelerator. Curves are shown for no rf input and for output levels of 1, 5, and 10 watts. With no input, the lowest permissible collector voltage is determined by the phenomenon of space charge blocking. With rf input, it is determined mainly by the velocity spread of the electrons. In all cases it was found that the alignment of the TWT with respect to the magnetic circuit becomes more critical as the knee of the curve is approached. For this reason the M1789 is usually operated with a collector voltage about 200 volts above the knee.

modulation, will be divorced from the discussion as outlined above and treated separately in Sections 4.4 and 4.5.

4.2 Operation Under Nominal Conditions

Basic Characteristics

By nominal conditions for the M1789 we mean the following:

frequency.....	6175 mc (band center)
beam current.....	40 ma
magnetic flux density.....	600 gauss
collector voltage.....	1200 volts

Fig. 23(a) shows representative curves of output power as a function of input power for several values of helix voltage. This information is replotted in Fig. 23(b) in terms of gain as a function of output power. We see that the TWT operates as a linear amplifier for low output levels. As the output level is increased, the tube goes into compression and finally a saturation level is reached. The maximum gain at low input levels is obtained with a helix voltage of 2,400 volts (about 10 per cent higher than the synchronous voltage because of space charge effects).

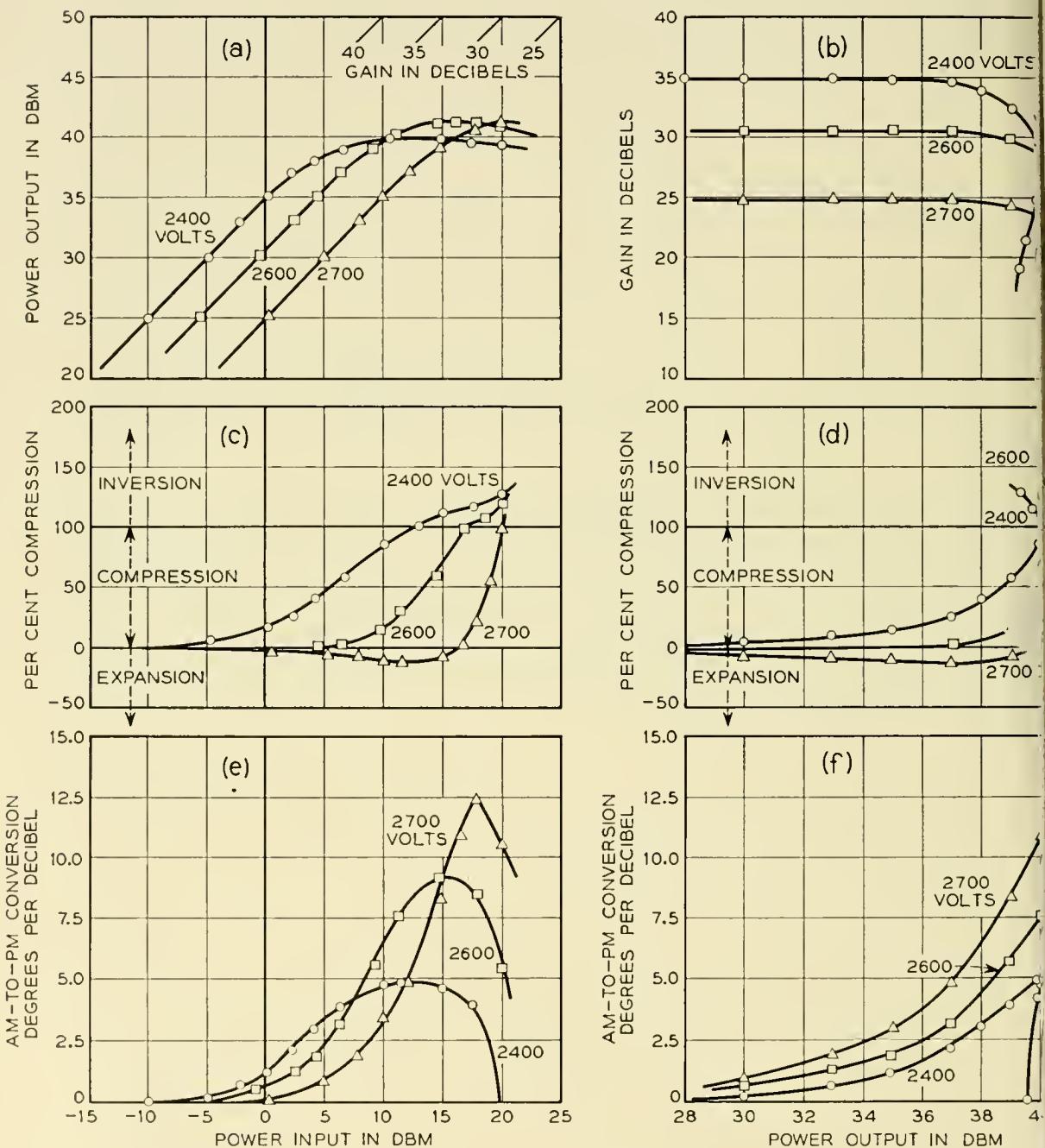


Fig. 23 — See opposite page for caption

The maximum output at saturation is obtained at a higher helix voltage as is common in TWT's. The helix voltage also affects the shape of the input-output curves — linear operation being maintained to higher output levels at higher helix voltages.

As a measure of the efficiency of electronic interaction in a TWT, we use an "electronic efficiency" which is defined as the ratio of the rf output power to the beam power (product of helix voltage and beam current). The "over-all efficiency" we define as the ratio of the rf output power to the total dc power (exclusive of heater power) delivered to the tube. With the collector operated at 1,200 volts, it is about twice the electronic efficiency. For the M1789, maximum efficiency occurs at the saturation level with a helix voltage of 2,600 volts. The electronic and over-all efficiencies there are equal to about 14 per cent and 28 per cent, respectively.

The curves of Figs. 23(a) and (b) were taken with sufficient time allowed for the tube to stabilize at each power level. If the TWT is driven to a high output level after having been operated for several minutes with no input signal, the output will be somewhat greater than is shown in the curves. It will gradually decrease until it reaches a stable level in a period of about two minutes. This "fade" is caused by an increase in the intrinsic attenuation of the helix near the output end. The increase is a result of heating from rf power dissipation. At maximum output the fade is about 0.6 db (about 15 per cent decrease in output power). At the five-watt output level the fade is about 0.1 db (about 2 per cent

Fig. 33 — See opposite page

(a) Output power as a function of input power. Both ordinate and abscissa are in dbm (db with respect to a reference level of one milliwatt). A straight line at 45° represents a constant gain. A gain scale is included along the top of the figure. For this tube a helix voltage of 2,400 volts gives maximum gain at low signal levels and a voltage of about 2,600 gives maximum output at saturation.

(b) Gain as a function of output power. This is an alternate way of presenting the information shown in (a).

(c) Compression as a function of input power. Three regions are shown in the figure. The "compression" region is that in which there is less than one db change in output level for a db change in input level. The "expansion" region is that in which there is more than one db change in output level for a db change in input level. The "inversion" region is that in which the output level decreases when the input level increases (or vice versa). It occurs for input levels greater than that necessary to drive the TWT to saturation. In this region the change in output is of opposite sign to the change in input. Using the definition in the text this gives rise to compression values in excess of 100 per cent.

(d) Compression as a function of output power.

(e) Conversion of amplitude modulation to phase modulation as a function of input power. This conversion arises because the electrical length of the TWT is a function of the input level. The effect can cause rather serious difficulties in certain types of low index FM systems.

(f) Conversion of amplitude modulation to phase modulation as a function of output power.

decrease in output power). We will present some additional data on this effect in Section 4.3.

Distortion of the Modulation Envelope

The curves of Figs. 23(a) and (b) tell what happens when a single frequency carrier signal is passed through the TWT. In addition we would like to know the effect on modulation which may be present on the signal. In particular, it is desirable to know the compression of the envelope of an AM signal and the amount of phase modulation generated in the output signal as a result of amplitude modulation of the input signal, (an effect commonly known as AM-to-PM conversion). As a measure of compression of an AM signal the quantity per cent compression will be used. This is defined as

$$\% \text{ Compression} = \left[1 - \frac{\Delta V_0/V_0}{\Delta V_i/V_i} \right] 100$$

where V_0 is the voltage of the output wave, V_i is the voltage of the input wave, and ΔV_0 is the change in output voltage for a small change ΔV_i in the input voltage. When $\Delta V/V$ is small it can be expressed in db as $8.68 \Delta V/V = \Delta V/V$ in db. From this it follows that

$$\% \text{ Compression} = \left[1 - \frac{\Delta P_0}{\Delta P_i} \right] \text{ in db } 100$$

where ΔP_0 is the change in output power for a change ΔP_i in input power, and the two powers are measured on a db scale. When the per cent compression is zero the TWT is operating as a linear amplifier; when it is 100 per cent the TWT is operating as a limiter.

From the above expression it may appear that the per cent compression could be determined directly from the slopes of the input-output curves. This would be the case were it not for fading effects. Since there is fading, however, the slope for rapid input level changes is different at high levels from the slope of the static curves. Thus it is necessary to determine compression from the resulting effect on an AM signal.

The electrical length of a TWT operated in the non-linear region is to some extent dependent on the input level. Therefore, an AM signal applied to the input of the TWT will produce phase modulation (PM) of the output signal. This effect may be of particular concern when a TWT operating at high output levels is used to amplify a low-index FM signal. If such a signal contains residual amplitude modulation, the TWT generates phase modulation with phase deviation proportional to the input amplitude variation. Under certain circumstances this can cause

severe interference with the signal being transmitted. We will discuss a particular example after consideration of the compression and AM-to-PM conversion characteristics of the M1789.

As in the case of compression, we must measure AM-to-PM conversion dynamically. This is necessary because point-by-point measurements of the shift in output phase as input level is changed include a component of phase shift caused by changes in temperature of the ceramic support rods and a consequent change in their dielectric constant. However, this thermal effect does not follow AM rates of interest and therefore does not produce AM-to-PM conversion.

Fig. 24 shows a simplified block diagram of the test set used to measure compression and AM-to-PM conversion. This equipment amplitude modulates the input signal to the TWT under test by a known amount and detects the AM in the output signal with a crystal monitor and the PM with a phase bridge. A more complete discussion of this measurement is given by Augustine and Slocum.³

Compression is given as a function of power input in Fig. 23(c) and as a function of power output in Fig. 23(d). We see that compression sets in more suddenly at higher helix voltages. Above about 2,500 volts

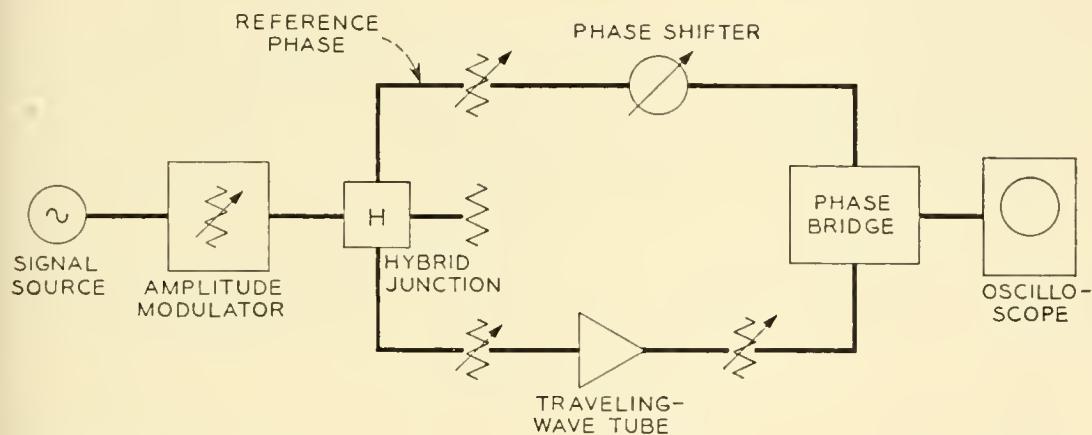


Fig. 24 — Simplified block diagram of test set used to measure compression and conversion of amplitude to phase modulation. A ferrite modulator introduces one db of 60 cps amplitude modulation into the test signal. The 60 cps rate is much higher than that which can be followed by thermal changes in the TWT. Half of the modulated signal serves as input to the TWT under test and half serves as a reference phase for a phase detector. The signals at the phase detector input are maintained equal and at constant level and nominally in phase quadrature. The detector is essentially a bridge circuit, the output of which is a dc voltage proportional to the phase difference of the two inputs. When operated with inputs in quadrature it is not sensitive to amplitude changes of as much as two db in either or both inputs. Phase modulation introduced by the amplitude modulator appears at both inputs and thus does not produce an indication. The output of the detector is therefore a direct measure of the phase modulation created in the TWT. Compression is determined by comparing the percentage amplitude modulation at the input and output crystal monitors.

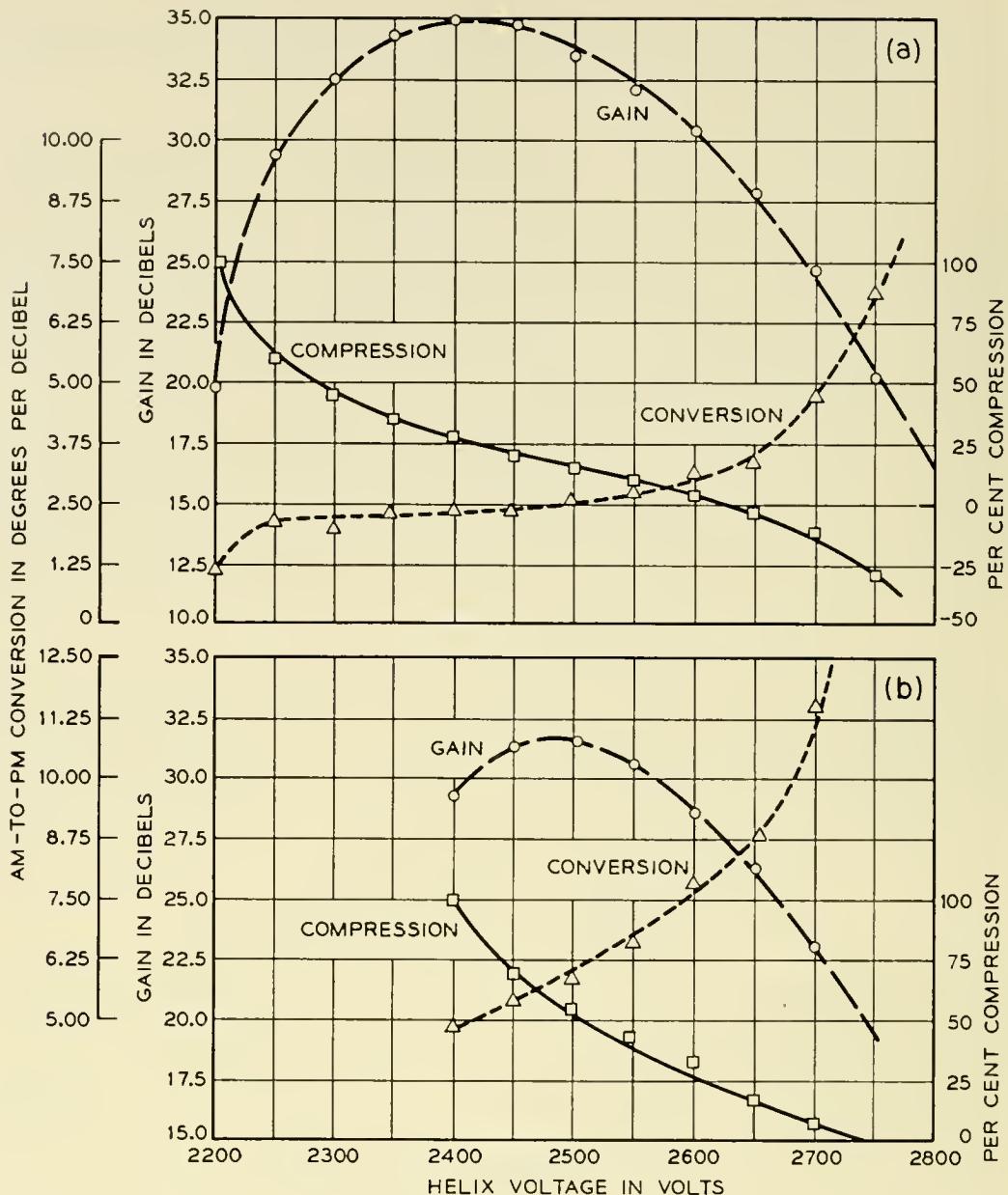


Fig. 25 — Gain, compression and amplitude to phase conversion as a function of helix voltage with the output power maintained constant at a level of five watts (a) and ten watts (b).

there is expansion for some values of power input. Figs. 23 (e) and 23(f) give the AM-to-PM conversion, as functions of input and output power respectively. These data indicate that the conversion is very much less if the tube is operated at lower helix voltages. For example, the conversion at the saturation level of the 2,700-volt curve is about $2\frac{1}{2}$ times that for the 2,400-volt curve.

A final method of plotting gain, compression, and AM-to-PM conversion data is shown in Fig. 25. The abscissa here is the helix voltage.

For these measurements power output was held constant by adjusting input level at each voltage. The figure shows that as helix voltage is increased, the compression decreases but the AM-to-PM conversion increases. The choice of a helix voltage at which to operate the tube must therefore represent a compromise between these quantities.

Phase Modulation Sensitivity

The equipment of Fig. 24 was also used to measure the phase modulation sensitivity of various electrodes by omitting the amplitude modula-

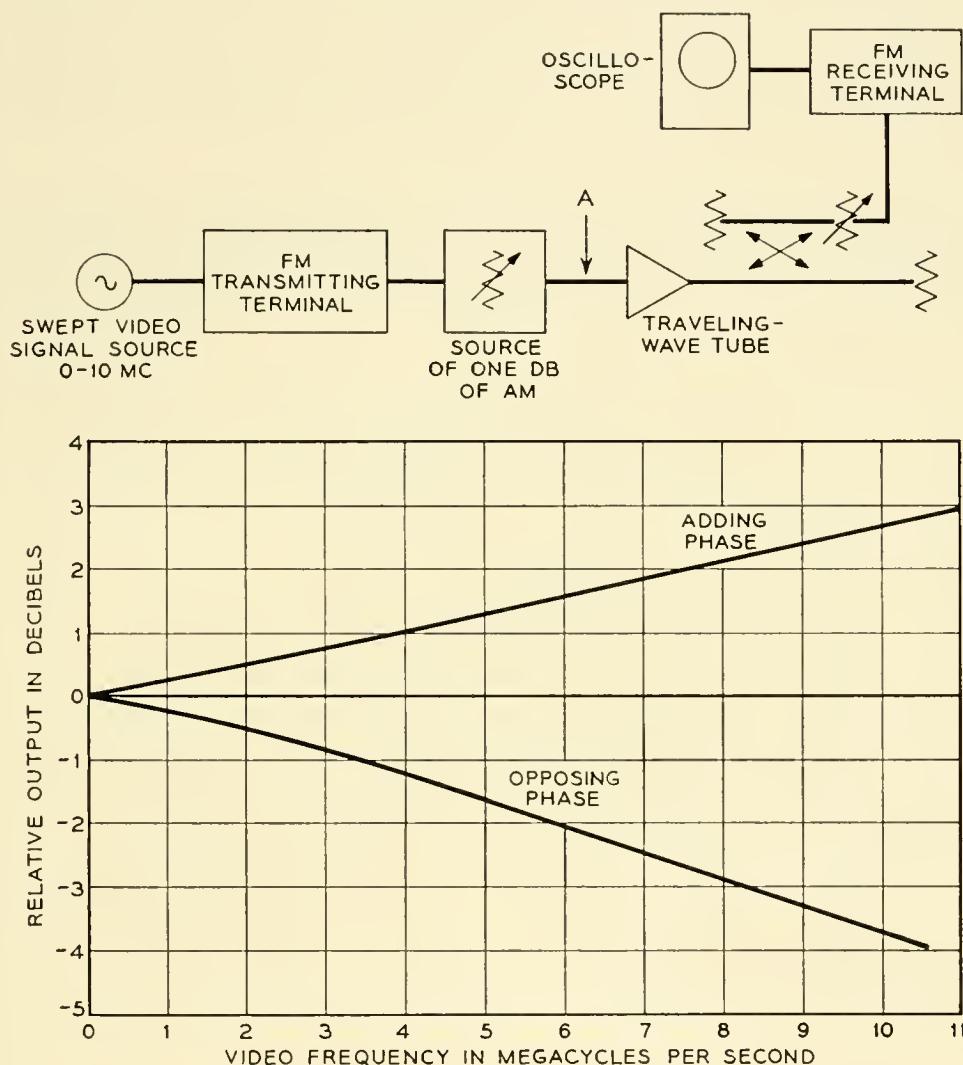


Fig. 26 — Example of frequency response shaping caused by AM-to-PM conversion. This figure shows the calculated frequency response viewed between FM terminals for the system shown in the block diagram. Curves are given for the case in which the phase modulation generated in the TWT both adds to and subtracts from that of the transmitted signal. Inclusion of a limiter at point A would result in a flat frequency response.

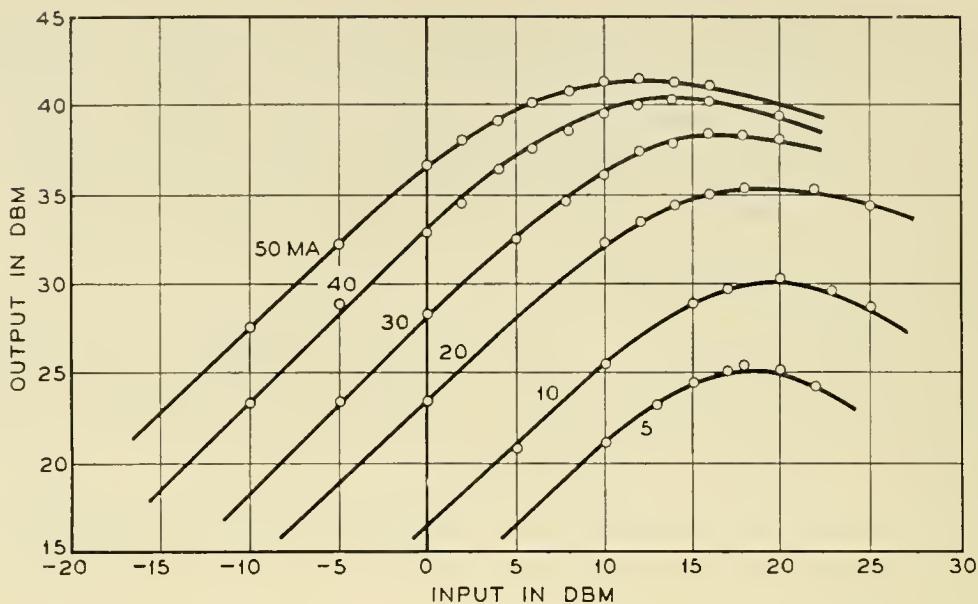


Fig. 27 — Output power as a function of input power at various beam currents. These curves were all taken with the helix voltage adjusted to give the maximum gain at low signal levels. At low beam currents (<20 ma) there is insufficient gain between the attenuator and the output so that at these currents the attenuator section is limiting the power output. This accounts for some of the difference in shape of the curves near maximum output.

tor and introducing small changes in electrode voltages. The modulation sensitivity of the helix is about two degrees per volt and that of the accelerator about 0.1 degree per volt with the TWT operating under nominal conditions.

Significance of AM-to-PM Conversion

Let us return briefly to a discussion of some consequences of AM-to-PM conversion. As an example, we will consider the case of a low-index FM signal. Assume the frequency deviation is ± 5 mc peak to peak. This gives a phase deviation of ± 0.5 radian for a 10 mc modulating signal. These values are typical of what might be found in a radio relay system. Let us also assume that there is a residual amplitude modulation of one db (about 13 per cent) in this signal and suppose further that the signal is amplified by a TWT having a value of AM-to-PM conversion of 10 degrees per db. The phase modulation thus created in the TWT can either add to or subtract from that of the original FM signal, thus changing its modulation index. At low modulation signal frequencies the phase deviation of the FM signal will be large compared to that of the PM interference and the interference will be of little consequence. At high modulation signal frequencies the phase deviation of the original FM and of the interfering PM signals will be comparable and the interference

can considerably change the net phase deviation of the overall signal. For the example we are considering the frequency responses in Fig. 26 show what would be seen at the output FM terminal. Curves are given both for the PM interference adding to and subtracting from the original FM signal. We see that a gain-frequency slope of about 4 db over 10 mc is introduced by AM-to-PM conversion. To prevent such an effect, a limiter should be used prior to the TWT in applications of this nature so as to remove the offending AM from the input signal.

The fact that compression and amplitude-to-phase conversion vary with input level means that in addition to the first order distortion just described, higher order distortions of the modulation envelope will occur. If, for example, the input signal is amplitude modulated at frequency f_1 , the output modulation envelope will contain amplitude and phase modulation both at f_1 and at harmonics of f_1 . The amount of higher order distortion can be estimated by expanding the compression and amplitude-to-phase conversion curves as a function of power input in a Taylor series about the operating point. Such an expansion shows that the greater the slope of these curves the greater will be the higher order distortions.

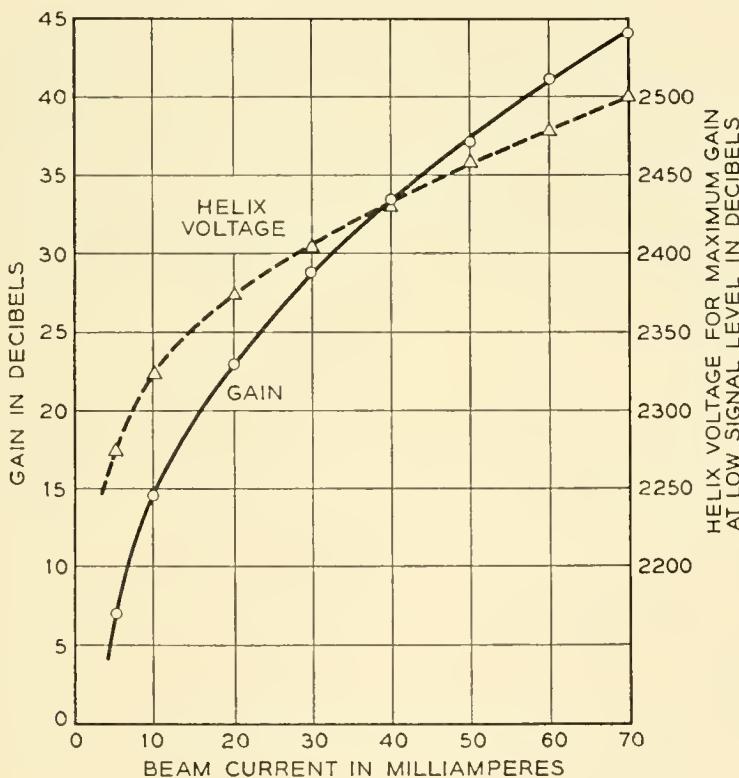


Fig. 28 — Low-level gain as a function of beam current. The helix voltage was adjusted for maximum gain at each current.

Reproducibility

The curves presented in this section are all for the same tube, one which is representative of a group of 50 which were built at the conclusion of the M1789 development program. The tubes in this group had characteristics falling within the following ranges. The numbers represent the range containing 90 per cent of the tubes tested.

Accelerator Voltage for 40 ma.....	2,500-2,700
Helix Voltage for maximum low-level gain.....	2,350-2,450
Low-level gain.....	33-37 db
Gain at 5 watts output.....	31-35 db
Maximum power output.....	{ 40.5-42 dbm (11.2-15.8 watts)

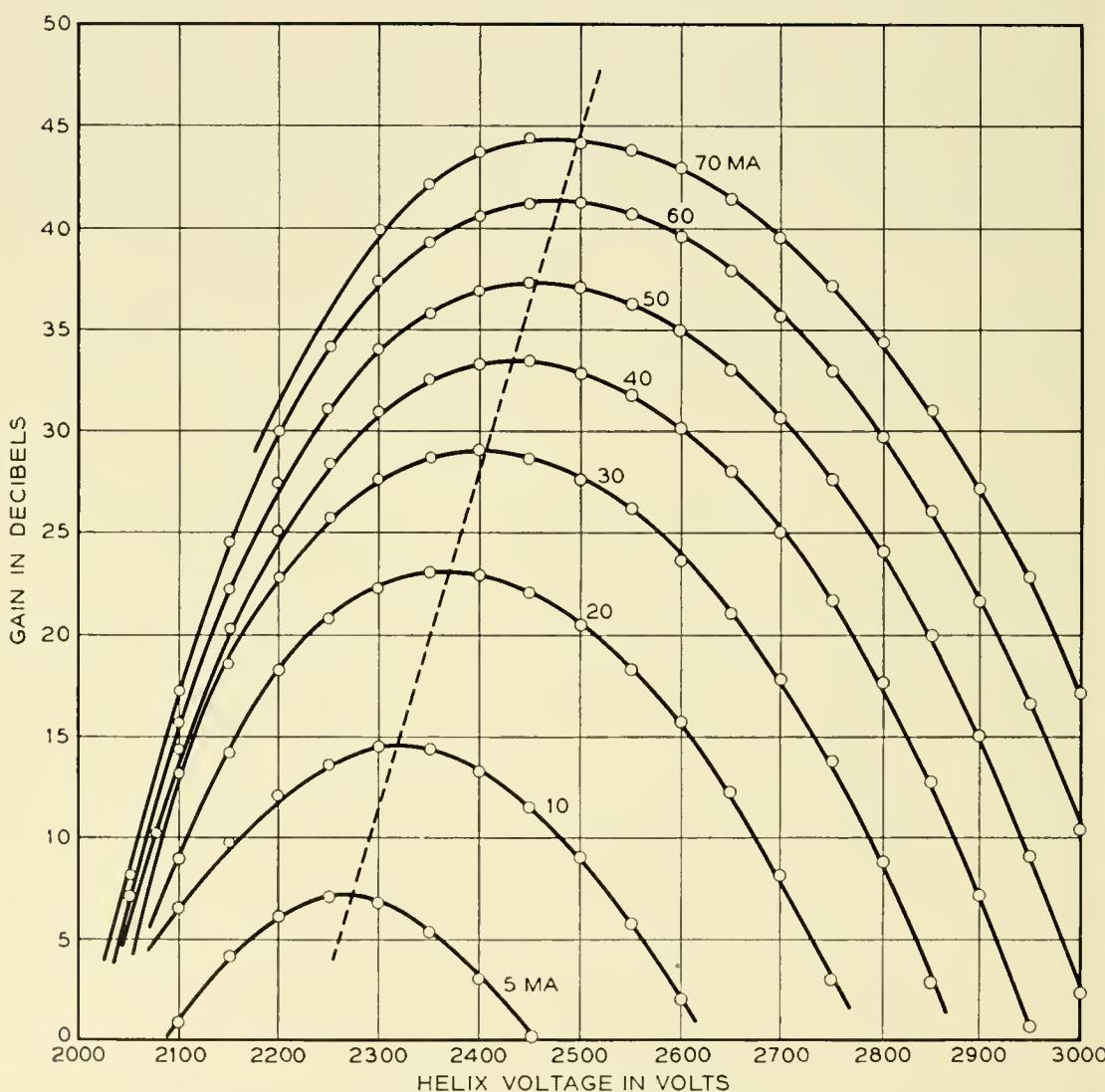


Fig. 29—Low-level gain as a function of helix voltage for various beam currents. The dotted line represents the locus of the maxima of the curves.

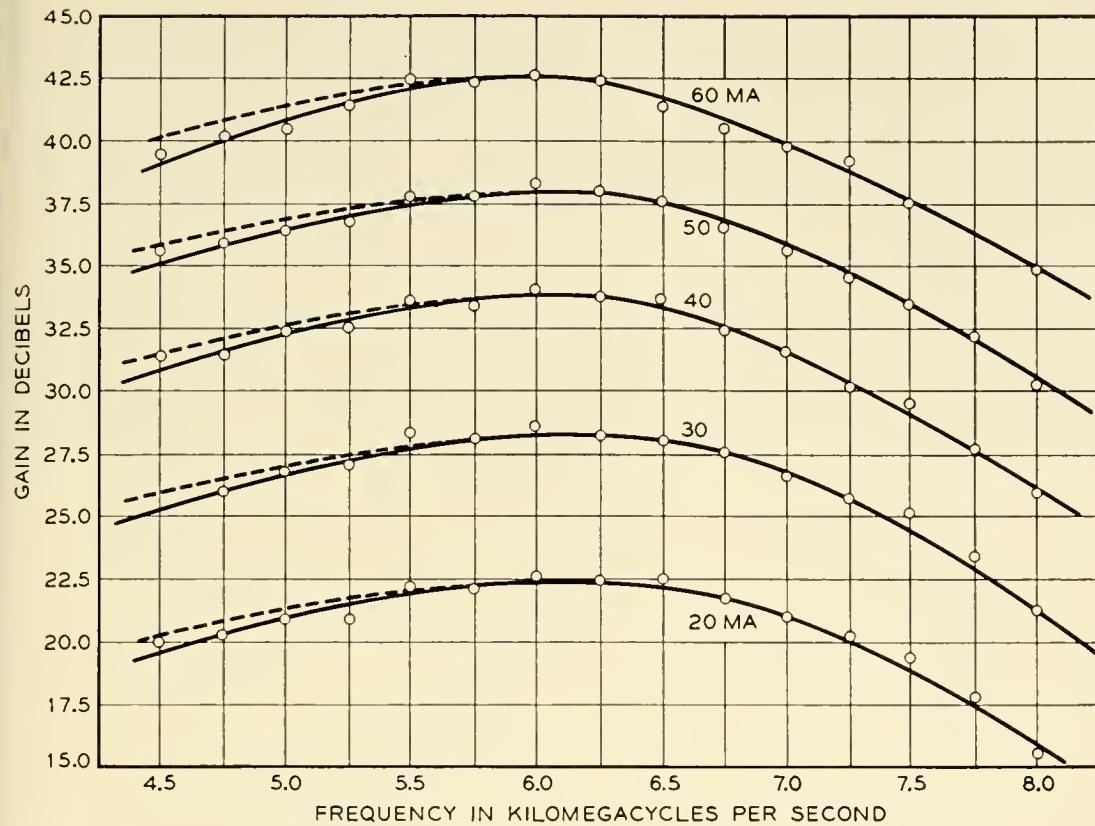


Fig. 30 — Low-level gain and helix voltage for maximum gain as functions of frequency for several beam currents. The TWT was matched to the waveguide (with tuners where necessary outside of the 5,925 to 6,425-mc range) at each frequency. The solid curves show the gain-frequency characteristic with the helix voltage adjusted for maximum gain at 6,000 mc for each beam current and then held constant as frequency was changed. Experimental points correspond to this condition. The dotted curves show how the characteristics change when helix voltage is optimized at each frequency. The optimum helix voltage increases by about 100 volts in going from 6,000 down to 4,500 mc because of slight dispersion in the phase velocity of the helix.

4.3 Operation Over an Extended Range

We now turn to a consideration of typical M1789 characteristics over an extended range of beam current, frequency, and magnetic field.* We shall concentrate on two items, the low-level gain and the maximum power output. From variations in these quantities the complete compression curves can be roughly deduced. This situation is illustrated in Fig. 27 which shows output as a function of input at different beam currents. While the shapes of these curves are slightly different, for the most part they can be derived from the 40-ma curve by shifting it along the abscissa

* The characteristics of the tube used for the low-level gain measurements in this Section were slightly different from those of the tube used for the maximum output measurements and both were slightly different from those of the tube used for the measurements of Section 4.2. All tubes, however, had characteristics falling within the ranges listed above.

by the amount the low-level gain changes, and along the ordinate by the amount the maximum output changes as beam current is varied. A similar procedure can be followed for variations with frequency and magnetic field. In all figures in this Section, parameters not being purposely varied were held at the nominal values given on page 1315.

Low-level Gain

Fig. 28 shows the variation in low-level gain with beam current and Fig. 29 shows its variation with helix voltage for several different beam currents. Fig. 30 shows the variation with frequency and Fig. 31 the variation with magnetic field.

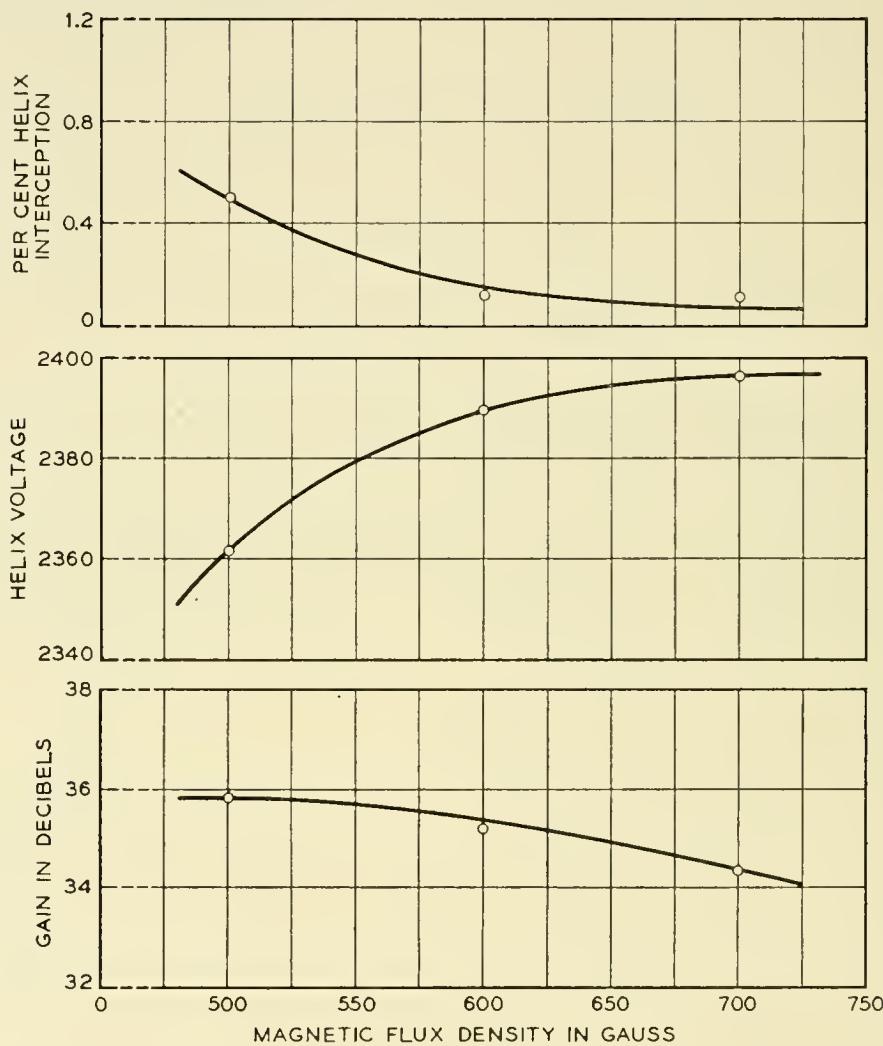


Fig. 31 — Low-level gain, helix voltage for maximum gain and helix interception at low signal level as functions of magnetic flux density. These measurements were made using different strength permanent magnet circuits. The gain varies with magnetic flux density mainly as a result of its effect on beam size and therefore on the degree of coupling between electron stream and helix. The helix voltage varies because of the effect of beam size on QC and therefore on the ratio of the optimum gain voltage to the helix synchronous voltage.

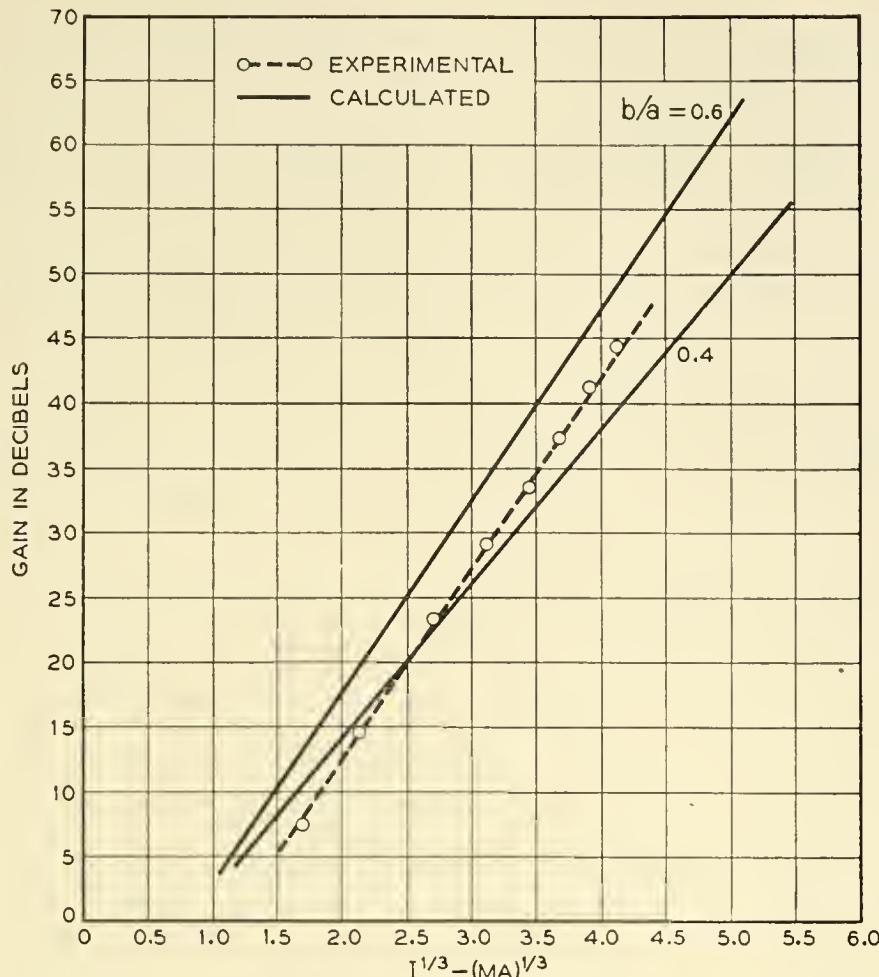


Fig. 32 — Measured and calculated low-level gain as a function of the one-third power of beam current. The parameter b/a is the ratio of effective beam diameter to mean helix diameter.

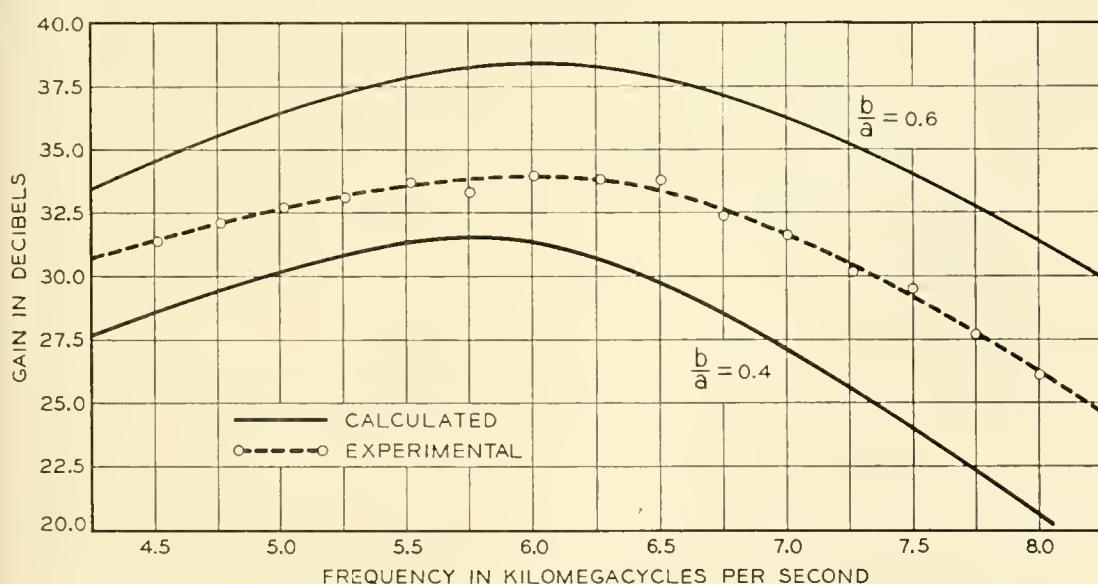


Fig. 33 — Measured and calculated frequency response for a current of 40 ma.

The observed gain compares well with that calculated from low-level TWT theory provided that we properly consider the effect of the helix attenuator and provided that we assume a b/a of one-half. The method we have used in calculating the M1789 gain is discussed further in Appendix I. Fig. 32 compares the measured and calculated gain as a function of beam current and Fig. 33 compares them as a function of frequency. Fig. 34 shows measured and calculated ratios of voltage for maximum gain to synchronous voltage as a function of beam current. In all these figures calculations are shown for several values of the ratio of effective beam diameter to mean helix diameter (b/a). We see that the effective value of b/a appears to be about one-half. On the basis of measurements made by probing the beam of a scaled up version of a

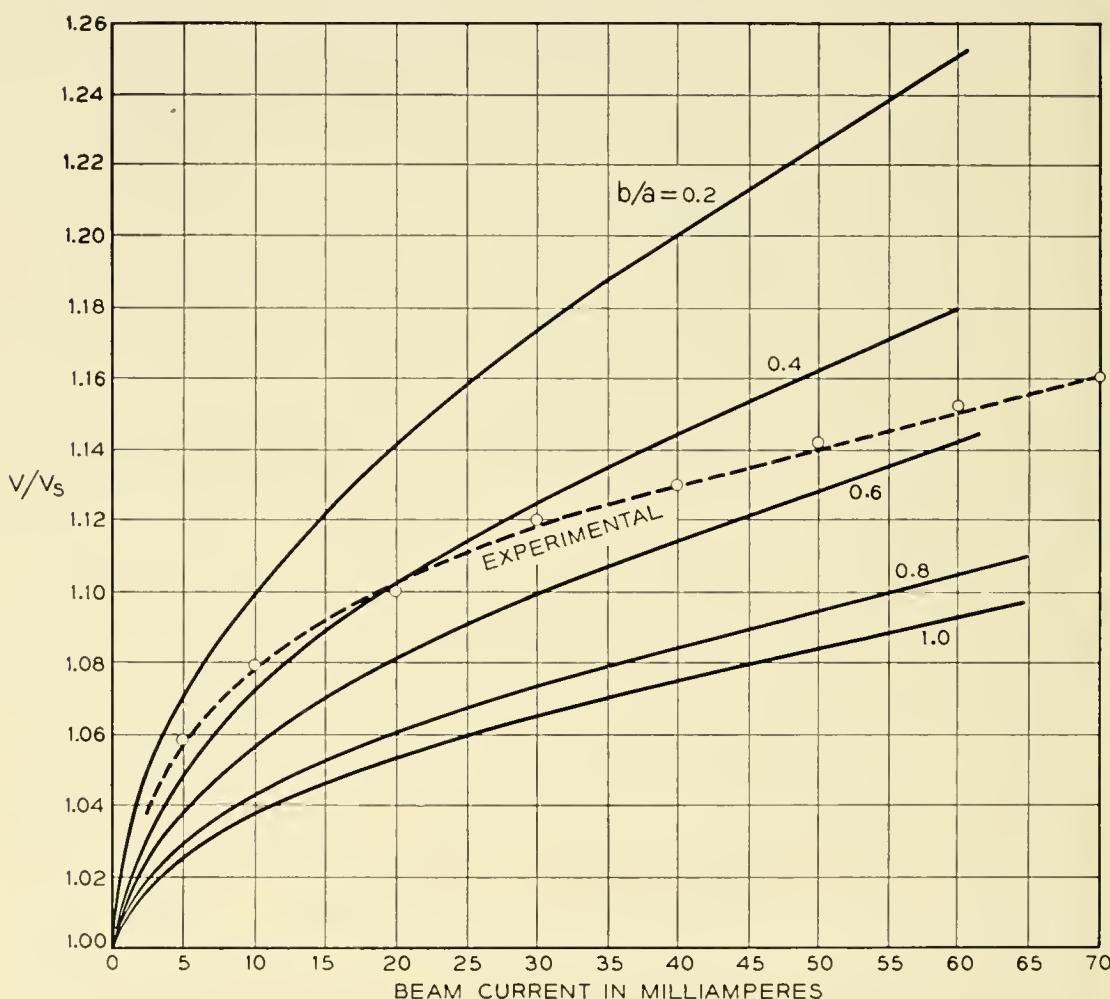


Fig. 34 — Measured and calculated ratio of voltage for maximum gain to synchronous voltage as a function of beam current. The calculated curves are shown for several values of the ratio of effective beam radius to mean helix radius (b/a). The location of the measured curve among the calculated ones is taken as an indication of the effective value of b/a in the M1789. At 40 ma it is about 0.5.

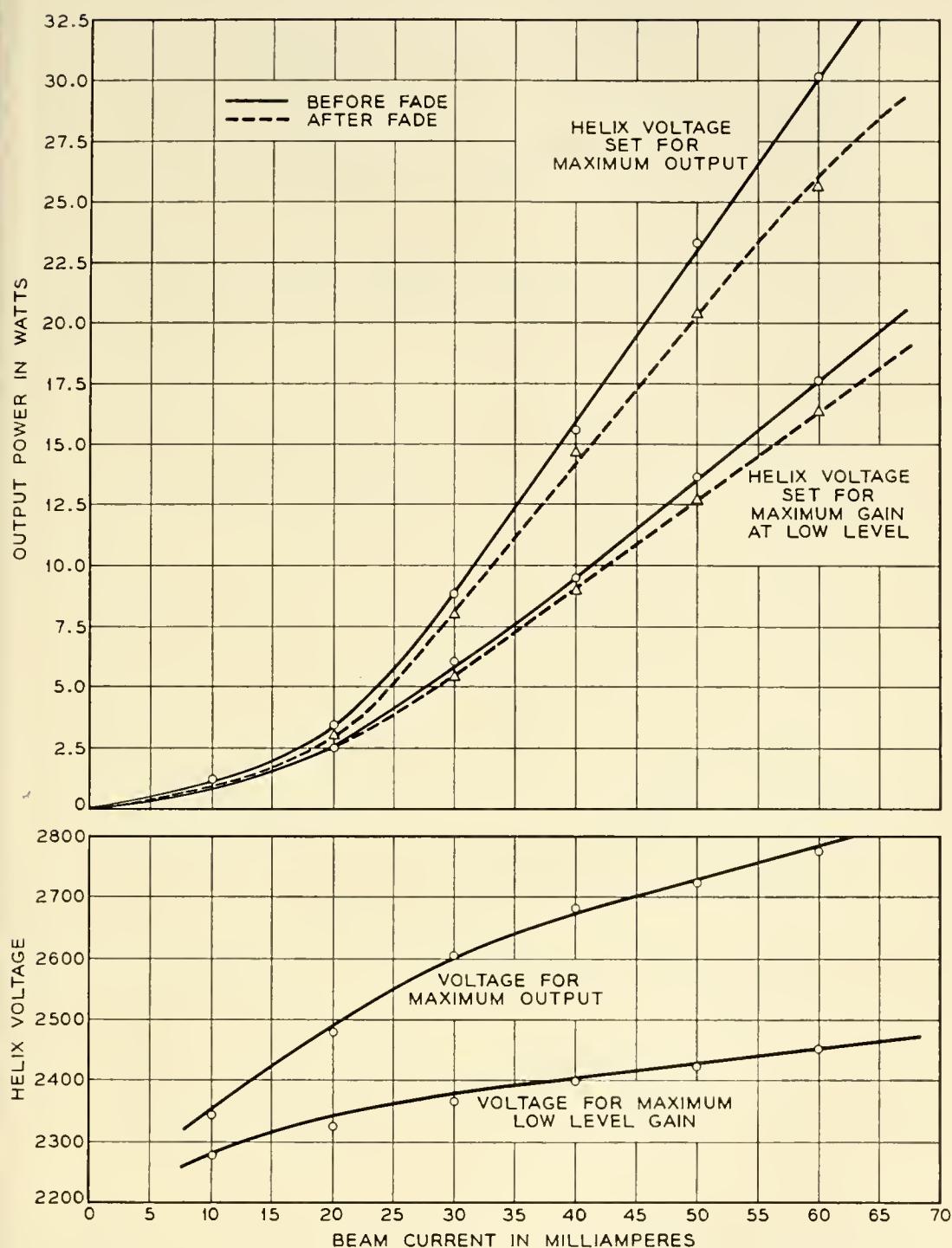


Fig. 35 — Maximum power output and helix voltage as functions of beam current. Curves are shown for before and after fading, and for the helix voltage adjusted for the maximum gain at low-level and for maximum output.

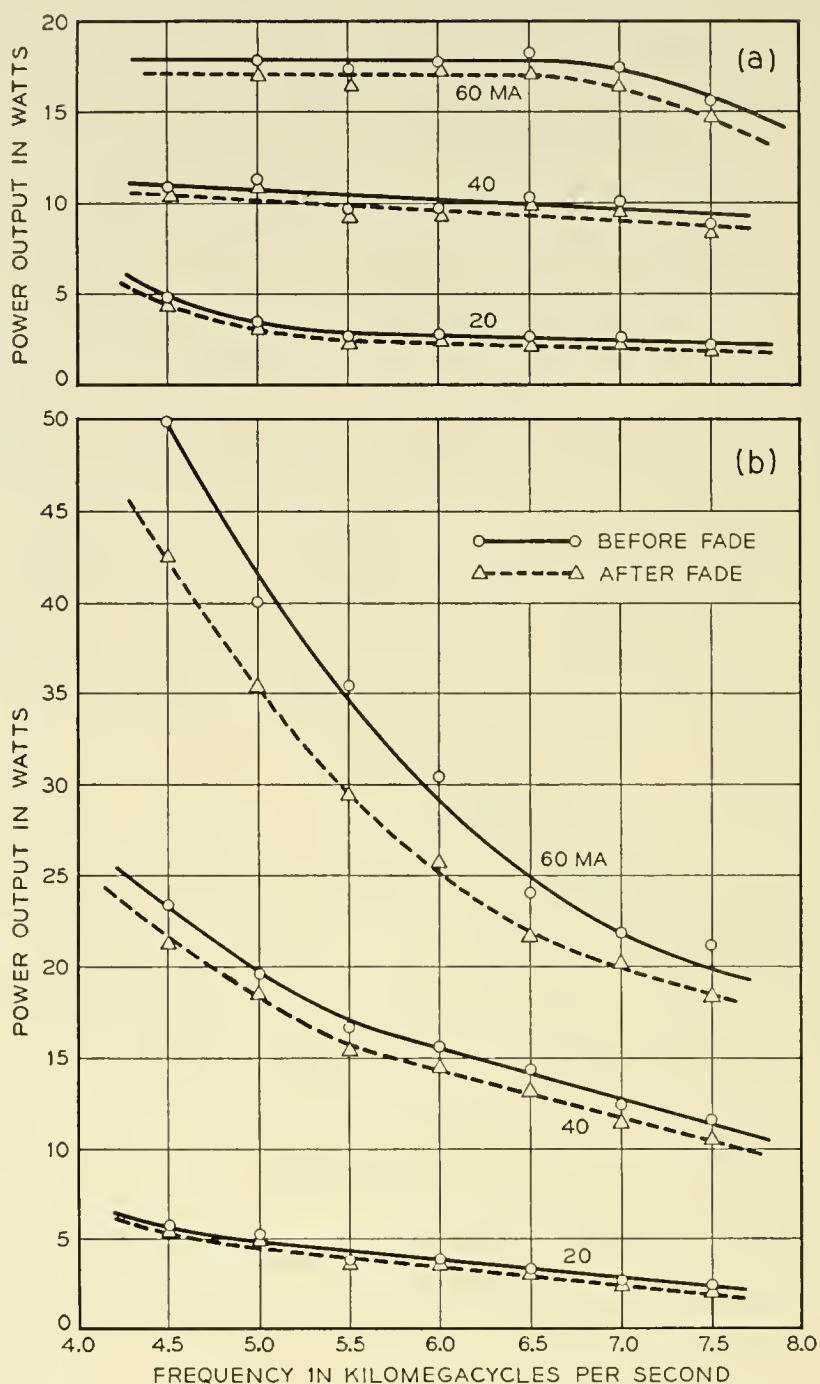


Fig. 36 — Maximum power output after fading as a function of frequency for several beam currents; in (a) with the helix voltage adjusted for maximum gain at low-level and in (b) with the helix voltage adjusted for maximum power output.

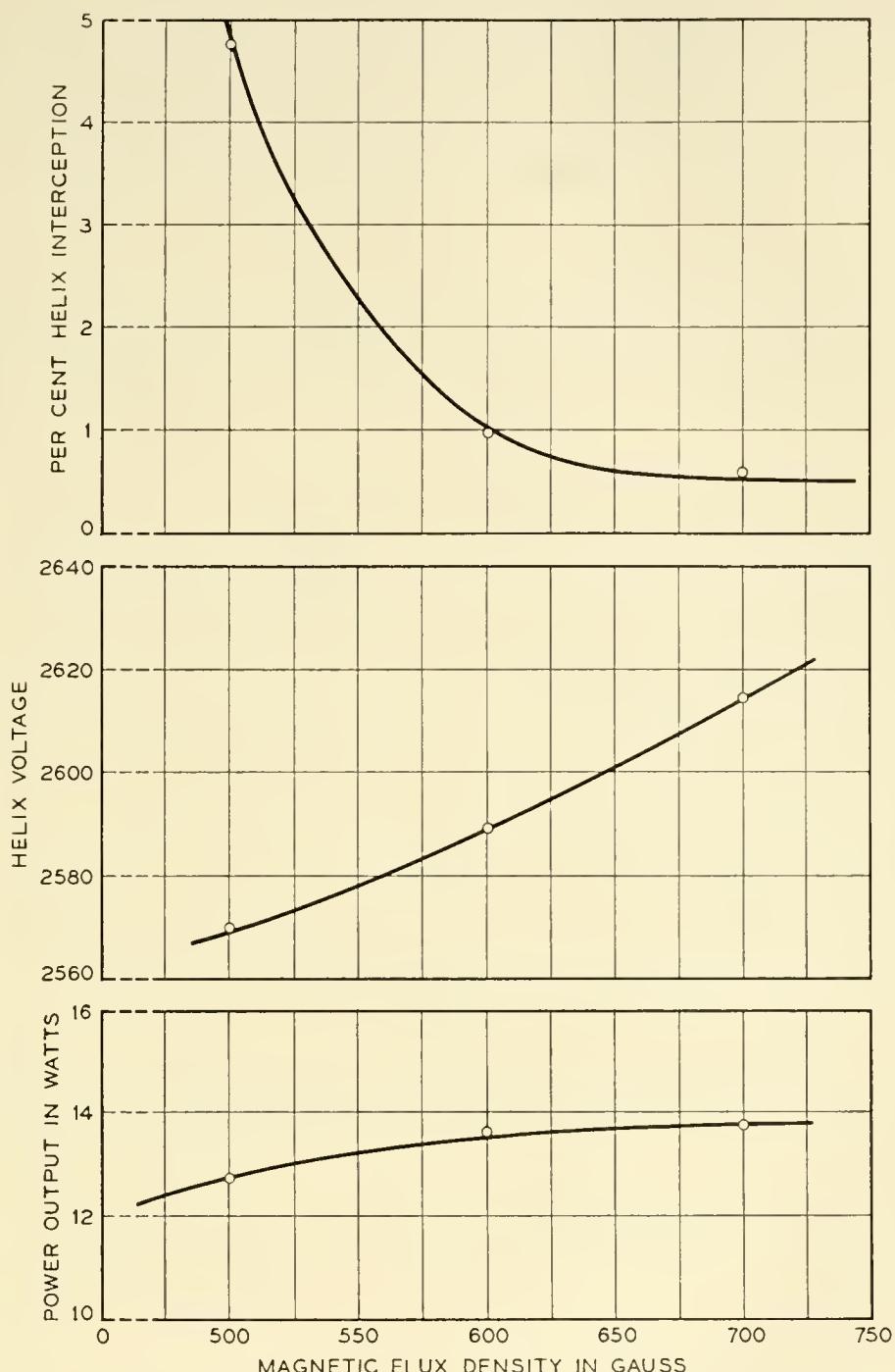


Fig. 37 — Maximum power output after fading, voltage for maximum output, and helix interception at maximum output as functions of magnetic flux density. These measurements were made using magnetic circuits charged to different strengths. Helix interception above about one per cent is undesirable if long tube life is required.

focusing system similar to that employed in the M1789, we estimate the actual beam diameter (for 99 per cent of the current) to be about 65 mils ($b/a = 0.7$). However, the current density distribution is peaked at the center of the beam because of the effect of thermal velocities of the electrons. Thus an effective b/a of 0.5 is not unreasonable.

Maximum Power Output

Fig. 35 shows the maximum power output as a function of beam current both immediately after rf drive is applied and after the tube has had time to stabilize. We see that at high rf power outputs the fading

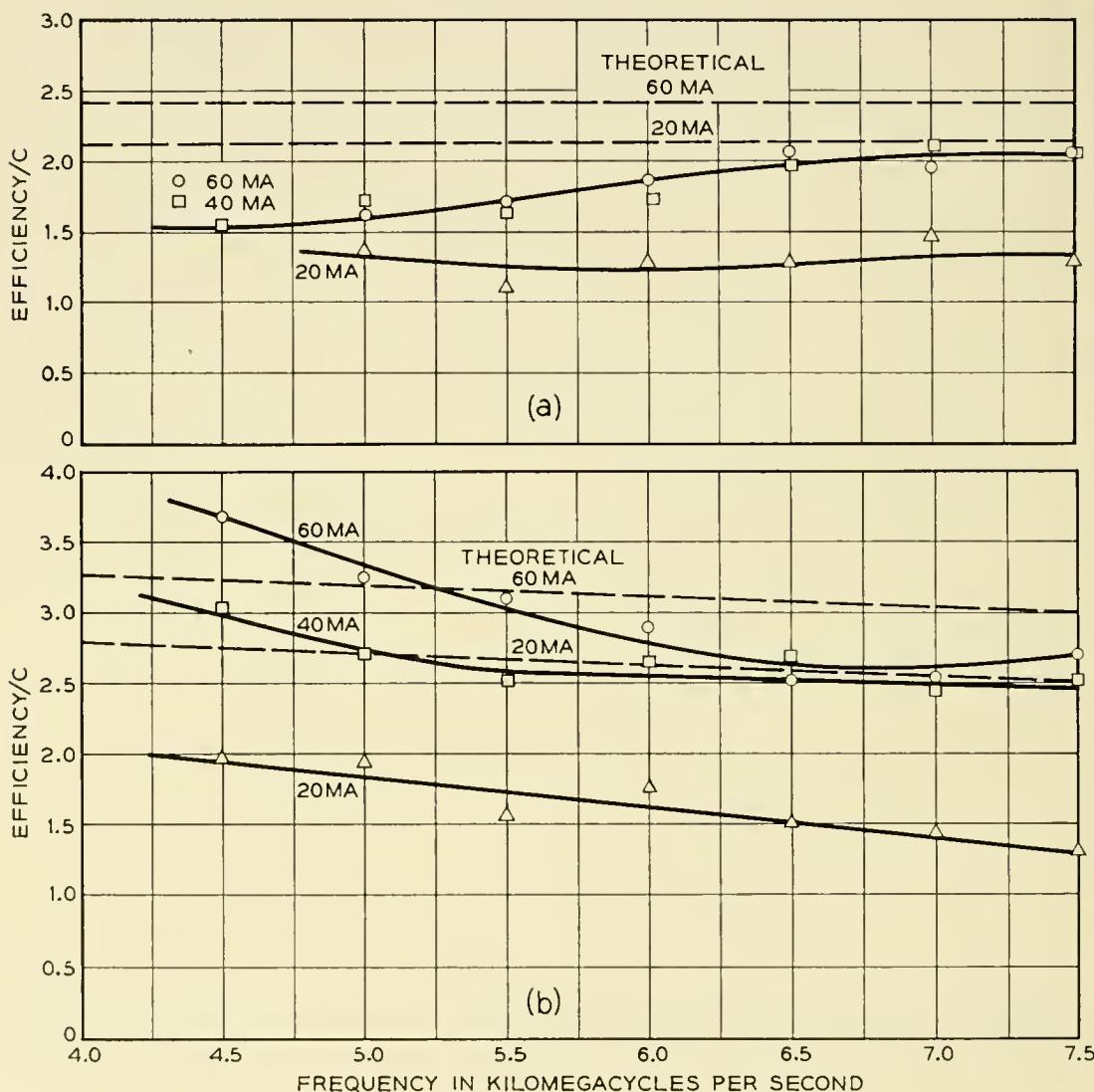


Fig. 38 — Ratio of electronic efficiency to gain parameter C as a function of frequency. The efficiencies used for this comparison are all before fading. The dotted lines are estimated from the Tien theory corrected for the intrinsic loss of the helix. The curves in (a) are for the case of the helix voltage adjusted for the maximum low-level gain and those in (b) for the case of the helix voltage adjusted for maximum power output.

becomes very serious and eventually limits the TWT output to about 30 watts. If it were necessary to reduce this fading, the envelope shrinking technique illustrated in Fig. 16 could be used. The maximum power output after fading is shown as a function of frequency for several beam currents in Fig. 36 and as a function of magnetic flux density in Fig. 37.

The theory of the high level behavior of a TWT⁴ predicts that the ratio of electronic efficiency (i.e., E = power output/beam power) to the gain parameter C should be a function of C , QC and γb (where b is the beam diameter). However, with the range of parameters encountered in the M1789, the variation in E/C should be small. Fig. 38(a) shows E/C as a function of frequency when the TWT is operating at the voltage for maximum gain at low signal levels. Fig. 38(b) shows the maximum value of E/C obtainable at elevated helix voltage. In both figures we show the efficiency as estimated using the results of Tien⁴ corrected for the effect of intrinsic loss following the procedure of Cutler and Brangaccio.⁵ All efficiencies in these two figures are the electronic efficiency before fading. It would be quite difficult to compare the efficiency after fading with theory because the intrinsic attenuation in this case varies along the helix in an unknown manner so that we cannot properly take it into account. From the figures we see that the calculated value of E/C at 6,000 me and 40 ma is not far from the experimental value but the experimental points show more variation with frequency than is predicted by theory. The low efficiency at 20 ma results from the fact that there is insufficient gain between the helix attenuator and the output. As a result, the TWT "overloads in the attenuation."

4.4 Noise Performance

A new and important noise phenomenon was observed in the course of the M1789 development. It was found that the noise figure is strongly dependent on the magnetic flux linking the cathode and on the rf output level of the TWT. For example, with the TWT operating near maximum output and with a cathode completely shielded from the magnetic field, noise figures of about 50 db were observed. By allowing 20 gauss at the cathode, the noise figure was reduced to 30 db. Fig. 39 shows the noise figure as a function of magnetic flux density at the cathode for several values of rf power output. We see that there is a peak of noise figure roughly symmetrical about zero flux at the cathode, and that the magnitude of this peak is considerably increased by operating the TWT at high output levels.

Some additional observed properties of the noise peak are:

- (1) The magnitude depends on the synchronous voltage of the helix. For a 1,600-volt helix it is about 10 db higher than shown in Fig. 39 and

for a 2,600-volt helix it is about 5 db lower. The noise figure for 25 gauss at the cathode remains constant, however.

(2) There appears to be a threshold level of about 15-ma beam current below which the peak does not occur. Between 15 and 25 ma the peak increases. Above 25 ma it is roughly constant in magnitude.

(3) The peak can be considerably reduced by intercepting some of the edge electrons before they reach the helix region.

For this discussion it has been necessary to extend the concept of noise figure to the case of non-linear operation of the TWT. Essentially this noise figure is defined by the means we use to determine it. A block diagram of the equipment is shown in Fig. 40. The outputs of a calibrated broad band noise source and a signal oscillator are combined and used for the input to the TWT under test. The noise output from the TWT is passed through a filter tuned about 100 mc away from the signal so as to reject the carrier. It is then detected by a receiver tuned to the filter frequency. The noise figure is measured by turning the noise source off and on, noting the change in receiver output level and calculating the noise figure in the conventional manner. This procedure reduces to an ordinary noise figure measurement in the absence of input signal.

There are other ways that could be used to measure noise figure of a non-linear amplifier. A method more closely related to the use of the

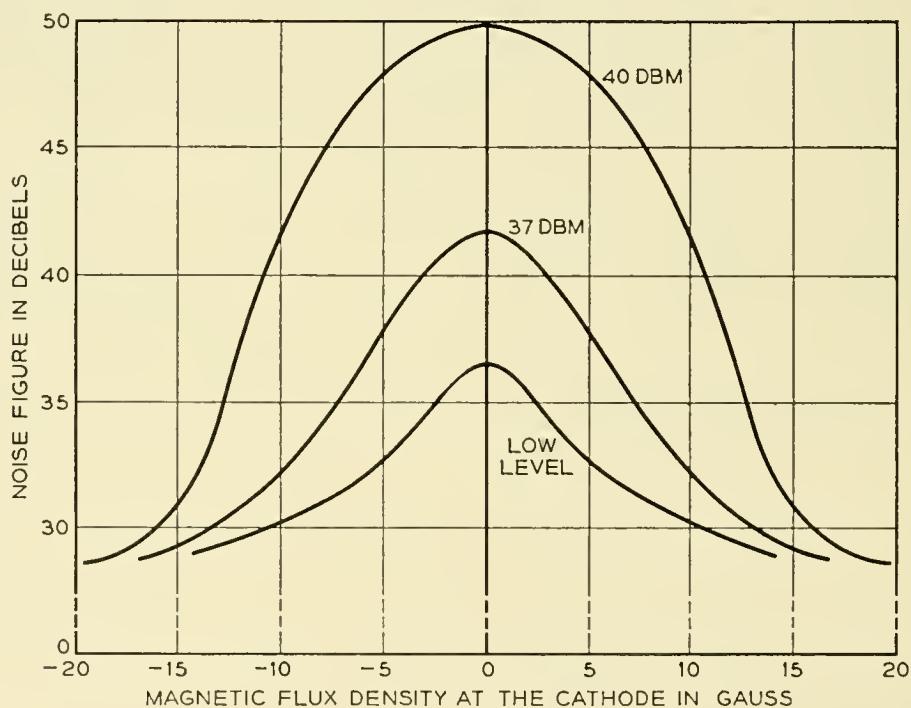


Fig. 39 — Noise figure as a function of magnetic flux density at the cathode for several values of rf power output. The flux density was varied by using an inductive heater through which ac current was passed. The present M1789 uses 19 gauss at the cathode, all of which is obtained from the focusing magnet — the heater now being non-inductive.

TWT in an FM radio relay was investigated briefly. In this measurement an FM receiver tuned to the carrier frequency was used to detect the noise modulation present in the TWT output. The noise figure was determined in the usual manner from the ratio of receiver outputs with the noise source turned off and on. When the TWT was operated in the linear region, this measurement gave the same result that our first method did. With the TWT operated in the non-linear region it gave a value within a few db of that obtained from the first method.

The cause of the high noise output observed for low magnetic flux densities at the cathode is at the present time not clearly understood. Fried at MIT and Ashkin and Rigrod at Bell Laboratories have all probed the beam formed by guns of the M1789 type and have found certain anomalous effects. Normally one would expect to find a standing wave of noise current along the electron beam. For the M1789 gun they find instead that after about two minima of the standing wave pattern, the noise current on the beam begins to grow and continues to do so until a saturation value is reached. The noise current at this saturation

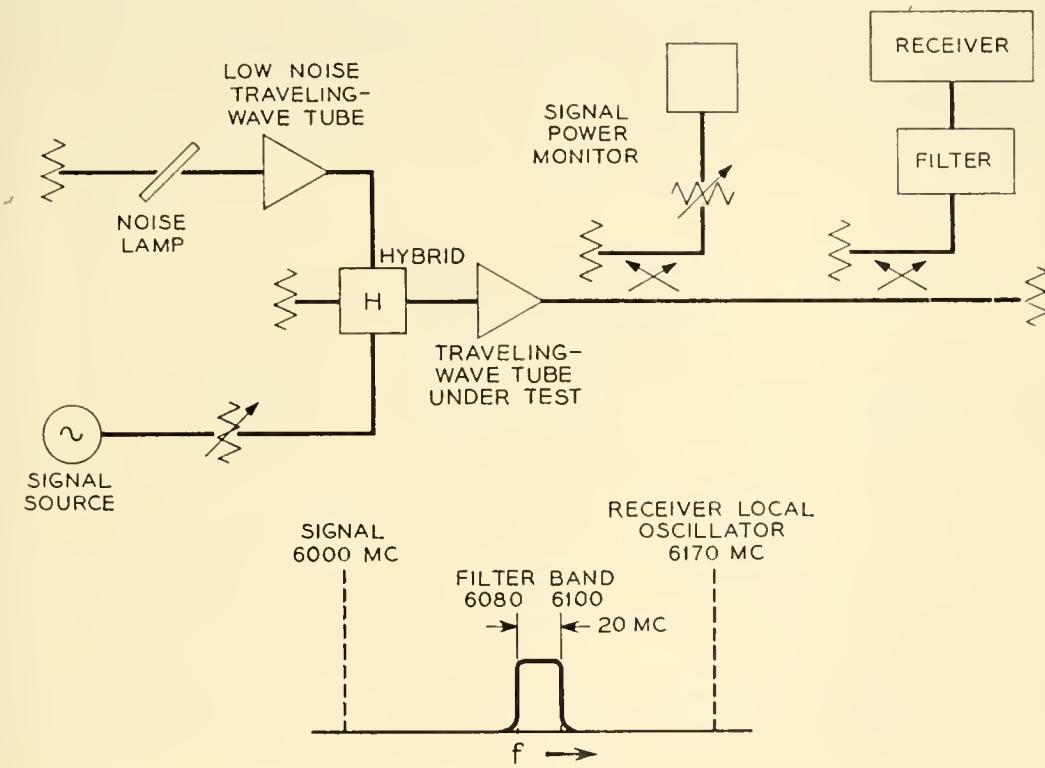


Fig. 40 — Block diagram of noise measuring equipment. The noise source consists of a fluorescent lamp the output of which is amplified by a low-noise TWT so as to bring the noise level to about 35 db above kTB at the M1789 input. The output from the M1789 is passed through a 20-me bandpass filter which eliminates both the single frequency test signal and the noise in the image band of the receiver. The noise figure is measured by noting the difference in noise level at the receiver output with the noise source off and on, in a manner similar to that used in a conventional noise figure measurement.

value may be considerably higher than the original average noise level. As is the case with the noise figure in the M1789, the growing noise current has been found to be very sensitive to magnetic field at the cathode. By allowing sufficient field to link the cathode, the growing noise current can be eliminated leaving the normal noise current standing wave pattern on the beam. This phenomenon is not peculiar to the M1789 gun. It has been observed by various workers at MIT⁶ and elsewhere on other guns producing beams with comparable current densities. A satisfactory explanation for it has not, at the time of this writing, been arrived at. It seems safe to say, however, that the growing noise current on the beam is the source of the high noise figures obtained in the M1789 when the cathode is completely shielded from the magnetic field.

4.5 Intermodulation

It has been found that certain intermodulation effects in the M1789 can be predicted from a knowledge of the compression and AM-to-PM conversion. Alternatively, these effects can be used to determine compression and AM-to-PM conversion. The procedure to be described has the advantage of being simple to implement as compared with the phase bridge arrangement of Fig. 24.

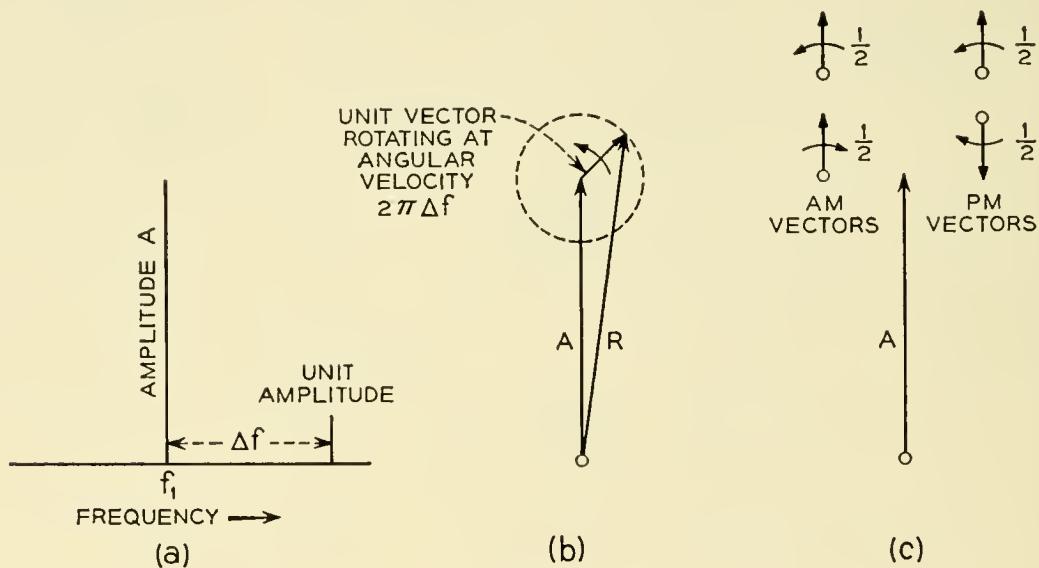


Fig. 41

- (a) Spectrum of input signal to amplifier.
- (b) Vector diagram of two input signals and the resultant signal (R) in a frame of reference rotating at an angular velocity $2\pi\Delta f$. Dotted line is the locus of the resultant signal.
- (c) The rotating vector of the preceding diagram can be broken down into a set of two vectors representing amplitude modulation and a set of two vectors representing frequency or phase modulation.

Intermodulation effects are ordinarily complicated and results are very hard to predict from single frequency measurements on an amplifier. For a TWT, however, one case — that in which two signals of very different amplitude are passed through the tube — can be treated simply. Consider an input to a TWT consisting of two signals at frequencies f_1 and $f_1 + \Delta f$ with the signal at f_1 being very much larger in amplitude. The composite signal applied to the amplifier will then be a signal at frequency f_1 which is amplitude and phase modulated at a rate Δf in an amount proportional to the relative magnitudes of the two signals. This can be represented vectorially as shown in Fig. 41(a) and b. In this figure the amplitude of the signal $f_1 + \Delta f$ has been normalized to unity. "A" thus represents the ratio of the larger to the smaller signal. The locus of the resultant signal is shown by the dotted line. The single rotating vector can be considered as the sum of vectors at $f_1 + \Delta f$ and $f_1 - \Delta f$ as shown in Fig. 41(c). One set of vectors produces PM and the other AM. The AM and PM vectors cancel at $f_1 - \Delta f$ and add at $f_1 + \Delta f$.

Suppose this signal is put through an amplifier operating in compression. For the time being let us assume this amplifier has no AM-to-PM conversion. The compression in the amplifier will operate on the AM sidebands of the signal but will leave the PM sidebands unaffected. Let us define the quantity c as a measure of compression in the amplifier by

$$c = 1 - \frac{\Delta V_0/V_0}{\Delta V_i/V_i} \quad (1)$$

where V_0 is the output voltage, V_i input voltage, and ΔV_0 is the change in output voltage for a change ΔV_i in the input voltage. This quantity is the per cent compression used in Section 4.2 divided by 100. If the signal in Fig. 41 is put through the amplifier while it is in compression, and the level of the signal at f_1 is subsequently brought back to amplitude A , we would then expect to have the situation shown in Fig. 42. Each AM sideband component has been multiplied by the factor $(1-c)$. The locus of the composite signal is now elliptical. Let S_1 and S_2 be the magnitude of the sidebands at $f_1 + \Delta f$ and $f_1 - \Delta f$ respectively. From Fig. 42 it is seen that

$$S_1 = \frac{1}{2} + \frac{1}{2}(1 - c) = 1 - c/2 \quad (2)$$

$$S_2 = \frac{1}{2} - \frac{1}{2}(1 - c) = c/2 \quad (3)$$

When $c = 0$, the amplifier is operating in the linear region and $S_1 = 1$,

$S_2 = 0$. This is the condition in Fig. 41. When the amplifier is operating as a perfect limiter, $c = 1$ and $S_1 = S_2 = 0.5$. Thus, in this case, the sideband S_1 is down 6 db from its value when the amplifier is operating in the linear region.

When there is conversion of AM-to-PM in the amplifier, the situation becomes somewhat more complex. Suppose an AM signal is fed into the amplifier and that its voltage is given by

$$V = V_1(1 + \alpha \sin \omega_m t) \sin \omega_c t \quad (4)$$

where ω_c and ω_m are the carrier and modulating radian frequencies and V_1 and α are constants. The outputs will be given by

$$V = KV_1[1 + \alpha(1 - c) \sin \omega_m t] \sin(\omega_c t + k_p \alpha \sin \omega_m t) \quad (5)$$

Here K is the amplification, c is the compression factor and k_p is a factor which is a measure of the AM-to-PM conversion. It is seen that k_p is the output phase change for a given fractional input change α . Thus

$$k_p = \frac{\Delta\theta}{\alpha} \quad (6)$$

where $\Delta\theta$ is the phase change in radians caused by a fractional input change α . Later on it will be desired to express k_p in terms of degrees phase shift per db change in input amplitude. To express α in db we

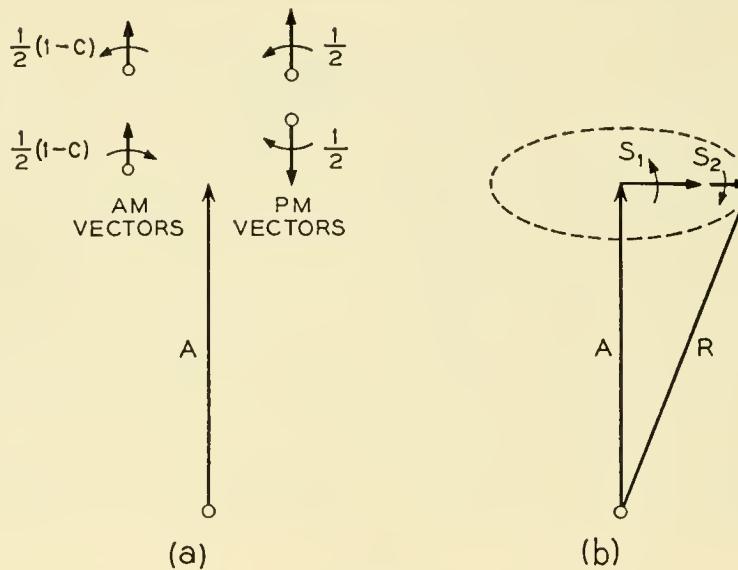


Fig. 42

(a) After passing through an amplifier in compression the AM sidebands are reduced in amplitude but the PM sidebands are unaffected. The lower two sidebands which represent a signal at frequency $f_1 - \Delta f$ no longer cancel and so there is a net signal at that frequency.

(b) The locus of the resultant signal now assumes an elliptical shape.

must evaluate $20 \log_{10} (1 + \alpha)$. The quantity $\log_e (1 + \alpha)$ can be expanded in a series to give

$$\log_e (1 + \alpha) = \alpha - \frac{1}{2} \alpha^2 + \frac{1}{3} \alpha^3 + \dots$$

As long as $\alpha \ll 1$, we can approximate it by taking only the first term of the above expression. Converting to the base ten and converting $\Delta\theta$ from radians as it appears in (6) to degrees, we find that

$$k_p = 0.152 \frac{\Delta\theta \text{ (in degrees)}}{\Delta \text{ input level (in db)}} \quad (7)$$

Now let us consider the case in which the signal of Fig. 41 is put through an amplifier having AM-to-PM conversion. Fig. 43 shows the vector picture of the resulting signal after the level of the signal at f_1 has been brought back to amplitude A . In this case the original PM sidebands and the compressed AM sidebands are the same as in Fig. 42, but there is now an additional set of PM sidebands as a result of the AM-to-PM conversion. Since the peak deviation of output phase due to this latter set of sidebands comes when the instantaneous amplitude is either a maximum or a minimum, they are 90 degrees out of phase with the other two sets of sidebands. From Fig. 43 it is seen that we can write

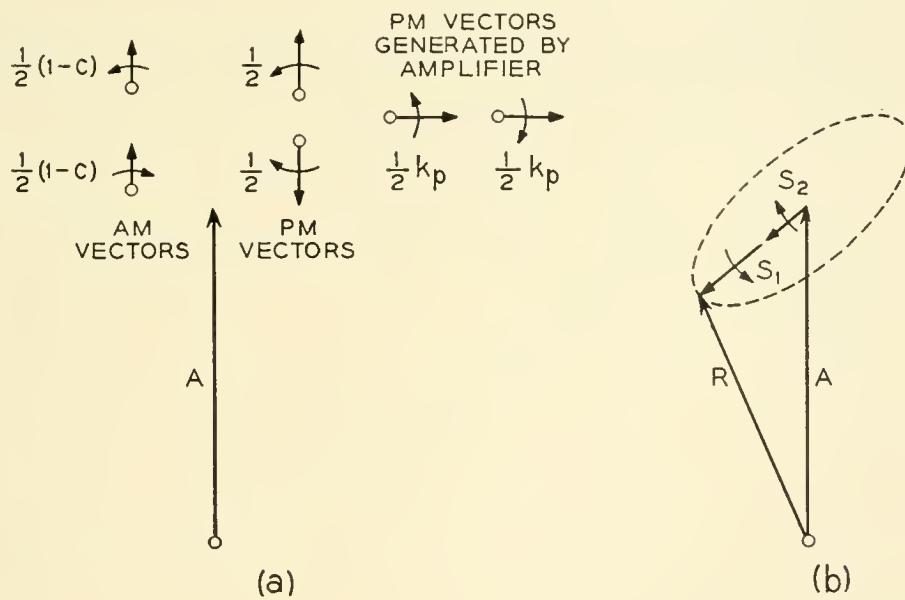


Fig. 43

(a) After passing through an amplifier having both compression and amplitude to phase conversion, the AM vectors are reduced in magnitude and a new set of PM vectors have appeared.

(b) The locus of the resultant signal of the vectors shown above is elliptical but the axis is tilted with respect to vector A.

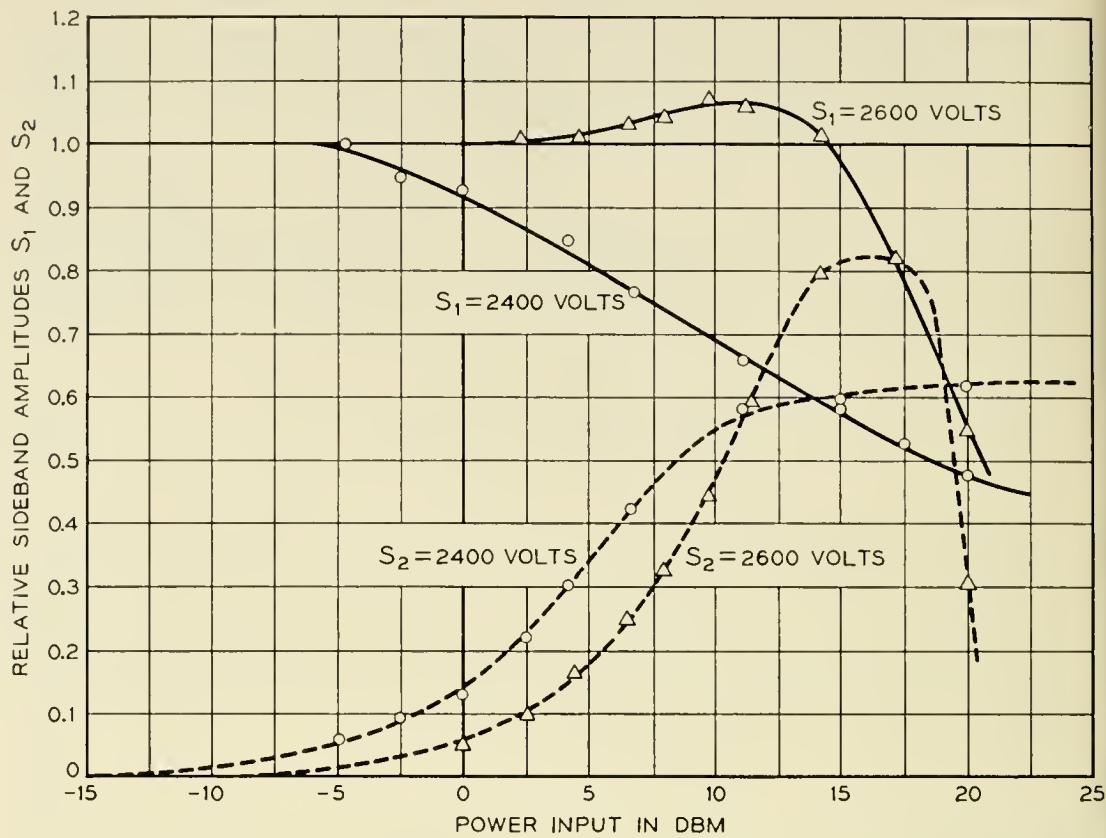


Fig. 44 — Relative side band amplitudes S_1 and S_2 for the M1789 as a function of power input for two values of helix voltage.

for the sideband amplitudes S_1 and S_2 at $f_1 + \Delta f$ and $f_1 - \Delta f$ respectively

$$S_1^2 = [\frac{1}{2} + \frac{1}{2}(1 - c)]^2 + \left[\frac{k_p}{2} \right]^2 = (1 - c/2)^2 + \left(\frac{k_p}{2} \right)^2 \quad (8)$$

$$S_2^2 = [\frac{1}{2} - \frac{1}{2}(1 - c)]^2 + \left[\frac{k_p}{2} \right]^2 = (c/2)^2 + \left(\frac{k_p}{2} \right)^2 \quad (9)$$

Solving for c and k_p we obtain

$$c = 1 - (S_1^2 - S_2^2) \quad (10)$$

$$k_p = 2 \left[S_1^2 - \left(\frac{1 + S_1^2 - S_2^2}{2} \right)^2 \right]^{1/2} \quad (11)$$

Thus we see that from a measurement of the amplitudes S_1 and S_2 the values of c and k_p can be determined.

To check the validity of this approach to intermodulation, we determined the values of compression and AM-to-PM conversion for an M1789 from an intermodulation measurement and compared them with values obtained using the phase bridge set-up described in Section 4.2. In the intermodulation measurement the two signals were 100 mc apart

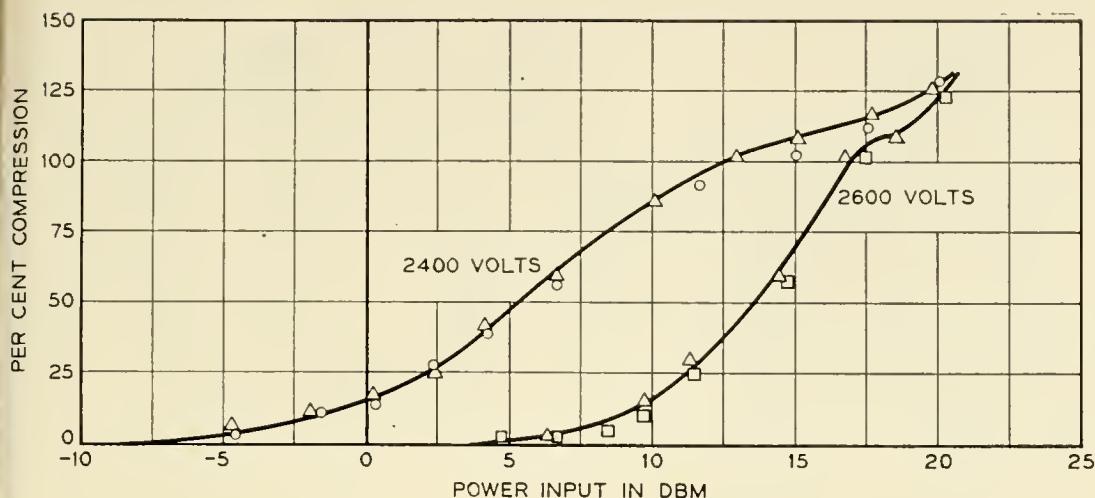


Fig. 45 — Compression as a function of input level for two values of helix voltage. Triangles represent data obtained with the test set of Fig. 24. Circles and squares represent data obtained by the two signal intermodulation measurement.

in frequency and 30 db different in level. From measurements of signal strength at the various frequencies involved, the magnitudes of S_1 and S_2 were determined with the results shown in Fig. 44. From these results the values of c and k_p were calculated and then converted to % compression and degrees per db in order to compare with the results of

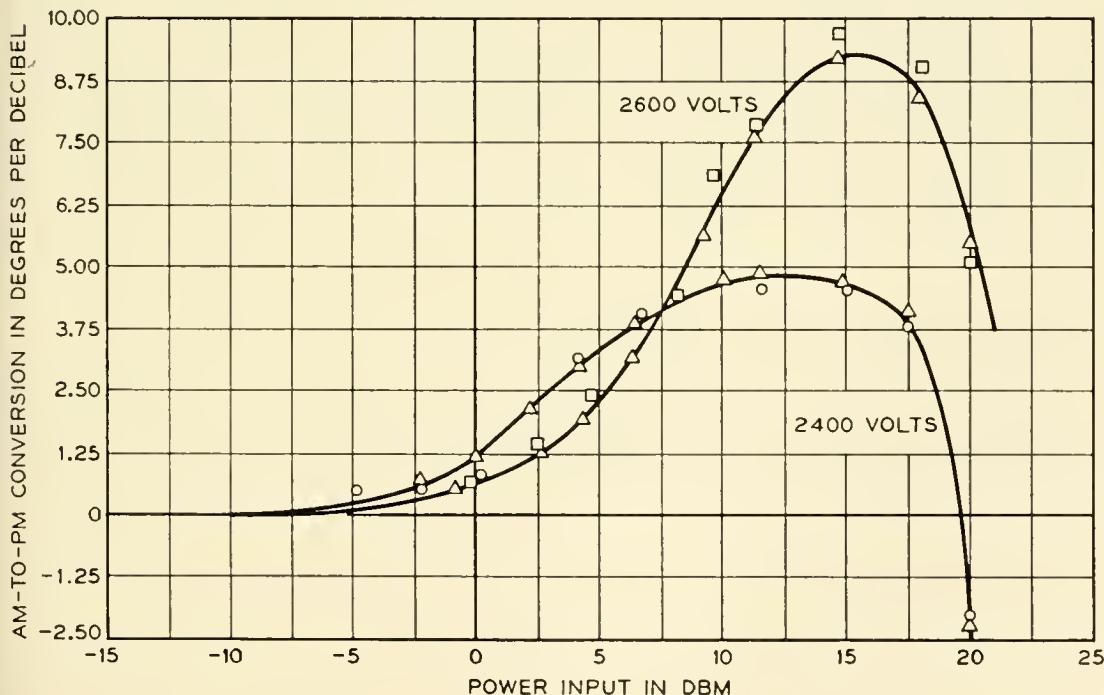


Fig. 46 — Conversion of amplitude modulation to phase modulation as a function of input level for two values of helix voltage. Triangles represent data obtained with the test set of Fig. 24. Circles and squares represent data obtained by the two signal intermodulation measurement.

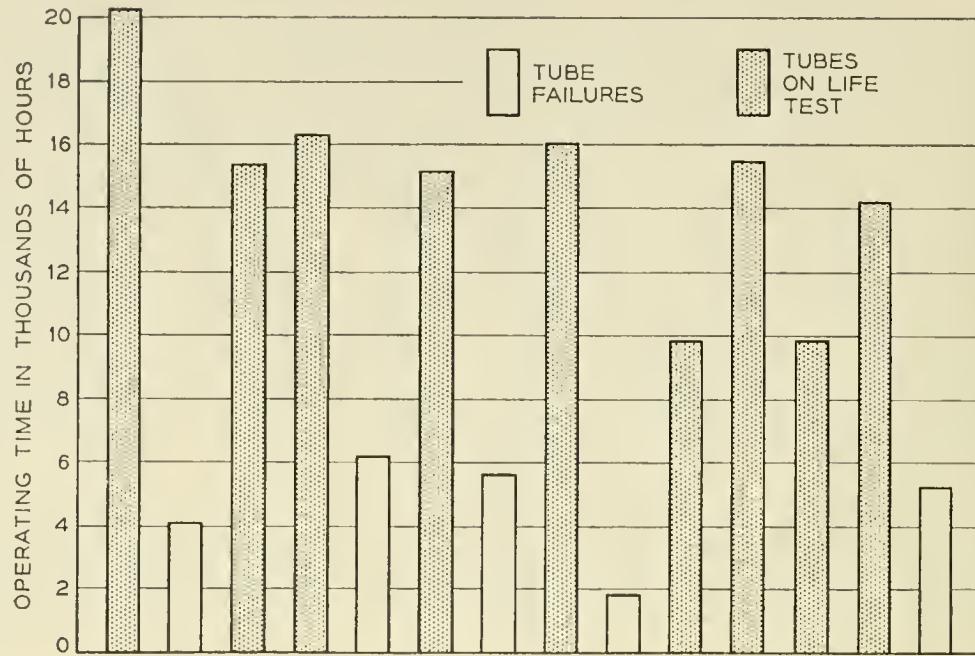


Fig. 47 — Life test results. The open bars indicate tubes that have failed; the solid bars tubes that were operating as of May 1, 1956. These tubes were operated with cathode temperatures between 720° and 760°C .

Figs. 23(c) and 23(e). The latter curves are repeated as Figs. 45 and 46 with the experimental points calculated from S_1 and S_2 shown. It is seen that the results of the two types of measurements compare remarkably well considering that the calculations of c and k_p both require the subtraction of nearly equal quantities. Thus we may conclude that our method of considering the intermodulation is substantially correct and that we can obtain compression and AM-to-PM conversion from an intermodulation measurement.

V. LIFE TESTS

We feel that sufficient data have been accumulated to indicate that tube life in excess of 10,000 hours can be expected. Fig. 47 summarizes our life test experience. All tube failures were caused by cathode failure and these were evidently the result of exhaustion of coating. End of life for these tubes comes comparatively suddenly i.e., in a few hundred hours after the cathode current begins to drop. At this time the emission becomes non-uniform over the cathode surface with consequent beam defocusing and helix interception. This in turn causes gas to be released into the tube which then accelerates the cathode failure through cathode poisoning. The rf performance remained good over the tube life — the gain and output power actually increasing slightly near the end of life as the beam started to defocus.

VI. ACKNOWLEDGMENTS

The M1789 TWT is the outcome of an intensive effort which has included many individuals in addition to the authors. R. Angle, J. S. Gellatly, E. G. Olson, and R. G. Voss all have contributed to the mechanical design of the tube and to its reduction to practice. R. W. DeVido has materially assisted with the electrical testing. M. G. Bodmer and J. F. Riley have been responsible for setting up the life test program and J. C. Irwin and J. A. Saloom contributed importantly to the design work on the electron gun. P. P. Cioffi and M. S. Glass have been largely responsible for the design of the magnetic circuits and P. I. Sandsmark for the helix-to-waveguide transducers. D. O. Melroy studied the effects of positive ions and performed the experiments on ion bombardment referred to in Section III. D. R. Jordan contributed to the studies on noise. In addition to the above, the authors would like to thank E. D. Reed for his very helpful criticism of this manuscript.

APPENDIX I — GAIN CALCULATIONS

The gain calculations for the M1789 follow the procedure outlined by Pierce⁷ with some minor modifications. The steps involved in the gain calculations for the loss free region of the helix are as follows:

- (1) The experimental synchronous voltage is used to determine γa and the dielectric loading factor as defined by Tien.⁸
- (2) From γa the value of helix impedance K is obtained from Appendix VI of Pierce.⁷
- (3) The value of K is corrected using Tien's⁸ results and C is then calculated in the usual manner.
- (4) The number of wavelengths N_1 per inch of helix is obtained using the experimentally determined (from synchronous voltage) wavelength.
- (5) The value of ω_q/ω is determined. In this calculation the curves for ω_p/ω_q from Watkins⁹ are employed.
- (6) QC is determined from

$$QC = \left(\frac{1}{2C} \frac{\omega_q}{\omega} \right)^2$$

- (7) From QC , B is determined from Fig. 8.10 of Pierce⁷ and the gain BCN_1 in the loss free region is calculated.

In calculating the effect of the attenuator section, we have had to make some rather gross assumptions. Fortunately, it turns out that the

gain in the attenuator is a small fraction of the total gain in the tube so that the over-all gain is not particularly sensitive to the means we use for treating the attenuator. Essentially what we have done is to consider the high loss part of the attenuator as a severed helix region and the low loss part of the attenuator as a lossy helix region.

Fig. 48 shows the value of the growing wave parameter as a function of the loss parameter d for various values of QC as calculated from theory. Because of discontinuity losses to the growing wave as it propagates in a region of gradually increasing loss, the actual gain will be less than that calculated from Fig. 48. Some rather crude probe measurements have indicated that the effective x vs. d curve can be approximated by a straight line through the $d = 0$ and $d = 1$ points — the dotted line in Fig. 48.

Since the helix is effectively severed by the high loss portion of the attenuator we must subtract some discontinuity loss from the gain in the attenuator region. The effective drift length in the severed region is unknown so this discontinuity loss cannot be accurately calculated from the low-level theory. The discussion in chapter nine of Pierce⁷ indicates that an average value of about 6 db is reasonable.

An alternate method of treating the attenuator was also tried. In this calculation, the x vs. d curves in Fig. 48 were assumed to be correct to

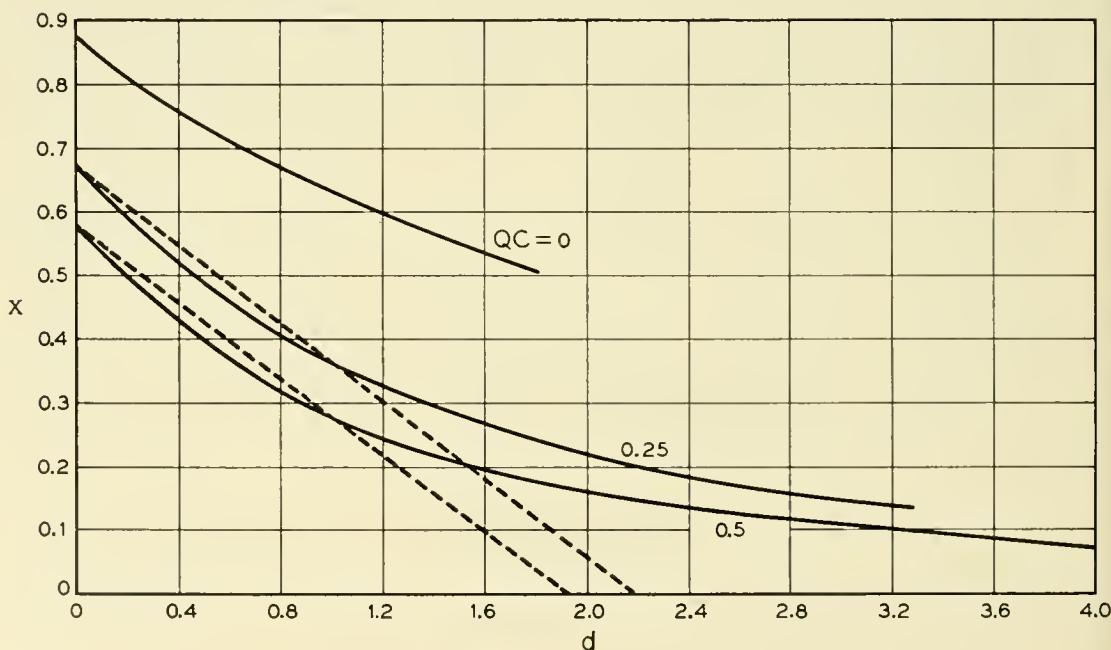


Fig. 48 — Curves of growing wave parameter x as a function of loss parameter d showing approximation (dotted lines) used in gain calculations for the M1789.

$d = 1$. The region for which $d > 1$ was considered as a severed helix region with 6-db discontinuity loss. Calculations using this procedure gave total gains for the TWT within a couple of db of the first method.

The remaining steps in calculating the gain of the TWT are therefore:

- (8) The quantity α is determined from the slope of the dotted lines in Fig. 48.
- (9) The length of helix, ℓ_e in the attenuator for which $x > 0$ is determined by using Fig. 48.
- (10) The total attenuation L , in the section of the attenuator effective in producing gain is calculated.
- (11) The initial loss parameter A is obtained from Fig. 94 of Pierce.⁷
- (12) The gain is calculated from

$$\text{Gain} = A - 6\text{db} + \alpha L + BCN_1 (3.5 + \ell_e)$$

where the six db is the discontinuity loss in the attenuator section and the 3.5 inches is the length of loss free helix.

GLOSSARY OF SYMBOLS

α	loss factor from Pierce ⁷
A	discontinuity loss parameter at input of helix from Pierce ⁷
B	magnetic flux density or the space charge parameter from Pierce ⁷
B_B	Brillouin flux density for a beam entirely filling the helix
C'	gain parameter from Pierce ⁷
a	helix radius
b	beam radius
d	loss parameter from Pierce ⁷
f	frequency
I_k	cathode current
I_a	accelerator current
I_h	helix current
I_c	collector current
k	$2\pi/\lambda_0$ where λ_0 is the free space wavelength
ℓ_e	length of helix attenuator in which gain is possible
L	loss in the part of the attenuator section which is capable of producing gain.
N	number of wavelengths in TWT
N_1	number of wavelengths on the helix per inch
QC	space charge parameter from Pierce ⁷
\underline{r}_a	anode radius of curvature of gun
r_c	cathode radius of curvature of gun
r_{min}	minimum beam radius from Pierce ¹⁰

r_c	cathode radius
r_{95}	radius at the beam minimum through which 95 per cent of the current flows
σ	standard deviation of electron trajectory
T_k	cathode temperature
V_a	accelerator voltage
V_h	helix voltage
V_c	collector voltage
x	growing wave parameter from Pierce
ω	radian frequency
ω_c	carrier radian frequency
ω_m	modulating signal radian frequency
ω_p	radian plasma frequency
ω_q	corrected radian plasma frequency
c	compression factor
k_p	AM-to-PM conversion factor
γ	radial propagation constant

REFERENCES

1. Cutler, C. C., Spurious Modulation of Electron Beams, Proc. I.R.E., **44**, pp. 61-64, Jan., 1956.
2. Danielson, W. E., Rosenfeld, J. L., and Saloom, J. A., A Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type, B.S.T.J. **35**, pp. 375-420, March, 1956.
3. Augustine, C. F., and Sloeum, A., 6KMC Phase Measurement System For Traveling-Wave Tubes, I.R.E. Trans. PGI-4, Oct., 1955.
4. Tien, P. K., A Large Signal Theory of Traveling-Wave Amplifiers, B.S.T.J., **35**, pp. 349-374, March, 1956.
5. Brangaccio, D. J., and Cutler, C. C., Factors Affecting Traveling-Wave Tube Power Capacity, I.R.E. Trans. PGED-3, June, 1953.
6. Smullin, L. D., and Fried, C., Microwave Noise Measurements on Electron Beams, I.R.E. Trans., PGED-4, Dec., 1954.
7. Pierce, J. R., Traveling-Wave Tubes, D. Van Nostrand, Inc., 1950.
8. Tien, P. K., Traveling-Wave Tube Helix Impedance, Proc. I.R.E., **41**, pp. 1617-1623, Nov., 1953.
9. Watkins, D. A., Traveling-Wave Tube Noise Figure, Proc. I.R.E., **40**, pp. 65-70, Jan., 1952.
10. Pierce, J. R., Theory and Design of Electron Beams, D. Van Nostrand, Inc., 1949.

Helix Waveguide

By S. P. MORGAN and J. A. YOUNG

(Manuscript received July 23, 1956)

Helix waveguide, composed of closely wound turns of insulated copper wire covered with a lossy jacket, shows great promise for use as a communication medium. The properties of this type of waveguide have been investigated using the sheath helix model. Modes whose wall currents follow the highly conducting helix have attenuation constants which are essentially the same as for copper pipe. The other modes have very large attenuation constants which depend upon the helix pitch angle and the electrical properties of the jacket. Approximate formulas are given for the propagation constants of the lossy modes. The circular electric mode important for long-distance communication has low loss for zero-pitch helices. The propagation constants of some of the lossy modes in helix waveguide of zero pitch have been calculated numerically, as functions of the jacket parameters and the guide size, in regions where the approximate formulas are no longer valid. Under certain conditions the attenuation constant of a particular mode may pass through a maximum as the jacket conductivity is varied.

GLOSSARY OF SYMBOLS

a	Inner radius of waveguide
$h = \beta - i\alpha$	Complex phase constant
n	Angular mode index
p	Denotes p_{nm} or p_{nm}' according to context
p_{nm}	m^{th} zero of $J_n(x)$
p_{nm}'	m^{th} zero of $J_n'(x)$
r, θ, z	Right-handed cylindrical coordinates
α	Attenuation constant
β	Phase constant
$\beta_0 = 2\pi/\lambda_0 = \omega(\mu_0\epsilon_0)^{1/2}$	Free-space phase constant
ϵ_0	Permittivity of interior medium
ϵ	Permittivity of exterior medium
ϵ'	ϵ/ϵ_0
ϵ''	$\sigma/\omega\epsilon_0$

ζ_1	$[\omega^2 \mu_0 \epsilon_0 - h^2]^{1/2}$
ζ_2	$[\omega^2 \mu_0 \epsilon_0 (\epsilon' - i\epsilon'') - h^2]^{1/2}$
λ_0	Free-space wavelength
$\lambda_c = 2\pi a/p$	Cutoff wavelength
μ_0	Permeability of interior and exterior media
$\nu = \lambda_0/\lambda_c = p\lambda_0/2\pi a$	Cutoff ratio
$\xi + i\eta$	$\frac{(\epsilon' - 1 + \nu^2 - i\epsilon'')^{1/2}}{\epsilon' - i\epsilon''}$
Π	Electric Hertz vector
Π^*	Magnetic Hertz vector
σ	Conductivity of exterior medium
ψ	Pitch angle of helix
ω	Angular frequency
$e^{i\omega t}$	Harmonic time dependence assumed throughout
$J_n(x)$	Bessel function of the first kind
$J_n'(x)$	$dJ_n(x)/dx$
$H_n^{(2)}(x)$	Hankel function of the second kind
$H_n^{(2)'}(x)$	$dH_n^{(2)}(x)/dx$

MKS rationalized units are employed throughout. Superscripts i and e are used to indicate the interior and exterior regions.

I. INTRODUCTION AND SUMMARY

Propagation of the lowest circular electric mode (TE_{01}) in cylindrical pipe waveguide holds great promise for low-loss long distance communication.^{1, 2} For example, the TE_{01} mode has a theoretical heat loss of 2 db/mile in waveguide of diameter 6 inches at a frequency of 5.5 kmc/s, and the loss decreases with increasing frequency. Increased transmission bandwidth, reduced delay distortion, and reduced waveguide size for a given attenuation are factors favoring use of the highest practical frequency of operation. An increased number of freely propagating modes and smaller mechanical tolerances are the associated penalties. Any deviation of the waveguide from a straight circular cylinder gives rise to signal distortions because of mode conversion-reconversion effects.

One solution to mode conversion-reconversion problems is to obtain a waveguide having the desired low attenuation properties of the TE_{01} mode in metallic cylindrical waveguide and very large attenuation for all other modes, the unwanted modes.^{1, 3} The low loss of the circular electric modes in ordinary round guide is the result of having only cir-

¹ S. E. Miller, B.S.T.J., **33**, pp. 1209-1265, 1954.

² S. E. Miller and A. C. Beck, Proc. I.R.E., **41**, pp. 348-358, 1953.

³ S. E. Miller, Proc. I.R.E., **40**, pp. 1104-1113, 1952.

cumferential current flow at the boundary wall. All other modes in round guide have a longitudinal current present at the wall. Thus the desired attenuation properties can be obtained by providing a highly conducting circumferential path and a resistive longitudinal path for the wall currents. This is done in the spaced-disk line by sandwiching lossy layers between coaxially arranged annular copper disks.⁴ Another possibility which has been suggested is a helix having a small pitch.

Helix waveguide, formed by winding insulated wire on a removable mandrel and coating the helix with lossy material, has been made at the Holmdel Radio Research Laboratory. Wires of various cross sections and sizes have been used to wind helices varying from $\frac{7}{16}$ to 5 inches in diameter, which have been tested at frequencies from 9 to 60 kmc/s. Pitch angles of from nearly 0° (wire in a plane perpendicular to the axis of propagation) to 90° (wire parallel to the axis of propagation) have been used. The helices having the highest attenuation for the unwanted modes while maintaining low loss for the TE_{01} mode are those wound with the smallest pitch from insulated wire of diameter 10 to 3 mils (American Wire Gauge Nos. 30 to 40). The high attenuation properties for unwanted modes also depend markedly on the electrical properties of the jacket surrounding the helix.

In this paper the normal modes of helix waveguide are determined using the sheath helix approximation, a mathematical model in which the helical winding is replaced by an anisotropic conducting sheath. A brief formulation of the boundary value problem leads to an equation which determines the propagation constants of modes in the helix guide. Since the equation is not easy to solve numerically, approximations are presented which show the effects of the pitch angle, the diameter, the conductivity and dielectric constant of the jacket, and the wavelength, when the conductivity of the jacket is sufficiently high.

By proper choice of the pitch angle and, in some instances, of the polarization, a helix waveguide can be made to propagate any mode of ordinary round guide, with an attenuation constant which should be essentially the same as in solid copper pipe. The pitch is chosen so that the wall currents associated with the desired mode follow the direction of the conducting wires. The losses to the other modes are in general much higher, and are determined by both the pitch angle and the jacket material.

Special attention is given in the present work to the limiting case of a helix of zero pitch, since the attenuation constant of the TE_{01} mode will be smallest when the pitch angle is as small as possible. To explore the

⁴ Reference 3, p. 1111.

region where the approximate formulas for the propagation constants of the lossy modes break down, some numerical results have been obtained for helices of zero pitch using an IBM 650 magnetic drum calculator. Tables and curves are given showing the propagation constants of various modes in such a waveguide, as functions of the electrical properties of the jacket and for three different ratios of radius/wavelength. In many cases it is found that the attenuation constant of a given mode passes through a maximum as the jacket conductivity is varied, the other parameters remaining fixed. The numerical calculations indicate that it is possible to get unwanted mode attenuations several hundred to several hundred thousand times greater than the TE₀₁ attenuation for the size waveguide that looks most promising for low-loss communication.

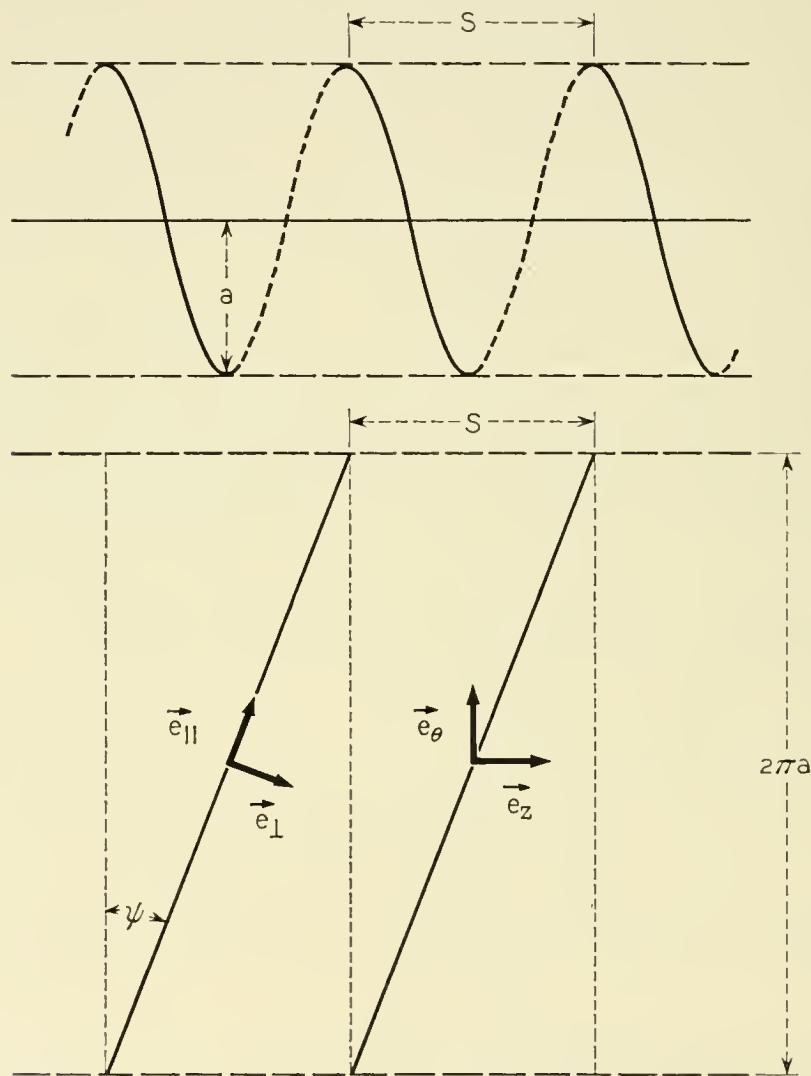


Fig. 1 — Schematic diagrams of the helical sheath and the helical sheath developed, showing the unit vectors and the periodicity.

II. SHEATH HELIX BOUNDARY VALUE PROBLEM

Ordinary cylindrical waveguide consists of a circular cylinder of radius a , infinite length, and zero (or very small) conductivity, imbedded in an infinite* homogeneous conducting medium. The sheath helix waveguide has the same configuration plus the additional property that at radius a dividing the two media, there is an anisotropic conducting sheath which conducts perfectly in the helical direction and does not conduct in the perpendicular direction. The attenuation and phase constants are determined by solving Maxwell's equations in cylindrical coordinates and matching the electric and magnetic fields at the wall of the guide.

The helix of radius a and pitch angle $\psi = \tan^{-1} s/2\pi a$ is shown in the upper part of Fig. 1. The developed helix as viewed from the inside when cut by a plane of constant θ and unrolled is shown in the lower part of the illustration. A new set of unit vectors \vec{e}_{\parallel} and \vec{e}_{\perp} parallel and perpendicular respectively to the helix direction is introduced. These are related to \vec{e}_r , \vec{e}_{θ} , and \vec{e}_z by

$$\begin{aligned}\vec{e}_r \times \vec{e}_{\parallel} &= \vec{e}_{\perp} \\ \vec{e}_{\parallel} &= \vec{e}_z \sin \psi + \vec{e}_{\theta} \cos \psi \\ \vec{e}_{\perp} &= \vec{e}_z \cos \psi - \vec{e}_{\theta} \sin \psi\end{aligned}$$

The boundary conditions at $r = a$ are

$$\begin{aligned}E_{\parallel}^i &= E_{\parallel}^e = 0 \\ E_{\perp}^i &= E_{\perp}^e \\ H_{\parallel}^i &= H_{\parallel}^e\end{aligned}$$

where the superscript i refers to the interior region, $0 \leq r \leq a$, and the superscript e refers to the exterior region, $a \leq r \leq \infty$. An equivalent set of boundary conditions in terms of the original unit vectors is

$$\begin{aligned}E_z^i \tan \psi + E_{\theta}^i &= 0 \\ E_z^e \tan \psi + E_{\theta}^e &= 0 \\ E_z^i &= E_z^e \\ H_z^i \tan \psi + H_{\theta}^i &= H_z^e \tan \psi + H_{\theta}^e\end{aligned}\tag{1}$$

We are looking for solutions which are similar to the modes of or-

* The assumption of an infinite external medium is made to simplify the mathematics. The results will be the same as for a finite conducting jacket which is thick enough so that the fields at its outer surface are negligible.

ordinary waveguide, i.e., "fast" modes as contrasted with the well-known "slow" modes used in traveling-wave tubes.^{5, 6} To solve the problem we follow the procedure set up by Stratton⁷ for the ordinary cylindrical waveguide boundary problem. The fields \vec{E} and \vec{H} are derived from an electric Hertz vector $\vec{\Pi}$ and a magnetic Hertz vector $\vec{\Pi}^*$ by

$$\begin{aligned}\vec{E} &= \vec{\nabla} \times \vec{\nabla} \times \vec{\Pi} - i\omega\mu\vec{\nabla} \times \vec{\Pi}^* \\ \vec{H} &= (\sigma + i\omega\epsilon)\vec{\nabla} \times \vec{\Pi} + \vec{\nabla} \times \vec{\nabla} \times \vec{\Pi}^*\end{aligned}\quad (2)$$

where

$$\begin{aligned}\vec{\Pi} &= \vec{e}_z \Pi_z \\ \vec{\Pi}^* &= \vec{e}_z \Pi^*\end{aligned}\tag{3}$$

and, assuming a time dependence $\exp(i\omega t)$,

$$\begin{aligned}
 \Pi_z^i &= \sum_{n=-\infty}^{\infty} a_n^i J_n(\xi_1 r) e^{-ihz-in\theta} \\
 \Pi_z^e &= \sum_{n=-\infty}^{\infty} a_n^e H_n^{(2)}(\xi_2 r) e^{-ihz-in\theta} \\
 \Pi_z^{*i} &= \sum_{n=-\infty}^{\infty} b_n^i J_n(\xi_1 r) e^{-ihz-in\theta} \\
 \Pi_z^{*e} &= \sum_{n=-\infty}^{\infty} b_n^e H_n^{(2)}(\xi_2 r) e^{-ihz-in\theta}
 \end{aligned} \tag{4}$$

In these expressions

$$\begin{aligned}\xi_1^2 &= \omega^2 \mu_0 \epsilon_0 - h^2 \\ \xi_2^2 &= \omega^2 \mu_0 \epsilon_0 (\epsilon' - i\epsilon'') - h^2 \\ \epsilon' - i\epsilon'' &= \epsilon/\epsilon_0 - i\sigma/\omega\epsilon_0\end{aligned}$$

where the interior region is assumed to have permittivity ϵ_0 and permeability μ_0 , while the exterior region has permittivity ϵ , permeability μ_0 , and conductivity σ . The superscripts i and e refer to the interior and exterior regions respectively, and the a 's and b 's are amplitude coefficients.

⁵ J. R. Pierce, Proc. I.R.E., 35, pp. 111-123, 1947.

⁶ S. Sensiper, Electromagnetic Wave Propagation on Helical Conductors, Sc.D. thesis, M.I.T., 1951. In Appendix B of this reference, Sensiper shows that when the interior and exterior media are the same, only slow waves will exist except in special cases. Fast guided waves become possible if the conductivity of the exterior medium is sufficiently high.

⁷ J. A. Stratton, Electromagnetic Theory, McGraw-Hill, New York, 1941, pp. 524-527. Note that Stratton uses the time dependence $\exp(-i\omega t)$.

Attention is restricted to waves traveling in the positive z -direction, which are represented by the factor $\exp(-ihz)$, where $h (= \beta - i\alpha)$ is the complex phase constant. However it is necessary to consider both right and left circularly polarized waves; this accounts for the use of both positive and negative values of n .

Substitution of (2), (3), and (4) into the boundary conditions (1) leads to the following set of equations:

$$\begin{aligned} & \left[\xi_1^2 \tan \psi - \frac{hn}{a} \right] J_n(\xi_1 a) a_n^i + i\omega\mu_0\xi_1 J_n'(\xi_1 a) b_n^i = 0 \\ & \left[\xi_2^2 \tan \psi - \frac{hn}{a} \right] H_n^{(2)}(\xi_2 a) a_n^e + i\omega\mu_0\xi_2 H_n^{(2)\prime}(\xi_2 a) b_n^e = 0 \\ & \xi_1^2 J_n(\xi_1 a) a_n^i - \xi_2^2 H_n^{(2)}(\xi_2 a) a_n^e = 0 \\ & -i\omega\epsilon_0\xi_1 J_n'(\xi_1 a) a_n^i + \left[\xi_1^2 \tan \psi - \frac{hn}{a} \right] J_n(\xi_1 a) b_n^i \\ & + (\sigma + i\omega\epsilon)\xi_2 H_n^{(2)\prime}(\xi_2 a) a_n^e - \left[\xi_2^2 \tan \psi - \frac{hn}{a} \right] H_n^{(2)}(\xi_2 a) b_n^e = 0 \end{aligned} \quad (5)$$

If the conductivity of the exterior region is infinite, it is possible to satisfy the boundary conditions with only one of the amplitude coefficients different from zero; for example

$$\begin{aligned} b_n^i = a_n^e = b_n^e = 0 & \quad a_n^i = a_n^e = b_n^e = 0 \\ a_n^i \neq 0 \quad \text{or} & \quad b_n^i \neq 0 \\ J_n(\xi_1 a) = 0 & \quad J_n'(\xi_1 a) = 0 \end{aligned}$$

The first case corresponds to TM modes and the second to TE modes in a perfectly conducting circular guide. Linearly polarized modes may be represented as combinations of terms in a_n^i and a_{-n}^i , or b_n^i and b_{-n}^i .

If the exterior region is not perfectly conducting, one can still find solutions having the fields confined to the interior region by properly choosing the angle of the perfectly conducting helical sheath. For example, it is easy to verify that equations (5) are satisfied under the following conditions:

$$\begin{aligned} a_n^i = a_n^e = b_n^e = 0 \\ b_n^i \neq 0 \\ \tan \psi = \frac{hn}{\xi_1^2 a} \\ J_n'(\xi_1 a) = 0 \end{aligned}$$

If $n \neq 0$, these conditions correspond to circularly polarized TE_{nm} waves, in which the wall currents follow the direction of the conducting sheath. If $n = 0$, then $\psi = 0$, and one has TE_{0m} modes with circumferential currents only.

The equations can also be satisfied with

$$\begin{aligned} b_n^i &= a_n^e = b_n^e = 0 \\ a_n^i &\neq 0 \\ \psi &= 90^\circ \\ J_n(\xi_1 a) &= 0 \end{aligned}$$

corresponding to the TM_{nm} modes (either circularly or linearly polarized) of a perfectly conducting pipe, which are associated with longitudinal wall currents only.

In the general case when the jacket is not perfectly conducting and the helix pitch angle is not restricted to special values, it is necessary to solve (5) simultaneously for the field amplitudes. The equations admit a nontrivial solution if and only if the determinant of the coefficients of the a 's and b 's vanishes. The transcendental equation which results from equating the determinant of the coefficients to zero is

$$\begin{aligned} &\xi_2 \left[\left(\xi_1 \tan \psi - \frac{hn}{\xi_1 a} \right)^2 \frac{J_n(\xi_1 a)}{J_n'(\xi_1 a)} - \omega^2 \mu_0 \epsilon_0 \frac{J_n'(\xi_1 a)}{J_n(\xi_1 a)} \right] \\ &= \xi_1 \left[\left(\xi_2 \tan \psi - \frac{hn}{\xi_2 a} \right)^2 \frac{H_n^{(2)}(\xi_2 a)}{H_n^{(2)\prime}(\xi_2 a)} - \omega^2 \mu_0 \epsilon_0 (\epsilon' - i\epsilon'') \frac{H_n^{(2)\prime}(\xi_2 a)}{H_n^{(2)}(\xi_2 a)} \right] \end{aligned} \quad (6)$$

The solution of this equation determines the propagation constant ih and therefore the attenuation and phase constants α and β . When ih has been obtained, it is a straightforward matter to determine the a and b coefficients from equations (5) and the electric and magnetic fields from (2), (3), and (4).

⁸It is well known⁸ that the only pure TE or TM modes that can exist in a circular waveguide with walls of finite conductivity are the circularly symmetric TE_{0m} and TM_{0m} modes. The other modes are all mixed modes whose fields are not transverse with respect to either the electric or the magnetic vector. In general the modes of helix waveguide are also mixed modes, and no entirely satisfactory scheme for labeling them has been proposed. In the present paper we shall call the modes TE_{nm} or TM_{nm} according to the limits which they approach as the jacket conductivity becomes infinite, even though they are no longer transverse and their

⁸ Reference 7, p. 526.

field patterns may be quite different when the jacket is lossy. This system is not completely unambiguous, because as will appear in Section IV the mode designations thus obtained are not always unique. However it is a satisfactory way to identify the modes so long as the jacket conductivity is high enough for the loss to be treated as a perturbation. Approximations derived on this basis are presented in the next section.

III. APPROXIMATE EXPRESSIONS FOR PROPAGATION CONSTANTS

If the jacket were perfectly conducting, the helix waveguide modes would be the same as in an ideal circular waveguide, with propagation constants given by

$$ih = i\beta_{nm} = i(2\pi/\lambda_0)(1 - \nu^2)^{1/2}$$

where

$$\nu = \lambda_0/\lambda_c = p\lambda_0/2\pi a$$

$p = m^{\text{th}}$ zero of $J_n(x)$ for TM_{nm} mode, or m^{th} zero of $J_n'(x)$ for TE_{nm} mode

If the jacket conductivity is sufficiently large, approximate solutions of (6) may be found by replacing $H_n^{(2)}(\xi_2 a)$ and $H_n^{(2)\prime}(\xi_2 a)$ with their asymptotic expressions, and expanding $J_n(\xi_1 a)$ or $J_n'(\xi_1 a)$ in a Taylor series near a particular zero. This calculation is carried out in the appendix. The propagation constant may be written in the form

$$ih = \alpha + i(\beta_{nm} + \Delta\beta)$$

where to first order the perturbation terms are

TM_{nm} modes

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}} \frac{1}{1 + \tan^2 \psi} \quad (7a)$$

TE_{nm} modes

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}} \frac{\nu^2 p^2}{p^2 - n^2} \frac{[\tan \psi - n(1 - \nu^2)^{1/2}/p\nu]^2}{1 + \tan^2 \psi} \quad (7b)$$

and

$$\xi + i\eta = (\epsilon' - i\epsilon'')^{-1/2}$$

$$\epsilon' = \epsilon/\epsilon_0, \quad \epsilon'' = \sigma/\omega\epsilon_0$$

The approximations made in deriving (7) are discussed in the appen-

dix. In practice, the range of validity of these expressions is usually limited by the criterion

$$\frac{a(1 - \nu^2)^{1/2}}{\nu} |\alpha + i\Delta\beta| \ll 1 \quad (8)$$

The numerical calculations described in Section IV indicate that the approximations are good so long as the left-hand side of (8) is less than about 0.1, and that they break down a little sooner for TE modes than for TM modes.

Inspection of (7) reveals three cases of particular interest, namely $\psi = 0^\circ$, $\psi = \tan^{-1} n(1 - \nu^2)^{1/2}/p\nu$, and $\psi = 90^\circ$. These cases, which were mentioned in Section II and are discussed again below, correspond to preferential propagation of certain modes, in which the wall currents follow the direction of the conducting helix. The preferred modes have zero attenuation in the present treatment because the helical sheath is assumed to be perfectly conducting. In practical helices wound from insulated copper wire the loss should be only slightly greater than in round copper pipe of the same diameter. The slight increase (of magnitude 10 per cent to 30 per cent) is due to the slightly nonuniform current distribution in the wires, an effect that can be kept small by keeping the gaps between the wires of the helix small. In general the attenuation constants of modes whose wall currents do not follow the helix are orders of magnitude larger than the attenuation constants of the preferred modes.

$$\psi = 0^\circ$$

The circular electric (TE_{0m}) modes have attenuation constants substantially the same as in solid copper pipe. The additional TE_{0m} loss if the pitch angle is not quite zero is proportional to $\tan^2 \psi$. This added loss can be made very small by using fine wire for winding the helix.

The losses for the unwanted modes can be made large by a proper choice of jacket material. When $\psi = 0$, equations (7) yield

TM_{nm} modes

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}} \quad (9a)$$

TE_{nm} modes

$$\alpha + i\Delta\beta = \frac{(1 - \nu^2)^{1/2}}{a} \frac{n^2}{p^2 - n^2} (\xi + i\eta) \quad (9b)$$

It may be of interest to compare the attenuation constants given by (9) with the results obtained by calculating the power dissipated in the walls of a pipe⁹ which has different resistances in the circumferential and longitudinal directions. If the wall resistance for circumferential currents is represented by R_θ and for longitudinal currents by R_z , the expressions for α are

TM_{nm} modes

$$\alpha = \frac{R_z}{(\mu_0/\epsilon_0)^{1/2}a(1 - \nu^2)^{1/2}}$$

TE_{nm} modes

$$\alpha = \frac{R_\theta\nu^2 + R_z(n/p)^2(1 - \nu^2)}{(\mu_0/\epsilon_0)^{1/2}a(1 - \nu^2)^{1/2}} \frac{p^2}{p^2 - n^2}$$

The results for ordinary metallic pipe are obtained by setting

$$R_\theta = R_z = R = (\omega\mu_0/2\sigma)^{1/2}$$

If $R_\theta = 0$, the expressions above agree with (9), inasmuch as $\xi = R(\epsilon_0/\mu_0)^{1/2}$ when the jacket conductivity is large.

$$\psi = \tan^{-1} n(1 - \nu^2)^{1/2}/p\nu, n \neq 0$$

For this value of ψ the circularly polarized TE_{nm} mode which varies as $\exp(-in\theta)$ has low attenuation. (We assume $n \neq 0$, since the case $n = 0$ has been treated above.) One of the properties of helix waveguide is the difference in propagation between right and left circularly polarized TE_{nm} modes. By properly designing the helix angle for the frequency, mode, and size of guide, the loss to one of the polarizations can be made very low. If the jacket is lossy enough the attenuation of the other polarization should be quite high. Thus only one of the circularly polarized modes should be propagated through a long pipe. Such a helix has features analogous to the optical properties of levulose and dextrose solutions, which distinguish between left and right circularly polarized light.

Let α_n be the attenuation constant of the mode which varies as $\exp(-in\theta)$, and α_{-n} the attenuation constant of the mode which varies

⁹ S. A. Schelkunoff, Electromagnetic Waves, van Nostrand, New York, 1943, pp. 385-387.

as $\exp(+in\theta)$. Then from (7b), for any pitch angle ψ ,

$$\begin{aligned}\alpha_{-n} &= \frac{\xi}{a} \frac{p^2}{p^2 - n^2} \frac{\nu^2}{(1 - \nu^2)^{1/2}} \frac{[\tan \psi + n(1 - \nu^2)^{1/2}/p\nu]^2}{1 + \tan^2 \psi} \\ \alpha_n &= \frac{\xi}{a} \frac{p^2}{p^2 - n^2} \frac{\nu^2}{(1 - \nu^2)^{1/2}} \frac{[\tan \psi - n(1 - \nu^2)^{1/2}/p\nu]^2}{1 + \tan^2 \psi} \\ \alpha_{-n} - \alpha_n &= 4 \frac{\xi}{a} \frac{np}{p^2 - n^2} \frac{\nu \tan \psi}{1 + \tan^2 \psi}\end{aligned}$$

The mode which varies as $\exp(-in\theta)$ has lower loss if ψ and n have the same sign.

The TM_{nm} attenuation constants are independent of polarization and are given by (7a).

$$\psi = 90^\circ$$

These "helices," with wires parallel to the axis of the waveguide, should propagate TM_{nm} modes with losses approximately the same as in copper pipe. For the TE_{nm} modes, (7b) gives

TE_{nm} modes

$$\alpha + i\Delta\beta = \frac{\nu^2}{a(1 - \nu^2)^{1/2}} \frac{p^2}{p^2 - n^2} (\xi + i\eta)$$

IV. NUMERICAL SOLUTIONS FOR ZERO-PITCH HELICES

The main interest in helix waveguide is for small pitch angles where the TE_{01} attenuation is very low. The propagation constants of various lossy modes in helix guides of zero pitch have been calculated by solving the characteristic equation (6) numerically. These calculations will now be described.

Equation (6) is first simplified by setting $\psi = 0$ and replacing the Hankel functions with their asymptotic expressions. The condition for validity of the asymptotic expressions, namely

$$|\xi_2 a| \gg |(4n^2 - 1)/8|$$

is well satisfied in all cases to be treated here. Equation (6) may then be rearranged in the dimensionless form

$$\begin{aligned}F_n(\xi_1 a) &= (\xi_2 a)^3 [(nha)^2 J_n^2(\xi_1 a) - (\beta_0 a)^2 (\xi_1 a)^2 J_n'^2(\xi_1 a)] \\ &\quad - i(\xi_1 a)^3 [(nha)^2 + (\beta_0 a)^2 (\epsilon' - i\epsilon'') (\xi_2 a)^2] J_n'(\xi_1 a) J_n(\xi_1 a) \\ &= 0\end{aligned}\quad (10)$$

There is no difference between the propagation constants of right and

left circularly polarized waves when $\psi = 0$. Using the relationships

$$\xi_2 a = [(\xi_1 a)^2 + (\beta_0 a)^2 (\epsilon' - i\epsilon'' - 1)]^{1/2}, \quad \text{Im} \xi_2 a < 0$$

$$ha = [(\beta_0 a)^2 - (\xi_1 a)^2]^{1/2}, \quad \text{Im } ha < 0$$

it is clear that $F_n(\xi_1 a)$ is an even function of $\xi_1 a$, involving the parameters $\beta_0 a$ ($= 2\pi a/\lambda_0$), ϵ' , ϵ'' , and n .

When specific values have been assigned to $\beta_0 a$, ϵ' , and ϵ'' , roots of (10) can be found numerically by the straightforward procedure of evaluating $F_n(\xi_1 a)$ at a regular network of points in the plane of the complex variable $\xi_1 a$, plotting the families of curves $\text{Re } F_n = 0$ and $\text{Im } F_n = 0$, and reading off the values of $\xi_1 a$ corresponding to the intersections of curves of the two families.

The procedure just outlined has been applied to the cases $n = 0$ and $n = 1$. When $n = 0$ one can take out of $F_0(\xi_1 a)$ the factor $J_0'(\xi_1 a)$, whose roots correspond to the TE_{0m} modes; the roots of the other factor are the TM_{0m} -limit modes. When $n = 1$ the function $F_1(\xi_1 a)$ does not factor, and its roots correspond to both TE_{1m} -limit and TM_{1m} -limit modes. If the jacket conductivity is high it is easy to identify the various limit modes, and a given mode can be traced continuously if the conductivity is decreased in sufficiently small steps.

The numerical calculations were set up, more or less arbitrarily, to cover the region $0 \leq \text{Re } \xi_1 a \leq 10$, $-10 \leq \text{Im } \xi_1 a \leq 10$, for each set of parameter values. A few plots of $\text{Re } F_n$ and $\text{Im } F_n$ made it apparent that for propagating modes the roots in this region are all in the first quadrant and usually near the real axis. The entire process of solution was then programmed by Mrs. F. M. Laurent for automatic execution on an IBM 650 magnetic drum calculator. The calculator first evaluated $F_n(\xi_1 a)$ at a network of points spaced half a unit apart in both directions, then examined the sign changes of $\text{Re } F_n$ and $\text{Im } F_n$ around each elementary square. If it appeared that a particular square might contain a root of F_n , the values of F_n at the four corner points were fitted by an interpolating cubic polynomial¹⁰ which was then solved. If the cubic had a root inside the given square, this was recorded as an approximate root of F_n . The normalized propagation constant $ih a = \alpha a + i\beta a$ was also recorded for each root.

The calculated roots $\xi_1 a$ and the normalized propagation constants are summarized in Tables I(a) to I(f), which relate to the following cases:

Table I(a) — $\beta_0 a = 29.554$, $\epsilon' = 4$, ϵ'' variable

Table I(b) — $\beta_0 a = 29.554$, $\epsilon' = 100$, ϵ'' variable

Table I(c) — $\beta_0 a = 29.554$, $\epsilon' = \epsilon''$, both variable

¹⁰ A. N. Lowan and H. E. Salzer, Jour. Math. and Phys., **23**, p. 157, 1944.

Table I(d) — $\beta_0 a = 12.930$, $\epsilon' = 4$, ϵ'' variable

Table I(e) — $\beta_0 a = 12.930$, $\epsilon' = \epsilon''$, both variable

Table I(f) — $\beta_0 a = 6.465$, $\epsilon' = 4$, ϵ'' variable

The three values of $\beta_0 a$ correspond to waveguides of diameter 2 inches, $\frac{7}{8}$ inch, and $\frac{7}{16}$ inch at $\lambda_0 = 5.4$ mm. The jacket materials (mostly carbon-loaded resins) which have been tested to date show a range of relative permittivities roughly from 4 to 100. There is some indication that the permittivity of a carbon-loaded resin increases as its conductivity increases; this suggested consideration of the case $\epsilon' = \epsilon''$.

The tables cover the range from $\epsilon'' = 1000$ down to $\epsilon'' = 1$ at small enough intervals so that the general course of each mode can be followed. It is worth noting that at 5.4 mm a resistivity ($1/\sigma$) of 1 ohm cm corresponds to $\epsilon'' = 32$. Copper at this frequency has an ϵ'' of approximately 2×10^7 .

In general the tables include the modes derived from $F_0(\xi_1 a)$ whose limits are TM_{01} , TM_{02} , and TM_{03} , and the modes derived from $F_1(\xi_1 a)$ whose limits are TE_{11} , TM_{11} , TE_{12} , TM_{12} , and TE_{13} (except that in the $\frac{7}{16}$ -inch guide TM_{03} , TM_{12} , and TE_{13} are cut off). Some results are given for the TM_{13} -limit mode, namely those which satisfy the arbitrary criterion $\text{Re } \xi_1 a \leq 10$; but these results are incomplete because for large ϵ'' the corresponding root of $F_1(\xi_1 a)$ approaches 10.173. Furthermore for small values of ϵ'' the attenuation constants of a few of the TM-limit modes become quite large and the corresponding values of $\xi_1 a$ move far away from the origin. Since our object was to make a general survey rather than to investigate any particular mode exhaustively, we did not attempt to pursue these modes outside the region originally proposed for study.

The results of the IBM calculations are recorded in Table I to three decimal places. Since the roots $\xi_1 a$ were obtained by cubic interpolation in a square of side 0.5, the last place is not entirely reliable; but spot checks on a few of the roots by successive approximations indicate that it is probably not off by more than one or two units. The propagation constants of some of the relatively low-loss modes (especially TE_{12} and TE_{13} , whose wall currents are largely circumferential) were calculated from the approximate formulas,* as noted in the tables. The attenuation

(Text continued on page 1375)

* The formulas used were (A9) and (A10) of the appendix, which are slightly more accurate than (7) of the text.

TABLE I(a) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$)
WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	ξa	$\alpha a + i\beta a$
TM ₀₁	∞	2.405	29.456i
	1000	2.154 + 0.384i	0.028 + 29.478i
	250	2.094 + 0.974i	0.069 + 29.496i
	100	2.408 + 1.679i	0.137 + 29.504i
	90	2.482 + 1.772i	0.149 + 29.503i
	80	2.579 + 1.878i	0.164 + 29.502i
	64	2.804 + 2.083i	0.198 + 29.495i
	40	3.519 + 2.547i	0.304 + 29.456i
	25	4.604 + 3.165i	0.496 + 29.369i
	16	5.870 + 3.763i	0.756 + 29.219i
	10	7.564 + 4.131i	1.082 + 28.887i
	8	8.464 + 4.158i	1.229 + 28.646i
TM ₀₂	∞	5.520	29.034i
	1000	5.399 + 0.127i	0.024 + 29.057i
	250	5.274 + 0.268i	0.049 + 29.081i
	100	5.109 + 0.445i	0.078 + 29.113i
	90	5.081 + 0.472i	0.082 + 29.118i
	80	5.047 + 0.504i	0.087 + 29.125i
	64	4.968 + 0.569i	0.097 + 29.139i
	40	4.716 + 0.701i	0.113 + 29.184i
	25	4.375 + 0.877i	0.101 + 29.237i
	16	4.172 + 0.551i	0.079 + 29.264i
	10	4.047 + 0.448i	0.062 + 29.279i
	8	4.004 + 0.412i	0.056 + 29.285i
	4	3.905 + 0.344i	0.046 + 29.297i
	1	3.820 + 0.310i	0.040 + 29.308i
TM ₀₃	∞	8.654	28.259i
	1000	8.577 + 0.078i	0.024 + 28.282i
	250	8.500 + 0.160i	0.048 + 28.306i
	100	8.408 + 0.260i	0.077 + 28.334i
	90	8.395 + 0.275i	0.081 + 28.338i
	80	8.378 + 0.293i	0.086 + 28.343i
	64	8.344 + 0.330i	0.097 + 28.354i
	40	8.253 + 0.424i	0.123 + 28.382i
	25	8.125 + 0.545i	0.156 + 28.421i
	16	7.943 + 0.678i	0.189 + 28.475i
	10	7.658 + 0.779i	0.209 + 28.556i
	8	7.511 + 0.780i	0.205 + 28.595i
	4	7.200 + 0.693i	0.174 + 28.673i
	1	6.986 + 0.612i	0.149 + 28.724i
TE ₁₁	∞	1.841	29.497i
	1000	1.703 + 0.234i	0.014 + 29.506i
	250	1.764 + 0.630i	0.038 + 29.508i
	100	2.465 + 0.963i	0.081 + 29.467i
	90	2.660 + 0.748i	0.068 + 29.444i
	80	2.633 + 0.604i	0.054 + 29.443i
	64	2.594 + 0.464i	0.041 + 29.444i
	40	2.546 + 0.312i	0.027 + 29.446i
	25	2.508 + 0.226i	0.019 + 29.448i
	16	2.481 + 0.176i	0.015 + 29.450i
	10	2.455 + 0.140i	0.012 + 29.452i
	8	2.445 + 0.129i	0.011 + 29.453i
	4	2.418 + 0.106i	0.009 + 29.455i
	1	2.394 + 0.095i	0.008 + 29.457i

TABLE I(a) — *Continued*

Limit Mode	ϵ''	ξ_{1a}	$\alpha a + i\beta a$
TM_{11}	∞	3.832	29.305i
	1000	3.652 + 0.197i	0.024 + 29.328i
	250	3.457 + 0.440i	0.052 + 29.355i
	100	2.978 + 0.880i	0.089 + 29.417i
	90	2.821 + 1.215i	0.116 + 29.445i
	80	2.945 + 1.476i	0.148 + 29.444i
	64	3.146 + 1.868i	0.200 + 29.446i
	40	3.728 + 2.564i	0.325 + 29.432i
	25	4.659 + 3.175i	0.504 + 29.361i
	16	5.921 + 3.727i	0.756 + 29.204i
	10	7.613 + 4.135i	1.090 + 28.875i
	8	8.487 + 4.153i	1.231 + 28.639i
TE_{12}	∞	5.331	29.069i
	1000		0.0008 + 29.070i*
	250		0.0016 + 29.071i*
	100		0.0026 + 29.072i*
	64		0.0033 + 29.072i*
	40		0.0042 + 29.073i*
	25		0.0055 + 29.074i*
	10		0.0092 + 29.075i*
	4	5.297 + 0.072i	0.013 + 29.076i
	1	5.322 + 0.096i	0.018 + 29.071i
TM_{12}	∞	7.016	28.710i
	1000	6.918 + 0.099i	0.024 + 28.733i
	250	6.821 + 0.203i	0.048 + 28.757i
	100	6.701 + 0.330i	0.077 + 28.786i
	90	6.683 + 0.349i	0.081 + 28.791i
	80	6.660 + 0.372i	0.086 + 28.796i
	64	6.612 + 0.419i	0.096 + 28.808i
	40	6.475 + 0.535i	0.120 + 28.841i
	25	6.253 + 0.655i	0.142 + 28.893i
	16	5.965 + 0.682i	0.141 + 28.954i
	10	5.719 + 0.590i	0.116 + 29.002i
	8	5.641 + 0.541i	0.105 + 29.016i
	4	5.471 + 0.419i	0.079 + 29.047i
	1	5.317 + 0.347i	0.063 + 29.074i
TE_{13}	∞	8.536	28.295i
	1000		0.0003 + 28.295i*
	250		0.0006 + 28.295i*
	100		0.0010 + 28.296i*
	64		0.0012 + 28.296i*
	40		0.0016 + 28.296i*
	25		0.0020 + 28.296i*
	10		0.0034 + 28.297i*
	4		0.0050 + 28.296i*
	1		0.0058 + 28.295i*
TM_{13}	∞	10.173	27.748i
	100	9.963 + 0.219i	0.078 + 27.825i
	90	9.952 + 0.231i	0.083 + 27.829i
	80	9.938 + 0.246i	0.088 + 27.834i
	64	9.911 + 0.277i	0.098 + 27.845i
	40	9.840 + 0.356i	0.126 + 27.870i
	25	9.746 + 0.460i	0.161 + 27.905i
	16	9.625 + 0.591i	0.204 + 27.950i
	10	9.433 + 0.757i	0.255 + 28.020i
	8	9.305 + 0.837i	0.278 + 28.065i
	4	8.836 + 0.898i	0.281 + 28.218i
	1	8.485 + 0.781i	0.234 + 28.322i

* Approximate formula.

TABLE I(b) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$)
 WITH $\epsilon' = 100$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	$\xi_1 a$	$\alpha a + i\beta a$	
TM_{01}	∞	2.405		29.456i
	1000	2.178 + 0.391i	0.029	+ 29.476i
	250	2.291 + 0.885i	0.069	+ 29.479i
	100	2.677 + 1.062i	0.097	+ 29.452i
	80	2.764 + 1.047i	0.098	+ 29.443i
	64	2.834 + 1.019i	0.098	+ 29.436i
	40	2.928 + 0.950i	0.094	+ 29.424i
	25	2.973 + 0.893i	0.090	+ 29.418i
	10	3.004 + 0.831i	0.085	+ 29.413i
	4	3.013 + 0.806i	0.083	+ 29.411i
	1	3.016 + 0.793i	0.081	+ 29.411i
TM_{02}	∞	5.520		29.034i
	1000	5.406 + 0.133i	0.025	+ 29.056i
	250	5.339 + 0.298i	0.055	+ 29.069i
	100	5.372 + 0.473i	0.087	+ 29.066i
	80	5.398 + 0.508i	0.094	+ 29.062i
	64	5.429 + 0.535i	0.100	+ 29.056i
	40	5.492 + 0.566i	0.107	+ 29.045i
	25	5.540 + 0.573i	0.109	+ 29.036i
	10	5.589 + 0.569i	0.109	+ 29.027i
	4	5.608 + 0.563i	0.109	+ 29.023i
	1	5.617 + 0.560i	0.108	+ 29.021i
TM_{03}	∞	8.654		28.259i
	1000	8.581 + 0.082i	0.025	+ 28.281i
	250	8.537 + 0.179i	0.054	+ 28.295i
	100	8.548 + 0.279i	0.084	+ 28.292i
	80	8.561 + 0.300i	0.091	+ 28.289i
	64	8.575 + 0.317i	0.096	+ 28.285i
	40	8.606 + 0.339i	0.103	+ 28.276i
	25	8.630 + 0.348i	0.106	+ 28.268i
	10	8.658 + 0.352i	0.108	+ 28.260i
	4	8.669 + 0.352i	0.108	+ 28.257i
	1	8.675 + 0.351i	0.108	+ 28.255i
TE_{11}	∞	1.841		29.497i
	1000	1.719 + 0.236i	0.014	+ 29.505i
	250	1.871 + 0.504i	0.032	+ 29.499i
	100	2.132 + 0.484i	0.035	+ 29.481i
	80	2.161 + 0.451i	0.033	+ 29.479i
	64	2.178 + 0.420i	0.031	+ 29.477i
	40	2.191 + 0.372i	0.028	+ 29.475i
	25	2.192 + 0.343i	0.026	+ 29.475i
	10	2.190 + 0.316i	0.023	+ 29.475i
	4	2.188 + 0.306i	0.023	+ 29.475i
	1	2.187 + 0.301i	0.022	+ 29.475i

TABLE I(b) — *Continued*

Limit Mode	ϵ''	ξ_{1a}	$\alpha a + i\beta a$
TM_{11}	∞	3.832	29.305i
	1000	3.663 + 0.204i	0.026 + 29.327i
	250	3.579 + 0.485i	0.059 + 29.341i
	100	3.715 + 0.788i	0.100 + 29.331i
	80	3.787 + 0.826i	0.107 + 29.322i
	64	3.856 + 0.843i	0.111 + 29.314i
	40	3.969 + 0.836i	0.113 + 29.299i
	25	4.043 + 0.817i	0.113 + 29.288i
	10	4.100 + 0.777i	0.109 + 29.279i
	4	4.119 + 0.759i	0.107 + 29.276i
	1	4.128 + 0.749i	0.106 + 29.274i
TE_{12}	∞	5.331	29.069i
	1000		0.0008 + 29.070i*
	250		0.0018 + 29.071i*
	100		0.0028 + 29.071i*
	64		0.0032 + 29.070i*
	40		0.0034 + 29.070i*
	25		0.0035 + 29.070i*
	10		0.0036 + 29.070i*
	4		0.0036 + 29.070i*
	1		0.0036 + 29.069i*
TM_{12}	∞	7.016	28.710i
	1000	6.923 + 0.103i	0.025 + 28.732i
	250	6.868 + 0.226i	0.054 + 28.746i
	100	6.885 + 0.355i	0.085 + 28.743i
	80	6.902 + 0.381i	0.092 + 28.740i
	64	6.922 + 0.403i	0.097 + 28.735i
	40	6.965 + 0.429i	0.104 + 28.725i
	25	7.000 + 0.440i	0.107 + 28.717i
	10	7.037 + 0.443i	0.109 + 28.708i
	4	7.051 + 0.441i	0.108 + 28.704i
	1	7.058 + 0.440i	0.108 + 28.703i
TE_{13}	∞	8.536	28.295i
	1000		0.0003 + 28.295i*
	250		0.0007 + 28.295i*
	100		0.0010 + 28.295i*
	64		0.0012 + 28.295i*
	40		0.0013 + 28.295i*
	25		0.0013 + 28.295i*
	10		0.0013 + 28.295i*
	4		0.0013 + 28.295i*
	1		0.0013 + 28.295i*

* Approximate formula.

TABLE I(c) — 2-INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 29.554$)
WITH $\epsilon' = \epsilon''$

Limit Mode	ϵ' and ϵ''	$\xi_1 a$	$\alpha a + i\beta a$	
ΓM_{01}	∞	2.405		29.456i
	1000	2.338 + 0.341i	0.027	+ 29.464i
	250	2.418 + 0.707i	0.058	+ 29.464i
	100	2.677 + 1.062i	0.097	+ 29.452i
	64	2.925 + 1.226i	0.122	+ 29.435i
	40	3.309 + 1.324i	0.149	+ 29.399i
	32	3.540 + 1.299i	0.156	+ 29.371i
	25	3.787 + 1.162i	0.150	+ 29.334i
	16	3.946 + 0.800i	0.108	+ 29.301i
	12	3.950 + 0.647i	0.087	+ 29.296i
	10	3.946 + 0.573i	0.077	+ 29.295i
	4	3.905 + 0.344i	0.046	+ 29.297i
	2	3.869 + 0.252i	0.033	+ 29.301i
	1	3.820 + 0.185i	0.024	+ 29.307i
$T M_{02}$	∞	5.520		29.034i
	1000	5.469 + 0.136i	0.026	+ 29.044i
	250	5.423 + 0.282i	0.053	+ 29.054i
	100	5.372 + 0.473i	0.087	+ 29.066i
	64	5.337 + 0.624i	0.115	+ 29.075i
	40	5.294 + 0.874i	0.159	+ 29.090i
	32	5.279 + 1.061i	0.193	+ 29.099i
	25	5.319 + 1.367i	0.250	+ 29.105i
	16	5.852 + 1.969i	0.397	+ 29.039i
	12	6.472 + 2.178i	0.487	+ 28.923i
	10	7.026 + 2.198i	0.536	+ 28.796i
$T M_{03}$	∞	8.654		28.259i
	1000	8.620 + 0.085i	0.026	+ 28.269i
	250	8.587 + 0.173i	0.052	+ 28.280i
	100	8.548 + 0.279i	0.084	+ 28.292i
	64	8.521 + 0.355i	0.107	+ 28.302i
	40	8.483 + 0.461i	0.138	+ 28.315i
	32	8.458 + 0.526i	0.157	+ 28.323i
	25	8.425 + 0.611i	0.182	+ 28.335i
	16	8.330 + 0.824i	0.242	+ 28.369i
	12	8.206 + 1.037i	0.300	+ 28.413i
	10	8.034 + 1.240i	0.350	+ 28.471i
	4	7.200 + 0.693i	0.174	+ 28.673i
	2	7.098 + 0.483i	0.120	+ 28.694i
	1	6.998 + 0.349i	0.085	+ 28.716i
$T E_{11}$	∞	1.841		29.497i
	1000	1.810 + 0.190i	0.012	+ 29.499i
	250	1.911 + 0.384i	0.025	+ 29.495i
	100	2.132 + 0.484i	0.035	+ 29.481i
	64	2.270 + 0.453i	0.035	+ 29.470i
	40	2.365 + 0.366i	0.029	+ 29.462i
	32	2.389 + 0.324i	0.026	+ 29.459i
	25	2.406 + 0.281i	0.023	+ 29.457i
	16	2.420 + 0.219i	0.018	+ 29.456i
	12	2.424 + 0.187i	0.015	+ 29.455i
	10	2.424 + 0.169i	0.014	+ 29.455i
	4	2.418 + 0.106i	0.009	+ 29.455i
	2	2.409 + 0.078i	0.006	+ 29.456i
	1	2.394 + 0.056i	0.005	+ 29.457i

TABLE I(c)—Continued

Limit Mode	ϵ' and ϵ''	$\xi_1 a$	$\alpha a + i\beta a$	
TM ₁₁	∞	3.832		29.305i
	1000	3.759 + 0.203i	0.026	+ 29.315i
	250	3.714 + 0.439i	0.056	+ 29.323i
	100	3.715 + 0.788i	0.100	+ 29.331i
	64	3.797 + 1.070i	0.139	+ 29.329i
	40	4.080 + 1.400i	0.195	+ 29.305i
	32	4.276 + 1.550i	0.226	+ 29.285i
	25	4.586 + 1.661i	0.260	+ 29.245i
	16	5.359 + 1.579i	0.291	+ 29.109i
	12	5.587 + 1.043i	0.201	+ 29.041i
	10	5.560 + 0.859i	0.164	+ 29.040i
	4	5.471 + 0.419i	0.079	+ 29.047i
	2	5.438 + 0.249i	0.047	+ 29.051i
	1	5.444 + 0.131i	0.025	+ 29.049i
TE ₁₂	∞	5.331		29.069i
	1000		0.0009	+ 29.070i*
	250		0.0018	+ 29.070i*
	100		0.0028	+ 29.071i*
	64		0.0035	+ 29.071i*
	40		0.0044	+ 29.071i*
	25		0.0055	+ 29.072i*
	10		0.0087	+ 29.073i*
	4	5.297 + 0.072i	0.013	+ 29.076i
	2	5.272 + 0.108i	0.020	+ 29.080i
	1	5.198 + 0.132i	0.023	+ 29.094i
TM ₁₂	∞	7.016		28.710i
	1000	6.971 + 0.107i	0.026	+ 28.721i
	250	6.931 + 0.217i	0.052	+ 28.731i
	100	6.885 + 0.355i	0.085	+ 28.743i
	64	6.852 + 0.457i	0.109	+ 28.753i
	40	6.801 + 0.610i	0.144	+ 28.768i
	32	6.768 + 0.708i	0.167	+ 28.778i
	25	6.720 + 0.850i	0.198	+ 28.793i
	16	6.562 + 1.359i	0.309	+ 28.850i
	12	6.869 + 2.095i	0.499	+ 28.825i
	10	7.322 + 2.374i	0.605	+ 28.737i
TE ₁₃	∞	8.536		28.295i
	1000		0.0003	+ 28.295i*
	250		0.0007	+ 28.295i*
	100		0.0010	+ 28.295i*
	64		0.0013	+ 28.295i*
	40		0.0016	+ 28.295i*
	25		0.0021	+ 28.295i*
	10		0.0032	+ 28.296i*
	4		0.0050	+ 28.296i*
	1		0.0094	+ 28.295i*
TM ₁₃	∞	10.173		27.748i
	25	9.981 + 0.497i	0.178	+ 27.823i
	16	9.910 + 0.652i	0.232	+ 27.852i
	12	9.841 + 0.785i	0.277	+ 27.880i
	10	9.776 + 0.893i	0.313	+ 27.907i
	4	8.836 + 0.898i	0.281	+ 28.218i
	2	8.656 + 0.596i	0.183	+ 28.265i
	1	8.523 + 0.409i	0.123	+ 28.302i

* Approximate formula.

TABLE I(d)— $\frac{7}{8}$ -INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 12.930$)
WITH $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	$\xi_1 a$	$\alpha a + i\beta a$	
TM ₀₁	∞	2.405		12.704i
	1000	2.286 + 0.140i	0.025	+ 12.727i
	250	2.183 + 0.324i	0.056	+ 12.749i
	100	2.113 + 0.595i	0.098	+ 12.771i
	64	2.114 + 0.800i	0.132	+ 12.782i
	40	2.185 + 1.072i	0.183	+ 12.790i
	25	2.377 + 1.369i	0.255	+ 12.786i
	10	3.212 + 1.699i	0.431	+ 12.647i
	6.4	3.694 + 1.440i	0.426	+ 12.482i
	4.0	3.765 + 1.029i	0.312	+ 12.416i
	2.5	3.700 + 0.853i	0.254	+ 12.421i
	1.0	3.624 + 0.733i	0.214	+ 12.435i
TM ₀₂	∞	5.520		11.692i
	1000	5.468 + 0.054i	0.025	+ 11.717i
	250	5.416 + 0.111i	0.051	+ 11.742i
	100	5.356 + 0.183i	0.083	+ 11.770i
	64	5.317 + 0.235i	0.106	+ 11.789i
	40	5.266 + 0.308i	0.137	+ 11.814i
	25	5.206 + 0.410i	0.180	+ 11.844i
	10	5.073 + 0.772i	0.328	+ 11.923i
	6.4	5.095 + 1.137i	0.485	+ 11.948i
	4.0	5.486 + 1.420i	0.664	+ 11.814i
	2.5	5.818 + 1.379i	0.689	+ 11.650i
	1.0	6.041 + 1.188i	0.624	+ 11.511i
TM ₀₃	∞	8.654		9.607i
	1000	8.620 + 0.034i	0.030	+ 9.637i
	250	8.587 + 0.069i	0.061	+ 9.667i
	100	8.550 + 0.111i	0.098	+ 9.701i
	64	8.525 + 0.141i	0.124	+ 9.723i
	40	8.494 + 0.183i	0.160	+ 9.752i
	25	8.459 + 0.239i	0.207	+ 9.785i
	10	8.393 + 0.411i	0.350	+ 9.851i
	6.4	8.386 + 0.532i	0.452	+ 9.866i
	4.0	8.426 + 0.668i	0.571	+ 9.847i
	2.5	8.515 + 0.769i	0.669	+ 9.784i
	1.0	8.676 + 0.824i	0.741	+ 9.651i
TE ₁₁	∞	1.841		12.798i
	1000	1.767 + 0.074i	0.010	+ 12.809i
	250	1.717 + 0.191i	0.026	+ 12.817i
	100	1.706 + 0.368i	0.049	+ 12.822i
	64	1.734 + 0.500i	0.068	+ 12.823i
	40	1.857 + 0.656i	0.095	+ 12.813i
	25	2.126 + 0.773i	0.129	+ 12.778i
	10	2.436 + 0.411i	0.079	+ 12.706i
	6.4	2.413 + 0.316i	0.060	+ 12.707i
	4.0	2.386 + 0.262i	0.049	+ 12.711i
	2.5	2.364 + 0.234i	0.043	+ 12.714i
	1.0	2.341 + 0.212i	0.039	+ 12.718i

TABLE I(d) — *Continued*

Limit Mode	ϵ''	$\xi_1 a$	$\alpha a + i\beta a$
TM ₁₁	∞	3.832	12.349i
	1000	3.750 + 0.081i	0.025 + 12.375i
	250	3.676 + 0.171i	0.051 + 12.398i
	100	3.588 + 0.290i	0.084 + 12.426i
	64	3.530 + 0.382i	0.108 + 12.445i
	40	3.447 + 0.516i	0.143 + 12.474i
	25	3.329 + 0.757i	0.201 + 12.519i
	10	3.749 + 1.664i	0.499 + 12.496i
	6.4	4.275 + 1.750i	0.606 + 12.343i
	4.0	4.701 + 1.553i	0.600 + 12.160i
	2.5	4.843 + 1.274i	0.511 + 12.067i
	1.0	4.844 + 1.031i	0.415 + 12.040i
TE ₁₂	∞	5.331	11.780i
	1000		0.0007 + 11.780i*
	250		0.0015 + 11.781i*
	100		0.0024 + 11.782i*
	64		0.0030 + 11.782i*
	40		0.0039 + 11.783i*
	25		0.0051 + 11.784i*
	10		0.0085 + 11.785i*
	4		0.0125 + 11.784i*
	1		0.0146 + 11.781i*
TM ₁₂	∞	7.016	10.861i
	1000	6.972 + 0.043i	0.027 + 10.889i
	250	6.930 + 0.087i	0.055 + 10.917i
	100	6.883 + 0.141i	0.088 + 10.947i
	64	6.853 + 0.179i	0.112 + 10.967i
	40	6.814 + 0.233i	0.144 + 10.992i
	25	6.769 + 0.305i	0.187 + 11.023i
	10	6.679 + 0.541i	0.326 + 11.090i
	6.4	6.670 + 0.718i	0.431 + 11.109i
	4.0	6.755 + 0.935i	0.570 + 11.080i
	2.5	6.942 + 1.061i	0.671 + 10.981i
	1.0	7.193 + 1.054i	0.700 + 10.819i
TE ₁₃	∞	8.536	9.712i
	1000		0.0002 + 9.712i*
	250		0.0005 + 9.712i*
	100		0.0008 + 9.712i*
	64		0.0010 + 9.713i*
	40		0.0012 + 9.713i*
	25		0.0016 + 9.713i*
	10		0.0027 + 9.713i*
	4		0.0040 + 9.713i*
	1		0.0048 + 9.712i*
TM ₁₃	∞	10.173	7.980i
	10	9.949 + 0.340i	0.409 + 8.276i
	6.4	9.943 + 0.436i	0.523 + 8.293i
	4.0	9.970 + 0.543i	0.655 + 8.277i

* Approximate formula.

TABLE I(e) — $\frac{7}{8}$ -INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 12.930$) WITH
 $\epsilon' = \epsilon''$

Limit Mode	ϵ' and ϵ''	$\xi_1 a$	$\alpha a + i\beta a$
TM ₀₁	∞	2.405	12.704i
	1000	2.360 + 0.141i	0.026 + 12.714i
	250	2.339 + 0.295i	0.054 + 12.720i
	100	2.351 + 0.482i	0.089 + 12.724i
	64	2.382 + 0.608i	0.114 + 12.724i
	40	2.450 + 0.766i	0.148 + 12.720i
	25	2.573 + 0.942i	0.191 + 12.708i
	10	3.052 + 1.244i	0.301 + 12.630i
	4	3.765 + 1.029i	0.312 + 12.416i
	2	3.841 + 0.653i	0.203 + 12.366i
	1	3.768 + 0.438i	0.133 + 12.378i
TM ₀₂	∞	5.520	11.692i
	1000	5.497 + 0.058i	0.027 + 11.704i
	250	5.475 + 0.118i	0.055 + 11.715i
	100	5.451 + 0.190i	0.088 + 11.727i
	64	5.435 + 0.241i	0.111 + 11.735i
	40	5.416 + 0.310i	0.143 + 11.746i
	25	5.393 + 0.402i	0.184 + 11.760i
	10	5.338 + 0.701i	0.317 + 11.802i
	4	5.486 + 1.429i	0.664 + 11.814i
	2	6.389 + 1.780i	0.996 + 11.425i
	1	6.901 + 1.040i	0.652 + 11.003i
TM ₀₃	∞	8.654	9.607i
	1000	8.639 + 0.037i	0.033 + 9.621i
	250	8.624 + 0.074i	0.067 + 9.635i
	100	8.607 + 0.118i	0.105 + 9.650i
	64	8.596 + 0.148i	0.132 + 9.661i
	40	8.581 + 0.189i	0.168 + 9.675i
	25	8.563 + 0.241i	0.213 + 9.694i
	10	8.512 + 0.393i	0.344 + 9.747i
	4	8.426 + 0.668i	0.571 + 9.847i
	2	8.320 + 1.094i	0.910 + 9.999i
	1	8.812 + 1.915i	1.721 + 9.806i
TE ₁₁	∞	1.841	12.798i
	1000	1.810 + 0.072i	0.010 + 12.803i
	250	1.807 + 0.161i	0.023 + 12.804i
	100	1.833 + 0.265i	0.038 + 12.802i
	64	1.870 + 0.330i	0.048 + 12.799i
	40	1.939 + 0.401i	0.061 + 12.790i
	25	2.047 + 0.459i	0.074 + 12.776i
	10	2.295 + 0.414i	0.075 + 12.732i
	4	2.386 + 0.262i	0.049 + 12.711i
	2	2.389 + 0.186i	0.035 + 12.709i
	1	2.369 + 0.129i	0.024 + 12.712i

TABLE I(e) — *Continued*

Limit Mode	ϵ' and ϵ''	$\xi_1 a$	$\alpha a + i\beta a$
TM ₁₁	∞	3.832	12.349i
	1000	3.794 + 0.086i	0.026 + 12.361i
	250	3.766 + 0.176i	0.054 + 12.371i
	100	3.739 + 0.288i	0.087 + 12.381i
	64	3.725 + 0.369i	0.111 + 12.388i
	40	3.711 + 0.485i	0.145 + 12.396i
	25	3.708 + 0.651i	0.195 + 12.406i
	10	3.893 + 1.161i	0.365 + 12.390i
	4	4.701 + 1.553i	0.600 + 12.160i
	2	5.319 + 1.062i	0.477 + 11.843i
	1	5.241 + 0.614i	0.272 + 11.840i
TE ₁₂	∞	5.331	11.780i
	1000		0.0008 + 11.780i*
	250		0.0016 + 11.780i*
	100		0.0026 + 11.781i*
	64		0.0032 + 11.781i*
	40		0.0041 + 11.781i*
	25		0.0051 + 11.782i*
	10		0.0081 + 11.783i*
	4		0.0125 + 11.784i*
	1		0.0236 + 11.782i*
TM ₁₂	∞	7.016	10.861i
	1000	6.996 + 0.047i	0.030 + 10.874i
	250	6.976 + 0.094i	0.060 + 10.887i
	100	6.955 + 0.149i	0.095 + 10.902i
	64	6.942 + 0.187i	0.119 + 10.911i
	40	6.924 + 0.238i	0.151 + 10.923i
	25	6.903 + 0.305i	0.192 + 10.939i
	10	6.841 + 0.509i	0.317 + 10.988i
	4	6.755 + 0.935i	0.570 + 11.080i
	2	7.053 + 1.730i	1.106 + 11.030i
	1	8.138 + 1.672i	1.325 + 10.272i
TE ₁₃	∞	8.536	9.712i
	1000		0.0003 + 9.712i*
	250		0.0005 + 9.712i*
	100		0.0008 + 9.712i*
	64		0.0010 + 9.712i*
	40		0.0013 + 9.712i*
	25		0.0016 + 9.712i*
	10		0.0025 + 9.713i*
	4		0.0040 + 9.713i*
	1		0.0076 + 9.713i*
TM ₁₃	∞	10.173	7.980i
	4	9.970 + 0.543i	0.655 + 8.277i
	2	9.863 + 0.826i	0.963 + 8.457i
	1	9.698 + 1.418i	1.561 + 8.808i

* Approximate formula.

TABLE I(f)— $\frac{7}{16}$ -INCH GUIDE AT $\lambda_0 = 5.4$ MM ($\beta_0 a = 6.465$) WITH
 $\epsilon' = 4$ AND ϵ'' VARIABLE

Limit Mode	ϵ''	$\xi_1 a$	$\alpha a + i\beta a$		
'M ₀₁	∞	2.405			6.001i
	1000		0.024	+	6.025i*
	250	2.287 + 0.141i	0.053	+	6.049i
	100	2.228 + 0.244i	0.090	+	6.074i
	64	2.197 + 0.324i	0.117	+	6.090i
	40	2.170 + 0.439i	0.156	+	6.107i
	25	2.169 + 0.594i	0.210	+	6.123i
	10	2.355 + 0.943i	0.364	+	6.105i
	4	2.740 + 1.040i	0.478	+	5.966i
	1	2.961 + 0.878i	0.446	+	5.830i
TM ₀₂	∞	5.520			3.365i
	1000		0.043	+	3.408i*
	250	5.468 + 0.054i	0.086	+	3.450i
	100	5.439 + 0.088i	0.137	+	3.499i
	64	5.420 + 0.112i	0.172	+	3.530i
	40	5.396 + 0.146i	0.221	+	3.570i
	25	5.370 + 0.191i	0.284	+	3.616i
	10	5.327 + 0.328i	0.471	+	3.707i
	4	5.369 + 0.512i	0.740	+	3.712i
	1	5.539 + 0.614i	0.965	+	3.524i
TE ₁₁	∞	1.841			6.197i
	1000		0.009	+	6.206i*
	250	1.772 + 0.069i	0.020	+	6.218i
	100	1.744 + 0.129i	0.036	+	6.227i
	64	1.731 + 0.176i	0.049	+	6.231i
	40	1.726 + 0.244i	0.068	+	6.235i
	25	1.744 + 0.334i	0.093	+	6.235i
	10	1.925 + 0.493i	0.153	+	6.193i
	4	2.121 + 0.425i	0.147	+	6.123i
	1	2.152 + 0.319i	0.112	+	6.106i
TM ₁₁	∞	3.832			5.207i
	1000		0.028	+	5.235i*
	250	3.751 + 0.082i	0.058	+	5.266i
	100	3.710 + 0.134i	0.094	+	5.297i
	64	3.683 + 0.173i	0.119	+	5.317i
	40	3.650 + 0.227i	0.155	+	5.343i
	25	3.615 + 0.303i	0.204	+	5.372i
	10	3.581 + 0.546i	0.360	+	5.422i
	4	3.763 + 0.810i	0.570	+	5.350i
	1	4.038 + 0.816i	0.639	+	5.154i
TE ₁₂	∞	5.331			3.657i
	1000		0.0005	+	3.657i*
	250		0.0009	+	3.658i*
	100		0.0015	+	3.658i*
	64		0.0019	+	3.659i*
	40		0.0024	+	3.659i*
	25		0.0031	+	3.659i*
	10		0.0052	+	3.660i*
	4		0.0079	+	3.660i*
	1		0.0097	+	3.658i*

* Approximate formula.

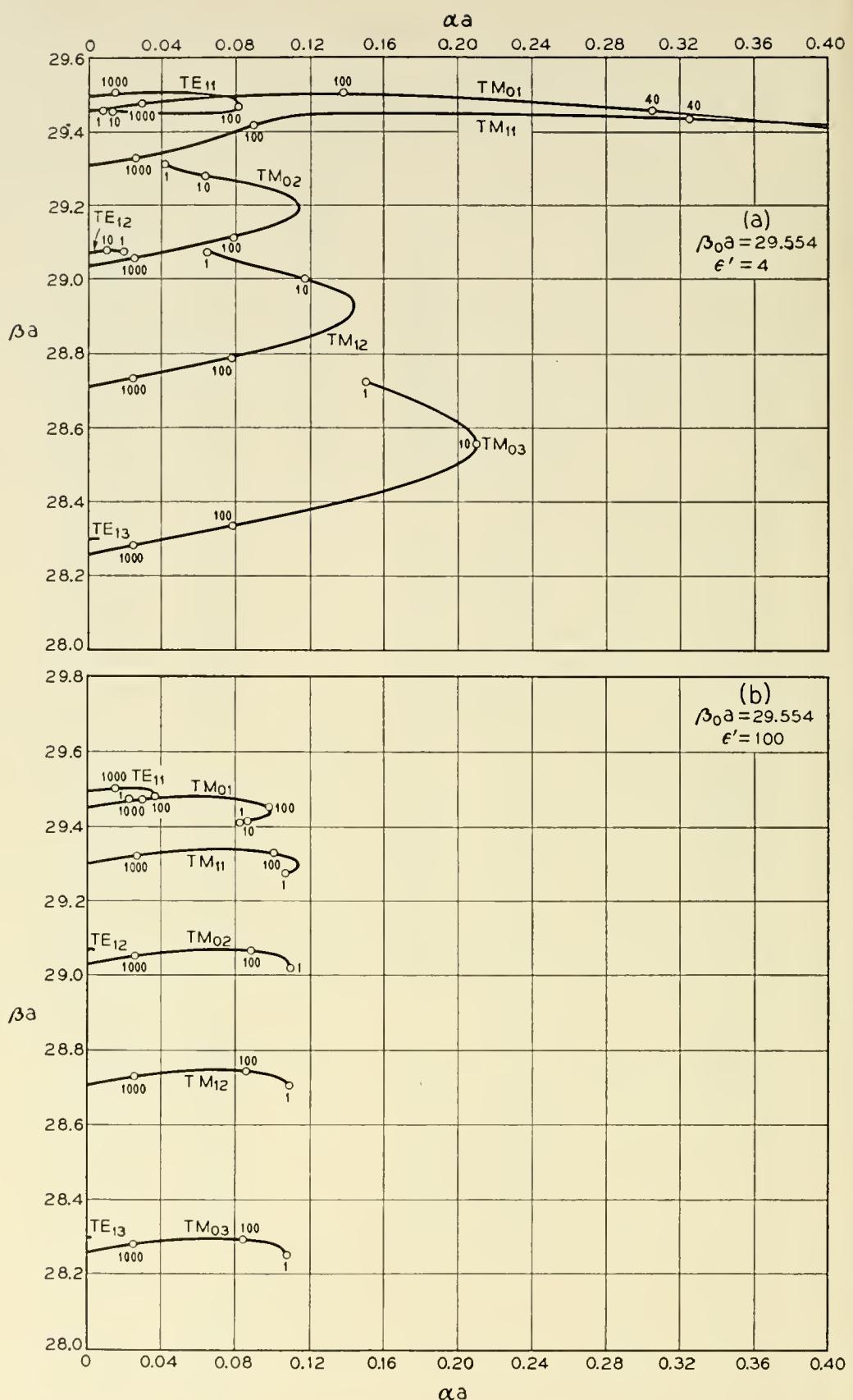


Fig. 2(a) and (b)

Fig. 2 — Plots of phase constant versus attenuation constant for modes in various helix waveguides. Representative values of ϵ'' are shown on the curves.

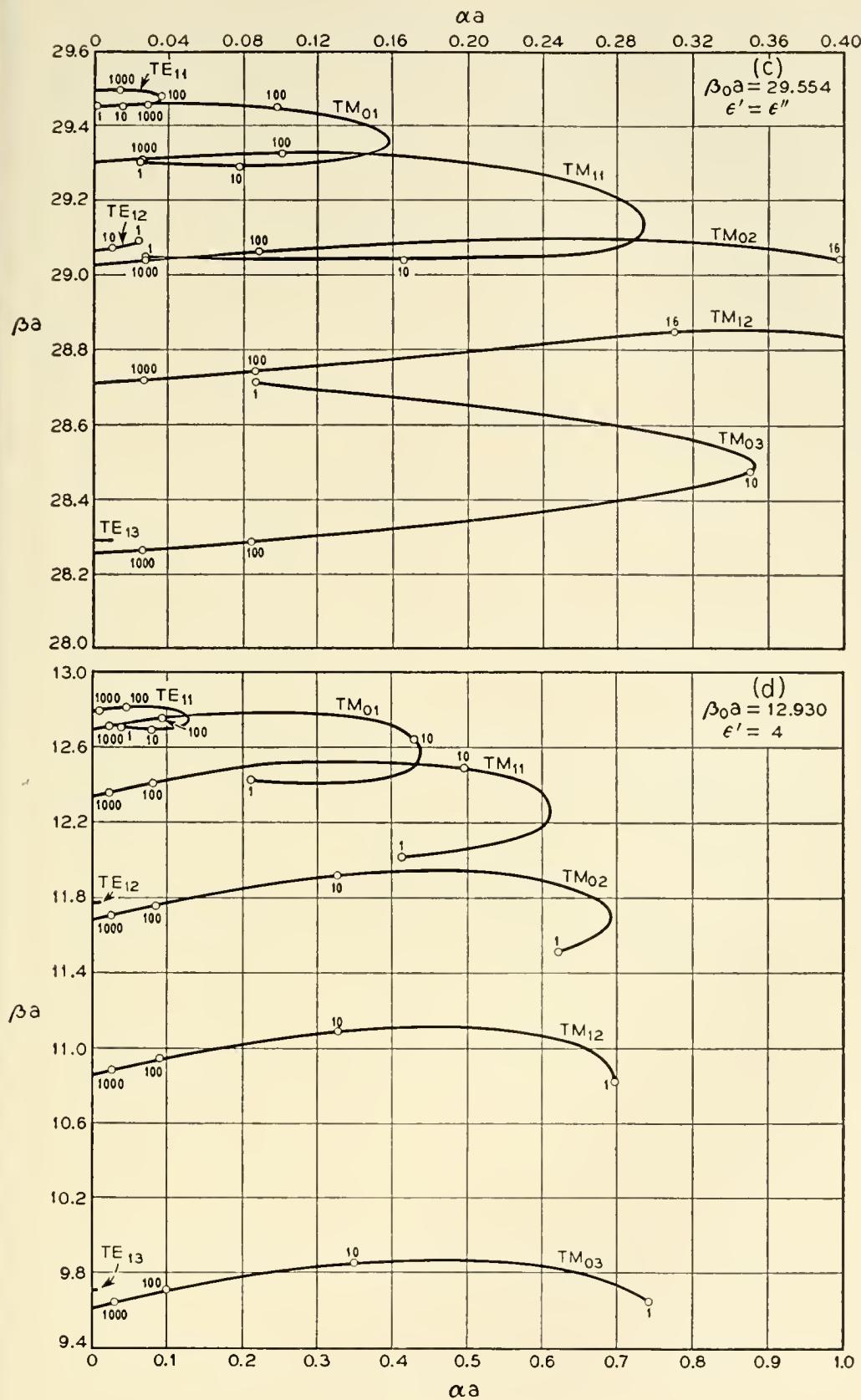


Fig. 2(c) and (d)

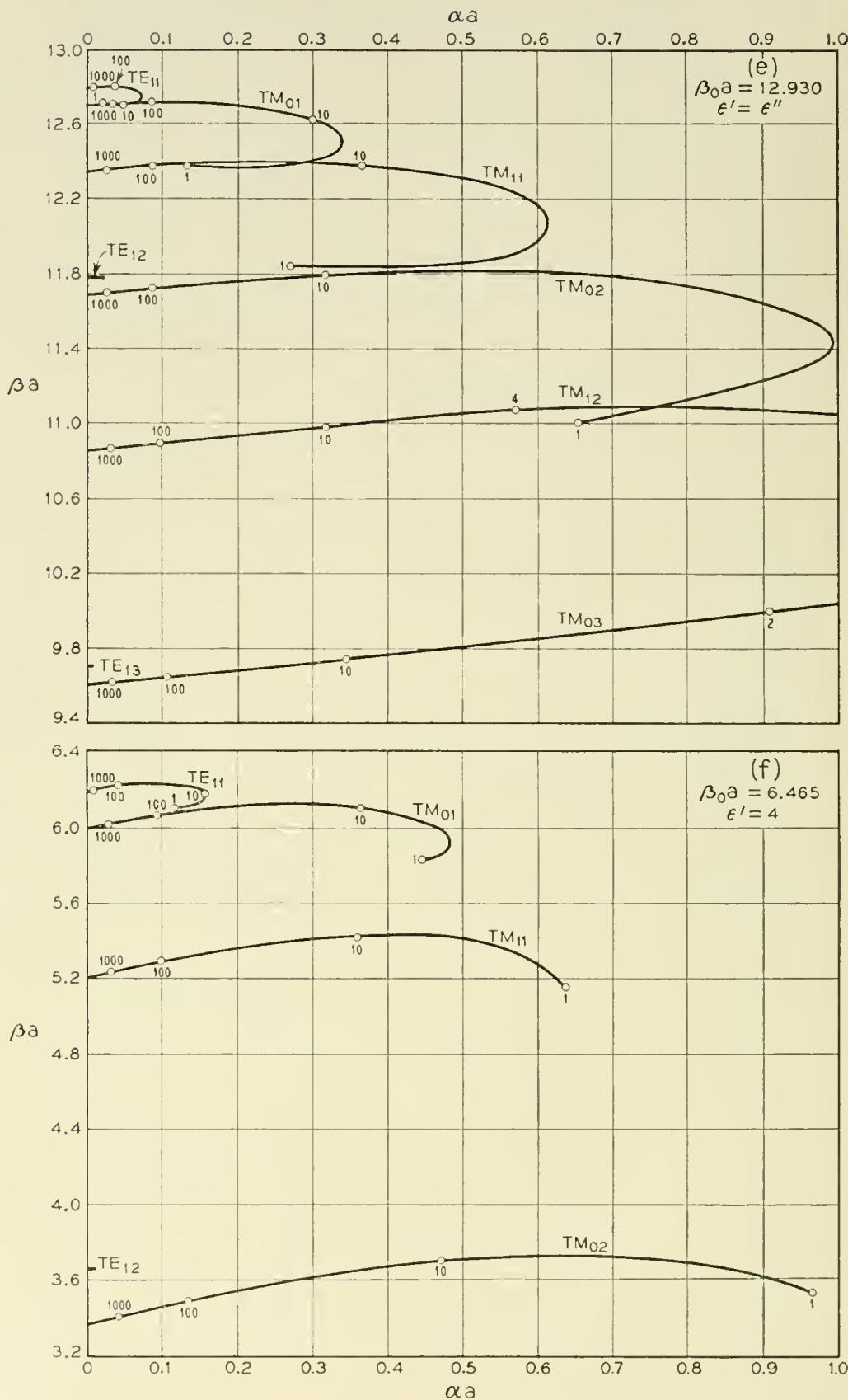


Fig. 2(e) and (f)

constants calculated from the approximate formulas are given to four decimal places, i.e., usually two significant figures.

The contents of Table I are displayed graphically in Figs. 2(a) through (f), which show plots of βa vs αa for all modes except TM_{13} . Representative values of ϵ'' are indicated on the curves. Note that the scales are different for the different guide sizes, and that the βa -scale is compressed in all cases. If αa and βa were plotted on the same scale, the curves would make an initial angle of 45° with the αa -axis when $\epsilon' = \text{constant}$, or 22.5° when $\epsilon' = \epsilon''$.

Figs. 3(a) to (f) show the normalized attenuation constants αa of various modes plotted against ϵ'' on a log-log scale. In Fig. 3(b) the curves for all TM modes would be similar to the two shown, and in Fig. 3(d) the TM_{03} curve is like TM_{12} . Although for some modes the attenuation constant increases steadily as the conductivity decreases over the range of our calculations, in many cases the attenuation passes through a maximum and then decreases as the conductivity is further decreased. This phenomenon will be discussed in Section V.

It may be noticed that in some instances the limit modes are not unique. For example, Tables I(a), with $\epsilon' = 4$, and I(c), with $\epsilon' = \epsilon''$, for the large guide have in common the case $\epsilon' = 4$, $\epsilon'' = 4$. For this case consider the circular magnetic mode corresponding to $\xi_1 a = 3.905 + 0.344i$. If ϵ' is constant ($= 4$) while ϵ'' tends to infinity, this mode approaches the TM_{02} mode in a perfectly conducting guide; but if ϵ' and ϵ'' tend to infinity while remaining equal to each other, the same mode approaches TM_{01} in a perfectly conducting guide. Presumably the TM_{01} -limit mode in the former case coincides with the TM_{02} -limit mode in the latter case; but the value of $\xi_1 a$ for this mode is outside the range of our calculations at $\epsilon' = \epsilon'' = 4$. A similar interchange occurs between the TM_{11} -limit and TM_{12} -limit modes in the large guide, depending on whether ϵ' is constant or ϵ' tends to infinity with ϵ'' . There is no evidence of any such phenomenon in the smaller guide of Tables I(d) and I(e); but the fact that it can occur means that the limit-mode designations of modes in a lossy waveguide are not entirely unambiguous. The phenomenon is not due to the presence of the helix, since a helix of zero pitch has no effect on circular magnetic modes.

Finally it is of interest to compare the propagation constants given by the approximate formula with those obtained by numerical solution of the characteristic equation. A reasonably typical case is provided by the TM_{02} -limit mode in a 2-inch guide at $\lambda_0 = 5.4$ mm with $\epsilon' = 4$, as in Table I(a). Exact and approximate results for βa vs αa and αa vs ϵ'' are plotted in Fig. 4. As the conductivity decreases, the attenuation con-

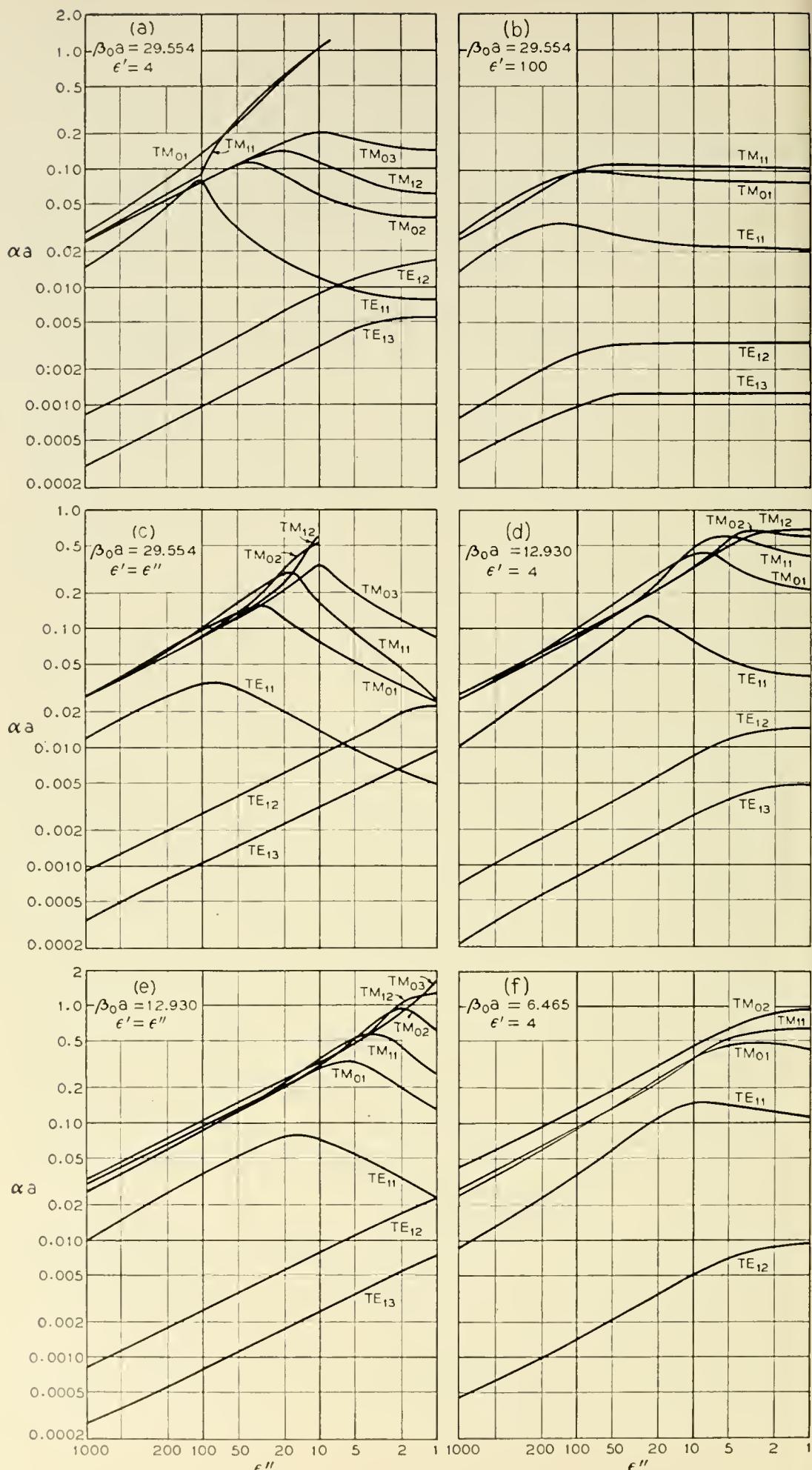


Fig. 3 — Attenuation constant as a function of jacket conductivity for modes in various helix waveguides.

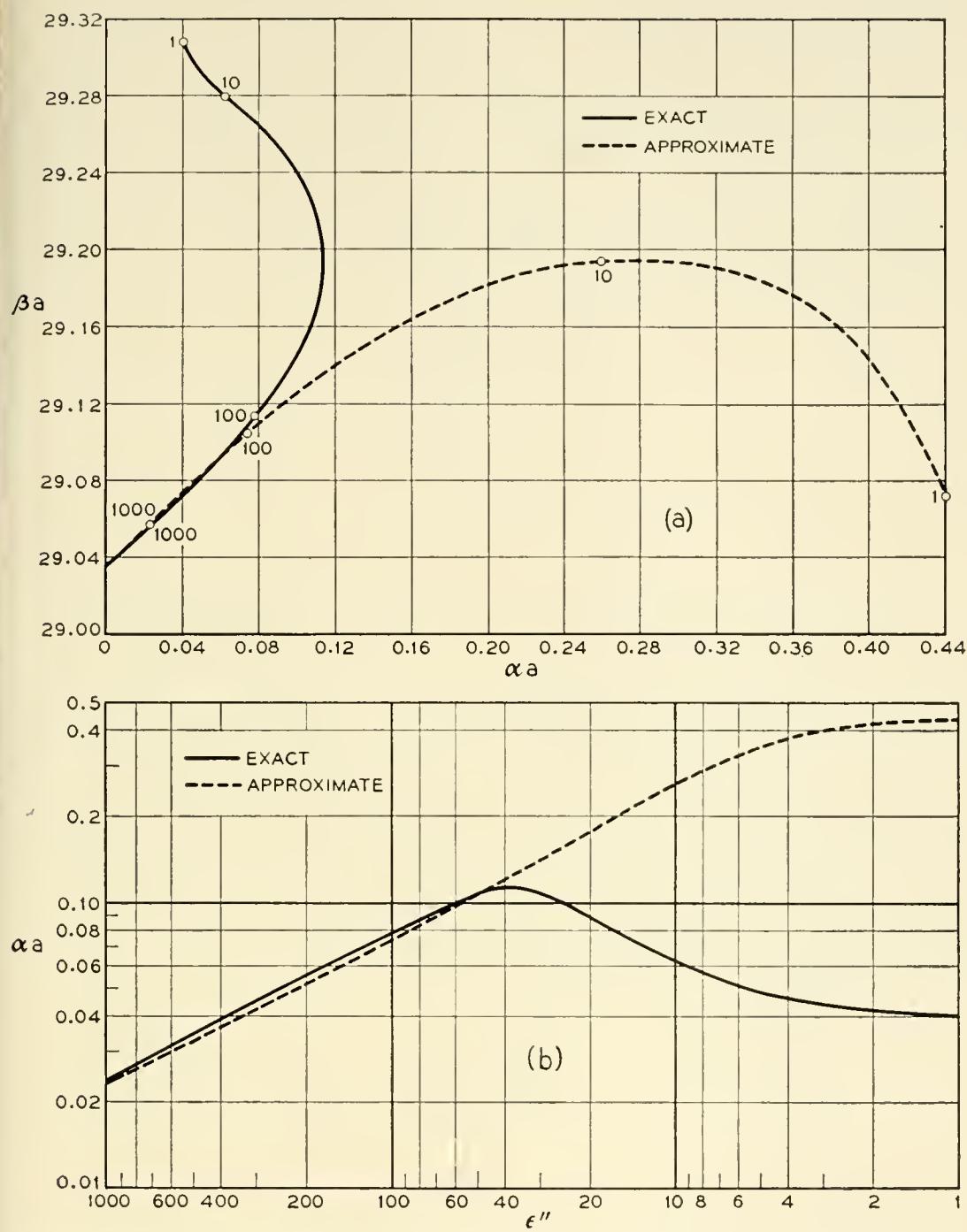


Fig. 4 — Comparison of exact and approximate formulas for the propagation constant of a typical mode (TM_{02} -limit in a guide with $\beta_0a = 29.554$ and $\epsilon' = 4$).

stant first becomes larger, in all cases, than predicted by the approximate formula. For still lower conductivities the attenuation constant may pass through a maximum, as in the present example, and decrease again. The existence of a maximum in the attenuation vs conductivity curve is not indicated by the approximate formula.

V. DISCUSSION OF RESULTS

The dimensionless results of Section IV may easily be scaled to any desired operating wavelength, and the attenuation constants and guide wavelengths expressed in conventional units. If λ_0 is the free-space wavelength in centimeters, then the guide diameter d in inches, the attenuation constant α in db/meter, and the guide wavelength λ_g in centimeters are given by the following formulas:

$$d_{\text{in}} = 0.12532 (\beta_0 a) (\lambda_0)_{\text{cm}}$$

$$\alpha_{\text{db/m}} = \frac{5457.5 (\alpha a)}{(\beta_0 a) (\lambda_0)_{\text{cm}}}$$

$$(\lambda_g)_{\text{cm}} = \frac{(\beta_0 a) (\lambda_0)_{\text{cm}}}{(\beta a)}$$

Table II lists the guide diameters and the conversion factors for α and λ_g for the three values of $\beta_0 a$ used in Section IV, at frequencies corresponding to free-space wavelengths of 3.33 and 0.54 cm. The table also lists the number of propagating modes in a perfectly conducting guide as a function of $\beta_0 a$ (different polarizations are not counted separately).

When helix waveguide is used to reduce mode conversions, an important parameter is the ratio of the attenuation constant of any given unwanted mode to the attenuation constant of the TE_{01} mode. The theoretical attenuation constants of the TE_{01} mode at $\lambda_0 = 5.4$ mm in copper guides of various sizes are listed below:

Diameter	αa	$\alpha_{\text{db/m}}$
2"	2.77×10^{-6}	9.47×10^{-4}
$\frac{7}{8}"$	1.50×10^{-5}	1.17×10^{-2}
$\frac{7}{16}"$	7.11×10^{-5}	1.11×10^{-1}

TABLE II — CONVERSION FACTORS FOR ATTENUATION CONSTANTS AND GUIDE WAVELENGTHS IN VARIOUS WAVEGUIDES

$\beta_0 a$	Propagating modes	$\lambda_0 = 3.33 \text{ cm}$			$\lambda_0 = 0.54 \text{ cm}$		
		Diameter (inches)	$\alpha \text{ db/meter}$	$\lambda_g \text{ cm}$	Diameter (inches)	$\alpha \text{ db/meter}$	$\lambda_g \text{ cm}$
29.554	227	12.33	55.5 αa	$98.41/\beta a$	2.000	342 αa	$15.959/\beta a$
12.930	44	5.40	127 αa	$43.06/\beta a$	0.875	782 αa	$6.982/\beta a$
6.465	12	2.70	253 αa	$21.53/\beta a$	0.4375	1563 αa	$3.491/\beta a$

Referring to the values of αa listed in Table I, we see that the unwanted mode attenuations can be made to exceed the TE_{01} attenuation by factors of from several hundred to several hundred thousand in the large helix guide. The attenuation ratios are somewhat smaller in the smaller guide sizes.

The attenuation versus conductivity plots of Fig. 3 show that for many of the modes there is a value of jacket conductivity, depending on the mode, the value of $\beta_0 a$, and the jacket permittivity, which maximizes the attenuation constant. Since one is accustomed to think of the attenuation constant of a waveguide as an increasing function of frequency for all sufficiently high frequencies (except for circular electric waves), or as an increasing function of wall resistance, it is worth while to see why one should really expect the attenuation constant to pass through a maximum as the frequency is increased indefinitely in an ordinary metallic guide, or as the wall resistance is increased at a fixed frequency. The argument runs as follows:

Guided waves inside a cylindrical pipe may be expressed as bundles of plane waves repeatedly reflected from the cylindrical boundary.¹¹ The angle which the wave normals make with the guide axis decreases as the frequency increases farther above cutoff; and the complementary angle, which is the angle of incidence of the waves upon the boundary, approaches 90° . If the walls are imperfectly conducting, the guided wave is attenuated because the reflection coefficient of the component waves at the boundary is less than unity. The theory of reflection at an imperfectly conducting surface shows that the reflection coefficient of a plane wave polarized with its electric vector in the plane of incidence first decreases with increasing angle of incidence, then passes through a deep minimum, and finally increases to unity at strictly grazing incidence.¹² For a metallic reflector, the angle of incidence corresponding to minimum reflection is very near 90° . Inasmuch as all modes in circular guide except for the circular electric family have a component of \vec{E} in the plane of incidence (the plane $\theta = \text{constant}$), one would expect the attenuation constant of each mode to pass through a maximum at a sufficiently high frequency. For example, the TM_{01} mode in a 2-inch copper guide should have maximum attenuation at a free-space wavelength in the neighborhood of 0.1 mm (100 microns), assuming the dc value for the conductivity of copper. To find the actual maximum, of course, would require the solution of a transcendental equation as in Section IV.

The circular electric waves all have \vec{E} normal to the plane of incidence.

¹¹ Reference 9, pp. 411-412.

¹² Reference 7, pp. 507-509.

For this polarization the reflection coefficient increases steadily from its value at normal incidence to unity at grazing incidence. Thus one has an optical interpretation of the anomalous attenuation-frequency behavior of circular electric waves.

If instead of varying the frequency one imagines the wall resistance varied at a fixed frequency, he can easily convince himself that there usually exists a finite value of resistance which maximizes the attenuation constant of a given mode. An idealized illustrative example has been worked out by Schelkunoff.¹³ He considers the propagation of transverse magnetic waves between parallel resistance sheets, and shows that if the sheets are far enough apart the attenuation constant increases from zero to a maximum and then falls again to zero, as the wall resistance is made to increase from zero to infinity. It may be instructive to consider that maximum power is dissipated in the lossy walls when their impedance is matched as well as possible to the wave impedance, looking normal to the walls, of the fields inside the guide.

In conclusion we mention a couple of theoretical questions which are suggested by the numerical results of Section IV.

(1) Limit modes. It has been seen that the limit which a given lossy mode approaches as the jacket conductivity becomes infinite may not be unique. Can rules be given for determining limit modes when the manner in which $|\epsilon' - i\epsilon''|$ approaches infinity is specified?

(2) Behavior of modes as $\sigma \rightarrow 0$. It is known¹⁴ that the number of true guided waves (i.e., exponentially propagating waves whose fields vanish at large radial distances from the guide axis) possible in a cylindrical waveguide is finite if the conductivity of the exterior medium is finite. The number is enormously large if the exterior medium is a metal; but the modes presumably disappear one by one as the conductivity is decreased. If the conductivity of the exterior medium is low enough and if its permittivity is not less than the permittivity of the interior medium, no true guided waves can exist. At what values of conductivity do the first few modes appear in a guide of given size, and how do their propagation constants behave at very low conductivities?

The complete theory of lossy-wall waveguide would appear to present quite a challenge to the applied mathematician. Fortunately the engineering usefulness of helix waveguide does not depend upon getting immediate answers to such difficult analytical questions.

¹³ Reference 9, pp. 484-489.

¹⁴ G. M. Roe, The Theory of Acoustic and Electromagnetic Wave Guides and Cavity Resonators, Ph.D. thesis, U. of Minn., 1947, Section 2.

APPENDIX

APPROXIMATE SOLUTION OF THE CHARACTERISTIC EQUATION

The characteristic equation (6) of the helix guide may be written in the dimensionless form

$$\begin{aligned} & \left(\xi_1 a \tan \psi - \frac{nha}{\xi_1 a} \right)^2 \frac{J_n(\xi_1 a)}{J_n'(\xi_1 a)} - (\beta_0 a)^2 \frac{J_n'(\xi_1 a)}{J_n(\xi_1 a)} \\ &= \frac{\xi_1 a}{\xi_2 a} \left[\left(\xi_2 a \tan \psi - \frac{nha}{\xi_2 a} \right)^2 \frac{H_n^{(2)}(\xi_2 a)}{H_n^{(2)\prime}(\xi_2 a)} - (\beta_0 a)^2 (\epsilon' - i\epsilon'') \frac{H_n^{(2)\prime}(\xi_2 a)}{H_n^{(2)}(\xi_2 a)} \right] \end{aligned} \quad (\text{A1})$$

If $|\epsilon' - \epsilon''|$ is sufficiently large, the right side of the equation is large and either $J_n(\xi_1 a)$ or $J_n'(\xi_1 a)$ is near zero. Let p denote a particular root of J_n or J_n' ; then to zero order,

$$\begin{aligned} \xi_1 a &= p \\ ha &= \beta_{nm} a = \beta_0 a (1 - \nu^2)^{1/2} \\ \xi_2 a &= \beta_0 a (\epsilon' - i\epsilon'' - 1 - \nu^2)^{1/2} \end{aligned} \quad (\text{A2})$$

where

$$\nu = p/\beta_0 a$$

Henceforth assume that

$$|\xi_2 a| \gg |(4n^2 - 1)/8| \quad (\text{A3a})$$

and

$$|\xi_2 a| \gg |n| \quad (\text{A3b})$$

It is convenient to postulate both inequalities, even though the first is more restrictive than the second unless $|n| = 1$ or $|n| = 2$.

If (A3a) is satisfied, the Hankel functions may be replaced by the first terms of their asymptotic expressions, and

$$\frac{H_n^{(2)\prime}(\xi_2 a)}{H_n^{(2)}(\xi_2 a)} = -i$$

Eq. (A1) becomes

$$\begin{aligned} & \left(\xi_1 a \tan \psi - \frac{nha}{\xi_1 a} \right)^2 \frac{J_n(\xi_1 a)}{J_n'(\xi_1 a)} - (\beta_0 a)^2 \frac{J_n'(\xi_1 a)}{J_n(\xi_1 a)} \\ &= \frac{i\xi_1 a}{\xi_2 a} \left[\left(\xi_2 a \tan \psi - \frac{nha}{\xi_2 a} \right)^2 + (\beta_0 a)^2 (\epsilon' - i\epsilon'') \right] \end{aligned}$$

It follows from (A3b), using the zero-order approximations (A2), that

$$|nha/\xi_2a| \ll |\beta_0a(\epsilon' - i\epsilon'')^{1/2}|$$

so the characteristic equation finally takes the approximate form

$$\begin{aligned} \left(\xi_1a \tan \psi - \frac{nha}{\xi_1a} \right)^2 \frac{J_n(\xi_1a)}{J_n'(\xi_1a)} &= (\beta_0a)^2 \frac{J_n'(\xi_1a)}{J_n(\xi_1a)} \\ &= \frac{i\xi_1a}{\xi_2a} [(\xi_2a \tan \psi)^2 + (\beta_0a)^2(\epsilon' - i\epsilon'')] \end{aligned} \quad (\text{A4})$$

Now let

$$\xi_1a = p + x, \quad |x| \ll 1$$

where x is a small complex number. The normalized propagation constant becomes, to first order,

$$\begin{aligned} iha &= [(\xi_1a)^2 - (\beta_0a)^2]^{1/2} \\ &= i\beta_0a(1 - \nu^2)^{1/2} - i\nu x(1 - \nu^2)^{-1/2} \\ &= \alpha a + i(\beta_{nm}a + \Delta\beta a) \end{aligned}$$

where β_{nm} is the phase constant of the mode in a perfectly conducting guide, and the perturbation terms are

$$\alpha a + i\Delta\beta a = - \frac{i\nu x}{(1 - \nu^2)^{1/2}} \quad (\text{A5})$$

For the TM_{nm} mode, let p be the m^{th} root of J_n ; then from Taylor's series, to first order in x ,

$$J_n(\xi_1a) = J_n(p + x) = xJ_n'(p) \quad (\text{A6})$$

Substituting (A6) into (A4), neglecting the first term on the left side of (A4), and replacing everything on the right side by its zero approximation according to (A2), one obtains

$$-\frac{(\beta_0a)^2}{x} = \frac{ip\beta_0a[(\epsilon' - i\epsilon'' - 1 + \nu^2)\tan^2 \psi + (\epsilon' - i\epsilon'')]}{(\epsilon' - i\epsilon'' - 1 + \nu^2)^{1/2}}$$

or

$$x = \frac{i(\xi + i\eta)}{\nu \left[1 + \left\{ 1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right\} \tan^2 \psi \right]} \quad (\text{A7})$$

where

$$\xi + i\eta = \frac{\left[1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right]^{1/2}}{(\epsilon' - i\epsilon'')^{1/2}} \quad (\text{A8})$$

It follows from (A5) and (A7) that for TM modes,

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2} \left[1 + \left\{ 1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right\} \tan^2 \psi \right]} \quad (\text{A9})$$

where $\xi + i\eta$ is given by (A8).

For the TE_{nm} mode, let p be the m^{th} root of J_n' ; then

$$J_n'(\xi_1 a) = J_n'(p + x) = \frac{(n^2 - p^2)x}{p^2} J_n(p)$$

Equation (A4) yields

$$x = \frac{ip^2\nu}{(p^2 - n^2)} \frac{\left[\tan \psi - \frac{n(1 - \nu^2)^{1/2}}{p\nu} \right]^2 (\xi + i\eta)}{\left[1 + \left\{ 1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right\} \tan^2 \psi \right]}$$

and, using (A5), we have for TE modes,

$$\alpha + i\Delta\beta$$

$$= \frac{p^2}{(p^2 - n^2)} \frac{\nu^2}{a(1 - \nu^2)^{1/2}} \frac{\left[\tan \psi - \frac{n(1 - \nu^2)^{1/2}}{p\nu} \right]^2 (\xi + i\eta)}{\left[1 + \left\{ 1 - \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right\} \tan^2 \psi \right]} \quad (\text{A10})$$

where $\xi + i\eta$ is given by (A8).

In view of (A5), the condition that $|x| \ll 1$ is equivalent to

$$\frac{(1 - \nu^2)^{1/2}}{\nu} |\alpha a + i\Delta\beta a| \ll 1 \quad (\text{A11})$$

In all the numerical cases treated in the present paper, the approximate formulas agree well with the exact ones provided that the left side of (A11) is not greater than about 0.1.

A condition which is usually satisfied in practice, although not strictly a consequence of the assumptions (A3) or (A11), is

$$\left| \frac{1 - \nu^2}{\epsilon' - i\epsilon''} \right| \ll 1$$

This final approximation leads to the simple equations (7a) and (7b) of Section III, namely:

TM_{nm} modes

$$\alpha + i\Delta\beta = \frac{\xi + i\eta}{a(1 - \nu^2)^{1/2}[1 + \tan^2\psi]}$$

TE_{nm} modes

$$\alpha + i\Delta\beta = \frac{(\xi + i\eta)}{a(1 - \nu^2)^{1/2}} \frac{\nu^2 p^2}{(p^2 - n^2)} \frac{[\tan \psi - n(1 - \nu^2)^{1/2}/p\nu]^2}{[1 + \tan^2\psi]}$$

where

$$\xi + i\eta = (\epsilon' - i\epsilon'')^{-1/2}$$

Wafer-Type Millimeter Wave Rectifiers*

By W. M. SHARPLESS

(Manuscript received June 18, 1956)

A wafer-type silicon point-contact rectifier and holder designed primarily for use as the first detector in millimeter wave receivers are described. Measurements made on a pilot production group of one hundred wafer rectifier units yielded the following average performance data at a wavelength of 5.4 millimeters: conversion loss, 7.2 db; noise ratio, 2.2; intermediate frequency output impedance 340 ohms. Methods of estimating the values of the circuit parameters of a point-contact rectifier are given in an Appendix.

INTRODUCTION

Point-contact rectifiers for millimeter waves have been in experimental use for several years. These units, for the most part, have been coaxial cartridges which were inserted in a fixed position, usually centered, in the waveguide. Impedance matching was accomplished by means of a series of matching screws preceding the rectifier and an adjustable waveguide piston following the rectifier. Tuning screws are generally undesirable because of the possibility of losses, narrow bandwidths and instability.

It is the purpose of this paper to describe a new type millimeter-wave rectifier and holder which were designed to eliminate the need for tuning screws and to provide a readily interchangeable rectifier of the flat wafer type. This wafer contains a short section of waveguide across which the point contact rectifier is mounted. The necessary low frequency output terminal (and the rectified current connection) together with the high-frequency bypass capacitor, are also contained within each wafer. The basic idea of the wafer-type rectifier is that the unit can be inserted in its holder and moved transversely to the waveguide to obtain a resistive match to the guide; the reactive component of the rectifier impedance is then tuned out by an adjustable waveguide plunger behind the rectifier.

* This work was supported in part by Contract Nonr-687(00) with the Office of Naval Research, Department of the Navy.

The wafer unit and holder were developed primarily for use as the first converter in double detection receivers operating in the 4- to 7-millimeter wavelength range. In order to check the practicability of the design and to supply rectifiers for laboratory use, a pilot production group of one hundred units was processed and measured. Performance data obtained with this group are presented. A balanced converter using wafer rectifiers is also described.

Methods of estimating the values of the various circuit parameters of a point-contact rectifier are outlined in an appendix. These calculations proved useful in the design of the wafer unit and in predicting the broadband performance of the converter.

DESCRIPTION OF WAFER UNIT AND HOLDER

Fig. 1 is a drawing of the wafer type rectifier. The unit is made from stock steel $\frac{1}{16}$ -inch thick and is gold plated after the milling, drilling and soldering operations are completed. To allow for the transverse impedance matching adjustment, the section of waveguide contained in the wafer is made wider than the RG98U input guide to the holder. By making the wafer thin ($\frac{1}{16}$ inch), the short sections of unused guide on either side will remain "cut-off" over the operating range of the rectifiers. The silicon end of the rectifier consists of a copper pin on which the silicon is press mounted, the assembly held in place with Araldite cement which also serves as the insulating material for a quarter-wavelength long high frequency bypass capacitor. The pin serving as the intermediate frequency and direct current output lead is also cemented in place with Araldite cement. A soft solder connection is made between this pin and the pin holding the silicon wafer. A nickel pin with a conical end on which a pointed tungsten contact spring is welded is pressed into place from the opposite side of the guide at the time of final assembly.

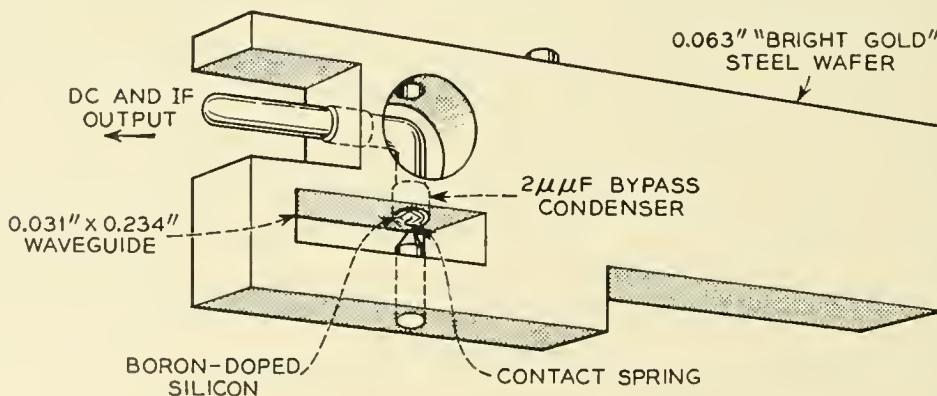


Fig. 1 — Millimeter-wave wafer unit.

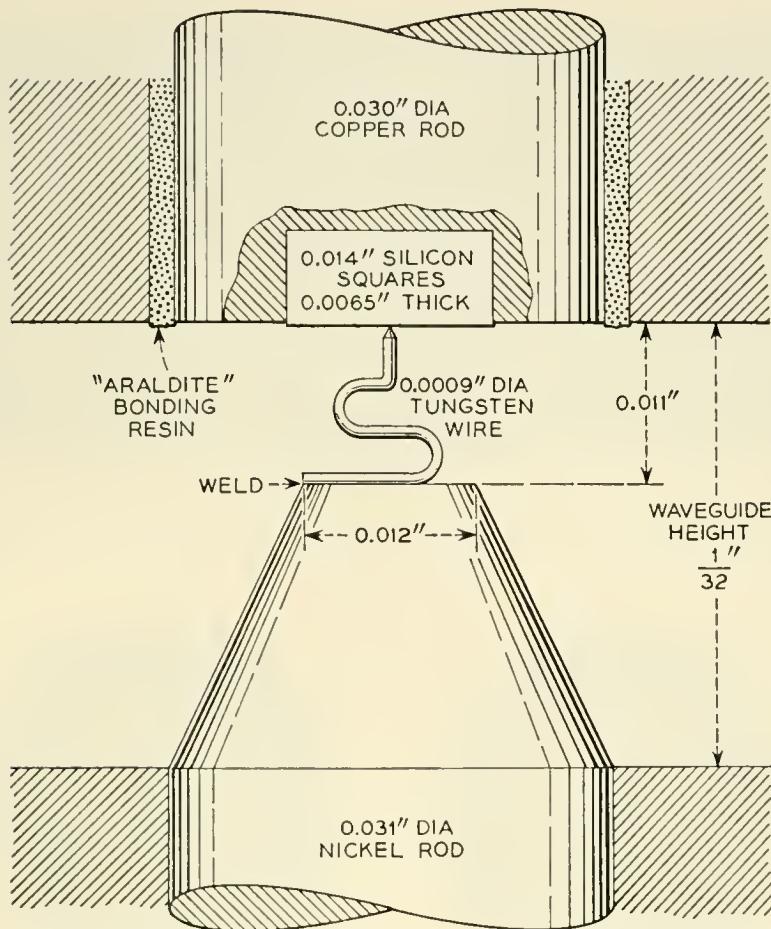


Fig. 2 — Millimeter-wave point-contact assembly.

The region of the wafer unit containing the silicon and point contact is shown in Fig. 2. The methods used in preparing the silicon wafer and the spring contact point are similar in many respects to the standard techniques used in the manufacture of rectifiers for longer wavelengths. Some modifications and refinements in technique are called for by a decrease in size and the increased frequency of operation.

A single-crystal ingot, grown from high purity DuPont silicon doped with 0.02 per cent boron, furnishes the material for the silicon squares used in the wafer unit. Slices cut from the ingot are polished and heat treated. Gold is evaporated on the back surface and the slices are diced into squares approximately 0.014-inch square and 0.0065-inch thick. These squares are pressed into indentations formed in the ends of the 0.030-inch copper pins which have previously been tin-plated. The rods are then cemented in place in the wafer. The spring contact points are made of pure tungsten wire that has been sized to 0.9 mil in diameter by an electrolytic etching process. A short length of this wire is spot welded on the conical end of the 0.031-inch nickel rod. The wire is then bent into

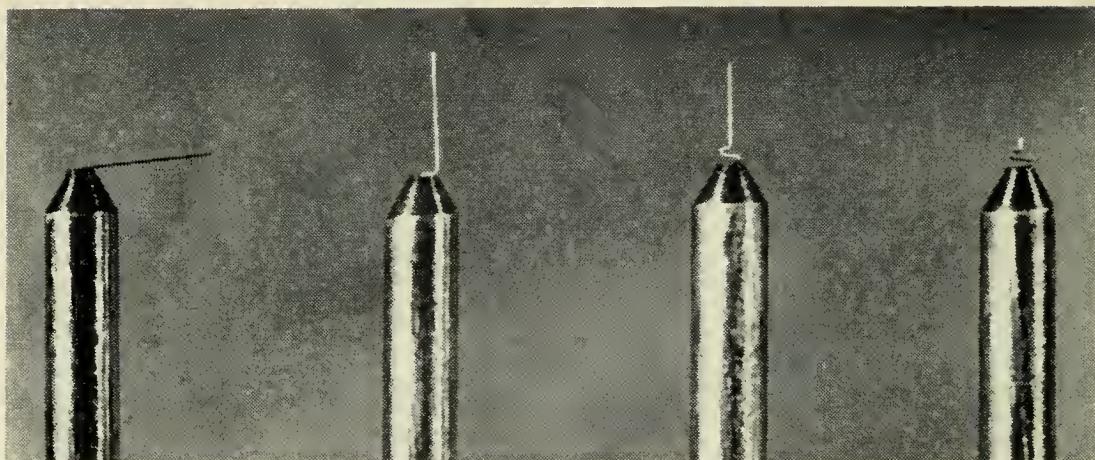


Fig. 3 — Micro-photograph showing successive stages in the formation of the contact spring. The posts are $\frac{1}{32}$ inch in diameter.

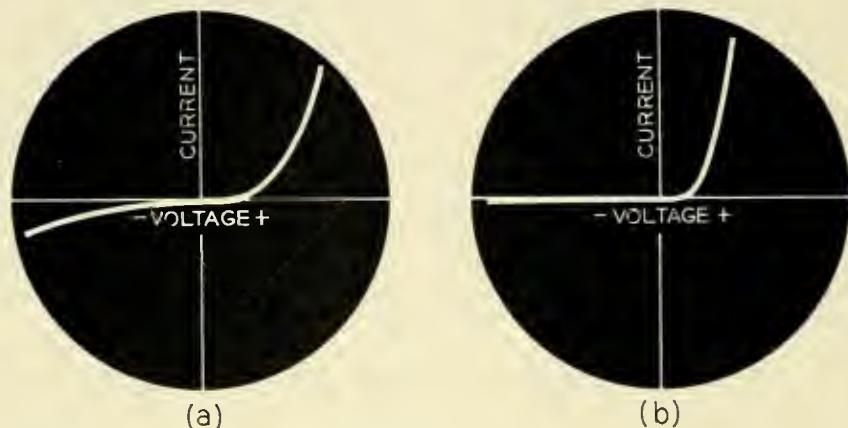


Fig. 4 — Cathode-ray oscilloscope display of wafer unit static characteristic: (a) before and (b) after tapping.

the "S" configuration in a forming jig. By an electrolytic process the spring is then cut to the proper length and pointed. The micro-photographs in Fig. 3 show successive stages in the formation of the contact spring.

In the final assembly of the unit the nickel rod with the contact spring is pressed into place until contact is made with the silicon. It is then advanced a half mil to obtain the proper contact pressure. The voltage-current characteristics as viewed at 60 cycles on a cathode-ray oscilloscope will then appear as shown in Fig. 4(a). The unit is "tapped" into final adjustment. This is done by clamping the unit in a holder and rapping it sharply on the top of a hard wood bench. This procedure requires experience as excessive "tapping" will impair the performance of the unit. Usually one vigorous "tap" is sufficient to produce the desired effect and the voltage-current characteristic will appear as

shown in Fig. 4(b). The static characteristic of a typical unit is shown in Fig. 5.

The conversion loss of each unit is measured before the end of the nickel rod carrying the contact point is cut off flush with the wafer. In the event that this initial measurement shows that the conversion loss exceeds an arbitrarily chosen upper limit (8.5 db), it is possible at this stage to withdraw the point and replace it with a new one. This procedure, which was necessary on only a few of the units processed, always resulted in an acceptable unit. The final operation is to cut off the protruding end of the nickel rod flush with the wafer.

A holder designed to use the wafer units is shown in Figs. 6 and 7. At the input end of the converter block is a short waveguide taper section to match from standard RG98U waveguide to the $\frac{1}{32}$ -inch high waveguide used in the wafer unit. As the wafer unit is moved in and out to match the conductance of the crystal to the waveguide, the output pin of the wafer unit slides in a chuck on the inner conductor of the coaxial

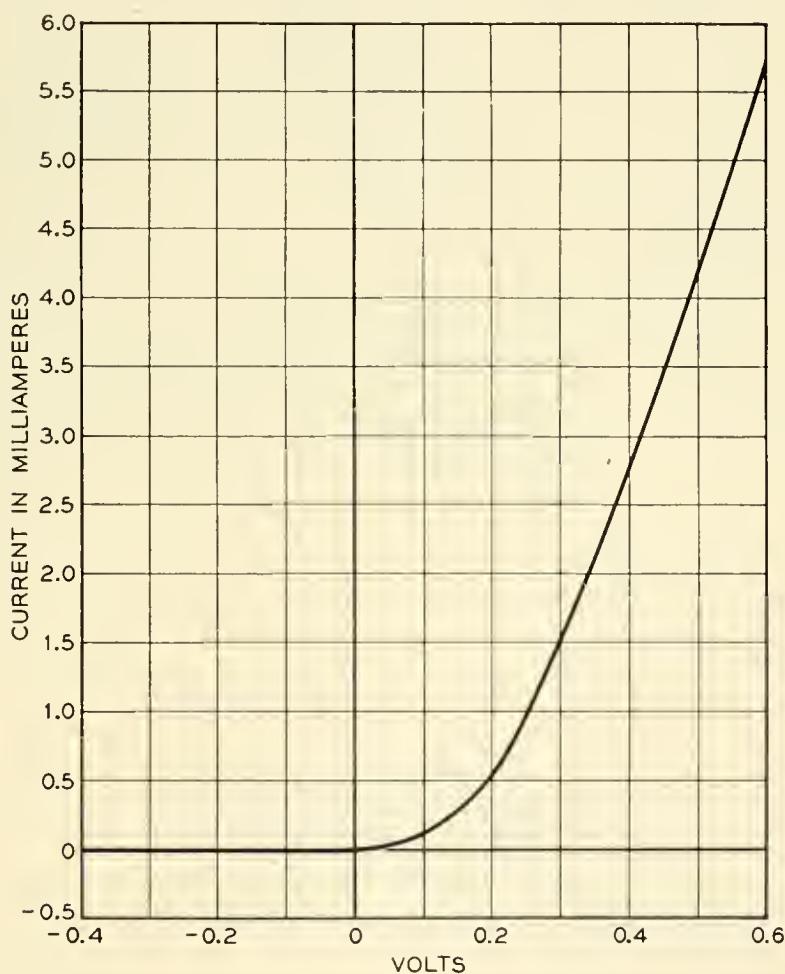


Fig. 5 — Static characteristic of typical millimeter-wave wafer unit.

output jack. The unit may be clamped in position after matching adjustments are made by tightening the knurled thumb screw which pushes a cylindrical slug containing an adjustable piston against the wafer unit. The piston is a short septum which slides in shallow grooves in the top and bottom of the $\frac{1}{32}$ -inch high waveguide, thus dividing the waveguide into two guides which are beyond cut-off. This septum is made of two pieces of thin beryllium copper bowed in opposite directions so that good contact is made to the sides of the grooves in the top and bottom of the waveguide. Since the piston with its connecting rod is very light in weight and is held firmly in place by the spring action of the bowed septum, no additional locking mechanism need be provided. Since the rectifier is essentially broadband by design, the adjustment of the piston is not critical and is readily made by hand. The piston rod is protected by a cap which is snapped in place over the thumb screw when all tuning adjustments are completed.

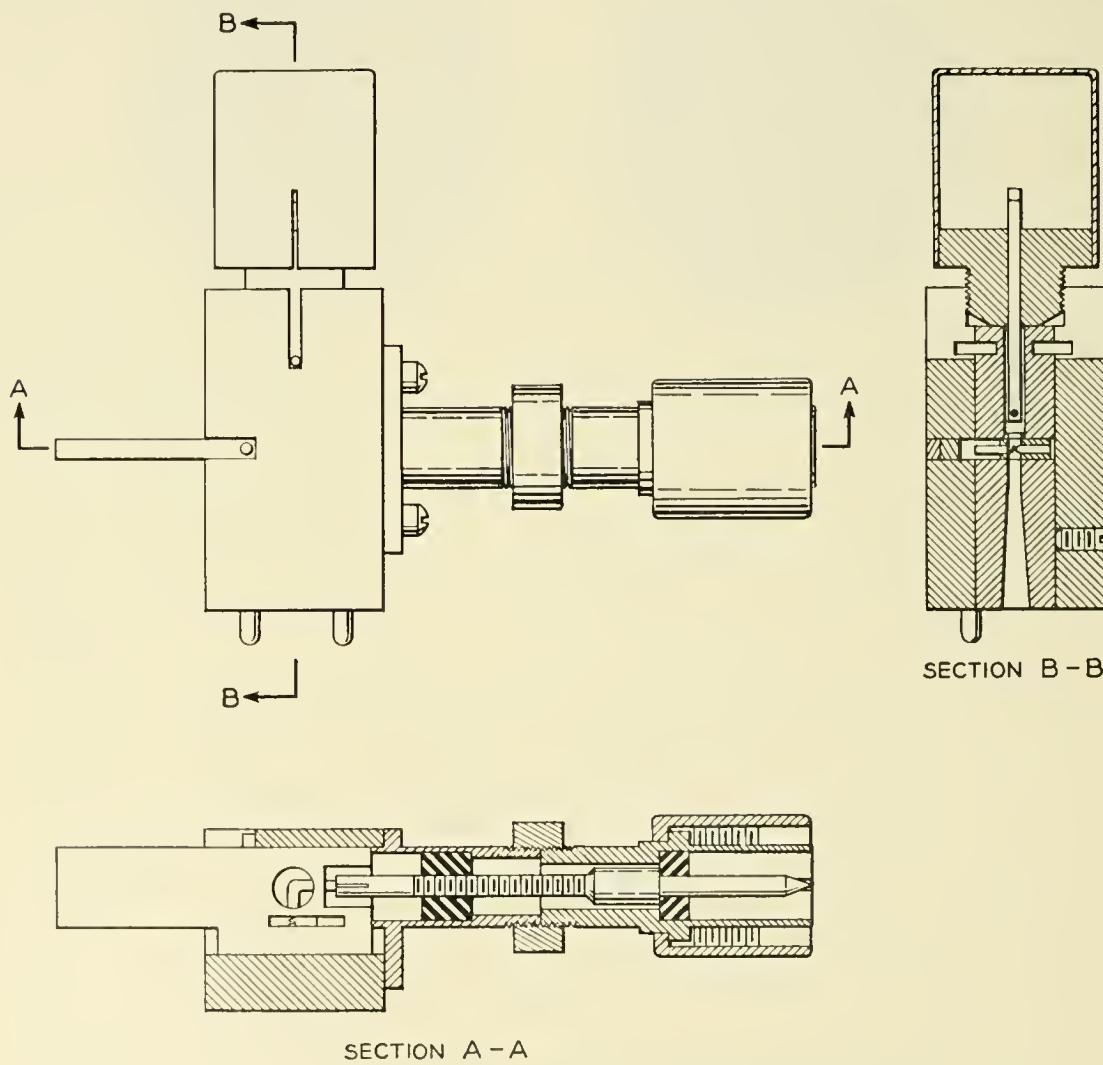


Fig. 6 — Assembly drawing of millimeter-wave converter.

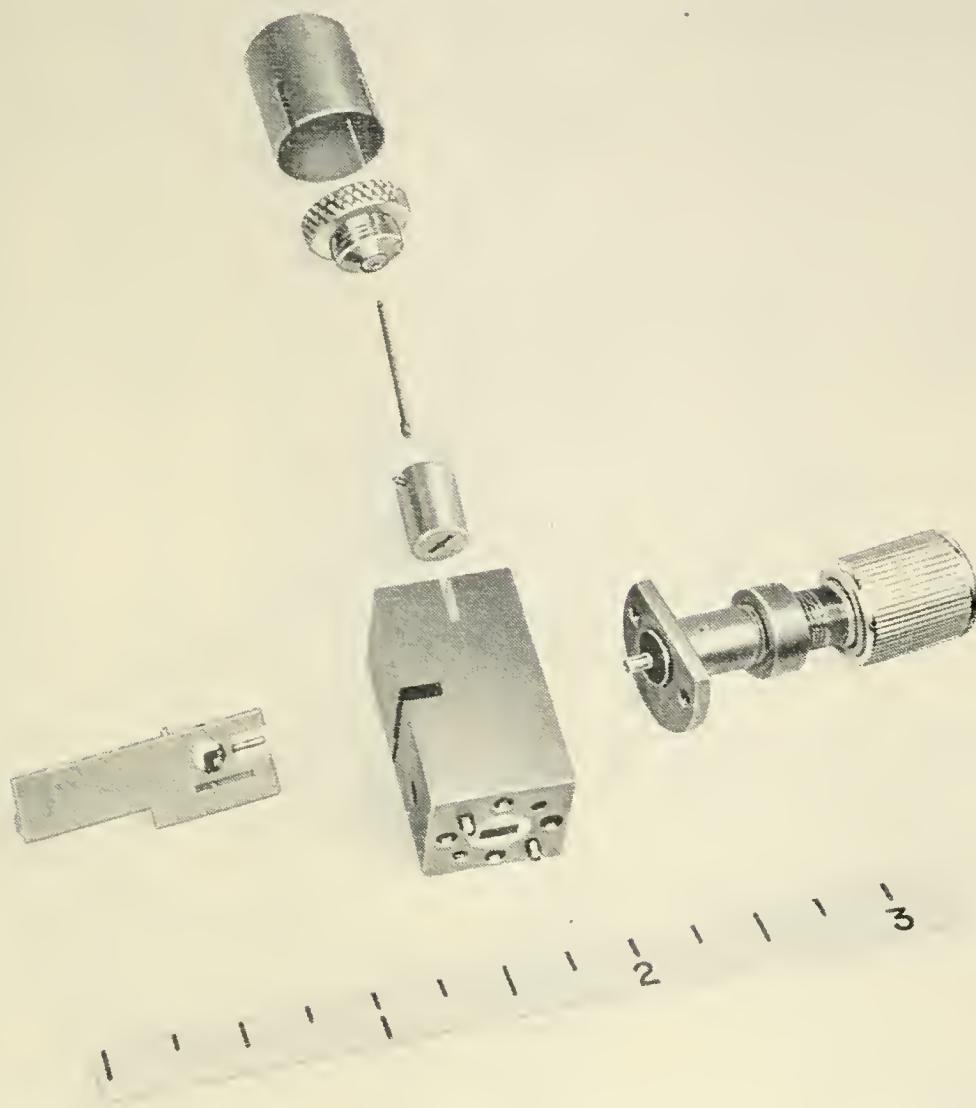


Fig. 7 — Exploded view of millimeter-wave converter.

With the converter fixed-tuned at 5.4 millimeters, a shift in wavelength to 6.3 millimeters (17 per cent change) produces a mismatch loss of from 1.6 to 4.0 db depending on the rectifier used.

PERFORMANCE DATA FOR WAFER-TYPE RECTIFIER UNIT

A pilot group of one hundred wafer units was processed and measured. Figs. 8, 9 and 10 are bar graphs of the distribution of the conversion loss L , and noise ratio N_R^* , and the 60 megacycle intermediate frequency output impedance Z_{IF} , for the hundred rectifiers measured in the

* N_R is the ratio of the noise power available from the rectifier to the noise power available from an equivalent resistor at room temperature.

mixer of Fig. 7 at a wavelength of 5.4 millimeters. In order that the measurements might be more readily compared with those made on commercially available rectifiers used at longer wavelengths, the available beating oscillator power was maintained at a level of one milliwatt for all measurements.* Further, in the case of the conversion loss, a

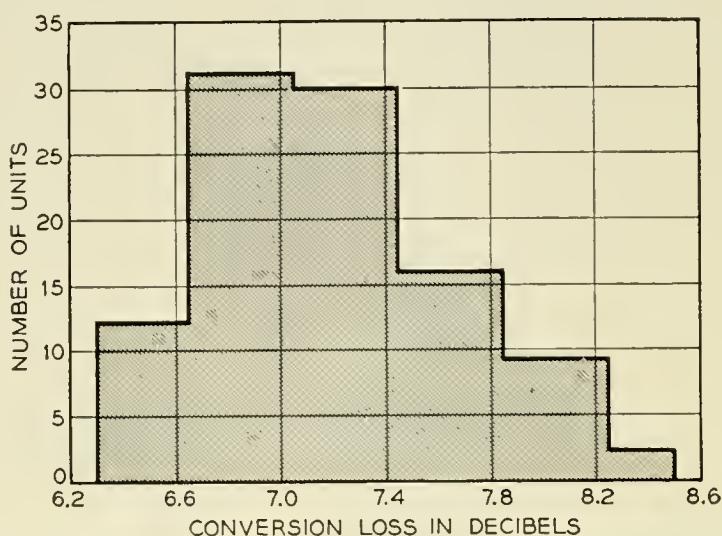


Fig. 8—Conversion loss (L) of 100 wafer units at a wavelength of 5.4 millimeters with one-milliwatt beating oscillator drive (average 7.2 db).

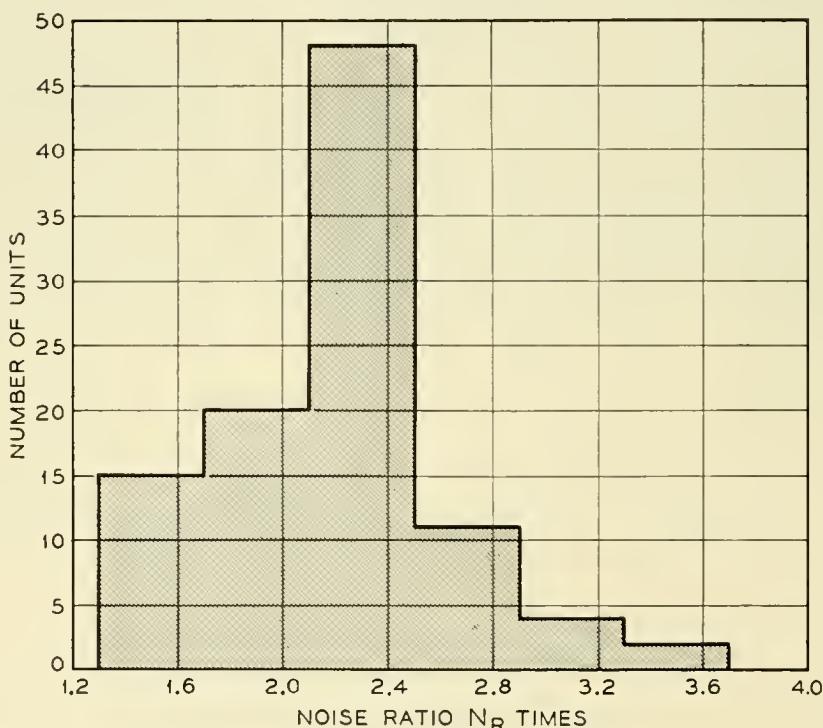


Fig. 9—Noise Ratio (N_r) for 100 wafer units at a wavelength of 5.4 millimeters with a one-milliwatt beating oscillator drive (average 2.21 times).

* Power levels were determined by the use of a calorimeter. See, A Calorimeter for Power Measurements at Millimeter Wavelengths, I. R. E. Trans., MTT-2, pp. 45-47, Sept., 1954.

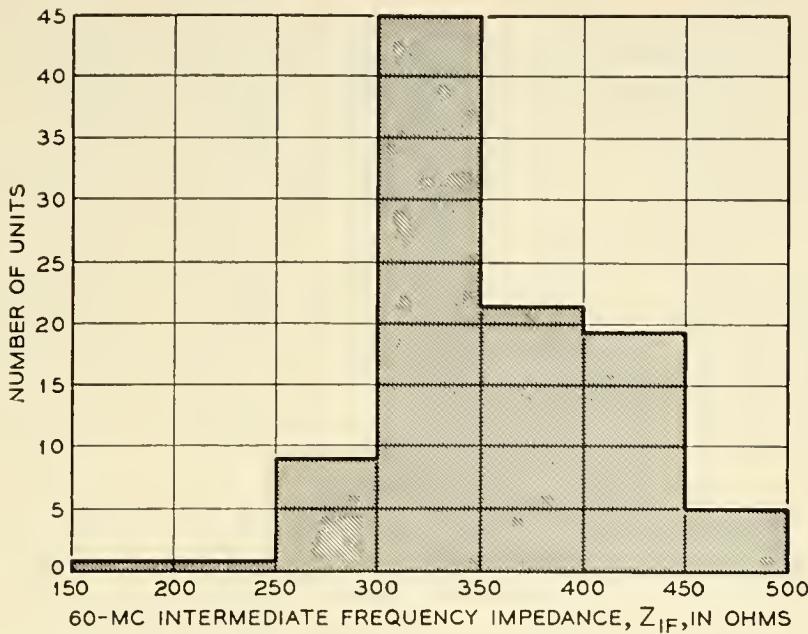


Fig. 10 — Sixty-megaycle intermediate-frequency output impedance (Z_{IF}) for 100 wafer units with one milliwatt beating oscillator drive (average 338 ohms)

limit of 8.5 db was arbitrarily adopted. This required the readjustment of eleven units, with a new point inserted in each case. No units were rejected because of high noise and none of the hundred units processed was lost.

From the bar graphs it may be seen that the wafer units have the average characteristics shown in the accompanying table at a wavelength of 5.4 millimeters.*

Conversion Loss L	7.2 db (5.3 times)
Noise Ratio N_R	2.2 times
IF Impedance (60 me) Z_{IF}	338 ohms

Knowing the noise figure, N_{IF} , of the *IF* amplifier intended for use with the rectifiers, the overall receiver noise figure, N_{REC} , may be calculated by the following formula (using numerical ratios):

$$N_{REC} = L(N_R - 1 + N_{IF})$$

Assuming an *IF* amplifier noise figure of 4.0 db ($2\frac{1}{2}$ times) and the average values of "L" and " N_R " given above for the millimeter wafer units, we have for the case of a noiseless beating oscillator;

$$N_{REC} = 5.3 (2.2 - 1 + 2.5) \approx 20 \text{ (13 db)}$$

* A few wafer units have also been measured at a wavelength of 4.16 millimeters. The conversion losses averaged about 1.6 db greater than those measured at a wavelength of 5.4 millimeters.

TABLE I—COMPARISON OF LOW-POWER CHARACTERISTICS OF CARTRIDGE-TYPE AND WAFER-TYPE RECTIFIERS

Test Conditions	JAN Specifications for Cartridge-Type Rectifiers		Performance of Wafer-Type Rectifiers
	IN26	IN53	
Frequency.....	23984 mc	34860 mc	55500 mc
Beating oscillator power level.....	1.0 milliwatts	1.0 milliwatts	1.0 milliwatts
Noise reference resistor.....	300 ohms	300 ohms	300 ohms
Conversion loss.....	8.5 db (max)	8.5 db (max)	8.5 db (max)*
Noise ratio.....	2.5 (max)	2.5 (max)	2.2 (average)†
Nominal IF impedance range.....	300 to 600 ohms	400 to 800 ohms	250 to 500 ohms

* Limit arbitrarily set on basis of 100 per cent yield as explained in the text.

† Limit not set. Actually in more recent production N_R has averaged 1.7 times.

In practice, the beating oscillator noise sidebands can be eliminated by the use of a matched pair of rectifiers in a balanced converter arrangement described later. The resulting overall noise figure of 13 db on an average compares quite favorably with the figures obtained at longer wavelengths.

In Table I it is seen that a high percentage of the group of one hundred units would be able to pass low-power JAN specifications similar to those set down for the commercially available IN26 and IN53 rectifiers used at longer wavelengths.

EFFECT OF VARYING THE BEATING OSCILLATOR POWER

When the optimum over-all receiver noise figure is desired, it may well turn out that a beating oscillator drive of one milliwatt (corresponding to a dc rectified current for different wafers of from $\frac{9}{10}$ to $1\frac{1}{4}$ milliamperes) is too large. Fig. 11 shows the effect on the performance of a typical unit as the beating oscillator drive is varied above and below the one milliwatt level as indicated by the change in the dc rectified current. It is seen that the value of N_R tends to increase rapidly for a beating oscillator drive much in excess of one milliwatt; with reduced drive, the over-all noise figure of the receiver, N_{REC} for the example taken, improves, reaching a minimum value near a rectified current of about $\frac{7}{10}$ milliampere corresponding to a drive of about $\frac{2}{3}$ of a milliwatt.

A BALANCED CONVERTER FOR WAFER UNITS

A broad-band balanced first converter has been developed which makes use of a pair of wafer-type millimeter-wave rectifiers. This converter

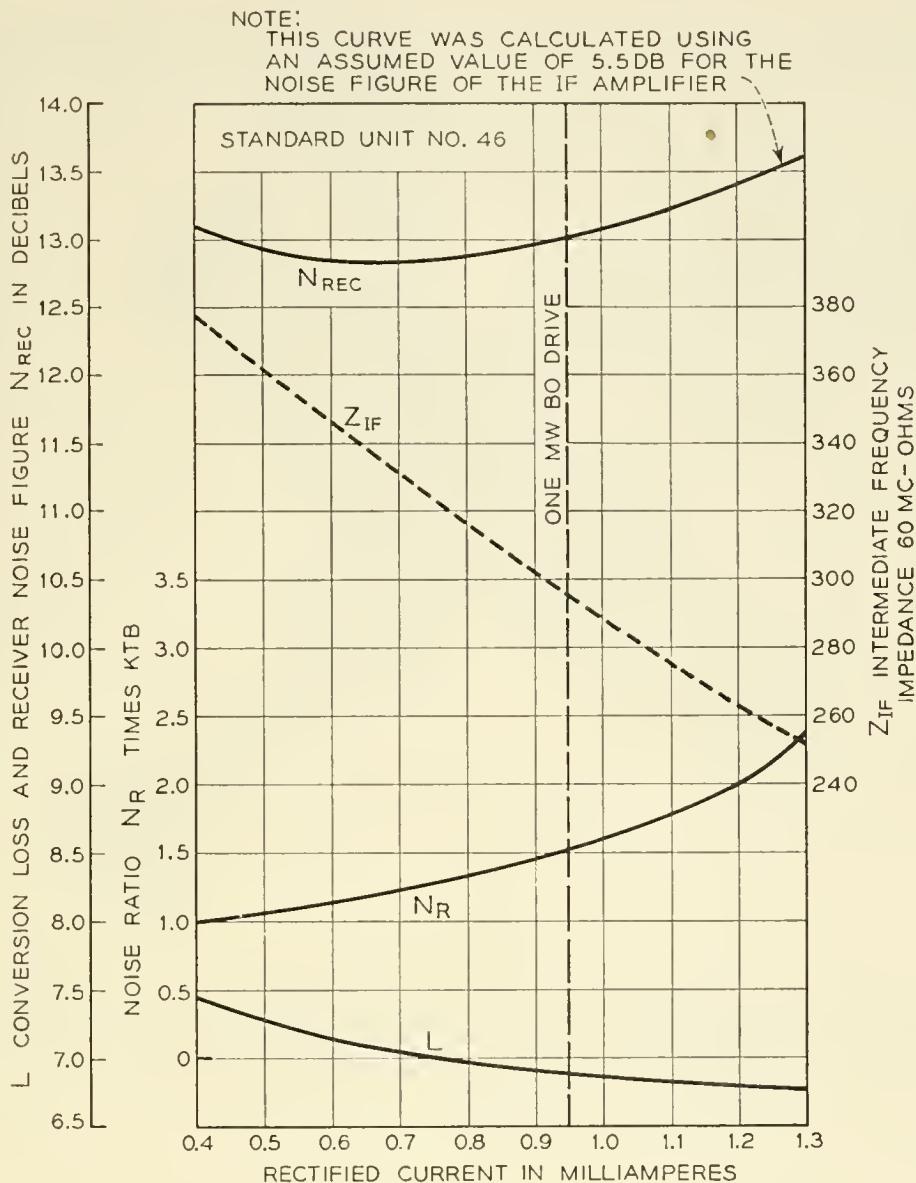


Fig. 11 — Typical performance curves for wafer-type rectifiers.

was designed to operate over the 4- to 7-millimeter band and is pictured in Fig. 12. A compact arrangement has been achieved which makes use of a waveguide finline-to-coaxial input circuit for the beating oscillator while the signal is introduced through a separate impedance-matched waveguide "Tee" section. Return loss measurements show that with a matched pair of wafer units, fixed-tuned in the center of the 5- to 6-millimeter band, an excess loss of about 1 db may be expected at the edges of a 15 per cent band. At midband, an improvement of 5 db in over-all receiver noise figure was obtained by substituting the balanced converter for an unbalanced one in a test receiver using an M1805 milli-

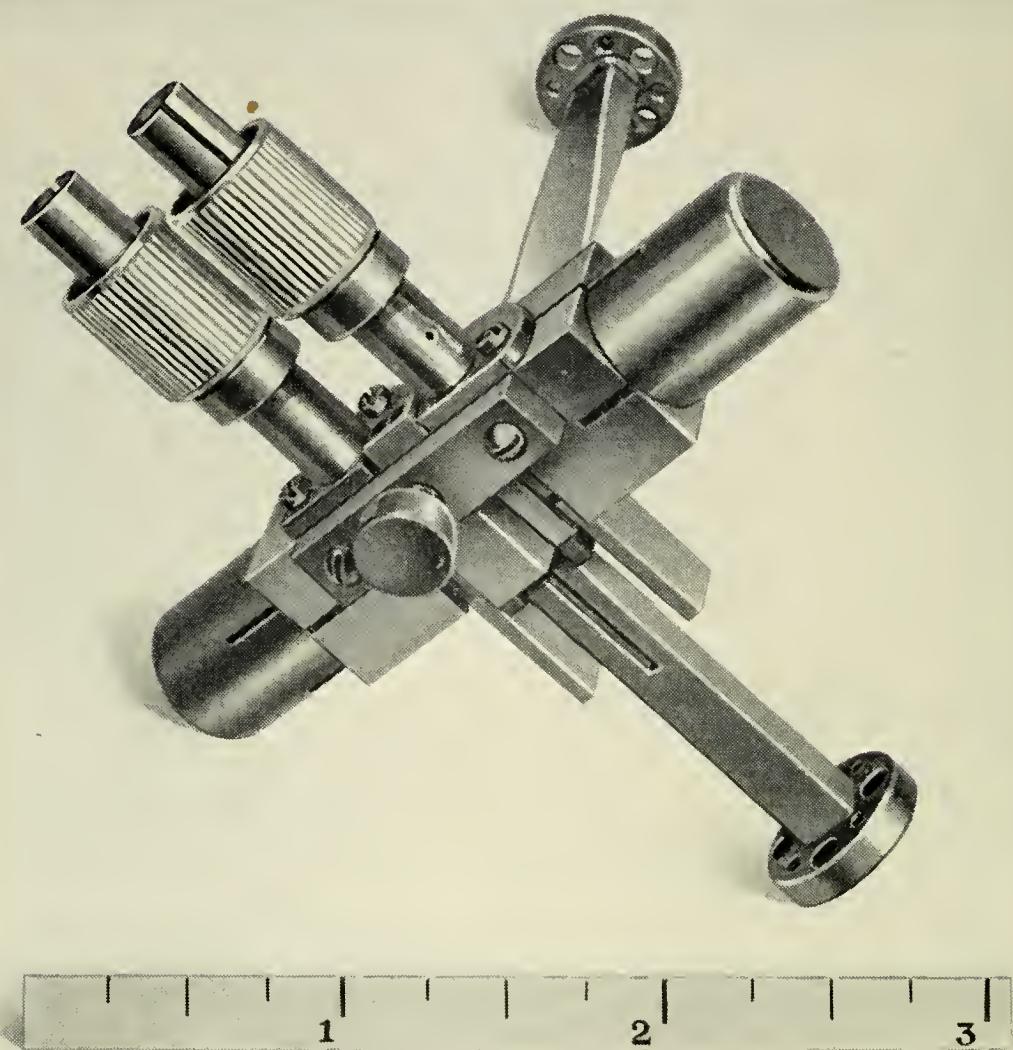


Fig. 12 — Balanced converter with wafer-type rectifiers.

meter-wave reflex klystron* as the beating oscillator and a 60-mc intermediate frequency amplifier with a 5-db noise figure.

REVERSED POLARITY WAFER UNIT

When using crystal rectifiers in a balanced converter arrangement, there is a distinct advantage, circuit-wise, in using two units of opposite polarity. For this reason, a reversed-polarity wafer type rectifier has also been developed. This was done by interchanging the silicon and the

* E. D. Reed, A Tunable, Low-Voltage Reflex Klystron for Operation in the 50 to 60 kmc Band, *B. S. T. J.*, **34**, pp. 563-599, May, 1955.

point contact spring in a standard unit. The standard and reverse-polarity wafer have the same outer physical dimensions and thus they may be used interchangeably in the holders as dictated by the specific problems at hand.

CONCLUDING REMARKS

Aside from their intended use as first detectors in double detection receivers, wafer units have been used for single detection measurements at frequencies as high as 107 kmc.

It is felt that the pilot production group of one hundred units is a sample of sufficient size to yield representative data and to demonstrate the practicability of the design. It should be pointed out that the units have not been filled with protective waxes and have not been subjected to temperature-humidity cycling tests. However, a few reference units have been in use in the laboratory for over a year and have shown no measurable deterioration. No attempt has been made to establish a burn-out rating for the rectifier, but units have withstood available cw input powers of the order of 15 milliwatts and narrow pulse discharges of the order of $\frac{1}{10}$ erg without causing noticeable changes in the conversion loss or noise ratio.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to H. T. Friis and A. B. Crawford for their helpful suggestions and guidance during the course of this work. Extensive use has also been made of the experience and techniques of R. S. Ohl. E. F. Elbert participated in the development of the wafer unit, being particularly concerned with the techniques of fabrication. H. W. Anderson and S. E. Reed were most helpful in solving mechanical problems encountered in the production of wafer units and holders.

APPENDIX

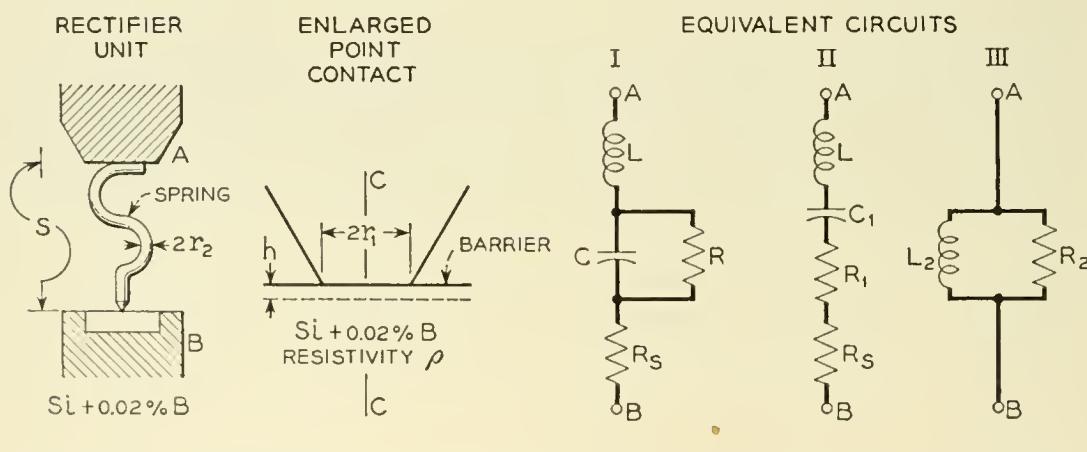
This section describes some calculations that were made for the purpose of estimating the values of the various parameters involved in the design of a high frequency point contact rectifier. These parameters are the barrier resistance, the spreading resistance, the capacitance of the barrier layer and the inductance of the contact spring. Knowing the approximate values of these parameters one can, by an equivalent circuit analysis, arrive at a simple parallel circuit for the rectifier which may

be used in designing an appropriate holder. Also, using this equivalent circuit, one may calculate the bandwidth expected for the converter.

Fig. 13 shows the point contact rectifiers under consideration and an enlarged view of the point contact region. On the right of the figure are shown equivalent circuits of the rectifier. Circuit I is the generally accepted circuit of a point contact rectifier. The true circuit for a rectifier operating at millimeter wavelengths is probably more complicated than that shown in the figure but, for an approximate analysis, the simplified circuit has been found to yield useful results. In the following paragraphs, values are derived for the parameters of this equivalent circuit. MKS units are used and values appropriate to the millimeter wave wafer unit are used as examples.

Spreading Resistance

The spreading resistance, R_s , may be calculated if we know the resistivity of the silicon used for the rectifier and the radius of the contact area formed when the units are assembled. For DuPont high-purity silicon, doped with 0.02 per cent boron by weight, W. Shockley* gives the resistivity, ρ , as 0.90×10^{-4} ohm meters. From numerous measurements on millimeter wave contact areas, R. S. Ohl finds the contact radius, r_1 , to be about 1.25×10^{-6} meters. The spreading resistance,



L = INDUCTANCE OF SPRING

C = CAPACITANCE OF BARRIER LAYER

R = RESISTANCE OF BARRIER LAYER

R_s = SPREADING RESISTANCE

$$\frac{1}{\omega C_1} = \frac{1}{\omega C \left(1 + \frac{1}{(\omega CR)^2}\right)}$$

$$\omega L_2 = \omega L - \frac{1}{\omega C_1} + \frac{(R_s + R_1)^2}{\omega L - \frac{1}{\omega C_1}}$$

$$R_1 = \frac{R}{1 + (\omega CR)^2}$$

$$R_2 = R_s + R_1 + \frac{\left(\omega L - \frac{1}{\omega C_1}\right)^2}{R_s + R_1}$$

Fig. 13 — Point contact rectifier and equivalent circuits.

* W. Shockley, Electrons and Holes in Semiconductors, New York: D. Van Nostrand Co., Inc., 1950, p. 284.

R_s , assuming a circular contact area, may be calculated from the formula, $R_s = \rho/4r_1$.* For the above example, $R_s = 18$ ohms.

Barrier Resistance

The approximate operating value of the barrier resistance, R , may be determined from a knowledge of the intermediate frequency impedance of a typical rectifier. A. B. Crawford has shown that the optimum intermediate frequency output impedance of a crystal mixer rectifier is a function of the exponent of the static characteristic of the rectifier and the impedance presented to the rectifier at the image and signal frequencies. This information is presented in Fig. 12.3-6 in G. C. Southworth's book.† In the millimeter wave case it is a good assumption that the impedances for the signal and image frequencies are equal; for this case and for matched conditions, the magnitude of the high frequency impedance is seen to be a simple multiple of the intermediate frequency impedance R_{IF} .

From numerous measurements on mixer rectifiers operating at different frequencies it is known that the intermediate frequency impedance of an average rectifier is very nearly 400 ohms. We also know from the DC static characteristics of our millimeter wave type rectifiers that the average exponent is about four. With this information, and the curves in Southworth's book, it is found that $R \approx R_{IF}/1.5$. Thus, the barrier resistance R is about 250 ohms.‡

Capacitance of Barrier Layer

From a knowledge of the point contact area, the barrier layer thickness, and the dielectric constant of the silicon, the capacitance of the point contact may be calculated. The radius of the contact point area is the same as that used for the calculation of the spreading resistance. The barrier layer thickness, h , for the heat treated silicon used for millimeter waves has been measured by R. S. Ohl to be about 10^{-8} meters. The dielectric constant of silicon is $\epsilon_r = 13$. The capacitance is given by the following formula

$$C = \frac{r_1^2 \epsilon_r}{3.6h \times 10^{10}} \text{ farads} \quad (1)$$

* J. H. Jeans, Mathematical Theory of Electricity and Magnetism, 5th Ed., Cambridge University Press, 1925.

† G. C. Southworth, Principles and Applications of Waveguide Transmission, New York: D. Van Nostrand Co., Inc., 1950.

‡ This resistance cannot be readily measured directly at millimeter waves.

For the above case $C = 5.7 \times 10^{-14}$ farads or $1/\omega C$ at 5.4 millimeters is about 50 ohms.

The accuracy of this capacitance calculation can be verified later when a completed rectifier is measured for its high frequency conversion loss. This is possible because we know the calculated low frequency conversion loss of the rectifier, for the case of zero spreading resistance from Southworth's book, Fig. 12.3-7. For an exponent of four this loss is given as 4.4 db. The additional loss at high frequency due to the capacitance, C , may be calculated (see Equivalent Circuit II) by the formula:

$$\text{Additional Loss} = 10 \log_{10} \frac{R_1 + R_s}{R_1} \text{ db} \quad (2)$$

From the text (Fig. 8), it is seen that the average wafer rectifier unit has a conversion loss at 5.4 millimeters of 7.2 db; thus, the difference between the low and high frequency conversion losses is very nearly 3 db. This means that about one-half the signal power is lost in the spreading resistance; hence R_1 and R_s are about equal. By transferring back to Equivalent Circuit I, the average value of the capacitance is found to be 4.1×10^{-14} farads, which is a reasonable check with the calculated value given by (1).

Inductance of the Contact Spring

The remaining parameter of the equivalent circuit to be determined is the inductance of the contact spring. The value of the equivalent parallel resistance, R_2 , depends on the inductance L (the other parameters being fixed), or conversely, for a given value of R_2 , the appropriate value for L may be calculated from the formula for Equivalent Circuit III. For an off-center match of the rectifier to the waveguide, R_2 must equal the guide impedance, Z_d , at the rectifier location. Also, for a match, the distance, ℓ , from the rectifier to the waveguide piston must

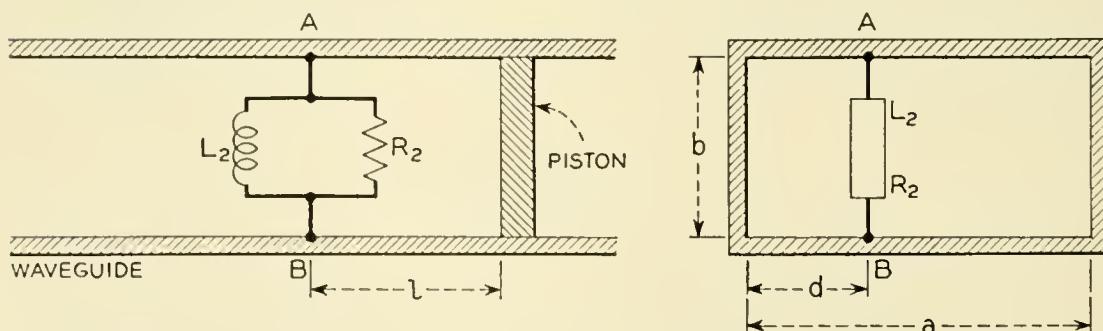


Fig. 14 — Matching circuit for rectifier offset in waveguide.

satisfy the relation, $Z_d \tan 2\pi\ell/\lambda_g = -\omega L_2$. (See Fig. 14.) The impedance of the guide as a function of d/a is given by,

$$Z_d = 240\pi \frac{b}{a} \frac{1}{\sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}} \sin^2 \frac{\pi d}{a} \quad (3)$$

As a compromise between electrical and mechanical requirements, a waveguide height, b , of $\frac{1}{32}$ inch was chosen for the wafer unit; the width of the guide was taken to be the same as RG98U. For $b = 7.88 \times 10^{-4}$, $a = 3.76 \times 10^{-3}$, $d/a = \frac{1}{4}$ and $\lambda = 5.4 \times 10^{-3}$, (3) gives a value of 113 ohms for Z_d (and R_2). The appropriate value for L then becomes 3.38×10^{-10} henries.

An estimate of the size of a contact spring having the inductance given above can be made from the formula below which gives the inductance of a straight thin wire of length S as a function of its sidewise position in the waveguide.* (See Fig. 15.)

$$L = 2S \log_e \frac{2a \sin \frac{\pi d}{a}}{\pi r_2} \times 10^{-7} \text{ henries} \quad (4)$$

$r_2 \ll d$

For $d/a = \frac{1}{4}$ and $2r_2 = 2.28 \times 10^{-5}$ (0.9×10^{-3} inches), the length, S , is found to be about 3.38×10^{-4} meters or about 0.013 inches.

Since the spring must be so very small, the circuit from the base of the spring to the waveguide wall is completed with a large low inductance conical post as shown in Fig. 2 of the text.

Bandwidth Calculation

Having assigned values to all the parameters of the equivalent circuit, it is now possible to calculate the mismatch loss of a fixed-tune

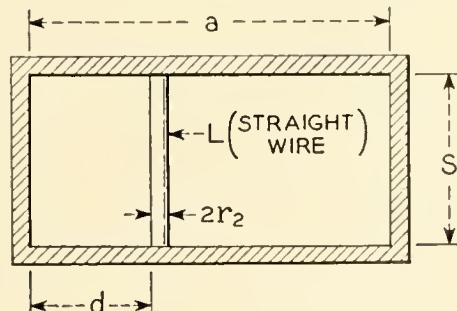


Fig. 15 — Thin wire in waveguide.

* Private communication from S. A. Schelkunoff.

converter for a given change in operating wavelength. This loss is given by the following formula:

Mismatch loss

$$= 10 \log_{10} \frac{4Z_d}{R_2 \left[\left(1 + \frac{Z_d}{R_2} \right)^2 + \left(\frac{Z_d}{\omega L_2} + \frac{1}{\tan 2\pi\ell/\lambda_g} \right)^2 \right]} \text{ db} \quad (5)$$

For the wafer unit, calculation shows that the rectifier is matched to the waveguide at a wavelength of 5.4×10^{-3} meters for $d/a = \frac{1}{4}$ and $\ell = 3.14 \times 10^{-3}$. If now the wavelength is changed to 6.3×10^{-3} meters, without retuning (17 per cent change) the mismatch loss calculated by (5) is 1.6 db. It was stated in the text that a number of wafer units gave measured mismatch losses of from 1.6 to 4.0 db for a 17 per cent change in wavelength without retuning. This is considered to be a reasonable correlation between calculations and measurements.

Frequency Conversion by Means of a Nonlinear Admittance

C. F. EDWARDS

(Manuscript received June 20, 1956)

This paper gives a mathematical analysis of a heterodyne conversion transducer in which the nonlinear element is made up of a nonlinear resistor and a nonlinear capacitor in parallel. Curves are given showing the change in admittance and gain as the characteristics of the nonlinear elements are varied. The case where a conjugate match exists at the terminals is treated.

It is shown that when the output frequency is greater than the input frequency, modulators having substantial gain and bandwidth are possible, but when the output frequency is less than the input frequency, the converter loss is greater than unity and is little affected by the nonlinear capacitor. The conditions under which a conjugate match is possible are specified and it is concluded that a nonlinear capacitor alone is the preferred element for modulators and that a nonlinear resistor alone gives the best performance in converters.

INTRODUCTION

Point contact rectifiers using either silicon or germanium are used as the nonlinear element in microwave modulators to change an intermediate frequency signal to an outgoing microwave signal and in receiving converters to change an incoming microwave signal to a lower intermediate frequency. Most point contact rectifiers now in use behave as pure nonlinear resistors as evidenced by the fact that in either of the above uses the conversion loss is the same. In recent experiments with heterodyne conversion transducers* using point contact rectifiers made with ion bombarded silicon this was found to be no longer true. The conversion loss of the modulator was found to be unusually low and

* This term is defined in American Standard Definitions of Electrical Terms — ASA C42 — as "a conversion transducer in which the useful output frequency is the sum or difference of the input frequency and an integral multiple of the frequency of another wave".

that of the converter was several decibels greater. In one instance the loss in a modulator used to convert a 70 mc signal to one at 11,130 mc was found to be only 2.3 db but when the direction of transmission through it was reversed and it was used as a converter, the loss was 7.8 decibels.

* Similar effects were observed several years ago in conversion transducers using welded contact germanium rectifiers.¹ In these early experiments substantial converter gain and negative conductance at the intermediate frequency terminals were also observed. These results were accounted for by assuming the presence of a nonlinear capacitance at the point contact in parallel with the nonlinear resistance. At that time attention was devoted mainly to the behavior of converters where noise is a vital factor. It was found that although the conversion loss could be reduced, the noise temperature increased and no improvement in noise figure resulted. However, the noise temperature requirements in modulators are much less severe and the nonlinear capacitance effect is useful and can substantially improve the performance.

THEORY

The mathematical analysis given here was undertaken in order to clarify the effect of the nonlinear capacitance in the frequency conversion process and to obtain an estimate of the usefulness of modulators exhibiting gain. The analysis is restricted to the simplest case in which signal voltages are allowed to develop across the nonlinear elements at the input and output frequencies only. This is not an unrealistic restriction since the conversion transducers used in microwave relay systems have filters associated with them which suppress the modulation products outside the signal band. The final results will be given only for those conditions which permit a conjugate match at the input and output of the transducer.

The procedure used to obtain expressions for the admittance and gain of conversion transducers utilizing a nonlinear element made up of a nonlinear resistance and a nonlinear capacitance in parallel follows the commonly used method of treating the nonlinear elements as local oscillator controlled linear time varying elements.² The current through the nonlinear resistor is a function of the applied voltage. The derivative of this function is the conductance as a function of the applied voltage. Thus when the local oscillator is applied, the conductance varies at the local oscillator frequency and the conductance as a function of time may be obtained. This is periodic and may be expressed as a Fourier series. The conductance is real and if we make the usual assumption that

it may be expressed as an even function of time, we may write

$$\gamma = \dots + G_2 e^{-j2\omega_0 t} + G_1 e^{-j\omega_0 t} + G_0 + G_1 e^{j\omega_0 t} + G_2 e^{j2\omega_0 t} + \dots \quad (1)$$

where $\omega_0/2\pi$ is the local oscillator frequency f_0 and the Fourier coefficients G_n are real. Similarly the charge on the nonlinear capacitor is a function of the applied voltage. The derivative of this function is the capacitance as a function of the applied voltage. The application of the local oscillator thus causes the capacitance to vary at the local oscillator frequency so that it also may be expressed as a Fourier series. The capacitance κ is real, and assuming it may be expressed as an even function of time, we have

$$\kappa = \dots + C_2 e^{-j2\omega_0 t} + C_1 e^{-j\omega_0 t} + C_0 + C_1 e^{j\omega_0 t} + C_2 e^{j2\omega_0 t} + \dots \quad (2)$$

It is assumed that the current and charge functions are single valued and that their derivatives are always positive.

When a small signal voltage v is applied to the nonlinear resistor, the signal current through the resistor is given by γv . When it is applied to the nonlinear capacitor the charge on the capacitor is κv . The total current i which flows through the two nonlinear elements connected in parallel thus becomes

$$i = \gamma v + \frac{d}{dt} (\kappa v) \quad (3)$$

v of course must be small and not affect the value of γ and κ .

Fig. 1 shows a heterodyne conversion transducer made up of a nonlinear resistor and a nonlinear capacitor in parallel driven by an internal local oscillator. f_1 is the signal frequency at the terminals 1-2, and y_1 is the external admittance connected to these terminals. The signal frequency at the terminals 3-4 is f_2 , and y_2 is the external admittance.

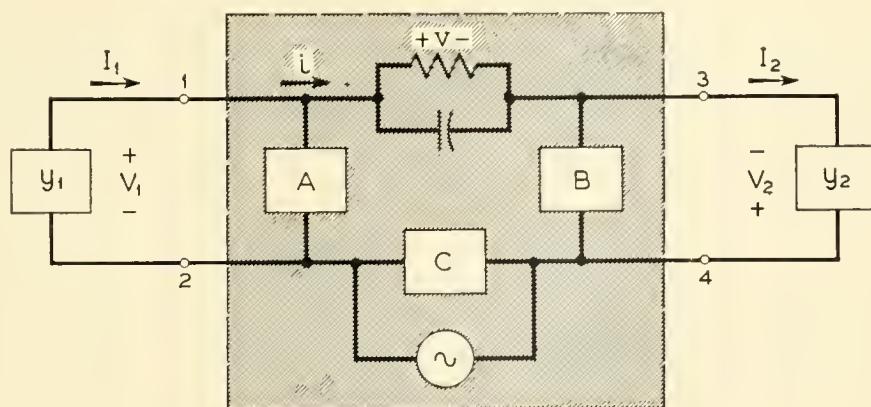


Fig. 1 — Heterodyne conversion transducer.

A , B and C are ideal frequency selective networks whose admittances are zero at f_1 , f_2 and f_0 respectively, and infinite at all other frequencies. This circuit permits the application of the local oscillator voltage to the nonlinear elements but permits signal voltages to develop across them at f_1 and f_2 only. Similarly, signal currents at frequencies other than f_1 and f_2 encounter no external impedance, so they cannot alter the signal voltage or contribute to the external power. This, of course, assumes that if the nonlinear element is a point contact rectifier the spreading resistance normally present is negligible.

If f_1 is a frequency less than half the local oscillator frequency f_0 (it is generally very much less), the network B can be selected to make f_2 either $f_0 + f_1$, or $f_0 - f_1$. To distinguish between the two cases, we will call the former a noninverting conversion transducer since an increase in one signal frequency causes an increase in the other. The latter will be called an inverting conversion transducer since an increase in one signal frequency results in a decrease in the other. When y_1 contains the generator and y_2 the load, the device becomes a modulator. When y_2 contains the generator and y_1 the load, it is a converter.

The real part of the signal voltage may be written

$$v = V_1 e^{j\omega_1 t} + V_1^* e^{-j\omega_1 t} + V_2 e^{j\omega_2 t} + V_2^* e^{-j\omega_2 t} \quad (4)$$

where V^* is the complex conjugate of V and $\omega = 2\pi f$. Similarly, the real part of the signal current may be written

$$i = I_1 e^{j\omega_1 t} + I_1^* e^{-j\omega_1 t} + I_2 e^{j\omega_2 t} + I_2^* e^{-j\omega_2 t} \quad (5)$$

If we multiply equations (1) and (4) and retain only those terms containing f_1 and f_2 we obtain, in the case of the non-inverting conversion transducer where $f_2 = f_0 + f_1$,

$$\begin{aligned} \gamma v &= [G_0 V_1 + G_1 V_2] e^{j\omega_1 t} + [G_1 V_1 + G_0 V_2] e^{j\omega_2 t} \\ &\quad + [G_0 V_1^* + G_1 V_2^*] e^{-j\omega_1 t} + [G_1 V_1^* + G_0 V_2^*] e^{-j\omega_2 t} \end{aligned} \quad (6)$$

Similarly, if we multiply (2) and (4) we get an expression like (6) with the G 's replaced by C 's. If we differentiate this expression we get

$$\begin{aligned} \frac{d}{dt} (\kappa v) &= j\omega_1 [C_0 V_1 + C_1 V_2] e^{j\omega_1 t} + j\omega_2 [C_1 V_1 + C_0 V_2] e^{j\omega_2 t} \\ &\quad - j\omega_1 [C_0 V_1^* + C_1 V_2^*] e^{-j\omega_1 t} - j\omega_2 [C_1 V_1^* + C_0 V_2^*] e^{-j\omega_2 t} \end{aligned} \quad (7)$$

When we perform the addition indicated by (3) and compare the result with (5) we obtain

$$\begin{aligned} I_1 &= [G_0 + j\omega_1 C_0] V_1 + [G_1 + j\omega_1 C_1] V_2 \\ I_2 &= [G_1 + j\omega_2 C_1] V_1 + [G_0 + j\omega_2 C_0] V_2 \end{aligned} \quad (8)$$

Going through the same steps for the inverting conversion transducer where $f_2 = f_0 - f_1$ we obtain

$$\begin{aligned} I_1 &= [G_0 + j\omega_1 C_0] V_1 + [G_1 + j\omega_1 C_1] V_2^* \\ I_2^* &= [G_1 - j\omega_2 C_1] V_1 + [G_0 - j\omega_2 C_0] V_2^* \end{aligned} \quad (9)$$

Equations (8) and (9) are in the form

$$\begin{aligned} I_1 &= Y_{11} V_1 + Y_{12} V_2 \\ I_2 &= Y_{21} V_1 + Y_{22} V_2 \end{aligned} \quad (10)$$

A heterodyne conversion transducer may thus be represented by a linear 4-pole, and the admittance and gain of the 4-pole may be expressed in terms of the admittance coefficients. In Fig. 1 we see that the admittance of the 4-pole y_1' at the terminals 1–2 is equal to I_1/V_1 and the admittance y_2 connected to terminals 3–4 is $-I_2/V_2$. Putting these in (10) we find

$$y_1' = Y_{11} - \frac{Y_{12} Y_{21}}{Y_{22} + y_2} \quad (11)$$

Similarly the admittance of the 4-pole y_2' at the terminals 3–4 is I_2/V_2 and the admittance y_1 connected to terminals 1–2 is $-I_1/V_1$. Putting these in (10) gives

$$y_2' = Y_{22} - \frac{Y_{12} Y_{21}}{Y_{11} + y_1} \quad (12)$$

To compute the gain of the 4-pole when y_1 contains the generator and y_2 the load, it is convenient to assume a current source connected across y_1 . If the current from this source is I_0 we have $I_1 = I_0 - y_1 V_1$. I_2 equals $-y_2 V_2$ as before. Putting these in (10) gives

$$\frac{I_0}{V_2} = Y_{12} - \frac{(Y_{11} + y_1)(Y_{22} + y_2)}{Y_{21}} \quad (13)$$

If we let $y_1 = g_1 + jb_1$ and $y_2 = g_2 + jb_2$, the power in the load is $V_2^2 g_2$ and the power available from the generator is $I_0^2/4g_1$. Therefore the transducer gain Γ_{12} defined as the ratio of the power in y_2 to that available from y_1 becomes

$$\Gamma_{12} = 4g_1 g_2 \frac{V_2^2}{I_0^2} = 4g_1 g_2 \left| \frac{Y_{21}}{Y_{12} Y_{21} - (Y_{11} + y_1)(Y_{22} + y_2)} \right|^2 \quad (14)$$

When y_2 contains the generator and y_1 the load, we may proceed in the same way (letting I_0 flow in terminal 4) and obtain

$$\Gamma_{21} = 4g_1 g_2 \left| \frac{Y_{12}}{Y_{12} Y_{21} - (Y_{11} + y_1)(Y_{22} + y_2)} \right|^2 \quad (15)$$

We may now obtain expressions for the admittance and gain of the 4-pole when the nonlinear element consists of a nonlinear resistor and a nonlinear capacitor in parallel. We shall do this for the case where a conjugate match exists at the terminals by letting $y_1' = y_1^*$ and $y_2' = y_2^*$. Equations (11) and (12) may thus be written

$$(Y_{11} - y_1^*)(Y_{22} + y_2) = (Y_{11} + y_1)(Y_{22} - y_2^*) = Y_{12}Y_{21} \quad (16)$$

When this is multiplied out, letting $Y_{mn} = G_{mn} + jB_{mn}$, and the real and imaginary parts set equal as indicated by the first equality we obtain $G_{11}g_2 = G_{22}g_1$ and $g_2(B_{11} + b_1) = g_1(B_{22} + b_2)$. In (8) and (9) it is seen that $G_{11} = G_{22} = G_0$ and that B_{22} is positive in equations (8) and negative in equations (9). We thus obtain

$$g_1 = g_2 \quad b_1 + \omega_1 C_0 = b_2 \pm \omega_2 C_0 \quad (17)$$

where the upper symbol of the \pm sign is used in the noninverting case and the lower symbol in the inverting case. When the real and imaginary parts are set equal as indicated by the second equality in (16) we obtain, using the results in (17),

$$g^2 = G_0^2 - G_1^2 \pm \omega_1 \omega_2 C_1^2 - B^2 \quad (18)$$

where

$$g = g_1 = g_2 \quad (19)$$

and

$$B = b_1 + \omega_1 C_0 = b_2 \pm \omega_2 C_0 = \pm \frac{G_1}{2G_0} (\omega_2 \pm \omega_1) C_1 \quad (20)$$

These results may be put in (14) to obtain the modulator gain. Since a conjugate match exists at the terminals of the 4-pole, this is the maximum available gain. The result is

$$MAG_{12} = \frac{G_1^2 + (\omega_2 C_1)^2}{(G_0 + g)^2 + B^2} \quad (21)$$

For the converter, using equation (15) we obtain

$$MAG_{21} = \frac{G_1^2 + (\omega_1 C_1)^2}{(G_0 + g)^2 + B^2} \quad (22)$$

These results are valid only when a conjugate match exists at the terminals. For this to be possible, the right side of (18) must be positive. If it is negative no combination of values of g_1 and g_2 will result in a match.

It may be shown that if the slope of the voltage-current characteristic of the nonlinear resistor is always positive, then G_1/G_0 can never be greater than unity. (Reference 1, p. 410.) It is therefore convenient to normalize the above results with respect to G_0 . If we let

$$\frac{\omega_1}{\omega_2} = \rho, \quad \frac{\omega_1 C_1}{G_0} = \rho x, \quad \frac{\omega_2 C_1}{G_0} = x, \quad \frac{G_1}{G_0} = y, \quad \frac{C_0}{C_1} = z \quad (23)$$

equations (18) through (22) become

$$\left(\frac{g}{G_0}\right)^2 = 1 - y^2 \pm \rho x^2 - \left[(1 \pm \rho) \frac{xy}{2}\right]^2 \quad (24)$$

$$\frac{b_1}{G_0} = \pm (1 \pm \rho) \frac{xy}{2} - \rho xz, \quad \frac{b_2}{g_0} = \pm (1 \pm \rho) \frac{xy}{2} \pm xz \quad (25)$$

$$MAG_{12} = \frac{y^2 + x^2}{\left(1 + \frac{g}{G_0}\right)^2 + \left[(1 \pm \rho) \frac{xy}{2}\right]^2} \quad (26)$$

$$MAG_{21} = \frac{y^2 + (\rho x)^2}{\left(1 + \frac{g}{G_0}\right)^2 + \left[(1 \pm \rho) \frac{xy}{2}\right]^2} \quad (27)$$

In these equations, ρ is less than $\frac{1}{3}$ in the noninverting case and less than 1 in the inverting case. Ordinarily it will be very much less than 1. The value of z will be determined by the shape of the nonlinear capacitor characteristic. However z appears only in (25) where it influences the values of the matching susceptances so that it does not affect the conductance or gain. While we can be certain that y will have values between 0 and 1, limitations on the value of x will depend on the particular device used. We will therefore assume that x may have any value.

EFFECT OF NONLINEAR CAPACITOR

We may now examine, in a general way, the manner in which the nonlinear capacitor influences the behavior of the 4-pole. Consider first the case where the nonlinear capacitor is absent. It is well known, and can be seen in the above equations by letting $C_0 = C_1 = 0$, that the noninverting and inverting cases are alike, that the 4-pole can always be matched and that the gain is the same in both directions and can never be greater than unity. In addition, the matching susceptances are zero and the gain is independent of frequency so that there is no limitation to the bandwidth. When the nonlinear capacitor is added, all but one of these conditions are changed. Equations (8) and (9) show that the non-

inverting and inverting cases are different, (24) may become negative so that the 4-pole cannot always be matched and (26) and (27) are different so that the gains through the 4-pole are not the same in the two directions. Furthermore, (26) can be greater than unity so that modulators may have gain. However, as will be shown, the converter gain given by (27) is still restricted to values less than unity. It is also seen that the matching susceptances are no longer zero and that the gain varies with frequency so that the bandwidth is limited.

If we remove the restriction that a conjugate match exists and operate the 4-pole between arbitrary admittances, it may be shown in (11) and (12) that the conductance of the 4-pole may become negative, and in (14) and (15) that the gain may have any value, however large. This is true for both noninverting and inverting modulators and converters. However, we see in (14) and (15) that the ratio of the modulator gain to the converter gain is $|Y_{21}/Y_{12}|^2$. This is greater than unity, so that for the same operating conditions the modulator gain will be greater than the converter gain. Although increased gain is possible, it is obtained at the expense of reduced bandwidth and increased sensitivity to changes in the terminating admittances, particularly in the case of converters. The present analysis will therefore be restricted to the case where a conjugate match exists.

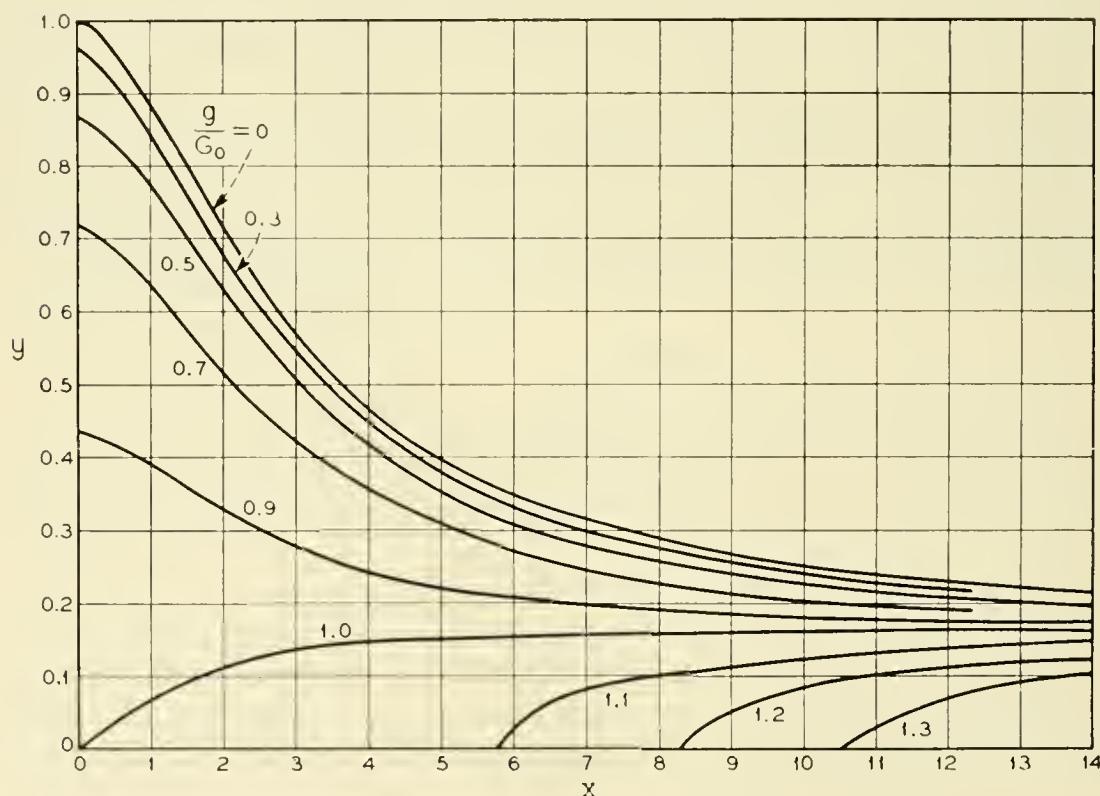


Fig. 2 — Conductance contours of noninverting transducer.

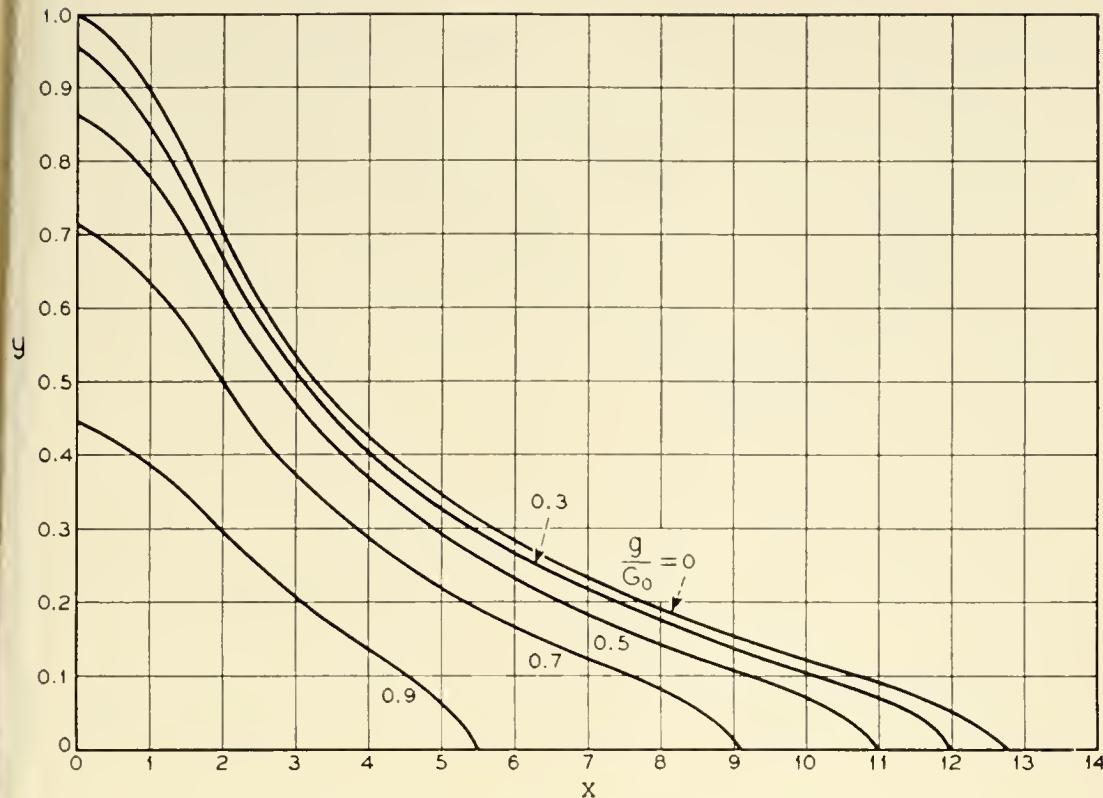


Fig. 3 — Conductance contours for inverting transducer.

CONDUCTANCE AND GAIN VERSUS x AND y

By assigning a value to ρ , curves may be plotted showing how the conductance and gain of the 4-pole change as the characteristics of the nonlinear resistor and nonlinear capacitor are varied. The particular case when f_2 is about 160 times f_1 will be treated. This corresponds, for example, to an intermediate frequency of 70 me and a local oscillator frequency of 11,200 me.

Figs. 2 and 3 show the normalized conductance contours as functions of x and y as given by (24) for the noninverting and inverting cases respectively. It will be seen that in most instances, increasing the value of x causes g/G_0 to decrease. An exception occurs in the noninverting case (Fig. 2) when y is less than $2\sqrt{\rho}/(\rho + 1)$ or 0.157 where it is seen that increasing x causes g/G_0 to increase. When x and y have values corresponding to points above the $g/G_0 = 0$ curve, the 4-pole cannot be matched and (23) through (27) are not applicable. However, it will be noted that connecting a resistor across either the nonlinear elements or across the input and output terminals has the effect of increasing G_0 . By this means the 4-pole can always be reduced to the condition where it can be matched.

Figs. 4 and 5 show the modulator gain contours as functions of x and y

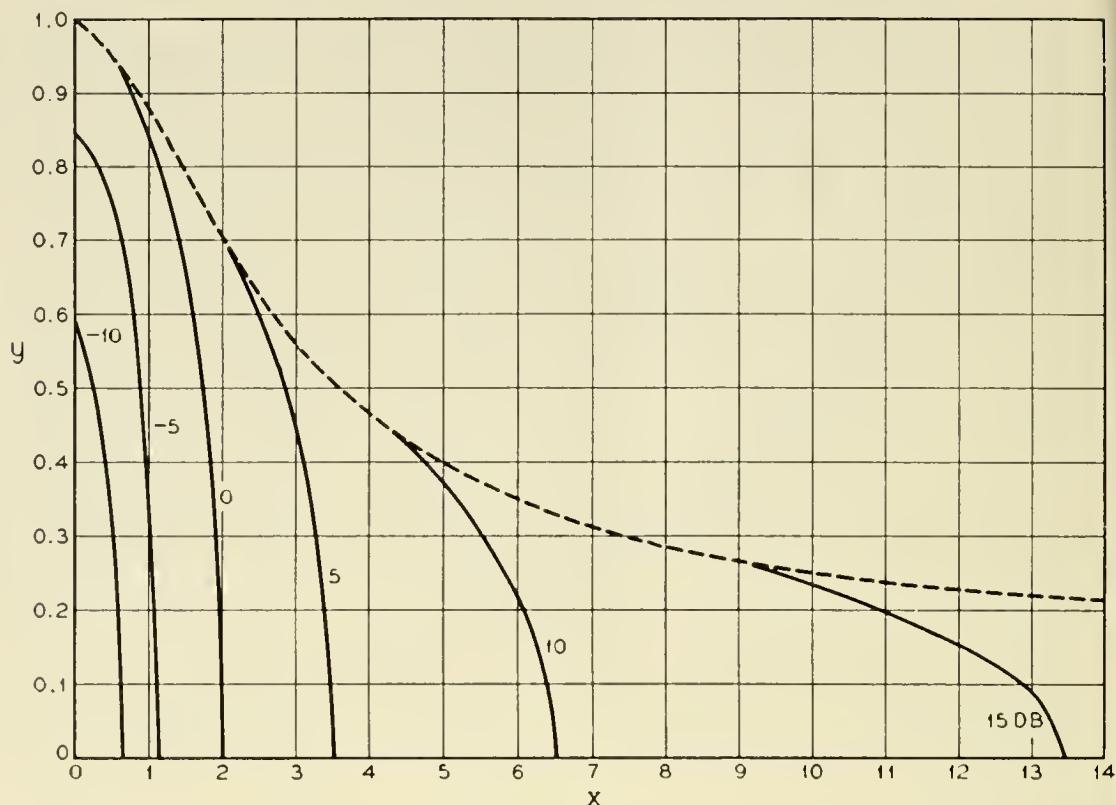


Fig. 4 — Gain contours for noninverting modulators.

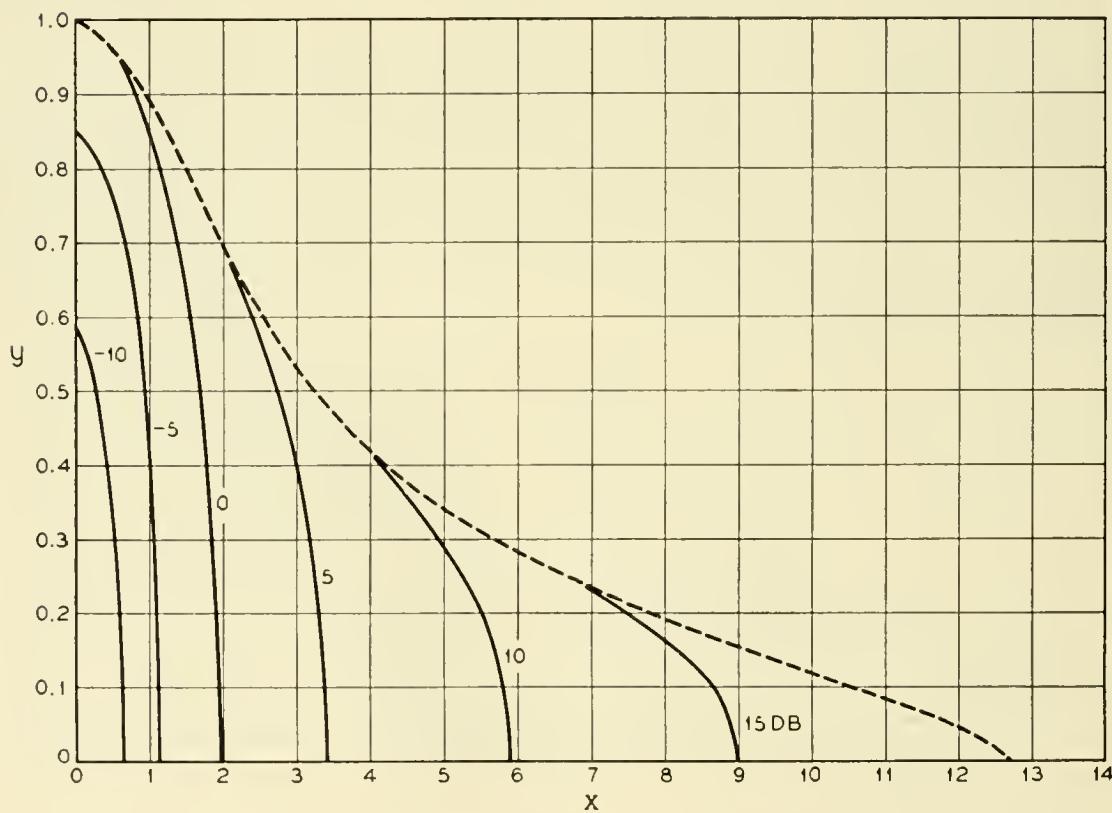


Fig. 5 — Gain contours for inverting modulator.

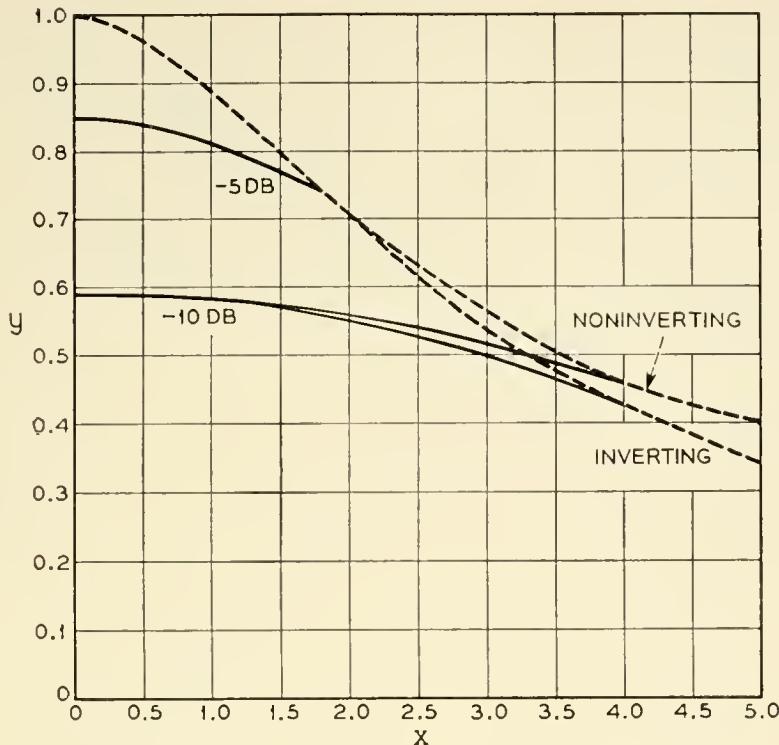


Fig. 6 — Gain contours for converter.

as given by (26). Here it is seen that increasing the value of x causes the gain to increase. For values of x less than about 3, the gains in the non-inverting and inverting cases are the same. In the noninverting case, x may increase indefinitely, provided y is less than 0.157, and a gain equal to the ratio of the output frequency to the input frequency eventually reached, 22.1 db in this case. In the inverting case, the maximum gain obtainable is 19.3 db, and it occurs when y is zero.

Fig. 6 shows the converter gain contours as given by equation (27). Here we see that increasing x causes a decrease in the loss, but the decrease is small and in no case can the gain be greater than 0 db. This occurs when x is zero. The nonlinear capacitor is thus of small benefit in the converter case. About the most benefit that can be obtained is a decrease in loss of perhaps 1 db. For example, if the nonlinear resistor alone has a loss of 6 db ($y = 0.8$), this could be reduced to 5 db by adding a nonlinear capacitor of such value as to make $x = 1.3$.

BANDWIDTH

Since both the admittance and gain of the 4-pole vary with frequency, the bandwidth over which it can be used is limited. Figs. 7 and 8 show the modulator gain as a function of x for input frequencies of 50, 70 and 90 mc, and a local oscillator frequency of 11,200 mc. These curves were

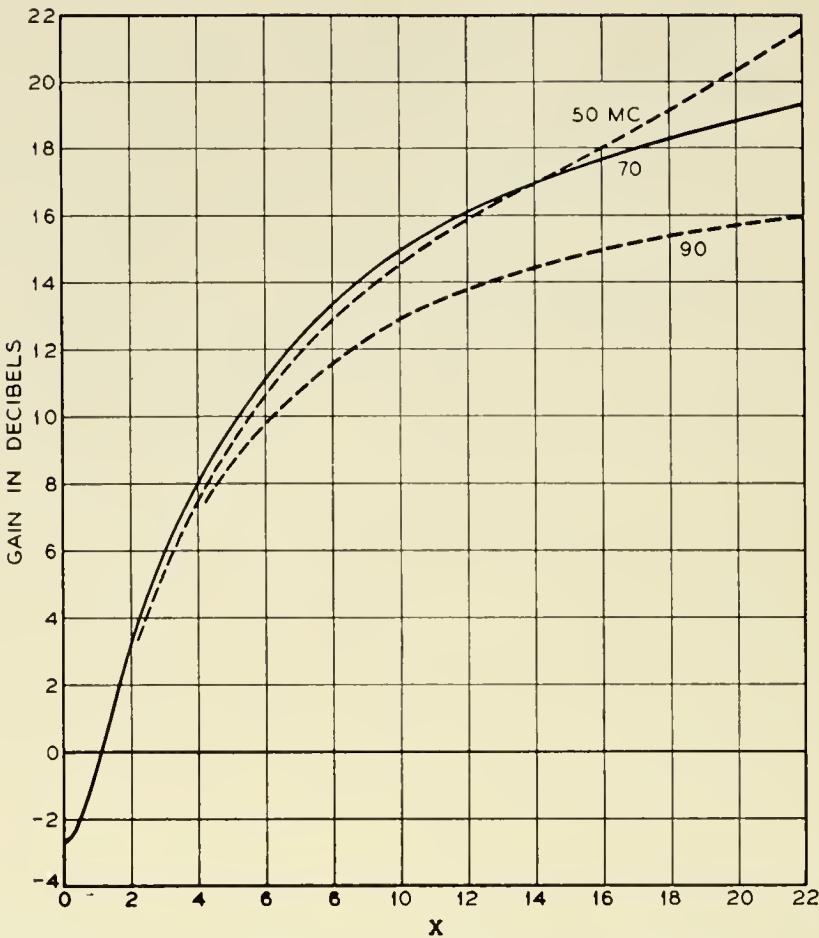


Fig. 7 — Gain of noninverting modulator, $g/G_0 = 0.3$.

computed using values of y which make $g/G_0 = 0.3$ at midband. They are thus near the largest gains obtainable for a given value of x . The matching susceptances were assumed to be a single inductance or capacitance connected across the terminating resistors. C_0/C_1 was arbitrarily assumed to have a value of 2. The procedure used was to compute y , b_1/G_0 , b_2/G_0 and the maximum available gain at midband using (24), (25) and (26); b_1/G_0 and b_2/G_0 were then multiplied by the appropriate frequency ratio to obtain the terminating susceptances at 50 and 90 mc and the gain at these frequencies was then computed using (14).

Figs. 7 and 8 show that with the simple matching susceptances used, the gain variation across the band increases as the gain increases. For the same midband gain, the variation in the inverting case is somewhat greater than in the noninverting case. The gain is thus limited by the bandwidth requirements.

When the gain at 50, 70 and 90 mc is calculated using larger values of g/G_0 it is found that as g/G_0 increases the gain variation across the band decreases. In the limit the least variation is obtained when y is

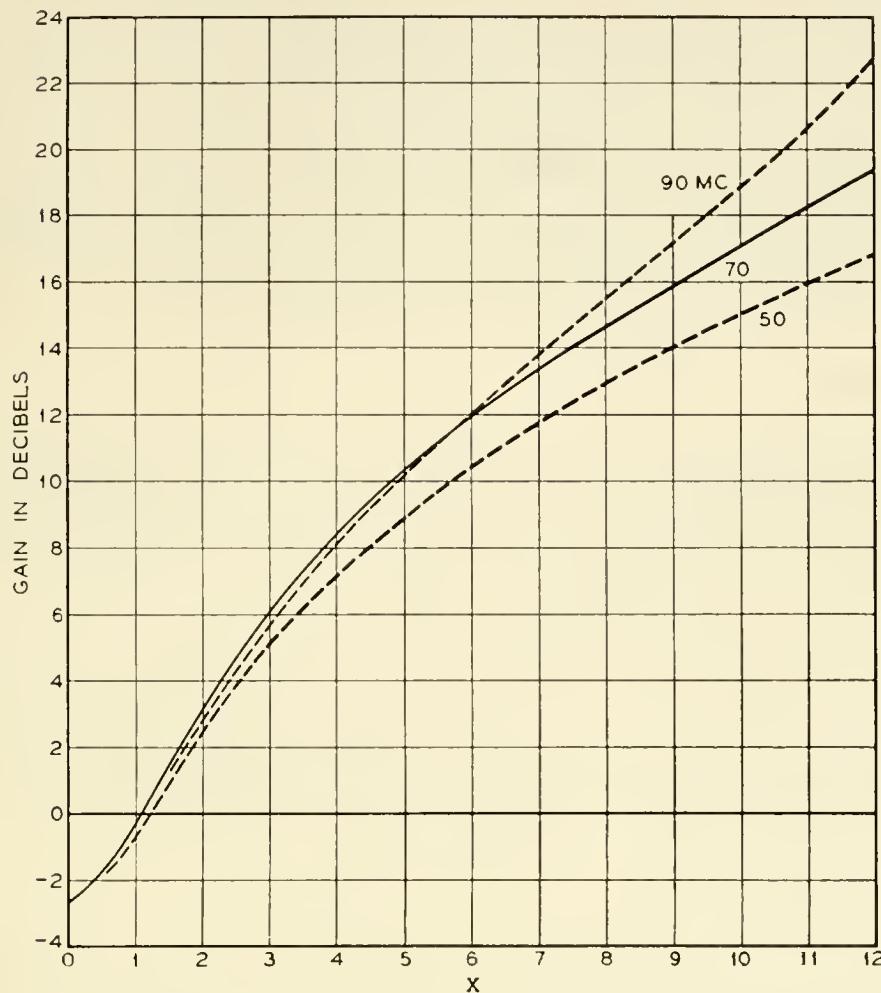


Fig. 8 — Gain of inverting modulator, $g/G_0 = 0.3$.

zero. When the midband gain is 15 db, Figs. 7 and 8 show that the gain variation is 2.0 db in the noninverting case and 2.7 db in the inverting case. When y is zero these variations are reduced to 0.8 db and 1.0 db respectively for the same midband gain. The nonlinear resistor therefore degrades the performance and, assuming complete freedom in the choice of x , a greater bandwidth can be obtained if it is absent.

PREFERRED NONLINEAR ELEMENTS

Thus we see that, under the requirement that a conjugate match exist at the terminals of the 4-pole, the nonlinear resistor contributes little to the gain of a nonlinear capacitor modulator while the nonlinear capacitor is of little benefit in a nonlinear resistor converter. In a modulator having appreciable gain, the degree of nonlinearity permissible in the nonlinear resistor is quite small. For gains exceeding 15 db, y must be less than 0.2. Such a nonlinear resistor used alone would have a con-

version loss exceeding 20 db. We thus find that, for the greatest bandwidth, the preferred nonlinear element for modulators is a nonlinear capacitor while the preferred nonlinear element for converters is a nonlinear resistor. In modulators, the nonlinear capacitor device should have as little resistance as possible, so that an external resistor could be used to control the value of x . It could be connected across the nonlinear capacitor or across the input and output terminals.

CONCLUSIONS

The results given above show that the preferred nonlinear element for use in modulators is a pure nonlinear capacitor while the preferred nonlinear element for use in converters is a pure nonlinear resistor. By shunting the nonlinear capacitor or the terminals of a nonlinear capacitor modulator with an appropriate resistance, an impedance match, adequate bandwidth, and a performance superior to that of a nonlinear resistor modulator can be obtained. Nonlinear capacitance effects are not useful in converters because of stability and bandwidth limitations and also because there is no evidence that an improved noise figure would result from a reduction in conversion loss.

ACKNOWLEDGMENT

The writer is indebted to H. E. Rowe for many helpful suggestions in the mathematical analysis and to R. S. Ohl for supplying the bombarded silicon rectifiers used in the experiments which lead to the ideas presented here.

REFERENCES

1. H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*, 15, Radiation Laboratory Series, McGraw-Hill, New York, 1948, Chapter 13.
2. L. C. Peterson and F. B. Llewellyn, The Performance and Measurement of Mixers in Terms of Linear Network Theory, Proc. I.R.E., 33, July, 1945.

Minimization of Boolean Functions*

E. J. McCLUSKEY, Jr.

(Manuscript received June 26, 1956)

A systematic procedure is presented for writing a Boolean function as a minimum sum of products. This procedure is a simplification and extension of the method presented by W. V. Quine. Specific attention is given to terms which can be included in the function solely for the designer's convenience.

1 INTRODUCTION

In designing switching circuits such as digital computers, telephone central offices, and digital machine tool controls, it is common practice to make use of Boolean algebra notation.^{1, 2, 3, 4} The performance of a single-output circuit is specified by means of a Boolean function of the input variables. This function, which is called the circuit transmission, is equal to 1 when an output is present and equals 0 when there is no output. A convenient means of specifying a transmission is a table of combinations such as that given in Table I. This table lists, in the column under T, the output condition for each combination of input conditions. If there are some combinations of input conditions for which the output is not specified (perhaps because these combinations can never occur), d-entries are placed in the T-column of the corresponding rows of the table of combinations. The actual values (0 or 1) assigned to these rows are usually chosen so as to simplify the circuit which is designed to satisfy the requirements specified in the table of combinations.

For each row of the table of combinations a transmission can be written which equals "one" only when the variables have the values listed in that row of the table. These transmissions will be called *elementary product terms* (or more simply, p-terms) since any transmission can always be written as a sum of these p-terms. Table I (b) lists the p-terms for Table I(a). Note that every variable appears in each p-term. The

* This paper is derived from a thesis submitted to the Massachusetts Institute of Technology in partial fulfillment of the requirements for the degree of Doctor of Science on April 30, 1956.

TABLE I — CIRCUIT SPECIFICATIONS

(a) Table of Combinations				
	x_1	x_2	x_3	T
0	0	0	0	0
1	0	0	1	1
2	0	1	0	1
3	0	1	1	1
4	1	0	0	1
5	1	0	1	1
6	1	1	0	1
7	1	1	1	0

(b) p-terms

$x_1' x_2' x_3'$
 $x_1' x_2' x_3$,
 $x_1' x_2 x_3'$
 $x_1' x_2 x_3$,
 $x_1 x_2' x_3'$
 $x_1 x_2' x_3$,
 $x_1 x_2 x_3'$
 $x_1 x_2 x_3$

(c) Canonical Expansion

$$T = x_1'x_2'x_3 + x_1'x_2x_3' + x_1'x_2x_3 + x_1x_2'x_3' + x_1x_2'x_3 + x_1x_2x_3'$$

p-term corresponding to a given row of a table of combinations is formed by priming any variables which have a "zero" entry in that row of the table and by leaving unprimed those variables which have "one" entries. It is possible to write an algebraic expression for the over-all circuit transmission directly from the table of combinations. This over-all transmission, T , is the sum of the p-terms corresponding to those rows of the table of combinations for which T is to have the value "one." See Table I(c). Any transmission which is a sum of p-terms is called a *canonical expansion*.

The decimal numbers in the first column of Table I(a) are the decimal equivalents of the binary numbers formed by the entries of the table of combinations. A concise method for specifying a transmission function is to list the decimal numbers of those rows of the table of combinations for which the function is to have the value one. Thus the function of Table I can be specified as $\sum(1, 2, 3, 4, 5, 6)$.

One of the most basic problems of switching circuit theory is that of writing a Boolean function in a simpler form than the canonical expansion. It is frequently possible to realize savings in equipment by writing a circuit transmission in simplified form. Methods for expressing a Boolean function in the "simplest" sum of products form were published by Karnaugh,¹ Aiken,⁵ and Quine.⁶ These methods have the common property that they all fail when the function to be simplified is reasonably complex. The following sections present a method for simplifying functions which can be applied to more complex functions than previous methods, is systematic, and can be easily programmed on a digital computer.

2 THE MINIMUM SUM

By use of the Boolean algebra theorem $x_1x_2 + x_1'x_2 = x_2$ it is possible to obtain from the canonical expansion other equivalent sum functions:

that is, other sum functions which correspond to the same table of combinations. These functions are still sums of products of variables but not all of the variables appear in each term. For example, the transmission of Table I, $T = x_1'x_2'x_3 + x_1'x_2x_3' + x_1'x_2x_3 + x_1x_2'x_3' + x_1x_2'x_3 + x_1x_2x_3' = (x_1'x_2'x_3 + x_1'x_2x_3) + (x_1'x_2x_3' + x_1x_2x_3') + (x_1x_2'x_3' + x_1x_2'x_3) = (x_1'x_2'x_3 + x_1x_2'x_3) + (x_1'x_2x_3' + x_1'x_2x_3) + (x_1x_2'x_3' + x_1x_2x_3')$ can be written as either $T = x_1'x_3 + x_2x_3' + x_1x_2'$ or $T = x_2'x_3 + x_1'x_2 + x_1x_3'$.

A *literal* is defined as a variable with or without the associated prime (x_1, x_2' are literals). The sum functions which have the fewest terms of all equivalent sum functions will be called *minimum sums* unless these functions having fewest terms do not all involve the same number of literals. In such cases, only those functions which involve the fewest literals will be called minimum sums. For example, the function

$$T = \sum(7, 9, 10, 12, 13, 14, 15)$$

can be written as either

$$T = x_4x_2x_1' + x_3x_2x_1 + x_4x_2'x_1 + x_4x_3x_1'$$

or as

$$T = x_4x_2x_1' + x_3x_2x_1 + x_4x_2'x_1 + x_4x_3$$

Only the second expression is a minimum sum since it involves 11 literals while the first expression involves 12 literals.

The minimum sum defined here is not necessarily the expression containing the fewest total literals, or the expression leading to the most economical two-stage diode logic circuit,¹ even though these three expressions are identical for many transmissions. The definition adopted here lends itself well to computation and results in a form which is useful in the design of contact networks. A method is presented in Section 9 for obtaining directly the expressions corresponding to the optimum two-stage diode logic circuit or the expressions containing fewest literals.

In principle it is possible to obtain a minimum sum for any given transmission by enumerating all possible equivalent sum functions then selecting those functions which have the fewest terms, and finally selecting from these the functions which contain fewest literals. Since the number of equivalent sum functions may be quite large, this procedure is not generally practical. The following sections present a practical method for obtaining a minimum sum without resorting to an enumeration of all equivalent sum functions.

3 PRIME IMPLICANTS

When the theorem $x_1x_2 + x_1x_2' = x_1$ is used to replace by a single term, two p -terms, which correspond to rows i and j of a table of combi-

nations, the resulting term will equal "one" when the variables have values corresponding to either row i or row j of the table. Similarly, when this theorem is used to replace, by a single term, a term which equals "one" for rows i and j and a term which equals "one" for rows k and m , the resulting term will equal "one" for rows i, j, k and m of the table of combinations. A method for obtaining a minimum sum by repeated application of this theorem ($x_1x_2' + x_1x_2 = x_1$) was first presented by Quine.⁶ In this method, the theorem is applied to all possible pairs of p -terms, then to all possible pairs of the terms obtained from the p -terms, and so on, until no further applications of the theorem are possible. It may be necessary to pair one term with several other terms in applying this theorem. In Example 3.2 the theorem is applied to the terms labeled 5 and 7 and also to the terms labeled 5 and 13. All terms paired with other terms in applying the theorem are then discarded. The remaining terms are called *prime implicants*.⁶ Finally a minimum sum is formed as the sum of the fewest prime implicants which when taken together will equal "one" for all required rows of the table of combinations. The terms in the minimum sum will be called *minimum sum terms* or *ms-terms*.

Example 3.1

$$T = \sum(3, 7, 8, 9, 12, 13)$$

Canonical Expansion:

$$\begin{aligned} T = & x_1'x_2'x_3x_4 + x_1'x_2x_3x_4 + x_1x_2'x_3'x_4' + x_1x_2'x_3'x_4 \\ & \left[\begin{smallmatrix} 0 & 0 & 1 & 1 \\ 3 \end{smallmatrix} \right] \left[\begin{smallmatrix} 0 & 1 & 1 & 1 \\ 7 \end{smallmatrix} \right] \left[\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 8 \end{smallmatrix} \right] \left[\begin{smallmatrix} 1 & 0 & 0 & 1 \\ 9 \end{smallmatrix} \right] \\ & + x_1x_2x_3'x_4' + x_1x_2x_3'x_4 \\ & \left[\begin{smallmatrix} 1 & 1 & 0 & 0 \\ 12 \end{smallmatrix} \right] \left[\begin{smallmatrix} 1 & 1 & 0 & 1 \\ 13 \end{smallmatrix} \right] \end{aligned}$$

The bracketed binary and decimal numbers below the sum terms indicate the rows of the table of combinations for which the corresponding term will equal "one." A binary character in which a dash appears represents the two binary numbers which are formed by replacing the dash by a "0" and then by a "1." Similarly a binary character in which two dashes appear represents the four binary numbers formed by replacing the dashes by "0" and "1" entries, etc.

$$\begin{aligned} x_1'x_2'x_3x_4 + x_1'x_2x_3x_4 &= x_1' x_3x_4 \\ \left[\begin{smallmatrix} 0 & 0 & 1 & 1 \\ 3 \end{smallmatrix} \right] \left[\begin{smallmatrix} 0 & 1 & 1 & 1 \\ 7 \end{smallmatrix} \right] \left[\begin{smallmatrix} 0 - 1 & 1 \\ 3, 7 \end{smallmatrix} \right] \end{aligned}$$

$$\begin{aligned}
 & x_1x_2'x_3'x_4' + x_1x_2'x_3'x_4 = x_1x_2'x_3' \\
 \left[\begin{matrix} 1 & 0 & 0 & 0 \\ 8 \end{matrix} \right] & \left[\begin{matrix} 1 & 0 & 0 & 1 \\ 9 \end{matrix} \right] \left[\begin{matrix} 1 & 0 & 0 & - \\ 8,9 \end{matrix} \right] \\
 & x_1x_2x_3'x_4' + x_1x_2x_3'x_4 = x_1x_2x_3' \\
 \left[\begin{matrix} 1 & 1 & 0 & 0 \\ 12 \end{matrix} \right] & \left[\begin{matrix} 1 & 1 & 0 & 1 \\ 13 \end{matrix} \right] \left[\begin{matrix} 1 & 1 & 0 & - \\ 12,13 \end{matrix} \right] \\
 & x_1x_2'x_3' + x_1x_2x_3' = x_1 x_3' \\
 \left[\begin{matrix} 1 & 0 & 0 & - \\ 8,9 \end{matrix} \right] & \left[\begin{matrix} 1 & 1 & 0 & - \\ 12,13 \end{matrix} \right] \left[\begin{matrix} 1 & - & 0 & - \\ 8,9,12,13 \end{matrix} \right]
 \end{aligned}$$

Prime Implicants:

$$\begin{matrix} x_1 & x_3', & x_1' & x_3x_4 \\ \left[\begin{matrix} 1 & - & 0 & - \\ 8,9,12,13 \end{matrix} \right] & \left[\begin{matrix} 0 & - & 1 & 1 \\ 3,7 \end{matrix} \right] \end{matrix}$$

Minimum Sum:

$$T = x_1x_3' + x_1'x_3x_4$$

Example 3.2

$$T = \sum(5, 7, 12, 13)$$

Canonical Expansion:

$$\begin{aligned}
 T &= x_1'x_2x_3'x_4 + x_1'x_2x_3x_4 + x_1x_2x_3'x_4' + x_1x_2x_3'x_4 \\
 \left[\begin{matrix} 0 & 1 & 0 & 1 \\ 5 \end{matrix} \right] & \left[\begin{matrix} 0 & 1 & 1 & 1 \\ 7 \end{matrix} \right] \left[\begin{matrix} 1 & 1 & 0 & 0 \\ 12 \end{matrix} \right] \left[\begin{matrix} 1 & 1 & 0 & 1 \\ 13 \end{matrix} \right] \\
 x_1'x_2x_3'x_4 + x_1'x_2x_3x_4 &= x_1'x_2 x_4 \\
 \left[\begin{matrix} 0 & 1 & 0 & 1 \\ 5 \end{matrix} \right] & \left[\begin{matrix} 0 & 1 & 1 & 1 \\ 7 \end{matrix} \right] \left[\begin{matrix} 0 & 1 & - & 1 \\ 5,7 \end{matrix} \right] \\
 x_1'x_2x_3'x_4 + x_1x_2x_3'x_4 &= x_2x_3'x_4 \\
 \left[\begin{matrix} 0 & 1 & 0 & 1 \\ 5 \end{matrix} \right] & \left[\begin{matrix} 1 & 1 & 0 & 1 \\ 13 \end{matrix} \right] \left[\begin{matrix} -1 & 0 & 1 \\ 5,13 \end{matrix} \right] \\
 x_1x_2x_3'x_4' + x_1x_2x_3'x_4 &= x_1x_2x_3' \\
 \left[\begin{matrix} 1 & 1 & 0 & 0 \\ 12 \end{matrix} \right] & \left[\begin{matrix} 1 & 1 & 0 & 1 \\ 13 \end{matrix} \right] \left[\begin{matrix} 1 & 1 & 0 & - \\ 12,13 \end{matrix} \right]
 \end{aligned}$$

Prime Implicants:

$$\begin{matrix} x_1'x_2 x_4, & x_2x_3'x_4, & x_1x_2x_3' \\ \left[\begin{matrix} 0 & 1 & - & 1 \\ 5,7 \end{matrix} \right] & \left[\begin{matrix} -1 & 0 & 1 \\ 5,13 \end{matrix} \right] & \left[\begin{matrix} 1 & 1 & 0 & - \\ 12,13 \end{matrix} \right] \end{matrix}$$

Minimum Sum:

$$T = x_1'x_2x_4 + x_1x_2x_3'$$

Quine's method, as illustrated in Examples 3.1 and 3.2, becomes unwieldly for transmissions involving either many variables or many p-terms. This difficulty is overcome by simplifying the notation and making the procedure more systematic. The notation is simplified by discarding the expressions involving literals and using only the binary characters. This is permissible because the expressions in terms of literals can always be regained from the binary characters. The theorem being used to combine terms can be stated in terms of the binary characters as follows: If two binary characters are identical in all positions except one, and if neither character has a dash in the position in which they differ, then the two characters can be replaced by a single character which has a dash in the position in which the original characters differ and which is identical with the original characters in all other positions.

TABLE II — DETERMINATION OF PRIME IMPLICANTS FOR TRANSMISSION

$$T = \sum (0, 2, 4, 6, 7, 8, 10, 11, 12, 13, 14, 16, 18, 19, 29, 30)$$

(a) I	(b) II	(c) III
$x_5x_4x_3x_2x_1$	$x_5x_4x_3x_2x_1$	$x_5x_4x_3x_2x_1$
0 0 0 0 0 ✓	0 2 0 0 0 - 0 ✓	0 2 4 6 0 0 - - 0 ✓
2 0 0 0 1 0 ✓	0 4 0 0 - 0 0 ✓	0 2 8 10 0 - 0 - 0 ✓
4 0 0 1 0 0 ✓	0 8 0 - 0 0 0 ✓	0 2 16 18 - 0 0 - 0
8 0 1 0 0 0 ✓	0 16 - 0 0 0 0 ✓	0 4 8 12 0 - - 0 0 ✓
16 1 0 0 0 0 ✓	2 6 0 0 - 1 0 ✓	2 6 10 14 0 - - 1 0 ✓
6 0 0 1 1 0 ✓	2 10 0 - 0 1 0 ✓	4 6 12 14 0 - 1 - 0 ✓
10 0 1 0 1 0 ✓	2 18 - 0 0 1 0 ✓	8 10 12 14 0 1 - - 0 ✓
12 0 1 1 0 0 ✓	4 6 0 0 1 - 0 ✓	
18 1 0 0 1 0 ✓	4 12 0 - 1 0 0 ✓	
7 0 0 1 1 1 ✓	8 10 0 1 0 - 0 ✓	
11 0 1 0 1 1 ✓	8 12 0 1 - 0 0 ✓	
13 0 1 1 0 1 ✓	16 18 1 0 0 - 0 ✓	
14 0 1 1 1 0 ✓	6 7 0 0 1 1 -	
19 1 0 0 1 1 ✓	6 14 0 - 1 1 0 ✓	
29 1 1 1 0 1 ✓	10 11 0 1 0 1 -	
30 1 1 1 1 0 ✓	10 14 0 1 - 1 0 ✓	
	12 13 0 1 1 0 -	
	12 14 0 1 1 - 0 ✓	
	18 19 1 0 0 1 -	
	13 29 - 1 1 0 1	
	14 30 - 1 1 1 0	
		(d) IV
		$x_5x_4x_3x_2x_1$
0 2 4 6 8 10 12 14		0 - - - 0

The first step in the revised method for determining prime implicants is to list in a column, such as that shown in Table II(a), the binary equivalents of the decimal numbers which specify the function. It is expedient to order these binary numbers so that any numbers which contain no 1's come first, followed by any numbers containing a single 1, etc. Lines should be drawn to divide the column into groups of binary numbers which contain a given number of 1's. The theorem stated above is applied to these binary numbers by comparing each number with all the numbers of the next lower group. Other pairs of numbers need not be considered since any two numbers which are not from adjacent groups must differ in more than one binary digit. For each number which has 1's wherever the number (from the next upper group) with which it is being compared has 1's, a new character is formed according to the theorem. A check mark is placed next to each number which is used in forming a new character. The new characters are placed in a separate column, such as Table II(b), which is again divided into groups of characters which have the same number of 1's. The characters in this new column will each contain one dash.

After each number in the first column has been considered, a similar process is carried out for the characters of column two. Two characters from adjacent groups can be combined if they both have their dashes in the same position and if the character from the lower group has 1's wherever the upper character has 1's. If any combinations are possible the resulting characters are placed in a third column such as Table II(c), and the Column II characters from which the new characters are formed are checked. All the characters in this third column will have two dashes. This procedure is repeated and new columns are formed, Table II(d), until no further combinations are possible. The unchecked characters, which have not entered into any combinations, represent the prime implicants.

Each binary character is labeled with the decimal equivalents of the binary numbers which it represents (see note in Example 3.1). These decimal numbers are arranged in increasing arithmetic order. For a character having one dash this corresponds to the order of its formation: When two binary numbers combine, the second number always contains all the 1's of the first number and one additional 1 so that the second number is always greater than the first. Characters having two dashes can be formed in two ways. For example, the character (0, 2, 4, 6) can be formed either by combining (0, 2) and (4, 6) or by combining (0, 4) and (2, 6) as given in Table III. Similarly, there are three ways in which a character having three dashes can be formed (in Table II the 0, 2, 4,

TABLE III—EXAMPLE OF THE TWO WAYS OF FORMING
A CHARACTER HAVING TWO DASHES

0	0 0 0 0	0 2	0 0 - 0	0 2 4 6	0 - - 0
2	0 0 1 0	0 4	0 - 0 0	(0 4 2 6	0 - - 0)
4	0 1 0 0	2 6	0 - 1 0		
6	0 1 1 0	4 6	0 1 - 0		

6, 8, 10, 12, 14 character can be formed from the 0, 2, 4, 6, and 8, 10, 12, 14 characters or the 0, 2, 8, 10, and 4, 6, 12, 14 characters or the 0, 4, 8, 12 and 2, 6, 10, 14 characters), four ways in which a character having four dashes can be formed, etc.

In general, any character can be formed by combining two characters whose labels form an increasing sequence of decimal numbers when placed together. It is possible to shorten the process of determining prime implicants by not considering the combination of any characters whose labels do not satisfy this requirement. For example, in Table II(b) the possibility of combining the (0, 4) character with either the (2, 6), (2, 10) or the (2, 18) character need not be considered. If the process is so shortened, it is not sufficient to place check marks next to the two characters from which a new character is formed; each member of all pairs of characters which would produce the same new character when combined must also receive check marks. More simply, when a new character is formed a check mark is placed next to all characters whose labels contain only decimal numbers which occur in the label of the new character. In Table II, when the (0, 2, 4, 6) character is formed by combining the (0, 2) and (4, 6) characters, check marks must be placed next to the (0, 4) and (2, 6) characters as well as the (0, 2) and (4, 6) characters. If the process is not shortened as just described, the fact that a character can be formed in several ways can serve as a check on the accuracy of the process.

It is possible to carry out the entire process of determining the prime implicants solely in terms of the decimal labels without actually writing the binary characters. If two binary characters can be combined as described in this section, then the decimal label of one can be obtained from the decimal label of the other character by adding some power of two (corresponding to the position in which the two characters differ) to each number in the character's label. For example, in Table IIb the label of the (4, 6)(0 0 1 - 0) character can be obtained by adding $4 = (2^2)$ to the numbers of the label of the (0, 2)(0 0 0 - 0) character. By searching for decimal labels which differ by a power of two, instead of binary characters which differ in only one position, the prime implicants can be

determined as described above without ever actually writing the binary characters.

4 PRIME IMPLICANT TABLES

The minimum sum is formed by picking the fewest prime implicants whose sum will equal one for all rows of the table of combinations for which the transmission is to equal one. In terms of the characters used in Section 3 this means that each number in the decimal specification of the function must appear in the label of at least one character which corresponds to a ms-term (term of the minimum sum).

The ms-terms are selected from the prime implicants by means of a prime implicant table,* Table IV. Each column of the prime implicant table corresponds to a row of the table of combinations for which the transmission is to have the value one. The decimal number at the top of each column specifies the corresponding row of the table of combinations. Thus the numbers which appear at the tops of the columns are the same as those which specify the transmission. Each row of the prime implicant table represents a prime implicant. If a prime implicant equals "one" for a given row of the table of combinations, a cross is placed at the intersection of the corresponding row and column of the prime implicant table. All other positions are left blank. The table can be written directly from the characters obtained in Section 3 by identifying each row of the table with a character and then placing a cross in each column whose number appears in the label of the character.

It is convenient to arrange the rows in the order of the number of crosses they contain, with those rows containing the most crosses at the top of the table. Also, horizontal lines should be drawn partitioning the table into groups of rows which contain the same number of crosses, Table IV. If, in selecting the rows which are to correspond to ms-terms, a choice between two equally appropriate rows is required, the row having more crosses should be selected. The row with more crosses has fewer literals in the corresponding prime implicant. This choice is more obvious when the table is partitioned as suggested above.

A minimum sum is determined from the prime implicant table by selecting the fewest rows such that each column has a cross in at least one selected row. The selected rows are called *basis rows*, and the prime implicants corresponding to the basis rows are the ms-terms. If any column has only one entry, the row in which this entry occurs must be a basis row. Therefore the first step in selecting the basis rows is to place

* This table was first discussed by Quine.⁶ However, no systematic procedure for obtaining a minimum sum from the prime implicant table was presented.

TABLE IV—PRIME IMPLICANT TABLE FOR THE
TRANSMISSION OF TABLE II

	0	2	4	8	16	6	10	12	18	7	11	13	14	19	29	30
A	x	x	x	x		x	x	x					x			*
B	x	x			x				x							*
C											x		x			*
D										x			x			*
E							x				x		x			*
F						x				x		x				*
G							x			x						*
H						x			x							*

an asterisk next to each row which contains the sole entry of any column (rows A, B, C, D, E, G, H, in Table IV). A line is then drawn through all rows marked with an asterisk and through *all* columns in which these rows have entries. This is done because the requirement that these columns have entries in at least one basis row is satisfied by selecting the rows marked with an asterisk as basis rows. When this is done for Table IV, all columns are lined out and therefore the rows marked with asterisks are the basis rows for this table. Since no alternative choice of basis rows is possible, there is only one minimum sum for the transmission described in this table.

5 ROW COVERING

In general, after the appropriate rows have been marked with asterisks and the corresponding columns have been lined out, there may remain some columns which are not lined out; for example, column 7 in Table V(b). When this happens, additional rows must be selected and the columns in which these rows have entries must be lined out until all columns of the table are lined out. For Table V(b), the selection of either row B or row F as a basis row will cause column 7 to be lined out. However, row B is the correct choice since it has more crosses than row F. This is an example of the situation which was described earlier in connection with the partitioning of prime implicant tables. Row B is marked with two asterisks to indicate that it is a basis row even though it does not contain the sole entry of any column.

The choice of basis rows to supplement the single asterisk rows becomes more complicated when several columns (such as columns 2, 3, and 6 in Table VI(a)) remain to be lined out. The first step in choosing these supplementary basis rows is to determine whether any pairs of rows exist such that one row has crosses only in columns in which the

TABLE V — DETERMINATION OF THE MINIMUM SUM FOR

$$T = \sum (0, 1, 2, 3, 7, 14, 15, 22, 23, 29, 31)$$

(a) Determination of Prime Implicants

	$x_5x_4x_3x_2x_1$		$x_5x_4x_3x_2x_1$		$x_5x_4x_3x_2x_1$
0	0 0 0 0 0 ✓	0 1	0 0 0 0 - ✓	0 1 2 3	0 0 0 - -
		0 2	0 0 0 - 0 ✓	7 15 23 31	- - 1 1 1
1	0 0 0 0 1 ✓				
2	0 0 0 1 0 ✓	1 3	0 0 0 - 1 ✓		
		2 3	0 0 0 1 - ✓		
3	0 0 0 1 1 ✓		3 7	0 0 - 1 1	
7	0 0 1 1 1 ✓				
14	0 1 1 1 0 ✓	7 15	0 - 1 1 1 ✓		
22	1 0 1 1 0 ✓	7 23	- 0 1 1 1 ✓		
			14 15	0 1 1 1 -	
15	0 1 1 1 1 ✓	22 23	1 0 1 1 -		
23	1 0 1 1 1 ✓				
29	1 1 1 0 1 ✓	15 31	- 1 1 1 1 ✓		
		23 31	1 - 1 1 1 ✓		
31	1 1 1 1 1 ✓	29 31	1 1 1 - 1		

(b) First Step in Selection of Basis Rows

	0	1	2	3	7	14	22	15	23	29	31	
A	x	x	x	x								*
B					x			x	x		x	***
C												*
D							x		x			*
E								x				*
F				x	x			x				*

(c) Minimum Sum

$$T = \sum [(0, 1, 2, 3), (7, 15, 23, 31), (29, 31), (22, 23), (14, 15)]$$

$$T = x_5'x_4'x_3' + x_3x_2x_1 + x_5x_4x_3x_1 + x_5x_4'x_3x_2 + x_6'x_4x_3x_2$$

other member of the pair has crosses. Crosses in lined-out columns are not considered. In Table VI(a), rows A and B and rows B and C are such pairs of rows since row B has crosses in columns 2, 3, and 6 and row A has a cross in column 6 and row C has crosses in columns 2 and 3. A convenient way to describe this situation is to say that row B *covers* rows A and C, and to write $B \supseteq A$, $B \supseteq C$. If row i is selected as a supplementary basis row and row i is covered by row j, which has the same total number of crosses as row i, then it is possible to choose row j as a basis row instead of row i since row j has a cross in each column in which row i has a cross.

The next step is to line out any *rows* which are covered by other rows in the same partition of the table, rows A and C in Table VI(a). If any

TABLE VI — PRIME IMPLICANT TABLES FOR
 $T = \sum (0, 1, 2, 3, 6, 7, 14, 22, 30, 33, 62, 64, 71, 78, 86)$

(a) Prime Implicant Table with Single Asterisk Rows and Corresponding Columns Lined Out

	0	1	2	64	3	6	33	7	14	22	30	71	78	86	62
A					x			x	x	x	x				
B			x		x	x		x							
C	x	x	x		x										
D										x				x	
E									x				x		
F								x				x			
G								x			x				
H		x					x								
I	x		x												

(b) Prime Implicant Table with Rows which are Covered by Other Rows Lined Out

	0	1	2	64	3	6	33	7	14	22	30	71	78	86	62
A					x			x	x	x	x				
B			x		x	x		x							
C	x	x	x		x										
D									x			x		x	
E									x			x		x	
F								x			x		x		
G		x					x				x				
H	x		x				x				x				
I		x						x							

column now contains only one cross which is not lined out, columns 2, 3, and 6 in Table VI(b), two asterisks are placed next to the row in which this cross occurs, row B in Table VI(b), and this row and all columns in which this row has crosses are lined out. The process of drawing a line through any row which is covered by another row and selecting each row which contains the only cross in a column is continued until it terminates. Either all columns will be lined out, in which case the rows marked with one or two asterisks are the basis rows, or each column will contain more than one cross and no row will cover another row. The latter situation is discussed in the following section.

6 PRIME IMPLICANT TABLES IN CYCLIC FORM

If the rows and columns of a table which are not lined out are such that every column has more than one cross and no row covers another row, as in Table VII(b), the table will be said to be in *cyclic form*, or, in short,

TABLE VII — DETERMINATION OF BASIS ROWS FOR A
CYCLIC PRIME IMPLICANT TABLE

(a) Selection of Single Asterisk Rows

0 4 16 12 24 19 28 27 29 31

A	x x								
B	x x								
C	x x								
D	x x								
E	x x								
F	x x								
G	x x								
H	x x								
I	x x								
J	x x								

(b) Selection of Double Asterisk Rows

0 4 16 12 24 19 28 27 29 31

A	x x								
B	x x								
C	x x								
D	x x								
E	x x								
F	x x								
G	x x								
H	x x								
I	x x								
J	x x								

(c) Selection of Row 1 as a Trial Basis Row (Column 0)

0 4 16 12 24 19 28 27 29 31

A	x x								
B	x x								
C	x x								
D	x x								
E	x x								
F	x x								
G	x x								
H	x x								
I	x x								
J	x x								

(d) Selection of Row 2 as a Trial Basis Row (Column 0)

0 4 16 12 24 19 28 27 29 31

A	x x								
B	x x								
C	x x								
D	x x								
E	x x								
F	x x								
G	x x								
H	x x								
I	x x								
J	x x								

to be *cyclic*. If any column has crosses in only two rows, at least one of these rows must be included in any set of basis rows. Therefore, the basis rows for a cyclic table can be discovered by first determining whether any column contains only two crosses, and if such a column exists, by then selecting as a trial basis row one of the rows in which the crosses of this column occur. If no column contains only two crosses, then a column which contains three crosses is selected, etc. All columns in which the trial basis row has crosses are lined out and the process of lining out rows which are covered by other rows and selecting each row which contains the only cross of some column is carried out as described above. Either all columns will be lined out or another cyclic table will result. Whenever a cyclic table occurs, another trial row must be selected. Eventually all columns will be lined out. However, there is no guarantee that the selected rows are actually basis rows. The possibility exists that a different choice of trial rows would have resulted in fewer selected rows. In general, it is necessary to carry out the procedure of selecting rows several times, choosing different trial rows each time, so

that all possible combinations of trial rows are considered. The set of fewest selected rows is the actual set of basis rows.

Table VII illustrates the process of determining basis rows for a cyclic prime implicant table. After rows G and J have been selected a cyclic table results, Table VII(b). Rows A and B are then chosen as a pair of trial basis rows since column 0 has crosses in only these two rows. The selection of row A leads to the selection of rows D and E as given in Table VII(c). Row A is marked with three asterisks to indicate that it is a trial basis row. Table VII(d) illustrates the fact that the selection of rows C and F is brought about by the selection of row B. Since both sets of selected rows have the same number of rows (5) they are both sets of basis rows. Each set of basis rows corresponds to a different minimum sum so that there are two minimum sums for this function.

Sometimes it is not necessary to determine all minimum sums for the transmission being considered. In such cases, it may be possible to shorten the process of determining basis rows. Since each column must have a cross in some basis row, the total number of crosses in all of the basis rows is equal to or greater than the number of columns. Therefore, the number of columns divided by the greatest number of crosses in any row (or the next highest integer if this ratio is not an integer) is equal to the fewest possible basis rows. For example, in Table VII there are ten columns and two crosses in each row. Therefore, there must be at least 10 divided by 2 or 5 rows in any set of basis rows. The fact that there are only five rows selected in Table VII(c) guarantees that the selected rows are basis rows and therefore Table VII(d) is unnecessary if only one minimum sum is required. In general, the process of trying different combinations of trial rows can be stopped as soon as a set of selected rows which contains the fewest possible number of basis rows has been found (providing that it is not necessary to discover *all* minimum sums). It should be pointed out that more than the minimum number of basis rows may be required in some cases and in these cases all combinations of trial rows must be considered. A more accurate lower bound on the number of basis rows can be obtained by considering the number of rows which have the most crosses. For example, in Table VI there are 15 columns and 4 crosses, at most, in any row. A lower bound of 4 ($\frac{15}{4} = 3\frac{3}{4}$) is a little too optimistic since there are only three rows which contain four crosses. A more realistic lower bound of 5 is obtained by noting that the rows which have 4 crosses can provide crosses in at most 12 columns and that at least two additional rows containing two crosses are necessary to provide crosses in the three remaining columns.

CYCLIC PRIME IMPLICANT TABLES AND GROUP INVARIANCE

It is not always necessary to resort to enumeration in order to determine all minimum sums for a cyclic prime implicant table. Often there is a simple relation among the various minimum sums for a transmission so that they can all be determined directly from any single minimum sum by simple interchanges of variables. The process of selecting basis rows for a cyclic table can be shortened by detecting beforehand that the minimum sums are so related.

An example of a transmission for which this is true is given in Table VIII. If the variables x_1 and x_2 are interchanged, one of the minimum sums is changed into the other. In the prime implicant table the interchange of x_1 and x_2 leads to the interchange of columns 1 and 2, 5 and 6, 9 and 10, 13 and 14, and rows A and B, C and D, E and F, G and H. The transmission itself remains the same after the interchange.

In determining the basis rows for the prime implicant table, Table VIII(d), either row G or row H can be chosen as a trial basis row. If row G is selected the i-set of basis rows will result and if row H is selected the ii-set of basis rows will result. It is unnecessary to carry out the procedure of determining both sets of basis rows. Once the i-set of basis rows is known, the ii-set can be determined directly by interchanging the x_1 and x_2 variables in the i-set. Thus no enumeration is necessary in order to determine all minimum sums.

In general, the procedure for a complex prime implicant table is to determine whether there are any pairs of variables which can be interchanged without effecting the transmission. If such pairs of variables exist, the corresponding interchanges of pairs of rows are determined. A trial basis row is then selected from a pair of rows which contain the only two crosses of a column and which are interchanged when the variables are permuted. After the set of basis rows has been determined, the other set of basis rows can be obtained by replacing each basis row by the row with which it is interchanged when variables are permuted. If any step of this procedure is not possible, it is necessary to resort to enumeration.

In the preceding discussion only simple interchanges of variables have been mentioned. Actually all possible permutations of the contact variables should be considered. It is also possible that priming variables or both priming and permuting them will leave the transmission unchanged. For example, if $T = x_4 x_3' x_2 x_1' + x_4' x_3 x_2' x_1$, priming all the variables leaves the function unchanged. Also, priming x_4 and x_3 and then interchanging x_4 and x_3 does not change the transmission. The general name for this property is *group invariance*. This was discussed by Shannon.⁴

A method for determining the group invariance for a specified transmission is presented in "Detection of Group Invariance or Total Symmetry of a Boolean Function."^{*}

8 AN APPROXIMATE SOLUTION FOR CYCLIC PRIME IMPLICANT TABLES

It has not been possible to prove in general that the procedure presented in this section will always result in a minimum sum. However, this procedure should be useful when a reasonable approximation to a minimum sum is sufficient, or when it is possible to devise a proof to show that the procedure does lead to a minimum sum for a *specific transmission* (such proofs were discussed in Section 6). Since this procedure is much simpler than enumeration, it should generally be tested before resorting to enumeration.

The first step of the procedure is to select from the prime implicant table a set of rows such that (1) in each column of the table there is a cross from at least one of the selected rows and (2) none of the selected rows can be discarded without destroying property (1). Any set of rows having these properties will be called a *consistent row set*. Each consistent row set corresponds to a sum of products expression from which no product term can be eliminated directly by any of the theorems of Boolean Algebra. In particular, the consistent row sets having the fewest members correspond to minimum sums. The first step of the procedure to be described here is to select a consistent row-set. This is done by choosing one of the columns, counting the total number of crosses in each row which has a cross in this column, and then selecting the row with the most crosses. If there is more than one such row, the topmost row is arbitrarily selected. The selected row is marked with a check. In Table IX, column 30 was chosen and then row A was selected since rows A and Z each have a cross in column 30, but row A has 4 crosses while row Z has only 2 crosses. The selected row and each column in which it has a cross is then lined out. The process just described is repeated by selecting another column (which is not lined out). Crosses in lined-out columns are not counted in determining the total number of crosses in a row. The procedure is repeated until all columns are lined out.

The table is now rearranged so that all of the selected rows are at the top, and a line is drawn to separate the selected rows from the rest. Table X results from always choosing the rightmost column in Table IX. If any column contains only one cross from a selected row, the single selected-row cross is circled. Any selected row which does not have any

* See page 1445 of this issue.

TABLE VIII—DETERMINATION OF THE MINIMUM SUMS FOR
 $T = \sum (0, 1, 2, 5, 6, 7, 9, 10, 11, 13, 14, 15)$

(a)

$x_4 x_3 x_2 x_1$				
0	0	0	0	0 ✓
1	0	0	0	1 ✓
2	0	0	1	0 ✓
5	0	1	0	1 ✓
6	0	1	1	0 ✓
9	1	0	0	1 ✓
10	1	0	1	0 ✓
7	0	1	1	1 ✓
11	1	0	1	1 ✓
13	1	1	0	1 ✓
14	1	1	1	0 ✓
15	1	1	1	1 ✓

(b)

$x_4 x_3 x_2 x_1$				
0, 1	0	0	0	-
0, 2	0	0	-	0
1, 5	0	-	0	1 ✓
1, 9	-	0	0	1 ✓
2, 6	0	-	1	0 ✓
2, 10	-	0	1	0 ✓
5, 7	0	1	-	1 ✓
5, 13	-	1	0	1 ✓
6, 7	0	1	1	- ✓
6, 14	-	1	1	0 ✓
9, 11	1	0	-	1 ✓
9, 13	1	-	0	1 ✓
10, 11	1	0	1	- ✓
10, 14	1	-	1	0 ✓
7, 15	-	1	1	1 ✓
11, 15	1	-	1	1 ✓
13, 15	1	1	-	1 ✓
14, 15	1	1	1	- ✓

(c)

$x_4 x_3 x_2 x_1$				
1	5	9	13	- - 0 1
2	6	10	14	- - 1 0
5	7	13	15	- 1 - 1
6	7	14	15	- 1 1 -
9	11	13	15	1 - - 1
10	11	14	15	1 - 1 -

(d)

0 1 2 5 6 9 10 7 11 13 14 15

A	x	x	x		x						
B	x	x	x		x						
C	x			x	x						
D		x		x	x						
E		x		x	x	x					
F			x	x	x	x					
G	x	x									
H	x	x									

(e)

$$(i) (0, 1) + (2, 6, 10, 14) + (5, 7, 13, 15) + (9, 11, 13, 15)$$

$$(ii) (0, 2) + (1, 5, 9, 13) + (6, 7, 14, 15) + (10, 11, 14, 15)$$

$$T_i = x_4' x_3' x_2' + x_2 x_1' + x_3 x_1 + x_4 x_1$$

$$T_{ii} = x_4' x_3' x_1' + x_1 x_2' + x_3 x_2 + x_4 x_2$$

of its crosses circled can be discarded without violating the requirement that each column should have at least one cross from a selected row. Rows with no circled entries are discarded (one by one, since removal of one row may require more crosses to be circled) until each selected row contains at least one circled cross. This completes the first step. The selected rows now correspond to a first approximation to a minimum sum. A check should be made to determine whether the number of selected rows is equal to the minimum number of basis rows. In Table X there are at most 4 crosses per row and 26 columns so that the minimum num-

TABLE IX—TABLE OF PRIME IMPLICANTS FOR TRANSMISSION

$$T = \sum (0, 1, 2, 4, 5, 6, 7, 8, 9, 11, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30)$$

The selection of row A is shown

	0	1	2	4	8	16	5	6	9	18	20	24	7	11	13	14	19	21	25	26	28	15	23	27	29	30
A							x											x	x							
B								x																	x	
C								x																		
D									x																	
E									x																	
F								x																		
G								x																		
H								x																		
I								x																		
J								x																		
K								x																		
L								x																		
M								x																		
N								x																		
O								x																		
P								x																		
Q								x																		
R								x																		
S	x								x																	
T	x							x	x																	
U	x	x						x	x																	
V	x	x	x					x																		
W	x	x	x	x				x																		
X	x	x	x	x	x			x																		
Y	x	x	x	x	x	x		x																		
Z	x	x	x	x	x	x	x	x																		

ber of basis rows is $\lceil \frac{2^6}{4} \rceil + 1 = 7$. Since the number of selected rows is 9 there is no guarantee that they correspond to a minimum sum.

If such an approximation to a minimum sum is not acceptable, then further work is necessary in order to reduce the number of selected rows. For each of the selected rows, a check is made of whether any of the rows in the lower part of the table (non-selected rows) have crosses in all columns in which the selected row has circled crosses. In Table X row E has a circled cross only in column 19; since row Y also has a cross in column 19 rows E and Y are labeled "a". Other pairs of rows which have the same relation are labeled with lower case letters, b, c, d, e in Table X. It is possible to interchange pairs of rows which are labeled with the same lower case letter without violating the requirement that each column must contain a cross from at least one selected row. If a non-selected row is labeled with two lower case letters then it may be possible to replace two selected rows by this one non-selected row and thereby reduce the

TABLE X—TABLE IX AFTER PARTITIONING

	0	1	2	4	8	16	5	6	9	18	20	24	7	11	13	14	19	21	25	26	28	15	23	27	29	30														
	(a)	(b)	(c)	(d)	(e)		(a)	(b)	(c)	(d)	(e)		(a)	(b)	(c)	(d)	(e)		(a)	(b)	(c)	(d)	(e)		(a)	(b)	(c)	(d)	(e)											
A	x						x						x					x						x																
E		x						x						x				x						x																
F			x						x						x			x						x																
G				x						x						x			x					x																
I					x						x						x			x			x																	
K						x						x						x			x			x																
T							x						x						x			x		x																
U								x						x						x			x		x															
W									x						x						x			x		x														
B										x						x					x			x		x														
C											x						x				x			x		x														
D												x						x			x			x		x														
H													x						x			x		x		x														
J														x						x			x		x		x													
L															x						x			x		x														
M																x						x			x		x													
N																	x						x			x		x												
O																		x						x			x		x											
P																			x						x			x		x										
Q																				x						x			x		x									
R																					x						x			x		x								
S																					x						x			x		x								
V																						x						x			x		x							
X																							x						x			x		x						
Y																								x						x			x		x					
Z																									x						x			x		x				

TABLE XI—TABLE X WITH ROWS E, F, AND K, REPLACED BY ROWS Y AND J

A J Y G I T U W E F K B C D H L M N O P Q R S V X Z

total number of selected rows (a check must be made that the two selected rows being removed do not contain the only two selected-row crosses in a column). In Table X no such interchange is possible.

Next a check should be made as to whether two of the labeled non-selected rows can be used to replace three selected rows, etc. In Table X rows Y(a) and J(b) can replace rows E(a), F(b) and K or rows Y(a) and P(d) can replace rows E(a), T(d) and K. The table which results from replacing rows E, F and K by rows Y and J is given in Table XI. The number of selected rows is now 8 which is still greater than 7, the minimum number possible. This table actually represents the minimum sum for this transmission even though this cannot be proved rigorously by the procedure being described.

If it is assumed that a minimum sum can always be obtained by exchanging pairs of selected and nonselected rows until it finally becomes possible to replace two or more selected rows by a single selected row, then it is possible to show directly that the Table XI does represent a minimum sum. The only interchange possible in Table XI is that of rows T and P. If this replacement is made then a table results in which only rows J and F can be interchanged. Interchanging rows J and F does not lead to the possibility of interchanging any new pairs of rows so that this process cannot be carried any further.

On the basis of experience with this method it seems that it is not necessary to consider interchanges involving more than one non-selected row. Such interchanges have only been necessary in order to obtain alternate minimum sums; however, no proof for the fact that they are never required in order to obtain a minimum sum has yet been discovered.

9 AN ALTERNATE EXACT PROCEDURE

It is possible to represent the prime implicant table in an alternative form such as that given in Table XII(b). From this form not only the minimum sums but also *all* possible sum of products forms for the transmission which correspond to consistent row sets can be obtained systematically. For concreteness, this representation will be explained in terms of Table XII. Since column 0 has crosses only in rows B and C, any consistent row set must contain either row B or row C (or both). Similarly, column 3 requires that any consistent row set must contain either row D or row E (or both). When both columns 0 and 3 are considered they require that any consistent row set must contain either row B or row C (or both) and either row D or row E (or both). This requirement can be expressed symbolically as $(B + C)(D + E)$ where

TABLE XII — DERIVATION OF THE MINIMUM SUMS
FOR THE TRANSMISSION
 $T = \sum (0, 3, 4, 5, 6, 7, 8, 10, 11)$

		(a) Table of Prime Implicants								
		0	3	4	5	6	7	8	10	11
$x_4'x_1$	A			x	x	x	x			
$x_4'x_2'x_1'$	B	x		x						
$x_3'x_2'x_1'$	C	x						x		
$x_4'x_2x_1$	D		x				x			
$x_3'x_2x_1$	E		x							x
$x_4x_3'x_2$	F							x	x	

(b) Boolean Representation of Table
 $(B + C)(D + E)(A + B)(A)(A + D)(C)(F)(E + F)$

(c) Consistent Row Sets

$$(A, C, F, D), \quad (A, C, F, E)$$

$$T = x_4'x_3 + x_3'x_2'x_1' + x_4x_3'x_2 + x_4'x_2x_1$$

$$T = x_4'x_3 + x_3'x_2'x_1' + x_4x_3'x_2 + x_3'x_2x_1$$

addition stands for "or" (non-exclusive) and multiplication signifies "and." This expression can be interpreted as a Boolean Algebra expression and the Boolean Algebra theorems used to simplify it. In particular it can be "multiplied out":

$$(B + C)(D + E) = BD + BE + CD + CE$$

This form is equivalent to the statement that columns 0 and 3 require that any consistent row set must contain either rows B and D, or rows B and E, or rows C and D, or rows C and E.

The complete requirements for a consistent row set can be obtained directly by providing a factor for each column of the table. Thus for Table XII the requirements for a consistent row set can be written as:

$$(B + C)(D + E)(A + B)(A)(A + D)(C)(F)(E + F)$$

By using the theorems that $A \cdot (A + D) = A$ and $A \cdot A = A$, this can be simplified to $AC(F(D + E))$. Thus the two consistent row sets for this table are A, C, F, D and A, C, F, E and since they both contain the same number of rows, they both represent minimum sums. This is true only because rows D and E contain the same number of crosses. In general, each row should be assigned a weight $w = n - \log_2 k$, where n is the number of variables in the transmission being considered and

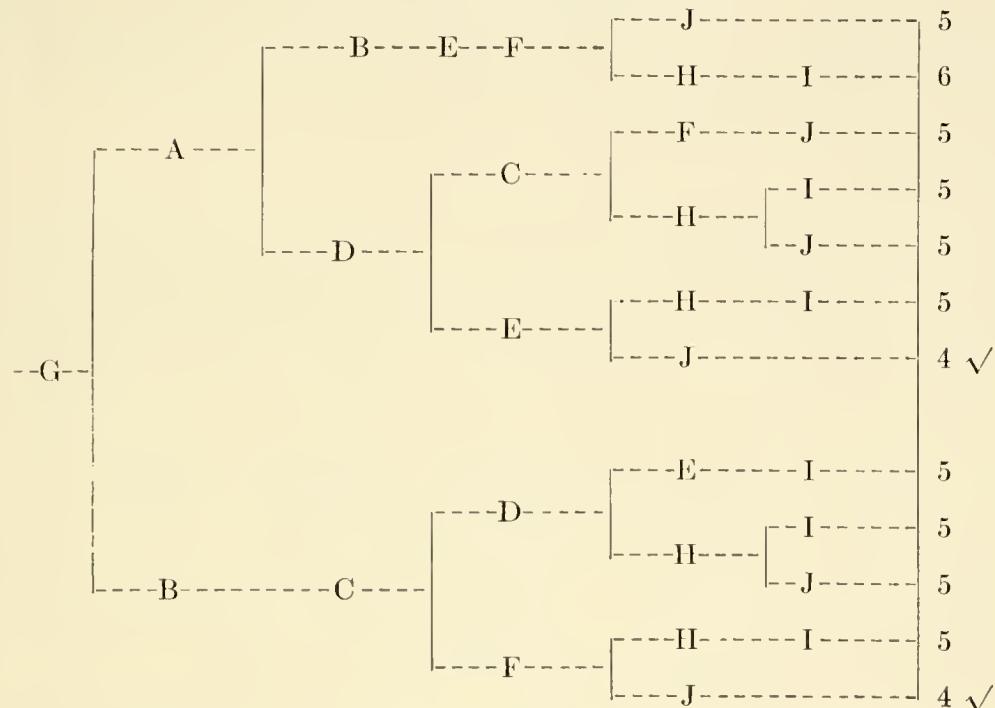
TABLE XIII — DETERMINATION OF THE MINIMUM SUMS FOR THE PRIME IMPLICANT TABLE OF TABLE VII BY MEANS OF THE BOOLEAN REPRESENTATION

- (a) Boolean representation of the Prime Implicant Table of Table VI
 $(A+B)(A+C)(B+D)(C+E)(D+F)(G)(E+F+H)(G+I)(H+J)(I+J)$

(b) The expression of (a) after multiplying out. (The terms in italic correspond to minimum sums)

$$ADEJG + ACDFJG + ACDHJG + ADEHIG + ACDHIG + ABEFJG + ABEFHIG + BCDEJG + BCDHJG + BCDHIG + BCFJG + BCFHIG$$

(c) Tree circuit equivalent of (b)



k is the number of crosses in the row.* To select the minimum sums, the sum of the weights of the rows should be calculated for each row set containing the fewest rows. The row sets having the smallest total weight correspond to minimum sums. If, instead of the minimum sum, the form leading to the two-stage diode-logic circuit requiring fewest diodes is desired, a slightly different procedure is appropriate. To each row set is assigned a total weight equal to the sum of the weights of the rows plus the number of rows in the set. The desired form then corresponds to the row set having the smallest total weight.

The procedure for an arbitrary table is analogous. A more complicated example is given in Table XIII. In this example the additional

* $n \log_2 k$ is the number of literals in the prime implicant corresponding to a row containing k crosses.

theorem $(A + B)(A + C) = (A + BC)$ is useful. This example shows that for a general table the expressions described in this Section and the multiplication process can become very lengthy. However, this procedure is entirely systematic and may be suitable for mechanization.

Since the product of factors representation of a prime implicant table is a Boolean expression, it can be interpreted as the transmission of a contact network. Each consistent row set then corresponds to a path through this equivalent network. By sketching the network directly from the product of factors expression, it is possible to avoid the multiplication process. In particular the network should be sketched in the form of a tree, as in Table XIII(c) and the Boolean Algebra theorems used to simplify it as it is being drawn. For hand calculations, this method is sometimes easier than direct multiplication.

10 *d*-TERMS

In Section 1 the possibility of having *d*-entries in a table of combinations was mentioned. Whenever there are combinations of the relay conditions for which the transmission is not specified, *d*-entries are placed in the *T*-column of the corresponding rows of the table of combinations.

TABLE XIV — DETERMINATION OF THE MINIMUM
SUM FOR THE TRANSMISSION

$$T = \sum(5, 6, 13) + d(9, 14) \text{ WHERE } 9 \text{ AND } 14 \text{ ARE THE } d\text{-TERMS}$$

(a) Determination of Prime Implicants

	$x_4 x_3 x_2 x_1$					$x_4 x_3 x_2 x_1$			
	5	0	1	0	1	✓		5	13
	6	0	1	1	0	✓		6	14
(d)	9	1	0	0	1	✓		9	13
	13	1	1	0	1	✓			
(d)	14	1	1	1	0	✓			

(b) Prime Implicant Table

	5	6	13
*	x		x
*		x	
			x

(c)

Basis rows: (5, 13), (6, 14)

(d)

$$T = x_3 x_2' x_1 + x_3 x_2 x_1'$$

The actual values (0 or 1) of these d -entries are chosen so as to simplify the form of the transmission. This section will describe how to modify the method for obtaining a minimum sum when the table of combinations contains d -entries.

The p -terms which correspond to d -entries in the table of combinations will be called d -terms. These d -terms should be included in the list of p -terms which are used to form the prime implicants. See Table XIV. However, in forming the prime implicant table, columns corresponding to the d -terms should *not* be included, Table XIV(b). The d -terms are used in forming the prime implicants in order to obtain prime implicants containing the fewest possible literals. If columns corresponding to the d -terms were included in forming the prime implicant table this would correspond to setting all the d -entries in the table of combinations equal to 1. This does not necessarily lead to the simplest minimum sum. In the procedure just described, the d -entries will automatically be set equal to either 0 or 1 so as to produce the simplest minimum sum. For the transmission of Table XIV the 14 d -entry has been set equal to 1 and the 9 d -entry has been set equal to 0.

11 NON-CANONICAL SPECIFICATIONS

A transmission is sometimes specified not by a table of combinations or a canonical expansion, but as a sum of product terms (not necessarily prime implicants). The method described in Section 3 is applicable to such a transmission if the appropriate table of combinations (decimal specification) is first obtained. However, it is possible to modify the procedure to make use of the fact that the transmission is already partly reduced. The first step is to express the transmission in a table of binary characters such as Table XVa. Then each pair of characters is examined to determine whether any different character could have been formed from the characters used in forming the characters of the pair. For example, in Table XV(a) a (1)(0 0 0 0 1) was used in forming the (0, 1)(0 0 0 0 -) character and a (3)(0 0 0 1 1) was used in forming the (3, 7)(0 0 - 1 1) character. These can be combined to form a new character (1, 3)(0 0 0 - 1). The new characters formed by this process are listed in another column such as Table XV(b). This process is continued until no new characters are formed.

In examining a pair of characters, it is sufficient to determine whether there is only one position where one character has a one and the other character has a zero. If this is true a new character is formed which has a dash in this position and any other position in which both characters have dashes, and has a zero (one) in any position in which either charac-

TABLE XV—DETERMINATION OF THE PRIME IMPLICANTS FOR THE TRANSMISSION OF TABLE XV SPECIFIED AS A SUM OF PRODUCT TERMS

(a) Specification		(b) Characters Derived from (a)	
		$x_5 x_4 x_3 x_2 x_1$	$x_5 x_4 x_3 x_2 x_1$
0	1	0 0 0 0 - ✓	1 3 0 0 0 - 1 ✓
0	2	0 0 0 - 0 ✓	2 3 0 0 0 1 - ✓
3	7	0 0 - 1 1	7 15 0 - 1 1 1 ✓
14	15	0 1 1 1 -	7 23 - 0 1 1 1 ✓
22	23	1 0 1 1 -	15 31 - 1 1 1 1 ✓
29	31	1 1 1 - 1	23 31 1 - 1 1 1 ✓
(c) Characters Derived from (a) and (b)		$x_5 x_4 x_3 x_2 x_1$	
0 1 2 3		0 0 0 - -	
7 15 23 31		- - 1 1 1	

ter has a zero (one). In Table XVa the (0, 1) character has a zero in the x_2 -position while the (3, 7) character has a one in the x_2 -position. A new character is formed (1, 3) which has a dash in the x_2 -position.

This rule for constructing new characters is actually a generalization of the rule used in Section 3 and corresponds to the theorem.

$$x_1 x_2 + x_1' x_3 = x_1 x_2 + x_1' x_3 + x_2 x_3 .$$

Repeated application of this rule will lead to the complete set of prime implicants. As described in Section 3, any character which has all of the numbers of its decimal label appearing in the label of another character should be checked. The unchecked characters then represent the prime implicants. The process described in this section was discussed from a slightly different point of view by Quine.⁷

12 SUMMARY AND CONCLUSIONS

In this paper a method has been presented for writing any transmission as a minimum sum. This method is similar to that of Quine; however, several significant improvements have been made. The notation has been simplified by using the symbols 0, 1 and - instead of primed and unprimed variables. While it is not completely new in itself, this notation is especially appropriate for the arrangement of terms used in determining the prime implicants. Listing the terms in a column which is partitioned so as to place terms containing the same number of 1's in the same partition reduces materially the labor involved in determining the prime implicants. Such a list retains some of the advantage of the arrangement of squares in the Karnaugh Chart without requiring a geometrical representation of an n -dimensional hypercube. Since the

procedure for determining the prime implicants is completely systematic it is capable of being programmed on a digital computer. The arrangement of terms introduced here then results in a considerable saving in both time and storage space over previous methods, making it possible to solve larger problems on a given computer. It should be pointed out that this procedure can be programmed on a decimal machine by using the decimal labels instead of the binary characters introduced.

A method was presented for choosing the minimum sum terms from the list of prime implicants by means of a table of prime implicants. This is again similar to a method presented by Quine. However, Quine did not give any systematic procedure for handling cyclic prime implicant tables; that is, tables with more than one cross in each column. In this paper a procedure is given for obtaining a minimum sum from a cyclic prime implicant table. In general, this procedure requires enumeration of several possible minimum sums. If a transmission has any non-trivial group invariances it may be possible to avoid enumeration or to reduce considerably the amount of enumeration necessary. A method for doing this is given.

The process of enumeration used for selecting the terms of the minimum sum from a cyclic prime implicant table is not completely satisfactory since it can be quite lengthy. In seeking a procedure which does not require enumeration, the method involving the group invariances of a transmission was discovered. This method is an improvement over complete enumeration, but still has two shortcomings. There are transmissions which have no nontrivial group invariances but which give rise to cyclic prime implicant tables. For such transmissions it is still necessary to resort to enumeration. Other transmissions which do possess nontrivial group invariances still require enumeration after the invariances have been used to simplify the process of selecting minimum sum terms. More research is necessary to determine some procedure which will not require any enumeration for cyclic prime implicant tables. Perhaps the concept of group invariance can be generalized so as to apply to all transmissions which result in cyclic prime implicant tables.

13 ACKNOWLEDGEMENTS

The author wishes to acknowledge his indebtedness to Professor S. H. Caldwell, Professor D. A. Huffman, Professor W. K. Linvill, and S. H. Unger with whom the author had many stimulating discussions. Thanks are due also to W. J. Cadden, C. Y. Lee, and G. H. Mealy for their helpful comments on the preparation of this paper.

This research was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

BIBLIOGRAPHY

1. Karnaugh, M., The Map Method for Synthesis of Combinational Logic Circuits, *Trans. A.I.E.E.*, **72**, Part I pp. 593-598, 1953.
2. Keister, W., Ritchie, A. E., Washburn, S., *The Design of Switching Circuits*, New York, D. Van Nostrand Company, Inc., 1951.
3. Shannon, C. E., A Symbolic Analysis of Relay and Switching Circuits, *Trans. A.I.E.E.*, **57**, pp. 713-723, 1938.
4. Shannon, C. E., The Synthesis of Two-Terminal Switching Circuits, *B.S.T.J.*, **28**, pp. 59-98, 1949.
5. Staff of the Harvard Computation Laboratory, *Synthesis of Electronic Computing and Control Circuits*, Cambridge, Mass., 1951, Harvard University Press.
6. Quine, W. V., The Problem of Simplifying Truth Functions, *The American Mathematical Monthly*, **59**, No. 8, pp. 521-531, Oct., 1952.
7. Quine, W. V., A Way to Simplify Truth Functions, *The American Mathematical Monthly*, **62**, pp. 627-631, Nov., 1955.

Detection of Group Invariance or Total Symmetry of a Boolean Function*

By E. J. McCLUSKEY, Jr.

(Manuscript received June 26, 1956)

A method is presented for determining whether a Boolean function possesses any group invariance; that is, whether there are any permutations or primings of the independent variables which leave the function unchanged. This method is then extended to the detection of functions which are totally symmetric.

1 GROUP INVARIANCE

For some Boolean transmission functions (transmissions, for short) it is possible to permute the variables, or prime some of the variables, or both permute and prime variables without changing the transmission. The following material presents a method for determining, for any given transmission, which of these operations (if any) can be carried out without changing the transmission.

The permutation operations will be represented symbolically as follows:

$S_{123\dots n}T$ will represent the transmission T with no variables permuted
 $S_{213\dots n}T$ will represent the transmission T with the x_1 and x_2 variables interchanged, etc.

Thus $S_{1432}T(x_1, x_2, x_3, x_4) = T(x_1, x_4, x_3, x_2)$

The symbolic notation for the priming operation will be as follows:

$N_{0000\dots 0}T$ will represent the transmission T with no variables primed
 $N_{0110\dots 0}T$ will represent the transmission T with the x_2 and x_3 variables primed, etc.

Thus $N_{1010}T(x_1, x_2, x_3, x_4) = T(x_1', x_2, x_3', x_4)$.

The notation for the priming operator can be shortened by replacing the binary subscript on N by its decimal equivalent. Thus N_9T is equiv-

* This paper is derived from a thesis submitted to the Massachusetts Institute of Technology in partial fulfillment of the requirements for the degree of Doctor of Science on April 30, 1956.

TABLE I — TRANSMISSION MATRICES SHOWING EFFECT OF
INTERCHANGING OR PRIMING VARIABLES

(a) Transmission Matrix	(b) Transmission Matrix with x_3 and x_4 columns interchanged	(c) Transmission Ma- trix with entries of the x_3 and x_4 columns primed
$x_1 \ x_2 \ x_3 \ x_4$	$x_1 \ x_2 \ x_4 \ x_3$	$x_1 \ x_2 \ x_3' \ x_4'$
0 0 0 0	0 0 0 0	3 0 0 1 1
1 0 0 1	2 0 0 1 0	2 0 0 1 0
2 0 0 1 0	1 0 0 0 1	1 0 0 0 1
9 1 0 0 1	10 1 0 1 0	10 1 0 1 0
10 1 0 1 0	9 1 0 0 1	9 1 0 0 1
11 1 0 1 1	11 1 0 1 1	8 1 0 0 0

alent to $N_{1001}T$. The permutation and priming operators can be combined. For example,

$$S_{2134}N_3T(x_1, x_2, x_3, x_4) = T(x_2, x_1, x_3', x_4')$$

The symbols S_iN_j form a mathematical group,¹ hence the term group invariance.

The problem considered here is that of determining which N_i and S_j satisfy the relation $N_iS_jT = T$ for a given transmission T . Since there are only a finite number of different N_i and S_j operators it is possible in principle to compute N_iS_jT for all possible N_iS_j and then select those N_iS_j for which $N_iS_jT = T$. If T is a function of n variables, there are $n!$ possible S_j operators and $2^n N_i$ operators so that there are $n!2^n$ possible combinations of N_iS_j . Actually, if $N_iS_jT = T$ then N_iT must equal $S_jT^{(2)}$ so that it is only necessary to compute all N_iT and all S_jT . For $n = 4$, $n! = 24$ and $2^n = 16$ so that the number of possibilities to be considered is quite large even for functions of only four variables. It is possible to avoid enumerating all N_iT and S_jT by taking into account certain characteristics of the transmission being considered.

The first step in determining the group invariances of a transmission is the same as that for finding the prime implicants.* The binary equivalents of the decimal numbers which specify the transmission are listed as in Table I(a). This list of binary numbers will be called the *transmission matrix*. When two variables are interchanged, the corresponding columns of the transmission matrix are also interchanged, Table I(b). When a variable is primed, the entries in the corresponding column of the transmission matrix are also primed, 0 replaced by 1 and 1 replaced by 0, Table I(c).

If an N_iS_j operation leaves a transmission unchanged then the cor-

* Minimization of Boolean Functions, see page 1417 of this issue.

responding matrix operations will not change the transmission matrix aside from possibly reordering the rows. In other words, it should be possible to reorder the rows of the modified transmission matrix to regain the original transmission matrix. The matrices of Table I(a) and (b) are identical except for the interchange of the 1 and 2 and the 9 and 10 rows. It is not possible to make the matrix of Table I(c) identical with that of Table I(a) by reordering rows; therefore the operation of priming the x_3 and x_4 variables does not leave the transmission $T = \sum (0, 1, 2, 9, 10, 11)$ unchanged.

If interchanging two columns of a matrix does not change the matrix aside from rearranging the rows, then the columns which were interchanged must both contain the same number of 1's (and 0's). This must

TABLE II — PARTITIONING OF THE STANDARD MATRIX FOR
 $T = \sum (4, 5, 7, 8, 9, 11, 30, 33, 49)$

(a) Transmission Matrix						(b) Standard Matrix for (a) Matrix								
	x_1	x_2	x_3	x_4	x_5	x_6		x_1	x_2	x_3	x_4	x_5	x_6'	Weight
4	0	0	0	1	0	0	4	0	0	0	1	0	0	1
8	0	0	1	0	0	0	8	0	0	1	0	0	0	1
	<hr/>						32	1	0	0	0	0	0	1
5	0	0	0	1	0	1	5	0	0	0	1	0	1	2
9	0	0	1	0	0	1	6	0	0	0	1	1	0	2
33	1	0	0	0	0	1	9	0	0	1	0	0	1	2
	<hr/>						10	0	0	1	0	1	0	2
7	0	0	0	1	1	1	48	1	1	0	0	0	0	2
11	0	0	1	0	1	1	31	0	1	1	1	1	1	5
49	1	1	0	0	0	1								
	<hr/>													
30	0	1	1	1	1	0								
Number of 0's	7	7	5	5	6	3								
Number of 1's	2	2	4	4	3	6								
(c) Second Partitioning of rows for (b) matrix						(d) Final Partitioning for (b) matrix								
x_1	x_2	x_3	x_4	x_5	x_6'	x_1	x_2	x_3	x_4	x_5	x_6'			
0	0	0	1	0	0	0	0	0	1	0	0			
0	0	1	0	0	0	0	0	1	0	0	0			
	<hr/>						1	0	0	0	0	0		
1	0	0	0	0	0	0	0	0	0	1	0			
0	0	0	1	0	1	0	0	0	1	0	1			
0	0	0	1	1	0	0	0	0	1	1	0			
0	0	1	0	0	1	0	0	0	1	0	0			
0	0	1	0	1	0	0	0	0	1	0	1			
	<hr/>						1	1	0	0	0	0		
1	1	0	0	0	0	0	0	1	1	1	1			
0	1	1	1	1	1	0	0	1	1	1	1			

be true since rearranging the rows of a matrix does not change the total number of 1's in each column. Similarly, if priming some columns of a matrix leaves the rows unchanged, either each column must have an equal number of 1's and 0's or else for each primed column which has an unequal number of 0's and 1's there must be a second primed column which has as many 1's as the first primed column has 0's and vice versa. Such pairs of columns must also be interchanged to keep the total number of 1's in each column invariant. For the matrix of Table II(a) the only operations that need be considered are either interchanging x_1 and x_2 or x_3 and x_4 or priming and interchanging x_5 and x_6 .

For the present it will be assumed that no columns of the matrix have an equal number of 0's and 1's. It is possible to determine all permuting and priming operations which leave such a matrix unchanged by considering only permutation operations on a related matrix. This related matrix, called the *standard matrix*, is formed by priming all the columns of the original matrix which have more 1's than 0's, the x_6 column in the matrix of Table II(a). Each column of a standard matrix must contain more 0's than 1's, Table II(b). The N_iS_j operations which leave the original matrix unchanged can be determined directly from the operations that leave the corresponding standard matrix unchanged. It is therefore only necessary to consider standard matrices.

Since no columns of a standard matrix have an equal number of 1's and 0's and no columns have more 1's than 0's it is only necessary to consider permuting operations. The number of 1's in a column (or row) will be called the *weight* of the column (or row). Only columns or rows which have the same weights can be interchanged. The matrix should be partitioned so that all columns (or rows) in the same partition have the same weight, Table II(b). It is now possible to interchange columns in the same column partition and check whether pairs of rows from the same row partition can then be interchanged to regain the original matrix. This can usually be done by inspection. For example, in Table II(b) if columns x_4 and x_3 are interchanged, then interchanging rows 4 and 8, 5 and 9, and 6 and 10 will regain the original matrix.

The process of inspection can be simplified by carrying the partitioning further. In the matrix of Table II(b), row 32 cannot be interchanged with either row 8 or row 4. This is because it is not possible to make row 32 identical with either row 8 or row 4 by interchanging columns x_1 and x_2 . Row 32 has weight 1 in these columns while rows 8 and 14 both have weight 0. In general, only rows which have the same weight in *each* submatrix can be interchanged. Permuting columns of the same partition does not change the weight of the rows in the corresponding submatrices.

The matrix can therefore be further partitioned by separating the rows into groups of rows which have the same weight in every column partition, Table II(c). Similar remarks hold for the columns so that it may then be necessary to partition the columns again so that each column in a partition has the same weight in each submatrix, Table II(d). Partitioning the columns may make it necessary to again partition the rows, which in turn may make more column partitioning necessary. This process should be carried out until a matrix results in which each row (column) of each submatrix has the same weight. Inspection is then used to determine which row and column permutations will leave the matrix unchanged. Only permutations among rows or columns in the same partition need be considered.

From the matrix of Table II(d) it can be seen that permuting either columns x_3 and x_4 or columns x_5 and x_6' will not change the matrix aside from reordering certain rows. This means that interchanging x_3 and x_4 or priming and interchanging x_5 and x_6 in the original transmission will leave the transmission unchanged. Interchanging x_6' and x_6 means replacing x_5 by x_6' and x_6 by x_5' which is the same as interchanging x_5 and x_6 and then priming both x_5 and x_6 . Thus for the transmission of Table II $S_{124356}T = T$ and $N_{000011}S_{123465}T = N_3S_{123465}T = T$.

A procedure has been presented for determining the group invariance of any transmission matrix which does not have an equal number of 1's and 0's in any column. This must now be extended to matrices which do have equal numbers of 0's and 1's in some columns, Table III(a). For such matrices the procedure is to prime appropriate columns so that there are either more 0's than 1's or the same number of 0's and 1's in each column, Table III(a). This matrix is then partitioned as described above and the permutations which leave the matrix unchanged are determined. The matrix of Table III(a) is so partitioned. Interchanging

TABLE III — TRANSMISSION MATRICES FOR $T = \sum (0, 6, 9, 12)$

(a) Transmission Matrix

(b) Transmission Matrix
with x_1 and x_2 primed

	x_1	x_2	x_3	x_4		x_1'	x_2'	x_3	x_4
0	0	0	0	0		0	0	0	0
6	0	1	1	0		10	1	0	1
9	1	0	0	1		5	0	1	0
12	1	1	0	0		12	1	1	0
Number of 0's	2	2	3	3			2	2	3
Number of 1's	2	2	1	1			2	2	1

both x_1 and x_2 , and x_3 and x_4 leave this matrix unchanged so that $S_{2143}T = T$. The possibility of priming different combinations of the columns which have an equal number of 0's and 1's must now be considered. Certain of the possible combinations can be excluded beforehand. In Table III(a) the only possibility which must be considered is that of priming both x_1 and x_2 . If only x_1 or x_2 is primed, there will be no row which has all zeros. No permutation of the columns of this matrix (with x_1 or x_2 primed) can produce a row with all zeros. Therefore this matrix cannot possibly be made equal to the original matrix by rearranging rows and columns. Priming both x_1 and x_2 must be considered since the 12-row will be converted into a row with all zeros. The operation of priming x_1 and x_2 is written symbolically as $N_{1100} = N_{12}$. In general, if the matrix has a row consisting of all zeros, only those N_i operations for which i is the number of some row in the matrix, need be considered. If the row does not have an all-zero row, only those N_i for which i is *not* the number of some row need be considered. Similarly, if the matrix has a row consisting of all 1's, only those N_i for which there is some row of the matrix which will be converted into an all-one row, need be considered. This is equivalent to considering only those N_i for which some row has a number $k = 2^n - 1 - i^*$ where n is the number of columns. If the matrix does *not* have an all-one row, only those N_i for which *no* row has a number $k = 2^n - 1 - i$ should be considered.

Each priming operation which is not excluded by these rules is applied to the transmission matrix. The matrices so formed are then partitioned as described previously. Any of these matrices that have the same partitioning as the original matrix are then inspected to see if any row and column permutations will convert them to the original matrix. For the matrix of Table III(a) the operation of priming both x_1 and x_2 was not excluded. The matrix which results when these columns are primed is shown in Table III(b). Inspection of this figure shows that interchange of either x_3 and x_4 or x_1' and x_2' will convert the matrix back to the matrix of Table III(a). Therefore, for the transmission of this table $S_{1243}N_{1100}T = T$ and $S_{2134}N_{1100}T = T$.

2 TOTAL SYMMETRY

There are certain transmissions whose value depends not on which relays are operated but only on how many relays are operated. For

* The number of the row which has all ones is $2^n - 1$. If N_i operating on some row, k , is to produce the all-one row, i must have 1's wherever k has 0's and vice versa. This means that

$$i + k = 2^n - 1 \quad \text{or} \quad k = 2^n - 1 - i.$$

TABLE IV — TRANSMISSION MATRIX FOR
 $T = \sum (3, 5, 6, 7) = S_{2,3}(x_1, x_2, x_3)$

	x_3	x_2	x_1
3	0	1	1
5	1	0	1
6	1	1	0
7	1	1	1

example, the transmission of Table IV equals 1 whenever two or three relays are operated. For such transmissions any permutation of the variables leaves the transmission unchanged. These transmissions are called *totally symmetric*.³ They are usually written in the form, $T = S_{a_1, a_2, \dots, a_m}(x_1, x_2, \dots, x_n)$, where the transmission is to equal 1 only when exactly a_1 or a_2 or \dots or a_m of the variables x_1, x_2, \dots, x_n are equal to one. The transmission of Table IV can be written as $S_{2,3}(x_1, x_2, x_3)$. This definition of symmetric transmissions can be generalized by allowing some of the variables (x_1, x_2, \dots, x_n) to be primed. Thus the transmission $S_3(x_1, x_2', x_3)$ will equal 1 only when $x_1 = x_2' = x_3 = 1$ or $x_1 = x_3 = 1$ and $x_2 = 0$. It is useful to know when a transmission is totally symmetric since special design techniques exist for such functions.⁴

It is possible to determine whether a transmission is totally symmetric from its matrix. Unless all columns of the standard matrix derived from the transmission matrix have the same weight, the transmission cannot possibly be totally symmetric. If all columns do have equal weights, the rows should be partitioned into groups of rows which all have the same weight. Whether the transmission is totally symmetric can now be determined by inspection. If there is a row of weight k ; that is, a row which contains k 1's, then every possible row of weight k must also be included in the matrix. This means that there must be ${}_n C_k$ rows of weight k where n is the number of columns (variables).* If any possible row of weight k was not included then the corresponding k literals could be set equal to 1 without the transmission being equal to 1. This contradicts the definition of a totally symmetric transmission. In Table V(b) there are 4 rows of weight 1 and 1 row of weight 4. Since ${}_4 C_1 = 4$ and ${}_4 C_4 = 1$ this transmission is totally symmetric and can be written as $S_{1,4}(x_1, x_2', x_3, x_4')$. The number of rows of weight 1 in Table V(d) is 2 and since ${}_4 C_1 = 4$ this transmission is *not* totally symmetric.

A difficulty arises if all columns of a transmission matrix contain equal

* ${}_n C_k$ is the binomial coefficient $\frac{n!}{(n - k)!k!}$

TABLE V — DETERMINATION OF TOTALLY SYMMETRIC TRANSMISSION

(a) Transmission Matrix for
 $T = \sum (1, 4, 7, 10, 13)$

	x_1	x_2	x_3	x_4
1	0	0	0	1
4	0	1	0	0
10	1	0	1	0
7	0	1	1	1
13	1	1	0	1

(b) Standard Matrix for
 $T = \sum (1, 4, 7, 10, 13)$
showing that
 $T = S_{1,4} (x_1, x_2', x_3, x_4')$

	x_1	x_2'	x_3	x_4'
1	0	0	0	1
2	0	0	1	0
4	0	1	0	0
8	1	0	0	0
15	1	1	1	1

Number of 0's 3 2 3 2
Number of 1's 2 3 2 33 3 3 3
2 2 2 2(c) Transmission Matrix for
 $T = \sum (3, 5, 10, 12, 13)$

	x_1	x_2	x_3	x_4
3	0	0	1	1
5	0	1	0	1
10	1	0	1	0
12	1	1	0	0
13	1	1	0	1

(d) Standard Matrix for
 $T = \sum (3, 5, 10, 12, 13)$
showing that it is not
totally symmetric

	x_1'	x_2'	x_3	x_4'
0	0	0	0	0
1	0	0	0	1
8	1	0	0	0
7	0	1	1	1
14	1	1	1	0

Number of 0's 2 2 3 2
Number of 1's 3 3 2 3TABLE VI — DETERMINATION OF TOTAL SYMMETRY FOR
 $T = \sum (0, 3, 5, 10, 12, 15)$ (a) Transmission Matrix
for $T(x_1, x_2, x_3, x_4)$

	x_1	x_2	x_3	x_4
0	0	0	0	0
3	0	0	1	1
5	0	1	0	1
10	1	0	1	0
12	1	1	0	0
15	1	1	1	1

(b) Standard Matrix
for $T(1, x_2, x_3, x_4)$

	x_2'	x_3'	x_4
1	0	0	0
0	1	0	0
0	0	1	0

Number of 0's 2 2 2
Number of 1's 1 1 1Number of 0's 3 3 3 3
Number of 1's 3 3 3 3 $T(1, x_2, x_3, x_4) = S_1(x_2', x_3', x_4)$ (c) Standard Matrix for $T(0, x_2, x_3, x_4)$

	x_2	x_3	x_4'
0	0	0	1
0	1	0	0
1	0	0	0

Number of 0's 2 2 2

Number of 1's 1 1 1

 $T(0, x_2, x_3, x_4) = S_1(x_2, x_3, x_4') = S_2(x_2', x_3', x_4)$

numbers of zeros and ones as in Table VI(a). For such a matrix it is not clear which variables should be primed. It is possible to avoid considering all possible primings by "expanding" the transmission about one of the variables by means of the theorem

$$T(x_1, x_2, \dots, x_n) = x_1 T(1, x_2, \dots, x_n) + x_1' T(0, x_2, \dots, x_n)^{2,3}$$

and then making use of the relation:

$$\begin{aligned} S_{a_1, a_2, \dots, a_m}(x_1, x_2, \dots, x_n) \\ = x_1 S_{a_1-1, a_2-1, a_2-1, \dots, a_{m-1}}(x_2, \dots, x_m) \\ + x_1' S_{a_1, a_2, \dots, a_m}(x_2, \dots, x_n)^5 \end{aligned}$$

This technique is illustrated in Table VI. The standard matrix for $T(1, x_2, x_3, x_4)$ has three rows each containing a single one so that

$$T(1, x_2, x_3, x_4) = S_1(x_2', x_3', x_4)$$

The transmission $T(0, x_2, x_3, x_4)$ has an identical standard matrix so that

$$T(0, x_2, x_3, x_4) = S_1(x_2, x_3, x_4')$$

This can be written in terms of x_2' , x_3' , and x_4 :

$$S_1(x_2, x_3, x_4') = S_2(x_2', x_3', x_4)^5.$$

Finally

$$\begin{aligned} T(x_1, x_2, x_3, x_4) &= x_1 T(1, x_2, x_3, x_4) + x_1' T(0, x_2, x_3, x_4) \\ &= x_1 S_1(x_2', x_3', x_4) + x_1' S_2(x_2', x_3', x_4) = S_2(x_1, x_2', x_3', x_4).^* \end{aligned}$$

The method just presented for detecting total symmetry is more systematic than the only other available method⁵ and applies for transmissions of any number of variables.

BIBLIOGRAPHY

1. Birkhoff, G., and MacLane, S., *A Survey of Modern Algebra*, The MacMillan Company, New York.
2. Shannon, C. E., The Synthesis of Two-Terminal Switching Circuits, *B.S.T.J.*, **28**, pp. 59-98, 1949.
3. Shannon, C. E., A Symbolic Analysis of Relay and Switching Circuits, *Trans. A.I.E.E.*, **57**, pp. 713-723, 1938.
4. Keister, W., Ritchie, A. E., Washburn, S., *The Design of Switching Circuits*, New York, D. Van Nostrand Company, Inc., 1951.
5. Caldwell, S. H., The Recognition and Identification of Symmetric Switching Circuits, *Trans. A.I.E.E.*, **73**, Part I, pp. 142-146, 1954.

* This technique for transmission matrices having an equal number of zeros and ones in all columns was brought to the author's attention by Wayne Kellner, a student at the Massachusetts Institute of Technology.

Bell System Technical Papers Not Published in This Journal

ANDERSON, O. L.¹

Effect of Pressure on Glass Structure, J. Appl. Phys., 27, pp. 943-949, Aug., 1956.

ANDERSON, P. W.¹

Ordering and Antiferromagnetism in Ferrites, Phys. Rev., 102, pp. 1008-1013, May 15, 1956.

ANDERSON, P. W., see Holden, A. N.

ARNOLD, S. M.,¹ and KOONCE, S. ELOISE¹

Filamentary Growths of Metals at Elevated Temperatures, J. Appl. Phys., Letter to the Editor, 27, p. 964, Aug., 1956.

BONNEVILLE, S., See Noyes, J. W.

BRIDGERS, H. E.¹

The Formation of P-N Junctions in Semiconductors by the Variation of Crystal Growth Parameters, J. Appl. Phys., 27, pp. 746-751, July, 1956.

BOZORTH, R. M.,¹ WILLIAMS, H. J.,¹ and WALSH, DOROTHY E.¹

Magnetic Properties of Some Orthoferrites and Cyanides at Low Temperatures, Phys. Rev., 103, pp. 572-578, August 1, 1956.

CHASE, F. H.¹

Power Regulation by Semiconductors, Elec. Engg., 75, pp. 818-822, Sept., 1956.

CHEN, W. H., see Lee, C. Y.

¹ Bell Telephone Laboratories, Inc.

CHYNOWETH, A. G.¹

Spontaneous Polarization of Guanidine Aluminum Sulfate Hexahydrate at Low Temperatures, *Phys. Rev.*, **102**, pp. 1021–1023, May 15, 1956.

COOK, R. K.¹ and WASILIK, J. H.¹

Anelasticity and Dielectric Loss of Quartz, *J. Appl. Phys.*, **27**, pp. 836–837, July, 1956.

DARROW, K. K.¹

Electron Physics in America, *Physics Today*, **9**, pp. 23–27, Aug., 1956

DAVID, E. E., JR.¹

Naturalness and Distortion in Speech Processing Devices, *J. Acous. Soc. Am.*, **28**, pp. 586–589, July, 1956.

DAVID, E. E., JR.,¹ and McDONALD, H. S.¹

A Bit-Squeezing Technique Applied to Speech Signals, *I.R.E. Convention Record*, **4**, Part 4, pp. 148–153, July, 1956.

DEWALD, J. F.¹ and LEPOUTRE, G.¹

I — The Thermoelectric Properties of Metal-Ammonia. II — The Thermoelectric Power of Sodium and Potassium Solutions at -78° and the Effect of Added Salt on the Thermoelectric Power of Sodium at -33° . III — Theory and Interpretation of Results, *J. Am. Chem. Soc.*, **78**, pp. 2953–2962, July 5, 1956.

EDER, M. J., see Veloric, H. S.

EMBREE, M. L.,¹ and WILLIAMS, D. E.¹

An Automatic Card Punching Transistor Test Set, Proc. 1956 Electronic Components Symposium, pp. 125–130, 1956.

FARRAR, H. K., see Maxwell, J. L.

FEHER, G.¹

Method of Polarizing Nuclei in Paramagnetic Substances, *Phys. Rev.*, Letter to the Editor, **103**, pp. 500–501, July 15, 1956.

¹ Bell Telephone Laboratories, Inc.

FEHER, G.,¹ and GERE, E.¹

Polarization of Phosphorus Nuclei in Silicon, Phys. Rev., Letter to the Editor, 103, pp. 501-503, July 15, 1956.

FREYNIK, H. S., see Gohn, G. R.

FTHENAKIS, E.¹

A Voltage Regulator Using High Speed of Response Magnetic Amplifiers With Transistor Driver, Proc. Special Tech. Conf. on Magnetic Amplifiers, T-86, pp. 185-199, July, 1955.

GAUDET, G., see Noyes, J. W.

GELLER, S.,¹ and WOOD, Mrs. E. A.,¹

Crystallographic Studies of Perovskite-Like Compounds. I—Rare Earth Orthoferrites and YFeO_3 , YCrO_3 , YAlO_3 , Acta Crys., 9, pp. 563-568, July 10, 1956.

GERE, E., see Feher, G.

GIANOLA, U. F.,¹ and JAMES, D. B.¹

Ferromagnetic Coupling Between Crossed Coils, J. Appl. Phys., 27, pp. 608-609, June, 1956.

GILBERT, E. N.¹

Enumeration of Labelled Graphs, Canadian J. of Math., 8, pp. 405-411, 1956.

GOHN, G. R.¹

Fatigue and Its Relation to the Mechanical and Metallurgical Properties of Metals, SAE Trans., 64, pp. 31-40, 1956.

GOHN, G. R.,¹ FREYNIK, H. S.,⁸ and GUERARD, J. P.¹

The Mechanical Properties of Wrought Phosphor Bronze Alloys, A.S.T.M. Special Tech. Pub., STP 183, pp. 1-114, Jan., 1956.

GUERARD, J. P., see Gohn, G. R.

HANNAY, N. B.¹

Recent Advances in Silicon — Progress in Semiconductors, Book, 1, pp. 1-35, 1956. (Published by Heywood & Co., Ltd., London)

¹ Bell Telephone Laboratories, Inc.

⁸ Riverside Metal Co., Div., H. K. Porter Co., Inc., Riverside, N. J.

HOLDEN, A. N.,¹ MATTHIAS, B. T.,¹ ANDERSON, P. W.,¹ and LEWIS, H. W.¹

New Low-Temperature Ferromagnets, Phys. Rev., **102**, p. 1463, June 15, 1956.

HUNTLEY, H. R.²

The Present and Future of Telephone Transmission, Elec. Engg., **75**, pp. 686-692, Aug., 1956.

JAMES, D. B., see Gianola, U. F.

JONES, H. L.³

A Blend of Operations Research and Quality Control in Balancing Loads on Telephone Equipment, Trans. Am. Soc. Quality Control (1956 Montreal Convention).

KAMINOW, I. P., see Kircher, R. J.

KIRCHER, R. J.¹ and KAMINOW, I. P.¹

Super-Regenerative Transistor Oscillator, Electronics, **29**, pp. 166-167, July, 1956.

KRETZMER, E. R.¹

Reduced-Alphabet Representation of TV Signals, I.R.E. Convention Record, **4**, Part 4, pp. 140-147, 1956.

KOONCE, S. ELOISE, see Arnold, S. M.

LEE, C. Y.,¹ and CHEN, W. H.⁴

Several-Valued Combinational Switching Circuits, Commun. and Electronics, **25**, pp. 278-283, July, 1956.

LEPOUTRE, G., see Dewald, J. F.

LEWIS, H. W.¹

Two-Fluid Model of an "Energy-Gap" Superconductor, Phys. Rev., **102**, pp. 1508-1511, June 15, 1956.

LEWIS, P. W., see Holden, A. N.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

⁴ University of Florida, Gainesville, Fla.

³ Illinois Bell Telephone Company, Chicago, Ill.

MANLEY, J. M.,¹ and ROWE, H. E.¹

Some General Properties of Non-Linear Elements. Part 1 — General Energy Relations, Proc. I.R.E., 44, pp. 904–913, July, 1956.

MATTHIAS, B. T., see Holden, A. N.; Wood, E. A.

MAXWELL, J. L.,⁶ and FARRAR, H. K.⁶

Automatic Dispatch System for Teletypewriter Lines, Elec. Engg., 75, p. 705, Aug., 1956.

MCDONALD, H. S., see David E. E.

MCLEAN, D. A.¹ and POWER, F. S.¹

Tantalum Solid Electrolytic Capacitors, Proc. I.R.E., 44, pp. 872–878, July, 1956.

McMAHON, W.¹

Dielectric Effects Produced by Solidifying Certain Organic Compounds in Electric or Magnetic Fields, J. Am. Chem. Soc., 78, pp. 3290–3294, July 20, 1956.

MERZ, W. J.¹

Effect of Hydrostatic Pressure on the Hysteresis Loop of Guanidine Aluminum Sulfate Hexahydrate, Phys. Rev., 103, pp. 565–566, Aug. 1, 1956.

MERZ, W. J.¹

Switching Time in Ferroelectric BaTiO₃ and Its Dependence on Crystal Thickness, J. Appl. Phys., 27, pp. 938–943, Aug. 1, 1956.

NELSON, L. S.¹

Windowed Dewar Vessels for Use at Low Temperatures, Rev. Sci. Instr., 27, pp. 655–656, Aug., 1956.

NOYES, J. W.,⁵ GAUDET, G.,⁵ and BONNEVILLE, S.⁵

Development of Transcontinental Communications in Canada, Commun. and Electronics, 25, pp. 342–352, July, 1956.

¹ Bell Telephone Laboratories, Inc.

⁵ Bell Telephone Company of Canada, Ltd., Montreal, Que., Canada.

⁶ Pacific Telephone and Telegraph Co., San Francisco, Calif.

PILLIOD, J. J.²

Clinton R. Hanna 1955 Lamme Medalist—History of the Metal, Elec. Engg., 75, p. 706, Aug., 1956.

POWER, F. S., see McLean, D. A.

PRINCE, M. B., see Veloric, H. S.

RINEY, T. D.¹

On the Coefficients in Asymptotic Factorial Expansions, Proc. of Am. Math. Soc., 7, pp. 245–249, Apr., 1956.

ROWE, H. E., see Manley, J. M.

SHULMAN, R. G.¹

Hole Trapping in Germanium Bombarded by High-Energy Electrons, Phys. Rev., 102, pp. 1451–1455, June 15, 1956.

SHULMAN, R. G.¹ and WYLUDA, B. J.¹

Copper in Germanium; Recombination Center and Trapping Center, Phys. Rev., 102, pp. 1455–1457, June 15, 1956.

SLICHTER, W. P.¹

On the Morphology of Highly Crystalline Polyethylenes, J. Poly. Sci., 21, pp. 141–143, July, 1956.

TIEN, P. K.¹

A Dip in the Minimum Noise Figure of Beam-Type Microwave Amplifiers, Proc. I.R.E., Correspondence Sec., 44, p. 938, July, 1956.

VELORIC, H. S.¹ EDER, M. J.¹ and PRINCE, M. B.¹

Avalanche Breakdown in Silicon Diffused P-N Junctions as a Function of Impurity Gradient, J. Appl. Phys., 27, pp. 895–899, August, 1956.

WALSH, DOROTHY E., see Bozorth, R. M.

WASILIK, J. H., see Cook, R. K.

¹ Bell Telephone Laboratories, Inc.

² American Telephone and Telegraph Company.

WERNICK, J. H.¹

Determination of Diffusivities in Liquid Metals by Means of Temperature-Gradient Zone-Melting, *J. Chem. Phys.*, 25, pp. 47-49, July, 1956.

WILKINSON, R. I.¹

Beginnings of Switching Theory in the United States, *Elec. Engg.*, 75, pp. 796-802, Sept., 1956.

WILLIAMS, D. E., see Embree, M. L.

WILLIAMS, H. J., see Bozorth, R. M.

WOOD, MRS. E. A.¹

Guanidinium Aluminum Sulfate Hexahydrate; Crystallographic Data, *Acta Crys.*, 9, pp. 618-619, July 10, 1956.

WOOD, MRS. E. A.¹

The Question of a Phase Transition in Silicon, *J. Phys. Chem.*, 60, p. 508, 1956.

WOOD, MRS. E. A.,¹ and MATTHIAS, B. T.¹

Crystal Structures of Nb₃Au and V₃Au, *Acta Crys.*, 9, pp. 534, June 10, 1956.

WOOD, E. A., see Geller, S.

WYLUDA, B. J., see Shulman, R. G.

¹ Bell Telephone Laboratories Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal

ALBRECHT, E. G., see Bullard, W. R.

ANDERSON, P. W.

Ordering and Antiferromagnetism in Ferrites, Monograph 2636.

BAKER, W. O., see Winslow, F. H.

BENNETT, W. R., see Pierce, J. R.

BOGERT, B. P.

The Vobanc — A Two-to-One Speech Bandwidth Reduction System, Monograph 2643.

BÖMMEL, H. E., MASON, W. P., and WARNER, A. W.

Dislocations, Relaxations, and Anelasticity of Crystal Quartz, Monograph 2618.

BOYET, H., see Weisbaum, S.

BULLARD, W. R., WEPPLER, H. E., ALBRECHT, E. G., DIETZ, A. E., CHRISTOFERSON, E. W., SLOTHOWER, J. E., ELLIS, H. M., PHELPS, J. W., ROACH, C. L., and TREEN, R. E.

Co-Ordinated Protection for Open-Wire Joint Use — Trends and Tests, Monograph 2662.

CHRISTOFERSON, E. W., see Bullard, W. R.

CHYNOWETH, A. G.

Spontaneous Polarization of Guanidine Aluminum Sulfate Hexahydrate at Low Temperatures, Monograph 2645.

* 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.

CHYNOWETH, A. G.

Surface Space-Charge Layers in Barium Titanate, Monograph 2628.

CHYNOWETH, A. G., and MCKAY, K. G.

Photon Emission from Avalanche Breakdown in Silicon, Monograph 2619.

DACEY, G. C., see Thomas, D. E.

DANIELSON, W. E., ROSENFELD, J. L., and SALOOM, J. A.

Analysis of Beam Formation with Electron Guns of the Pierce Type, Monograph 2609.

DARLINGTON, S.

A Survey of Network Realization Techniques, Monograph 2620.

DIETZ, A. E., see Bullard, W. R.

DITZENBERGER, J. A., see Fuller, C. S.

DUDLEY, H. W.

Fundamentals of Speech Synthesis, Monograph 2648.

ELLIS, H. M., see Bullard, W. R.

FULLER, C. S., and DITZENBERGER, J. A.

Diffusion of Donor and Acceptor Elements in Silicon, Monograph 2651.

GIANOLA, U. F., and JAMES, D. B.

Ferromagnetic Coupling Between Crossed Coils, Monograph 2653.

HARROWER, G. A.

Auger Electrons in Energy Spectra of Secondary Electrons from Mo and W, Monograph 2621.

HEIDENREICH, R. D.

Thermionic Emission Microscopy of Metals, Monograph 2445.

HOLDEN, A. N., MERZ, W. J., REMEIKA, J. P., and MATTHIAS, B. T.

Properties of Guanidine Aluminum Sulfate Hexahydrate and Some of Its Isomorphs, Monograph 2580.

HUTSON, A. R.

Effect of Water Vapor on Germanium Surface Potential, Monograph 2623.

JAMES, D. B., see Gianola, U. F.

KAMINOW, I. P., see Kircher, R. J.

KATZ, D.

A Magnetic Amplifier Switching Matrix, Monograph 2654.

KELLY, M. J.

The Record of Profitable Research at Bell Telephone Laboratories,
Monograph 2663.

KIRCHNER, R. J., and KAMINOW, I. P.

Superregenerative Transistor Oscillator, Monograph 2664.

LOGAN, R. A., see Thurmond, C. D.

MASON, W. P., see Bömmel, H. E.

MATTHIAS, B. T., see Holden, A. N.

MCKAY, K. G., see Chynoweth, A. G.

McSKIMIN, H. J.

Propagation of Longitudinal and Shear Waves in Rods at High Frequencies, Monograph 2637.

MERZ, W. J., see Holden, A. N.

PEARSON, G. L.

Electricity from the Sun, Monograph 2658.

PHELPS, J. W.

Protection Problems on Telephone Distribution Systems, Monograph 2631.

PHELPS, J. W., see Bullard, W. R.

PIERCE, J. R., and BENNETT, W. R.

Noise — Physical Sources; and Methods of Solving Problems, Monograph 2624.

PRINCE, E.

Neutron Diffraction Observation of Heat Treatment in Cobalt Ferrite, Monograph 2632.

REISS, H.

P-N Junction Theory by the Method of δ -Functions, Monograph 2638

REMEIKA, J. P., see Holden, A. N.

RICE, S. O.

A First Look at Random Noise, Monograph 2659.

ROACH, C. L., see Bullard, W. R.

ROSENFELD, J. L., see Danielson, W. E.

SALOOM, J. A., see Danielson, W. E.

SLOTHOWER, J. E., see Bullard, W. R.

THEUERER, H. C.

Purification of Germanium Tetrachloride by Extraction with Hydrochloric Acid and Chlorine, Monograph 2639.

THOMAS, D. E., and DACEY, G. C.

Application Aspects of Germanium Diffused Base Transistor, Monograph 2660.

THURMOND, C. D., and LOGAN, R. A.

Copper Distribution Between Germanium and Ternary Melts Saturated with Germanium, Monograph 2640.

TREEN, R. E., See Bullard, W. R.

WARNER, A. W., see Bömmel, H. E.

WEISBAUM, S., and BOYET, H.

Broadband Nonreciprocal Phase Shifts — Two Ferrite Slabs in Rectangular Guide, Monograph 2642.

WEPPLER, H. E., see Bullard, W. R.

WINSLOW, F. H., BAKER, W. O., and YAGER W. A.

The Structure and Properties of Some Pyrolyzed Polymers, Monograph 2572.

YAGER, W. A., see Winslow, F. H.

Contributors to This Issue

C. F. EDWARDS, B.A. 1929 and M.A. 1930, Ohio State University; A. T. & T. Co. 1930-34; Bell Telephone Laboratories, 1935-. Research in transoceanic short wave transmission, transoceanic short wave transmission using multiple unit steerable antenna receiving system, waveguide circuit design, frequency converters for microwave radio relay systems and time division multiplex telephone system. Author of articles published in I.R.E. Proceedings. Member of I.R.E.

JOSEPH P. LAICO, M.E., Brooklyn Polytechnic Institute, 1933; General Drafting Company, 1920-23; American Machine and Foundry Company, 1923-29; Bell Telephone Laboratories, 1929-. Supervision in the field of mechanical design and development of electronic devices is Mr. Laico's occupation at the Laboratories. He holds some twenty patents, all in electronic devices, and is a member of Tau Beta Pi.

E. J. McCLUSKEY, JR., A.B., 1953, Bowdoin College, B.S. and M.S. 1953 and Sc.D. 1956, M.I.T.; Bell Telephone Laboratories, co-operative student, 1950-52; M.I.T. research assistant and instructor, 1953-55; Bell Telephone Laboratories, 1955-. Research in connection with electronic switching systems. Non-resident instructor at M.I.T., summer 1956. Lecturer at C.C.N.Y., 1956. Member of I.R.E., Phi Beta Kappa, Tau Beta Pi, Eta Kappa Nu and Sigma Xi.

HUNTER L. McDOWELL, B.E.E., Cornell University, 1948; Bell Telephone Laboratories, 1948-. At the Laboratories, Mr. McDowell has been principally engaged in vacuum tube development, particularly traveling wave amplifiers. He is a member of I.R.E.

SAMUEL P. MORGAN, B.S. 1943, M.S. 1944 and Ph.D. 1947, California Institute of Technology; Bell Telephone Laboratories, 1947-. A research mathematician, Dr. Morgan specializes in electromagnetic theory. Studies in problems of waveguide and coaxial cable transmission and microwave antenna theory. Member of the American Physical Society, Tau Beta Pi, Sigma Xi and I.R.E.

CLARENCE R. MOSTER, B.E.E., Alabama Polytechnic Institute, 1942; S.M., Massachusetts Institute of Technology, 1947; Naval Research Laboratory, 1942-45; Bell Telephone Laboratories, 1947-. Mr. Moster's main work at the Laboratories has been in vacuum tube development, specializing in microwave tubes. Member of Institute of Radio Engineers, Sigma Xi, Eta Kappa Nu and Phi Kappa Phi.

W. T. READ, JR., B.S. 1944, Rutgers and M.S. 1948, Brown University; National Defense Research Committee, 1943-46; Engaged in air-blast and earth-shock tests at Princeton University Station and measurements of air blast at Bikini atom bomb tests; Bell Telephone Laboratories, 1947-. Photoelastic and mathematical stress analysis. Dislocation theory and problems of plastic deformation were early studies. Later involved with theory of flow and space charge of holes and electrons and with electrical and mechanical effects of dislocations and other imperfections in semiconductors. Author of "Dislocations in Crystals," McGraw-Hill, 1953. Member of Phi Beta Kappa.

WILLIAM MERLIN SHARPLESS, B.S. in E.E. 1928 and Professional Engineering in E.E. 1951, University of Minnesota; Bell Telephone Laboratories, 1928-. Studies of optical behaviors of the ground for short radio waves, artificial ground systems for short wave reception, angle of arrival of transatlantic short wave signals, multiple unit steerable antenna system, microwave radio circuits, noise factors in microwave silicon rectifiers, broad band balanced and unbalanced crystal converters, radar, propagation of microwaves over land paths, angle of arrival of microwaves, and antenna systems and artificial dielectrics for microwaves. Several patents. Published papers on short radio waves and microwaves. Member of American Physical Society and Scientific Research Society of America. Senior member of I.R.E.

JAMES A. YOUNG, JR., B.S. 1943, California Institute of Technology; Radio Officer, U. S. Army Signal Corps, 1943-1946; Jet Propulsion Laboratory of California Institute of Technology, 1946-1947; Ph.D. 1953, University of Washington; Bell Telephone Laboratories, 1953-. Concerned primarily with low loss circular electric mode waveguide. Member of American Physical Society, Sigma Xi and I.R.E.

THE BELL SYSTEM

Technical Journal

D VOTED TO THE SCIENTIFIC AND ENGINEERING
PECTS OF ELECTRICAL COMMUNICATION

ADVISORY BOARD

A. B. GOETZE

M. J. KELLY

E. J. MCNEELY

EDITORIAL COMMITTEE

B. McMILLAN, *Chairman*

S. E. BRILLHART

E. I. GREEN

A. J. BUSCH

R. K. HONAMAN

L. R. COOK

H. R. HUNTLEY

A. C. DICKIESON

F. R. LACK

R. L. DIETZOLD

J. R. PIERCE

K. E. GOULD

G. N. THAYER

EDITORIAL STAFF

J. D. TEBO, *Editor*

R. L. SHEPHERD, *Production Editor*

KANSAS CITY, MO.
PUBLIC LIBRARY

INDEX

VOLUME XXXV

FEB 13 1957

1956

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

LIST OF ISSUES IN VOLUME XXXV

No.	Month	Pages
1	January	1-248
" 2	March	249-534
" 3	May	535-766
" 4	July	767-990
" 5	September	991-1238
" 6	November	i-iv, 1239-1466

Index to Volume XXXV

A

AM *See* Amplitude Modulation

Adam, Armand O.

biographical material 531

Crossbar Tandem as a Long Distance Switching System 91-108

Adda, L. P.

zone leveler

development 660

ADMINISTRATION EQUIPMENT

translator

magnetic drum 741-44; *illus* 740

block diagram 742

ADMITTANCE

nonlinear

frequency

conversion 1403-16

AKRON, OHIO

toll traffic graph 429

ALBANY, NEW YORK

toll traffic graphs 427-28

ALGEBRA *See* Boolean Algebra

ALLOY

silicon

diode

announcement 661

ALLOY JUNCTION TRANSISTOR *See* Transistor: junction

ALPHABET

signaling

binary

group 203-34

best 212-15

defined 207

properties 204-19

special features 203

ALTERNATE ROUTING *See* Routing

AMERICAN TELEPHONE AND TELEGRAPH COMPANY

functions, primary 422-23

operating companies, *see* Operating Companies

AMPLIFIER

feedback, negative, design 296-308

pulse

regenerative

described 1085

transistorized 1085-1114

reliability 1085-86

signal

binary

transistorized 1059-84

summing 308-13

transistor

junction

tetrode

design 813-40

traveling wave *See* Electron Tube

AMPLITUDE MODULATION

electron tube

traveling wave

M1789 1321-22

ANALOG SYSTEMS

transistor

junction

applications 295-332

Anderson, H. W.

antenna

parabolic

design 1208

rectifier

wave

millimeter

wafer-type 1397

Angle, R.

electron tube

traveling wave

M1789 1343

ANTENNA

microwave

testing

pulses, millimicrosecond 45-48

parabolic

60-foot diameter 1199-1208; *illus*

- ANTENNA, continued**
- pulse
 - millimicrosecond 45-48
- APPARATUS**
- reliability studies
 - experiment time
 - reduction, by statistical techniques 179-202
- AQUEOUS SOLUTIONS**
- semiconductors, analogy 537
- ATLANTIC TELEPHONE CABLE** *See* Transatlantic Telephone Cable
- ATTENUATION**
- atmospheric
 - wavelengths
 - millimeter
 - measurement
 - radar 907-16
 - slope
 - unit, semi-infinite
 - phase, tables 747-49
- wave**
- circular
 - pipes
 - medium-sized
 - 5-6 mm 1115-28
 - small
 - 5-6 mm 1115-28
- waveguide**
- helix 1358
- ATTENUATOR**
- coupled helix 165-67
- Automatic Machine for Testing Capacitors and Resistance-Capacitance Networks* (C. C. Cole, H. R. Shillington) 1179-98
- Automatic Manufacturing Testing of Relay Switching Circuits* (L. D. Hansen) 1155-78
- Automatic Testing in Telephone Manufacture* (D. T. Robb) 1129-54
- Automatic Testing of Transmission and Operational Functions of Intertoll Trunks* (H. H. Felder, A. J. Pasarella, H. F. Shoffstall) 927-54
- B**
- Babcock, Wallace C.
- biographical material 531
 - Crosstalk on Open-Wire Lines* 515-18
- Bardeen, John *illus* ii
- biographical material iii-iv
 - Nobel Prize in Physics, 1956 i-iv
- BASE**
- diffused
 - high-frequency
 - transistor
 - junction
 - p-n-p
 - germanium 23-34
- BEAM** *See* Electron Beam
- Beaton, Daniel
- antenna
 - parabolic
 - design 1208
- Beek, A. C.
- biographical material 244
 - Waveguide Investigations with Millimicrosecond Pulses* 35-65
- Bell, J. W.
- wave
 - electric
 - circular
 - attenuation 1128
- BELL LABORATORIES TYPE M1789 TUBE**
- See* Electron Tube: traveling wave
- BELL SYSTEM**
- intertoll trunks 423
 - outside plant, *see* Outside Plant Department
- BELL SYSTEM TECHNICAL JOURNAL**
- advisory board, *see* inside front cover
 - editorial committee, *see* inside front cover
 - editorial staff, *see* inside front cover
- BELL TELEPHONE LABORATORIES**
- Nobel Prizes in Physics i-iv
- Bennett, A. L.
- testing
 - automatic 1154
- Bennett, Donald C.
- biographical material 762
 - Single Crystals of Exceptional Perfection and Uniformity by Zone Leveling* 637-60
- Bennett, W. R.
- amplifier
 - transistor
 - junction
 - tetrode 840

- regenerator
 pulse
 binary
 transistor 1084
- Bergwall, F. W.
 zone leveler
 development 660
- BIFILAR HELIX** *See* Helix; coupled
- BINARY MICROWAVE PULSE** *See* Pulse
- BINARY PULSE TRANSMISSION** *See* Transmission
- BINARY SIGNALING ALPHABET** *See* Alphabet
- Blecher, Franklin H.
 biographical material 531
Transistor Circuits for Analog and Digital Systems 295-332
- BLOOMSBURG, PENNSYLVANIA**
 automatic alternate routing
 schematic 440
- Bodmer, M. G.
 electron tube
 traveling wave
 M1789 1343
- Bond, W. L.
 biographical material 1233
Use of an Interference Microscope for Measurement of Extremely Thin Surface Layers 1209-21
- BOOLEAN ALGEBRA**
 circuits
 switching
 design 1417
- invariance
 group
 detection 1445-53
- symmetry
 total
 detection 1445-53
- Bosworth, R. H.
 amplifier
 transistor
 junction
 tetrode 840
- Brannen, Miss M. J.
 isolator
 field displacement 896
- Brattain, Walter H. *illus* ii
 biographical material ii-iii, 1233
- Combined Measurements of Field Effect, Surface Photo-Voltage and Photo-Conductivity* 1019-40
- Distribution and Cross-Sections of Fast States on Germanium Surfaces* 1041-58
- Nobel Prize in Physics, 1956 i-iv
- transistor
 point-contact
 experiments 770
- BREAKDOWN VOLTAGE** *See* Voltage
- Brooks, C. E.
 concentrator
 line, remote controlled
 development 293
- BUFFALO, NEW YORK**
 toll traffic graphs 427-28
- Buhrendorf, F. G.
 biographical material 762
Laboratory Model Magnetic Drum Translator for Toll Switching Offices 707-4
- Burke, P. J.
 toll traffic study 506
- C**
- CABLE**
 coaxial
 equalization
 phase, tables
 tabulation 747-49
- transatlantic *See* Transatlantic Telephone Cable
- CAPACITOR**
 nonlinear
 frequency
 conversion 1409-11
- testing
 machine
 automatic 1179-98
- CARD-O-MATIC TEST SET** *See* Test Set
- CARD TRANSLATOR** *See* Translator
- Carthage, N.
 transistor
 point-contact
 surface effects 810
- CENTER** *See* Wire Center

- CENTRAL OFFICE
concentrator
line
remote controlled circuits 274-86
defined 250
- Chapman, A. G.
transposition theory 515
- Chemical Interactions among Defects in Germanium and Silicon* (C. S. Fuller, F. J. Morin, H. Reiss) 535-636
- Cioffi, P. P.
electron tube
traveling wave M1789 1343
- CIRCUIT
concentrator
line
remote controlled 261-70
- switching
design
Boolean algebra 1417
- relay
testing
automatic 1155-78
- transistor
junction
analog systems 295-332
digital systems 295-332
- translator
magnetic drum 725-41
- CIRCULAR WAVE *See* Wave
- Class of Binary Signaling Alphabets* (D. Slepian) 203-34
- Clausen, C. P.
antenna
parabolic
design 1208
- Clos, C.
toll traffic study 431, 470
- COAXIAL CABLE *See* Cable
- COIL
relay
U-Type
testing
automatic 1141-48
- UA-type
testing
automatic 1141-48
- Y-type
testing
automatic 1141-48
- Cole, C. C.
Automatic Machine for Testing Capacitors and Resistance-Capacitance Networks 1179-98
- biographieal material 1233
- COLLECTOR
electron tube
traveling wave M1789 1311-13
- Combined Measurements of Field Effect, Surface Photo-Voltage and Photo-Conductivity* (W. H. Brattain, C. G. B. Garrett) 1019-40
- COMPARATOR
voltage 320-27
- COMPLEX ION *See* Ion
- COMPONENT(S)
reliability studies
experiment time
reduction, by statistical techniques 179-202
- COMPUTER *See* Analog Systems; Digital Systems
- CONCENTRATOR
line
remote controlled
experimental 249-93
illus 277, 287, 289
circuits 261-70
central office 274-86
economy 249-93
field trials 286-93
operation 270-74
power supply 269-70
relay
reed switch 252-53
traffic loading 249-93
- CONDUCTIVITY
modulation
rectifier
series resistance 666-70
See also Photoconductivity
- CONTACT
transistor
point-contact
formed 770-83
unformed 783-96

- CONVERSION**
- frequency
 - admittance
 - noulinear 1403-16
- COOK, J. S.**
- biographical material 244
 - Coupled Helices* 127-78
- COONCE, H. E.**
- amplifier
 - pulse
 - regenerative
 - transistor 1114
- COPPER**
- plating
 - transistor
 - point-contact
 - surface 776-81
- COPPER OXIDE RECTIFIER** *See Rectifier*
- COST**
- concentrator, line 249-93
 - drums, magnetic 707
 - outside plant 249-93
 - switching 249-93
 - See also Economy*
- Coupled Helices** (J. S. Cook, R. Komppner, C. F. Quate) 127-78
- COUPLED HELIX** *See Helix*
- COUPLED HELIX ATTENUATOR** *See Attenuator*
- COUPLED HELIX TRANSDUCER** *See Transduseer*
- COUPLER**
- stepped
 - helices, coupled 158-59
 - tapered
 - helices, coupled 157-58
- Crawford, Arthur B.**
- biographical material 985, 1234
 - Measurement of Atmospheric Attenuation at Millimeter Wavelengths* 907-16
 - rectifier
 - wave
 - millimeter
 - wafer-type 1397
 - 60-Foot Diameter Parabolic Antenna for Propagation Studies* 1199-1208
- CROSSBAR SYSTEMS**
- 4-type
 - development 423
 - 5-type
 - concentrator
 - line
 - remote 251-93
 - switching plan 257-61
- tandem**
- switching system, long distance
 - major toll switching features 91
- See also Switching Systems*
- Crossbar Tandem as a Long Distance Switching System* (A. O. Adam) 91-108
- CROSSTALK**
- coupling
 - types 515
 - measurement 516
- Crosstalk on Open-Wire Lines* (W. C. Babcock, Miss E. Rentrop, C. S. Thaeler) 515-18
- CRYSTAL**
- defects
 - interaction 535-636
 - diffusion *See Diffusion*
 - germanium
 - acceptor content
 - zone leveling 638-60
 - defects
 - interactions, chemical 535-636
 - donor content
 - zone-leveling 638-60
 - etched
 - field effect 1019-40
 - measurements
 - combined 1019-40
 - photoconductivity 1019-40
 - photo-voltage
 - surface 1019-40
 - semiconductor applications
 - requirements 641-55
 - shaping
 - electrolytic 333-47
 - surface
 - fast states
 - cross-sections 1041-58
 - distribution 1041-58
 - transistor forming, relation 796-SOS
 - testing 642-43
 - lattice, *see Lattice*

CRYSTAL, continued

silicon

defects

interactions, chemical 535-636

diffusion 664-66

rectifiers 661-84

diode, *see* Diode

shaping

electrolytic 333-47

shaping

electrolytic 333-47

CUSTOMER DIRECT DISTANCE DIALING

See Dial Telephone: nationwide

Cutler, C. Chapin

biographical material 985

electron beam formation, theory 375

generator, pulse, regenerative, development 36

Nature of Power Saturation in Traveling Wave Tubes 841-76

D

Danielson, W. E.

biographical material 531

Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type 375-420DATA SYSTEM *See* Digital Systems

Davisson, Clinton J.

biographical material iii

Nobel Prize in Physics, 1937 i, iii

DEFECT

crystal

interaction 535-636

DeLange, O. E.

biographical material 244

Experiments on the Regeneration of Binary Microwave Pulses 67-90

DELAY

distortion

phase

tables

tabulation 747-49

DELAY DISTORTION REPEATER *See* Repeater

DESIGN

amplifier

transistor

junction

tetrode 813-40

circuit

switching

Boolean algebra 1417

electron gun

Pierce-type 378-79, 399, 402-13
418-20

electron tube

traveling wave 867-68

M1789 1289-91

isolator

field displacement 884-91

relays 991

switching systems

electronics in 991-1018

transistor

junction

n-p-n

base, diffused 14-21

emitter, diffused 14-21

point-contact 769

Design of Tetrode Transistor Amplifier
(J. G. Linvill, L. G. Schimpf) 813-40*Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type*
(W. E. Danielson, J. L. Rosenfeld,
J. A. Saloom) 375-420*Detection of Group Invariance or Total Symmetry of a Boolean Function* (E.
J. McCluskey, Jr.) 1445-53

DeVido, R. W.

electron tube

traveling wave

M1789 1343

DIAL TELEPHONE, DIALING

crosstalk, *see* Crosstalk

direct distance 955-72

crossbar tandem systems 107-08

lines, *see* Transmission Lines

nationwide

aspects, general 93-94

crossbar tandem switching system
91-108

customer direct

crossbar tandem systems 107-08

expansion 423

routing, *see* Routing

service requirements 436-37

translator

card 716-19

- magnetic drum 707-45
 trunks
 intertoll
 testing, automatic 927-54
 operator distance 955-72
 United States statistics 423
 testing, automatic 1129-54
 traffic, *see* Traffic
 transmission lines, *see* Transmission Lines
 Dickten, E.
 amplifier
 transistor
 junction
 tetrode 840
DIELECTRICS
 helices, coupled, between 148-50
 Dietzold, R. L.
 phase, tables
 computation 749
Diffused Emitter and Base Silicon Transistors (M. Tannenbaum, D. E. Thomas) 1-22
DIFFUSED JUNCTION SILICON DIODE *See* Diode
Diffused p-n Junction Silicon Rectifiers (M. B. Prince) 661-84
DIFFUSION
 crystal
 silicon 664-66
 rectifiers 661-85
DIGITAL SYSTEMS
 drums, magnetic 707-45
 transistor
 junction
 applications 295-332
DIGITAL TRANSMISSION *See* Transmission
DIODE
 junction
 germanium
 large area
 announcement 661
 temperature 661
 silicon
 diffused
 current-voltage characteristic equations, basic 688-706
 forward 685-706
 silicon alloy
 announcement 661
 temperature 661
 PIN *See* Diode; junction; silicon; diffused
 voltage
 breakdown 685
DIODE RECTIFIER *See* Rectifier
DIRECT DISTANCE DIALING *See* Dial Telephone
DISTORTION
 delay
 phase
 tables
 tabulation 747-49
Distribution and Cross-Sections of Fast States on Germanium Surfaces (W. H. Brattain, C. G. B. Garrett) 1041-58
DOMINANT MODE WAVEGUIDE *See* Waveguide
DRUM
 magnetic *illus* 1007
 access time 707
 applications 707, 745
 digital-data storage 707
 features 709-16
 geography 712
 memory units 707
 reading 713-16
 speed 107
 switching
 toll
 translator 707-45
 writing 712-13
E
ECONOMY
 concentrator, line 249-93
 outside plant 249-93
 switching 249-93
 See also Cost
 Edwards, C. F.
 biographical material 1465
Frequency Conversion by Means of a Nonlinear Admittance 1403-16
Effect of Surface Treatments on Point-Contact Transistors (J. H. Forster, L. E. Miller) 767-811

- Elbert, E. F.
- rectifier
- wave
 - millimeter
 - wafer-type 1397
- ELECTRIC FIELD
- helices, coupled 139-42, 144-46
 - junction
 - PIN
 - bias
 - reversed 1239-84
- ELECTROLYTIC ETCHING *See* Etching
- Electrolytic Shaping of Germanium and Silicon* (A. Uhlig, Jr.) 333-47
- ELECTRON(S) AND HOLES
- chemical entities 537-46
 - interaction 537-57
 - junction
 - PIN
 - bias
 - reversed 1239-84
- ELECTRON BEAM
- amplifier
 - traveling wave
 - coupling, finite, effects 349-74
 - space charge, effects 349-74
 - wave
 - backward 351-55
 - forward 351-55
 - electric field 859-63
 - electron tube
 - traveling wave
 - M1789 1298-1303
 - formation
 - electron gun
 - Pierce-type 375-420
 - size
 - electron tube
 - traveling wave 856-58
 - spent
 - electron tube
 - traveling wave 846-54
 - spreading 388-401
 - waves
 - growing, due to transverse velocities 109-25
- ELECTRON FLOW
- waves
 - growing, due to transverse velocities 109-25
- ELECTRON GUN
- electron tube
- traveling wave
 - M1789 1298-1303; *illus* 1296-97
- Pierce-type
- anode lens 379-88
 - beam
 - current densities 413-16
 - formation 375-420
 - spreading 388-401
 - design 378-79, 399, 402-13, 418-20
 - development 377
- ELECTRON TUBE
- gun, *see* Electron Gun
 - microwave
 - electron gun, *see* Electron Gun
 - traveling wave
 - M1789 1285-1346
 - illus* 1286, 1292-93
 - AM to PM conversion 1321-22
 - collector 1311-13
 - description 1291-1313
 - design 1289-91
 - electron beam 1298-1303
 - electron gun 1298-1303; *illus* 1296-97
 - gain calculations 1343-44
 - helix 1303-11
 - intermodulation 1335-42
 - life expectancy 1342-43
 - noise 1328-35
 - performance 1313-42
 - relay systems
 - radio 1285-1346
 - amplifier
 - equations 355-59
 - non-linear behavior 349-74
 - signal, large, theory 349-74
 - applications 1285
 - circuit elements 129-30
 - design 867-68
 - dispersive
 - helices, coupled 159-61
 - efficiency 841
 - measurements 844-46
 - electron beam
 - spent 846-54
 - helices, coupled 127-78
 - operating characteristics
 - non-linear 841-76

- power saturation 841-76
 research 1285
 space charge 854-56
See also Amplifier
- Electronics in Telephone Switching Systems* (A. E. Joel, Jr.) 991-1018
- ELECTROPLATING
- transistor
- point contact
 - surface
 - copper 776-81
- EMITTER
- transistor
- junction
 - n-p-n
 - diffused 1-22
- ENCODER
- voltage
- transistor 327-29
- EQUALIZATION
- cable
- coaxial
 - phase, tables
 - tabulation 747-49
- delay distortion
- pulses, millimicrosecond 54-57
- EQUIPMENT
- administration, *see* Administration
 - Equipment
 - reliability studies
 - experiment time
 - reduction, by statistical techniques 179-202
- Erhart, D. L.
- zone leveler
 - development 660
- ETCHING
- electrolytic
- crystal
 - germanium 333-47
 - silicon 333-47
- EXPERIMENT TIME
- reliability studies
- reduction
 - statistical methods 179-202
- Experimental Remote Controlled Line Concentrator* (A. E. Joel, Jr.) 249-93
- Experiments on the Regeneration of Binary Microwave Pulses* (O. E. DeLange) 67-90
- F**
- 4-TYPE CROSSBAR SYSTEM *See* Crossbar Systems
- 4A TOIL SWITCHING SYSTEM *See* Switching Systems
- 5-TYPE CROSSBAR SYSTEMS *See* Crossbar Systems
- 56A OSCILLATOR *See* Oscillator
- 425B NETWORK *See* Network
- FABRICATION
- transistor
- junction
 - n-p-n
 - silicon
 - base, diffused 2-6
 - emitter, diffused 2-6
 - p-n-p
 - germanium
 - base, diffused 23-24
- FEEDBACK AMPLIFIER *See* Amplifier
- Felder, Harry H.
- Automatic Testing of Transmission and Operational Functions of Intertoll Trunks* 927-54
 - biographical material 985
 - Intertoll Trunk Net Loss Maintenance under Operator Distance and Direct Distance Dialing* 955-72
- Feldman, C. B.
- regenerator
 - pulse
 - binary
 - transistor 1084
- Felker, J. H.
- amplifier
 - pulse
 - regenerative
 - transistor 1114
- FIELD *See* Electric Field
- Field Displacement Isolator* (H. Siedel, S. Weisbaum) 877-98
- FIELD EFFECT
- germanium
 - etched
 - measurements
 - combined 1019-40

- Finch, T. R.
 amplifier
 pulse
 regenerative
 transistor 1114
 transistor circuit research 329
- Fleming, C. C.
 trunks
 intertoll
 testing
 automatic 954
- FLOW OF ELECTRONS *See* Electron Flow
- Forster, J. H.
 biographical material 985
Effect of Surface Treatments on Point-Contact Transistors 767-811
- Forward Characteristic of the PIN Diode* (D. A. Kleinman) 685-706
- Foster, F. G.
 semiconductors
 defects
 chemical interactions 613
- Fox, A. G.
 ferrite devices
 nonreciprocal 877
- FREQUENCY
 conversion
 admittance
 nonlinear
 mathematical analysis 1403-16
- helices, coupled
 strength of coupling versus frequency 142-44
- microwave
 pulses, binary
 regeneration 67-90
- Frequency Conversion by Means of a Nonlinear Admittance* (C. F. Edwards) 1403-16
- Friis, Harold T.
 biographical material 1234
 rectifier
 wave
 millimeter
 wafer-type 1397
- 60-Foot Diameter Parabolic Antenna for Propagation Studies 1199-1208
- Frisbee, S. E.
 testing machines 1198
- Fuller, Calvin S.
 biographical material 762
Chemical Interactions among Defects in Germanium and Silicon 535-636
- FUNCTION
 Boolean
 minimization 1417-44
 prime implicants 1419-40
 sum, minimum 1418-19
 writing
 products, sum of 1417-44
- transmission
 Boolean
 invariance
 group
 detection 1445-53
 symmetry
 total
 detection 1445-53
- G**
- Garrett, C. G. B.
 biographical material 1234
Combined Measurements of Field Effect, Surface Photo-Voltage and Photo-Conductivity 1019-40
- Distribution and Cross-Sections of Fast States on Germanium Surfaces* 1041-58
- Gellatly, J. S.
 electron tube
 traveling wave
 M1789 1343
- GENERATOR
 pulse, regenerative
 block diagram 37
 development 36-38
See also Regenerator
- GERMANIUM
 crystal, *see* Crystal
 defects
 interactions, chemical 535-636
 diode, *see* Diode
 zone leveling 638-68
 apparatus 655-60
 technique 655-60
 zone refining 637
- GERMANIUM P-N-P TRANSISTOR *See* Transistor: junction
- Germer, L. H.
 electron diffusion studies i

- Gibney, R. B.
semiconductor studies i
- Glass, M. S.
electron tube
traveling wave
M1789 1343
- Glezer, L. L.
trunks
intertoll
testing
automatic 954
- Goeltz, Miss J. D.
phase, tables
tabulation 749
- Graham, R. E.
information rate
interpretation 926
- Grant, D. W.
magnetotor, construction 329
- Gray, Miss M. C.
semiconductors
defects
chemical interactions 613
- Grossman, A. J.
amplifier
pulse
regenerative
transistor 1114
- GROUP ALPHABET *See* Alphabet
- Growing Waves due to Transverse Velocities (J. R. Pierce, L. R. Walker) 109-25
- GUN *See* Electron Gun
- H**
- Hall, W. J.
toll traffic study 506
- Hamming, R. W.
phase, tables
tabulation 749
- Hannay, N. B.
semiconductors
defects
chemical interactions 613
- Hansen, L. D.
Automatic Manufacturing Testing of Relay Switching Circuits 1155-78
biographical material 1235
- Harris, J. R.
amplifier
pulse
regenerative
transistor 1114
- Harris, W. B.
magnetotor, construction 329
- Hayward, W. S.
toll traffic study 506
- Heilos, L. J.
electron tube
traveling wave
power saturation 867
- HELIX
coupled 127-78
applications, Bell System 154-67
attenuator 165-67
bifilar
dispersion 146-48
coupler
stepped 158-59
tapered 157-58
dielectrics between, effect 148-50
field equations 169-73
fields 139-42, 144-46
power transfer, maximum 151-52
solutions, non-synchronous 137-39
strength of coupling versus frequency 142-44
transducer 161-65
transmission line equations 133-37
- electron tube
traveling wave
M1789 1303-11
- HELIX TRANSDUCER *See* Transducer
- Helix Waveguide (S. B. Morgan, J. A. Young) 1347-84
- Henning, H. A.
biographical material 762
- Laboratory Model Magnetic Drum Translator for Toll Switching Offices 707-45
- Herbert, N. J.
transistor
point-contact
surface effects 810
- HETERODYNE CONVERSION TRANSDUCER
See Transducer
- High-Frequency Diffused Base Germanium Transistor (C. A. Lee) 23-34

- Hines, M. E.
electron beam formation, theory 375
- Hogg, David C.
biographical material 986
- Measurement of Atmospheric Attenuation at Millimeter Wavelengths* 907-16
- HOLDER
rectifier
wave
millimeter 1385
- HOLE(S) *See Electron(s) and Holes*
- Howard, L. F.
trunks
intertoll
testing
automatic 954
- I**
- IMPEDANCE
heliees, coupled
modes 152-54
- IMPURITY
semiconductors
diffusion into 1-34
- INDIANAPOLIS WORKS (Western Electric)
network
425B
testing
automatic 1135-41
- INFORMATION
storage
drums, magnetic 707-45
See also Digital Systems
- INFORMATION RATE
interpretation
new 917-26
- INSULATION *See Dielectric(s)*
- INTEGRATOR
transistor 313-20
- INTERCONNECTING NETWORK *See Network*
- INTERFERENCE *See Crosstalk; Noise*
- INTERFERENCE MICROSCOPE *See Microscope*
- INTERMODULATION
electron tube
traveling wave
M1789 1335-42
- Intertoll Trunk Net Loss Maintenance under Operator Distance and Direct Distance Dialing* (H. H. Felder, E. N. Little) 955-72
- ION(s)
complex
formation 557-65
pairing 565-636
calculations 578-82
carrier
mobility
effect 601-07
(by) diffusion 591-601
energy levels 610-11
relaxation time 582-91, 607-10
semiconductors
phenomena 575-78
solubility, effect 613-17
theories 567-75
- Irwin, J. C.
electron tube
traveling wave
M1789 1343
- [SOLATOR
field displacement 877-98
illus 878, 890
design 884-91
- J**
- Jakes, William C., Jr.
biographical material 1235
60-Foot Diameter Parabolic Antenna for Propagation Studies 1199-1208
- Joel, Amos E., Jr.
biographical material 532, 1235
Electronics in Telephone Switching Systems 991-1018
Experimental Remote Controlled Line Concentrator 249-93
- Johnston, R. L.
rectifier
junction
p-n
silicon
development 684
- Jones, M. S.
transistor
point-contact
surface effects 810

- Jordan, D. R.
 electron tube
 traveling wave
 M1789 1343
- JUNCTION
 NP 1241-42
- PIN
 bias
 reversed
 electrons and holes 1239-84
- JUNCTION DIODE *See* Diode
- JUNCTION SILICON DIODE *See* Diode
- JUNCTION TETRODE TRANSISTOR *See*
 Transistor
- JUNCTION TRANSISTOR *See* Transistor
- K**
- KEARNEY WORKS (Western Electric)
 coil
 relay
 testing
 automatic 1141-48
- Kelly, John L., Jr.
 biographical material 986
New Interpretation of Information Rate
 917-26
- King, Arehie P.
 biographical material 986, 1235
Observed 5-6mm Attenuation for the Circular Electric Wave in Small and Medium-Sized Pipes 1115-28
Transmission Loss Due to Resonance of Loosely-Coupled Modes in a Multi-Mode System 899-906
- Kingsbury, B. A.
 phase, tables
 computation 749
- Kleinman, David A.
 biographical material 763
Forward Characteristic of the PIN Diode 685-706
- Kleinman, D. A.
 rectifier
 junction
 p-n
 silieon
 development 684
- Komper, R.
 biographical material 244
Coupled Helices 127-78
- Kosten, L.
 toll traffic study 431
- L**
- LABORATORIES *See* Bell Telephone Laboratories
- Laboratory Model Magnetic Drum Translator for Toll Switching Offices* (F. J. Buhrendorf, H. A. Henning, O. J. Murphy) 707-45
- Laieu, Joseph P.
 biographical material 1465
Medium Power Traveling-Wave Tube for 6000-Mc Radio Relay 1285
 1346
- Lamont, J.
 testing
 automatie 1154
- Large-Signal Theory of Traveling-Wave Amplifiers* (P. K. Tien) 349-74
- LATTICE
 crystal
 germanium
 zone leveling 638
 perfeetion
 zone leveling 649-55
- Leagus, Miss D. C.
 amplifier, traveling wave
 large signal theory 373
- Lee, Charles A.
 biographical material 245
High-Frequency Diffused Base Germanium Transistor 23-34
- Lennon, Miss C. A.
 toll traffic study 506
- LEVELER *See* Zone Leveler
- LIFE EXPECTANCY
 eletron tube
 traveling wave
 M1789 1342-43
- rectifier
 junction
 p-n
 silieon 680-83
- reliability studies
 experiment time
 reduction, by statistical techniques 179-202
- LINE(s), transmission *See* Transmission Lines

- LINE CONCENTRATOR** *See* Concentrator
- Linville, J. G.**
- biographical material 986
 - Design of Tetrode Transistor Amplifiers* 813-40
- Little, Edward N.**
- biographical material 987
 - Intertoll Trunk Net Loss Maintenance under Operator Distance and Direct Distance Dialing* 955-72
- LOCAL SWITCHING** *See* Switching
- LONG DISTANCE TRAFFIC** *See* Traffic: toll
- M**
- M1789 ELECTRON TUBE** *See* Electron Tube: traveling wave
- McCluskey, E. J., Jr.**
- biographical material 1465
 - Detection of Group Invariance or Total Symmetry of a Boolean Function* 1445-53
 - Minimization of Boolean Functions* 1417-44
- McDowell, Hunter L.**
- biographical material 1465
 - Medium Power Traveling-Wave Tube for 6000-Mc Radio Relay* 1285-1346
- McKim, B.**
- trunks
 - intertoll
 - testing
 - automatic 954
- MAGNETIC DRUM** *See* Drum
- MAGNETIC DRUM TRANSLATOR** *See* Translator
- MAINTENANCE**
- switching systems 1014-16
 - trunks
 - intertoll
 - testing
 - automatic 927-54
- Maita, J. P.**
- semiconductors
 - defects
 - chemical interactions 613
- Mareatili, Enrique A. J.**
- biographical material 987
- Transmission Loss due to Resonance of Loosely-Coupled Modes in a Multi-Mode System* 899-906
- MATRIX**
- Boolean
 - transmission
 - invariance 1445-53
 - symmetry 1445-53
- Mead, Mrs. Sallie P.**
- toll traffic study 506
- Measurement of Atmospheric Attenuation at Millimeter Wavelengths* (A. B. Crawford, D. C. Hogg) 907-16
- Medium Power Traveling-Wave Tube for 6000-Mc Radio Relay* (J. P. Laico, H. L. McDowell, C. R. Moster) 1285-1346
- Melroy, D. O.**
- electron tube
 - traveling wave
 - M1789 1343
- MELTING** *See* Zone Melting
- MIAMI, FLORIDA**
- toll traffic map 439
- MICROSCOPE**
- interference
 - surface layers
 - measurement 1209-21
- MICROWAVE ANTENNA** *See* Antenna
- MICROWAVE MODULATOR** *See* Modulator
- MICROWAVE PULSE** *See* Pulse
- MICROWAVE PULSE REGENERATOR** *See* Regenerator
- MICROWAVE TRANSMISSION** *See* Transmission
- MICROWAVE TUBE** *See* Electron Tube
- MILITARY APPLICATIONS**
- transistor
 - point-contact 768
- Miller, Lewis E.**
- biographical material 987
 - Effect of Surface Treatments on Point-Contact Transistors* 767-811
- Miller, S. E.**
- ferrite devices
 - nonreciprocal 877
- MILLIMETER WAVE** *See* Wave
- MILLIMICROSECOND PULSE** *See* Pulse

- Minimization of Boolean Functions* (E. J. McCluskey, Jr.) 1417-44
- MINNESOTA
intertoll trunk groups, principal map 424
- MODE
loosely-coupled
resonance
transmission
loss 899-906
- spurious
resonance 899-906
- MODULATION *See* Amplitude Modulation; Intermodulation; Phase Modulation
- MODULATOR
microwave
noise temperature requirements 1404
rectifier
point contact 1403-16
- Molina, E. C.
toll traffic study 506
- Moll, J. L.
rectifier
junction
p-n
silicon
development 684
- MONOGRAPHS, recent, of Bell System Technical Papers not published in this Journal 242-43, 527-30, 759-61, 979-84, 1230-32, 1461-64
- Moore, H. R.
rectifier
junction
p-n
silicon
development 684
- Morgan, S. O.
semiconductor studies i
- Morgan, Samuel P.
biographical material 1465
Helix Waveguide 1347-84
- Morin, F. J.
biographical material 763
Chemical Interactions among Defects in Germanium and Silicon 535-636
- Moster, Clarenée R.
biographical material 1467
- Medium Power Traveling-Wave Tube for 6000-Mc Radio Relay 1285-1346
- MULTI-MODE TRANSMISSION SYSTEM.
See Transmission Systems
- Murphy, O. J.
biographical material 763
Laboratory Model Magnetic Drum Translator for Toll Switching Offices 707-45
- N**
- No. 4 CROSSBAR SYSTEM *See* Crossbar System
- No. 4A TOLL SWITCHING SYSTEM *See* Switching Systems
- No. 5 CROSSBAR SYSTEM *See* Crossbar Systems
- No. 56A OSCILLATOR *See* Oscillator
- No. 425B NETWORK *See* Network
- NP JUNCTION *See* Junction
- N-P-N TRANSISTOR *See* Transistor; junction
- NATIONWIDE DIALING *See* Dial Telephone
- NATIONWIDE SWITCHING *See* Switching
- Nature of Power Saturation in Traveling Wave Tubes* (C. C. Cutler) 841-76
- Neely, T. H.
trunks
intertoll
testing
automatic 954
- NEGATIVE FEEDBACK AMPLIFIER *See* Amplifier
- NETWORK
425B
testing
automatic 1135-41
- interconnecting
switching systems 994-98
- resistance-capacitance
testing
machine
automatic 1179-98
- New Interpretation of Information Rate (J. L. Kelly, Jr.) 917-26
- NEW YORK CITY
toll traffic graph 431

- N**EWARK, NEW JERSEY
 toll traffic graph 429
- N**OBEL PRIZE IN PHYSICS
 awards
 1937
 Davisson, Clinton J. i, iii
 Thompson, G. P. iii
 1956
 Bardeen, John i-iv
 Brattain, Walter H. i-iv
 Shockley, William i-iv
- N**OISE
 electron tube
 traveling wave
 M1789 1328-35
- modulator
 microwave
 temperature requirements 1404
- N**ONLINEAR ADMITTANCE *See* Admittance
- N**ONLINEAR CAPACITOR *See* Capacitor
- N**ONRECIPROCAL FERRITE DEVICE *See* Isolator
- N**UMBER 4 CROSSBAR SYSTEM *See* Crossbar System
- N**UMBER 4A TOLL SWITCHING SYSTEM
See Switching Systems
- N**UMBER 5 CROSSBAR SYSTEM *See* Crossbar System
- N**UMBER 56A OSCILLATOR *See* Oscillator
- N**UMBER 425B NETWORK *See* Network
- Nyquist, H.
 toll traffic study 506
- O**
- Observed 5-6mm Attenuation for the Circular Electric Wave in Small and Medium-Sized Pipes* (A. P. King) 1115-28
- O**FICE *See* Central office
- Ohl, R. S.
 frequency conversion 1416
 rectifier
 wave
 millimeter
 wafer-type 1397
 semiconductor studies i
- O**HMIC RESISTANCE *See* Resistance
- Olsen, K. M.
 zone leveler, design 655
- Olson, E. G.
 electron tube
 traveling wave
 M1789 1343
- O**PEN-WIRE LINES *See* Transmission Lines
- O**PERATE SPEED
 drum, magnetic 707
 electronics 993
 relays 995
- O**PERATING COMPANIES
 functions, primary 422-23
- O**PERATOR DISTANCE DIALING *See* Dial Telephone
- O**SCILLATOR
 56A
 film scales
 calibration 1148-54
- O**UTSIDE PLANT DEPARTMENT
 costs 249
 concentrator
 line
 remote controlled
 economy 249-93
- P**
- P**IN DIODE *See* Diode; junction: silicon: diffused
- P**IN JUNCTION *See* Junction
- P**M *See* Phage Modulation
- P**-N-P TRANSISTOR *See* Transistor: junction
- P**AIRING
 ions 565-636
 calculations 578-82
 carrier
 mobility
 effect 601-07
 (by) diffusion 591-601
 energy levels 610-11
 relaxation time 582-91, 607-10
 semiconductors
 phenomena 575-78
 solubility, effect 613-17
 theories 567-75
- P**ARABOLIC ANTENNA *See* Antenna

- PARAMETER**
 design
 transistor
 junction
 n-p-n
 silicon
 base, diffused
 calculation 14-21
 emitter, diffused
 calculation 14-21
- rectifier, millimeter wave
 wafer-type 1397-1402
- transistor
 amplifier
 performance, relation 815-26
- Pascarella, A. J.
Automatic Testing of Transmission and Operational Functions of Intertoll Trunks 927-54
- biographical material 988
- PHASE MODULATION**
 electron tube
 traveling wave
 M1789 1321-22
- PHOTOCONDUCTIVITY**
 germanium
 etched
 measurements
 combined 1019-40
- PHOTO-VOLTAGE**
 surface
 germanium
 etched
 measurements
 combined 1019-40
- PHYSICS PRIZE** *See* Nobel Prize in Physics
- Pierce, John R.
 amplifier, traveling wave
 large signal theory 373
 biographical material 245
Growing Waves due to Transverse Velocities 109-25
- PIERCE-TYPE ELECTRON GUN** *See* Electron Gun
- Pietruszkiewicz, A. J., Jr.
 semiconductors
 defects
 chemical interactions 613
- PIPE *See* Waveguide
- PLANT *See* Outside Plant Department
- PLATING *See* Electroplating
- POINT CONTACT RECTIFIER *See* Rectifier
- POINT CONTACT TRANSISTOR *See* Transistor
- POWER SUPPLY
 concentrator, line, remote controlled, experimental 269-70
- Prince, M. B.
 biographical material 764
Diffused p-n Junction Silicon Rectifiers 661-84
 diode
 PIN 706
- PRIZE *See* Nobel Prize in Physics
 transmission
 rate 917-26
- PROPAGATION
 waveguide, helix 1355-58
- PULSE
 microwave
 binary
 regeneration 67-90
 testing 69-73
 millimicrosecond
 antenna 45-48
 generation 36-41
 waveguide
 dominant mode
 testing 35-65
 apparatus 36-43
 regenerative
 generator, *see* Generator
- PULSE REGENERATOR *See* Regenerator
- PULSE TRANSMISSION *See* Transmission
- Q**
- Quate, C. F.
 biographical material 245
Coupled Helices 127-78
- R**
- Rack, A. J.
 regenerator
 pulse
 binary
 transistor 1084

RADAR

frequency-modulation
attenuation
atmospheric
wavelengths
millimeter
measurement 907-16

RADIO

propagation
beyond-the-horizon
antenna
parabolic
60-foot diameter 1199-1208;
illus
relay systems
6000 mc
electron tube
traveling wave
M1789 1285-1346

RADIO DETECTION AND RANGING *See* Radar

Raisbeck, Gordon

regenerator
pulse
binary
transistor 1084

RATE *See* Information Rate

Read, W. T., Jr.

biographical material 1466
Theory of Swept Intrinsic Structure
1239-84

READING

magnetic drum 713-16

RECTIFIER

characteristics, ideal 662-63
graph 662

copper oxide
introduction 661
development
problems 662-63

diode
junction
p-n
silicon
diffused 681-83
silicon
diffused 661-84; *illus* 671
design 678-80
electrical characteristics 671-

78

fabrication 670-71
life expectancy 680-83
reliability 680-83

point-contact
wafer-type
silicon
waves
millimeter 1385-1402

selenium
introduction 661

semiconductor
characteristics, ideal 662-63
graph 662

series resistance
control
conductivity
modulation 666-70

wave
millimeter
wafer-type 1385-1402
converter *illus* 1390, 1391, 1396
description 1386-91
parameters 1397-1402
performance 1391-94
wave-wafer unit *illus* 1386

Reed, E. F.

electron tube
traveling wave
M1789 1343

Reed, S. E.

rectifier
wave
millimeter
waver-type 1397

REED SWITCH RELAY *See* Relay

REFINING *See* Zone Refining

REGENERATIVE PULSE GENERATOR *See* Generator

REGENERATIVE REPEATER *See* Repeater

REGENERATOR

pulse
binary
transistorized 1059-84
microwave 73-82
description 83-89

signal

binary
transistorized 1059-84

- Reiss, Howard
 biographical material 764
Chemical Interactions among Defects in Germanium and Silicon 535-636
- RELAXATION TIME
 ions
 pairing 582-91, 607-10
- RELAY
 design 991
 U-type
 coils
 testing
 automatic 1141-48
- UA-type
 coils
 testing
 automatic 1141-48
- Y-type
 coils
 testing
 automatic 1141-48
- reed switch *illus* 253
 concentrator
 line
 remote controlled 252-53
- speed 995
- RELAY, Radio *See* Radio; relay systems
- RELIABILITY
 amplifier
 pulse
 regenerative
 transistorized 1085-86
- Bell System standards 708
- rectifier
 junction
 p-n
 silicon 680-83
- switching systems
 electronic 1016
- transistor 295, 1085
 point-contact 768
- RELIABILITY STUDIES
 experiment time, reduction
 statistical methods 179-202
- Rentrop, Esther M.
 biographical material 532
Crosstalk on Open-Wire Lines 515-18
- REPEATER
 delay distortion
- phase, tables
 tabulation 747-49
- regenerative
 block diagram 68
- RESISTANCE
 ohmic
 diodes 685
- RESISTANCE-CAPACITANCE NETWORK
See Network
- RESONANCE
 modes
 loosely-coupled 899-906
 spurious 899-906
- Richardson, P. H.
 phase, tables
 computation 749
- Riley, J. F.
 electron tube
 traveling wave
 M1789 1343
- Riordan, J.
 toll traffic study 507
- Robb, D. T.
Automatic Testing in Telephone Manufacture 1129-54
 biographical material 1236
- Rosenfeld, Jack L.
 biographical material 532
Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type 375-420
- Ross, I. M.
 diode
 PIN 706
- rectifier
 junction
 p-n
 silicon
 development 684
- ROUND WAVEGUIDE *See* Waveguide
- ROUTING
 alternate
 toll
 administration 441-42
 economics 437-41
 engineering 441-42
 methods, practical 487-505
- dialing, nationwide 97-99
 map, 1965, 96

- Rowe, H. E.
 frequency conversion 1416
- RUGGEDNESS
 Bell System standards 708
 transistor
 point-contact 768
- Rulison, R.
 rectifier
 junction
 p-n
 silicon
 development 684
- S**
- Saloom, Joseph A., Jr.
 biographical material 532
- Saloom, J. A.
Detailed Analysis of Beam Formation with Electron Guns of the Pierce Type
 375-420
- Saloom, J. A.
 electron tube
 traveling wave
 M1789 1343
- Sandsmark, P. I.
 electron tube
 traveling wave
 M1789 1343
- Sansalone, F. J.
 isolator
 field displacement 896
- Sawyer, Baldwin
 biographical material 764
Single Crystals of Exceptional Perfection and Uniformity by Zone Leveling
 637-60
- Scuff, J. H.
 semiconductor studies i
- Seattaglia, J. V.
 regenerator
 pulse
 binary
 transistor 1084
- Scheideler, C. E.
 amplifier
 transistor
 junction
 tetrode 840
- Schimpf, L. G.
 biographical material 988
Design of Tetrode Transistor Amplifiers
 813-40
- Schramm, F. W.
 testing
 automatic 1154
- Seidel, Harold
 biographical material 988
Field Displacement Isolator 877-98
- SELENIUM RECTIFIER *See Rectifier*
- SEMICONDUCTOR(S), SEMICONDUCTING MATERIALS
 aqueous solutions, analogy 537
 impurities
 diffusion into 1-34
 ions
 pairing
 phenomena 575-78
- leveling, *see Zone Leveling*
- Nobel Prize in Physics, 1956 i-iv
 regions
 extrinsic 1239-84
 intrinsic 1239-84
- shaping
 electrolytic 333-47
 mechanical 333
- structure
 swept intrinsic
 theory 1239-84
- surface
 layers
 measurement 1209-21
- traps
 cross sections 1041-58
 distribution 1041-58
- zone leveling, *see Zone Leveling*
- See also Crystal; Diode; Junction*
- SEMICONDUCTOR RECTIFIER *See Rectifier*
- SERIES RESISTANCE RECTIFIER *See Rectifier*
- SERVICE MAINTENANCE *See Maintenance*
- Shannon, C. E.
 information rate
 interpretation 926
- SHAPING
 electrolytic

- crystal
 germanium 333-47
 silicon 333-47
 semiconductors 333-47
- Sharpless, William M.
 biographical material 1466
Wafer-Type Millimeter Wave Rectifiers
 1385-1402
- Shillington, Harry R.
Automatic Machine for Testing Capacitors and Resistance-Capacitance Networks 1179-98
 biographical material 1236
- Shockley, William *illus* ii
 biographical material iv
 Nobel Prize in Physics, 1956 i-iv
- Shoffstall, H. F.
Automatic Testing of Transmission and Operational Functions of Interroll Trunks 927-54
 biographical material 988
- SIGNAL
 binary
 amplification
 transistorized 1059-84
 regeneration
 transistorized 1059-84
- SIGNALING ALPHABET *See* Alphabet
- SILICON
 crystal, *see* Crystal
 defects
 interactions, chemical 535-636
- SILICON DIODE *See* Diode
- SILICON N-P-N TRANSISTOR *See* Transistor
- SILICON RECTIFIER *See* Rectifier
- Single Crystals of Exceptional Perfection and Uniformity by Zone Leveling* (D. C. Bennett, B. Sawyer) 637-60
- 60-Foot Diameter Parabolic Antenna for Propagation Studies (A. B. Crawford, H. T. Friis, W. C. Jakes, Jr.) 1199-1208
- Slepian, David
 biographical material 245
Class of Binary Signaling Alphabets
 203-34
- Smith, K. D.
 rectifier
 junction
- p-n
 silicon
 development 684
- Smith, S. V.
 testing machines 1198
- Smits, Friedolf M.
 biographical material 1236
Use of an Interference Microscope for Measurement of Extremely Thin Surface Layers 1209-21
- Sobel, Milton
 biographical material 246
Statistical Techniques for Reducing the Experimental Time in Reliability Studies 179-202
- SOLUTION *See* Aqueous Solutions
- SPEED *See* Operate Speed
- SPENT BEAM *See* Electron Beam
- SPURIOUS MODES *See* Mode
- STATISTICAL METHODS
 reliability studies
 experiment time, reduction 179-202
Statistical Techniques for Reducing the Experiment Time in Reliability Studies (M. Sobel) 179-202
- STEPPED COUPLER *See* Coupler
- Stiles, G. J.
 electron tube
 traveling wave
 power saturation 867
- STORAGE SYSTEMS *See* Digital Systems
- SUMMING AMPLIFIER *See* Amplifier
- SURFACE
 semiconductor
 layers
 measurement 1209-21
- transistor
 point-contact
 treatments
 effects 767-811
- SURFACE PHOTO-VOLTAGE *See* Photo-Voltage
- Swenson, R. C.
 rectifier
 junction
 p-n
 silicon
 development 684
- SWEPT INTRINSIC STRUCTURE *See* Semiconductor(s); structure

- SWITCH**
- waveguide
 - structure *illus* 1121
- SWITCHING**
- background 991
 - concentrator
 - line
 - remote controlled
 - economy 249-93
 - knowledge 991
 - local
 - crossbar tandem
 - application 91-93
 - nationwide
 - crossbar tandem
 - application 94-97
 - toll
 - translator
 - card 716-19
 - magnetic drum 707-45
- SWITCHING SYSTEMS**
- 4A toll
 - translation
 - card 716-19
 - magnetic drum 708-45
 - concentration 249
 - control 998-1012
 - defined 993
 - design
 - electronics in 991-1018
 - electronic
 - reliability 1016
 - electronics in 991-1018
 - equipment concepts 1012-14
 - interconnecting network 994-98
 - long distance
 - crossbar tandem as 91-108
 - maintenance 1014-16
 - See also* Crossbar Systems
- SYRACUSE, NEW YORK**
- toll traffic map 439
- T**
- TWT** *See* Electron Tube: traveling wave
- Tables of Phase of a Semi-Infinite Unit Attenuation Slope* (D. E. Thomas) 747-49
- TANDEM CROSSBAR** *See* Crossbar Systems
- Tannenbaum, Morris**
- biographical material 246
 - Diffused Emitter and Base Silicon Transistors* 1-22
- TAPE-O-MATIC TEST SET** *See* Test Set
- TAPERED COUPLER** *See* Coupler
- TECHNICAL PAPERS**, Bell System, not published in this Journal 235-41, 519-26, 751-58, 973-78, 1223-29, 1454-60
- TELEPHONE** *See* Dial Telephone
- TEMPERATURE**
- diode
 - junction
 - germanium 661
 - silicon alloy 661
 - modulator
 - microwave
 - noise 1404
- Tendick, Frank H., Jr.**
- biographical material 1236
 - Transistor Pulse Regenerative Amplifiers* 1085-1114
- TEST(s), TESTING**
- antenna
 - microwave
 - pulses, millimicrosecond 45-48
 - circuits
 - relay
 - switching
 - automatic 1155-78
 - crystal
 - germanium 642-43
 - defined 1129-30
 - dial telephone
 - (in) manufacture
 - automatic 1129-54
 - electron tube
 - traveling wave
 - M1789 1342-43
 - manual
 - cost 1129-54
 - network
 - 425B
 - automatic 1135-41
 - oscillator
 - 56A
 - film scales
 - calibration 1148-54

- pulse
 microwave
 binary
 regeneration 69-73
- rectifier
 diode
 junction
 p-n
 silicon 681-83
- relay
 U-type
 automatic 1141-48
- UA-type
 automatic 1141-48
- Y-type
 automatic 1141-48
- reliability studies
 experiment time
 reduction, by statistical techniques 179-202
- translator
 magnetic drum 744-45
- trunks
 intertoll
 automatic
 operation 927-54
 equipment 929-34; *illus* 933
 scheme, basic 937-52
 transmission 927-54
 equipment 929-34; *illus* 933
 scheme, basic 937-52
- waveguide
 dominant mode
 pulses, millimicrosecond 35-65
 apparatus 36-43
- See also* Life expectancy; Reliability; Ruggedness
- TEST SET
 card-o-matic
 relay circuits 1155-78
- tape-o-matic
 relay circuits 1155-78
- TEST MACHINES
 capacitors 1179-98
 development 1129-35
 networks
 resistance-capacitance 1179-98
 requirements 1132-33
- TETRODE TRANSISTOR *See* Transistor:
 junction
- Thaeler, Charles S.
 biographical material 533
Crosstalk on Open-Wire Lines 515-18
Theories for Toll Traffic Engineering in the U. S. A. (R. I. Wilkinson) 421-514
Theory of Swept Intrinsic Structure (W. T. Read, Jr.) 1239-84
- Theurer, H. C.
 semiconductor studies i
- Thomas, Donald E.
 biographical material 246, 765
Diffused Emitter and Base Silicon Transistors 1-22
Tables of Phase of a Semi-Infinite Unit Attenuation Slope 747-49
- Thomas, L. C.
 amplifier
 pulse
 regenerative
 transistor 1114
- Thompson, G. P.
 Nobel Prize in Physics, 1937 iii
- Tien, Ping King
 biographical material 533
Large-Signal Theory of Traveling-Wave Amplifiers 349-74
- TOLL ALTERNATE ROUTING *See* Routing
- TOLL SWITCHING *See* Switching
- TOLL TRAFFIC *See* Traffic
- TRAFFIC
 concentrator
 line
 remote controlled 249-93
- routing, *see* Routing
- toll
 Clos, C., study 431, 470
 engineering
 United States 421-514
- expansion 423
- Kosten, L., study 431
- overflow
 moments 507-11
 peakedness 443-46; *graph* 444
- Wyckoff, Miss E. V., study 445
- trunking 421-514
- TRANSATLANTIC TELEPHONE CABLE
 repeaters
 delay distortion
 phase, tables
 tabulation 747-49

TRANSDUCER

helix

coupled 161-65

heterodyne conversion

nonlinear element

admittance 1403-16

gain 1403-16

TRANSISTOR

base, *see* Basecontacts, *see* Contact

high-frequency operation

base layers, thin 1

junction

alloy

germanium

concentrator, line

remote controlled 253

temperature, effect 253

analog systems, applications 295-332

digital systems, applications 295-332

n-p-n

silicon

base, diffused 1-22

design 14-21

electrical characteristics 6-10

fabrication 2-6

parameters, design 14-21

structure 10-14

emitter, diffused 1-22

design 14-21

electrical characteristics 6-10

fabrication 2-6

parameters, design 14-21

structure 10-14

p-n-p

germanium

base, diffused 23-34

electrical characteristics 26-33

fabrication 23-24

physical characteristics 24-26

properties 295

parameter, *see* Parameter

point-contact

applications 768

characteristics

surface treatments
effect 767-811

demand 768

dependability 768

design 769

electrical characteristics 768

knowledge

empirical 769

military applications 768

physical properties 769-811
processing

surface effects 767-811

reliability 768

ruggedness 768

power consumption 295, 1085

reliability 295, 1085

size 295, 1085

surface, *see* SurfaceTRANSISTOR AMPLIFIER *See* Amplifier
Transistor Circuits for Analog and Digital Systems (F. H. Blecher) 295-332TRANSISTOR INTEGRATOR *See* Integrator*Transistor Pulse Regenerative Amplifiers*
(F. H. Tendick, Jr.) 1085-1114TRANSISTOR VOLTAGE ENCODER *See* Encoder*Transistorized Binary Pulse Regenerator*
(L. R. Wrathall) 1059-84

TRANSLATOR

card

4A illus 1006

magnetic drum alternative 707, 709

magnetic drum illus 710

administration equipment 741-44;
illus 740

block diagram 742

block diagram 720-21

card translator

alternative 707, 709

circuit design 725-41

equipment 725-41

functions 719-25

interchangability 719

switching

toll 707-45

testing program 744-45

TRANSMISSION

data, *see* Digital Systems
digital
 history 1059-84
loss
 resonance
 modes
 loosely-coupled 899-906
trunk
 intertoll 955
 dialing
 direct distance 955-72
 operator distance 955-72

MICROWAVE

pulse, binary
 advantages 67-68
 regeneration 67-90

RATE

routing, *see* Routing

TRUNKS

 intertoll
 testing
 automatic 927-54

See also Information Rate

TRANSMISSION LINES

concentrator, *see* Concentrator

DISPERSION

 helix, bifilar 146-48

HELICES, COUPLED

 equations 133-37

OPEN-WIRE

 crosstalk 515-18

 transpositions 515-18

Transmission Loss due to Resonance of Loosely-Coupled Modes in a Multi-Mode System (A. P. King, E. A. J. Mareatili) 899-906

TRANSMISSION SYSTEMS**MULTIMODE**

 loss
 modes
 loosely-coupled 899-906

TRANSPOSITION

 transmission lines
 open-wire 515-18

TRAVELING-WAVE TUBE *See* Electron Tube

TRUNK(S), TRUNKING

 intertoll

Bell System statistics 423

LOSS**NET****Maintenance****DIALING**

 direct distance 955-72

 operator distance 955-72

OPERATION**TESTING**

 automatic 927-54

TRANSMISSION**TESTING**

 automatic 937-54

ROUTING, ALTERNATE**TRAFFIC ENGINEERING**

TUBE, ELECTRON *See* Electron Tube
TYPE M1789 ELECTRON TUBE *See* Electron Tube: traveling wave

U**UHLIR, ARTHUR, JR.**

biographical material 533

Electrolytic Shaping of Germanium and Silicon 333-47

UNITED STATES

telephone statistics 423

Use of an Interference Microscope for Measurement of Extremely Thin Surface Layers (W. L. Bond, F. M. Smits) 1209-21

V**VACUUM TUBE** *See* Electron Tube**VOLTAGE****BREAKDOWN**

 diodes 685

 photo, *see* Photo-Voltage

VOLTAGE COMPARATOR *See* Comparator

VOLTAGE ENCODER *See* Encoder

VOSS, R. G.

 electron tube

 traveling wave

 M1789 1343

W

Wafer-Type Millimeter Wave Rectifiers (W. M. Sharpless) 1385-1402

- Walker, Laurence R.
- amplifier, traveling wave
 - large signal theory 373
 - biographical material 247
 - Growing Waves due to Transverse Velocities* 109-25
- Wallace, R. L., Jr.
- amplifier
 - transistor
 - junction
 - tetrode 840
- WAVE
- backward
 - amplifier, traveling wave 351-55
 - circular
 - attenuation
 - 5-6mm
 - pipes
 - medium-sized 1115-28
 - small 1115-28
 - forward
 - amplifier, traveling wave 351-55
 - millimeter
 - rectifier
 - point-contact
 - wafer-type 1385-1402
 - slow
 - propagation
 - electron flow 109-25
- WAVE COUPLER *See Coupler*
- WAVE RECTIFIER *See Rectifier*
- WAVEGUIDE
- coupler, *see* Coupler
 - dominant mode
 - testing
 - pulses
 - millimicrosecond 35-65
 - apparatus 36-43
 - helix 1347-84
 - attenuation 1358
 - boundary value problem 1351-55
 - composition 1347-84
 - equation 1381-84
 - formation 1348-50
 - propagation constants 1355-58
 - helices
 - zero-pitch 1358-78
 - mode, *see* Mode
- round
 - attenuation
 - 5-6mm 1115-28
 - medium-sized
 - wave
 - circular
 - attenuation
 - 5-6mm 1115-28
 - small
 - wave
 - circular
 - attenuation
 - 5-6mm 1115-28

switch, *see* Switch

transmission, *see* Transmission

Waveguide Investigations with Millimicro-second Pulses (A. C. Beck) 35-65

Weeks, G. E.

 - testing machines 1198

Weisbaum, S.

 - biographical material 989
 - Field Displacement Isolator* 877-98

Weiss, M. T.

 - ferrite devices
 - nonreciprocal 877

Weiss, Miss R. A.

 - phase, tables
 - tabulation 749

WESTERN ELECTRIC

 - test machines 1129-54
 - testing
 - automatic
 - facilities 1154

Whitaere, W. E.

 - wave
 - electric
 - circular
 - attenuation 1128

Wilkinson, Roger I.

 - biographical material 533
 - Theories for Toll Traffic Engineering in the U. S. A.* 421-514

WIRE CENTER

 - defined 250

WISCONSIN

 - intertoll trunk groups, principal
 - map 424

Wolfertz, W. F.

amplifier
transistor
junction
tetrode 840

Wolontis, V. M.

amplifier, traveling wave
large signal theory 373

Wrathall, Leishman R.

biographical material 1237
Transistorized Binary Pulse Regenerator 1059-84

WRITING

magnetic drum 712-13

Wyckoff, Miss E. V.

toll traffic study 445

Y

Young, James A., Jr.
biographical material 1466
Helix Waveguide 1347-84

Z

ZONE LEVELER *illus* 656
ZONE LEVELING
germanium
apparatus 655-60 *illus* 656
technique 655-60
principles, basic 638-41
ZONE MELTING
defined 637
ZONE REFINING
germanium 637



