THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXVIII JANUARY 1959 NUMBER 1

Logic for a Digital Servo System

By R. W. KETCHLEDGE

(Manuscript received June 27, 1958

Methods are described for performing comparisons of fast binary numbers. These techniques have proved useful in the positioning of cathode ray tube beams in a photographic memory. A binary address is compared with a digital indication of the present position in circuitry called digital servo logic. The output of the servo logic is an analog indication of the positional error. Logics are described for obtaining sign only, sign plus magnitude and sign plus approximate magnitude.

1. INTRODUCTION

Digital storage of information on photographic emulsion is characterized by the large amounts of information that can be stored on a small physical area.[1] This same advantage also implies the need for exceptionally high precision in the access facilities. A memory system of this type has been developed which uses a cathode ray tube to interrogate simultaneously a group of photographic plates.[2,3,4] The access problem in this store is to position an electron beam in accordance with a binary number to an accuracy of a fraction of a thousandth of an inch with microsecond positioning times. The binary address calls for digital circuitry, while the high accuracy implies the use of feedback techniques.

The positioning technique adopted uses a fraction of the storage capacity to indicate, in parallel digital form, the present position of the

cathode ray spot. This digital indication is then compared with the binary number representing the desired position. The comparison is performed in circuitry called the digital servo logic. The output of the servo logic is an analog error signal which drives the electron beam to eliminate the positional error. The use of this feedback technique permits the beam position to be determined by the mechanical edges of bar patterns on a group of photographic plates. This relaxes the mechanical tolerances on the optical system and reduces the need for high precision in the electrical circuits.

The digital servo logic problem is a number comparison problem. The number representing the present position must be compared with the binary number representing the desired address in order to extract the sign of the difference or, preferably, the sign and magnitude of the difference. The number comparison must remain valid at all possible transition values of the present position number, since this number varies continuously during the servo process.

The transition problem is one of the more difficult aspects of digital servoing. The actual position takes on a continuous range of values during the servo operation and, consequently, one desires a continuous indication of the position, even though the position is being represented in "digital" form. Thus, transition values for the position digits must operate the circuit satisfactorily. In binary codes where many digits change simultaneously difficulties occur because these changes are not exactly simultaneous. This has led to the use of standard Gray code to represent the present position. In addition, a "pseudo-binary" translation of the Gray number is required.

Conventional digital adding and subtracting circuits are unsuitable for high positioning speeds because of their dependence upon digits of low significance. In normal subtraction logic the least significant digits are compared first and the carry process moves towards digits of greater significance. One departure of this digital servo logic from a conventional parallel subtractor is the use of carries that proceed from the most significant digit towards the least. This permits the subtraction process to ignore digits which are changing too rapidly to be read.

Finally, the output of the logic need not be digital. The typical requirement is only for an analog signal representing the number difference. In most cases, this signal is an error signal, and a rather rough approximation of its magnitude is sufficient. However, the analog signal must be a continuous representation of the error and, near the desired address, fractional cell errors must be corrected.

II. SIGN-ONLY LOGIC

2.1 *General*

Sign-only methods are less complex than the sign-plus-magnitude methods. Also, the sign-only methods are a useful introduction to some of the concepts involved. Thus they are described first. The simplest of these is binary-binary sign-only logic — logic that gives the sign of the difference between two numbers when both are expressed in conventional binary code.

2.2 *Binary-Binary Sign-Only*

The logic circuit described here takes two binary numbers as voltages or no voltage on two sets of leads and produces an output whenever the first number is equal to or larger than the second. Modifications can also be used to recognize as separate signals "larger," "equal" or "smaller."

The method is to observe the most significant error and to disregard all errors of lesser significance. This can be accomplished by comparing the numbers digit by digit, starting with the most significant digits and disregarding all but the first error found. In other words, the sign of the most significant mismatch indicates the sign of the difference between the two numbers.

The logic circuit of Fig. 1 is one which will perform the desired operations and yield an output if and only if $a_1a_2a_3 \cdots a_n$ is equal to or larger

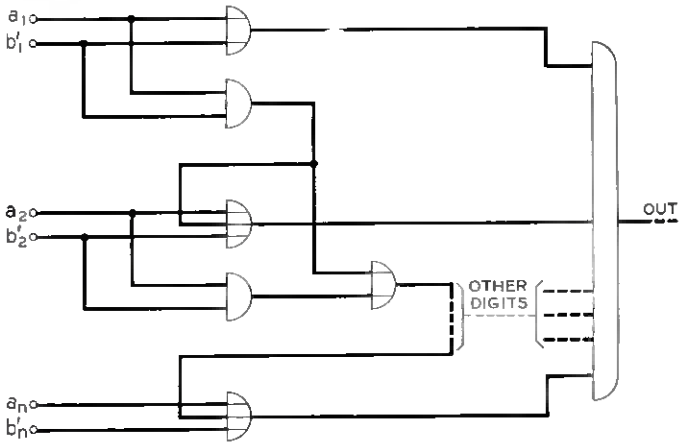


Fig. 1 — Comparison circuit for binary numbers.

than $b_1b_2b_3 \cdots b_n$. Note that this logic is driven by the logical complement of the b number. Thus, if the a digits are the output of the optical beam position decoder and the b digits are the address of the desired beam position, the deflection amplifier may be properly driven by the output of the logic circuit. This will servo the beam to the edge of the match position. The transition problem limits the practicality of this method for servo purposes where transitional values of one of the numbers must be faithfully decoded. This suggests the use of Gray code for the present-position numbers.

2.3 *Gray-Binary Sign-Only*

Having binary-binary logic indicates that, for Gray-binary, all that is needed is to translate the Gray to binary and then compare. However, the translation introduces the same transition difficulties that required the use of Gray code initially. Fortunately, the difficulty is avoidable by the use of a pseudo-translation of the Gray number.

To translate a Gray number to binary, reverse any Gray digit which is preceded by a 1 in the binary translation. Thus, a Gray 111 translates to a binary 101. The most significant Gray 1 is not reversed since it is "preceded" by a binary 0. Thus the most significant binary digit is a 1. This reverses the second Gray digit, making it 0 in binary. The least significant Gray digit is thus preceded by a 0 and is not reversed.

The normal translation of Gray to binary is based on a binary number that changes as the Gray number changes. The pseudo-translation of Gray to binary is based on a fixed binary number. It is useful because of the following, somewhat surprising, fact: Choose any Gray number and any binary number. Reverse those Gray digits which are immediately preceded by a 1 in the binary number. This forms a pseudo-binary number. If the original Gray number was larger than the binary number, the pseudo-binary number, interpreted as binary, will also be larger. If the Gray was smaller, the pseudo-binary will be smaller; if equal, equal. In other words, pseudo-translation of a Gray number to pseudo-binary does not change the sign of the comparison with the controlling binary. In the servo logic the binary address is used to reverse those Gray digits which immediately follow 1's in the address. The new number so formed can then be compared with the address in binary-binary logic to obtain the sign of the difference. Note that the pseudo-binary number is still characterized by only one digit transition at a time. Therefore, the transition difficulties associated with multiple simultaneous digit changes do not occur.

TABLE I

Present Position (Gray)	Desired Address (Binary)							
	000	001	010	011	100	101	110	111
000	000	000	001	001	010	010	011	011
001	001	001	000	000	011	011	010	010
011	011	011	010	010	001	001	000	000
010	010	010	011	011	000	000	001	001
110	110	110	111	111	100	100	101	101
111	111	111	110	110	101	101	100	100
101	101	101	100	100	111	111	110	110
100	100	100	101	101	110	110	111	111

Table I shows the pseudo-binary translation for three-digit numbers

An investigation of Table I shows the following features:

i. Corresponding to a given input address, each column yields a code where only one digit changes at a time.

ii. For the numbers lying along the main diagonal in the table, the Gray number corresponding to the input address has been transformed into the input address number. Thus, a match between the pseudo-binary number and the input address indicates that the beam is at the desired address.

iii. All positions above the desired one are translated into numbers which, in a binary sense, are smaller than the input number; those below are translated into larger ones.

Property iii indicates that the final step merely involves determining whether a particular binary number is greater or smaller than a given one. A simple circuit which effects such a determination has already been presented. Fig. 2 shows a logic structure for such a Gray-binary comparison.

2.4 *Gray-Gray Sign-Only*

Logics have been found which compare two Gray numbers directly with polarity reversals controlled by the address. The sign of the difference is determined by the most significant mismatch between the numbers, except that the sign of this mismatch is reversed if there is an odd number of preceding 1's in the address. Alternatively, one can merely translate the address Gray number to binary, and then use Gray-binary logic.

A third method is to treat the Gray address as if it were a binary number. With random addressing, it often does not matter where the

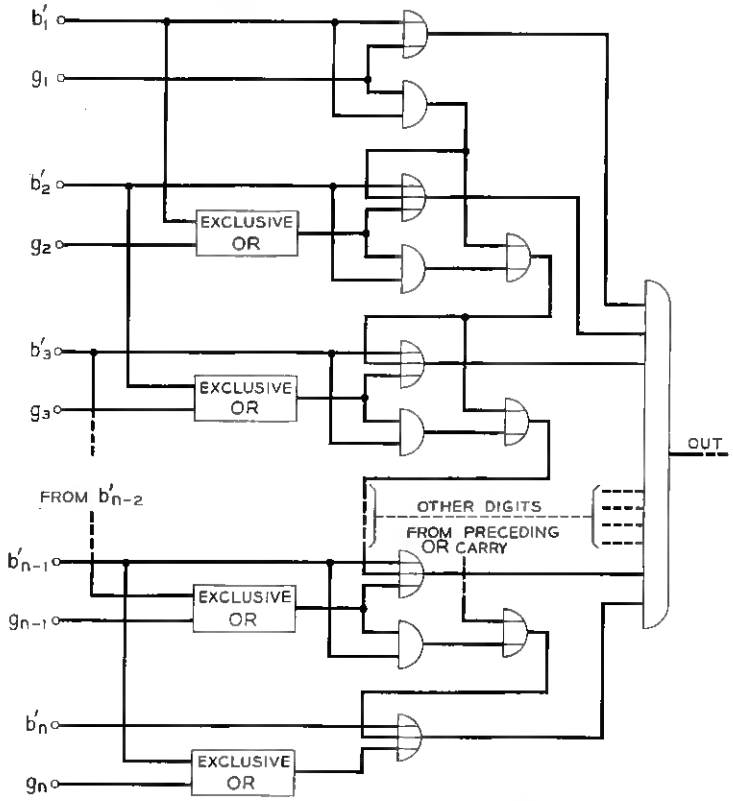


Fig. 2 — Gray-binary sign-only logic.

information is stored on the storage surface. Thus, transpositions occur in information location, but an independent position is still obtained for every address.

III. LOGIC FOR MAGNITUDE AND SIGN

3.1 *General Binary-Binary Considerations*

If both the address number and the position number are in binary code, a simple subtraction will establish the difference. In this subtraction,

$$\begin{aligned} 1 - 1 &= 0, \\ 1 - 0 &= +1, \\ 0 - 1 &= -1, \\ 0 - 0 &= 0. \end{aligned}$$

This forms a difference number and three possible values for each of its digits, namely, $+1$, 0 or -1. If these digits are weighted as in ordinary binary notation (1, 2, 4, 8, etc.) and the indicated sign is applied, the resultant number is an accurate measure of the magnitude and sign of the difference. However, certain combinations lead to a plus digit followed by a succession of minus digits. This cancellation effect makes it very difficult to obtain an accurate analog voltage for the difference number. This accuracy problem makes it desirable, therefore, to consider schemes in which large cancellations are avoided.

In any of the magnitude-determining circuits, regardless of the codes used, some form of quantizing or offset should be added to define an exact balance point. An added digit can be used to permit a small additional drive that will settle the servo in the center of the zero-output region of the logic. Alternatively, a small fixed drive in a fixed direction can be used so that the net drive in the balance region of the logic is small but finite. Typically, it should be equal to a half-cell error. This would cause the servo to drive to the transition between a zero output for the logic and a one-cell output of sign opposite to the fixed drive. This latter is probably the easiest method of quantizing, since it does not require an additional digit and its corresponding circuitry.

3.2 *A Typical Binary-Binary Logic*

A number of binary-binary logics have been found for obtaining the analog magnitude and sign. In general, these compare the two numbers digit by digit, developing at each digit outputs and/or carries. The carries propagate towards digits of less significance. The outputs feed a digital-to-analog converter of a conventional sort. The direction of carry in certain cases can result in some outputs being developed from digits of high significance which represent too large an error signal. This is corrected by the following outputs being generated in the opposite sign. So long as this cancellation does not exceed 50 per cent, no large loss in accuracy results. In effect, this approach takes a $+$, 0, $-$ form of the binary difference number and translates it into a form where severe cancellations cannot occur. In the logic to be described next, all cancellations have been eliminated by increasing the possible outputs per digit from three ($+$, 0, $-$) to five (-2, -1, 0, $+1$, $+2$).

The point of departure is to take one binary number, subtract a second from it, and obtain a binary difference number whose digits may be $+1$, 0 or -1 but whose digits have the same magnitude significance as a conventional binary number. This binary difference number is then modified by logic circuits. Finally, conventional binary-to-analog

conversion is used, but each binary digit can drive separately in either plus or minus sign and in single or double strength. Therefore, the possible outputs on any binary position are +2, +1, 0, −1, −2. Plusses and minuses of output are never generated simultaneously on the comparison of any two binary numbers (no cancellations).

The first mismatch starts a carry, but this carry cannot be stopped by a succeeding mismatch of opposite sign. If the first mismatch (most significant) is plus, a plus carry is started. This inhibits the initiation of any minus carry in less significant digits. A following 0 (match) produces a single-strength plus output at that digit's weight, but a following minus mismatch inhibits this single-strength plus output. A following plus mismatch combined with the plus carry produces a second single-strength plus output and, since in this case the other output is not inhibited, a double-strength total is generated. Equivalent rules apply on a minus carry. The Boolean algebra representing these rules is given below. Each V_n represents a "unity" output, and double weight occurs when both V_n's are active; A = position digit, B = address digit:

$$(+\Delta B_n) \equiv B_n A_n',$$

$$(-\Delta B_n) \equiv B_n' A_n,$$

$$\text{(following } C_+) = (+\Delta B)(C_+')(C_-') + C_+ = C_+ + (+\Delta B)(C_-'),$$

$$\text{(following } C_-) = (-\Delta B)(C_+')(C_-') + C_- = C_- + (-\Delta B)(C_+'),$$

$$+V_{n_1} = C_+(-\Delta B_n)',$$

$$-V_{n_1} = C_-(+\Delta B_n)',$$

$$+V_{n_2} = C_+(+\Delta B_n),$$

$$-V_{n_2} = C_-(-\Delta B_n).$$

A logic circuit having these characteristics is shown on Fig. 3.

3.3 *A Use for Binary-Binary Logic*

Transition problems limit the usefulness of binary-binary logics when transitional values must be faithfully decoded. However, it can be used as an applique in a sign-only servo. A sign-only servo is relatively slow for large address changes because of the limited error signal. If more speed is desired, it is possible to add a forward-acting positioning circuit. This can be a conventional digital-to-analog converter driven directly from the address, in which case the servo merely mops up for converter inaccuracies. Alternatively, binary-binary logic can be used to add to

the existing deflection voltage an additional voltage proportional to the change in address. In this case, the old binary address is compared with the new address, and approximate positioning occurs on a forward-acting basis. The use of binary-binary logic to obtain an analog indication of the address change has an accuracy advantage over the use of the new address alone for small address changes. This is because the digital-to-analog converter is required to convert a smaller number when decoding only the change in address.

3.4 *General Gray-Binary Considerations*

The comparison of a binary and a Gray number can take a variety of forms, but, as indicated earlier, the objective of high speed has sharply limited the techniques considered. The problem again is to find a fast, simple method, avoiding carries except toward digits of less significance,

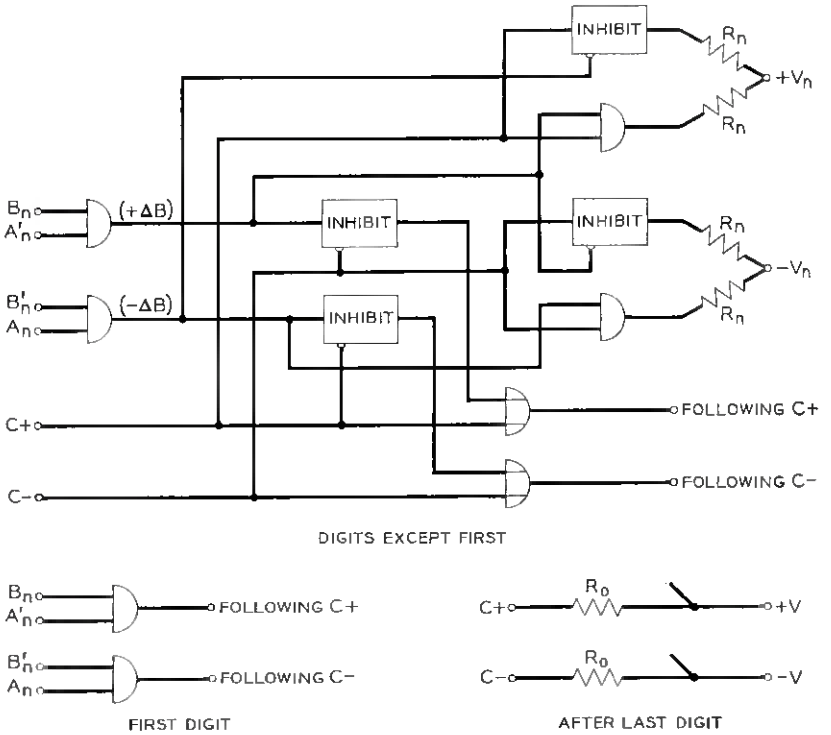


Fig. 3 — Binary-binary logic for magnitude and sign.

etc. Several methods have been devised. All of the schemes derive from the same general observation: merely an observation of the geometrical properties of a Gray code and the numerical significance of a change in a single digit.

Consider a conventional Gray code and what happens to its numerical value if the first digit is reversed. If the initial number were 100 (7), the inversion produces 000 (0) and a change in value of 7 in the example chosen. If the initial number were 110 (4), the inversion yields 010, or 3, and the change is only 1. Inspection of the geometry of the Gray code points up that the position in the sequence of numbers images about the natural reversal point of the digit in question. For this three-digit Gray code, the first digit changes between 3 and 4, and taking any number in the sequence and reversing the first digit merely moves the number to the image point and then an equal distance on the other side. Thus, changing from 100 or 7 to 000 or 0 means moving from 7 to $3\frac{1}{2}$ and then moving the same distance past the image point. Therefore, if we know how far we are from the image point to start with, we know that the digit change represents twice this distance. The obvious conclusion is that, for a change in a single digit of a Gray code, the change is twice the distance from the image point.

A further consideration of the geometric properties of the Gray code indicates its symmetry about the image point of the preceding digit. Take the image point of any digit and observe that the following Gray digits are symmetical about the image point. They can thus be used to measure the distance from the image point and consequently the effect of reversing the preceding digit. The distance to the image point is merely the following Gray digits interpreted as Gray but with the first following digit reversed. If one uses the binary translation of the Gray number being changed, it is possible to use it as a measure of the distance to the image point in the Gray code. Thus, the following binary digits (or their complement) are also a measure of the image distance. Therefore, the magnitude significance of a change in a Gray digit can be obtained from either the following Gray or following binary digits.

None of the foregoing has considered the important case where more than one digit is changed at once. However, this is a matter of signs rather than magnitudes and is described best in connection with the particular schemes. The important point is that, if one is given two Gray numbers which differ in only a single digit, the magnitude of the difference is obtainable from either the following Gray digits or from the corresponding binary digits of one of the numbers.

3.5 *A Typical Gray-Binary Servo Logic*

Determination of the mismatches can be obtained by a comparison of the two numbers in Gray code. It can also be done in the following way. Use the binary address to pseudo-translate the Gray position to pseudo-binary. Then compare this pseudo-binary number with the binary address. The mismatches will occur in the same digits, although the directions of the mismatches (1/0 vs. 0/1) will be reversed following a 1 in the binary address. Recall that this same pseudo-translation was used in the binary-Gray method for sign alone. This produces a set of mismatches having values of +, −, 0, as shown in Table II.

The difference between the address and position is developed by adding the components of the difference represented by each mismatch. This requires the determination of a sign and a magnitude for each mismatch. Table III lists the rules for the magnitudes.

The weights of the mismatches are obtained as noted above following the properties described in Section 3.3. The signs of the mismatches are reversed following an odd number of preceding mismatches. This is not as bad as it might seem. Consider the first mismatch (most significant). A plus mismatch has a weight equal to the following binary address digits in double their binary weight plus one. In other words, gate out

TABLE II

Address	Binary..... Gray.....	111 100	110 101	101 111	100 110	011 010	010 011	001 001	000 000
Position (Gray)	100	0 0 0	0 0 −	0 − +	0 − 0	− + 0	− + −	− 0 +	− 0 0
	101	0 0 +	0 0 0	0 − 0	0 − −	− + +	− + 0	− 0 0	− 0 −
	111	0 + +	0 + 0	0 0 0	0 0 −	− 0 +	− 0 0	− − 0	− − −
	110	0 + 0	0 + −	0 0 +	0 0 0	− 0 0	− 0 −	− − +	− − 0
	010	+ + 0	+ + −	+ 0 +	+ 0 0	0 0 0	0 0 −	0 − +	0 − 0
	011	+ + +	+ + 0	+ 0 0	+ 0 −	0 0 +	0 0 0	0 − 0	0 − −
	001	+ 0 +	+ 0 0	+ − 0	+ − −	0 + +	0 + 0	0 0 0	0 0 −
	000	+ 0 0	+ 0 −	+ − +	+ − 0	0 + 0	0 + −	0 0 +	0 0 0

TABLE III

Gray Position (after reversal if preceding $B_n = 1$)	Binary Address	Mismatch	Weight
1	1	0	0
0	0	0	0
0	1	+	1 + twice following binary number
1	0	−	1 + twice the complement of the following binary number

TABLE IV

First Mismatch	Second Mismatch	Reverse	Action
+	−	yes	double next output, stop carry
+	+	yes	stop carry

all following binary 1's into a converter with binary weighting and add 1. Now consider the second mismatch. If the first and second mismatches are of the same sign they subtract, because the sign of the second is reversed. This says, "do not energize any more outputs beyond the second mismatch." If the signs oppose, the reversal of the second says, "put out in the same polarity both the following binary and its complement." But that is the same as putting out the next following binary weight at double value or the corresponding one at normal. The latter is used here. Table IV may clarify this point.

The corresponding case for an initial minus is obvious.

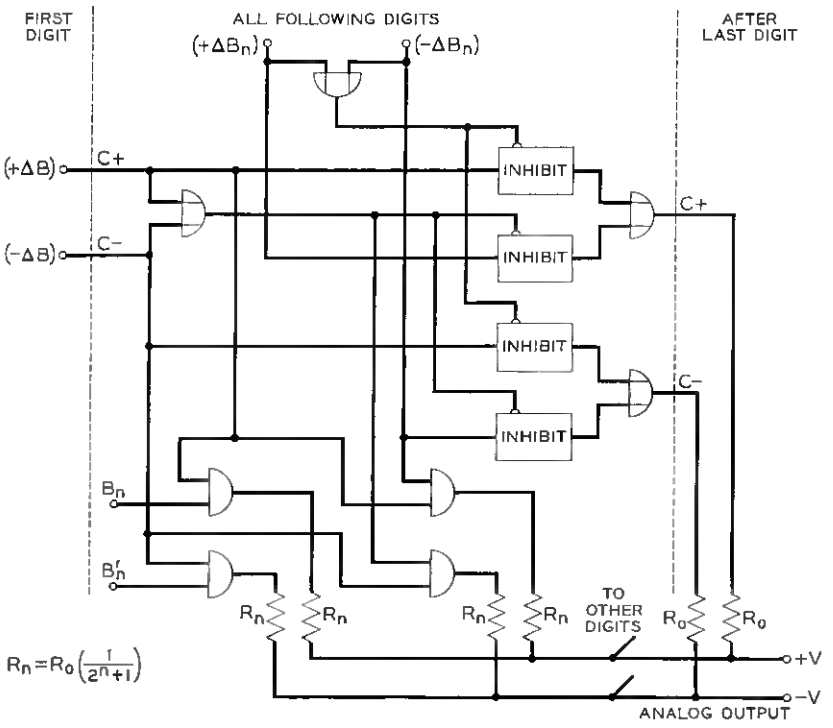


Fig. 4 — Gray-binary logic for magnitude and sign.

This leads rather simply to the following logic and the circuit of Fig. 4, in which $(+\Delta B)$ and $(-\Delta B)$ are plus and minus mismatches respectively:

$$\begin{aligned}
(\text{following } C_+) &= C_+(+\Delta B_n)'(-\Delta B_n)' + (C_-')(C_+')(+\Delta B_n),\\
(\text{following } C_-) &= C_-(+\Delta B_n)'(-\Delta B_n)' + (C_+')(C_-')(-\Delta B_n),\\
(+\Delta B_n) &= G_n(B_{n+1})\, B_n + G_n'(B_{n+1}')\, B_n,\\
(-\Delta B_n) &= G_n'(B_{n+1})\, B_n' + G_n(B_{n+1}')B_n',\\
(+V_{n1}) &= C_+B_n,\\
(-V_{n1}) &= C_-B_n',\\
(+V_{n2}) &= C_+(-\Delta B_n),\\
(-V_{n2}) &= C_-(+\Delta B_n).
\end{aligned}$$

Alternatively,

$$\begin{aligned}
(+V_n) &= C_+[B_n + (-\Delta B_n)],\\
(-V_n) &= C_-[B_n' + (+\Delta B_n)].
\end{aligned}$$

The first digit has no preceding carries so

$$\begin{aligned}
(\text{following } C_+) &= (+\Delta B),\\
(\text{following } C_-) &= (-\Delta B),\\
(+V_{n1}) &= (-V_{n1}) = (+V_{n2}) = (-V_{n2}) = 0.
\end{aligned}$$

The last digit (least significant) must be followed by a ± 1 contributor but no digit mismatch can occur, so

$$\begin{aligned}
(+V_0) &= C_+,\\
(-V_0) &= C_-.
\end{aligned}$$

3.6 *Other Gray-Binary and Gray-Gray Logics*

As noted in Section 3.4, the magnitude significance of Gray-to-Gray or pseudo-binary-to-binary mismatches can be determined from the following Gray digits. These digits can be used in a variety of ways to derive appropriate analog signals. Other logic structure variations involve the number of carries used and their significance. Additional variations depend upon the form of the drives to the digital to analog converter — three-valued, five-valued, with or without cancellation,

etc. It is also possible to form a set of analog signals and switch these to a common output under control of the mismatches. This particular approach probably is of value only in relay circuitry, because of the need for switching a wide range of analog signals.

In spite of all these possible variations there are common factors. Some of these are: the use of mismatches located by Gray-to-Gray or pseudo-binary-to-binary, carries which propagate toward digits of less significance, digital outputs which are controlled by mismatches of digits of equal or greater significance, use of the following digits of one of the numbers to develop the magnitude significance of a mismatch, creation of an analog signal accurately representing the number difference and accurate number comparisons even for transitional values of at least one of the numbers.

3.7 *The Edge Problem*

In any digital servo, and particularly in those using magnitude as well as sign, there is an edge problem. If the digital position information does not extend beyond the edge of the servo area, the circuit may fail if it gets out of the area. For example, a transient overshoot beyond the limits of the normal servo area should produce just as big an error signal as would a similar overshoot in the center of the servo area. Otherwise, overshoots beyond an edge may fail to produce any drive (or perhaps a very small drive) to return the beam to its desired edge position. The foregoing implies that it is mandatory to extend the coding of the position of the beam to the extreme limits to which the beam may be deflected. It is always possible to introduce limiters into the deflection circuit which will prevent extreme values of excess deflection. However, because of drift, these limiters must operate at some distance from the actual edge of the servo area.

One solution to the edge problem appears to be the following: Add one digit to the present position number in the most significant position, thereby extending it over double the normal range. This does not change the width of any digit but merely adds a digit in front of the previous number. The binary address is then modified as follows: The first (most significant) digit is moved one place in the more significant direction. This digit thereby is matched against the digit that was added to the code plate. The gap left in the binary address is filled by using the complement of the first binary address digit. It will be seen that this restricts the binary address so formed to numbers beginning with either 10 or 01. Thus, this binary sequence covers the center half of the range generated by the present position. With this arrangement, overshoots of up to 50

per cent in either direction will still generate accurate error signals. One other advantage is that the addition of a fixed half-cell drive provides full use of all address combinations.

IV. APPROXIMATE LOGICS

4.1 *General Considerations of Approximation*

All the magnitude-determining circuits mentioned in the preceding section are exact. This is a valuable property. However, an ordinary servo need not always determine the exact magnitude of the error. Often, only the approximate error is required. This may necessitate somewhat more loop gain margin, but this is not serious if the error variations are not too large. Much of the complexity of the exact error methods is attributable to the handling of signs. The signs of each of the individual contributions must be properly manipulated to obtain the correct total. Considerable simplification is possible if one merely determines the magnitude of the most significant error contributor and, in addition, the over-all sign of the error signal. In such cases, it is possible thereby to determine the over-all sign exactly and to determine the magnitude to within ± 6 db.

To approximate the difference between a Gray and a binary number to within 2-to-1 accuracy, it is sufficient to know the position and direction of the two most significant pseudo-binary-to-binary mismatches. If these first two mismatches are both of the same sign, only the most significant is necessary. When the first two mismatches are of opposite sign, the most significant mismatch dominates the sign, but the second most significant mismatch may sometimes dominate the magnitude. These considerations permit the construction of approximate magnitude and sign logics. Such logics can be derived independently or can be obtained by simplification of exact magnitude logics.

4.2 *A Typical Approximation Logic*

Fig. 5 shows one type of approximation logic. Here the sign and magnitude have been generated separately. A set of outputs, V_n, is energized to indicate the magnitude, although only the most significant V_n is to be used. This can be accomplished by decoding each V_n separately into an analog signal of the appropriate strength and feeding all of these analog signals through an ordinary diode OR circuit. The output of the OR circuit will equal the largest input and indicates the approximate magnitude of the error. Note that the digit weights used

are not quite binary. Instead of the usual 1, 2, 4, 8, etc., the weights follow the "1 + twice following binary" trend, namely, 1, 3, 5, 9, etc.

Since pseudo-translation of Gray does not change the mismatch pattern but only changes signs, a Gray-Gray method can be used with inputs composed of the binary address and the pseudo-binary position. This simplifies the determination of sign, since the sign is simply that of the most significant mismatch. The logic is:

$$\Delta_n = \text{mismatch of either sign},$$

$$C_{n-1} = C_n + \Delta_{n+1},$$

$$V_n = C_n G_n + \Delta_{n+1} G_n',$$

$$S_+ = \Delta_n B_n C_{n-1}' \text{ with OR from all digits.}$$

Note that G_n is the Gray position without any reversals by the binary address. The reversal of G_n by a preceding binary 1 is used only to form Δ.

Note also that only one carry is required and that only OR logic is used. This permits the carry circuit to be made of a group of cascaded cathode followers. This form of carry has been found to be extremely fast.

V. CONCLUSIONS

The type of digital servo logic described has permitted the attainment of precise high-speed positioning of electron beams. These logics use relatively few active elements compared with many other types of access to a comparable complex of addresses. This is because the logic

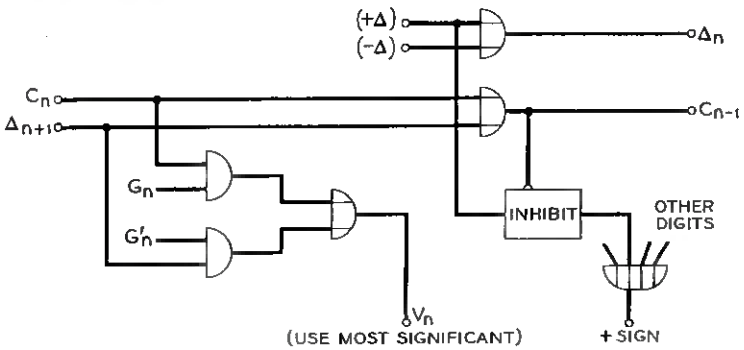


Fig. 5 — Approximation logic.

deals with the binary or Gray forms of the address instead of forms such as 1 out of N, which are less efficient.

These logics are characterized by carries that proceed toward digits of less significance. Thus, the digital outputs are not in the form of ordinary binary numbers but, in general, contain digits having more than two possible states. Such numbers are still suitable, however, for driving an ordinary digital-to-analog converter. Also, the weightings may depart from the normal binary values. A further characteristic of these logics is their ability to decode transitional values of Gray numbers. In other words, if the Gray digits are all 1 or 0 except one, and that digit is, say, 0.5, the logic still gives appropriate analog outputs. The key to this characteristic is the use of pseudo-binary translation of the Gray number. Finally, in the magnitude-determining logics, the following digits of the address are used to establish the magnitude significance of a mismatch.

VI. ACKNOWLEDGMENT

The author is indebted to his associates for both the experimental confirmation of these digital servo concepts as well as for numerous mathematical proofs and extensions. These concepts and various of their embodiments were obtained on a highly intuitive basis, and the author's associates were responsible for bringing mathematical coherence to a group of seemingly different logics. In particular the author wishes to mention L. E. Gallaher, M. Nesenbergs and V. O. Mowery.

REFERENCES

1. King, G. W., Brown, G. W. and Ridenour, L. N., Photographic Techniques for Information Storage, Proc. I.R.E., **41**, October 1953, p. 1421.
2. Staehler, R. E. and Davis, R. C., U. S. Patent No. 2,830,285.
3. Ketchledge, R. W., An Introduction to the Bell System's First Electronic Switching Office, Proc. of Eastern Joint Computer Conf., December 1957, p. 204.
4. Hoover, C. W., Jr., Staehler, R. E. and Ketchledge, R. W., Fundamental Concepts in the Design of the Flying Spot Store, B.S.T.J., **37**, September 1958, p. 1161.

H. S. Renne Is New Editor of B.S.T.J.

H. S. Renne, Technical Information Supervisor, now has editorial responsibility for the Bell System Technical Journal, in addition to his other responsibilities. The former Editor, W. D. Bulloch, has transferred to the Public Relations Department of the American Telephone and Telegraph Company.

Mr. Renne received a bachelor's degree from Kalamazoo College and a Master of Science degree in Physics from Syracuse University. He taught basic electronics, pre-radar and physics at Illinois Institute of Technology during the early years of World War II, and was Editor of the magazine *Radio-Electronic Engineering* for a number of years before joining the staff of Bell Telephone Laboratories.

Logic Synthesis of Some High-Speed Digital Comparators

By M. NESENBERGS and V. O. MOWERY

(Manuscript received September 10, 1958)

Logical schemes for realizing high-speed digital comparators are derived by Boolean algebra methods. Requirements for speed and precision place serious restrictions on the switching circuits. In particular, the precision requirement makes direct subtraction by the use of analog devices undesirable; the speed requirement dictates that any carry structure should propagate from the most significant digit toward the least significant digits. Such schemes have obvious advantages when only an approximate magnitude is desired. Changing numbers in binary code introduces the common transition problem due to multiple digit changes; this problem is avoided by use of the Gray code.

Circuits satisfying the synthesis requirements and giving the sign and exact magnitude of the difference are derived first. These schemes are then modified and simplified to give the sign and approximate magnitude. Circuits giving only the sign of the difference are also derived.

I. INTRODUCTION

1.1 *Applications*

A digital comparator compares two numbers presented in digital form and obtains a measure of the difference between them. Comparison may consist of detecting only the sign of the difference or the direction of mismatch of the two numbers, or the result of the comparison may be both magnitude and sign of the difference. The comparator is essentially a subtracter suitably modified to fulfill requirements of the intended application.

An immediate need for a dependable, high-speed digital comparator is in the feedback control loop for the flying spot store of an experimental electronic switching system.[1] R. W. Ketchledge has derived several methods of implementing such comparators.[2] In this application the comparator functions as an error detector, giving an output depending on the difference between the desired input address position and the fed-

back present position. By servo action, this difference then acts to eliminate the positional error. The input address is presented in parallel digital form and the present position is digitally encoded in parallel. For proper holding action of the servo the comparator output should be a linear error signal when close to a zero difference. A similar technique could be used in applications requiring fast and accurate positioning through a precise digital servo action.

The use of reliable, high-speed comparators in a digital computer allows greater flexibility of operation. Comparison of numbers, in the sequence of a computation, can be used to determine the choice of further operations. Since the programmer cannot always know the results of a certain operation in advance, a built-in comparison scheme can initiate judgement to proceed automatically with subsequent routines. Comparison of numbers is frequently employed in sorting and determining square roots and dividends.[3]

In a broad sense, all measurements can be considered comparisons. If the quantity to be measured appears in digital form the types of comparators to be considered here could be used, especially for rapidly varying quantities. Unlike an analog device, where precision is determined largely by the components and accuracy of driving potentials, the precision of a digital device is limited only by the number of digits used.

1.2 *Synthesis Requirements*

Intended applications of the comparators discussed here place serious restrictions on the logical form of the switching circuits. For example, the extremely high access precision necessary in the flying spot store prohibits the use of analog open-loop positioning and requires the use of digital closed-loop control. This precision requirement also eliminates the possibility of direct, complete analog subtraction of the digital signals in the error detector of the servo loop. High-speed operation of the servo system implies the use of electronic combinational switching circuits with simultaneous operation on all of the digits.

In a servo operating on the magnitude of the difference between the input address and the feedback signal, the value of the feedback signal fed into the comparator may be rapidly changing for large differences. For example, in the flying spot store servo presently operating only on the sign of the error or difference, the beam position moves at a velocity of three spots per microsecond. Each spot, corresponding to an address point on the cathode ray tube face, is designated by a digital number. With the feedback signal appearing in a binary code, one cycle of the least significant digit corresponds to two spots. The frequency or rate of change of this least significant digit is therefore 1.5 mc. If the feedback

signal appears in a Gray code (discussed in Section 2.2) one cycle of the least significant digit corresponds to four spots and the rate of change of this digit is then half of that for the binary code. For a servo controlled by the magnitude of the difference, the velocity of the change of beam position is proportional to the difference. As a specific example, consider such a proportional servo operating with a Gray code and with an error of about 200 spots. The bandwidth required for the least significant digit would then be about $200 \times \frac{3}{4}$ mc = 150 mc.

It is apparent that in such applications the digits of lower significance can be changing at such a rate that their use in the comparison or subtraction scheme becomes impractical. The digits of higher weight will be better defined and those of lower weight will be blurred, due to band limitations. Logical operations, whenever possible, should therefore be performed on the more significant digits, and any carries necessary should propagate from digits of higher significance to digits of lower significance. This synthesis requirement prohibits the use of conventional parallel subtracters with a borrow propagating from the least significant digits.

In many applications only an approximate, or order of magnitude, difference may be required. For such applications it is also advantageous to have the carry or carries in the subtraction operation propagate from digits of higher significance to digits of lesser significance. Since the magnitude of the mismatch between digital numbers is usually determined by digits of higher significance, an approximate difference can often be obtained without the necessity of the carries propagating through all of the digits. This is not possible in a conventional subtracter. In either case, however, it is necessary to examine all digits to obtain an exact difference.

There is little loss of generality in assuming that the input address number, designated by A, appears in two-rail parallel form, as, for example, from a flip-flop register, and that the second number, which can be rapidly changing, appears in a parallel binary or Gray-code form, designated by B or G. For speed of operation, it is desirable to perform as much of the logic as possible on the digits of the fixed number A and to have the digits of the changing number B or G travel through a minimum number of series gates. This does not mean that minimal switching circuits will always be used, however, since there may be advantages to either combining or sharing functions in a slightly expanded circuit. Only functional forms of the switching circuits containing AND, OR and EXCLUSIVE-OR gates plus inverters will be derived here. For economy, flexibility and ease of replacement, an iterated logic structure is desirable. The actual electronic circuits used to realize the

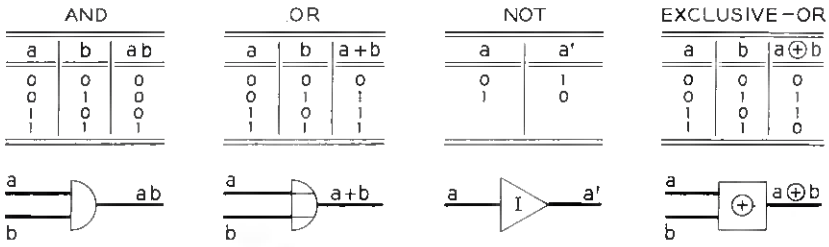


Fig. 1 — Truth tables and circuit symbols for logical operations.

various functions will not be discussed. It should be mentioned that the comparators giving both magnitude and sign of the difference are intended to drive digital-to-analog converters giving an analog output. In all cases when the difference is within ± 1 a linear indication of the difference is required if the comparator is to perform properly in a feedback control loop.

II. NOMENCLATURE

2.1 *Algebraic Operations*

Synthesis of the logic circuits to be used in the comparators is frequently simplified by the use of algebraic expressions. Boolean or switching algebra[4] has proved to be a convenient notation permitting manipulation of series-parallel two-terminal networks into a variety of equivalent circuits, often resulting in simplified or more appropriate forms. The quantities involved in Boolean algebra can be represented by letters or symbols. These symbols represent signals having discrete on-or-off values, represented by 1 or 0 respectively. This convention differs from that sometimes used (as, for example, in Ref. 4). Logical operations employed here will be briefly defined.

The AND operation is a logical conjunction or intersection resulting in 1 only when both variables are 1. The OR operation is the logical disjunction or union, indicated by $+$, resulting in 1 when either or both of the variables are 1. Truth tables and circuit symbols of these operations are shown in Fig. 1. As a consequence of these rules, both the OR and AND functions are commutative, associative and distributive. For example:

$$\begin{aligned} (a + b) + c &= (b + c) + a, \\ a(b + c) &= ab + ac, \\ (a + b)(a + c) &= a + ab + ac + bc, \\ &= a + bc. \end{aligned}$$

An additional concept is the NOT, complement, or negation indicated by a prime. The truth table and circuit symbol for this operation is also shown in Fig. 1. Since effectively the NOT function is an inversion of the signal, the symbol represents an inverting amplifier. Several important relations follow:

$$\begin{aligned} a + a' &= 1, \\ aa' &= 0, \\ (a')' &= a, \\ (ab)' &= a' + b'. \end{aligned}$$

The validity of such equations can always be verified by the method of perfect induction, or substituting for the variables the two possible values, 0 and 1, in all combinations.

A Boolean algebra is closed under the operations of negation and either the OR or AND operations; however, for convenience, we will allow both the AND and OR operations. In addition, we will find it convenient to use a fourth operation called the EXCLUSIVE-OR or "ring-sum" defined as

$$a \oplus b = ab' + a'b. \tag{1}$$

This ring sum is 1 if either a or b, but not both, are 1; it is 0 if both a and b are either 1 or 0. The ring-sum therefore detects a mismatch between the two digits and is the algebraic expression for the common half-adder used in conventional digital adders and subtracters (Ref. 3, Ch. 4). Fig. 1 also shows the truth table and circuit symbol for this operation. Note in particular that

$$a \oplus 0 = a \quad \text{and} \quad a \oplus 1 = a'. \tag{2}$$

2.2 *Codes and Translations*

The type of synthesis used in developing the comparators is dependent upon the types of codes used to represent the numbers. For this reason a brief discussion of codes employed and the translations between them is included.

One of the most convenient number systems for logical operations is the binary system. Since the number $B = b_m b_{m-1} \cdots b_1 b_0$ is represented to the base 2, and therefore each digit b_i takes on the value 0 or 1, the Boolean algebra described in the previous section can be conveniently applied to the digits. The magnitude of the integral number B is represented in this binary code by

$$B = \sum_{i=0}^{m} b_i 2^i,$$

TABLE I

Decimal	Binary			Gray		
	b_2	b_1	b_0	g_2	g_1	g_0
0	0	0	0	0	0	0
1	0	0	1	0	0	1
2	0	1	0	0	1	1
3	0	1	1	0	1	0
4	1	0	0	1	1	0
5	1	0	1	1	1	1
6	1	1	0	1	0	1
7	1	1	1	1	0	0

where m is the most significant place. In this conventional binary system multiple digit changes occur for every increase by two, and half-way through the code all of the digits change. Table I shows the three-digit binary code.

Such multiple digit changes cause difficulties whenever all of the changing digits do not change simultaneously. Nonsimultaneous changes of the digits may be due to variations of bias, gain, delay or operating levels of the individual stages or to misalignment of the coding devices. This is the familiar problem encountered, for example, in digitally encoding shaft positions or other analog-to-digital conversions, in digital positional servomechanisms,[5] and in pulse code communication.[6] To avoid the difficulty of incorrectly reading a rapidly changing number, a Gray[7] or reflected binary code[8] may be introduced, in which only one digit changes between successive numbers of the code.

In our algebraic synthesis of various comparators it is convenient first to consider both numbers in the conventional binary code and then to translate to Gray code the input number, which may be rapidly varying. The cyclic reflected binary code, which we will call simply Gray, has the convenient property of simple translation to and from the conventional binary equivalent. Table I also shows the three-digit Gray code.

The method for finding the magnitude of a number G written in the Gray code is more involved than is evaluating a number written in the binary code. Each digit g_i again has the value 0 or 1, but the weight of the ith digit is now $(2^{i+1} - 1)$ and the sign of each digit not a zero is now alternated, starting with + for the most significant digit g_m (where $g_m \neq 0$) and alternating in sign for each digit which is not zero. This could be termed the decimal translation, and can be written

$$G = \sum_{i=0}^{m} g_i(2^{i+1} - 1)(-1)^{(g_m+g_{m-1}+\cdots+g_{i+1})}.$$

The usual rule for translating the digits of a number in binary code to the digits of a number in Gray code is the following: If the binary digit b_i is preceded by a 1 ($b_{i+1} = 1$), change the ith digit ($g_i = b_i'$); if it is preceded by a 0 ($b_{i+1} = 0$), use the same digit ($g_i = b_i$). This can be written in the shorthand notation of the ring-sum operation:

$$g_i = b_i \oplus b_{i+1} . \tag{3}$$

To convert a Gray digit to its equivalent binary form the rule is to reverse those Gray digits which are preceded by an odd number of 1's in the Gray digits of higher significance. This can be expressed in terms of repeated ring-sums which, in effect, counts the number of preceding 1's:

$$b_i = g_i \oplus g_{i+1} \oplus g_{i+2} \oplus \cdots .$$

Repeated ring-sum operations effectively performs the same function as the modulo 2 notation in determining whether a set of digits is odd or even. An equivalent Gray-to-binary translation can be obtained from the previous binary-to-Gray translation by ring-summing both sides of (3) with b_{i+1} :

$$g_i \oplus b_{i+1} = b_i \oplus b_{i+1} \oplus b_{i+1} = b_i \oplus 0 = b_i . \tag{4}$$

III. EXACT PROPORTIONAL COMPARATORS

3.1 *Analog Precision Problem*

Two main requirements imposed on these comparator syntheses are speed of operation and precision. Perhaps the fastest method of comparison of digital numbers is simple and direct analog subtraction. A typical analog subtracter for obtaining the difference between two binary numbers A and B is shown in Fig. 2. An output voltage (or current) corresponding to the difference is obtained by shunting suitably weighted currents into the summing resistance R_s (or the load). Currents of proper magnitude and polarity are obtained from the AND gates controlled by the individual digits.

Although it is possible to build fast analog subtracters, it may be difficult to meet stringent precision requirements. Difficulties may arise when subtracting large numbers having a small difference. An error as small as one half of one part in the maximum value of the numbers can possibly even result in an incorrect sign for the difference. Analog methods of avoiding these difficulties are not entirely satisfactory. For accurate results, precision components and well-regulated supplies are

necessary. Further complications may result when one of the numbers is in Gray code.

Application of comparators as the error-detecting element in a digital feedback control loop usually implies an analog difference output in order to drive the control elements. What is desired is a method of avoiding direct analog subtraction of two numbers of nearly the same size. This problem is illustrated by the examples:

A	1000	A	0111
B	0111	B	1000
Diff.	+0001	Diff.	−0001

We call such a grouping of digits, where a mismatch in one direction is followed immediately by consecutive mismatches in the opposite direction, a run. Such a run ends when it is followed by a match of the digits or a mismatch in the original direction. If the two original input numbers can be operated on digitally to eliminate such runs then the precision problem can be avoided and analog subtraction can be retained.

3.2 *A One-Carry Binary-Binary Comparator*

Since runs of consecutive digits of the type described in the previous section lead to difficulties in analog subtraction, it is desirable to transform the two binary input numbers A and B into an equivalent set W and V, in which these runs are eliminated. The equivalence implied

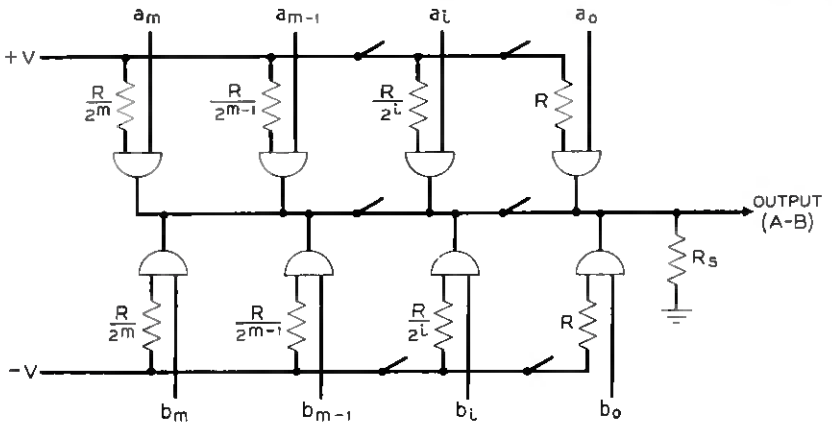


Fig. 2 — Typical analog subtracter.

TABLE II — TRUTH TABLE FOR CARRY FUNCTION OF ONE-CARRY BINARY-BINARY LOGIC

a_i	b_i	a_{i-1}	b_{i-1}	q_i	
				If $q_{i+1} = 0$	If $q_{i+1} = 1$
0	1	0	0	0	0
0	1	0	1	0	1
0	1	1	0	1	0
0	1	1	1	0	0
1	0	0	0	0	0
1	0	0	1	1	0
1	0	1	0	0	1
1	0	1	1	0	0

here is that of $A - B = W - V$ (see example on p. 31). Furthermore, it is also undesirable to perform analog subtraction of equal-weight digits, since a similar precision problem is likely. In this section a simple logic scheme for eliminating these runs and avoiding subtraction of equal digits is illustrated. Resulting circuits are not the most practical for this purpose but will serve to demonstrate the method. The logic will be derived in detail to further illustrate the method.

For reasons outlined in Section 1.2, comparison of the two numbers A and B should start from the most significant digit and proceed toward the lower significant digits. Whenever a run starts, i.e., a mismatch followed by a mismatch of the opposite kind, the outputs should be prohibited. To do this, we can start a carry, or inhibit, function, with the ith digit represented by q_i. This carry should then propagate through the run and stop at the last digit, so that an output can be permitted. In other words, if there is a carry coming in from the next more significant digit, i.e., $q_{i+1} = 1$, then the carry should continue only if any mismatch of the following digits a_{i-1} and b_{i-1} is of the same sign. Or, if there is no carry coming in, i.e., $q_{i+1} = 0$, then a carry should start only if a mismatch of the ith digits is followed by an opposite mismatch in the a_{i-1} and b_{i-1} digits. A carry is possible only when there is a mismatch in the digits a_i and b_i. Table II is a modified truth table giving the carry digit for the two conditions $q_{i+1} = 0$ and $q_{i+1} = 1$. The disjunctive canonical form for this truth table is obtained from the OR of the minimal polynomials for each row. The carry function can then be written:

$$q_i = q_{i+1}'(a_i'b_ia_{i-1}b_{i-1}' + a_ib_i'a_{i-1}'b_{i-1}) + q_{i+1}(a_i'b_ia_{i-1}'b_{i-1} + a_ib_i'a_{i-1}b_{i-1}').$$

TABLE III — TRUTH TABLE FOR OUTPUT FUNCTIONS OF ONE-CARRY BINARY-BINARY LOGIC

q_{i+1}	a_i	b_i	w_i	v_i
0	0	0	0	0
0	0	1	0	1
0	1	0	1	0
0	1	1	0	0
*1	0	0	1	0
1	0	1	1	0
1	1	0	0	1
*1	1	1	0	1

* From (5) or the truth table shown in Table II, these conditions cannot occur and the outputs have been chosen to simplify the final expressions.

Using the ring-sum or EXCLUSIVE-OR operation introduced in Section 2.1, this carry can be rewritten as

$$q_i = (a_i \oplus b_i)(a_{i-1} \oplus b_{i-1})(q_{i+1} \oplus a_i \oplus a_{i-1}). \tag{5}$$

Outputs are permitted only when there is no carry or inhibit function; therefore, a condition for any output is $q_i = 0$. The output digits w_i carry positive weight and the digits v_i carry negative weight. If we are at the end of a run, with $q_{i+1} = 1$, then a positive output is required if $a_i = 0$ and a negative output if $a_i = 1$. These conditions are apparent from the examples given in Section 3.1. If we are not in a run, i.e., $q_{i+1} = 0$, then an output is allowed only if there is a mismatch: a positive output for $a_i = 1$ and $b_i = 0$, and a negative output for $a_i = 0$ and $b_i = 1$. Conditions for the outputs are summarized in Table III. Recalling the condition $q_i = 0$, the outputs can be obtained from the truth table in disjunctive canonical forms:

$$w_i = q_i'(q_{i+1}'a_ib_i' + q_{i+1}a_i'b_i' + q_{i+1}a_i'b_i),$$

$$v_i = q_i'(q_{i+1}'a_i'b_i + q_{i+1}a_ib_i' + q_{i+1}a_ib_i).$$

Again, using the ring-sum notation, these outputs can be manipulated to the equivalent expressions:

$$\begin{aligned} w_i &= q_i'(a_i' + b_i')(q_{i+1} \oplus a_i), \\ v_i &= q_i'(a_i + b_i)(q_{i+1} \oplus a_i)'. \end{aligned} \tag{6}$$

This form is convenient because forming the ring-sum of the ith digits by the combination $a_i \oplus b_i = (a_i + b_i)(a_i' + b_i')$ allows some of the operations in the carries and outputs to be shared.

Fig. 3 shows the ith digit of a one-carry comparator utilizing the logic of (5) and (6). Since all runs have been eliminated, it is not possible to have adjacent output digits of opposite polarity. Nor is it possible in have scheme to have both polarity outputs occurring together. These properties follow from the equations because the logical products $w_i v_i$, $w_i v_{i-1}$ and $w_{i-1} v_i$ vanish identically. Under these conditions, subtraction of the w_i and v_i digits can be performed in an analog subtracter such as shown in Fig. 2 without encountering the precision problem discussed previously.

3.3 *A Two-Carry Binary-Binary Comparator*

It should be apparent, after careful examination of the subtraction process for two binary numbers and paying particular attention to the runs described in Section 3.1, that many schemes are possible for translating the input numbers A and B into an equivalent set W and V meeting the requirement demanded by analog subtraction. In the previous section we derived a comparator logic using a single carry operating as an inhibit function. In this section we will outline the synthesis of a comparator using two carries which perform the function of permitting

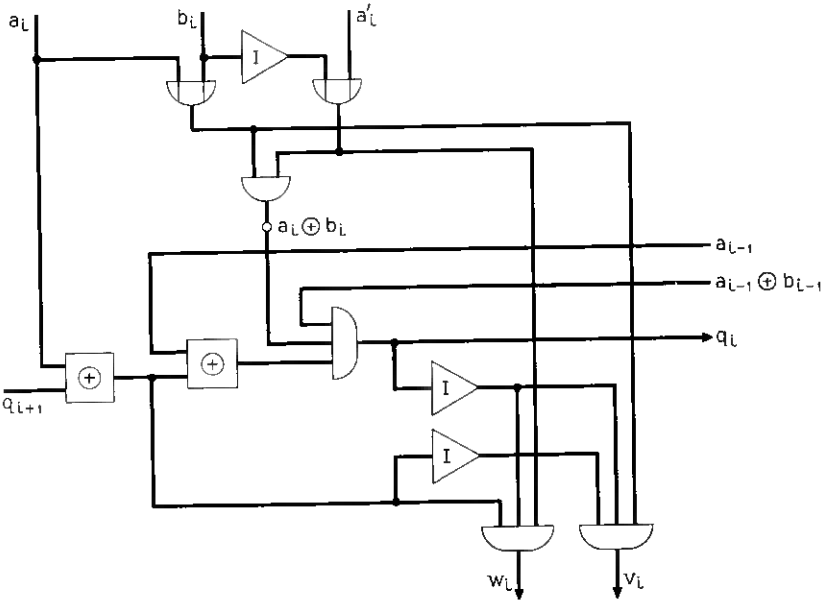


Fig. 3 — One-carry binary-binary comparator.

the outputs under proper conditions. This scheme has a somewhat simpler circuit and has the advantage that the logic can be modified to operate with one of the input numbers in the Gray code. Also, if the difference is required to be only approximately proportional to the exact difference, while still meeting the precision requirements of analog subtraction, then the logic can be further simplified.

Since an output should be permitted at the ith digit only when there is a mismatch between the digits a_i and b_i, a necessary condition for any carry which permits an output should be the ring-sum $a_i \oplus b_i = 1$. As described in Section 2.1, this logical operation detects a mismatch of either polarity. Let the ith digit of the carry which permits a positive output w_i be designated n_i and the ith digit of the carry permitting a negative output v_i be designated m_i.

If there are no carries coming into the ith digit from the next more significant digit, i.e., $n_{i+1}'m_{i+1}' = 1$, and if the mismatch is of the type $a_ib_i' = 1$, then the positive carry should be started. Similarly, if there are no carries coming in and the mismatch is of the type $a_i'b_i = 1$, then the negative carry should be started. Also, if there is a mismatch in the ith digit, i.e., $a_i \oplus b_i = 1$, with a carry coming in from the next more significant digit, then it should be propagated through this ith digit in order to permit detection of the runs discussed in Section 3.1.

These rules could again be summarized in a truth table from which the required functions could be derived. However, it should be apparent from the previous description that the carries satisfy the following functions:

Positive carry:

$$n_i = (a_i \oplus b_i)(n_{i+1}'m_{i+1}'a_ib_i' + n_{i+1}),$$

Negative carry:

$$m_i = (a_i \oplus b_i)(n_{i+1}'m_{i+1}'a_i'b_i + m_{i+1}).$$

These functions can be simplified by algebraic manipulations. In particular, using the identities $x + x'y = x + y$ and $xy'(x \oplus y) = xy'$, the carries can be rewritten as

$$\begin{aligned} n_i &= n_{i+1}(a_i \oplus b_i) + m_{i+1}'a_ib_i', \\ m_i &= m_{i+1}(a_i \oplus b_i) + n_{i+1}'a_i'b_i. \end{aligned} \tag{7}$$

Note from the development of the carry structure that both carries cannot exist together. We can show this by noting from (7) that

$$\begin{aligned} n_km_k &= n_{k+1}m_{k+1}(a_k \oplus b_k) \\ &= n_{k+2}m_{k+2}(a_k \oplus b_k)(a_{k+1} \oplus b_{k+1}) \\ &= n_mm_m(a_k \oplus b_k) \cdots (a_{m-1} \oplus b_{m-1}). \end{aligned}$$

One more substitution gives a factor $n_{m+1}m_{m+1}$ which is always zero, since the subscript m denotes the most significant digit. Therefore $n_k m_k = 0$, for all k.

Outputs of a particular polarity are possible only when the carry permitting that polarity is present. Thus, a necessary condition for a positive output w_i is $n_i = 1$ and a condition for a negative output v_i is $m_i = 1$. Whenever we are in a run, no output should occur until the run is terminated. A positive run is terminated at the ith digit when $n_{i-1} = 0$ or $a_{i-1} = 1$ and a negative run is terminated at the ith digit when $m_{i-1} = 0$ or $a_{i-1} = 0$. If we are not in a run but a carry is present, then there must also be a mismatch. When $n_i = 1$ and $a_{i-1} = 1$ a positive output w_i is needed, and when $m_i = 1$ and $a_{i-1} = 0$ a negative output v_i is needed. Forcing as much logic as possible on one of the input numbers will be to our advantage when this comparator is modified to operate with one of the input numbers changing rapidly. The output functions satisfying the previous description are

$$\begin{aligned} w_i &= n_i(n_{i-1}' + a_{i-1}), \\ v_i &= m_i(m_{i-1}' + a_{i-1}'). \end{aligned} \tag{8}$$

The following example illustrates the formation of the carries N and M and the outputs W and V:

A	1	0	0	1	1	0	0	1	1
B	0	1	1	0	1	1	0	0	0
N	1	1	1	1	0	0	0	1	1
M	0	0	0	0	0	1	0	0	0
W	0	0	1	1	0	0	0	1	1
V	0	0	0	0	0	1	0	0	0

Fig. 4 shows a typical digit of a two-carry comparator using the logic of (7) and (8).* This circuit, with carries propagating from the more significant digits, is suitable for driving an analog subtracter, since all runs have been eliminated and outputs of equal weight and opposite polarity are not possible. The circuit also has the advantage that much of the logic is performed on only one of the input numbers.

3.4 *A Two-Carry Binary-Gray Comparator*

For those applications in which one of the input numbers can be changing rapidly the multiple digit changes of the binary code lead to the

* Equations (7) and (8) can be manipulated to an equivalent scheme obtained by Ketchledge[2] from other considerations.

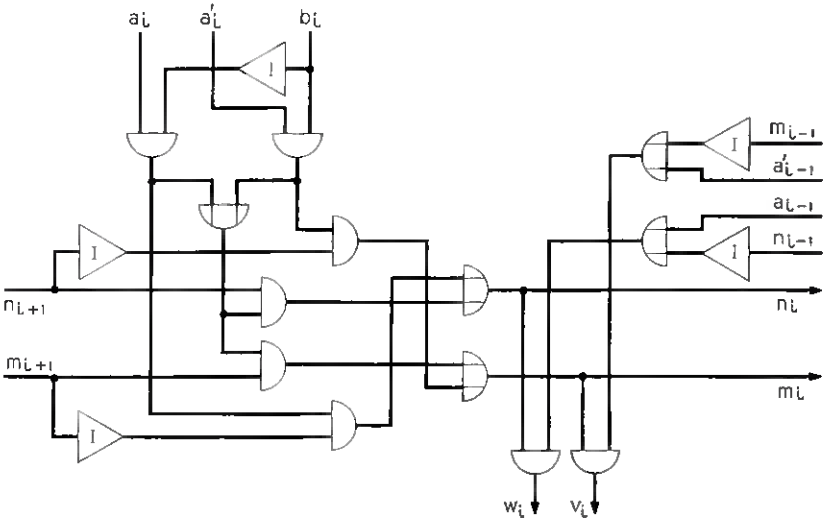


Fig. 4 — Two-carry binary-binary comparator.

difficulties discussed in Section 2.2. In this case, it is desirable to use the Gray code for the changing number and, since the cyclic reflected-binary type Gray code has the simple translation property given by (3) and (4), a translation of the logic of Section 3.3 can be easily carried out. If we make the proper translation of the binary input number B into a Gray number G, then the outputs W and V remain unaltered.

Equation (3) gives the translation from binary code to the Gray code used here. From the logic of Section 3.3, if there is a carry coming into the ith digit, i.e., $n_{i+1} = 1$ or $m_{i+1} = 1$, then there must have been a mismatch of the previous digits, i.e., $a_{i+1} \oplus b_{i+1} = 1$. Using the ring-sum properties given by (2), the ith Gray digit under these conditions can be expressed as

$$\begin{aligned} g_i' &= b_i \oplus b_{i+1} \oplus a_{i+1} \oplus b_{i+1} \\ &= b_i \oplus a_{i+1}\,. \end{aligned}$$

Rearranging terms gives

$$b_i = (g_i \oplus a_{i+1})' \qquad \text{if} \qquad a_{i+1} \oplus b_{i+1} = 1.$$

Similarly, if there is a match of the preceding digits, or $a_{i+1} \oplus b_{i+1} = 0$, then the carries coming in must also be zero, i.e., $n_{i+1} = 0$ and $m_{i+1} = 0$.

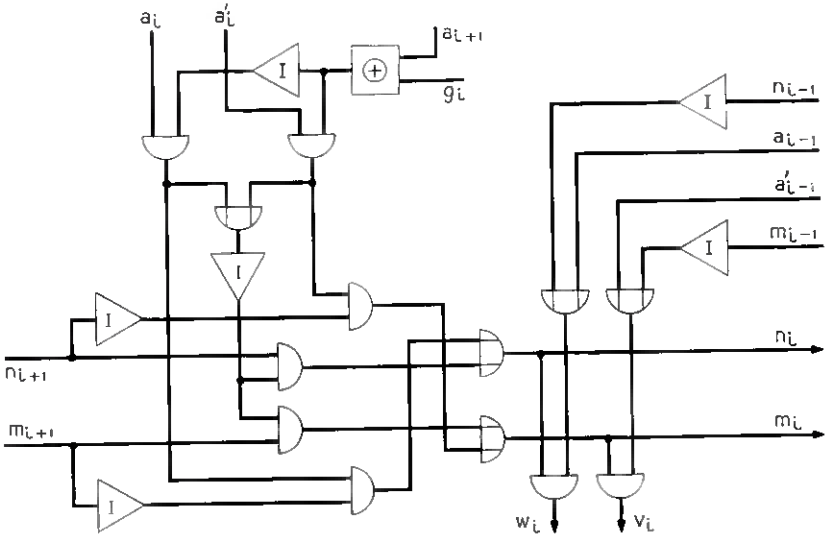


Fig. 5 — Binary-Gray comparator.

Under these conditions, the ith Gray digit can be expressed as

$$g_i = b_i \oplus a_{i+1},$$

or

$$b_i = g_i \oplus a_{i+1} \qquad \text{if} \qquad a_{i+1} \oplus b_{i+1} = 0.$$

These translations can then be substituted in the carries of (7) under the proper conditions determined by the incoming carries. The resulting logic scheme for the binary-Gray comparator is given by the following equations* and appears in Fig. 5:

$$\begin{aligned} n_i &= n_{i+1}(a_i \oplus a_{i+1} \oplus g_i)' + m_{i+1}'a_i(a_{i+1} \oplus g_i)', \\ m_i &= m_{i+1}(a_i \oplus a_{i+1} \oplus g_i)' + n_{i+1}'a_i'(a_{i+1} \oplus g_i), \\ w_i &= n_i(n_{i-1}' + a_{i-1}), \\ v_i &= m_i(m_{i-1}' + a_{i-1}'). \end{aligned} \tag{9}$$

* These equations can also be manipulated to an equivalent scheme obtained by Ketchledge[2] from other considerations.

IV. APPROXIMATE PROPORTIONAL COMPARATORS

4.1 *Approximate Binary-Binary Comparator*

In some applications of digital comparators it is not necessary to obtain the exact difference. For example, a digital servo will operate satisfactorily if the comparator or error detector supplies a control signal which increases with increasing error and decreases for decreasing error, but which is not necessarily equal to the true error difference, except for small errors. It is difficult to build circuits using the logic of the previous sections which meet very high speed requirements and it is desirable to simplify the circuits as much as possible. The logic schemes derived in this and the following section are attempts to simplify the circuitry of the exact proportional comparators synthesized previously. Again, many variations are of course possible. The particular binary-binary approximate comparator synthesized in this section meets the further requirement of simple translation to a binary-Gray approximate comparator.

Consider the subtraction of two binary numbers A and B where the most significant output occurs in the kth digit. Then $|A - B|$ will be maximum if all the following digits $(k-1, k-2, \cdots, 1, 0)$ have the same polarity outputs. With the run structure determining the outputs as described in Section 3.1, it is not possible to have an uninterrupted sequence of opposite polarity outputs start in the next digit. Therefore $|A - B|$ will be minimum if all the digits starting two lower $(k-2, \cdots, 1, 0)$ have the opposite polarity outputs. From these considerations,

$$|A - B|_{\max} = 2^k + \sum_{i=0}^{k-1} 2^i = 2^k + (2^k - 1) = 2^{k+1} - 1,$$

$$|A - B|_{\min} = 2^k - \sum_{i=0}^{k-2} 2^i = 2^k - (2^{k-1} - 1) = 2^{k-1} + 1,$$

and therefore

$$\tfrac{1}{2} \cdot 2^k < |A - B| < 2 \cdot 2^k \tag{10}$$

whenever the most significant difference output occurs in the kth digit. This means that, if only the most significant difference digit is allowed, the result will always be within a factor of two of the exact difference. The same result will be obtained if the output digits of one polarity, say w_i, are cancelled by the appearance of digits of the opposite polarity v_i for $i < k$. These should of course not be strict rules in the synthesis

of approximate comparators since allowing some lower-order output digits may frequently improve the approximation.

When forming carries starting from the more significant digits it is still necessary to detect combinations of digits forming runs, as discussed in Section 3.1. This restriction avoids some of the difficulties when analog subtraction is to be performed on the new output difference digits. Two carries will again be used, with the property of determining outputs of the correct polarity in the proper digits.

The logic of this scheme is simplified if all carries are allowed to propagate unaltered through digits of lower significance. One of the conditions for a positive carry n_i is then an incoming carry, or $n_{i+1} = 1$, and a condition for a negative carry m_i is $m_{i+1} = 1$. If the most significant mismatch occurs in the ith digit it must necessarily be preceded by $a_{i+1} \oplus b_{i+1} = 0$. Under these conditions, a positive carry is started if $a_i b_i' = 1$ and a negative carry is started if $a_i' b_i = 1$.

Information concerning the end of a run, and therefore the need for an output, in the exact comparators discussed previously, was contained in the carries and the address number. Examination of the conditions ending a positive-type run, e.g., (1000) − (0111), at the ith digit indicates that it is necessary to have $a_{i-1}' b_{i-1} = 0$. One of the conditions is therefore $a_{i-1} = 1$. This condition will be used in the output expressions derived later. The other necessary condition for ending a positive type run at the ith digit is $a_{i-1}' b_{i-1}' = 1$. This condition will be used to start a carry of the opposite polarity. In other words, we start a negative carry m_i if there is a mismatch of the previous digits, i.e., $a_{i+1} \oplus b_{i+1} = 1$, and $a_i' b_i' = 1$. Similarly, the conditions for ending a negative-type run, e.g., (0111) − (1000), at the ith digit is $a_{i-1} = 0$ or $a_{i-1} b_{i-1} = 1$. The first condition will also be used in forming the outputs and the second condition will be used to start a positive carry. Thus, *after* the end of a run both carries may be present. The presence of a carry at the ith digit is given by the OR of the above conditions.

$$n_i = n_{i+1} + a_i b_i'(a_{i+1} \oplus b_{i+1})' + a_i b_i (a_{i+1} \oplus b_{i+1}),$$

$$m_i = m_{i+1} + a_i' b_i (a_{i+1} \oplus b_{i+1})' + a_i' b_i' (a_{i+1} \oplus b_{i+1}).$$

And, since $x'y' + xy = (x \oplus y)'$, these expressions can be rewritten as

$$\begin{aligned} n_i &= n_{i+1} + a_i (a_{i+1} \oplus b_{i+1} \oplus b_i)', \\ m_i &= m_{i+1} + a_i' (a_{i+1} \oplus b_{i+1} \oplus b_i). \end{aligned} \tag{11}$$

Outputs are again formed only when carries are present. That is, for a positive output w_i we need n_i and for a negative output v_i we need m_i.

However, first outputs in the case of a run are never allowed until the last digit of a run. The end of a positive-type run at the ith digit, as explained above, can be detected by $a_{i-1} = 0$, or by the start of the opposite type carry in the next digit, i.e., $m_{i-1} = 1$. Similarly, the end of the first negative-type run at the ith digit can be detected by $a_i = 0$ or n_{i-1}. Outputs are therefore given by

$$\begin{aligned} w_i &= n_i(a_{i-1} + m_{i-1}), \\ v_i &= m_i(a_{i-1}' + n_{i-1}). \end{aligned} \tag{12}$$

Note that the logic of (11) and (12) also allows a proper output at the most significant mismatch which is not the start of a run. Such a scheme will give an approximate difference when the outputs are fed into an analog subtracter of the type shown in Fig. 2. This difference will always be within a factor of two of the exact difference, and the difficulties inherent in analog subtraction when a run occurs in the original input numbers, as explained in Section 3.1, do not appear. The following examples showing formation of the most significant outputs may help to clarify the process:

A	1 0 0 1 0	A	1 0 0 1 0	A	1 0 0 0 0	A	1 0 0 1 1
B	0 1 1 1 0	B	0 1 1 0 0	B	0 1 1 0 0	B	0 0 1 1 1
N	1 1 1 1 1	N	1 1 1 1 1	N	1 1 1 1 1	N	1 1 1 1 1
M	0 0 0 0 0	M	0 0 0 0 1	M	0 0 0 1 1	M	0 1 1 1 1
W	0 0 1 1 1	W	0 0 1 1 1	W	0 0 1 1 1	W	1 1 1 1 1
V	0 0 0 0 0	V	0 0 0 0 1	V	0 0 0 1 1	V	0 1 1 1 1

Fig. 6 shows the circuit for the ith digit of this approximate proportional binary-binary comparator. Note that it is possible in this scheme to have outputs of both polarities appearing together. This may also lead to difficulties in the analog subtraction of the W and V numbers. A simple way to avoid this is to allow outputs only when there is exactly one carry present and to inhibit all outputs when both carries are present. The output expressions given by (12) should then be modified as follows:

$$\begin{aligned} w_i &= n_i m_i'(a_{i-1} + m_{i-1}), \\ v_i &= m_i n_i'(a_{i-1}' + n_{i-1}). \end{aligned} \tag{13}$$

For this type of output the circuit of Fig. 6 should be modified by adding the dotted-line inputs to the output gates. This modified scheme still gives an approximate difference within a factor two as determined

by (10). It also avoids subtracting digits of equal weight in an analog subtracter. In fact, if an analog difference is not required, the numbers W and V can be used directly as the approximate digital difference.

4.2 *Approximate Binary-Gray Comparator*

Difficulties of using the binary code for an input number which is rapidly changing were pointed out in Section 2.2. These difficulties can be avoided by translating the varying number into the Gray code, as was demonstrated in the synthesis of exact comparators. Again we assume only the input B to be rapidly changing.

Examination of the logic synthesized for the binary-binary comparator of Section 4.1 shows that digits of the number B appear only in the expressions for the carries and that the only combination of these digits is $b_i \oplus b_{i+1}$. From (3), this is exactly the expression used for translating from binary to Gray. By direct substitution, the expressions for the carries therefore become

$$\begin{aligned} n_i &= n_{i+1} + a_i(a_{i+1} \oplus g_i)' \\ m_i &= m_{i+1} + a_i'(a_{i+1} \oplus g_i), \end{aligned} \tag{14}$$

and the equations for the outputs are not altered.

Using (12) for the outputs and (14) for the carries results in the approximate binary-Gray comparator circuit shown in Fig. 7. As with the

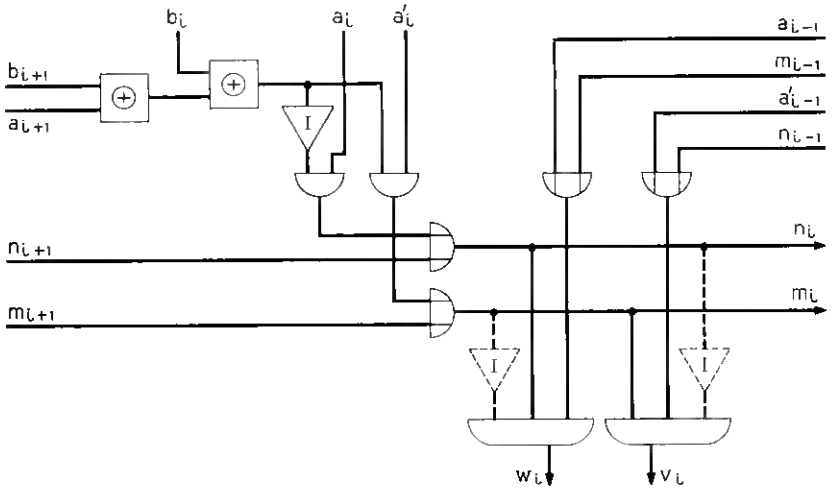


Fig. 6 — Approximate binary-binary comparator.

binary-binary comparator of the previous section, we can again inhibit all outputs during the presence of both carries by using the dotted connections shown.

V. SIGN-ONLY COMPARATORS

5.1 *Modifications of Approximate Comparators*

The approximate comparators of Section IV were largely the result of simplifying the exact comparators of Section III. In this section we will examine the approximate comparators and attempt to simplify them further. All of the previous schemes gave an output which was proportional to the difference between the two input numbers and also indicated the sign of this difference. If only the sign is required, with no measure of the magnitude of the difference, then it is possible to synthesize simple schemes giving an output on a single lead when one of the input numbers is less than or equal to the other input number, or when one number is greater than the other. No analog conversion is necessary in this case, since the output gives the desired indication directly. However, if the sign-only comparator is used as an error detector in a digital servo it is still necessary to have a linear error signal when close to zero difference.

Considering again the schemes having both inputs in the binary code, it is apparent that the sign of the difference is determined completely

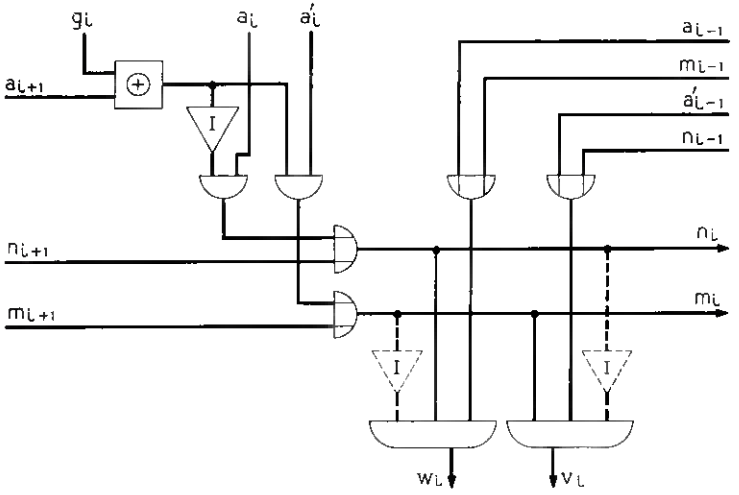


Fig. 7 — Approximate binary-Gray comparator.

by the direction of the most significant mismatch. In the proportional comparators it was desirable to examine the adjacent lower significant digits for the start of a possible run and to prohibit any output until the end of a run. Since the sign is determined only by the sign of the first mismatch it is not necessary to account for any run structure in sign-only comparators and the logic is therefore simpler. Whenever the most significant mismatch is detected, an output should be formed immediately. In addition, we should avoid any contrary action due to opposite polarity mismatches in the lower significant digits. This function can be performed by carries propagating toward the lower significant digits.

These ideas can be illustrated by an example with both inputs in the binary code. Let the output of the ith digit, designated u_i, be zero for all digits which precede the first mismatch and also be zero during and after the first mismatch if it is of the form $a_i'b_i = 1$. Let $u_i = 1$ for the first mismatch if it is of the form $a_ib_i' = 1$. Then an OR over all u_i will provide the proper output:

$$\begin{aligned} \underset{0 \leqq i \leqq m}{\text{OR}} \ (u_i) &= 1 \quad \text{if} \quad A > B \\ &= 0 \quad \text{if} \quad A \leqq B. \end{aligned} \tag{15}$$

Each u_i is determined by the carries present in that digit. A positive carry n_i should be formed for a positive mismatch, i.e., $a_ib_i' = 1$; and a negative carry m_i formed for a negative mismatch, i.e., $a_i'b_i = 1$. If these carries are allowed to propagate through lower significant digits and an output is formed in the ith digit only if a positive carry is present but not a negative carry, then the comparator output given by (15) will result. The expressions for the carries in the ith digit and the output at the ith digit are

$$\begin{aligned} u_i &= n_im_i', \\ n_i &= n_{i+1} + a_ib_i', \\ m_i &= m_{i+1} + a_i'b_i. \end{aligned} \tag{16}$$

The circuit for this logic is very simple, requiring three AND gates and two OR gates, each with two inputs for each digit.

A similar scheme when one of the input numbers is in the Gray code can be obtained from the same type of reasoning by using the carries of (14) in Section 4.2. Again, the output in the ith digit is $u_i = n_im_i'$. The circuit for this scheme is given in Fig. 8.

An improvement of the logic used in Fig. 8 follows from a close ex-

amination of the carries n_i and m_i. The functions of these carries in this sign-only comparator can be summarized:

i. For all initial match digits there are no carries and therefore no outputs.

ii. Only one carry is present at the *first* mismatch. This carry determines the output.

iii. If an opposite polarity mismatch occurs *after* the first mismatch a second carry is formed. This second carry inhibits all outputs.

Evidently then, one carry has been used to permit outputs and the other carry to inhibit outputs. These two operations could be performed equally well by one carry if the output function were properly chosen. When a positive mismatch occurs first at the ith digit, we require a positive output, i.e., $u_i = 1$. From the positive carry of Section 4.2, a positive mismatch at the ith digit is detected by $a_i(a_{i+1} \oplus g_i)' = 1$; however, this output should be inhibited if a previous negative mismatch has occurred. Therefore, we could use the negative carry of Section 4.2

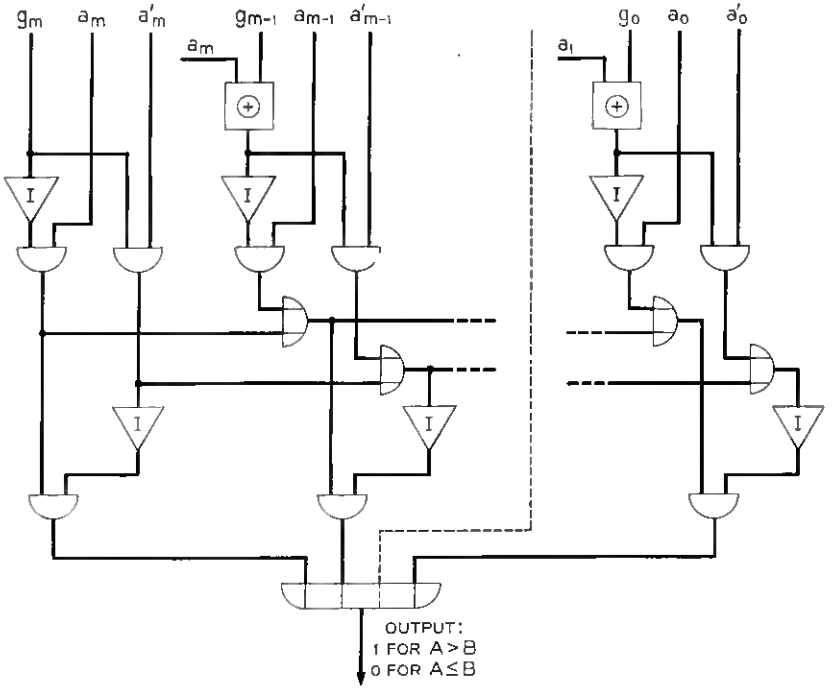


Fig. 8 — Sign-only binary-Gray comparator.

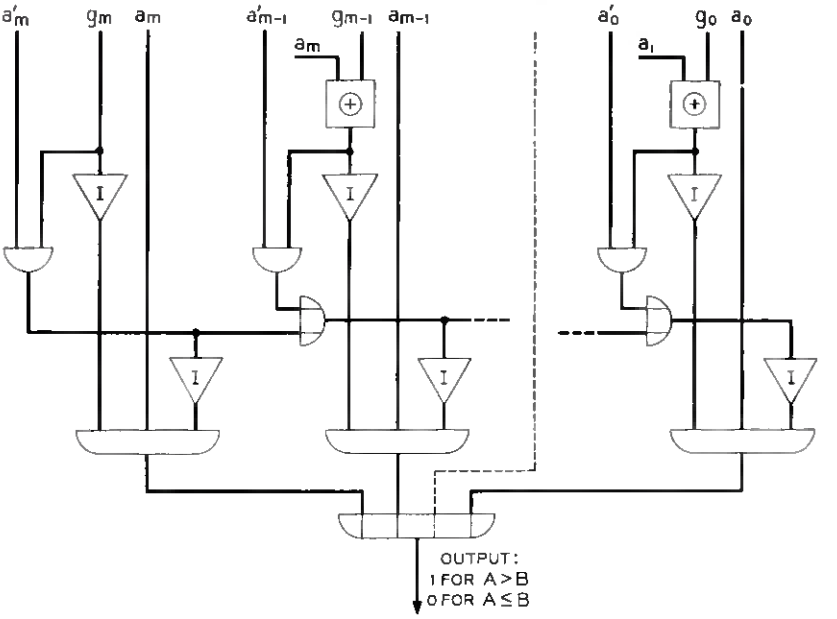


Fig. 9 — Alternate sign-only binary-Gray comparator.

as an inhibit function on u_i. The logic for this scheme then appears as

$$\begin{aligned} u_i &= a_i(a_{i+1} \oplus g_i)'m_i', \\ m_i &= m_{i+1} + a_i'(a_{i+1} \oplus g_i). \end{aligned} \tag{17}$$

These equations, together with (15), have the circuit shown in Fig. 9.

5.2 *Other Sign-Only Comparators**

Assume that the first mismatch occurs between a_j and b_j; that is, all the digits a_i and b_i for $i > j$ match in corresponding places. If $A > B$, then this first mismatch will be of the form $a_j = 1$ and $b_j = 0$. Therefore

$$a_j b_j' = 1, \quad \text{or} \quad a_j + b_j' = 1, \quad \text{if } A > B. \tag{18}$$

Similarly, if $A < B$, then this first mismatch will be of the form $a_j = 0$ and $b_j = 1$. Therefore

$$a_j b_j' = 0, \quad \text{or} \quad a_j + b_j' = 0, \quad \text{if } A < B. \tag{19}$$

* The comparators derived in this section were previously obtained by Ketchledge[2] from other considerations.

For all digits of higher significance, that is, for all $i > j$, we have

$$a_i b_i + a_i' b_i' = 1.$$

This is equivalent to

$$(a_i + b_i')(a_i b_i')' = 1. \tag{20}$$

Since (20) holds for all $i > j$ up to m, we can write

$$(a_m + b_m')(a_{m-1} + b_{m-1}') \cdots (a_{i+1} + b_{i+1}')(a_i + b_i') = 1 \tag{21}$$

and

$$a_m b_m' + a_{m-1} b_{m-1} + \cdots + a_{i+1} b_{i+1}' + a_i b_i' = 0. \tag{22}$$

Now consider the function

$$\Phi_k = (a_k + b_k') + a_m b_m' + a_{m-1} b_{m-1}' + \cdots + a_{k+1} b_{k+1}' \tag{23}$$

for all possible values of k for the conditions $A = B$, $A > B$, and $A < B$.

Then, if $A = B$, from (21), $\Phi_k = 1$ for all k. That is, the digits match in each place so that (20) is 1 for all i.

If $A > B$ from (21) and (22), $\Phi_k = 1$ for all k. That is, the first mismatch will be of the form $a_j b_j' = 1$, by (18). Therefore, $\Phi_j = 1$ and, since the digits match for all $k > j$, we have $a_k = b_k$ or $a_k + b_k' = 1$, which is the first term in Φ_k. Also, Φ_k for all $k < j$ will include the term $a_j b_j' = 1$ as an OR term and will therefore be 1.

If $A < B$ from equations (21) and (22), $\Phi_k = 0$ for $k = j$. That is,

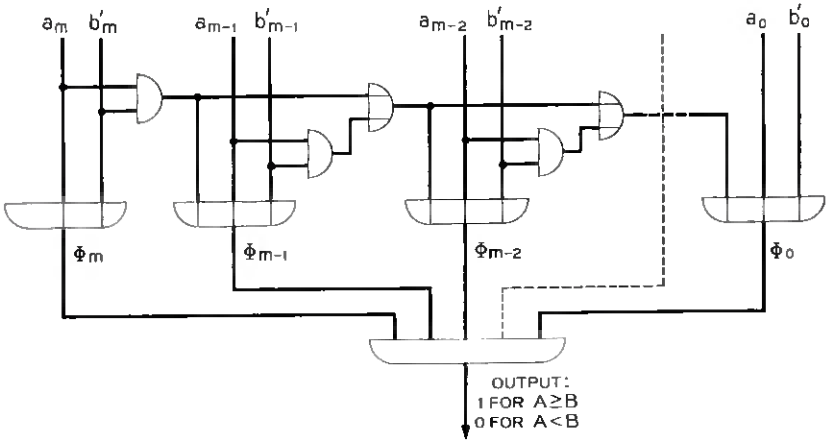


Fig. 10 — Sign-only binary-binary comparator.

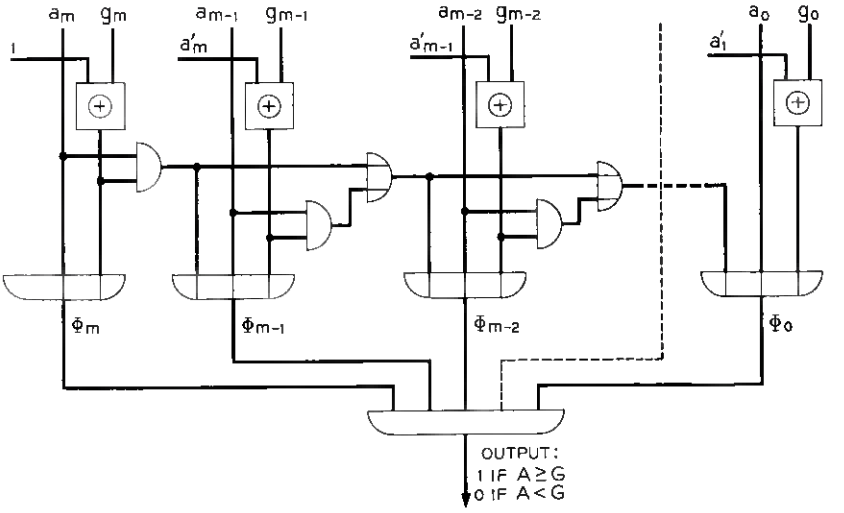


Fig. 11 — Sign-only binary-Gray comparator.

from (19), $a_j + b_j' = 0$, and, from (22), $a_m b_m' + \cdots + a_{j+1} b_{j+1}' = 0$ if the first mismatch occurs in the jth digit.

If Φ_k is generated in each digit we then have a function which is always 1 if $A \geqq B$ and which will be 0 for at least one digit if $A < B$. Therefore, the desired sign detector can be obtained by forming the AND over all the Φ_k :

$$\begin{aligned} \underset{m \geqq k \geqq 0}{\text{AND}} \ (\Phi_k) &= 1 \quad \text{for} \quad A \geqq B \\ &= 0 \quad \text{for} \quad A < B. \end{aligned} \tag{24}$$

Equations (23) and (24) are implemented in the circuit shown in Fig. 10. As in the previous developments, it is desirable to modify this scheme to operate with one of the input numbers in the Gray code. To transform b_i to g_i, we note that, if the first mismatch occurs in the jth digit, then all the preceding digits match. Thus, for $i > j$

$$b_i = g_i \oplus b_{i+1} = g_i \oplus a_{i+1}$$

and

$$b_i' = g_i \oplus a_{i+1}'. \tag{25}$$

Substituting in (23) gives

$$\begin{aligned} \Phi_k = {} & [a_k + (g_k \oplus a_{k+1}')] + a_m g_m' \\ & + a_{m-1}(g_{m-1} \oplus a_m') + \cdots + a_{k+1}(g_{k+1} \oplus a_{k+2}'). \end{aligned} \tag{26}$$

As previously, if $A = G$ in magnitude, then either $a_k = 1$ and $g_k \oplus a_{k+1}' = 0$ or $a_k = 0$ and $g_k \oplus a_{k+1}' = 1$ for $0 \leqq k \leqq m$. That is, $\Phi_k = 1$ for all k. If $A > G$ and the first mismatch occurs for $i = j$, then $\Phi_k = 1$ for $k > j$, $a_j = 1$ and $g_j + a_{j+1}' = 1$. Therefore, $\Phi_j = 1$, since the term $a_j(g_j \oplus a_{j+1}') = 1$ for all k if $A > G$. If $A < G$ and the first mismatch occurs for $i = j$, then, by transforming (19), we have $a_j + (g_j \oplus a_{j+1}') = 0$, and therefore $\Phi_j = 0$.

By forming the AND over all k of the function given in (26) we have the desired sign detection:

$$\begin{aligned} \underset{m \geqq k \geqq 0}{\text{AND}} \ (\Phi_k) &= 1, \quad \text{for} \quad A \geqq G \\ &= 0, \quad \text{for} \quad A < G. \end{aligned} \tag{27}$$

A circuit performing the operations of (26) and (27) is shown in Fig. 11. The circuits of Figs. 10 and 11 could be modified by using the duals of the above expressions and changing the output gate to an OR over all k. The individual digit functions Φ_k would then be changed to AND's of each of the dual terms.

REFERENCES

1. Hoover, C. W., Jr., Staehler, R. E. and Ketchledge, R. W., Fundamental Concepts in the Design of the Flying Spot Store, B.S.T.J., **37**, September 1958, p. 1161.
2. Ketchledge, R. W., this issue, pp. 1–17.
3. Richards, R. K., *Arithmetic Operations in Digital Computers*, D. Van Nostrand Co., New York, 1955, Ch. 9 and 10.
4. Keister, W., Ritchie, A. E. and Washburn, S. H., *The Design of Switching Circuits*, D. Van Nostrand Co., New York, 1951, Ch. 5.
5. Foss, F. A., The Use of a Reflected Code in Digital Control Systems, Trans. I.R.E., **EC-3**, December 1954, p. 1.
6. Gray, F., Patent No. 2,632,058, March 17, 1953.
7. Gilbert, E. N., Gray Codes and Paths on the n-Cube, B.S.T.J., **37**, May 1958, p. 815.
8. Flores, I., Reflected Number Systems, Trans. I.R.E., **EC-5**, June 1956, p. 79.

The Laddic — A Magnetic Device for Performing Logic

By U. F. GIANOLA and T. H. CROWLEY

(Manuscript received August 18, 1958)

The Laddic is a ladder-like structure cut out of a rectangular hysteresis-loop ferrite. The sides of the ladder and all of the rungs are equal in minimum cross section so that all possible paths are flux-limited. The structure presents a large number of possible flux paths. By controlling the actual switching path through the structure any Boolean function of n variables can be produced.

A number of methods of operation are discussed, and design formulae and experimental results presented. One of the attractive features of this device is that the operating currents are not critical. Therefore, it can be operated at speeds limited essentially only by the current drives available. The output may be taken during the input variable phase or during a subsequent reset phase. Switching speeds of a few tenths of a microsecond and repetition rates of a few hundred kilocycles have been achieved.

I. INTRODUCTION

Toroidal cores of magnetic material having a rectangular hysteresis loop are widely used as memory and switching elements in logic circuitry.[1] The possibility of simplifying core circuitry by using more complicated cores has occurred to a number of people, and several multihole core devices for specific applications have been described in the literature.[2, 3, 4] The present work was initiated to explore the possible systematic use of the magnetic "linkages" between flux patterns in a multihole magnetic structure. In particular, it was hoped that core circuits could be simplified by replacing the function of coupling windings between individual cores by the magnetic "linkages". A structure containing a continuous network of flux-limited paths was considered, and a generalized technique was developed for realizing any class of Boolean switching function* in combinational logic by controlling the switching

* The use of Boolean notation and algebra is described, for example, in Ref. 1.

path through the network. For specific switching functions there are a large number of geometries which can be used. In the present paper one structure that can be used generally for all combinational logic will be described. This specific structure resembles a ladder, Fig. 1, and has been named "Laddic," an abbreviation for "ladder-logic."

II. OUTLINE OF PAPER

The general principle of operation of the Laddic is discussed in Section III. Briefly, the structure shown in Fig. 1 is made out of a rectangular hysteresis loop material, for example, a memory-core ferrite. The cross sections of its rungs are all equal and the cross sections of the side rails are preferably equal to that of the rungs, but may be greater. It is found that, starting from a suitable saturation flux pattern, a drive applied so as to switch flux in the first rung will switch the flux almost entirely through the closest available rung rather than split it among all available rungs.

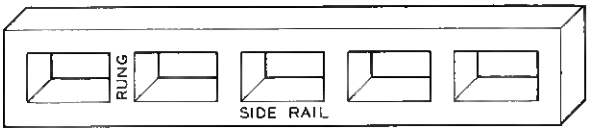


Fig. 1 — Basic Laddic structure.

In Section IV it is shown how the Laddic is used to generate Boolean functions. Briefly, the procedure is as follows. A suitable remanent flux pattern is first established by a current pulse through a reset winding. One unit of flux, corresponding to the saturation flux through one rung, is then reversed in the first rung by applying a clock-pulse current to an appropriate winding. An output winding is placed on some other rung — the last rung, for example. The reset flux pattern was such that the reversed flux preferentially chooses return paths other than through the output rung. Thus, the output is normally zero. However, if all the alternative paths are blocked by inhibiting fields produced by current pulses representing input variables, the switched flux must return through the output rung, and an output will be obtained. It will be shown that any Boolean function can be realized as the output of a single Laddic of suitable length. Of course, there are practical limitations on the size of the structure and, therefore, to the number of variables that can actually be handled. Modified circuits and modified structures will also be discussed.

In Section V some experimental results and design formulae are pre-

sented. Because all paths are flux-limited, the operating current margins in the Laddic are very broad. The variable currents must exceed a minimum value proportional to the clock current, but, practically speaking, they have no upper limit, which is a considerable advantage. Furthermore, the variable currents perform only an inhibiting function; that is, they are not required to switch flux. Thus, the back voltages induced in the variable input windings are very small, so that relatively low power sources may be used for the variable inputs. Because the input drives have no set maximum, the speed of operation is limited mainly by the abitrary maximum set for the drives. Using available materials and transistorized driving circuits, this means that switching speeds are normally in the region of 1 to 10 microseconds. The materials used are those developed for memory cores and, in general, the Laddic is compatible with core circuitry.

Section VI presents a general discussion of practical considerations.

III. PRINCIPLE OF THE LADDIC

The basic structure is that shown in Fig. 1. Three states of the material are considered. The first two are the remanent points for saturation in positive and negative senses, and correspond to the "1" and "0" states of a memory core. The third is the point of zero remanent magnetization, shown for illustration as state "2" in Fig. 2. It should be realized that

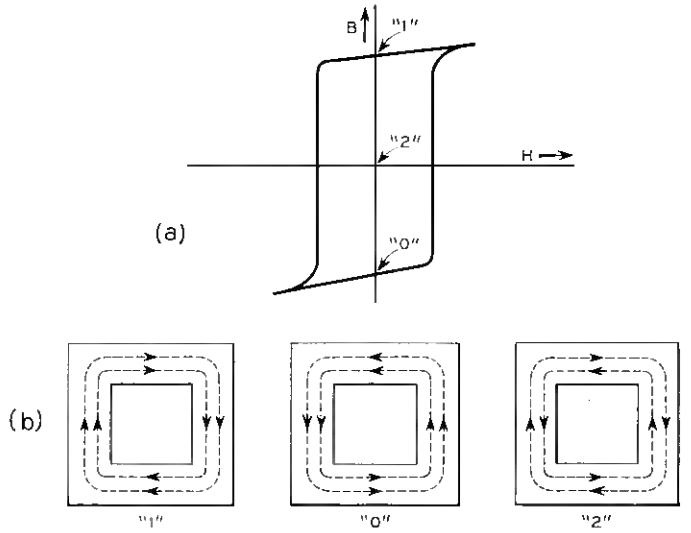


Fig. 2 — (a) Hysteresis loop showing the three states of the magnetization considered; (b) graphical representation of the three states.

state "2" is not arrived at by a sinusoidal demagnetization but by partial switching of the core from a saturated state. In the equi-flux networks being considered, a convenient model is to represent the state of magnetization of the material graphically by means of two parallel arrows, each in the direction of magnetization, and each representing one-half of the remanent saturation flux. Thus, if the arrows are in the same direction, the material is considered to be in one of the two saturated states. If they are in opposite directions, the resultant magnetization is zero and the material is considered to be in state "2". This enables one to represent the state of magnetization by closed flux patterns, as illustrated in Fig. 2(b). However, it should be kept in mind that the actual domain structure has not been observed experimentally. It is undoubtedly more complex than would follow from the simple flux patterns described, which are used solely for the purposes of a working model. Nevertheless, the model is found to be adequate for practical usage. The additional assumption that flux paths are closed within the structure, i.e., air-leakage is small, has also been found to be sufficient for the structures and materials considered.

Because of the flux-limited nature of the Laddic structure, when a reversal field is applied to the first rung the switched flux is returned

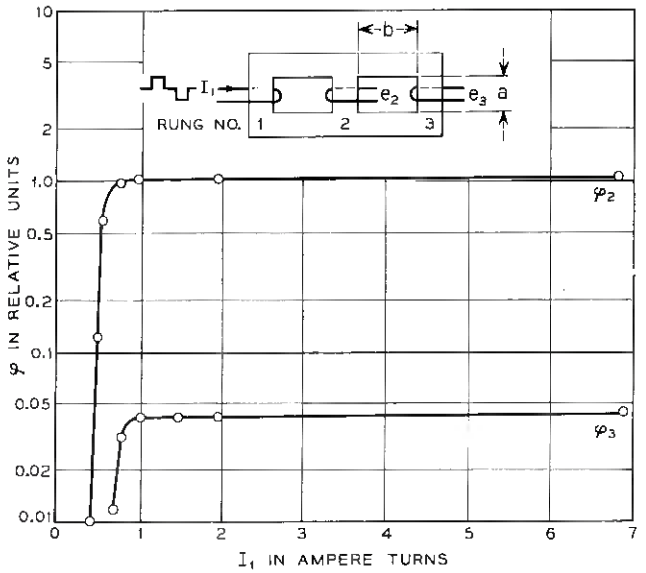


Fig. 3 — Proportions of the total flux $\omega = \int e dt$ switched through rungs 2 and 3 by a drive applied to rung 1.

TABLE I — φ_2/φ_3 FOR DIFFERENT GEOMETRIES

	b/a		
	1	2	3.5
φ_2/φ_3	21	83	710

almost entirely through the closest available path, no matter how large the applied drive. To illustrate this, consider the three-rung Laddic shown in the inset of Fig. 3. When alternately positive and negative current pulses are applied to the winding on rung 1, flux is switched alternately up and down. The switched flux is returned via rungs 2 and 3, the flux returned through each rung being φ_2 and φ_3 respectively. Fig. 3 shows experimental values of φ_2 and φ_3 versus the applied drive I_1. It will be seen that the flux ratio φ_2/φ_3 remains virtually constant once the minimum drive current, which is necessary to produce a full reversal, is exceeded. The ratio φ_2/φ_3 depends upon the geometry factor b/a, the ratio of rung spacing to side-rail spacing, as illustrated by Table I, which gives experimental values for a manganese magnesium ferrite. These ratios are much larger than might be expected. For example, considering the case $b/a = 1$, the mmf acting on rung 2 is three times as large as that acting on rung 3. Thus, because the rate of switching in a rectangular loop ferrite is proportional to the applied field, it might be expected that φ_2/φ_3 be more nearly 3:1 at drives much larger than threshold, rather than the 21:1 found experimentally. Obviously, the flux-splitting mechanism is complicated by the dynamic magnetic reluctances of the two paths. Unfortunately, the theoretical understanding of the switching mechanism in these materials is incomplete, so that a quantitative interpretation cannot be given. The important thing is that, under these conditions, the shortest return path containing flux that can be switched acts virtually as a magnetic short circuit.

It follows that the flux paths in the three-rung Laddic for the two cases considered can be represented approximately as in Figs. 4(a) and 4(b). As discussed previously, the flux in rung 3 is represented as being in the "2" state. Clearly, this is only an approximation to the true flux pattern because, in accordance with Table I, rung 2 will not in fact be fully saturated.

It is of interest to note that the flux pattern shown in Fig. 4(c) gives the same flux distribution in the rungs as does that in Fig. 4(a). However, there is a physical difference. The flux pattern of Fig. 4(c) is that obtained following several flux reversals by a drive applied to a winding on rung 2,

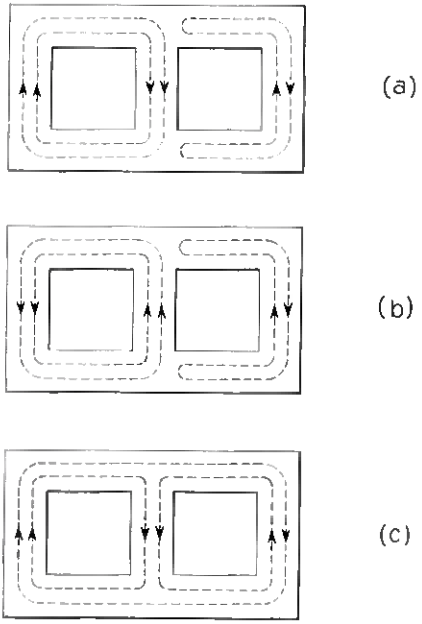


Fig. 4 — Three stable flux patterns in a three-rung Laddic.

whereas the flux pattern of Fig. 4(a) was obtained following several flux reversals by a drive applied to a winding on rung 1. If the initial flux pattern is symmetrical, as in Fig. 4(c), and a further reversal pulse is now applied to the rung 2 winding, the switched flux will divide equally between rung 1 and rung 3, as expected because of the symmetry. However, if there is a local flux closure, as in the initial flux pattern shown in Fig. 4(a), the switched flux shows a preference for the rung 1 return, so that there is a small but significant difference in the flux switched through rungs 1 and 3. This difference is small enough that it need not normally be taken into account in describing the Laddic as a device.

IV. USE OF THE LADDIC IN LOGIC CIRCUITS

4.1 *Basic Procedures*

The operation of the Laddic as an AND gate can now be explained in detail. A symmetrical flux pattern is first established by saturating odd-numbered rungs in the upward direction and even-numbered rungs in the downward direction, by pulsing a current through the reset wind-

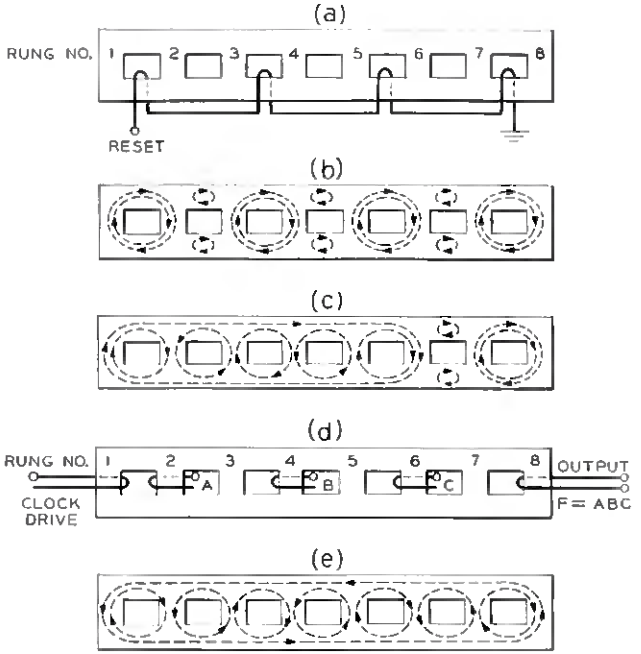


Fig. 5 — Applied drives and flux patterns for the normal mode of operation of the Laddic.

ing shown in Fig. 5(a).* The resulting flux pattern may take a number of forms, depending on the previous flux pattern. Figs. 5(b) and 5(c) illustrate two of these possibilities. The direction of magnetization of the rungs is the same in both cases, but the flux closures in the side rails are different. This difference produces only small differences in subsequent flux switching, as discussed previously, and need not be considered here.

If, following the reset, a current pulse is applied to a winding on rung 1, Fig. 5(d), in a direction to switch flux down, the flux return will be through the closest available path, i.e., through rung 2. However, if a current pulse corresponding to an input variable A is simultaneously present on rung 2, and is in a direction to hold it down, this rung is not available. Rung 3 is already saturated upwards, as are rungs 5 and 7, so these paths are also not available. Similarly, if inputs B and C are also present and hold the flux in their respective rungs, rungs 4 and 6

* The reset winding shown in Fig. 5(a) does not necessarily produce the exact flux pattern described. However, this turns out to be unimportant for normal use of the Laddic and, for simplicity, we have chosen to consider only the pattern shown. In any case, if desired, this pattern could be produced by a suitable winding.

are not available. Thus, if variables A, B and C are all present, then the return for flux switched in rung 1 must be through rung 8, and an output will result. It may be observed that the ouput voltage corresponds to a half reversal of the saturation flux, because of flux limiting by the side rail.

An output of the reverse polarity will be obtained during the subsequent reset phase; in this sense, the device also has memory.

The flux pattern obtained after a Boolean input ABC is shown in Fig. 5(e).

Obviously the "hold" currents that are necessary to prevent flux from being switched in the variable rungs, must exceed a certain minimum. However, because they serve only to hold an already saturated rung, there is no definite maximum. It follows that, if several separate windings are present on a hold rung, a current through one or more of them will suffice to hold the rung. For example, in Fig. 5(d) rung 2 could be held by individual currents through separate windings representing A_1, A_2, A_3, $\cdots$, A_n, so that the output would be obtained for inputs satisfying the Boolean equation $F = (A_1 + A_2 + A_3 + \cdots + A_n)BC$. It follows that a single Laddic can be used to generate any Boolean function of the form

$$(X_{11} + X_{12} + \cdots + X_{1n})(X_{21} + \cdots + X_{2n}) \cdots (X_{m1} + \cdots + X_{mn}). \quad (1)$$

Equation (1) is a general form that can represent any Boolean function if the X's may represent either the variables or their negations. In other words, any Boolean function can be generated by a single Laddic if current pulses are available for all of the variables and their negations.

Since a Boolean function can also be written as a sum of products, namely,

$$(X_{11}\, X_{21} \cdots X_{m1}) + \cdots + (X_{1n}\, X_{2n} \cdots X_{mn}), \quad (2)$$

it follows that another way to generate the function is to use one Laddic to generate each term in the sum, and connect the output windings in series so that a pulse on any one Laddic appears as an output.

This second procedure can sometimes lead to a considerable simplification of the Laddic circuitry. As an example, consider the alternating symmetric function of four variables, which expresses the condition that an output should be obtained if, and only if, one or three of the four variables are present as inputs. If w, x, y and z represent the four variables, the appropriate Boolean function may be written as follows:

$$= wx'y'z' + w'xy'z' + w'x'yz' + w'x'y'z$$
$$+ wxyz' + wxy'z + wx'yz + w'xyz.$$

This expression can be factorized as follows:

$$F = (w+x+y+z)(w+x+y'+z')(w+x'+y+z')(w+x'+y'+z)$$
$$\cdot(w'+x+y+z')(w'+x+y'+z)(w'+x'+y+z)(w'+x'+y'+z').$$

This is in the form of (1), so that, using the first design procedure, the function could be generated on a single Laddic having eight "held" rungs, with four variable windings on each.

The function can also be rewritten in the following form:

$$F = (w+x)(w'+x')(y'+z)(y+z') + (y+z)(y'+z')(w+x')(w'+x)\cdot$$

Thus, using the second design procedure, the function could be obtained by combining the outputs of two Laddics, each having four "held" rungs with two variable windings on each, as in Fig. 6. In this case, a total of only 16 variable windings is required, compared to the total of 32 needed when using the first method. Thus, the wiring is considerably simplified. It should be noted that when two Laddics are used to generate the variable, as in the present case, the two separate outputs can conveniently be combined in one magnetic circuit, as illustrated in Fig. 7. Thus, the alternating symmetric function of four variables could be generated as conveniently on a single Laddic, using a drive at each end and taking the output from the middle.

Up to the present it has been assumed that current sources are available for both the variables and their negations. Occasionally this may not be practicable. In this connection, it should be noted that terms like

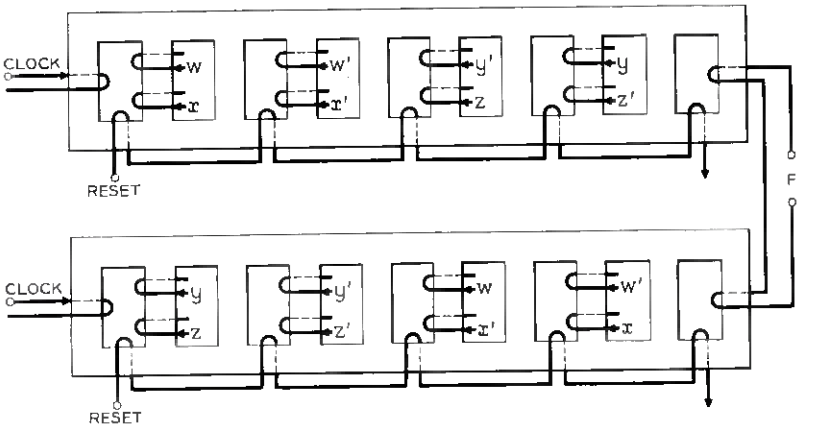


Fig. 6 — Two-Laddic circuit for generating an alternating symmetric function of four variables.

$x(y_1'y_2' \cdots y_n')$ can be produced on a single "held" rung, using currents to represent only the variables but not their primes. The procedure is to apply the x current, which could either represent a variable or be the clock current, in a direction to "hold" the rung. The currents representing the variables y are applied in a direction to counteract the "holding" drive of x, each being sufficient to prevent the rung from being "held". Thus, x will "hold" the rung when y_1 to y_n are absent, but not when one or more of them is present, so that the required Laddic output will be obtained. This procedure can always be applied if the second design procedure, using several Laddics with a common series output, is used, because the Boolean expression (2) can always be reduced to a suitable form. Therefore, it is not essential to have current sources for both a variable and its negation.

4.2 *Modified Procedures*

In Section 4.1 an outline of the basic procedures used in designing Laddic circuits was given. In practice, it is often possible to simplify the circuitry considerably by modifying the drive and output windings. For example, in connection with the alternating symmetric function of four variables it was pointed out that the single Laddic circuit could be simplified by using two drive windings in place of the conventional single drive winding. Since the optimization depends largely on the system requirements, a final design may vary from one application to another. Thus, it is not practicable to list all possibilities; rather, in this section, a number of representative examples will be given.

Using the first design procedure, a single Laddic having a total of n "held" rungs would be required to generate a term of the type

$$(x_1x_2 \cdots x_{n-1})\ (x_n + x_{n+1} + \cdots + x_m);$$

that is, n-1 "held" rungs for the first term $(x_1 \cdots x_{n-1})$ and one "held" rung for the second term $(x_n + \cdots + x_m)$. The same function can be gen-

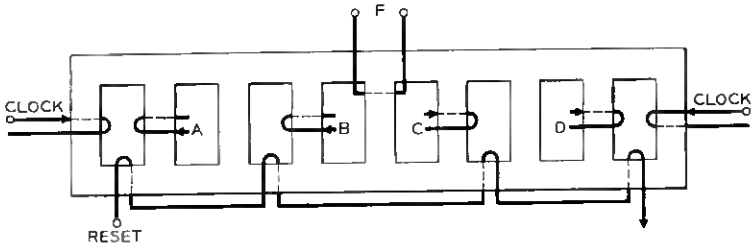


Fig. 7 — Illustrating use of a more complicated drive winding; $F = AB + CD$

erated using the four-rung Laddic shown in Fig. 8. In this, the primes of the variables x_1 to x_{n-1} are all used in opposition to the clock drive to prevent it from switching at all, and the variables x_n to x_m are used to "hold" rung 2. The result is a considerable shortening of the Laddic required, and therefore a proportionately faster switching speed for a given drive. Of course, there is no reduction of the total number of windings necessary.

Occasionally, both the function and its negation are required. The negation may be taken from the same Laddic by using an output winding which links flux changes in all rungs except the first and output rungs. Thus, an output will occur on this winding if and only if no pulse occurs on the output winding. In some cases, it may even be simpler to generate a function by using the Laddic to generate its negation according to the customary method, and to take the output from the negation winding.

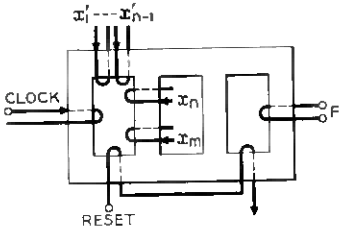


Fig. 8 — Modified Laddic circuit; $F=(x_1\ x_2 \cdots x_{n-1})\ (x_n + x_{n+1} + \cdots + x_m)$

For example, in order to generate $F = x_1'x_2' + x_2x_3' + x_1'x_3'$, using the conventional output winding, the Laddic circuit shown in Fig. 9(a) is required. Using the negation winding, the simpler circuit shown in Fig. 9(b) can be used. The use of more complicated output windings of this nature is not generally recommended because the switching speeds will vary with the return path, so that the output amplitudes will differ for the different combinations of inputs. This is not the case using the conventional output. In addition, the signal-to-noise ratio is usually worsened, because of the possibility of undesired coupling. If the output requirements are not rigorous, output circuits of this kind can be satisfactory, but generally the conventional output is preferred.

A different mode of operation is based upon the following observation. If a short-circuited winding is placed around an even-numbered variable rung, for example, rung 2, 4, or 6 in Fig. 5(d), an output is obtained from the winding on the next available rung, e.g., rung 8 in Fig. 5(d). This occurs because, as flux attempts to switch through a "shorted" rung, the emf induced in the winding produces a current which opposes the applied switching drive in exactly the same manner as the hold currents do in

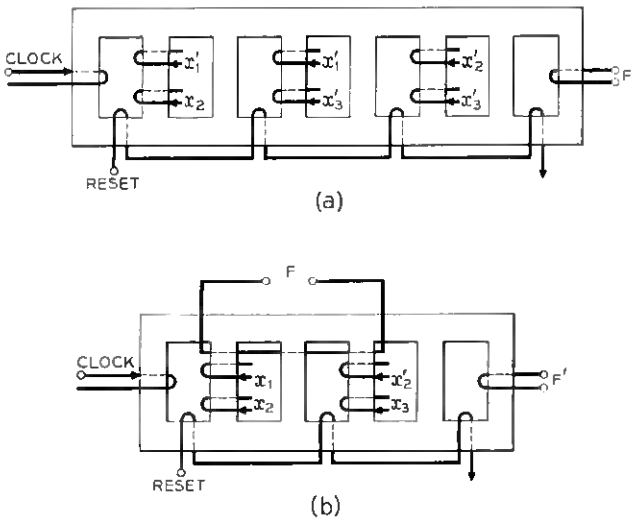


Fig. 9 — Laddic circuits illustrating the use of a negation output winding; $F = x_1'x_2' + x_1'x_3' + x_2\,x_3' = (x_1' + x_2)\,(x_1' + x_3')\,(x_2' + x_3')$; $F' = x_1\,x_2' + x_1\,x_3 + x_2\,x_3 = (x_1 + x_2)\,(x_2' + x_3)$.

the normal mode. However, if a suitable opposing emf is introduced in the shorted winding, no current will flow, so that the flux can switch through the rung and there will be no output. The opposing emf can represent an input variable. In other words, an output will be obtained when the input variables are absent. Thus, the operation is almost an exact dual to the normal mode, with the constant current sources representing variables being replaced by constant voltage sources. In practice, a rung will not be entirely "held" by the self-induced current, because the short-circuited winding will have a finite resistance. Thus, not all of the flux will be returned through the output rung, and there will be a small attenuation of the output signal.

In all of the examples up to now, the switching drive during the variable input phase has been considered to be a clock drive, and has not represented a variable input. In some applications this is an advantage, because the "hold" currents representing variables are never required to switch flux, provided that the flux pattern is reset during a subsequent reset phase, so that quite low impedance sources may be used. Furthermore, the timing of the variable pulses is not critical, the only restriction being that they be applied before or at the same time as the clock pulse, and remain at least until the end of the switching period. For other applications, these considerations may be unimportant and, in this case,

windings which represent any one of the factors in (1), may replace the normal clock winding on rung 1.

4.3 *Modified Structures*

It will be apparent from the previous section that the Laddic structure makes a very versatile circuit element. For specific applications, an economy in size and windings can sometimes be obtained by using a more complicated, less general structure in place of the Laddic. Broadly speaking, these alternative structures may be based upon the following operating principles as used in the Laddic:

i. The structure has a number of stable flux patterns.

ii. A normal or original flux pattern is set up in this structure.

iii. The tendency of this pattern to change to other fixed flux patterns according to the presence or absence of various applied fields is utilized to determine the actual switching path through the structure.

It is not the purpose of this paper to outline a design procedure for producing an optimum structure. Instead, one elementary example will be given. In Fig. 10(a) is shown the conventional Laddic circuit for generating the function $x(z + wy)$. Because both rungs 2 and 4 can be held by variable z, it follows that this part of the magnetic circuit

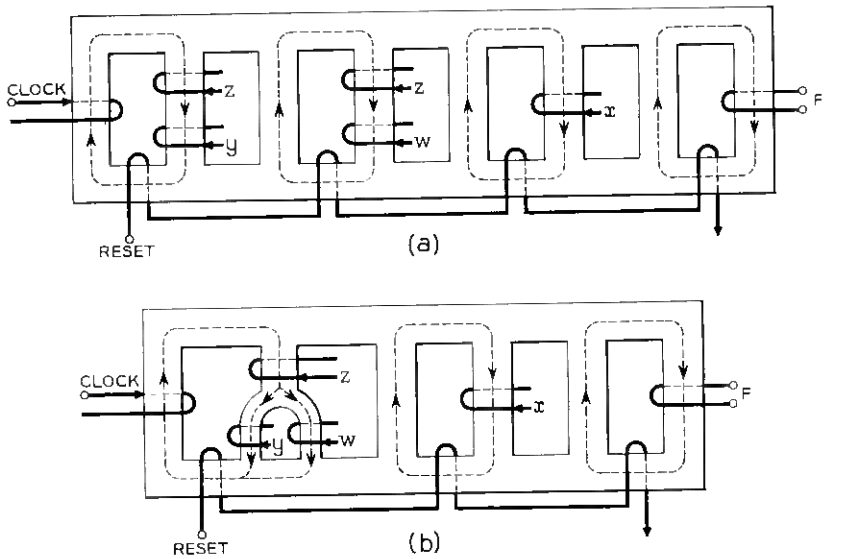


Fig. 10 — Simple illustration of the use of a more complicated structure; $F = x(z + wy)$. The main features of the reset flux pattern are shown.

can be combined as shown in the modified structure of Fig. 10(b). It will be noticed that the cross sections of the split rungs carrying variables y and w are shown as one-half of that carrying variable z in order to preserve a saturated flux structure. This modified structure has one less variable winding, and may be shorter than the conventional Laddic circuit.

In the Laddic structure, no use is normally made of the odd-numbered rungs except for the first. These rungs serve only to maintain flux continuity in the particular flux patterns used. The same result could be achieved by combining them elsewhere in the magnetic circuit, at the same time enlarging the side rails to maintain flux continuity, as in Fig. 11, for example. In Fig. 11 the rungs 1, 2, 4, 6, 8 and 10 are labelled to correspond to the like rungs of the Laddic, Fig. 5, and the remaining odd-numbered rungs are considered to be collected together on the left-hand side of rung 1 to provide the flux returns for the reset flux pattern as illustrated. Operation would then be as in the Laddic, the flux in rung 1 being reversed by the clock drive, rungs 2, 4, 6 and 8, being "held" by variable inputs, and an output being taken off rung 10.

It might be thought that this structure would give a gain in switching speed because the distance between rung 1 and rung 10 is reduced. However, experimentally the switching speed is found to be characterized by a switching path length such as $ABCD$, rather than the shorter path $abcd$. The distance aA is approximately equal to the sum of the widths of the intermediate rungs. Thus, there is little actual gain in switching speed. More serious is the fact that the flux limiting action of the side rails is reduced, and the signal-to-noise ratio degenerates. Accordingly, the conventional Laddic structure is considered preferable.

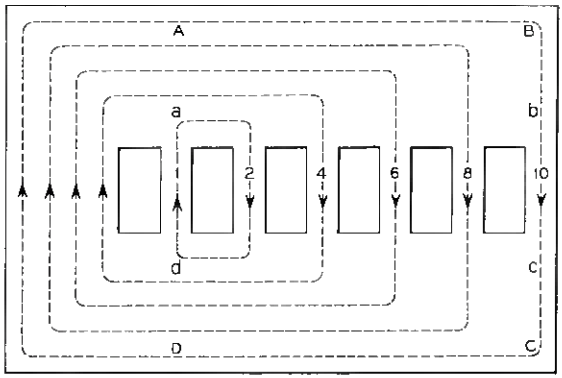


Fig. 11 — Modified Laddic structure.

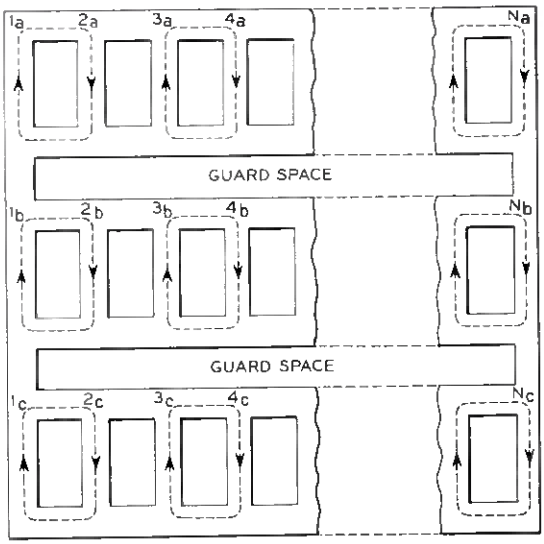


Fig. 12 — Combination of Laddics on a single sheet. The normal reset flux pattern is shown.

Another advantage of the Laddic structure is that it is compact laterally, so that it is practicable to combine several adjacent Laddics on a continuous sheet, as illustrated in Fig. 12. The guard spaces shown are sufficient to prevent any interaction between the adjacent Laddics.

4.4 *Cascading Laddics*

There is never any necessity for cascading Laddics for combinational logic, since the desired result can always be realized using single Laddic circuitry. In fact, the latter is inherently more efficient, because a cascade circuit must provide additional power to allow for dissipation in the coupling loops. A cascade circuit can produce some simplification in cases where the output of one Laddic can provide a common input for a number of others. The design problems here are reasonably straightforward and need no discussion.

In the following, two methods for coupling Laddics for sequential operation will be described.

The first method is illustrated in Fig. 13. The output of the first Laddic is used to provide the switching drive for the second. During phase Φ_1, the first Laddic is "set" by its input variables, and the second Laddic is reset. During phase Φ_2, the second Laddic is set, the first Laddic being simultaneously reset to provide the appropriate advance current. Clearly

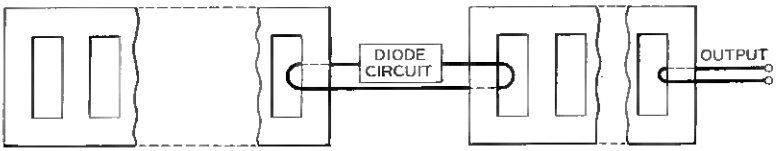


Fig. 13 — First method for cascading Laddics.

the intermediate "hold" rungs play a secondary part only in the advance operation, since they are not included in the coupled switching path. Thus, when considering the advance operation, the Laddic may be treated as a conventional core of the same peripheral length. It follows that Laddics may replace cores in conventional core circuits for sequential operation, and that an intermediate diode circuit or a transistor is necessary to prevent back propagation.[5] Because additional logical inputs may be inserted at each stage of the cascade, the Laddic circuit can be more versatile than the corresponding core circuit.

A second possibility for coupling Laddics is illustrated in Fig. 14. In this case, the output of the first Laddic is used to provide the hold current for a rung of the second Laddic. As before, the clock phase Φ_1 of the first Laddic coincides with the reset phase of the second, and *vice versa* for Φ_2. This procedure appears promising at first sight because there need be no diode in the coupling loop. Furthermore, a high coupling efficiency might be anticipated, because the advance current is not required to switch flux in the held rung of the second Laddic, and so the back emf is small. However, the coupling efficiency is actually limited, because a resistance must be included in the coupling loop. Otherwise, as discussed in Section 4.2, the loop would act as a short-circuited turn

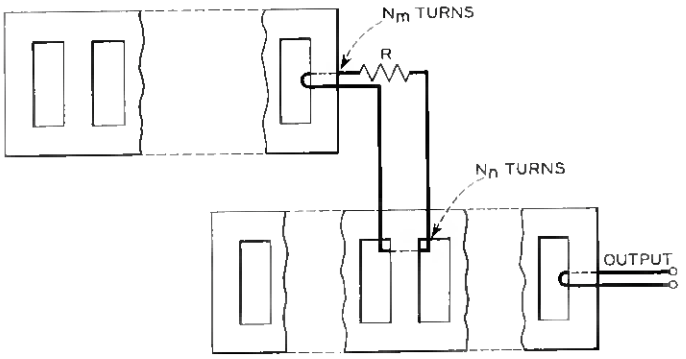


Fig. 14 — Second method for cascading Laddics.

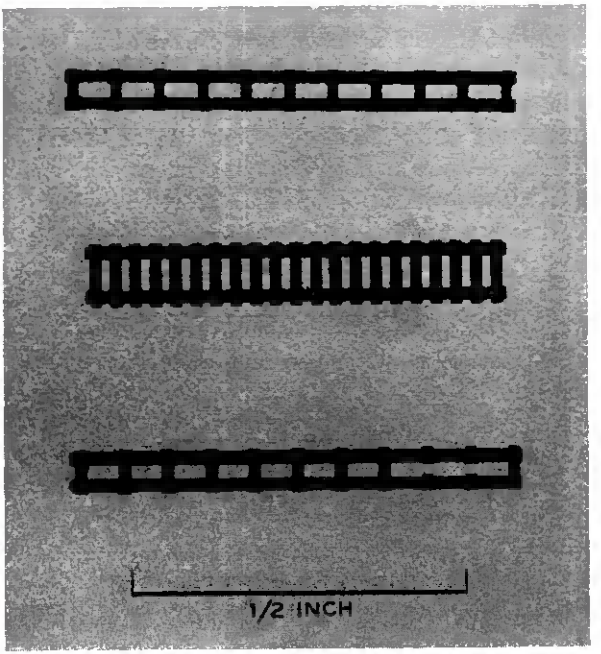


Fig. 15 — Photograph of experimental Laddics.

and hold the coupled rung of the second Laddic even when there is no output from the first Laddic, thus making the output of the second Laddic independent of the first. To allow for the power dissipated in R the turns ratio Nm/Nn in the coupling loop must exceed unity. A detailed analysis has shown that these requirements are necessarily different for each stage of the cascade, becoming more rigorous with each successive stage. It is considered that a two-stage diodeless cascade is the only case worthy of practical consideration.

V. EXPERIMENTAL RESULTS AND DESIGN FORMULAE

5.1 *Experimental Laddics*

For experimental purposes, Laddics are cut out of a ferrite sheet, using an ultrasonic cutter. Fig. 15 illustrates the size of Laddics which have been used. The smallest unit was made from a sheet of cadmium manganese ferrite 30 mils thick, the width of the rungs and side rails being 15 mils, the spacing between rungs 15 mils and the spacing between side rails 50 mils. In order to improve the signal-to-noise ratio,

the output window is sometimes enlarged by cutting out some of the rungs as in one of the units shown in Fig. 15. The actual dimensions used were chosen somewhat arbitrarily, keeping in mind the convenience of fabrication and handling. As will be shown later, from the point of view of minimizing drives for a given switching speed, the over-all length should be as small as possible. When used with single turn windings, the units shown can be driven by a transistor pulser. The output for a given drive—that is, for a given switching speed—is approximately proportional to the cross section of the rung (A cm^2), and to the remanent flux density (B_r gauss) of the material. For an output waveform approximately rectangular in shape, the mean output voltage per turn E is $E \cong B_r A \times 10^{-8}/\tau$ volts, where τ is the switching time in seconds. For the experimental units, $B_r A \cong 5.8$ and a normal range for τ was 1 to 5 microseconds.

The requirements on the structure and material are not very stringent. The best signal-to-noise ratios are obtained with close dimensional control, although a relative dimensional tolerance of 5 per cent is found to be adequate in practice. For the same reason, the material should be homogeneous and have a good squareness ratio B_r/B_s . The best material from the point of view of minimizing drives has a low threshold field for switching, H_0 , and a small switching time constant, s. From the point of view of maximizing the output, B_r should be as large as possible

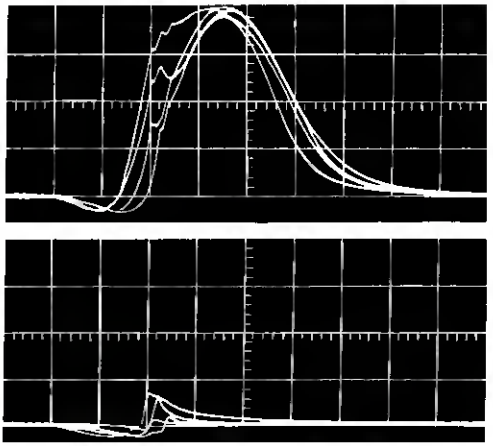


Fig. 16 — Signal and noise outputs for the circuit shown in Fig. 17. The vertical calibration is 0.01 volt/turn/large division. The horizontal calibration is 1 μs/large division. The variable currents were all equal and equal to the clock current, which was 0.8 ampere. Single-turn windings were used.

In practice, available memory core ferrites for which $B_r \cong 2400$ gauss, $H_0 \cong 0.17$ oersted, $B_r/B_s \cong 0.92$, and $s \cong 0.8$ oersted microsecond are reasonably satisfactory.

Representative output waveforms are shown in Fig. 16, which is discussed in the next section in connection with Fig. 17.

The important parameters of the Laddic are the values of the minimum currents needed to hold the variable rungs and the switching speed, for a given drive, switching path and material. Useful design formulae are derived in the following sections.

5.2 *Hold Currents*

As discussed previously, the hold currents necessary to prevent flux from being switched in the corresponding rungs must exceed a certain minimum. This minimum is proportional to, and obviously less than, the switching drive current, and decreases with the distance between the held rung and the driven rung. The following simple treatment gives a relation between the minimum hold currents, drive current and switching path, which agrees satisfactorily with experiment. These assumptions are made:

i. During switching the amount of flux by-passed from the desired switching path by the held rungs and the saturated rungs of the Laddic is negligible.

ii. The reluctance of the side rails, and of the input and output rungs, is linearly proportional to their length.

iii. The concepts of static magnetic circuitry can be applied to the dynamic switching problem for the particular problem considered here.

The first assumption can be justified because there is a difference of at least two orders of magnitude between the relative permeabilities of switching and nonswitching paths in a rectangular loop material. The second assumption can be only a first-order approximation, because the reset flux pattern is such that the initial reluctance of a side rail may

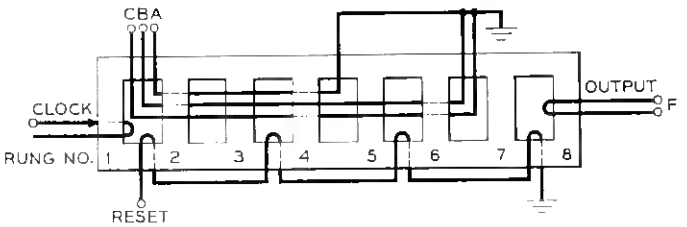


Fig. 17 — Experimental Laddic circuit; $F = (A + B)\ (A + C)\ (B + C)$.

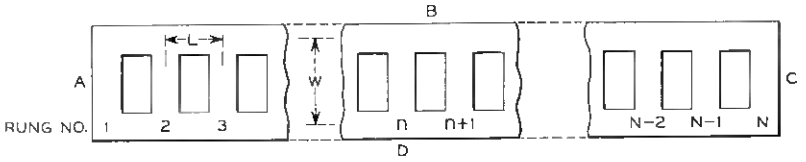


Fig. 18 — Diagram to show peripheral lengths referred to in the text.

not be uniform along its length. The justification for the third assumption is that it has been shown to be useful in a previous study of a dynamic magnetic circuit.[6]

Let L be the mean length per window of the Laddic, and let W be the mean width, Fig. 18. Let a switching mmf M_1 be applied to rung 1, and let holding mmf's M_2, M_4, $\cdots$ M_n, $\cdots$ M_{N-2}, which exactly balance the switching mmf's appearing across them, be applied to the even-numbered rungs 2 to $(N-2)$, inclusive. Thus the switching flux return is through rung N. In this case, according to assumptions i, ii and iii, the ratio M_n/M_1 will be equal to the ratio of the reluctance or length of the portion of switching path BCD beyond rung n to the total switching path $ABCDA$, Fig. 18. Thus,

$$\frac{M_n}{M_1} = \frac{BCD}{ABCDA} = \frac{W + 2(N-n)L}{2W + 2(N-1)L} = \frac{N - n + W/2L}{N - 1 + W/L}, \quad (3)$$

For the smallest Laddic dimensions shown in Fig. 15, $W \cong 2L$, and M_n/M_1 is tabulated for this case in Table II.

Experimentally, the minimum hold drives $(M_n)_{\text{exp}}$ necessary to operate the Laddic were determined by adjusting the holding currents to the minimum values necessary to produce maximum switching of flux in rung N, when a switching drive was applied to rung 1. The ratios M_n/M_1

Table II — Ratios of Minimum Hold Drive M_n to Clock Drive M_1 for the Case $W = 2L$.

N	n						
	2	4	6	8	10	12	14
4	0.60						
6	0.71	0.43					
8	0.78	0.56	0.33				
10	0.82	0.64	0.46	0.27			
12	0.85	0.69	0.54	0.38	0.23		
14	0.87	0.73	0.60	0.47	0.33	0.20	
16	0.88	0.76	0.65	0.53	0.41	0.29	0.18

were measured for all of the input-output conditions covered by Table II, and for two typical values of the rung 1 drive, $M_1 = 0.3$ ampere turn, and $M_1 = 0.65$ ampere turn. Experimental accuracy was about ± 10 per cent. It was found that the experimental ratios agreed with the predicted values to within ± 15 per cent. When comparing the experimental and predicted values, it was assumed that $(M_n)_{exp}$ was equal to $M_n - M_c$, where M_c is the effective bias mmf in the rung due to the coercive field of the material. For the present case, $H_c \cong 0.17$ oersted, and rung length is 50 mils, so that $M_c \cong 0.02$ ampere turns.

It is concluded that (3) is an adequate representation for design purposes.

5.3 *Signal-to-Noise-Ratio*

In principle, there is no upper limit to the hold currents. However, the signal-to-noise ratio degenerates with increasing hold currents because of lack of squareness of the *B-H* loop, the noise signal corresponding to the zero output of a memory core. The noise pulse decreases as the distance of a hold drive from the output winding increases, and in the Laddic only the final stages of the hold currents are serious sources of noise. An extreme practical case occurs when the hold currents are all equal to the first. Table II shows that, in this case, the final hold current is more than five times its minimum value in the 16-rung Laddic. Fig. 19(a) shows the signal and noise outputs that were obtained experimentally, using a 16-rung Laddic, for the condition where all of the hold currents are at their minimum values, as defined by (3). The remaining outputs shown were those obtained as a progressively increasing number of hold currents were made equal to the minimum value for rung 2, leaving the remainder at their previous values. The noise signal shown was the maximum that could be obtained, that is, when M_2 alone was missing. Single-turn windings were used throughout. It will be seen that the signal-to-noise ratio is excellent when all hold currents are at their minimum values, Fig. 19(a), but that the noise signal deteriorates as an increasing number of hold drives are made equal to M_1. However, for many applications, the signal-to-noise ratio is tolerable even for the extreme case of all hold currents equal. If necessary, the ratio may be improved by enlarging the effective size of the output window, as in the modified Laddic shown in Fig. 15, or by permanently holding the final rungs by means of direct currents through the hold windings. Both methods have been shown experimentally to reduce the influence of the hold currents on the output noise signal.

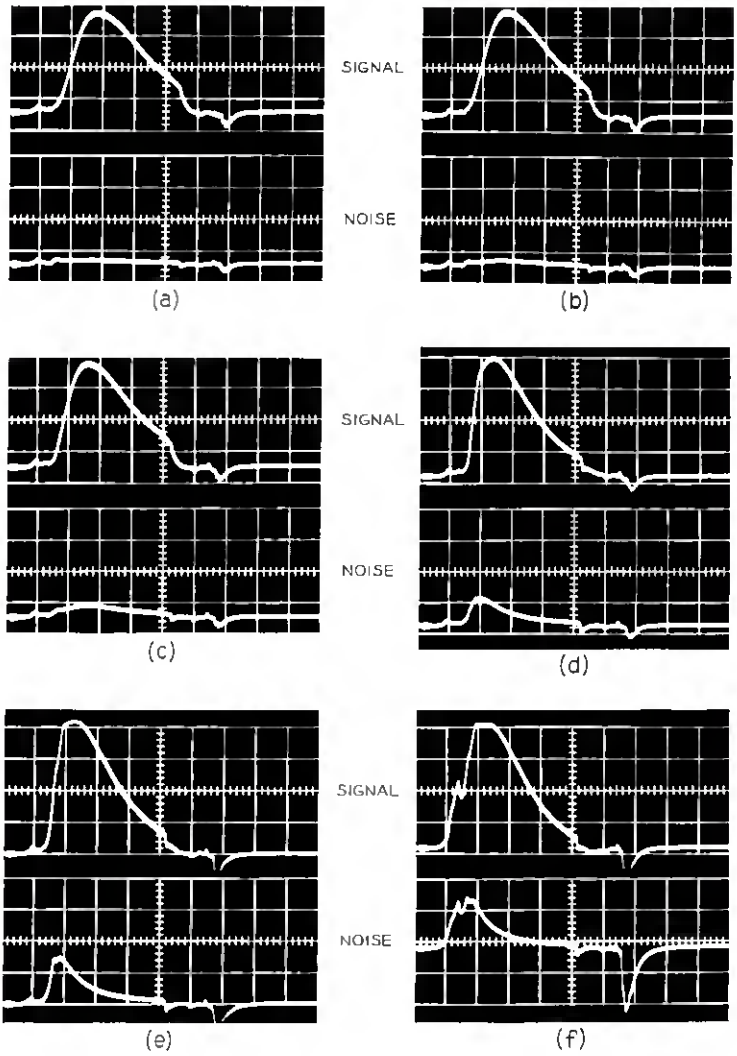


Fig. 19 — Signal and noise outputs taken from the sixteenth rung of a Laddic showing the effect of increasing the hold currents beyond their minimum values, (a) all hold currents at their minimum values; (b) M_2, M_8, M_{10}, M_{12}, and M_{14} at their minimum values, remaining hold currents equal to M_2; (c) M_2, M_{10}, M_{12} and M_{14} at their minimum values, remaining hold currents equal to M_2; (d) M_2, M_{12} and M_{14} at their minimum values, remaining hold currents equal to M_2; (e) M_2 and M_{14} at their minimum values, remaining hold currents equal to M_2; (f) all hold currents equal to M_2.

An illustrative case of practical interest, where equal hold currents must be used, is that of producing the carry (K) of a full binary addi-

tion. In Boolean algebra notation, $K = AB + AC + BC$. This can also be written in the form $K = (A + B)(A + C)(B + C)$. This function can be produced by the Laddic circuit shown in Fig. 17. In this circuit, rung 2 is held by $(A + B)$, rung 4 by $(A + C)$ and rung 6 by $(B + C)$, so that an output results only if two or three of the input variables are simultaneously present. If single-turn windings are to be used, all the hold currents are necessarily approximately equal. The outputs that were obtained experimentally for the separate inputs, ABC', $AB'C$, $A'BC$ and ABC, are shown superimposed in Fig. 16(a). The outputs obtained for the remaining possible inputs, $A'B'C'$, $A'B'C$, $A'BC'$ and $AB'C'$, are shown superimposed in Fig. 16(b). It is clear that, even when all the hold currents are equal, the signal-to-noise ratio is adequate for most purposes. Thus, apart from satisfying the condition for the minimum hold current, the margin requirements are lax.

The foregoing discussion is based on the assumption that the Laddic structure is uniform in geometry and material. If this is not the case, the reset flux pattern may be affected and, as a result, irreversible flux changes may contribute to the noise pulse. This contribution will only be large when the available flux from the hold rungs cannot be entirely returned by rungs other than the output rung. This condition can be avoided by making the odd-numbered rungs large enough and/or providing a bias winding which links odd-numbered rungs to ensure more complete saturation of all rungs.

A complete description of Laddic noise is complicated and will not be attempted here. It should be remarked that, in many circuits, no special noise suppression techniques appear to be necessary.

5.4 *Switching Speed*

The convention[7] will be adopted here that the switching time be measured between the points of 10 per cent of maximum amplitude of the output waveform.

If it is assumed, as in the last section, that the rungs of the Laddic may be ignored unless they are included in a switching path, then, for the purpose of determining switching speeds, the Laddic may be treated as a memory core of the same peripheral length. The length of the switching path when the output is taken off rung N is equal to $2(N - 1)L + 2W$, Fig. 18. Thus, since the switching time τ is related to the applied drive by the usual relation,[7] $\tau(H - H_0) = s$, for the Laddic

$$\frac{1}{\tau} = \frac{1}{s}\left[\frac{0.2\pi M_1}{W + (N - 1)L} - H_0\right]; \tag{4}$$

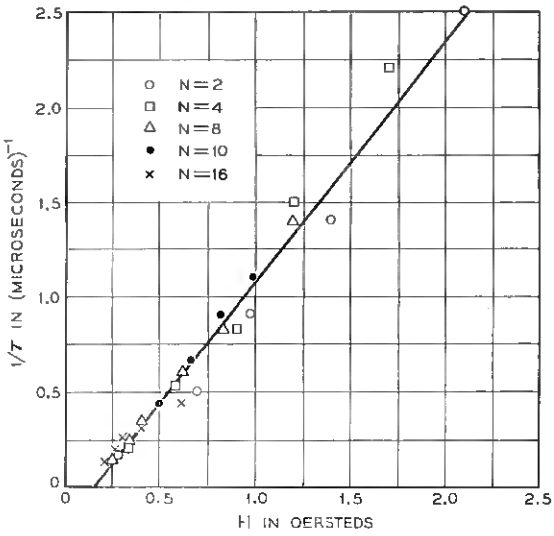


Fig. 20 — Experimental plot of inverse switching speed versus applied field, $H = (0.2\pi M_1)/[W + (N - 1)L]$, for outputs taken from the 2nd, 4th, 8th, 10th and 16th rung of the Laddic.

M_1 is in ampere turns, W and L in centimeters, H_0 in oersteds and s in oersted microseconds.

The experimental relations for the 20 hole Laddic shown in Fig. 15, ($W = 2L$), are shown in Fig. 20, for N in the range 2 to 16. It is concluded that (4) is a satisfactory representation. The values of s and H_0, derived from Fig. 20, are $s = 0.77$ oersted microsecond and $H_0 = 0.15$ oersted.

It should be remarked that, for these measurements, all the hold currents were maintained at their minimum values. If this is not the case, the output waveforms may be modified, and the effective switching times may change. However, the data presented give the approximate magnitudes.

5.5 *Impedances*

Since the hold current is not called upon to switch flux, the impedance of a hold winding is quite small. It is equal to $r + j\omega l$, where r is the resistance of the winding, and l its inductance. For a single-turn winding typical values are $r \sim 0.02$ ohm, $l \sim 0.01$ microhenry.

The impedances for the clock and reset drives are approximately resistive and may be derived from the usual core formula.[3]

5.6 *Cascading*

Two methods for cascading Laddics were described in Section 4.4. The first method was shown to be similar to the cascading of conventional cores, and so need not be considered further. As stated, the second method has limited applicability, but it is still of practical interest. For this reason the characteristics of a two-stage cascade will be described.

Refer to Fig. 14. Approximate relations for the minimum values of R and N_m/N_n are the following:

$$R_{\min} = \frac{4\pi\varphi_r N_n^{\,2}}{s(I_{10} - I_{1n})}, \tag{5}$$

$$\left[\frac{N_m}{N_n}\right]_{\min} = \frac{L_{1m}}{L_{10}}\frac{R}{R_{\min}}. \tag{6}$$

In deriving (5) and (6) it was assumed, as a first-order approximation, that the rate of change of flux $\dot{\varphi}$ is constant during the switching period. Equations (3) and (4) were used for the minimum hold currents and switching speeds respectively; L_{1m} and L_{10} are the mean peripheral lengths of the first and second Laddics, i.e., $ABCDA$ in Fig. 18, and L_{1n} corresponds to the peripheral length DAB; φ_r is the total flux available for switching, and was assumed to be equal in the two Laddics; R is the actual loop resistance. All units are in cgs.

Equations (5) and (6) have been found to provide a reasonable guide to practical design. As an example, consider the circuit shown in Fig. 21. For $N_n = 1$ the theoretical $R_{\min}$ is 0.2 ohm. Experimental output wave-

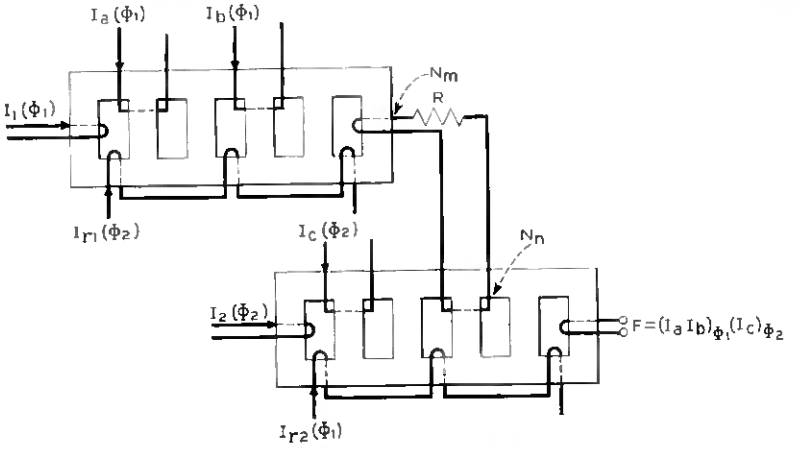


Fig. 21 — A two-Laddic cascade.

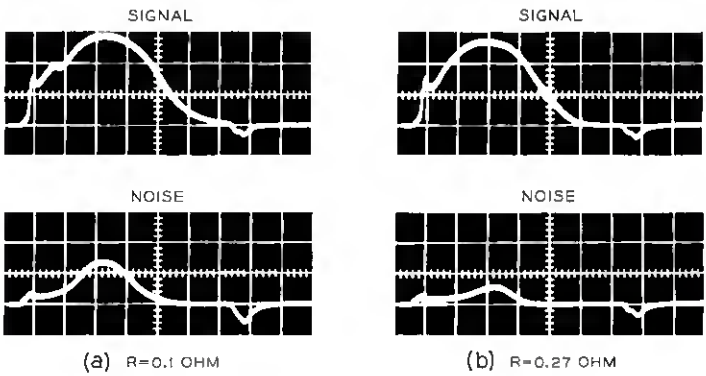


Fig. 22 — Experimental waveforms from the circuit of Fig. 21, with $N_m = 4$.

forms for one value of R on either side of $R_{\min}$, $N_n = 1$ and $N_m = 4$, are shown in Fig. 22. It will be seen that, as R is increased through the theoretical minimum, the main effect is to decrease the amplitude of the "0" signal, as is to be expected.

For a given N_m/N_n and R, the ratio of the current drives I_{r1}/I_2 must exceed a certain minimum to provide an adequate advance hold current; I_{r1}/I_2 also has a maximum limit, for otherwise the advance hold current will not persist for the full switching period of the second Laddic. The margin between maximum and minimum vanishes if N_m/N_n is equal to or less than the theoretical minimum. In practice, R should be chosen close to $R_{\min}$, and N_m/N_n made as large as practicable. Table III illustrates the maximum and minimum current ratios obtained experimentally using the circuit of Fig. 21, for different parametric values. It will be seen that, for a suitable choice of R and N_m/N_n, the operating margins are broad.

TABLE III — RATIOS OF MAXIMUM AND MINIMUM DRIVE CURRENTS FOR DIFFERENT EXPERIMENTAL CONDITIONS

N_2I_2 ampere-turns	N_m turns	N_n turns	R ohms	$(I_{r1}/I_2)_{\min}$	$(I_{r1}/I_2)_{\max}$
0.3	6	1	0.43	2.8	>4
0.3	6	1	0.26	3.0	>4
0.3	6	1	0.1	3.0	>4
0.3	4	1	0.42	2.0	3.0
0.3	4	1	0.25	2.0	4.0
0.3	4	1	0.1	2.0	>4
0.3	2	1	0.1	min. and max. overlap	

VI. DISCUSSION

The Laddic structure is a versatile one for use in logic circuitry. More complicated structures, as discussed, may offer particular advantages when certain functions have to be realized, but for many applications the Laddic structure is sufficient.

The details of Laddic behavior are not yet completely understood. Their understanding probably requires a more thorough explanation of the switching process in ferrites. However, by making a number of simple assumptions it has been possible to give simple formulae and design techniques for Laddic circuits, which appear in general to be satisfactory.

The Laddic is basically a device for combinational logic and, as shown, a single Laddic can be used to realize any switching function of n variables. For systems applications where sequential operation is necessary, it may be used in conjunction with intermediate diode or transistor circuitry. In certain cases, the intermediate circuitry is not necessary.

The Laddic is a simple device to make. Suitable materials are available, and their properties are not very critical, provided that the material is reasonably homogeneous. For experimental purposes these devices have been cut out of solid ferrite sheets, but for larger scale fabrication the green ferrite would more conveniently be pressed into the final form. Experience with other ferrite devices suggests that pressing into the final form will slightly improve the material properties. Because of the use of single-turn windings for the variables, the wiring of a Laddic is fairly simple. It involves dropping a hairpin-shaped conductor across a rung, and a number of simple assembly schemes can be thought of. The reset winding is more complicated, because it involves threading a number of holes with a single wire. Printed wiring is very suitable for this purpose.

The speed of the Laddic is basically the same as that of other magnetic core devices, being limited primarily by the properties of available materials. Switching speeds of a few tenths of a microsecond and repetition rates of a few hundred kilocycles have been achieved. For many applications in the telephone system, speed is not a prime requirement.

No attempt will be made in this paper to compare Laddic circuits with conventional core logic circuits, or with other multi-aperture devices. A useful comparison would be one that covered all possible applications, and this is not practicable. The main merits of the Laddic are its probable low cost, versatility and compatibility with existing core circuits, and the convenience of its design from the point of view of fabrication.

VII. ACKNOWLEDGMENTS

We particularly wish to thank D. B. Armstrong, P. Mallery and H. J. Schulte for many discussions and contributions. R. A. Chegwidden, F. Monforte and F. J. Schnettler have kindly provided suitable materials; L. K. Degen, A. W. Koenig and J. F. Muller have supervised the cutting procedures for the experimental models and E. M. Walters has actively assisted in all stages of the experimental work.

REFERENCES

1. Karnaugh, M., Pulse-Switching Circuits Using Magnetic Cores, Proc. I.R.E., **43**, May 1955, p. 570.
2. Rajchman, J. A. and Lo, A. W., The Transfluxor, Proc. I.R.E., **44**, March 1956, p. 321.
3. Abbott, H. W. and Suran, J. J., Multihole Ferrite Core Configurations and Applications, Proc. I.R.E., **45**, August 1957, p. 1081.
4. Lockhart, N. F., Logic by Ordered Flux Changes in Multipath Ferrite Cores, I.R.E. Nat. Conv. Rec., **6**, 1958, Part 4, p. 268.
5. Guterman, S., Kodis, R. D. and Ruhman, S., Logical and Control Functions Performed with Magnetic Cores, Proc. I.R.E., **43**, March 1955, p. 291.
6. Gianola, U. F., Switching in Rectangular Loop Ferrites Containing Air Gaps, J. Appl. Phys., **29**, July 1958, p. 1122.
7. Menyuk, N. and Goodenough, J. B., Magnetic Materials for Digital-Computer Components, J. Appl. Phys., **26**, January 1955, p. 8.
8. Sands, E. A., The Behavior of Rectangular Hysteresis Loop Magnetic Materials Under Current Pulse Conditions, Proc. I.R.E., **40**, October 1952, p. 1246.

Radio Attenuation at 11 kmc and Some Implications Affecting Relay System Engineering

By S. D. HATHAWAY and H. W. EVANS

(Manuscript received May 14, 1958)

Radio waves at 11 kmc are attenuated by rain. In order to derive rules for engineering radio relay systems at 11 kmc, a one-year experiment was conducted in a region of frequent heavy rainfall. The attenuation of paths 27 and 12 miles long was measured, together with rainfall at two-mile intervals along the paths. The instrumentation and the test results are described, and some implications related to systems engineering are pointed out.

I. INTRODUCTION

Increasing use of the common-carrier microwave frequency bands at 4 and 6 kmc has directed attention to the next higher band at 11 kmc. All three bands are subject to atmospheric fading, but propagation at 11 kmc differs from that at the lower frequency bands chiefly in its vulnerability to rain. Knowledge of the statistics of the excess path loss caused by rain is a necessary prerequisite to 11-kmc system design, and therefore an experiment was undertaken to extend the modest body of available knowledge.

The effects of rain on microwave radio propagation have been calculated by Ryde and Ryde.[1,2] The radio energy is absorbed and scattered by the rain drops, and these effects become more pronounced at the higher microwave frequencies where the wavelength and the raindrop diameter become more nearly comparable.

The excess attenuation caused by rainfall depends on the number of drops per unit volume in the radio path, the square of the drop diameter and a complex factor representing the ratio of the total energy absorbed and scattered by a single drop to the energy in that area of the wavefront equal to the projected area of the drop.

Laws and Parsons[3] observed the distributions of drop sizes for various

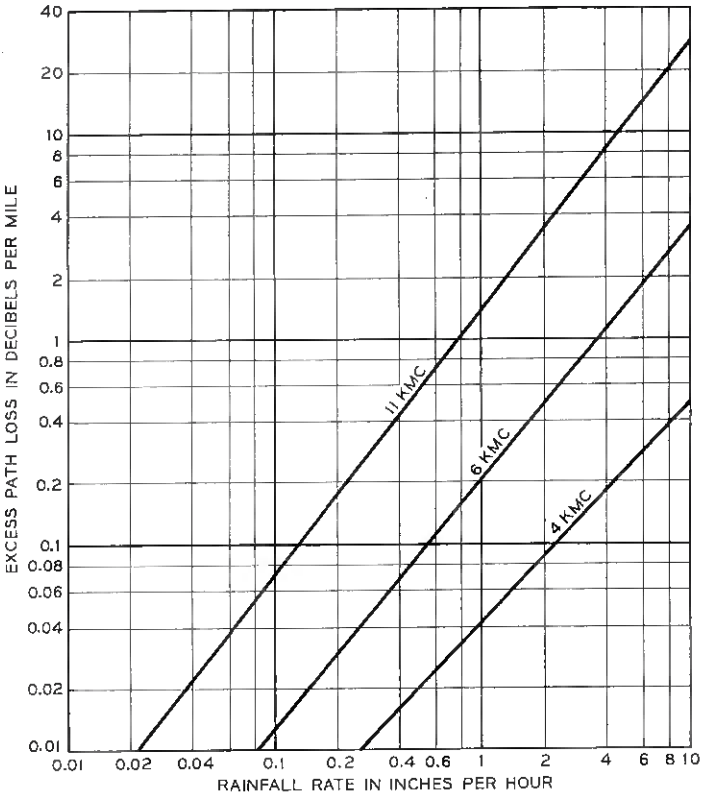


Fig. 1 — Rain attenuation vs. rainfall rate (theoretical, after Ryde and Ryde[1]).

rates of fall on a horizontal surface, using the method of Bentley.[4]* The higher the rainfall rate, the larger the drops, and also the greater the spread in size of drops. Ryde computes the number of drops per unit volume from the data of Laws and Parsons by applying the terminal velocity appropriate to the drop mass.

The excess path loss per mile according to Ryde for the three common-carrier frequency bands — 4, 6 and 11 kmc — is shown on Fig. 1 for various rates of rainfall.†

* The Bentley method involves exposing trays of sifted flour to the rainfall, baking the flour to solidify the pellets formed by impinging raindrops and then sorting the pellets by size. The flour has been calibrated by generating drops of known size, so that drop size can be determined from the pellet size.

† It is interesting to note that, from his computations, Ryde concludes that the excess attenuation caused by hail is in the order of one-hundredth that caused by rain, that ice crystal clouds cause no sensible excess attenuation, and that snow produces very small attenuation, even at the excessive rate of fall of five inches per hour.

Much rainfall data is available for point locations, but very little is known about the relationship between the rate of fall at a single rain-gauging point and the profile of rate of rainfall along a radio path. Furthermore, most rainfall data are in terms of fairly long discrete intervals, such as 30 minutes or one hour, and the relationship between hourly and instantaneous rates of fall is not perfectly known.

Bussey[5] has analyzed rainfall data for one year from the Muskingum River watershed in Ohio,[6] and finds that the annual distribution of one-hour point rates is approximately the same as an annual distribution of instantaneous 50-km path rates. He further suggests that 10-minute point data may apply to an 8-km path, 30-minute data to a 25-km path, etc.

It was the purpose of the experiment described here to seek confirmation of Ryde's relationship between excess path attenuation and instantaneous rate of rainfall, and to measure the profile of rate of rainfall along a radio path in hopes of finding correlation with rainfall measured at a single point. It was expected that this information would be useful in determining design parameters for 11-kmc radio relay systems and for suggesting the conditions under which they be used.

The experiment consisted of operating a radio path of a length typical of short-haul radio relay systems in a heavy rain area for a year. Instrumentation included devices for measuring excess radio path loss and rain gauges along the path at intervals short enough to define the rainfall profile.

II. RADIO PATH

The requirements that determined the choice of the radio path were:

(a) Heavy rainfall, both in rate and depth.

(b) Length of about 25 miles, which is considered typical of possible 11-kmc application, with the possibility of a second receiver midway in the path so that some feel of interpolation versus length would result.

(c) All-weather highway parallel to and very near the path, to permit access to rain gauges.

(d) Existing structures and buildings for antennas and radio equipment.

(e) Preferably a path equipped with an operating 4-kmc radio relay system, so that some comparison could be made between 4- and 11-kmc propagation.

Literally hundreds of possible paths were examined. The choice narrowed quickly to the Gulf Coast region because of the high incidence of heavy rainfall and the great total rainfall, which is in the order of 60

in. per year.* Finally, a path was selected between Mobile and Mount Vernon, Alabama, 27.7 miles long, approximately north and south, and parallel to a good highway, as shown in Fig. 2.

Arrangements were made to locate the transmitter at the Mount Vernon TD-2 4-kmc radio relay station, and an 8- by 12-ft plane reflector, specially made to be flat to $\frac{1}{16}$ in., was mounted at the 300-ft level of the TD-2 antenna tower. The reflector was illuminated by a 5-ft parabolic antenna mounted on top of the transmitter equipment housing, using a button-hook feed constructed of commercially available waveguide pieces. The gain of the antenna system at Mount Vernon was 44.2 db, 1.2 db greater than the gain of the parabolic antenna alone.

At Mobile, a similar antenna system for the receiver was placed on the TD-2 tower atop the telephone building, but, because this tower was only 85 ft tall, a 6- by 8-ft plane reflector was used with the 5-ft parabolic antenna. The gain of the antenna system at Mobile was 42.8 db, 0.2 db less than that of the parabolic antenna alone because of the small spacing between the antenna and the reflector.

At a point 12.6 miles south of the transmitter at Mount Vernon, near Axis, Alabama, a second receiving station was constructed. It used a 104-ft path-loss testing tower,[7] with a 3-ft parabolic antenna having a gain of 37.1 db, mounted directly on the receiver front-end, which could be run up and down the tower on a carriage. The tower was located directly in the path from Mount Vernon to Mobile.

The antenna sizes were chosen to produce roughly equal received signal levels at Mobile and Axis, and to produce as large a received signal as was consistent with physical stability of the towers and the flatness of commercially available reflectors.

The path loss (between isotropic antennas) from Mount Vernon to Mobile is 146.3 db, and from Mount Vernon to Axis 139.4 db. When the antenna gains are included, the net loss is 59.4 db from the Mount Vernon transmitter to the Mobile receiver and 58.1 db to the Axis receiver, in the absence of rain and atmospheric fading.

The terrain near Mount Vernon is gently rolling, and south of Axis the path traverses the broad swampy valley of the Mobile River, the southernmost three miles being partly over water. The Mobile-Mount Vernon path had been tested previously[8] at 4 kmc and was found to be free of strong ground reflections, so it was considered unnecessary to retest at 11 kmc. A profile of the path is shown in Fig. 3. The path was engineered at 4 kmc to have one-third first Fresnel zone clearance over

* Annual depths of about 150 in. occur in the North Pacific Coast rain forests, but, surprisingly, the rate of fall is quite low.

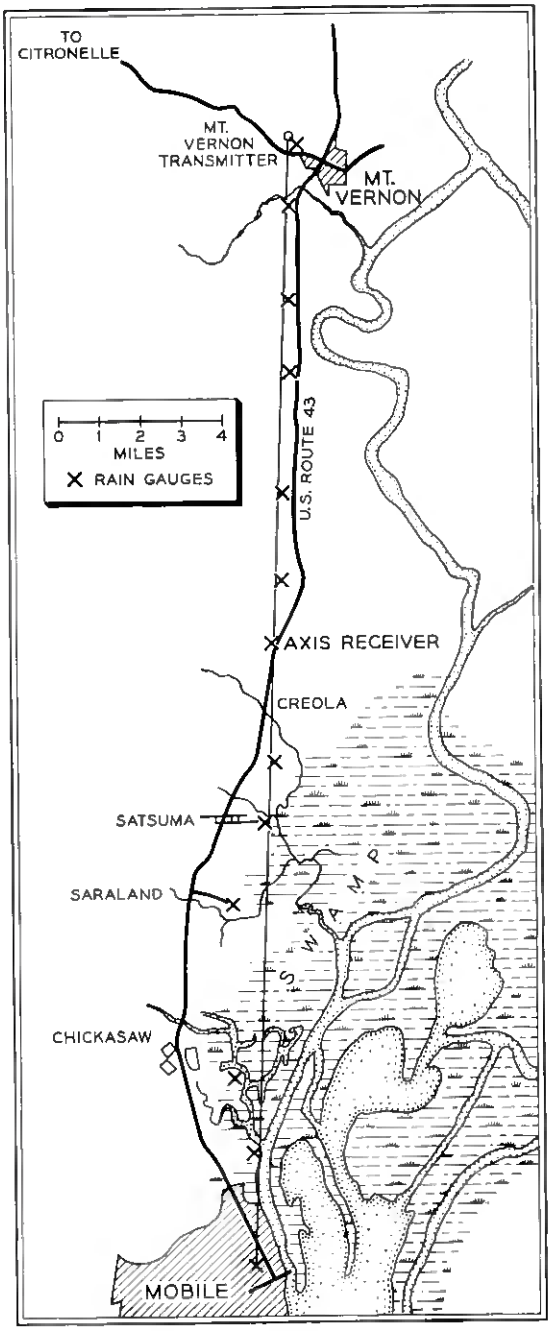


Fig. 2 — Map of Mobile-Axis-Mount Vernon radio path.

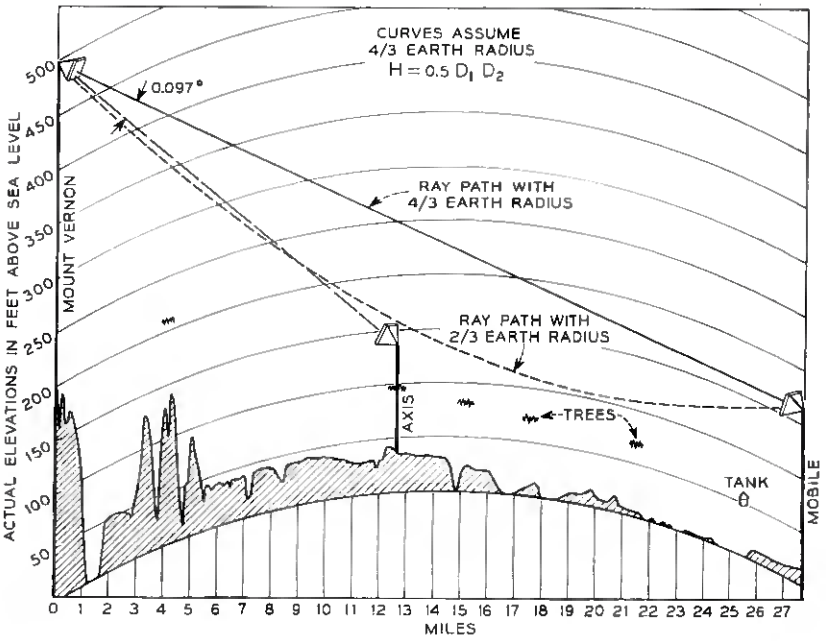


Fig. 3 — Profile of Mobile-Axis-Mount Vernon radio path.

an apparent earth radius equal to two-thirds of its true value. The passive reflectors for the 11-kmc tests had to be mounted below the existing TD-2 system antennas for physical reasons. Because of the lower antenna heights, the path clearance was 15 ft less than at 4 kmc. The limiting point in the path was 18 miles from the Mount Vernon end and, with 60-ft trees, the clearance was 36 ft when the effective earth radius was two-thirds of the true earth radius. At 11 kmc this was approximately 0.7 first Fresnel zone, which is considered adequate even for this area. The shorter path, Axis-Mount Vernon, had clearance in the order of four Fresnel zones, which was far more than sufficient.

III. RESULTS

3.1 *Fading*

Fig. 4 shows the signal level distributions of both paths due to multipath fading for a four-month period, omitting the effects of rain. The long path distribution exceeds Rayleigh for fades greater than 20 db. This would seem to indicate a strong stable reflection condition due either to ground reflections or to layer stratifications in the atmosphere.

Since no path-loss tests were made at 11 kmc, ground reflections cannot be ruled out. However, the path for the most part traverses land covered with low vegetation and pine forests, usually thought to be nonreflective. The TD-2 system suffered similar fading, and it had been established that at 4 kmc the path was essentially nonreflective. Unfortunately, only a slow-speed strip recorder was monitoring the 4-kmc system and comparative distribution data are not available.

Figs. 5 and 6 show typical 4-kmc and 11-kmc signal strengths during periods of multipath activity. In general, multipath fading on 4 kmc

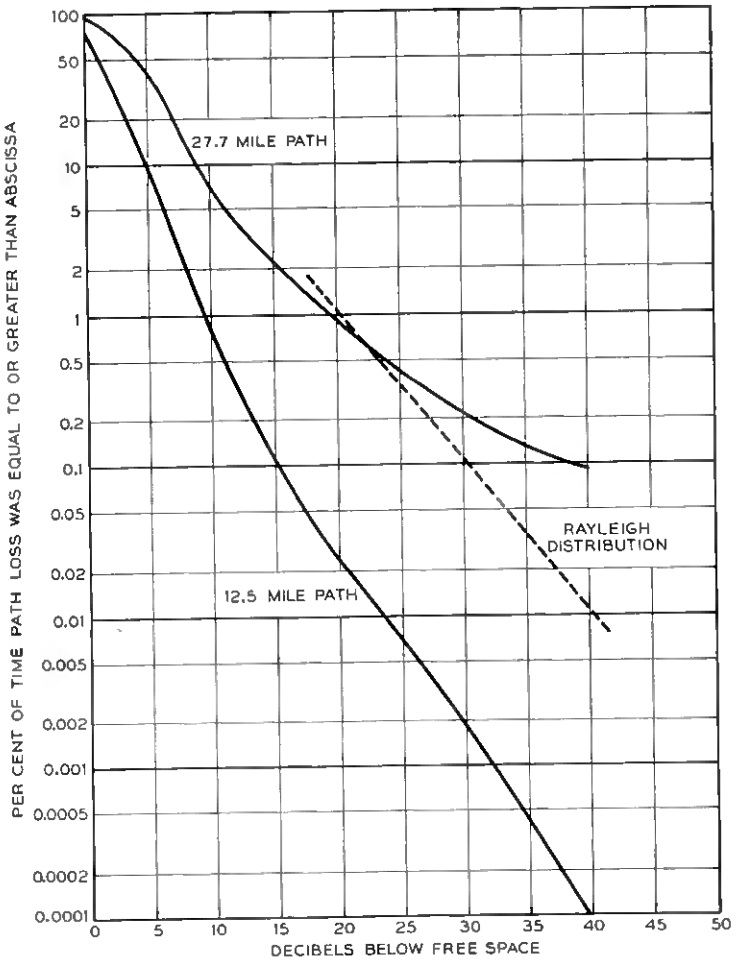


Fig. 4 — Distribution of selective fading, March 1 through July 31, 1956.

and 11 kmc began and ended at approximately the same time and followed the same over-all pattern. However, the number of fades was greater at 11 kmc than at 4 kmc. The necessity of having diversity protection for such systems is apparent if they are to meet long distance telephone circuit standards.

In addition to selective fading there were several long periods of depressed fields caused by earth bulge or obstructive-type fading. Atmospheric conditions in the Gulf region are favorable for fading of this type because high humidity and stable conditions exist at night and

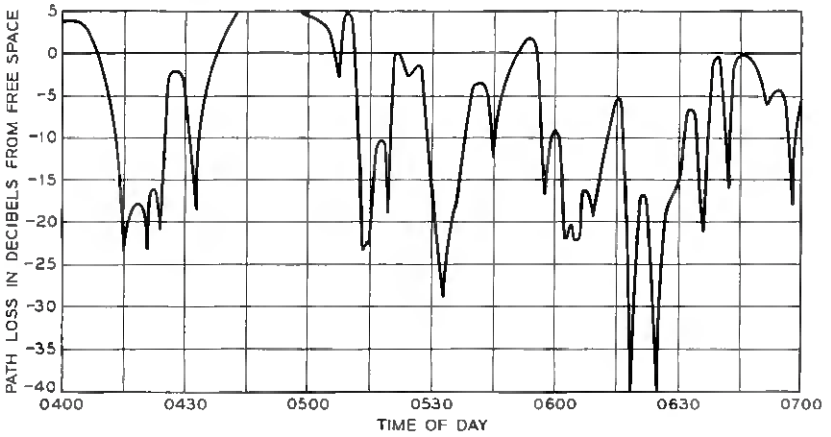


Fig. 5 — Typical selective fading at 4 kmc.

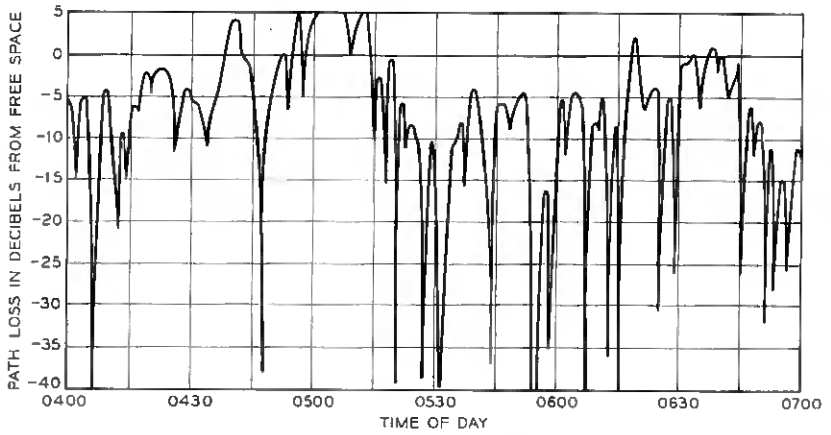


Fig. 6 — Typical selective fading at 11 kmc.

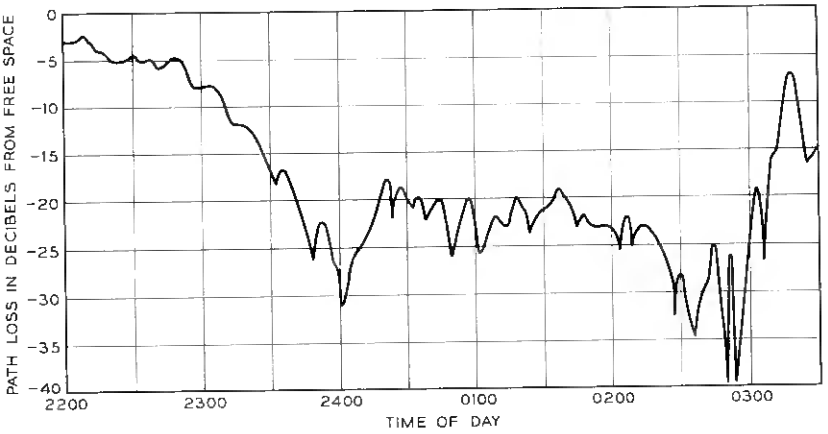


Fig. 7 — Earth bulge fading at 4 kmc.

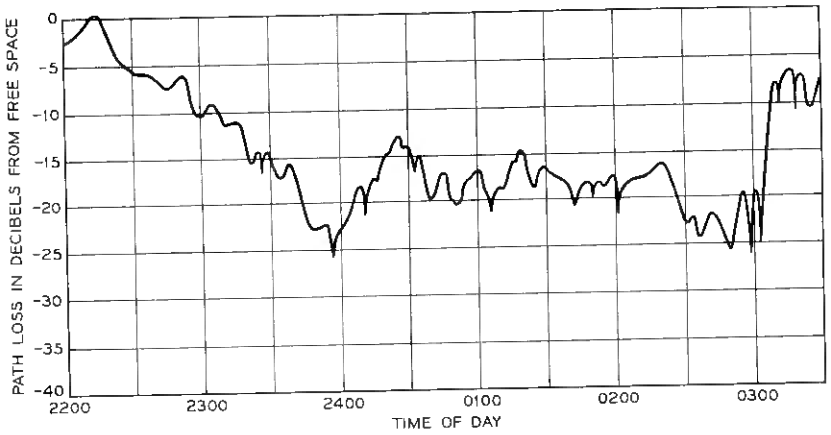


Fig. 8 — Earth bulge fading at 11 kmc.

during the early morning hours. Figs. 7 and 8 show depressed fields that occurred on November 18 and 19. Since fades of this type are insensitive to frequency, protection can be accomplished only by providing adequate clearance and restricting the lengths of the radio paths. The received signal strength at 11 kmc on the Mobile-Mount Vernon path was 40 db or more below the normal received level during less than 0.03 per cent of the year, due to obstructive-type fading. The shorter path was unaffected.

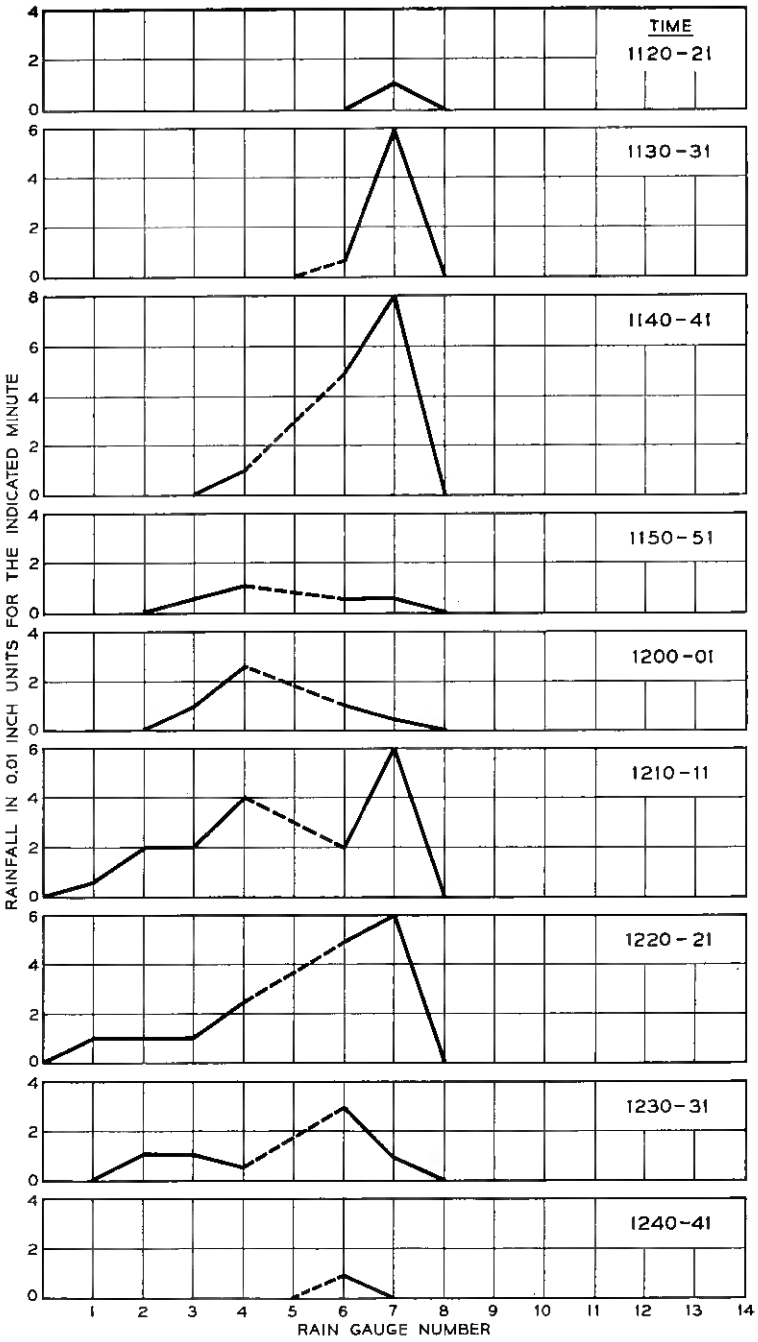


Fig. 9 — Rainfall distribution over Axis-Mount Vernon path, March 15, 1956.

3.2 *Rain Attenuation*

Frontal storms of short duration and high rates of rainfall are common in the Delta area of the United States. In general, these storms originate in the Gulf of Mexico and travel landward in a northeasterly direction. A typical storm arrived in Mobile on March 15, 1956, and passed diagonally across the radio test path between the small towns of Creola and Mount Vernon, Alabama. Rain fell over this area of the test path from about 11:15 A.M. to 12:45 P.M. Fig. 9 shows the rainfall rate distribution at ten-minute intervals as the storm progressed across the path. In analyzing the data, such profiles were constructed for each minute of each significant rain event.

Fig. 10 shows the correlation between the measured and calculated signal levels during the progress of the March 15 storm. The calculated signal level is based on the effective two-mile rainfall rate measured along the path during the storm. Ryde[2] has indicated that the attenuation due to rainfall can be approximated by

$$\text{db} = k \int_0^r R^\alpha \, dr,$$

where

R = rainfall rate,

r = length of propagation path.

Hitschfeld, Gunn and East of McGill University, Montreal, Canada,[9]

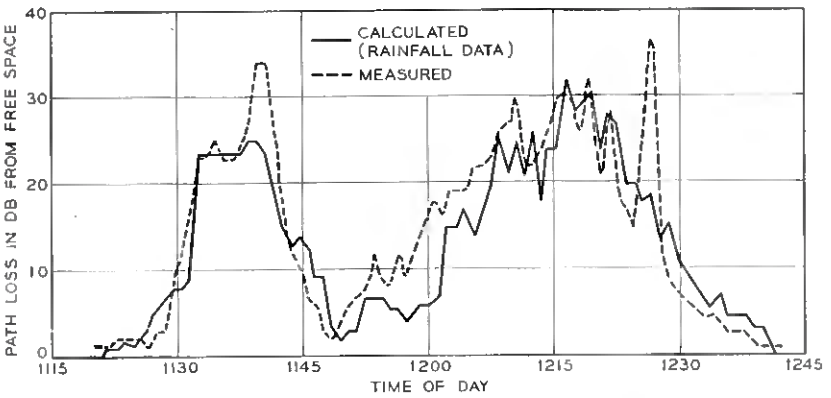


Fig. 10 — Correlation between rainfall and path loss, March 15, 1956.

have made computations of k and α for various wavelengths at 18°C. Extrapolation of their work yields for 11 kmc:

$$k = 1.395 \text{ db/mile/in./hr},$$

$$\alpha = 1.3.$$

The rain gauges on the Mobile-Mount Vernon path were placed at approximately two-mile intervals. The rainfall rate within one mile either side of a rain gauge has been assumed uniform. Then, over any two-mile path the attenuation due to rainfall is approximated by

$$\text{db} = 2kR^{\alpha}.$$

The attenuation over the entire path is then the sum of the attenuations due to the two-mile segments of the path:

$$\text{db} = 2k(R_1^{\alpha} + R_2^{\alpha} + \cdots + R_n^{\alpha}).$$

The assumption of uniform rainfall within one mile either side of a rain gauge is most surely inaccurate. However, it permits an approximate solution to the problem of attenuation due to rainfall that is not inconsistent with the measured values. Certainly a better correlation would have been obtained if the rain gauges had been spaced closer together.

A number of rain events were analyzed and the data reduced to the equivalent two-mile rate assuming uniform rainfall over the two-mile spans. Fig. 11 is a scatter diagram showing transmission loss in db per mile due to precipitation versus precipitation in inches per hour. For comparison, Ryde's equation is plotted using the constant values suggested earlier.

The recording equipment at Mobile and Axis was arranged to record receiver input levels from 0 to 40 db below the normal input level (approximately −32 dbm). The received signal strength was 40 db or more below the normal received level due to rainfall for 0.106 per cent of the year on the Mobile-Mount Vernon path and 0.020 per cent of the year on the Axis-Mount Vernon path. These figures indicate the expected order of outage time due to rainfall for single-hop 11-kmc radio systems having a 40 db fading margin and operating over similar paths in the Gulf Coast region. Fading margin is taken to mean the number of db the receiver input level can be reduced before the noise exceeds the system objective; outage time is defined as the time the noise does exceed the objective. Any predictions based on the above figures for outage time would be pessimistic for most other areas of the United

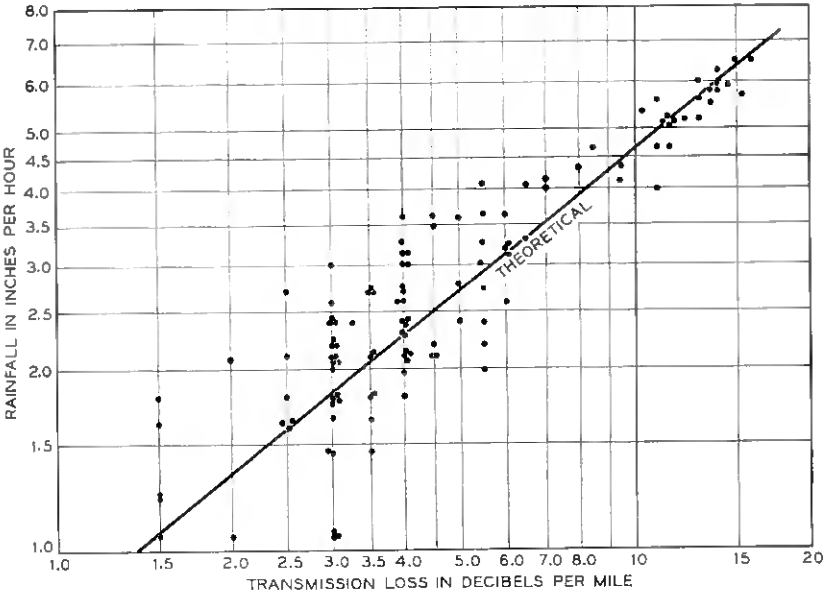


Fig. 11 — Scatter plot — transmission loss vs. rainfall rate.

States since they reflect the attenuations due to severe rainfall found along the Gulf.

IV. IMPLICATIONS AFFECTING SYSTEM ENGINEERING

Rain attenuation is obviously a large factor in determining system reliability, and hence it reacts strongly on both the design and the application of the system. Since rainfall varies greatly in frequency and intensity from one region to another, it is important to be able to predict performance in any region, so that the system as designed will have the widest possible application consistent with cost, and so that the applications engineer will know how to tailor those system parameters at his command to produce the degree of reliability desired.

It is not feasible, for reasons of cost, to measure rain attenuation in all parts of the country, so it is necessary to use what rainfall data are available, and to couple the data, through what are thought to be reasonable assumptions, to the relationships between rainfall and attenuation. The validity of the predictions rests clearly on the validity of the assumptions, and it is to be expected that further refinements in predicting rainfall outage will result from observing the performance of early 11-kmc systems.

As an approach to the problem of predicting outage time due to rainfall for all areas of the country, it has been assumed that the annual distribution of one-hour point rates is indicative of the annual distribution of instantaneous 30-mile path rates, along the lines suggested by Bussey.[5] This is equivalent to assuming a fixed storm pattern moving at 30 miles per hour in the direction of the path. Furthermore, it has been assumed that the frequency of occurrence of severe rainfall of the type measured in the Mobile area will be reduced in other parts of the country in proportion to the distribution of annual point rates of one inch or more per hour. Fig. 12, based on these assumptions and the work of Dych and Mattice,[10] illustrates contours of constant path lengths for fixed outage times for different areas of the United States. Fig. 13 shows the expected outage time due to rainfall for various path lengths in different rain areas of the United States. Curves A through H of Fig. 13 correspond to the general areas described by the contours in Fig. 12. The longer paths have been weighted somewhat to take account of less severe rainfall covering larger areas than do storms typical of the Gulf region.

In engineering a complete 11-kmc radio relay system, the rain outages of the individual hops must be added to obtain the performance for the system. Also, it is desirable to lay out the system in such a manner that

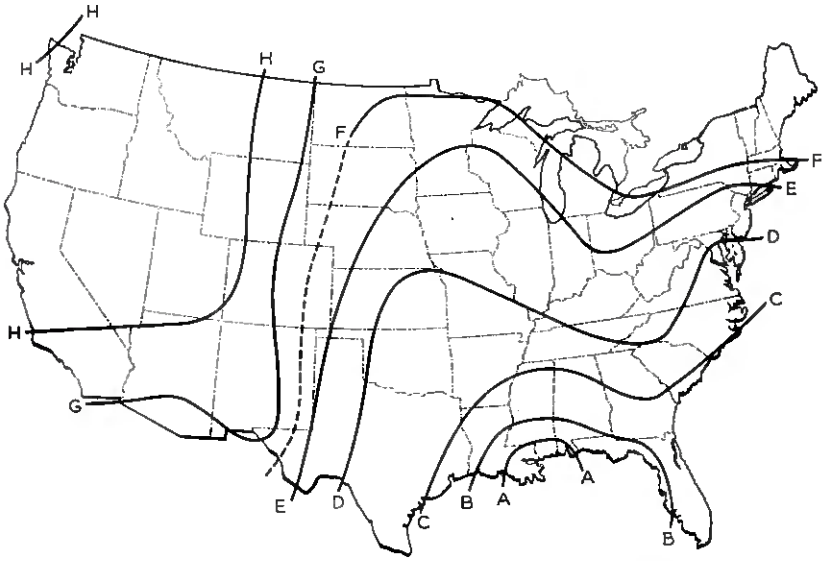


Fig. 12 — Contours of constant path length for fixed outage time.

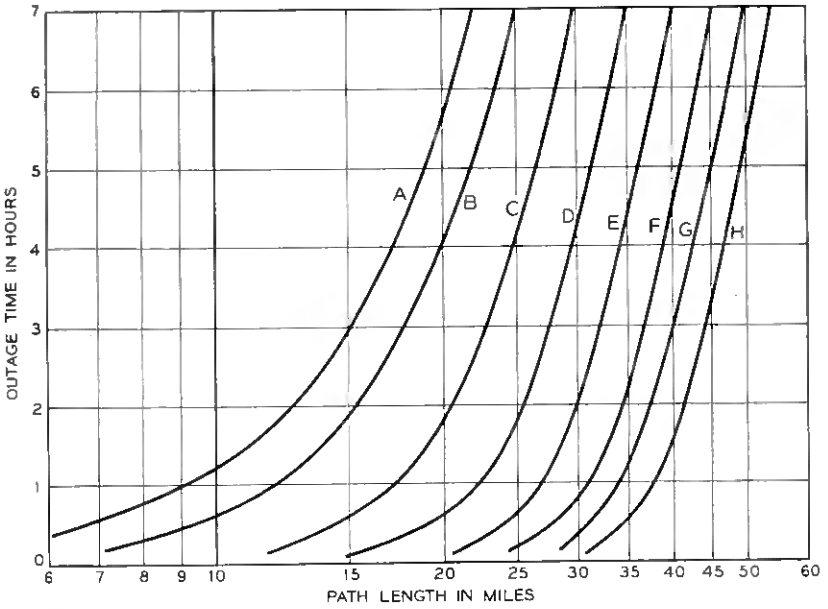


Fig. 13 — Expected outage time in hours per year vs. path length in miles for various areas of the United States.

the individual hops meet the same objective. From a practical standpoint, this will not always be possible. Sometimes it is necessary to have one or more hops of a system electrically long; they will have insufficient fading margin and hence contribute more than their share of outage time. From the over-all system viewpoint, this "excess" must be made up by imposing tighter requirements on the remaining hops.

To meet the over-all system objective, it becomes necessary to know the contributions of the long hops—those having a fading margin less than 40 db. Fig. 14 shows excess path loss due to rain versus hours per year for the Mobile-Mount Vernon path. The shape of this curve is nearly identical with Bussey's curve of cumulative distribution for point rates at Washington, D. C. If we assume the shape of this curve to be representative for other areas of the country, then the additional outage time for path lengths given by Fig. 13 can be estimated for hops having a fading margin less than 40 db. The data shown in Fig. 14 have been rationalized and are shown in Fig. 15 as an estimate of additional outage time.

Sometimes it is practical to shorten a proposed path to bring the fading margin up to 40 db. An approximation of the necessary reduction

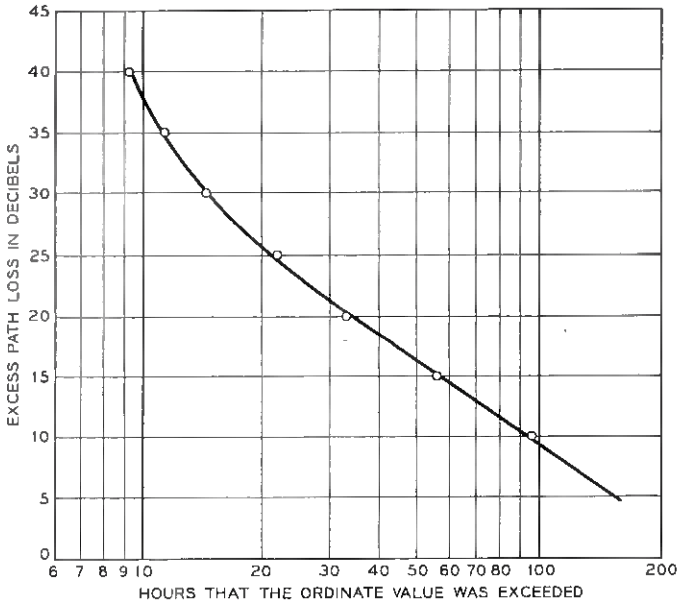


Fig. 14 — Excess path loss due to rainfall vs. hours per year (at Mobile).

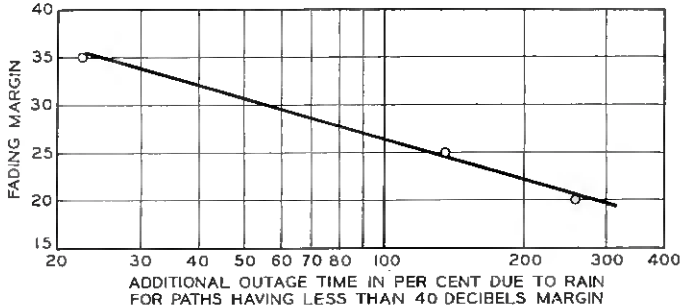


Fig. 15 — Additional outage time expected for 11-kmc systems having a fading margin less than 40 db.

in path length can be made if uniform rainfall rate is assumed over the path. Under this condition, Ryde shows the attenuation due to rainfall to be directly proportional to the path length. Thus the path length given in Fig. 13 can be shortened to correct for insufficient fading margin.

Rainfall in the extreme southeastern region of the United States will limit 11-kmc radio systems having a 40-db fading margin to path lengths

of approximately 10 to 15 miles, depending on the number of hops, if normal reliability objectives are to be met. Path lengths of 20 to 30 miles should be acceptable in the central area and paths as long as 35 miles should be acceptable in the northwestern part of the country. However, in existing short-haul radio systems, the paths average 22 to 24 miles, due to considerations other than those of propagation. It would then appear that 11-kmc systems will not be penalized unduly except in the extreme southeastern part of the United States.

V. INSTRUMENTATION

5.1 *Transmitter*

The transmitter employed a small commercially available klystron whose output was 0.5 watt. A variable probe was used to match the klystron to the waveguide to the antenna, and a 20-db directional coupler sampled the output so that frequency could be measured and the output power monitored. The frequency stability was such that it was not necessary to use automatic frequency control. Since the ac line was subject to frequent failure (a natural result of the thunderstorms whose rainfall provided the reason for the experiment), a strip-chart recorder with a mechanical clock drive was used to monitor the transmitter output.

The transmitter with its power supply was mounted in a weather resistant cabinet, as shown in Fig. 16.

5.2 *Receiver*

The receiver was adapted from equipment designed to record path loss at 4 kmc,* which was, in turn, adapted from equipment designed to measure path reflections.[8] It is shown in block schematic form in Fig. 17. The normal received signal level was −32.2 dbm at Mobile and −31.1 dbm at Axis. The receiver was arranged to record signals from 0 to −40 db relative to the normal signal.

A balanced converter supplied by a local klystron oscillator modulated the incoming 11.4-kmc signal to a 60-mc intermediate frequency. In the preamplifier the 60-mc signal was amplitude modulated with 1000-cps. The output of the preamplifier was divided between an IF amplifier feeding a frequency discriminator, which provided automatic frequency

* An extensive path loss measuring program was carried out in 1947–1950 at 4 kmc prior to commercial use of those frequencies by the Bell System. This 4-kmc equipment was designed by H. C. Franke and was converted to 11 kmc by S. D. Hathaway.

Fig. 16 — 11-kmc transmitter, power supply and antenna, Mount Vernon, Alabama.

control to the local klystron oscillator, and an amplitude detector, where a 1000-cps signal reasonably proportional to the microwave input signal over a 50-db range was recovered. The 1000-cps signal was amplified and rectified at a level suitable to operate the display equipment. The receiver at Axis (except for the converter, which was tower-mounted) and the display equipment are shown in Fig. 18.

5.3 *Display Equipment*

Two types of display equipment were used:

(a) A level distribution recorder with a range of 40 db, which operated

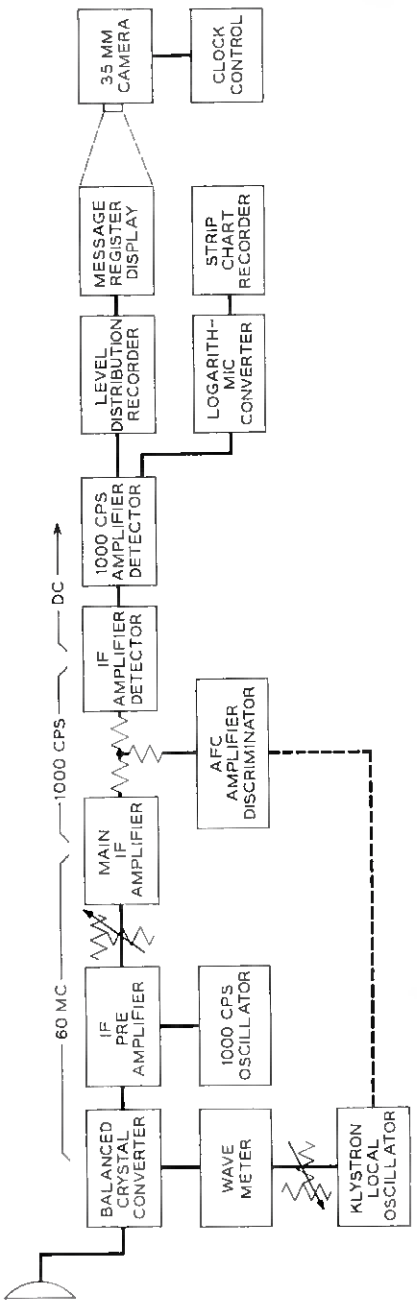


Fig. 17 — Block schematic of radio receiver and recording equipment.

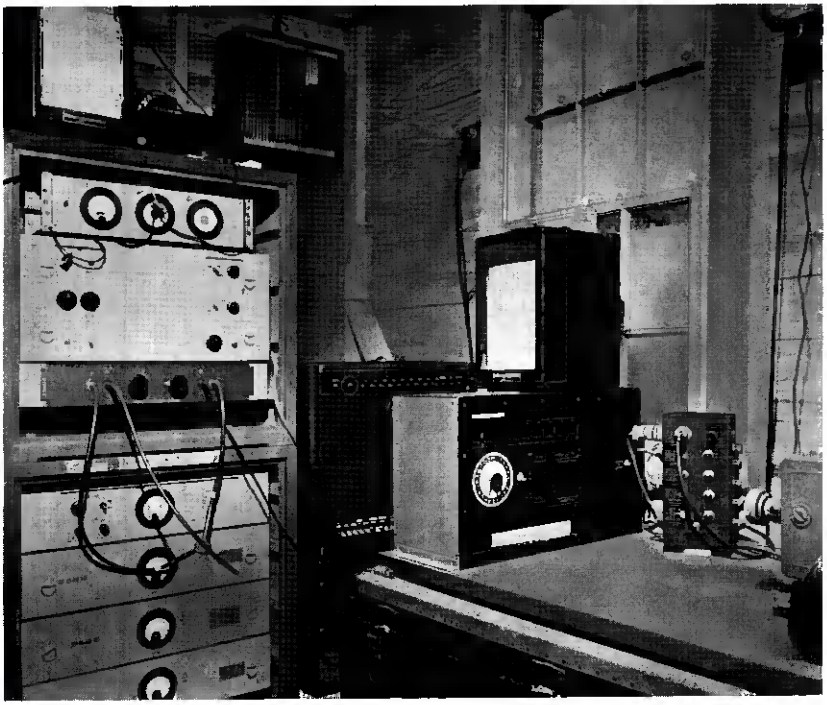


Fig. 18 — Axis station — IF portion of radio receiver, display equipment.

message registers which were photographed automatically on 35-mm film every 0.1 hour.

(b) A strip-chart recorder with a logarithmic converter to produce a scale linear in db over a 50-db range.

5.4 *Level Distribution Recorder**

A series of nine dc slicer circuits, each arranged to operate at an input of one volt, was connected to a voltage divider at 5-db intervals, thus covering a range of 40 db, as shown in Fig. 19.

A 2-mf capacitor was connected by a relay actuated by a synchronous timer to a cathode follower output of the 1000-cps rectifier for 0.85 sec-

* This electronic level-distribution recorder was developed in 1946 to replace a relay device that had been used for many years to study distributions of talker volume and telephone circuit noise. This circuit was conceived by L. Y. Lacy and developed by C. R. Eckberg for mobile use in connection with investigations of VHF transmission to vehicles, and later modified by H. C. Franke for use in microwave propagation measurements.

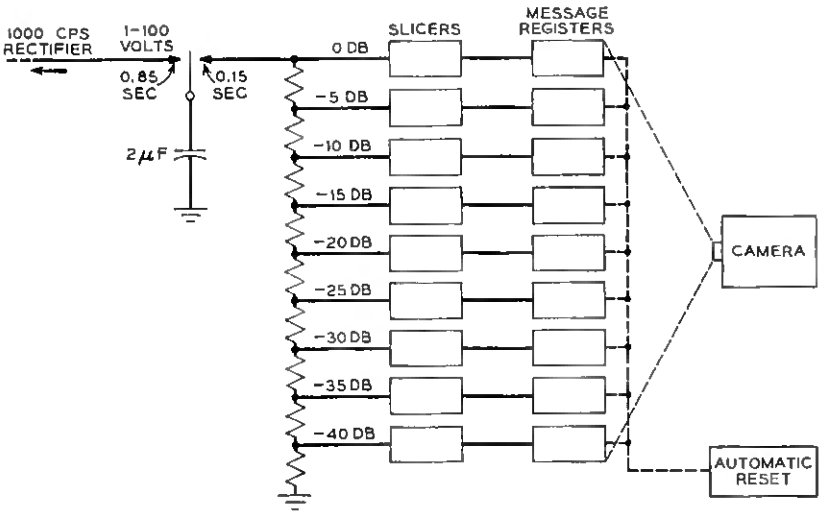


Fig. 19 — Block schematic of level distribution recorder.

ond. Then, for 0.15 second, the capacitor was connected to the voltage-divider slicer circuit. The slicer circuits were arranged to drive message registers (counters) with the polarity such that, if the slicer threshold was exceeded, the message register would not count; if the slicer threshold was not exceeded, the message register counted. Thus, an input of 100 volts (normal signal) caused no counts. An input of 0.9 volt (41-db fade) caused all message registers to count.*

Standard 14-type telephone message registers were used, and a neon tube was connected across each message register for easy observation of individual counts.†

Means were provided to calibrate the level distribution recorder in a preliminary way from an accurate dc source, but final calibration was always made from an accurate signal generator connected to the radio receiver input, so as to reduce the effects of nonlinearity in the radio receiver.

* Other level distribution recorders have been built with crossgating between the slicers, so that only the message register corresponding to the level just exceeded will operate. This type yields a histogram data presentation, whereas radio fading data are usually presented as a cumulative distribution.

† Some message register units were arranged with an automatic reset mechanism that reduced all message register readings to zero hourly. This simplified reducing the data, since smaller numbers had to be dealt with, but the delicate mechanism of the resettable registers available at the time the equipment was designed led to maintenance problems in the field, so these units were rebuilt to use the simpler registers.

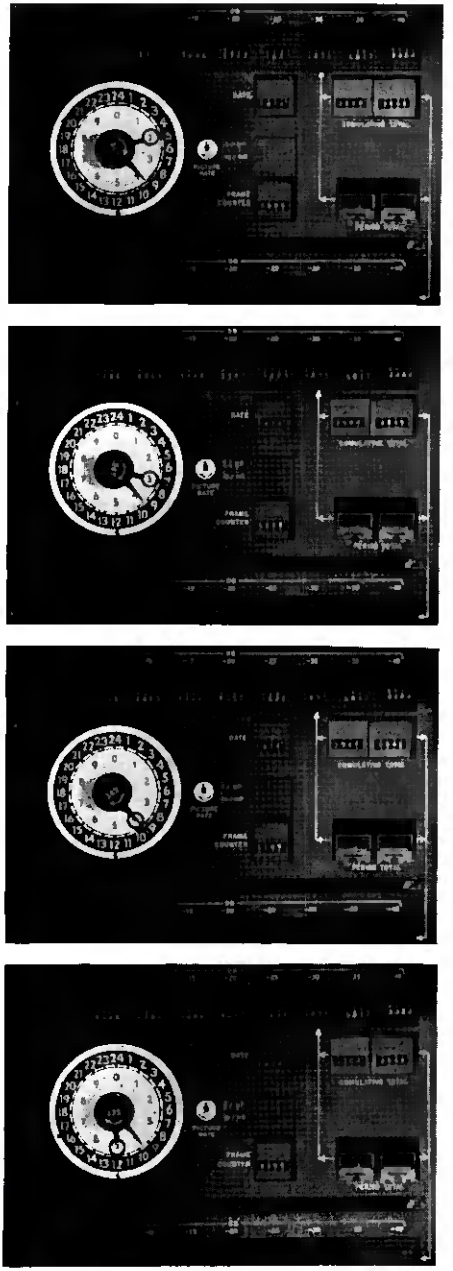


Fig. 20 — Level distribution recorder camera film (enlarged two times).

5.5 *Camera Equipment*

Pictures of the message registers were taken every 0.1 hour by a 35-mm camera adapted from a movie camera, under the control of a 24-hour electric clock which was included in the message register panel so that the time of each picture was recorded. A series of pictures is shown in Fig. 20. Auxiliary message registers recorded the total number of measurements and the date, the latter register being actuated by the 24-hour clock. The message register units and cameras were enclosed in a trunk to reduce spurious illumination. A projection film reader was used to transfer the data to a data book for analysis.

5.6 *Logarithmic Converter*

The logarithmic converter accepted a dc input voltage in the range -0.3 to -100 volts (as developed by the 1000-cps detector) and changed it to a direct current proportional to the logarithm of the input voltage, with a range of 0–1 ma for the operation of a strip-chart recorder.

The dc input voltage was chopped at a 60-cps rate and applied to a differentiating circuit followed by a dc slicer circuit. The time that the differentiated wave exceeded the threshold of the slicer was proportional to the logarithm of the input voltage, so that the output of the slicer was a 60-cps wave with pulse length modulation proportional to the input voltage in decibels. This output was filtered* and applied to a strip-chart recorder to provide a linear 0–50 db recording.

This display was limited by the slow response of the strip-chart recorder, whose time constant was about 0.5 second, so that the strip-charts were used chiefly for monitoring and quick scanning of data.

5.7 *Rain Gauges*

Automatic recording tilt-bucket rain gauges† were placed approximately every two miles along the radio path, as shown in Fig. 2. The exact locations were determined by considerations of accessibility, since many of the roads across the radio path were little more than swamp traces. Also, an effort was made to minimize the effect of nearby objects such as trees and buildings.

The rain gauge mechanisms, shown in Fig. 21, were proportioned so that the bucket tilted after each 0.01 in. of rain fell. A magnet attached

* It was found that imperfect filtering was desirable, in that a small amount of the 60-cps component improved the response of the strip-chart recorder to small changes in input.

† The rain gauges were designed by L. E. Hunt.

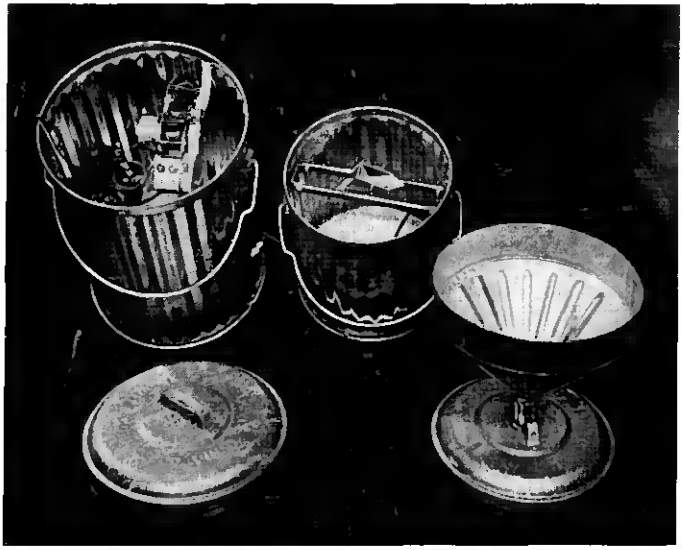


Fig. 21 — Rain gauge assembly: (a) recorder; (b) tilt bucket; (c) funnel.

to the tilt bucket closed the contacts of a glass-sealed switch* during the instant of tilting.

5.8 *Rain Gauge Recorders*

Because of sparse wire facilities along the path, it was impossible to bring each rain gauge circuit to a central recorder, so individual strip-chart recorders were used at each rain gauge. These used teletypewriter tape driven† by a clock‡ at the speed of 0.1 in. per minute (a little over 11 ft per day), and a stylus driven by an electromagnet punched a tiny hole in the tape each time the rain gauge bucket tilted. Thus at the cloudburst rate of fall of eight inches per minute, the punch marks were just over 0.01 in. apart, so that careful examination under a strong glass was required.

* Earlier models used open contacts which worked well in the laboratory but proved unsatisfactory under field conditions.

† The usual trouble of changes of paper dimension caused much experimentation with chart drive. Pins in the drive drum were superseded by a neoprene band friction drive.

‡ Several types of clocks were tried — automobile electric clocks, precision automobile electric clocks, large spring-wound clocks and, finally, governor-controlled dc motors. None yielded the precision of timing desired, so that it was necessary to interpolate to avoid errors in correlating the individual rain gauge records for particular rain events. The improvements in the recorders were made by K. J. Frolund.

VI. ACKNOWLEDGMENTS

The authors are indebted to K. J. Frolund, G. H. Swan, D. K. White and others, without whose help this paper would not have been possible. The assistance of the Southern Bell Telephone Company and the Central Area of the Long Lines Department of the American Telephone and Telegraph Company is deeply appreciated.

REFERENCES

1. Ryde, J. W. and Ryde, D., *Attenuation of Centimeter Waves by Rain, Hail, Fog and Clouds*, General Electric Co., Ltd., Wembley, England, 1945.
2. Ryde, J. W., *Meteorological Factors in Radio Wave Propagation*, The Physical Society, London, England 1946, pp. 169–188.
3. Laws, J. O. and Parsons, D. A., The Relationship of Raindrop Size to Intensity, Trans. Amer. Geophysical Union, 24th Annual Meeting, 1943, pp. 452–460.
4. Bentley, W. A., Studies of Raindrops and Raindrop Phenomena, Monthly Weather Rev., **32,** 1904, pp. 450–456.
5. Bussey, H. E., Microwave Attenuation Statistics Estimated from Rainfall and Water Vapor Statistics, Proc. I.R.E., **38,** July 1950, pp. 781–785.
6. *Precipitation on the Muskingum River Watershed by 30-Minute Intervals*, U. S. Dept. of Agriculture, Soil Conservation Service, Washington, D. C., 1938.
7. Portable Microwave Tower, Bell Labs. Record, **2,** Jan. 1948, pp. 6–8.
8. Campbell, R. D., Path Testing for Microwave-Radio Routes, Trans. A.I.E.E., Pt. I, **72,** July 1953, pp. 326–334.
9. Hitschfeld, W., private communication, November 22, 1954.
10. Dyck, H. D. and Mattice, W. A., A Study of Excessive Rainfall, Monthly Weather Rev., **69,** October 1941, pp. 293–301.

Space-Charge Wave Excitation in Solid-Cylindrical Brillouin Beams

By W. W. RIGROD and J. R. PIERCE

(Manuscript received April 24, 1958)

The voltage and current modulation of ideal cylindrical electron beams in Brillouin flow, as well as beams in zero magnetic field, are studied by means of Laplace transforms. With a large-diameter beam of this class, suddenly accelerated from a temperature-limited cathode and without transverse velocities, the minimum noise figure of an amplifier is found to be smaller than it would be for a narrow, essentially one-dimensional (filament or sheet) beam, or for a confined-flow beam with the same diameter, longitudinal velocity and direct current.

Certain space-charge wave solutions obtained in field analyses of beams from shielded diodes, which have never been detected experimentally, are found to be nonexistent in the sense that no phenomenon taking place in a vacuum tube excites them.

I. INTRODUCTION

When a beam only partly fills the space within a concentric drift tube, the field patterns of the modes derived by small-signal slow-wave analysis are not orthogonal to one another. This makes it difficult to find the amplitude of any single mode excited by an arbitrary initial disturbance. The cases of ion-neutralized beams in the absence of a magnetic field and of Brillouin flow are even more difficult, for in these cases infinite groups of modes assume the same phase velocity and degenerate into a wave of arbitrary transverse distribution, which, we shall show, cannot be excited at all.

In treating the excitation of a confined-flow beam, Scotto and Parzen[1] have circumvented such difficulties by means of a Laplace transform procedure. More recently, Bresler, Joshi and Marcuvitz[2] have succeeded in formulating a complete set of orthogonal modes for such unidirectional electron beams, at the cost of some increased complexity in description.

In this paper, a technique similar to that of Scotto and Parzen will be employed to solve several problems in the excitation of a solid-cylin-

drical beam, focused in ideal Brillouin flow. The method consists of transforming the exciting current or voltage with respect to the axial coordinate z, and finding the beam response by means of a transfer function which satisfies the transverse boundary conditions. The relative amplitudes of each of the various modes could be found, in this way, by using transfer functions evaluated in terms of each such mode. Here, only the fundamental mode, having axial symmetry, will be considered. The solutions so obtained will also apply to the beam in zero magnetic field, as the mode patterns are the same in both cases.[3]

The first problem treated, of field modulation by means of an annular gap in a concentric drift tube, will illustrate the general technique. The remaining three calculations deal with different aspects of the problem of noise excitation of a finite-diameter beam in a shielded diode, in which the effect of transverse electron motions is disregarded. These calculations show that the "noisiness" of such a beam falls to half that for a narrow beam or a one-dimensional beam as the diameter is increased (as βb is made larger). An additional calculation shows that certain space-charge waves obtained in field analyses of such beams,[4, 5] which are independent of transverse boundary conditions, cannot be excited and therefore do not exist.

The prospects of producing low-noise amplifiers with large-diameter beams in Brillouin flow are not very good, because of large transverse electron excursions near the cathode. However, it is possible that a similar noise-reduction mechanism may be present in confined-flow beams abruptly hollowed-out (relative to the cathode surface) close to the cathode. The extremely low noise figures reported[6, 7] for TWT amplifiers using beams of this sort are chiefly due to other noise-reduction processes,[8, 9] but the effect of large beam size may perhaps be important at higher frequencies.

II. MODULATING VOLTAGE ACROSS GAP IN DRIFT TUBE

At the input plane, $z = 0$, an ac voltage V is impressed across a very short gap in a drift tube of radius a, concentric with and enclosing a Brillouin-flow beam of radius b. The response is sought in the form of the total current in the drift tube to the right of this plane, $i_t(z, a)$. Polar cylindrical coordinates (r, θ, z) and MKS units will be employed, consistent with the notation of Ref. 5, in which axial-symmetric space-charge waves in beams of this type are described. All of the ac quantities associated with any such wave are assumed to propagate as

$$\exp(j\omega t) \cdot \exp(-j\beta z) \tag{1}$$

with the time variation suppressed.

As the amplitudes of all ac quantities are zero to the left of the input plane, it is convenient to use the Laplace transform pairs in the form[10]

$$F(\beta) = \int_0^\infty f(z) \exp(j\beta z)\, dz, \tag{2}$$

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\beta) \exp(-j\beta z)\, d\beta. \tag{3}$$

The integration contour for the inverse transform (3) is along the real axis of β, indented above any poles on that axis, and enclosing the third and fourth quadrants of the complex β-plane. When $F(\beta)$ has n simple, distinct poles within this contour, the last integral can be evaluated by means of Cauchy's residue theorem, for $z > 0$, as

$$f(z) = -j \sum_n [(\beta - \beta_n) F(\beta) \exp(-j\beta z)]_{\beta=\beta_n}. \tag{4}$$

Accordingly, the transform of the impressed field, in a gap of nominal (but negligible) width d, is

$$A(\beta) = \int_0^d (-V/d) \exp(j\beta z)\, dz \cong -V. \tag{5}$$

The response current is found by multiplying this quantity by a transfer function $Y(\beta)$ to obtain the transform of that current, and then its inverse transform. Any transfer function relating two ac quantities with the same (z, t) variation will, in general, be a function of the propagation constant β and the transverse properties of the electron beam and its cylindrical enclosure. The transfer function $Y(\beta)$ relating the ac amplitudes $i_t(z, a)$ and $E_z(z, a)$ will therefore be the same as that relating their transforms $\hat{\imath}_t(\beta, a)$ and $\hat{E}_z(\beta, a)$. In the present instance, the z-component of the field equation for curl $\mathbf{H}$ provides the desired relation defining $Y(\beta)$:

$$\hat{\imath}_t(\beta, a) = 2\pi a \hat{H}_\theta = 2\pi a \hat{E}_z Y(\beta), \tag{6}$$

and the response current is given by

$$\begin{aligned} i_t(z, a) &= -aV \int_{-\infty}^{\infty} Y(\beta) \exp(j\beta z)\, d\beta \\ &= j2\pi a V \sum_n [(\beta - \beta_n) Y(\beta) \exp(-j\beta z)]_{\beta=\beta_n}. \end{aligned} \tag{7}$$

The boundary equations at the surface of a drifting Brillouin-flow beam[5] must be solved in terms of $Y(\beta)$, rather than of the infinite ad-

mittance of a conducting wall. The axial electric field inside and outside of the beam, respectively, can be written

$$E_z = AI_{0r} \qquad (0 \leqq r \leqq b), \tag{8}$$

$$E_z = BI_{0r} + CK_{0r} \qquad (b \leqq r \leqq a), \tag{9}$$

where A, B, C are arbitrary constants, and I and K stand for the modified Bessel functions of the first and second kinds, respectively, the first subscript denoting the order number and the second the radius. The propagation factor, $\exp j(\omega t - \beta z)$, as well as the argument (βr) are omitted for brevity, and will be omitted elsewhere when they are unambiguous.

The surface ripple due to ac radial electron motions can be represented by a surface charge density,

$$\sigma = -R\epsilon E_r = -\frac{j\epsilon R}{\beta}\frac{\partial E_z}{\partial r}, \tag{10}$$

evaluated in terms of the fields just inside of the beam, where R is the square of the reciprocal of the space-charge reduction factor, p, defined in terms of the radian plasma frequency ω_p, the excitation frequency ω and the beam drift velocity u, as follows:

$$R = \frac{1}{p^2} = \frac{\omega_p^{\,2}}{(\omega - \beta u)^2} = \frac{\beta_p^{\,2}}{(\beta - \beta_e)^2}. \tag{11}$$

In the last expression, $\beta_p = \omega_p/u$ and $\beta_e = \omega/u$.

The boundary equations at $r = b$ can then be written

$$AI_{0b} = BI_{0b} + CK_{0b}, \tag{12}$$

$$A(1 - R)I_{1b} = BI_{1b} - CK_{1b}, \tag{13}$$

and the admittance function is

$$\begin{aligned} Y(\beta) &= \left(\frac{H_\theta}{E_z}\right)_{r=a} = \frac{\omega\epsilon}{\beta}\left(\frac{E_r}{E_z}\right)_{r=a} \\ &= \frac{j\omega\epsilon}{\beta}\frac{I_{1a}}{I_{0a}}\left[\frac{(B/C) - (K_{1a}/I_{1a})}{(B/C) + (K_{0a}/I_{0a})}\right]. \end{aligned} \tag{14}$$

Substitution here of (B/C), found by solving the two boundary equations, yields:

$$Y(\beta) = \frac{j\omega\epsilon}{\beta}\frac{I_{1a}}{I_{0a}}\frac{p^2 - wn}{p^2 - wm}, \tag{15}$$

where

$$w = \beta b I_{1b} K_{0b}, \tag{16}$$

$$n = 1 + \frac{I_{0b} K_{1a}}{K_{0b} I_{1a}}, \tag{17}$$

$$m = 1 - \frac{I_{0b} K_{0a}}{K_{0b} I_{0a}}. \tag{18}$$

By writing

$$\frac{p^2 - wn}{p^2 - wm} = \frac{(\beta - \beta_e)^2 - \beta_p{}^2 wn}{(\beta - \beta_e)^2 - \beta_p{}^2 wm} \tag{19}$$

it is readily verified that, along the real β-axis, the only poles of $Y(\beta)$ are

$$\beta_{1,2} = \beta_e \pm \beta_p \sqrt{wm} = \beta_e \pm \beta_q, \tag{20}$$

in terms of which

$$p^2 - wm = (\beta - \beta_1)(\beta - \beta_2). \tag{21}$$

In addition, as the integration contour encloses the third and fourth quadrants of the β-plane, the term $I_0(\beta a)$ contributes poles at each of the zeros of $J_0(x)$ along the negative imaginary β-axis. For each root x_n, the pole is $\beta = (jx_n/a)$, so that the corresponding residue contains a factor $\exp(-x_n z/a)$. Since such terms decay rapidly with distance z from the input plane, and we are solely interested in propagating waves, they will not be considered further.

If the changes in w, m and n due to changes of β are neglected, by evaluating all Bessel-function arguments at β_e, the expression for the current response reduces to the following:

$$\begin{aligned} i_t(z, a) &\doteq -\pi\omega\epsilon \frac{\beta_p}{\beta_e} aV \frac{I_{1a}}{I_{0a}} \frac{w(m - n)}{\sqrt{wm}} [\exp(-j\beta_1 z) - \exp(-j\beta_2 z)] \\ &\doteq \frac{-jV 2\pi\epsilon\omega_q \sin \beta_q z \exp(-j\beta_e z)}{\beta_e I_{0a}{}^2 \left(\frac{K_{0b}}{I_{0b}} - \frac{K_{0a}}{I_{0a}}\right)}. \end{aligned} \tag{22}$$

In klystron theory, it is customary to write this quantity in another form, by introducing the dc beam current I_0 and voltage V_0. For a beam with negligible potential depression,

$$\frac{I_0}{V_0} = \frac{2\pi\epsilon\omega_p{}^2 b^2}{u}. \tag{23}$$

With this and the reduction factor $p = \sqrt{wm}$, we obtain

$$i_1(z, a) \doteq \frac{-jVI_0}{V_0}\left(\frac{I_{0b}I_{1b}}{I_{0a}{}^2}\right)_{\beta=\beta_e} \frac{\sin \beta_q z \exp(-j\beta_e z)}{\beta_q b}. \tag{24}$$

Beck[11] has treated this problem in a slightly different way, by introducing several additional approximations. His result consists of the above expression, followed by a smaller second term. The present derivation shows that this latter term should be simply zero.

III. MODULATION BY INJECTED FILAMENT OF NOISE CURRENT

The response of a one-dimensional beam to injected noise current has been computed by one of the authors[10] with the Laplace transform technique described above. Within the framework of its assumptions, this computation led to results in agreement with the work of Rack, Llewellyn and Peterson,[12] thereby establishing its validity as an alternative procedure. It is now proposed to extend this treatment to the noise excitation of a finite-diameter beam in Brillouin flow or in zero magnetic field, and with an infinitely remote outer conducting tube. The treatment will be for a source of electrons with no transverse velocities. This may be unrealistic, but it is not unphysical, for such a source can be approximated by collimating the electron flow from a cathode by means of an array of holes, such as a thick hexagonal grid. First, the response will be found to a slender filament of noise charge injected at the axis of this beam, and later on the response will be calculated for noise-charge modulation over the entire beam area. Comparison of the results with those for the one-dimensional beam should reveal the effect of beam diameter on its noisiness.

The approximations used in the one-dimensional computation[10] are to be adopted here as well, and the reader is referred to Ref. 10 for a detailed discussion of their meaning. Effects due to the multivelocity nature of the beam and the inertial effects of a space-charge cloud during acceleration are avoided by assuming the beam to be abruptly accelerated from a temperature-limited cathode. The modes of propagation of the beam are assumed to be indistinguishable from those for a beam without thermal velocities. Excitation of Landau-type damped plasma oscillations,[13] which tend to decelerate fast-entering charges, is neglected.

The noise excitation due to injected charge in each velocity class is calculated in a narrow frequency band, and its mean square summed over all velocity classes, restricted to a small spread about the mean beam velocity. The beam is thus regarded as a linear impedance through which the exciting charges flow. The entering charges are treated as

current filaments with discrete velocities, which are modulated by the noise field due to all the other charges, but have no separate identities with respect to entering times.

The Brillouin beam is taken to have radius b, and to be drifting in free space. In a narrow frequency band about ω, the injected filament with velocity v can be regarded as a circular electron stream of radius δ, carrying a convection current

$$i_1 = i_0 S(z) \exp(-j\gamma z), \tag{25}$$

$$\gamma = \frac{\omega}{v}, \tag{26}$$

where S is the unit step function, and z is measured from the entering plane. This current corresponds to an ac charge density

$$\rho_1 = \frac{\gamma}{\omega} \frac{i_1}{\pi\delta^2}. \tag{27}$$

The total charge density ρ_t at the input plane must satisfy Poisson's equation

$$\rho_t = \rho + \rho_1 = \epsilon \operatorname{div} \mathbf{E}, \tag{28}$$

where ρ is the induced charge density in the driven beam, consistent with the dynamics and charge-conservation equations for axial-symmetric space-charge waves in Brillouin-flow beams:

$$\rho = R\epsilon \operatorname{div} \mathbf{E}. \tag{29}$$

Thus the total charge density at the input plane is related to the injected charge density ρ_1 as follows:

$$\rho_t = \frac{\rho_1}{1 - R} = \frac{\gamma i_1}{\omega(\pi\delta^2)(1 - R)}. \tag{30}$$

Outside of the radius δ the charge density is zero, and the axial electric field up to the rim of the beam can be written:

$$E_{z1} = A\, I_{0r} + B\, K_{0r}, \tag{31}$$

omitting the propagation factor $\exp(-j\beta z)$ for brevity. In terms of these constants, the total charge per unit length within the very small radius δ is then

$$q_t = j2\pi\delta\epsilon(AI_{1\delta} - BK_{1\delta}) \leqq -\frac{j2\pi\epsilon B}{\beta} \tag{32}$$

for $\beta\delta \ll 1$.

In the unbounded space outside of the beam, the longitudinal electric field must have the form

$$E_{z2} = C\,K_{0r}. \tag{33}$$

Taking the surface charge of the beam into account, the boundary equations at radius b are

$$A\,I_{0b} + B\,K_{0b} = C\,K_{0b}\,, \tag{34}$$

$$(1 - R)(A\,I_{1b} - B\,K_{1b}) = -C\,K_{1b}\,. \tag{35}$$

The total current inside of a cylinder of radius $r > b$ is

$$i_t(r) = 2\pi r H_\theta = \frac{j2\pi r\omega\epsilon}{\beta^2}\frac{\partial E_{z2}}{\partial r} = -\frac{j2\pi r\omega\epsilon}{\beta}CK_{1r}\,. \tag{36}$$

To obtain the transfer function needed in this problem, a relation between the injected current i_1 and the total induced current $i_t(r)$, or between their transforms $\hat{\imath}_1$ and $\hat{\imath}_t(r)$, the boundary equations must be solved for the constant C, as follows:

$$B = \frac{j\beta q_l}{2\pi\epsilon} = \frac{j\gamma\beta i_1}{2\pi\omega\epsilon(1 - R)}\,, \tag{37}$$

$$C = \frac{B(1 - R)}{1 - R\beta bI_{1b}K_{0b}} = \frac{j\gamma\beta i_1}{2\pi\omega\epsilon(1 - R\beta bI_{1b}K_{0b})}\,, \tag{38}$$

$$i_t(r) = \frac{\gamma rK_{1r}i_1}{1 - R\beta bI_{1b}K_{0b}} = F(\gamma, \beta)i_1 \tag{39}$$

and

$$\hat{\imath}_t(\beta, r) = F(\gamma, \beta)\hat{\imath}_1(\beta), \tag{40}$$

where

$$\hat{\imath}_1(\beta) = i_0\int_0^\infty \exp j(\beta - \gamma)z\,dz = \frac{ji_0}{\beta - \gamma}\,. \tag{41}$$

The response current within the radius r is thus

$$i_t(r, z) = -\frac{i_0}{2\pi j}\int_{-\infty}^{\infty}\frac{F(\gamma, \beta)\,\exp\,(-j\beta z)\,d\beta}{\beta - \gamma} \tag{42}$$

$$= i_0\sum_n\left[\frac{(\beta - \beta_n)F(\gamma, \beta)\,\exp\,(-j\beta z)}{\beta - \gamma}\right]_{\beta=\beta_n}. \tag{43}$$

The integrand

$$\frac{F(\gamma, \beta)}{\beta - \gamma} = \frac{(\gamma r)(\beta - \beta_e)^2K_{1r}}{(\beta - \gamma)[(\beta - \beta_e)^2 - \beta_p{}^2\beta bI_{1b}K_{0b}]} \tag{44}$$

has four poles:

$$\beta_{1,2} = \beta_e \pm \beta_p(\beta b I_{1b} K_{0b})^{1/2} = \beta_e \pm \beta_q \tag{45}$$

and

$$\beta_3 = \gamma, \qquad \beta_4 = 0. \tag{46}$$

The pole of $K_1(\beta r)$ at $\beta = 0$ contributes a residue $-i_0$, which serves to make $i_t(r, z)$ zero at zero frequency. This is consistent with the formulation of the problem, in which the dc component of the entering charge is neglected, and the beam itself manifests its dc current only in the plasma wave number. However, as the calculation is only valid for slow-wave propagating modes ($\beta > k$), this residue will be disregarded.

As before, the resultant expression is simplified by neglecting the small rate of change of the Bessel functions with β, replacing β by β_e where this error is small. With the time factor suppressed, the result is

$$\begin{aligned} i_t(r, z) = {} & i_0\gamma r K_1(\beta_e r)\left[\frac{\beta_q \exp(-j\beta_1 z)}{2(\beta_1 - \gamma)} - \frac{\beta_q \exp(-j\beta_2 z)}{2(\beta_2 - \gamma)}\right] \\ & + i_0\gamma r K_1(\gamma r)\left[\frac{(\gamma - \beta_e)^2 \exp(-j\gamma z)}{(\gamma - \beta_e)^2 - \beta_q^2}\right]. \end{aligned} \tag{47}$$

The assumption of small velocity spread in the entering charges, centered about the mean velocity u of the beam, permits the definition of a small quantity associated with each value of v:

$$\epsilon = \frac{v - u}{v} \ll 1, \tag{48}$$

$$(\gamma - \beta_e)^2 = (-\epsilon\beta_e)^2 \approx 0, \tag{49}$$

such that only terms up to first order in ϵ need be retained, to a good approximation. The expression for total current response then reduces to

$$i_t(r, z) \cong i_0(\gamma r)K_1(\beta_e r)\exp(-j\beta_e z)\left(\cos\beta_q z + j\epsilon\frac{\omega}{\omega_q}\sin\beta_q z\right). \tag{50}$$

The total current in the drifting beam, $i_t(b)$, is related to the total convection current, $i_c(b)$, by the ratio:[5]

$$\frac{i_c(b)}{i_t(b)} = \frac{R}{R - 1} = \frac{1}{1 - \beta b I_{1b} K_{0b}} = \frac{1}{\beta b I_{0b} K_{1b}}. \tag{51}$$

Thus,

$$i_c(b, z) = \frac{i_0 \gamma b \exp(-j\beta_e z)}{\beta_e b I_{0b}} \left(\cos \beta_q z + j\epsilon \frac{\omega}{\omega_q} \sin \beta_q z \right), \tag{52}$$

the argument of the Bessel function understood to be $(\beta_e b)$ here.

The beam responses due to electrons in different velocity ranges are assumed to add in a mean square manner. In each velocity class, the impressed current has only shot noise. Thus, using the subscript n for each velocity class, the mean square impressed current in each class is

$$i_n^2 = 2eI_n \Delta f, \tag{53}$$

where e is the electronic charge, Δf the bandwidth about $f = \omega/2\pi$, and I_n the direct current in the nth velocity class. The mean square convection current response in the beam, due to i_n, is

$$| i_c^2 |_n = \frac{2eI_n \Delta f}{I_{0b}^2} (\cos^2 \beta_q z + \epsilon_n^2 \sin^2 \beta_q z), \tag{54}$$

where ϵ_n is associated with v_n as in (48) and, approximately,

$$\overline{\gamma_n^2} \cong \beta^2 \cong \beta_e^2. \tag{55}$$

The total mean square convection current is then

$$| i_c^2 | = \frac{2eI_0 \Delta f}{I_{0b}^2} (\cos^2 \beta_q z + \bar{\epsilon}^2 \sin^2 \beta_q z), \tag{56}$$

where I_0 is the total direct current in the injected filament, and

$$\bar{\epsilon}^2 = \frac{\sum_n I_n \epsilon_n^2}{I_0} \cong \frac{1}{u^2} \sum_n \frac{I_n}{I_0} (v_n - u)^2, \tag{57}$$

assuming that

$$\epsilon_n^2 = \frac{(v_n - u)^2}{(v_n)^2} \cong \frac{(v_n - u)^2}{u^2}, \tag{58}$$

where u is the average velocity, given by

$$u = \frac{1}{I_0} \sum_n I_n v_n . \tag{59}$$

The expression for $| i_c^2 |$ in the finite beam is the same as that previously obtained in the one-dimensional analysis,[10] except for the presence of β_q in place of β_p within the brackets, and the term I_{0b}^2 in the denominator. Thus the maximum value of $| i_c^2 |$ is less than the total

impressed shot-noise current by the factor $1/I_{0b}^2$, which is smaller, the larger the beam diameter.

IV. NOISE-CURRENT MODULATION OVER ENTIRE BEAM AREA

The beam of the previous section is now supposed to be uniformly modulated by impressed noise current over its entire area, subject to all of the assumptions and conditions stipulated earlier. Since the space-charge mode of interest has axial symmetry, the contribution to the total induced current by any entering charge filament is independent of its angular position. The elementary areas of excitation can be taken to be thin rings (r to $r + \delta r$), for which the transfer function relating the induced to the exciting current is the same for noise-current modulation in each velocity class as for coherent rings of injected charge, of the same velocity.

The rms charge in a ring of current with velocity v is related to the rms current in the nth velocity class by

$$dq_n = \frac{di_n}{v}, \tag{60}$$

where

$$di_n = (\overline{di_n^2})^{1/2} = [(J_n 2\pi r \delta r)(2e\Delta f)]^{1/2}, \tag{61}$$

J_n being the portion of the uniform current density with this velocity. As in the previous section, the total ring of charge at the input plane is related to this current element by

$$dq_t = \frac{dq_n}{1 - R} = \frac{\gamma\, di_n}{\omega(1 - R)} \tag{62}$$

and

$$di_n = |\, di_n \,| \exp(-j\gamma z), \tag{63}$$

where, as before, $\gamma = \omega/v$.

To evaluate the transfer function giving the current within some radius a, outside of the beam (radius b), the cross section is divided into three regions, separated by the rings of charge at radii r and b:

$$E_{z1} = AI_{0r'} \qquad (0 \leqq r' \leqq r_-), \tag{64}$$

$$E_{z2} = BI_{0r'} + CK_{0r'} \qquad (r_+ \leqq r' \leqq b_-), \tag{65}$$

$$E_{z3} = DK_{0r'} \qquad (r' \geqq b_+). \tag{66}$$

The first expression holds inside of the injected charge ring; the second between that radius, r, and the beam boundary, b; and the last in free space outside of the beam.

The boundary equations at r and b, respectively, are:

$$AI_{0r} = BI_{0r} + CK_{0r}\,, \tag{67}$$

$$AI_{1r} - \frac{j\,dq_t}{2\pi r\epsilon} = BI_{1r} - CK_{1r}\,, \tag{68}$$

$$BI_{0b} + CK_{0b} = DK_{0b}, \tag{69}$$

$$(1 - R)[BI_{1b} - CK_{1b}] = DK_{1b}. \tag{70}$$

The total current within radius a, due to the injected charge ring at r, is

$$di_t(a) = \frac{j2\pi a\omega\epsilon}{\beta^2}\frac{\partial E_{z3}}{\partial r} = -\frac{j2\pi a\omega\epsilon}{\beta}DK_{1a}\,, \tag{71}$$

where

$$D = \frac{j\beta r I_{0r}\,dq_t}{2\pi\epsilon r\left[1 + \left(\dfrac{R}{1-R}\right)\beta b I_{0b}K_{1b}\right]}. \tag{72}$$

Thus, we obtain the transfer function $F(\gamma, \beta)$ relating the transform of the total induced current $d\hat{\imath}_t(\beta, a)$ to that of the injected current ring $d\hat{\imath}_n(\beta, r)$:

$$d\hat{\imath}_t(\beta, a) = \frac{\gamma a K_{1a}I_{0r}(\beta - \beta_e)^2\,d\hat{\imath}_n}{(\beta - \beta_1)(\beta - \beta_2)} = F(\gamma, \beta)\,d\hat{\imath}_n(\beta, r), \tag{73}$$

where

$$\beta_{1,2} = \beta_e \pm \beta_p(\beta b I_{1b}K_{0b})^{1/2} = \beta_e \pm \beta_q\,. \tag{74}$$

The inverse transform of $d\hat{\imath}_t(\beta, a)$, describing the total current in the propagating wave, is evaluated as before with the approximations

$$\gamma \cong \beta \cong \beta_e \tag{75}$$

in terms that are not sensitive to changes in β:

$$di_t(z, a) = |\,di_n\,|\int_{-\infty}^{\infty}\frac{F(\gamma, \beta)\,\exp\,(-j\beta z)\,d\beta}{\beta - \gamma}\,, \tag{76}$$

$$= |\,di_n\,|\sum_n\left[\frac{(\beta - \beta_n)F(\gamma, \beta)\,\exp\,(-j\beta z)}{\beta - \gamma}\right]_{\beta=\beta_n}, \tag{77}$$

$$\cong |\,di_n\,|\,(\beta_e a)K_{1a}I_{0r}\exp\,(-j\beta_e z)\left[\cos\,\beta_q z + j\epsilon\frac{\omega}{\omega_q}\sin\,\beta_q z\right]. \tag{78}$$

Following the same summation procedure as in the case of the single

injected noise filament, the total mean square current due to charge rings at r in all of the velocity classes is

$$|\, di_t^{\,2}(z, a)\,| = (2eJ_0\Delta f)(\beta_e a)^2 K_{1a}^{\,2} I_{0r}^{\,2} 2\pi r\delta r \cdot \left[\cos^2 \beta_q z + \bar{\epsilon}^2 \left(\frac{\omega}{\omega_q}\right)^2 \sin^2 \beta_q z\right], \tag{79}$$

where

$$J_0 = \sum_n J_n \tag{80}$$

is the total direct current density. The square of the total response current is found by integrating this quantity over the beam radius:

$$|\, i_t^{\,2}(z, a)\,| = (2eI_0\Delta f)(\beta_e a K_{1a})^2 (I_{0b}^{\,2} - I_{1b}^{\,2}) \cdot \left[\cos^2 \beta_q z + \bar{\epsilon}^2 \left(\frac{\omega}{\omega_q}\right)^2 \sin^2 \beta_q z\right], \tag{81}$$

where I_0 is the total direct current.

The mean square noise convection current in the drifting beam is consequently

$$|\, i_c^{\,2}(z)\,| = (2eI_0\Delta f)\left(\frac{I_{0b}^{\,2} - I_{1b}^{\,2}}{I_{0b}^{\,2}}\right)\left[\cos^2 \beta_q z + \bar{\epsilon}^2 \left(\frac{\omega}{\omega_q}\right)^2 \sin^2 \beta_q z\right] \tag{82}$$

The noise convection current at the maxima and minima of this standing wave are, respectively,

$$|\, i_c^{\,2}\,|_{\max} = 2eI_0\Delta f\left[1 - \left(\frac{I_{1b}}{I_{0b}}\right)^2\right], \tag{83}$$

$$|\, i_c^{\,2}\,|_{\min} = 2eI_0\Delta f\left[1 - \left(\frac{I_{1b}}{I_{0b}}\right)^2\right]\bar{\epsilon}^2 \left(\frac{\omega}{\omega_q}\right)^2. \tag{84}$$

The product of maximum and minimum rms amplitudes of the noise convection current can therefore be written in the form

$$\frac{|\, i_{\max} i_{\min}\,|_B}{2eI_0\Delta f} = \left(1 - \frac{I_{1b}^{\,2}}{I_{0b}^{\,2}}\right)\frac{\omega}{\omega_q}\,(\bar{\epsilon}^2)^{1/2}, \tag{85}$$

where the subscript B stands for the Brillouin-flow beam. If all of the electrons are accelerated by the same dc voltage V_0, such that $(eV_0/kT_c) \gg 1$, where T_c is the cathode temperature,

$$\bar{\epsilon}^2 = \tfrac{1}{2}(kT_c/eV_0), \tag{86}$$

and

$$\frac{|\, i_{\max} i_{\min}\,|_B}{2eI_0\Delta f} = \left(1 - \frac{I_{1b}^{\,2}}{I_{0b}^{\,2}}\right)\frac{\omega}{2\omega_q}\,\frac{kT_c}{eV_0}. \tag{87}$$

By comparison, the result of the same analysis applied to the one-dimensional beam,[10] which is identical with that obtained by the Rack-Llewellyn-Peterson method,[12] is

$$\frac{|\, i_{\max} i_{\min} \,|_T}{2eI_0\Delta f} = \frac{\omega}{2\omega_q}\frac{kT_c}{eV_0}, \tag{88}$$

the subscript T standing for the "thin" beam; or

$$\frac{|\, i_{\max} i_{\min} \,|_B}{|\, i_{\min} i_{\max} \,|_T} = 1 - \frac{I_{1b}^2}{I_{0b}^2} \tag{89}$$

if the two types of beam are compared on the basis of the same I_0, V_0, T_c and ω/ω_q. This ratio is less than unity for finite βb.

Although the "noisiness" $|\, i_{\max}\, i_{\min} \,|$ of a thin beam is a measure of the least attainable noise figure of any amplifier using that beam,[14, 15, 16, 17] it does not follow from this result that the Brillouin-flow beam is necessarily less noisy than a thin beam with the same direct current and voltage. For instance, in a thin beam the shot-noise current is $2eI_0\Delta f$ and all of I_0 is effective in interaction with the longitudinal RF field of an amplifier circuit. In the Brillouin-flow beam, however, the RF field has both longitudinal and transverse components, and varies in intensity over the beam cross section. The effective part of the total beam current, therefore, may be less than I_0.

In single-velocity thin-beam theory, the kinetic power P_k accounts for virtually all of the power transported by the space-charge waves, and may be defined by[15]

$$\mathrm{Re}(P_k) = \tfrac{1}{2}K(i_f^2 - i_s^2), \tag{90}$$

where i_f and i_s are the convection currents in the "fast" and "slow" traveling waves, respectively, and

$$K = 2\frac{\omega_q}{\omega}\frac{V_0}{I_0}. \tag{91}$$

In terms of K the noise-current expression for the thin beam may be rewritten as

$$P_s = \tfrac{1}{2}K\,|\, i_{\max} i_{\min} \,| = kT_c\Delta f. \tag{92}$$

This noise quantity has the dimensions of power; we may call it noisiness. It is invariant in all beam transformations not involving loss of RF power.[14] The minimum attainable noise figure F_T of any amplifier depending on RF interaction between a circuit and the slow space-charge wave has been shown to be[15, 16]

$$F_T = 1 + P_s/(kT\Delta f) = 1 + T_e/T, \tag{93}$$

where T is the ambient temperature. This summary of thin-beam theory applies to a thin hollow beam as well as a filamentary beam, as in both such beams the RF field acts equally on all of the direct current I_0.

From this it follows that the minimum noise figure F_B of any amplifier using the ideal Brillouin beam we have discussed can be evaluated by finding the noise kinetic power of an equivalent thin beam. Both beams will be equivalent with respect to interaction with any external RF circuit if both produce the same fields (or wave admittances) in free space just outside of the thick beam, at $r = b$.

Just outside of the Brillouin beam, with current I_0 and voltage V_0, the TM wave admittance looking into the beam is[5]

$$Y = \frac{H_\theta}{E_z} = \frac{j\omega\epsilon}{\beta}\left[1 - \frac{\omega_p^2}{(\omega - \beta u)^2}\right]\frac{I_{1b}}{I_{0b}}. \tag{94}$$

The portion Y_d of Y due to displacement current i_d in the volume occupied by the beam is given by the same expression, with $\omega_p^2 = 0$:

$$Y_d = \left(\frac{i_d}{2\pi b E_z}\right)_{r=b} = \frac{j\omega\epsilon}{\beta}\frac{I_{1b}}{I_{0b}}. \tag{95}$$

The remainder of the total admittance is due to the convection current i_c in the beam:

$$Y_e = \left(\frac{i_c}{2\pi b E_z}\right)_{r=b} = -\frac{j\omega\epsilon}{\beta}\frac{\omega_p^2}{(\omega - \beta u)^2}\frac{I_{1b}}{I_{0b}}. \tag{96}$$

The equivalent beam is chosen to be a thin hollow beam, of the same radius b as the Brillouin beam, with current i_0 not yet specified, and the same voltage V_0. We can take the ac convection current of the thin beam as equal to the total convection current of the Brillouin flow beam, because it can be shown that the total convection current of the Brillouin beam is equal to the surface current to within a small fraction ω_q/ω.

The relation between total convection current i_e and longitudinal field E_z in this hollow beam is

$$\left(\frac{i_e}{E_z}\right)_{r=b} = -\frac{j\beta_e}{(\beta_e - \beta)^2}\left(\frac{i_0}{2V_0}\right). \tag{97}$$

Its electronic admittance in space just outside of this beam is

$$Y_e = \left(\frac{H_\theta}{E_z}\right)_{r=b} = -\frac{j\beta_e}{2\pi b(\beta_e - \beta)^2}\left(\frac{i_0}{2V_0}\right). \tag{98}$$

Near $\beta_e u = \omega$, the admittance Y_d due to displacement current in the space inside of this cylinder is the same as for the Brillouin-flow beam:

$$Y_d = \frac{j\omega\epsilon}{\beta}\frac{I_{1b}}{I_{0b}}. \tag{99}$$

The two beams will then be equivalent if their electronic admittances are the same at $r = b$:

$$-\frac{j\beta_e}{2\pi b(\beta_e - \beta)^2}\left(\frac{i_0}{2V_0}\right) = -\frac{j\omega\epsilon}{\beta}\frac{\omega_p^2}{(\omega - \beta u)^2}\frac{I_{1b}}{I_{0b}}, \tag{100}$$

$$\frac{2V_0}{i_0} = \frac{\beta u}{2\pi b\epsilon\omega_p^2}\frac{I_{0b}}{I_{1b}}. \tag{101}$$

As this expression changes relatively slowly with βb, the admittances of the thin hollow beam and of the Brillouin beam vary in essentially the same way with β. This approximation, therefore, is fairly good over a small range of β about ω/u.

The noisiness of the equivalent hollow beam is

$$P_s = \tfrac{1}{2}K \mid i_{\max} i_{\min} \mid , \tag{92}$$

where

$$K = \frac{2\omega_q}{\omega}\frac{V_0}{i_0} = \frac{\omega_q}{\omega}\frac{\beta u}{2\pi b\epsilon\omega_p^2}\frac{I_{0b}}{I_{1b}} \tag{102}$$

and

$$\mid i_{\max} i_{\min} \mid = \left(1 - \frac{I_{1b}^2}{I_{0b}^2}\right)\frac{\omega I_0(kT_e\Delta f)}{\omega_q V_0}, \tag{103}$$

as found above for the thick beam. Since the direct current density and longitudinal velocity of this beam are constant over its cross section,

$$\frac{I_0}{2V_0} = \frac{e}{m}\frac{I_0}{u^2} = \frac{\pi b^2}{u}\omega_p^2\epsilon. \tag{104}$$

With these substitutions, the expression for noisiness P_s in the Brillouin-flow beam reduces to

$$P_s = (I_{0b}^2 - I_{1b}^2)\left(\frac{\beta b}{2I_{1b}I_{0b}}\right)kT_e\Delta f. \tag{105}$$

Another way to state this result is to express the minimum attainable noise figure F_B of the Brillouin-flow beam in terms of that of the thin beam (whose noisiness is $kT_e\Delta f$):

$$\frac{F_B - 1}{F_T - 1} = (I_{0b}^2 - I_{1b}^2)\left(\frac{\beta b}{2I_{1b}I_{0b}}\right). \tag{106}$$

This ratio, plotted in Fig. 1, varies rather slowly from unity at $\beta b = 0$, to one-half at $\beta b \to \infty$. With $F_T = 4$, corresponding to about 6 db, the predicted value[16] for a univelocity thin beam, the least noise figure of the infinitely broad beam, for example, would be 4 db.

We should, of course, recall that this result applies for the unusual but not unphysical case of a beam with no transverse velocities.

Haus[18] has demonstrated formally that an amplifier with a thick beam in confined flow cannot have a lower noise figure than one with a thin beam, when the input conditions are full shot-noise current and the Rack equivalent velocity fluctuations. His proof depends on expansion of the excitation in terms of a complete orthogonal set of functions at the input plane. In the absence of mode coupling in the acceleration region, each mode can be treated as though it were along a single thin beam, independent of the other modes. The opposite point of view has been advanced by Beam and Bloom[19] and by Paschke.[20] They have argued, essentially, that a lower noise figure can be obtained with a thicker beam (in confined flow), because the field of the RF circuit couples less effectively to the beam interior than to its surface, whereas

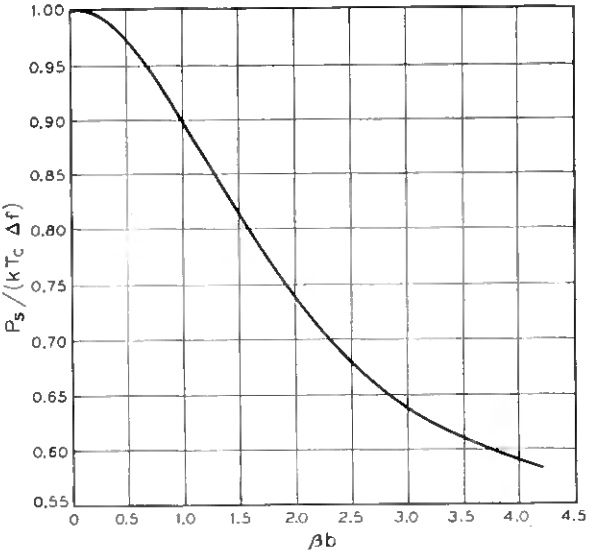


Fig. 1. — The ratio of the noisiness P_s of an idealized Brillouin-flow beam to that of an equivalent thin hollow beam in confined flow, as a function of the product of propagation constant β and beam radius b. The ordinates also represent the ratio $(F_B - 1)/(F_T - 1)$, where F_B and F_T are the minimum noise figures attainable with the two types of electron beam, respectively, when they are abruptly accelerated from temperature-limited cathodes [see (106)].

the noise excitation is uniform over the entire cross section. This argument, however, assumes that the circuit field in the presence of the beam is the same as in its absence—an assumption open to question.

In connection with the fact that we have found a noisiness less than that prescribed by Haus, we can only note that, for the beam with zero magnetic field and for the Brillouin flow beam, in whose interior the ac space-charge density is zero, the set of propagating space-charge modes is incomplete. (There are no slow space-charge modes with radial periodicity.) It may be that the axial-symmetric mode fails to propagate all of the axial-symmetric noise excitation and the higher-order modes fail to carry all of the excitation with angular periodicity. Recent calculations by Bobroff and Haus[21] point to the same conclusions—that the space-charge wave modes in such beams do not form a complete set, and therefore that an arbitrary initial excitation cannot be expanded in terms of these modes.

The noisiness of beams produced by shielded guns is actually much greater than that calculated for the idealized beam, because of the transverse thermal electron velocities near the cathode, neglected in the calculation. Their principal effect, as Beam has shown,[22] is to increase the velocity fluctuations near the potential minimum due to "mixing" of electrons from different parts of the cathode. The increase in noisiness due to this effect probably outweighs any possible decrease due to increase in beam diameter. However, the noise reduction mechanism described by the calculations may perhaps play a role in low-noise beams of a special type.

Noise figures considerably less than the 6-db minimum for an abruptly accelerated thin beam have been observed by a number of workers. Using a hollow confined-flow beam in a backward-wave amplifier, Currie and Forster[6] have measured a noise figure of less than 4 db. More recently, St. John and Caulton[7] have attained a noise figure of 4.5 db with a fairly conventional gun and, by using a solid-circular gun similar in cross section to that of Currie and Forster's annular gun, they attained a 3.5-db noise figure at microwave frequencies. Noise reduction due to a gradual acceleration allowing drifting[9] has been put forward as a plausible explanation of such low noise figures.

It should be noted, however, that in both instances the beams were found to have current density profiles sharply peaked at the surface, so as to resemble to some degree the case of Brillouin flow, in which the ac current is at the surface of the beam. Their low noisiness, therefore, might, at least in part, have been due to the noise-reduction mechanism described by the calculations of this paper.

V. SPACE-CHARGE WAVES INDEPENDENT OF BOUNDARY CONDITIONS

In analyses of slow-wave propagation along electron beams produced by magnetically shielded guns,[4, 5] two pairs of space-charge waves are found. In one of these, the field distributions and propagation constants depend in the usual way on the transverse boundary conditions. The waves of the second pair, however, are not accompanied by any field outside of the beam; they have never been detected experimentally and they are not found when magnetic flux, however slight, threads the cathode.[23, 24] These very singular waves appear to have first been described in 1946 by Feenberg and Feldman.[4]

For simplicity, the properties of such waves will be examined in the case of axial-symmetric fields in a Brillouin-flow beam.[5] At the surface of this beam, the boundary conditions are (i) that E_z be continuous, and (ii) that

$$[(1 - R)E_r]^{\text{beam}} = [E_r]^{\text{space}}, \tag{107}$$

where $R = \omega_p^2/\omega_q^2$ as defined in (11). For these waves, $R = 1$. It follows that the fields are zero outside of the beam, and E_z is zero at the common boundary. The waves, therefore, cannot be excited by fields outside of the beam.

Within the beam, if excited somehow, they would propagate with arbitrary radial field distribution and the longitudinal propagation constants

$$\beta_{1,2} = \beta_e \pm \beta_p , \tag{108}$$

which are characteristic of waves with *purely longitudinal fields*. (In ordinary space-charge waves, the plasma oscillation frequency is reduced, because of transverse fields coupling the current filaments to one another and to other currents.) However, if E_r were zero everywhere inside of the beam, E_z would also be zero, as it is zero at the boundary. This leads one to suspect that these waves do not really exist at all.

It was shown that, when a Brillouin-flow beam is current-modulated, the total charge density ρ_t at any point in the excitation plane is related to the injected charge density ρ_1, and to that induced in the smoothed-out beam, ρ, by the equations

$$\rho_t = \rho + \rho_1 = \epsilon \operatorname{div} \mathbf{E}, \tag{28}$$

$$\rho = R\epsilon \operatorname{div} \mathbf{E}. \tag{29}$$

When $R = 1$, therefore, the initial conditions are $\rho_t = \rho$ for all values of the injected charge ρ_1. This means that the $R = 1$ modes cannot be excited by charge modulation or, since the charge-injection velocity is

arbitrary, by either current or velocity modulation within the beam. As, in addition, they cannot be excited by external voltage modulation, the $R = 1$ modes are physically nonexistent.

REFERENCES

1. Scotto, M. and Parzen, P., Excitation of Space-Charge Waves in Drift Tubes, J. Appl. Phys, **27,** April 1956, p. 275.
2. Bresler, A. D., Joshi, G. H. and Marcuvitz, N., Orthogonality Properties for Modes in Passive and Active Uniform Waveguides, J. Appl. Phys., **29,** May 1958, p. 794.
3. Ramo, S., Space Charge and Field Waves in an Electron Beam, Phys. Rev., **56,** August 1, 1939, p. 276.
4. Feenberg, E. and Feldman, D., Theory of Small-Signal Bunching in a Parallel Electron Beam of Rectangular Cross Section, J. Appl. Phys., **17,** December 1946, p. 1025.
5. Rigrod, W. W. and Lewis, J. A., Wave Propagation Along a Magnetically Focused Cylindrical Electron Beam, B.S.T.J., **33,** March 1954, p. 399.
6. Currie, M. R. and Forster, D. C., Low-Noise Tunable Preamplifiers for Microwave Receivers, Proc. I.R.E., **46,** March 1958, p. 570; Currie, M. R., Letter to the Editor, Proc. I.R.E., **46,** May 1958, p. 911.
7. St. John, G. E. and Caulton, M., S-Band TWT with Noise Figure Below 4 db, Proc. I.R.E., **46,** May 1958, p. 911.
8. Siegman, A. E., Analysis of Multivelocity Electron Beams by the Density-Function Method, J. Appl. Phys., **28,** October 1957, p. 1132.
9. Siegman, A. E., Watkins, D. A. and Shieh, H., Density-Function Calculations of Noise Propagation on an Accelerated Multivelocity Electron Beam, J. Appl. Phys., **28,** October 1957, p. 1138.
10. Pierce, J. R., A New Method of Calculating Microwave Noise in Electron Streams, Proc. I.R.E., **40,** December 1952, p. 1675.
11. Beck, A. H. W., High-Order Space-Charge Waves in Klystrons, J. of Elect., **2,** March 1957, p. 489.
12. Pierce, J. R., *Traveling-Wave Tubes*, D. Van Nostrand Co., New York, 1950, Ch. 10.
13. Landau, L., On the Vibrations of the Electronic Plasma, J. Phys. (U.S.S.R.), **10,** 1946, p. 25.
14. Pierce, J. R., A Theorem Concerning Noise in Electron Streams, J. Appl. Phys., **25,** August 1954, p. 931.
15. Haus, H. A. and Robinson, F. N. H., The Minimum Noise Figure of Microwave Beam Amplifiers, Proc. I.R.E., **43,** August 1955, p. 981.
16. Pierce, J. R. and Danielson, W. E., Minimum Noise Figure of Traveling-Wave Tubes with Uniform Helices, J. Appl. Phys., **25,** September 1954, p. 1163.
17. Bloom, S. and Peter, R. W., A Minimum Noise Figure for the Traveling-Wave Tube, R.C.A. Rev., **15,** June 1954, p. 252.
18. Haus, H. A., Limiting Noise Figure of a Microwave Tube with a Beam of Finite Diameter, Quart. Prog. Rep., M. I. T. Res. Lab. of Elect., **32,** January 15, 1957.
19. Beam, W. R. and Bloom, S., Minimum Noise Figure of Traveling-Wave Tubes, Including Higher Space-Charge Wave Modes, I.R.E.-A.I.E.E. Electron Tube Research Conf., Boulder, Colo., June 1956.
20. Paschke, F., Die Wechselseitigkeit der Kopplung in Wanderfeldröhren, Arch. Elek.Über., **11,** April 1957, p. 137.
21. Bobroff, D. L. and Haus, H. A., Uniqueness and Orthogonality of Small Signal Solutions in Electron Beams, to be published.
22. Beam, W. R., Noise Wave Excitation at the Cathode of a Microwave Beam Amplifer, Trans. I.R.E., **ED-4,** July 1957, p. 226.
23. Brewer, G. R., Some Effects of Magnetic Field Strength on Space-Charge Wave Propagation, Proc. I.R.E., **44,** July 1956, p. 896.
24. Rigrod, W. W., this issue, pp. 119–139.

Space-Charge Wave Harmonics and Noise Propagation in Rotating Electron Beams

By W. W. RIGROD

(Manuscript received May 2, 1958)

Higher-order space-charge waves on solid cylindrical electron beams produced by shielded or nearly shielded guns have only azimuthal periodicity, as in hollow beams. Because of beam rotation, they are members of a broad class of space-charge waves which can travel faster than the beams themselves, either forwards or backwards. The properties of such waves for the beam in a drift tube and in a concentric sheath helix are derived from a slow-wave, small-signal analysis and the appropriate boundary equations. Experimental observations of their interaction with harmonic fields of a helix, as well as of their role in noise propagation, tend to confirm the results of these computations.

I. INTRODUCTION

Interest in the ac behavior of cylindrical electron beams issuing from magnetically shielded or partly shielded guns has been stimulated in recent years by their increasing application in medium- and high-power traveling-wave tubes. As yet, however, such beams have received considerably less attention in the literature than have those in confined flow. The properties of the fundamental (axial-symmetric) space-charge mode in the former type of beam have been studied by Rigrod and Lewis,[1] and by Brewer.[2] Waves of this type provide a first-order description of the beam interaction with its environment, such as a drift tube or helix. The present paper will supplement this work by considering higher-order modes of wave propagation in such beams, in which the fields have azimuthal, but not radial, periodicity. Following an analysis of the waves themselves, several problems will be discussed in which they play important roles: the excitation in a helix of spatial-harmonic modes, the propagation of noise excitation and possible new

applications of these space-charge-wave "harmonics." Experimental confirmation of their interaction with the harmonic fields of a helix, observed by Kiryushin[3,4] will be described, as well as some interesting noise measurements obtained by Ashkin and White,[5] which illustrate their participation in noise propagation.

II. NATURE OF HIGHER-ORDER MODES

The formation of ripple-free beams from convergent electron guns is often facilitated by letting some magnetic flux thread the cathode. Although this paper is primarily concerned with waves along Brillouin-flow beams, the computations of this section will include provision for arbitrary flux density at the cathode, for greater generality.* The ratio α of flux encircled at the cathode to that in the drift region, is assumed constant for any ring of electrons. The steady-state electron flow is then laminar, and can be described by the following equations:

$$\dot{\theta} = \frac{\omega_c}{2}(1 - \alpha), \tag{1}$$

$$\eta \frac{\partial V_0}{\partial r} = r\dot{\theta}(\omega_c - \dot{\theta}), \tag{2}$$

$$\omega_p^{\,2} = \omega_c^{\,2} \frac{(1 - \alpha^2)}{2}, \tag{3}$$

$$\dot{z} = u, \quad \dot{r} = 0. \tag{4}$$

Here (r, θ, z) are polar cylindrical coordinates; V_0 the dc potential due to the uniform space-charge density ρ_0 ; η the charge-mass ratio for the electron (a positive quantity) and ω_c and ω_p the angular cyclotron and plasma frequencies, respectively. A dot indicates time differentiation, and MKS units are used.

The problem is to find the properties of small-signal ac waves which propagate along the beam as

$$\exp j(\omega t - n\theta - \beta z), \tag{5}$$

with $n = 0, 1, 2, \cdots$, subject to the slow-wave condition

$$\frac{\omega^2 \mu\epsilon}{\beta^2} = \frac{k^2}{\beta^2} \cong 0. \tag{6}$$

With this condition, the scalar wave equation

* The basic equations of this section were first derived by J. R. Pierce of Bell Telephone Laboratories.

$$(\Delta + k^2)E_z = \frac{\beta}{\omega\epsilon}\,\mathrm{div}\,\mathbf{J} + j\omega\mu J_z \tag{7}$$

reduces to the following equivalent forms

$$\Delta E_z = \frac{\beta}{\omega\epsilon}\,\mathrm{div}\,\mathbf{J}, \tag{8}$$

$$\Delta E_z = -\frac{j\beta\rho}{\epsilon}, \tag{9}$$

using the charge-conservation equation. Here Δ is the Laplacian operator, $\mathbf{J}$ the ac convection current density, ρ the ac space-charge density, ϵ and μ the dielectric constant and permeability of free space, respectively, ω the angular excitation frequency, k the free-space wave number and β the axial propagation constant.

The terms which drop out of this wave equation due to the slow-wave assumption are precisely those arising from curl $\mathbf{E}$. That is, the slow-wave condition is equivalent to setting curl $\mathbf{E}$ to zero, or to neglecting the contribution to $\mathbf{E}$ made by the ac magnetic fields (provided $\mathbf{J}$ does not exceed $j\omega\epsilon\mathbf{E}$ by a factor approaching β^2/k^2 in magnitude). The electric field can therefore be derived from a scalar potential, or

$$E_r = \frac{j}{\beta}\frac{\partial E_z}{\partial r}, \qquad E_\theta = \frac{n}{\beta r}E_z. \tag{10}$$

Another consequence of the slow-wave restriction is that the contribution of the ac magnetic field to the force on electrons can be disregarded, as it is negligible compared with that exerted by the electric field.

With this and the assumption of single-valued velocities at each point in the beam, the electron dynamics equation can be expressed in Eulerian coordinates as follows:

$$\frac{d}{dt}(\mathbf{v}_0 + \mathbf{v}) = -\eta[-\mathrm{grad}\ V_0 + \mathbf{E} + (\mathbf{v}_0 + \mathbf{v}) \times \mathbf{B}_0], \tag{11}$$

where

$$\mathbf{v}_0 = (0, r\dot{\theta}, u), \tag{12}$$

$$\mathbf{v} = (v_r, v_\theta, v_z)\exp j(\omega t - n\theta - \beta z), \tag{13}$$

and $\mathbf{B}_0$ is the axial magnetic field, the zero subscript being used wherever necessary to distinguish the steady-state quantities. Expansion of this equation yields the components of the ac velocity amplitude:

$$v_r = \frac{j\eta P}{\omega_n^2}\left[E_r + j\left(\frac{\alpha\omega_c}{P}\right)E_\theta\right], \tag{14}$$

$$v_\theta = \frac{j\eta P}{\omega_n^2}\left[E_\theta - j\left(\frac{\alpha\omega_c}{P}\right)E_r\right], \tag{15}$$

$$v_z = \frac{j\eta E_z}{P}, \tag{16}$$

where

$$P = \omega - n\dot{\theta} - \beta u, \tag{17}$$

and

$$\omega_n^2 = P^2 - (\alpha\omega_c)^2. \tag{18}$$

From the charge-conservation equation,

$$\rho = \frac{j\rho_0 \operatorname{div} \mathbf{v}}{P} = \frac{j\epsilon\omega_p^2}{\omega_n^2}\left[\Delta E_z + \left(\frac{\alpha\omega_c}{P}\right)^2 \beta^2 E_z\right], \tag{19}$$

and the wave equation for E_z reduces to the Bessel equation

$$\left(1 - \frac{\omega_p^2}{\omega_n^2}\right)\Delta E_z - \frac{\omega_p^2}{\omega_n^2}\left(\frac{\alpha\omega_c}{P}\right)^2 \beta^2 E_z = 0, \tag{20}$$

whose solution has the form

$$E_z = \sum_n A_n I_n(\gamma_n r) \exp j(\omega t - n\theta - \beta_n z). \tag{21}$$

Here A_n is a constant, and I_n the nth order modified Bessel function of the first kind, with transverse propagation constant γ_n defined by

$$\frac{\gamma_n^2}{\beta_n^2} = 1 + \left(\frac{\alpha\omega_c}{P}\right)^2 \left(\frac{\omega_p^2}{\omega_n^2 - \omega_p^2}\right). \tag{22}$$

The ac space-charge density can conveniently be re-expressed in terms of the above ratio (for any chosen n):

$$\rho = \frac{j\epsilon\Delta E_z}{\beta} = j\epsilon\beta\left(\frac{\gamma^2}{\beta^2} - 1\right)E_z, \tag{23}$$

showing that ρ becomes zero when $\alpha\omega_c$ is zero.

Since, for slow waves, the electric field is irrotational both inside and outside of the beam, it can be determined by the boundary conditions for E_z and E_r at the beam surface:

$$\left(\frac{E_r + \tilde{\sigma}/\epsilon}{E_z}\right)_{r=b-} = \left(\frac{E_r}{E_z}\right)_{r=b+}, \tag{24}$$

where $b-$ and $b+$ refer to the regions just inside and outside of the beam surface, respectively, and $\tilde{\sigma}$ is the ac surface-charge density due to unbalanced radial electron motions:

$$\tilde{\sigma} = -\left(\frac{j\rho_0 v_r}{P}\right)_{r=b-} = -\frac{j\epsilon\omega_p^2}{\beta\omega_n^2}\left[\frac{\partial E_z}{\partial r} + \left(\frac{\alpha\omega_c}{P}\right)\frac{n}{r}E_z\right]_{r=b-}. \tag{25}$$

The simplest boundary-value problem is that of the beam drifting in a concentric conducting tube of radius a. In the space between beam and tube wall, the field is of the form

$$E_z = B_n[I_n(\beta r)K_n(\beta a) - K_n(\beta r)I_n(\beta a)], \tag{26}$$

where B_n is a constant, and K_n the nth order modified Bessel function of the second kind. Thus, the boundary equation at the beam surface, $r = b$, can be written

$$\left[\left(1 - \frac{\omega_p^2}{\omega_n^2}\right)\frac{\gamma b I_n'(\gamma b)}{I_n(\gamma b)} - \frac{\omega_p^2}{\omega_n^2}\frac{n\alpha\omega_c}{P}\right] = \beta b\left[\frac{I_n'(\beta b)K_n(\beta a) - K_n'(\beta b)I_n(\beta a)}{I_n(\beta b)K_n(\beta a) - K_n(\beta b)I_n(\beta a)}\right], \tag{27}$$

the primes denoting differentiation with respect to the total argument. For any set of values of n, α and b/a, this equation can be solved for the square of the plasma-frequency reduction factor $p_n = P/\omega_p$. For each frequency, there are two values of the propagation constant:

$$\beta_{1,2} = \beta_e - n\dot{\theta}/u \pm p_n\beta_p, \tag{28}$$

where $\beta_e = \omega/u$, $\beta_p = \omega_p/u$, and p_n is a function of βb. The two traveling waves in each such solution interfere with one another to form a standing wave, with half-wavelength

$$\frac{\lambda_s}{2} = \frac{2\pi}{\beta_1 - \beta_2} \cong \frac{\pi}{p_n\beta_p}. \tag{29}$$

Brewer[2] has solved this admittance equation (27) for the fundamental mode, $n = 0$, using a flux parameter Ω related to α by

$$\left(\frac{\Omega}{\omega_p}\right)^2 = \frac{\alpha^2}{2(1 - \alpha^2)}. \tag{30}$$

His results show that, for α below about 0.5, the solution p_0 differs little from its value for $\alpha = 0$, the rate of change $dp_0/d\alpha$ being less as βb and b/a decrease.

The influence of cathode flux on the reduction factor p_n for the higher-order modes is quite different from that for the fundamental. This is

illustrated in Fig. 1, showing how this factor varies with βb for the $n = 1$ mode, for $\alpha = 0.2$ and 0.4, and $b/a = 0$ and 0.6. For $0 < \alpha < 1$, and small b/a, p tends to increase as βb decreases, reaching some finite value at the limit $\beta b = 0$. Calculations indicate, moreover, that p becomes infinite for $\alpha = 1$ (confined flow), for all values of βb. (In confined flow, there is an infinite set of solutions for p, but the one described here is that which blends continuously into that for Brillouin flow as α is varied from unity to zero.) In general, $dp/d\alpha$ decreases as βb increases or α decreases.

In most cases when beams are produced by shielded or nearly shielded diodes, the flux parameter α ranges from zero to at most about 0.4. Except for very small βb and b/a, the reduction factor for $\alpha = 0.2$ differs negligibly from that for $\alpha = 0$, and for $\alpha = 0.4$ it ranges mostly between 0.85 and unity. Over this range of α, then, it would appear that the

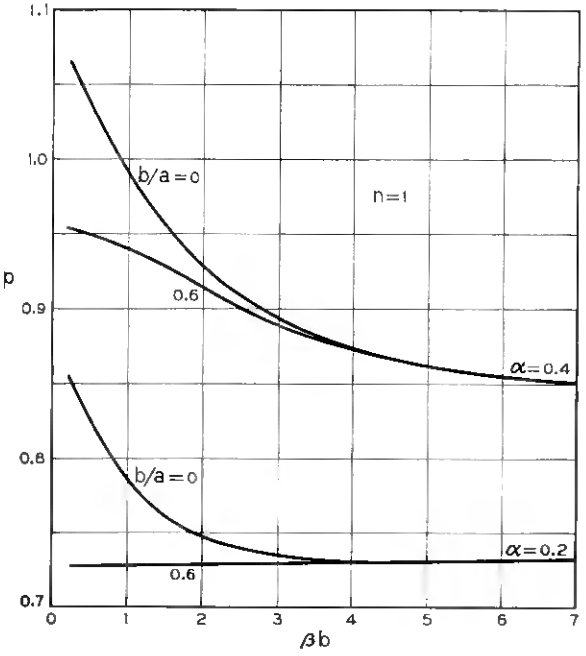


Fig. 1 — Plasma-frequency reduction factors $p = \omega_q/\omega_p$ for space-charge waves with azimuthal periodicity, $n = 1$, along a solid-cylindrical beam with small amounts of flux threading the cathode; α is the ratio of flux at the cathode to that flooding the beam, β is the axial propagation constant, b the beam radius and a the radius of a concentric drift tube.

properties of higher-order space-charge waves do not differ markedly from those on a Brillouin-flow beam (Figs. 2 and 3).

To obtain the equations for Brillouin flow, it is only necessary to set α equal to zero. It should be noted that, in all of the functional relations among ac quantities, the flux parameter α appears explicitly only in the *product* $\alpha\omega_c$, proportional to the flux density at the cathode. Because of this, all the equations determining the current and field patterns of the higher-order modes, including the TM boundary-matching equation, are the same for the Brillouin-flow beam and the beam in zero magnetic field, both of which have zero flux at the cathode. The only wave properties affected by the rotation of the Brillouin-flow beam are the θ-directed surface current (which excites TE fields), and the axial phase velocity of higher-order modes.

The reduced space-charge wavelength, however, is the same in both types of beams. In a shielded diode with small convergence angle, therefore, the accelerated beam throughout the univelocity region can be regarded approximately as a chain of short sections of drifting beams, each with its own velocity and geometry, in which the allowed mode patterns are the same as in Brillouin flow. When the beam enters the magnetic focusing field, these patterns rotate with the rotating beam, each thereby acquiring a higher axial phase velocity.

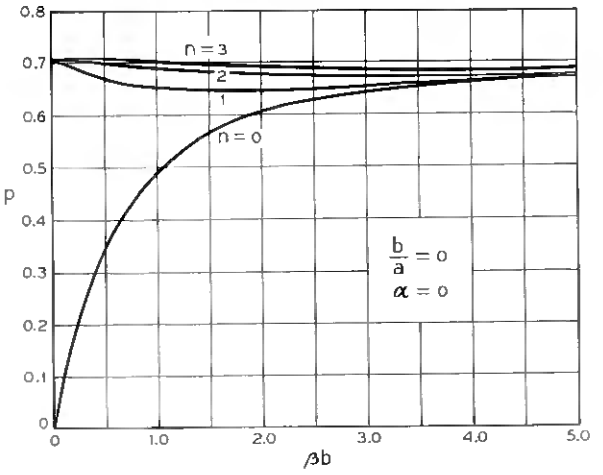


Fig. 2. — Plasma-frequency reduction factors $p = \omega_q/\omega_p$ for the fundamental ($n = 0$) and the first three higher-order modes of space-charge waves, along a solid-cylindrical Brillouin-flow beam ($\alpha = 0$), of radius b in free space.

When $\alpha\omega_c$ is zero, the wave equation for E_z has two sets of solutions. The first is

$$\omega_p^2 = P^2 \quad \text{or} \quad p_n = 1. \tag{31}$$

It has been shown in the accompanying article[6] that this solution is spurious, as there is no way in which the corresponding waves can be excited. The second solution is

$$\Delta E_z = 0. \tag{32}$$

The transverse propagation constant is now $\gamma = \beta$, and the ac space-charge density in the beam is zero. The boundary equation can be reduced to an explicit expression for the space-charge reduction factor:

$$p_n^2 = \beta b I_{nb}' I_{nb} \left[\frac{K_{nb}}{I_{nb}} - \frac{K_{na}}{I_{na}} \right]. \tag{33}$$

Here, and wherever else they are unambiguous, the arguments (βa) and (βb) are replaced by the subscripts a and b, the radii of drift tube and beam, respectively.

For very small arguments, $\beta a < n^2$, and $n > 0$,

$$p_n^2 \cong \tfrac{1}{2} - \tfrac{1}{2}(b/a)^{2n}, \tag{34}$$

whereas, for very large arguments, $\beta b > n^2$, and $n \geqq 0$,

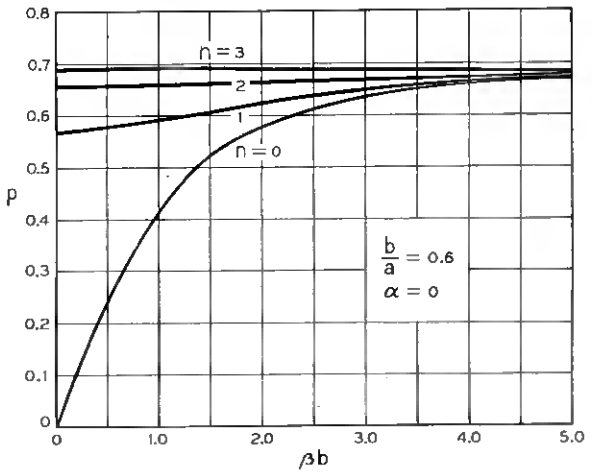


Fig. 3. — Plasma-frequency reduction factors $p = \omega_q/\omega_p$ for the same modes as in Fig. 2, when the Brillouin-flow beam is in a concentric drift tube of radius a, where $b/a = 0.6$.

$$p_n^2 \cong \tfrac{1}{2} - \tfrac{1}{2} \exp\left[-2\beta(a - b)\right]. \tag{35}$$

These limiting expressions, and the curves of p_n plotted in Figs. 2 and 3 for $n = 1$, 2 and 3, for two values of b/a, show that the smaller b/a is and the larger n is the closer the reduction factor p_n clings to the asymptotic value 0.707. To a good approximation, then, for all $n > 1$, and for $n = 1$ when b/a is small,

$$\beta_n \cong \beta_e - \tfrac{1}{2}(n \mp 1)\beta_c \tag{36}$$

and

$$\frac{\lambda_s}{2} \cong \frac{2\pi}{\beta_c} = \lambda_c\,; \tag{37}$$

that is, the distance between current minima equals the cyclotron wavelength.

Since the convection current carried by the beam is chiefly due to transport of ac surface charge, the field pattern can best be visualized by examining the locus of maximum surface current density in each standing wave:

$$\mathrm{Re}(G_s + G_f) = |\,G\,| \cos(\omega t - \beta_e z) \cos n\left(\theta - \frac{\dot{\theta} z}{u}\right) \cos \beta_q z. \tag{38}$$

The subscripts s and f refer to the slow and fast waves, respectively, and $|\,G\,|$ is the amplitude of surface current density at $t = z = \theta = 0$. For most of the high-order modes, for which $p_n \cong 0.707$,

$$\beta_q z \cong \frac{1}{2}\beta_c z = \frac{\dot{\theta} z}{u}\,. \tag{39}$$

The surface-current maxima follow spiral loci, therefore, with the same "pitch" as the rotating beam itself, increasing and decreasing along these loci with a period equal to one beam rotation. For the nth order mode, there are $2n$ such loci, resembling the conductors of a multifilar helix, in which the lines of force start and end on ac charges in adjacent parts of the beam. This is why the reduction factor tends to be independent of beam or wall geometry when n is large.

III. MODE COUPLING BETWEEN BEAM AND HELIX

The coupling impedance, measuring the interaction between waves on a Brillouin-flow beam and a concentric sheath helix, both with the same azimuthal periodicity, will be evaluated in this section, with some simplifying assumptions. Somewhat weaker interaction should also be

possible when the beam and helix waves have the same axial phase velocities, but different azimuthal periodicities. However, this problem will not be treated here.

The four boundary-matching equations at a sheath helix[1, 7] involve E_z, E_θ, H_z and H_θ, as well as the helix pitch angle Ψ. In addition to matching E_z and its radial derivative at the beam boundary, therefore, it would seem necessary to introduce another two equations in terms of H_z and its radial derivative at the beam boundary:

$$(H_z + G_\theta)_{b-} = (H_z)_{b+}, \tag{40}$$

$$\left(\frac{\partial H_z}{\partial r} + J_\theta\right)_{b-} = \left(\frac{\partial H_z}{\partial r}\right)_{b+}. \tag{41}$$

From the results of the previous section, it is readily found that

$$\Delta H_z = -\text{curl}_z \mathbf{J} = 0. \tag{42}$$

Thus the radial propagation constant for H_z is β, inside as well as outside of the beam.

The eliminant of the eight boundary-matching equations can be combined in the form of a single wave-admittance equation, similar in form to that for the beam in a drift tube:

$$\left(1 - \frac{\omega_p^2}{P^2}\right)\frac{I_{nb}'}{I_{nb}} = \frac{I_{nb}' + \delta K_{nb}'}{I_{nb} + \delta K_{nb}}, \tag{43}$$

where δ stands for an expression which depends on the helix geometry and the amplitude of H_z. When the slow-wave assumption is invoked, however, it turns out that this term differs negligibly from its value when the beam is absent; i.e., the TE-TM wave coupling at the helix is negligibly small:

$$\delta \cong \delta_0 = -\frac{1}{K_{na}^2}\left[I_{na}K_{na} + \left(\frac{ka\cot\Psi}{\beta a + n\cot\Psi}\right)^2 I_{na}'K_{na}'\right]. \tag{44}$$

The TE fields excited by J_θ and G_θ in the beam, therefore, do not affect the TM wave admittance presented to the beam by the helix; the expression on the right-hand side of (43) would be the same whether or not the beam rotated. In addition, due to the absence of ac space charge inside of the beam, the field there has the same radial variation as it would have in free space. These two circumstances suggest the possibility of evaluating the normal-mode parameters[7] of the Brillouin-flow beam in terms of an "equivalent" thin hollow beam in confined flow, by the same method employed earlier for the axial-symmetric mode.[1]

The procedure consists of reformulating (43) by replacing the electronic admittance of the Brillouin beam with that of a thin hollow confined-flow beam of the same radius b, direct current I_0 and longitudinal velocity u. (It is not convenient to compare these beams on the basis of the same dc voltage.) The altered circuit admittance Y_B, consistent with this new electronic admittance and (43), is compared with the actual circuit admittance of a thin hollow beam, in a narrow range of propagation constants near that of the empty helix. The normal-mode parameters of each beam, which can then be compared, indicate how the different distributions of electron flow in the two beams affect their interaction with the helix field near synchronism. (The compared beams are "equivalent" only in their common dc properties, not their RF behavior.)

The beam admittance on the left-hand side of (43) has two components, one due to the displacement current and another due to the electron current, within radius b. The latter portion, the electronic beam admittance, is Y_e in the following restatement of (43):

$$Y_e = Y_c \tag{45}$$

$$-\frac{\omega_p^2}{(\omega - \beta u - n\dot{\theta})^2}\frac{I_{nb}'}{I_{nb}} = \frac{I_{nb}' + \delta_0 K_{nb}'}{I_{nb} + \delta_0 K_{nb}} - \frac{I_{nb}'}{I_{nb}}. \tag{46}$$

The expression on the right side, Y_c, is the net circuit admittance due to displacement current both inside and outside of the Brillouin-flow beam.

The boundary equation for a thin hollow beam (thickness dr) at its outer radius b is obtained by matching the free-space values of E_z inside and outside of the beam, and equating the change in H_θ, between these surfaces, to $J_z dr$. (Inside of this beam, the field E_z is taken to be the same as in free space.) With $b-$ to identify the fields just inside of this beam, and $b+$ the fields outside of it, this boundary equation is

$$\frac{J_z\, dr}{E_z} = \left(\frac{H_\theta}{E_z}\right)_{b+} - \left(\frac{H_\theta}{E_z}\right)_{b-}. \tag{47}$$

For confined flow and fields with azimuthal periodicity n, this reduces to

$$-\frac{\omega_{pH}^2 \beta dr}{(\omega - \beta u)^2} = \frac{I_{nb}' + \delta_0 K_{nb}'}{I_{nb} + \delta_0 K_{nb}} - \frac{I_{nb}'}{I_{nb}} = Y_c, \tag{48}$$

where ω_{pH} is the angular plasma frequency in the hollow beam. The expression on the left-hand side is the electronic admittance of the thin hollow beam, and that on the right is the net circuit admittance due to the helix, the same as in (46) for the solid Brillouin beam.

If the hollow beam is assigned the same direct current and longitudinal velocity as the Brillouin-flow beam, the plasma frequencies in the two beams are related by

$$\frac{\omega_{pH}^2}{\omega_{pB}^2} = \frac{\pi b^2}{2\pi b\, dr} = \frac{b}{2\, dr}. \tag{49}$$

Then the admittance equation (46) for the Brillouin-flow beam can be rewritten as though it were a thin hollow beam whose circuit admittance was

$$Y_B = \left[\frac{\beta b}{2}\left(\frac{\omega - \beta u - n\dot{\theta}}{\omega - \beta u}\right)^2 \frac{I_{nb}}{I_{nb}'}\right] Y_c. \tag{50}$$

The solid-cylindrical Brillouin beam is thus equivalent to a confined-flow hollow beam whose helix admittance is Y_B. The normal-mode parameters of this and a true hollow beam depend on the behavior of Y_B and Y_c in the neighborhood of the synchronous phase velocity, i.e., of their common zero and pole. As the latter are very close together, the admittance functions can be represented in this region by a Weierstrass (algebraic) approximation. Then, just as in the case of the axial-symmetric mode,[1,8] the two types of beam have the same space-charge parameter:

$$Q_B = Q_H, \tag{51}$$

where B and H stand for the Brillouin-flow and hollow beam, respectively. Their impedance parameters, identified similarly, are related as follows:

$$\frac{K_B}{K_H} = \left(\frac{Y_c}{Y_B}\right)_{\beta_{0n}} = \left[\frac{2}{\beta b}\left(\frac{\omega - \beta u}{\omega - \beta u - n\dot{\theta}}\right)^2 \frac{I_{nb}'}{I_{nb}}\right]_{\beta_{0n}}, \tag{52}$$

where β_{0n} is the zero of Y_c, i.e., the empty-helix propagation constant. It is found by putting δ_0 to zero:

$$\frac{I_{na}'K_{na}'}{I_{na}K_{na}} = -\left[\frac{\beta^2 a^2 + n\beta a \cot \Psi}{ka\beta a \cot \Psi}\right]^2. \tag{53}$$

This is the determinantal equation of the empty sheath helix given by Sensiper,[9] modified by the slow-wave approximation. Sensiper has shown that, for $\beta a > 2$, and in the nondispersive region of the helix, the cold-helix propagation constant is given to a good approximation by

$$\beta_{0n} \cong \beta_0 + \frac{2\pi n}{p}, \tag{54}$$

where β_0 is the propagation constant of the fundamental and p is the helix pitch. Sensiper has also evaluated the impedance of the sheath helix in the nth order mode, at a radius $b < a$, as follows:

$$K_H(n, b) \cong \frac{30}{| n + ka |} \frac{I_{nb}^2}{I_{na}^2} \text{ ohms} \tag{55}$$

when $(k/\beta) < 0.4$ and $| n | > 0$. This is the same as K_H in (52) for the thin hollow beam in confined flow at radius b.

In expression (52) for K_B/K_H there is a factor dependent on the beam velocity components $\dot{\theta}$ and u, arising from the comparison of a rotating with a nonrotating beam. When both beams have the same value of βu, this factor is greater than unity for positive n, indicating that the angular component of beam motion contributes to field-wave interaction in the rotating beam, i.e., to interaction with E_θ as well as E_z and E_r. The remaining terms express, on the other hand, the superior efficiency of the hollow beam due to its concentration in a region of nonzero field.

It will be shown below that the space-charge reduction factor for the $n = 0$ mode is very nearly the same in the presence of a sheath helix as that of a drift tube at the same radius, when the slow wave is in synchronism with the cold-helix propagation constant. Without proof, it seems reasonable to assume that the same equivalence is true for a high-order mode as well. With this assumption, and using the approximation $p_n \cong 0.707$, the impedance ratio can be simplified further as follows:

$$\beta_s \cong \beta_e - \left(\frac{n-1}{2}\right)\beta_c = \beta_{0n}\,, \tag{56}$$

$$\frac{K_B}{K_H} \cong (n-1)^2 \left[\frac{2}{\beta b}\frac{I_{nb}'}{I_{nb}}\right]_{\beta_{0n}}. \tag{57}$$

This, combined with expression (55) for K_H, yields K_B for the comparable Brillouin beam at synchronism.

The sign of n is positive for a wave which spirals in the same sense as the beam, since both n and $\dot{\theta}$ are referred to the same set of cylindrical coordinates. Thus, β_n is less than β_0 when $n\dot{\theta}$ is positive. For the spatial harmonics of an empty helix, the opposite convention has been established; i.e., $\beta_n < \beta_0$ when n and the pitch p have opposite signs. Aside from this distinction, however, there is a close analogy between the spatial harmonic waves on a helix and those on the Brillouin-flow or hollow rotating beam.

The above expression for K_B/K_H is only valid for $n = 0$, and for $n > 1$ whenever the approximation $p_n \cong 0.707$ is valid. The coupling impedance K_B, therefore, is not necessarily zero when $n = 1$. For negative n, the phase constant β is greater than for zero or positive n, and K_B decreases very rapidly as β increases. (This might be expected, as the larger β is, the more rapidly the field decays radially away from the helix.) Thus, none of the negative-order beam harmonics have appreciable coupling impedances, but those of positive order greater than unity may have very large coupling impedances.

Evidence of interaction between a number of these beam hamonics and those of a bifilar helix has been reported by V. P. Kiryushin.[3, 4] Operating a backward-wave oscillator with its electron gun in a field-free region, he found narrow-band gaps in the output spectrum of the tube (and corresponding peaks in the starting current) at a number of discrete values of ω_c/ω, which he attributed to loss of energy to various harmonic modes. The values of ω_c/ω at which these disturbances were noted were found by Kiryushin to correspond to the "ratios of small integers", and appear to show interaction between beam and helix modes of the same as well as of different azimuthal periodicities (when their axial phase velocities are the same). In the latter case, it seems likely that the coupling impedance K_B would lack the factor $(n - 1)^2$ expressing interaction with the azimuthal electric field.

IV. PLASMA FREQUENCY REDUCTION FACTOR FOR BEAM IN SHEATH HELIX

For axial-symmetric waves on a solid-cylindrical beam in confined flow, Branch[10] has found the space-charge reduction factor to be nearly the same in a drift tube as in a helix of the same diameter. A similar computation can be made for the Brillouin-flow beam.

For any beam in a concentric helix, the relation derived by Branch is

$$p^2 = (QK)\,\frac{(\beta b)^2 V_0^{1/2}}{(174.1)^2}\,, \tag{58}$$

which reduces, when the beam is at synchronous velocity, to

$$p^2 = (QK)\,\frac{k}{60\beta_0}\,. \tag{59}$$

Here Q and K are Pierce's[7] normal-mode parameters, properly evaluated for the finite-diameter beam in question, and V_0 is the dc beam potential. The Q and K values for the Brillouin beam will be identified as before by the subscript B, and those for the thin hollow beam at the bounding radius b by the subscript H.

As shown in the preceding section of this paper,

$$Q_B = Q_H \tag{60}$$

for space-charge modes of any order number, including zero. Fletcher[8] has shown, in curves reproduced in Fig. A6.1 of Ref. 7, that Q_H differs very little from its value in a drift tube of the same diameter ($2a$) as the helix

$$Q_H = \frac{60}{F^3(\beta a)}\left[\frac{K_{0b}}{I_{0b}} - \frac{K_{0a}}{I_{0a}}\right], \tag{61}$$

where $F^3(\beta a)$ is given by Equation (41), p. 232, of Ref. 7. The impedance parameter K_B of the Brillouin-flow beam is related to that of the thin beam at the axis, K_T, as follows:[1]

$$\left(\frac{K_B}{K_T}\right)_{n=0} = \left[\frac{2I_{1b}I_{0b}}{\beta b}\right]_{\beta_0}, \tag{62}$$

where

$$K_T \cong \frac{\beta}{2k} F^3(\beta a). \tag{63}$$

Thus, when the beam is at synchronous velocity,

$$p^2 = \beta b I_{1b} I_{0b}\left[\frac{K_{0b}}{I_{0b}} - \frac{K_{0a}}{I_{0a}}\right], \tag{64}$$

an expression identical with that for p^2 when the beam is in a drift tube of diameter $2a$.[1]

Paschke[11] has questioned the results of Branch's computations for the solid-cylindrical confined-flow beam, on the grounds that QK was computed in terms of an equivalent hollow beam—an equivalence of rather restricted validity. This objection does not apply to the present computation. Since its ac convection current is almost entirely carried by the moving surface charge, the Brillouin-flow beam very closely resembles the thin hollow beam on which the calculation is based.

V. INTERCEPTION NOISE DUE TO IMMERSED GRID

Ashkin and White[5] have obtained a series of periodic noise patterns along a drifting cylindrical beam, by means of an axial-symmetric cavity trailing in the wake of a moving, immersed grid. In addition, they were able to observe changes in beam structure with the aid of Ashkin's beam analyzer,[12] mounted behind the cavity. The beam was produced

by a convergent, shielded gun, and focused by a uniform axial magnetic field in the drift region. The cathode flux could be varied by an auxiliary coil near the cathode.

Some of these observations can be explained on the basis of (a) a general description of the nature of interception noise at microwave frequencies, given by Beam[13] and (b) the nature of higher-order space-charge waves in beams produced by partly-shielded guns, as described in the first section of this paper. This explanation will apply to the periodic noise patterns obtained with the pickup cavity located half a plasma wavelength (in the fundamental mode) behind the moving grid, which fall roughly into two groups:

i. When the fields were adjusted to obtain clear images of the cathode region on the analyzer (22 to 36 gauss at the cathode), the beam was rippled and the noise-current pattern had sharp dips. Within the accuracy of measurement, these dips appeared to coincide with the image planes, and were spaced a cyclotron wavelength apart.

ii. When the fields were adjusted for maximum beam transmission through the gun anode (well below 10 gauss at the cathode), no cathode images were observed and the noise current varied sinusoidally, with large amplitude and the cyclotron period. The beam was comparatively smooth; i.e., its ripple was insufficient to account for the observed noise variations by variations in coupling to the cavity or in intercepted current.

5.1 *Sources of Interception Noise*

When a filamentary electron stream in a finite magnetic field is partially intercepted by a grid, Beam[13] has shown that the transmitted filament contains four uncorrelated noise components: the incident noise current reduced by the transmission factor, plus the incident axial-velocity fluctuations, and, in addition, two new independent fluctuation sources, partition velocity and current, which are due to the uncertainty of electron position at the grid plane or the randomness of interception. The first two components of interception noise, therefore, are produced by the noise space-charge *waves* in the incident beam, whereas the latter two components are due to the behavior of the *particles* in that beam. The latter components arise because of transverse thermal velocities which are uncorrelated with the longitudinal ones; they would be absent in confined flow.

The beam of *finite area* is equivalent to a bundle of many filamentary streams, whose space-charge waves are coupled to one another. Thus, all of the propagating space-charge modes are involved in transporting the current and velocity fluctuations. (The question of the completeness of these modes, in the mathematical sense, does not enter here;[14] it is only necessary that all of the propagating modes be known.) Since the transverse distribution of each incident mode is distinct and unlike the nearly uniform distribution of partition components due to random interception at a grid, each incident mode contributes differently to each of the transmitted modes.

An additional complication that usually besets beams in finite magnetic fields, as Robinson and Kompfner[15] have shown, is an increased spread in longitudinal velocities over the beam area, and increased transverse electron excursions, due to electron-optical defects in beam focusing. For a strongly rippled beam which is alternately focused and defocused Herrmann[16] has shown how the transverse thermal excursions wax and wane along the beam. Increased transverse excursions correspond to increased current partition noise, whereas increased spread in longitudinal velocities means increased velocity partition noise.

Despite the complexity of this description, some general conclusions may be drawn relevant to the Ashkin-White observations:

i. Due to nonlinear mode conversion at an immersed grid, the noise current in the fundamental transmitted mode will depend on all of the propagating modes in the incident beam.

ii. The amplitude of noise current induced in the axial-symmetric cavity, a half-plasma-wavelength behind the grid, will depend chiefly on two factors in the incident beam: the current amplitudes of all the space-charge modes and the transverse excursions of electrons in the incident beam.

5.2 *Noise Modes in Imperfect Brillouin-Flow Beam*

Electron beams ordinarily obtained in the laboratory with incompletely shielded, convergent guns are known to depart considerably from the models assumed in space-charge-wave computations. Thermal electron motions, gas ions and haphazard focusing usually conspire to produce a rippled beam with more or less nonlaminar flow.[16,17] Nevertheless, there is experimental evidence that space-charge waves in such beams closely resemble those predicted for the idealized model with the same average velocity field.

The writer, for example, has found that, in just such an imperfect beam, the radial distribution of ac current density corresponded closely to that calculated for Brillouin flow.[18] In addition, the measured space-charge wavelength (for the fundamental mode) was found to agree closely with the calculated values, based on the average diameter of the usually rippled beam, for field strengths ranging from below the nominal "Brillouin field" to several times that value. Good agreement between measurements and these calculations has also been reported by Winslow[19] for a gap-excited 10-kilovolt beam of microperveance one.

The reason is that the space-charge waves are not dependent on the individual electron trajectories (which may intercept the axis regularly or not[16, 17] in a rippled beam) but only on the net motion of the charge assemblage. Over a considerable range of field strengths, the average beam diameter varies inversely with the field, so that the average plasma frequency remains proportional to the cyclotron frequency over that range, just as in ideal Brillouin flow. When magnetic flux threads the cathode, the field distribution at any cross-section plane of a rippled beam will depart from that in a smooth beam due to (a) the ripple itself, and (b) nonlaminar flow. The former condition causes the angular velocity $\dot{\theta}$ of the smoothed-out charge to vary from plane to plane *along* the beam, whereas the latter causes $\dot{\theta}$ to vary with radius *inside* of the beam. The ensuing field distortion in both cases is periodic along the rippled beam, however, and for relatively small cathode flux or ripple is not likely to produce marked changes in the space-charge wavelength (relative to that in a comparable smooth beam in laminar flow, with the same cathode flux and average beam diameter).

In another set of relevant observations, Ashkin[20] has excited such a beam in the $n = 1$ and $n = 2$ modes, respectively, by means of cavities with the appropriate angular periodicities, and then traced the spiral loci of the current minima along the beam by means of similar pick-up cavities. In each case, the current minima were found to follow the computed axial and rotational fluid, or average, velocities of the beam, in agreement with the description of such waves in the first section of this paper.

When the cathode of such a beam is shielded, the field pattern for any mode is essentially the same in the diode and drift regions. (For small values of the flux parameter α, the transverse field distribution is only slightly different from that in Brillouin flow, and the space-charge wavelength is slightly smaller.) As nearly all such mode-pairs but the fundamental have the same standing-wave periodicity (in both diode and drift regions), and are initially excited at the same plane near the cathode,

they will preserve phase coherence along the axis even after partial energy interchange among the modes in the diode. In the Ashkin-White experiment, therefore, *nearly all of the high-order modes in the beam striking the grid have the same current-minimum planes, spaced a cyclotron wavelength or so apart.*

As primary current fluctuations can occur in an infinitesimally small area, the modal distribution of noise currents is probably nearly flat The sum of the squared moduli of all but the fundamental mode should then greatly outweigh that of the latter alone; and the same should be true of their net contribution to noise current in the fundamental mode, excited at the grid. In a relatively smooth beam, therefore, in which the electron-interception probability is independent of grid position, the cavity-detected noise current should vary sinusoidally and with the cyclotron period, at any frequency. This was the pattern observed under such conditions by Ashkin and White at both 400 and 4000 mc.

When the fields were adjusted for sharp cathode images, the flux parameter α ranged from $\frac{1}{4}$ to $\frac{1}{2}$, and the beam was strongly rippled. Both factors helped to minimize the transverse thermal excursions at the image planes, and thereby the contribution of random interception to partition noise there. If these planes are, in addition, made to coincide with those of noise-current minima for the high-order modes, the observed noise dips should be very much sharper than in the smooth beam, again as observed. As the variation along the beam of partition noise due to random interception is very large in rippled beams with periodic imaging of the cathode, the sharp noise dips are primarily due to such variations, rather than to current variations in the noise standing waves.

The two groups of noise patterns, therefore, illustrate the dual nature of the sources of interception noise at the grid. In the smooth beam, the variations are chiefly due to noise current variations in the space-charge *waves*, whereas in the rippled beam with periodic cathode images they are chiefly due to uncorrelated transverse *particle* excursions of thermal origin.[16] Both processes sometimes happen to have the same axial periodicity, but they are otherwise distinct and independent of one another.

VI. CONCLUSIONS

The higher-order modes of slow space-charge waves on beams produced by shielded or partly shielded cathodes have azimuthal, but no

radial, periodicity. Another feature that distinguishes these waves from those in cylindrical confined-flow beams is that their axial phase velocities increase rapidly with the order number, n, and angular velocity, $\dot{\theta}$, of the rotating beam. With suitable means for exciting such modes, therefore, it should be possible to achieve interaction between a relatively low-voltage beam and the field of a structure with high phase velocity. This should be equally possible for (a) hollow rotating beams, focused in any way whatever, provided they are stable and (b) solid-cylindrical beams produced by shielded guns, with arbitrarily strong focusing fields (since only the net angular velocity of the beam, not the particle trajectories, affects the wave velocity). More generally, the same type of interaction should be possible with stable beams of any geometry, when they have a transverse velocity component parallel to the beam surface and are excited by RF fields which are periodic in that transverse direction.

Another interesting property of harmonic waves on a Brillouin-flow beam is that, because the axial and radial propagation constants are equal, the rate of decay of fields with distance from an enclosing RF structure can be smaller, the smaller this constant is. A computation indicates that, consequently, the coupling of this beam to the harmonic fields of a sheath helix can be quite large. Experimental evidence of such interaction has been reported.[4]

When a Brillouin-flow beam is at synchronous velocity inside a concentric sheath helix, its plasma-frequency reduction factor in the fundamental mode has been found to be the same as if the beam were in a drift tube of the same diameter as the helix.

The computations also show that, for nearly all of the higher-order space-charge modes on beams from shielded or nearly-shielded guns, the space-charge wavelength is close to twice the cyclotron wavelength. This feature, together with a multimode description of interception noise given by Beam,[13] has helped explain some periodic noise patterns obtained with a cavity behind an immersed grid.[5]

VII. ACKNOWLEDGMENTS

The writer wishes to thank Mrs. C. Lambert, Miss M. C. Gray and Mrs. I. Leopold for computing the curves of Figs. 1–3. He is also indebted to J. R. Pierce for permission to publish his derivation of (27), and to C. F. Quate and L. D. White for invaluable advice and criticism.

REFERENCES

1. Rigrod, W. W. and Lewis, J. A., Wave Propagation Along a Magnetically Focused Cylindrical Electron Beam, B.S.T.J., **33,** March 1954, p. 399.
2. Brewer, G. R., Some Effects of Magnetic Field Strength on Space-Charge-Wave Propagation, Proc. I.R.E., **44,** July 1956, p. 896.
3. Kiryushin, V. P., Experimental Investigation of Bifilar B.W.O., Radiotekhnika i Elektronika, **1,** June 1956, p. 798.
4. Kiryushin, V. P., Influence of Azimuthal Components of Spatial-Harmonic Electric Field on Performance of Tubes with Spiral Delay Circuits, Radiotekhnika i Elektronika, **2,** October 1957, p. 1310.
5. Ashkin, A. and White, L. D., private communication.
6. Rigrod, W. W. and Pierce, J. R., this issue, pp. 99–118.
7. Pierce, J. R., *Traveling-Wave Tubes*, D. Van Nostrand Co., New York, 1950.
8. Fletcher, R. C., Helix Parameters in Traveling-Wave Tube Theory, Proc. I.R.E., **38,** April 1950, p. 413.
9. Sensiper, S., Electromagnetic Wave Propagation on Helical Conductors, D. Sc. Thesis, M. I. T., 1951.
10. Branch, G. M., Reduction of Plasma Frequency in Electron Beams by Helices and Drift Tubes, Proc. I.R.E., **43,** August 1955, p. 1018.
11. Paschke, F., Die Wechselseitigkeit der Kopplung in Wanderfeldröhren, Arch. Elek. Über., **11,** April 1957, p. 137.
12. Ashkin, A., Electron Beam Analyzer, J. Appl. Phys., **28,** May 1957, p. 564.
13. Beam, W. R., Interception Noise in Electron Beams at Microwave Frequencies, R.C.A. Rev., **16,** December 1955, p. 551.
14. Bobroff, D. L. and Haus, H. A., Uniqueness and Orthogonality of Small-Signal Solutions in Electron Beams, to be published.
15. Robinson, F. N. H. and Kompfner, R., Noise in Traveling-Wave Tubes, Proc. I.R.E., **39,** August 1951, p. 918.
16. Herrman, G., Optical Theory of Thermal Velocity Effects in Cylindrical Electron Beams, J. Appl. Phys., **29,** February 1958, p. 127.
17. Harker, K. J., Nonlaminar Flow in Cylindrical Electron Beams, J. Appl. Phys., **28,** June 1957, p. 645.
18. Rigrod, W. W., Noise Spectrum of Electron Beam in Longitudinal Magnetic Field. Part II-The U.H.F. Noise Spectrum, B.S.T.J., **36,** July 1957, p. 855.
19. Winslow, D. K., The Current Distribution in Magnetically Focused Modulated Electron Beams, M. L. Rep. No. 380, Stanford Univ., April 1957.
20. Ashkin, A., private communication.

An Experimental Visual Communication System

By F. K. BECKER, J. R. HEFELE and W. T. WINTRINGHAM

(Manuscript received March 10, 1958)

Substantial technical and economic benefits are obtainable by fitting a visual communication system to a specific application. Some of the considerations involved in such adaptation are discussed in this paper.

An experimental system to demonstrate the specific adaptation to the problem of signature verification in a savings bank is described. It is shown that satisfactory images can be transmitted over a 5-kilocycle sound program circuit in 5 seconds. This result is obtained by reducing both the area scanned by the transmitter and the resolution of the reproduction to the minimum required for this application.

I. INTRODUCTION

Facilities for the transmission of visual material by facsimile have been offered to the public for over 30 years. Even so, facsimile has not become a very widely used service. In contrast, the post-war development of broadcast television has excited much interest in the use of television as a means for transmitting visual material. Industrial television equipment has been developed and sold for all sorts of applications. However, there is increasing awareness that television may not be the most suitable and least expensive way of filling some of the needs for visual communication for which its use has been suggested.

Television as a private visual communication means is attractive from many points of view. Much of the terminal equipment is similar or identical to that developed for broadcast purposes. Consequently, the costs of terminal equipment can be kept low through the benefits of mass production for broadcasting. In many installations the receiving terminal may be nothing more than a standard broadcast receiver. This has the added advantage that little training is required for operation of the receiver.

On the other hand, transmission of television signals produced under

broadcasting standards requires a bandwidth of the order of four megacycles. Wherever distances greater than a few hundred feet are involved, suitable circuits of this bandwidth are expensive to provide.

These facts suggest the wisdom of examining the general problem of transmitting visual material. Such an examination should weigh the requirements of particular communication problems against the advantages offered by facsimile and by television.

In a broad way, one can distinguish facsimile and television by the character of the received image. This image is a permanent copy of the transmitted material in facsimile; it is transient when the transmission is by television. Since facsimile produces a permanent image, the material need be sent only once. In contrast, the transient character of the television images necessitates the transmission of the material often enough to avoid flicker and for as long a time as examination of the material is required.

These distinctions between facsimile and television suggest the application for which each system is most suited. Facsimile is the more appropriate medium if the material to be transmitted is itself in permanent form. Television is the more appropriate medium if the material to be transmitted is in transient form. That is, television should be used if motion is an important attribute of the original.

This seemingly clear-cut distinction between facsimile and television becomes blurred if one introduces the possibility that a permanent copy of a document may not be required and even may be undesirable. In such cases, the transient character of the television image is attractive.

A transient image, however, would be useful only if it were available for study for several minutes. This suggests the possibility that some system intermediate between facsimile and television might be useful. Since the person receiving the information could use no more than one document during the time required for his study of it, the system need have only the capacity to transmit a single document at a time. It would seem desirable that this system complete a transmission in a few seconds.

This speculative thinking only discloses that a hybrid visual communication system might have some advantages over either facsimile or television. But the difficulties of system design are not revealed. It was decided that a complete system should be built to gain some measure of the complexity of such hybrid systems. This experimental system was intended to demonstrate the feasibility of combining techniques from television and from facsimile to produce a visual communication system for a specific field of use. For this reason, the experimental equipment was built without regard for the usual engineering limitations

of size or of cost. The sole aim was the investigation of the feasibility of tailoring a visual communication system to a single communication problem.

II. FEASIBILITY STUDY

Our preliminary thoughts and discussions about a new intermediate visual communication system revealed the fact that a study completely divorced from application would be of little value. Through the cooperation of Albert F. Kendall, Comptroller of the New York Savings Bank, it was decided that our feasibility study would be related to one of the more serious problems of larger savings banks.

With the growth both of individual savings banks and of branch banking it has become increasingly difficult to give each teller access to the complete file. Savings banks have been active in experimenting with the application of modern visual communication techniques to this problem. Facsimile, industrial television and slow-scan television all have been tried. None of these techniques has been found to be completely satisfactory. Therefore, our study was based on the problem of transmitting signature and account information from a central file to individual bank tellers at remote locations.

Fundamentally there are two requirements to be applied to a visual communication system for this service. The time required for transmission should be reasonably short and the rental fees for the communication circuits should be kept small. This latter requirement may be interpreted as meaning that the transmission bandwidth should be minimized and should be within the capabilities of a standard Bell System facility.

In order to keep the time of transmission short and the bandwidth small, the time-bandwidth product must also be minimized. This result is attained when no more than the necessary area of the copy is scanned and when the resolution of the received picture is no greater than is acceptable.

III. SYSTEM DESIGN

The information to be transmitted to the teller from the New York Savings Bank files is contained on a signature card and an account card, of which samples are shown in Fig. 1. The 3- x 5-in. signature card may be any one of several types. For individual accounts, the account number and the signature may appear at the top of the card or the signature may appear below a printed agreement. Another common variant provides for two signatures for a joint account, written below an extended

form of printed agreement. In the case of every one of these cards, the signatures fall within an area 1 in. high. The account number, used for identifying the signature card, can be restamped to the left of the signature(s) in this 1-in. area. The significant information on the card may be contained therefore within a 1- x 5-in. rectangle.

The information of interest to a teller on the account card is contained

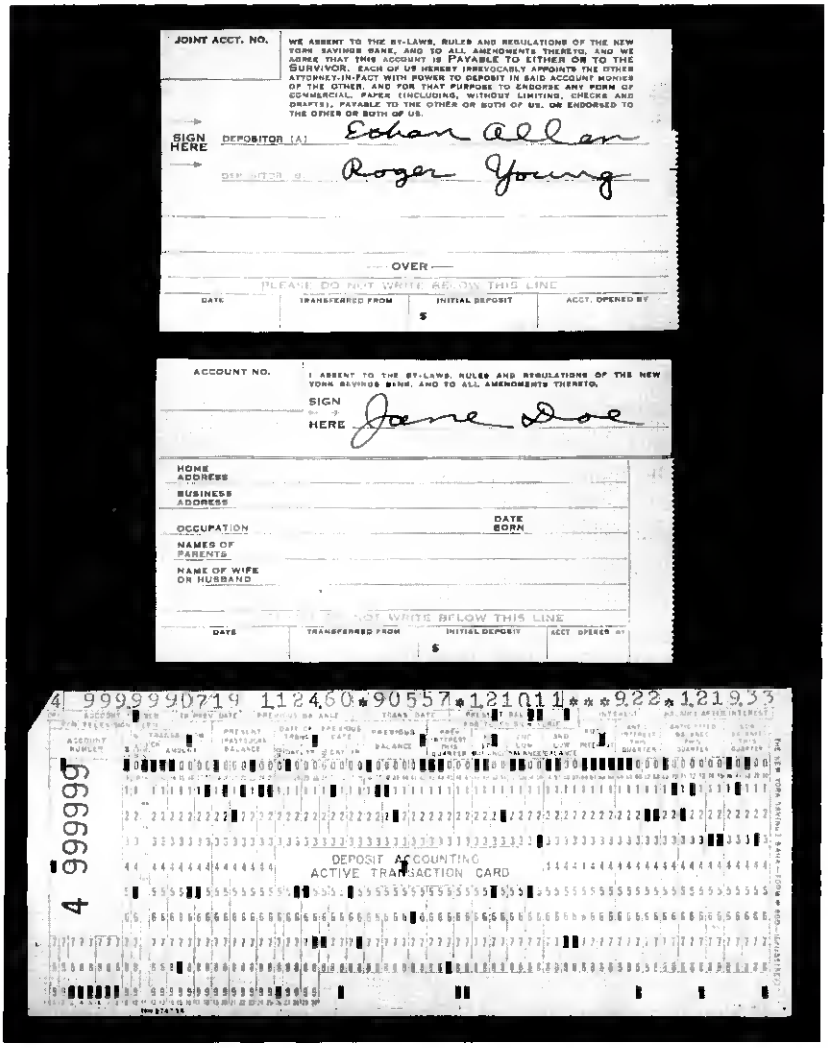


Fig. 1 — Signature cards and accounting machine card.

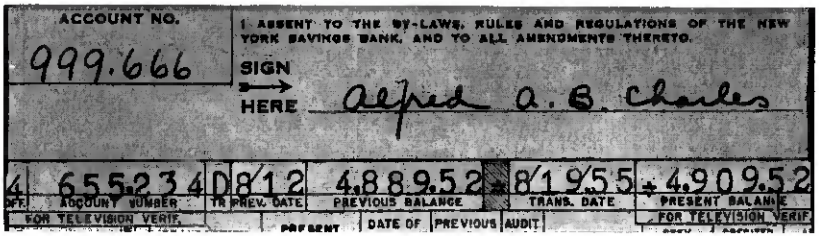


Fig. 2 — Bank signature and account card positioned so as to present only essential information.

in the left-hand 5 in. of a single line of type within the top $\frac{1}{4}$ in. of the card. As one step in optimizing the visual transmission of these data, the scanned area should be limited to the important 1- x 5-in. area of the signature card and the $\frac{1}{4}$ x 5-in. area of the account cards. A suitable arrangement of the two cards for such scanning is illustrated in Fig. 2, where the total scanned area of the two cards need be no greater than $1\frac{1}{4}$ x 5 in.

To maintain the two cards in this position, several types of card holders were considered from an operating and also a convenience point of view. Fig. 3 shows a laminated metal card holder in which the cards are placed into two depressions and held in place with a hinged cover (shown in right side of Fig. 4). Fig. 4 (left) shows a stiff plastic envelope which could be economically made and readily stacked. It is cut so that the positioning of the cards would be maintained when the holder was inserted in the transmitter. These two holder designs showed significant differences in loading and unloading times when used by a file clerk.

Fig. 5 shows the holder adopted for the experimental system, which can be loaded and unloaded more quickly than the previous designs. The thin holders are made from a laminate of thin fibre, each section cut with the appropriate mask to fit the particular card and assembled into a sandwich. Two types were made: one from brown fibre (top) to fit the single signature cards; the other from black fibre (bottom) to fit the single and double signature cards bearing the bank agreement. The cards are held flat with a spring clip on the right side which clamps both cards firmly in place.

The second factor to be minimized in reducing the time-bandwidth product is the resolution of the received picture. Experiments have been carried out at Bell Telephone Laboratories[1] with a television system to show how few scanning lines might be required to display signatures. Fig. 6, taken from that study, shows that two signatures can be recog-

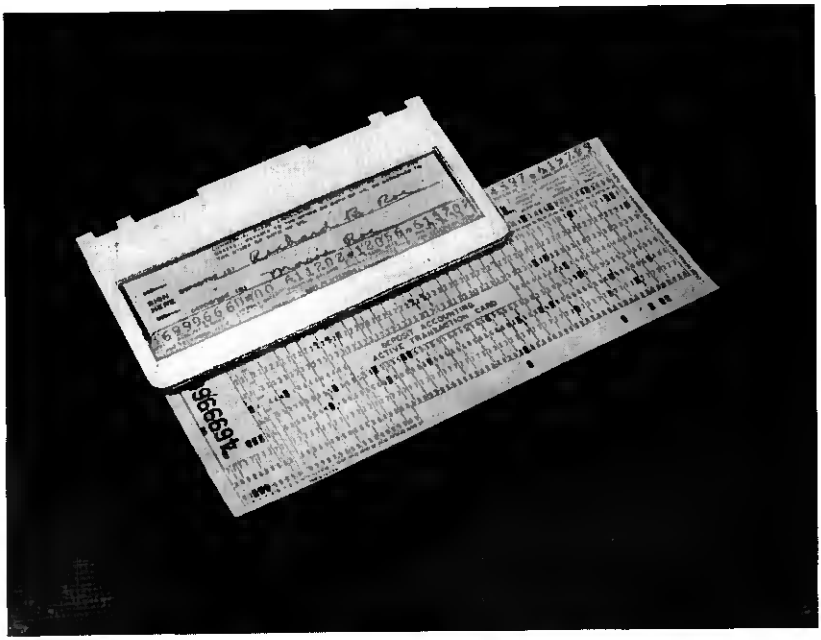


Fig. 3 — Metal card holder with hinged cover.

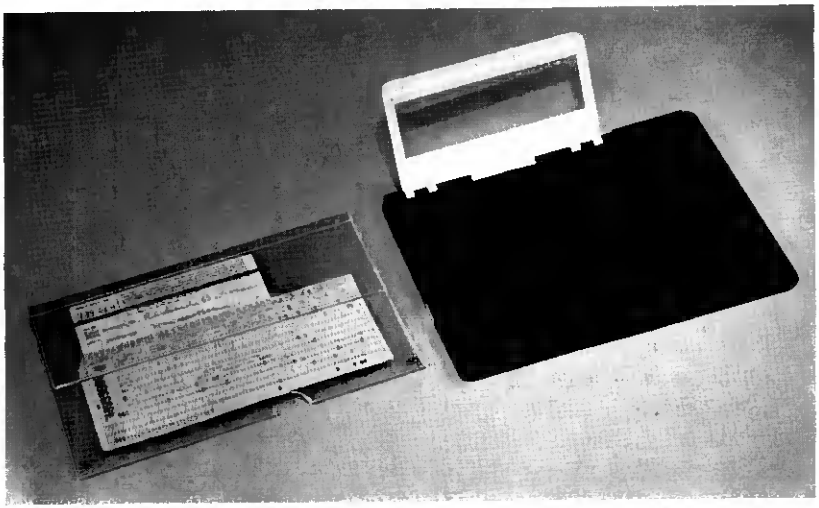


Fig. 4 — Experimental plastic and metal card holders.

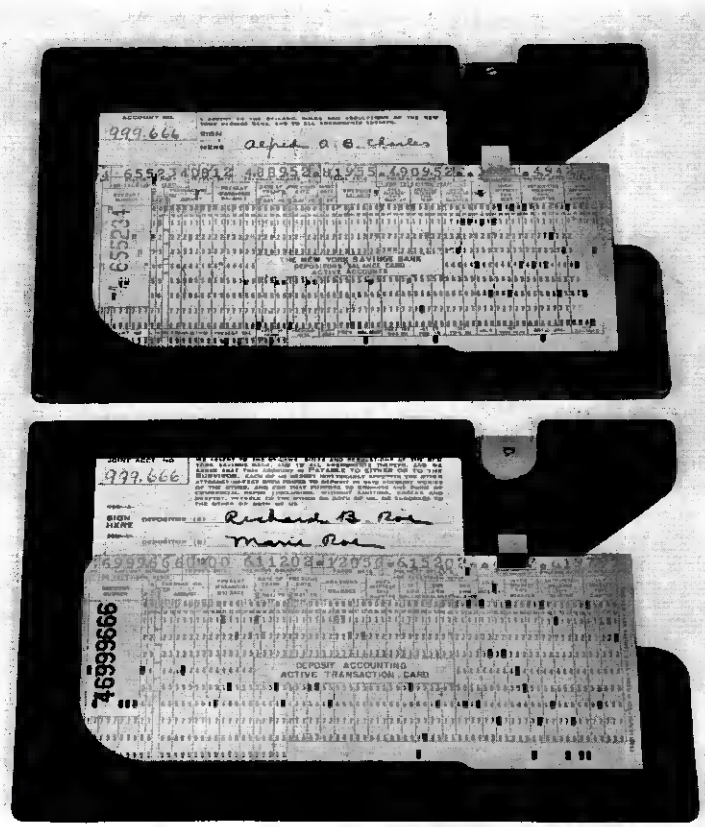


Fig. 5 — Card holder used in bank signature verification transmitter.

nized with little masking of their essential characteristics when 80 scanning lines are used. Fig. 7, taken from the same study, shows that a single signature and the numerals on an account card are quite readable when 120 scanning lines are used. In this illustration, there are 9 or 10 lines reproducing the line of numbers.

The printed digits on the account card are 0.150 in. high and 0.095 in. wide. Hence, the scanning line pitch is about 0.015 in. in the case where the digits are easily read. If the vertical dimension of the important area of the signature card — 1 in. — is scanned at this same pitch, it would be covered by only 67 scanning lines. However, it was concluded

that 80 scanning lines were required for the area containing two signatures.

The more stringent of these two requirements, 80 scanning lines per inch, must be adopted for the whole area. With a scanning pitch of 0.0125 in., the signature area requires 80 lines, and the top $\frac{1}{4}$ in. of the account card requires 20 lines, or a total of 100 lines across the entire

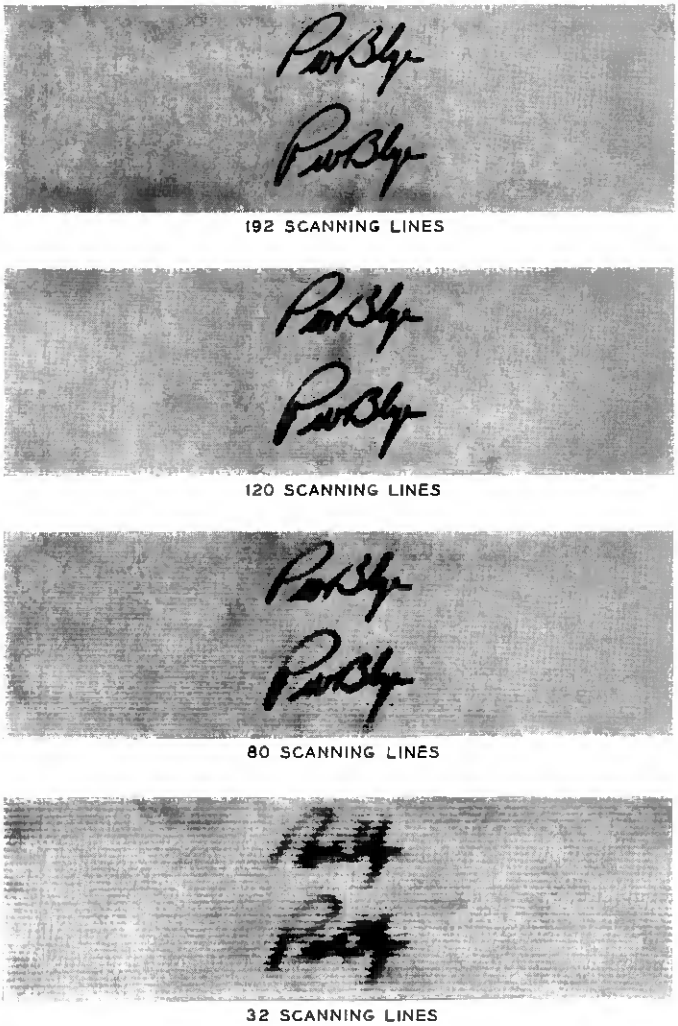


Fig. 6 — Two signatures transmitted over a TV system with scanning line structures as indicated.

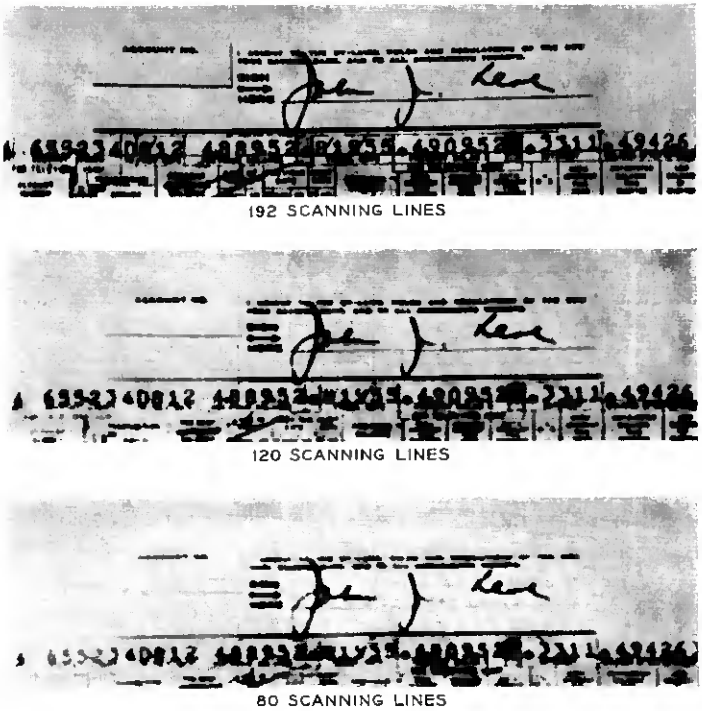


Fig. 7 — Bank signature and account card transmitted with scanning line structures as indicated.

$1\frac{1}{4}$-in. dimension of the two cards. It is of interest to compare this scanning line density with the 90 to 100 scanning lines per inch used in commercial facsimile systems.

It seems desirable to equalize the apparent horizontal and vertical resolutions in the reproduction. This condition is reached when the number of cycles of picture signal per scanning line is given by

$$f_l = l K N_v/2, \tag{1}$$

where

f_l = number of cycles per line,
l = length of line (inches),
N_v = number of scanning lines per inch,
K = vertical resolution factor.

The factor K takes into account the fact that the apparent vertical

resolution is smaller than is given by the number of scanning lines. This decrease arises because scanning is a sampling process and the samples may or may not coincide with detail in the copy. M. W. Baldwin has suggested that the best value of K is 0.7.[2]

In the case of the signature and account cards, the lines are 5 in. long, and there are 80 scanning lines per inch. Substituting these values in (1), we have

$$f_l = 5 \times 0.7 \times 80/2 = 140 \text{ cycles/line.}$$

Since a total of 100 scanning lines is required to cover the significant areas of the signature and account cards, one transmission will consist of 14,000 cycles of picture signal.

This figure, 14,000 cycles, is a measure of the time-bandwidth product required for the transmission of the important information from the bank's file. We now have the problem of dividing this quantity into its two factors, time and bandwidth. The significant parameter at our disposal is the bandwidth of the type of Bell System facility which might be used for this service.

To give proper consideration to the various characteristics of the several types of circuits available in the Bell System, we must digress slightly to consider the waveform of signals which should be transmitted. It seems that the gradations of tone in signatures, arising from changes of pen pressure and writing speed, are not significant in the recognition of signatures. Consequently it is necessary to transmit only two values of picture signal, one corresponding to the background on which the signature is written and the second corresponding to the inked lines. It follows that we may make use of experimental results which bear on the transmission of telegraph-like signals.

Horton and Vaughan have reported that telephone message circuits can be used for the transmission of digital information satisfactorily in the band from 700 to 1700 cycles per second.[3] Carrier transmission is required to handle picture signals in this band, so that the baseband may be no wider than, say, 500 to 800 cycles per second. The more optimistic of the figures leads to the conclusion that nearly 20 seconds would be required to transmit our minimum picture signal over a message circuit.

Program circuits for sound broadcasting are available from the Bell System with bandwidths of 5 kc, 8 kc and 15 kc. By far the most widely available of these is the 5 kc circuit, the gain characteristics of which are substantially flat between 100 and 5000 cycles. The delay distortion of these circuits is at a minimum between about 1 kc and 4 to $4\frac{1}{2}$ kc.

This suggests that a baseband bandwidth of 3 kc might be used with vestigial sideband transmission about a carrier near 4 kc.

Under these conditions, it would require 4.7 seconds to transmit our minimum picture signal. This figure is sufficiently small in comparison with the total time required to order the information from the file clerk, locate the cards in the file and load the card holder that there seems to be little advantage in considering wider band circuits.

Actually, we have chosen a transmission time of 5 seconds instead of 4.7 seconds. This increase in time tends to compensate slightly for the imperfections of a practical terminal equipment.

IV. TRANSMITTER

In addition to the general characteristics of the system which have been discussed in the preceding paragraphs, certain peculiarities of the file cards place restrictions on the transmitter. The New York Savings Bank uses signature cards of several different colors for special purposes. In addition, some of these cards have been used for a good many years. Consequently, the background reflectance of the original copy varies significantly, and it was necessary to introduce an automatic gain control circuit in the scanner to compensate for the variation.

In addition to the color of the background, the bank's customers exercised their individualities and wrote their signatures with inks of almost every conceivable color. Fortunately it was found possible to provide sufficient gain in the scanner circuits to allow the signal corresponding to the ink to be sliced at a fixed amplitude.

The individuality of the bank's customers displayed itself in another way. It was found that the width of the lines in their signatures varied from 0.005 in. upwards.

The length of a reproduced picture element was fixed when the transmission bandwidth and the scanning standards were selected. No more than 6,000 picture elements per second can be transmitted in a 3,000 cps baseband. The scanning of a complete image with 100 lines in a time of 5 seconds means that each line is scanned in $\frac{1}{20}$ second. Hence, there can be no more than 300 picture elements per line. This means that each picture element in the reproduction can be no shorter than $\frac{1}{60}$ in. (or 0.017 in.), since each line is 5 in. long.

It is obvious that the signals corresponding to lines in the signatures as narrow as 0.005 in. would be reduced in amplitude after transmission through this system. It was learned, however, that there was no objection to artificially broadening these lines in the electrical equipment.

However, to resolve 5-mil lines on a 5-in. card demands good 1000-line

resolution capability. Cathode ray tube scanners, as developed for entertainment television, fail to meet this requirement by nearly a factor of two.[4] Thus, because the required line scan rate is low, a mechanical scanner was suggested. Then, scanning spot size, and therefore resolution, would be only a question of precision machine work and lens capabilities. The problem of illuminating the subject material to obtain a useful electrical signal-to-noise ratio could then be divorced from any consideration of scanning spot size, as it cannot be in cathode ray tube or flying spot scanners.

4.1 *Scanner*

Fig. 8 is a photograph of the entire transmitter. The mechanical optical scanner or analyzer (which was constructed to Bell Laboratories specifications by Hogan Laboratories, Inc.) is located in the middle of the table top. The signal processing equipment for broadening the fine lines in copy, the frequency translation equipment and the power supplies are enclosed below the table top. A monitoring picture tube which reproduces the transmitted signal is mounted above the scanner.

A schematic of the optical path of the mechanical scanner is shown in Fig. 9. An optical system forms a reduced image of the illuminated card at the intersection of a spiral slit on a disc and a straight slit which is fixed parallel to the long side of the image. The intersection of the spiral and the straight slit forms a nearly rectangular aperture which sweeps across the image at a constant speed when the disc is rotated at a constant rate. The optical path is folded to save space, so that the objective lens actually views the cards through two front-surfaced mirrors.

A 60-cycle synchronous motor rotates the disc at 20 revolutions per second, causing the aperture to sweep across the slit and thereby scan a single horizontal line of the image for each revolution. The slit apertures are of such a size that the resolution in the horizontal direction is greater than 1000 lines.

A second motor, energized when the transparent lid is closed and an address button is pushed, moves the carriage forward. The carriage travel causes the image of the subject material to traverse across the fixed linear slit in 5 seconds, thus producing the vertical scan. At the finish of the scan, the carriage motor is automatically stopped, and the carriage releases and snaps forward into a position convenient for unloading and reloading the card holder. The light from each successive point of the image, as determined by the scanning aperture, is converted by a multiplier phototube to a proportional electrical signal.

Line-synchronizing pulses are derived directly from the rotating

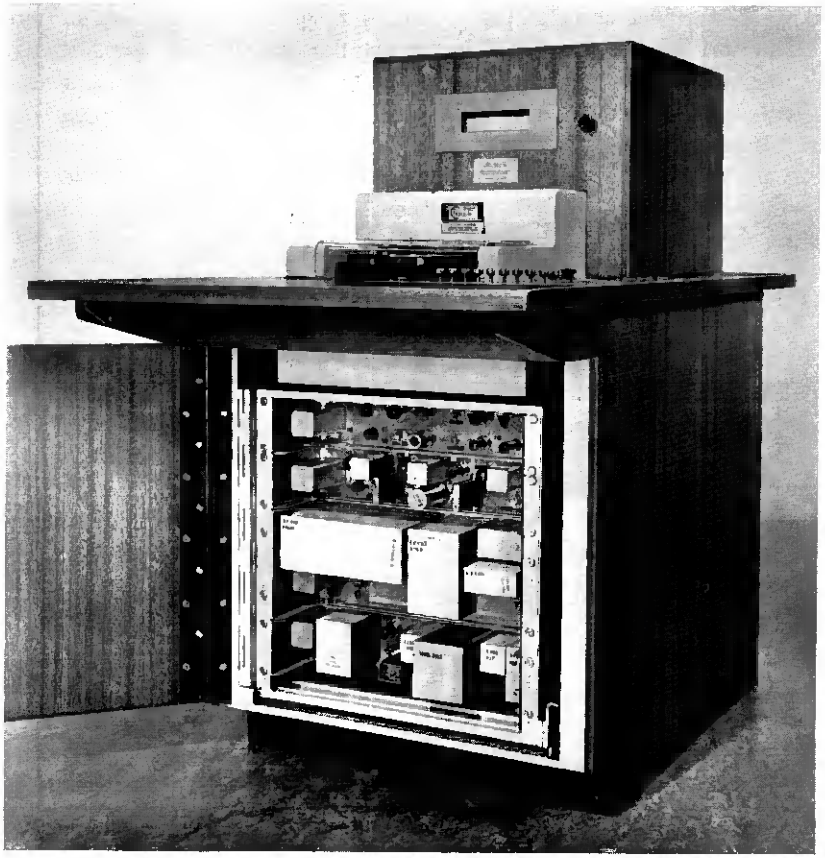


Fig. 8 — Experimental transmitter unit of the bank signature verification system (door open).

scanning disc. Through a cleared circular slit on the opaque scanning disc, a small source of light is focussed onto a photo diode. An opaque section, properly positioned, interrupts this light, thus producing an electrical pulse once per revolution. This pulse, when processed and added to the outgoing signal, produces the line-scan sync signal for receiver synchronization.

A block diagram of the signature verification transmitter system is shown in Fig. 10. From the multiplier phototube and cathode follower in the mechanical scanner, the signal baseband frequencies are applied to a signal processing circuit, in which the analog signals, derived from scanning, are clipped either to black or to white signal level. The circuit

likewise stretches the duration of fine line signals to a minimum value satisfactory for transmission. The rectangular output pulses pass through a low-pass filter and then modulate a carrier at 26.88 kc. A second stage of modulation and filtering produces a vestigial sideband signal for transmission having its carrier at 3.84 kc.

A portion of the output signal is demodulated at the transmitter to operate the picture display tube, which serves as a check on the outgoing signals.

4.2 *Signal-Processing Circuits*

A feature of the signal-processing circuit is its nonlinear pulse-stretching operation. Variable-duration pulses with a minimum duration of 160 microseconds can be transmitted through a band limited system of 3 kc and can be utilized satisfactorily with a proper receiver. The scanner, however, may generate signal pulses as short as 50 microseconds when it scans signatures or other material written with a very fine pen. The combination slicer and pulse-stretcher circuit operates on these pulses in the following manner. It recognizes first whether there is a signal pulse present or not, producing at its output a two-level signal free from the effect of background noise. Moreover, signal pulses arriving at the

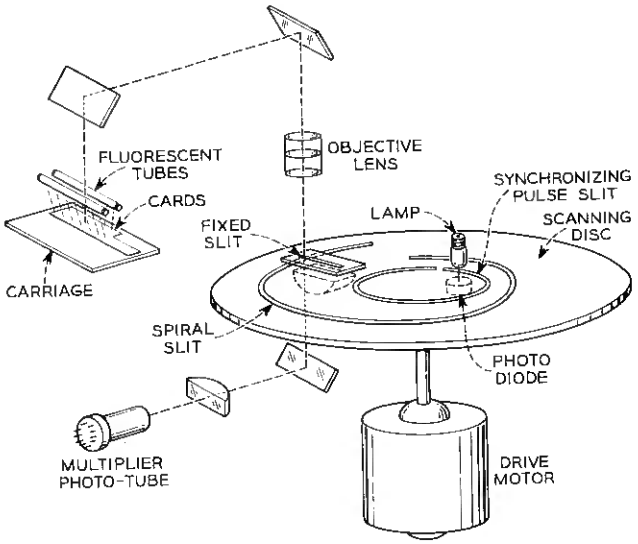


Fig. 9 — Optical path of the mechanical scanner.

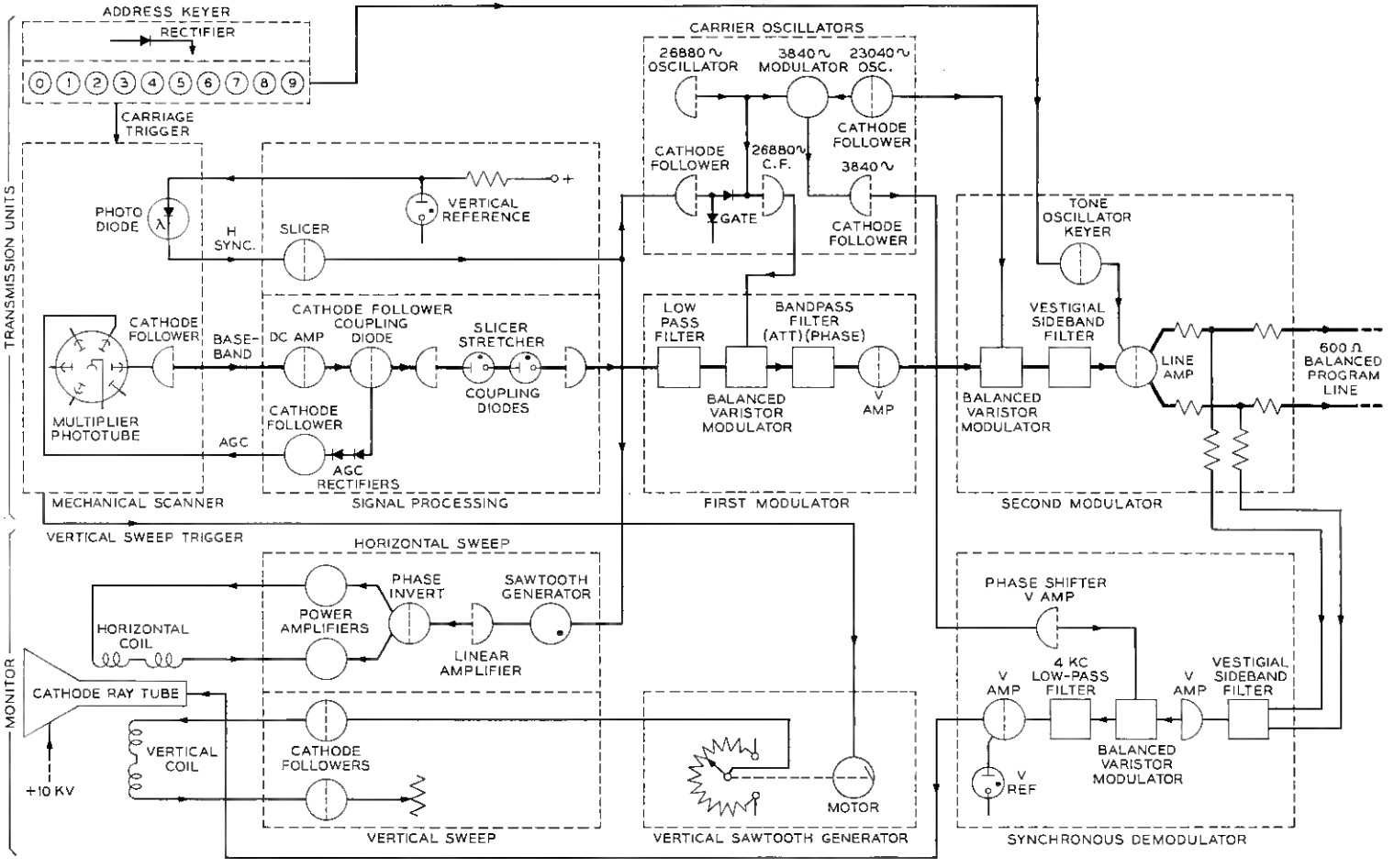


Fig. 10 — Block diagram of the transmitter terminal.

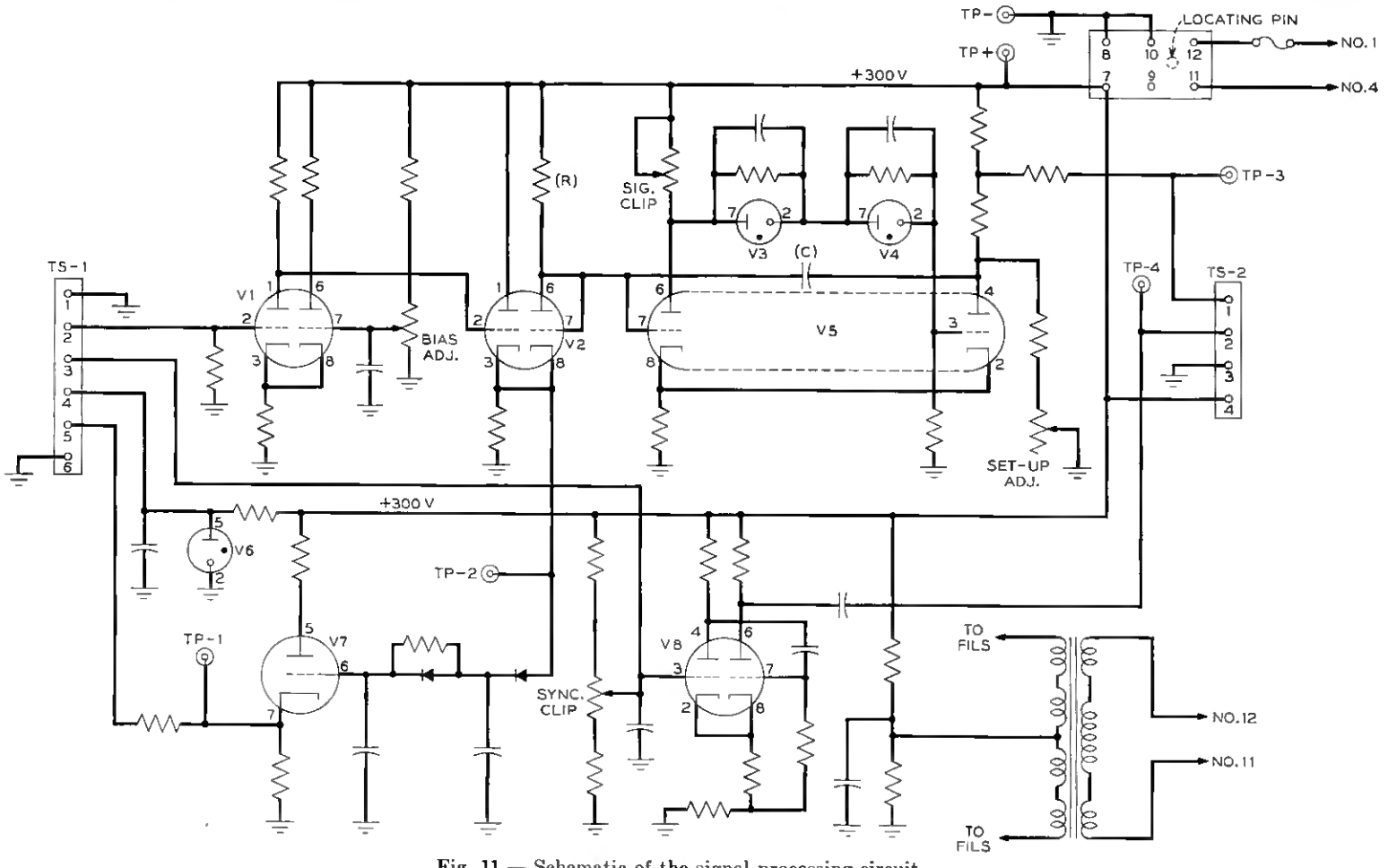


Fig. 11 — Schematic of the signal-processing circuit.

slicer input having duration less than 160 microseconds are lengthened and appear at the output as clipped rectangular pulses of 160-microsecond duration; the signal pulses of longer duration are merely regeneratively clipped into rectangular pulses with their longer durations undisturbed. The output signals of this unit are clean rectangular pulses corresponding to the inked lines or figures on the subject material to be transmitted.

The processing circuit is illustrated in detail in Fig. 11. This circuit, consisting of seven tubes, performs baseband signal amplification, automatic level control of the output signal, regenerative clipping and pulse stretching. The eighth tube on this chassis is used as a conventional slicer for the line sync signal.

The picture signal, as derived from a cathode follower in the scanner, has a maximum peak-to-peak value of 0.5 volt and a bandwidth of 10 kc. It is first amplified to a value of 40 volts by a dc amplifier and cathode follower. Tube V1-2 is a bias control to adjust the no-signal or black-signal level at the output to 60 volts. The white level signal, corresponding to the card background, is peak detected at this point, and through a fast-operate and slow-decay RC time constant, adjusts the voltage applied to the multiplier phototube through the cathode follower V7. The automatic level control action sets the background signal level at this point at a value of 100 volts, regardless of the color or the luminance of the card, within reasonable limits.

The signal is then sliced and shaped by the tubes V2-2 and V5-1. The slicer is direct-coupled, and the level of signal slicing is adjusted by changing the grid voltage on V5-2 through the signal-clipping control and the reference voltage tubes V3 and V4. The operation of this slicer has been modified by the positive feedback capacitor C.

When a negative signal pulse, due to the scanning spot crossing a black line on the card, arrives at the output of the cathode follower, it is transmitted through the diode V2-2 to the grid of V5-1, which tube has been conducting current. Since the cathodes of V5-1 and V5-2 are common, plate current in V5-2, previously zero, is quickly turned on by the simultaneous action of cathode and grid signals. The negative voltage pulse, developed across its plate resistor, is fed back through capacitor C to the grid of V5-1. The grid potential of V5-1, reduced below cutoff by the signal applied through the diode, is reduced still further by this positive feedback, since the pulse disconnects the grid from the low-impedance signal driver by inactivating the diode V2-2. Capacitor C now charges through resistor R to the potential of the cathode follower, at which point the diode V2-2 becomes conducting, returning the control of the slicer condition to the signal cathode follower.

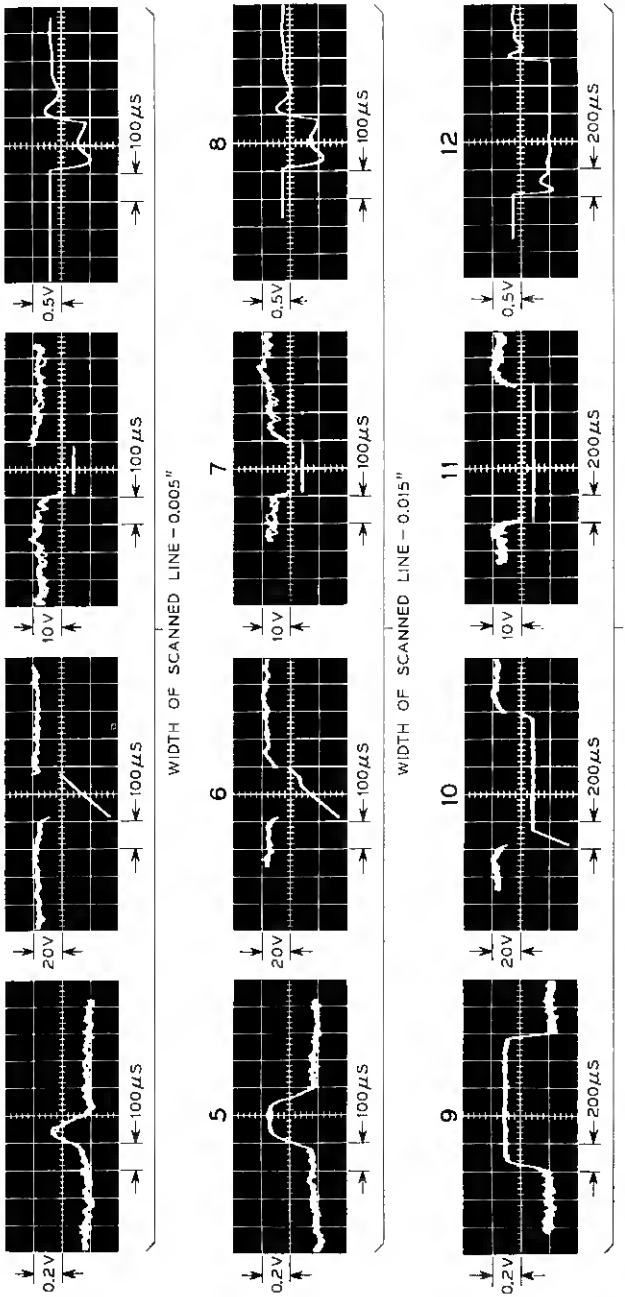


Fig. 12 — Waveforms of signals in the processing circuit.

If the signal pulse duration is longer than the time constant of the circuit composed of R and C, the charge cycle of the capacitor is terminated at the potential of the cathode follower, which remains at the triggering potential. The slicer circuit is therefore held in the triggered condition for the full duration of the pulse, since the voltage on the cathode follower is now controlling.

If the signal pulse duration is shorter than the time constant of the RC circuit — the cathode follower potential having become more positive due to the termination of the signal — the charge cycle of the capacitor terminates the triggered cycle, just as in a monostable multivibrator. The output pulse length is therefore determined by the RC circuit for signal pulses shorter in duration than the time constant of the clipper; it is determined by the length of the pulses themselves, if their time duration is greater than the time constant of the clipper.

By connecting the charging resistor R to the positive supply potential, a more linear portion of the timing exponential curve is used, and there is little lengthening of signal pulses of the same or longer durations than that of the RC circuit.

By setting the time constant of the slicer circuit to 160 microseconds, signal pulses arriving at the slicer input having duration less than 160 microseconds are lengthened and appear at the output as clipped rectangular pulses of 160-microsecond duration; the signal pulses of greater duration are regeneratively clipped into rectangular pulses with their durations undisturbed. The ouput signals of this unit are clear rectangular pulses corresponding to the inked lines or figures on the subject material to be transmitted.

The action of this combined slicer and pulse-lengthener circuit is illustrated in Fig. 12 by the waveform photos taken during its operation.

Photos 1, 5 and 9 show signal pulses of 50, 150 and 1000 microseconds respectively derived from scanning lines of the widths indicated.

Photos 3, 7 and 11 indicate the voltage across the common cathodes of the circuit and shows that the sliced pulses are of 160, 160 and 1000 microseconds duration respectively.

The output pulses, applied to the low-pass filter circuit are shown in photos 4, 8 and 12. The effect of the impedance characteristic of this filter on the extremely sharp transitions of the output waves is noticeable.

Photos 2, 6 and 10 show the wave shape at the input grid of the slicer. Photo 2 indicates that, after the slicing operation has been begun by the front edge of the short duration input pulse, the duration of its operation is controlled by the decay of the RC circuit.

Photo 10 shows that the slicer operation is controlled by the length

of the signal pulse when it is longer than the time constant of the *RC* circuit.

Photo 6 indicates the slicer operation when the signal pulse is approximately of the same duration as the *RC* time constant. (The loading of the oscilloscope at this point of the circuit shortens somewhat the decay time of the *RC* circuit.)

A second conventional slicer circuit, V8 and associated components, operates on the pulse from the photo diode in the scanner. It produces an output rectangular pulse synchronized in time with the black-signal pulse produced at the termination of each scanning line by the scanning spiral. The negative pulse is passed to the carrier-oscillator unit where it operates a gate circuit consisting of one-half of a 12AU7 and associated diodes. The 26.880-kc carrier is interrupted during the time the gate is open. This carrier zero is used for blanking the reproducer beam and, after separation, for horizontal synchronization of receiver scanning circuits.

4.3 *Carrier Modulation*

Modulation of the carrier at 3.84 kc with the picture signals is done in two balanced modulator stages (see Fig. 10). The picture signals, having been limited to a nominal bandwidth of 3 kc by a low-pass phase-equalized filter, modulate a 26.88-kc carrier in a lattice copper-oxide varistor. The output balance has been adjusted so that signals corresponding to inked lines on the original cards produce maximum modulation. Signals corresponding to background information produce half-maximum modulation and synchronizing signals gate out or produce zero carrier.

The double sideband frequencies around the 26.88-kc carrier are then filtered with a band-pass filter having flat loss and linear phase characteristics through the wanted-frequency band from 23.88 to approximately 28 kc. Tolerances are unspecified over other frequencies in the two sidebands, since they are eliminated in the vestigial filter which follows the second stage of modulation.

The resulting band of frequencies then modulates a 23.04-kc carrier, resulting in a 3.84-kc carrier with double sidebands. The result is shaped by a linear-phase low-pass filter which partially reduces the upper sideband and produces half of the vestigial shaping requirement of the system.

An identical filter located in the terminal receiving equipment completes the vestigial sideband shaping. Since it is located in the receiver,

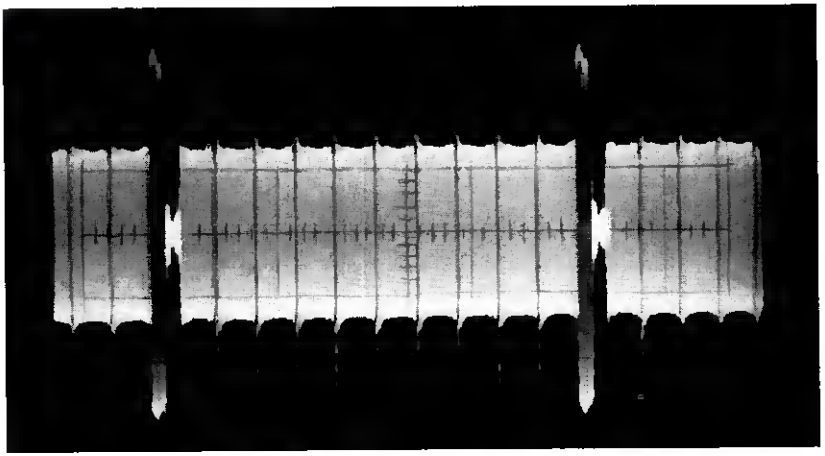


Fig. 13 — Oscillogram of transmitted signal.

this arrangement assists in reducing noise produced by the transmission lines.

The outgoing signal occupies the band from 840 to nearly 5,000 cycles and is transmitted to the balanced program circuit through an output line amplifier.

The waveform of the transmitted signal, corresponding to slightly more than one line, is shown in Fig. 13. The carrier amplitudes corresponding to the horizontal sync signals, to background and to inked lines are easily recognized.

4.4 *Addressing*

Since it requires 5 seconds to transmit a single message to one teller, means must be included to direct each message to a particular teller, so that the transmitter and the wire circuit can be released for the use of other tellers. This switching is accomplished by sending a short pulse of tone of a different frequency for each receiver at the start of each transmission. The frequency of the tone, which serves as the address selector, is chosen by the transmitter operator by means of a key switch mounted on the operating console. This address tone, a single frequency between 400 and 775 cycles per second, likewise serves as the vertical-start signal and activates the vertical scan for the chosen receiver. A momentary-contact relay, activated when the carriage on the scanner begins to move, gates a 40-millisecond burst of the address and vertical-start frequency, which is mixed with the picture signals.

The scanner carriage motor is started by two switches in cascade, one on the carriage cover and the other on the address key. Both must be closed to send the picture information, but the sequence is unimportant.

The processing and modulating circuits are located below the top of the operating console. The monitor unit, with its cover removed, is visible in Fig. 14.

4.5 *Operation*

The bank signature card and the corresponding account card are positioned in a card holder.

The card holder is pushed into position on the top of the scanner carriage with the transparent lid lifted.

Closing the lid and pushing the address button, in any order, initiates the transmission to a particular receiver.

A reproduction of the subject material is displayed for a short time on the picture monitor.

After the 5 seconds required for transmission, the carriage auto-

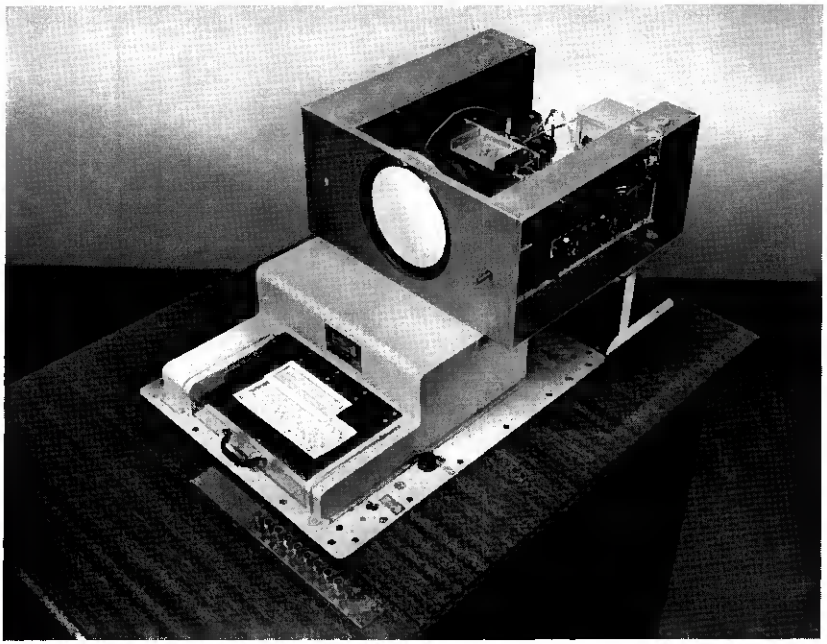


Fig. 14 — Transmitter unit, showing monitoring tube assembly.

matically returns to its "unloading" position, where the card holder can be easily removed.

A second loaded card holder can be inserted immediately for transmission of another message to another teller.

V. RECEIVER

It has been pointed out that the receiver for this experimental system should have several unusual characteristics. The selection of only one of a group of receivers, through the use of a burst of tone of a particular frequency at the onset of a transmission, does not present any very difficult problems. However, our desire to present the received picture in nonpermanent form is a different matter. Here the problem is that of displaying a picture for possibly as long as a few minutes from a single transmission lasting only 5 seconds. Obviously, storage of some form is required.

5.1 *Storage Devices*

A study of several storage devices was made to determine their relative ability to convert the electrically stored signal into a visual form. Magnetic storage drums were discarded because of circuit complexity required to avoid display flicker. The transmitted signal could be recorded on a magnetic drum turning one revolution in 5 seconds and reproduced at the same speed repetitively for display on a long-persistence-phosphor kinescope. However, it was found that the intense writing beam of such a kinescope is distracting to the viewer. Familiarity with this type of presentation does enable some persons to disregard the beam's presence, but others always find it annoying. The signal might be recorded on a drum rotating at $\frac{1}{5}$ rps and reproduced at 60 rps for flickerless display, but the accelerating time of the drum would have to be added to the file-search and transmission time. Existing sampling techniques could enable the magnetic drum to run continuously at the high speed but at the expense of added circuit complexity.

Electric-to-electric or signal-converter storage tubes, and electric-to-visual or direct-view storage tubes were investigated. The electric-to-electric variety of tube has little to recommend it over the electric-to-visual model as to relative performance in this application, and the direct-view type requires much simpler circuitry.

Tallying the advantages and disadvantages of these forms of storage, it appeared that, if large enough direct-view storage tubes could be made to accommodate a full-size presentation of the $1\frac{1}{4}$- by 5-in. scanning field,

such tubes would permit the simplest form of receiver. The bright-trace form of direct-view storage tubes available at the outset were 5 in. in size. These tubes have a working diameter of 3.5 in., and, for full-scale reproduction of the scanned area, we required 5.2 in. Both the Radio Corporation of America and the Farnsworth Electronics Company agreed to supply experimental quantities of 7 in. tubes with a working diameter of 5.5 in. The literature should be consulted for descriptions of the functioning of this tube.[5, 6]

A second form of electric-to-visual storage tube, best known as a dark-trace cathode ray tube, has been produced in 7-in. diameter models by the Skiatron Corporation and its licensees.[7]

All these storage tubes have their relative merits and faults, but very little comparative data are available. In order to ascertain the operational characteristics of the tubes that were — or could become — available, it was determined to construct three separate display units.

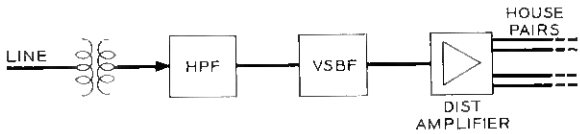


Fig. 15 — Common equipment block diagram.

5.2 *Receiving Equipment*

Since there may be several teller's display units at a single terminal location, it was logical to split the receiving equipment into two units. One unit consists of the equipment common to all of the receiving installation, in which the incoming signal is partially processed and then distributed to the several display units. This common-equipment unit is shown in Fig. 15. An isolating transformer feeds filters which complete the vestigial-sideband spectrum shaping and eliminate power frequency harmonics. The signal is then amplified for distribution. The individual display units comprise the remaining portion of the equipment.

A block diagram of the control units for each display tube is shown in Fig. 16. The received signal is processed in three branches to achieve the required functions. The first branch includes a narrow band-pass amplifier for display unit selection. The tone burst at the beginning of a message intended for this station is passed through this amplifier to initiate the vertical deflection and simultaneously open a gate in the picture signal branch. The tone burst for other display units will be

rejected by the selectivity of the narrow band-pass amplifier. Consequently, the only display unit responding to a transmission is the one to which it is addressed.

The second branch processes the received signal to provide a highly stable, continuous, horizontal sweep. The signal is envelope-detected to regain baseband, and the horizontal sync pulse is separated and stabilized by standard television methods. The stabilized sync pulse drives the horizontal sweep for the display tube.

The third branch processes the picture signal. The received signal is translated to a higher-frequency region, for more accurate envelope detection. This is accomplished by a single modulation and band-pass filtering. After suitable amplification, the high-frequency signal is limited to three levels. This limiting effectively removes the central signal portion corresponding to the bank card background, and peak-limits the signal levels denoting the pen strokes of the signature or the typed account information. The signal at the ouput of the slicer is a high-frequency carrier modulated with the picture signal. Therefore, it may be coupled easily to the writing grid of the direct-view storage tube, which is at high dc potential. The curvature of the grid characteristic of the storage tube is sufficient for demodulation of the signal. An identical slicer is used in the circuits for the dark-trace cathode ray tube. However, an additional rectifier is provided at the grid of the tube in order to gain the greater writing beam currents required by this tube for acceptable contrast of the display.

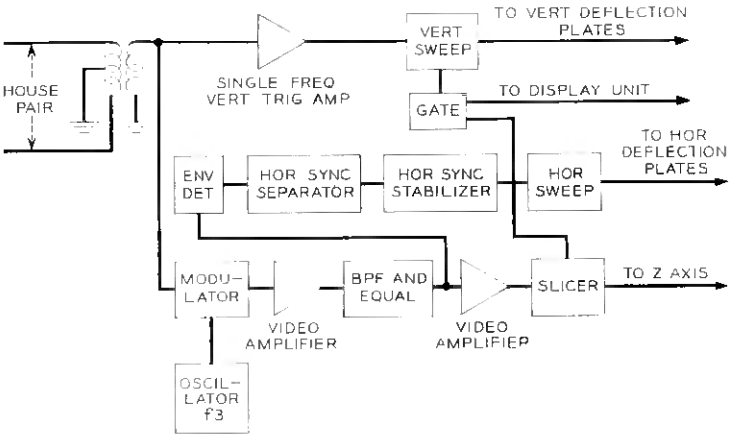


Fig. 16 — Teller's control and display unit block diagram.

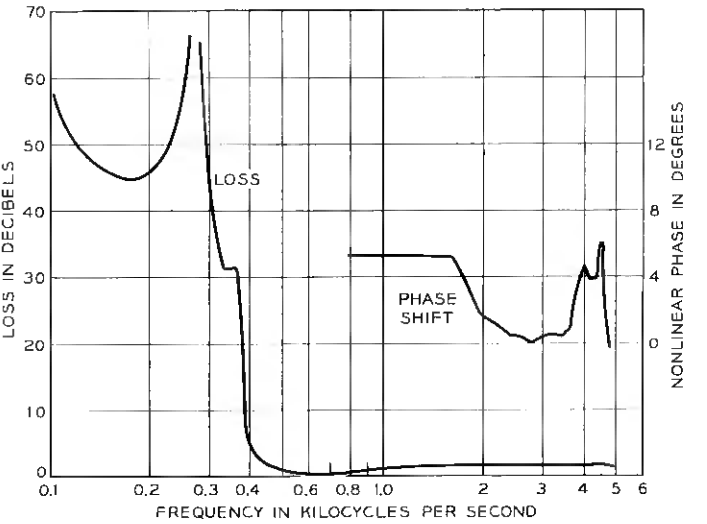


Fig. 17 — Amplitude and phase characteristics of high-pass filter of common equipment unit.

Circuits are provided to protect the storage tubes against loss of deflection voltages.

5.3 *Common Receiving Equipment*

The incoming program line is terminated in the common receiving equipment unit. This unit is intended to be installed in a location convenient to a building's telephone-distribution terminal. The received composite signals are filtered by a 400-cps high-pass filter to remove the induced 60-cps power frequency and its first six harmonics. The amplitude and phase characteristics of this filter are shown in Fig. 17. A vestigial sideband "half-filter" completes the vestigial sideband shaping and also acts to reduce noise outside the useful frequency band. The over-all characteristics of the two parts of the vestigial sideband filter (one part is included in the transmitting equipment) are shown in Fig. 18.

The filtered signal is amplified. Provision is made at the ouput of the distribution amplifier to feed as many as four balanced 600-ohm lines to as many teller's display units.

The tones for selection of a teller's display unit can be located every 75 cps from 400 to 775 cps. The principal picture sideband is 0.84 to 3.84 kc and the vestigial sideband extends to nearly 5 kc.

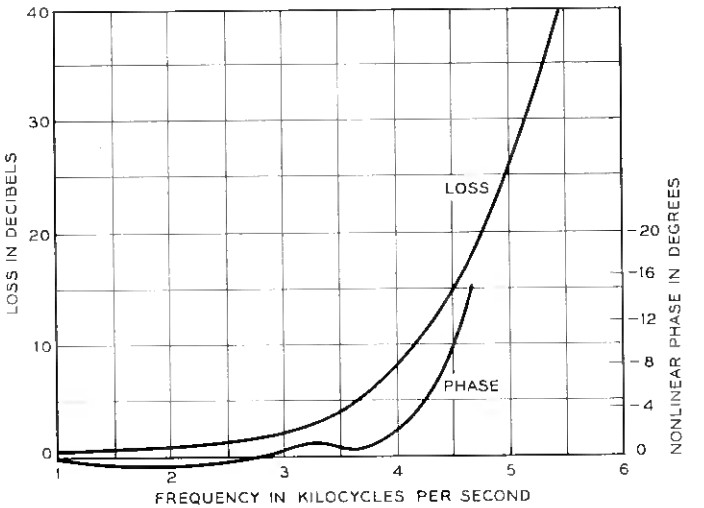


Fig. 18 — Amplitude and phase characteristics of vestigial sideband shaping in the system.

5.4 *Teller's Display Unit*

The signal from the common receiving equipment is applied to several teller's display units. The circuits for carrying out the necessary switching, writing, and display operations are described in the following paragraphs.

5.4.1 *Frame Start and Address Tone Identification*

The line-isolation transformer of Fig. 16 delivers the composite picture signal to a sharply tuned active filter. This filter will respond only to the 50-millisecond address-tone burst of the proper frequency. At the start of a transmission, the appropriate tone burst is superimposed on the composite picture signal. This address tone, as amplified and filtered, is used to trigger a single-shot vertical sweep generator and a gating circuit.

5.4.2 *Vertical Sweep Generation*

The address tone is detected by a voltage-doubler circuit but no smoothing is employed. The detected burst triggers a conventional, monostable, screen-coupled phantastron[8] (Miller run-down). The plate output of this circuit is a linearly decreasing voltage, which is an ideal generator source for a vertical deflection circuit.

5.4.3 *Gate Control of Picture Signal*

Two or more teller's display units may be used in a single installation. It is desirable, therefore, to disconnect the picture signal from all but the display tube for which the message is intended. Also, such switching avoids any possibility of damage to the storage surfaces of vertically undeflected display tubes. A suitable gating pulse is found in the screen-grid circuit of the phantastron vertical-sweep generator tube. This is a positive pulse of the same duration as the sweep voltage. The details of the gating circuit making use of this pulse are described in Section 5.4.6.

5.4.4 *Translation of Received Picture Signal from 3.84 kc to 26.88 kc*

The vestigial-sideband composite picture signal at the input of the receiver may be envelope-detected to regain the baseband frequencies; however, this will result in distortion of the high-frequency components of the picture signal, since accurate resolution of the envelope is impractical at these frequencies. In addition, the writing grid of the direct-view storage tubes is operated at a high dc potential, requiring the application of signals through a capacitor. This capacitive coupling would produce distortion of the reproduced picture if the baseband picture signal were transmitted through it. These limitations are avoided by translating the carrier before detection or display to a frequency considerably greater than the highest baseband frequency.

The picture signal is modulated on a 23.04-kc carrier in a balanced modulator to produce a double-sideband, suppressed-carrier output. The upper sideband produced by this modulation is transmitted and the lower sideband is rejected by a bandpass filter. The picture carrier at 3.84 kc is translated in frequency to 26.88 kc by this modulation process.

5.4.5 *Three-Level Limiting of the 26.88-kc Carrier Picture Signal*

In order to realize the advantages of binary transmission, it is advantageous to insert a threshold detector in the signal path. This slicer separates the signal from average-level noise and also reduces signal distortions caused by delay distortion. The additional requirement for high-frequency coupling to the grid of the direct-view storage tubes led to the adaptation of a slicer which would operate at the 26.88-kc carrier frequency and not contain baseband signal components in its output. Such a slicer circuit is shown in Fig. 19. Both positive and negative alternations of the input signal are sliced at discrete levels

between black and white. Fixed-amplitude square waves are produced at the output. The only signal-level values permitted are positive maxima, zero and negative maxima. The ternary-valued output is a completely modulated carrier. This output is suitably amplified and capacitively coupled to the control grid of the display tube. The cutoff characteristic of the grid effects half-wave envelope detection.

5.4.6 *Vertical Sweep-Period Gating of Picture Signal*

To insure no response in the unselected teller's display units, as mentioned earlier, the picture signal circuits must be gated.

During the vertical sweep period, the gating pulse available at the screen-grid of the phantastron circuit is applied to adjust the bias of tube V1 in Fig. 19, for proper slicing action. This bias value is such that, with zero signal level input, diodes D1 and D4 are conducting. The voltage difference between the anodes of D1 and D3 is set so that the input signal level corresponding to background is not sufficient to cause D3 to conduct or D1 to cease conduction. The signal level corresponding to signature information is sufficient to cause D3 to conduct on the positive peaks and D1 to cease conduction on the negative peaks. This gives rise to an output waveform at V2 restricted to positive and negative peaks and zero.

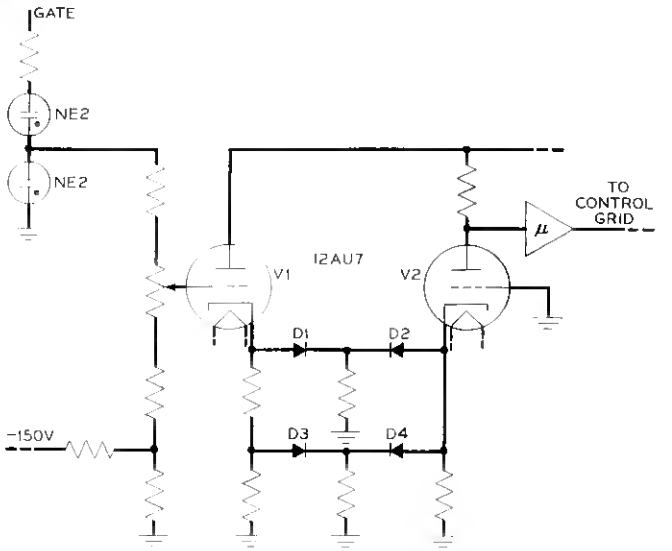


Fig. 19 — Three-level limiter schematic.

Upon completion of the vertical sweep, cessation of the positive pulse reduces the bias on V1 to a negative value and signals are unable to influence further the output of V2.

5.4.7 *Horizontal Sync Pulse Separation*

The composite picture signal, after frequency translation, is envelope-detected and the horizontal sync pulse is derived from the picture signal by the familiar amplitude-separation technique of broadcast television. After separation, the sync pulse is amplitude-limited to remove amplitude components of the transmission-line noise.

5.4.8 *Horizontal Sync Pulse Stabilization*

The limiting of the horizontal sync pulse results in a pulse with fast rise and decay times, but the timing of the leading and trailing edges is perturbed by noise. A further refinement is necessary to minimize the disturbance to sweep timing arising from this cause. The clipped horizontal sync pulses are phase-detected against locally generated pulses and the time difference between these two signals results in an error signal whose value is a function of such difference. This error signal is integrated and applied as a correcting voltage to change the phase of the local pulse source. These locally generated pulses drive the horizontal sweep generator. Therefore, short-term errors, such as are representative of noise, cause only minor perturbations of the horizontal sweep.

5.4.9 *Horizontal Sweep Generation*

The stabilized horizontal sync pulses trigger a saw-tooth generator which, in turn, drives a deflection amplifier. The average voltage with respect to the cathode ray tube gun at the output of both the horizontal and vertical deflection amplifiers is adjustable for optimum focusing of the storage tube displays. Adjustments are also provided for centering the sweeps in the display.

5.5 *Features of the Teller's Display Units*

Three separate display units were constructed for study of the different storage tubes. Each unit provides physical mounting for the tube, and for its operating potential supplies as well as switching and protection circuits.

Fig. 20 — Direct-view storage tubes.

5.5.1 *Direct-View Storage Tubes*

Fig. 20 is a photograph of the direct-view storage tubes as manufactured by (from left to right) the Farnsworth Electronics Company, the Radio Corporation of America and the National Union Electric Corporation, as a licensee of Skiatron Corporation. Fig. 21 shows a representative tube mounted in its housing. These frameworks are designed for mounting in a teller's desk. For demonstration purposes, they are fitted with covers and mounted on mobile carts.

5.5.2 *Switching Circuits*

Switching circuits are required for properly cycling the operations of the direct-view storage tubes. Some means must be provided to maintain these tubes in a state of readiness to store an incoming signal. This is achieved by maintaining the storage mesh at erase potential during any period when there is no signal stored and being viewed. Additional switching disables the phosphor-screen high voltage during all but the viewing period. One further switching circuit is provided to permit either manual erasure of the stored information or automatic erasure after a total viewing time of two minutes. The automatic feature is provided to insure readiness to store new information in the event of a teller's failure to erase a previous message.

The switching functions required for the dark-trace cathode ray tubes are somewhat different. Upon completion of the short cycle of erasure performed by heating the scotophor with an internal filament and then

permitting the surface to cool, this tube is in a condition to record another message, and will continue in such a state indefinitely without further manipulation. In order to conserve power, the high voltage supply is enabled only during the writing process, as it is only during this period that the electron gun of the tube is active. Again, both manual and automatic erasure are provided.

The erasure filament must be maintained at high voltage (13 kv) during the writing cycle for proper beam focus. Consequently, erasure is carried out by disconnecting the high voltage from the erasure filament before the heating power is applied to it. Vacuum switches were used in this equipment, but the need for switching would be eliminated if a suitable isolation transformer were used for the heating power.

5.5.3 *Tube Protection Circuits*

A circuit is provided to protect the storage tubes against excessive beam current density in the event of a horizontal sweep failure. This

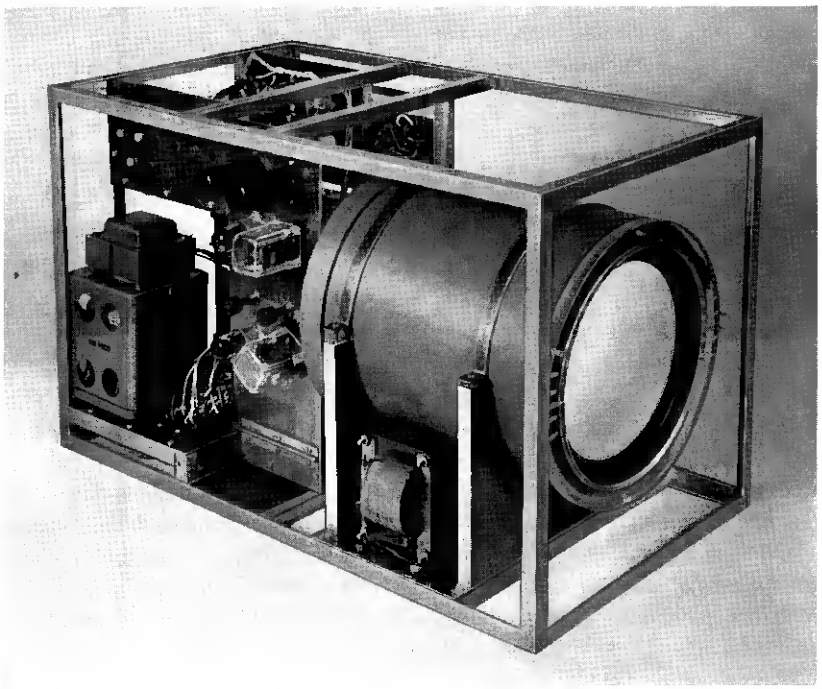


Fig. 21 — Display unit employing a Farnsworth Electronics Company 7-in. direct-view storage tube.

circuit is also self-protecting against the most common types of vacuum tube failures.

5.5.4 *Operating Potentials for Display Devices*

Potentials for the critical anodes of the direct-view storage tubes are provided through cathode followers to improve the voltage regulation. The writing gun and screen voltages are provided by separate supplies for independent operation and switching.

The dark-trace cathode ray tube may have all required potentials supplied by a single 15-kv high-voltage power pack. Separate anode potentials are derived from taps on a resistive voltage divider.

VI. EXPERIMENTAL RESULTS

By constructing separate receivers for each of the three storage tubes subjective evaluation was accomplished. All three tubes performed adequately to accomplish the desired purpose and, fortunately, there is no need to designate a "best" tube, since this equipment was not intended for Bell System manufacture.

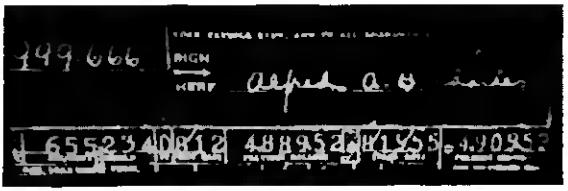
Graphic presentation of the over-all results in the form of photographs of the storage tube viewing screens is given in Fig. 22. The signals shown stored in these photographs were transmitted over a 40-mile loop of a 5-kc program channel. It is evident from Fig. 22 that the system design produced the desired results. It is also evident that all elements of the system have been taxed to their limits of performance. If the time-bandwidth product were appreciably increased the available storage tubes could not adequately display the increased information. Similarly, if the quality of the storage tubes were significantly better only marginal improvement of the system would result. Only an increase in both the time-bandwidth product — by appropriate changes in the transmitting equipment — and an improvement in the resolution and quality of the storage tube can improve the net results.

The most serious distortion of the transmitted signal resulting from transmission was found to be delay distortion. The delays encountered, however, were well within tolerable limits for program use. Noise introduced by typical long circuits was found to have little effect on the picture in the presence of the larger distortions produced by delay distortion. A series of tests was run to show the transmission distance limit of some representative forms to Bell System 5-kc program circuits without additional delay equalization.

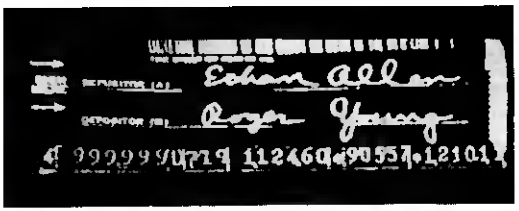
Fig. 23(a) shows the output of the transmitter to the program line

after the half-filter has partially shaped the vestigial sideband. Fig. 23(b) shows the same signal after transmission from Murray Hill to Philadelphia via New York and return over a 5-kc program circuit on an audio-frequency cable system. Fig. 23(c) shows the same signal after frequency translation and limiting. It is seen that the limiting process was capable of separating the delay distortion components from the desired signal. Fig. 23(d) is a storage-tube viewing screen photograph of a sample transmission.

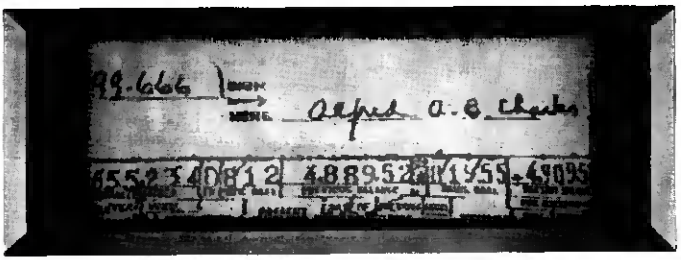
Similar data also are shown in Fig. 23 for various other loops. An audio-frequency cable channel to Washington, D. C., and return showed no loss in picture detail. The doublet and triplet output of the three-level limiter resulting from circuits with large amounts of delay distortion produce unusable results at some threshold level. All lines with less distortion produce no visible distortion in the display.



MANUFACTURER A

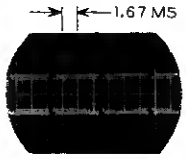


MANUFACTURER B



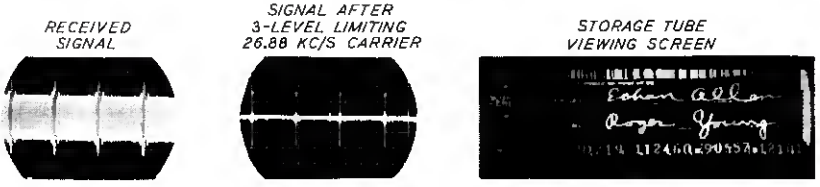
MANUFACTURER C

Fig. 22 — Photographs of storage tube viewing screens with stored signal.



(a)

PROGRAM CIRCUIT INPUT

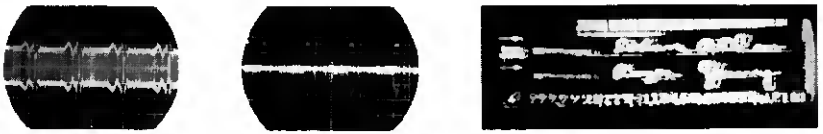


(b) (c) (d)

MURRAY HILL-PHILADELPHIA-MURRAY HILL LOOP*
B-22 CHANNEL

(e) (f) (g)

MURRAY HILL-WASHINGTON D.C.-MURRAY HILL LOOP
K-CARRIER CHANNEL



(h) (i) (j)

MURRAY HILL-CHICAGO-MURRAY HILL LOOP
B-22 CHANNEL

(k) (l) (m)

MURRAY HILL-CHICAGO-MURRAY HILL LOOP
L-CARRIER CHANNEL

*AN IDENTICAL TEST TO WASHINGTON D.C. AND RETURN OVER A B-22 CHANNEL YIELDED SIMILAR RESULTS

Fig. 23 — Results over various transmission loops.

VII. CONCLUSIONS

It has been shown that it is indeed feasible to build a visual communication system combining to advantage features of television and facsimile. While the results represent a solution of a particular communication problem, it is apparent that the techniques used may be applied to many similar problems in the business field. It should be obvious that the physical form of equipment for use in any specific application would differ in many details from that used in our experiment.

In considering the results of this experiment, the reader must keep our goals in mind. One aim was to demonstrate the use of a system intermediate between facsimile and television and providing features of both types of visual communication facility. Another aim was to demonstrate that, for some purposes, the product of transmission bandwidth and time could be reduced to a minimum without impairing the readability of the reproduced image. It is obvious that the resulting images have very little esthetic appeal; the use of a larger over-all time-bandwidth product coupled with display devices capable of increased resolution would result in more pleasing images.

VIII. ACKNOWLEDGMENTS

It is impossible to give individual credit to all of our associates who have contributed to this experiment. We do wish, however, to express our appreciation of the enthusiastic liaison carried on with the New York Savings Bank by M. B. Long, since retired from the Laboratories; and to Albert F. Kendall of the New York Savings Bank for his patient explanations of the workings of a banking house. Furthermore, we are indebted to John W. Smith and his associates of the Hogan Laboratories for their cooperation in producing the mechanical scanner employed in our experiment.

REFERENCES

1. Christopher, H. N., unpublished memorandum.
2. Baldwin, M. W. Subjective Sharpness of Simulated Television Images, B.S.T.J., **19,** October 1940, pp. 563–586.
3. Horton, A. W., Jr. and Vaughan, H. E., Transmission of Digital Information over Telephone Circuits, B.S.T.J., **34,** May 1955, pp. 511–528.
4. Bliss, W. H. and Young, C. J., Facsimile Scanning by Cathode Ray Tubes, R.C.A. Rev., **15,** September 1954, pp. 275–290.
5. Knoll, M. and Kazan, B., *Storage Tubes*, John Wiley & Sons, New York, 1952, Part V, p. 66.
6. Knoll, M., Rudnick, P. and Hook, H., Viewing Storage Tube with Half-tone Display, R.C.A. Rev., **14,** December 1953, p. 492.
7. Nottingham, W. B., *Cathode Ray Displays*, McGraw-Hill Book Co., New York, 1948, Ch. 18.
8. Chance, B., Hughes, V., et al, *Waveforms*, McGraw-Hill Book Co., New York, 1949, p. 197.

The Z Transformation

By H. A. HELM

(Manuscript received April 15, 1958)

The Stieltjes integral is used to develop a rigorous derivation of the z transform. Sufficient properties of the transformation are included to form a reasonably complete basis for the operational solution of constant coefficient, linear, finite difference equations.

I. INTRODUCTION

The desire to use digital computers in automatic control loops created the need for methods with which to analyze systems that are partly continuous and partly discrete. Since the methods of network theory could be applied to the analysis of the continuous part of such a hybrid system, it was natural that such methods should be extended to include the discrete case. This resulted in the z transform introduced by Raggazini and Zadeh.[1] There is today an extensive literature devoted to the z transform.[2, 3, 4] However, the fundamental assumption of the z transform derivation is that the process of instantaneous sampling is equivalent to the amplitude modulation of a train of unit impulses by the "sampled" function. But the unit impulse as commonly defined has infinite height and zero width, and the process of amplitude modulating such a function is not intuitively clear. While it is true that such a process may be considered as an approximation to the behavior of a linear network with an amplifier and sampling switch, "impulse sampling" bears no simple relation to the manner in which the digital computer operates. The digital computer, in the type of real time operation typical of control system applications, works with sequences of numbers which represent a continuous function evaluated at particular instance of time. Since these numbers must of necessity be finite, "impulse sampling" is not an obvious mathematical model for describing the working of the computer. It is the intention of this paper to define the problem from the point of view of operations within the computer and to develop a rigorous and appealing derivation of the z transform.

In place of impulse modulation, the alternate approach is taken

of generalizing the definition of the Laplace transformation by means of the Stieltjes integral. This approach has the advantage of rigor and of more closely relating the operational solutions of continuous and discrete systems. As developed below, only rational transforms are considered. Also, in general, functions involving derivatives of impulses cannot be represented by the Stieltjes integral. Practically, these restrictions do not limit the applications greatly. Derivation of the principal properties of the z transform, based upon the Stieltjes integral definition, are given in the Appendix.

II. THE LAPLACE-STIELTJES TRANSFORMATION

The Stieltjes integral has the important property of including both sums and limits of sums (integrals). For the reader's convenience the definition of the Stieltjes integral as given in Widder[5] is repeated.

Let an interval (a, b) be divided into sub-intervals in the following manner:

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

and let Δ equal the largest of these subintervals. Then the Stieltjes integral of $f(x)$ with respect to $\alpha(x)$ from a to b is

$$\int_a^b f(x)\, d\alpha(x) = \lim_{\Delta \to 0} \sum_{k=1}^{n} f(\zeta_k)[\alpha(x_k) - \alpha(x_{k-1})], \tag{1}$$

where $x_{k-1} \leqq \zeta_k \leqq x_k$. The left-hand side of (1) is the usual notation for a Stieltjes integral. The integral itself is defined only when the limit on the right exists. It can be shown[5] that the integral exists if $f(x)$ is continuous and if $\alpha(x)$ is monotonic but not necessarily continuous, i.e., nonincreasing or nondecreasing. We shall assume that both these conditions apply to all functions to be considered. However, these conditions are quite strong and will be somewhat relaxed subsequently.

It is now possible to generalize the Laplace transformation by making the defining integral a Stieltjes integral. Thus,

$$L_s[f(t)] = \int_0^\infty f(t)e^{-st}\, d\alpha(t), \tag{2}$$

where

$$s = \sigma + j\omega.$$

It is assumed that (2) is subject to all the above restrictions. As defined by (2) the Laplace-Stieltjes transformation actually defines a different

transformation for each different selection of the function $\alpha(t)$. If $\alpha(t) = t$, (2) reduces to the usual definition of the Laplace transform. If $\alpha(t)$ is continuous and has a continuous derivative, (2) reduces to the Laplace transform of $\alpha'(t)f(t)$, which may be handled by the usual theorems of the Laplace transform. However, if $\alpha(t)$ is not continuous a new class of transformations results. For the purpose of this paper, the function $\alpha(t)$ will be defined as the "staircase" function which increases by unity at integral multiples of T but remains constant between such points. This function is shown in Fig. 1. The constant T is equivalent to the sampling period of modulation theory. From this point on, the function $\alpha(t)$ will be assumed to be that of Fig. 1. Thus, (2) may be evaluated for this $\alpha(t)$ by means of (1) as

$$\int_0^\infty f(t)e^{-st}\,d\alpha(t) = \sum_{n=0}^{\infty} f(nT)e^{-nTs} = F(e^{sT}), \tag{3}$$

where $\alpha(t)$ is given by Fig. 1, $s = \sigma + j\omega$ and $f(nT)$ is the function $f(t)$ evaluated at $t = nT$. Since $\alpha(t)$ is monotonic, and we have restricted $f(t)$ to continuous functions, (2) clearly *exists*. This does not imply that the series on the right converges or diverges. Since no part of $f(t)$ be-

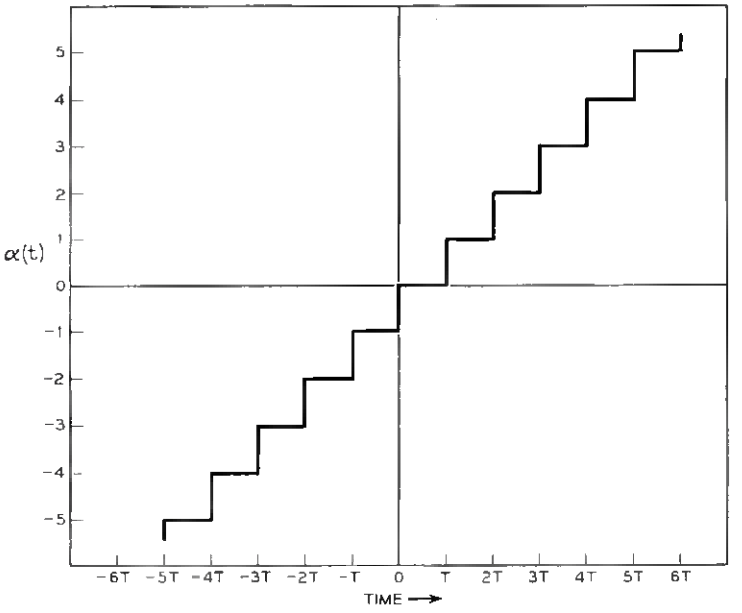


Fig. 1 — "Staircase" function $\alpha(t)$.

tween multiples of T affects (3), $f(t)$ need only be continuous and well defined in the neighborhood of the points $t = nT$; i.e., have no discontinuities or "jumps" at multiples of T.

A simple change of variable in (3) introduces the z transform. Let $z = e^{sT}$ and substitute in (3):

$$L_s[f(t)] = \int_0^\infty f(t)z^{-t/T}\, d\alpha(t) = \sum_{n=0}^{\infty} f(nT)z^{-n} = f(z). \tag{4}$$

The expression (4) emphasizes the power series nature of the z transform. Since the functions with which we shall deal are convergent, they can almost always be written in closed form. In fact, to allow the order of certain limit functions to be interchanged in the development of various theorems, absolute convergence of (4) will be assumed. That is,

$$\sum_{n=0}^{\infty} | f(nT)z^{-n} | \leqq M, \tag{5}$$

where M is finite although possibly very large. The magnitude of z is $e^{-nT\sigma}$, and if (5) holds for some σ, say σ_0, it will obviously hold for any $\sigma > \sigma_0$. In the following sections, the functions $f(t)$ are now restricted to those functions for which (5) holds. Rewriting (5) as a Stieltjes integral leads to:

$$\int_0^\infty | f(t) | e^{-\sigma t}\, d\alpha(t) \leqq M, \tag{5a}$$

which holds for the Laplace transform when $\alpha(t) = t$.[6] The restriction of absolute convergence has little practical effect upon the utility of the transform.

III. THE INVERSE TRANSFORM

An inverse transform is necessary for operational completeness. The derivation is straightforward, making use of Cauchy's method for evaluating the coefficients of a power series.[7] The proof proceeds from the definition of the z transform in (4):

$$I_s[f(t)] = \sum_{n=0}^{\infty} f(nT)z^{-n} = F(z).$$

Expanding,

$$\begin{aligned} f(z) = f(0) + f(T)z^{-1} + \cdots + f(nT - T)z^{-(n-1)} + f(nT)z^{-n} \\ + f(nT + T)z^{-(n+1)} + \cdots, \end{aligned} \tag{6}$$

which is absolutely convergent for $|z| \geqq e^{\sigma_0 T}$. Since $F(z)$ is a power series, absolute convergence also implies uniform convergence within the radius of convergence. Hence, (6) may be integrated term by term along a contour such that $|z| > e^{\sigma_0 T}$. Multiplying by z^{n-1} and so integrating gives:

$$\begin{aligned}\int_\Gamma z^{n-1}F(z)\,dz = \int_\Gamma f(0)z^{n-1}\,dz + \int_\Gamma f(T)z^{n-2}\,dz + \cdots \\ + \int_\Gamma f(nT - T)\,dz + \int_\Gamma f(nT)z^{-1}\,dz + \cdots \qquad (7) \\ + \int_\Gamma f(nT + mT)z^{-(m+1)} \cdots,\end{aligned}$$

where $m = 1, 2, 3, \cdots$ and, Γ the contour of integration, is a circle enclosing the origin of the z plane and whose radius is greater than $e^{\sigma_0 T}$. It is obvious that all integrals except $\int_\Gamma f(nT)z^{-1}\,dz$ either have no singularity within the contour of integration, and hence are zero, or else have a pole at the origin of order greater than unity, and hence also are zero. Therefore, (7) reduces to

$$\int_\Gamma z^{n-1}F(z)\,dz = f(nT)\int_\Gamma z^{-1}\,dz, \qquad (8)$$

since $f(nT)$ is the function $f(t)$ evaluated at $t = nT$, and hence a constant. Thus, the inversion formula becomes

$$f(nT) = \frac{1}{2\pi j}\int_\Gamma z^{n-1}F(z)\,dz. \qquad (9)$$

It is necessary to add a word of caution here. Equation (9) only gives the value of the function at *one* point, $t = nT$. Obviously, by assigning any given integer to n the value at any point may be obtained and, usually, this is sufficient. However, in the case of certain summations this will lead to a very confusing notation. To avoid this, we introduce the notation

$$\{f(nT)\}_m$$

to indicate the sequence

$$f(0), f(T), f(2T) \cdots, f(nT), \cdots, f(mT).$$

Omission of the subscript m will denote an infinite sequence.

It is now possible to summarize the Laplace-Stieltjes pairs:

$$L_s[f(t)] = \int_0^\infty f(t)e^{-st}\,d\alpha(t) = F(e^{sT}) \tag{3}$$

or

$$L_s[f(t)] = \int_0^\infty f(t)z^{-t/T}\,d\alpha(t) = F(z), \tag{3a}$$

$$L_s^{-1}[F(z)] = \frac{1}{2\pi j}\int_\Gamma z^{n-1}F(z)\,dz = f(nT). \tag{9}$$

A closer inspection of the above pairs indicates a very interesting property of the Laplace-Stieltjes transformation. A continuous function $f(t)$ is transformed and then the inverse operation performed. However, the function is then only defined at integral multiples of T, i.e., at $t = nT$. This is exactly equivalent to "sampling"; that is, the computer, by means of some encoding device, evaluates instantly a continuous function of time (commonly represented by voltages, shaft positions, etc.) at periodic intervals. The L_s transformation is a mathematical model or abstraction which represents this process. It is very often the case that the function which is to be sampled (a voltage, for instance) is applied to a linear network and then sampled, and it would be very convenient to be able to analyze the system directly from the Laplace transform. Such results follow from the relationships between the Laplace and Laplace-Stieltjes transforms which we develop below.

IV. THE RELATIONSHIP BETWEEN THE LAPLACE AND LAPLACE-STIELTJES TRANSFORMATIONS

The transform pairs (3) or (3a) and (9) are sufficient to derive operational methods for discrete linear systems which are similar to those of continuous linear systems. However, there are several reasons for examining the relationship between the two transforms. In the first place, many of the most interesting problems arise from the analysis of systems which are partly continuous and partly discrete. Also, the relationship between the Laplace-Stieltjes and Laplace transforms may be expressed as a convolution integral, which historically was used first in the study of these systems and which is still a very handy computational formula.

Returning to (4), we have

$$L_s[f(t)] = \sum_{n=0}^{\infty} f(nT)e^{-nTs}. \tag{4}$$

The inverse of the normal Laplace transform, with the simple change of variable $t = nT$ and using p in place of the usual s, is

$$f(nT) = \frac{1}{2\pi j}\int_{\Gamma} F(p)e^{pnT}\,dp, \tag{10}$$

where Γ, the contour of integration, encloses the poles of $F(p)$, which is assumed rational. Substitution of (10) into (4) gives

$$\begin{aligned} L_s[f(t)] &= \sum_{n=0}^{\infty} e^{-nTs}\frac{1}{2\pi j}\int_{\Gamma} F(p)e^{pnT}\,dp \\ &= \frac{1}{2\pi j}\sum_{n=0}^{\infty}\int_{\Gamma} F(p)e^{(p-s)nT}\,dp. \end{aligned} \tag{11}$$

Since the series (4) and the integral are both uniformly convergent, the order of summation and integration may be interchanged:

$$L_s[f(t)] = \frac{1}{2\pi j}\int_{\Gamma} F(p)\sum_{n=0}^{\infty} e^{(p-s)nT}\,dp. \tag{12}$$

If $|e^{(p-s)T}| < 1$, we may sum the geometric series in (12) to

$$\sum_{n=0}^{\infty} e^{(p-s)nT} = \frac{1}{1 - e^{(p-s)T}}, \tag{13}$$

where the condition that this summation be valid is that Re $p <$ Re s. Thus, (12) becomes

$$L_s[f(t)] = \frac{1}{2\pi j}\int_{\Gamma}\frac{F(p)}{1 - e^{(p-s)T}}\,dp. \tag{14}$$

The contour of integration, Γ, is the usual one parallel to the imaginary axis extending from $-j\infty$ to $+j\infty$ to include all possible poles. However, since we do not wish to exclude functions which do not vanish as s(or p) approaches infinity [in particular, $F(p) =$ constant] the contour is closed by an infinite semicircle to the left from $+j\infty$ to $-j\infty$. The requirement that Re $p <$ Re s is equivalent to stating that Γ shall include the poles fo $F(p)$ but exclude the poles of

$$\frac{1}{1 - e^{(p-s)T}}.$$

If the inverse Laplace transform of

$$F(s) = \frac{1}{1 - e^{-sT}}$$

be interpreted as the sum of a sequence of unit impulses, $\delta_T(t)$, a distance T apart, then the amplitude modulation of $\delta_T(t)$ by some function $f(t)$ may be found, from the complex convolution formula of Laplace transform theory, to be:

$$L[f(t)\delta_T(t)] = \frac{1}{2\pi j}\int_\Gamma \frac{F(p)}{1 - e^{(p-s)T}}\,dp,$$

which is, of course, the Laplace-Stieltjes transform of $f(t)$ as given by (14). Hence, the Laplace-Stieltjes transform is formally equivalent to the results of impulse modulation in the sense that both lead to the same transform. However, the definition of the Laplace-Stieltjes transform is rigorous and it directly relates discrete and continuous systems. Intuitively, one would expect that, as the interval between samples approaches zero, the Laplace-Stieltjes transform should approach the Laplace; i.e., the discrete system should look more like the continuous. This follows from the definition of the Laplace-Stieltjes transform. Note that, for the $\alpha(t)$ of Fig. 1,

$$\lim_{T\to 0} T\alpha(t) = t.$$

Thus,

$$\lim_{T\to 0} TL_s[f(t)] = \lim_{T\to 0}\int_0^\infty f(t)e^{-st}\,dT\alpha(t)$$

$$= \int_0^\infty f(t)e^{-st}\,dt = L[f(t)],$$

which is the desired relation, the Laplace-Stieltjes transformation approaching the Laplace in the same manner as the staircase distribution function $T\alpha(t)$ approaches the straight line t.

Equation (14) is a very useful computational tool and can be used to prepare a table of Laplace-Stieltjes transforms from the common tables of Laplace transforms. Some elementary functions are given in Table I, where a direct comparison between the two transforms can be made, the Laplace-Stieltjes transform being written in the e^{sT} form. Such a comparison is interesting, but the relationship between the two transforms can be better shown by a closer examination of the transforms of some elementary functions. Consider first $f(t) = e^{-\alpha t}$:

$$L(e^{-\alpha t}) = \frac{1}{s + \alpha} = F(s), \tag{15}$$

$$L_s(e^{-\alpha t}) = \frac{e^{sT}}{e^{sT} - e^{-\alpha T}} = F(e^{sT}). \tag{16}$$

The single pole at $s = -\alpha$ of $F(s)$ is shown in Fig. 2(a) in the usual manner. However, $F(e^{sT})$ has an infinite number of singularities occurring at $s = -\alpha \pm 2\pi n/T (n = 0, 1, 2, \cdots)$. Thus, the effect of sampling is to

multiply the single real pole of the L transform into an infinite number of complex poles as shown in Fig. 2(b). The case of imaginary roots in the s plane is even more interesting. Let $f(t) = \sin \beta t$

$$L[\sin \beta t] = \frac{\beta}{s^2 + \beta^2} = F(s), \tag{17}$$

$$L_s[\sin \beta t] = \frac{e^{sT} \sin \beta T}{e^{2sT} - 2e^{sT} \cos \beta T + 1}. \tag{18}$$

The singularities of $F(s)$ are shown in Fig. 3(a) and those of $F(e^{sT})$ in Fig. 3(b). Here, the sampling multiplies the original *pair* of poles into an infinite number of such pairs. The center of each pair is separated from the next by a distance of $2\pi/T$. The distance or period T may be identified with a radian sampling frequency $\omega_s = 2\pi/T$. From Fig. 3(b) it is apparent that, if $\beta \geqq \omega_s/2$, the pairs of poles overlap each other and form a new configuration which is indistinguishable from a configuration resulting from some function of radian frequency less than $\omega_s/2$. This result also follows from Shannon's sampling theorem which, in effect, states that, if the original signal is to be recovered after sampling, then the sampling frequency must be greater than twice the highest frequency sampled.

If the L transformation of a function has a singularity in the right half of the s plane, the L_s transformation will have an infinite number in the right half plane.

TABLE 1

$f(t)$	$F(s)$	$f(nT)$	$F(e^{sT})$
$u(t)$	$\frac{1}{s}$	$u(nT)$	$\frac{e^{sT}}{e^{sT} - 1}$
$e^{-\alpha t}$	$\frac{1}{s + \alpha}$	$e^{-\alpha nT}$	$\frac{e^{sT}}{e^{sT} - e^{-\alpha T}}$
$\sin \beta t$	$\frac{\beta}{s^2 + \beta^2}$	$\sin \beta nT$	$\frac{e^{sT} \sin \beta T}{e^{2sT} - 2e^{sT} \cos \beta T + 1}$
$e^{-\alpha t} \sin \beta t$	$\frac{\beta}{(s + \alpha)^2 + \beta^2}$	$e^{-\alpha nT} \sin \beta nT$	$\frac{e^{(s+\alpha)T} \sin \beta T}{e^{2(s+\alpha)T} - 2e^{(s+\alpha)T} \cos \beta T + 1}$
$e^{-\alpha t} \cos \beta t$	$\frac{s}{(s + \alpha)^2 + \beta^2}$	$e^{-\alpha nT} \cos \beta nT$	$\frac{(e^{(s+\alpha)T} - \cos \beta T)e^{(s+\alpha)T}}{e^{2(s+\alpha)T} - 2e^{(s+\alpha)T} \cos \beta T + 1}$

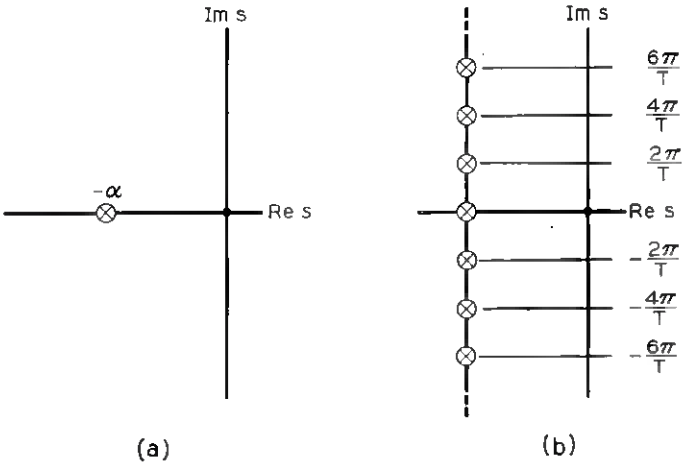


Fig. 2 — (a) Single S plane pole of $F(s) = 1/(s + \alpha)$; (b) infinite number of complex S plane poles of $F(e^{sT}) = e^{sT}/(e^{sT} - e^{-\alpha T})$.

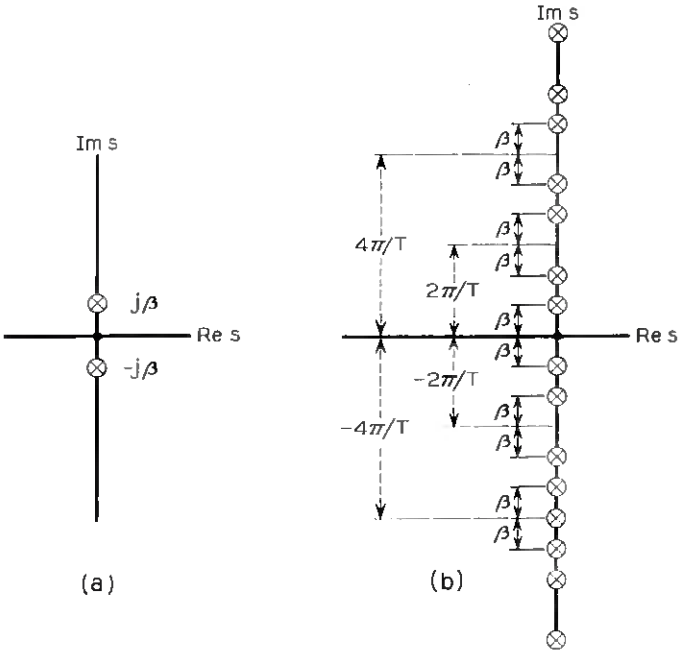


Fig. 3 — Pair of S plane poles of $F(s) = \beta/(s^2 + \beta^2)$; (b) infinite number of S plane pole pairs of $F(e^{sT}) = e^{sT} \sin \beta T/(e^{2sT} - 2e^{sT} \cos \beta T + 1)$.

The change of variable, $z = e^{sT}$, which introduces the z transform, simplifies the above mappings. It is well known that the transformation, $z = e^{sT}$, maps the imaginary axis of the complex s plane into the unit circle about the origin in the z plane. The left half of the s plane maps wholely within the unit circle, the right half maps exterior to it. Since the z plane mapping is repetitive for multiples of $2\pi/T$, the infinite number of roots and root-pairs in the s plane map into single roots and root-pairs in the z plane.

V. APPLICATION OF THE L_s TRANSFORMATION TO THE SOLUTION OF FINITE DIFFERENCE EQUATIONS

Many of the concepts of sampling can be applied to the solution of linear *finite difference* equations, with constant coefficients. These equations are simply linear combinations of sequences of numbers shifted forward and backward in time by integral multiples of some fixed interval. In the case of a digital computer operating in a control loop, the sequences are actually generated by sampling some continuous function of time. If the assumption is made that all the sequences in a finite difference equation result from such sampling, then the L_s transformation offers a very useful method for the operational solution of such an equation. The resulting solution is the "smooth" curve which, when sampled, will give the sequence satisfying the difference equation.

A finite difference may be defined in either of two ways. One could be called a *backward* difference, defined as

$$\Delta_b\{y(nT)\} = \{y(nT)\} - \{y(nT - T)\}. \tag{19}$$

That is, the backward difference is simply the difference of two sequences, one of which is the other shifted backward one interval in time. Since there is no possibility of ambiguity, the braces may be omitted and (19) written in the more usual form

$$\Delta_b y_{nT} = y_{nT} - y_{nT-T}\,. \tag{19}$$

In similar fashion, the *forward* difference may be written as:

$$\Delta_f y_{nT} = y_{nT+T} - y_{nT}\,. \tag{20}$$

Higher differences of course are formed by taking "differences of differences." That is,

$$\Delta_b{}^n y_{nT} = \Delta_b{}^{n-1} y_{nT} - \Delta_b{}^{n-1} y_{nT-T}$$

or

$$\Delta_f{}^n y_{nT} = \Delta_f{}^{n-1} y_{nT+T} - \Delta_f{}^{n-1} y_{nT}\,.$$

In order to eliminate possible confusion between differences, a difference equation can always be expanded and written in the ordinate form:

$$\sum_{j=1}^{n} b_j y_{nT+jT} + \sum_{i=0}^{m} a_i y_{nT-iT} = x_{nT}.$$

If the assumption is made that the sequences $\{y(nT)\}$, $\{x(nT)\}$ result from a sampling of some continuous function which has an L_s transform, then the finite difference equations of the above form are readily solved by the L_s transformation. This follows from the results of Property I of the L_s transform, which is proved in the Appendix. The property is repeated below without proof.

Property I: If $L_s[f(t)] = F(z)$ *and* a *is a nonnegative integer, then:*

$$\begin{aligned} L_s[f(t - aT)] &= L_s[\{f(nT - aT)\}] \\ &= z^{-a}\left[F(z) + \sum_{m=1}^{a} f(-mT)z^m\right] \end{aligned} \tag{21}$$

and

$$\begin{aligned} L_s[f(t + aT)] &= L_s[\{f(nT + aT)\}] \\ &= z^{a}\left[F(z) - \sum_{m=0}^{a-1} f(mT)z^{-m}\right]. \end{aligned} \tag{22}$$

Application of (21) to the *backward* difference equation (14) leads to

$$L_s[\Delta_b Y_{nT}] = \frac{z - 1}{z} Y(z) - y(-T) \tag{23}$$

and to the forward difference

$$L_s[\Delta_f t_{nT}] = (z - 1)Y(z) - zy(0). \tag{24}$$

The terms $y(0)$ in (23) and $y(-T)$ in (24) are the usual initial conditions, and allow the specification of arbitrary boundary conditions in a manner completely analogous to the insertion of initial conditions in the solution of differential equations. However, for simplicity, zero initial conditions will be assumed in the problem below.

As an example, the above can be applied to the simultaneous difference equations:

$$\begin{aligned} 0.25y_{n-1} + w_n &= x_n, \\ -1.0y_n - 0.25y_{n-1} + w_n + 0.5w_{n-1} &= 0, \end{aligned} \tag{25}$$

with T taken as unity for convenience. Letting

$$Y(z) = L_s[y_n],$$
$$W(z) = L_s[w_n],$$
$$X(z) = L_s[x_n],$$

and performing the indicated operation on (25) leads to:

$$\begin{aligned} 0.25Y(z) + zW(z) &= zX(z), \\ -(z + 0.25)Y(z) + (z + 0.5)W(z) &= 0, \end{aligned} \tag{26}$$

whence

$$W(z) = \frac{z(z + 0.25)X(z)}{z^2 + 0.5z + 0.125}. \tag{27}$$

As a "test function" let $x(n)$ be the *unit sample* defined as unity for $n = 0$ and zero elsewhere. The L_s transform of x_n is then $L_s[x_n] = 1$, and it follows that

$$W(z) = \frac{z(z + 0.25)}{[(z + 0.25)^2 + .0625]}. \tag{28}$$

By the application of (9),

$$w_n = 4(0.35)^{n+1} \sin (n + 1)\frac{3\pi}{4} + (0.35)^n \sin n\frac{3\pi}{4}. \tag{29}$$

In a similar manner w_n(or y_n) may be determined for any function x_n which is L_s-transformable.

In electrical network theory based upon the Laplace transform, the weighting function of a network is the time response of that network to a unit *impulse*. The analogy to the above response of a difference equation to a unit sample is clear. The fact that the difference equations could describe the operation of a digital computer performing linear operations in real time lends physical reality to the analogy and introduces the concept of a digital network. The designer of a servomechanism or feedback amplifier is concerned that his device shall be stable. The designer of a program for a digital computer is likewise concerned that his machine behave in a stable manner. Here, stability is defined in the sense of the electronic network designer that, for any bounded input, the output shall not continually increase. This is more elegantly and precisely stated in terms of the complex s plane. Here, the criterion for stability is that the characteristic equation have no roots in the right half of the s plane. Since the defining transformation $z = e^{sT}$ of the z plane maps the right

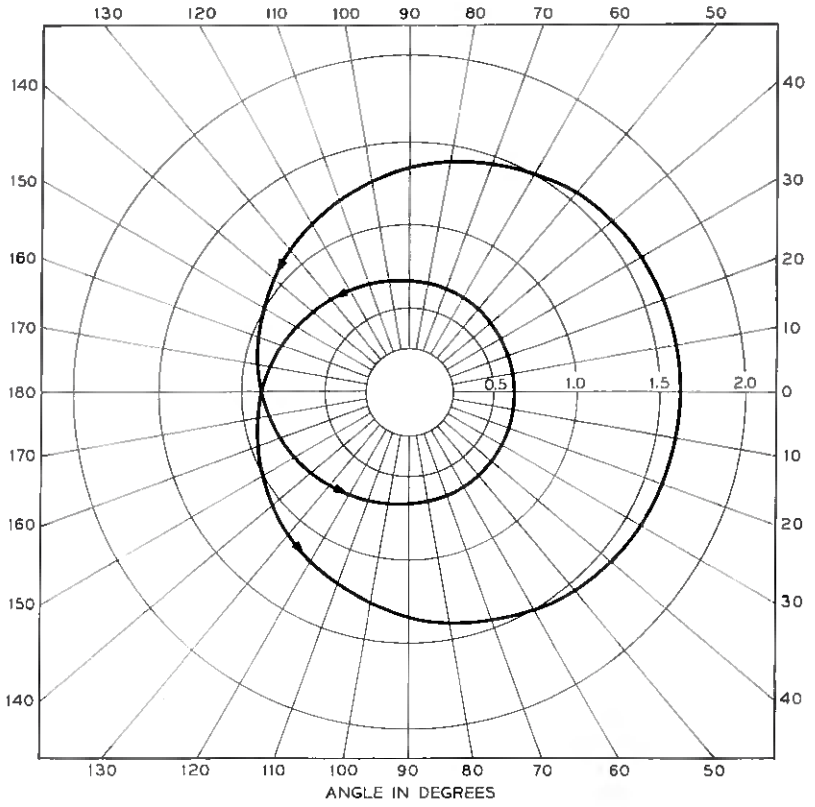


Fig. 4 — Nyquist plot of $z^2 + 0.5z + 0.125$.

half of the s plane onto the exterior of the unit circle, the criterion for stability of finite difference equations is that the characteristic equation shall have no roots exterior to the unit circle in the z plane. Intuitively, it would seem that Nyquist's criterion could be applied in one form or another directly in the z plane to determine stability. However, the application is not as attractive as in the continuous case. The finite difference equations (25) will be used to illustrate the point. As above, the roots of the denominator of (27), the L_s transformation of (25), determine the stability of the difference equations. Since all roots do indeed lie within the unit circle, as z takes on values along the unit circle in the *positive sense*, the plot of $z^2 + 0.5z + 0.125$ will encircle the origin in the *positive sense* a number of times equal to the number of zeros of the polynomial which lie within the unit circle (two in this case). Fig. 4 is such a plot and it is readily apparent that there are two encirclements.

In principle, the method can be extended to a polynomial of any order. As a practical matter, counting the number of encirclements from a polynomial of high order without making a mistake would be difficult. However, this is not the most important point. One of the great attractions of the Nyquist criterion to the practical designer of feedback devices is that not only does it determine stability but it also indicates at a glance the margins against instability. To know the allowable variation in the gain of an amplifier, as a specific example, is of great value to the designer (and manufacturer) of a feedback amplifier. Unfortunately, in the discrete case no such information is apparent and the indication is simply one of "go" or "no go".

If the change of variable $w = z^{-1} = e^{-sT}$ is made, the *right* half of the s plane is now mapped into the interior of the unit circle, and hence the criterion for stability becomes the more usual one of not enclosing the origin. Illustrating, the characteristic equation of (27) becomes;

$$w^{-2} + 0.5w^{-1} + 0.125. \tag{30}$$

As w takes on values on the unit circle in the positive sense, (30) has exactly the same values as shown in Fig. 4 with the exception that the curve now encloses the origin in the *negative* sense and, hence, stability depends also upon the *sense* of enclosure. The difficulties above will usually require that stability analysis still be made in the s plane. However, some advantage can be taken of the angle preserving properties of the change of variable $s = 1/T \ln z$.

VI. CONCLUSION

The Laplace-Stieltjes derivation of the z transform is straightforward and rigorous. As a mathematical model of the sampling process it has the advantage of also describing some of the operations possible within the computer. In particular, since it can be used as a basis for the operational solution of linear finite difference equations it closely relates the solution of discrete linear systems and continuous linear systems. It is of considerable advantage to be able to apply the methods of network analysis to this type of computer operation. The Laplace-Stieltjes transformation forms the connection in a very clear manner.

VII. ACKNOWLEDGMENT

The author would like to express his sincere appreciation to J. G. Tryon of Bell Telephone Laboratories for his constant encouragement

and help and to N. J. Rose of the Stevens Institute of Technology for the proof of the property of the transform for a finite sum.

APPENDIX

PROPERTIES OF THE L_s TRANSFORMATION

The utility of the L_s transform is increased by various properties pertaining to its use. The more important of these properties are derived below with some discussion of the area of application. Such discussion must of necessity be brief.

Real Translation

This property is the basis upon which the operational solution of linear finite difference equations is based.

Property I: If $L_s[f(t)] = F(z)$ and a is a nonnegative integer, then:

$$L_s[f(t - aT)] = z^{-a}\left[F(z) + \sum_{m=1}^{a} f(-mT)z^m\right],$$

and

$$L_s[f(t + aT)] = z^{a}\left[F(z) - \sum_{m=0}^{a-1} f(mT)z^{-m}\right].$$

The proof of the first part follows:

By definition

$$L_s[f(\tau)] = \int_0^\infty f(\tau)e^{-s\tau}\, d\alpha(\tau) = F(e^{sT}).$$

Dividing the range of integration into two parts:

$$F(e^{sT}) = \int_{-aT}^{\infty} f(\tau)e^{-s\tau}\, d\alpha(\tau) - \int_{-aT}^{0-} f(\tau)e^{-s\tau}\, d\alpha(\tau).$$

We now let $\tau = t - aT$ in the first integral on the right-hand side:

$$F(e^{sT}) = \int_0^\infty f(t - aT)e^{-s(t-aT)}\, d\alpha(t - aT) - \int_{-aT}^{0-} f(\tau)e^{-s\tau}\, d\alpha(\tau). \quad (31)$$

From Fig. 1 it is apparent that

$$d\alpha(t - aT) = d\alpha(t).$$

Hence,

$$F(e^{sT}) = e^{aTs}\int_0^\infty f(t - aT)e^{-st}\, d\alpha(t) - \int_{-aT}^{0-} f(\tau)e^{-s\tau}\, d\alpha(\tau). \quad (32)$$

But the second integral may be evaluated as:

$$\int_{-aT}^{0-} f(\tau)e^{-s\tau}\,d\alpha(\tau) = \sum_{m=1}^{a} f(-mT)e^{mTs}. \tag{33}$$

Substitution of (33) into (32) and rearrangement leads to

$$e^{-aTs}\left[F(e^{sT}) + \sum_{m=1}^{a} f(-mT)e^{mTs}\right] = \int_{0}^{\infty} f(t - aT)e^{-st}\,d\alpha(t).$$

But the right-hand side of the above is by definition $L_s[f(t - aT)]$. Making the usual substitution of $z = e^{sT}$ now leads to the desired result:

$$L_s[f(t - aT)] = z^{-a}\left[F(z) + \sum_{m=1}^{a} f(-mT)z^{m}\right]. \tag{34}$$

We note in particular that, for $a = 1$,

$$L_s[f(t - aT)] = z^{-1}[F(z) + f(-T)].$$

For proof of the second part, the range of integration is divided into two parts:

$$L_s[f(\tau)] = \int_{aT-}^{\infty} f(\tau)e^{-s\tau}\,d\alpha(\tau) + \int_{0}^{(a-1)T+} f(\tau)e^{-s\tau}\,d\alpha(\tau).$$

Letting $\tau = t + aT$ in the first integral on the right-hand side, we have

$$F(e^{sT}) = \int_{0-}^{\infty} f(t + aT)e^{-s(t+aT)}\,d\alpha(t + aT) + \int_{0}^{(a-1)T+} f(\tau)e^{-s\tau}\,d\alpha(\tau).$$

Rearranging and substituting $d\alpha(t + aT) = d\alpha(t)$,

$$e^{aTs}\left[F(e^{sT}) - \int_{0}^{(a-1)T+} f(\tau)e^{-st}\,d\alpha(t)\right] = \int_{0-}^{\infty} f(t + aT)e^{-s\tau}\,d\alpha(t). \tag{35}$$

The integral on the left is obviously

$$\int_{0}^{(a-1)T+} f(\tau)e^{-s\tau}\,d\alpha(\tau) = \sum_{m=0}^{a-1} f(mT)e^{-mTs},$$

and substitution into (35) gives

$$e^{aTs}\left[F(e^{sT}) - \sum_{m=0}^{a-1} f(mT)e^{-mTs}\right] = L_s[f(t + aT)],$$

whence

$$L_s[f(t + aT)] = z^{a}\left[F(z) - \sum_{m=0}^{a-1} f(mT)z^{-m}\right]. \tag{36}$$

For the special case of $a = 1$, we have

$$L_s[f(t+T)] = z[F(z) - f(0)] = L_s[f(t+T)]. \tag{37}$$

Finite Differences

An immediate consequence of the real translation property is that for finite differences. If finite differences are defined as in the text, we have:

Property II: If the sequence $\{f(nT)\}$ resulting from the sampling of the continuous function $f(t)$ has the L_s transform $F(z)$, then:

$$L_s[\Delta_b\{f(nT)\}] = L_s[\Delta_b f_{nT}] = \frac{z-1}{z} F(z) - f(-T),$$

$$L_s(\Delta_f f_{nT}) = (z-1)F(z) - zf(0).$$

By definition,

$$L_s(\Delta_b f_{nT}) = L_s(f_{nT} - f_{nT-T}).$$

By linearity of the transform and (34),

$$\begin{aligned} L_s(\Delta_b f_{nT}) &= F(z) - z^{-1}[F(z) + f(0)] \\ &= \left(\frac{z-1}{z}\right) F(z) - f(-T). \end{aligned} \tag{38}$$

Again, by definition,

$$L_s(\Delta_f f_{nT}) = L_s(f_{nT+T} - f_{nT})$$

and, by linearity of the transform and (37),

$$L_s(\Delta_f f_{nT}) = z[F(z) - f(0)] - F(z) = (z-1)F(z) - zf(0). \tag{39}$$

Finite Summation

As integration is the inverse operation of differentiation, summation can be considered the inverse operation of taking differences. This is demonstrated more clearly in the property below. The process of finite summation is best demonstrated in the case of a computer operating in real time. The computer samples a function which is continuous and well defined at the sampling instants. At each sample the computer adds that sample to the sum of all preceding samples. If the result of this operation is a sequence $\{g(nt)\}$, we have as the value of $g(nt)$ at any time nT:

$$g(nT) = \sum_{k=0}^{n} f(kT),$$

where $\{f(kT)\}_k$ is the sequence of sampled inputs. The analogy to integration with respect to time is clear. The L_s transform of such a summation is given below.

Property III: If $\{g(nT)\}$ is an infinite sequence such that each value of it is given by $g(nT) = \sum_{k=0}^{n} f(kT)$, then $G(z) = z/(z - 1)F(z)$, where $F(z)$ is the L_s transform of the sequence $\{f(nT)\}$.

By definition:

$$G(z) = \sum_{n=0}^{\infty} g(nT)z^{-n}.$$

Substitution of the value for $g(nT)$ gives

$$G(z) = \sum_{n=0}^{\infty} z^{-n}\left[\sum_{k=0}^{n} f(kT)\right].$$

Since we are dealing with uniformly convergent series, the order of summations may be interchanged, provided a suitable change in the limits is made in a manner equivalent to the change in limits when the order of integration is interchanged in double integration. Thus,

$$G(z) = \sum_{k=0}^{\infty} f(kT) \sum_{n=k}^{\infty} z^{-n},$$

which may be written as

$$G(z) = \sum_{k=0}^{\infty} f(kT)z^{-k} \sum_{n=0}^{\infty} z^{-n}.$$

However, the series in n may be summed as $1/(1 - z^{-1})$, and hence

$$G(z) = \sum_{k=0}^{\infty} f(kT)z^{-k}\left[\frac{z}{z - 1}\right],$$

and hence

$$G(z) = \frac{z}{z - 1} F(z). \tag{40}$$

Complex Multiplication

The superposition property of electrical networks is a very elegant and useful result of their linearity. For continuous linear networks, superposition is most concisely represented as a convolution integral, which has a particularly important Laplace transform. The same ideas also apply in the discrete case with the integral replaced by a summation.

Property IV: If $f(t)$ and $w(t)$ have the L_s transforms $F(z)$ and $W(z)$, then:

$$W(z)F(z) = L_s\left[\sum_{k=0}^{n} w(kT)f(nT - kT)\right].$$

By definition,

$$W(z) = \sum_{k=0}^{\infty} w(kT)z^{-k},$$

$$W(z)F(z) = \sum_{k=0}^{\infty} w(kT)z^{-k}F(z),$$

but

$$z^{-k}F(z) = L_s[f(t - kT)] = L_s\{[f(nT - kT)]\}.$$

Hence,

$$\begin{aligned} W(z)F(z) &= \sum_{k=0}^{\infty} w(kT)L_s[f(t - kT)] = \sum_{k=0}^{\infty} L_s[w(kT)f(t - kT)] \\ &= L_s\left[\sum_{k=0}^{\infty} w(kT)f(t - kT)\right]. \end{aligned} \tag{41}$$

At time $t = nT$ we have

$$W(z)F(z) = L_s\left[\sum_{k=0}^{\infty} w(kT)f(nT - kT)\right]$$

but $f(t) = 0; t < 0$ and therefore

$$W(z)F(z) = L_s\left[\sum_{k=0}^{n} w(kT)f(nT - kT)\right]. \tag{42}$$

Scale Change

Property V: If $L_s[f(t)] = F(z)$, *then* $L_s[e^{-at}f(t)] = F(kz)$, *where* $k = e^{+aT}$. From definition,

$$\begin{aligned} L_s[e^{-aT}f(t)] &= \int_0^{\infty} f(t)e^{-(sTa)T}\, d\alpha(t) \\ &= \sum_{n=0}^{\infty} f(nT)[e^{-aT}e^{-sT}]^n, \\ L_s[e^{-aT}f(t)] &= \sum_{n=0}^{\infty} f(nT)[kz]^{-n} = F(kz). \end{aligned} \tag{43}$$

REFERENCES

1. Ragazzini, J. R. and Zadeh, L. A., The Analysis of Sampled Data Systems, Trans. A.I.E.E., **71**, Part II, November 1952, pp. 225–234.
2. Barker, R. H., The Pulse Transfer Function and Its Application to Sampling Servo-Systems, Institution of Electrical Engineers (British) Monograph No. 43, July 1952.
3. Truxall, J. G., *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955.
4. Jury, E. I., Analysis and Synthesis of Sampled Data Control Systems, Trans. A.I.E.E., **73**, Part I, September 1954, pp. 332–346.
5. Widder, D. V., *Advanced Calculus*, Prentice-Hall, New York, 1942.
6. Gardner, M. F. and Barnes, J. L., *Transients in Linear Systems*, John Wiley & Sons, New York, 1942.
7. Widder, D. V., *The Laplace Transform*, Princeton University Press, Princeton, N. J., 1946.

Radio Transmission into Buildings at 35 and 150 mc

By L. P. RICE

(Manuscript received April 29, 1958)

Investigations of radio propagation at 35 and 150 mc into large city buildings have disclosed that, on the average, a loss in the order of 20 to 25 db may be encountered on the first floor. This loss, which represents the reduction from the median field in the city streets at the same distance from the transmitter, is known as building loss. Losses were found to be slightly smaller and more uniform at 150 mc than at 35 mc. Losses also were found to be appreciably less on higher floors in a building.

Methods of using this information for engineering radio systems to serve people in buildings are described. Some sample problems demonstrate that, with equal receiver performance, the effective coverage range in buildings for a 150-mc system will be greater than that for a 35-mc system.

I. INTRODUCTION

1.1 *Background*

With the advent of mobile telephone service has come a considerable fund of information concerning the nature of VHF radio propagation in city and suburban streets.[1]

Plans are now being made to extend the use of the mobile land transmitters to provide one-way personal radio signaling services. In these services, the transmitter signals will be detected by small pocket-carried receivers issued to subscribers. Coverage will be desired not only in the streets but also in the various buildings and other structures which subscribers might normally be expected to frequent.

The extent of useful coverage from a mobile land transmitter will be somewhat less for personal signaling than it is for mobile voice transmission. This is primarily due to two factors: (1) the inherently poorer sensitivity of a pocket-carried receiver due to its small antenna and (2) the increase in path loss to a location inside a building in comparison with

that to an outside location at the same distance from the transmitter. To offset this reduction in coverage, satellite transmitters will be required in large metropolitan areas to assure that reliable service is offered throughout. For a signaling system in which the receiver sensitivity is known, the spacing of the transmitters is largely a function of path loss.

Estimates of path loss can be made using the results of the measurements made by W. R. Young, Jr.,[1] if the additional losses in propagation caused by buildings are known.

1.2 *Scope of Study*

The losses encountered in propagating an RF field into a building were measured at eleven different locations in downtown New York City. Two of the mobile telephone channels, one in the 35-mc highway band and the other in the 150-mc urban band, were chosen for these measurements.

Most of the field-strength measurements were taken at various points on the main floor of each building. This was done because the first floor has been found to be the most difficult portion of a building to cover. The number of measurements taken varied from building to building, depending on the amount of floor space and the complexity of the floor plan.

A number of measurements at each of the two frequencies were also made in the streets adjacent to each of the buildings. For a given distance from the transmitter, the difference between the *median* field intensity in the *streets* and the field intensity at a location on the main floor of a building is defined as the *building loss* for that location. Thus, building loss is a factor which can be applied to the field intensity in the streets to assist in the prediction of the performance of a radio service in buildings.

II. OBSERVATIONS

The heterogeneous nature of the environment — both inside and outside the buildings — has been found to create extensive and erratic space variations in the RF field; accordingly, the measurements of building loss are presented statistically. The fields in the upper stories of buildings were generally found to be stronger than those near street level. Therefore, measurements made on the main floor of a building would give limiting values of building loss.

An approximate relationship between the architectural characteristics of a building — e.g., the height of the ceilings or the area of external glass — appeared to exist in certain cases.

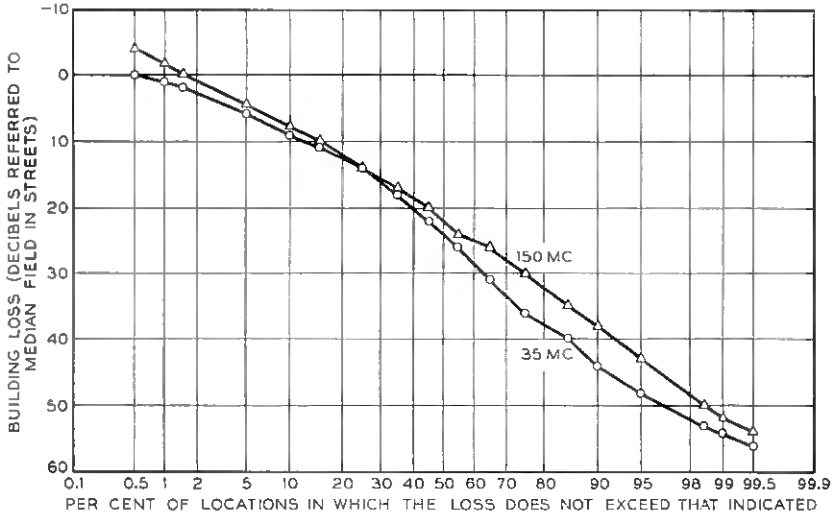


Fig. 1 — Over-all distribution of building losses at 35 and 150 mc.

The building losses at 35 and 150 mc tended to follow a log-normal distribution (see Fig. 1). At 35 mc, the over-all average building loss was found to be about 24 db; at 150 mc, it was found to be about 22 db.

Variations in signal at the lower frequency were found to be slightly greater than at the higher. Thus, the standard deviation of the building losses was found to be about 14 db at 35 mc and about 12 db at 150 mc. These variations are reversed from signal variations in the city streets, where the standard deviation of the field distributions appears to be about 7 db at 35 mc and 9 db at 150 mc.

A comparison of the useful ranges* in New York City, from transmitters of equal power to receivers of equal sensitivity, in terms of field strength in microvolts per meter, shows that the expected range of coverage into buildings is somewhat greater at 150 mc than at 35 mc. Expected ranges into buildings of almost one mile at 35 mc and almost one and one-half miles at 150 mc appear reasonable between a 250-watt transmitter and a pocket-carried signaling receiver with a sensitivity of 30 db greater than one microvolt per meter. In contrast, the useful range in city streets was found to be greater at 35 mc than at 150 mc. Service could be provided in streets over a radius of about eight miles at 35 mc and four to five miles at 150 mc.

* Useful range is defined as the distance at which there is a certain specified probability, such as 99 per cent, of successful signaling.

III. DISCUSSION OF RESULTS

3.1 *Nature of RF Field in Buildings*

It became apparent as the RF field intensities were being measured that their geometry was exceedingly complex. Variations sometimes as great as 20 db were encountered between locations a few feet apart. Since it was apparent that a point-by-point display of the field intensities would be neither useful nor meaningful, a statistical analysis of these data has been carried out to emphasize their trends.

The wide variations in field intensity can be attributed to the nature of the physical surroundings. The RF field may enter the building directly from the transmitting antenna or may be bounced in off the many reflecting surfaces presented by the surrounding buildings. Once inside, the field encounters a heterogeneous array of objects, such as walls, ceilings, floors, furniture and equipment of many kinds. Such items present lossy, shielding or reflecting media to the RF field. As a result, the field not only encounters varying degrees of attenuation in reaching a specific location, but it also arrives over a multiplicity of paths with random phase and random polarization.

Spot checks of polarization have been made by comparing the field measured at several points in a building with the antenna oriented vertically and horizontally. Differences of 10 db or more were found between the vertical and horizontal components of the field when compared on a point-by-point basis. However, when the median of the vertical components, measured at several locations, was compared with

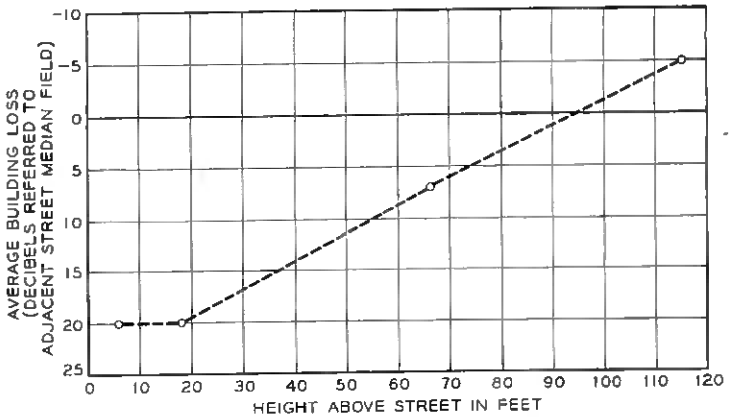


Fig. 2 — Average building loss at 150 mc on various floors in a building (463 West Street, New York City).

the median of the horizontal components, the difference was found to be negligible. This points to the interesting possibility that an omnidirectional and nonpolarized antenna might be best for reception in buildings.

Some preliminary measurements were made on various floors in a couple of buildings to determine what effect the height in a building might have on the field strength. It was found that the interference caused by adjacent structures diminished with increasing height so that the RF field was commensurately stronger on the upper floors (see Fig. 2). Therefore, it was felt that concentration could be made on the first floors with confidence that, if an adequate radio field for a system existed there, coverage in the rest of the building would be generally assured.

All the buildings surveyed were constructed of reinforced concrete or brick. Some had large window areas on the first floor. Some had large open corridors and vestibules with high ceilings. Others were more confined, with smaller external apertures on the first floor. These characteristics probably had a tendency to affect the field intensity inside the building.

A thumbnail description of characteristics which might affect propagation into each of the buildings is given in Table I, with arbitrary building identification numbers being used.

TABLE I — LOCATION AND ARCHITECTURAL CHARACTERISTICS OF BUILDINGS

Building Number	Location	First Floor Characteristics
1	463 West Street	Low ceiling height, below average window area, many halls and partitions
2	Broadway and Bowling Green	High ceilings, average window area, very thick walls
3	140 West Street	High ceilings, above average window area
4	1 Peck Slip	High ceilings, large window area, large unobstructed areas
5	130 East Broadway	High ceilings, average window area, large unobstructed areas
6	395 Hudson Street	Warehouse type building, medium ceiling height, small window area
7	432 East 14th Street	High ceilings, large window area, large unobstructed open areas
8	40 Irving Place	High ceilings, average window area, many halls
9	26 Cortlandt Street	High ceilings, large window area, large unobstructed areas
10	195 Broadway	Very high ceilings, large window area
11	220 Church Street	Medium ceiling height and window area

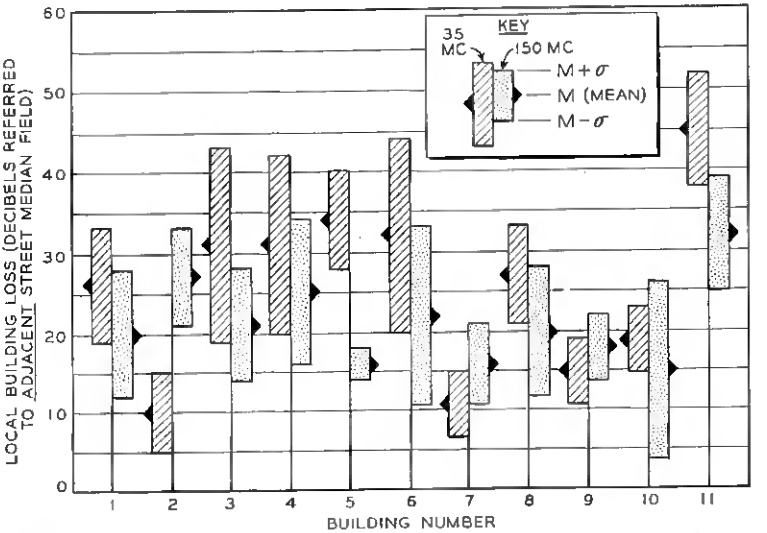


Fig. 3 — Distribution of local building losses at 35 and 150 mc for eleven buildings.

It was found that, in certain cases, a qualitative prediction might be made concerning the lossiness of a building based on the architectural characteristics just cited. For example, it may be seen from Fig. 3 that buildings 7, 9 and 10 were all found to have mean *local building losses** below 20 db at both 35 mc and 150 mc. These three buildings all had high ceilings, large windows and large unobstructed areas on their main floors. Conversely, building 11, the only one found to have an average loss exceeding 30 db at both frequencies, has lower ceilings, smaller window area and an abundance of furniture.

However, such guesses as these must necessarily be considered inconclusive because other buildings have loss effects which appear to be in direct contradiction with this hypothesis. Building 4 is an example. This building was found to present a high loss at both frequencies. Yet it is characterized in the table as being a building in which the losses might be expected to be low.

* Local building loss, distinct from the building loss defined on page 198, is defined as the difference between the median RF field in the streets *adjacent* to the individual building and the field intensity at a location on the main floor of the building. Building loss is a concept useful for the estimation of service range. Local building loss is a concept useful in evaluating the coverage of an individual building. As will be shown, the local building losses for all buildings measured together with the known variations in path losses into the streets have been combined to provide an estimate of the over-all *building loss*.

3.2 *Local Building Losses*

The local building losses at 150 mc were found to be somewhat lower than those at 35 mc. The average of these losses in the buildings ranged from 15 to 32 db at 150 mc, while at 35 mc the average ranged from 10 to 45 db. The over-all average of the local losses for the 11 buildings was found to be about 20 db at 150 mc and 25 db at 35 mc.

The distribution of the measurements at both frequencies was found to be roughly log-normal. The standard deviation in the various buildings ranged from 2 to 11 db at 150 mc and from 4 to 12 db at 35 mc. The combined standard deviation for the 11 buildings was found to be about 9 db at 150 mc and 14 db at 35 mc. Medians and standard deviations for the individual buildings are presented in Fig. 3. The distributions of local loss for the group as a whole are shown in Figs. 4 and 5.

3.3 *Determination of Building Loss*

In the preceding section, the discussion has been confined to the local building loss — referred to the median field around the particular building in question. However, a person who wishes to estimate the limiting range at which a given transmitter will propagate a field of a certain

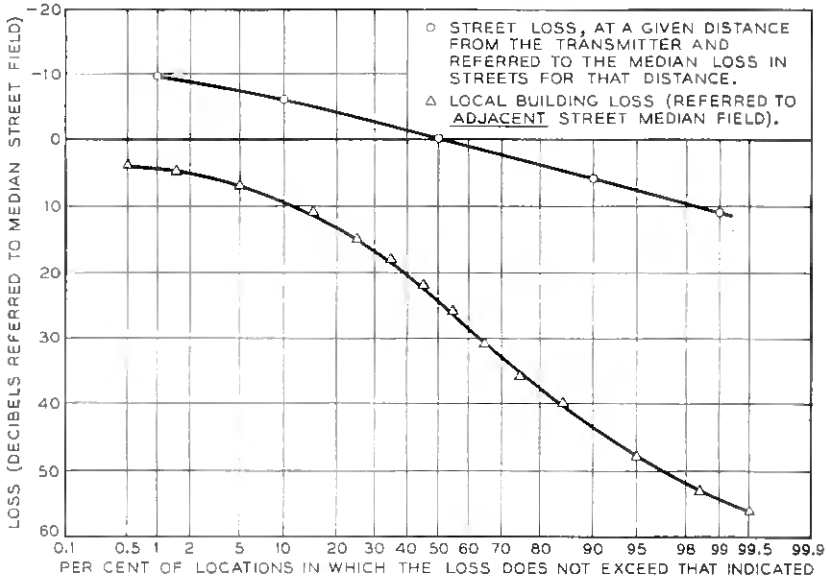


Fig. 4 — Over-all distribution of local building and street losses at 35 mc.

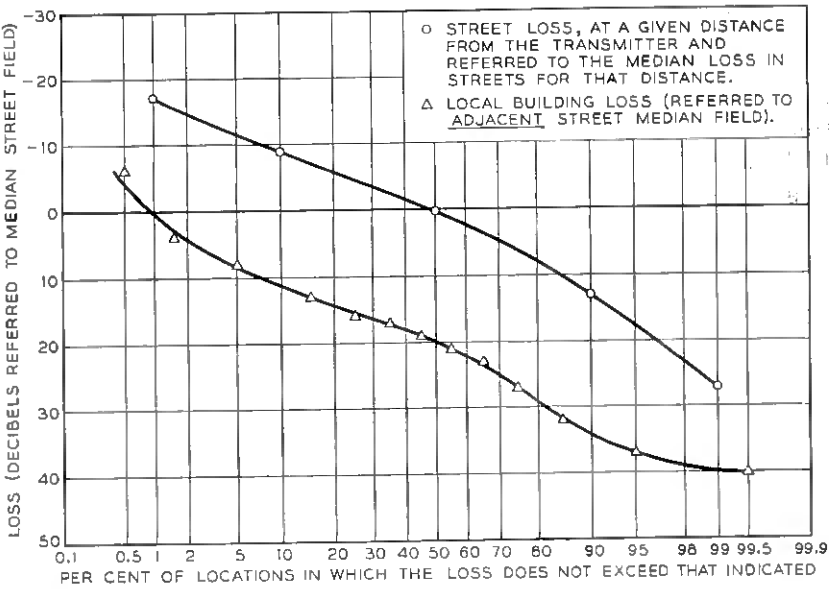


Fig. 5 — Over-all distribution of local building and street losses at 150 mc.

minimum intensity with a certain degree of reliability is concerned with building loss on the over-all basis. He would be interested in the field intensities within the buildings on the periphery of a circle. The radius of this circle around the transmitter would be the useful range of the system.

Each one of these buildings on the circle would have local building losses with respect to the median field in the streets adjacent to it. The variations in these local losses would differ from building to building. However, the variations in the local losses of a "typical building" could be approximated by combining the measurements taken in the eleven buildings. This has been done graphically to obtain the lower curves in Fig. 4 for 35 mc and in Fig. 5 for 150 mc. It is possible to determine from these figures the probability that the field at any point on the main floor of any building in a heavily built-up metropolitan area will be equal to or greater than some given level with respect to the median field intensity in the streets adjacent to that building. So, if the median of the adjacent street field is known, the coverage in the building can be estimated.

As a general rule, however, the median field in the streets adjacent to any particular building will not be known, whereas the over-all charac-

teristics of propagation into city streets have already been determined (see Figs. 6 and 7). The median field intensities in the streets adjacent to various buildings will vary from building to building, due to interference caused by the local terrain and nearby structures.

If the field intensities were measured in the streets adjacent to a large number of buildings, all equidistant from the transmitter, it is expected that the distribution of the medians of these groups of measurements would approach the distribution of all street measurements at that distance, i.e., the upper curves in Figs. 4 and 5. Now, if it is assumed that each of these many buildings is one of the "typical buildings" for which the local loss characteristics are shown in the lower curves of Figs. 4 and 5, it follows that the expected building losses for all the buildings on the circle with respect to the median field in streets at that distance from the transmitter may be determined by statistically combining the two curves on each of these two figures. This was done graphically in Fig. 1 for each of the two frequencies. The percentage of locations within all buildings, on a circle of a given radius, in which the field intensity will not be lower than a given level with respect to the median street field at the same radius may be determined directly from this figure for either frequency.

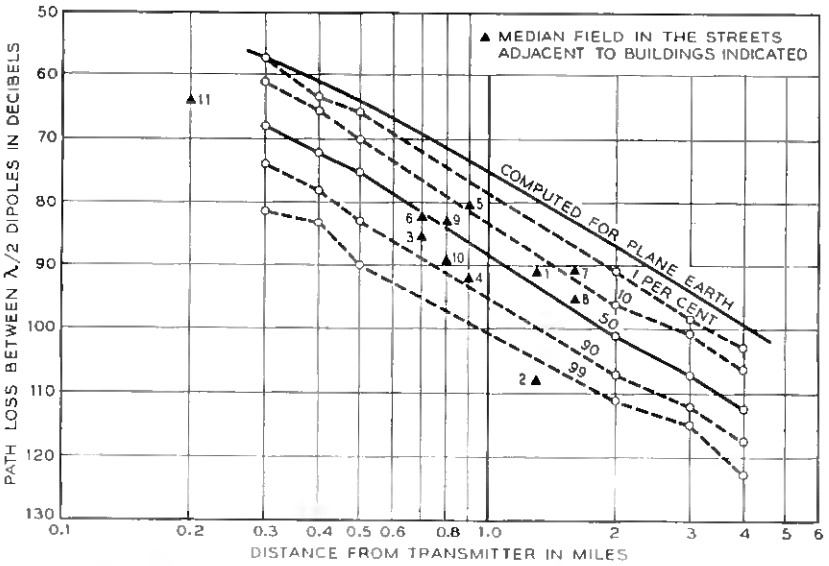


Fig. 6 — Measured path loss at 35 mc between half-wave dipoles in city streets in Manhattan. Antenna heights: transmitter—450 ft.; receiver—6 ft.

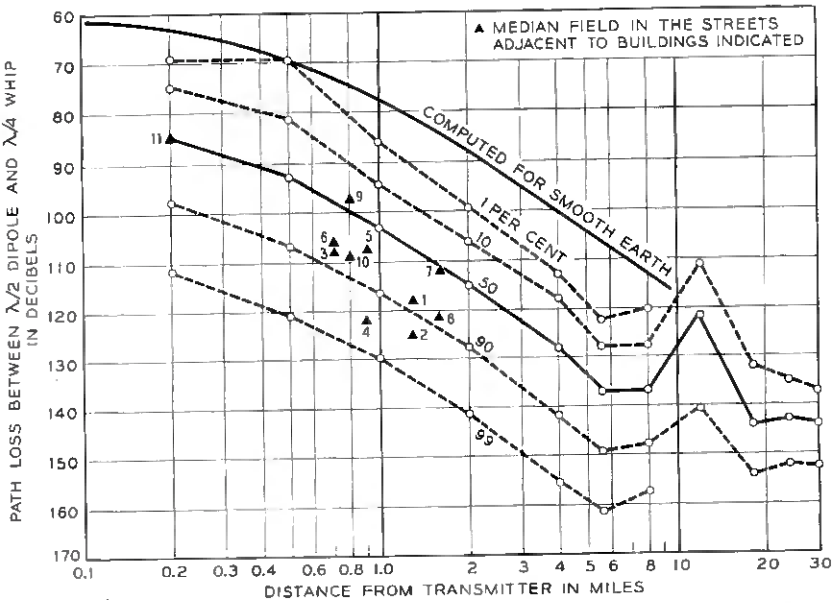


Fig. 7 — Measured path loss at 150 mc between a half-wave dipole and a quarter-wave whip in Manhattan and the Bronx and suburbs. Antenna heights: transmitter – 450 ft.; receiver – 6 ft. (All data except the 11 adjacent-street medians taken with permission of W. R. Young, Jr. from Ref. 1).

3.4 *Test Equipment Arrangements*

The New York Telephone Company's mobile telephone facilities at 32 Avenue of the Americas were used as a signal source for measuring building losses.

The field measuring equipment for work in buildings had to meet three principal requirements: portability, stability and selectivity. The available commerical field-intensity measuring apparatus was not selective enough to reject the adjacent mobile channels in New York City. Therefore, the limiter grid current in standard, battery-powered, crystal-controlled receivers was used as an indication of field strength. Provisions were made to insure that the battery aging did not upset the calibrations of the grid current meter. Prior to use, each receiver was equipped with the antenna to be used during the measurements and was calibrated in a known field by varying the field and noting the limiter current for each field intensity. The same receivers were used to measure the fields in the streets adjacent to the buildings. Antennas mounted on automobiles were connected to the receivers and the sets were recalibrated.

IV. APPLICATION TO THE ENGINEERING OF RADIO SYSTEMS

Building loss can be utilized in the engineering of a radio system in much the same way as other propagation losses. One aspect of building loss — its amplitude distribution — has an important effect on the range of reliable coverage into buildings. Inasmuch as building loss has been defined as the difference between the levels of RF field in the building and the *median* field in the streets at a given range from the transmitter, the distribution of the field intensity in the buildings must be the same as the distribution of the building loss.

The amount of building loss that can be tolerated by a system depends on the required degree of reliability. This reliability is numerically equal

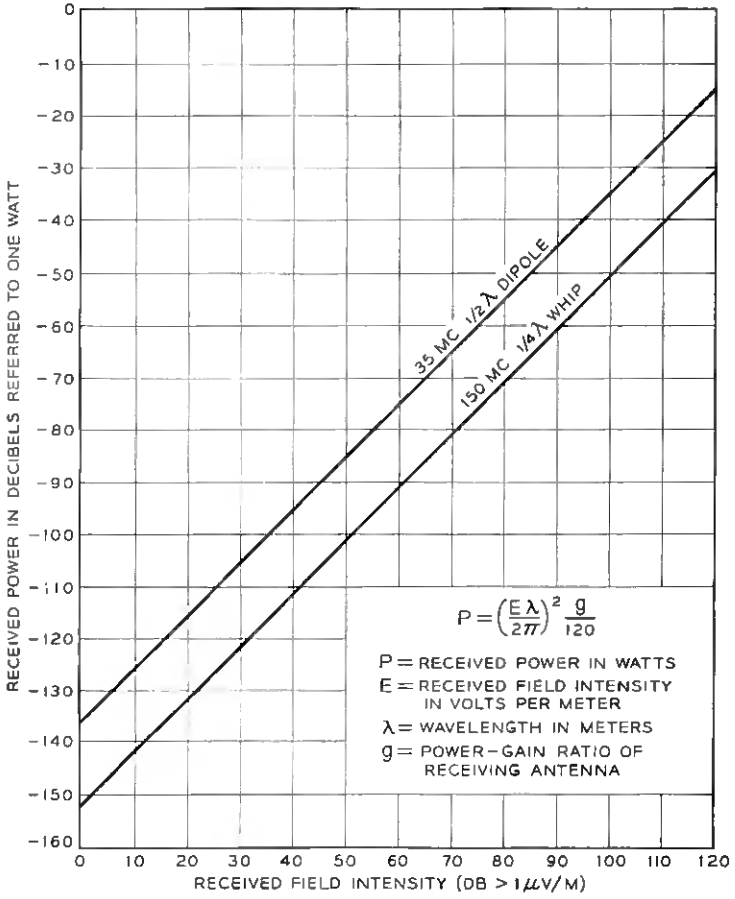


Fig. 8 — Received power at 35 and 150 mc versus received field intensity (Ref. 2).

to the percentage of the locations in the peripheral buildings in which the building loss must not exceed a certain threshold value. This threshold loss may be determined directly from the ordinate of Fig. 1 for any given per cent of reliability on the abscissa. When the maximum allowable path loss to the receiver and the threshold building loss are known, their difference represents the *median* path loss in streets that can be tolerated and still provide the minimum acceptable coverage in the buildings. The determination of such factors as required transmitter power or maximum range of coverage can be handled in any convenient manner, in terms of median losses to the adjacent streets.

The following five steps describe one method for determining the service range of a transmitter in a large metropolitan area such as New York City. The procedure consists of the determination of first the maximum allowable path loss and then the range at which this path loss is not exceeded for a given system reliability.

1. Determine the *minimum usable received power* from a half-wave

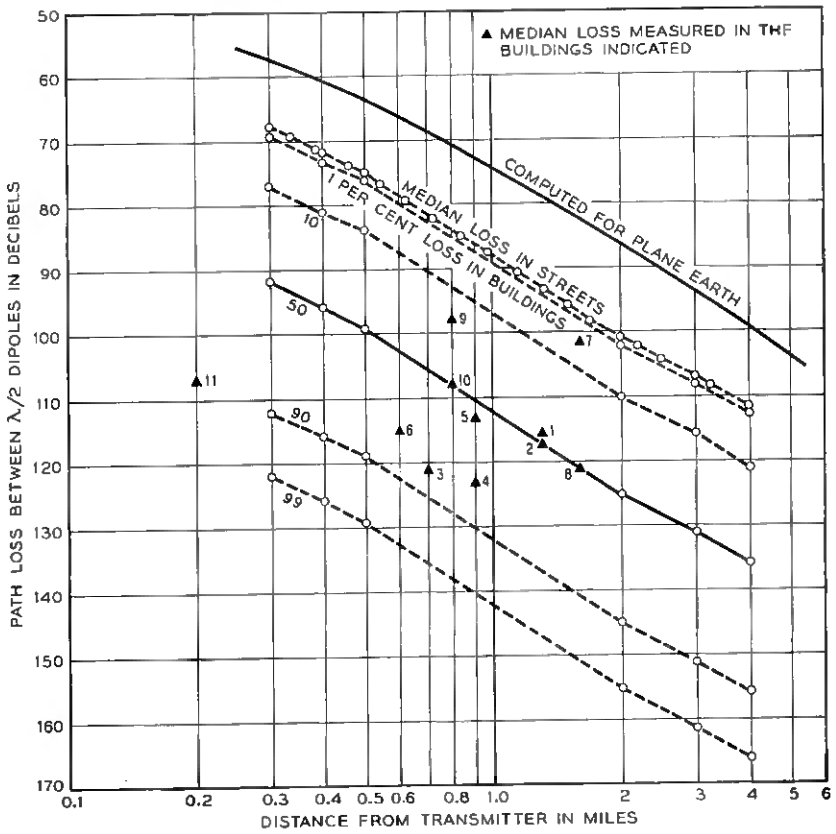


Fig. 9 — Path loss at 35 mc between half-wave dipoles into large city buildings.

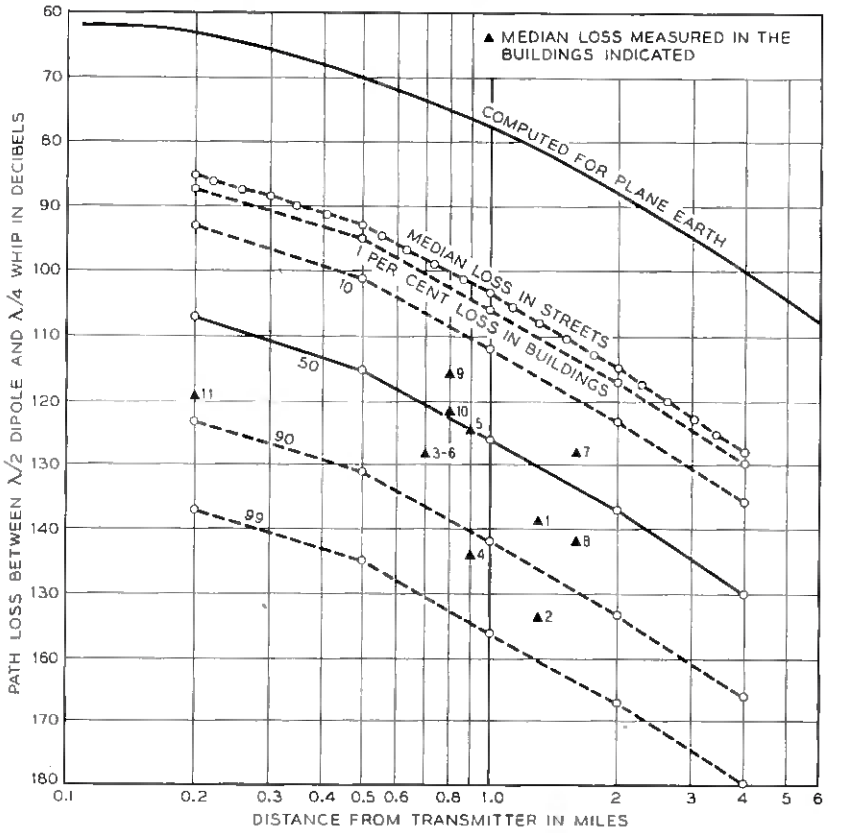


Fig. 10 — Path loss at 150 mc between a half-wave dipole and a quarter-wave whip into large city buildings.

dipole (for 35 mc) or a quarter-wave whip (for 150 mc). Fig. 8 can be used for this purpose when receiver sensitivity is known in terms of minimum required field intensity.

2. Subtract the minimum usable received power (in dbw) from the equivalent transmitted power from a half-wave dipole to determine the *maximum allowable path loss* between the two pairs of antenna terminals.

3. Determine the *building loss* from the ordinate on Fig. 1 which corresponds to the required system reliability. (The system reliability in per cent is numerically equal to the scale on the abscissa of this figure.)

4. Subtract the building loss from the maximum allowable path loss to determine the equivalent *median loss in streets.*

5. Determine from Fig. 9 (for 35 mc) or 10 (for 150 mc) the *range* at which this median loss in streets occurs. Use the curve labeled "Median Loss in Streets". This is the useful range of the system for coverage into buildings.

TABLE II — EXAMPLES OF ESTIMATION OF RANGE AT WHICH RADIO SERVICE CAN BE OFFERED IN METROPOLITAN BUILDINGS

Assumptions:
Receiver sensitivity — 30 db > 1 microvolt per meter
Effective radiated power from dipole 450 ft above ground — 24 dbw (250 watts)
System reliability* — 90 per cent

* This is the system reliability for a nonrepetitive system. Signals may be sent out more than once in a personal signaling system. If two signals are sent to a subscriber moving about in a marginal field, the system reliability for this problem would be 99 per cent; if three signals are sent, 99.9 per cent.

Frequency	Step Number						
	(1)		(2)		(3)	(4)	(5)
	Receiver Sensitivity, in db < 1 μv per meter	Minimum Usable Received Power, in dbw (Fig. 8)	Radiated Power (Dipole), in dbw	Maximum Allowable Path Loss in db	Building Loss for 90% Reliability, in db (Fig. 1)	Equivalent Median Loss in Streets, in db	Estimated Service Ranges in miles (Figs. 9 & 10)
35 mc	30	−106 (dipole)	24	130	44	86	0.9
150 mc	30	−122 (whip)	24	146	38	108	1.3

If only a rough estimate is required, steps 3 and 4 may be eliminated by interpolating between the 50 per cent, 90 per cent and 99 per cent curves on Figs. 9 and 10 and determining directly the range corresponding to the maximum allowable path loss found in step 2. Here again, the percentages are numerically equal to the system reliability.

Some numerical examples of range estimation are given in Table II in order to illustrate the use of this procedure. The step numbers correspond to those listed above.

By comparison, for the conditions given in the table, the expected coverage in streets may be in the order of 8 miles at 35 mc and 4 to 5 miles at 150 mc.* It is of interest to note that, while better coverage may be expected in streets at 35 mc than at 150 mc, the higher building losses at 35 mc attenuate the field so much that better coverage in buildings can be expected at 150 mc.

V. ACKNOWLEDGMENTS

The author wishes to acknowledge the participation of A. Bader, W. Mitchell, and J. Franzblau in the planning, conduct, and evaluation of these measurements.

REFERENCES

1. Young, W. R., Jr., Comparison of Mobile Radio Transmission at 150, 450, 900 and 3700 mc, B.S.T.J., **31,** November 1952, pp. 1068–1085.
2. Bullington, K., Radio Propagation at Frequencies above 30 Megacycles, Proc. I.R.E., **35,** October 1947, Equation (3), p. 1123.

* From the 90 per cent curves of Figs. 6 and 7.

On Trunks with Negative Exponential Holding Times Serving a Renewal Process

By VÁCLAV E. BENEŠ

(Manuscript received June 27, 1958)

A group of N trunks serves calls arriving in a renewal process, and lost calls are cleared. The number, $N(k)$, of trunks found busy by the kth arriving customer is studied as a Markov process imbedded in a (usually) non-Markov process $N(t)$, the number of trunks busy at t. Results of C. Palm and F. Pollaczek on the distribution of $N(k)$ are generalized, and a study is made of bounds for, and approximations to, the probability of loss. The probability of loss is studied as a functional of the interarrival distribution function, and certain extremal properties are proven. Formulas for the mean of $N(k)$ and for the covariance function are given, together with equilibrium curves for the probability of loss, for the mean and variance of $N(k)$, and for the first four values of the covariance function. Some applications to switch counting are discussed.

1. INTRODUCTION

We shall study a mathematical model for the random behavior of the occupancy of trunk groups. The principal results are complete descriptions (in principle) of (a) the variations of the traffic in time, (b) the equilibrium probabilities and (c), the covariance function of the traffic found by arriving customers. These mathematical results have practical application in engineering trunk groups to have a given probability of loss, and in estimating the sampling error incurred in certain ways of measuring traffic.

A "trunk group" is a set of transmission channels (trunks) between central offices. The trunks in a group are often equivalent in the sense that a call handled on one idle trunk could as well have been assigned another. A "holding time" of a trunk is a length of time during which it is continuously unavailable because it is being seized and used as a talk-

ing path. By "interarrival times" we mean the time intervals elapsing between successive epochs at which attempts are made to place a call on the trunk group. With these definitions in mind, the theoretical model we use to describe the trunk group involves four assumptions:

i. The holding times of trunks are independent random quantities having a negative exponential distribution, with mean value, γ^{-1} (γ is the hang-up rate). This means that if a trunk is in use at time x, the chance that it is still in use at $(x + dx)$ is $1 - \gamma dx - o(dx)$, $o(dx)$ denoting a quantity of order smaller than dx, irrespective of how long the trunk has been in use. The probability that a holding time is less than t is then $1 - \exp\{-\gamma t\}$ for $t \geqq 0$, and 0 otherwise.

ii. The interarrival times of calls are independent positive variates; each has the general distribution $A(u)$, where $A(u)$ is arbitrary except for the condition $A(0) = 0$. If t_k and t_{k+1} are successive arrival times, then

$$\Pr\{t_{k+1} - t_k \leqq u\} = A(u),$$

for all k, independently. This assumption covers Poisson (or completely random) arrivals as a special case. In accordance with the usage in the literature, we call a sequence of mutually independent, identically distributed, positive variates, a "renewal process." The interarrival times in our model then form a renewal process. It has been shown by Palm[1] and noted by Feller[2] that non-Poisson renewal processes arise in their own right in the study of overflow traffic from a trunk group, even when the original offered traffic is Poisson in character.

iii. There are $N < \infty$ trunks in the group.

iv. Calls which find all N trunks busy are lost, and are cleared from the system.

A model like the above, but without the strong simplifying assumption of exponential holding time, was studied by Pollaczek.[3] The model described in (i) through (iv) above has been considered by Palm,[1] and also by Takács,[4] who used a functional equation. Takács' paper was apparently written without knowledge of the prior work of Palm and Pollaczek; in a recent paper,[5] Takács thanks R. Syski for calling his attention to Refs. 1 and 3. The same model has also been treated by Cohen.[6] For convenience and unity of exposition, some of the results of these authors shall be rederived here, and attributed to the appropriate author as they arise.

II. SUMMARY OF RESULTS

It is natural to use the number $N(t)$ of calls in progress on the trunk group at time t as an indicator of traffic; $N(t)$ is a random step function, fluctuating in unit steps from 0 to N.

Unless the arrivals form a Poisson process; that is, unless

$$A(u) = 1 - \exp\{-u/\mu_1\}$$

for $u \geqq 0$ and $\mu_1 > 0$, $N(t)$ is not a Markov process. However, let t_k be the epoch of the kth arrival, and suppose that $N(t_k - 0)$ is known. Thus, we know how many busy trunks were found by the kth call. Until the next call arrives at t_{k+1}, the number of calls in progress forms essentially a simple death process, with death rate γ per head of population. The conditional distribution of $N(t_{k+1} - 0)$, given $N(t_k - 0)$, can then be calculated from the known transition probabilities of the death process (see Feller[7]). No additional knowledge of $N(t)$ for $t < t_k$ is of prognostic relevance to $N(t)$ for $t > t_k$, when $N(t_k - 0)$ is known. We define

$$N(k) = N(t_k - 0),$$

where $N(k)$ is the number of trunks found busy by the kth arriving call. The variates $N(k)$ form a Markov chain imbedded in the non-Markov process $N(t)$. This Markov chain is the basic random process considered in this paper.

Let the numbers a_n, $n = 1, \cdots, N$ be defined by

$$a_n = \int_0^\infty e^{-n\gamma u}\, dA(u),$$

so that a_n is the Laplace-Stieltjes transform of the interarrival distribution $A(u)$, evaluated at the point $n\gamma$, where γ is the hangup rate. The principal theoretical result of this paper is Theorem 1 in Section IV. This result gives formulas for the generating functions

$$\psi_n(z) = \sum_{k \geqq 0} z^k \Pr\{N(k) = n\}$$

for an arbitrary initial distribution of $N(0)$. These formulas depend only on the numbers $a_1, \cdots, a_N$ defined previously, so the entire Markov process $N(k)$ depends only on these numbers. Theorem 1 determines, in principle, the transition probabilities of $N(k)$ purely in terms of $a_1, \cdots, a_N$, and so provides a complete description of the statistical variations of the traffic found by arriving customers. For $N(0) = 0$, the formulas were obtained by Pollaczek;[3] the formulas to be given coincide with those of Pollaczek in this case.

In Section V the limiting probabilities

$$p_n = \lim_{k \to \infty} \Pr\{N(k) = n\},$$

already considered by Palm, Pollaczek and Takács, are briefly discussed. The quantity p_n is the equilibrium chance that an arriving customer find

n trunks busy; in particular, p_N is the probability of loss. It should be kept in mind that p_n is not the probability that, if we inspect the trunk group at a random moment in equilibrium, we will find n trunks busy; the moments of inspection must be those immediately preceding arrivals. In Section V, also, various moments (such as the ordinary, binomial and factorial) and the variance of the limit distribution $\{p_n\}$ are presented. Curves of the probability of loss, the fraction of trunks found busy by an arrival and the variance of $\{p_n\}$ are plotted as functions of the offered erlangs for three choices of the interarrival distribution $A(u)$.

Sections VI and VII discuss bounds for, and approximations to, the probability p_N of loss. The results of Section VI are general; those of Section VII are restricted to the case of regular arrivals. Consideration of the unrealistic (for telephone trunking) special case of regular arrivals is justified (in Section VIII) by the fact that regular arrivals form a limiting best case.

In Section VIII we treat p_N as a functional of the interarrival distribution $A(u)$. The chief results can be summarized informally as follows:

i. For a fixed mean interarrival time and a fixed hang-up rate, the minimum loss is achieved when arrivals are regular.

ii. Arriving customers can, without changing either their mean arrival rate or their hang-up rate, still make the telephone company give them arbitrarily bad service (high loss) by a proper choice of $A(u)$.

iii. The maximum number of erlangs that N trunks can carry at a fixed loss probability p [the maximum being over $A(u)$ that achieve p], is a number depending only on N and p.

Section IX is a brief discussion of $\Pr\{N(k) = N\}$, the chance that the kth arrival suffers loss, as a function of k. The case $N = 2$ is described in detail, and curves are included for one choice of $A(u)$.

Finally, Section X is devoted to the mean value $E\{N(k)\}$ of $N(k)$ as a function of k, and to the covariance function of $N(k)$ defined as

$$R(n) = \lim_{k\to\infty} E\{N(k)N(k+n)\} - E^2\{N(k)\}.$$

General formulas for both $E\{N(k)\}$ and $R(n)$ are derived, together with a recurrence relation for the latter to facilitate computation. The chief practical application of the covariance function is to theoretical estimates of sampling error in traffic measurement. Discussions of the use of our results to estimate sampling error in certain possible kinds of switch counting are given, together with some curves of the covariance. We stress that our results are for a finite, not an infinite, number of trunks. In particular, we show that a natural exponential approximation to the covariance, valid for $N = \infty$, can be several times too large for small N.

III. SUMMARY OF PRINCIPAL NOTATIONS AND DEFINITIONS

'E' is used to denote mathematical expectation

N = number of trunks in the group

γ = hang-up rate = (mean holding time)$^{-1}$

$A(u)$ = Pr{interarrival time $\leqq u$}

μ_i = ith ordinary moment of the interarrival distribution $A(u)$

$$a_n = \int_0^\infty e^{-n\gamma u}\, dA(u),\ n = 1, 2, \cdots, N$$

$N(t)$ = number of trunks busy at t

t_k = epoch of the kth arrival

$N(k)$ = $N(t_k - 0)$ = number of trunks found busy by the kth arrival

p_n = $\lim_{k\to\infty} \Pr\{N(k) = n\}$ = equilibrium probability of finding n trunks busy

p_N = equilibrium probability of loss

$$b_n = \sum_{m=n}^{N} \binom{m}{n} p_m = n\text{th binomial moment of the distribution } \{p_m\}$$

$$M_{(n)} = n!b_n = \sum n(n-1)\cdots(n-m+1)p_m = n\text{th factorial moment of } \{p_m\}$$

$$m_n = \sum_m m^n p_m = n\text{th ordinary moment of } \{p_m\}$$

$$\sigma^2 = m_2 - m_1^2 = \text{variance of } \{p_m\}$$

$$P_u(x) = 1 + (x-1)e^{-\gamma u}$$

$$E\{x^{N(k)}\} = \sum_{m=0}^{N} x^m \Pr\{k\text{th call find } m \text{ trunks busy}\}$$

$$\varphi(x, z) = \sum_k z^k E\{x^{N(k)}\}$$

$$\psi_n(z) = \sum_k z^k \Pr\{N(k) = n\}$$

$$b_n(z) = \sum_{m=n}^{N} \binom{m}{n} \psi_m(z) = (n!)^{-1} \times \text{factorial moment generating function}$$

$$k_n = \sum_{m=n}^{N} \binom{m}{n} \Pr\{N(0) = m\} = n\text{th binomial moment of initial distribution}$$

$$D(a_1, a_2, \cdots, a_N, z) = \sum_{0}^{N} \binom{N}{i} (1 - za_1) \cdots (1 - za_i)\, a_{i+1} \cdots a_N\, z^{N-i}$$

$$L_k^{(N)} = \prod_{m=1}^{N} (1 - a_k + a_{k+m})$$

$$U_k^{(N)} = \prod_{m=o}^{N-1} (1 - a_{k+m} + a_{k+N})$$

$$f(x_1, x_2, \cdots, x_N) = 1 + \binom{N}{1} \frac{1 - x_1}{x_1} + \cdots + \binom{N}{N} \frac{(1-x_1)\cdots(1-x_N)}{x_1 \cdots x_N}$$

$$R(n) = \lim_{k\to\infty} E\{N(k)N(k+n)\} - E^2\{N(k)\} = \text{covariance function of } N(k)$$

$$Q_k = \sum_{m=0}^{N} m p_m \Pr\{N(k) = N \mid N(0) = m\}$$

IV. DERIVATION OF GENERATING FUNCTIONS

The behavior of a trunk group with (a) independent holding times, (b) independent interarrivals and (c) N trunks with lost calls cleared has been studied by Pollaczek[3], who derived the generating functions

$$\sum_k z^k \Pr\{k\text{th arrival finds } n \text{ trunks busy}\},$$

on the condition that the first arrival found all trunks idle.

Palm and Takács derived the limit probabilities

$$p_n = \lim_{k\to\infty} \Pr\{k\text{th arrival finds } n \text{ trunks busy}\}$$

for the case of exponential holding times, to which we are also limiting ourselves here. Takács used the equilibrium equations for the same Markov process $N(k)$ as we have introduced. We shall show that his functional equation approach can be used to generalize Pollaczek's results, and to obtain further formulas of practical importance in traffic engineering.

We let $P_u(x) = 1 + (x - 1)e^{-\gamma u}$. Then, by the argument of Takács,[4]

$$E\{x^{N(k+1)} \mid N(k)\} = \int_0^\infty [P_u(x)]^{1+N(k)-\delta_{N,N(k)}}\, dA(u),$$

with the δ symbol indicating that lost calls are cleared. Hence

$$E\{x^{N(k+1)}\} = \int_0^\infty \left[\sum_{n<N} \Pr\{N(k) = n\} P_u{}^{1+n}(x) + \Pr\{N(k) = N\} P_u{}^{N}(x) \right] dA(u).$$

Let

$$\varphi(x,z) = \sum_{k\geq 0} z^k E\{x^{N(k)}\},$$

$$\psi_n(z) = \sum_{k\geq 0} z^k \Pr\{N(k) = n\}, \qquad n = 0, 1, \cdots, N.$$

Then φ satisfies the functional equation

$$\varphi(x,z) = E\{x^{N(0)}\} + z\int_0^\infty \left\{ \varphi[P_u(x), z] P_u(x) - \psi_N(z)[P_u{}^{N+1}(x) - P_u{}^{N}(x)] \right\} dA(u). \tag{1}$$

This is a discrete time-dependent analog of Takács' functional equation. To solve it, set $x = 1 + w$ and define the functions b_n by

$$b_N(z) = \sum_{m=n}^{N} \binom{m}{n} \psi_m(z), \qquad n = 0, 1, \cdots, N.$$

Note that

$$b_n(z) = \psi_N(z), \tag{2}$$

$$\varphi(x,z) = \sum_{n=0}^{N} x^n \psi_n(z). \tag{3}$$

If we now equate coefficients of like powers of w in the functional equation (1), we obtain the following recurrence for the functions $b_n(z)$:

$$b_n(z) = z a_n \left[b_n(z) + b_{n-1}(z) - \binom{N}{n-1} \psi_N(z) \right] + k_n, \quad n \geq 1, \tag{4}$$

where

$$k_n = \sum_{m=n}^{N} \binom{m}{n} \Pr\{N(0) = m\},$$

$$a_n = \int_0^\infty e^{-n\gamma u}\, dA(u).$$

The terms k_n are the binomial moments of the distribution of $N(0)$, and represent initial conditions. Since

$$\sum_{n=0}^{N} \Pr\{N(k) = n\} = 1$$

for each $k \geqq 0$, we find $b_0(z) = (1 - z)^{-1}$.

The solution of the recurrence (4) is

$$b_n(z) = \prod_0^N \frac{za_j}{1 - za_j} \cdot \left\{(1 - z)^{-1} - \sum_{j=1}^{n} \left[\binom{N}{j-1} b_N(z) - \frac{k_j}{za_j}\right] \prod_0^{j-1} \frac{1 - za_i}{za_i}\right\}, \tag{5}$$

where the first term of the products is always taken to be 1. From this and (2) one can determine $b_N(z)$ and hence all the $\psi_n(z)$. The complete result is

Theorem 1: *The generating function* $\psi_n(z)$ *of* $\Pr\{N(k) = n\}$, *defined by*

$$\psi_n(z) = \sum_{k \geqq 0} z^k \Pr\{N(k) = n\}$$

is given by the formula

$$\psi_n(z) = \sum_{j=0}^{N-n} (-1)^j \binom{n+j}{n} b_{n+j}(z),$$

where the $b_n(z)$ *are solutions of* (4). *In particular, the generating function of the probabilities that the kth arrival find all N trunks busy is*

$$\psi_N(z) = b_N(z) = \sum_{k \geqq 0} z^k \Pr\{N(k) = N\}$$

$$= (1 - z)^{-1} \frac{k_0 + \dfrac{k_1(1 - z)}{za_1} + \cdots + \dfrac{k_N(1 - z)(1 - za_1) \cdots (1 - za_{N-1})}{z^N a_1 a_2 \cdots a_N}}{1 + \binom{N}{1}\dfrac{1 - za_1}{za_1} + \cdots + \binom{N}{N}\dfrac{(1 - za_1)\cdots(1 - za_N)}{z^N a_1 a_2 \cdots a_N}}.$$

This reduces to Pollaczek's result (Ref. 3, p. 1470) when the system starts empty with $N(0) = 0$, since $k_0 \equiv 1$, and $N(0) = 0$ implies that $k_i = 0$ for $i > 0$. Let us set

$$D_N(x_1, x_2, \cdots x_N, z) = \sum_{j=0}^{N} \binom{N}{j} (1 - zx_1) \cdots (1 - zx_j) x_{j+1} \cdots x_N z^{N-j}.$$

In this notation we can write

$$D_N(a_1, a_2, \cdots a_N, z) = a_1a_2 \cdots a_Nz^N \text{ [denominator of } \psi_N(z)].$$

Lemma 1: *The functions* $D_N(x_1, x_2, \cdots, x_N, z)$ *satisfy the recurrence relations*

$$D_{N+1}(x_k, \cdots, x_{k+N}, z) = zx_{k+N}D_N(x_k, \cdots, x_{k+N-1}, z) + (1 - zx_k)D_N(x_{k+1}, \cdots, x_{k+N}, z).$$

Proof of this is from the formula

$$\binom{N+1}{i} = \binom{N}{i} + \binom{N}{i-1}.$$

V. THE STATIONARY DISTRIBUTION

In the terminology of Feller,[7] the variates $N(k)$ form an aperiodic, irreducible Markov chain; hence the limits

$$p_n = \lim_{k\to\infty} \Pr\{N(k) = n\}$$

exist, and can be evaluated from the generating functions $\psi_n(z)$ by Abel's theorem. The result is

Theorem 2: *The stationary distribution of* $N(k)$ *is* $\{p_n\}$, *given by*

$$p_n = \sum_{j=0}^{N-n} (-1)^j \binom{n+j}{n} b_{n+j},$$

with $b_0 = 1$, *and*

$$b_n = \prod_0^n \frac{a_j}{1-a_j}\left\{1 - p_N \sum_{m=1}^{n}\binom{N}{m-1}\prod_0^{m-1}\frac{1-a_i}{a_i}\right\}, \tag{6}$$

$$p_N = \text{probability of loss} = \frac{a_1a_2 \cdots a_N}{D_N(a_1, a_2, \cdots, a_N, 1)} = \left\{1 + \binom{N}{1}\frac{1-a_1}{a_1} + \cdots + \binom{N}{N}\frac{(1-a_1)\cdots(1-a_N)}{a_1a_2\cdots a_N}\right\}^{-1}. \tag{7}$$

Theorem 2, and the loss formula (7) are due to Palm[1] and Pollaczek;[3] these results have been rederived independently by L. Takács, H. Scarf, the present author — and doubtless many others.

The quantities b_n of Theorem 2 are the binomial moments of $\{p_n\}$, defined as

$$b_n = \sum_{m=n}^{N}\binom{m}{n}p_m,$$

and they satisfy the recurrence

$$b_0 = 1,$$

$$b_n = a_n\left[b_n + b_{n-1} - p_N\binom{N}{n-1}\right], \qquad n > 0,$$

which can be solved to give formulas (6) and (7). The factorial moments $M_{(n)}$ are then given by

$$M_{(n)} = n!\, b_n = \sum_{m=n}^{N} m(m-1)\cdots(m-n+1)p_m,$$

and they satisfy the recurrence

$$M_{(0)} = 1,$$

$$M_{(n)} = a_n[M_{(n)} + M_{(n-1)} - np_N N(N-1)\cdots(N-n+2)], \qquad n \geqq 1.$$

In Fig. 1, the probability p_N of loss has been plotted as a function of the average offered load, a, in erlangs, for three separate choices of the interarrival distribution $A(u)$, for values of N from 1 to 8. The choices have been intentionally made so that the crucial quantities a_n depend on γ and $A(u)$ only via the offered load, a. The choices are as follows:

i. Poisson arrivals are represented in Fig. 1 by a dashed line. In this case, $a_n = a/(a+n)$.

ii. Suppose that the times between successive arrivals are uniformly distributed in the interval $(\mu_1 - b, \mu_1 + b)$ for $0 < b \leqq \mu_1$. The mean interarrival time is μ_1, and a simple calculation gives

$$a_n = e^{-n\gamma\mu_1}\frac{\sinh n\gamma b}{n\gamma b}. \tag{8}$$

We choose $b = \mu_1$; then a_n depends only on $\gamma\mu_1 = a^{-1}$, and

$$a_n = e^{-n/a}\frac{\sinh n/a}{n/a}.$$

This choice of $A(u)$ we shall call "uniformly distributed interarrivals;" it is represented in Fig. 1 by alternating long and short dashes.

iii. Regular arrivals are represented in Fig. 1 by a solid line. For regular arrivals, $a_n = e^{-n/a}$, which is the limiting form of (8) as b tends to zero.

The curve for regular arrivals ($a_n = e^{-n/a}$) always falls below the curves

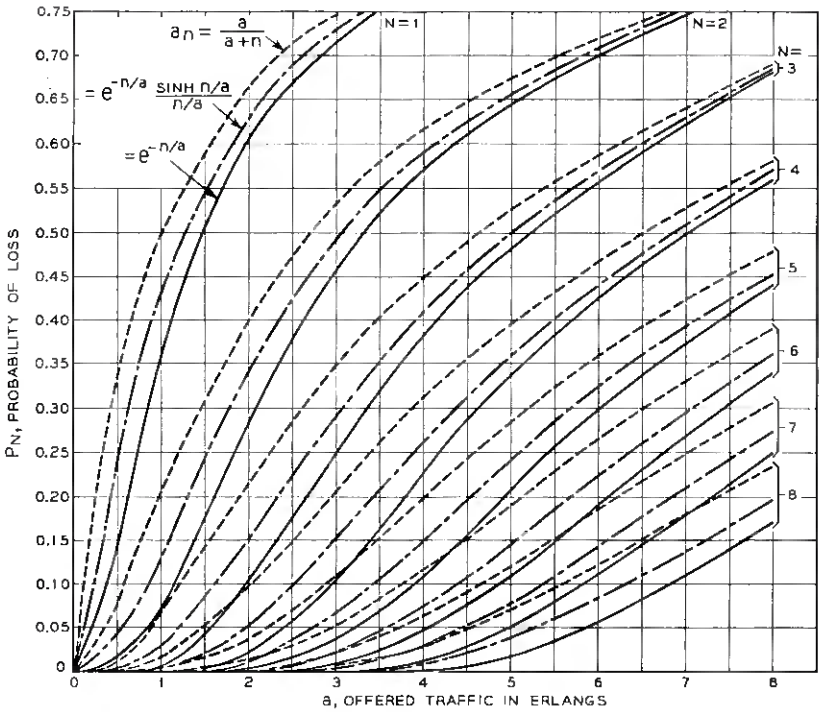


Fig. 1 — The probability of loss: (i) Poisson arrivals, $a_n = a/(a + n)$, dashed line; (ii) uniformly distributed interarrivals, $a_n = e^{-n/a}(\sinh n/a)/(n/a)$, long-and-short dashed line; (iii) regular arrivals, $a_n = e^{-n/a}$, solid line.

for the other two choices. This is a consequence of Theorem 9 of Section VIII, according to which regular arrivals form a limiting best case, for which p_N assumes its lower bound for fixed offered traffic a. On the other hand, the curve for Poisson arrivals, although always above the curves for the other two choices in Fig. 1, is by no means the limiting worst case, since there is none. For Theorem 10 of Section VIII says that, for given $\epsilon > 0$ and offered traffic a, we can always find an interarrival distribution $A(u)$ for which $p_N > 1 - \epsilon$.

The differences in p_N for the various choices of $A(u)$ in Fig. 1 are possibly explainable by considering the amount of mass that $A(u)$ concentrates in the neighborhood of 0. For regular arrivals there is no mass, so that the system always has a "breathing spell" before the next arrival. For uniformly distributed interarrivals, there is always mass in a neighborhood of zero, but the density at 0 is no larger than anywhere else. For Poisson arrivals, however, not only is there mass in any neigh-

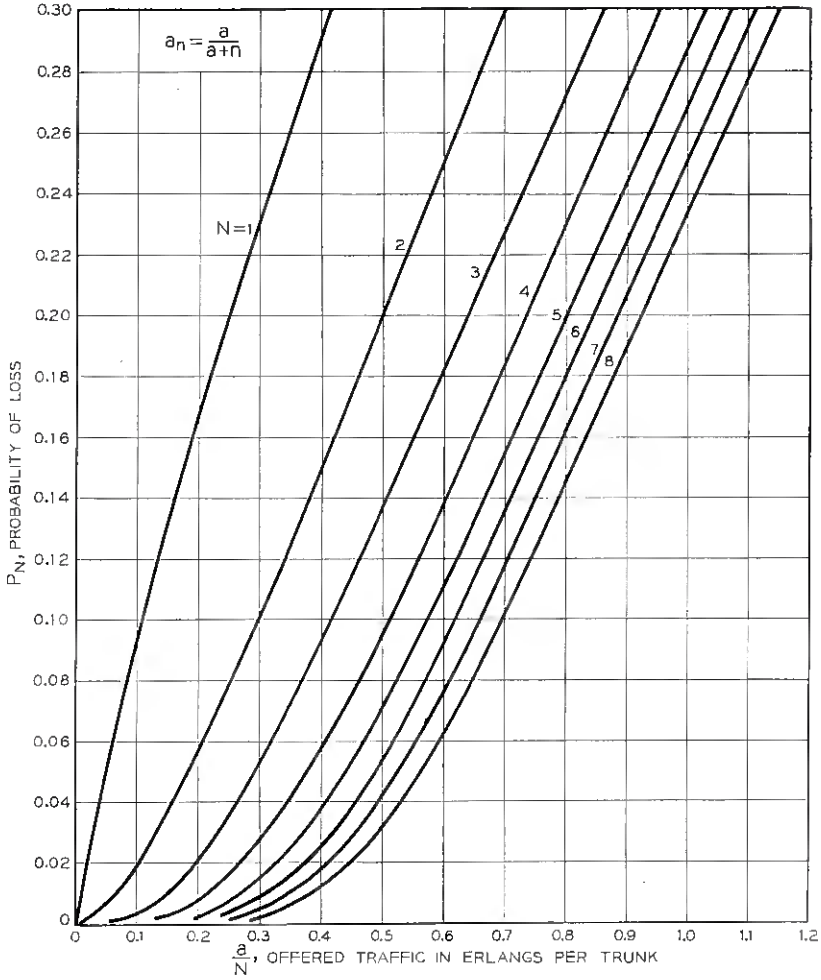


Fig. 2 — The probability of loss as a function of load per trunk for Poisson arrivals, $a_n = a/(a + n)$.

borhood of 0, but the density is a maximum at 0, so that the damaging short interarrivals are, in a sense, the most likely.

From Theorem 13 and the Palm formula (7) it can be verified that, as $a \to \infty$, the curves for the different choices of $A(u)$ must approach each other and 1. But for small values of a there are substantial differences among them. For this reason, they have been replotted in the separate Figs. 2, 3 and 4 as functions of a/N, the offered load per trunk.

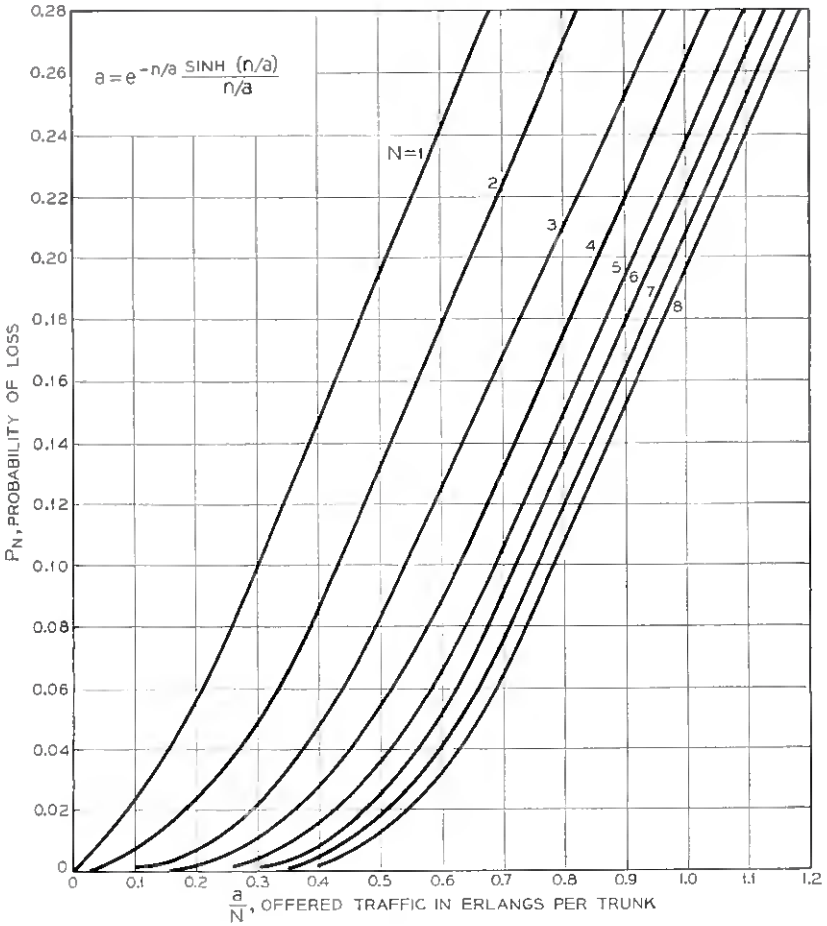


Fig. 3 — The probability of loss as a function of load per trunk for uniformly distributed interarrivals, $a_n = e^{-n/a}\ (\sinh n/a)/(n/a)$.

The first two ordinary moments m_1 and m_2 of $\{p_n\}$ are respectively given by

$$m_1 = M_{(1)} = b_1 = \sum_{n=0}^{N} np_n = \frac{a_1(1 - p_N)}{1 - a_1},$$

$$m_2 = M_{(2)} + M_{(1)} = 2b_2 + b_1 = \sum_{n=0}^{N} n^2 p_n$$

$$= \frac{a_1 a_2(1 - p_N)}{(1 - a_1)(1 - a_2)} - \frac{2a_2 N p_N - m_1}{1 - a_2}, \qquad \text{for } N > 1,$$

$$= a_1, \qquad \text{for } N = 1.$$

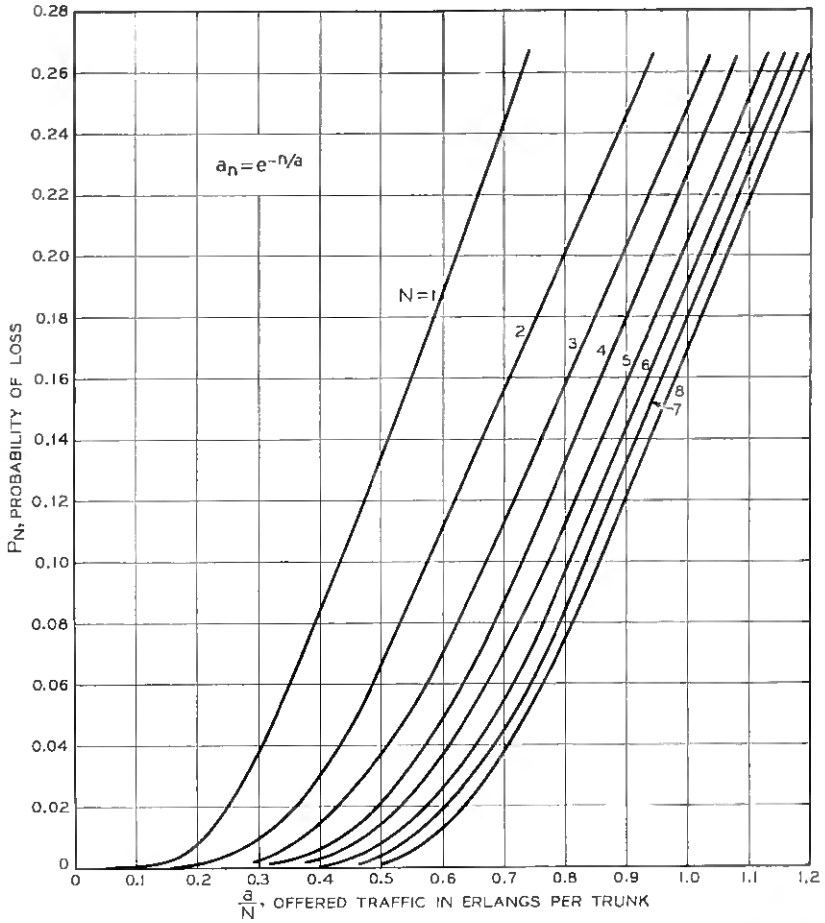


Fig. 4 — The probability of loss as a function of load per trunk for regular arrivals, $a_n = e^{-n/a}$.

The variance associated with $\{p_n\}$ is then

$$\sigma^2 = R(0) = m_2 - m_1^2 = 2b_2 + b_1 - b_1^2,$$

where $R(n)$ is the covariance function.

Because of the bias introduced by defining $N(k)$ to be the number of busy trunks found by the kth arriving customer, it is not in general true that m_1 equals lim $E\{N(t)\}$ as t tends to ∞, even when this limit exists. In Fig. 5, the ratio

$$\frac{m_1}{N} = \text{fraction trunks found busy}$$

$$= \frac{\text{expected number found busy by an arrival}}{\text{number of trunks}}$$

is plotted as a function of offered load a for Poisson arrivals. In Figs. 6 and 7 the same ratio is plotted for uniformly distributed interarrivals, and regular arrivals, respectively.

In Figs. 8, 9 and 10, the variance σ^2 of $N(k)$ in equilibrium is shown plotted against the offered load a for Poisson arrivals, uniformly distributed interarrivals and regular arrivals, respectively. The variance is also the value of the covariance function $R(n)$ for $n = 0$. In all cases, as the load a increases, the variance increases to a unique maximum, and then decreases to zero.

VI. BOUNDS FOR, AND APPROXIMATIONS TO, p_N FOR GIVEN $a_1, \cdots, a_N$

This section is devoted to inequalities which may be useful in estimating the loss probability p_N without too much computation. Since $1 > a_1 > \cdots > a_N$, we have

$$\frac{1 - a_n}{a_n} < \frac{1 - a_{n+1}}{a_{n+1}},$$

so that, from (7), we find

$$\sum_0^N \binom{N}{j} \left(\frac{1 - a_1}{a_1}\right)^j \leqq {p_N}^{-1} \leqq \sum_0^N \binom{N}{j} \left(\frac{1 - a_N}{a_N}\right)^j.$$

This proves:

Theorem 3: *The probability p_N of loss satisfies* $(a_N)^N \leqq p_N \leqq (a_1)^N$.

To obtain a sharper result, write

$${p_N}^{-1} = (a_1 a_2 \cdots a_N)^{-1} \sum_0^N \binom{N}{j} (1 - a_1) \cdots (1 - a_j) a_{j+1} \cdots a_N.$$

Then, in view of $1 > a_1 > \cdots > a_N$,

$$\sum \binom{N}{j} (1 - a_1)^j (a_N)^{N-j} \leqq \frac{a_1 a_2 \cdots a_N}{p_N} \leqq \sum \binom{N}{j} (1 - a_N)^j (a_1)^{N-j}.$$

From this we conclude:

Theorem 4: *The probability p_N of loss satisfies*

$$(1 - a_1 + a_N)^{-N} \leqq \frac{p_N}{a_1 a_2 \cdots a_N} \leqq (1 + a_1 - a_N)^{-N}.$$

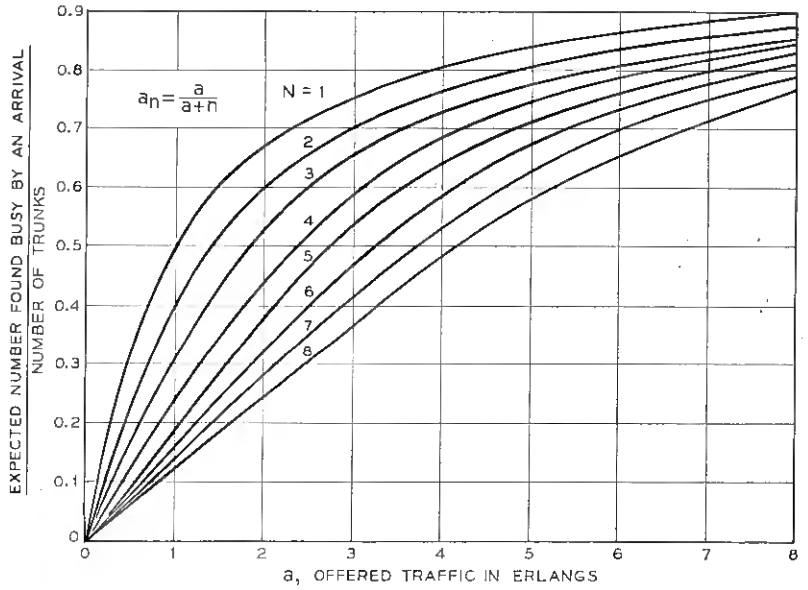


Fig. 5 — $\lim_{k\to\infty} E\{N(k)\}/N = m_1/N$ as a function of offered traffic a for Poisson arrivals.

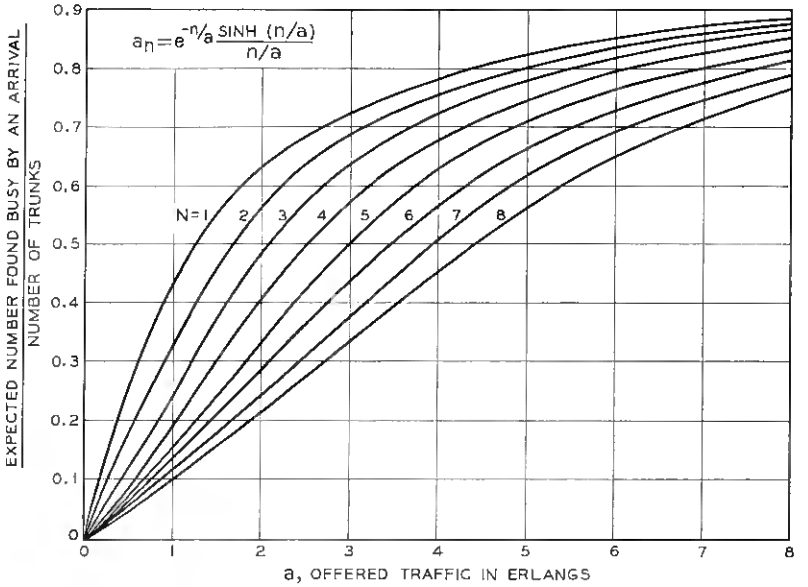


Fig. 6 — $\lim_{k\to\infty} E\{N(k)\}/N = m_1/N$ as a function of offered traffic a for uniformly distributed interarrivals.

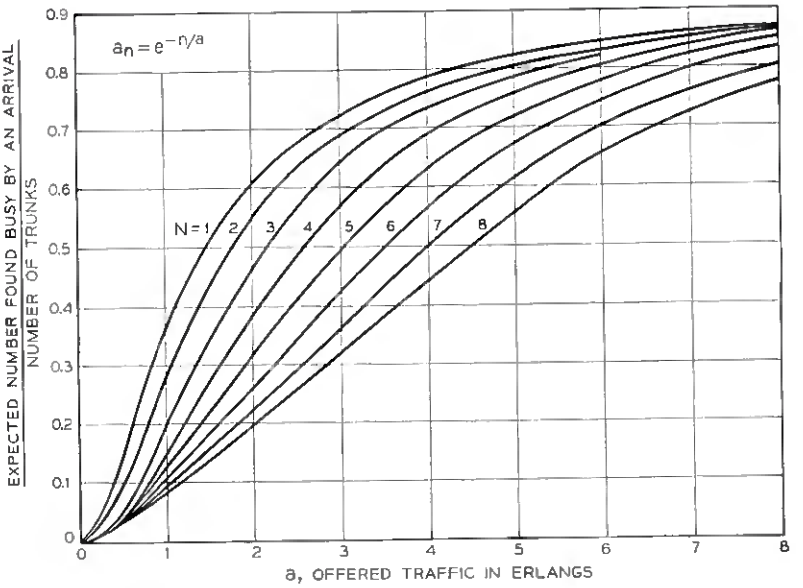


Fig. 7 — $\lim_{k\to\infty} E\{N(k)\}/N = m_1/N$ as a function of offered traffic a for regular arrivals.

This result suggests that, if $a_1 - a_N$ is sufficiently small, then the product $a_1a_2 \cdots a_N$ can serve as an approximation to p_N. There are cases, to be exemplified later, in which this is a good approximation. However, the next theorem shows that the product $a_1a_2 \cdots a_N$ always *underestimates* the loss.

Theorem 5: *For* $N = 1$, $p_N = a_1$; *for* $N \geqq 2$, $p_N > a_1a_2 \cdots a_N$.*
To prove this, we write p_N in the notation of Lemma 1 as

$$p_N = \frac{a_1a_2 \cdots a_N}{D_N(a_1, \cdots, a_N, 1)},$$

so that it suffices to prove that $D_N(a_1, \cdots, a_N, 1) < 1$. We shall actually prove the stronger result that $D_N(a_k, \cdots, a_{k+N-1}, 1) < 1$ for $k \geqq 1$. First we note

$$\begin{aligned} D_2(a_k, a_{k+1}, 1) &= a_ka_{k+1} + 2(1 - a_k)a_{k+1} + (1 - a_k)(1 - a_{k+1}) \\ &= 1 - a_k + a_{k+1} < 1. \end{aligned}$$

Now, because $1 > a_1 > \cdots > a_k > \cdots$, we find

$$\frac{D_N(a_k, \cdots, a_{k+N-1}, 1)}{a_ka_{k+1} \cdots a_{k+N-1}} < \frac{D_N(a_{k+1}, \cdots, a_{k+N}, 1)}{a_{k+1}a_{k+2} \cdots a_{k+N}}.$$

* A. J. Goldstein has pointed out that Theorem 4 implies directly that $p_N > a_1a_2 \cdots a_N$ for $N \geqq 2$.

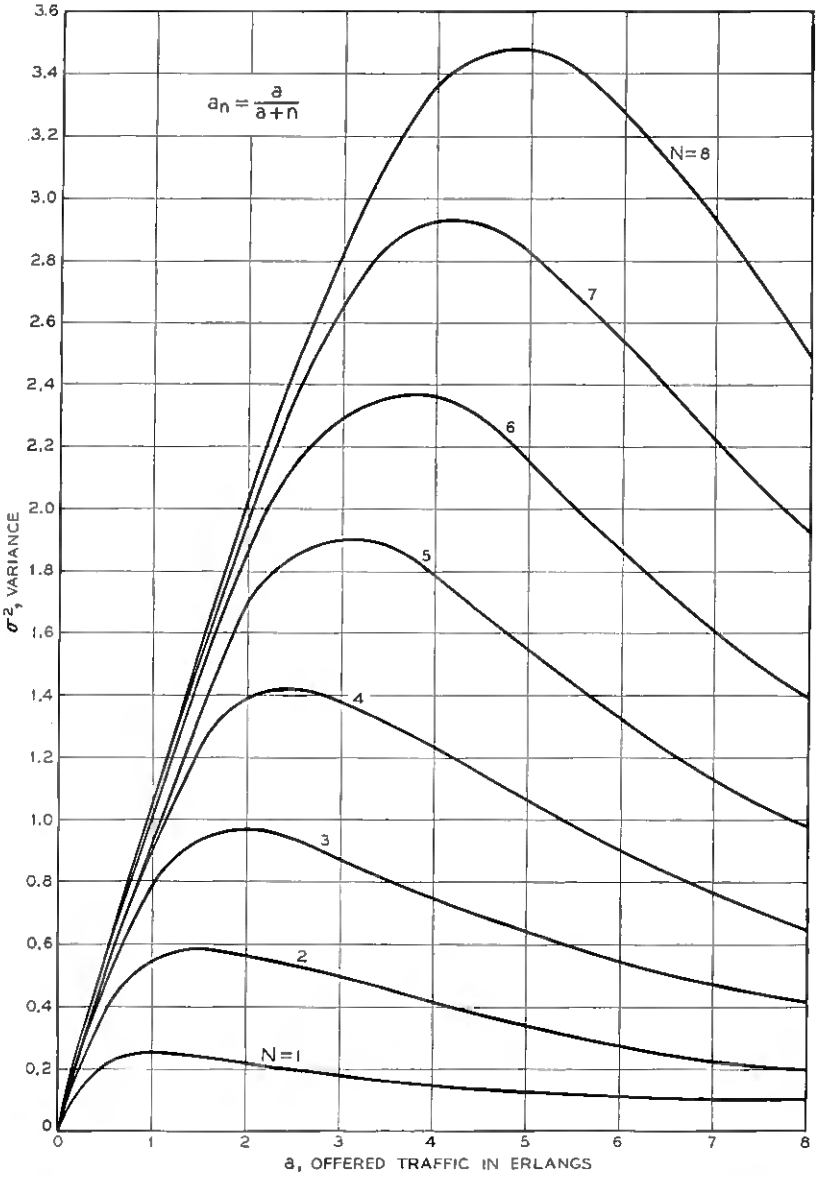


Fig. 8 — The variance σ^2 [= $R(O)$] of $N(k)$ in equilibrium for Poisson arrivals.

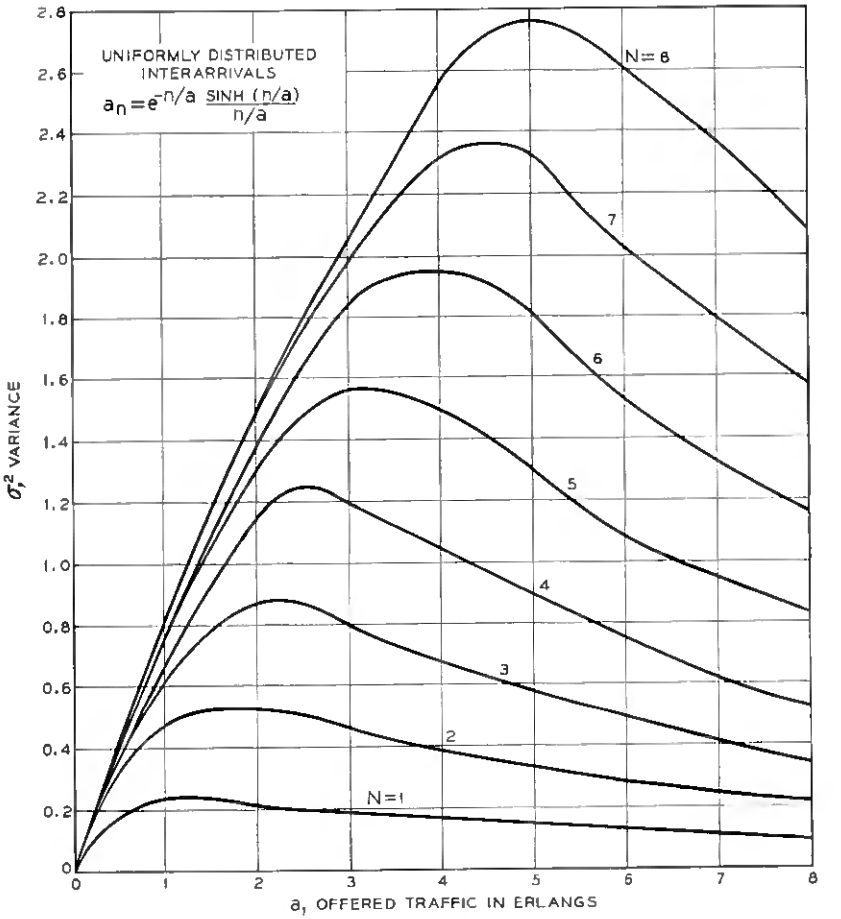


Fig. 9 — The variance σ^2 [$= R(0)$] of $N(k)$ in equilibrium for uniformly distributed interarrivals.

Therefore, the recurrence of Lemma 1 gives, for $z = 1$,

$$D_{N+1}(a_k , \cdots , a_{k+N} , 1) < D_N(a_{k+1} , \cdots , a_{k+N} , 1) < 1,$$

and the result follows by induction.

We now discuss the approximation $p_N \sim a_1 a_2 \cdots a_N$. Since $1 > a_1 > a_N$, two cases in which $a_1 - a_N$ is small are as follows: (a) a_1 is close to 0 and p_N is very small; (b) a_N is close to 1 and p_N is very high. The quantity $a_1 - a_N$ determines the excellence of the approximation, as meas-

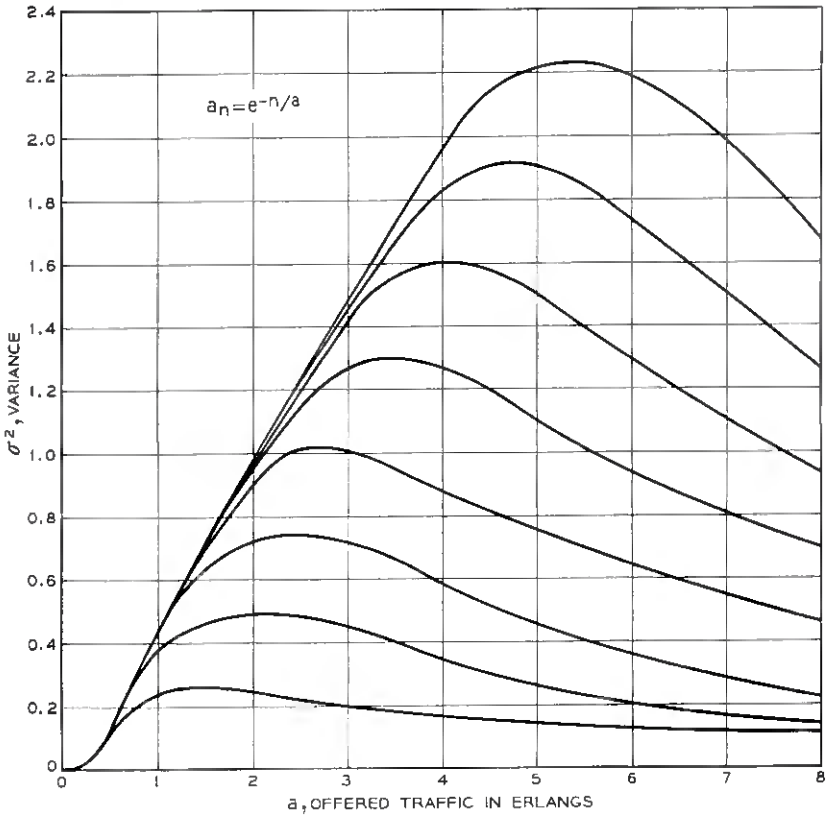


Fig. 10 — The variance σ^2 [$= R(0)$] of $N(k)$ in equilibrium for regular arrivals.

ured by Theorem 4. The value of $a_1 - a_N$ may be estimated from below in terms of a_1 alone by the inequality $a_1 - a_N \geqq a_1 - a_1^N$. From Theorems 4 and 5 we see that

$$(1 - a_1 + a_N)^N \leqq r = \frac{a_1 a_2 \cdots a_N}{p_N} < 1,$$

and this inequality indicates the values of $a_1 - a_N$ for which $p_N \sim a_1 a_2 \cdots a_N$ is justified.

To put the matter more intuitively, we note that a_1 is the chance that a conversation, in progress at one arrival epoch, is still in progress at the next arrival epoch, i.e.,

$$a_1 = \Pr\{\text{holding time} > \text{interarrival time}\}.$$

Similarly, if $h_1, \cdots, h_N$ are N (independent) holding times,

$$a_N = \Pr\left\{\min_{1 \leq j \leq N} h_j > \text{interarrival time}\right\}.$$

So the approximation is likely to be good at least when the chance that one holding time exceeds an interarrival time is not very different from the chance that each of N holding times exceeds an interarrival time (the same one for all N). As a tentative conclusion we may say that $p_N \sim a_1 a_2 \cdots a_N$ is good when the loss is very high or very low.

The ratio $r = a_1 a_2 \cdots a_N / p_N$ has been plotted as a function of the average offered traffic a in Figs. 11, 12 and 13 for Poisson arrivals, uniformly distributed interarrivals and regular arrivals, respectively. The curves bear out the conclusions of the previous paragraph, that the

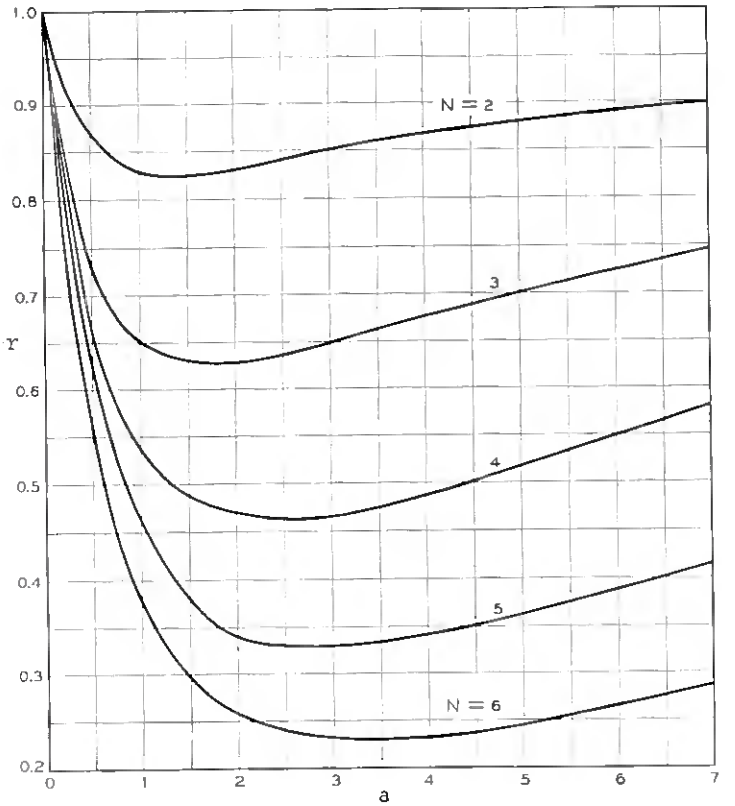


Fig. 11 — The ratio $r = (a_1 a_2 \cdots a_N)/p_N$ as a function of traffic a for Poisson arrivals.

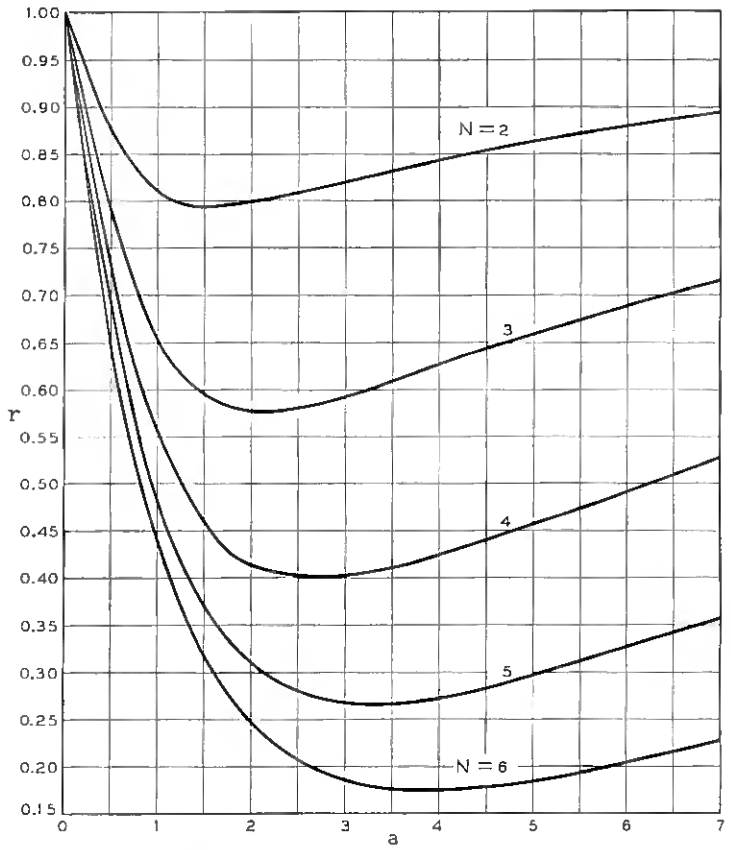


Fig. 12 — The ratio $r = (a_1 a_2 \cdots a_N)/p_N$ as a function of traffic a for uniformly distributed interarrivals.

approximation $p_N \sim a_1 a_2 \cdots a_N$ is good for low and high traffic. Fig. 14 shows a detail of r for very low traffic, for all cases at once.

Lemma 2: *For* $m, k \geqq 1$, $a_{k+1} + a_{k+m} \leqq a_k + a_{k+m+1}$.

Proof: the case $m = 1$ holds by convexity; for the same reason,

$$a_k + a_{k+2} \geqq 2a_{k+1}.$$

Assume that the lemma holds for a given m and all $k \geqq 1$. Then

$$a_{k+2} + a_{k+1+m} \leqq \frac{a_k + a_{k+2}}{2} + a_{k+2+m},$$

$$a_{k+1} + a_{k+m+1} \leqq a_{k+1} + \frac{a_k - a_{k+2}}{2} + a_{k+m+2}.$$

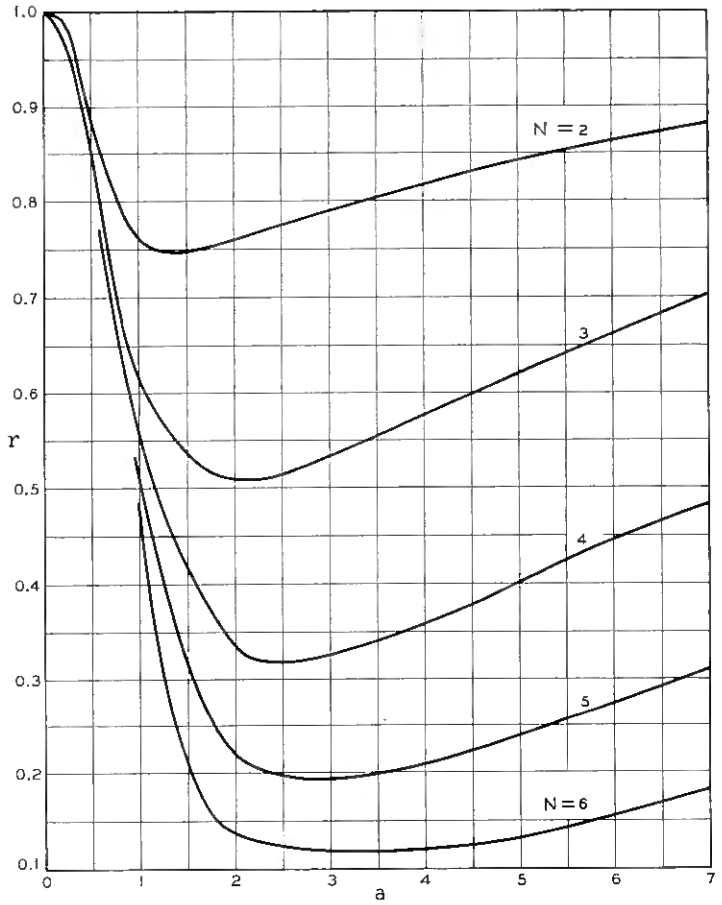


Fig. 13 — The ratio $r = (a_1 a_2 \cdots a_N)/p_N$ as a function of traffic a for regular arrivals.

But $a_k + a_{k+2} \geqq 2a_{k+1}$ implies

$$a_{k+1} + \frac{a_k - a_{k+2}}{2} \leqq a_k ,$$

so the lemma follows by induction.

Theorem 6: *Let*

$$L_k^{(N)} = \prod_{m=1}^{N} [1 - a_k + a_{k+m}] ,$$

$$U_k^{(N)} = \prod_{j=0}^{N-1} [1 - a_{k+j} + a_{k+N}] .$$

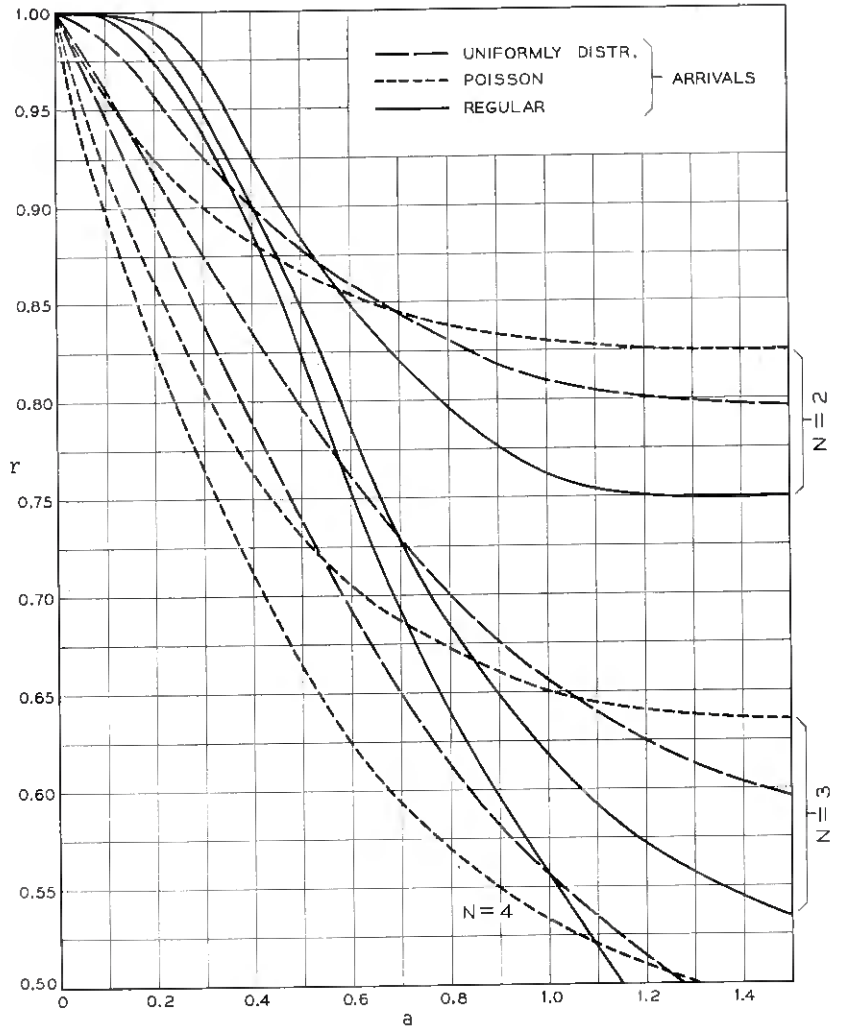


Fig. 14 — Detail of $r = (a_1 a_2 \cdots a_N)/p_N$ for low traffic a for Poisson arrivals, uniformly distributed interarrivals and regular arrivals.

Then

$$L_k^{(N-1)} \leqq D_N(a_k , \cdots , a_{k+N-1} , 1) \leqq U_k^{(N-1)}$$

and, also, the chance of loss satisfies

$$\frac{a_1 a_2 \cdots a_N}{U_1^{(N-1)}} \leqq p_N = \frac{a_1 a_2 \cdots a_N}{D_N(a_1 , a_2 , \cdots , a_N , 1)} \leqq \frac{a_1 a_2 \cdots a_N}{L_1^{(N-1)}} .$$

Proof: For $N = 2$, we have for $k \geqq 1$:

$$D_2(a_k , a_{k+1} , 1) = 1 - a_k + a_{k+1} = U_k^{(1)} = L_k^{(1)}.$$

Now, assume that for all $k \geqq 1$

$$L_k^{(N-1)} \leqq D_N(a_k , \cdots , a_{k+N-1} , 1) \leqq U_k^{(N-1)}.$$

Then, by Lemma 1,

$$a_{k+N} L_k^{(N-1)} + (1 - a_k) L_{k+1}^{(N-1)} \leqq D_{N+1}(a_k , \cdots , a_{k+N} , 1) \leqq a_{k+N} U_k^{(N-1)} + (1 - a_k) U_{k+1}^{(N-1)}.$$

By convexity and Lemma 2,

$$L_{k+1}^{(N-1)} \geqq L_k^{(N-1)},$$
$$U_{k+1}^{(N-1)} \geqq U_k^{(N-1)}.$$

Therefore

$$L_k^{(N-1)}(1 - a_k + a_{k+N}) \leqq D_{N+1}(a_k , \cdots , a_{k+N} , 1) \leqq (1 - a_k + a_{k+N}) U_{k+1}^{(N-1)}.$$

But

$$(1 - a_k + a_{k+N}) L_k^{(N-1)} = L_k^{(N)},$$
$$(1 - a_k + a_{k+N}) U_{k+1}^{(N-1)} = U_k^{(N)},$$

so the theorem follows by induction.

VII. BOUNDS AND APPROXIMATIONS WHEN ARRIVALS ARE REGULAR

In telephony, it is unrealistic to expect regular arrivals. Nevertheless, the results of Section VIII indicate that regularity of arrivals represents in a definite sense a limiting best case, for which the loss assumes a lower bound. For this reason we devote some effort to approximating the loss p_N in this case.

For regular arrivals the loss p_N is given by

$$(p_N)^{-1} = \sum_j \binom{N}{j} \frac{1 - x}{x} \frac{1 - x^2}{x^2} \cdots \frac{1 - x^j}{x^j},$$

where $x = \exp\{-(1/a)\}$ and a = offered erlangs. A simple procedure for obtaining an upper bound on p_N is as follows: we note that, since $x < 1$,

$$(p_N)^{-1} = \sum_j \binom{N}{j} (1 - x)^j (1 + x^{-1})(1 + x^{-1} + x^{-2}) \cdots \left(\sum_0^{j-1} x^{-1}\right)$$

$$\geqq \sum_j \binom{N}{j} (1 - x)^j j! = N!(1 - x)^N \sum_j \frac{(1 - x)^{j-N}}{(N - j)!}.$$

The term on the right of the last inequality is seen to be the reciprocal of $B[N, 1/(1 - x)]$, where $B(c, a)$ is the classical Erlang B function; that is,

$$B(c, a) = \frac{\dfrac{a^{c/N}}{(c/N)!}}{\sum_{j=0}^{c/N} \dfrac{a^j}{j!}} = \frac{e^{-a}}{1 - P(c + 1, a)},$$

where $P(c, a)$ is the cumulative term $\sum_{n \geq c} a^n e^{-a}/n!$ of the Poisson distribution. This proves:

Theorem 7: *If arrivals are regular and a erlangs are offered, then* $p_N \leqq B(N, \eta)$, *where B is Erlang's function, and* $\eta = (1 - x)^{-1} = (1 - e^{-1/a})^{-1}$.

From Theorem 9 of Section VIII we know that $p_N \leqq B(N, a)$; that is, we overestimate the loss for regular arrivals if we pretend that arrivals are Poisson. Let us therefore see whether the bound of Theorem 7 is better than $B(N, a)$. Let $a = (1 - \zeta)^{-1}$, so that $\eta = (1 - e^{\zeta-1})^{-1}$. Now ζ is tangent to $e^{\zeta-1}$ at $\zeta = 1$, i.e., at $a = \infty$, and $e^{\zeta-1}$ is convex; hence $e^{\zeta-1} \geqq \zeta$, and $1 - \zeta \geqq 1 - e^{\zeta-1}$, so that

$$a = (1 - \zeta)^{-1} < (1 - e^{\zeta-1})^{-1} = \eta$$

for finite a. Since B is monotone increasing in the offered erlangs we conclude that $B(N, a) < B(N, \eta)$. Thus the bound of Theorem 7 is nowhere as good as the overestimate $B(N, a)$ for p_N.

However, there is a systematic way of obtaining a useful upper bound on p_N for regular arrivals. This bound again has the functional form of Erlang's formula $B(N, \eta)$. However, η, instead of being chosen equal to a, is chosen to correspond to a Poisson process, which gives the right value of a_1, $\exp\{-(1/a)\}$, and involves fewer offered erlangs $\eta < a$. Now

$$a_1 = \begin{cases} e^{-1/a} \text{ for regular arrivals at } a \text{ erlangs} \\ \eta/(1 + \eta) \text{ for Poisson arrivals at } \eta \text{ erlangs.} \end{cases}$$

So η erlangs will give the right value of a_1 if and only if

$$\eta = \frac{y}{1-y} \quad \text{for} \quad y = e^{-1/a}.$$

We first show that, if η is defined in this way, then $\eta < a$. For $u > 0$, we have $u + 1 < e^u$, so that for $u = a^{-1}$ we find

$$e^{-1/a} < a(1 - e^{-1/a}),$$

$$\eta = \frac{y}{1-y} < a \quad \text{for} \quad y = e^{-1/a}.$$

For this choice of η, then, $B(N, \eta) < B(N, a)$. Now, from formula (7), it is apparent that if the a_i are replaced term by term with quantities a_i', with $a_i \leqq a_i'$, the result will be $\geqq p_N$. We choose

$$a_i' = \eta/(\eta + i), \qquad i = 1, 2, \cdots, N.$$

The a_i' correspond to Poisson arrivals with η erlangs offered. To obtain a bound it remains to be shown that, for $i = 2, 3, \cdots, N$,

$$a_i = e^{-i/a} \leqq a_i' = \frac{\eta}{\eta + i}.$$

This is equivalent to

$$y + i = iy \leqq y^{1-i}, \quad \text{for} \quad y = e^{-1/a},$$

which is seen to be true because $y + i - iy$ is tangent to y^{1-i} at $y = 1$. The result of replacing a_i by the chosen a_i' is just $B(N, \eta)$. This proves:

Theorem 8: *If arrivals are regular and a erlangs are offered, then* $p_N \leqq B(N, \eta) < B(N, a)$, *where B is Erlang's function, and*

$$\eta = \frac{y}{1-y} = \frac{e^{(-1/a)}}{1 - e^{(-1/a)}}.$$

This result suggests use of $B(N, \eta)$ as an approximation to p_N. Two numerical cases illustrate this approximation:

i. $N = 8$, 8 erlangs are offered; then $y = e^{-0.125}$ and $\eta = 0.747$. We find $p_N = 0.17$, $B(N, \eta) = 0.20$, $B(N, a) = 0.235$.

ii. $N = 5$, 8 erlangs are offered; again, $\eta = 0.747$, and $p_N = 0.437$, $B(N, \eta) = 0.450$, $B(N, a) = 0.478$.

VIII. THE LOSS AS A FUNCTIONAL OF $A(u)$

For each N, and each hang-up rate γ, the loss p_N can be regarded as a mapping from the set of distributions $A(u)$ of positive variates to the interval (0, 1). We write $p_N(A)$ in this section for the loss resulting from

the interarrival distribution $A(u)$, and we study the loss as a functional of $A(u)$.

First, it is instructive to keep the mean interarrival time μ_1 fixed and to vary the interarrival distribution $A(u)$. Since $e^{-n\gamma u}$ is convex in u, we find

$$e^{-n\gamma\mu_1} \leqq \int_0^\infty e^{-n\gamma u}\, dA(u) = a_n\,,$$

and hence

$$\frac{1 - e^{-n\gamma\mu_1}}{e^{-n\gamma\mu_1}} \geqq \frac{1 - a_n}{a_n}.$$

But $e^{-n\gamma\mu_1} = a_n$ for the case where arrivals are regular, and μ_1 apart. This proves:

Theorem 9: *If* γ, μ_1 are positive constants, then

$$\inf_A \left\{ p_N(A) \,\middle|\, \int u\, dA(u) = \mu_1 \right\}$$

is achieved for the unit step distribution

$$A(u) = \begin{cases} 1, & u \geqq \mu_1\,, \\ 0, & u < \mu_1\,. \end{cases}$$

Thus, the probability of loss assumes a minimum, for fixed γ and μ_1, when the arrivals are regularly spaced at epochs μ_1 apart.

We next show that, if the mean interarrival time μ_1 and the hang-up rate γ are kept fixed, then the probability of loss can still be made arbitrarily close to unity by a proper choice of $A(u)$.

Theorem 10: *If* γ, μ_1 *are positive constants, then*

$$\sup_A \left\{ p_N(A) \,\middle|\, \int u\, dA(u) = \mu_1 \right\} = 1.$$

To prove this, let $1 > \epsilon > 0$ be given, and consider those distributions which have a mass $(1 - p)$ at $y_0 > 0$, and a mass p at $y_1 > 0$. For such an $A(u)$ we have

$$a_n = (1 - p)e^{-n\gamma y_0} + pe^{-n\gamma y_1}.$$

Let $q = (1 - p)\exp\{-N\gamma y_0\}$, so that, for each n,

$$\frac{1 - q}{q} \geqq \frac{1 - a_n}{a_n}.$$

Then, since $a_N < a_{N-1} < \cdots < a_1 < 1$, we find from Palm's formula (7) that

$$p_N(A) \geqq \left(1 + \frac{1-q}{q}\right)^{-N}.$$

We can now choose p and y_0 so small that $p_N(A) \geqq 1 - \epsilon$, independently of y_1, which can then be chosen to satisfy $\mu_1 = y_0(1-p) + y_1 p$. This proves the theorem.

It is natural to use

$$\frac{1}{\mu_1} = \frac{1}{\int_0^\infty u\, dA(u)}$$

as a measure of the calling rate, and to use

$$r_N(A) = \frac{1 - p_N(A)}{\mu_1} = \text{fraction served times calling rate}$$

as a measure of the rate of service, the rate at which calls are actually being completed. Suppose now that we are willing to tolerate a probability p of loss. Can we find an interarrival distribution $A(u)$ which achieves p and for which the rate of service is a maximum for a given hang-up rate γ? To answer this question, define the function

$$f(x_1, x_2, \cdots, x_N) = 1 + \binom{N}{1}\frac{1-x_1}{x_1} + \cdots + \binom{N}{N}\frac{(1-x_1)\cdots(1-x_N)}{x_1 \ldots x_N},$$

so that $p_N(A) = [f(a_1, a_2, \cdots, a_N)]^{-1}$.

Theorem 11: *If* $\gamma > 0$ *and* $0 < p < 1$, *then*

$$\sup_A \{r_N(A) \mid p_N(A) = p\} = \frac{\gamma(1-p)}{-\log x},$$

where x *is the unique solution of the equation* $f(x, x^2, \cdots, x^N) = p^{-1}$ *in the unit interval. The supremum is achieved by the unit step distribution* $A(u)$ *defined by*

$$A(u) = \begin{cases} 1, & u \geqq -\gamma^{-1} \log x, \\ 0, & u < -\gamma^{-1} \log x. \end{cases} \tag{9}$$

The function $f(x, x^2, \cdots, x^N)$ is monotone, decreasing from ∞ to 1 in the unit interval. Since f is continuous, and $0 < p < 1$, there exists a solution x of the equation $f(x, x^2, \cdots, x^N) = p^{-1}$. Obviously, for $A(u)$ defined by (9), we have $p_N(A) = p$. Now let $B(u)$ be any other interarrival distribution with a finite mean, so that the service rate $r_N(B)$

exists. Suppose that $B(u)$ achieves the probability p of loss; i.e., that $p_N(B) = p$. We show that

$$\int_0^\infty u\, dB(u) \geqq -\gamma^{-1} \log x.$$

For, suppose the contrary and set

$$y = \exp\left\{-\gamma \int_0^\infty u\, dB(u)\right\};$$

then, by Theorem 9, $[f(y, y^2, \cdots, y^N)]^{-1} \leqq p$, and

$$\int_0^\infty u\, dB(u) < -\gamma^{-1} \log x$$

implies $y > x$, so that $[f(x, x^2, \cdots, x^N)]^{-1} < [f(y, y^2, \cdots, y^N)]^{-1} \leqq p$, which is impossible. This proves that

$$\inf_B \left\{ \int u\, dB(u) \,\middle|\, p_N(B) = p \right\} = -\gamma^{-1} \log x,$$

and also Theorem 11. Note that the supremum in Theorem 11 is a linear function of the hang-up rate, γ.

Let $N(t)$ be the number of trunks busy at time t, and let $E\{N(t)\}$ be its average. It is not always true that $\lim E\{N(t)\}$ exists as $t \to \infty$. However, if $A(u)$ is not a lattice distribution, then

$$\lim_{t\to\infty} E\{N(t)\} = \frac{1 - p_N(A)}{\gamma \mu_1},$$

where μ_1 may be ∞. This limit is the number of erlangs carried by the trunk group in equilibrium (see Takács[4]). Now a lattice distribution can be approximated arbitrarily closely by absolutely continuous distributions. Thus, an immediate consequence of Theorem 11 is:

Theorem 12: *If* $0 < p < 1$, *and* x *is as in Theorem* 11, *then*

$$\sup_A \lim_{t\to\infty} E\{N(t)\} = \frac{1 - p}{-\log x},$$

where the supremum is taken over $A(u)$ *such that* $p_N(A) = p$ *and such that lim* $E\{N(t)\}$ *as* $t \to \infty$ *exists.*

This theorem means, intuitively, that the maximum number of erlangs that N trunks can carry at a fixed loss probability p [the maximum being over the appropriate $A(u)$ which achieve a loss p] is a number depending only on N and p.

It may also be of interest sometimes to know what is the least probability of loss incurred by offering a traffic of a erlangs to N trunks, with $A(u)$ being varied, and γ as well. The answer is given by:

Theorem 13: *If* $a > 0$, *then*

$$\inf_{A,\gamma} \left\{ p_N(A) \,\middle|\, \int \gamma u \, dA(u) = a^{-1} \right\} = [f(x, x^2, \cdots, x^N)]^{-1},$$

where $x = e^{-1/a}$, *and the inf is achieved by any unit step distribution* $A(u)$ *and* $\gamma > 0$ *such that*

$$A(u) = \begin{cases} 1, & u \geqq (a\gamma)^{-1}, \\ 0, & u < (a\gamma)^{-1}. \end{cases}$$

The proof is essentially that of Theorem 9, and is omitted.

IX. $\Pr\{N(k) = N\}$ AS A FUNCTION OF k

The time-dependent behavior of the process $N(k)$ is only touched on here, since a complete treatment requires the detailed investigation of the roots of the polynomial $D_N(a_1, a_2, \cdots, a_N, z)$ occurring in the generating function $\psi_N(z)$. Such a study is still incomplete.

Nevertheless, some hints of the rate of approach to the limit p_N can be obtained from Theorem 1 and $\psi_N(z)$ as they stand. For instance, if $N(0) = 0$, then

$$\psi_N(z) = \frac{(1 - z)^{-1} a_1 a_2 \cdots a_N z^N}{\sum_{i=0}^{N} \binom{N}{i} (1 - a_1 z) \cdots (1 - a_i z) a_{i+1} \cdots a_N z^{N-i}}.$$

From this it can be seen directly that

$$\Pr\{N(k) = N \mid N(0) = 0\} = \begin{cases} 0 & \text{for } k < N, \\ a_1 a_2 \cdots a_N & \text{for } k = N, \\ a_1 a_2 \cdots a_N \left[1 + \sum_{j=1}^{N-1} (a_j - a_N) \right] & \text{for } k = n + 1. \end{cases} \tag{10}$$

More terms may be computed from the generating function, but the labor involved increases rapidly. It is to be noted that (10), together with Theorem 5, suggests that the approach to p_N is monotone; also, the first nonzero term is the approximating product $a_1 a_2 \cdots a_N$ discussed in Section VI.

For $N = 2$ trunks, it is possible to discuss $\Pr\{N(k) = N \mid N(0)\}$ in a particularly simple way. The results are given here, together with a numerical illustration, for the light they shed on the time development of the process. From Theorem 1 we find that

$$\sum_k z^k \Pr\{N(k) = 2 \mid N(0) = 0\} = \frac{a_1 a_2 z^2}{(1 - z)(1 - za_1 + za_2)},$$

so that

$$\Pr\{N(k) = 2 \mid N(0) = 0\} = \begin{cases} 0 & \text{for } k = 0, 1 \\ \dfrac{a_1 a_2}{1 - a_1 + a_2}[1 - (a_1 - a_2)^{k-1}] & \text{for } k \geqq 2. \end{cases}$$

Here p_N is $a_1 a_2/(1 - a_1 + a_2)$, and is approached exponentially.

Similarly, the generating function of $\Pr\{N(k) = 2 \mid N(0) = 1\}$ is

$$\frac{a_1 a_2 z^2}{(1 - z)(1 - za_1 + za_2)} + \frac{za_2}{1 - za_1 + za_2},$$

so that

$$\Pr\{N(k) = 2 \mid N(0) = 1\} = \begin{cases} 0 & \text{for } k = 0 \\ a_2 & \text{for } k = 1 \\ \dfrac{a_1 a_2}{1 - a_1 + a_2}[1 - (a_1 - a_2)^{k-1}] + a_2(a_1 - a_2)^{k-1} & \text{for } k \geqq 2. \end{cases}$$

Finally, the generating function of $\Pr\{N(k) = 2 \mid N(0) = 2\}$ is

$$\frac{a_1 a_2 z^2}{(1 - z)(1 - za_1 + za_1)} + \frac{za_2}{1 - za_1 + za_2} + 1,$$

from which we find

$$\Pr\{N(k) = 2 \mid N(0) = 2\} = \begin{cases} 1 & \text{for } k = 0 \\ a_2 & \text{for } k = 1 \\ \dfrac{a_1 a_2}{1 - a_1 + a_2}[1 - (a_1 - a_2)^{k-1}] + a_2(a_1 - a_2)^{k-1} & \text{for } k \geqq 2. \end{cases}$$

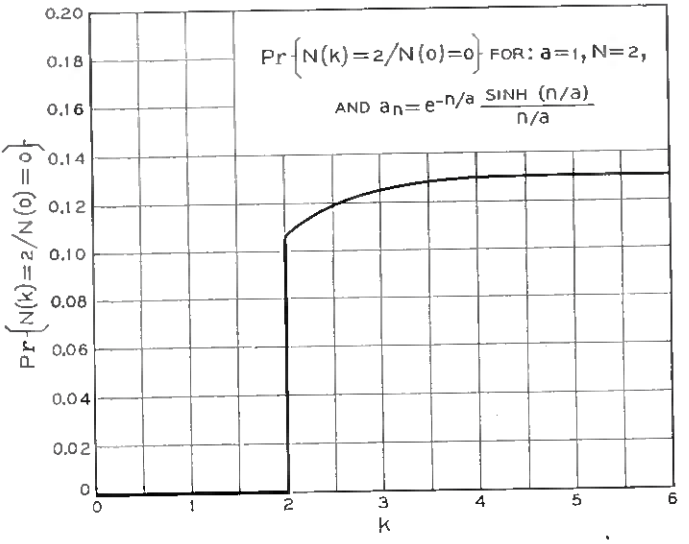


Fig. 15 — $\Pr\{N(k) = 2|N(O) = 0\}$ for $a = 1$ erlang, $N = 2$ trunks and uniformly distributed interarrivals.

This agrees with the previous conditional probability for $k \geqq 1$, as it should.

The three conditional probabilities $\Pr\{N(k) = 2 \mid N(0) = m\}$ for $m = 0, 1, 2$ have been plotted as functions of k for uniformly distributed interarrivals in Figs. 15, 16 and 17, respectively. The probabilities have been drawn continuously, but of course the functions are only defined for integers k. The example chosen exhibits a very rapid approach to equilibrium in terms of numbers of arriving calls, since the third arriving call finds essentially the equilibrium situation.

X. THE EXPECTATION OF $N(k)$ AND THE COVARIANCE

The next result gives a formula for the mean value $E\{N(k)\}$ in terms of the initial value $E\{N(0)\}$, and the probabilities $\Pr\{N(j) = N\}$ for $j \leqq k - 1$.

Theorem 14: *The mean value of* $N(k)$ *is*

$$E\{N(k)\} = \frac{a_1(1 - a_1^k)}{1 - a_1} + a_1^k E\{N(0)\} - \sum_{j=0}^{k} a_1^{j+1} \Pr\{N(k - j - 1) = N\}.$$

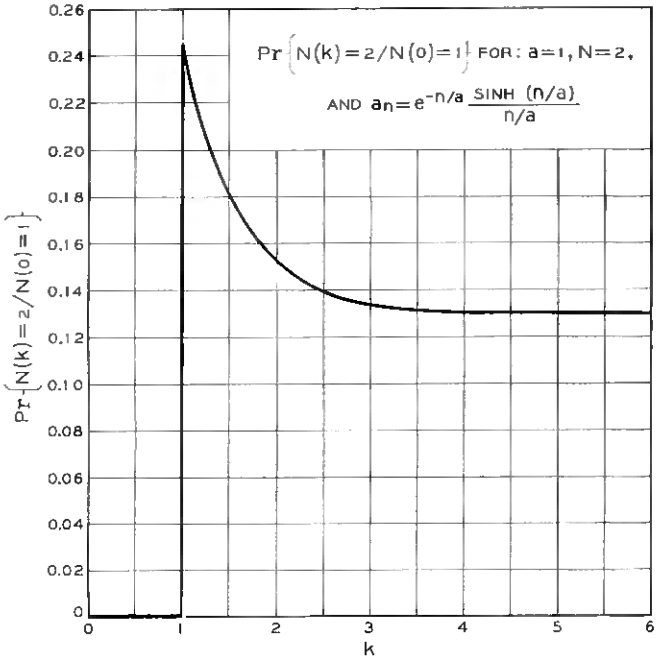


Fig. 16 — $\Pr\{N(k) = 2|N(O) = 1\}$ for $a = 1$ erlang, $N = 2$ trunks and uniformly distributed interarrivals.

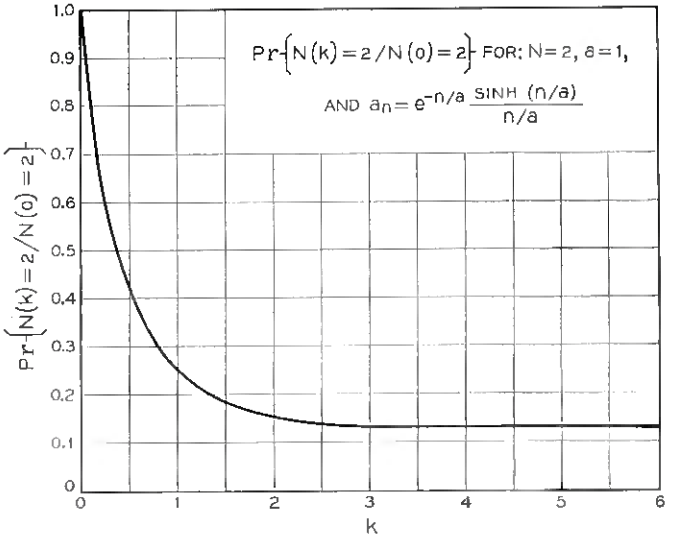


Fig. 17 — $\Pr\{N(k) = 2|N(O) = 2\}$ for $a = 1$ erlang, $N = 2$ trunks and uniformly distributed interarrivals.

To prove this we first obtain the generating function $b_1(z)$, either by differentiation from (1) or directly from the recurrences (4). We find

$$\sum_{k\geqq 0} z^k E\{N(k)\} = b_1(z) = \frac{za_1[(1-z)^{-1} - \psi_N(z)] + E\{N(0)\}}{1 - za_1},$$

and this gives Theorem 14 upon expansion in powers of z.

We define the covariance function $R(n)$ of the random process $N(k)$ by

$$R(n) = \lim_{k\to\infty} E\{N(k)N(k+n)\} - E^2\{N(k)\}.$$

From Theorem 14 we can derive a formula for the covariance function $R(n)$.

Theorem 15: *If* $\psi_{n,N}(z)$ *is* $\psi_N(z)$ *for the initial condition* $N(0) = n$, $|z| < 1$, $\{p_m\}$ *is the stationary distribution of* $N(k)$, *and* $m_i = \sum_m m^i p_m$ *for* $i = 1, 2$, *then*

$$\sum_{n\geqq 0}^{\infty} z^n R(n) = \sum_m p_m m \left\{\frac{za_1[(1-z)^{-1} - \psi_{m,N}(z)] + m}{1 - za_1}\right\} - (1-z)^{-1}\left(\frac{a_1 - a_1 p_N}{1 - a_1}\right)^2,$$

and

$$R(n) = R(-n) = \frac{m_1(a_1 - a_1^{n+1})}{1 - a_1} + a_1^n m_2 - m_1^2 - \sum_m m p_m \sum_{j=0}^{n-1} a_1^{j+1} \Pr\{N(n-j-1) = N \mid N(0) = m\}.$$

Before developing the results of Theorem 15 into a form useful for computation, we shall sketch the reasons for interest in the covariance function $R(n)$. The function expresses quantitatively the cohesiveness of the process, the extent to which $N(k+n)$ and $N(k)$ are correlated. Besides this theoretical role, the covariance is involved in the practical matter of evaluating (theoretically) the sampling error in a certain kind of switch count (traffic measurement.) For a concrete example, suppose that

$$S = \sum_1^n N(k)$$

is used to estimate the average traffic encountered by arriving custom-

ers. Here n is the number of successive observations of the random process $N(k)$. The variance of S is

$$\begin{aligned}\text{var}\{S\} &= E\left\{\sum_i \sum_j N(i)N(j)\right\} - n^2 m_1^2 \\ &= \sum_i \sum_j \text{cov}\{N(i), N(j)\} \\ &= \sum_i \sum_j R(i-j) \\ &= nR(0) + 2\sum_{j=1}^{n-1} (n-j)R(j),\end{aligned}$$

where we have assumed that the observations began in a condition of equilibrium. Thus var$\{S\}$ can be expressed in terms of the covariance function $R(n)$.

The formula for $R(n)$ can be made more useful for computation by turning it into a recurrence relation for successive values of a certain linear function of $R(n)$. We define auxiliary quantities Q_k by

$$Q_k = \sum_{m=0}^{N} m p_m \Pr\{N(k) = N \mid N(0) = m\} \tag{11}$$

and note that

$$R(n) + m_1^2 - \frac{m_1 a_1}{1 - a_1} = a_1^n \left(m_2 - \frac{m_1 a_1}{1 - a_1}\right) - \sum_{j=0}^{n-1} a_1^{j+1} Q_{n-j-1}.$$

Hence also

$$\begin{aligned}R(n+1) + m_1^2 - \frac{m_1 a_1}{1 - a_1} &= a_1 \left\{a_1^n \left(m_2 - \frac{m_1 a_1}{1 - a_1}\right) - Q_n - \sum_{j=0}^{n-1} a_1^{j+1} Q_{n-j-1}\right\} \qquad (12) \\ &= a_1 \left\{R(n) + m_1^2 - \frac{m_1 a_1}{1 - a_1} - Q_n\right\}\end{aligned}$$

Thus, if the Q_k are known, the $R(n)$ may be calculated by a simple recursive procedure from $R(0)$, which is the variance. The calculation of the Q_k is simplified by the fact that, for small k, (a region of principal interest), many terms of the sum defining Q_k are 0. For example, if $0 \leqq m < N - k$, the conditional probability $\Pr\{N(-k) = N \mid N(0) = m\}$ is 0, since it is not possible for the kth man to find all trunks busy if the 0th man found fewer than $N - k$ busy. The first few correction

TABLE I. — $\Pr\{N(k) = N \mid N(0) = m\}$.

m	k			
	0	1	2	3
N	1	a_N	$Na_Na_{N-1} - (N-1)a_N^2$	$(N^2 - N + 1)a_N^3 + \left(N - 2N^2 + \frac{N^2 - N}{2}\right)a_{N-1}a_N^2 + Na_{N-1}^2a_N + \frac{N(N-1)}{2}a_{N-2}a_{N-1}a_N$
N-1	0	a_N	$Na_{N-1}a_N - (N-1)a_N^2$	same as above
N-2	0	0	$a_{N-1}a_N$	$a_{N-1}[a_{N-1} + (N-1)(a_{N-2} - a_N)]$
N-3	0	0	0	$a_{N-2}\,a_{N-1}\,a_N$

terms Q_k as defined above may be computed (by summation) for $k = 0, 1, 2, 3$ from Table I, which shows $\Pr\{N(k) = N \mid N(0) = m\}$, valid for $m \geqq 0$:

Curves of the covariance function $R(n)$ for $n = 1, 2$ and 3 are plotted as functions of the offered traffic a for trunk group sizes $N = 2, \cdots, 8$, as follows: in Figs. 18 through 20 for Poisson arrivals; in Figs. 21 through 23 for uniformly distributed interarrivals; and in Figs. 24 through 26 for regular arrivals. The curve for $N = 1$ is not shown in any of Figs. 18 through 26 because, in this case, $R(n) = 0$ for $|n| > 0$ (see below).

The following conclusions seem to be reasonable after examination of the curves:

i. $R(n)$ is nonnegative and monotone decreasing in $|n|$.

ii. For n and traffic a fixed, the covariance $R(n)$ for Poisson arrivals exceeds the covariance $R(n)$ for both the other two interarrival distributions (uniform and fixed) we have considered. Similarly, the covariance $R(n)$ for regular arrivals falls below the value of $R(n)$ for both Poisson arrivals and uniform interarrivals. We conjecture that $R(n)$ for regular arrivals is less than or equal to $R(n)$ for any other distribution of interarrivals, for the same traffic.

A particularly simple but important case arises when $N = 1$; the case is simple because $R(n) = 0$ except for $n = 0$; the case is important, not because groups consisting of a single trunk are common (they are not), but because the case $N = 1$ corresponds to making a measurement only on the first trunk of a group (of arbitrary size) in which the trunks are tried in a fixed order. For $N = 1$ it is easy to see (from Theorem 1) that

$$\begin{aligned} \Pr\{N(k) = 1 \mid N(0)\} &= \delta_{N(0),1} \quad \text{for } k = 0, \\ &= a_1 \qquad \text{otherwise,} \end{aligned}$$

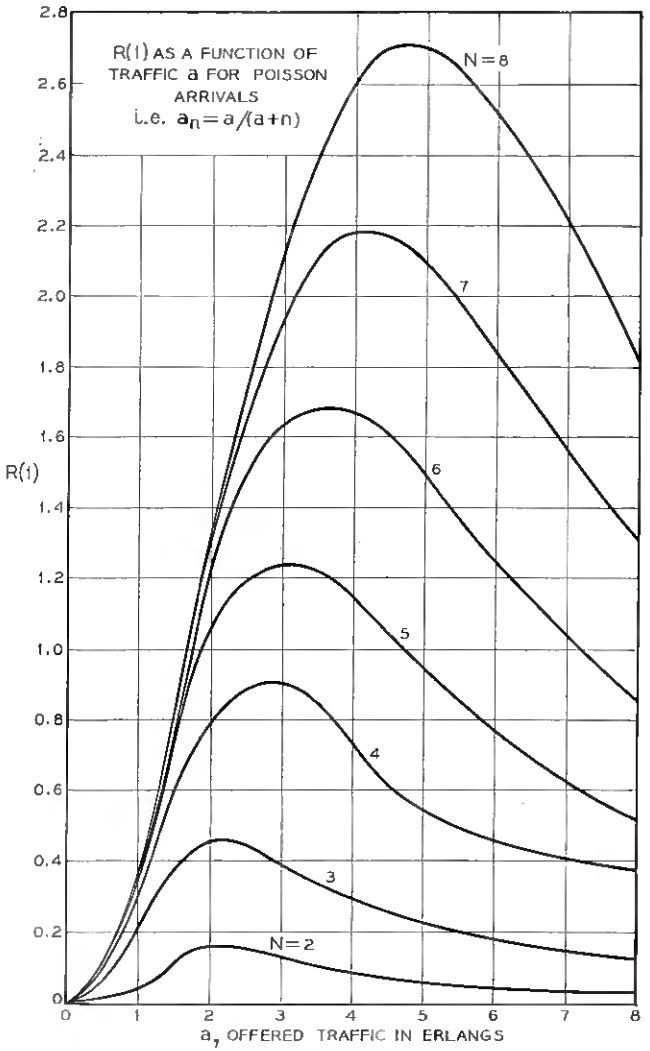


Fig. 18 — The covariance value $R(1)$ as a function of traffic a for Poisson arrivals.

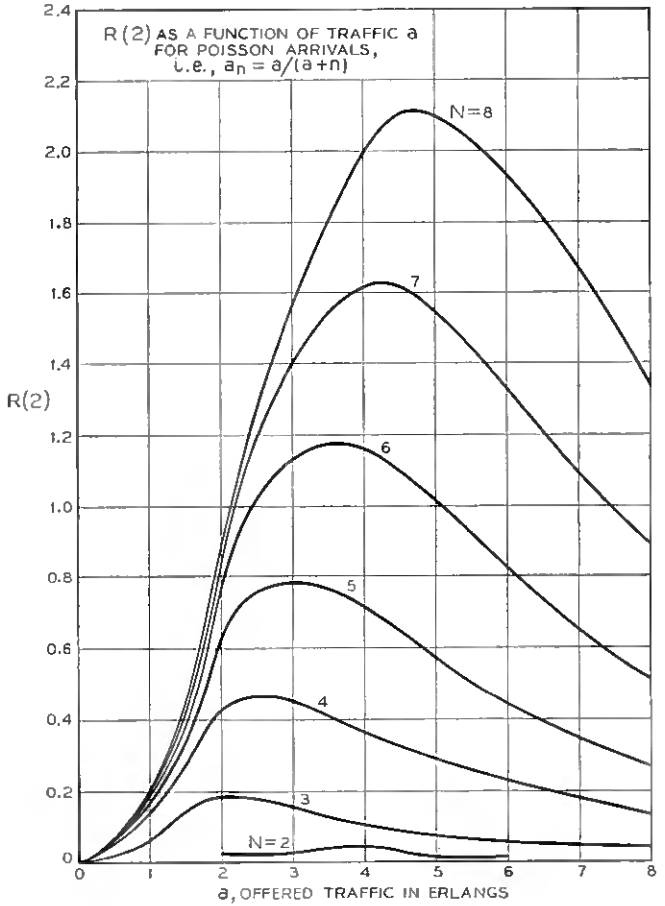


Fig. 19 — The covariance value $R(2)$ as a function of traffic a for Poisson arrivals.

so that $N(k)$ is *independent* of $N(0)$ for $k > 0$. Thus, in this case,

$$R(0) = \text{var}\,\{N(k)\} = a_1 - a_1^2,$$

$$R(n) = 0, \qquad \text{for } n \neq 0,$$

$$E\{S/n\} = E\{N(k)\} = a_1,$$

$$\text{var}\{S/n\} = \frac{a_1 - a_1^2}{n},$$

so that S/n is a consistent and unbiased estimator for a_1. It is to be

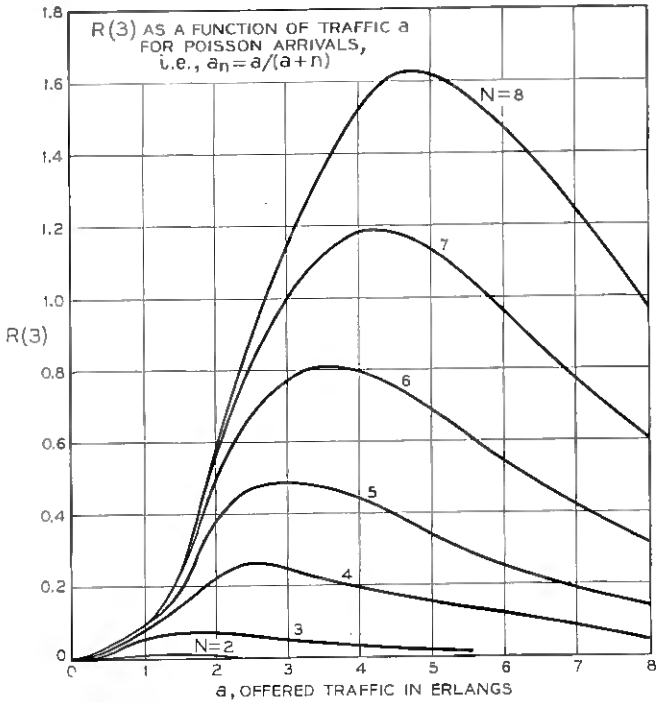


Fig. 20 — The covariance value $R(3)$ as a function of traffic a for Poisson arrivals.

emphasized that in this case S is the sum of n independent identically distributed random variables, each equal to 1 with probability a_1, and to 0 with probability $1 - a_1$. Thus, S has a binomial distribution with "success" parameter a_1.

The method of traffic measurement (on a group) outlined in the preceding paragraph has the disadvantage that it collects information very slowly. But it is relatively cheap, since all that has to be recorded is whether the first trunk is busy at arrival epochs or not, and it has the additional advantage that its statistical theory is relatively simple and has been well developed in the literature. It must be kept in mind that the sampling error estimates we develop are limited to measurements made at epochs just preceding arrivals.

Often the traffic engineer needs to estimate the *load offered* to a group, rather than the *load carried* by it. The use of S to estimate a_1 tells him what fraction of the time the first trunk is busy. However, there are cases in which the knowledge of a_1 determines the offered load. This

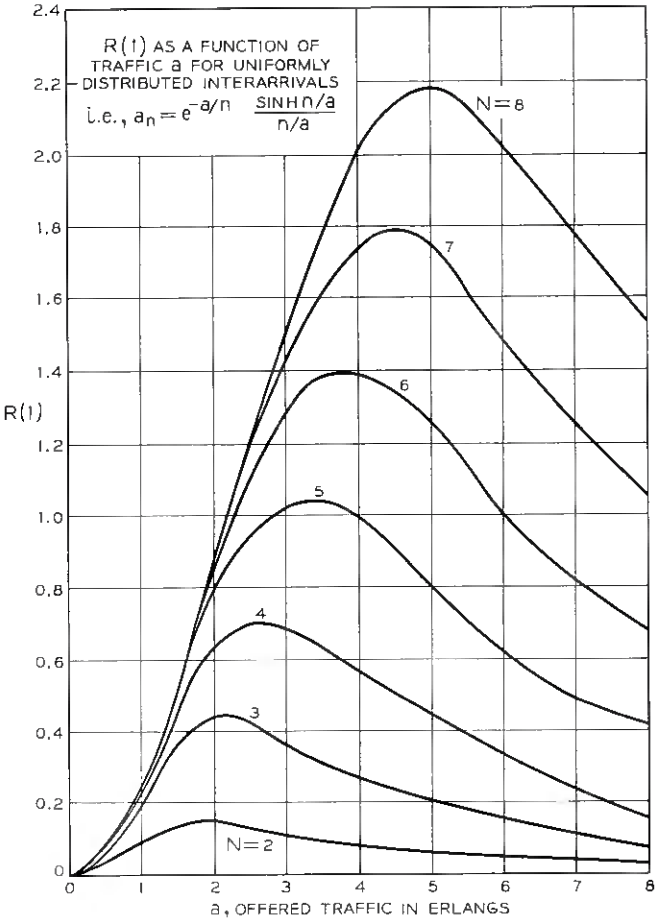


Fig. 21 — The covariance value $R(1)$ as a function of traffic a for uniformly distributed interarrivals.

occurs, in fact, whenever a_1 is a monotone function of the offered load a only. For example, when arrivals are Poisson, we have $a_1 = a/(1 + a)$, so it is reasonable to use $S/(n - S)$ as an estimator of the offered load a. When arrivals are regular, $a_1 = e^{-1/a}$, so a reasonable estimate of a is $1/(\log n - \log S)$.

In the Poisson example, this method of estimating a can be evaluated readily if we estimate a^{-1} instead by means of $(n + 1)/(S + 1) - 1$, whose stochastic limit is obviously a^{-1}.

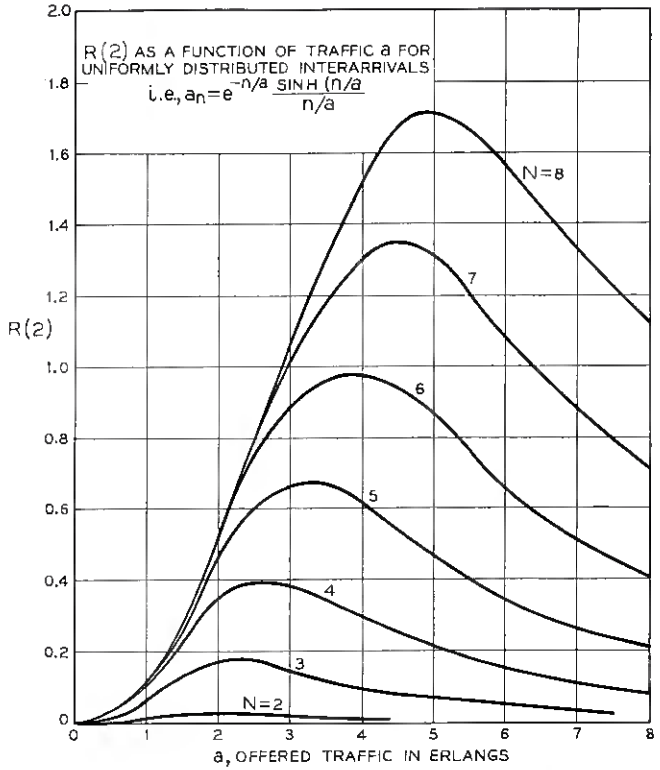


Fig. 22 — The covariance value $R(2)$ as a function of traffic a for uniformly distributed interarrivals.

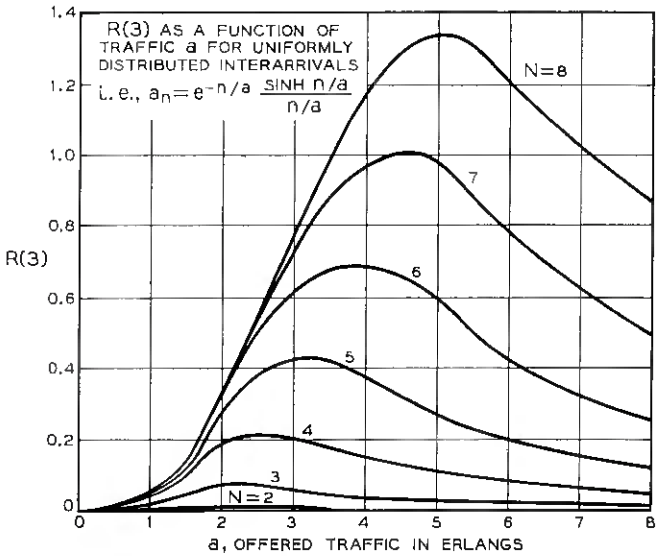


Fig. 23 — The covariance value $R(3)$ as a function of traffic a for uniformly distributed interarrivals.

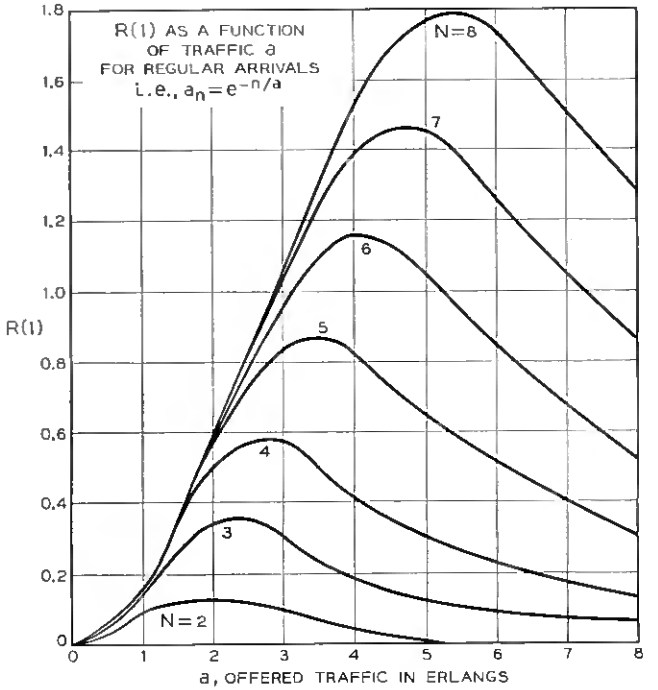


Fig. 24 — The covariance value $R(1)$ as a function of traffic a for regular arrivals.

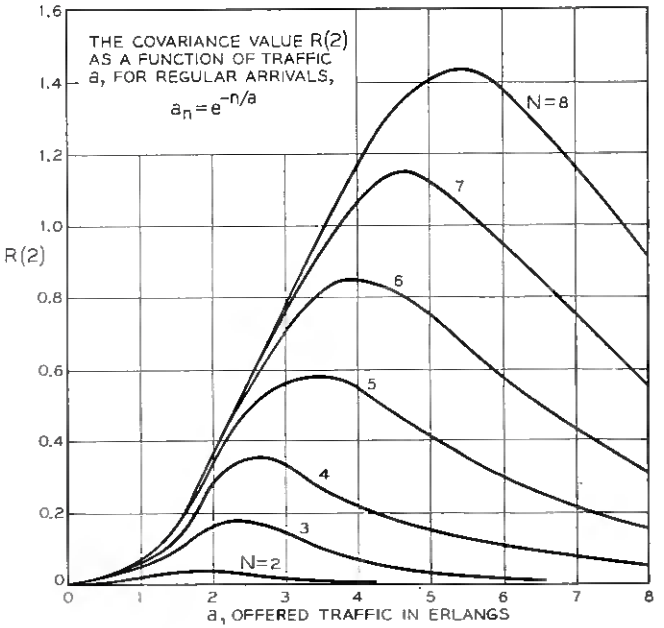


Fig. 25 — The covariance value $R(2)$ as a function of traffic a for regular arrivals.

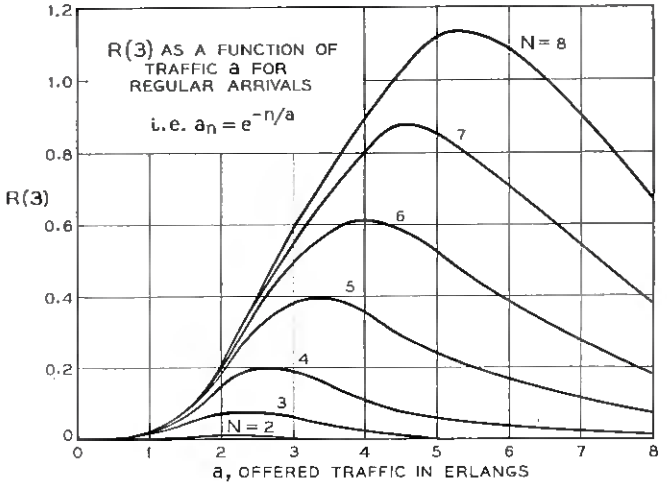


Fig. 26 — The covariance value $R(3)$ as a function of traffic a for regular arrivals.

The generating function of S is

$$E\{x^S\} = [1 + (x - 1)a_1]^n = \sum_{j=0}^{n} x^j \Pr\{S = j\}.$$

Hence

$$E\{(S + 1)^{-1}\} = \int_0^1 E\{x^S\}\, dx = \frac{1 - (1 - a_1)^{n+1}}{(n + 1)a_1},$$

$$E\left\{\frac{n + 1}{S + 1} - 1\right\} = \frac{1 - a_1}{a_1}[1 - (1 - a_1)^n].$$

There seems to be no simple formula for the second moment of this estimator, nor for that of n/S. However, noting that

$$\frac{(n + 1)^2}{(S + 1)(S + 2)} \leqq \frac{(n + 1)^2}{(S + 1)^2},$$

we can verify (by the same method as above) that

$$E\{(S^2 + 3S + 2)^{-1}\} = \int_0^1 \int_0^y E\{x^S\}\, dx\, dy = \frac{1 + a_1(1 - a_1)^{n+1}}{(n + 1)(n + 2)a_1^2},$$

$$E\left\{\frac{(n + 1)^2}{(S + 1)(S + 2)}\right\} = \frac{n + 1}{n + 2}\,\frac{1 + a_1(1 - a_1)^{n+1}}{a_1^2},$$

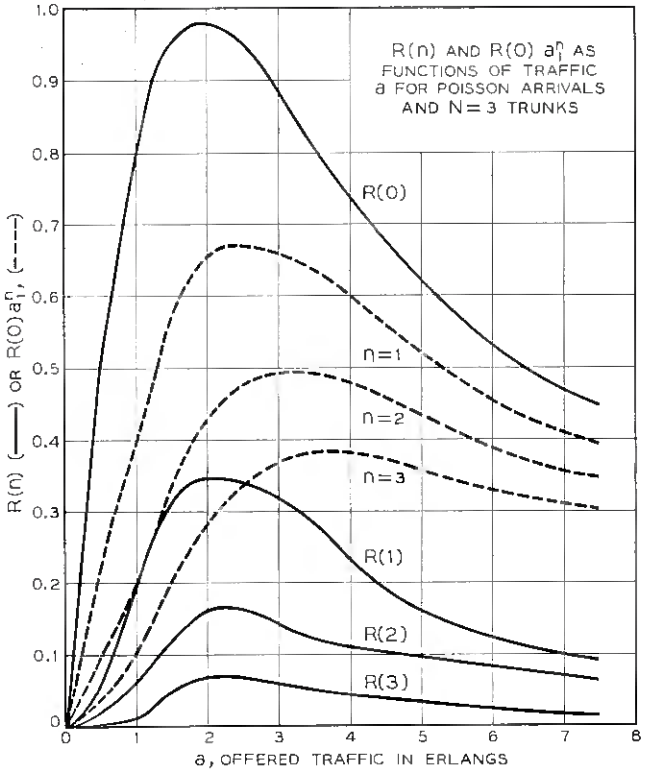


Fig. 27 — $R(n)$ and $R(O)a_1^n$ as functions of traffic a for Poisson arrivals and $N = 3$ trunks.

and so conclude that

$$\text{var}\left\{\frac{n+1}{S+1} - 1\right\} \geqq \frac{n+1}{n+2}\,\frac{1 + a_1(1-a_1)^{n+1}}{a_1^2} - \frac{[1-(1-a_1)^{n+1}]^2}{a_1^2}.$$

This lower bound is likely to be very close to the variance on the left for large n, so that, in this region,

$$\text{var}\left\{\frac{n+1}{S+1} - 1\right\} \sim \frac{(a_1+2)(1-a_1)^{n+1} - (1-a_1)^{2n+2}}{a_1^2}.$$

It can easily be shown (by the methods of Section IV) that, if $N = \infty$; i.e., if the trunk group is unlimited in size, the covariance function is exponential in character:

$$R(n) = R(0)a_1^n.$$

This suggests that, in some cases, $R(n) \sim R(0)a_1^n$ is a good approximation to the covariance for $N < \infty$. This approximation is equivalent to ignoring the correction terms Q_k in the recurrence relation (12) for the covariance. Since the sign of the Q_k in (12) is negative, it is clear that the approximation is an *overestimate.*

The covariance $R(n)$ for $n = 0, 1, 2, 3$ and the overestimate $R(0)a_1^n$ for $R(n)$ have been plotted together in Figs. 27, 28 and 29 for 3, 5 and 8 trunks, respectively, and Poisson arrivals. The curves suggest the following conclusions:

i. The approximation $R(n) \sim R(0)a_1^n$ is likely to be good if the load per trunk a/N is low.

ii. If the load per trunk a/N is high, e.g., $a/N = 1$, the approximation $R(n) \sim R(0)a_1^n$ may give a figure for the covariance (between

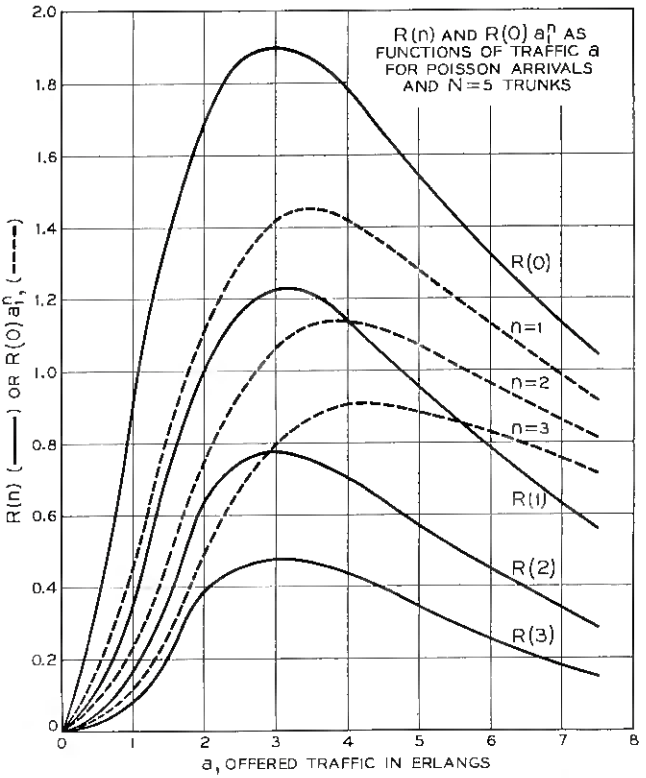


Fig. 28 — $R(n)$ and $R(0)a_1^n$ as functions of traffic a for Poisson arrivals and $N = 5$ trunks.

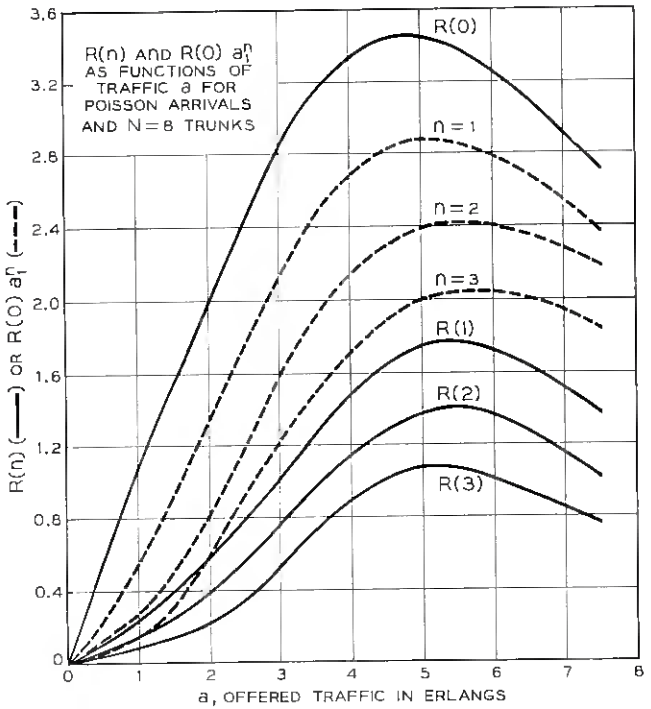


Fig. 29 — $R(n)$ and $R(O)a_1^n$ as functions of traffic a for Poisson arrivals and $N = 8$ trunks.

separate observations of $N(k)$) that is several times the actual value. This effect seems to increase with the separation, n.

iii. Variances, such as that of

$$S = \sum_{1}^{n} N(k),$$

computed on the basis of the approximation $R(n) \sim R(0)a_1^n$ are *overestimates*, so that use of this approximation in estimating sampling error is *conservative*.

iv. The value of a at which $R(n)$ has its (apparently unique) maximum seems to be the same for all n, depending only on N, the size of the group.

XI. ACKNOWLEDGMENT

The author expresses his gratitude to A. W. Horton, Jr., for criticism and encouragement, and to Miss Joyce Wiltshire for performing or programming computations.

REFERENCES

1. Palm, C., Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, **44,** 1943.
2. Feller, W., On the Theory of Stochastic Processes with Particular Reference to Applications, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability,* Univ. of California Press, Berkeley, Calif., 1949.
3. Pollaczek, F., Genéralization de la théorie probabiliste des systémes téléphoniques sans dispositif d'attente, Comptes Rendus, **236,** 1953, pp. 1469–1470.
4. Takács, L., On the Generalization of Erlang's Formula, Acta Math. Acad. Sci. Hungar., **7,** 1956, pp. 419–433.
5. Takács, L., On a Probability Problem Concerning Telephone Traffic, Acta Math. Acad. Sci. Hungar, **8,** 1957, pp. 319–324.
6. Cohen, J. W., The Full Availability Group of Trunks with an Arbitrary Distribution of the Interarrival Times and a Negative Exponential Holding-Time Distribution, Phillips Report, 1956, pp. 1–10.
7. Feller, W., *An Introduction to Probability Theory and Its Applications,* John Wiley & Sons, New York, 1950.

High-Frequency Gallium Arsenide Point-Contact Rectifiers

By W. M. SHARPLESS

(Manuscript received September 11, 1958)

Gallium arsenide, one of the Group III–V intermetallic compounds, appears to be an excellent semiconductor for use in point-contact devices. This paper describes some recent work in which single-crystal gallium arsenide, with resistivity adjusted to fit the application, is used for point-contact rectifiers which operate efficiently as frequency converters at frequencies as high as 60 kmc, and for switching diodes which show no minority carrier storage effects for switching time of the order of 10^{-10} seconds. These devices will operate over a considerable range in temperature.

I. INTRODUCTION

Silicon and germanium semiconductor materials have been used in point-contact rectifiers for many years and numerous types of rectifiers employing these two materials are commercially available today. Technical papers too numerous to mention have been published covering the important features of these Group IV semiconductor materials.

More recently, there has been increased interest in some of the semiconductor materials generally referred to as the intermetallic compounds. These are formed by a combination of some of the Group III and Group V elements and tend to possess some of the better properties of both silicon and germanium.* Due to the higher energy gaps, higher electron mobilities and, in some cases, the lower dielectric constants of some of these III–V compounds, theoretically they should make efficient high-frequency rectifiers and should be able to operate at higher temperatures than either silicon or germanium.†

Gallium arsenide (GaAs), one of the III–V intermetallic compounds, appears to be very attractive for high-frequency point-contact rectifiers.

* A good review of the work that has been done on the Group III-V compounds appears in a recent book.[1]

† The importance of the semiconducting compounds was perhaps first discerned by H. Welker in Germany early in the 1950's.[2]

In a recent paper, Jenny[3] reports that GaAs point-contact rectifiers have operated efficiently as first detectors at frequencies as high as 6 kmc.

This paper describes some of the work on gallium arsenide point-contact rectifiers which is currently in progress at Bell Telephone Laboratories, Holmdel, N. J. By controlling the resistivity of the single-crystal gallium arsenide and by the selection of the proper point material and processing technique, rectifiers intended for either high-frequency first detectors or lower-frequency high-speed switching devices have been produced. Measurements have been made of conversion loss, output noise ratio and intermediate-frequency impedance of GaAs rectifiers operating as first detectors in the millimeter wave band (55 kmc) and in the X-band (11 kmc). High-speed switching diodes have been made which showed no carrier storage effects for switching times of the order of 10^{-10} seconds. Rectifying characteristics have been taken on test diodes over a temperature range between $-320°$ F and $+237°$ F.

II. GENERAL PROCESSING OF GaAs RECTIFIERS

Some variations in the general processing techniques have been found necessary in order to produce the several different types of rectifiers desired. There are, however, several steps in the processing that are common to all types and these will be discussed first.

It is of prime importance to obtain a good ohmic back contact to the GaAs sample. Experience at our laboratory has indicated that one of the best ways to accomplish this is to deposit a thin tin-and-nickel coating on the flat, clean back surface of the GaAs. The sample is then heated in a vacuum furnace to a temperature at which the tin will start to diffuse into the GaAs. This forms an excellent ohmic back contact to the GaAs and leaves a tough nickel external surface which may be used for subsequent soldering. Back contacts made in this way are very uniformly adherent. This becomes most important when the samples are diced into miniature squares suitable for soldering to the small supporting structures needed in very high frequency devices.

The surface of the sample which is later to be used for the point contact is finished either by grinding with M305 abrasive or by polishing to a smooth, mirror-like finish with a one-micron sapphire dust abrasive. We have found that the polished surface results in a more reproducible, lower-capacity point-contact area than any of the ground surfaces tested. In either case, just before the rectifiers are assembled, the GaAs contact surfaces are given a light chemical etch with a dilute solution of hydrofluoric and nitric acids.

An exhaustive study of all the materials which might be used for point-contact springs was not made, but, of the metals tested, our best results have been obtained with spring-tempered phosphor bronze wires. The "S" springs are welded to their supports and sharply pointed electrolytically. For very high frequency work the use of a very sharp point is desirable and the pressure applied to the contact area is kept small. For lower frequency and higher power switching applications, the strength of the springs is increased and the sharpness of the points becomes of lesser importance.

When a metal point is brought into contact with a prepared semiconductor surface, the initial rectification pattern is usually poor compared to the desired static characteristic but can be improved by further contact conditioning.* In the case of p-type silicon rectifiers, it has generally been found possible to bring about this improved rectification ratio by mechanically tapping the rectifier case. For n-type germanium rectifiers this conditioning or forming may be accomplished by applying electrical pulses directly to the rectifier terminals.

N-type GaAs responds to contact area forming in much the same way as do n-type germanium rectifiers. The forming technique consists of applying a series of fairly high-level pulses of energy to the rectifier terminals after point contact has been established. We have found that 60-cycle sine-wave pulses are quite satisfactory for this forming, and the low-frequency static characteristic may be observed on an oscilloscope while the forming is taking place. Arrangements are provided for separately controlling the magnitude of the voltage applied in either the positive or negative direction, or both voltages may be applied simultaneously. The resulting current is controlled by adjusting the value of a series resistor. A considerable amount of latitude in forming is thus provided.

III. CONTROLLED-RESISTIVITY GALLIUM ARSENIDE MATERIAL

The rectifiers described in this paper have all been made from specially doped GaAs material obtained from single-crystal ingots prepared by J. M. Whelan of Bell Telephone Laboratories at Murray Hill, N. J. Preparation of the compound and growth of single crystals by the floating zone method have been previously described.[4] In a private communication, Whelan describes the method used to prepare single crystal GaAs of controlled resistivity as follows: Zone refining was used to

* An exception is found for semiconductor materials which have had their surfaces previously conditioned by ionic bombardment. In such cases further surface conditioning is not necessary or, in general, desirable.

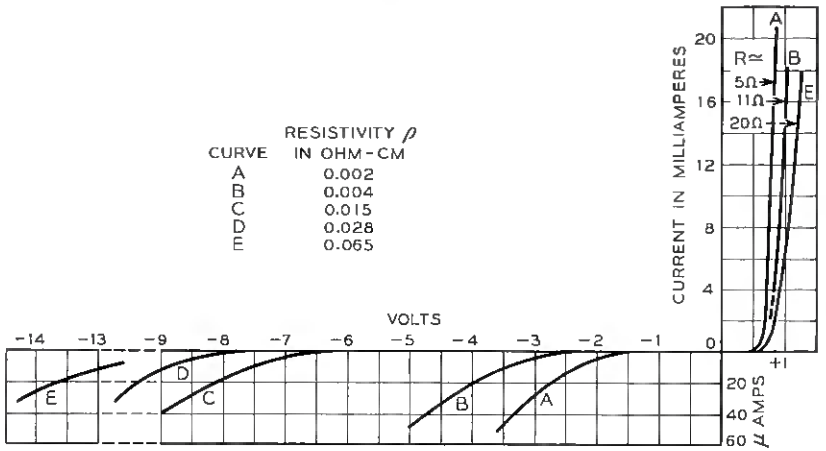


Fig. 1 — Typical static characteristics of GaAs point-contact rectifiers as a function of semiconductor resistivity.

increase the resistivity of the GaAs above that required for diodes. The purified material was then doped to the desired resistivity, 0.002 to 0.07 ohm-cm, by regrowing the crystal in an arsenic atmosphere containing one of the following donor impurities: sulphur, selenium or tellurium. Overdoping was corrected, when necessary, by subsequent floating zone passes in a "pure" arsenic atmosphere.[5]

The effect of varying the resistivity of the GaAs material used in a point-contact rectifier is shown in Fig. 1. Typical static characteristics are shown for rectifiers made from doped GaAs materials, ranging from 0.065 ohm-cm for a lightly doped sample to 0.002 ohm-cm for heavily doped material. The data presented in Fig. 1 were obtained using the same size pointed phosphor bronze springs and the same contact pressure in each case. The contact surface preparations were also the same. The forming techniques were optimized insofar as possible for each rectifier, and thus the curves show the typical static characteristics that result when rectifiers are processed using different resistivity GaAs materials. Depending on the particular application intended, the spreading resistance and other characteristics of the rectifiers may thus be varied over a considerable range by the selection of the properly doped material. Further, depending on the type of rectifier desired, the frequency characteristic and power handling capacity may be varied by controlling the size of the point-contact area and the contact pressure.

3.1 *Millimeter Wave Rectifiers*

As mentioned earlier, GaAs rectifiers have been prepared for use as first detectors in both the millimeter and X-band range. Measurements have been made of conversion loss, output noise ratios and IF impedance for typical operating conditions.

GaAs rectifiers intended for operation as first detectors at millimeter waves (50 to 90 kmc) are assembled in the wafer-type millimeter wave mounting shown in Fig. 2. For very high frequency first detector rectifiers, it is important to keep the product of the barrier capacity and the spreading resistance as small as possible; thus, the lowest resistivity ($\rho = 0.002$ ohm-cm) material is selected for this application (see Fig. 1). A 0.001-in. diameter phosphor bronze wire spring is selected to give the low contact pressure desired and the wire is sharply pointed to give the small contact area needed. The rectifier is mechanically assembled in much the same way as are the silicon millimeter wave wafer rectifiers described in a previous paper.[6] A light contact pressure is applied by advancing the point one-half mil after contact between the polished surface of the GaAs and the phosphor bronze point has been established. A low ac voltage (4–6 volts peak) is then applied for viewing the static characteristics on an oscilloscope. Arrangements are provided for limiting the voltage and current, as mentioned earlier. Rectifiers intended for operation at millimeter waves are given no contact forming other than that which takes place when the low-voltage ac is applied through a 1000-ohm series resistor while the static characteristic is being viewed.

Table I gives the measured performance figures for the best point-contact wafer-type millimeter wave rectifiers we have made using either silicon, germanium or gallium arsenide as semiconductor materials. Conditions of operation have been optimized for the best output signal-to-noise ratio for each type of rectifier measured. Both the germanium

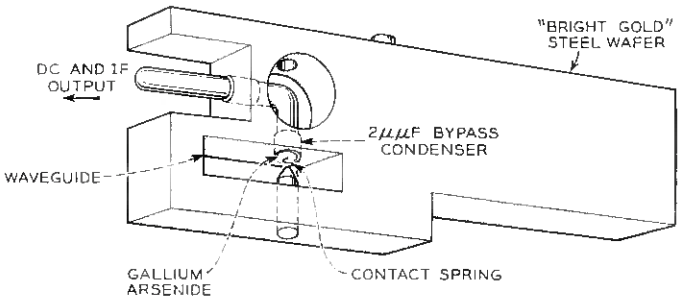


Fig. 2 — Millimeter-wave wafer unit using GaAs.

TABLE I

Rectifier Type	Semiconductor Material	Conversion Loss, db	Noise Ratio, N_R	Z_{IF}, ohms (60 mc)
Wafer	p-type silicon	6.4	1.6	260
Wafer	n-type germanium	6.6	2.0	300
Wafer	n-type gallium arsenide	5.6	2.1	325

and GaAs units were given a few tenths of a volt positive bias. Measurements were made at a frequency of 55.5 kmc using a 60-mc intermediate frequency.

From the table it can be seen that the lowest first detector conversion loss figure measured at millimeter waves was obtained using a GaAs rectifier. One would expect low conversion losses for GaAs units due to the high mobility and low spreading resistance of the basic GaAs materials used, but the actual measurement of a conversion loss below 6 db at a frequency of 55 kmc is gratifying. The noise outputs from millimeter wave rectifiers all tend to be higher than those measured for similar types of rectifiers designed for and used at longer waves. It is believed that this is at least partly due to the lighter point-contact pressures that are required in these very high frequency units. Further experience may be helpful in reducing these noise ratios.

3.2 *Microwave Rectifiers*

GaAs rectifiers intended for operation in the microwave bands have been assembled in the small low-capacity cartridge case shown in Figs. 3 and 4. These units use contact points and semiconductor wafers of approximately the same size as the millimeter wave units described

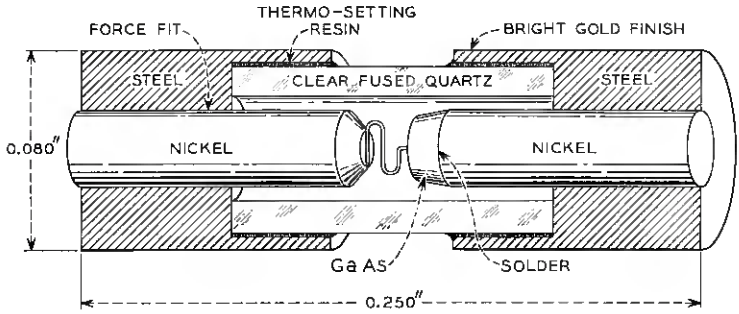


Fig. 3 — Cross section of microwave low-reactance cartridge-type GaAs rectifier holder.

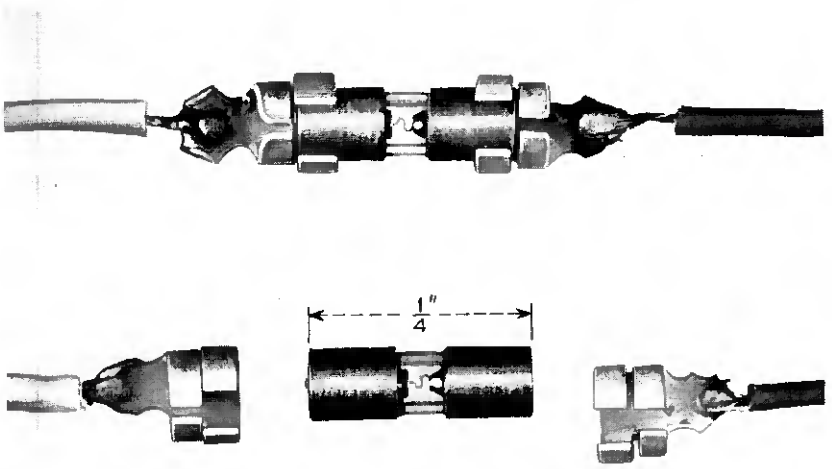


Fig. 4 — Photograph of cartridge-type GaAs rectifier holder, with pigtail leads for use at low frequencies (magnified 8 times).

above, and since the inductance and capacity of the rectifier components are very small — as they must be for millimeter waves — the unit is easily adaptable to broadband designs in the kilomegacycle range.

The cartridge rectifiers intended for operation at *X*-band (11 kmc) are assembled by first pressing in the GaAs detail and then the phosphor bronze point detail until contact is established. The pointed wire "S" spring, approximately 0.002 in. in diameter, is then advanced seven-tenths of a mil and the contact area lightly formed in the same manner as for the millimeter wave first detectors. GaAs material in the low resistivity range between 0.002 and 0.004 ohm-cm is used for these rectifiers.

In the microwave region, the cartridge-type rectifiers may be mounted directly in waveguides in the conventional manner. The cartridge, because of its small size, also adapts itself to special types of broadband arrangements such as those provided by coaxial lines or by ridged or stepped waveguides. Fig. 5 suggests one such possible microwave mounting arrangement, where the impedance of the regular waveguide is lowered by reducing the E-plane guide height until it presents an impedance the same as the resistive component of an average rectifier. The remaining reactance may be tuned out with a waveguide shorting section, which follows the rectifier. Since both ends of the cartridge are the same physical size, the rectifiers may be turned end for end when a change in polarity is desired.

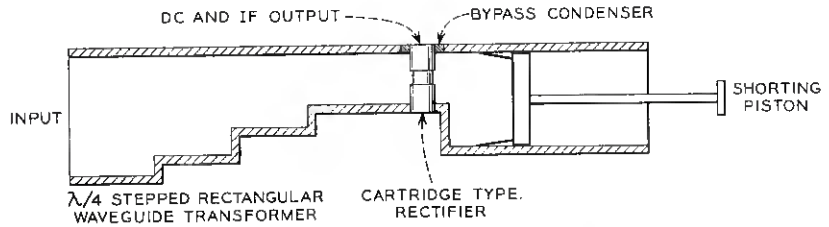


Fig. 5 — Broadband waveguide arrangement for use with cartridge-type rectifiers.

Table II gives some measured broadband first detector performance figures for several point-contact cartridge-type rectifiers that we have made using samples of our best silicon and best gallium arsenide materials. Conditions for operation were optimized insofar as possible for each group of rectifiers measured, with the GaAs units given a few tenths of a volt positive bias. Measurements were made at a frequency of 11 kmc employing a 60-mc intermediate frequency. A beating oscillator drive of one milliwatt was used in all cases.

From the table below it can be seen that the measured output noise ratios of GaAs point-contact rectifiers operating as first detectors at 11 kmc are at least as good as those for similar units employing silicon as the semiconductor material. Further, the average conversion loss of the GaAs units is at least one db better than that of the silicon group measured, which would mean that, conservatively, the over-all noise figure of an 11-kmc receiver would be improved at least one db by using GaAs in place of the silicon. If we used a GaAs first detector having the best conversion loss and noise ratio in Table II together with an IF amplifier having a noise figure of $1\frac{1}{2}$ db, the resulting over-all receiver noise figure at 11 kmc would be 6.0 db. It appears that the reproducibility of the GaAs units is good and that the variation between one unit and the next will be small; this is evidenced by the small difference between the average and best conversion loss listed in Table II.

TABLE II

Semiconductor Material	Number of Units	Conversion Loss, db		Noise Ratio, N_R		Range of Z_{IF}, ohms
		Average	Best	Average	Best	
p-type silicon	18	5.8	5.1	1.47	1.36	300–500
n-type gallium arsenide	8	4.2	4.0	1.35	1.20	275–580

IV. SWITCHING RECTIFIERS

GaAs rectifiers intended for high-speed switching applications are assembled in much the same way as are the microwave diodes, using the same low-capacity case shown in Fig. 4. In general, these units operate at higher power levels and at frequencies in the lower kmc region; thus a heavier contact pressure may be applied, and a 0.003-in. diameter phosphor bronze wire spring is used with a spring advance after contact of up to one and one-half mils. These units are given a more intensive ac forming by increasing the peak driving voltage to about 20 volts and reducing the value of the series current limiting resistor to 300 ohms.

GaAs material in the resistivity range from 0.02 to 0.04 ohm-cm is used for general-purpose switching rectifiers. It is obvious that, depending on the exact requirements regarding forward and reverse impedances at a given driving voltage, one might use resistivity values different from the range mentioned above. In general, the GaAs material for switching applications is selected on the basis of using the lowest resistivity material that will allow a satisfactory reverse characteristic (see Fig. 1). During the forming process, the forward resistance tends to decrease rapidly, as desired, but, at the same time, the contact area tends to increase, which is in the direction of limiting the efficiency of operation at the higher frequencies. Thus, as in all point-contact devices, several factors must be considered in the processing of the rectifiers and, in general, a compromise is adopted which will arrive at the best rectifier for a particular application. In the case of the switching rectifiers, a resistivity of about 0.03 ohm-cm is used, together with a forming technique which gives a good compromise between low capacity, low forward resistance and very high back impedance up to, say, -10 volts. Measurements made on switching rectifiers of this type show no minority carrier storage effects up to switching times of the order of 10^{-10} seconds, which is the limit of our present measuring equipment.

V. TEMPERATURE EFFECTS

Since GaAs possesses a relatively high energy gap (1.34 ev) it tends to be more suitable for stable operation at higher temperatures than are possible for either silicon or germanium. Fig. 6 is a multiple photograph of a cathode ray oscilloscope display showing the low-frequency static characteristic of a point-contact GaAs rectifier as the temperature was varied from $-320°$ F to $+237°$ F. It will be noticed that the static characteristics changed only slightly over this large temperature range, the

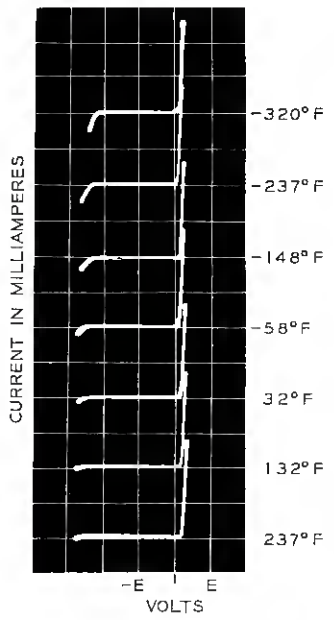


Fig. 6 — Effect of temperature variation on the static characteristic of a GaAs point-contact rectifier.

principal change being in the location of the knee of the reverse current characteristic. It is obvious that the temperature limit for good rectification was not reached at either extreme of the range covered. Experimental evidence has been published[1] which indicates that certain samples of n-type GaAs were found to show very small changes in Hall coefficient, resistivity or mobility up to a temperature near 600° F. This indicates that GaAs rectifiers would probably operate efficiently at temperatures this high if a temperature-stable mounting were provided.

Groups of GaAs point-contact rectifiers have now been in storage at room temperature in our laboratory for periods of several months. There is no evidence to date that the normal changes in relative humidity and temperature experienced in the laboratory have had any effect on the rectifying properties of the units.

VI. CONCLUDING REMARKS

It appears that gallium arsenide semiconductor devices may well enjoy a very bright future. The relative insensitivity of gallium arsenide point-contact rectifiers to rather large changes in operating tempera-

tures will be important in many diode applications. Extensive burn-out tests were not made, but gallium arsenide rectifiers appear to possess as good burn-out properties as similar types of rectifiers made with either silicon or germanium. Efficient operation of gallium arsenide rectifiers as first detectors at frequencies extending upward into the millimeter wave band has been demonstrated. Gallium arsenide high-speed switching diodes have been made which show no carrier storage effects for switching times of the order of 10^{-10} seconds.

VII. ACKNOWLEDGMENTS

The author is greatly indebted to J. M. Whelan, who developed the process for making the single-crystal gallium arsenide and prepared all the GaAs materials used in the work described. Helpful suggestions regarding details of the processing techniques were received from R. S Ohl. Assisting in the work were E. F. Elbert and S. E. Reed.

REFERENCES

1. Cunnell, F. A. and Saker, E. W., *Progress in Semiconductors, Vol. II*, John Wiley & Sons, New York, 1957, pp. 37–65.
2. Welker, H., Z. Natur., **7a**, 1952, p. 744.
3. Jenny, D. A., A Gallium Arsenide Microwave Diode, Proc. I.R.E., **46**, April 1958, pp. 717–722.
4. Whelan, J. M. and Wheatley, G. H., J. Phys. Chem. Solids, **6**, 1958, p. 169.
5. Struthers, J. D., Whelan, J. M. and Ditzenberger, J. A., to be published.
6. Sharpless, W. M., Wafer-Type Millimeter Wave Rectifiers, B.S.T.J., **35**, November 1956, pp. 1385–1402.

Paramagnetic Spectra of Substituted Sapphires—Part I: Ruby*

By E. O. SCHULZ-DU BOIS

(Manuscript received September 29, 1958)

The paramagnetic resonance properties of Cr^{+++} ions in Al_2O_3 (ruby) were investigated theoretically and experimentally in order to obtain information necessary for the application of this material as active material in a three-level solid-state maser (3LSSM). Numerically computed energy levels, together with their associated eigenvectors, are presented as a function of applied magnetic field for various orientations of the magnetic field with respect to the crystalline symmetry axis. A more detailed discussion is devoted to energy levels, eigenvectors and transition probabilities at angles 0°, 54.74° and 90°, where certain simple relations and symmetries hold. Paramagnetic spectra for signal frequencies between 5 and 24 kmc are shown; agreement between computed and measured resonance fields is satisfactory.

I. INTRODUCTION

Among the paramagnetic salts that have been used as active materials in three-level solid-state masers (3LSSM),[1, 2, 3] ruby shows rather desirable properties. While maser action of this material has been achieved at microwave signal frequencies of 3 to 10 kmc,[4] it should be possible to cover more than the whole centimeter microwave range. Perhaps even more important from a practical point of view are the bulk physical properties. Extremely good heat conductivity at low temperatures allows handling of relatively high microwave power dissipation. Industrial growth of large single crystals by the flame fusion technique and machinability with diamond tools make it possible to fabricate long sections of ruby to very close tolerances, a necessity in travelling-wave maser (TWM) development. Also, ruby can be bonded to metals, thus allowing a high degree of versatility in maser structural design. While the use of ruby in 3LSSM, in particular in nonreciprocal TWM, will be described

* This work is partially supported by the Signal Corps under Contract Number DA-36-039 sc-73224.

in forthcoming papers by members of Bell Telephone Laboratories, this paper is intended to give some background on paramagnetic resonance behavior of ruby.

In general, the paramagnetic resonance properties of an ion in a crystal can be completely described by a spin Hamiltonian containing a relatively small number of constants. In the case of ruby, these include the spectroscopic splitting factors parallel and perpendicular to the crystalline axis, $g_\parallel$ and $g_\perp$, the total spin $S = 3/2$ and the sign and magnitude of $2D$, the zero field splitting. Nuclear interactions can be neglected since the most abundant isotope, Cr^{52}, is nonmagnetic ($I = 0$) whereas the magnetic isotope, Cr^{53}, ($I = 3/2$) has small abundance (9.5 per cent) and leads to negligible line broadening only. Taking this into account, one can even predict on the basis of the total spin and the crystalline symmetry surrounding the Cr^{+++} ion that no other terms can occur in the spin Hamiltonian.

However, in order to predict operating conditions of this or other materials in a 3LSSM, it is necessary to know the separation of energy levels for supplying the proper pump and signal frequencies, the order of magnitude of the associated transition probabilities and perhaps other circumstances, such as coincidence of transition frequencies, which, by spin-spin interaction, may lead to shortening of the associated relaxation times (self-doping condition). In this paper, this information is evaluated by the formalism of the spin Hamiltonian and, at least in part, compared with experiment. The data presented graphically are intended to form an "atlas" of the ruby paramagnetic resonance properties. In the paper which follows, some general viewpoints are presented on modes in which paramagnetic materials can be operated as active materials in a 3LSSM. In further papers, paramagnetic spectra of other substitutional ions such as Co^{++} and Fe^{+++} in sapphire will be presented in order to furnish sufficient information to find coincidences of transition frequencies of Cr^{+++} with Co^{++} or Fe^{+++} lines resulting in reduced relaxation times (impurity-doping condition).

For a derivation of the method of spin Hamiltonians, reference should be made to such review articles as those by Bleaney and Stevens[5] and Bowers and Owen.[6] Knowledge of the associated formalism is perhaps desirable but not necessary for utilization of the results reported in this paper. Briefly, the spin Hamiltonian describes the energy of a paramagnetic ion arising from interaction with host crystal environment and applied magnetic field. Obeying quantum laws, the ion can exist in one of several states associated with discrete energy levels. Transitions between such states can occur if the energy balance ΔE is supplied to or

extracted from the ion. Given some probability for radiative transitions, these can be induced by applying a magnetic field of radio frequency $\nu = \Delta E/h$ (h = Planck's constant). If there are more transitions to the higher state, net absorption will be observed such as is normally observed with a spectrometer. If there are more transitions to the lower state, stimulated emission of energy will be observed such as is utilized for amplification in a 3LSSM.

II. THE SPIN HAMILTONIAN

The spin Hamiltonian of Cr^{+++} in Al_2O_3 was first published by Manenkov and Prokhorov[7], and later by Geusic[8] and Zaripov and Shamonin.[9] It was given in the form

$$\mathcal{H} = g_{\parallel}H_zS_z + g_{\perp}(H_xS_x + H_yS_y) + D[S_z^2 - \tfrac{1}{3}S(S+1)]. \qquad (1)$$

The effective spin $S = 3/2$ is identical with the true spin. All Cr^{+++} ions in the crystal lattice show identical paramagnetic behavior, with the magnetic z-axis being the same as the trigonal symmetry axis of the crystal. The best values for the constants seem to be

$$2D = -2D' = -0.3831 \pm 0.0002 \text{ cm}^{-1} = -11.493 \pm 0.006 \text{ kmc},$$

$$g_{\parallel} = 1.9840 \pm 0.0006,$$

$$g_{\perp} = 1.9867 \pm 0.0006.$$

While it is customary in spectroscopy to express energy in units of cm^{-1} omitting a factor hc (h = Planck's constant, c = velocity of light), units of kmc are used simultaneously, omitting a factor of 10^9 h, because this allows direct interpretation in observed spectra.

In particular, the negative sign of D was obtained by Geusic.[8] He deduced this from the fact that $g_{\parallel} < g_{\perp}$, since in less than half-filled d-shell ions, such as Cr^{+++}, the spin-orbit coupling term λ is positive, and D is given by $2D = \lambda(g_{\parallel} - g_{\perp})$. Sign and magnitude of D are in agreement with results of low-temperature static susceptibility measurements by Bruger.[10] In this work also, the negative sign of D was confirmed by comparing the relative intensities of two lines at liquid nitrogen and helium temperatures.

The spin Hamiltonian (1) can more conveniently be written in spherical coordinates:

$$\mathcal{H} = g_{\parallel}H\cos\theta S_z + \tfrac{1}{2}g_{\perp}H\sin\theta(e^{-i\varphi}S_+ + e^{i\varphi}S_-) - D'[S_z^2 - \tfrac{1}{3}S(S+1)]. \qquad (2)$$

Here $S_\pm = S_x \pm iS_y$. In both representations (1) and (2) the crystalline axis was chosen to be the z-axis. While the choice of reference system is immaterial to obtaining eigenvalues (energy levels), this choice shows up in the associated eigenvectors. The eigenvectors have no direct physical interpretation; they must be evaluated in order to obtain transition probabilities. The transition probabilities most naturally obtained from eigenvectors of the Hamiltonian (2) are those which correspond to excitation by RF magnetic fields whose polarization is either linear and parallel to, or circular and perpendicular to, the *crystalline axis*.

In 3LSSM design, however, it seems more appropriate to analyze the performance in terms of RF magnetic fields whose polarization is either linear and parallel to, or circular and perpendicular to, the *applied field*. The corresponding eigenvectors and transition probabilities can, of course, be obtained from those belonging to the Hamiltonian (2) by a 4-by-4 transformation matrix. But it is more efficient to obtain them directly through a transformation of the original spin Hamiltonian (1) or (2) into a coordinate system with the z-axis parallel to the applied field. The result of this transformation is

$$\begin{aligned}\mathcal{H} = {} & (g_\parallel \cos^2\theta + g_\perp \sin^2\theta)\beta H S_z \\ & - D'(\cos^2\theta - \tfrac{1}{2}\sin^2\theta)[S_z^2 - \tfrac{1}{3}S(S+1)] \\ & - D'\tfrac{1}{2}\cos\theta\sin\theta[e^{-i\varphi}(S_zS_+ + S_+S_z) + e^{i\varphi}(S_zS_- + S_-S_z)] \\ & - D'\tfrac{1}{4}\sin^2\theta(e^{-2i\varphi}S_+^2 + e^{2i\varphi}S_-^2).\end{aligned} \tag{3}$$

III. ENERGY LEVELS AND EIGENVECTORS

From the Hamiltonian $\mathcal{H}$ (3), its energy eigenvalues W are found numerically by solving the fourth-order secular equation

$$\begin{gathered}\| \langle n \mid \mathcal{H} - W \mid m \rangle \| = 0, \\ n, m = 3/2,\ 1/2,\ -1/2,\ -3/2.\end{gathered} \tag{4}$$

The eigenvalues W are functions of H and θ, but not of φ since, because of the symmetry of the Hamiltonian, rotation about the z-axis does not change the physical situation. On the following plots (left-hand sections of Figs. 1 through 11) diagrams of energy levels W (in units of kmc) are shown as a function of applied field H (in units of kilogauss). Plots are given for angles θ from 0° to 90° in steps of 10° and, in addition, for 54.74°.

Also, by change of scales, dimensionless eigenvalues $y = W/D'$ are shown as functions of the dimensionless quantity $x = G/D'$, where

$$G = (g_\parallel \cos^2\theta + g_\perp \sin^2\theta)\beta H.$$

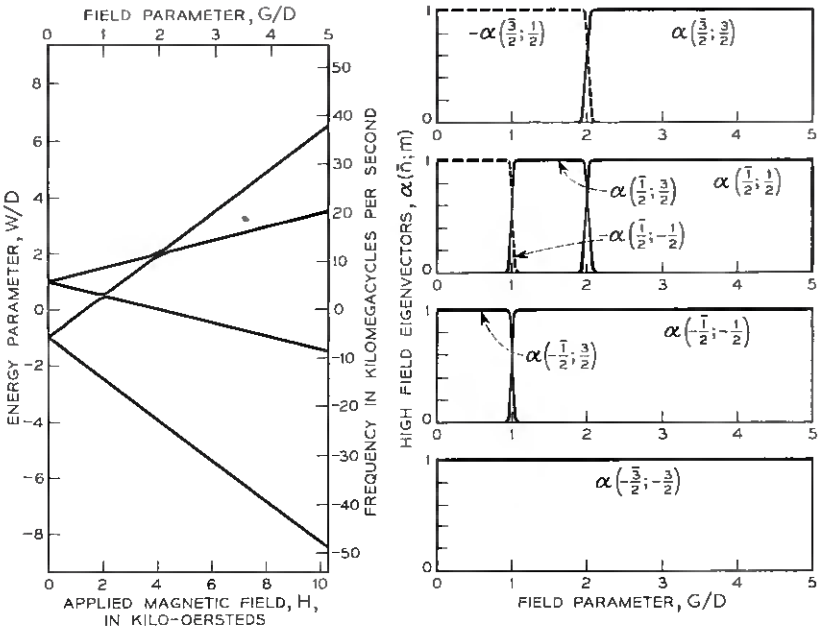


Fig. 1 — Energy levels and eigenvectors of the Cr^{+++} paramagnetic ion in ruby at angle $\theta = 0°$ between crystalline symmetry axis and applied magnetic field.

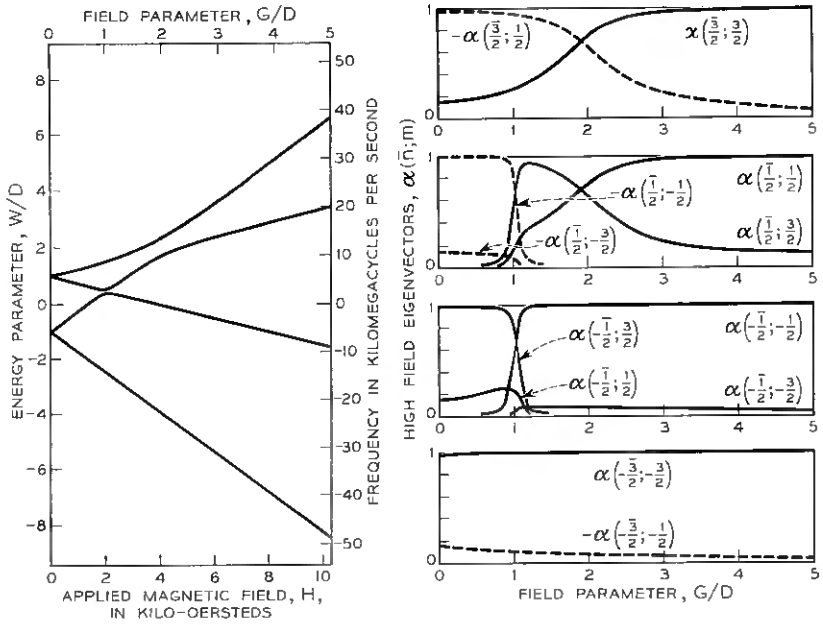


Fig. 2 — Energy levels and eigenvectors at 10°.

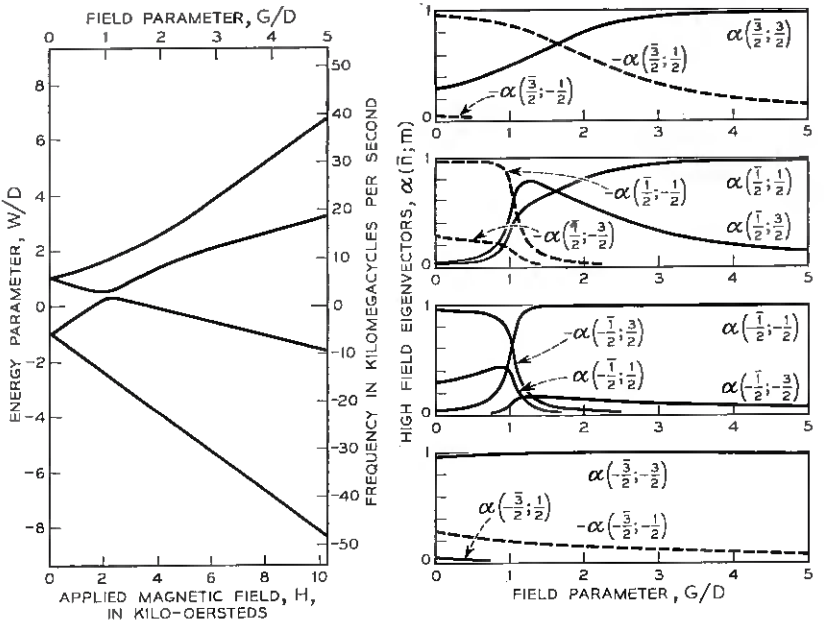


Fig. 3 — Energy levels and eigenvectors at 20°.

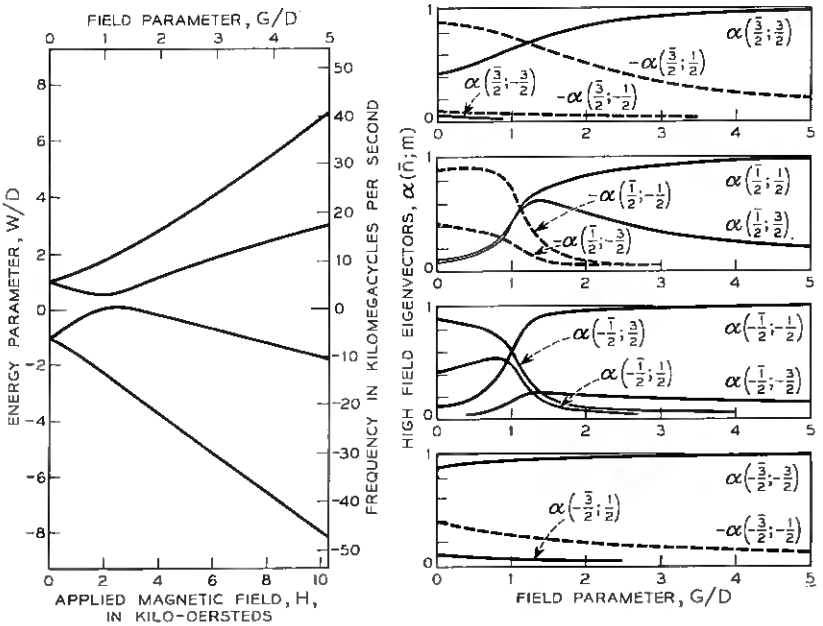


Fig. 4 — Energy levels and eigenvectors at 30°.

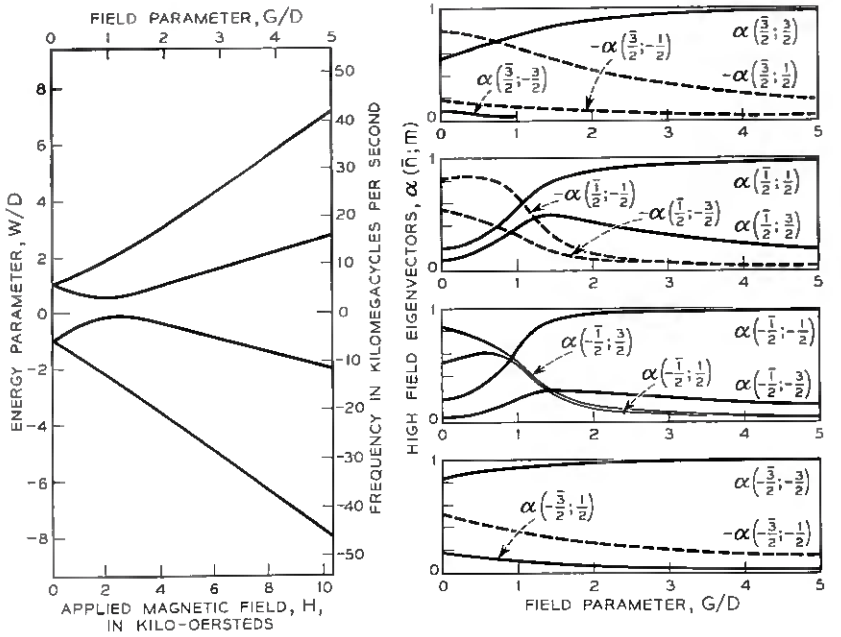


Fig. 5 — Energy levels and eigenvectors at 40°.

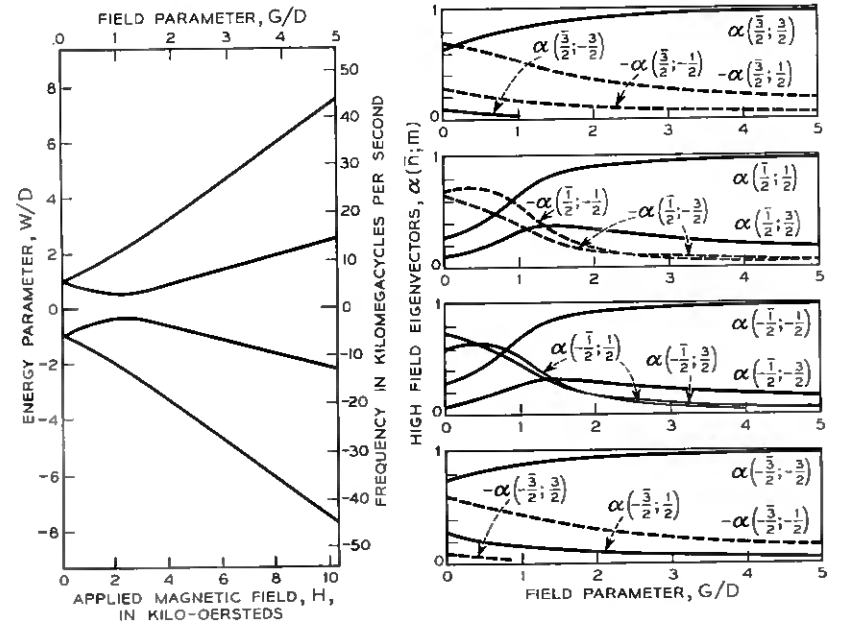


Fig. 6 — Energy levels and eigenvectors at 50°.

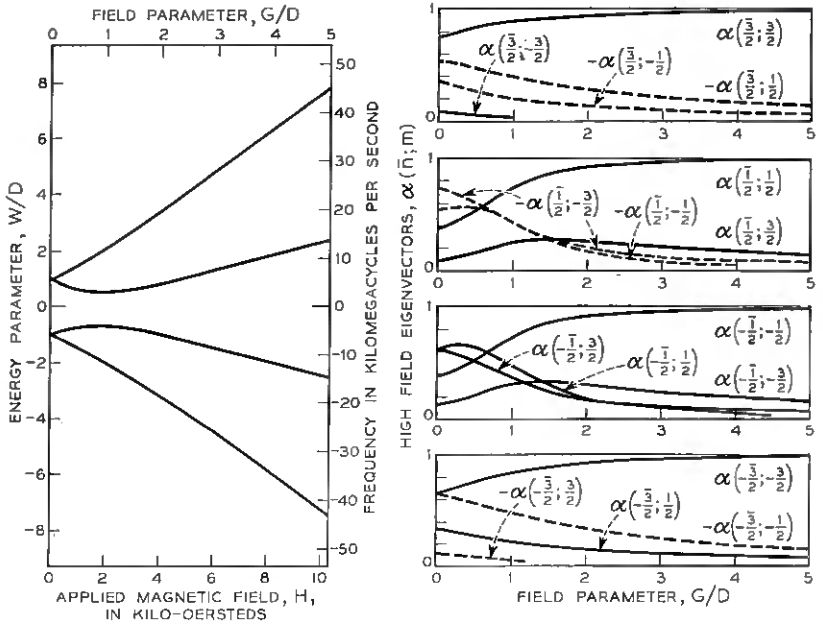


Fig. 7 — Energy levels and eigenvectors at 54.7°.

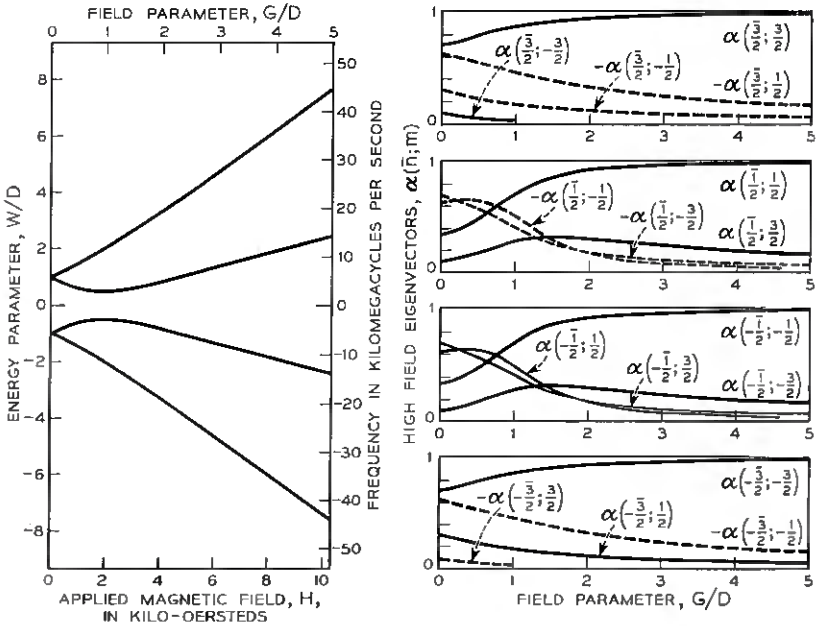


Fig. 8 — Energy levels and eigenvectors at 60°.

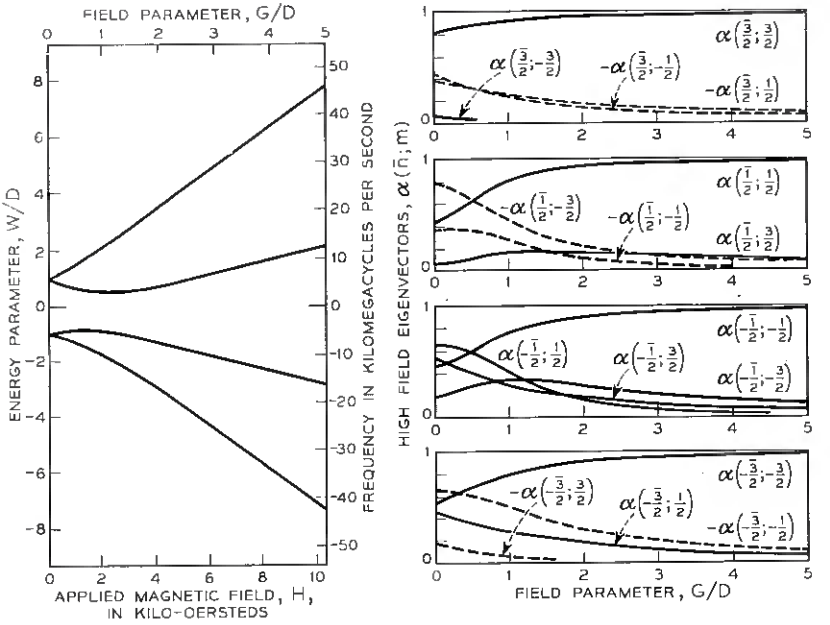


Fig. 9 — Energy levels and eigenvectors at 70°.

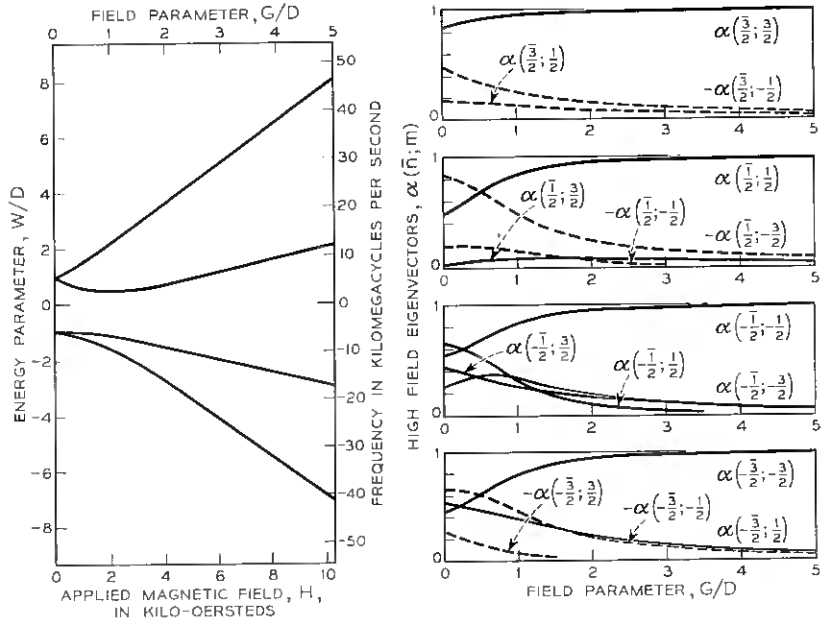


Fig. 10 — Energy levels and eigenvectors at 80°.

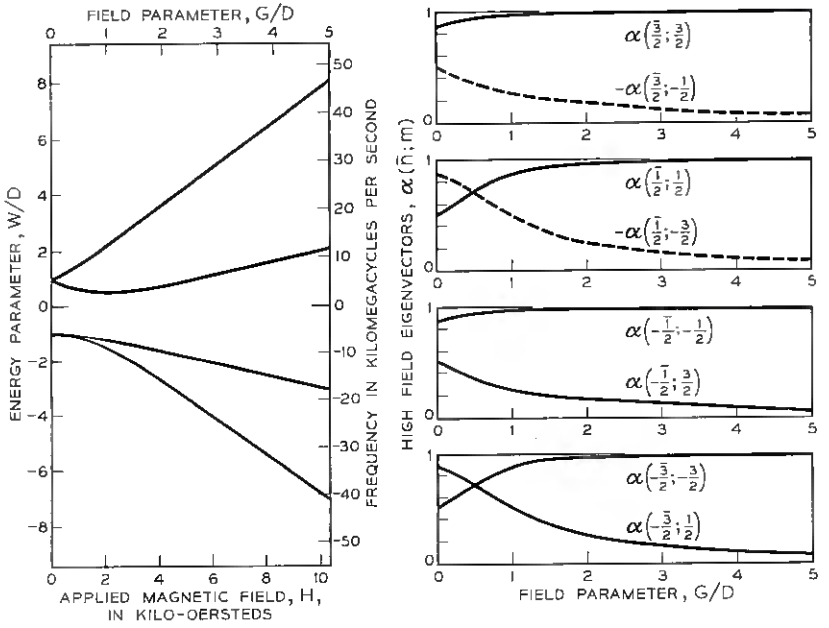


Fig. 11 — Energy levels and eigenvectors at 90°.

This dimensionless representation facilitates computations and reveals more clearly symmetries and singular relations in the energy level scheme. It also permits the use of the same diagrams for ions having the same Hamiltonian but different zero field splitting $2D$. Similar energy level diagrams were computed by P. M. Parker[11] for the case of nuclear spin resonance with nuclear quadrupole splitting present which is described by the same type of Hamiltonian.

As a convenient way to identify the energy levels W, a quantum number $\bar{n}$ ranging from $-\frac{3}{2}$ to $+\frac{3}{2}$ is used *in order of increasing energy*. Thus $W(-\frac{3}{2})$ is the lowest, $W(\frac{3}{2})$ the highest energy level. It is easily shown that, for all angles θ and $x = 0$, $y(-\frac{3}{2}) = y(-\frac{1}{2}) = -1$ and $y(\frac{1}{2}) = y(\frac{3}{2}) = 1$. As a matter of mathematical curiosity, it may be mentioned that, irrespective of θ at $x = 1$, $y(\frac{1}{2}) = 1/2$.

The eigenstates $|\,\bar{n}\rangle$ (using Dirac's "ket" notation) associated with energy levels $W(\bar{n})$ can be expanded in the form

$$|\,\bar{n}\rangle = \sum_{m=-3/2}^{3/2} \alpha(\bar{n};\, m)\,|\,m\rangle. \tag{5}$$

Here, $|\,m\rangle$ are eigenstates of a Zeeman Hamiltonian $\mathcal{H} = g\beta HS_z$. The $\alpha(\bar{n};\, m)$ are amplitudes of eigenvector components or, more briefly,

eigenvectors and form a normalized and orthogonal system of coefficients. With high applied magnetic field H, $|\bar{n}\rangle \rightarrow |n\rangle$ and $\alpha(\bar{n}; n) \rightarrow 1$; therefore, $|m\rangle$ are termed high-field eigenstates and $\alpha(\bar{n}; m)$ high-field eigenvectors.

Eigenvectors $\alpha(\bar{n}; m)$ are obtained as solutions of linear homogeneous equation systems, the matrix of which forms the secular equation (4) with the particular eigenvalue $W(\bar{n})$ inserted. Since this matrix depends on φ, the $\alpha(\bar{n}; m)$ are also functions of φ. The computations were carried out for $\varphi = 0$, with θ restricted to $0 < \theta < \pi/2$ and negative sign of $D = -D'$.

This choice implies that the crystalline axis lies in the positive quadrant of the x-z plane and it results in real eigenvectors $\alpha(\bar{n}; m)$. These are plotted in the right-hand sections of Figs. 1 through 11, adjacent to plots of the corresponding eigenvalues $W(\bar{n})$. Negative $\alpha(\bar{n}; m)$ are indicated by dashed lines.

A nonzero φ would, in general, result in new complex eigenvectors $\alpha'(\bar{n}; m) = [\exp i(m - \bar{n})\varphi]\alpha(\bar{n}; m)$. Taking $\pi/2 < \theta < \pi$ or $\varphi = \pi$ would change the sign of every second eigenvector, that is, of those with $m = n \pm 1$ and $m = n \pm 3$. The same is true for a change of sign of D, but then, in addition, every $\bar{n}$ and m and energy eigenvalue has to be replaced by its negative. It is obvious that such transformations do not change the physical situation as far as transition probabilities are concerned.

IV. TRANSITION PROBABILITIES

There are several ways in which transition probabilities could be evaluated and plotted. One way would be to consider transitions induced by radiation of given polarization. With eigenvectors belonging to the Hamiltonian (3), the obvious RF magnetic field polarizations to consider are those with RF H-field linear and parallel to, or circular and perpendicular to, the applied field. But transitions due to any other polarization could be evaluated as well. Perhaps more natural from a theoretical point of view would be an evaluation of the maximum transition probability. This requires a particular—in general—elliptical, polarization for excitation, which of course should be evaluated, too. All polarizations orthogonal to this (which in general are elliptical as well), and which describe a plane in space having complex components, are associated with zero transition probability. Taking into account these different viewpoints and the six transitions which are possible between four energy levels, it appears that an unrealistically high number of graphs would be necessary to describe the transition probabilities properly.

Furthermore, in maser design it is usually sufficient to know the order of magnitude of transition probabilities of particular lines, because often other factors may be more important. Therefore, no plots of transition probabilities are presented. On the other hand, enough of the pertinent formalism is given below so that any transition probability can be evaluated from the eigenvectors plotted.

Following essentially Bloembergen, Purcell and Pound,[12] with slight generalization, the transition probability w describing the rate of transitions per ion from a lower state $\bar{n}$ to a higher state $\bar{n}' > \bar{n}$ is given by

$$w_{\bar{n}\to\bar{n}'} = \frac{1}{4}\left(\frac{2\pi g\beta H_1}{h}\right)^2 g(\nu - \nu_0) \mid \langle \bar{n}' \mid S_1 \mid \bar{n} \rangle \mid^2. \tag{6}$$

Here H_1 is the amplitude of the exciting RF magnetic field, $g(\nu - \nu_0)$ is a normalized function describing the line shape $\int g(\nu - \nu_0)\, d\nu = 1$, and S_1 is a spin operator reflecting the polarization of the inducing RF magnetic field. If the RF magnetic field is described by the real parts of $H_x = H_1 a e^{i\omega t}$, $H_y = H_1 b e^{i\omega t}$, $H_z = H_1 c e^{i\omega t}$ with "complex direction cosines" a, b, c accounting for elliptical polarization,

$$a^*a + b^*b + c^*c = 1,$$

then

$$S_1 = a^*S_x + b^*S_y + c^*S_z\,. \tag{7}$$

Matrix elements for S_1 occurring squared in (6) are linear combinations of the following three:

$$\langle \bar{n}' \mid S_z \mid \bar{n} \rangle = \sum_{m=-3/2}^{+3/2} m\alpha(\bar{n}'; m)\alpha(\bar{n}; m), \tag{8}$$

$$\langle \bar{n}' \mid S_+ \mid \bar{n} \rangle = \sum_{m=-3/2}^{+1/2} [S(S+1) - (m+1)m]^{1/2}\alpha(\bar{n}'; m+1)\alpha(\bar{n}; m), \tag{9}$$

$$\langle \bar{n}' \mid S_- \mid \bar{n} \rangle = \sum_{m=-1/2}^{+3/2} [S(S+1) - (m-1)m]^{1/2}\alpha(\bar{n}'; m-1)\alpha(\bar{n}; m). \tag{10}$$

The square root in (9) and (10) takes on the values $\sqrt{3}$, 2 and $\sqrt{3}$. For example, with linear polarization in the z-direction, $H_z = H_1 \cos \omega t$ and $S_1 = S_z$. For circular polarization perpendicular to the z-direction,

$H_x = (1/\sqrt{2})H_1 \cos \omega t$, $H_y = \pm(1/\sqrt{2})H_1 \sin \omega t$ and $S_1 = (1/\sqrt{2})S_\pm$. For linear polarization in the x direction, $H_x = H_1 \cos \omega t$ and $S_1 = S_x = \frac{1}{2}(S_+ + S_-)$. Similarly, in the y direction $H_y = H_1 \cos \omega t$ and $S_1 = S_y = (1/2i)(S_+ + S_-)$.

The expression (7), or more correctly, the associated matrix element, can be interpreted as a scalar product of (a^*, b^*, c^*) with $\langle \bar{n}' \mid S \mid \bar{n} \rangle$. It should be noted that, in general, all components can be complex. As a consequence of this interpretation, the maximum transition probability occurs if H_{rf} or (a, b, c) is parallel in space and conjugate complex in phase to $\langle \bar{n}' \mid S \mid \bar{n} \rangle$. Since for real eigenvectors the matrices (8), (9), (10) are all real, it follows that $\langle \bar{n}' \mid S_x \mid \bar{n} \rangle$ and $\langle \bar{n}' \mid S_z \mid \bar{n} \rangle$ are real, whereas $\langle \bar{n}' \mid S_y \mid \bar{n} \rangle$ is imaginary. Thus, for all ruby lines, the polarization for maximum transition probability will be a linear combination of H_x and H_z components with an H_y component in quadrature. In a similar fashion, a set of complex direction cosines can be found which causes the scalar product of (a^*, b^*, c^*) with $\langle \bar{n}' \mid S \mid \bar{n} \rangle$, and hence the transition probability, to vanish. These vectors (a, b, c) describe a plane orthogonal to the vector for maximum transition probability.

It should be noted that frequently the complete formula (6) is not used to evaluate and compare transition probabilities. Instead, usually only the squared matrix element $| \langle \bar{n}' \mid S_1 \mid \bar{n} \rangle |^2$ is computed and this is then compared with a simple standard transition. The obvious standard is the transition $-1/2 \rightarrow +1/2$ of an $S = 1/2$ Zeeman doublet induced by circular polarization. This is described, in our notation, by $| \langle +1/2 \mid (1/\sqrt{2})S_+ \mid -1/2 \rangle |^2 = 1/2$. Accordingly, transitions involving a squared matrix element of order 1 or greater are considered strong, while perhaps 1/100 is typical of weak transitions.

V. SPECIAL CASES

5.1. $\theta = 0°$.

The energy levels are parts of straight lines $y = 1 \pm \frac{1}{2}x$, $-1 \pm \frac{3}{2}x$ with change of slope for some of them at $x = 1$ and 2. Eigenvectors are ± 1 and 0 only, again joined for some levels at $x = 1$ and 2. The minus sign of eigenvectors at 0° has no significance; it is only used to preserve continuity to neighboring angles.

At $\theta = 0°$ and $x < 2$, the labeling of energy levels by high field quantum numbers *in order of increasing energy* is perhaps not the usual one. In this paper, however, it seems appropriate because, with this terminology, in going from $\theta = 0°$ to other orientations, the notation of states

stays the same. It may be pointed out that energy levels defined in this fashion should be considered as continuous functions of applied field without cross-overs (see Fig. 1). The reason is that any off-diagonal perturbation will indeed prevent levels from intercepting by perturbation theory arguments.

Only three transitions are allowed:

$$0 < x < 1: \langle +\tfrac{3}{2} \mid S_+ \mid +\tfrac{1}{2}\rangle^2 = 4$$

$$\langle +\tfrac{3}{2} \mid S_- \mid -\tfrac{1}{2}\rangle^2 = \langle +\tfrac{1}{2} \mid S_+ \mid -\tfrac{3}{2}\rangle^2 = 3,$$

$$1 < x < 2: \langle +\tfrac{3}{2} \mid S_+ \mid -\tfrac{1}{2}\rangle^2 = 4$$

$$\langle +\tfrac{3}{2} \mid S_- \mid +\tfrac{1}{2}\rangle^2 = \langle +\tfrac{1}{2} \mid S_+ \mid -\tfrac{3}{2}\rangle^2 = 3,$$

$$2 < x: \quad \langle +\tfrac{1}{2} \mid S_+ \mid -\tfrac{1}{2}\rangle^2 = 4$$

$$\langle +\tfrac{3}{2} \mid S_+ \mid +\tfrac{1}{2}\rangle^2 = \langle -\tfrac{1}{2} \mid S_+ \mid -\tfrac{3}{2}\rangle^2 = 3.$$

It is interesting to note that, for $0 < x < 2$, one transition requires opposite polarization from the others. This was verified in an experiment. Resonance absorption was measured for this and another transition in a propagating comb-type slow-wave structure having regions of predominantly circular polarization. Reversal of applied magnetic field results in drastic increase of one and reduction of the other line.

5.2. $\theta = 54.74°$, $\cos^2\theta = 1/3$.

For this angle, the fourth-order secular equation reduces to a biquadratic one. The four eigenvalues are $y = \pm[1 + \frac{5}{4}x^2 \pm (3x^2 + x^4)^{1/2}]^{1/2}$. This implies an up-down symmetry $y(-\bar{n}) = -y(\bar{n})$. The closest approach of the two middle eigenvalues is $y(+\frac{1}{2}) - y(-\frac{1}{2}) = 1$ at $x = 1$. A similar symmetry relation holds for eigenvectors $\alpha(-\bar{n}; -m) = (\bar{n}m/|\bar{n}m|)\alpha(\bar{n}; m)$. As a consequence, some transition probabilities for linear polarization are identical, namely

$$\langle -\tfrac{1}{2} \mid S_z \mid -\tfrac{3}{2}\rangle = \langle +\tfrac{3}{2} \mid S_z \mid +\tfrac{1}{2}\rangle$$

and

$$\langle +\tfrac{1}{2} \mid S_z \mid -\tfrac{3}{2}\rangle = -\langle +\tfrac{3}{2} \mid S_z \mid -\tfrac{1}{2}\rangle.$$

The analogous is not true for other polarizations.

5.3. $\theta = 90°$.

The secular equation can be factorized into two quadratic equations with the solutions

$$y(\tfrac{3}{2}) = \frac{x}{2} + (1 + x + x^2)^{1/2},$$

$$y(\tfrac{1}{2}) = -\frac{x}{2} + (1 - x + x^2)^{1/2},$$

$$y(-\tfrac{1}{2}) = \frac{x}{2} - (1 + x + x^2)^{1/2},$$

$$y(-\tfrac{3}{2}) = -\frac{x}{2} - (1 - x + x^2)^{1/2}.$$

Each state contains only two eigenvectors, namely $\alpha(\bar{n};\ n)$ and $\alpha(\bar{n};\ n \pm 2)$. In addition, $\alpha(\bar{n};\ n) = \alpha(\overline{n \pm 2};\bar{n} \pm 2)$ and $\alpha(\bar{n};\ n \pm 2) = -\alpha(\overline{n \pm 2};\ n)$. As a result, transition probabilities between adjacent

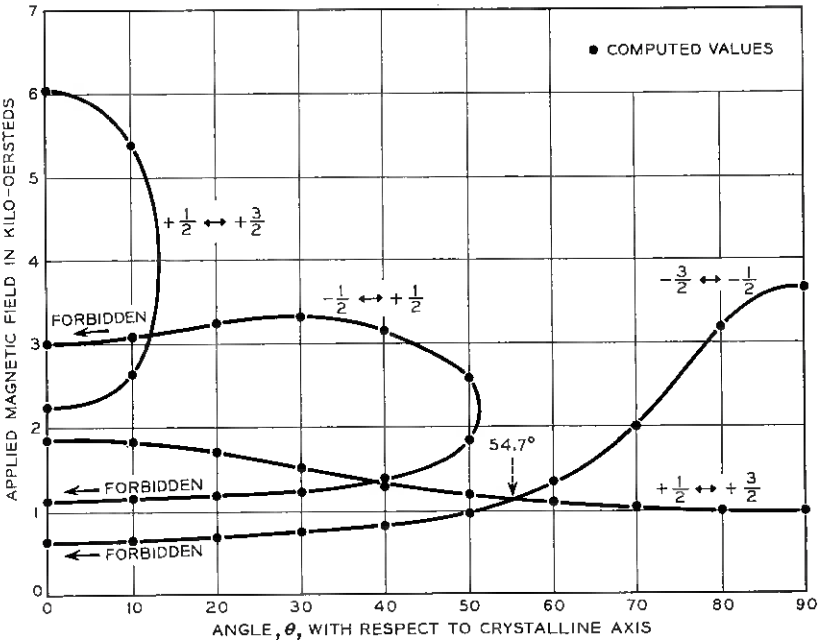


Fig. 12 — Paramagnetic resonance spectrum of Cr^{+++} ions in ruby at signal frequency 5.18 kmc.

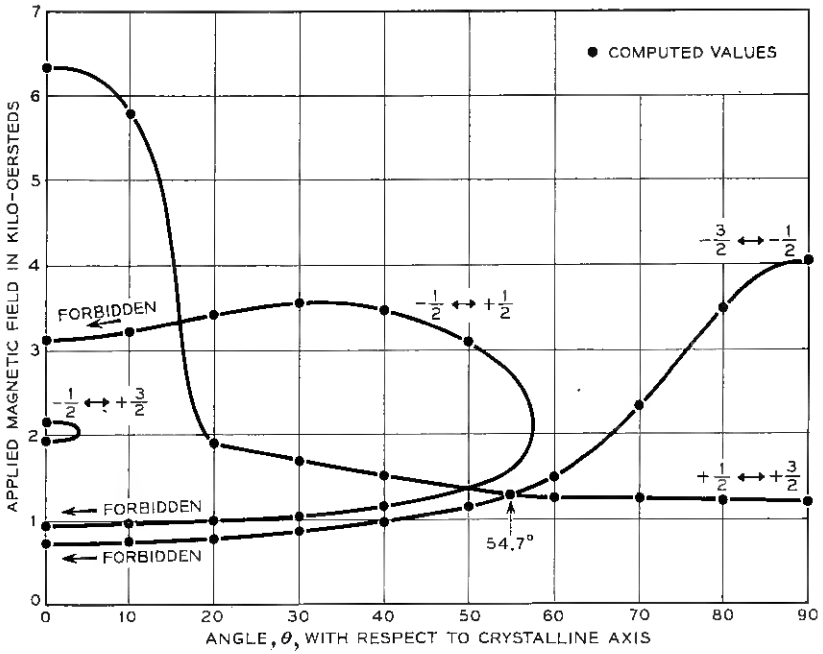


Fig. 13 — Resonance spectrum at 6.08 kmc.

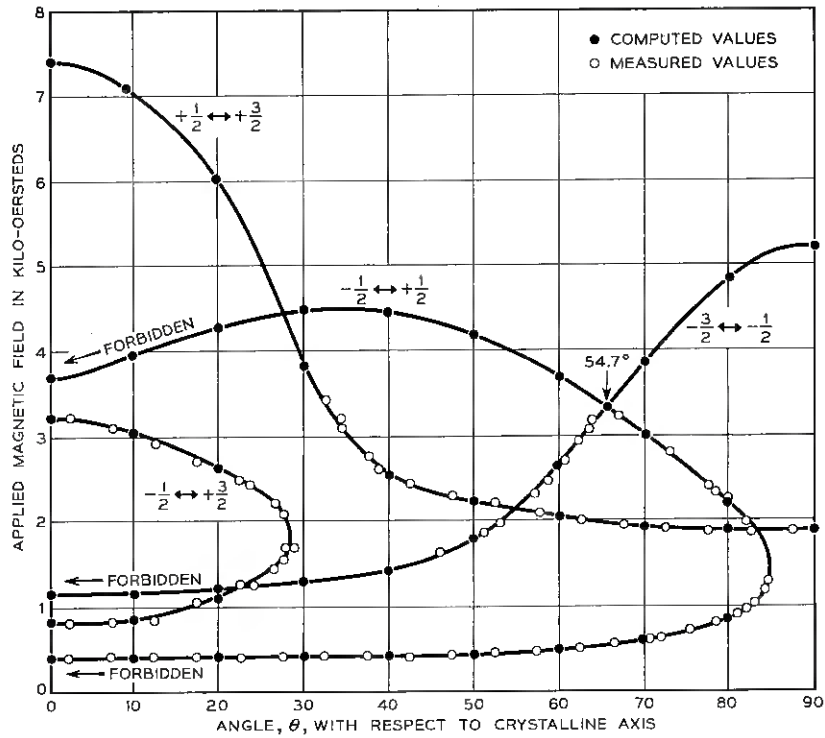


Fig. 14 — Resonance spectrum at 9.30 kmc.

levels $\bar{n} \rightarrow \overline{n+1}$ contain only matrix elements of S_+ and S_-, the same being true for $-\frac{3}{2} \rightarrow +\frac{3}{2}$. Double jumps $\bar{n} \rightarrow \overline{n+2}$ are described by nonvanishing elements of S_z only.

VI. PARAMAGNETIC RESONANCE SPECTRA

In Figs. 12 through 17 some resonance spectra are shown for signal frequencies of 5.18, 6.08, 9.30, 12.33, 18.2 and 23.9 kmc. The plots show resonance fields as functions of the angle between crystalline axis and applied field. Measurements have been carried out at all of these frequencies to varying extents, although measured values are recorded only on Figs. 14 and 15. Generally, these spectra have been used in the laboratory to align ruby crystals by resonance for maser experiments. They have proved accurate to about ±50 gauss.

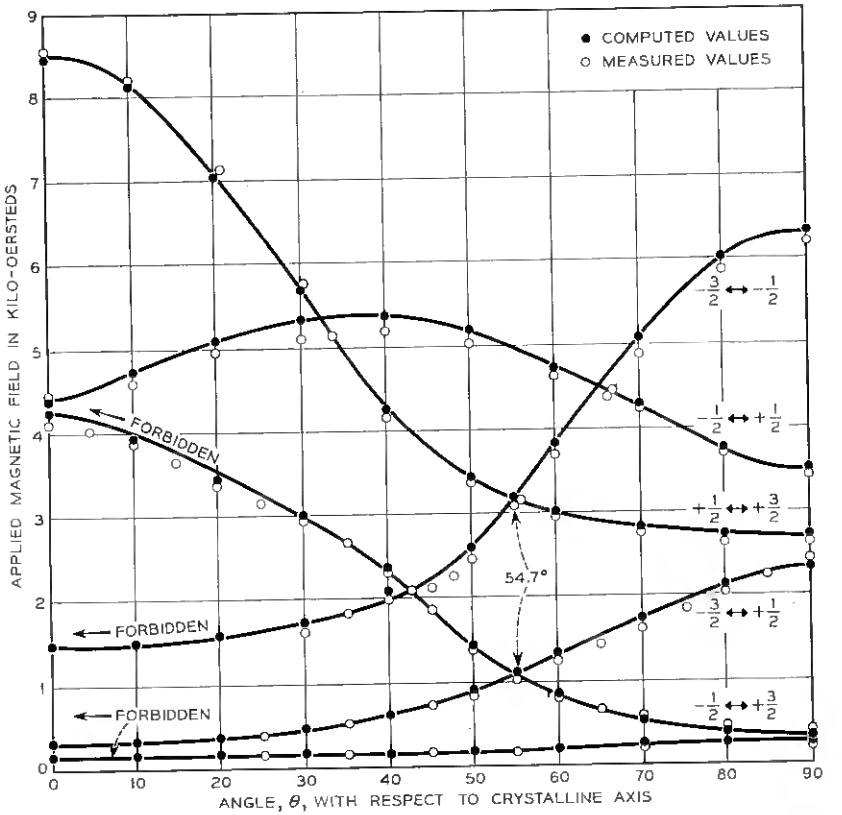


Fig. 15 — Resonance spectrum at 12.33 kmc.

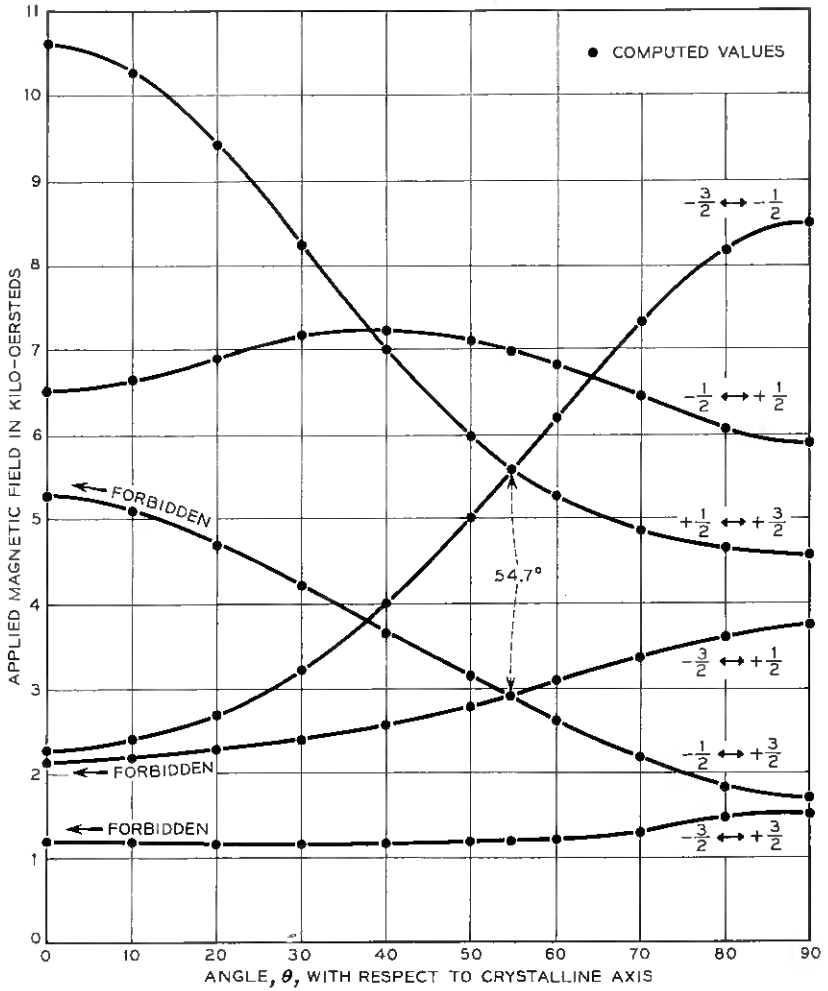


Fig. 16 — Resonance spectrum at 18.2 kmc.

Measurements at 9.3 kmc are an extension of Geusic's work[4] and confirm his results. Results at 12.33 kmc show some discrepancy between theory and experiment, which, however, is believed to be caused by inadequate magnetic field measuring equipment used in an experiment designed for other purposes. As a general rule, the spectra show two looping lines if $\nu < 2D$. Lines marked "forbidden" are strictly forbidden at 0° only. Usually, however, they can be followed quite close to 0° by

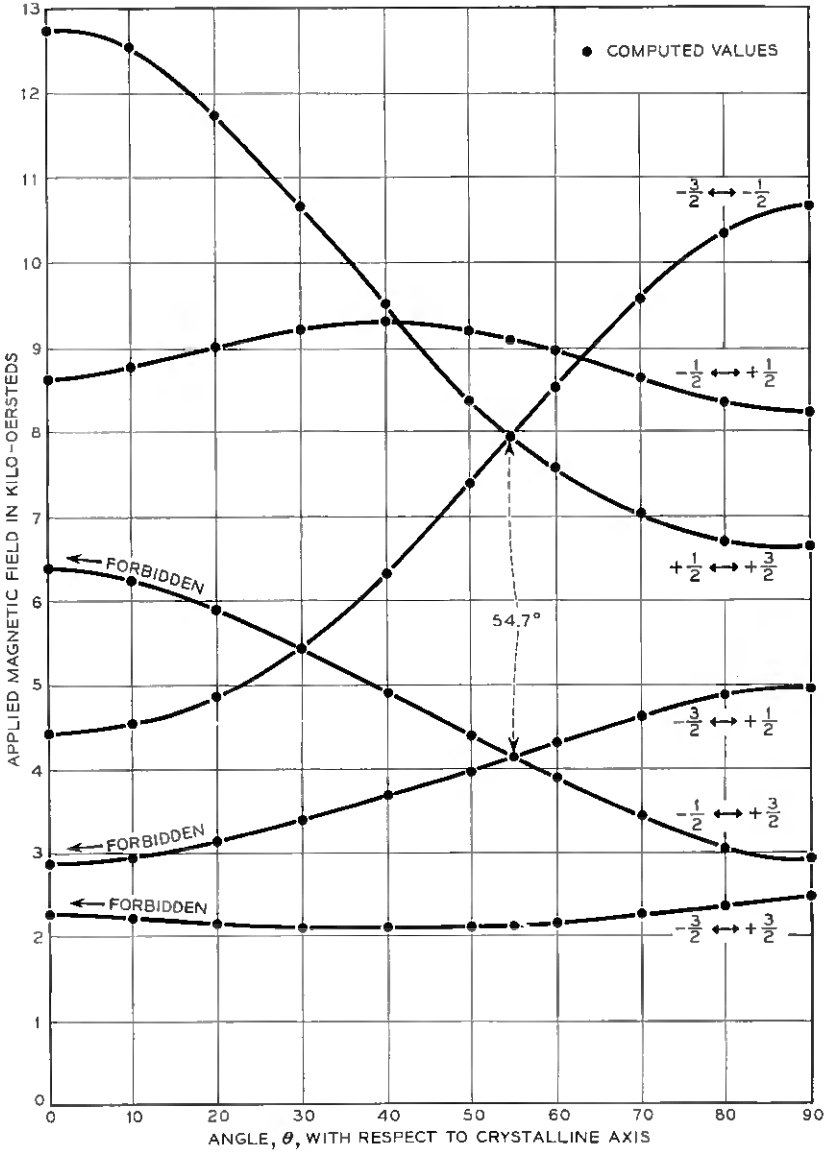


Fig. 17 — Resonance spectrum at 23.9 kmc.

use of more sensitivity in the spectrometer. An exception is the line shown on the graphs having the lowest resonance field at 0° if $\nu < \frac{6}{5}D'$. It has the second lowest resonance field if $\frac{6}{5}D' < \nu < \frac{3}{2}D'$ and the third lowest if $\frac{3}{2}D' < \nu < 3D'$. It originates between $-\frac{3}{2}$ and $+\frac{3}{2}$ eigenstates at 0° and is more strongly forbidden than the other forbidden lines; hence it usually ceases to be measurable at about 30°.

For reasons of symmetry, all lines approach 0° and 90° with zero slope $dH/d\theta$. Experimentally, it has been found that most lines are rather narrow at 90° and similarly at 0°, whereas they broaden in proportion with $dH/d\theta$. This behavior is expected from crystalline imperfections if these can be interpreted as fluctuations throughout the crystal of the direction of the crystalline axis.

VII. ACKNOWLEDGMENT

The author wishes to thank many colleagues at Bell Telephone Laboratories for suggestions in the course of this study. He is particularly indebted to H. E. D. Scovil and J. E. Geusic. Miss M. C. Gray programmed and supervised the numerical calculations.

REFERENCES

1. Bloembergen, N., Proposal for a New-Type Solid-State Maser, Phys. Rev., **104,** October 15, 1956, p. 324.
2. Scovil, H. E. D., Feher, G. and Seidel, H., Operation of a Solid-State Maser, Phys. Rev., **105,** January 15, 1957 p. 762.
3. Scovil, H. E. D., The Three-Level Solid-State Maser, Trans. I.R.E., **MTT-6,** January 1958, p. 29.
4. Makhov, G., Kikuchi, C., Lambe, J. and Terhune, R. W., Maser Action in Ruby, Phys. Rev., **109,** February 15, 1958, p. 1399.
5. Bleaney, B. and Stevens, K. W. H., Paramagnetic Resonance, Rep. Prog. Phys., **16,** 1953, p. 107.
6. Bowers, K. D. and Owen, J., Paramagnetic Resonance II, Rep. Prog. Phys., **18,** 1955, p. 304.
7. Manenkov, A. A. and Prokhorov, A. M., J. Exp. Theor. Phys. (U.S.S.R.), **28,** 1955, p. 762.
8. Geusic, J. E., Phys. Rev., **102,** June 15, 1956, p. 1252; also Ph.D. dissertation, Ohio State Univ., 1958.
9. Zaripov, M. and Shamonin, I., J. Exp. Theor. Phys. (U.S.S.R.), **30,** 1956, p. 291.
10. Bruger, K., Ph.D. dissertation, Ohio State Univ., 1958.
11. Parker, P. M., J. Chem. Phys., **24,** 1956, p. 1096.
12. Bloembergen, N., Purcell, E. M. and Pound, R. V., Relaxation Effects in Nuclear Magnetic Resonance Absorption, Phys. Rev., **73,** April 1, 1948, p. 679.

Paramagnetic Resonance Spectrum of Cr^{+++} in Emerald*

By J. E. GEUSIC, MARTIN PETER and E. O. SCHULZ-DU BOIS

(Manuscript received October 6, 1958)

Paramagnetic resonance for the Cr^{+++} ion in emerald has been observed at X-band (8.2 to 12.4 kmc), K-band (18 to 26.5 kmc) and M-band (50 to 75 kmc). From spectra observed at these frequencies, the spectroscopic splitting factors $g_{\parallel}$, $g_{\perp}$ and D have been determined. The large value of D observed suggests the possible use of emerald as an active material in relatively high microwave-frequency solid-state masers.

I. INTRODUCTION

Survey articles by Bleaney and Stevens[1] and Bowers and Owens[2] list paramagnetic resonance data for crystals containing ions of the transition groups. A careful study of these tabulated data reveals that most of the crystals studied are hydrated or contain several magnetically non-equivalent ions and, therefore, discourages the use of many of these crystals in a practical three-level solid-state maser (3LSSM).

The present study was undertaken with a view to investigating crystals doped with paramagnetic ions which possess good chemical stability and which might be expected to have energy-level schemes suitable for extending the design of solid-state masers to higher microwave frequencies. In this article, paramagnetic resonance spectra of emerald (Cr-doped beryl) are reported. The large zero field splitting observed for the Cr^{+++} ion in this crystal might suggest emerald as a possible material for use in the design of solid-state masers for high microwave-frequency applications.

II. CRYSTAL STRUCTURE AND SPIN HAMILTONIAN OF EMERALD

The structure of beryl[3, 4, 5] is hexagonal with two molecules of $(Be_3Al_2Si_6O_{18})$ per unit cell. In the crystal, SiO_4 tetrahedra share oxygens

* This work is partially supported by the Signal Corps under Contract Number DA 36-039 sc-73224.

to form Si_6O_{18} rings, with each Al linked to six Si_6O_{18} rings. In the lattice, all Al sites are identical and the symmetry at each Al site includes a three-fold axis parallel to the hexagonal or c-axis of the crystal. In emerald, it is found that Cr substitutionally replaces Al in the beryl lattice and is present as Cr^{+++}.

For the Cr^{+++} ion in such a crystalline electric field of three-fold symmetry, it is well known[1] that the paramagnetic resonance spectrum is described by a spin Hamiltonian of the form

$$\mathcal{H} = \beta[g_\parallel H_z S_z + g_\perp(H_x S_x + H_y S_y)] + D[S_z^2 - \tfrac{1}{3}S(S+1)], \quad (1)$$

with $S = \frac{3}{2}$ and with the z-axis taken parallel to the three-fold symmetry axis which in emerald is the c-axis. In the preceding paper[6] the energy levels of the spin Hamiltonian above were discussed for arbitrary values of the parameters $g_\parallel$, $g_\perp$ and D and for arbitrary orientation of the magnetic field H with the z-axis. The notation of the previous paper is adopted for labeling the energy levels of (1). The energy levels are labeled $W(\bar{n})$ where $\bar{n}$ is used to enumerate the levels in order of their energy and is just the high magnetic field quantum number which takes on the values $-\frac{3}{2}, \cdots, +\frac{3}{2}$. For example, $W(-\frac{3}{2})$ represents the lowest energy level and $W(\frac{3}{2})$ represents the highest energy level.

III. EXPERIMENTAL WORK

Initial paramagnetic resonance measurements of Cr^{+++} in a single emerald crystal were made at 9.309 kmc. The spectrometer used for these X-band measurements is similar in design to that described by Feher.[7] At 9.309 kmc and magnetic fields which were available, the spectrum of Cr^{+++} in emerald consisted of a single anisotropic line. The effective g-value, $g^e(\theta)$, of this line is plotted in Fig. 1 as a function of θ, where θ is the angle between the magnetic field H and the c-axis of the crystal. The extreme values of g^e at 9.309 kmc. are $g^e(0°) = 1.973 \pm 0.002$ and $g^e(90°) = 3.924 \pm 0.004$. This line was identified (taking the sign of D negative) as the transition $W(\frac{1}{2}) \rightarrow W(\frac{3}{2})$. The fact that $g^e(90°) \cong 4$ suggests that at 9.309 kmc the frequency of observation is much less than the splitting of the energy levels of (1) in zero field. Under the condition that ν, the frequency of observation, is small compared to the zero field splitting, a perturbation expression for the g^e of the $W(\frac{1}{2}) \rightarrow W(\frac{3}{2})$ transition is given by

$$g^e = [g_\parallel^2 + (4g_\perp^2 - g_\parallel^2)\sin^2\theta]^{1/2}\left[1 - \tfrac{1}{2}\left(\frac{g_\perp \beta H}{2D}\right)^2 F(\theta)\right], \quad (2)$$

where

$$F(\theta) = \frac{3\sin^2\theta(\sin^2\theta - \frac{1}{3})}{\sin^2\theta + \frac{1}{3}}.$$

Specialization of (3) to $\theta = 0°$ and $\theta = 90°$ gives

$$g^e(0°) = g_\parallel,$$

$$g^e(90°) = 2g_\perp\left[1 - \tfrac{3}{4}\left(\frac{g_\perp \beta H}{2D}\right)^2\right],$$

from which it is seen that the zero field splitting $|2D|$ can be computed from measurements of $g^e(90°)$ at two frequencies small compared to $|2D|$. Measurements at 23.983 kmc gave $g^e(90°) = 3.814 \pm 0.004$. From the measurements of $g^e(0°)$ and $g^e(90°)$ at X-band and $g^e(90°)$ at K-band, the constants in (1) were found to be

$$g_\parallel = 1.973 \pm 0.002,$$

$$g_\perp = 1.97 \pm 0.01,$$

$$2D = -52.0 \pm 2.0 \text{ kmc}.$$

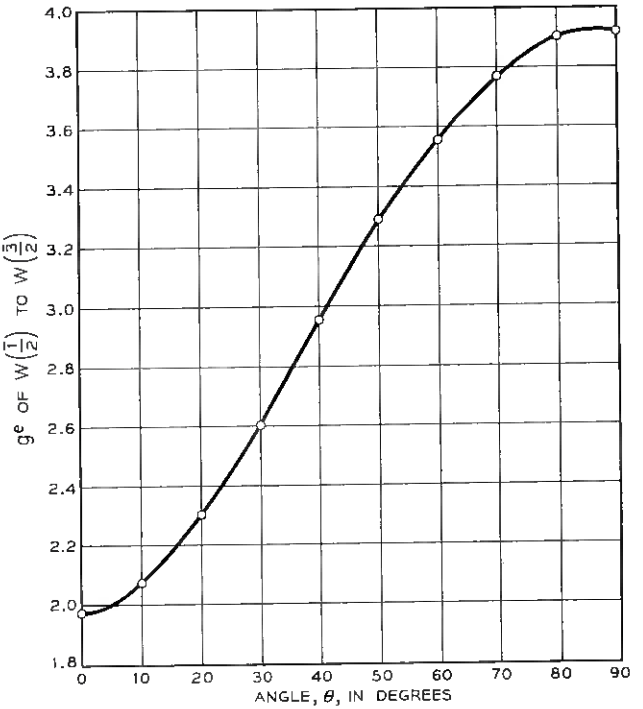


Fig. 1 — Variation of the effective g-value, g^e, of the $W(\frac{1}{2}) \rightarrow W(\frac{3}{2})$ transition at 9.309 kmc.

In order to establish the sign of D, the intensity of the $W(\frac{1}{2}) \rightarrow W(\frac{3}{2})$ line of Cr^{+++} in emerald was compared at two temperatures with the intensity of the same line of Cr^{+++} in ruby. The measurements were carried out with both crystals simultaneously mounted in the X-band spectrometer using temperatures of 78°K and 1.6°K, respectively. The variation of line intensity with temperature can be predicted from Boltzmann statistics if the sign of D is known. Computations were carried out, therefore, on this variation, assuming positive and negative signs of D for both ruby and emerald. The results of these calculations and of the measurement are summarized in Table I. It thus can be concluded that the sign of D for both emerald and ruby is negative.

In order to obtain $2D$ more accurately, measurements were made on Cr^{+++} in emerald at M-band. The millimeter wave paramagnetic resonance spectrometer used was constructed by one of the authors (M. Peter); a block diagram is shown in Fig. 2. Microwave power for this spectrometer is generated by free-running backward wave oscillators (BWO's). Three such BWO's are used to cover the frequency range of 48 to 82 kmc. No resonant cavity is employed in this spectrometer; instead, the sample is situated in "straight" waveguide so that the remarkably wide tuning range of the BWO's can be used. The sensitivity of this spectrometer is comparable to those at low microwave frequencies employing resonant cavities. This is because the loss of sensitivity due to the absence of a cavity is compensated for by higher microwave susceptibility and higher filling factors at these frequencies.

At M-band and for $\theta = 0°$, the two allowed transitions are, in our notation, $W(-\frac{1}{2}) \rightarrow W(\frac{3}{2})$ and $W(-\frac{3}{2}) \rightarrow W(\frac{1}{2})$; they are illustrated in Fig. 3. Both these transitions merge at zero field with the transition frequency being $|2D|$. Both transitions were studied as a function of magnetic field, as shown on Fig. 4. By following them to zero field, the value of the zero field splitting was determined to be $|2D| = 53.6 \pm 0.1$ kmc. From these X-band, K-band and M-band measurements, the best values for the constants in the spin Hamiltonian (1) for

TABLE I

$$\text{Ratio } R = \frac{I_{\text{ruby, } 1.6°\text{K}}}{I_{\text{ruby, } 78°\text{K}}} \times \frac{I_{\text{emerald, } 78°\text{K}}}{I_{\text{emerald, } 1.6°\text{K}}}$$

Computed with sign of D for			Measured R_{exp}
Emerald	Ruby	R_{theor}	
negative	negative	1.8	1.9 ± 0.1
negative	positive	2.6	
positive	negative	0.5	
positive	positive	0.7	

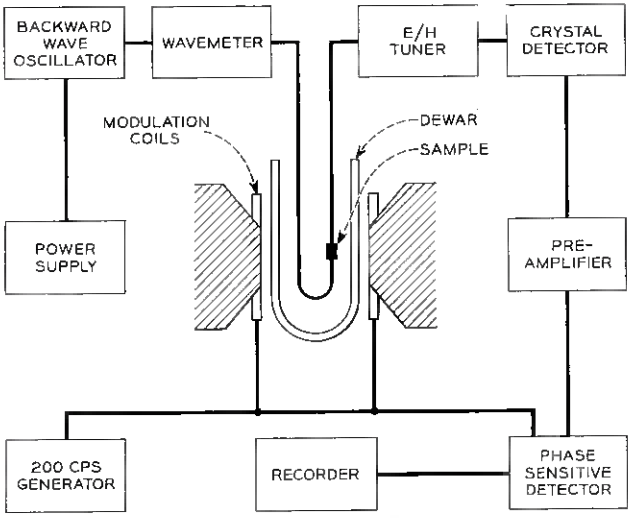


Fig. 2 — Block diagram of the M-band spectrometer.

Cr^{+++} in emerald are

$$2D = -53.6 \pm 0.1 \text{ kmc},$$
$$g_{\parallel} = 1.973 \pm 0.002,$$
$$g_{\perp} = 1.97 \pm 0.01.$$

This value of $2D$ for Cr^{+++} in emerald is, to date, the largest zero field splitting which has been reported for the Cr^{+++} ion in any crystal. Measurements on spin lattice relaxation times in this crystal are planned.

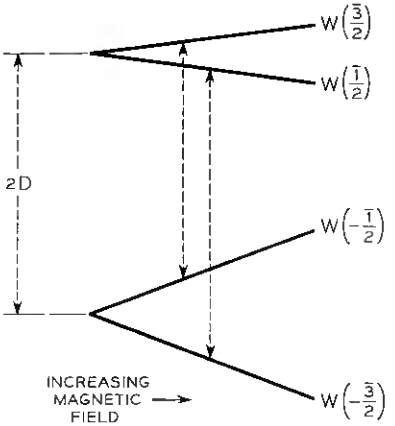


Fig. 3 — Energy level diagram of Cr^{+++} in emerald with applied magnetic field parallel to crystalline axis.

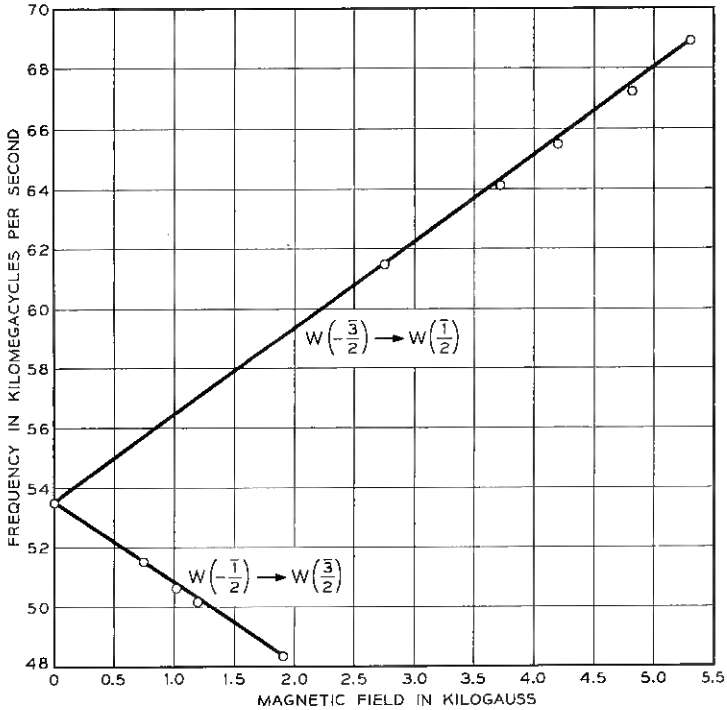


Fig. 4 — Plot of frequencies associated with transitions $W(-\frac{1}{2}) \rightarrow W(\frac{3}{2})$ and $W(-\frac{3}{2}) \rightarrow W(\frac{1}{2})$ at low applied magnetic fields and $\theta = 0°$.

IV. ACKNOWLEDGMENTS

The authors wish to thank H. E. D. Scovil for suggestions throughout the course of the work. We are indebted to S. Geschwind and D. Linn for making the K-band measurements, and to J. B. Mock for valuable assistance with the M-band measurements.

REFERENCES

1. Bleaney, B. and Stevens, K. W. H., Paramagnetic Resonance, Rep. Prog., Phys. **16,** 1953, p. 107.
2. Bowers, K. D. and Owen, J., Paramagnetic Resonance II, Rep. Prog. Phys., **18,** 1955, p. 304.
3. Bragg, W. L. and West, J., Proc. Roy. Soc. (London), **A111,** 1926, p. 691.
4. Bragg, W. L., *Atomic Structure of Minerals*, Cornell Univ. Press, Ithaca, N. Y., 1937.
5. Wyckoff, R. W. G., *The Structure of Crystals*, The Chemical Catalog Co., New York, 1931.
6. Schulz-DuBois, E. O., this issue, pp. 271–290.
7. Feher, G., Sensitivity Considerations in Microwave Paramagnetic Resonance Absorption Techniques, B.S.T.J., **36,** March 1957, p. 449.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

Allison, H. W. and Moore, G. E.

Diffusion of Tungsten in Nickel and Reaction at Interface with SrO, Monograph 3089.

Anderson, P. W.

Absence of Diffusion in Certain Random Lattices, Monograph 3008.

Anderson, P. W. and Talman, J. D.

Pressure Broadening of Spectral Lines at General Pressures, Monograph 3117.

Archer, R. J.

Optical Constants of Germanium: 3600 A to 7000 A, Monograph 3091.

Ballman, A. A., see Laudise, R. A.

Fairbanks, G. and Guttman, N.

Effects of Delayed Auditory Feedback Upon Articulation, Monograph 3022.

Flanagan, J. L. and Saslow, M. G.

Pitch Discrimination for Synthetic Vowels, Monograph 3094.

Fleischer, I. and Kooharian, A.

On the Statistical Treatment of Stochastic Processes, Monograph 3095.

---

* 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.

Nondestructive Memory Employing a Domain Oriented Steel Wire, Monograph 3093.

GRAHAM, R. E.

Communication Theory Applied to Television Coding, Monograph 3096.

GUPTA, S. S. and SOBEI, M.

On the Distribution of a Statistic Based on Ordered Uniform Chance Variables, Monograph 3027.

GUTTMAN, N., see Fairbanks, G.

JOEL, A. E., JR.

Communication Switching Systems as Real-Time Computers, Monograph 3098.

KETCHLEDGE, R. W.

An Introduction to the Bell System's First Electronic Switching Office, Monograph 3099.

KOOHARIAN, A., see Fleischer, I.

LAUDISE, R. A. and BALLMAN, A. A.

Hydrothermal Synthesis of Sapphire, Monograph 3100.

LAX, M.

Generalized Mobility Theory, Monograph 3097.

LAX, M. and PHILLIPS, J. C.

One-Dimensional Impurity Bands, Monograph 3101.

MAYS, J. M., MOORE, H. R. and SHULMAN, R. G.

Improved Nuclear Magnetic Resonance Spectrometer, Monograph 3102.

MIRANKER, W. L.

Reduced Wave Equation With a Variable Index of Refraction, Monograph 3118.

MOORE, G. E., see Allison, H. W.

MOORE, H. R., see Mays, J. M.

MORRISON, J.

Gas Collection and Analysis System in Vacuum Tube Problems, Monograph 3103.

NELSON, L. S. and SPINDLER, G. P.

Sealing Glass to Sapphire, Monograph 3104.

PHILLIPS, J. C., see Lax, M.

PIERCE, J. R.

Proposal for an Explanation of Limens of Loudness, Monograph 3105.

SASLOW, M. G., see Flanagan, J. L.

SHULMAN, R. G., see Mays, J. M.

SLICHTER, W. P.

Study of High Polymers by Nuclear Magnetic Resonance, Monograph 3049.

SOBEL, M., see Gupta, S. S.

SPINDLER, G. P., see Nelson, L. S.

SUHL, H.

Nonlinear Behavior of Ferrites at High Microwave Signal Levels, Monograph 2734.

TALMAN, J. D., see Anderson, P. W.

TURNER, D. R.

Electropolishing Silicon in Hydrofluoric Acid Solutions, Monograph 3107.

WHITE, D. L.

High Q Quartz Crystals at Low Temperatures, Monograph 3108.

Contributors to This Issue

Floyd K. Becker, B.S. (E.E.), 1945, University of Colorado; M.S. (E.E.), 1947, California Institute of Technology; Mountain States Telephone and Telegraph Company, 1947–51; Bell Telephone Laboratories, 1951—. Mr. Becker's work at Mountain States was in dial equipment engineering. Since joining Bell Laboratories he has been concerned with research in acoustics and underwater sound transmission and more recently with development of experimental picture transmission systems. Member I.R.E., Acoustical Society of America, Tau Beta Pi, Eta Kappa Nu.

Václav E. Beneš, A.B., 1950, Harvard College; M.A., Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical systems research, involving stochastic processes describing the passage of traffic through a switching system. Member Association for Symbolic Logic, Mind Association, Institute of Mathematical Statistics, American Mathematical Society, Phi Beta Kappa.

T. H. Crowley, B.E.E., 1948, M.A., 1950, and Ph.D., 1954, Ohio State University; Bell Telephone Laboratories, 1954—. He has been engaged in studies of switching systems problems, including application of magnetic devices to switching circuits and studies of time-varying networks for time-division switching. He took part in the design of an analog-to-digital encoder and is at present in charge of a group engaged in theoretical analysis of switching systems problems. Member I.R.E., Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Hayden W. Evans, B.A., 1934, Ohio Wesleyan University; B.S. in E.E., 1936, University of Michigan; Bell Telephone Laboratories, 1936—. Mr. Evans' early work was on transmission engineering problems on open wire and cable circuits. During World War II he was engaged in development of radar, radar test equipment and countermeasures equipment. After the war he was concerned with mobile radio systems engineering and later with planning and evaluation of new radio relay systems. He is at present in charge of a group engaged in broadband systems engineer-

ing. Senior member I.R.E.; member Acoustical Society of America, Tau Beta Pi, Sigma Xi, Phi Kappa Phi.

J. E. Geusic, B.S., 1953, Lehigh University; M.S., 1955, and Ph.D., 1958, Ohio State University; Bell Telephone Laboratories, 1958—. He is engaged in research and development work on the solid state maser. Member Sigma Xi, Pi Mu Epsilon.

U. F. Gianola, B.Sc., 1948 and Ph.D., 1951, University of Birmingham (England); Royal Aircraft Establishment, 1951; post-doctoral fellow, University of British Columbia, 1951–53; Bell Telephone Laboratories, 1953—. As a member of the transmission research department he took part in experimental and theoretical studies of transmission line structures, analyses of a new magnetostrictive transducer, the application of the solar battery to communications channels and fundamental studies of the effects of ion bombardment on semiconductors. Since transferring to communications techniques research he has been engaged in studies of solid-state memory and logic devices. Member American Physical Society, Research Society of America.

S. D. Hathaway, B.E.E., 1947, University of Virginia; M.S.E.E., 1950, Virginia Polytechnic Institute; M.S.E.E., 1952, University of Illinois; Bell Telephone Laboratories, 1952—. Mr. Hathaway has been engaged in systems engineering on microwave radio relay systems. He is at present in charge of a group engaged in light-route radio systems engineering. Member I.R.E., Tau Beta Pi, Eta Kappa Nu.

John R. Hefele, B.S. in E.E., 1929, E.E., 1932, Cooper Union; Fordham University, 1931–1932; Columbia University, 1932–33; Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925—. Mr. Hefele has specialized in television systems research. He has participated in all of the television developments and demonstrations of Bell Laboratories including the first demonstration in 1927. During World War II he was engaged in investigation and development of airplane detection and automatic-following radar devices. In 1956 he transferred to the visual systems group in the Transmission Research Department, where he is currently concerned with special problems of visual transmission over restricted-bandwidth channels. Member I.R.E., Society of Motion Picture and Television Engineers, Electronics Industries Association.

H. A. Helm, B.S. in physics, 1942, Massachusetts Institute of Technology; M.S. in E.E., 1956, Stevens Institute; Bell Telephone Laboratories, 1945—. His first work was in military systems development in-

volving radar and analog computers. He was later concerned with design, systems analysis and programming for digital computers, and he was an instructor in the Digital Techniques Laboratory. He is now engaged in military systems research. Member I.R.E., Association for Computing Machinery.

Raymond W. Ketchledge, B.S. and M.S., 1942, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1942—. Mr. Ketchledge first worked on infra-red detection and underwater sound systems. Later he assisted in the development of the Key West-Havana submarine cable system. His next assignment was the responsibility for equalization, regulation and other system aspects of the L3 coaxial carrier system. In 1953 he was appointed Electron Tube Development Engineer responsible for gas tube and storage tube development. He was named Switching Systems Development Engineer in 1954 responsible for memory systems and switching networks for electronic switching. In 1956 he was appointed Assistant Director of Switching Systems Development III. Senior Member I.R.E.; member New York Academy of Sciences, Sigma Xi.

V. O. Mowery, B.E., 1954, Johns Hopkins University; M.S., 1956, California Institute of Technology; Bell Telephone Laboratories, 1956—. Mr. Mowery has had CDTP assignments in semiconductor surface studies and development of PCM repeaters. His present assignment is in development of the flying spot store for electronic switching systems. Member I.R.E., American Physical Society, Tau Beta Pi, Sigma Xi.

Martin Nesenbergs, Technische Hochschule, Karlsruhe and Stuttgart (Germany); B.S., 1952, University of Denver; M.S., 1958, New York University; Bell Telephone Laboratories, 1953—. Since completing the CDTP course he has been engaged in development work on the flying spot store for electronic switching. Member American Physical Society, Sigma Pi Sigma, Pi Mu Epsilon.

Martin Peter, Dipl., 1952, Swiss Federal Institute of Technology; Ph.D., 1955, Massachusetts Institute of Technology; research associate, M.I.T., 1955–57; Bell Telephone Laboratories, 1957—. Mr. Peter has been engaged in fundamental research involving microwave spectroscopy in the physics of solids. Member American Physical Society, Sigma Xi.

John R. Pierce, B.S., 1933, M.S., 1934, and Ph.D., 1936, California Institute of Technology; Bell Telephone Laboratories, 1936—. Director

of Research-Communications Principles at Bell Laboratories. He has specialized in research on electron tubes, microwave research, electronic devices for military applications and communications circuits. Mr. Pierce has been granted 55 patents and is the author of four books. For his research leading to the development of the beam traveling wave tube, he was awarded the 1947 Morris Liebmann Memorial Prize. Fellow I.R.E., American Physical Society; member Acoustical Society of America, American Institute of Electrical Engineers, National Academy of Sciences, British Interplanetary Society.

Lincoln P. Rice, B. of E.E., 1953, and M.S. in E.E., 1955, Georgia Institute of Technology; Bell Telephone Laboratories, 1954—. After completing the Communications Development Training Program course, Mr. Rice worked on personal radio signaling systems in the Special Systems Engineering Department. He is now engaged in work on data transmission systems. Member I.R.E., Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi, Sigma Xi.

W. W. Rigrod, B.S. in E.E., 1934, Cooper Union Institute of Technology; M.S. in Eng., 1941, Cornell University; D.E.E., 1950, Polytechnic Institute of Brooklyn; State Electrotechnical Institute (U.S.S.R.), 1935–39; Westinghouse Electric Corporation, 1941–51; Bell Telephone Laboratories, 1951—. His work has been related principally to the study and development of electron tubes, both the gaseous-discharge and the high-vacuum types. Member I.R.E., American Physical Society, Sigma Xi.

E. O. Schulz-DuBois, Dipl. phys., 1950, and Dr. phil. nat., 1954, Johann Wolfgang Goethe University (Germany); Purdue University, 1954–55; Raytheon Manufacturing Co., 1956–57; Bell Telephone Laboratories, 1957—. At Purdue Mr. Schulz-DuBois was engaged in solid state research, particularly on paramagnetic resonance in irradiated semiconductors. At Raytheon he was concerned with the development of ferrite materials and devices. After joining Bell Laboratories he worked for a short time on low-frequency ferrite isolators. His present work is with paramagnetic materials and slow-wave structures for application to solid state maser devices.

W. M. Sharpless, B.S. (E.E.). 1928 and E.E., 1951, University of Minnesota; Bell Telephone Laboratories, 1928—. As a member of the Radio Research Department he worked for several years on problems associated with transatlantic short-wave radio reception and took part in studies of the angle of arrival of microwaves. During World War II

he was concerned with development of radar systems and later with design of artificial dielectrics and microwave antennas. More recently he has been associated with studies of point-contact rectifiers and low-level power measurements in the millimeter-wave field. Fellow I.R.E.; member American Physical Society, Scientific Research Society of America.

W. T. WINTRINGHAM, S.B., 1924, Harvard Engineering School; American Telephone and Telegraph Company, 1924–34; Bell Telephone Laboratories, 1934—. At A. T. & T. Mr. Wintringham was engaged in studies of radio telephone systems and transatlantic radio telephone. He later worked on UHF and VHF systems and development and installation of special short wave antennas. During World War II he worked on military projects and then took part in studies of television transmission systems, application of information theory to television and color television. Since 1956 Mr. Wintringham has been in charge of the visual systems research group involved in such projects as facsimile, slow-scan television and picturephone. Fellow I.R.E., American Association for the Advancement of Science, Society of Motion Picture and Television Engineers; member Acoustical Society of America, Optical Society of America, Tau Beta Pi.