THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 50 July-August 1971 Number 6

Mathematical Models of Computation Using Magnetic Bubble Interactions

By A. D. FRIEDMAN and P. R. MENON

(Manuscript received February 10, 1971)

This paper considers the computational capabilities of different mathematical models of magnetic bubble interactions. A specific model was studied earlier by R. L. Graham, who showed that there exist combinational functions of 11 or more variables that cannot be computed by this model. This paper extends his results by introducing different types of interactions which seem to be practical and enable the computation of all combinational functions. The problem of efficient computation from the points of view of time and space requirements and the geometrical requirements imposed by the fact that interactions can occur only between physically adjacent locations are also examined. Finally, a model in which computations are carried out by applying uniform magnetic fields to the entire platelet, with individual access limited to locations along the periphery, is presented.

I. INTRODUCTION

Cylindrical magnetic domains in certain orthoferrite materials have been investigated extensively in recent years.[1–3] These domains, commonly referred to as bubbles, have the property that they can be moved within the material by the application of suitable external magnetic fields. The motion of a bubble is also dependent on other bubbles in its vicinity. The locations of bubbles can be restricted to a finite set of possible positions in an orthoferrite platelet. The presence or absence of a bubble at a particular location may be treated as representing the values of a binary variable. Thus, magnetic bubbles seem to have natural applications in performing memory and logic functions.

In a recent paper, R. L. Graham[4] considered the computational capabilities of a particular mathematical model of magnetic bubble interaction. The only type of interaction allowed in this model is represented by an instruction of the type (x, y) where x and y are distinct adjacent locations in the orthoferrite platelet. Application of this instruction results in moving the bubble in location x to location y if the latter does not contain a bubble prior to the application of the instruction. If x does not contain any bubble or y already has a bubble, no transfer of bubbles takes place. Two locations, only one of which contains a bubble, are used to represent each binary variable and its complement. Graham has shown that only a small fraction of all combinational functions of 11 or more variables can be computed by this model.

In this paper, we review Graham's results and extend his model by introducing different types of interactions which seem to be practical and enable the computation of all combinational functions. We then examine the problem of efficient computation from the points of view of space and time requirements. The geometrical requirements, necessitated by the fact that bubble interactions can occur only between physically neighboring locations, are also examined. Finally, we consider a model in which computations are carried out by applying magnetic fields to the entire platelet and individual access is limited to locations along the periphery.

II: MODELS OF BUBBLE INTERACTIONS

Let us first consider the model studied by Graham.[4] All locations are assumed to be adjacent to one another so that interaction between any pair of locations is possible. If S is the set of all possible bubble locations and n is the number of such locations, let Q be the set of all 2^n subsets of S. If a subset $X \varepsilon Q$ containing bubbles represents the values of the variables of a function f, then f is a mapping from Q onto Q. The function f is realized by applying a program P to S with bubbles in locations contained in X. If the set of resultant bubble locations is

X^P, then $X^P = f(X)$ and $X^P \varepsilon Q$. The program P consists of a sequence of instructions of the form $e = (a, b)$ such that $X^e = (X - \{a\}) \cup \{b\}$, if $a \varepsilon X$, $b \not\varepsilon X$, and $X^e = X$ otherwise. The nondecreasing overlap (NDO) theorem due to Shockley and proven in Ref. 4 points out some of the limitations of this model of interaction. The NDO theorem is repeated below in the interest of completeness.

Theorem 1 (NDO Theorem): Let X_1 and X_2 be two initial sets of locations of bubbles and let P be any program (of the type discussed above). Then $| X_1^P \cap X_2^P | \geqq | X_1 \cap X_2 |$, where $| X |$ is the cardinality of the set X.

The NDO theorem precludes the possibility of certain types of computation using this model of bubble interaction. For example, it is impossible to have a program P such that $\{a, b\}^P = \{c, d\}$ and $\{a\}^P = \{e\}$. Another consequence of the NDO theorem is the nonexistence of a replicating program. That is, there exists no program P^* such that if S and S' are nonintersecting sets of vertices and $X \subset S$, P^* creates a set $X' \subset S'$ without disturbing X, so that there is a one-to-one correspondence between X and X'. In order to be able to compute all combinational functions, it may be necessary to do one or both of the following: (*i*) Include other types of interactions. (*ii*) Code the inputs and the outputs.

Let us consider the following set of instruction types, where the first type is the one we have discussed above.

Name	Notation	Resultant Bubble Locations
I. Bubble transfer	$e = (a, b)$	$X^e = \begin{cases} (X - \{a\}) \cup \{b\} \text{ if } a \varepsilon X, b \not\varepsilon X; \\ X \text{ otherwise.} \end{cases}$
II. Bubble splitting	$e = (a, b)_S$	$X^e = \begin{cases} X \cup \{b\} \text{ if } a \varepsilon X; \\ X \text{ otherwise.} \end{cases}$
III. Bubble annihilation	$e = (0, a)$	$X^e = X - \{a\}$.
IV. Bubble creation	$e = (1, a)$	$X^e = X \cup \{a\}$.

In the above set, type II instructions are necessary for replication. Type III instructions have the property that $|X^e| < |X|$ and are therefore necessary to compute $\{a, b\}^P = \{c\}$. Instructions of type IV are necessary for performing computations of the type $\varphi^P = \{a\}$, where φ is the null set.

Theorem 2: The four types of instructions (viz., bubble transfer, splitting, annihilation, and creation) are not sufficient for performing all computations.

Proof: Consider a program P such that $X^P = S - X$. Let the first instruction of the program be e_1. We show that a program P for computing $S - X$ using only the four types of instructions discussed above does not exist, by showing that for each type of instruction there exist two sets $X_1 \neq X_2$ such that $X_1^{e_1} = X_2^{e_1}$. Since $S - X_1 \neq S - X_2$, the program cannot compute $S - X$ for all X.
Case 1: Let $e_1 = (a, b)$.

Let $X_2 = (X_1 - \{a\}) \cup \{b\}$; $X_1, X_2 \subset S$; $a \in X_1$; $b \notin X_1$. Then

$$X_1^{e_1} = X_2^{e_1} = X_2 \quad \text{and} \quad X_1^P = X_2^P .$$

Case 2: Let $e_1 = (a, b)_S$.
Let $a \in X_1$, $b \notin X_1$, and $X_2 = X_1 \cup \{b\}$. Then

$$X_1^{e_1} = X_2^{e_1} = X_2 \quad \text{and} \quad X_1^P = X_2^P .$$

Case 3: Let $e_1 = (0, a)$.
Let $a \in X_1$, and $X_2 = X_1 - \{a\}$. Then

$$X_1^{e_1} = X_2^{e_1} = X_2 \quad \text{and} \quad X_1^P = X_2^P .$$

Case 4: Let $e_1 = (1, a)$.
Let $a \in X_1$, and $X_2 = X_1 - \{a\}$. Then

$$X_1^{e_1} = X_2^{e_1} = X_1 \quad \text{and} \quad X_1^P = X_2^P .$$

Thus there exists no program P using only the four types of instructions that can compute $S - X$ for all $X \subset S$. This implies that complementation cannot be done (using only these types of instructions) if the value of a Boolean variable is represented by the presence or absence of a bubble in a particular location.

The other approach mentioned earlier for improving the computational capabilities of magnetic bubbles is the use of suitable codes. Graham[4] has used two bubble locations x_0 and x_1 to represent the value of a binary variable x. $x = 0$ is represented by $x_0 = 1$, $x_1 = 0$ and $x = 1$ by $x_0 = 0$, $x_1 = 1$. However, Graham has shown that this representation and the bubble transfer instruction are not sufficient for realizing all Boolean functions. By enumerating the number of distinct programs, he has proven the following theorem.
Theorem 3: There exists a Boolean function of 11 variables which cannot be realized by a program of bubble transfer instructions.

It is not known whether there exists some coding of the variables which permits the realization of all Boolean functions using only the bubble transfer instruction. However, the four types of instructions together with a coding of the type used by Graham are sufficient for realizing all Boolean functions.

Theorem 4: All combinational functions can be computed using the four types of instructions and a suitable coding of inputs.

Proof: Any combinational function can be realized using the following method. Let the set of all possible bubble locations be denoted by M, where $M = S \cup T$ and $S \cap T = \varphi$. As before, values of the input variables are represented by bubbles in subsets of S, and let each variable be represented by two locations x_0 and x_1. Let T be the set of possible bubble locations that serves as temporary storage. Any given combinational function f and its complement $\bar{f}$ are expressed in the sum of products form. The following procedure is used to realize the function f:

(*i*) Clear the output locations f_0 and f_1 by $(0, f_0)(0, f_1)$.

(*ii*) Clear temporary storage by $(0, y_{i0})(0, y_{i1})$ for all $y_{i0}, y_{i1} \varepsilon T$.

(*iii*) Copy the values of the input variables into temporary storage using bubble splitting instructions $(x_{i0}, y_{i0})_S(x_{i1}, y_{i1})_S$ for all $x_{i0}, x_{i1} \varepsilon S$.

(*iv*) If x_i^* is a variable appearing in a product term, where x_i^* represents x_i or $\bar{x}_i$, the term $\prod x_j^*$ is computed by the set of instructions (y_{1a}, y_{ib}) for all $j \neq i$ such that x_j^* is contained in the product term and $a = 0$ if $x_i^* = \bar{x}_i$, $a = 1$ if $x_i^* = x_i$, $b = 0$ if $x_j^* = \bar{x}_j$, and $b = 1$ if $x_j^* = x_j$.

(*v*) (y_{ia}, f_1) puts the OR of all terms computed so far in f_1.

(*vi*) If there are additional terms, go to step *ii* and repeat for next term.

(*vii*) If there are no more terms in f, repeat steps *ii* through *vi* for $\bar{f}$, replacing step *v* by (y_{ia}, f_0).

The above procedure merely shows that all combinational functions can be computed using the four types of instructions and the particular type of coding used. Both the program and the coding could be made more efficient.

The property of the code that made the computation of all combinational functions possible is the elimination of the need for complementation, which was shown to be impossible to perform with the four types of instructions used. A more general code which has this property is an *m*-out-of-*n* (m/n) code[5] where n bubble locations, out of which exactly m contain bubbles, are used to represent each input combination. A procedure analogous to that given above may be used for computing any combinational function using this code. The case discussed above, where each variable is represented by two bubble locations, is a special case of the m/n code and is referred to as the autosynchronous code.[5]

The m/n code is more efficient than the autosynchronous code in terms of the number of bubble locations required for encoding a given number of binary variables. Whereas $2k$ locations are necessary to represent k binary variables, the number of locations required with an m/n code is such that $\binom{n}{m} \geqq 2^k$. Since $m = [n/2]$, where $[x]$ is the integral part of x, gives the largest value of $\binom{n}{m}$, we have $\binom{n}{[n/2]} \geqq 2^k$. Using Stirling's approximation, we have $[n - \frac{1}{2} \log_2 n] \geqq k$. For large values of k, the $[n/2]/n$ code requires fewer locations than the autosynchronous code as the following sample values indicate:

k	$n([n/2]/n$ code)	$2k$ (autosynchronous code)
14	16	28
29	32	58
61	64	122

A given type of instruction is said to be *computationally complete* if all combinational functions can be computed using programs containing only instructions of the given type, assuming that any location may be initially cleared and a bubble may be initially inserted in any location. That is, instructions of the type $(0, x)$ and $(1, x)$ may be used for initialization, prior to the application of the program. Note that the bubble transfer instruction discussed earlier is not computationally complete under our definition, although bubble transfer and bubble splitting together are sufficient to compute any combinational function. (The need for clearing locations within the program in the procedure discussed earlier can be eliminated by using a sufficient number of temporary locations.) We shall now consider some computationally complete instruction types that may be realizable by bubble interactions.

Consider an instruction $e = (x, y)_d$, defined by

$$\begin{aligned} X^e &= X - \{y\} \qquad \text{if } x, y \in X \\ &= (X - \{x\}) \cup \{y\} \qquad \text{if } x \in X,\ y \notin X \\ &= X \text{ otherwise.} \end{aligned}$$

If the presence or absence of a bubble in a location is treated as the value of a binary variable, and if the value of a variable after the application of an instruction $(x, y)_d$ is denoted by the corresponding primed variable, the instruction can be represented by the following equations:

$$x' = xy \quad \text{and} \quad y' = x \oplus y$$

where $\oplus$ represents exclusive-OR.

We prove that the instruction $(x, y)_d$ is computationally complete by showing that we can perform duplication and also realize the NAND function using only instructions of this type. It is well known that any combinational function can be realized using only two-input NAND gates if duplication (fan-out) is allowed. Denoting locations that initially contain bubbles by lower-case letters with the subscript 1, the NAND of two variables x and y can be realized by the sequence of instructions $(x, y)_d(x, a_1)_d$; which results in, $x' = xy$, $y' = x \oplus y$, $a_1' = 1 \oplus xy = \overline{xy}$. Duplication can be performed by the sequence $(x, b_1)_d(b_1, c_1)_d$, resulting in $x' = x$, $b_1' = 1 \oplus x = \bar{x}$, $c_1' = 1 \oplus \bar{x} = x$. Thus all combinational functions can be computed without special encodings.

A timing problem associated with this instruction[*] must be noted. Two applications of the instruction $(x, y)_d$ results in $xy(x \oplus y) = 0$ in location x and $xy \oplus x \oplus y = x + y$ in location y. Physically, if $x = y = 1$ when the instruction $(x, y)_d$ is applied, the bubble in y is destroyed. If the field effecting the execution of the instruction is still applied, the bubble in location x is transferred to y. Thus, this instruction requires application of the field for a fixed interval of time. We may say that the instruction is pulse-width dependent. For each of the other instruction types considered, application of the same instruction a second time does not result in any further change in either location.

Another type of computationally complete instruction is the conditional transfer[†] denoted by $e = (x, y)z$ and defined by

$$\begin{aligned} X^e &= (X - \{x\}) \cup \{y\} \qquad \text{if } x, z \in X,\ y \notin X \\ &= X \text{ otherwise.} \end{aligned}$$

Duplication can be performed by the instruction $(a_1, a_0)x$ where a_0 and a_1 denote an empty location and a location with a bubble respectively. After the application of the instruction, locations a_0 and x will contain the variable x. Similarly, the sequence $(x, y)b_1$, $(b_1, b_0)x$,

[*] In practice, there exists an interaction involving three locations x, y, z, such that if $y = 0$, $y' = x \oplus z$; $x' = z' = xz$. The timing problem mentioned above does not occur in this case.

[†] The interaction involving three locations, mentioned in the previous footnote, may be represented by the following sequence of conditional transfer instructions, if $y = 0$ initially: $(x, y)\bar{z}$; $(z, y)\bar{x}$. The order of execution of these two instructions is immaterial.

yields the function $\overline{xy}$ in location b_1. All combinational functions can be computed using only conditional bubble transfers without special encodings.

III. TIME AND SPACE CONSIDERATIONS

In the preceding section, we showed how any combinational function can be computed using magnetic bubble interactions, assuming that interactions are possible between any pair of bubble locations. Since interactions can occur only between bubbles in physically adjacent locations, this implies that bubbles that are required to interact are brought into adjacent locations prior to the application of the instruction. The time taken for computing a function depends not only on the number of instructions in the program but also on the layout of the bubble locations. It is also necessary to be able to move a bubble from any arbitrary location to any other location without affecting the bubbles in other locations. We shall examine these problems in this section.

3.1 *Bounds on Memory Requirements*

Consider the layout shown in Fig. 1. If the locations in the shaded area are used as bubble locations for a program, the locations in the unshaded part can be used as transit points. These memory locations are initialized so as not to contain bubbles. Denoting the shaded and unshaded regions by S and T respectively, an instruction of the form (a, b) $a, b \varepsilon S$ will be replaced by a sequence of instructions

$$(a, x_i)(x_i , x_{i+1}) \cdots (x_{j-1} , x_j)(x_j , b)(x_j , x_{j-1}) \cdots (x_{i+1} , x_i)(x_i , a)$$

where

$$x_i , x_{i+1} , \cdots x_j$$

are adjacent locations in T so as not to affect any other location in S. With this layout $3n/2 - 1$ locations are required for n locations which are useful for computation. (We shall refer to these as data points.) If p is the total number of locations, the maximum number of data points can be shown to approach $2p/3$ asymptotically (K. C. Knowlton, private communication).

The maximum path length (one-way) between two data points is $n/2$. The maximum path length can be greatly reduced at the expense of a slight increase in the required memory (i.e., total number of locations) by the layout shown in Fig. 2a. The total number of locations

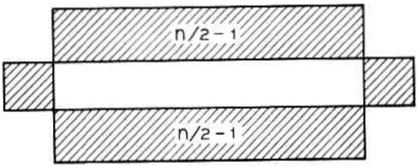


Fig. 1—Memory layout.

is $(n/2k - 3) \cdot 3k + 12k = 3n/2 + 3k$, and the number of data points is $2 \cdot 3k + (k - 1) \cdot (n/k - 6) + 2(n/2k - 3) = n$. The maximum path length (one-way) is $n/2k + 3k - 1$. [By changing the corners as shown in Fig. 2b, the maximum one-way path length can be reduced by 2, and the total number of locations can be reduced by 4 (M. D. McIlroy, private communication).] The maximum path length is dependent on k. Since k must be an integer and $2k$ must divide n, the maximum path length is minimal when $k = \sqrt{n/6}$ if this is an integer. Thus if $n = 6k^2$ the maximum path length for the layout of Fig. 2a is $(\sqrt{6}\,\sqrt{n} - 1)$. Some typical values of n and the minimum maximal path length for various values of k are shown below:

k	n	path length
2	24	11
3	54	17
4	96	23
5	150	29

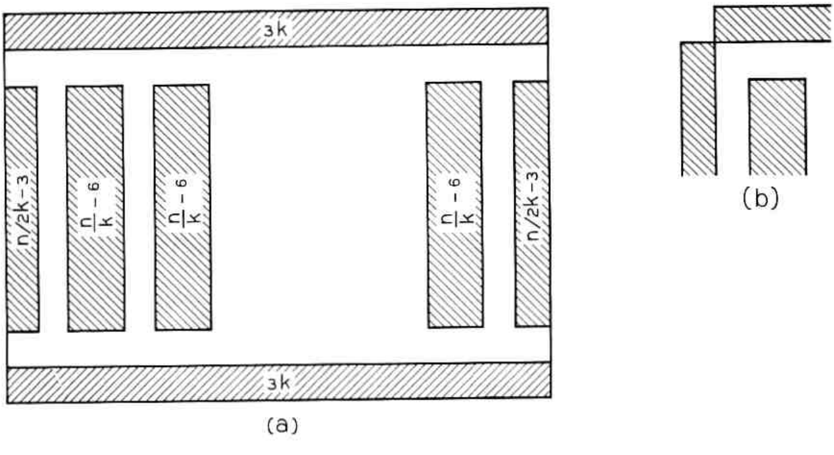


Fig. 2—Memory layout to reduce maximum path length.

In order to obtain a lower bound on the maximum path length for a memory with n locations, allowing interactions between any pair of locations, consider the closest packing of these locations in two-dimensional space, which is a circle. If the radius of the circle is r, $\pi r^2 = n$ or $r = \sqrt{n/\pi}$. The maximum distance between two points is $2r = \sqrt{4n/\pi}$. Clearly, this bound cannot be attained because all points in the circle are treated as data points in deriving the bound. Thus the best we can hope to do is to obtain a maximum path length of $C\sqrt{n}$ where $C > \sqrt{4/\pi}$. The coding of Fig. 2 has maximum path length $< \sqrt{6}\sqrt{n}$ and hence requires approximately twice as much time as the lower bound.

The memory layouts discussed above allow interactions between all pairs of data points and are therefore suitable for realizing any arbitrary program. For realizing any specific function, interactions would be necessary only between a subset of these pairs of data points and consequently a smaller memory would suffice. The maximum path length can also be reduced in general. In this case, the memory requirements and maximum path lengths derived above serve only as upper bounds.

3.2 *Number of Instructions for Realizing a Function*

The number of instructions and bubble locations required for realizing any given function using any given type(s) of instructions may vary over a wide range depending upon the realization, as shown in the following examples.

Example 1: Consider the computation of $f = x \oplus y$, using bubble transfer instructions and bubble splitting, if necessary. Let two locations x_0 and x_1 be used to represent each variable x, as discussed earlier. Using the canonical realization in the sum of products form discussed earlier, four minterms (two each for the function and its complement) have to be computed. Each minterm is computed using four bubble splitting instructions and one bubble transfer instruction. Finally, two bubble transfer instructions are required to compute the sum of two minterms. Thus the total number of instructions in the program will be $5 \times 4 + 2 \times 2 = 24$. The total number of data points required is 10 (4 for x and y, 2 for f, and 4 temporary storage). Using the layout of Fig. 1, 14 bubble locations are required in the memory.

The following program containing only six instructions also computes this function, if f_0 and f_1 are both initially empty:

$$P = (x_1, y_1)\ (x_0, y_0)\ (x_1, f_0)\ (x_0, f_0)\ (y_1, y_0)\ (y_1, f_1).$$

Six data points are necessary, and these can be obtained using eight locations and the layout of Fig. 1. Since interaction between all pairs of data points is not required by the program, it can be shown that seven locations are necessary and sufficient to permit the required interactions. However, the following program which also contains six instructions and requires six data points can be realized with only six bubble locations:

$$P' = (x_1\,,\,y_1)(x_0\,,\,y_0)(x_1\,,\,x_0)(x_0\,,\,f_0)(y_0\,,\,y_1)(y_0\,,\,f_1).$$

Figure 3 shows the memory layout for this program.

This example shows that the canonical realization may be inefficient in the number of instructions, computation time, and also memory requirements. The inefficiency of the canonical realization becomes even more apparent if we compare the canonical realization of $f = x_1 \oplus x_2 \oplus \cdots \oplus x_n$ with the realization in the form $[(x_1 \oplus x_2) \oplus x_3 \cdots \oplus x_n]$, where each exclusive-OR operation is performed by a program of the type given above.

The above example points out several open problems associated with the design of efficient programs for computing combinational functions. One problem is that of obtaining a program with the smallest number of instructions for computing a given function. For a given program, the assignment of locations so as to minimize the total number of locations required for its implementation is also an important problem. Since the number of memory locations required is dependent not only on the number of instructions, but also on the specific program, another open problem is that of obtaining a program with minimum memory requirements.

The problem of minimizing the number of instructions in a program for computing a combinational function corresponds to that of finding a realization of the function using the smallest number of modules of given types. For example, the bubble transfer instruction can be represented by the module shown in Fig. 4a. The bubble location at which each output appears is given in parentheses. Note that fan-out is not

x_1	x_0	f_0
y_1	y_0	f_1

Fig. 3—Memory layout for program P'.

allowed in the realization. Bubble splitting is represented by the module of Fig. 4b. Instructions of the type $(x, y)_d$ and the conditional transfer $(x, y)_z$ can be represented by modules of Fig. 4c and d respectively. Creation and annihilation of bubbles can be represented by appropriate constant inputs (0 or 1). If a realization of the function can be obtained using a subset of these module types, then the realization can be readily translated into a program which computes the function. All bubble locations used in the program will correspond to inputs of the circuit. Although efficient realizations of functions using only modules representing the allowed bubble interactions lead directly to efficient bubble programs, there are at present no known algorithms for the former. The following example shows how a program is obtained from a modular realization of a function.

Example 2: Consider the function $f = a\bar{c} + bcd + \bar{b}\bar{c}\bar{d}$, which is to be computed using the bubble transfer instruction only. Let each variable be represented by two locations. Thus, we wish to realize f and $\bar{f}$ using only modules of Fig. 4a, with no fan-out. The inputs to the circuit consist of the variables and their complements. $\bar{f} = c\bar{d} + \bar{b}c + \bar{a}\bar{c}d + \bar{a}b\bar{c} = c(\bar{b} + \bar{d}) + \bar{a}\bar{c}(b + d)$. We obtain the circuit of Fig. 5 (by trial) which realizes both f and $\bar{f}$. Translating each module of Fig. 5 to the corresponding instruction, we obtain the following program:

$$P = (b_0\,, d_0)\ (d_0\,, c_1)\ (b_1\,, d_1)\ (a_0\,, c_0)\ (d_1\,, a_0)\ (d_0\,, d_1)$$
$$(c_1\,, b_1)\ (b_0\,, a_0)\ (c_0\,, a_1)\ (b_0\,, c_0)\ (c_1\,, c_0).$$

At the end of the program f and $\bar{f}$ will be in c_0 and d_1 respectively. In translating the circuit to the program, no module should be translated to the corresponding instruction until the modules connected

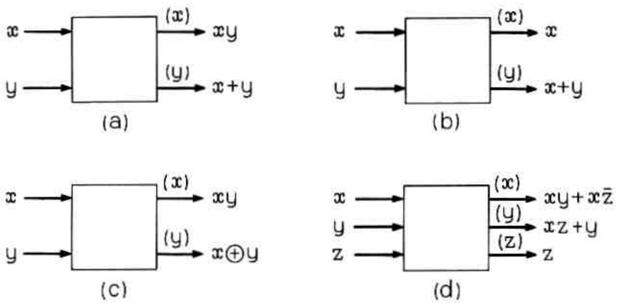


Fig. 4—Modules representing instructions: (a) Bubble transfer, (x, y). (b) Bubble splitting, $(x, y)_S$. (c) Type $(x, y)_d$. (d) Conditional transfer $(x, y)_z$.

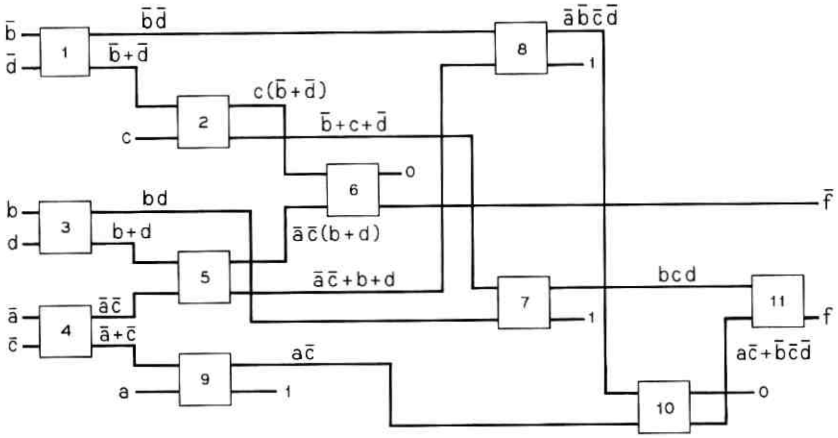


Fig. 5—Circuit for example 2.

to its input terminals have been translated. It is also important to observe the proper correspondence between bubble locations and the output terminals of a module. After applying an instruction (x, y), xy appears at location x, and $x + y$ appears at location y.

The other types of modules shown in Fig. 4 can be used in a similar manner for deriving programs using other types of instructions. However, there are no known algorithms for realizing any arbitrary function using only modules of a given type, so as to minimize the number of modules used. When such algorithms become available they can also be used for obtaining efficient programs of magnetic bubble interactions.

3.3 *Speed-up of Computations*

We have seen how a combinational function can be computed by different programs having different numbers of instructions and memory requirements. If the instructions in a program are executed one at a time, as was implicitly assumed so far, then the time for computing a function depends on the number of instructions in the program. However, it may be possible to speed up the computation by executing several instructions simultaneously.

An obvious method of obtaining parellelism in a program is by executing independent instructions simultaneously. Instructions that may be executed simultaneously can be determined from the circuit corresponding to the program, discussed in the preceding section. The modules in the circuit are arranged according to levels as follows: Modules whose inputs are only input variables (or their complements) are assigned to level 1. The level of any module is defined to be the smallest integer greater than the levels of the modules connected to its input terminals. With levels assigned to all modules in the circuit in this manner, the instructions corresponding to these modules are executed in the order of the level numbers, and the instructions corresponding to all modules in one level may be executed simultaneously. However this may necessitate a further restriction on a valid memory allocation scheme.

Example 3: Consider the realization of the function $f = a\bar{c} + bcd + \bar{b}\bar{c}\bar{d}$ in Fig. 5. The modules can be arranged according to levels as follows: level 1–(1, 3, 4); level 2–(2, 5, 9); level 3–(6, 7, 8); level 4–(10); level 5–(11). The program P can be executed in five steps as follows:

(*i*) (b_0, d_0) (b_1, d_1) (a_0, c_0).

(*ii*) (d_0, c_1) (d_1, a_0) (c_0, a_1).

(*iii*) (d_0, d_1) (c_1, b_1) (b_0, a_0).

(*iv*) (b_0, c_0).

(*v*) (c_1, c_0).

Another situation where simultaneous execution is possible becomes obvious by considering the instructions $(a, b)(a, c)$. If these instructions are executed simultaneously, the location a will contain abc, whereas the contents of locations b and c are indeterminate because they depend on which of the two instructions is actually executed first. However, if we are interested only in the product term abc, the two instructions may be executed simultaneously. In general, all $(n - 1)$ instructions for forming a product of n variables can be executed simultaneously. Similarly, the simultaneous execution of the instructions (a, s) (b, s) $\cdots$, leaves the sum $a + b + \cdots + s$ in location s with the contents of $a, b, \cdots$ indeterminate. Using this technique, the program of Example 3 can be executed in four steps by replacing the instructions of steps *iv* and *v* by $(b_0, c_0)(c_1, c_0)$, executed simultaneously. Instructions of the type $(x, y)_d$ can also be executed simultaneously with similar results. For instance, the simultaneous execution of $(a, b)_d$ and $(a, c)_d$ leaves the product abc in location a, with the contents of locations b and c indeterminate. Similarly, if $(a, c)_d$ and $(b, c)_d$ are executed simultaneously, the location c will contain $a \oplus b \oplus c$, but the contents of locations a and b will be indeterminate.

It may also be possible to speed up computations in a manner similar to the use of completion signals in asynchronous circuits.[5–7] A combinational function f can be expressed as a sum of disjoint minterms. The function can be computed by computing these minterms, one at a time. As soon as a 1 is generated, the computation is, in effect, complete because the value of the function is known to be 1. The computation can be terminated if the presence of a bubble in the output location can be detected by the control circuitry. The conditional transfer instruction $(x, y)z$ discussed earlier may be useful for this purpose.

3.4 *Iterative Array Realizations*

Our discussions so far have been restricted to realizations which minimize the number of instructions or the computation time. In this section, we shall consider some techniques for obtaining regular structures, which are capable of performing computations by the application of uniform magnetic fields. In such a structure, the same type of instruction is executed simultaneously in all parts by the application of a uniform magnetic field, the type of instruction being controlled possibly by the direction of the magnetic field.

Since regularity of interconnections is a highly desirable feature in integrated circuits also, a great deal of effort has been directed toward the realization of functions as iterative arrays.[8,9] Such arrays may be classified as uniform cell function arrays (in which the function realized by each cell is identical and different functions are realized by the array by changing the inputs to some cells) and nonuniform cell function arrays in which different cells realize different functions, depending on the position of the cell in the array. The operation of uniform cell function arrays can be simulated by magnetic bubble interactions with uniform magnetic fields.

The rectangular array shown in Fig. 6a may be used for realizing any combinational function of n variables by merely changing the connections to the last row of cells. A typical cell in the array is shown in Fig. 6b. It can be shown that the vertical outputs of the n^{th} row consist of the 2^n minterms of n variables. Since at most one of the minterms will be 1 at any time, the last row will compute the sum (OR) of all minterms connected to it. Any function can be realized by connecting to the cells in the last row only those vertical outputs of the n^{th} row that correspond to 1-points of the function.

Note that the number of cells in the array is $(n + 1)2^n$. The longest

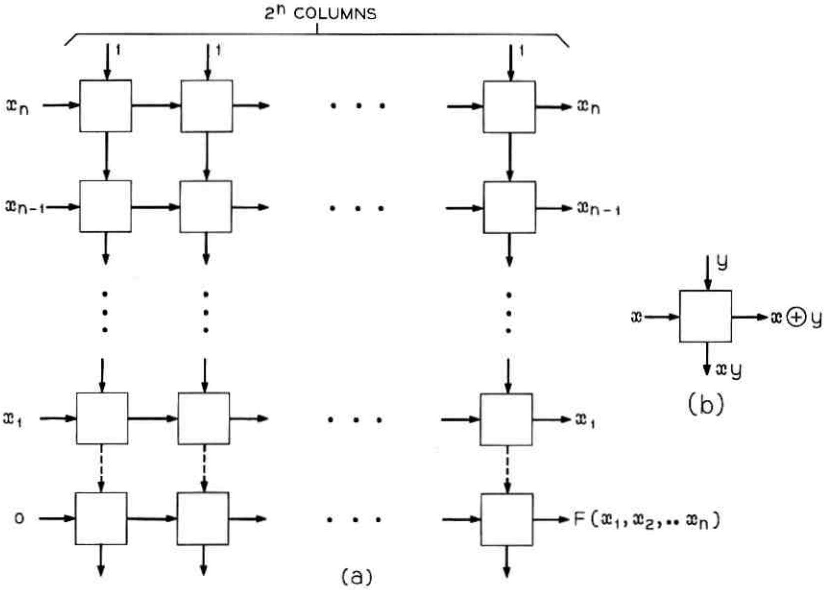


Fig. 6—(a) Rectangular array. (b) Typical cell in array.

path in the array is $2^n + n + 1$. Therefore, the time required for computing a function of n variables is proportional to $2^n + n + 1$.

The operation of this array can be simulated by magnetic bubble interactions as follows: Let each cell in the array be replaced by two bubble locations as shown in Fig. 7. Let the application of a magnetic field to a cell cause the execution of the instruction $(x, y)_d$ and transfer of the resultant bubbles as shown by the arrows in Fig. 7. (It may be necessary to apply a magnetic field in one direction followed by a field in a different direction to accomplish this.) Let the cells in the ith row and jth column be denoted by C_{ij}. If the left and right locations in C_{ij} represent x and y respectively, application of the field to C_{ij} results in xy in the right location of $C_{i+1,j}$ and $x \oplus y$ in the left location of $C_{i,j+1}$. Initially, we set the left locations of all cells in the first column, except the cell in the last row, to correspond to the values of the n variables. The left location of the first cell in the last row should be empty. The right locations of all cells in the first row initially contain bubbles. Paths from the nth row to the $(n + 1)$st row which correspond to minterms for which the function is 0 are inhibited. At $t = 1$, the field is applied to C_{11}. At $t = 2$ the field is applied to C_{12} and C_{21};

at $t = 3$ to C_{13}, C_{22}, and C_{31}; and so on. That is, at $t = k$, $1 \leqq k \leqq 2^n + n$, the field is applied to all cells C_{ij} such that $i + j = k + 1$. At $t = 2^n + n + 1$, the value of the function will be in the right location of $C_{n+1,2^n}$.

Instead of applying fields to successive diagonals, the fields may be applied to the entire platelet. Now, the right locations of all cells in the top row should be connected to "bubble generators," so that they always contain bubbles. The values of the n variables, contained in the left locations in the first column, may be changed every n units of time. The output corresponding to any input combination will appear at the output location $2^n + n + 1$ units of time after the application of the input combination. If input changes occur n units of time apart, the correct outputs will also appear n units of time apart. However, the contents of the output location will not be correct between these fixed instants of time. Therefore, it is necessary to read the output at the appropriate instants of time.

The preceding discussion shows how a uniform cell function iterative array can be simulated by magnetic bubbles by applying identical sequences of instructions to successive sets of bubble locations. This provides us with a systematic way of realizing all combinational functions. If conditional operations are permitted, nonuniform cell function arrays can also be simulated. Such arrays can be made smaller than uniform cell function arrays. The number of cells required for realizing any arbitrary function of n variables is proportional to 2^n with nonuniform arrays compared to $n \cdot 2^n$ required for uniform cell arrays.[8] However, the cells of the nonuniform array are likely to be more complex than those in a uniform array.

It is interesting to compare the size of array realizations to the size of realizations without any restrictions on the interconnection structure. Bounds on the latter will provide us with an indication of the space and time requirements for computing an arbitrary combina-

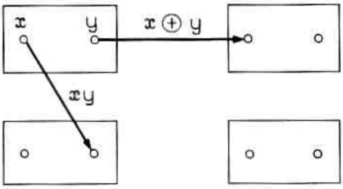


Fig. 7—Simulation of array by magnetic bubble interactions.

tional function with magnetic bubbles. D. E. Muller[10] has shown that for any set of elements capable of realizing all combinational functions, the number of elements N for realizing any arbitrary function of n variables satisfies the inequality

$$\frac{C_1 2^n}{n} \leqq N \leqq \frac{C_2 2^n}{n},$$

where C_1 and C_2 are constants whose values depend on the set of elements used. The proof of the upper bound is constructive and utilizes only one type of cell. It is outlined below because of its possible applicability to magnetic bubble computations.

The cell used realizes a function of three inputs $ab + \bar{a}c$. Any function of k variables can be expressed as

$$f(x_1, x_2, \cdots, x_k) = x_k f(x_1, x_2, \cdots, x_{k-1}, 1) + \bar{x}_k f(x_1, x_2, \cdots, x_{k-1}, 0).$$

If all functions of $k - 1$ variables are available, every function of k variables can be realized using exactly one cell. There are 2^{2^k} functions of k variables (which also include all functions of $k - 1$ variables, $k - 2$ variables, etc). Thus 2^{2^k} cells are sufficient for realizing all functions of k variables.

By expanding about any variable, the output function can be formed from two functions of $n - 1$ variables. Repeating the procedure, it may be formed from 2^{n-k} functions of k variables, using $2^{n-k} - 1$ elements. Since 2^{2^k} elements are sufficient for realizing all functions of k variables,

$$N = 2^{2^k} + 2^{n-k} - 1.$$

The value of k may now be chosen so as to minimize N. An approximate minimum is $N < C_2(2^n/n)$, for some constant C_2. Thus the bound on the number of cells required is greatly reduced for nonregular structures.

IV. SUMMARY

Three different mathematical models of magnetic bubble interactions were studied and each of them was shown to be sufficient for computing all combinational functions. Some methods of speeding up computations were presented. Geometrical patterns in which it is possible to move bubbles between any pair of data locations and also minimize the maximum path length were examined. A uniform structure for computing arbitrary functions was also presented. These structures have the advantage that computations can be carried out by the application of uniform magnetic fields to the entire platelet.

REFERENCES

1. Bobeck, A. H., "Properties and Device Applications of Magnetic Domains in Orthoferrites," B.S.T.J., *46,* No. 8 (October 1967), pp. 1901–1925.
2. Bobeck, A. H., Fischer, R. F., Perneski, A. J., Remeika, J. P., and Van Uitert, L. G., "Applications of Orthoferrites to Domain-Wall Devices," IEEE Trans. Magnetics, *MAG-5,* No. 3 (September 1969), pp. 544–553.
3. Perneski, A. J., "Propagation of Cylindrical Magnetic Domains in Orthoferrites," IEEE Trans. Magnetics, *MAG-5,* No. 3 (September 1969), pp. 554–557.
4. Graham, R. L., "A Mathematical Study of a Model of Magnetic Domain Interactions," B.S.T.J., *49,* No. 8 (October 1970), pp. 1627–44.
5. Armstrong, D. B., Friedman, A. D., and Menon, P. R., "Design of Asynchronous Circuits Assuming Unbounded Gate Delays," IEEE Trans. Computers, *C-18,* No. 12 (December 1969), pp. 1110–1120.
6. Muller, D. E., "Asynchronous Logics and Application to Information Processing," Proc. Symposium on Application of Switching Theory in Space Technology, Stanford University Press (1963), pp. 289–297.
7. Miller, R. E., *Switching Theory,* Vol. 2, New York: John Wiley and Sons, 1965, Chap. 9, 10.
8. Minnick, R. C., "A Survey of Microcellular Research," J. ACM, *14,* No. 2 (April 1967), pp. 203–241.
9. Arnold, T. F., Tan, C. J., and Newborn, M. M., "Iteratively Realized Sequential Circuits," IEEE Trans. Computers, *C-19,* No. 1 (January 1970), pp. 54–66.
10. Muller, D. E., "Complexity of Electronic Switching Circuits," IRE Trans. Electronic Computers, *EC-5,* No. 1 (March 1956), pp. 15–19.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer

By R. J. STRAIN and N. L. SCHRYER

(Manuscript received January 25, 1971)

In those charge-coupled devices (CCD's) which have regions under each electrode which are substantially free of externally applied tangential electric fields, charge motion takes place by space charge assisted diffusion, and this relatively slow process represents a limitation to the operating frequency of CCD's with large plates. In this paper, subject to certain approximations, the equations of motion for CCD charge transfer have been derived, yielding a nonlinear diffusion equation. The solution of this equation by a stable finite difference scheme is described, and the solutions are applied to predicting the operating characteristics of CCD's. The results in synopsis are:

- *(i) The n-channel devices will have lower losses at a given frequency than the p-channel devices, and a higher upper frequency limit.*
- *(ii) Higher amplitude signals (more charge) yield lower losses.*
- *(iii) Losses can be reduced an order of magnitude by using* ZERO*'s which carry a substantial amount of charge. For example, with 2-MHz clocking of a 25-μm plate (pad), p-channel 2-phase device, a 2-volt* ONE *is diminished by 4 percent per transfer with empty* ZERO*'s, but with 1-volt* ZERO*'s, the diminution is 0.26 percent per transfer.*
- *(iv) Reasonably efficient CCD operation should be possible up to the 50-MHz range using contemporary design tolerances.*
- *(v) Diffusion is important for reaching high transfer efficiencies.*

The frequency limitations described in this paper can be overcome by using a structure in which the distance between the electrodes and transferring charge is comparable to the electrode width and spacing.

I. INTRODUCTION

The charge-coupled device, as conceived by Boyle and Smith[1] and realized by Amelio, Tompsett, and Smith,[2,3] consists of a series of metal oxide semiconductor (MOS) capacitors that are driven by clock pulses into deep depletion. In this condition, the MOS structures are capable of holding minority carriers in the potential wells beneath the plates (pads). In order to establish a transfer of charge from one plate to another, it is necessary to make the potential well beneath the second plate deeper than that beneath the first. This is illustrated in Fig. 1. In the absence of charge, the potential configuration beneath any one plate would be essentially flat, and finite electric fields would exist only in that span between two CCD electrodes. When charge is present, its transfer is driven predominately by the electrostatic forces associated with the presence of the charge in the CCD and by the thermal forces responsible for diffusion. It is the purpose of this paper to describe the transport of charge under the influence of these forces and, using this description, to make predictions concerning the operation of CCD's. In order to do this, it will be necessary first to derive the equation governing the transport of charge and solve it. Then the solutions will be applied to some particular CCD situations.

II. DERIVATION OF TRANSPORT EQUATION

The derivation of the transport equation will utilize a number of approximations. The first of these is a linear approximation of

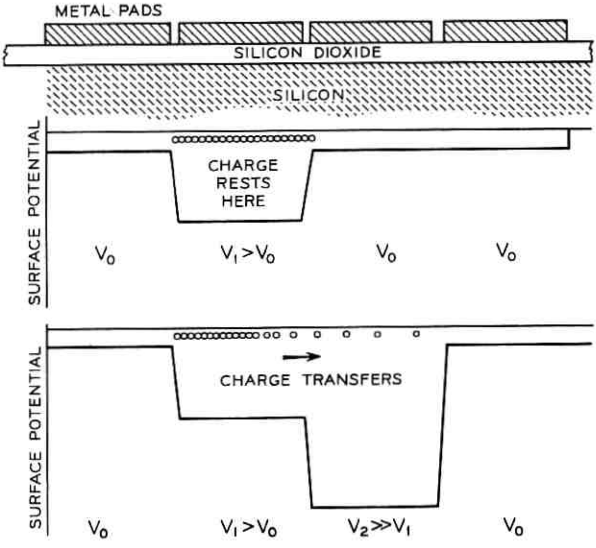


Fig. 1—A diagram of a CCD cross section and surface potential diagrams illustrating charge storage and charge transfer in CCD's.

the surface potential versus applied voltage relationship. The one-dimensional relationship between surface potential φ and applied voltage V_A is described by this equation:

$$\begin{aligned} \varphi &= V' + V_0 - \sqrt{V_0^2 + 2V_0V'} \\ V' &= V_A - V_{FB} = \frac{q}{C_{ox}} \\ V_0 &= \delta^2 e N_D \frac{\epsilon_s}{\epsilon_{ox}} \\ C_{ox} &= \frac{\epsilon_{ox}}{\delta}. \end{aligned} \tag{1}$$

In this equation V_{FB} is the MOS flatband voltage, and q is the surface density of mobile charge stored at the interface between the silicon and silicon dioxide. The constant V_0 depends upon the oxide thickness δ, the doping density N_D, the dielectric constant ε_{ox} of the oxide, and the dielectric constant ε_s of the silicon. In the subseqeunt calculations, however, an approximation to this relationship will be used which is linear in V_A and q. This approximation is

$$\varphi = \varphi_0 + \gamma\left(V_A - V_{FB} + \frac{q}{C_{ox}}\right). \tag{2}$$

In this case γ is a number, typically between 0.7 and 1, which matches equation (2) to equation (1) over the range of anticipated operation. Such a match is shown in the example of Fig. 2, which shows the surface potential versus applied voltage reduced by the flatband voltage and q/C_{ox}, both in the precise form and in the linear approximation. It may be seen in this case that the linear approximation is rather good over the range from approximately 4 to 15 volts. If the electric field parallel to the Si-SiO$_2$ interface is calculated from the approximate equation, it is found to depend upon the derivative of q. There is, however, another term to the tangential electric field; this term arises from the electrostatic repulsion of the carriers. In a two-dimensional approximation, the tangential electric field at x will have a component due to the summation of the fields of line charges located at other points x' along the surface. Because there is a metallic electrode on the surface of the SiO$_2$, each elemental line charge has an image, and the result is a shielding of the tangential field. Taking account of the two different dielectrics and the image charge, the repulsion field E_R can be written as this integral:

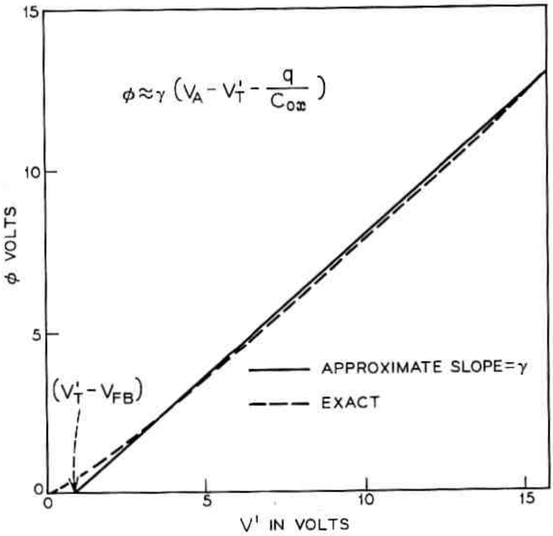


Fig. 2—The exact variation of deep depletion surface potential in an MOS with 1000Å of SiO_2 over a substrate with 5×10^{14} ionized impurities per cm^3, and the linear approximation $\Phi \propto \gamma V'$.

$$E_R(x) = \pi(\epsilon_{ox} + \epsilon_s) \int q(x')\left\{\frac{1}{x - x'} - \frac{x - x'}{(x - x')^2 + 4\delta^2}\right\} dx'. \tag{3}$$

Since the principal contribution to this integral comes from the small region within a few δ of x, it is appropriate to expand $q(x')$ as a Taylor series about x. If this is done, and the limits of integration are extended over all x', all even terms in the expansion vanish by symmetry, and to the extent that $\partial^3 q/\partial x^3$ and higher derivatives are negligible, the repulsion field becomes a local function of $\partial q/\partial x$.

$$E_R(x) = -\frac{2\delta}{\epsilon_{ox} + \epsilon_s}\frac{\partial \delta}{\partial x}. \tag{4}$$

If one takes the gradient of equation (2) and adds the repulsion field in equation (4), the total electric field acting on the charge under the CCD plate is given by

$$\begin{aligned} E_x &= -S\frac{\partial q}{\partial x} \\ S &= \left(\frac{\gamma\delta}{\epsilon_{ox}} + \frac{2\delta}{\epsilon_s + \epsilon_{ox}}\right). \end{aligned} \tag{5}$$

Typical values of both γ and the net elastance S are presented in Figs. 3 and 4 as functions of the oxide thickness δ. Ordinarily S is approximately equivalent to the reciprocal of the oxide capaçitance; consequently, Fig. 4 has been plotted in such a way that the value of S is compared with the oxide capacitance.

The transport equation for CCD's will be based on the divergence equation for current

$$\frac{\partial q}{\partial t} = -\nabla \cdot J, \tag{6}$$

using the relationship

$$J = q\mu E - D\frac{\partial q}{\partial x} \tag{7}$$

for the value of the current density J. In this equation D is the diffusion coefficient appropriate to the surface, and μ is the surface mobility. Combining equations (5), (6), and (7) gives the basic equa-

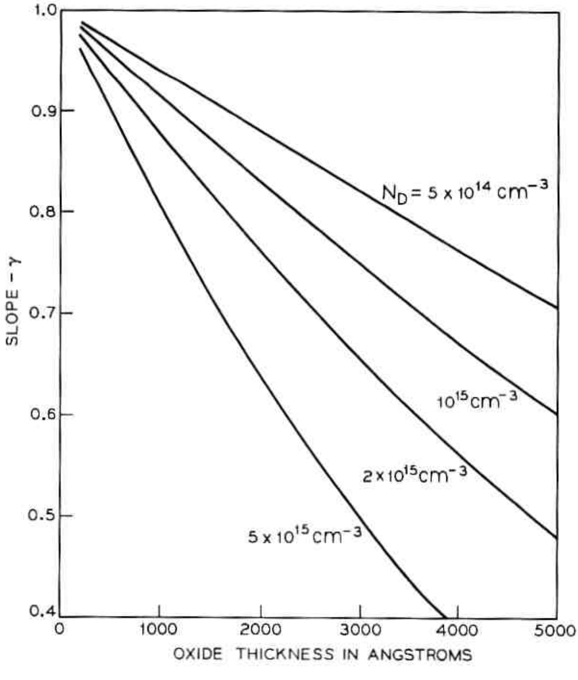


Fig. 3—The proportionality constant γ as a function of oxide thickness, using doping density as a parameter.

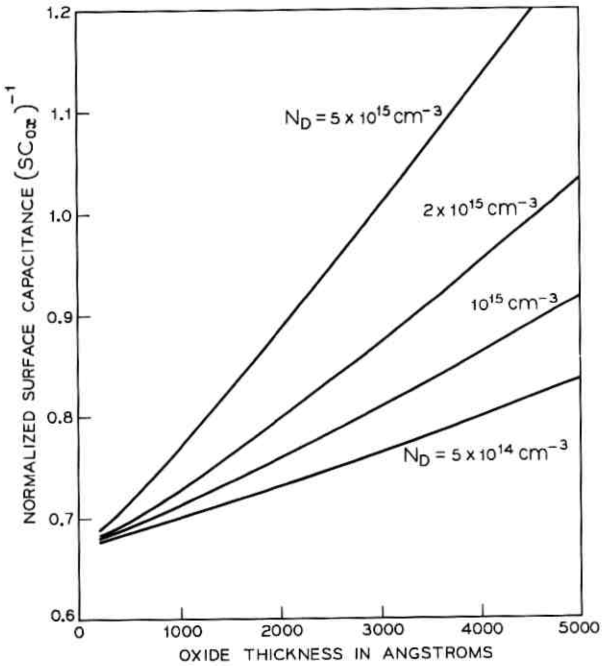


Fig. 4—The effective surface capacitance, $1/S$ (the reciprocal of elastance S) normalized by C_{ox} and presented as a function of oxide thickness.

tion for charge transport in a CCD:

$$\frac{\partial q}{\partial t} = \mu S\left(\frac{\partial q}{\partial x}\right)^2 + q\mu S\,\frac{\partial^2 q}{\partial x^2} + D\,\frac{\partial^2 q}{\partial x^2}. \tag{8}$$

This may be simplified somewhat by noting the Einstein relationship between diffusion and mobility.

$$D = \frac{\kappa T}{e}\,\mu. \tag{9}$$

Here κT and e all assume their traditional values as Boltzmann's constant, absolute temperature, and electronic charge. Using this substitution, one reaches this equation:

$$\frac{\partial q}{\partial t} = \mu S\left(\frac{\partial q}{\partial x}\right)^2 + \mu Sq\,\frac{\partial^2 q}{\partial x^2} + \frac{\kappa T}{e}\,\mu\,\frac{\partial^2 q}{\partial x^2}. \tag{10}$$

The last term on the right of equation (10) is simply the diffusion term. The resemblance between this diffusion term and the other term involving the second partial of charge is quite apparent, because the product Sq, like $\kappa T/e$, is a voltage.

Without boundary conditions, the problem is only half specified. The physical situation on which the boundary conditions are based is shown in Fig. 5. In Fig. 5a, a CCD plate is shown storing charge. The potential beneath the plate and the charge density are both uniform. In Fig. 5b, another potential well has been impressed immediately adjacent to the first potential well. The second well is considerably deeper than the first; as a consequence, charge will start to flow from the first potential well into the second under the control of equation (8). No charge, however, flows in the opposite direction. Between the two potential wells there is an abrupt step in potential. This means that there is an extremely high electric field, and charge in that vicinity will move with a very high velocity. Using the preceding facts, it is possible to approximate the physical boundary conditions in the

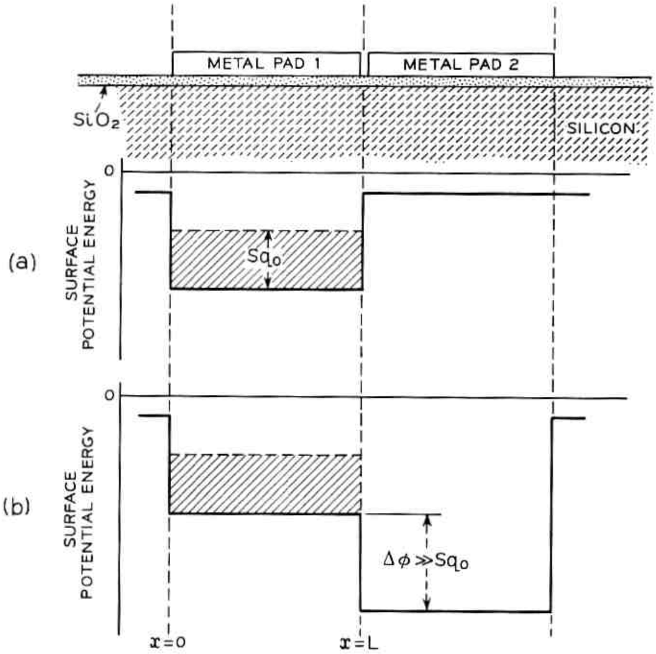


Fig. 5—An illustration of the length definition and the surface potential configuration in equilibrium, and at $t = 0$, when the transfer is initiated.

following way.

$$\frac{\partial q}{\partial x} \to 0 \quad \text{as} \quad x \to 0 \tag{11}$$

$$q = 0 \quad \text{at} \quad x = L.$$

Both of these boundary conditions apply for all time. The first expresses the fact that current is unable to flow to the left in the device, and the second presumes either an ideal sink for charge at $x = L$, or alternatively that charge moves away from that point at an infinite velocity. The initial condition is a value of q at time $t = 0$ for all points except $x = L$; it will be assumed that the charge is evenly distributed with the magnitude q_0.

Before proceeding with the solution to this equation, it will be convenient to scale the variables. The scaling can be effected by applying the following definitions:

$$\tau = \frac{t}{\dfrac{L^2}{\mu\tau_0}} = \frac{t}{\tau_0} \tag{12a}$$

$$u_0 = 1 \text{ volt}$$

$$y = \frac{x}{L} \tag{12b}$$

$$u = \frac{Sq}{u_0}. \tag{12c}$$

Time is now represented by the variable τ, which scales time against τ_0, and y is the dimensionless variable representing length. In the definitions of τ and τ_0, the voltage unit u_0 is introduced. If this equation had a natural voltage unit, u_0 would be defined as that unit. However, there is no natural voltage scale to equation (10), even though the thermal voltage $\kappa T/e$ appears in one term; consequently, it has proven convenient to define u_0 to be unity. The quantity u representing charge comes about because the product of the elastances and charge q is equivalent to a voltage. In the solutions which follow τ will typically range from 10^{-2} to 10^2, y will range from 0 to 1, and u will range from 10^{-3} to 10.

III. SOLUTION OF THE EQUATION

The scaled equation has the form

$$u_\tau = u_y^2 + (u + a)u_{yy} \tag{13}$$

where $a = \kappa T/eu_0$ is a constant. The problem may be simplified by letting $w = u + a$ to obtain

$$w_\tau = w_y^2 + w_{yy}w \qquad (14)$$
$$0 \leqq y \leqq 1$$
$$0 \leqq \tau \leqq \infty$$

subject to

$$w_y(0, \tau) = 0$$
$$w(1, \tau) = a$$
$$w(y, 0) = \begin{cases} u_i + a, & 0 \leqq y < 1 \\ a, & y = 1 \end{cases} \qquad (15)$$
$$u_i = \frac{Sq_0}{u_0}$$

where u_i is determined by the initial uniform charge density q_0.

The nonlinear parabolic initial boundary value problem given by equation (14) and equation (15) may be solved using a finite difference scheme. The scheme comes from a more general technique for solving systems of coupled nonlinear parabolic equations which results in *linear* difference equations. This more general scheme will be published at a later date.

Let h be the space mesh size and $\Delta\tau$ be the time step. Let $w_j^k = w(jh, k\Delta\tau)$. The indices j and k describe position and time respectively. The difference equation is then:

$$\frac{w_j^{k+1} - w_j^k}{\Delta\tau}$$
$$= \frac{1}{2}\left[\frac{w_{j+1}^{k+1} - w_{j-1}^{k+1}}{2h}\,\frac{w_{j+1}^k - w_{j-1}^k}{2h} + w_j^{k+1}\,\frac{w_{j+1}^k + w_{j-1}^k - 2w_j^k}{h^2}\right.$$
$$\left. + \frac{w_{j+1}^k - w_{j-1}^k}{2h}\,\frac{w_{j+1}^{k+1} - w_{j-1}^{k+1}}{2h} + w_j^k\,\frac{w_{j+1}^{k+1} + w_{j-1}^{k+1} - 2w_j^{k+1}}{h^2}\right], \qquad (16)$$

for $j = 1, \cdots, J - 1$, where $h = 1/J$. The boundary conditions reduce to $w_0^{k+1} = w_1^{k+1}$ and $w_J^{k+1} = a$.

This is a direct generalization of the Crank-Nicholson scheme for solving the heat equation, $w_t = w_{xx}$. For a description of that scheme and its properties, see Ref. 4, pp. 185–193.

The most important properties of (16) come from the time and space averaging which is the hallmark of the Crank-Nicholson scheme. The left-hand side of (16) is close to the value of $w_t(jh, (k + 1/2)\Delta t)$ and the right-hand side of (16) has been averaged to be close to

$$(w_y(jh, (k + 1/2)\Delta t))^2 + w(jh, (k + 1/2)\Delta t)w_{yy}(jh, (k + 1/2)\Delta t).$$

The averaging has been done in such a way as to make w_{j+1}^{k+1} , w_{j-1}^{k+1} and w_j^{k+1} appear linearly. This averaging gives (16) two nice properties. First, if a solution of (14) is inserted in (16), and everything is expanded in a Taylor series about $(jh, (k + 1/2)\Delta t)$, then equality of the right- and left-hand sides follows up to terms of order $0(h^2)$ and $0(\Delta t^2)$. That is, the scheme is accurate to second-order in both time and space. In practice, this means that the difference between the computed w_j^k and $w(jh, k\Delta t)$, the true solution at the mesh points, will be $0(h^2) + 0(\Delta t^2)$. Second, (16) is a linear system of equations for the w_j^{k+1}. In fact, equation (16) is equivalent to

$$\begin{aligned} w_{j+1}^{k+1}\left(-\frac{\Delta\tau}{4h^2}(w_{j+1}^k - w_{j-1}^k) - \frac{\Delta\tau}{2h^2}w_j^k\right) \\ + w_j^{k+1}\left(1 - \frac{\Delta\tau}{2h^2}(w_{j+1}^k + w_{j-1}^k - 2w_j^k) + \frac{\Delta\tau}{h^2}w_j^k\right) \\ + w_{j-1}^{k+1}\left(\frac{\Delta\tau}{4h^2}(w_{j+1}^k - w_{j-1}^k) - \frac{\Delta\tau}{2h^2}w_j^k\right) = w_j^k\ . \end{aligned} \tag{17}$$

This has the form

$$A_jw_{j+1} + B_jw_j + C_jw_{j-1} = D_j \qquad j = 1, 2, \cdots, J - 1, \tag{18}$$

with $w_J = a$ and $w_0 = w_1$, reflecting the boundary conditions. This is a tridiagonal system of equations for the w_J and may be solved quite efficiently using Gaussian elimination (see Ref. 4, pp. 198–201). The method is easily described. Let

$$w_j = E_jw_{j+1} + F_j \tag{19}$$

for $j = 0, 1, \cdots , J - 1$. Replacing w_{j-1} by $E_{j-1}w_j + F_{j-1}$ in (18) gives w_{j+1} in terms of w_j . Comparing this relation with (19) gives

$$\begin{aligned} E_j &= \frac{-A_j}{B_j + C_jE_{j-1}} \\ F_j &= \frac{D_j - C_jF_{j-1}}{B_j + C_jE_{j-1}} \end{aligned} \tag{20}$$

for $j = 0, \cdots , J - 1$. Since $w_1 = w_0 = E_0w_1 + F_0$ we see that $E_0 = 1$

and $F_0 = 0$. This completely determines $w_0, \cdots, w_J$ using (20) for $j = 1, \cdots, J - 1$, and then (19) for $j = J - 1, \cdots, 0$.

IV. RESULTS AND INTERPRETATION

In applying the results of this analysis to CCD operation, the most important operating characteristic is the amount of charge remaining behind after transfer from one plate to the next in a limited time. The remaining charge is plotted as a function of time in Fig. 6, using two different initial conditions, $Sq_0 = 10$ volts and $Sq_0 = 2$ volts. The nonlinear equation which has been solved reduces essentially to the diffusion equation after most of the initial q has been dissipated, and the solutions in this domain approach straight lines of log(charge) versus time when the average value of Sq is much less than $\kappa T/e$. However, at short time $\tau \lesssim 1$ the solution is effectively driven by the charge so the time rate of change for $Sq_0 = 10$ volts is roughly five times the time rate of change associated with $Sq_0 = 2$ volts. Between these two domains is a transition region where the average value of Sq lies between 0.2 $\kappa T/e$ and 5 $\kappa T/e$. Here the curves are very nearly

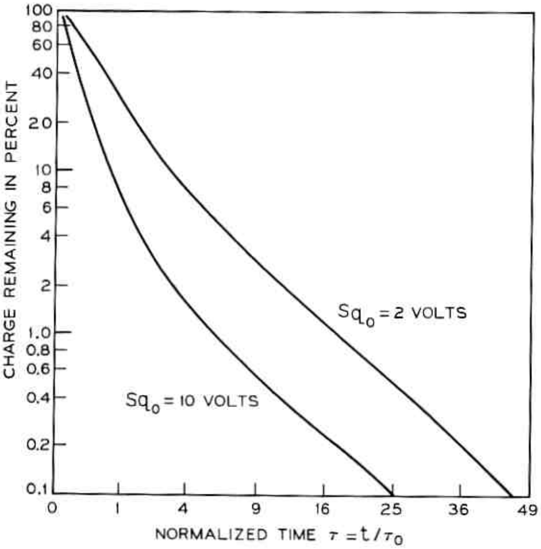


Fig. 6—The fraction of the charge remaining under the first plate as a function of the normalized time $\tau = t/\tau_0$. The abscissa is plotted as $\tau^{1/2}$, which spreads the region where $\tau < 1$, and leads to straight lines over the transition portion of the characteristic.

straight lines of log(charge) versus $\tau^{\frac{1}{2}}$. The different regimes reflect a feature mentioned when the equation was scaled; i.e., there is no single natural potential to use in forming the time scale. For very short times, the natural potential is the initial value of Sq. For times much greater than one, the natural potential is $\kappa T/e$. Figure 7 shows the transfer of charge with and without the aid of diffusion. It is clear that the last 0.1 volt of scaled charge receives a considerable boost from $\kappa T/e$.

In Fig. 6 it can be seen that in the vicinity of $\tau = 10$ the charge remaining diminishes to approximately 1 percent of its initial value, and in the vicinity of $\tau = 30$ the charge diminishes to 0.1 percent. These results are meaningless without some concept of the size of τ. Figure 8 helps to alleviate this shortcoming by showing τ_0 as a function of the plate width L for n-channel and p-channel devices. The n-channel mobility has been taken to be 750 and the p-channel mobility 150 cm²/volt-second.[5] From Fig. 8 it is possible to see that the values of τ_0 range from approximately 1 ns to 1 μs for plates ranging from 10 to 100 μm in width. To take a specific example, a 50-μm plate in a p-channel device would have a τ_0 of one-sixth of a microsecond. If operation were desired at a value of $\tau = 10$, aiming for a transfer

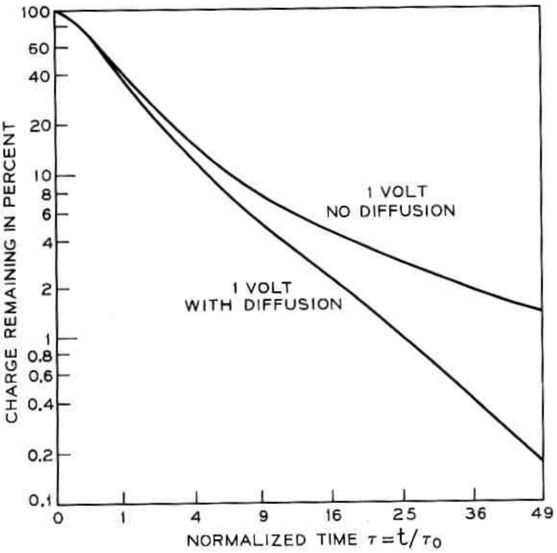


Fig. 7—The fraction of the charge remaining for initial values of $Sq_0 = 1$ volt with room temperature diffusion and no diffusion ($\kappa T/e = 0$).

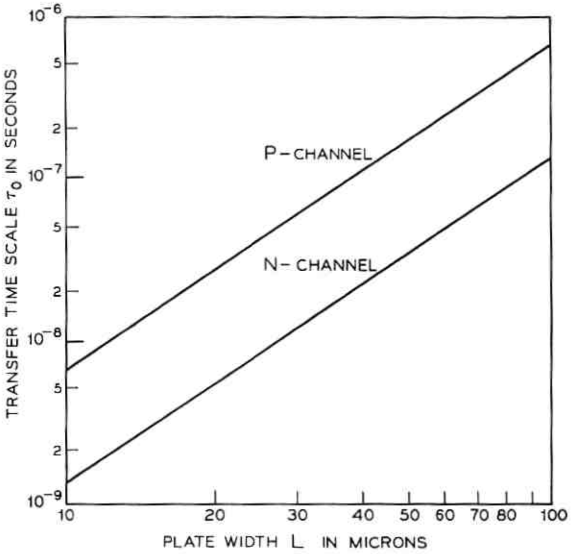


Fig. 8—A plot of the time scale factor τ_0 as a function of the CCD plate length L.

efficiency near 99 percent, then the operating frequency of a 2-phase CCD would be roughly 300 kHz, and the same loss per transfer would be achieved in a 3-phase CCD at 200 kHz. In an n-channel device with the same loss, those frequencies would increase to 1.5 and 1 MHz respectively.

In order to illustrate the effect of frequency, another example is shown in Fig. 9, where the loss per transfer is plotted as a function of clock frequency for a p-channel 2-phase CCD operating with square waves. In this type of operation, higher signal levels result in higher transfer efficiencies. The plate width in this case was chosen to be 25 μm, and at 1 MHz the loss per transfer ranges from about 0.3 percent for $Sq_0 = 10$ to 1.5 percent for $Sq_0 = 2$ volts.

The figures quoted so far must be regarded as worst-case loss figures. They apply to an isolated cluster of charge being propagated through a CCD. It will be possible to reduce this loss by approximately an order of magnitude by arranging the information propagated through the device in such a way that no information ever leads to a totally empty CCD cell. This will be effective, because for long times, i.e., $\tau > 10$, the charge remaining behind is substantially independent of the amount of charge initially put under a pad. This

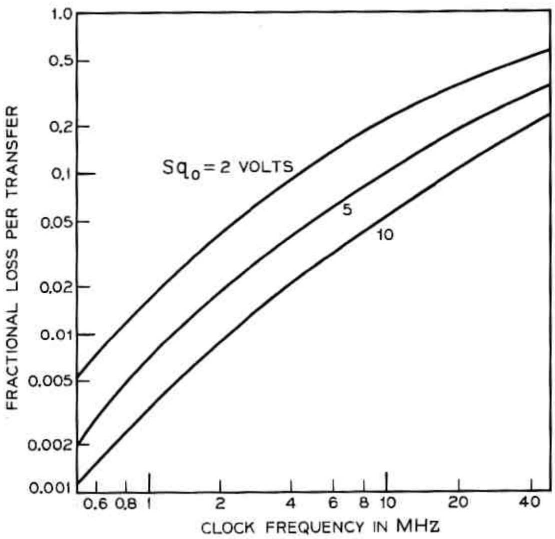


Fig. 9—The fractional loss per transfer of a P-channel 2-phase CCD, with a field-free length L of 25 μm, plotted against clock frequency with the initial ONE signal as a parameter. The ZERO's have no charge.

is illustrated by Fig. 10, which shows the absolute magnitude of remaining charge as a function of the charge which was initially placed in the device. In this figure, the amount of charge is described as its voltage equivalent Sq. The impact on device loss is illustrated in Fig. 11, where the loss per transfer for the same device shown in Fig. 8 is plotted as a function of the charge existing in a ZERO when a ONE contains 2 volts. It is presumed that the information carried is an alternating series of ONE's and ZERO's. This represents a worst-case condition when nonempty or "fat" ZERO's are used. From this figure, it is readily seen that a 2 to 1 ONE to ZERO ratio leads to practically an order of magnitude reduction in loss at a given frequency. Operation in this way requires the use of a threshold detector to ascertain whether a ONE or a ZERO was present; the success of such operation depends in large part upon the realization of a practical threshold detector.

The solution of the nonlinear diffusion equation also gives charge distributions in the CCD as a function of time. The charge distribution at selected times is shown in Fig. 12. Initially the left-hand end is almost totally undisturbed by the presence of an accepting well at the right-hand end of the CCD, but then as charge transfers out of

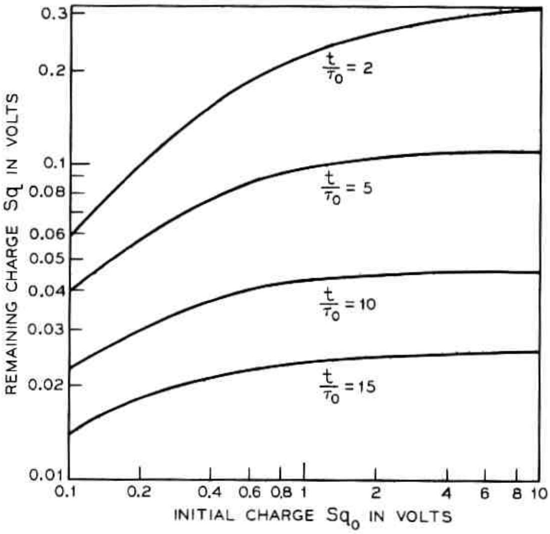


Fig. 10—The remaining charge at several times after the initiation of transfer presented as a function of the initial charge Sq_0.

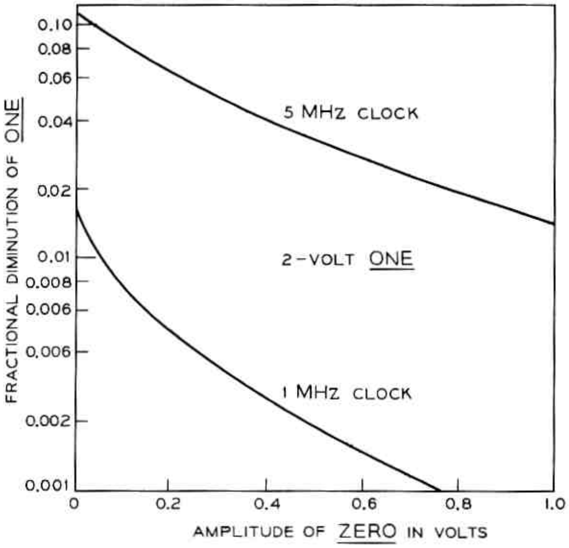


Fig. 11—The fractional diminution of a ONE as a function of the amplitude of a "fat" ZERO in an alternating series of ONE's and ZERO's. This example also refers to a 25-μm 2-phase p-channel CCD.

the device, the left end follows, but the profile which is established relatively early in the device cycle remains at least superficially unchanged through the course of the transfer.

If the potential of the source well is being changed at the same time the transfer is taking place, it is valuable to know the time development of the highest charge density in the source well, because it is necessary to prevent this point from ever reaching an injection condition. In this approximation injection occurs when $Sq \geq \gamma(V_A - V_{FB})$, and the well is no longer capable of holding the charge; the well and the substrate then form a forward biased p-n junction. For the example discussed before, the 25-μm p-channel CCD, the highest voltage in the well has been plotted as a function of time, assuming initial charges of 10 volts, 2 volts, and 1 volt; Fig. 13 shows the real time for the specified device and the general, normalized time. This figure shows that the potential on the plate must not drop to zero in a time less than 100 ns, if the injection of stored CCD charge into the bulk of the semiconductor is to be avoided.

It is unlikely that the initial charge configuration in a CCD will be perfectly square as assumed. To test the effect of this assumption

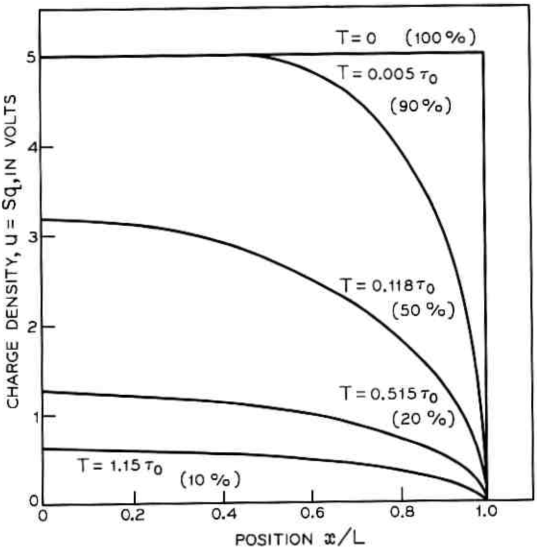


Fig. 12—The charge density as a function of position under the CCD plate, and the parameter is the time after initiation of transfer. The percentages refer to the total fraction of charge remaining under the plate.

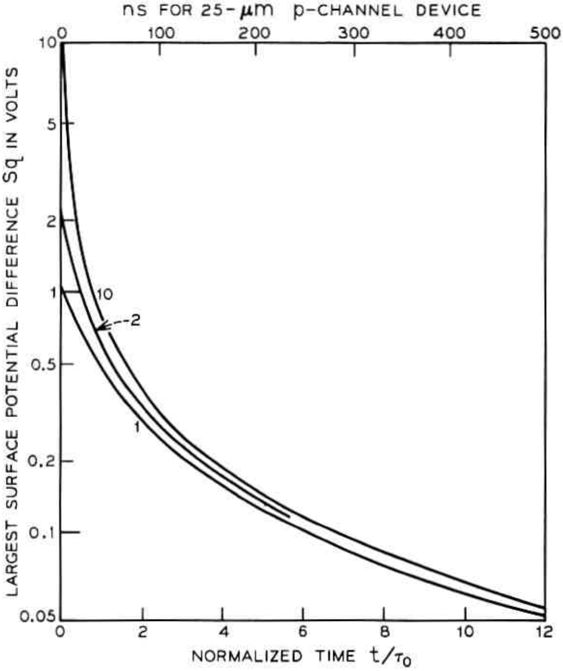


Fig. 13—The amplitude of the largest surface potential difference $Sq(0)$ as a function of time. Two time-axes have been provided, the generalized τ, and t for a 25-μm p-channel device.

on the results presented here, two initial conditions were used; one had charge distributed uniformly to an Sq value of 8 volts, and the other had charge distributed nonuniformly but having the same average charge density, that is $\overline{Sq} = 8$ volts. Figure 14 shows the difference in the charge remaining in the potential well as a function of time for these two distributions, and it is easily seen that beyond the time of $\tau = 0.25$ the two charge quantities are substantially identical.

The effect of finite carrier velocity at the transfer point was also tested. If infinite velocity cannot be achieved, the charge density at the point L cannot be 0, but it must be that finite value which will allow current to flow out of the potential well at the saturation velocity. Using this fact, it is possible to write the charge density at L as a function of the time rate of change of the total charge in the well dQ/dt.

$$u(L) = Sq(L) = \frac{1}{v_s}\frac{dQ}{dt} = \frac{L}{v_s}\frac{d\bar{u}}{dt}. \qquad (21)$$

Here v_s is a saturation velocity and, assuming its value to be 10^7 centimeters per volt-second, the solution of equation (10) was reprogramed to reflect a value $u(L)$ in accordance with equation (21). The difference caused in the result was negligible, but because of the L^2/μ time scaling, it was more significant for very small plates than for the larger plates. The time error is shown in Fig. 15 as a function of charge remaining beneath the plate.

V. DISCUSSION

We have developed an equation governing charge transfer in flat-plate CCD's and shown how it can be solved numerically by an unconditionally stable finite difference routine. These solutions have been applied to the calculation of the properties of CCD's subject to certain constraints; the most serious of these limitations is the assumption of the instantaneous creation of the transferring condition. This assumption transforms into physical reality as the use of square-

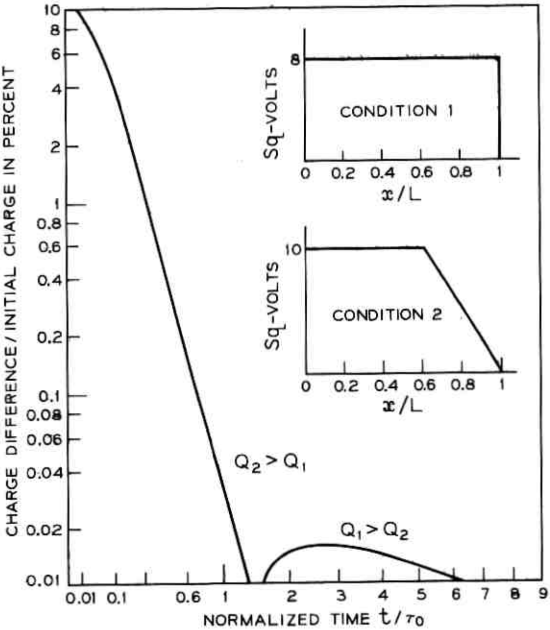


Fig. 14—Two initial conditions are compared to evaluate the impact of nonuniform distribution of a given total charge on the transfer predictions. The ordinate is the fractional difference in the total remaining charge and the abscissa is normalized time.

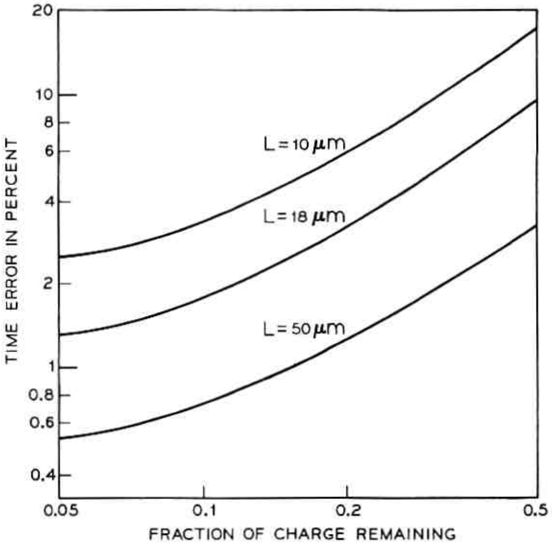


Fig. 15—An illustration of the effect of limiting the carrier velocity at $x = L$ to 10^7 cm/s, expressed as an error between the time predicted in the idealized case and the time in the velocity-limited case.

wave driving. Under these conditions, we have shown that with empty ZERO's losses of the order of 1 percent can be achieved in the vicinity of 1 MHz in p-channel CCD's, and the same order of loss at 5 MHz in n-channel CCD's. Further it has been shown that the use of "fat" ZERO's can effectively reduce this loss by as much as an order of magnitude. The problem which has not been solved here is that which arises when the accepting well is not created instantaneously but moves down with a finite time-rate of potential change. In this situation, a ONE will have a longer time in which to transfer than a ZERO will. This may tend to reduce the amount of charge left behind by the ONE as contrasted with that left behind by a ZERO. This would diminish the loss. At this stage, it is evident that losses of realizable CCD's, i.e., those with plates in the range of 15 to 25 μm, will be modest in the 5 to 10 MHz range, particularly if they are made using an n-channel technology. Using properly loaded ONE's and ZERO's, short shift registers should deliver useful signals, even clocked at 50 MHz.

This analysis has also shown that the existence of field-free regions in CCD's represents a considerable limitation on their performance. It will be possible to further lower losses by arranging substantial penetration of transverse electric fields under the plates so that the charge transfer is limited by drift rather than diffusion. In this case, the scattering limited velocity assumes an important role.

CCD's on 10 Ω-cm substrates appear, on the basis of preliminary measurements,[3,6] to behave according to the predictions of the space charge assisted diffusion theory.

VI. ACKNOWLEDGMENT

The authors are indebted to G. E. Smith and J. McKenna for their stimulating suggestions, and to J. A. Morrison for his constructive comments.

REFERENCES

1. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.
2. Amelio, G. F., Tompsett, M. F., and Smith, G. E., "Experimental Verification of the Charge Coupled Device Concept," B.S.T.J., *49*, No. 4 (April 1970), pp. 593–600.
3. Tompsett, M. F., Amelio, G. F., and Smith, G. E., Appl. Phys. Letters, *17*, 1970, p. 111.
4. Richtmeyer R. O., and Morton, K. W., *Difference Methods for Initial Value Problems*, 2nd Ed., New York: Interscience, 1970.
5. Murphy, N. St. J., Berg, F., and Flinn, I., Solid State Elect., *12*, 1969, p. 775.
6. Berglund, C. N., and Powell, R. J., private communication.

(*Added in proof.*) A related analysis with more limited scope than that above has been published by W. E. Engeler, J. J. Tiemann, and R. D. Baertsch in *Applied Physics Letters*, *17*, 1970, p. 469.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Linearized Dispersion Relation and Green's Function for Discrete-Charge-Transfer Devices with Incomplete Transfer

By W. B. JOYCE and W. J. BERTRAM

(Manuscript received February 19, 1971)

A small-signal (linearized) theory of discrete-charge-transfer-device performance is presented for the case of incomplete charge transfer. Specifically, the dispersion relation is derived which relates the charge-transfer efficiencies presently characterizing these discrete (in space and time) devices to the usual measures of device or transmission-line performance based on the attenuation, dispersion, phase velocity, etc., of sine waves. In a more general sense this emphasizes the applicability of conventional signal theory to these new devices. The impulse solution or Green's function is then shown to be the equivalent of a bivariate distribution in probability theory. More generally the utility of (deterministically interpreted) probability theory is emphasized by showing the equivalence of a general small-signal theory to a random-walk process.

I. INTRODUCTION

In an important new class of discrete-charge-transfer devices including charge-coupled devices[1] (CCD's), bucket-brigade shift registers,[2-3] and other shift-register or image-detection or display devices, externally applied time-dependent voltages step captive charge along a chain of equivalent discrete storage stations. In some of these devices the charge transfer is imperfect with a fraction of the charge failing to advance and a fraction lost altogether during each step. Explicit expressions are constructed here for the dispersion relations and Green's functions which describe this imperfect performance under conditions when the fractions of charge that go astray can be described by constant parameters characteristic of the particular device.

The theory of analog signal processing is based on the properties of sine waves, and analog devices are conventionally characterized by the dispersion, attenuation, etc., which they cause. A second standard method of characterization is based on the distortion and attenuation of pulses. In contrast, discrete-charge-transfer devices are presently characterized by their charge-transfer efficiencies (fractions). The solutions given here were chosen to facilitate the application of conventional signal processing theory to these new discrete devices. That is, they show the equivalence of these three methods of characterization in the small-signal limit and provide the appropriate interrelating formulas. Specifically, Section III treats the sinusoidal representation, while Section IV considers pulses.

Although the physical interpretation of our equations is deterministic, it is also an objective in Section IV to show the direct applicability of many established results of probability theory. In the Appendix our basic equation is re-derived as the simplest nontrivial example of a discrete small-signal theory. In turn, the small-signal theory is seen to be the deterministic limit of a random-walk process.

Whether the mode of device operation be digital or analog, one is ordinarily interested in utilizing the full information capability; i.e., maximum bandwidth or wavelengths approaching twice the station separation in shortness. Thus, we at no time approximate the discrete equation by a (continuous) differential equation.

In summary our objective is to emphasize by way of a case of present interest that standard methods of device characterization, as well as established probability theory, can be applied to discrete charge-transfer devices.

II. BASIC EQUATION

Discrete-charge-transfer devices are often, to a good approximation if not exactly, discrete in both space and time. That is, the information-bearing charge is moved in discrete bursts in time from one spatially discrete storage station (e.g., capacitor or potential well) to the next along a line of stations. Consider $q_{x,t}$ the charge in station x at time t where x and t assume only integer values; i.e., the unit of time is taken as the stepping interval, and the unit of distance is taken as the station separation (center-to-center). Perfect charge transfer implies

$$q_{x,t} = q_{x-1,t-1} \tag{1}$$

and hence unit signal speed.

Real devices are typically experimentally characterized by the fraction α of the charge which is successfully advanced per step.[4,5] If a fraction ϵ fails to advance and remains in its original station, then the process is described by

$$q_{x,t} = \alpha q_{x-1,t-1} + \epsilon q_{x,t-1} , \tag{2}$$

which reduces to equation (1) in the ideal case of $\alpha = 1$ and $\epsilon = 0$. In other words, equation (2) states that the charge in station x at time t, q_{xt} , is the successfully transferred fraction of the charge in the previous station at the previous time, $\alpha q_{x-1,\ t-1}$, plus a fraction ε of the charge $q_{x,t-1}$ in station x at the previous time which failed to advance. From equation (2) it follows that a fraction $\ell \equiv 1 - \alpha - \epsilon$ of the charge is lost per step. Theoretical expressions for α and/or ε are contained explicitly or implicitly in the work of several authors on different devices.[6-9] [In the appendix, equation (2) is generalized slightly to include an inhomogeneous term and nonconstant values of α and ϵ.] In this context equation (2) was introduced by Berglund to describe incomplete transfer in CCD and IGFET bucket-brigade shift registers, and he showed an approximate equivalence at low frequencies to an (analog) matched transmission line.[6] Traditionally equation (2) is associated with probability theory (Bernoulli trials) where $\alpha + \epsilon = 1$.[10] In the case $\alpha = \epsilon = 1$ (outside our primary interpretation) equation (2) is usually known as Pascal's triangle although he was preceded by Cardan in 1540 who, in turn, cited earlier sources.[11] We give here a number of exact and exact limiting ($\epsilon \rightarrow 0$) solutions to equation (2) useful both in image-detection and shift-register contexts. It should be noted that our results can also be applied to Berglund's model[12] of bipolar bucket-brigade shift registers which differs from equation (2) only through an interchange of the physical interpretation of x and t.

III. SINUSOIDAL REPRESENTATION

The dispersion relations can be found from the space and time Fourier transforms of equation (2) or by the separation-of-variables method. Either approach amounts to seeking a running-wave solution of the form

$$\operatorname{Im} e^{i(\omega t - kx + \varphi)} \qquad |x|, |t| = 0, 1, 2, \cdots \tag{3}$$

where k and ω may be complex to account for the attenuation, and φ is any constant phase factor. Since x takes on only integer values, equation (3) is not affected by adding multiples of 2π to the real part of k, and thus without loss of generality we require $|\operatorname{Re} k| \leqq \pi$. In the theory of lattice vibrations (where time is continuous but the physical system is discrete) this range limitation on Re k is usually expressed by saying that the wavelength with one atomic species must be at least two lattice spacings.[13] Similarly, since time is discrete, all frequencies outside the fundamental range $|\operatorname{Re} \omega| \leqq \pi$ are redundant. In the theory of sampled-data control systems (where the physical system is usually continuous but time is discrete) this is usually expressed by saying that the sampling frequency must be at least twice the maximum frequency to be detected.[14] (Here the sampling frequency f_s is once per unit time, i.e., $f_s = 1$ or $\omega_s = 2\pi$). Substituting equation (3) into equation (2) yields for the dispersion relation between ω and k

$$e^{i\omega} = \alpha e^{ik} + \epsilon. \tag{4}$$

Two important special cases of equation (4) are discussed in the rest of this section. Since equation (4) shows that the dispersion relations are independent of φ, we take $\varphi = 0$ hereafter.

In the case of image detection (or projection), k is real corresponding to a term in the spatial Fourier representation of the initial image. Equation (4) then reduces to

$$\omega(k) \equiv \omega' + i\omega'' = -\omega^*(-k) \tag{5}$$

$$= k - \tan^{-1} \frac{\sin k}{\cos k + \alpha/\epsilon}$$
$$- i(\ln \alpha + \tfrac{1}{2} \ln [1 + (\epsilon/\alpha)^2 + 2(\epsilon/\alpha) \cos k]); \tag{6}$$

i.e., the wave is attenuated in time but remains spatially sinusoidal. The identity

$$\tan^{-1} \frac{\sin \theta}{c + \cos \theta} + \tan^{-1} \frac{\sin \theta}{c^{-1} + \cos \theta} = \theta$$

was used to express the effect of nonzero ϵ as a separate correction term in equation (6). The real and imaginary parts, ω' and ω'' are plotted in Fig. 1. Although not considered here, it is sometimes useful to regard α and ϵ as k-dependent (i.e., wavelength-dependent) quantities. In the practical case of small ϵ/α, equation (6) is usefully approximated by truncating its expansion in powers of ϵ/α after the linear term, i.e.,

$$\omega(k) \to k - (\epsilon/\alpha) \sin k - i[\ln \alpha + (\epsilon/\alpha) \cos k], \qquad \epsilon/\alpha \to 0$$
$$\to k - \epsilon \sin k + i[\ell + \epsilon(1 - \cos k)], \qquad \epsilon, \ell \to 0. \tag{7}$$

A phase velocity v'_φ can be defined which describes the apparent speed of the crests and troughs of the attenuating sine wave; i.e.,

$$v'_\varphi(k, \epsilon/\alpha) \equiv \frac{\omega'}{k} = 1 - k^{-1} \tan^{-1} \frac{\sin k}{\alpha/\epsilon + \cos k} \tag{8}$$

$$\rightarrow 1 - \frac{\epsilon}{\alpha} \frac{\sin k}{k}, \qquad \epsilon/\alpha \rightarrow 0. \tag{9}$$

In particular

$$v'_\varphi(k, \epsilon/\alpha) + v'_\varphi(k, \alpha/\epsilon) = 1 \tag{10}$$

$$v'_\varphi(k, 1) = \tfrac{1}{2} \tag{11}$$

$$\begin{aligned} v'_\varphi(k, \epsilon/\alpha) &\rightarrow 1 \quad \text{as} \quad k \rightarrow \pi \quad \text{if} \quad \epsilon/\alpha < 1 \\ &\rightarrow 0 \quad \text{as} \quad k \rightarrow \pi \quad \text{if} \quad \epsilon/\alpha > 1. \end{aligned}$$

At long wavelengths (i.e., $k \rightarrow 0$) $\omega' \sim k$ while $\omega'' - \text{const} \sim k^2$. Thus we ignore the attenuation and find an infinite-wavelength group velocity

$$\frac{d\omega'(0)}{dk} = v'_\varphi(0, \epsilon/\alpha) = (1 + \epsilon/\alpha)^{-1}.$$

In the case of shift-register operation ω is real, corresponding to a term in the Fourier representation of a time-dependent signal introduced into one station. Equation (4) then reduces to

$$k(\omega) = k' - ik'' = -k^*(-\omega), \qquad 0 \leqq \omega \leqq \pi \tag{12}$$

$$= \omega + \tan^{-1} \frac{\sin \omega}{\epsilon^{-1} - \cos \omega} + i[\ln \alpha - \tfrac{1}{2} \ln (1 + \epsilon^2 - 2\epsilon \cos \omega)] \tag{13}$$

$$\begin{aligned} &\rightarrow \omega + \epsilon \sin \omega + i(\ln \alpha + \epsilon \cos \omega), \qquad \epsilon \rightarrow 0 \\ &\rightarrow \omega + \epsilon \sin \omega - i[\ell + \epsilon(1 - \cos \omega)], \qquad \epsilon, \ell \rightarrow 0; \end{aligned} \tag{14}$$

i.e., the wave is attenuated as a function of x but is purely sinusoidal in time at a given x. The identity following equation (6) was again used. The quantity ik corresponds to the *propagation function* of (continuous-time) lossy transmission lines; similarly k' and k'' correspond to the *phase function* and *attenuation function* respectively.[15] When Fig. 1 is rotated through 180 degrees in its plane, it becomes a plot of equation (13). Although not considered here it is sometimes useful to regard α and ϵ as ω-dependent quantities.

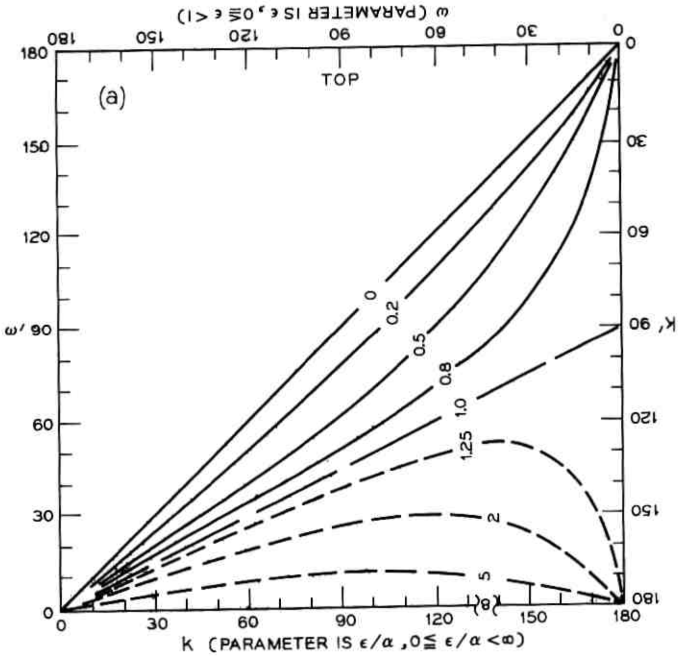


Fig. 1a

Figs. 1a and 1b—For a spatial sine wave of the form $\exp[i(\omega t - kx)]$ with $|x|$, $|t| = 0, 1, 2, \cdots$ the complex angular frequency $\omega = \omega' + i\omega''$ is plotted versus the (real) angular wave number k. After conversion from degrees to radians, k is measured in units of (2π times the stations spacing)$^{-1}$. The curves are parameterized by the ratio ϵ/α where α is the fraction of the charge successfully advanced per stepping operation and ϵ is the fraction of the charge which remains in a station per step [From equation (6)].

Fig. 1a—The real part of ω. (Note that $\omega'(k, \epsilon/\alpha) + \omega'(k, \alpha/\epsilon) = k$.)

Fig. 1b—The imaginary part of ω.

After rotating the figures by 180 degrees: For a sinusoidal signal of the form $\exp[i(\omega t - kx)]$ with $|x|$, $|t| = 0, 1, 2, \cdots$ (temporal sine wave) the complex angular wave number $k = k' - ik''$ is plotted versus the (real) angular signal frequency ω where ω is measured in degrees per stepping interval. The curves are parameterized by ϵ where $0 \leq \epsilon < 1$ is the physically significant range [From equation (13)].

Fig. 1a—The real part of k.

Fig. 1b—The imaginary part of k.

The phase velocity is given by

$$v_\varphi(\omega, \epsilon) = \frac{\omega}{k'} = \left[1 + \omega^{-1} \tan^{-1} \frac{\sin \omega}{\epsilon^{-1} - \cos \omega}\right]^{-1}$$
$$\to 1 - \epsilon \frac{\sin \omega}{\omega}, \qquad \epsilon \to 0. \tag{15}$$

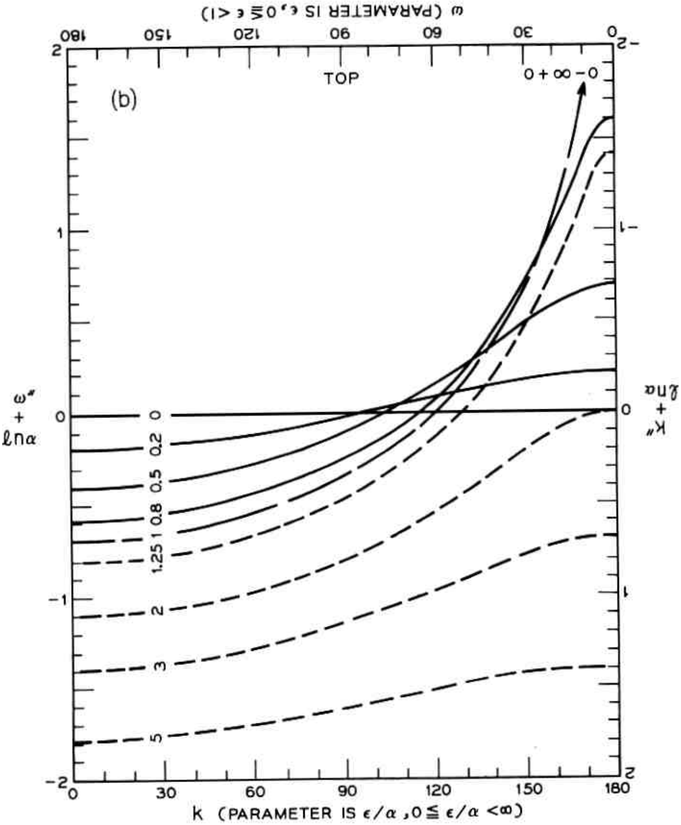


Fig. 1b

k'', and hence the attenuation, is frequency dependent for any $\epsilon > 0$ (just as ω'' is k dependent for any $\varepsilon/\alpha > 0$); however, as one would expect[16] from causality (Hilbert Transform; Kramers-Kronig Relation) in a continuous-time system, the apparent phase velocity ω/k' has no dispersionless case for $0 < \epsilon < 1$ analogous to equation (11). At low frequencies $k' \sim \omega$ while $k'' - \text{const} \sim \omega^2$. Thus we ignore the attenuation and find a zero-frequency group velocity

$$\left(\frac{dk'(0)}{d\omega}\right)^{-1} = v_\varphi(0, \epsilon) = 1 - \epsilon.$$

For a numerical example in the small-ϵ approximation, let a time-dependent signal be introduced into a shift register at $x = 0$ and extracted at station x. If x is large, the product ϵx is not necessarily small compared to unity; consider the case $\epsilon x = 0.2$. From equation (15) the phase transit time through the shift register is $x/v_\varphi = xk'/\omega = x + x\epsilon \sin \omega/\omega$. Thus the highest frequency component of the signal ($\omega = \pi$) is transmitted in the ideal (i.e., $\epsilon = 0$) time x. Lower frequencies take increasingly long with the $\omega = 0$ component arriving $x\epsilon = 0.2$ stepping intervals after the $\omega = \pi$ component. The amplitude is attenuated by the factor $e^{-k''x} = \alpha^x e^{\epsilon x \cos \omega}$ which is a decreasing function of frequency. Thus the cut-off frequency ($\omega = \pi$) is attenuated by an additional factor of $e^{-2\epsilon x} = e^{-0.4} = 0.67$ over the attenuation of the $\omega = 0$ component.

Equation (4) is invariant under the transformation

$$\omega \rightarrow k \qquad k \rightarrow \omega$$

$$\alpha \rightarrow 1/\alpha \qquad \epsilon \rightarrow -\epsilon/\alpha$$

which can therefore be used to derive from each other the two parallel sets of relations developed in this section.

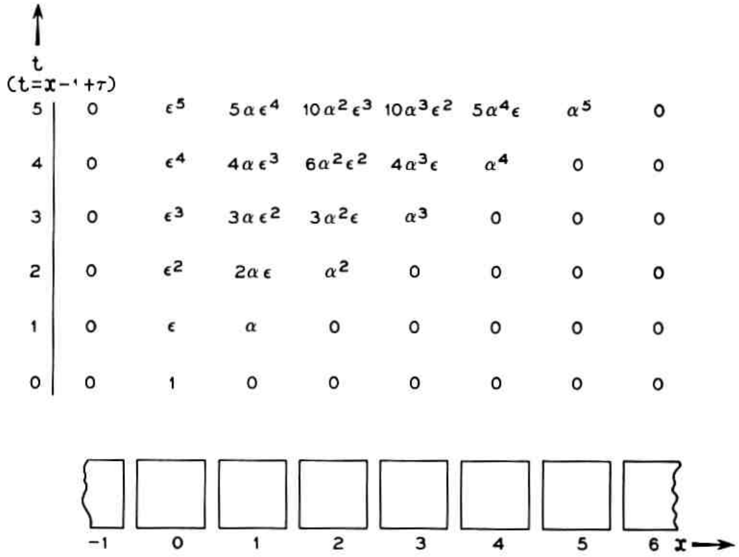
If ω is eliminated from equations (3) and (4), equation (25) results; if k is eliminated, equation (33) results.

IV. IMPULSE REPRESENTATION

Having established the connection with the usual basis for characterizing continuous linear systems in terms of their effect on sine waves, we turn to the impulse representation and investigate the solution of a unit charge placed in station $x = 0$ at $t = 0$ with all other stations initially uncharged. By considering the spatial distribution after the first few steppings as shown in Fig. 2, the solution to equation (2) with this boundary condition can be recognized as the binomial expansion of $(\alpha + \varepsilon)^t$; i.e.,

$$\begin{aligned} q_{x,t} = G_{x,t} &\equiv \binom{t}{x} \alpha^x \epsilon^{t-x}, \qquad 0 \leqq x \leqq t \\ &\equiv 0 \qquad \text{otherwise} \end{aligned} \tag{16}$$

where $\binom{t}{x}$ is the widely tabulated binomial coefficient. Later we will regard $G_{x,t}$ as the Green's function for more general initial or boundary conditions. Although our interpretation is deterministic rather than probabilistic, the terminology and results of probability theory will be quite useful. For example, equation (16) is a discrete bivariate



t								
5	0	ϵ^5	$5\alpha\epsilon^4$	$10\alpha^2\epsilon^3$	$10\alpha^3\epsilon^2$	$5\alpha^4\epsilon$	α^5	0
4	0	ϵ^4	$4\alpha\epsilon^3$	$6\alpha^2\epsilon^2$	$4\alpha^3\epsilon$	α^4	0	0
3	0	ϵ^3	$3\alpha\epsilon^2$	$3\alpha^2\epsilon$	α^3	0	0	0
2	0	ϵ^2	$2\alpha\epsilon$	α^2	0	0	0	0
1	0	ϵ	α	0	0	0	0	0
0	0	1	0	0	0	0	0	0

Fig. 2—Evolution of a unit charge initially in station $x = 0$ at $t = 0$. A fraction α of the charge is advanced per step and a fraction ϵ remains behind.

(x and t independent discrete variables) distribution, defined by a parabolic (partial) finite difference equation [Equation (2)].

In the description of practical devices one is usually interested in limiting expressions as ℓ and/or ϵ approach zero. Rather than writing several limiting forms at each point in the development, we emphasize once and for all an approximation which, in one variation or another, is frequently useful,

$$(\alpha + \epsilon)^t = (1 - \ell)^t = e^{t \ln (1-\ell)} \approx e^{-t\ell}, \quad t\ell^2 \ll 1. \tag{17}$$

Although in a real device the loss per transfer ℓ may be orders of magnitude less than unity, the number of stations and hence transfers may be so large that $t\ell$ is comparable to unity, and the further approximation $e^{-t\ell} \approx 1 - t\ell$ is not generally accurate.

Consider first the spatial distribution of charge with time regarded merely as a parameter, i.e., a horizontal row in Fig. 2. In probability theory it is usually assumed that $\alpha + \epsilon = 1$ (probability of heads plus probability of tails are unity). To make direct use of extensively compiled properties define

$$\alpha' = \alpha/(1 - \ell), \quad \epsilon' = \epsilon/(1 - \ell) \tag{18}$$

so that $\alpha' + \epsilon' = 1$. Then equation (16) becomes

$$G_{x,t} = (1 - \ell)^t \binom{t}{x} \alpha'^x \epsilon'^{t-x} = (1 - \ell)^t b_x , \qquad x = 0, 1, \cdots , t \tag{19}$$

which defines b_x the binomial distribution[17] with normalization

$$\mu_0 \equiv \sum_{x=0}^{t} b_x = (\alpha' + \epsilon')^t = 1 \tag{20}$$

and $(1 - \ell)^t$ is just a position-independent scale factor. (The intermediate steps involved in carrying out the numerous sums which follow can be found in many texts on probability theory.[17])

If the speed of the pulse is defined as the speed of the center of mass of that charge remaining at any time, then from

$$\langle x \rangle \equiv \sum_{x=0}^{t} x b_x = \alpha' t \tag{21}$$

it follows that the speed is $\alpha' = (1 + \epsilon/\alpha)^{-1}$, i.e., the same as the infinite-wavelength group velocity following equation (11). In general, the pulse can be characterized by its central moments μ of which the rth

$$\mu_r = \sum_{x=0}^{t} (x - \langle x \rangle)^r b_x , \qquad r = 0, 1, \cdots , \infty \tag{22}$$

can be obtained explicitly as the coefficient of $s^r/r!$ in the Taylor-series expansion in s of the *central-moment-generating function*[17]

$$e^{-s\langle x \rangle} \sum_{x=0}^{t} e^{sx} b_x = e^{-s\langle x \rangle} (\epsilon' + \alpha' e^s)^t . \tag{23}$$

Thus in particular the variance of the pulse, a useful measure of its spread, is $\mu_2 = t\alpha' \epsilon' = t(2 + \alpha/\epsilon + \epsilon/\alpha)^{-1}$ where the last form again explicitly exhibits the fact that normalization-independent properties depend on α and ϵ only through their ratio. Higher central moments are conveniently obtained from Romanovsky's recursion relation.[17,18]

$$\mu_{r+1} = \alpha' \epsilon' (tr\mu_{r-1} + d\mu_r/d\alpha'). \tag{24}$$

Some further useful relations are as follows: When the pulse spreads excessively for digital applications, one alternative is to utilize $2p + 1$ consecutive stations for one pulse. It then follows directly from the parallel-axis theorem for the moment of inertia that the variance is increased by (only) the additive term $p(p + 1)/3$ over the previous result.

At time t the ratio of the charge in the next-to-leading station to that in the leading station of the pulse is $t\epsilon/\alpha$. Thus the leading station of the pulse is also the most heavily charged until time $t = \alpha/\epsilon$ when the peak starts to drop back from the lead. At all times the charge distribution falls off monotonically from the peak. At any time the fraction of the remaining charge confined to the leading station is $(1 + \epsilon/\alpha)^{-t}$.

The results in Section III can be derived as consequences of those of Section IV, and conversely. For example the single-spatial-Fourier-transform solution of equation (2) can be obtained via the binomial distribution regarded as a Green's function (in the terminology of mathematical physics). Thus the initial sinusoidal spatial charge distribution $q_{x,0} = \text{Im}\, e^{-ikx}$ evolves into $q_{x,t}$ at time t where

$$\begin{aligned} q_{x,t} &= \text{Im} \sum_{x=0}^{t} G_{x',t} e^{-ik(x-x')} \\ &= \text{Im}\,(\epsilon + \alpha e^{ik})^t e^{-ikx} = \text{Im}\,(\alpha + \epsilon e^{-ik})^t e^{ik(t-x)}. \end{aligned} \tag{25}$$

Equation (25) is equivalent to equations (3) and (4) with ω eliminated. Equation (6) results if the real and imaginary parts of equation (25) are extracted. Apart from the normalization, the evolution factor which carries the initial spatial distribution e^{-ikx} into the later distribution can be recognized as $(\epsilon' + \alpha' e^{ik})^t$, the *characteristic function* (Fourier transform) of the binomial distribution in probability theory.[17] In the last form of equation (25), $(\alpha + \epsilon e^{-ik})^t$ evolves the unperturbed or ideal $(\epsilon = 0,\ \alpha = 1)$ solution $e^{ik(t-x)}$ into the actual solution. Finally $(1 + (\epsilon/\alpha)e^{-ik})^t$ ($\rightarrow \exp(e^{-ik}t\epsilon/\alpha)$ as $\epsilon/\alpha \rightarrow 0$) evolves the $\epsilon = 0$ solution $\alpha^t e^{ik(t-x)}$.

As before [cf. equation (7)] we seek an approximation of the Green's function in the limit that ϵ/α, but not necessarily $t\epsilon/\alpha$, is small compared to unity. If ℓ is also small, repeated use of equation (17) yields

$$G_{x,t} \rightarrow e^{-t\ell} p(t - x), \qquad \epsilon, \ell \rightarrow 0, \quad t \rightarrow \infty, \quad t\epsilon, t\ell = \text{const.}$$

$$\lambda \equiv t\epsilon, \quad x = t, t - 1, \cdots \tag{26}$$

where $p(u) = e^{-\lambda}\lambda^u/u!$ $(u = 0, 1, \cdots)$ is the Poisson distribution.[17] Sum rules similar to those illustrated for the binomial distribution can be carried out yielding in particular a pulse speed $\langle x \rangle/t = 1 - \lambda/t = 1 - \epsilon$ and a variance $\lambda = t\epsilon$. As long as the pulse peaks near or at the leading station, the Poisson distribution is usefully accurate for the more strongly charged stations even when t is only modestly large. A comparison of the binomial distribution and its Poisson approximation are given in Table I.

It is impractical to summarize the very many relevant exact and approximate properties of the binomial distribution here, particularly since a thorough compilation is already available.[17] We have, however, given a few of the most important properties to illustrate the usefulness of this connection with the extensive literature of probability theory.

Turning now to a shift-register interpretation of equation (16), we regard x as a parameter (the number of the stations which are used as the register) and study the time-dependence of the charge in station $x - 1$, i.e., a vertical column in Fig. 2. The development proceeds in close analogy to that of the binomial distribution and most of the motivating remarks need not be repeated.

Table I—A Comparison of the Binomial and Negative Binomial Distributions with their Poisson Approximation

1	2	3	4	5	6
x					τ
9	0	0	0	0	−1
8	0	0.430	0.411	0.430	0
7	0	0.383	0.365	0.344	1
6	0	0.149	0.162	0.155	2
5	0	0.033	0.048	0.052	3
4	0	0.0046	0.011	0.014	4
3	0	4.1×10^{-4}	1.9×10^{-3}	3.4×10^{-3}	5
2	0	2.3×10^{-5}	2.8×10^{-4}	7.4×10^{-4}	6
1	0	7.2×10^{-7}	3.6×10^{-5}	1.5×10^{-4}	7
0	1	1×10^{-8}	4.0×10^{-6}	2.8×10^{-5}	8
−1	0	0	3.9×10^{-7}	4.9×10^{-6}	9
−2	0	0	3.5×10^{-8}	8.4×10^{-7}	10
−3	0	0	—	—	11

For the case $\alpha = 0.9$ (fraction of charge successfully advanced per step), $\epsilon = 0.1$ (fraction which remains in its station per step), and $\ell = 0$ (fraction of charge lost per step).

Column 1: Station number x for columns 2, 3, and 4.

Column 2: Initial distribution of charge among the stations.

Column 3: Binomial distribution of charge among stations eight steps later [from equation (16) and (19)].

Column 6: Detection time τ for Columns 5 and 4 as measured from time of initial detection, $\tau = 0$.

Column 5: α times the negative binomial distribution of charge nondestructively observed in station 7 (α times equation (27)) as a result of initial spatial distribution of Column 2 = distribution of charge observed by a charge-removing detector in station 8 (cf. Appendix).

Column 4: Poisson approximation to Columns 3 and 5 [from equation (26) or (34)].

It is notationally convenient to use a new time τ which starts at time $t = x - 1$ when some of the charge placed initially in station zero is first observed in station $x - 1$, i.e., $\tau = t - x + 1$. Then equation (16) becomes

$$G_{x-1,\tau} = \binom{\tau + x - 1}{x - 1}\alpha^{x-1}\epsilon^{\tau}, \qquad \tau = 0, 1, \cdots, \infty. \tag{27}$$

If $\alpha^{\dagger}$ is defined as $1 - \epsilon$ (so that $\alpha^{\dagger} + \epsilon = 1$), equation (27) becomes

$$G_{x-1,\tau} = \alpha^{-1}\left(\frac{\alpha}{1-\epsilon}\right)^{x}\binom{\tau + x - 1}{x - 1}\alpha^{\dagger x}\epsilon^{\tau} \equiv \alpha^{-1}\left(\frac{\alpha}{1-\epsilon}\right)^{x} n_{\tau} \tag{28}$$

and defines n_τ the negative binomial distribution[17] with normalization

$$\mu_0 \equiv \sum_{\tau=0}^{\infty} n_{\tau} = \alpha^{\dagger x}(1 - \epsilon)^{-x} = 1. \tag{29}$$

Since charge which fails to advance is observed again, the sum of all charge (nondestructively) observed in station $x - 1$, i.e., $\alpha^{x-1}(1 - \epsilon)^{-x}$, may exceed unity even if $\ell \geqq 0$ (cf. the discussion of boundary conditions in the appendix). Whereas all normalization-independent properties of the binomial distribution depend on α and ϵ only through their ratio ϵ/α, here all normalization-independent properties depend only on ϵ (cf. Figs. 1 and 2).

The charge in station $x - 1$ is observed at a mean time $\langle t \rangle = x - 1 + \langle \tau \rangle$ where

$$\langle \tau \rangle = \sum_{\tau=0}^{\infty} \tau n_{\tau} = x\epsilon(1 - \epsilon)^{-1} \tag{30}$$

and thus in the shift-register context the pulse speed can be defined as $(x - 1)/\langle t \rangle$ which approaches $1 - \epsilon$ for large x; i.e., the same as the zero-frequency group velocity following equation (15) but faster than the definition following equations (11) and (21) except in the case of no charge loss ($\ell = 0$) when all definitions agree.

The rth central moment of the charge sequence in station $x - 1$ is

$$\mu_r = \sum_{\tau=0}^{\infty} (\tau - \langle \tau \rangle)^r n_{\tau} \tag{31}$$

which can be obtained by direct calculation or as the coefficient of $s^r/r!$ in the Taylor expansion in powers of s of the central-moment generating function[17]

$$e^{-s\langle\tau\rangle} \sum_{\tau=0}^{\infty} e^{s\tau} n_\tau = e^{-s\langle\tau\rangle}(1 - \epsilon e^{s})^{-x}(1 - \epsilon)^{x}. \tag{32}$$

In particular, the variance is $\mu_2 = x\epsilon\,(1 - \epsilon)^{-2}$.

The first charge observed in station $x - 1$ is a fraction $(1 - \epsilon)^x$ of all that will be observed in that station. It is also the largest charge observed in $x - 1$ if $x\epsilon$, the ratio of the second amount observed to the first, is less than unity. Thus the input, $x\epsilon$, for the example at the end of Section III can be obtained experimentally from the ratio of the first two nonzero charges present in $x - 1$ due to a single initial charge in $x = 0$. More accurate graphical methods actually developed for statistical contexts can be directly applied here to infer the parameters from the final distribution resulting from an initial pulse.[17]

According to equation (2) a sinusoidal signal $q_{0,t} = \text{Im}\, e^{i\omega t}$ present in station zero at time t adds an additional charge $\Delta q_{1,t+1}$ to station 1 at $t + 1$ where $\Delta q_{1,t+1} = \alpha q_{0,t}$ (cf. the Appendix). Thus from equation (28) the charge present in station x at time t is

$$\begin{aligned} q_{x,t} &= \text{Im} \sum_{\tau=0}^{\infty} G_{x-1,\tau}\alpha e^{i\omega(t-x-\tau)} \\ &= \text{Im}\left(\frac{\alpha}{e^{i\omega} - \epsilon}\right)^{x} e^{i\omega t} = \text{Im}\left(\frac{\alpha}{1 - \epsilon e^{-i\omega}}\right)^{x} e^{i\omega(t-x)} \end{aligned} \tag{33}$$

which is equivalent to equations (3) and (4) with k eliminated. Equation (13) results if the real and imaginary parts of equation (33) are extracted. Apart from the normalization the transfer factor which multiplies the initial signal $e^{i\omega t}$ in equation (33) is the characteristic function of the negative binomial distribution.

In the limit of small ϵ and ℓ, but not necessarily small ϵx or $\epsilon\ell$, $G_{x,\tau}$ becomes exp $(-\ell x)$ times the Poisson distribution; i.e.,

$$G_{x,\tau} \to e^{-\ell x} e^{-\Lambda}\Lambda^{\tau}/\tau!, \qquad \ell, \epsilon \to 0;\, x \to \infty;\, \ell x,\, \epsilon x = \text{const}\ \Lambda = x\epsilon \tag{34}$$

which is functionally equivalent to equation (26). Thus as illustrated in Table I, the distinction between the spatial and temporal projections of the impulse solution disappears in the Poisson limit [cf. equations (7) and (14)].

V. SUMMARY

Without reviewing any one of the three specialized fields individually, we have emphasized by way of an explicit case of current interest the connection between, on the one hand, the present practice of characterizing discrete charge-transfer devices in terms of their charge-transfer efficiencies and, on the other hand, the two well-developed fields of (analog) sinusoidal signal analysis and probability theory (interpreted deterministically).

Parenthetically we note that the model of this paper [equation (2)] is of some tutorial interest in that it permits the basic ideas of a traveling-wave description of discrete-space-and-time linear systems (including therefore as special cases the limits of continuous space and time) to be exemplified immediately in a unified manner via a simple device. The usual textbook vehicles such as sampled-data control systems, lattices, or transmission lines suffer tutorially in varying degrees from being too restricted (discrete in only space or time; no attenuation) and requiring a lengthy and specialized physical explanation of the origin of the basic equation in a less easily visualized system.

VI. ACKNOWLEDGMENTS

We would like to thank L. A. Shepp for a number of helpful conversations and in particular for recognizing equation (27) as the negative binomial distribution. We have also benefited from the comments of C. N. Berglund, R. W. Dixon, E. I. Gordon, J. McKenna, G. E. Smith, H. A. Watson, and S. H. Wemple.

APPENDIX

In the main portion of this paper the boundary conditions were deemphasized. For example, in shift-register operation the charge distribution was assumed to be unaffected by its own detection. Let $q_{x,t}$ be the nondestructively observed charge as distinguished from $q'_{x,t}$, the charge measured by a destruction process that removes all of the charge from station x. Then by equation (2) these two limiting cases are simply related as follows

$$q'_{x,t} = \alpha q_{x-1,t-1} ; \tag{35}$$

i.e., the expressions given for a nondestructively observed shift-register need only be multiplied by α to obtain their destructively measured counterparts in a register with one more station. In the case of a detection process which removes a fixed fraction of the charge, the signal can be obtained from the difference between successive measurements.

Similarly the two limiting cases at the input station of the register are Dirichlet (voltage) and Neumann (charge) boundary conditions. In the Neumann case a fixed amount of charge is added to whatever charge may already be in the station. In the Dirichlet case a fixed applied voltage adds charge as necessary to what is already in the station thereby achieving a total charge predetermined by the voltage level. By equation (2), and in analogy to equation (35), these two cases are simply related because a fixed total charge $q_{x,t}$ in x at t injects a fixed additional charge $\alpha q_{x,t}$ into $x + 1$ at $t + 1$ (cf. equation (33)).

To generalize equation (2) consider a constant inhomogeneous term q_d which might, for example, represent the amount of dark-current charge added per station per stepping interval. Then (2) becomes

$$q_{x,t} = \alpha q_{x-1,t-1} + \epsilon q_{x,t-1} + q_d . \tag{36}$$

An important particular solution of equation (36) is the spatially-uniform rising-in-time solution which starts from a spatially uniform distribution q_0 at $t = 0$:

$$\begin{aligned} q_{x,t} &= \beta^t q_0 + \frac{1-\beta^t}{1-\beta}\, q_d, \qquad \beta = \alpha + \epsilon, \qquad t = 0, 1, \cdots \\ &= q_0 + q_d t, \qquad \ell = 0. \end{aligned} \tag{37}$$

This could describe a recirculating memory where the output of the last station is fed back into the initial station. The constant-in-time rising-in-space solution is

$$\begin{aligned} q_{x,t} &= \gamma^x q_0 + \frac{1}{1-\epsilon}\,\frac{1-\gamma^x}{1-\gamma}\, q_d, \qquad \gamma \equiv \frac{\alpha}{1-\epsilon}, \qquad x = 0, 1, \cdots \\ &= q_0 + \frac{x}{1-\epsilon}\, q_d, \qquad \ell = 0, \end{aligned} \tag{38}$$

where q_0 is now the charge in the (initial) station $x = 0$. This steady-state distribution could be maintained in a shift register by removing a charge $q_d - (1 - \epsilon) q_0$ from the initial station per step. Equation (2) is now understood to describe a signal (homogeneous solution) imposed on some particular solution to equation (36) describing the background charge.

Next assume, as is certainly true in some devices, that the transfer fractions α and ϵ depend upon at least some of the q's; i.e., the process is nonlinear. Then in the small-signal case (whether or not q_d is further generalized to be a function of x and t) α and ϵ in equation (2) would still be independent of the magnitude of the signal charge but no longer independent of x and t even though all stations were physically equivalent apart from their momentary charges. In the case of equation (37) α and ϵ would still be x independent; in the case of equation (38) still t independent[19] (cf. nonuniform transmission lines).

Before proceeding we switch to the suitably developed terminology of probability theory (without necessarily implying a probabilistic interpretation). Actually, this has some plausible justification. To the extent that equation (2) applies independently to each of the indivisible electrons or holes in the charge, α and ϵ must be interpreted as transition probabilities, and our deterministic interpretation results only because the large number of electrons or holes permits fluctuations to be ignored. In this probabilistic sense one can define the information content of a discrete charge distribution.[20] Similarly in equation (36), $q_d(x, t)$ could be a random variable describing the introduction of extraneous noise.

To complete the small-signal or linear model for the signal we note that in practice the stations are driven by n-phase time-dependent voltages such that in a coordinate system y which moves at the ideal signal speed ($y = x - t$) the voltage distribution among the stations appears constant in discrete time and periodic in discrete space with a period of n stations. In each cycle of n stations one station is the designated potential well which, in the ideal case, carries all of the charge in that spatial cycle. The distortion of the signal corresponds to the transitions which the electrons or holes make to neighboring stations and, eventually, to neighboring wells. In a causal system the homogeneous equation for the signal charge takes the form

$$q_{y,t} = \sum_{t'=-\infty}^{t-1} \sum_{y'=-\infty}^{\infty} T_{y,y'\,,\,t,t'}\, q_{\,y',t'} \tag{39}$$

where the transition elements of the T matrix are q-independent in a small-signal theory. (In some cases, the upper limit of Σ is t rather than $t - 1$.)

If the system is memoryless such that the spatial charge distribution at one time is sufficient to determine the distribution at the next time and hence for all time, then equation (39) reduces to the Markoffian form

$$q_{y,t} = \sum_{y'=-\infty}^{\infty} T_{y,y',t,t-1} q_{y',t-1}\ . \tag{40}$$

This simplification corresponds, for example, to neglecting traps which accept signal charge at a rate proportional to the local signal and then emit the charges into the signal over a time scale long compared to the stepping interval.

When the background charge distribution, if any, is time independent, $T_{y,y't,t'}$ depends on t and t' only through their difference. If the background distribution is spatially constant, T is a periodic function of $y - y'$; i.e., T describes a random-walk[21] problem on a one-dimensional lattice with anisotropic and spatially periodic transition probabilities (cf. Floquet's theorem[13]).

Finally, of course, large signals under nonlinear conditions are of interest. However, it is not yet clear that there is any dependence of T on the q's which is more general than the specific device where it arises.

In the moving coordinate system equation (2) becomes

$$q_{y,t} = \alpha q_{y,t-1} + \epsilon q_{y+1,t-1} \tag{41}$$

which is functionally equivalent to equation (2) and distinct only in that the roles of α and ϵ, as well as the sense of spatial direction, are reversed. If the transition probabilities (T matrix elements) are small out of the wells and large out of the other stations between wells, then this 1-phase equation can be used to approximate an n-phase system. The unit of time is taken as the effective stepping interval equal to n actual stepping intervals, and successive integer values of x or y label the n-cycle groups or the wells. The matrix elements ϵ and α in equation (41) then describe a net transfer of charge between, or retention of charge by, the n-cycle moving groups during one (effective) time unit.

REFERENCES

1. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49,* No. 4 (April 1970), pp. 587–593.
2. Sangster, F. L. J., and Teer, K., "Bucket-Brigade Electronics-New Possibilities for Delay, Time-Axis Conversion, and Scanning," IEEE J. Solid State Circuits, *SC-1* (1969), pp. 131–136, and citations therein.
3. Berglund, C. N., and Boll, H. J., "Performance Limits of IGFET Bucket Brigade Shift Registers," Paper 2.3 presented at IEEE Int. Elec. Devices Meeting, Washington, D. C., October 28–30, 1970.
4. Amelio, G. F., Tompsett, M. F., and Smith, G. E., "Experimental Verification of the Charge Coupled Device Concept," B.S.T.J., *49,* No. 4 (April 1970), pp. 593–600.
5. Tompsett, M. F., Amelio, G. F., and Smith, G. E., "Charge Coupled 8-Bit Shift Registers," Appl. Phys. Letters *17,* 1970, pp. 111–115.
6. Berglund, C. N., "Analog Equivalent Circuit for Charge-Transfer Dynamic Shift Registers," IEEE J. Solid State Circuits, to be published.
7. Thornber, K. K., "Incomplete Charge Transfer in IGFET, Bucket-Brigade Shift Registers," IEEE Trans. Elec. Devices, to be published.
8. Strain, R. J., and Schryer, N. L., "A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer," B.S.T.J., this issue, pp. 1721–1740.

9. Amelio, G. F., "Computer Modeling of Charge-Coupled-Device Characteristics," unpublished work.

10. Goldberg, S., *Introduction to Difference Equations,* New York: John Wiley & Sons, Inc., 1958, pp. 196–200.
Jordan, C., *Calculus of Finite Differences,* New York: Chelsea Publishing Co., reprinted 1960, 2nd ed., pp. 615, 612.

11. An interesting history of Pascal's triangle by C. B. Boyer, "Cardan and the Pascal Triangle," appears in Amer. Math. Monthly *57*, pp. 387–390 (1950). Bernoulli also considered Pascal's triangle: Bernoulli, J., *Ars Conjectandi* (*De Seriebus Infinitis*), Basel, Switzerland, 1713, p. 87. The nonzero elements of our Fig. 2 reduce to Pascal's triangle when $\alpha = \epsilon = 1$.

12. Berglund, C. N., "The Bipolar Bucket Brigade Shift Register," unpublished work.

13. Brillouin, L., *Wave Propagation in Periodic Structures,* New York: Dover Publications, Inc., 1953.

14. Truxal, J. G., *Automatic Feedback Control System Synthesis,* New York: McGraw-Hill, Inc., 1955, Chap. 9.

15. Magnusson, P. C., *Transmission Lines and Wave Propagation,* Boston: Allyn and Bacon, 1970, 2nd ed., Chap. 4.

16. Faulkner, E. A., *Introduction to the Theory of Linear Systems* London: Chapman and Hall Ltd., 1969, equation (5.6.11).
Bode, H. W., *Network Analysis and Feedback Amplifier Design,* New York: D. Van Nostrand Co. Inc., 1945, Chap. 14.

17. Johnson, N. L., and Kotz, S., *Distributions in Statistics; Discrete Distributions,* Boston: Houghton Mifflin Co., 1969.

18. Romanovsky, V., "Note on the Moments of the Binomial $(p + q)^N$ about its Mean," Biometrika *15* (1923), pp. 410–412.

19. In the context of a probabilistic theory of learning equation (2) is analyzed with x-dependent α and ϵ ($\alpha + \epsilon = 1$), see Miller, G. A., and McGill, W. J., Psychometrika *17,* 369, 1952, and Goldberg, S., *Introduction to Finite Difference Equations* New York: John Wiley and Sons, Inc., 1958, pp. 200–203. See also W. Feller, *An Introduction to Probability Theory and Its Applications,* (John Wiley and Sons Inc., New York, 1968), 3rd ed., p. 230, 282.

20. Golomb, S. W., "The Information Generating Function of a Probability Distribution," IEEE Trans. Inform. Theory, *IT-11,* January 1966, pp. 75–77.

21. Spitzer, F., *Principles of Random Walk,* Princeton, N. J.: D. Van Nostrand Co., Inc., 1964.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Micromachining and Image Recording on Thin Films by Laser Beams

By D. MAYDAN

(Manuscript received March 22, 1971)

A theoretical and experimental description of laser machining of thin metallic films is given. Calculations are carried out for the temperature rise of a thin film in response to a pulse of incident light energy and for the dependence of the temperature rise on the different thermal constants of the substrate and film and on the illumination conditions.

Experiments have been conducted on the pulsed machining of thin bismuth films with glass and mylar as substrates using a lowest-order transverse mode argon laser with an average power output capability of 20 milliwatts. A fast intracavity acoustooptical modulation system produced optical pulses with durations controllable from 25 ns to several ms. The pulse repetition rate could be varied from a single pulse up to several MHz. The peak power depended on the duty cycle but was limited to 2 watts for low duty cycles. Experiments were done with both front and back illumination. (In back illumination the laser beam is incident on the film through the substrate.) For optimal machining conditions in which the optical beam diameter is adjusted to produce the maximum diameter of transparent area for a given pulse energy, less than 50 percent of the removed material left the surface of the substrate. The remainder was displaced so as to leave some areas free of bismuth while the thickness of bismuth in other areas was increased. With a pulse duration of 25 ns and a 600-Å-thick bismuth film on mylar, the peak power required to machine a 6-μm-diameter spot was about 0.7 watt.

Variable amplitude light pulses produced by the intracavity modulation system with a 1-MHz repetition rate and duration of 25 ns were used to write images on a 600-Å-thick bismuth film, deposited on a mylar substrate, by deflecting the laser beam in raster fashion over the surface of the bismuth film. The average laser power output was 20 *mW. The film area was 8 × 10 mm^2. Both positive and negative continuous-tone images were recorded. The images consisted of an array of 1200 by 2000 completely independent spots of varying diameter and a spot density of 4 × 10^6/cm^2. Images of documents which were originally of size 8½ × 11 inches showed a limiting resolution of over 175 lines per inch when magnified to the original size.*

I. INTRODUCTION

Images can be recorded on an opaque metallic film on a transparent substrate by using an amplitude-modulated laser beam to machine the metallic film. The modulated laser beam is scanned in raster fashion over the surface of the film, and the thermal energy dissipated in the film causes a local displacement or removal of the metallic material. In this way a permanent image suitable for immediate display or storage is created.

Experiments along this line have been reported[1,2] but reveal a number of shortcomings from the standpoint of practical application. Generally, there is no gray scale, only negative line images are reproduced, resolution is limited, and the required laser power is large.

A new laser machining technique is described in this paper which overcomes these shortcomings and permits the generation of high-quality, continuous-tone, permanent images using low-power gas lasers. The significant departure from previous work rests in the use of very short laser pulses to record the image in the form of an array of discrete holes in the metal film and in the modulation of the intensity of the laser pulses so as to vary the size of the holes. Each laser pulse serves to machine a single, nearly circular hole in the metal film by displacing and removing metal from the transparent substrate. In this way a transparency is created, and incandescent light can be directed through the film to display the image on a screen. The delivery of the laser energy to the metal film in short (20–30 nanosecond) pulses reduces the energy density required to raise the metal film to a given temperature by at least an order of magnitude over that which would be required for the same writing speed if a CW laser were used. Consequently, the average laser power required to displace or remove a given amount of metal per unit time from the transparent substrate is correspondingly reduced. In addition, the area of the transparent spot machined in the film varies in a nearly linear fashion with the pulse height, so that a good gray scale results. Equally good positive or negative images are produced, and successive, isolated machined spots have a transmission factor independent of adjacent spots or position along the scan line. The spots were machined at the rate of 10^6 spots/second, and the frames consisted of about 2000 lines of spots with about 1200 spots in each line.

The short laser pulses were generated with an argon ion laser in conjunction with an intracavity acoustooptic modulator[3,4]. The laser had an average power capability of 20 mW. The intracavity modulator served to deflect very short pulses from the laser cavity with much higher peak power but with an average power of approximately 20 mW.

In what follows, a model of the machining process will be developed based on electron micrographs of the machined spot. A calculation of the temperature rise of the film as a function of time is described and used to predict the laser power required to produce threshold machining for pulses of a given length.

II. DISCUSSION

2.1 *Temperature Rise of Thin Films Due to Laser Radiation*

A laser beam incident on a thin, optically absorbing film will cause the film temperature to rise.[5-7] Assuming a laser power density P_o transmitted through the film surface, the light intensity absorbed per unit thickness of the film at depth x may be described by

$$P_{\text{abs}}(x) = -\frac{d}{dx}(P_0 e^{-x/\delta}) = \frac{P_o}{\delta} e^{-x/\delta} \tag{1}$$

where δ is the skin depth of the film, that is, the depth at which the light intensity drops by a factor e from its initial value. If the laser beam diameter is very large compared with the film thickness, the heat flow may be treated as a one-dimensional problem. It will be assumed that the thermal and optical* properties of the absorbing material are independent of its temperature and phase.† Under these conditions and assuming no heat loss by radiation, the temperature rise of the thin film and the substrate is described by the differential equa-

* The optical properties of thin bismuth films were measured and no significant change was detected below the vaporization temperature. The measurements were carried out by monitoring the amount of transmitted and reflected light from the bismuth film during a laser machining process.

† While this assumption clearly limits the accuracy of the calculations, the experimental results presented later show reasonable agreement with the calculations (See Fig. 14, for instance) and hence the error involved may not be very great.

tions of heat conduction:[8]

$$\frac{\partial^2 T_1}{\partial x^2} - \frac{1}{k_1}\frac{\partial T_1}{\partial t} = -\frac{1}{K_1}\frac{P_o}{\delta}e^{-x/\delta} \tag{2}$$

$$\frac{\partial^2 T_2}{\partial x^2} - \frac{1}{k_2}\frac{\partial T_2}{\partial t} = 0 \tag{3}$$

where T_1 is the film temperature, T_2 is the substrate temperature measured relative to the ambient temperature, k_1 and k_2 are the thermal diffusivity of the film and substrate respectively, ($k = K/\rho c$), ρ is the density, c is the specific heat, K_1 and K_2 are the thermal conductivity of the film and substrate respectively, and t is time. The initial conditions, corresponding to the onset of illumination, are:

$$\text{at} \quad t = 0, \qquad T_1 = T_2 = 0.$$

The boundary conditions are:

$$\text{at} \quad t > 0 \quad \text{and} \quad x = 0,$$

$$\frac{\partial T_1}{\partial x} = 0;$$

$$\text{at} \quad t > 0 \quad \text{and} \quad x = d$$

(where d is the thickness of the metal film),

$$T_1 = T_2$$

and

$$K_1 \frac{\partial T_1}{dx} = K_2 \frac{\partial T_2}{\partial x}$$

(equal temperature and heat fluxes at film-substrate boundary). The substrate is assumed to be infinitely thick, so that $T_2 = 0$ at $x = \infty$. The temperature at depth x within the thin film can then be shown to be approximately equal to

$$\begin{aligned} T_1(x, t) \cong \frac{P_o}{K_1} \sum_{n=0}^{\infty} \Bigg\{ &+ 2\sqrt{k_1 t}\, \alpha^{n+1} \operatorname{ierfc}\left(\frac{2(n+1)d \pm x}{2\sqrt{k_1 t}}\right) \\ &- 2\sqrt{k_1 t}\, e^{-d/\delta} \frac{\Lambda\alpha^n}{1+\Lambda} \operatorname{ierfc}\left(\frac{(2n+1)d \pm x}{2\sqrt{k_1 t}}\right) \\ &+ 2\sqrt{k_1 t}\, e^{-d/\delta} \frac{\alpha^n}{1+\Lambda} \operatorname{erfc}\left(\frac{(2n+1)d \pm x}{2\sqrt{k_1 t}}\right) \\ &- \delta e^{-x/\delta} + 2\sqrt{k_1 t} \operatorname{ierfc}\left(\frac{x}{2\sqrt{k_1 t}}\right)\Bigg\} \end{aligned} \tag{4}$$

in which erfc is the complementary error function and ierfc is its integral over the argument between limits x and ∞.

$$\begin{aligned} \alpha &\equiv \frac{\Lambda - 1}{\Lambda + 1} \\ \Lambda &\equiv \left(\frac{K_1}{K_2}\right)\sqrt{\frac{k_2}{k_1}}. \end{aligned} \tag{5}$$

All the terms in equation (4) which have $\pm x$ should be repeated once with $+x$ and a second time with $-x$. The first term for example is equal to:

$$\sum_{n=0}^{\infty} 2\sqrt{k_1 t}\,\alpha^{n+1}\left\{\text{ierfc}\,\frac{2(n+1)d + x}{2\sqrt{k_1 t}} + \text{ierfc}\,\frac{2(n+1)d - x}{2\sqrt{k_1 t}}\right\}.$$

For the case of back illumination, where the laser beam is incident on a transparent substrate, the differential equations are the same as equations (2) and (3) with x replaced by $d - x$ in the exponential term. The boundary and initial conditions are the same as in the previous case. The temperature of the thin film in this case is approximately equal to

$$\begin{aligned} T_1(x, t) \cong \frac{P_o}{K_1} \sum_{n=0}^{\infty} \Bigg\{ &- 2\sqrt{k_1 t}\, e^{-d/\delta}\, \alpha^{n+1}\, \text{ierfc}\left(\frac{2(n+1)d \pm x}{2\sqrt{k_1 t}}\right) \\ &+ 2\sqrt{k_1 t}\,\frac{\Lambda\alpha^n}{\Lambda + 1}\,\text{ierfc}\left(\frac{(2n+1)d \pm x}{2\sqrt{k_1 t}}\right) \\ &- 2\sqrt{k_1 t}\, e^{-d/\delta}\,\text{ierfc}\left(\frac{x}{2\sqrt{k_1 t}}\right) \\ &+ \delta\,\frac{\alpha^n}{1 + \Lambda}\,\text{erfc}\left(\frac{(2n+1)d \pm x}{2\sqrt{k_1 t}}\right) - \delta e^{-(d-x/\delta)}\Bigg\}. \end{aligned} \tag{6}$$

The model used above refers only to illumination by a step function of optical intensity. For comparison with experimental pulses having finite rise and fall times, a correction factor depending on the time dependence of the incident light pulse should be included in equations (4) and (6). For most practical cases in which $d > \delta$ all the terms in equations (4) and (6) which contain the expression $e^{-d/\delta}$ are smaller than the other terms and can be ignored. In addition, for the cases in which the illumination duration t is 10 ns or longer, $\sqrt{kt} \gg \delta$ so that all other terms with δ can be ignored. Numerical results based on equations (4) and (6) were obtained for the temperature distribution across the film and for the temperature dependence on the time from the onset of illumination. Of a number of metallic films examined, bismuth films required the least energy to machine a given size of spot at a given optical density. For this reason, the results presented herein all relate to bismuth films on glass and mylar substrates.

The thermal constants of bismuth, glass, and mylar are given in Table I. The light absorption in bismuth was measured, and δ was found to be ~140 Å for the 4880-Å wavelength of the argon ion laser. The calculated temperature rise of the bismuth film as a function of the light pulse duration, with the film thickness as a parameter, is shown in Figs. 1, 2, and 3. Figures 1 and 2 illustrate the case of front illumination with mylar and glass substrates, respectively. The results in Figs. 1 and 3 are given for bismuth film thicknesses of 330, 670, and 1000 Å and in Fig. 2 for film thicknesses of 330 and 1000 Å. The temperature rise is shown both at the bismuth-substrate boundary ($x = d$) and the bismuth-air ($x = 0$) surface. Figure 3, which refers to a bismuth film on mylar with the laser beam incident onto the mylar substrate, is seen to be essentially identical to Fig. 1. For long times, the temperature is proportional to $\sqrt{t}$ (slope equal to that of the broken line on the figures). This same proportionality is obtained for the surface temperature of a semi-infinite solid with constant flux of heat at the surface.[9]

As seen in Figs. 1, 2, and 3, for short times (<50 ns), the temperature rise approaches a linear dependence on time such as one might expect if the thermal loss to the substrate were small compared with the absorption of thermal energy by the thin metallic film. This is especially so with the mylar substrate due to the lower thermal conductivity of the mylar as compared to glass. For the same reason, the temperature rise of the bismuth film on glass is smaller than that of the same film on mylar.

The temperature distribution through the thickness of the film in the case of a 1000-Å-thick bismuth film on mylar is shown in detail in

TABLE I—THERMAL CONSTANTS

Substrate	Conductivity [cal/(s)(cm^2) (°C/cm)]	Diffusivity [cm^2/s]	α	Λ
Bismuth	2×10^{-2}	7×10^{-2}		
Glass	28×10^{-4}	58×10^{-4}		
Mylar	4×10^{-4}	1.1×10^{-3}		
Bismuth on glass			0.35	2.06
Bismuth on mylar			0.72	6.30

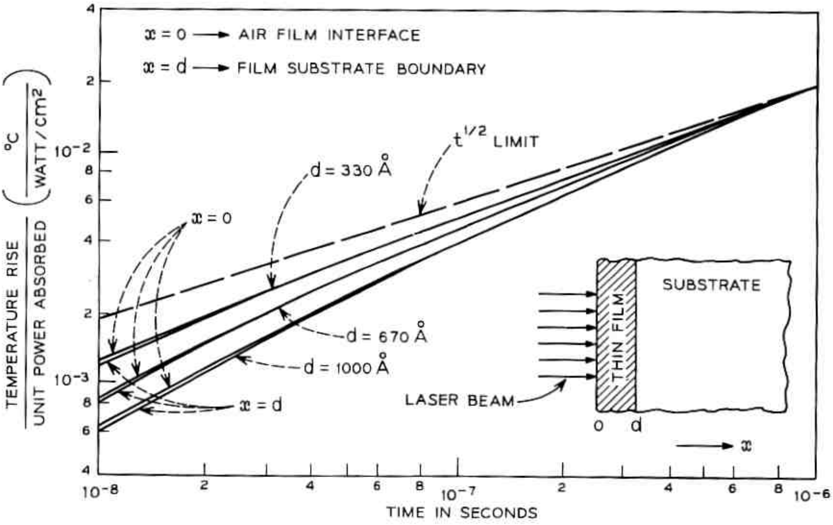


Fig. 1—Temperature rise of a bismuth film on a mylar substrate caused by front illumination.

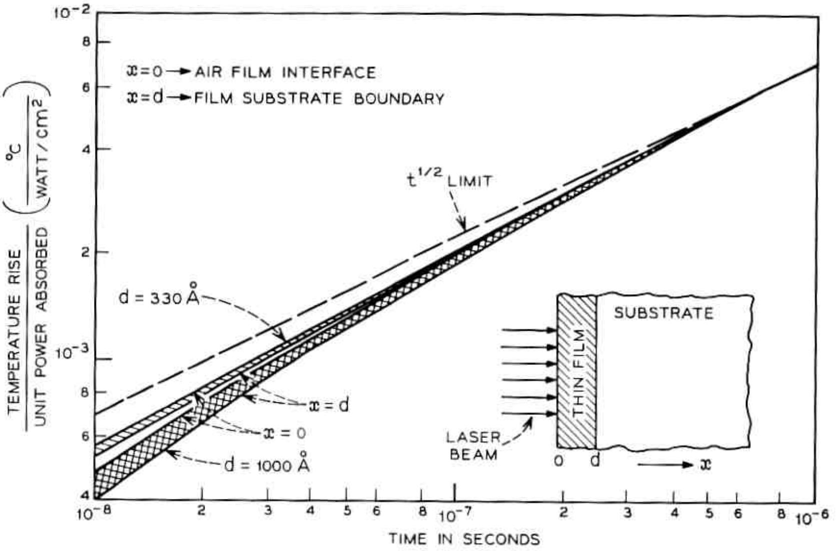


Fig. 2—Temperature rise of a bismuth film on a glass substrate caused by front illumination.

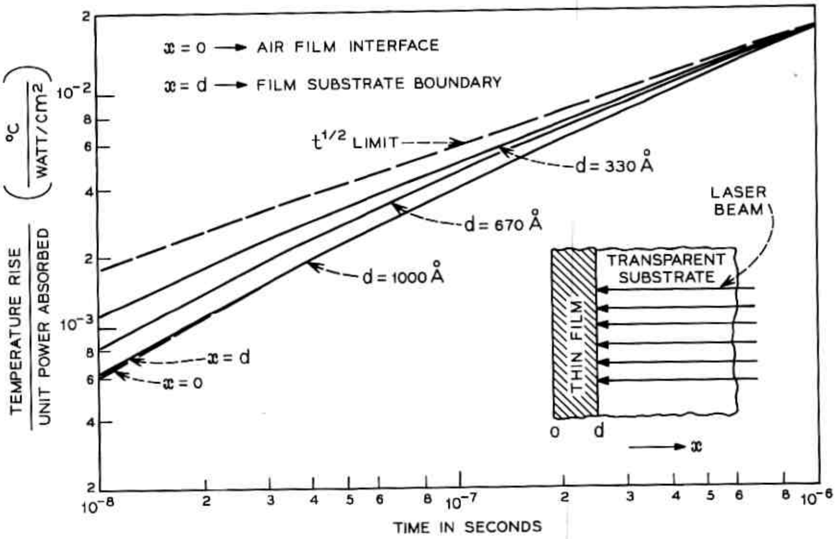


Fig. 3—Temperature rise of a bismuth film on a mylar substrate caused by back illumination.

Fig. 4a and b. The temperature distribution is given for the case of front illumination at the end of 10 ns (Fig. 4a) and at the end of 100 ns (Fig. 4b). The temperature gradient decreases with longer illumination, which can also be seen from the previous figures.

The temperature gradient across a 670-Å bismuth film on mylar, illuminated from the back, is shown in Fig. 5. Results are shown after 10 ns (Fig. 5a), after 100 ns (Fig. 5b), and after 1 μs (Fig. 5c). As expected, for short illumination duration, the temperature is higher near the substrate interface, where most of the laser intensity is absorbed. Because of loss of heat to the substrate, the location of maximum temperature shifts in time towards $x = 0$, and the temperature gradient decreases (Fig. 5c).

The temperature gradient across the film thickness is always very small (about 2 percent of surface temperature for the case described in Fig. 4b). For this reason, the model used for the foregoing calculations, which assumes that a single phase of the film is in existence at any instant, is adequate up to the vaporization temperature.

2.2 *Optimal Machining Considerations*

It is assumed that the laser beam power density absorbed by the thin film has an azimuthally symmetric Gaussian shape of the form

$$I(z, r) = I_o(z)e^{-(r/w(z))^2} \tag{7}$$

in which r is the radial distance from the axis, $I(z, r)$ is the laser intensity at a distance z from the focal plane along the optical axis, $I_o(z)$ is the laser intensity at $r = 0$, and $w(z)$ is the radius of the waist at the $1/e$ points of the intensity. For a beam propagating in the z direction, H. Kogelnik and T. Li[10] have shown that

$$\left(\frac{w(z)}{w_o}\right)^2 - 1 = \left(\frac{\lambda z}{2\pi w_o^2}\right)^2 \tag{8}$$

where λ is the light wavelength and w is the waist radius at the focal plane, $w_o \equiv w(o)$. Assuming a laser beam with a specified total power P, one can write

$$P = \int_0^\infty I_0 e^{-(r/w)^2} \cdot 2\pi r \, dr = \pi I_o w^2 \tag{9}$$

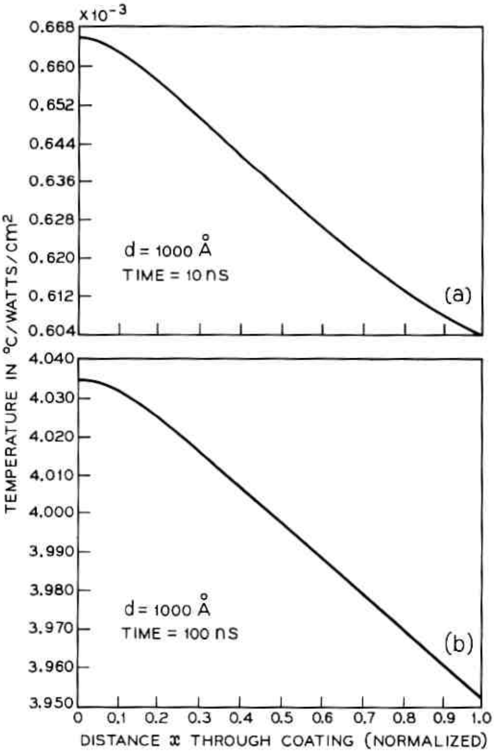


Fig. 4—Temperature distribution across 1000 Å of a bismuth film on mylar: (a) after 10 ns of front illumination; (b) after 100 ns of front illumination.

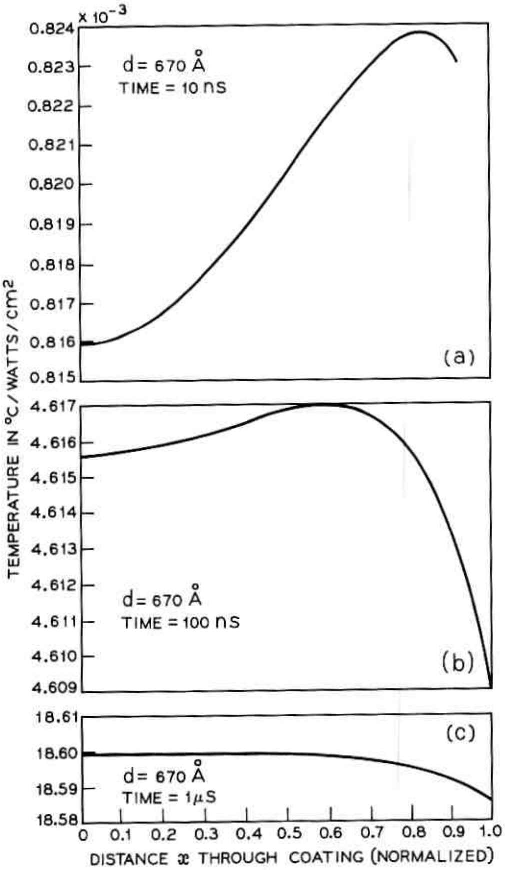


Fig. 5—Temperature distribution across 670 Å of a bismuth film on mylar after 10 ns (a), 100 ns (b), and 1 μs (c) of back illumination.

which defines

$$I_o(z) = \frac{1}{\pi} \frac{P}{w(z)^2}. \tag{10}$$

Machining or evaporation of the thin films with a spot diameter D and a given pulse duration requires a certain threshold intensity $I_T = I(z, D/2)$. At the focal plane ($z = 0$), it follows from equation (7) that

$$I_T = \frac{1}{\pi} P \frac{1}{w_o^2} e^{-(D/2w_o)^2} \tag{11}$$

or

$$P = \pi I_T w_o^2 e^{(D/2w_o)^2}$$

where D is the machined spot diameter. By taking the derivative of P with respect to w_o and setting $dP/dw_o = 0$, it can be shown that the laser power required to machine a spot of given diameter is minimum when

$$\frac{I_o}{I_T} = e. \tag{12}$$

It follows that the optimal beam waist diameter machining a spot of diameter D is

$$2w_{\text{opt}} = D. \tag{13}$$

As an example, consider the machining of a 1000-Å bismuth film with a mylar substrate. The laser power density absorbed in the film is assumed to have a Gaussian shape with a peak intensity I_o equal to 3×10^5 watts/cm². The melting temperature of bismuth is about 271°C. From Fig. 3 it is found that starting from room temperature the thin bismuth film will reach the melting temperature after 16 ns of illumination time. The latent heat of fusion of the bismuth is $L = 10.2$ cal/g, and the specific heat is $C = 0.03$ cal/g°C. The additional time Δt used in supplying the latent heat of fusion is estimated from the time required to raise the temperature by the amount $\Delta T = L/C =$ 300°C. From Fig. 3 this time is found to be $\Delta t \cong 20$ ns. Following that, the temperature will continue to rise toward the evaporation temperature. The time required to reach the evaporation temperature after supplying the latent heat of fusion is found from Fig. 3 to be equal to 124 ns. Assuming that machining is done under optimal conditions, as previously described, the temperature rise was calculated at different points across the diameter of the Gaussian beam. Results are shown in Fig. 6 for the temperature of the bismuth film after 160 ns illumination duration with a laser peak intensity of 3×10^5 watts/cm². The four lines shown correspond to the temperature at $a = 0$ (the center of the machined spot, at $a = 0.25D$, at $a = 0.50D$ (the edge of the machined spot), and at $a = 0.67D$, which is the radius where the film just reached the melting point.

When the laser beam is focused to a radius different from w_{opt}, the machined spot diameter is reduced. The ratio between the diameter of the machined spot and the optical beam diameter is found from equa-

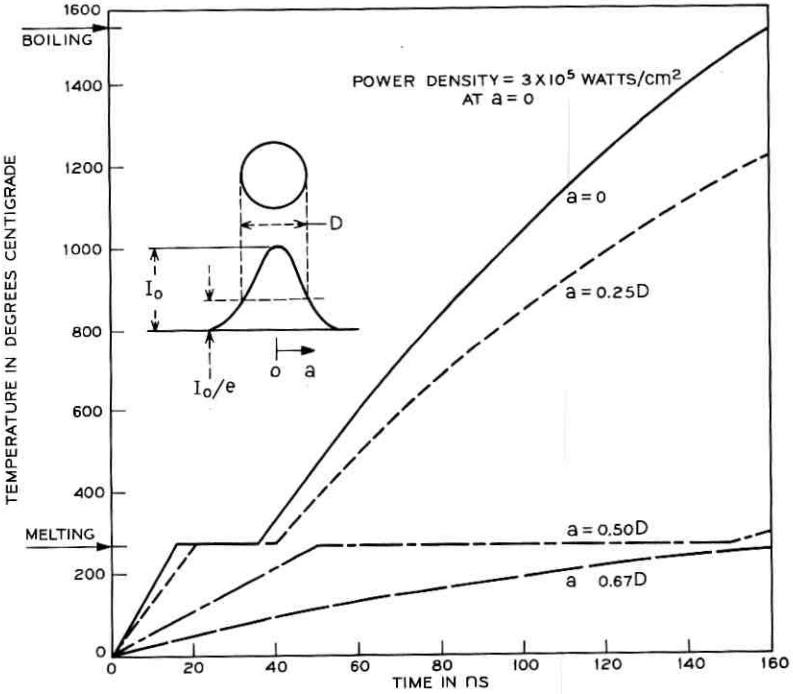


Fig. 6—Temperature distribution across the machine spot diameter.

tions (7), (10), and (12) to be:

$$\frac{D(z)}{2w(z)} = \sqrt{1 - 2\ \ln\ (w(z)/w_{\text{opt}})}. \tag{14}$$

Suppose a beam is focused to a waist w_{opt} which is optimal for machining a spot of diameter $D_{\text{opt}} = 2w_{\text{opt}}$. Let the depth of field be the distance l from the focal plane for which the machining spot diameter is reduced by a factor of $\sqrt{2}$. It is found from equations (8) and (14) that

$$l = \pm \frac{1.7D^2_{\text{opt}}}{\lambda}. \tag{15}$$

As an example, to machine a spot diameter of 10μm, with an argon ion laser which has a wavelength of 0.488 Å, the field depth according to (15) is

$$l = \pm \frac{1.7D^2_{\text{opt}}}{\lambda} = \pm 348\mu\text{m}.$$

III. EXPERIMENTAL RESULTS

Machining experiments have been conducted on thin bismuth films using glass and mylar as substrate materials. These films were prepared by vacuum evaporation techniques in thickness of 200, 400, 500, 600, 800, and 1000 Å respectively. A folded-cavity, CW argon laser was used in conjunction with a fast acoustooptic modulator[3] to provide light pulses whose length could be varied from 25 ns to several ms. The repetition rate of the light pulses could be controlled and varied from a single pulse up to several million pulses/second. The focused spot could be swept in one dimension using deflectors, such as galvanometers,[11] rotating mirrors, etc. However, for most of the data reported in this section, the focused spot was kept fixed, and the film was moved in a direction perpendicular to the optical axis of the light pulses, as shown in Fig. 7. Tilting the film at a small angle α with respect to its direction of travel (shown in Fig. 7) causes the bismuth film to pass from inside the plane of focus to outside in its travel past the focused spot. With sufficient power in the light pulse a line of circular spots was machined.

Figure 8 shows four pairs of lines machined with four different light intensities. The spot diameters along these lines exhibit two maxima corresponding to the optimum case described in Section II [equations (12) and (13)]. The two maxima are at equal distances from the spot machined when the film is at the beam waist. The pulse repetition rate

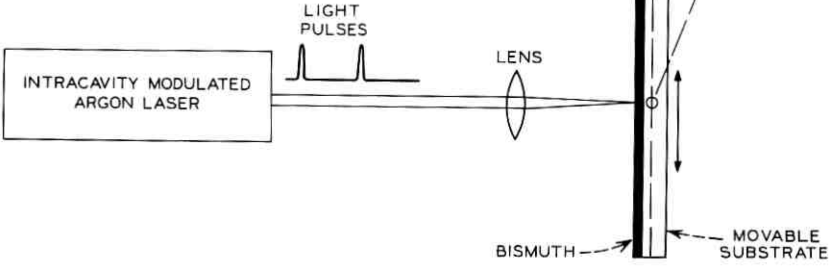


Fig. 7—Experimental setup.

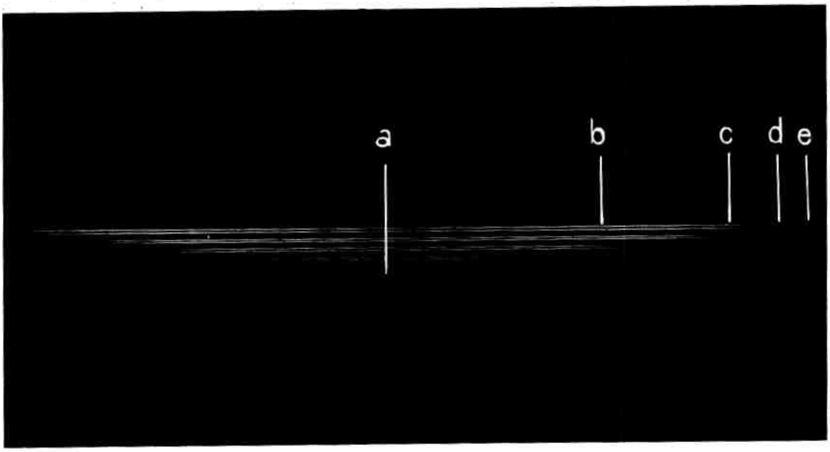


Fig. 8—Four sets of lines machined with different intensities.

and film speed were adjusted to produce nonoverlapping but nearly contiguous spots along each line; i.e., each spot was formed by a single optical pulse. The film in this case was a bismuth film of 1000 Å thickness on a glass substrate. The optical beam was focused to a diameter $2w_o$ of about 4μm. The light pulse duration was 25 ns with peak intensities of 1.6, 0.8, 0.5, and 0.25 watts for the first, second, third, and fourth pairs of lines, respectively. The maximum spot diameter (which occurs at the brightest part of each line in Fig. 8) and the separation between the two maxima in each individual line increase with the intensity of the incident light. The machining obtained when the beam was in focus on the bismuth film is marked by the vertical line in the center. The disappearance of machining at the two ends of each line occurs where the peak power at the center of the beam falls below the machining threshold power I_T.

A more detailed picture of some individual spots machined with the same laser intensity is given in Fig. 9. These photographs were obtained from a scanning electron microscope at 10,000 and 2000 times magnification. The approximate locations of these spots are indicated by the letters a, b, c, d, and e in Fig. 8. All spots shown in Fig. 9 are located on the top line of machined spots in Fig. 8. The individual machined spot magnified by 10,000 times in Fig. 9 is also marked by a small arrow on each 2000 magnification photograph. The machined spot obtained when the laser beam was focused in the plane of the bismuth film is shown in (a) of Fig. 9. Machining obtained with a power

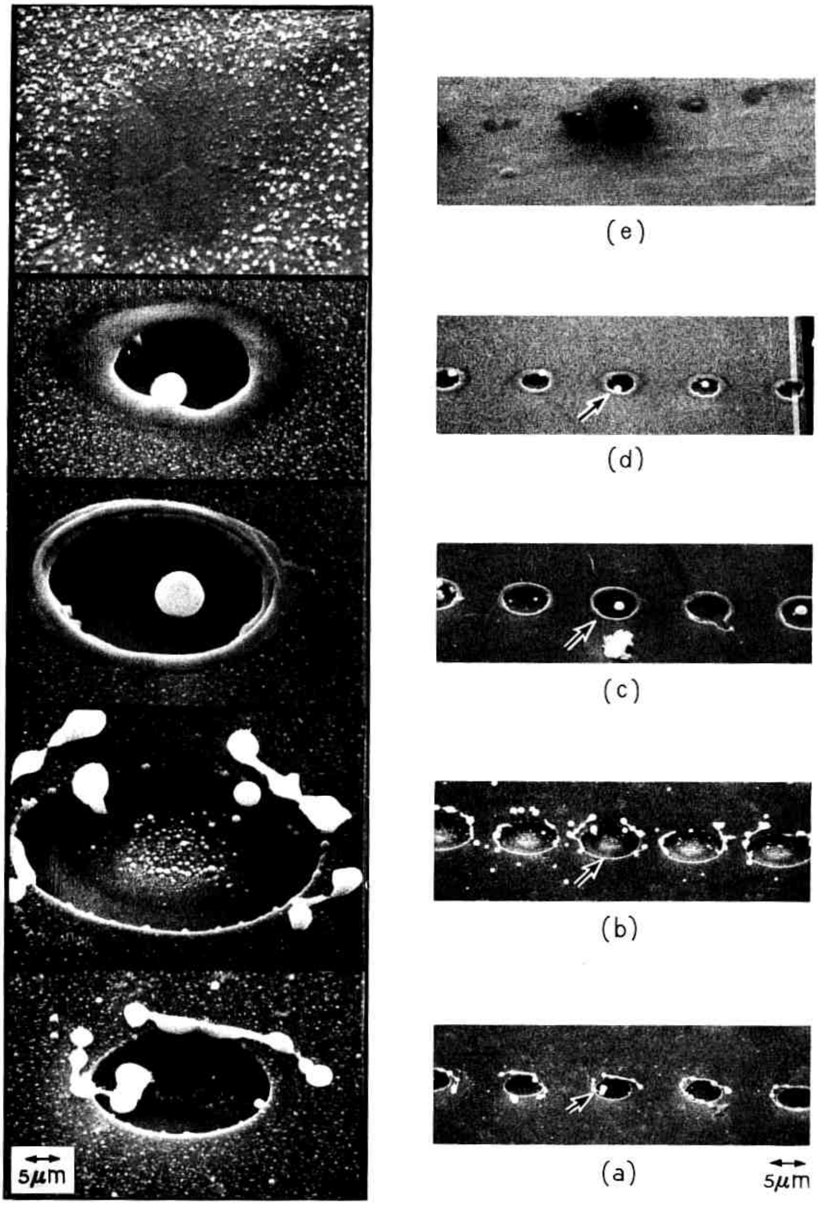


Fig. 9—Photographs of individual machining spots obtained from a scanning electron microscope.

just above threshold is shown in (d). The largest machined spot obtained is shown in (b). As clearly seen from these photographs, very little, if any, of the bismuth film was evaporated at the lower power densities, as shown in (c) and (d). Near the threshold power, the metal was only melted and the surface tension pulled the metal back from the center forming a crater, as shown in (c). In the machined spot at the center of (c), surface tension caused a large percentage of the liquid metal to resolidify into a sphere within the crater. In (e) the laser power was just below the threshold level, and while the metal appears to have been melted at each spot, only in a few cases were craters formed.

In (a) and (b) of Fig. 9 the optical beam intensity at the center of the spot is greater, and it is likely that some evaporation of material occurred in this region. The metal vapor pressure has forced some of the liquid metal away from the substrate, and surface tension has caused the remainder of the liquid bismuth to resolidify in the irregular form evident in (a) and (b). By estimating the volume of material accumulated around the crater in (b), the amount of material evaporated for the optimum machining is found to be less than 50 percent of the total metal displaced from the crater.

In order to measure the machining field depth, a similar experiment was carried out. Argon laser pulses with 25 ns duration and peak power of 0.75 watt were focused to a minimum diameter of 2μm. The bismuth sample was again moved perpendicular to the optical axis and was tilted at a small angle α with respect to its direction of travel. Assuming that the beam has a Gaussian shape, the optical beam diameter could be calculated for each machined spot. The relation between the machined spot diameter and the light beam diameter is given in Fig. 10. The results are given for an 800-Å-thick bismuth film on a glass substrate. The solid line is the theoretical line given by equation (14). As seen in this figure, the observed machining field depth is extended beyond that of the theoretical curve, especially at the lower power densities. This may result from the surface tension effect near the threshold power.

Figure 11 shows the effect of varying the laser intensity. A 400-Å-thick bismuth film on a mylar substrate was used in this case. The optical beam was in focus all along the machined line, and the machining was obtained by back illumination. As indicated on this figure, the beam intensity was varied from 100 percent to 24 percent, the 100-percent intensity corresponding to a light pulse of 1 watt peak intensity with 25 ns duration. The machining spots varied in diameter according

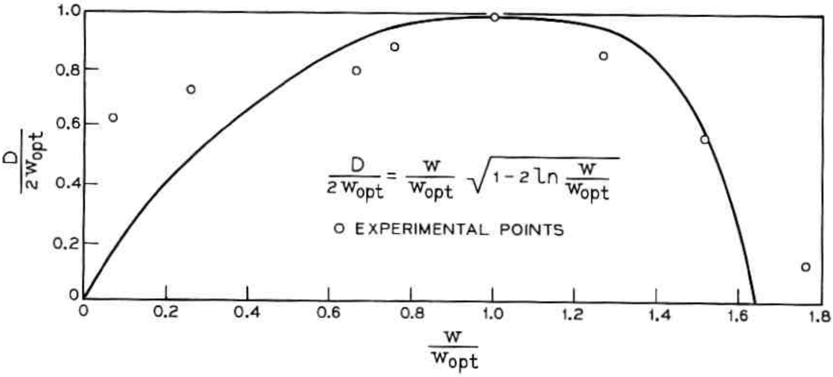


Fig. 10—Relation between machining spot and light beam diameters. The notation is that used in equations (13) and (14).

to the incident light intensity, from about 8 to 1μm. The optical beam diameter was $2\ w_o \approx 6\mu$m. As shown in Fig. 12, above threshold the measured machining area varied linearly with the light beam intensity for these particular machining conditions.

The variation of threshold machining power density with pulse duration was measured for pulse durations in the range from 25 to 4000 ns. The light pulse shapes for several durations are shown in Fig. 13. The results obtained with a 600-Å-thick bismuth film on mylar (illuminated from the substrate side) are given in Fig. 14. The solid line is the theoretical curve based on the results given in Section II. The small circles are the experimental results. Both the theoretical and the experimental results in Fig. 14 have been normalized so that they coincide at the 1-μs illumination period. The small deviations from the theoretical results for the short durations are not yet explained.

The peak pulse power required to machine 6-μm-diameter spots on 500-Å bismuth films on mylar with 25 ns pulse duration was found to be of the order of 0.7 watt. This corresponds to an average energy density of 0.06 joule/cm^2 over the area of the spot. As theory predicted (Figs. 1 and 3), the same power requirements were found for both back and front illumination.

Figure 15 shows the dependence of the energy density, required to machine 6-μm-diameter spots with laser pulses of 25 ns duration, on the film thickness. The energy density plotted in Fig. 15 is the peak pulse power times the pulse duration divided by the spot area. For optimal machining, this is also equal to the peak power density I_o at the center of the beam times the pulse duration [from equations (10)

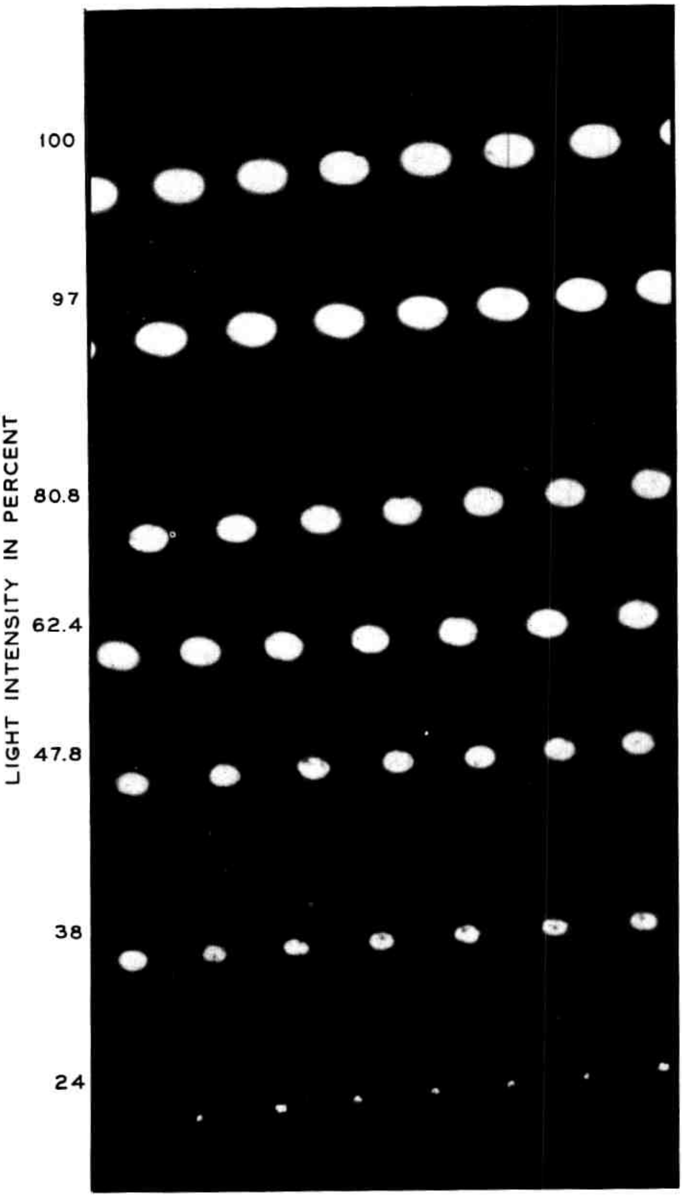


Fig. 11—Machining spots obtained with different light intensities. (The 100-percent value corresponds to a peak intensity of 1 watt with 25 ns pulse duration.)

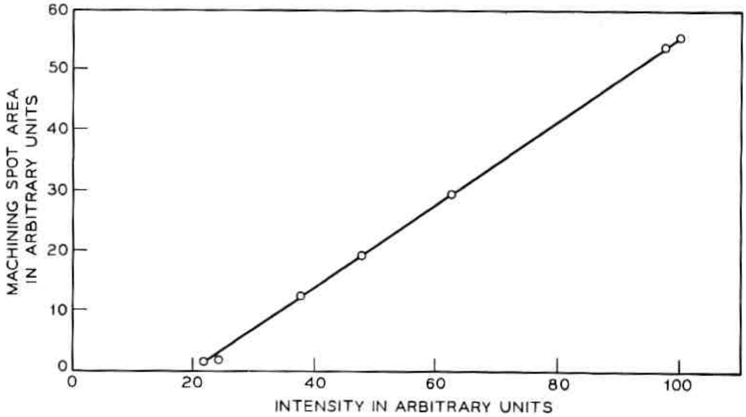


Fig. 12—Relation between machining spot area and illumination intensity.

and (13), $I_o = P/(\pi D^2/4)$]. The films were on mylar substrates and the machining was from the substrate side. As expected from the theoretical results presented in Figs. 1, 2, and 3, due to some loss of energy to the substrate, the required machining power increases relatively slowly with film thickness in the range from 200 to 1000 Å.

The power densities required to machine different thicknesses of bismuth films with different illumination duration can be estimated from Figs. 14 and 15. For the example given at the end of Section II, the peak power density required to machine a 1000-Å film with a 160-ns illumination duration, according to Fig. 14, is 4.5 times smaller than the peak power density with a 25-ns illumination duration. From Fig. 15 it is found that the energy density required to machine a 1000-Å film with a 25-ns pulse duration is 0.085 joule/cm². Assuming that due to reflectivity only 50 percent of the laser light is absorbed in the film, the peak power density I_o is then of the order of $0.5 \times 0.085/(4.5 \times 25 \times 10^{-9}) = 3.8 \times 10^5$ watts/cm². This is in agreement with the theoretical prediction presented in Fig. 6 and with the experimental results of Fig. 9 which indicate that only a small portion of the film is evaporated.

Back illumination machining experiments were carried out and compared for Mylar S* and Mylar D* substrates. Both of these materials have similar thermal constants. The Mylar S however has a better optical quality. The same power densities were required for machining in the two cases. As experimentally observed, due to the poorer optical

* Trademarks of DuPont.

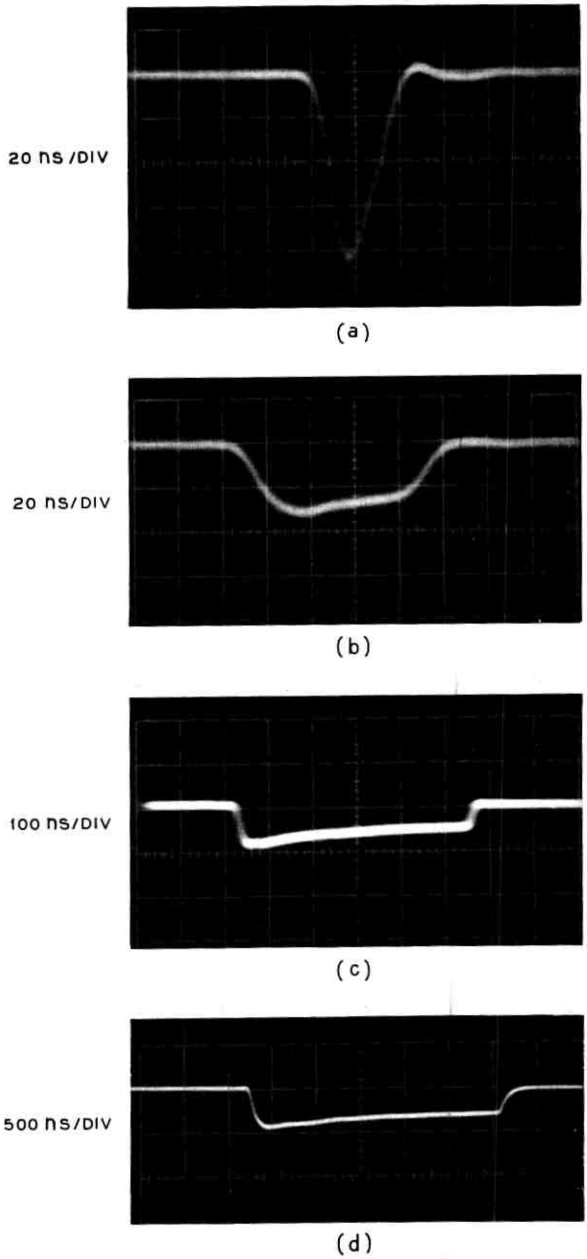


Fig. 13—Different light pulse durations obtained by the modulation system.

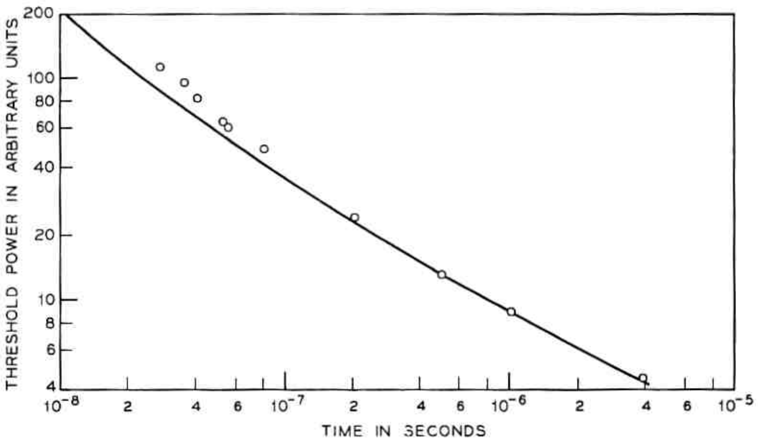


Fig. 14—Relation between threshold machining powers and illumination duration.

quality of the Mylar D, the optical beam shape was distorted causing a distortion in the shape of the machining. In Fig. 16a, b, and c, machining spot shapes are shown and compared for front and back illumination on Mylar D substrates and back illumination on Mylar S. As seen in Fig. 16a, the spot shape is distorted for the back illumination on Mylar D.

Many bismuth particles in the vicinity of the machining area can be observed in Fig. 16. These apparently were displaced as molten drops by the vapor pressure during machining.

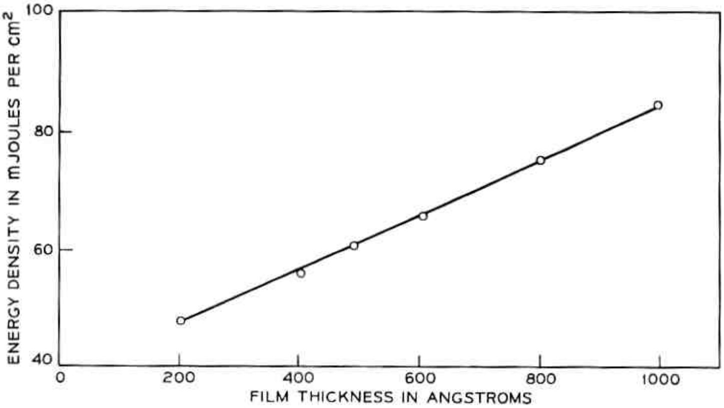


Fig. 15—Dependence of the energy densities, required to machine 6-μm-diameter spots, on the bismuth film thickness.

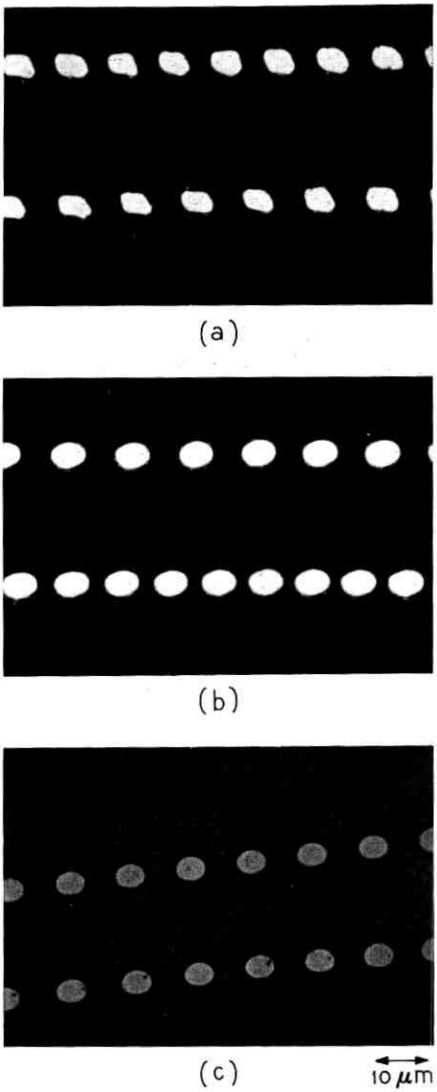


Fig. 16—Machining spot shapes with different substrates for front and back illumination: (a) Bismuth on Mylar D, back illumination. (b) Bismuth on Mylar D, front illumination. (c) Bismuth on Mylar S, back illumination.

IV. IMAGE RECORDING ON THIN BISMUTH FILM

The use of laser beams to record information on thin metallic film has been demonstrated by several authors.[1,2,12,13] C. O. Carlson, et al.[1] used a 38-mW CW He-Ne laser to write video information on various kinds of thin films on a glass substrate. The thin films used were 500 Å of lead and tantalum and 1μm of a triphenylmethane. A negative image of a page was formed by micromachining the thin film with the laser beam which was modulated at a video frequency. Machining lines had a width of 2μm with a total image size of 0.24 × 0.24 mm. The writing speed was equivalent to about 10^6 spots per second. The energy density required was 1.2 joules/cm^2.

A 100-Å-thick bismuth film on a glass substrate has been utilized to record holograms of various patterns with a pulsed ruby laser.[2]

In the present work, an intracavity modulated argon laser of average power capability, 20 mW, has been utilized to record images on a 600-Å-thick bismuth film on a mylar substrate. The recorded image size was 10 by 12 mm and the writing speed was 10^6 spots per second. Images of documents which were originally 8.5 × 11 inches have a resolution of over 175 lines/inch when magnified to the original size. The picture quality obtained is comparable with a good black and white photograph with 8 to 10 shades of gray. The energy density (given by the pulse peak energy divided by the spot area) had a maximum value of about 0.060 joule/cm^2.

The system is illustrated in Fig. 17. An acousto-optical intracavity modulation system was used to obtain 25-ns-duration light pulses from a CW argon laser producing 20 mW average power in the Gaussian mode. The modulation technique preserved the average output power of the laser. The video information was acquired by a He-Ne laser flying-spot scanner. A balanced mixer amplitude modulated a 5-mW 450-MHz RF signal with the video signal. By supplying a dc bias in addition to the video signal, the modulation depth could be adjusted to obtain the proper gray scale. A second balanced mixer, triggered with 30-ns-duration base band pulses at a 1-MHz repetition rate, was used to obtain the RF pulses required to dump the cavity. The video-modulated 30-ns-duration RF pulses, after being amplified by a 20-watt linear amplifier, were fed directly into the ZnO transducer sputtered on the fused silica modulator. As a result of a Bragg interaction between the acoustic pulses and the light beam, light pulses with the duration of 25–30 ns were dumped out of the cavity at a 1-MHz repetition rate. The pulses were intensity modulated at the video frequency. The light

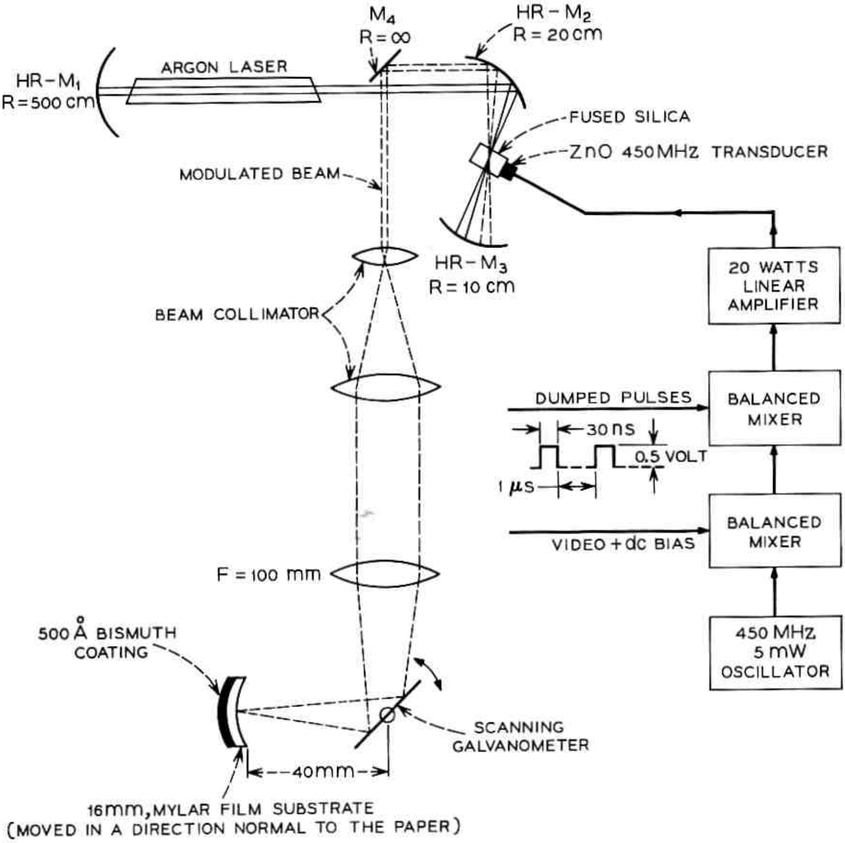


Fig. 17—The system setup for recording video information by laser machining.

pulses after being collimated were focused with an f/5.6 lens onto the bismuth film so as to machine 6-μm-diameter spots. The light pulses were deflected in the horizontal direction by a galvanometer mirror. The 16-mm mylar substrate with the bismuth coating moved at a constant speed in the vertical direction. To prevent the bismuth vapor from coating the optical components, the laser beam was incident on the thin film through the mylar substrate. The writing speed was 2 ms per line including 40-percent retrace time. The number of lines printed was about 2000 lines per frame (4-second frame with 2.4×10^6 total number of resolvable spots). Figure 18 shows a photograph of a micromachined copy of an IEEE facsimile test chart. As seen in this figure, about 8 to 10 shades of gray were obtained. A magnified

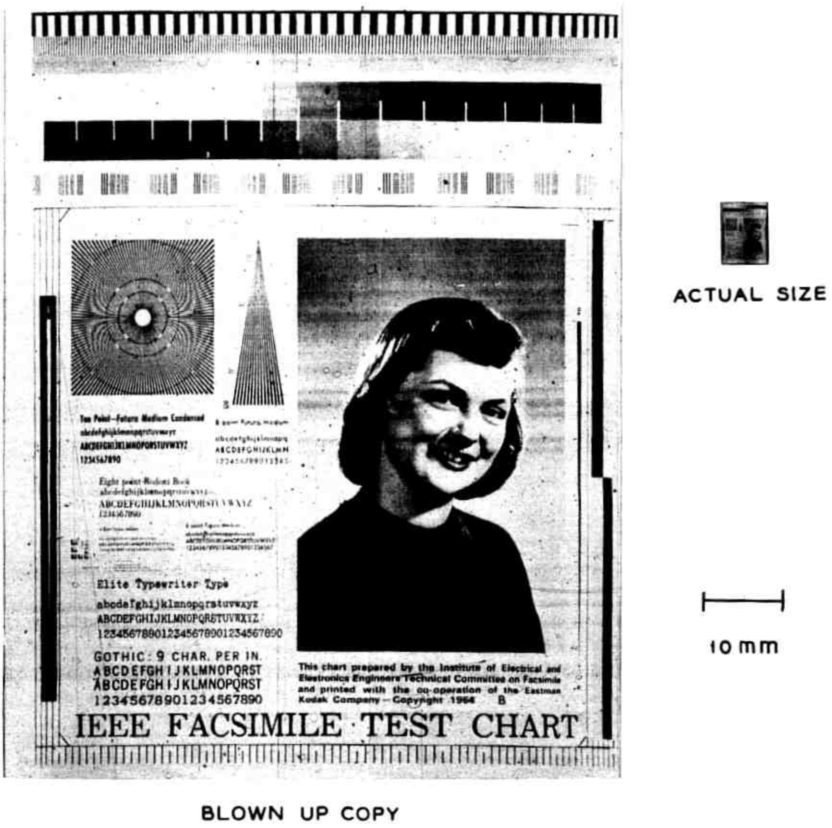

ACTUAL SIZE
10 mm
BLOWN UP COPY
IEEE FACSIMILE TEST CHART
Elite Typewriter Type
GOTHIC: 9 CHAR. PER IN.
This chart prepared by the Institute of Electrical and Electronics Engineers Technical Committee on Facsimile and printed with the co-operation of the Eastman Kodak Company—Copyright 1964 B

Fig. 18—Microfilm recording of a facsimile test chart.

region of an all-white area is shown in Fig. 19. A region of several gray scales (a detail of the eye in Fig. 18) is shown in Fig. 20.

By changing the polarity of the video signal (Fig. 17) either positive or negative images could be recorded. In Fig. 21, magnification of both positive and negative images of newspaper clippings are shown.

V. CONCLUSION

Calculations and experiments have been carried out for the machining of thin metallic films by a laser beam and for the associated temperature rise of the film. For a long illumination duration t, depending on the thermal constants of the film and substrate and on the film thickness, the temperature to which the film has risen by the end of the light pulse is proportional to $\sqrt{t}$. For a very short illumination duration,

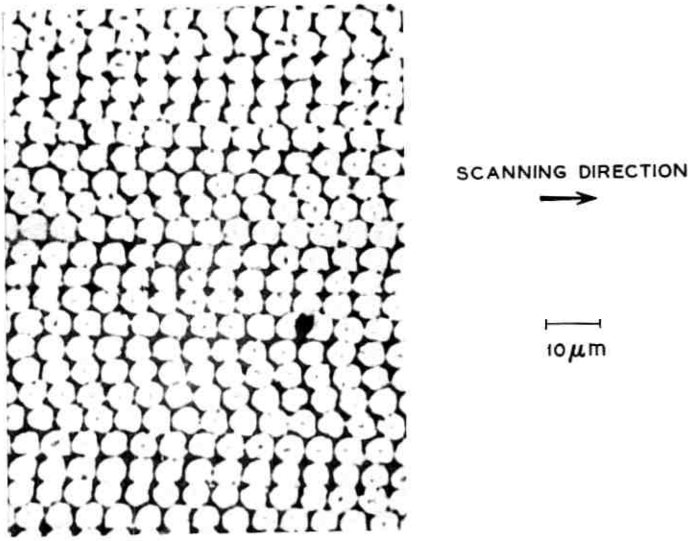


Fig. 19—1000 times magnification of a white area.

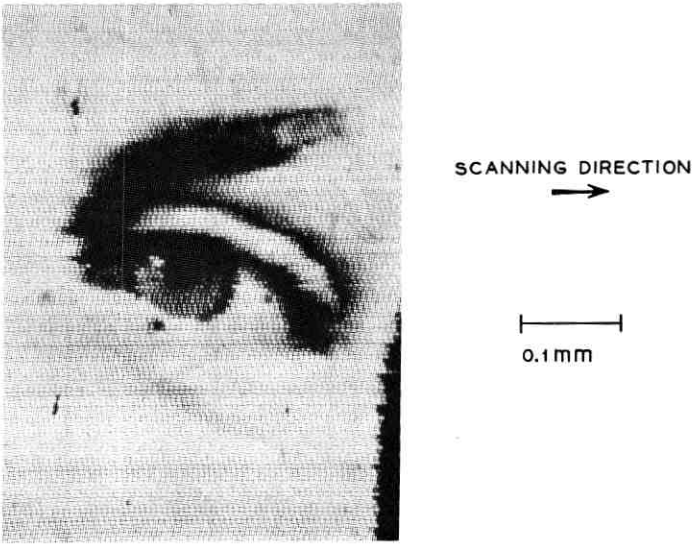
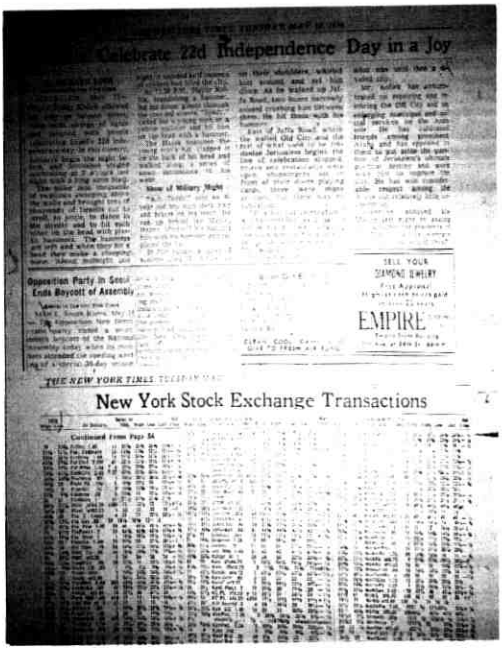
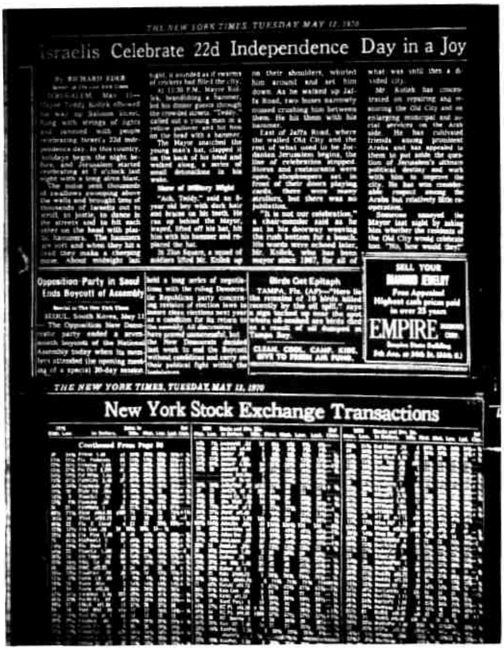


Fig. 20—Magnification of a gray scale area.


Celebrate 22d Independence Day in a Joy
Opposition Party in Seoul Ends Boycott of Assembly
SELL YOUR DIAMOND JEWELRY
EMPIRE
THE NEW YORK TIMES, TUESDAY, MAY
New York Stock Exchange Transactions

MAGNIFICATION
OF POSITIVE
MICROFILM IMAGE

1 mm


THE NEW YORK TIMES, TUESDAY, MAY 12, 1970
Israelis Celebrate 22d Independence Day in a Joy
Opposition Party in Seoul Ends Boycott of Assembly
Birds Get Epitaph
SELL YOUR DIAMOND JEWELRY
EMPIRE
THE NEW YORK TIMES, TUESDAY, MAY 12, 1970
New York Stock Exchange Transactions

MAGNIFICATION
OF NEGATIVE
MICROFILM IMAGE

Fig. 21—Negative and positive microfilm recording of a newspaper.

however, the temperature rise approaches a linear dependence on time. For longer light pulses, a larger portion of the laser energy is lost to the substrate by heat conduction during the pulse. Machining of thin films on substrates of very low thermal conductivity is most efficiently obtained when the pulse duration is short enough that the temperature rise is linearly proportional to the pulse duration.

With optically transmitting substrates, machining could be obtained either by front illumination or by back illumination. The machining efficiency is approximately equal for the two methods. In many cases, however, back illumination is advantageous since the vapor from the surface is directed away from the focusing and deflecting optical components. By varying the laser pulse intensity, the machining spot area varies nearly linearly with the pulse intensity over a certain range above some threshold value. A better understanding of the dynamics of the machining process, however, is still required and further work is planned.

The use of very short laser pulses permits the machining of images of a given size and resolution in a given length of time with much lower average laser power than would be required with a CW laser. By varying the intensity of the laser pulses and hence the area of the machined spots, images with a continuous range of shades of gray were obtained. Images of printed material having a resolution of about 175 lines per inch of an original $8\frac{1}{2} \times 11$ inch document were machined.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge the expert collaboration of R. Sard in obtaining the electron microscope pictures of machined spots. He wishes also to acknowledge the invaluable help of R. G. Chemelli in the laboratory investigation of laser machining and of C. W. Van Hise in preparing the thin films and the acoustooptic transducers, and to express appreciation to W. J. Bertram and R. L. Field for the use of the laser scanner. Acknowledgments go to M. Feldman for many useful discussions, and to M. I. Cohen and R. E. Kerwin, who contributed some of the initial concepts which led to this work. The author also wishes to thank A. J. Chick for the numerical calculations, and R. C. Miller, H. A. Watson, and E. I. Gordon for encouragement and valuable discussions.

REFERENCES

1. Carlson, C. O., et al., "He-Ne Laser: Thermal High-Resolution Recording," Science, *154* (1966), pp. 1550–1551.
2. Harris, A. L., et al., "CW Laser Recording on Metallic Thin Film," Image Technology, *12*, No. 3 (April–May 1970), pp. 31–35.
3. Maydan, D., "Fast Modulator for Extraction of Internal Laser Power," J. Appl. Phys., *41*, No. 4 (March 1970), pp. 1552–1559.
4. Maydan, D., and Chesler, R. B., "Q-Switching and Cavity Dumping of Nd:YA1G Lasers," J. Appl. Phys., *42*, No. 3 (March 1971), pp. 1031–1034.
5. Ready, J. F., "Effects Due to Absorption of Laser Radiation," J. Appl. Phys., *36*, No. 2 (February 1965), pp. 462–468.
6. Cohen, M. I., Unger, B. A., and Milkosky, J. F., "Laser Machining of Thin Films and Integrated Circuits," B.S.T.J., *47*, No. 3 (March 1968), pp. 385–405.
7. Cohen, M. I., and Epperson, J. P., *Application of Lasers to Microelectronic Fabrication, Electron Beam and Laser Beam Technology*, New York: Academic Press, 1968, pp. 139–185.
8. Carslaw, H. S., and Jaeger, J. C., *Conduction of Heat in Solids*, 2nd Edition, New York: Oxford University Press, 1959, p. 10.
9. Ibid., p. 75.
10. Kogelnik, H., and Li, T., "Laser Beams and Resonators," Proc. IEEE, *54* (1966), pp. 1312–1329.
11. Feldman, M., private communication.
12. Amodei, J. J., and Mezrich, R. S., "Holograms in Thin Bismuth Films," Appl. Phys. Letters, *15*, No. 2 (15 July 1969), pp. 45–46.
13. Hamisch, H., "Aufzeichnung und Speicherung von Informationen mit Laserstrahlen," Clearinghouse, U.S. Department of Commerce, N 69-34646 (May 1969).

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

The Coupling of Degenerate Modes in Two Parallel Dielectric Waveguides

By D. MARCUSE

(Manuscript received February 10, 1971)

Two parallel dielectric waveguides can exchange energy if the field carried by one guide reaches the other guide. We consider only the case of coupling between degenerate modes of dielectric waveguides (optical fibers, etc). Degenerate modes have equal phase velocities, but their transverse field distributions need not be identical.

The coupling theory presented in this paper applies to dielectric waveguides of arbitrary shape and arbitrary distribution of refractive index. The dielectric media of the guides as well as the surrounding medium are allowed to be lossy. The coupling coefficient is obtained by means of perturbation theory. It is shown that whereas lossless degenerate modes can exchange their power completely, lossy modes tend to equalize their power.

The theory is applied to the problem of crosstalk between cladded dielectric slab waveguides and cladded optical fibers embedded in a lossy medium.

Since a lossy surrounding medium also causes an increase in the loss of the guided modes, formulas for this additional loss are presented. It is shown that additional mode loss results even for a lossless surrounding medium if $n_3k > \beta$. (n_3 = index of surrounding medium, k = free space propagation constant, β = propagation constant of guided mode.)

I. INTRODUCTION

Optical fibers appear attractive as waveguides for the transmission of light. Since many such fibers are likely to be bunched together to form a cable, the problem of crosstalk between the different fibers of the cable arises. There are several sources of crosstalk. Light scattered in one fiber may excite a guided mode of an adjacent fiber. This double scattering mechanism is discussed in the next paper in this issue.

The present paper is concerned with the direct coupling of power between two parallel dielectric waveguides. Even though it addresses itself to crosstalk in optical fibers the theory is more general and applies to dielectric waveguides of arbitrary distribution of the refractive index. The modes carried by the two guides need not be identical. However, we assume that the phase velocities of the modes in either guide are identical. We call modes with identical phase velocity degenerate modes. We furthermore admit the possibility that the media of the waveguides as well as the surrounding medium in which the guides are embedded may be lossy. The restriction to degenerate modes is no serious limitation for cases of practical interest. Provided that the coupling coefficients are constant, appreciable coupling of power from a mode in one guide to a mode in the other guide is possible only if the modes have identical phase velocities.

The coupling theory is based on finding the change of the propagation constant of the mode of one guide caused by the presence of the other guide. This perturbation theory is directly based on Maxwell's equations and is thus independent of the composition and geometry of the coupled waveguides.

We obtain the coupling coefficient in form of an integral over the cross section of the waveguides. The integral is taken over the scalar product of the electric field vectors of the two guides. The general coupling formula is then used to obtain the coupling between two cladded slab waveguides as well as the coupling between two cladded round optical fibers. The crosstalk between these two types of waveguides is discussed. The crosstalk can be reduced by increasing the separation between the two guides and by increasing the loss of the medium in which the guides are embedded. Coupling between dielectric fibers has previously been treated by A. L. Jones[1] and by R. Vanclooster and P. Phariseau.[2] However, these authors consider only uncladded lossless fibers.

II. COUPLED LINE EQUATIONS

The general properties of coupled waveguides can be studied with the help of coupled line equations. Neglecting the possibility of coupling to modes traveling in the opposite direction we can write the general coupled line equations of two degenerate modes as follows:[3]

$$\frac{dA}{dz} = -i\beta A + c_1 B, \tag{1a}$$

$$\frac{dB}{dz} = -i\beta B + c_2 A. \tag{1b}$$

The propagation constant β is assumed to be the same for either mode; A is the amplitude of the mode of one guide and B that of the mode of the other guide. If each guide can support more than one mode these modes would have to be included in the coupled line equations. However, it is well known that only modes with equal phase velocity exchange a significant amount of energy if they are coupled together by a length-independent coupling mechanism. The restriction to only two modes is thus an excellent approximation. The coupling coefficients c_1 and c_2 are allowed to be different since the two modes can be different even though they are degenerate. However, if the waveguides are lossless, conservation of power imposes the condition[3]

$$c_1 = -c_2^*. \tag{2}$$

The asterisk indicates complex conjugation.

In the following analysis we assume that β as well as c_1 and c_2 may be complex quantities since we want to include the case of lossy modes. Equation (2) is thus not assumed to hold exactly. Since β as well as c_1 and c_2 are constants we can immediately solve the coupled line equations and obtain solutions in the form

$$A(z) = \frac{1}{2}\left\{A_0(e^{-i\Delta\beta z} + e^{i\Delta\beta z}) + \left(\frac{c_1}{c_2}\right)^{\frac{1}{2}} B_0(e^{-i\Delta\beta z} - e^{i\Delta\beta z})\right\}e^{-i\beta z}, \tag{3a}$$

$$B(z) = \frac{1}{2}\left\{B_0(e^{-i\Delta\beta z} + e^{i\Delta\beta z}) + \left(\frac{c_2}{c_1}\right)^{\frac{1}{2}} A_0(e^{-i\Delta\beta z} - e^{i\Delta\beta z})\right\}e^{-i\beta z}, \tag{3b}$$

with

$$\Delta\beta = i\sqrt{c_1 c_2}\ . \tag{4}$$

The coefficients A_0 and B_0 are the field amplitudes A and B at $z = 0$.

It is apparent that equations (3a) and (3b) are composed of the superposition of two new normal modes with propagation constants

$$\beta_1 = \beta + \Delta\beta \tag{5}$$

and

$$\beta_2 = \beta - \Delta\beta. \tag{6}$$

For complex values of c_1 and c_2 we obtain also a complex value of the change of the propagation constant $\Delta\beta$. In that case, it is clear that the two new normal modes do not only have slightly different propagation velocities but also suffer different amounts of loss.

For the study of crosstalk we need the solution with $B_0 = 0$. That means we assume that at $z = 0$ only mode A is excited. If we restrict ourselves to the case

$$| \Delta\beta z | \ll 1 \tag{7}$$

we find that the ratio of the amplitude B to amplitude A for distances $z = L$ for which (7) is valid is given by

$$\frac{B(L)}{A(L)} = -i\left(\frac{c_2}{c_1}\right)^{\frac{1}{2}} \Delta\beta L. \tag{8}$$

The absolute square of the ratio of B/A is the power ratio at $z = L$ in the two guides. Using the letter V for this power ratio we obtain the following useful formula for the distance L at which the power ratio in the two guides is V:

$$L = \left| \frac{c_1}{c_2} \right|^{\frac{1}{2}} \frac{(V)^{\frac{1}{2}}}{| \Delta\beta |}. \tag{9}$$

For low-loss modes equation (2) must hold at least approximately so that we obtain from (9) in the low-loss case

$$L = \frac{V^{\frac{1}{2}}}{| \Delta\beta |}. \tag{10}$$

The low-loss assumption refers only to the actual loss suffered by the modes propagating inside of the guides. The loss of the surrounding medium in which the guides are embedded can be arbitrarily high.

Finally we consider two special cases. If we assume that the two guides are lossless and assume again $B_0 = 0$ we obtain from equation (3)

$$A(z) = A_0 \cos (\Delta\beta z)e^{-i\beta z}, \tag{11a}$$

$$B(z) = -i\left(\frac{c_2}{c_1}\right)^{\frac{1}{2}} A_0 \sin (\Delta\beta z)e^{-i\beta z}. \tag{11b}$$

In the lossless case assumed here $\Delta\beta$ is real and it is clear that the two guides exchange power after a distance D given by

$$D = \frac{\pi}{2\Delta\beta}. \tag{12}$$

If, on the other hand, $\Delta\beta$ is complex, as it would be in the case of lossy guides, and if we make z sufficiently large that the exponential factor $\exp(i\Delta\beta z)$ is much less than unity, we obtain again for $B_0 = 0$ from equation (3)

$$A(z) = \tfrac{1}{2}A_0 e^{-i(\beta+\Delta\beta)z} \tag{13a}$$

and

$$B(z) = \frac{1}{2}\sqrt{\frac{c_2}{c_1}}\, A_0 e^{-i(\beta+\Delta\beta)z}. \tag{13b}$$

Since the absolute value of the ratio c_2/c_1 is nearly unity we see that the power carried by the two guides equalizes in the lossy case provided that we let both modes travel far enough. This power equalization may occur after many power exchange cycles or it may occur even before one cycle is completed depending on the magnitude of the imaginary part of $\Delta\beta$. Of course both modes suffer additional loss because of the complex value of the propagation constant β. The magnitude of β is much larger than that of $\Delta\beta$ for this discussion to be meaningful.

The preceding discussion shows that the crosstalk between the two coupled guides is determined by the change in propagation constant $\Delta\beta$ and not by the individual coupling constants c_1 and c_2. It is thus sufficient to determine $\Delta\beta$.

III. MODE COUPLING THEORY

We assume that two dielectric waveguides are positioned parallel to each other. Each guide is a cylindrical structure with a refractive index distribution that is independent of the z coordinate (z is the direction parallel to the axis of the structure) but which can have an arbitrary dependence on the transverse x and y coordinates. However, it is assumed that the index distribution is such that it produces a dielectric waveguide capable of supporting guided modes. We assign a refractive index distribution n_1 to guide one assuming that n_1 has the constant value of the refractive index n_3 of the background material in the region occupied by guide two. A similar assumption applies to the index distribution of guide two. It is assumed to have the constant value n_3 of the background in the region occupied by guide one. The square of the index distributions n_1 and n_2 is shown in Fig. 1. Since it is the square of the refractive index that enters Maxwell's equations we use the following expression

$$n^2 = (n_1^2 - n_3^2) + (n_2^2 - n_3^2) + n_3^2 . \tag{14}$$

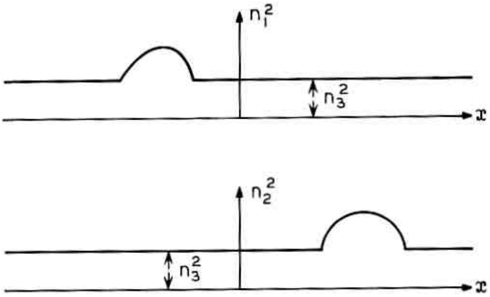


Fig. 1—Distribution of the square of the refractive index used to describe the two dielectric waveguides.

Owing to the definitions of $n_1(x, y)$ and $n_2(x, y)$ it is clear that (14) becomes $n^2 = n_1^2$ in the region of guide one, $n^2 = n_2^2$ in the region of guide two, and $n^2 = n_3^2$ outside the regions of either guide. The refractive indices n_1, n_2, and n_3 need not be real quantities but are allowed to be complex if the media are lossy.

The starting point of the perturbation theory is Maxwell's equations

$$\nabla \times \mathbf{H} = i\omega\epsilon_0 n^2 \mathbf{E} \tag{15a}$$

and

$$\nabla \times \mathbf{E} = -i\omega\mu_0 \mathbf{H}. \tag{15b}$$

The time dependence of the fields was assumed to be given as exp $(i\omega t)$. The constants ϵ_0 and μ_0 are the permittivity and permeability of the vacuum and the square of the refractive index is given by (14). Next we introduce the perturbation assumption that the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ are very nearly given by the superposition of the mode fields of each guide in the absence of the other. We thus write

$$\mathbf{E} = a\mathbf{E}_1 + b\mathbf{E}_2 + \mathbf{e} \tag{16}$$

and

$$\mathbf{H} = a\mathbf{H}_1 + b\mathbf{H}_2 + \mathbf{h}. \tag{17}$$

The coefficients a and b are being determined by the theory. The field $\mathbf{E}_\nu$, $\mathbf{H}_\nu (\nu = 1, 2)$ being a mode of guide ν has the z dependence exp $(-i\beta_0 z)$. It is assumed that the propagation constant is the same for the modes of either one of the guides in the absence of the other. As indicated in Section I, such modes are here called degenerate. The electric and magnetic fields of the unperturbed modes of either guide satisfy the equations

$$\nabla_t \times \mathbf{H}_\nu - i\beta_0(\hat{z} \times \mathbf{H}_\nu) - i\omega\epsilon_0 n_\nu^2 \mathbf{E}_\nu = 0 \quad \nu = 1, 2 \tag{18a}$$

and

$$\nabla_t \times \mathbf{E}_\nu - i\beta_0(\hat{z} \times \mathbf{E}_\nu) + i\omega\mu_0 \mathbf{H}_\nu = 0. \tag{18b}$$

$\hat{z}$ is a unit vector in z direction. The symbol ∇_t indicates the transverse part of the ∇ operator. The electric vector **e** and the magnetic vector **h** appearing in (16) and (17) are assumed to be small perturbations that are needed to express the change in the field shape resulting from the close proximity of the two guides. In addition to the changes in the field shape we must allow for the possibility of a change in the propagation constant which can be assumed to be small if the coupling of the two guides is sufficiently loose. We thus assume that the new fields **E** and **H** of (16) and (17) have a z dependence of the form $\exp(-i\beta z)$ with

$$\beta = \beta_0 + \Delta\beta. \tag{19}$$

We have seen in the last section that the change $\Delta\beta$ of the propagation constant determines the coupling of the two guides. It is thus the quantity we are most interested in.

We substitute (16) and (17) into (15a) and (15b) using (14). After cancellation of a number of terms with the help of (18a) and (18b) we are left with the following system of equations:

$$\begin{aligned}\nabla_t \times \mathbf{h} - i\beta_0(\hat{z} \times \mathbf{h}) - i\omega\epsilon_0[(n_1^2 - n_3^2) + (n_2^2 - n_3^2) + n_3^2]\mathbf{e} \\ = i\Delta\beta(a\hat{z} \times \mathbf{H}_1 + b\hat{z} \times \mathbf{H}_2) \\ + i\omega\epsilon_0[(n_2^2 - n_3^2)a\mathbf{E}_1 + (n_1^2 - n_3^2)b\mathbf{E}_2]\end{aligned} \tag{20}$$

and

$$\nabla_t \times \mathbf{e} - i\beta_0(\hat{z} \times \mathbf{e}) + i\omega\mu_0\mathbf{h} = i\Delta\beta(a\hat{z} \times \mathbf{E}_1 + b\hat{z} \times \mathbf{E}_2). \tag{21}$$

Products of $\Delta\beta$ with **e** and **h** have been omitted from (20) and (21) since they are small of second order.

Since we are interested in obtaining $\Delta\beta$ from our theory we must eliminate the dependence on the spatial coordinates from (20) and (21). Concentrating for the moment on the left-hand sides of these equations we obtain by scalar multiplication of (20) with $\mathbf{E}_1$ and of (21) with $\mathbf{H}_1$ and by subtracting the two equations and integration over the entire cross-sectional area

$$\int \{\mathbf{E}_1\cdot[\nabla_t \times \mathbf{h} - i\beta_0(\hat{z}\times\mathbf{h}) - i\omega\epsilon_0((n_1^2 - n_3^2) + (n_2^2 - n_3^2) + n_3^2)\mathbf{e}] \\ - \mathbf{H}_1\cdot[\nabla_t\times\mathbf{e} - i\beta_0(\hat{z}\times\mathbf{e}) + i\omega\mu_0\mathbf{h}]\}\,dx\,dy. \qquad (22)$$

The terms inside of the square brackets are small of first order. We now assume that the field intensity of $\mathbf{E}_1$ in the region of guide two has decayed sufficiently to render it also small of first order. Since $n_2^2 - n_3^2$ is different from zero only in a small area in the vicinity of guide two the product of $\mathbf{E}_1$ with $(n_2^2 - n_3^2)\mathbf{e}$ is a term that is small of second order. In first order of perturbation theory this term can be neglected. Performing a partial integration on the terms containing the ∇_t operator and rearranging the terms of the mixed products allows us to write

$$\int \{\mathbf{h}\cdot[\nabla_t\times\mathbf{E}_1 + i\beta_0(\hat{z}\times\mathbf{E}_1) - i\omega\mu_0\mathbf{H}_1] \\ - \mathbf{e}\cdot[\nabla_t\times\mathbf{H}_1 + i\beta_0(\hat{z}\times\mathbf{H}_1) + i\omega\epsilon_0 n_1^2\mathbf{E}_1]\}\,dx\,dy. \qquad (23)$$

Comparison of (23) with (18) shows that this expression would vanish if instead of $\mathbf{E}_1$ and $\mathbf{H}_1$ we had used the complementary fields $\mathbf{E}_1^-$ and $\mathbf{H}_1^-$ that result from the original field if β_0 and ω are replaced by $-\beta_0$ and $-\omega$. We have thus shown that by scalar multiplication with $\mathbf{E}_1^-$ and $\mathbf{H}_1^-$, subtraction, and integration over the entire cross-sectional area the field terms $\mathbf{e}$ and $\mathbf{h}$ can be made to vanish from (20) and (21). A similar procedure can, of course, be carried through using $\mathbf{E}_2^-$ and $\mathbf{H}_2^-$. We thus obtain from (20) and (21) the following set of equations:

$$a\left[\Delta\beta\hat{z}\cdot\int(\mathbf{H}_1^-\times\mathbf{E}_1 + \mathbf{H}_1\times\mathbf{E}_1^-)\,dx\,dy \right. \\ \left. + \omega\epsilon_0\int(n_2^2 - n_3^2)\mathbf{E}_1^-\cdot\mathbf{E}_1\,dx\,dy\right] \\ + b\left[\Delta\beta\hat{z}\cdot\int(\mathbf{H}_1^-\times\mathbf{E}_2 + \mathbf{H}_2\times\mathbf{E}_1^-)\,dx\,dy \right. \\ \left. + \omega\epsilon_0\int(n_1^2 - n_3^2)\mathbf{E}_1^-\cdot\mathbf{E}_2\,dx\,dy\right] = 0 \qquad (24a)$$

and

$$a\left[\Delta\beta\hat{z}\cdot\int(\mathbf{H}_2^-\times\mathbf{E}_1 + \mathbf{H}_1\times\mathbf{E}_2^-)\,dx\,dy\right.$$

$$+ \omega\epsilon_0 \int (n_2^2 - n_3^2)\mathbf{E}_2^- \cdot \mathbf{E}_1 \, dx \, dy \Big]$$

$$+ b\Big[\Delta\beta\hat{z}\cdot \int (\mathbf{H}_2^- \times \mathbf{E}_2 + \mathbf{H}_2 \times \mathbf{E}_2^-) \, dx \, dy$$

$$+ \omega\epsilon_0 \int (n_1^2 - n_3^2)\mathbf{E}_2^- \cdot \mathbf{E}_2 \, dx \, dy \Big] = 0. \tag{24b}$$

Equations (24a) and (24b) still contain a number of second-order terms. Products of field terms with different indices are small of first order since the modes of the two guides are supposed to overlap only slightly. Products of field terms with different indices that are multiplied by the first-order quantity $\Delta\beta$ can thus be neglected as being of second order. The term $(n_2^2 - n_3^2)\mathbf{E}_1^- \cdot \mathbf{E}_1$ is also of second order since it involves the square of the amplitude of $\mathbf{E}_1$ at the location of the opposite guide. Neglecting it and a similar term results in a consistent equation system containing only first-order terms.

$$a\Delta\beta\hat{z}\cdot \int (\mathbf{H}_1^- \times \mathbf{E}_1 + \mathbf{H}_1 \times \mathbf{E}_1^-) \, dx \, dy$$

$$+ b\omega\epsilon_0 \int (n_1^2 - n_3^2)\mathbf{E}_1^- \cdot \mathbf{E}_2 \, dx \, dy = 0 \tag{25a}$$

and

$$a\omega\epsilon_0 \int (n_2^2 - n_3^2)\mathbf{E}_2^- \cdot \mathbf{E}_1 \, dx \, dy$$

$$+ b\Delta\beta\hat{z}\cdot \int (\mathbf{H}_2^- \times \mathbf{E}_2 + \mathbf{H}_2 \times \mathbf{E}_2^-) \, dx \, dy = 0. \tag{25b}$$

The equation system (25) allows us to determine the coefficients a and b. Since we have a homogeneous system of equations we must require that the system determinant vanishes. This latter condition leads to a determination of $\Delta\beta$.

$$\Delta\beta = \pm\omega\epsilon_0$$

$$\cdot\left\{\frac{\int (n_1^2 - n_3^2)\mathbf{E}_1^- \cdot \mathbf{E}_2 \, dx \, dy \int (n_2^2 - n_3^2)\mathbf{E}_2^- \cdot \mathbf{E}_1 \, dx \, dy}{\int \hat{z}\cdot(\mathbf{E}_1 \times \mathbf{H}_1^- + \mathbf{E}_1^- \times \mathbf{H}_1) \, dx \, dy \int \hat{z}\cdot(\mathbf{E}_2 \times \mathbf{H}_2^- + \mathbf{E}_2^- \times \mathbf{H}_2) \, dx \, dy}\right\}^{\frac{1}{2}}. \tag{26}$$

The ratio of a/b is of no interest to us so that we do not need to write down the solution of (25).

Equation (26) is the final result of our coupling theory. In the general case of lossy dielectric media $\Delta\beta$ is a complex quantity. Equation (26) is very general and holds for any two dielectric waveguides carrying degenerate modes. The field expressions $\mathbf{E}_i^-$ and $\mathbf{H}_i^-$ are obtained by changing the sign of β and ω in the regular expressions for $\mathbf{E}_i$ and $\mathbf{H}_i$. In the special case of lossless media $\mathbf{E}_i^-$ and $\mathbf{H}_i^-$ become simply the complex conjugates of the original fields.

$$\left.\begin{aligned}\mathbf{E}_i^- &= \mathbf{E}_i^*\\ \mathbf{H}_i^- &= \mathbf{H}_i^*\end{aligned}\right\} \text{for real } n_i\,. \tag{27}$$

Equation (26) can be simplified in another special case. If both waveguides are identical, carrying the same mode, the two types of integrals appearing in the numerator as well as in the denominator of (26) become identical for reasons of symmetry. The expression for identical modes of two coupled identical waveguides can thus be written more simply

$$\Delta\beta = \pm\omega\epsilon_0 \frac{\int (n_2^2 - n_3^2)\mathbf{E}_2^-\cdot\mathbf{E}_1\,dx\,dy}{\int \hat{z}\cdot(\mathbf{E}_1 \times \mathbf{H}_1^- + \mathbf{E}_1^- \times \mathbf{H}_1)\,dx\,dy}. \tag{28}$$

For real refractive indices the denominator of (28) is equal to $4P$, P being the power carried by the mode of one of the guides. It is necessary to normalize the mode fields of each guide so that field 1 as well as field 2 carry equal power P. This normalization was already implied in converting (26) to (28). Even if the medium surrounding the two waveguides is lossy it is possible to use the relations (27) and the identification of the denominator of (28) as $4P$ provided that the surrounding medium has a very small influence on the mode fields carried by each guide. It is thus permissible to use (28) in the form

$$\Delta\beta = \pm\frac{\omega\epsilon_0}{4P}\int (n_2^2 - n_3^2)\mathbf{E}_2^*\cdot\mathbf{E}_1\,dx\,dy \tag{29}$$

in the special case of lossless dielectric media or, at least approximately, even in the case that the surrounding medium is lossy if only the fields are so well guided that only a very small amount of power reaches the lossy surrounding medium.

IV. APPLICATION TO CLADDED SLAB WAVEGUIDE

The dielectric slab is the simplest type of dielectric waveguide. Because of its simplicity much insight into the performance of dielectric waveguides can be gained by studying the slab waveguide. We consider two identical cladded slab waveguides as shown in Fig. 2a. Instead of solving the mode problem of the cladded slab waveguide exactly, we assume that the cladding is thick enough so that only a small amount of power reaches the surrounding medium. This assumption allows us to obtain approximate field solutions that are much simpler than the exact solutions. This case is, moreover, of the most practical interest since the cladding is meant to protect the guided mode from the disturbing influence of the surroundings. Our approximation thus coincides with the case of real practical interest. Without going into the details of the calculation we state only the results. The field of one of the two guides can be expressed as follows: (For simplicity the field of one guide is expressed in the coordinate system shown in Fig. 2b.)

$$E_y = \begin{cases} A \cos \kappa x & 0 \leqq x \leqq d \\ Be^{\gamma x} + Ce^{-\gamma x} & d \leqq x \leqq D \\ Fe^{-\rho x} & D \leqq x < \infty . \end{cases} \tag{30}$$

The field is continued symmetrically for $x < 0$. Equation (30) gives the E component of the symmetric TE modes of the slab waveguide.

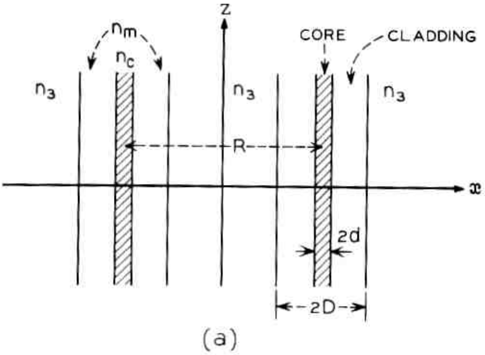


Fig. 2a—Sketch of two coupled cladded slab waveguides.

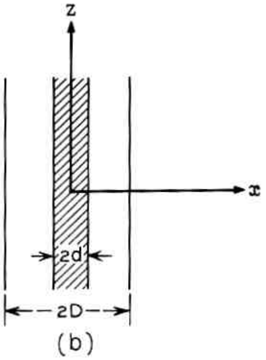


Fig. 2b—Sketch of a single slab waveguide used for equation (30).

The other field components are obtained from E_y by the relations

$$H_x = -\frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial z} \tag{31}$$

and

$$H_z = \frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial x}. \tag{32}$$

The dependence of the field on time t and on the z coordinate is, as always, given by the factor

$$e^{i(\omega t - \beta z)} \tag{33}$$

which is being omitted from the equations. The parameters appearing in (30) are defined as follows: ($k = \omega\sqrt{\epsilon_0\mu_0}$)

$$\kappa = (n_c^2k^2 - \beta^2)^{\frac{1}{2}}, \tag{34}$$

$$\gamma = (\beta^2 - n_m^2k^2)^{\frac{1}{2}}, \tag{35}$$

$$\rho = (\beta^2 - n_3^2k^2)^{\frac{1}{2}}. \tag{36}$$

n_c, n_m, and n_3 are the refractive indices of core, cladding, and outer medium. The boundary conditions lead to relations between the constants A, B, C, and F:

$$B = \frac{A}{2}\frac{\kappa}{(\kappa^2 + \gamma^2)^{\frac{1}{2}}}\frac{\gamma - \rho}{\gamma + \rho}e^{-\gamma d}e^{-2\gamma(D-d)}, \tag{37}$$

$$C = Ae^{\gamma d}\cos\kappa d, \tag{38}$$

$$F = \frac{2\kappa\gamma A}{(\gamma + \rho)(\kappa^2 + \gamma^2)^{\frac{1}{2}}} e^{\rho D} e^{-\gamma(D-d)}. \tag{39}$$

The coefficient A can be related to the power P carried by the mode,

$$A = \left[\frac{2\omega\mu P}{\beta d + \dfrac{\beta}{\gamma}}\right]^{\frac{1}{2}}. \tag{40}$$

The coefficient F depends exponentially on the factor $-\gamma(D - d)$. For increasing cladding thickness this factor becomes very small. However, the factor B vanishes even more rapidly since it depends on the square of the same exponential decay factor. This shows that the field inside of core and cladding is altered only very slightly if the cladding is sufficiently thick. The eigenvalue equation differs from the eigenvalue equation for infinite cladding only by a term that also vanishes as the square of the aforementioned exponential decay factor. To a good approximation we are thus justified to use the eigenvalue equation for the infinite cladding guide.

$$\tan \kappa d = \frac{\gamma}{\kappa}. \tag{41}$$

Once the mode fields are known it is an easy matter to evaluate the coupling coefficient (29). We obtain

$$\Delta\beta = \frac{4\kappa^2\gamma^3\rho}{\beta(1 + \gamma d)(\gamma + \rho)^2(\kappa^2 + \gamma^2)} e^{-2\gamma(D-d)} e^{-\rho(R-2D)}. \tag{42}$$

(R is the distance between the core centers of the two slab waveguides.) In case that the cladding material is identical to the surrounding medium we have $\rho = \gamma$ and (42) reduces to the expression for coupled slab waveguides to be found in N. S. Kapany's book.[4]

The same mechanism that is responsible for the coupling between the two slab waveguides also causes loss to the modes in each of the guides. Coupling occurs because some of the field in one guide reaches the region of the other guide. However, fields reaching out into the lossy medium between the guides also cause power loss to the mode traveling inside of the guide. Only if the surrounding medium is truly lossless can there be no additional loss to the guided mode. I calculated the mode loss contributed by the surrounding medium assuming that it is infinitely extended and also ignoring the presence of the other guide. The power loss can be obtained to a good approximation by computing the power flow in transverse direction to the guide axis at the boundary between the cladding and the surrounding medium. The validity of this approach hinges again on the fact that most of the power is carried inside of core and cladding. The approximate calculation leads to the following expression for the power loss 2α of the mode (α is the amplitude loss coefficient).

$$2\alpha = \frac{8\kappa^2\gamma^3 \operatorname{Im}(\rho)e^{-2\gamma(D-d)}}{\beta(1+\gamma d)(\kappa^2+\gamma^2)\mid\gamma+\rho\mid^2}. \tag{43}$$

It is assumed that n_c and n_m are both real; n_3 is allowed to assume complex values. This loss expression has two interesting features. It shows that the loss decreases exponentially with increasing cladding thickness. It also shows that the loss is dependent only on the imaginary part $\operatorname{Im}(\rho)$ of ρ. For ρ to be real, two conditions must be satisfied. First of all, the refractive index of the surrounding medium, n_3, must be purely real. This means that the mode suffers loss whenever the surrounding medium is lossy. However, even if the surrounding medium is lossless, it is still possible for ρ to be purely imaginary. This happens when $n_3k > \beta$. In this case there is no total internal reflection between the cladding and the surrounding medium and the evanescent field tail reaching into the surrounding medium does no longer decay exponentially but begins to radiate. Power is thus lost from the guided mode even though all the materials of the guide and the surrounding medium are lossless.

Finally, we state the result of the coupling calculation and mode loss for the cladded slab waveguide carrying the symmetric TM mode. For the details of the field distribution of this mode in an infinite cladding guide see Ref. 5. The result of the evaluation of (29) in this case is

$$\Delta\beta = \frac{n_c^2n_m^2\kappa^2\gamma^2[2e^{(\gamma-\rho)(D-d)}-1]e^{-2\gamma(D-d)}e^{-\rho(R-2D)}}{\beta[(n_m^4\kappa^2+n_c^4\gamma^2)\gamma d+n_c^2n_m^2(\kappa^2+\gamma^2)]}. \tag{44a}$$

The power loss caused by the lossy surrounding medium is

$$2\alpha = \frac{8n_c^2n_m^4\mid n_3\mid^4\kappa^2\gamma^3\operatorname{Im}\left(\dfrac{\rho}{n_3^2}\right)e^{-2\gamma(D-d)}}{\beta[\gamma d(n_m^4\kappa^2+n_c^4\gamma^2)+n_c^2n_m^2(\kappa^2+\gamma^2)]\mid n_m^2\rho+n_3^2\gamma\mid^2}. \tag{44b}$$

It is apparent that the coupling expression (42) for the TE mode differs from that for the TM mode. The coupling as well as the loss, thus turn out to be polarization dependent.

V. COUPLING OF THE HE$_{11}$ MODES OF ROUND CLADDED FIBERS

In complete analogy to the cladded slab waveguide we now proceed to a discussion of coupling between HE$_{11}$ modes in two round cladded

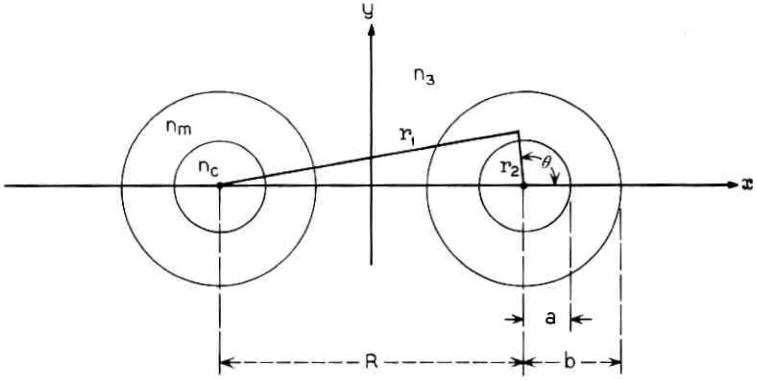


Fig. 3—Cross section of two coupled cladded fibers.

fibers, Fig. 3. Again, we assume that the cladding is sufficiently thick to allow us to treat the field inside of core and cladding as being affected only very little by the presence of the surrounding medium. With this approximation in mind, we state the relevant equations for the mode field of a cladded fiber.

The field inside of the fiber core is described by the set of equations[6]

$$\left.\begin{aligned} E_z &= AJ_1(\kappa r)\cos\phi \\ H_z &= BJ_1(\kappa r)\sin\phi \end{aligned}\right\} \qquad 0 \leqq r \leqq a, \tag{45}$$

the field inside of the cladding is given by

$$\left.\begin{aligned} E_z &= [CH_1^{(1)}(i\gamma r) + DH_1^{(2)}(i\gamma r)]\cos\phi \\ H_z &= [FH_1^{(1)}(i\gamma r) + GH_1^{(2)}(i\gamma r)]\sin\phi \end{aligned}\right\} \qquad a \leqq r \leqq b, \tag{46}$$

and the field in the surrounding medium is

$$\left.\begin{aligned} E_z &= MH_1^{(1)}(i\rho r)\cos\phi \\ H_z &= NH_1^{(1)}(i\rho r)\sin\phi \end{aligned}\right\} \qquad b \leqq r < \infty. \tag{47}$$

The parameters κ, γ and ρ are given by the equations (34), (35), and (36). The other field components are obtained from the longitudinal components by differentiation

$$E_r = -\frac{i}{\Gamma^2}\left(\beta\frac{\partial E_z}{\partial r} + \omega\mu_0\frac{1}{r}\frac{\partial H_z}{\partial\phi}\right), \tag{48}$$

$$E_\phi = -\frac{i}{\Gamma^2}\left(\beta\frac{1}{r}\frac{\partial E_z}{\partial\phi} - \omega\mu_0\frac{\partial H_z}{\partial r}\right), \tag{49}$$

$$H_r = -\frac{i}{\Gamma^2}\left(\beta\,\frac{\partial H_z}{\partial r} - \omega\epsilon_0\,\frac{n^2}{r}\,\frac{\partial E_z}{\partial\phi}\right), \tag{50}$$

$$H_\phi = -\frac{i}{\Gamma^2}\left(\beta\,\frac{1}{r}\,\frac{\partial H_z}{\partial\phi} + \omega\epsilon_0 n^2\,\frac{\partial E_z}{\partial r}\right). \tag{51}$$

The symbol Γ stands either for κ, $i\gamma$, or $i\rho$ depending on the medium in which the field is evaluated. The same is true for n, it represents either n_c, n_m, or n_3. The boundary conditions establish the following relations between the amplitude coefficients:

$$\frac{B}{A} = \left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}}\frac{ka\kappa^2\gamma^2}{\beta(\kappa^2+\gamma^2)}\left\{\frac{n_c^2}{\kappa}\left[\frac{1}{\kappa a} - \frac{J_0(\kappa a)}{J_1(\kappa a)}\right]\right.$$
$$\left. + \frac{n_m^2}{\gamma a}\left[\frac{1}{\gamma a} - \frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)}\right]\right\}, \tag{52}$$

$$\frac{C}{A} = \frac{F}{B} = \frac{J_1(\kappa a)}{H_1^{(1)}(i\gamma a)}, \tag{53}$$

$$D \approx G \approx 0, \tag{54}$$

$$M = \left(\frac{\rho}{\gamma}\right)^{\frac{3}{2}} e^{(\rho-\gamma)b}$$
$$\cdot\left\{C\left[\left(\frac{\beta}{b}\right)^2\left(1 - \frac{\gamma^2}{\rho^2}\right)^2\left(\frac{\gamma}{\rho} - 1\right) - \gamma^2 k^2\left(1 + \frac{\gamma}{\rho}\right)\left(n_m^2 + n_3^2\,\frac{\gamma^2}{\rho^2}\right.\right.\right.$$
$$\left.\left. + \frac{\gamma}{\rho}\,(n_m^2 - n_3^2)\right)\right] + 2\omega\mu_0 F\gamma\,\frac{\beta}{b}\left(1 - \frac{\gamma^2}{\rho^2}\right)\Bigg\}$$
$$\cdot\left\{\left[\frac{\beta}{b}\left(1 - \frac{\gamma^2}{\rho^2}\right)\right]^2 - k^2\gamma^2\left(1 + \frac{\gamma}{\rho}\right)\left(n_m^2 + n_3^2\,\frac{\gamma}{\rho}\right)\right\}^{-1}, \tag{55}$$

$$N = -\left(\frac{\rho}{\gamma}\right)^{\frac{1}{2}} e^{(\rho-\gamma)b}\,\frac{2Fk^2\gamma^2\left(n_m^2 + n_3^2\,\frac{\gamma}{\rho}\right) + 2\omega n_m^2\epsilon_0\gamma\,\frac{\beta}{b}\left(\frac{\gamma^2}{\rho^2} - 1\right)C}{\left[\frac{\beta}{b}\left(1 - \frac{\gamma^2}{\rho^2}\right)\right]^2 - k^2\gamma^2\left(1 + \frac{\gamma}{\rho}\right)\left(n_m^2 + n_3^2\,\frac{\gamma}{\rho}\right)}. \tag{56}$$

The symbols J_0 and J_1 indicate Bessel functions of zero and first order; $H_0{}^{(1)}(i\gamma a)$ and $H_1{}^{(1)}(i\gamma a)$ are Hankel functions of zero and first order and of the first kind. We use here the notation of a Hankel function with imaginary argument used by Jahnke and Emde.[7] The coefficients M and N given by (55) and (56) do not contain any Hankel functions because the approximation for large arguments was used. The relation between the power P carried by the HE_{11} mode in the waveguide and the amplitude coefficient A is given by the equation

$$P = \frac{\pi}{4}\left\{\frac{k\beta_0}{\kappa^4}\left[(a\kappa)^2(J_0^2(\kappa a) + J_1^2(\kappa a)) - 2J_1^2(\kappa a)\right]\left(n_c^2 + \frac{\mu_0}{\epsilon_0}\frac{B^2}{A^2}\right)\right.$$
$$+ \frac{k\beta}{\gamma^4}\left[(a\gamma)^2\left\{1 - \left(\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)}\right)^2\right\} + 2\right]J_1^2(\kappa a)\left(n_m^2 + \frac{\mu_0}{\epsilon_0}\frac{B^2}{A^2}\right)$$
$$\left. + 2\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{B}{A}\left[\frac{\beta_0^2 + n_c^2k^2}{\kappa^4} - \frac{\beta_0^2 + n_m^2k^2}{\gamma^4}\right]J_1^2(\kappa a)\right\}\left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}}A^2. \tag{57}$$

For sufficiently large cladding radius b the propagation constant of the mode can still be obtained from the eigenvalue equation for a rod embedded in a medium with refractive index n_m.

$$\left\{n_c^2\frac{a\gamma^2}{\kappa}\left(\frac{J_0(\kappa a)}{J_1(\kappa a)} - \frac{1}{\kappa a}\right) + n_m^2\left(\gamma a\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} - 1\right)\right\}\left\{n_m^2\frac{a\gamma^2}{\kappa}\left(\frac{J_0(\kappa a)}{J_1(\kappa a)} - \frac{1}{\kappa a}\right)\right.$$
$$\left. + n_m^2\left(\gamma a\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} - 1\right)\right\} = \left[(n_c^2 - n_m^2)\frac{\beta n_m k}{\kappa^2}\right]^2\cdot \tag{58}$$

This collection of formulas is sufficient to calculate the coupling term (29) for the coupling of the HE_{11} modes in two parallel cladded optical fibers. We distinguish two cases. We assume that the modes are either both polarized perpendicular to the plane connecting the two fiber core centers or that they are polarized parallel to this plane. The coupling is different in the two cases, corresponding to the difference in coupling between the TE modes and the TM modes of the slab waveguide treated in the preceding section. It must be pointed out that a perpendicularly polarized mode does not couple with a horizontally polarized mode. The two-mode assumption is thus still valid.

The integral occurring in (29) cannot be solved generally for two coupled HE_{11} modes. However, if we assume that the distance R separating the two cores is large compared to the core radius "a," we are justified in expanding the expression for the radial coordinate r (reaching from the center of guide one to the vicinity of guide two) in the following way (Fig. 3),

$$r_1 = \sqrt{R^2 + r_2^2 + 2r_2R\cos\theta} \approx R + r_2\cos\theta. \tag{59}$$

The angle θ describes the departure of the direction of r_2 from the line connecting the core centers, r_2 is the distance from the center of guide two to the endpoint of r_1. When the expansion (59) is used, the angular integration can be performed leading to Bessel functions of the imaginary argument $i\rho r_2$. The remaining integration over r_2 now involves only products of cylinder functions and can be carried out. We obtain the following result for the case of horizontal polarization:

$$\Delta\beta_h = \left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}} \frac{A^2}{2P}\frac{M}{A}\frac{1}{k}\left(\frac{2\pi}{\rho R}\right)^{\frac{1}{2}} e^{-\rho R}\left\{\frac{e^{(\rho-\gamma)b}}{\pi(\rho\gamma)^{\frac{1}{2}}}\frac{J_1(\kappa a)}{H_1^{(1)}(i\gamma a)}\left[\left(1+\frac{\beta^2}{\gamma\rho}\right)(\gamma+\rho)\right.\right.$$

$$\left. + \frac{\beta^2}{b\gamma^2\rho^2}(\rho^2-\gamma^2)\left(1-\frac{k}{\beta}\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{B}{A}\right)\right]$$

$$+ iJ_1(i\rho a)\left[a\left(\frac{\beta^2}{\kappa}-\kappa\right)J_0(\kappa a) + a\left(\frac{\beta^2}{\gamma}+\gamma\right)\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)}J_1(\kappa a)\right.$$

$$\left.\left. - \frac{\beta^2}{\kappa^2\gamma^2}(\kappa^2+\gamma^2)\left(1-\frac{k}{\beta}\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{B}{A}\right)J_1(\kappa a)\right]\right\}. \tag{60}$$

For the case of vertically polarized modes we get

$$\Delta\beta_v = \frac{A^2}{2P}\frac{N}{A}\frac{1}{\rho}\left(\frac{2\pi}{\rho R}\right)^{\frac{1}{2}} e^{-\rho R}\left\{\frac{e^{(\rho-\gamma)b}}{\pi\gamma b(\rho\gamma)^{\frac{1}{2}}}\left[k\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{B}{A}\left((\rho+\gamma)b+\frac{\rho^2-\gamma^2}{\gamma\rho}\right)\right.\right.$$

$$\left. - \frac{\beta}{\gamma\rho}(\rho^2-\gamma^2)\right]\frac{J_1(\kappa a)}{H_1^{(1)}(i\gamma a)}$$

$$+ i\rho J_1(i\rho a)\left[ka\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{B}{A}\left(\frac{1}{\kappa}J_0(\kappa a)+\frac{1}{\gamma}\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)}J_1(\kappa a)\right)\right.$$

$$\left.\left. + \frac{\beta}{\kappa^2\gamma^2}(\kappa^2+\gamma^2)\left(1-\frac{k}{\beta}\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{B}{A}\right)J_1(\kappa a)\right]\right\}. \tag{61}$$

Finally, we present also the expression for the power loss 2α of the HE_{11} mode caused by the lossy surrounding medium. The calculation follows the idea described in connection with the slab waveguide.

$$2\alpha = \left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}}\frac{e^{-2\rho_r b}}{|\rho| P}\left\{\frac{\beta}{b}\frac{1}{|\rho|^4}\,\mathrm{Im}\left[\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} MN^*(\rho^{*2}-\rho^2)\right]\right.$$

$$\left. - k\,|M|^2\,\mathrm{Im}\left(\frac{n_3^2}{\rho}\right) - \frac{\mu_0}{\epsilon_0}k\,|N|^2\,\mathrm{Im}\left(\frac{1}{\rho}\right)\right\}. \tag{62}$$

The coefficients M and N are given by (55) and (56), ρ_r is the real part of ρ, and the symbol Im( ) indicates the imaginary part of the expression in parenthesis (or brackets).

VI. NUMERICAL RESULTS

To illustrate the consequences of the theory presented in this paper a few sample cases have been computed. For comparison purposes I have checked the results of my theory against the curves for the coupling of uncladded fibers published by Jones.[1] The agreement between his theory and similar curves obtained from (60) and (61) is excellent.

The amount of coupling between two optical fibers depends on the type of mode they are carrying, on the separation between the fiber cores, and on the loss of the surrounding medium in which the fibers are embedded. Figure 4 shows a set of curves describing the coupling parameter $a\ |\Delta\beta|$ as a function of the loss in the surrounding medium. The abscissa labeled "loss" measured in db indicates the plane wave loss in going through the surrounding medium for a distance $R - 2b$; that means it is the loss that is encountered in going perpendicular to the fiber axis from the cladding boundary of one fiber to the other. There are two curve parameters in Fig. 4. The two sets of curves differ by the cladding thickness normalized with respect to the core radius,

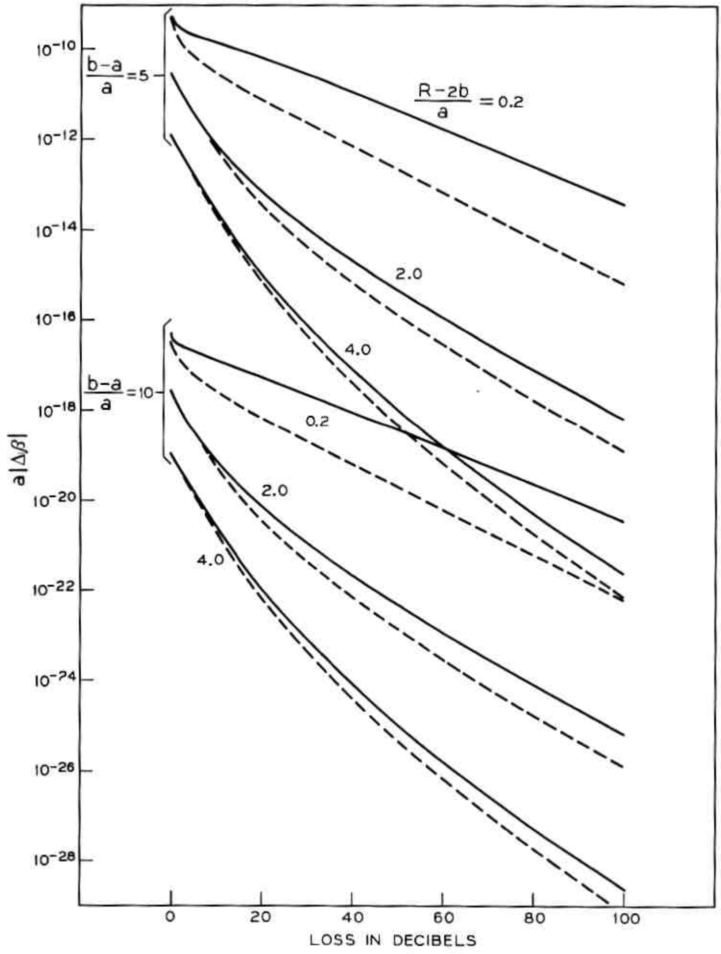


Fig. 4—Coupling of round fibers. $a|\Delta\beta|$ is the coupling parameter of equations (60) and (61), the abscissa indicates the loss (in db) that a plane wave would experience in going from the boundary of one fiber to the boundary of the other fiber. The solid lines hold for horizontal polarization, equation (60), while the dotted lines apply to the case of vertical polarization, equation (61). $n_m ka = 16$, $\gamma a = 1.594$, $n_c/n_m = 1.01$, $\mathrm{Re}(n_a) = n_m$.

$(b - a)/a$. The coupling is, of course, reduced for increasing cladding thickness since the fiber cores are farther removed from each other. The second parameter is the separation between the fiber claddings normalized with respect to the fiber core radius, $(R - 2b)/a$. Increasing values of this parameter again cause the fiber cores to be separated farther from each other. The loss values that form the abscissa hold regardless of fiber separation, always indicating the total loss that exists between the fiber claddings on their closest approach. The curves show clearly the decrease in coupling with increasing loss in the surrounding medium. The solid curves apply to the horizontal polarization of the two modes and are obtained from (60). The dotted curves apply to vertical polarization and were computed from (61). Figure 4 holds, of course, only for one particular pair of round optical fibers. It was assumed that $n_m ka = 16$, and that the ratio of core index to cladding index is $n_c/n_m = 1.01$. For simplicity, it was furthermore assumed that the real part of the index n_3 is identical to the index n_m of the cladding material. The imaginary part of n_3 is varied in order to obtain the variable loss values of the abscissa. The abscissa can be expressed by $8.68 \text{ Im } (n_3)k(R - 2b)$.

With the help of (10), Fig. 4 can be used to evaluate the crosstalk between the two optical fibers to which the conditions of Fig. 4 apply. Let us assume that we can tolerate a power ratio $V = 10^{-6}$ at the end, $z = L$, of the two fibers. If we use, for example, $a = 1$ μm and allow $L = 1$ km we find that we must not exceed

$$a \mid \Delta\beta \mid \leqq 10^{-12}. \tag{63}$$

Any value appearing in Fig. 4 below this line thus leads to acceptable crosstalk levels. It is apparent that all curves with $(b - a)/a = 10$ are acceptable. For $(b - a)/a = 5$ a certain amount of loss is required to reduce the coupling and consequently the crosstalk below the desired level.

Figure 5 presents the same data as Fig. 4 applied to two identical slab waveguides. It was assumed that the refractive indices as well as the guide dimensions of the slab waveguides correspond exactly to the round optical fibers of Fig. 4. ($a = d$, $b = D$). However, the two slab waveguides are less tightly coupled than the corresponding round fibers. The reason for this behaviour is the difference in the decay parameter γa (or γd). For the round fibers, we obtained from the eigenvalue equation (58) the value $\gamma a = 1.594$ while (41) results in $\gamma d = 1.997$ for the slab waveguides. A larger value of the decay parameter indicates that exponential decrease of the field amplitude outside of the fiber core is more rapid. The more rapid decrease of the field of the slab waveguide leads to less coupling. The solid and dotted lines have the same meaning as in Fig. 4. The solid lines apply to the TM mode which is polarized horizontally while the dotted lines apply to the vertically polarized TE mode. For simplicity we have assumed that the values of γd are identical for the two modes which, strictly speaking, is not quite true. However, we see from Fig. 4, as well as

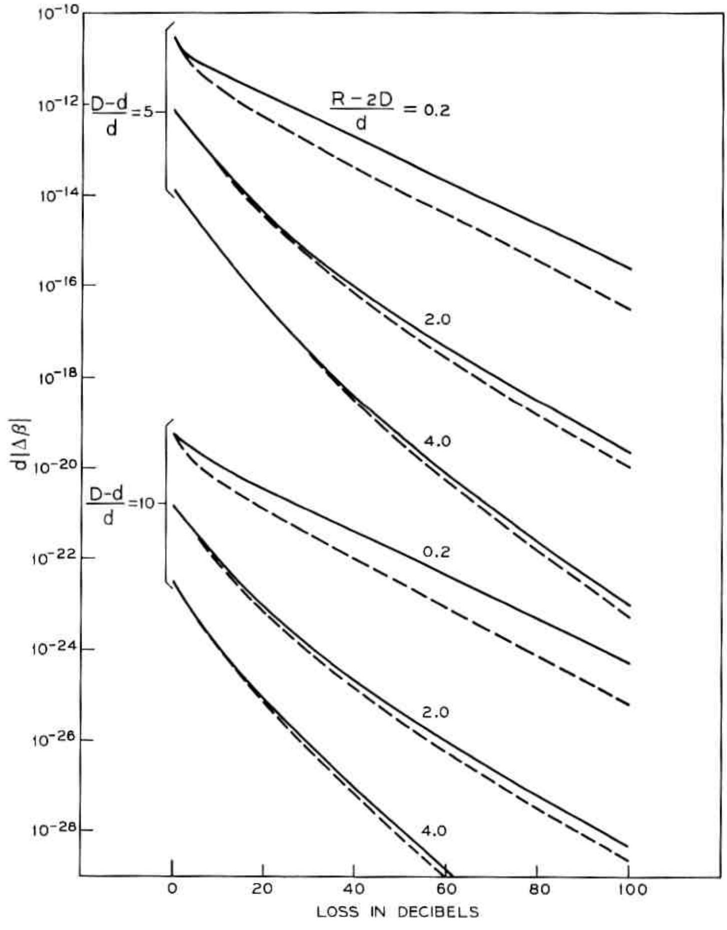


Fig. 5—Coupling of slab waveguides. Coupling parameter and loss have the same meaning as in Fig. 4. The solid lines indicate the TM mode (horizontal polarization) while the dotted lines apply to the TE mode (vertical polarization). $n_m ka = 16$, $\gamma a = 1.997$, $n_c/n_m = 1.01$.

from Fig. 5, that the horizontally polarized modes are coupled more tightly than the vertically polarized modes.

Figure 6 presents a comparison between the coupling strength of round fibers and slab waveguides. Since equal geometry does not lead to identical values of the decay parameter we have arbitrarily used equal values of $\gamma a = \gamma d = 1.594$ for both types of waveguide. The solid lines indicate the results for the round fiber while the dotted lines apply to the slab waveguides. It is apparent that the slab waveguides are more tightly coupled when the conditions are such that both guides have equal decay parameters. This appears reasonable if we consider that the slab waveguides have constant distance from each other over an infinite length while the round fibers have a point of closest approach so that the total amount of field overlap is less in this case.

The lossy medium does not only serve to reduce the coupling between the waveguides but it also has the adverse effect of increasing the loss of the modes traveling in the guides. Figure 7 is a plot of

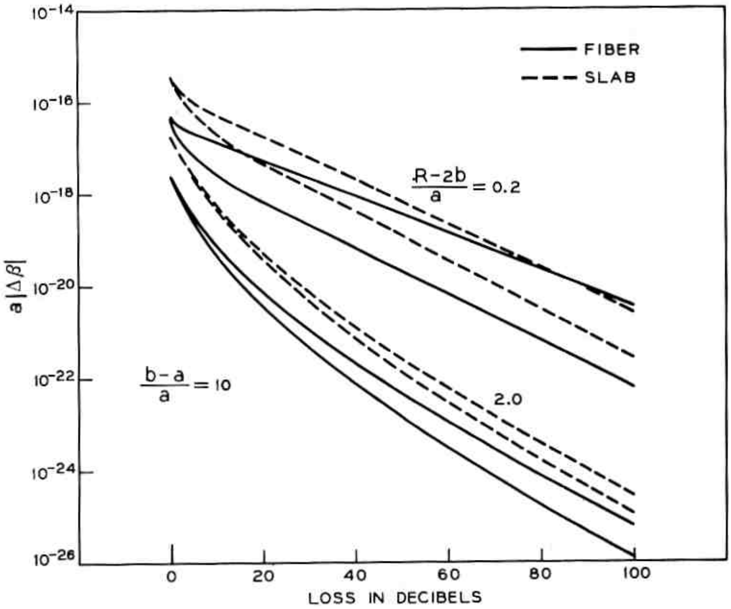


Fig. 6—Comparison between the coupling of round fibers and slab waveguides ($a = d$, $b = D$). The parameter value $\gamma a = 1.594$ is correct for the round fibers. It has also been used for the slab waveguides to enable a realistic comparison. $n_m ka = 16$, $n_e/n_m = 1.01$.

mode power loss versus plane wave power loss $2k[\mathrm{Im}\ (n_3)a]$ in the surrounding medium obtained from (43), (44b), and (62). Both types of waveguides are represented in this figure. The loss values of abscissa and ordinate are both expressed in db. The decay parameters γa and γd (we use $a = d$ and $b = D$ in Fig. 7) are solutions of the respective eigenvalue equations and are not adjusted arbitrarily. We have seen that the coupling of the slab waveguides was weaker than the coupling of the fibers under these conditions. It is thus not surprising to see from Fig. 7 that the losses of the fiber are larger than the slab waveguide losses. It is interesting to see that the symmetric TM mode of the slab waveguide is slightly lossier than the symmetric TE mode.

Assuming again that $a = d = 1\ \mu\mathrm{m}$ and limiting the allowable loss caused by the lossy surrounding medium to 1 db/km, we find that we can accept those curves with losses of less than $2\alpha a = 10^{-9}$ db. Since the loss curves are nearly independent of the actual value of the loss in the surrounding medium (after an initial rise from zero mode loss in a lossless surrounding medium) we can exclude all curves above the 10^{-9} line as unacceptable in case of a lossy surrounding medium. We thus see that the slab waveguide works well under all conditions shown. For the more important round fiber we must exclude the case of a cladding with $(b - a)/a = 5$. This means that the curves of Fig. 4 that lead to excessive coupling in the absence of sufficient loss in the surrounding medium are also unacceptable from the point of view of additional mode loss. We see that additional mode loss and mode coupling have similar trends. When the coupling between the fibers is too large we also have to contend with additional mode loss in excess of what we want to tolerate. A guide with low mode loss on the other hand is also sufficiently protected by its cladding so that coupling between adjacent fibers is not critical.

The two remaining curves, Figs. 8 and 9, show the coupling parameter and the additional mode loss for the case of a guide with lower core-to-cladding ratio. Only the case of the higher coupling for horizontally polarized modes has been plotted in Fig. 8. The remarks about coupling strength and additional mode loss hold true also for these conditions as can be seen from the figures.

VII. CONCLUSIONS

We have presented a perturbation theory for the coupling between two arbitrary dielectric waveguides made of lossy materials and embedded in a lossy environment. Only the case of two degenerate

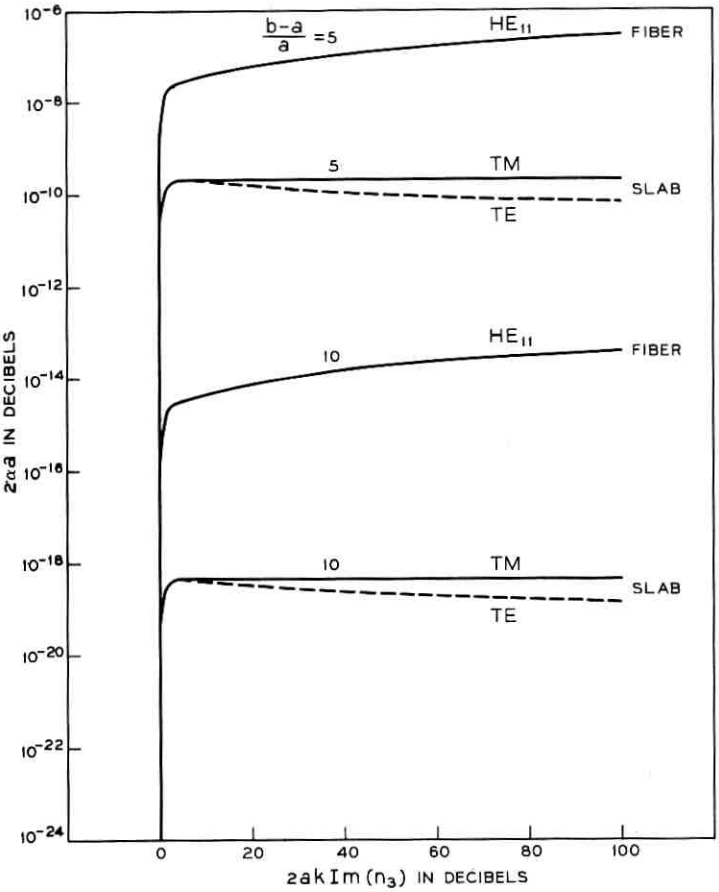


Fig. 7—Additional mode loss caused by the lossy surrounding medium. The abscissa indicates the loss (in db) that a plane wave would suffer by traveling a distance equal to the core radius in the surrounding medium. $n_m ka = 16$, $n_c/n_m = 1.01$.

modes has been treated. The general theory has been applied to cladded slab waveguides and cladded round optical waveguides (fibers) with lossless cladding material but in the presence of lossy surroundings. We have found that the coupling between fibers depends very strongly on the separation between the two fiber cores. Far separated cores result in loose coupling. The coupling between the fibers can be reduced by increasing the losses of the surrounding medium. However, a lossy surrounding medium introduces losses to the waveguide modes to such an extent that the mode losses are unacceptably high if the loss in the surrounding medium is necessary to uncouple the fibers. From this point of view it appears as though losses in the surrounding medium are unsuitable to achieve uncoupling or reduction of crosstalk between optical fibers. However, the coupling mechanism discussed in this paper is not the only contributor to crosstalk. Imperfections in the fibers cause light scattering that can also be a source of crosstalk.[8] This mechanism is less strongly dependent on the separation of the fiber cores so that it may not be possible to uncouple fibers sufficiently simply by protecting them with a cladding of sufficient thickness.

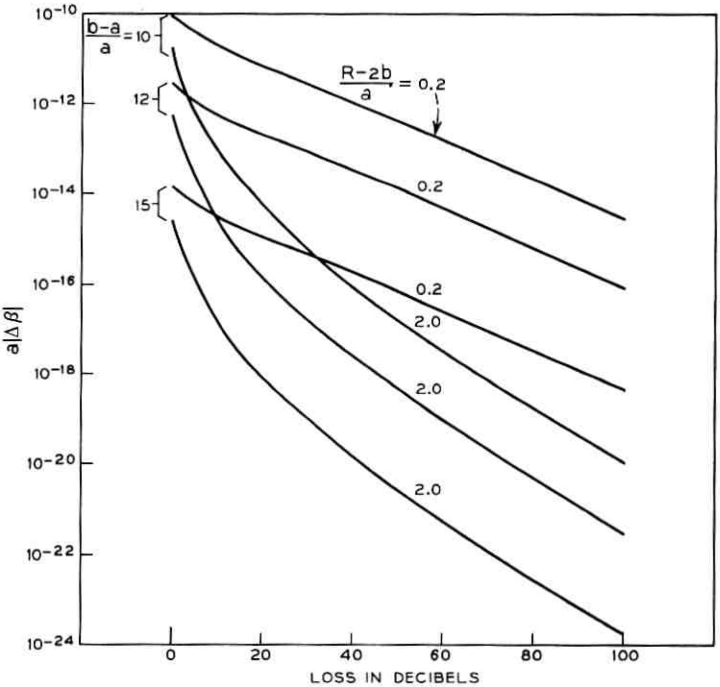


Fig. 8—Coupling of round fibers. Only the case of horizontal polarization is shown. $n_m ka = 21$, $\gamma a = 0.857$, $n_c/n_m = 1.003$.

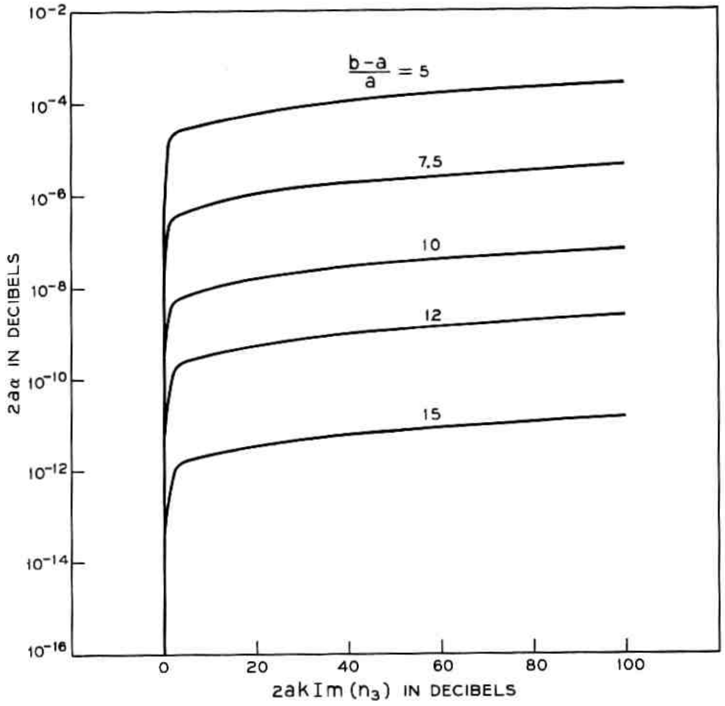


Fig. 9—Additional mode loss caused by the lossy surrounding medium. $n_m ka = 21$, $n_c/n_m = 1.003$.

REFERENCES

1. Jones, A. L., "Coupling of Optical Fibers and Scattering in Fibers," J. Opt. Soc., *55*, No. 3 (March 1965), pp. 261–271.
2. Vanclooster, R., and Phariseau, P., "The Coupling of Two Parallel Dielectric Fibers, I. Basic Equations," Physica, *47*, No. 4 (June 30, 1970), pp. 485–500.
3. Miller, S. E., "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 3 (May 1954), pp. 661–719.
4. Kapany, N. S., *Fiber Optics*, New York: Academic Press, 1967.
5. Marcuse, D., "Radiation Losses of Tapered Dielectric Slab Waveguides," B.S.T.J., *49*, No. 2 (February 1970), pp. 273–290.
6. Collin, R. E., *Field Theory of Guided Waves*, New York: McGraw-Hill, 1960.
7. Jahnke, E., and Emde, F., *Tables of Functions*, New York: Dover, 1945.
8. Marcuse, D., "Crosstalk Caused by Scattering in Slab Waveguides," B.S.T.J., this issue, pp. 1817–1831.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Crosstalk Caused by Scattering in Slab Waveguides

By D. MARCUSE

(Manuscript received March 1, 1971)

Light scattered in one optical fiber can be scattered back into a guided mode in a neighboring optical fiber. This type of scattering crosstalk is investigated in this paper for slab waveguides. However, the results for the slab waveguide case are upper limits on the crosstalk between round optical fibers. Scattering crosstalk is expressed in terms of guided mode radiation loss whose cause is the same scattering mechanism. It is thus not necessary to know the details of the scattering mechanism as long as the scattering loss, which is a measurable quantity, is known. In addition to the total radiation loss the crosstalk depends on the width of the radiation lobe in which the scattered energy escapes from the waveguide. The crosstalk is inversely proportional to the width of the radiation lobe and thus is larger when the radiation lobe is narrow. Scattering crosstalk can become serious if both guides have a systematic sinusoidal imperfection of the same mechanical frequency resulting in a radiation lobe of nearly zero width. In the absence of a systematic sinusoidal imperfection it can be concluded that scattering crosstalk between dielectric waveguides (optical fibers) is negligible if the radiation losses are tolerable. In this case it appears unnecessary to suppress crosstalk by means of a lossy medium between the waveguides.

I. INTRODUCTION

In an earlier paper[1] I discussed the crosstalk between optical dielectric waveguides based on the directional coupler mechanism. In that case, crosstalk was caused by the fact that the exponentially decaying field of one guide reaches the region of a neighboring guide. This type of crosstalk exists even if both guides can be considered to be perfect dielectric cylinders.

The present paper is devoted to a crosstalk mechanism of a different type. A dielectric waveguide or optical fiber loses power by radiation if the guiding structure is imperfect in any way. This imperfection may consist of deviations in the perfect cylindrical geometry of the core-cladding interface or it may consist in random variations of the refractive index distribution. Some of the radiation, causing loss to the guided mode of one fiber, may reach a neighboring fiber and can there be scattered back into one of its guided modes. This type of crosstalk is thus caused by waveguide imperfections. However, just as in the case of the directional coupler effect, it is possible to link predictions of the effectiveness of this type of crosstalk mechanism to the loss it causes to the guided mode. This makes it possible to give very simple rules that link crosstalk to radiation loss (see equation (50)). It turns out that scattering crosstalk is no serious problem provided that the radiation loss of the guided mode (that is caused by the same mechanism) remains within acceptable limits.

Our treatment of the scattering crosstalk problem is simplified by limiting it to the case of the slab waveguide. It is intuitively clear that the crosstalk between slab waveguides must be stronger than crosstalk between round optical fibers. In fact, the crosstalk between parallel slab waveguides is independent of their separation while the crosstalk between round fibers must vary inversely proportional to their distance. Furthermore, our treatment of crosstalk starts out by assuming a definite scattering mechanism. That is, we assume that the core of either guide varies in width. However, this special assumption allows us to describe the more general case. No matter what the mechanism, scattering of light can be attributed to the individual Fourier components of either the core width variations or of the departure of any other quantity from its perfect value. Regardless of the particular mechanism each Fourier component causes a narrow radiation lobe to escape in a definite direction. The same mechanism that produced the radiation lobe in one guide is responsible for the capture of some of this radiation in the neighboring guide. It is thus immaterial by what mechanism the radiation and reconversion was produced. In particular, if we express the strength of the scattering mechanism by the radiation loss that it causes to the guided mode, we have a description of the crosstalk effect that is independent of the actual scattering mechanism. Realization of these few basic facts greatly simplifies the treatment of the scattering crosstalk problem.

The deterministic scattering theory can easily be extended to include statistical assumptions about the scattering mechanism. We are thus able to express the scattering crosstalk in terms of three parameters: the scattering mode loss, the ratio of waveguide length to free space wavelength of the guided radiation, and the width of the radiation lobe.

II. CROSSTALK CAUSED BY SINUSOIDAL CORE THICKNESS VARIATION

As discussed in the introduction, we are basing the treatment of scattering crosstalk on the mechanism of core width variations. To begin with, we concentrate on crosstalk caused by sinusoidal variations of the core of both slab waveguides. The more general case of arbitrary core width variations can be treated by decomposing the arbitrary functions into Fourier series and utilizing the result for a single sine wave component. Finally, we express the crosstalk in terms of radiation losses suffered by each guided mode and are thus able to free ourselves of the particular scattering mechanism.

It is well known that dielectric waveguides possess two types of modes. At a given operating frequency there are a finite number of guided modes and a continuum of radiation modes. The coupling between these two types of modes caused by variations of the core thickness has been explored in earlier papers.[2,3] We can thus simply draw on these earlier results for our present purposes.

The even and odd radiation modes of the dielectric slab waveguide [equations (24) and (25) of Ref. 3] can be expressed in the following simple form outside of the core, $x_1 > d$ (Fig. 1):

$$\mathcal{E}_y^{(e)}(\rho) = \left(\frac{2\omega\mu P}{\pi\beta}\right)^{\frac{1}{2}} \cos\left[\rho(|x_1| - d) + \phi\right]e^{-i\beta z}, \tag{1}$$

$$\mathcal{E}_y^{(o)}(\rho) = \frac{x_1}{|x_1|}\left(\frac{2\omega\mu P}{\pi\beta}\right)^{\frac{1}{2}} \cos\left[\rho(|x_1| - d) + \Psi\right]e^{-i\beta z}. \tag{2}$$

As usual, the factor $\exp(+i\omega t)$ has been suppressed. The phases appearing in (1) and (2) are given by

$$\tan\phi = \frac{\sigma}{\rho}\tan\sigma d \tag{3}$$

and

$$\tan\psi = -\frac{\sigma}{\rho}\cot\sigma d, \tag{4}$$

with $(k = \omega\sqrt{\epsilon_o\mu_o})$

$$\rho = (n_2^2k^2 - \beta^2)^{\frac{1}{2}} \tag{5}$$

and

$$\sigma = (n_1^2k^2 - \beta^2)^{\frac{1}{2}}. \tag{6}$$

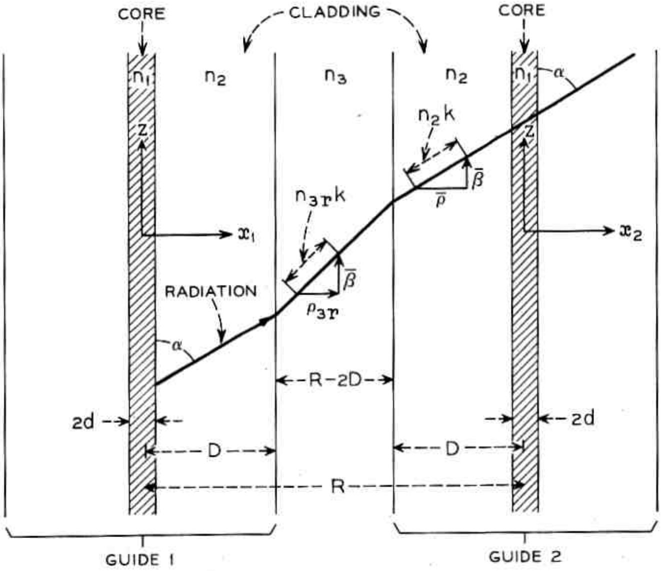


Fig. 1—Geometry of the two slab waveguides. The figure shows the definition of several parameters and indicates a scattered light beam (as a ray) going from guide 1 to guide 2.

The slab half-width is d; n_1 and n_2 are the refractive indices of core and cladding. The power carried by the radiation mode is defined by the orthogonality condition of the radiation modes:

$$\frac{\beta}{2\omega\mu}\int_{-\infty}^{\infty} \mathcal{E}_y^{(e)}(\rho)\,\mathcal{E}_y^{(o)}(\rho')\,dx = P\delta_{eo}\,\delta(\rho-\rho'). \tag{7}$$

δ_{eo} is the Kronecker delta symbol while $\delta(\rho-\rho')$ is Dirac's delta function. An even guided TE mode is coupled to only even radiation modes by a sinusoidal variation of the core thickness of the form

$$d_1 = d + a_1 \sin\theta_1 z. \tag{8}$$

According to Ref. 3 the radiation field excited by the sinusoidal thickness variation of the slab core is given by

$$E_y = -\frac{a_1 k^2 (2\omega\mu_o P)^{\frac{1}{2}}(n_1^2 - n_2^2)\cos\kappa_1 d}{\pi\left(\beta_1 d + \dfrac{\beta_1}{\gamma_1}\right)^{\frac{1}{2}}}$$

$$\cdot\int_{\theta_1-\beta_1-n_2k}^{\theta_1-\beta_1+n_2k} \frac{\cos\sigma d\,\cos\left[\rho(x_1-d)+\phi\right]}{\{\rho^2\cos^2\sigma d+\sigma^2\sin^2\sigma d\}^{\frac{1}{2}}}\,e^{-i(\beta-\Delta/2)z}\,\frac{\sin\Delta\dfrac{z}{2}}{\Delta}\,d\Delta. \tag{9}$$

This formula holds for $d < x_1 < \infty$. The variable Δ is given by

$$\Delta = \theta_1 - (\beta_1 - \beta). \tag{10}$$

The parameter β_1 is the propagation constant of the even guided TE mode. The other parameters appearing in (9) are

$$\kappa_1 = (n_1^2k^2 - \beta_1^2)^{\frac{1}{2}} \tag{11}$$

and

$$\gamma_1 = (\beta_1^2 - n_2^2k^2)^{\frac{1}{2}}. \tag{12}$$

There are two differences between (9) and equation (27) of Ref. 3. We are now dealing with core thickness variations instead of a variation of only one of the faces of the slab core. The guide length L had to be replaced with the length coordinate z since we assume that the guide is very long and that we are looking at a field close to the fiber core and not at $z \gg L$ infinitely far from the sinusoidally perturbed section of length L. Finally, we have replaced the differential $d\rho$ of equation (27), Ref. 3 by

$$d\rho = -\frac{\beta}{\rho}\, d\Delta. \tag{13}$$

For large values of z the factor $(1/\pi)(\sin \Delta z/2)/\Delta$ approaches a delta function. We are thus permitted to take slowly varying functions out of the integral and are left with

$$\lim_{z\to\infty} \int_{-\infty}^{\infty} \cos\,[\rho(x_1 - d) + \phi]e^{-i(\beta - \Delta/2)z}\, \frac{\sin \Delta \dfrac{z}{2}}{\Delta}\, d\Delta$$
$$= \frac{\pi}{2} \exp\,\{-i[\bar{\beta}z + \bar{\rho}(x_1 - d) + \phi]\}. \tag{14}$$

The limits on the integral were extended from $-\infty$ to $+\infty$. This does not change its value since only the immediate vicinity of $\Delta = 0$ gives a contribution. The dependence of ρ and β on Δ follows from (5) and (10). The propagation constant $\bar{\beta}$ is given by $\bar{\beta} = \beta_0 - \theta_1$ (or $\Delta = 0$). It is more convenient, however, to express it in terms of the angle α at which the radiation is leaving the slab core.[3] We have

$$\bar{\beta} = n_2k \cos \alpha \tag{15}$$

and, similarly,

$$\bar{\rho} = n_2k \sin \alpha. \tag{16}$$

The radiation field that is caused by scattering of light from the even TE guided mode by a sinusoidal variation of the core thickness is thus simply a plane wave.

$$E_y = -\frac{a_1 k^2 (2\omega\mu_o P)^{\frac{1}{2}} (n_1^2 - n_2^2) \cos \kappa_1 d \cos \bar{\sigma} d}{2\left\{\left(\beta_1 d + \dfrac{\beta_1}{\gamma_1}\right)(\bar{\rho}^2 \cos^2 \bar{\sigma} d + \bar{\sigma}^2 \sin \bar{\sigma}^2 d)\right\}^{\frac{1}{2}}} e^{i(\bar{\rho} d - \phi)} e^{-i(\bar{\beta} z + \bar{\rho} x_1)}. \tag{17}$$

The phase terms ϕ is given by (3), the parameter $\bar{\sigma}$ follows from (6) by a replacement of β with $\bar{\beta}$ of (15).

This calculation proves once more that the standing-wave radiation modes result in traveling waves when properly superimposed.

We have so far ignored the finite width of the cladding. We could easily take reflections and refraction at the cladding boundary into account. However, our theory is meant as an order-of-magnitude estimate of crosstalk so that the small reflection losses at the cladding boundary will simply be ignored.

As the radiation reaches the core of the neighboring slab waveguide, reflection and refraction are taking place. It is intuitively clear that the plane wave impinging on the core of the neighboring slab waveguide can be expressed as a superposition of an even and an odd radiation mode of the guide. In fact, it can be shown that the plane wave impinging on the second guide can be expressed by

$$E_y = G\{e^{-i\phi} \mathcal{E}_y^{(e)}(\bar{\rho}) - e^{-i\Psi} \mathcal{E}_y^{(o)}(\bar{\rho})\}, \tag{18}$$

with the even and odd radiation modes (1) and (2) being expressed in a coordinate system centered at the core of the second guide. The coefficient is given by

$$G = -e^{-i\phi} \frac{a_1 k^2 (\pi\beta)^{\frac{1}{2}} (n_1^2 - n_2^2) \cos \kappa_1 d \cos \bar{\sigma} d}{2\left\{\left(\beta_1 d + \dfrac{\beta_1}{\gamma_1}\right)(\bar{\rho}^2 \cos^2 \bar{\sigma} d + \bar{\sigma}^2 \sin^2 \bar{\sigma} d)\right\}^{\frac{1}{2}}} e^{-i\bar{\rho}(R - 2d)}. \tag{19}$$

R is the distance between the centers of the two waveguides. We now know how the radiation originating at the first guide excites the radiation modes of the second guide. If we assume that the second guide also has no other imperfections than thickness changes of the core and if we are concerned with the excitation of another even TE mode, only even radiation modes of the second guide can contribute to the crosstalk problem.

The question of how an even radiation mode excites an even guided TE mode has been solved (in principle) in Ref. 2. The excitation of the guided mode is expressed by an excitation coefficient $c(z)$. However, we are interested only in the value of c at $z = L$. To the first order of perturbation theory we obtain

$$c(L) = ie^{-i2\bar{\beta}(D-d)}e^{-i\rho_{3r}(R-2D)}e^{-\alpha_3(n_3k/\rho_{3r})(R-2D)}e^{-2i\phi}$$
$$\cdot \frac{a_1a_2\bar{\rho}k^4(n_1^2 - n_2^2)^2 \cos \kappa_1 d \cos \kappa_2 d \cos^2 \bar{\sigma} d e^{-\alpha_2 L}}{2\left\{\left(\beta_1 d + \frac{\beta_1}{\gamma_1}\right)\left(\beta_2 d + \frac{\beta_2}{\gamma_2}\right)\right\}^{\frac{1}{2}}(\bar{\rho}^2 \cos^2 \bar{\sigma} d + \bar{\sigma}^2 \sin^2 \bar{\sigma} d)}$$
$$\cdot \int_0^L e^{-(\alpha_1-\alpha_2)z}\left(\sin \theta_2 z\right)e^{i(\beta_2-\bar{\beta})z}\, dz. \tag{20}$$

Several new parameters have been introduced in (20) (see Fig. 1). R is the separation between the core centers of the two guides. The first exponential factor of (20) accounts for the phase shift of the plane wave inside the claddings of the two guides with cladding thickness $D - d$. The second exponential factor accounts for the phase shift in the medium of index n_3 between the two guides (separation $R - 2D$). The value of $\bar{\rho}$ in the medium between the guides was designated by ρ_{3r} (see Fig. 1). The subscript r indicates the real part of ρ_3. The imaginary part of ρ_3 attenuates the plane wave as it traverses the space between the two guides. We have assumed that the core and cladding of both guides are lossless but allow for the possibility that the medium separating the two guides may be lossy. The loss suffered by the plane wave in the medium between the guides is given by the third exponential function in (20). The real part of the refractive index of the medium between the guides is n_3, the amplitude loss coefficient of the wave is α_3. The ratio n_3k/ρ_{3r} adjusts for the slant angle of the wave and $R - 2D$ is the separation of the two guides (cladding-to-cladding distance). The subscripts 1 and 2 refer to quantities belonging to guide 1 and 2. Finally, we have added the amplitude loss coefficients α_1 and α_2 to account for the losses suffered by each guided mode. These losses can be thought of as the sum of heat losses in each guide plus radiation losses caused by the variation of the core thickness. The factors a_1 and a_2 indicate the amplitudes of the sinusoidal thickness variation of the cores of each guide. However, it is assumed that the perfect geometry and the refractive indices of both guides are identical. The mechanical frequencies θ_1 and θ_2 are not independent of each other. Only when the propagation constant β_2 of the guided mode of the second guide and the mechanical frequency θ_2 of the thickness variation of its core are related to each other by the equation

$$\theta_2 = \beta_2 - \bar{\beta} \tag{21}$$

can any amount of power be coupled from one guide to the other.

From (10) (with $\beta = \bar{\beta}$ and $\Delta = 0$) and (21) we obtain the condition

$$\theta_2 = \theta_1 - (\beta_1 - \beta_2). \tag{22}$$

The range of θ_1 values contributing to radiation is given by[3]

$$\beta_1 - n_2 k < \theta_1 < \beta_1 + n_2 k. \tag{23}$$

In the more general case of a Fourier spectrum of mechanical frequencies of the core thickness variation, we see that each Fourier component of guide 2 combines with a certain Fourier component of guide 1 to provide crosstalk.

III. GENERAL CORE THICKNESS VARIATION

Having determined the amplitude coefficient of the guided mode in guide 2 that is excited by purely sinusoidal variations of the thickness of the fiber cores we can immediately generalize to the case of arbitrary core thickness variation. We expand the core thickness of each guide in a Fourier series:

$$d_1 = d + \left(\frac{2}{L}\right)^{\frac{1}{2}} \sum_{\nu=1}^{\infty} a_{1\nu} \sin \theta_\nu z, \tag{24}$$

$$d_2 = d + \left(\frac{2}{L}\right)^{\frac{1}{2}} \sum_{\mu=1}^{\infty} a_{2\mu} \sin \theta_\mu z, \tag{25}$$

with $\theta_\nu = \nu\pi/L$. We choose the Fourier sine series because we already know the result for a single sinusoidal thickness variation. The expansions (24) and (25) involve a complete set of orthogonal functions. By replacing

$$a_1 \rightarrow \left(\frac{2}{L}\right)^{\frac{1}{2}} a_{1\nu} \tag{26}$$

and

$$a_2 \rightarrow \sqrt{\frac{2}{L}}\, a_{2\mu}\ , \tag{27}$$

and summing over ν we convert (20) to the general case. Only one summation is required since each Fourier component of guide 1 combines with only one definite Fourier component of guide 2. However, before writing down the resulting formula we will make two more changes. *First,* we can convert the sum to an integral with the help of the relation

$$\sum_{\nu} \rightarrow \frac{L}{\pi} \int d\theta. \tag{28}$$

Second, we introduce the Fourier coefficient $\phi(\theta)$ that was used to express the radiation losses in Ref. 3. By definition we have

$$\phi_j(\theta) = \frac{1}{L}\int_0^L (d_j - d)e^{-i\theta z}\,dz \qquad j = 1, 2. \tag{29}$$

The relation between $\phi_j(\theta)$ and $a_{j\nu}$ is given by the following equation

$$\phi_j^{(i)}(\theta_\nu) = \mathrm{Im}\,[\phi_j(\theta_\nu)] = -\frac{1}{(2L)^{\frac{1}{2}}}\,a_{j\nu} \qquad j = 1, 2. \tag{30}$$

The notation Im indicates the imaginary part. Then we form the ratio of power at the end of guide 2, $P_2(L)$, to the power at the end of guide 1, $P_1(L) = e^{-2\alpha_1 L}P$,

$$\frac{P_2(L)}{P_1(L)} = \frac{|\,c\,|^2 P}{P_1(L)} = e^{2\alpha_1 L}\,|\,c\,|^2. \tag{31}$$

With the help of (21) the integral in (20) can easily be performed and we finally obtain

$$\frac{P_2(L)}{P_1(L)} = \frac{L^2(n_1^2 - n_2^2)^4 k^8 \cos^2 \kappa_1 d \cos^2 \kappa_2 d e^{2(\alpha_1-\alpha_2)L}}{\pi^2\left(\beta_1 d + \dfrac{\beta_1}{\gamma_1}\right)\left(\beta_2 d + \dfrac{\beta_2}{\gamma_2}\right)}\left(\frac{1 - e^{-(\alpha_1-\alpha_2)L}}{\alpha_1 - \alpha_2}\right)^2$$
$$\cdot\left|\int_{\beta_1-n_3k}^{\beta_1+n_3k}\frac{\phi_1^{(i)}(\theta_1)\phi_2^{(i)}(\theta_2)\bar{\rho}\cos^2\bar{\sigma}d}{\bar{\rho}^2\cos^2\bar{\sigma}d + \bar{\sigma}^2\sin^2\bar{\sigma}d}\right.$$
$$\left.\cdot e^{-2i\phi}e^{-i[2\bar{\beta}(D-d)+\rho_{3r}(R-2D)]}e^{-\alpha_3(n_3k/\rho_{3r})(R-2D)}\,d\theta_1\right|^2. \tag{32}$$

Equation (32) is the expression for the crosstalk between two identical slab waveguides. The coupling mechanism is the thickness variation of the two waveguide cores. The even guided TE modes in either waveguide need not be the same.

In its general form the crosstalk ratio is not very useful since the spectral functions ϕ_1 and ϕ_2 are usually not known. However, it is possible to provide estimates of the crosstalk by using the radiation loss of the guided mode that is caused by the same mechanism that provides the crosstalk. The scattering loss was given in Ref. 3 $(i = 1, 2)$.

$$2\alpha_{ri} = \frac{\Delta P_i}{PL} = \frac{L(n_1^2 - n_2^2)^2 k^4 \cos^2 \kappa_i d}{\pi\left(\beta_i d + \dfrac{\beta_i}{\gamma_i}\right)}$$
$$\cdot\int_{-n_3k}^{n_3k}\frac{|\,\phi_i(\theta_i)\,|^2\,\rho\cos^2\sigma d}{\rho^2\cos^2\sigma d + \sigma^2\sin^2\sigma d}\,d\beta. \tag{33}$$

[There is an error in equation (14a), Ref. 3. The left-hand side should read $\Delta Pd/(PL)$.] Equation (33) differs from equation (14) of Ref. 3 since we have here assumed that the core thickness varies while in Ref. 3 it was assumed that only one side of the core cladding interface is distorted.

For simplicity we consider now the case that the two modes of the coupled guides are identical and that only one sine wave component of the Fourier spectrum exists. We thus have ($\phi_1 = \phi_2$):

$$| \phi_1(\theta) |^2 = | \phi_2 (\theta)|^2 = \frac{\pi}{2L} a^2 \, \delta(\theta - \theta') \tag{34a}$$

and

$$[\phi_1^{(i)}(\theta)]^2 = \phi_1^{(i)}(\theta)\phi_2^{(i)}(\theta) = \frac{\pi}{4L} a^2 \, \delta(\theta - \theta'), \tag{34b}$$

and obtain from (32) and (33) with $\alpha_1 = \alpha_2$

$$\frac{P_2(L)}{P_1(L)} = (1/4)(2\alpha_r L)^2 e^{-2\alpha_3(n_3k/\rho_{3r})(R-2D)}. \tag{35}$$

The index r is meant to indicate that α_r is the radiation loss. In this special case the crosstalk is simply proportional to the square of the mode power loss coefficient $2\alpha_r L$ times the power loss that the plane wave suffers in going from guide 1 to guide 2 through the lossy medium between the guides.

IV. RANDOM VARIATIONS OF THE CORE THICKNESS

The average radiation loss is simply obtained by replacing $|\phi_i|^2$ with the ensemble average $\langle|\phi_i|^2\rangle$ in (33). Obtaining the ensemble average of (32) is a little more complicated. To simplify the discussion we write the integral expression occurring in (32) in the following form

$$I = \left| \int_{\beta_1 - n_2 k}^{\beta_1 + n_2 k} \phi_1^{(i)}(\theta_1)\phi_2^{(i)}(\theta_2)F(\theta) \, d\Theta \right|^2$$
$$= \left| \frac{\pi}{L} \sum_{\nu=1}^{\infty} \phi_1^{(i)}(\theta_\nu)\phi_2^{(i)}(\theta_\mu)F(\theta_\nu) \right|^2, \tag{36}$$

with

$$\theta_\mu = \theta_\nu - (\beta_1 - \beta_2). \tag{37}$$

With the help of (30) we can also write for the ensemble average of I

$$\langle I\rangle = \frac{\pi^2}{4L^4}\sum_{\nu\nu'}\langle a_{1\nu}a^*_{1\nu'}a_{2\mu}a^*_{2\mu'}\rangle F(\theta_\nu)F^*(\theta_{\nu'}). \tag{38}$$

The asterisk indicates complex conjugation. It appears reasonable to assume that there is no correlation between guide 1 and 2. We thus assume that we can write

$$\langle a_{1\nu}a^*_{1\nu'}a_{2\mu}a^*_{2\mu'}\rangle = \langle a_{1\nu}a^*_{1\nu'}\rangle\langle a_{2\mu}a^*_{2\mu'}\rangle. \tag{39}$$

It is shown in Ref. 4 that for a stationary random process with $L \rightarrow \infty$ we have

$$\langle a_{1\nu}a^*_{1\nu'}\rangle = \langle\, |a_{1\nu}|^2\rangle\delta_{\nu\nu'}\,. \tag{40}$$

[The proof presented in Ref. 4 for the usual complex Fourier series can easily be extended to the Fourier series (24). The conclusion (40) remains correct.]

We thus have

$$\langle I\rangle = \frac{\pi^2}{4L^4}\sum_{\nu}\langle|\,a_{1\nu}\,|^2\rangle\langle|\,a_{2\mu}\,|^2\rangle\,|\,F(\theta_\nu)\,|^2. \tag{41}$$

Returning to the integral notation we have

$$\langle I\rangle = \frac{\pi}{L}\int_{\beta_1-n_2k}^{\beta_1+n_2k}\langle|\,\phi_1^{(i)}(\theta_1)\,|^2\rangle\langle|\,\phi_2^{(i)}(\theta_2)\,|^2\rangle\,|\,F(\theta_1)\,|^2\,d\theta_1\,. \tag{42}$$

For stationary random processes we can assume $\langle|\phi^{(i)}(\theta)|^2\rangle = \frac{1}{2}\,\langle|\phi(\theta)|^2\rangle$. We thus obtain the ensemble average of the crosstalk in the form:

$$\frac{\langle P_2(L)/P_1(L)\rangle}{\langle 2\alpha_{r1}L\rangle\langle 2\alpha_{r2}L\rangle} = \frac{\pi}{4n_2kL}\,\frac{e^{2(\alpha_1-\alpha_2)L}\left(\dfrac{1-e^{-(\alpha_1-\alpha_2)L}}{\alpha_1-\alpha_2}\right)^2\dfrac{1}{L^2}\displaystyle\int_0^\pi\frac{\langle|\,\phi_1(\theta_1)\,|^2\rangle\langle|\,\phi_2(\theta_2)\,|^2\rangle\bar{\rho}^3n_2k\cos^4\bar{\sigma}d}{(\bar{\rho}^2\cos^2\sigma d+\bar{\sigma}^2\sin^2\bar{\sigma}d)^2}\,e^{-2\alpha_2(n_2k/\rho_{2r})(R-2d)}\,d\alpha}{\displaystyle\int_0^\pi\frac{\langle|\,\phi_1(\theta)\,|^2\rangle\rho^2\cos^2\sigma d}{\rho^2\cos^2\sigma d+\sigma^2\sin^2\sigma d}\,d\alpha\int_0^\pi\frac{\langle|\,\phi_2(\theta)\,|^2\rangle\rho^2\cos^2\sigma d}{\rho^2\cos^2\sigma d+\sigma^2\sin^2\sigma d}\,d\alpha}. \tag{43}$$

We used (15), (16), and (10) which, with $\Delta = 0$, assumes the form

$$\theta_1 = \beta_1 - \bar{\beta} = \beta_1 - n_2k\cos\alpha \tag{44}$$

to obtain

$$d\theta_1 = n_2k\sin\alpha d\alpha = \bar{\rho}d\alpha$$

for the integral in the numerator of (43). The angle α indicates the direction in which the plane wave radiation from a given Fourier component escapes from the waveguide core. (The angle α must not be confused with the loss coefficients α_1 or α_r.) In the integrals in the denominator we have likewise used the transformation

$$\beta = n_2 k \cos \alpha, \tag{45}$$

so that we obtain from (5)

$$\rho = (n_2^2 k^2 - \beta^2)^{\frac{1}{2}} = n_2 k \sin \alpha \tag{46}$$

and

$$d\beta = -\rho d\alpha. \tag{47}$$

For a discussion of (43) it is much more convenient to consider the case that mode 1 of guide 1 and mode 2 of guide 2 are identical. The conclusions for the more general case are very similar. We can furthermore assume that the power spectra of the two guides are identical since they are supposed to be statistically similar,

$$\langle |\phi_1|^2 \rangle = \langle |\phi_2|^2 \rangle = \langle |\phi|^2 \rangle. \tag{48}$$

We thus obtain the much simplier formula (For simplicity we use $\alpha_3 = 0$)

$$\frac{\langle P_2(L)/P_1(L) \rangle}{\langle 2\alpha_r L \rangle^2} = \frac{\pi}{4n_2 kL} \frac{\displaystyle\int_0^\pi \frac{\langle |\phi(\theta)|^2 \rangle^2 n_2 k \rho^3 \cos^4 \sigma d}{(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)^2} d\alpha}{\left\{\displaystyle\int_0^\pi \frac{\langle |\phi(\theta)|^2 \rangle \rho^2 \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} d\alpha\right\}^2}. \tag{49}$$

V. DISCUSSION OF THE SCATTERING CROSSTALK

The crosstalk for identical modes, equation (49), can be expressed approximately by very simple formulas. Let us assume that the Fourier spectrum of the core thickness variation is so narrow that only a narrow beam of radiation leaves guide 1. The radiation is centered around an angle α and fills a range $\Delta\alpha$. Provided this range of angles is sufficiently narrow and if we also assume that the Fourier spectrum is constant in the region contributing radiation and zero outside of this region, we can write (49) in the following simple form.

$$\left\langle \frac{P_2(L)}{P_1(L)} \right\rangle = \frac{\pi}{4n_2 kL} \frac{1}{(\Delta\alpha) \sin \alpha} \langle 2\alpha_r L \rangle^2 e^{-2\alpha_3 (R-2D)/\sin \alpha}. \tag{50}$$

(The angle α and the loss coefficients α_r and α_3 must not be confused.) We have reinserted the loss in the medium between the two guides in (50). We thus have the interesting result that the crosstalk depends on the square of the loss coefficient of the guided mode, on the loss in the medium between the waveguides, and on the width of the radiation lobe. The crosstalk decreases with increasing lobe width. In fact, by comparing (50) with (35) we see that we can let $\Delta\alpha$ become as small as

$$\Delta\alpha = \frac{\pi}{n_2 kL \sin\alpha}, \tag{51}$$

because in this limit (50) applies to a single plane wave coupling the two guides together.

In the other extreme it can be shown that (49) can be approximated quite well by the expression

$$\left\langle \frac{P_2(L)}{P_1(L)} \right\rangle = \frac{3}{4n_2 kL} \langle 2\alpha_r L\rangle^2, \tag{52}$$

if we assume that $\langle|\phi|^2\rangle = \text{const}$, which corresponds to the case of random scattering of radiation in all directions. We assumed $\alpha_3 = 0$ in (52). The error between the approximation (52) and the equation (49) is always less than a factor of two. This is not a bad approximation for an order-of-magnitude estimate. It is interesting to note that we obtain (52) formally from (50) by taking

$$\Delta\alpha \sin\alpha = \frac{\pi}{3}. \tag{53}$$

Since it is certainly reasonable to assign $\Delta\alpha = \pi$ to the case of isotropic radiation, we see that the crude approximation (50), which holds precisely only when $\Delta\alpha$ is very small, can actually be used to estimate the scattering crosstalk for all cases of interest.

It is furthermore important to note that the model of scattering used to derive our equations is very likely to be of no importance to the results. Since we could express the crosstalk in terms of the radiation loss of the guided mode and some features of the radiation spectrum we can be confident that our equation (50) holds at least to order of magnitude for any type of waveguide imperfection that causes radiation.

If we ask ourselves what relation the slab waveguide theory may have to the case of round optical fibers we can only draw some very general conclusions. The plane wave radiation produced by the slab waveguide does not diminish with distance. It is thus not surprising that the crosstalk formula (50) is independent of distance except for the obvious distance dependence of the loss term containing α_3. The radiation of a round optical fiber has the form of cylindrical waves so that we must expect that, in addition to the features expressed by (50), there will be an additional factor d/R multiplying the crosstalk expression. It is hard to say whether this adjustment will yield an expression that describes the crosstalk for round fibers to a fair approximation. However, it appears safe to assume that the crosstalk predicted by the slab waveguide theory is actually larger than the crosstalk for round optical fibers. We can thus consider the results of this paper as an upper bound of the crosstalk of round optical fibers. Any design features that are predicted on the basis of this theory are likely to be conservative if applied to the round fiber case.

We conclude our discussion with a numerical example. Let us assume that the length of the guides is $L = 1$ km. Taking $d = 1$ μm, we have $L/d = 10^9$. Next, we assume that $n_2kd = 10$ and that the radiation loss of the waveguides is 4 dB/km corresponding to $2\alpha_1L = 1$. In the absence of loss in the medium between the two guides ($\alpha_3 = 0$) we thus obtain from (52)

$$\left\langle\frac{P_2(L)}{P_1(L)}\right\rangle = 3/4 \times 10^{-10}. \tag{54}$$

The crosstalk is thus insignificant even without trying to isolate the guides from each other by providing loss in the medium between the guides. This example is actually quite representative. The assumed mode loss is typical for Rayleigh scattering losses in good solid materials. Since Rayleigh scattering is the theoretical limit of scattering loss, one should consider the crosstalk caused by this mechanism. Our conclusion is thus that Rayleigh scattering does not cause appreciable crosstalk between two optical fibers. The situation worsens somewhat if the radiation is bunched. However, we can stand a much smaller value of $\Delta\alpha$ than $\Delta\alpha = \pi$ before crosstalk becomes a problem. It appears that the only real danger is the possibility of a pure sinusoidal thickness variation of the fiber core such as may be caused by a systematic flaw in the fiber pulling process. Such an imperfection can cause serious crosstalk even if the radiation loss attributable to this mechanism is quite small. We see from (35) that a purely sinusoidal distortion of the waveguide core causing only $2\alpha_1L = 10^{-2}$ can cause a crosstalk ratio of 2.5×10^{-5}. The theory of Reference 2 predicts that a loss of $2\alpha_1L = 10^{-2}$ is caused by a sinusoidal core thickness variation with an amplitude as small as $\approx 10^{-5}$ μm.

VI. CONCLUSIONS

Using the model of two slab waveguides with core thickness variations we have computed the crosstalk that is caused by light scattering in one guide and reconversion of this scattered light in the other guide. Since the results of the theory can be expressed in terms of the radiation loss that is caused by the same scattering mechanism, it is assumed that the theory has general validity. It is also concluded that the crosstalk of round optical fibers is smaller than that of two slab waveguides so that the theory can be considered as an upper bound on the crosstalk between fibers.

Scattering crosstalk can be substantial if both waveguides have a sinusoidal imperfection of equal mechanical frequency that persists throughout the entire length of the guides. If such an imperfection should exist, it would be necessary to isolate the waveguides from each other by providing a surrounding medium with loss. However, it appears unlikely that such a systematic sinusoidal deformation should exist since it would require that a definite mechanical period with a frequency inside of the range given by (23) should have been generated by the manufacturing process.

Any other imperfections that are of a statistical nature are much less serious even though they might lead to radiation patterns with narrow radiation lobes. It appears likely that scattering crosstalk will not present a problem particularly since the radiation losses of the waveguides must be kept low in order for the guides to be useful. Unless a definite sinusoidal imperfection does exist it seems unnecessary to provide loss in the medium separating the two waveguides in order to reduce the scattering crosstalk.

REFERENCES

1. Marcuse, D., "The Coupling of Degenerate Modes in Two Parallel Dielectric Waveguides," B.S.T.J., this issue, pp. 1791–1816.
2. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10, part 3 (December 1969), pp. 3187–3215.
3. Marcuse, D., "Radiation Losses of Dielectric Waveguides in Terms of the Power Spectrum of the Wall Distortion Function," B.S.T.J., *48*, No. 10, part 3 (December 1969), pp. 3233–3242.
4. Rowe, H. E., *Signals and Noise in Communications Systems,* New York: D. Van Nostrand Company, 1965.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Dynamic Channel Assignment in High-Capacity Mobile Communications Systems

By D. C. COX and D. O. REUDINK

(Manuscript received January 4, 1971)

Computer simulations of high-capacity mobile radio systems using different channel assignment philosophies are described. These simulations initiate call attempts and move vehicles about randomly according to prescribed statistical distributions. Base stations and radio channels are assigned to serve mobiles and system operating statistics are accumulated. Relationships between systems parameters obtained from the simulation are presented. Performance of a dynamic channel assignment system (DYNSYS) *which has all channels available at all base stations is compared with performance of a fixed-channel assignment system* (FIXSYS) *which reserves channel subsets for use at specific base stations.*

For uniform spatial distributions of call attempts and 40-channel systems with reuse intervals of four base station radio coverage areas, the DYNSYS *outperforms the* FIXSYS *at blocking rates up to 13 percent. For example, at a 3 percent blocking rate the* DYNSYS *provides 20 percent more calls "on" in the system.*

I. INTRODUCTION

High-capacity mobile radio systems with large numbers of radio channels have been proposed to relieve the over-crowded conditions existing in mobile radio communications today. In order to make efficient use of frequency spectrum, radio coverage from base stations would be limited to small zones and channels would be reused several times within a particular urban area.[1–4] Figure 1 depicts such a channel-reuse situation along a single line of base stations. Computer control of such systems certainly would be necessary. Bounds on the behavior of simplified systems have been obtained previously[2] but the

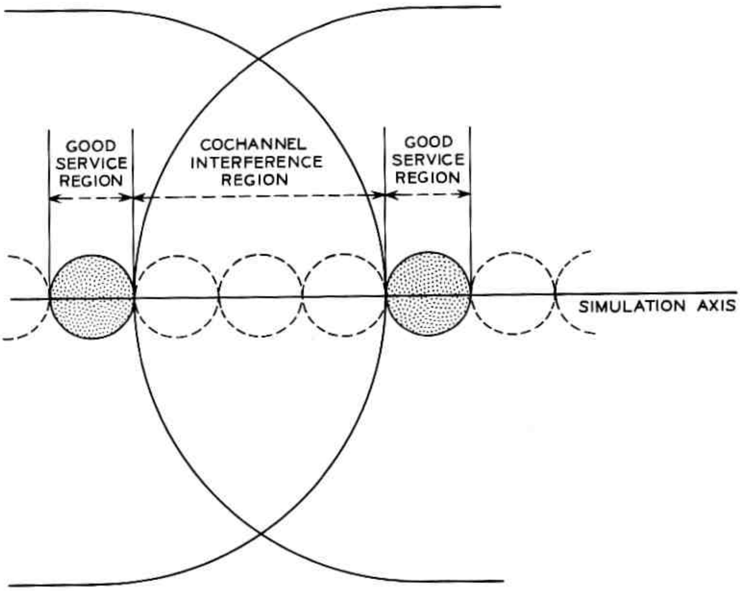


Fig. 1—Illustration of channel reuse.

problem of determining the performance of these complex systems by rigorous analytical methods does not appear tractable. Since little is known about the overall performance of dynamic channel assignment systems computer simulations were run for postulated systems whose parameters were chosen somewhat arbitrarily; however, these parameters (e.g., coverage cell size, call durations, number of channels, velocities, etc.) were selected within the wide range of values which might be experienced in an actual system. This paper presents some performance characteristics of two dynamic mobile radio systems obtained by simulation on a high-speed digital computer. This provides a means for comparing two basically different system disciplines. It should be noted that more work is required to simulate performance characteristics for actual system configurations designed to operate optimally within a specific set of conditions. No consideration has been given in this study to either the time required to develop the different system configurations or the economic factors necessary for final system design.

Channel assignment in the first system is made on a dynamic basis, that is, a channel is assigned to a requesting vehicle and its covering base station by a method which considers the channels in use at that instant at that base station and at base stations within some region surrounding the vehicle. Once a channel is assigned it will generally stay with the vehicle for the duration of the call. This may require the assignment of other base stations to that channel to cover the vehicle as it moves about.

Channel assignment in the second system is made on the basis that only a fixed subset of the channels available to the radio system are available for use at a given base station. The essential difference between these systems is illustrated in Fig. 2 where M-N indicates the channels available at each base station for a system example with 15 channels and a reuse interval of every third base station.

The first part of the paper describes a computer simulation of both systems. The second part presents performance characteristics for two different spatial-demand profiles for 40-channel systems. Throughout the paper a channel is defined as a two-way duplex radio channel.

II. OVERALL MOBILE RADIO SYSTEM OPERATION

This section describes briefly the functions which must be performed in the operation and control of a high-capacity mobile radio system and indicates which portions of the system are simulated. The major characteristics of a system are illustrated in Fig. 3. The box labeled "call initiation and vehicle motion" may be regarded as the driving function. A radio system must react to the situations created by this driver. These situations involve such parameters as the rate of occurrence and the distribution of locations of new call-attempts. Because coverage areas will be crossed as vehicles move, direction of travel and speed distributions are other influencing parameters controlled by the driver.

Since the radio path is usually not line-of-sight, propagation effects must be taken into account. The goals in this case are to assign the best base station or base stations to provide good service to the mobile and to make efficient use of the radio spectrum. The way of dealing with propagation effects can be quite varied and is closely tied to the

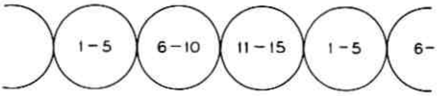


Fig. 2—Fixed-channel assignment system example.

philosophy under which base station assignments are made. One approach could be to attempt to force propagation effects to have a minimal influence by restricting the use of signal from each transmitter out to a fixed radius and then providing a margin to handle the statistical fluctuations. This would produce strong signals beyond that radius in some regions and thus would be bad from the standpoint of frequency reuse. Another extreme would be to have stored the exact coverage areas from every base station. In this case all propagation effects are known. The problem here is that of being able to store, retrieve and make intelligent decisions based on the information in a short period of time. The criterion for selection of the proper base station to serve a mobile is to provide the maximum channel reuse consistent with good service.

When a mobile requests service, a channel must be selected to serve it from the selected base station. Choosing the most distant channel currently being used may not always be the proper choice. This approach may minimize cochannel interference problems, but would be poor from the spectrum reuse point of view.

Since many of the important performance parameters of a mobile radio system cannot be obtained analytically and competing systems would be far too costly to build up in order to make the comparisons, computer simulation can be a powerful tool in providing answers for system design. At this stage the simulation to be described has simplified some of the operations illustrated in Fig. 3, but was written with the flexibility necessary for greater sophistication. Presently the model is one-dimensional. This causes some difficulty in extrapolating results to a two-dimensional mobile radio system but the results have direct application to a system laid out along an expressway or possibly to air-ground telephony. It also serves as a valid framework within which to make comparisons of different channel assignment schemes. Call initiations and vehicle movements are assumed to be random processes which will be described in detail in Section III. The propagation path between base and mobile was assumed to be such

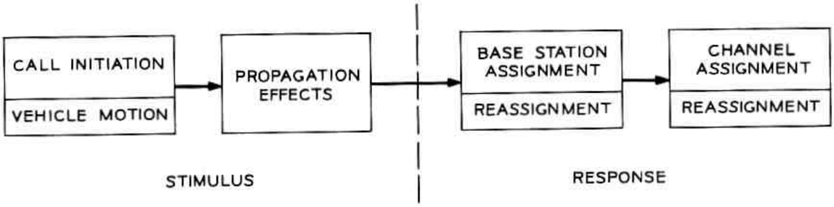


Fig. 3—Functions of a high-capacity mobile radio system.

that there is a smooth variation ($1/R^{\alpha}$, for example) of signal strength as the vehicle moves with respect to the base station. This assumption is of course unrealistic in many cases, but it permits the study of channel assignment methods separately from the effects of random propagation. It is assumed that there exists a distance interval over which the same frequency may not be reused. For each mobile it is further assumed that at any given instant there is only one base station which will provide a signal greater than the required thermal noise threshold and simultaneously a signal-to-cochannel interference ratio greater than some acceptable value. A discussion of the simulation of the channel assignments and of the system updating is presented in Sections IV and V.

III. CALL INITIATION AND VEHICLE MOVEMENT IN THE SIMULATIONS

Call initiation and vehicle movement in the simulation are handled in a single subroutine which is used to drive several different operating systems subroutines. The flow-chart in Fig. 4 illustrates the major sequences in this driver subroutine. For each system cycle, the subroutine steps through each subscriber in sequence and checks his activity status. Depending on events which have occurred in previous system cycles, a subscriber will be either "on" in the system or "off."

If the subscriber is not "on" in the system, the right-hand branch of the chart is followed. The first step along this branch is to generate a random number which determines whether or not a call-attempt will be made. If the result is not to make an attempt, that fact is noted and the subroutine steps to the next subscriber. If the result is to attempt a call, then another random number is generated from a uniformly distributed set to determine the subscriber's location in the one-dimensional universe of the simulation. This method of generating call-attempts produces call arrivals which are Poisson processes in time at each base station and are uniformly distributed in space. This uniform distribution in space can be weighted to produce any desired spatial distribution by using counters which reject certain attempts at some locations. A velocity is assigned to the subscriber for each call-attempt produced. This velocity is assigned from a random number set which has a prescribed density function. The simulation currently uses a truncated Gaussian velocity distribution. If the call-attempt is located near an "edge" of the one-dimensional space, it is checked against criteria used to compensate for the effects of the edges and simulate an infinite system. This compensation then

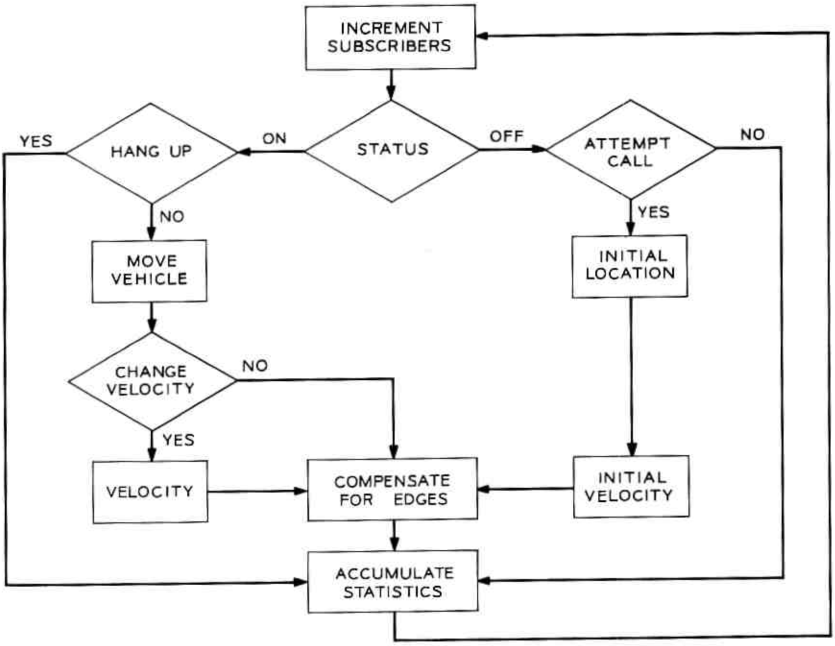


Fig. 4—Driver program.

cancels some attempts near the edges before they actually enter the system.

The left-hand branch of Fig. 4, starting at the status block, indicates the procedure followed for subscribers that are already "on" in the system. The first check is whether or not the subscriber hangs up. Call terminations are determined in a way that permits specifying the density function for call durations. The simulation currently uses a truncated Gaussian call duration distribution. Subscribers terminating calls are accounted for in the system statistics and flagged for the system operating subroutine. A subscriber who does not hang up is moved according to his predetermined velocity and the length of time between system cycles. His new location is then passed on to the system operating subroutine.

After a vehicle is moved, a random number is generated which determines if the velocity should be changed. If a velocity change is indicated, the new velocity is assigned at random following the same procedure used to determine vehicle velocities for call attempts. After vehicles are moved, those near the "edges" are again checked against an edge compensation criterion and appropriate statistical values are accumulated.

IV. A DYNAMIC CHANNEL-ASSIGNMENT OPERATING SYSTEM

In the current simulation a single operating system subroutine determines the appropriate base station to serve a call-attempt, assigns a radio channel if one is available or refuses service otherwise, and updates channel assignments as vehicles move about. The flowchart in Fig. 5 illustrates the major sequences in this operating system. For each system cycle, this subroutine steps through each subscriber who is either "on" in the system or is attempting a call and checks his activity status.

4.1 *New Call-Attempts*

New call-attempts are processed along the center branch of the flowchart. The first step in processing an attempt is to determine

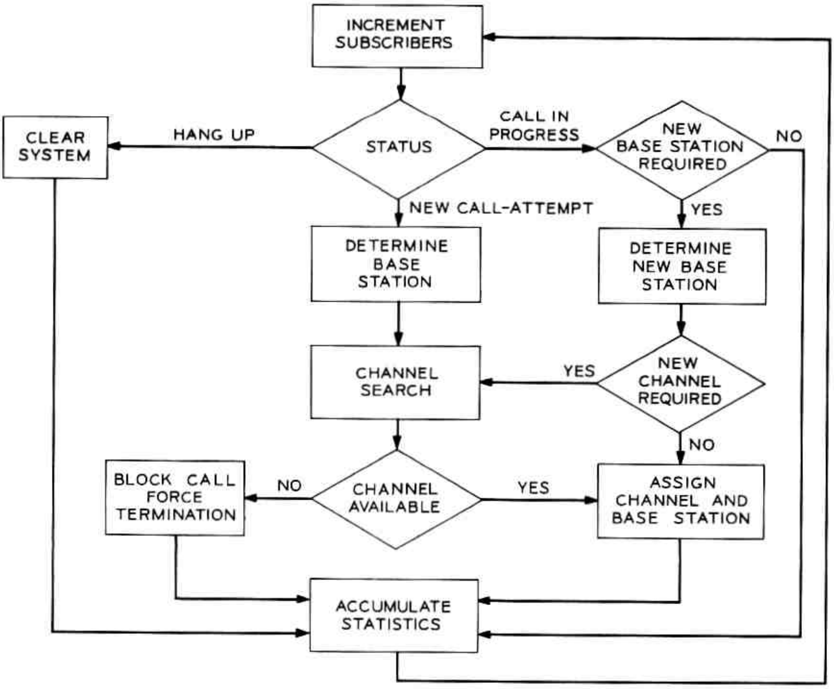


Fig. 5—System control.

which base station (or stations) should be used to serve the call. Because of the simple propagation characteristics assumed so far in the simulation, the base station nearest to the subscriber has the strongest received signal from the subscriber and will be the only one with adequate signal strength to serve. It is the station assigned to serve the call-attempt. If the base station is near an "edge" of the one-dimensional space, additional checks of the calls "on" at the base station are made. Some call-attempts near the edges of the one-dimensional space are refused service at this point if the total calls "on" at the base station to be assigned exceed some value determined by the edge compensation criterion.

The next step in processing a new call-attempt is to search for a radio channel which satisfies some channel assignment criteria. The simulation will permit many different assignment criteria to be investigated. Because of the simple propagation model assumed at this time, the current channel assignment criterion assumes that the base stations separated by greater than a certain reuse interval will not interfere with each other when they use the same radio channel (i.e., there will be no cochannel interference). The current dynamic channel assignment criterion assumes that all radio channels are available at all base stations. It also assumes no limit on available radio equipment at the base stations. This last assumption is somewhat unrealistic but it will permit the determination of upper bounds on actual radio equipment requirements. These equipment limits will then be included in later simulations. Channel search is done channel by channel starting with Channel 1. If the channel being checked is in use at the base station to be assigned, then that channel is rejected and the next channel is checked. If a channel is not in use at that base station, a check is made base station by base station on both sides of that base station. If the channel is in use at a base station closer than a reuse interval away from the base station to be assigned, the channel is rejected and the next channel is checked starting again at the beginning of the channel check procedure. If the channel is not in use at any base station less than a reuse interval from the base station to be assigned, then further checks on that channel are made and it is compared with other channels, if any, which also have met this reuse interval criterion. If, at any time during the channel searching, a channel is found which is not in use at any other base station in the system, that channel is flagged for assignment and the channel search is terminated. Otherwise, channels which meet the reuse criteria are intercompared to find the "most desirable one." The most desirable one will depend on the system operating parameter which is to be minimized, but the optimum choice has not been determined yet. For this simulation, the channel flagged for assignment is the one whose next "in use" base station is closer to the base station to be assigned than any of the other channels satisfying the reuse criterion but is at least a reuse interval plus an integer number, J, of base stations away from the base station to be assigned. The use of $J = 0$ packs assignments as close as possible on one side of the base station to be assigned but increases forced terminations of calls when boundaries are crossed. The use of $J = 1$ relieves the forced terminations slightly but results in a slightly greater separation between base stations using the same channel. In any case, if the only channel satisfying the reuse criterion is "on" at a base station exactly one reuse interval from the base station to be assigned, that channel is still flagged for assignment. This channel search strategy never allows a channel to be assigned at base stations less then a reuse interval apart, but attempts to pack channel assignments close together for efficient channel usage.

If, after checking all radio channels in the system, no available channel is found, the call-attempt is refused service (the call is blocked) as indicated on Fig. 5. Blocked call-attempts are completely removed from the system and no other attempt to serve them is made. On the other hand, if an available channel is found, it is assigned to the vehicle and the base station to be assigned and the call setup is complete. Statistics are accumulated and the next subscriber checked.

4.2 *Calls in Progress*

The processing of calls in progress can be followed by returning to the status check block at the top of Fig. 5 and following the right-hand branch. The first step along this branch checks the suitability of the current base station assignment. This is done by checking to see if the vehicle has moved (during its last pass through the driving subroutine) closer to some other base station than the one currently serving it (and thus has a stronger signal at the other base station). If it has not moved into such a situation, then its current base station and channel assignment are acceptable and no further action is taken except to account for the call in the system statistics. If, however, the vehicle has moved closer to another base station, that "now closer" station must be used to serve the call. As indicated on the flowchart, the assigned channel is then checked at the new base station to be used and at all base stations within a reuse interval of the new one. If the channel passes this check, that is, is not in use at those stations, the new base station and old channel are assigned to serve the call and the old base station is cleared of the call. If the old channel is in use within a reuse interval of the new base station, then a whole new channel search is initiated identical to the one used for new call-attempts. In the case of a call in progress, the only difference is that, if a substitute channel is not available, the call is forced to terminate instead of being refused. Only one attempt is made to switch channels for such a call in progress. In either case, the old base station is cleared of the call, statistics are accumulated, and the subroutine steps to the next subscriber.

4.3 *Call Terminations*

The only actions required to process call terminations are to clear the channel at the assigned base station and accumulate the appropriate statistics. These actions are indicated on the left-hand branch of Fig. 5.

The dynamic channel-assignment operating system used in the simulation and just described follows a simple, straight-forward operating procedure. No doubt, more efficient channel usage could be achieved if the system were continuously checked call by call and channels changed on calls in progress to more optimally pack the calls in the system and free more channels for new calls. Similarly, more efficient initial channel assignments could be achieved if attempts were made to change channels of calls in progress to try to free a channel for serving a call attempt which would now be blocked. Fewer new calls would be blocked if new callers were given several attempts over a short time period to obtain a channel rather than the one attempt allowed in the current simulation.

V. A FIXED-CHANNEL ASSIGNMENT OPERATING SYSTEM

A single subroutine simulates the fixed-channel assignment operating system. This subroutine, which is similar to the dynamic channel assignment subroutine, determines the appropriate base station to serve a call attempt, assigns a radio channel or blocks the attempt depending on the availability of a radio channel, and updates channel assignment as vehicles move about. Fig. 5 also illustrates the major sequences in this operating system but some of the block functions are different for this system. Again, for each system cycle, the subroutine steps through each subscriber and takes appropriate action.

5.1 *New Call-Attempts*

Processing of new call-attempts for the fixed-channel system is identical to that for the dynamic system down to the channel search procedure. The base station assigned to serve a call-attempt is the one nearest to the subscriber as discussed for the dynamic system.

In the fixed-channel system, a subset of the radio channels available to the system is permanently reserved for each base station. The subsets are reused at base stations separated by the reuse interval described previously. The number of channel subsets is the same as the number of base stations in a reuse interval, *I*. Every *I*th base station uses the same channel set. (See Fig. 2.) The number of channels, C_s, in a subset is

$$C_s = C_t/I$$

where C_t is the total number of channels available to the whole system. Thus, each base station has C_s channels reserved for its exclusive use. Channel search, then, only involves searching through the C_s channels of the subset reserved for the base station to be assigned. If a channel from this subset is not in use, then it is assigned to serve the call. If no channel from this subset is available, the call attempt is refused service (the call is blocked).

5.2 *Calls in Progress and Call Terminations*

Processing of these calls also proceeds as in the dynamic system. A check is made to see if the vehicle has moved so that a different base station is required to provide service. If a different base station is not required, no system action is necessary. If, however, a different base station is now required, that base station is determined and for this fixed-channel system the only recourse is another channel search. The channel assigned at the previous base station is not available at adjacent stations so the alternative, available to the dynamic system, of continuing its use is not possible. This alternative, indicated on Fig. 5, is thus not applicable to the fixed system. The channel search at the new base station to be assigned is identical to the one described for new call-attempts. In this case, if no vacant channel is found, the call is terminated.

Subscribers who hang up are cleared from the system and statistics are accumulated as in the dynamic system.

VI. THE DRIVING PARAMETERS

The operating systems were tested for a 24-base-station 40-radio-channel network where the base station separation was assumed to be 2 miles; the minimum reuse spacing was 8 miles. This simulation assumed 1000 subscribers and operated with a 10-second cycle time, that is, within 10 seconds (in simulated time) the status of each subscriber was examined. In a true system only subscribers making or attempting calls are of interest. For simulation purposes, however, it was found more convenient to gather statistics and create demand by cycling through all subscribers. Statistics were taken from only the central 14 base stations to avoid any effects from the edges of the finite system.

The spatial location of call-attempts will vary with traffic intensity. To provide a reference system, Section 7.1 considers a uniform spatial distribution of call-attempts. In Section 7.2, an example of a non-uniform call-attempt distribution is simulated.

Call-attempts occur at random points in time (i.e., call-attempts are a Poisson process), therefore, the time difference between call-attempts should have an exponential density function with its mean equal to its variance. A histogram of the time differences for a typical data run is plotted in Fig. 6 along with a theoretical exponential curve.

The distribution of the duration of mobile telephone calls for a high-capacity system is unknown. There is, however, some minimum length of time which a mobile caller will hold a channel. For example, even a call where no one answers will create a demand on the mobile radio system of perhaps 30 seconds. The distribution of call durations for this simulation was assumed to be a truncated Gaussian with a minimum call of 30 seconds and a maximum call of 10 minutes. The mode call time was chosen to be 1-1/2 minutes with a standard deviation of 1 minute. The mean for this assumed distribution is 103.5 seconds. Figure 7 shows both a theoretical curve of the density function and a histogram of call lengths taken from a 1000-cycle data run.

The distribution of vehicle velocity was also chosen to be a truncated Gaussian. In this example, the maximum speed was 60 miles per hour; the distribution had a zero mean and a variance of 30 miles per hour. A theoretical velocity distribution and a histogram from the simulation are plotted in Fig. 8.

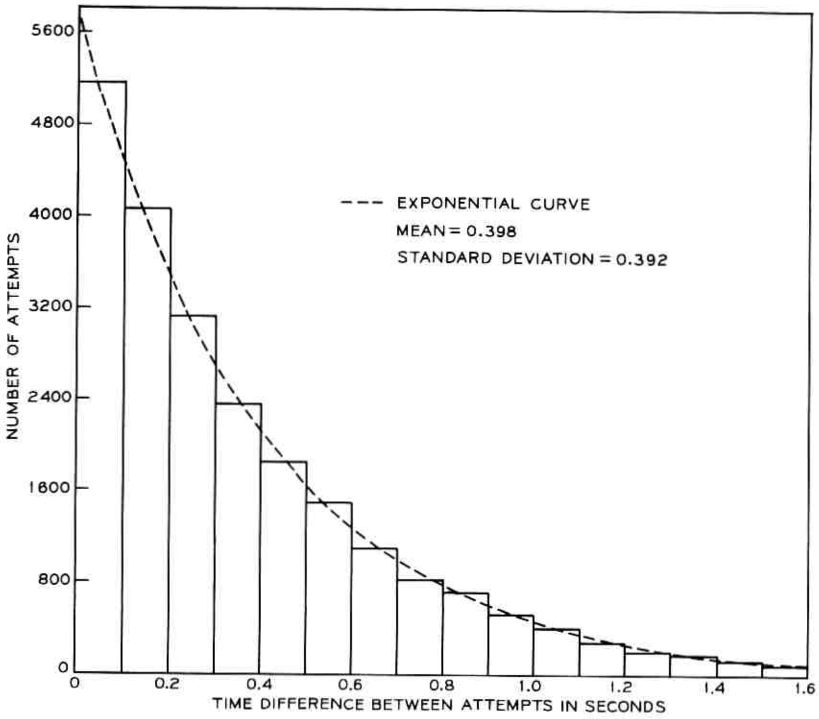


Fig. 6—Time difference between call-attempts histogram.

VII. EXAMPLES OF SYSTEM PERFORMANCE

7.1 *Uniform Call-Attempts*

The performance of a mobile radio system is measured knowing the interrelations between such parameters as the call-attempt rate, percent of call-attempts blocked, number of calls being handled, number of vehicles crossing cell boundaries, and number of calls being forced to terminate.

The location of call-attempts for this part of the simulation was chosen to be a flat distribution in space with appropriate edge tapering. Figure 9 shows the number of call-attempts at each base station for two typical data runs. Since it was desirable to take averages over this ensemble of 14 base stations the number of system cycles was

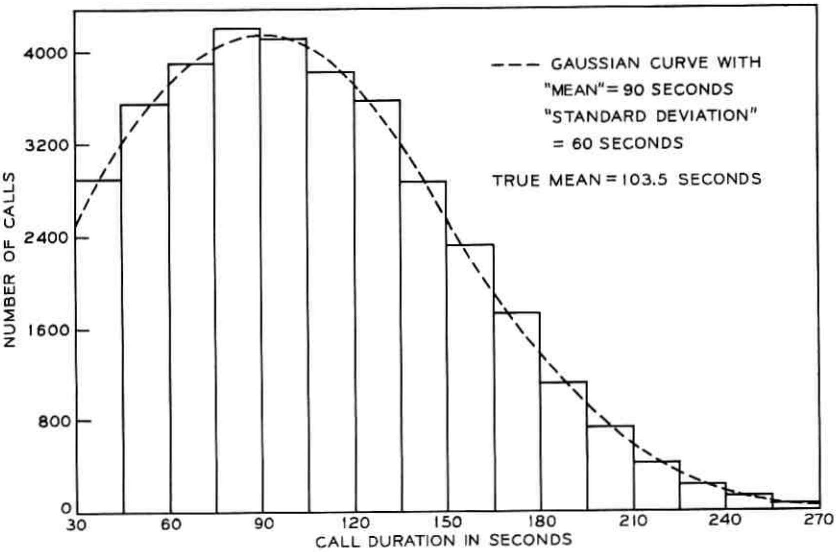


Fig. 7—Call duration histogram.

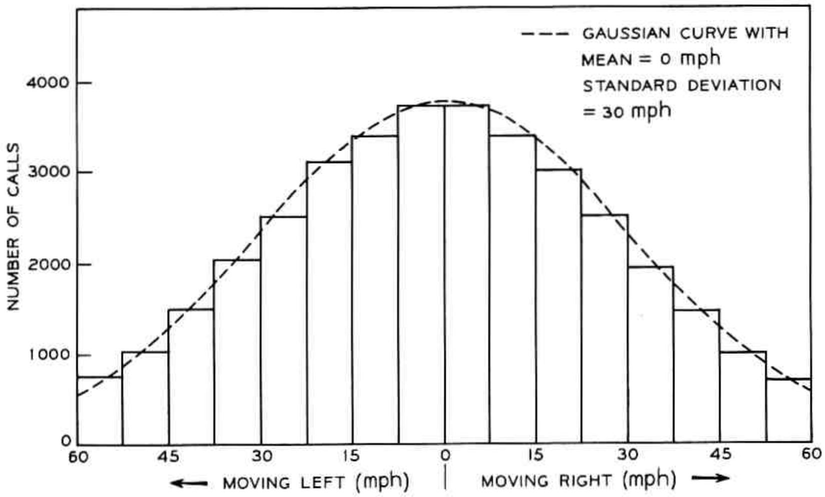


Fig. 8—Velocity histogram.

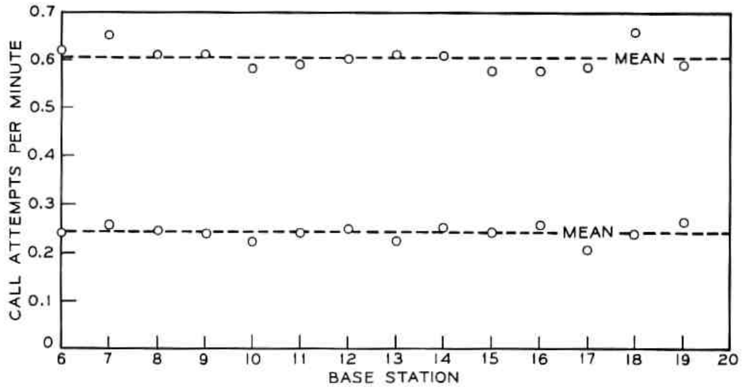


Fig. 9—Average call-attempt rate at 14 base stations for two attempt rates.

increased until the deviation from the mean number of call-attempts by an individual base station was small.

As an aid in understanding the details of the system operation, a special program was written to view at regular instants of time the status of the mobile radio system. Specifically, a computer printout was generated which indicates the locations where radio channels are being used and prints a symbol unique to that channel. The appropriate symbol for each of the 40 two-way channels is shown in Table I. In addition, two other special symbols are used. The $ is printed if, during the time interval under consideration (usually taken to be 10 seconds), a request for service was initiated but no channel was available and service was refused. The $ is printed at the location where the call was attempted. An * is printed when a vehicle crosses a cell boundary and the user is forced to terminate his call prematurely because a channel is no longer available for him.

Figure 10 illustrates the dynamic channel assignment system, DYNSYS, activity for 500 seconds for base stations 11 through 16. In the example, about 10 percent of the call-attempts are being blocked. The circled track 1 shows a typical call within the coverage area of base station 13, where the caller is assigned channel 39, designated by an open parenthesis. After 80 seconds a change in velocity occurred and the call terminated after 140 seconds. Recalling 2-mile base station spacing and 10-second time intervals, the velocity can be roughly calculated from the figure. In this example, the initial velocity was about 45 miles per hour and the final velocity was about 30 miles per hour in the opposite direction. Track 2 in this same figure shows a

TABLE I—CHANNEL IDENTIFICATION

Channel Number	Symbol
1	A
2	B
3	C
4	D
5	E
6	F
7	G
8	H
9	I
10	J
11	K
12	L
13	M
14	N
15	O
16	P
17	Q
18	R
19	S
20	T
21	U
22	V
23	W
24	X
25	Y
26	Z
27	1
28	2
29	3
30	4
31	5
32	6
33	7
34	8
35	9
36	0
37	+
38	−
39	(
40	)

vehicle with channel C crossing a boundary and maintaining the same channel. That is, the channel assigned to the call "floats" with the vehicle. When the system activity is quite high, not all callers crossing boundaries can keep the same channel. Track 3 shows a vehicle, originally assigned channel G, reassigned to channel S as it crosses the boundary between base stations 11 and 12. Note, the channel change was required because channel G was in use (circle 3′) at base station 15 which is less than a reuse interval from base station 12.

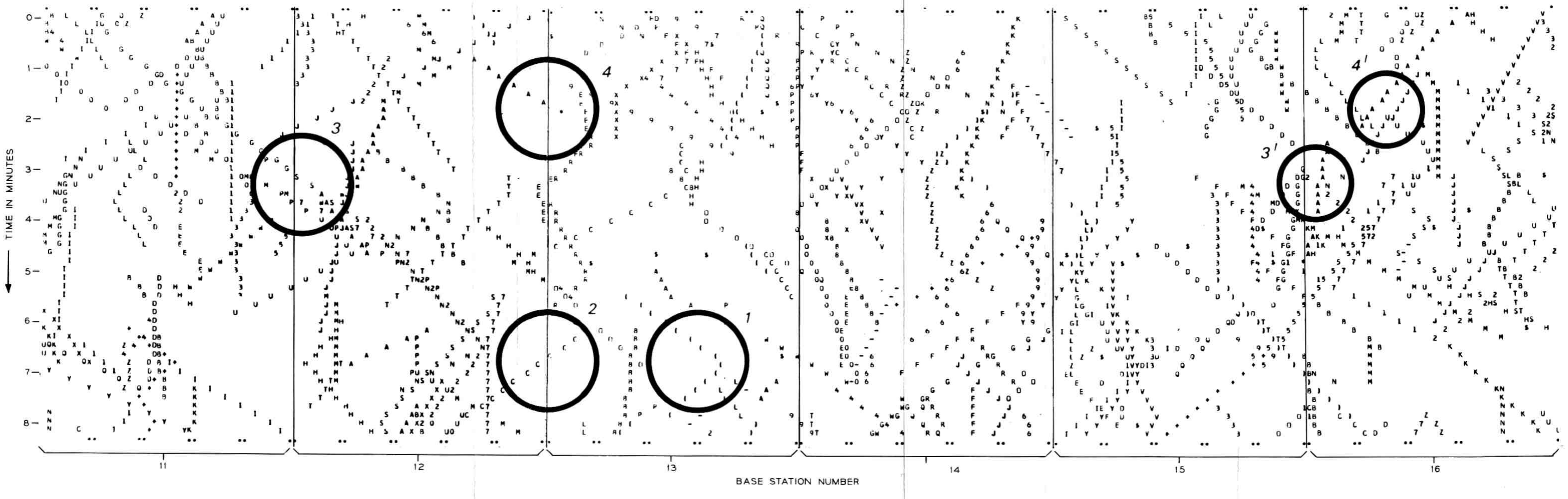


Fig. 10—Examples of dynamic channel assignment system operations.

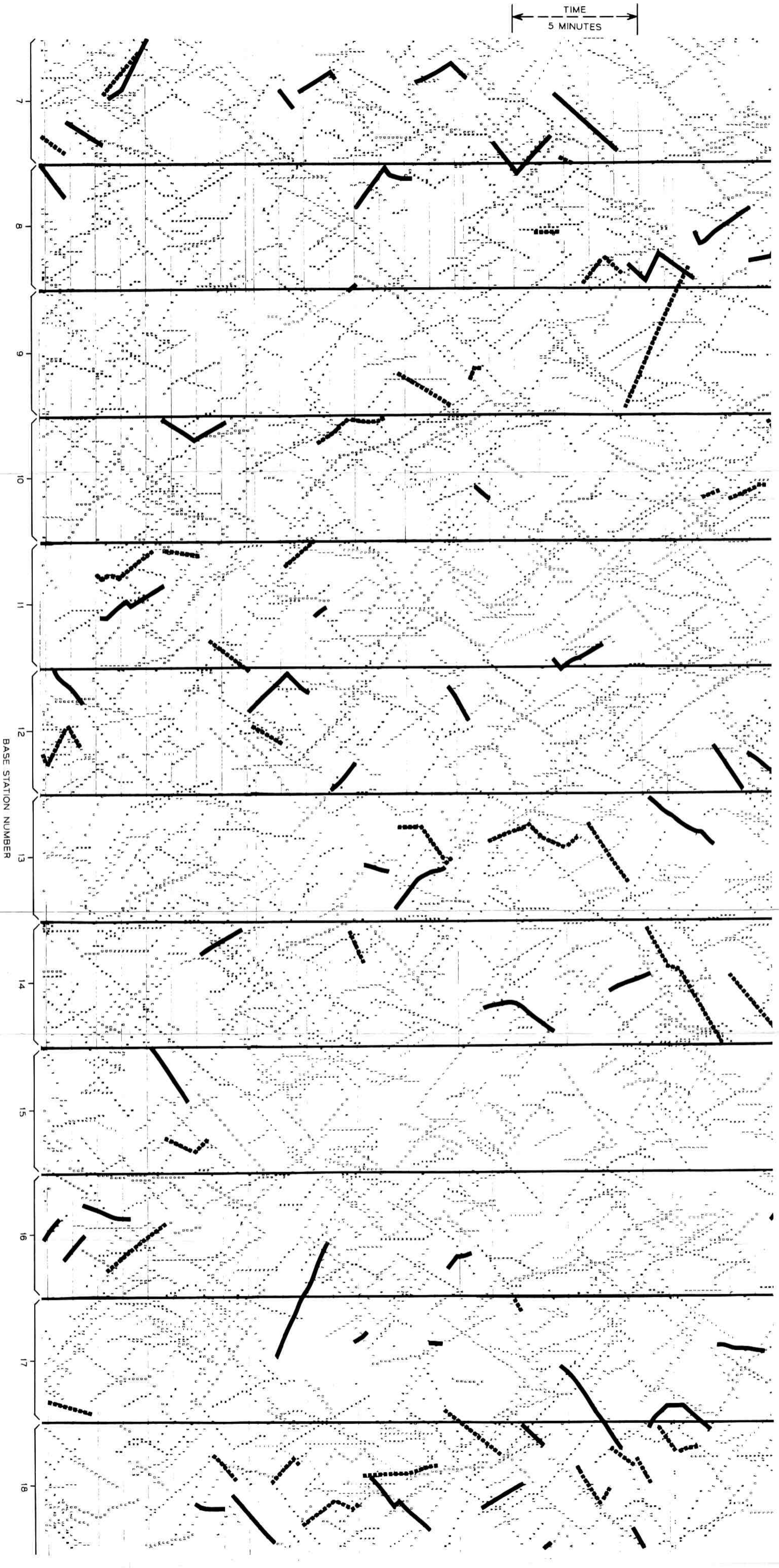


Fig. 11—Examples of channel reuse with dynamic channel assignment.

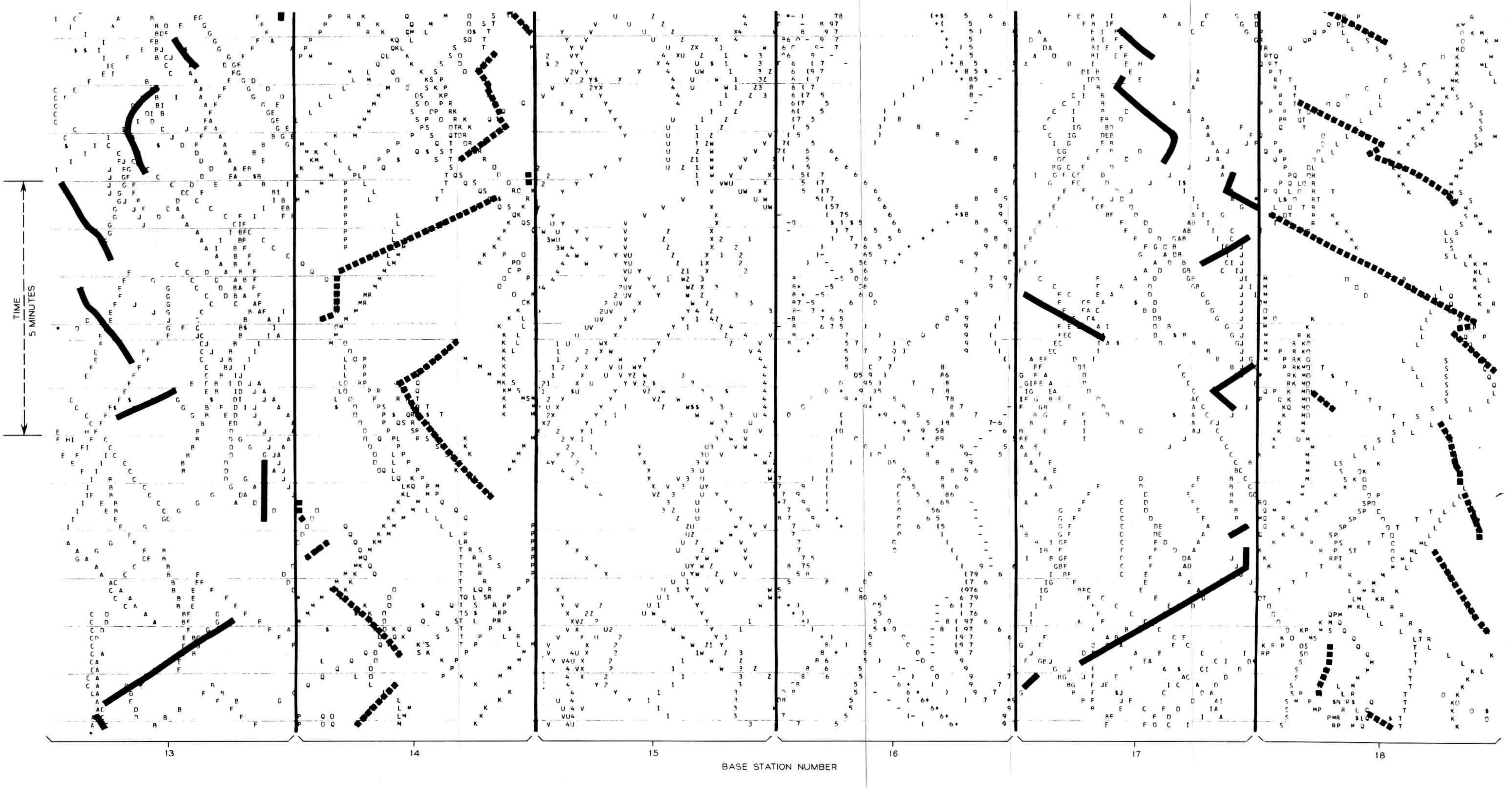


Fig. 12—Examples of channel reuse with fixed-channel assignment.

Track 4 shows a forced termination (of channel A) occurring at the boundary between base stations 12 and 13 because the channel is in use at base station 16 (circle 4′), and no other channels are available.

In the DYNSYS, any channel is available at any base station. In the fixed-channel assignment system, FIXSYS, with a reuse interval of 4, the 40 available channels were apportioned as shown in Table II. In the FIXSYS, vehicles are not allowed to keep frequencies when crossing boundaries. Therefore, a track corresponding to track 2 in Fig. 10 is not possible.

Channel reuse for the two types of systems is illustrated in Figs. 11 and 12. In the DYNSYS, calls are served on a first come first serve basis regardless of location. This system must hunt for an available channel which is not being used within a reuse interval on either side of the new caller. The highlighted tracks of Fig. 11 illustrate system usage of channels X and H during a time period of $\frac{1}{2}$ hour. The FIXSYS usage of channels H and N is illustrated in Fig. 12.

As the systems were running, data were taken on operating parameters such as the number of call-attempts, hang-ups, boundary crossings, etc., which occurred. These performance statistics were accumulated over runs of 1000 system cycles. Blocking is expressed as the ratio of call-attempts not finding a channel to the total call-attempts for a given time period. Forced terminations, which occur when a vehicle crosses a radio coverage boundary and cannot find an available channel, are expressed as the ratio of calls forced off to calls served for a given time period. Calls-keeping-channel are expressed as the ratio of calls keeping the same channel when crossing a coverage boundary to the total number of boundary crossings for a given time period. This parameter indicates the degree to which channel coverage "floats" with vehicles.

Figures 13 and 14 characterize the number of calls "on," blocking, forced terminations, and calls retaining their channel as a function

TABLE II—CHANNEL ALLOCATION FOR FIXED-CHANNEL ASSIGNMENT SYSTEM

Base Station	Channels Available
1, 5, 9, 13, 17, 21	1–10
2, 6, 10, 14, 18, 22	11–20
3, 7, 11, 15, 19, 23,	21–30
4, 8, 12, 16, 20, 24	31–40

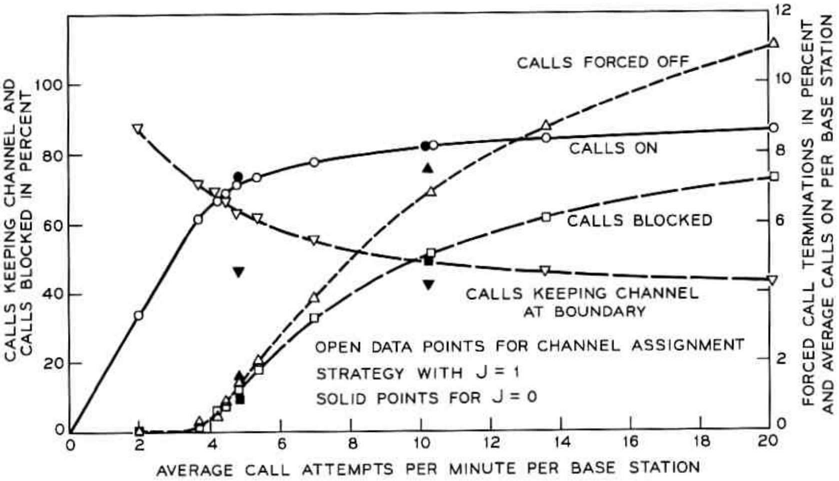


Fig. 13—Performance of a dynamic channel assignment system (flat loading).

of the call-attempt rate for the DYNSYS and FIXSYS. The values are averages per base station and the spatial call-attempt distribution is flat over all base stations.

The channel assignment philosophy currently used in DYNSYS selects as first choice a channel being used five base stations (a reuse interval plus one base station) on one side or the other of the new attempt location. In Fig. 13 the solid points are for two attempt rates where the first choice for channel assignment is a channel which is being used only four base station intervals (one reuse interval) away. As expected, the calls "on" were increased, the blockage was lowered slightly, forced terminations increased, and fewer vehicles kept the same channel after crossing a boundary.

The calls-keeping-channel curve shows that at low loading most vehicles are able to maintain their channels as cell boundaries are crossed while at very high loading, channel usage is packed more nearly at reuse intervals apart and fewer vehicles can maintain their originally assigned channel. Even at high loading, however, about half of the vehicles keep their channels as boundaries are crossed. Note, this parameter is the one most affected by the different assignment philosophies (preferred 4- or 5-station separation).

Figure 14 presents similar performance curves for the FIXSYS. Since channels must always be switched at boundary crossings the number of forced terminations is increased over the DYNSYS. When no motion is allowed and call durations are assumed to be exponentially distributed, it is possible to calculate (given the average call duration) from trunking formulas a theoretical expression for the blocking probability versus the number of calls "on" at a base station. The FIXSYS simulation, which allows motion and assumes truncated Gaussian call time distributions, agrees almost exactly with the trunking formula curves (Calls On and Calls Blocked) in Fig. 14 and Fig. 15 (dashed curve). Under the current operating conditions, motion and call time distributions have negligible effect on these two parameters.

The DYNSYS performance is compared to the FIXSYS performance in Fig. 15. In the low blocking region (below 10 percent) which is where a radio system should normally operate, the DYNSYS handles more calls than the FIXSYS. It provides about one extra call per base station (20 percent increase) at 3 percent blocking. At high blocking rates, the FIXSYS operates more efficiently; the reason is that this system can achieve a maximum number of calls "on" when there is always a call waiting at each base station. The DYNSYS could, of course, be programmed to switch over to this mode of operation as the system saturates.

As shown in Fig. 9, call-attempts were chosen to be a flat distribution in space. The corresponding reaction to the call-attempts is the

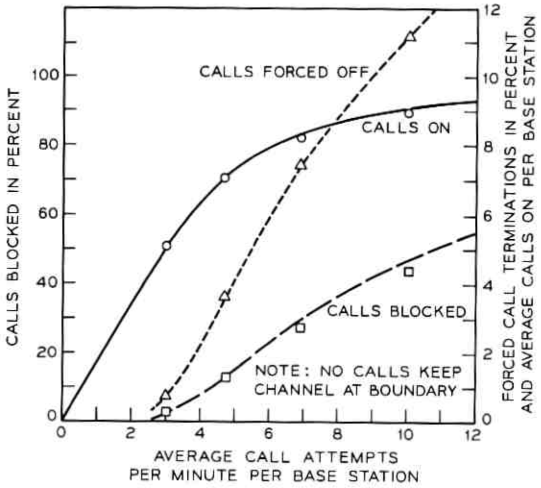


Fig. 14—Performance of a fixed-channel assignment system (flat loading).

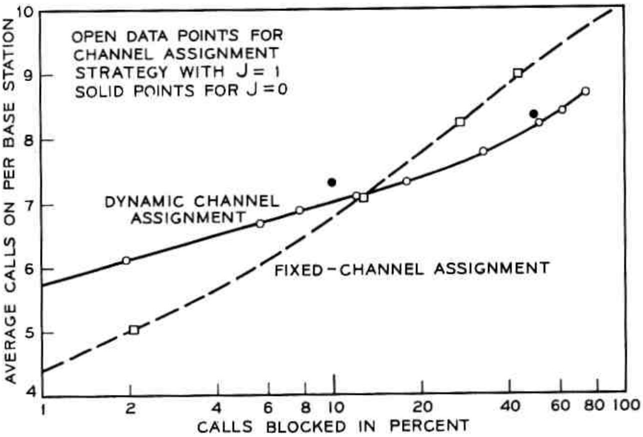


Fig. 15—Channel usage (flat loading).

average-calls-on distribution which is plotted in Fig. 16 as a function of base station for two demand rates for the DYNSYS.

The DYNSYS simulation allowed for the possibility that all channels could be used simultaneously at any base station. A histogram of the actual number of calls "on" for two different attempt rates is shown in Fig. 17a and b. It is immediately obvious that for a flat spatial distribution of call-attempts, halving the number of radio transmitters would not affect the system performance, since the number of channels

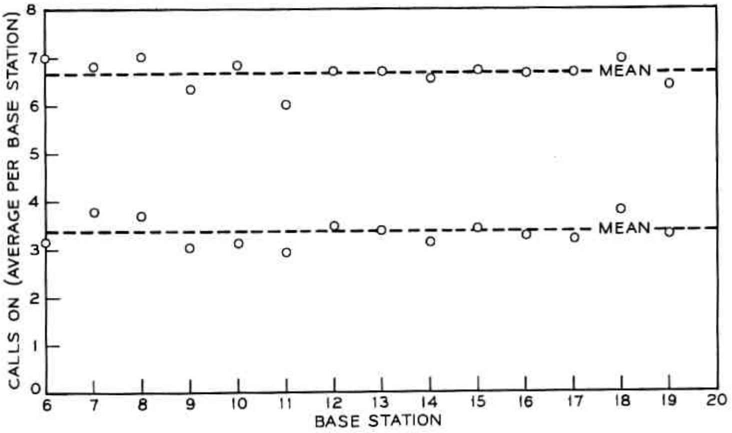


Fig. 16—Average number of calls "on" for two attempt rates.

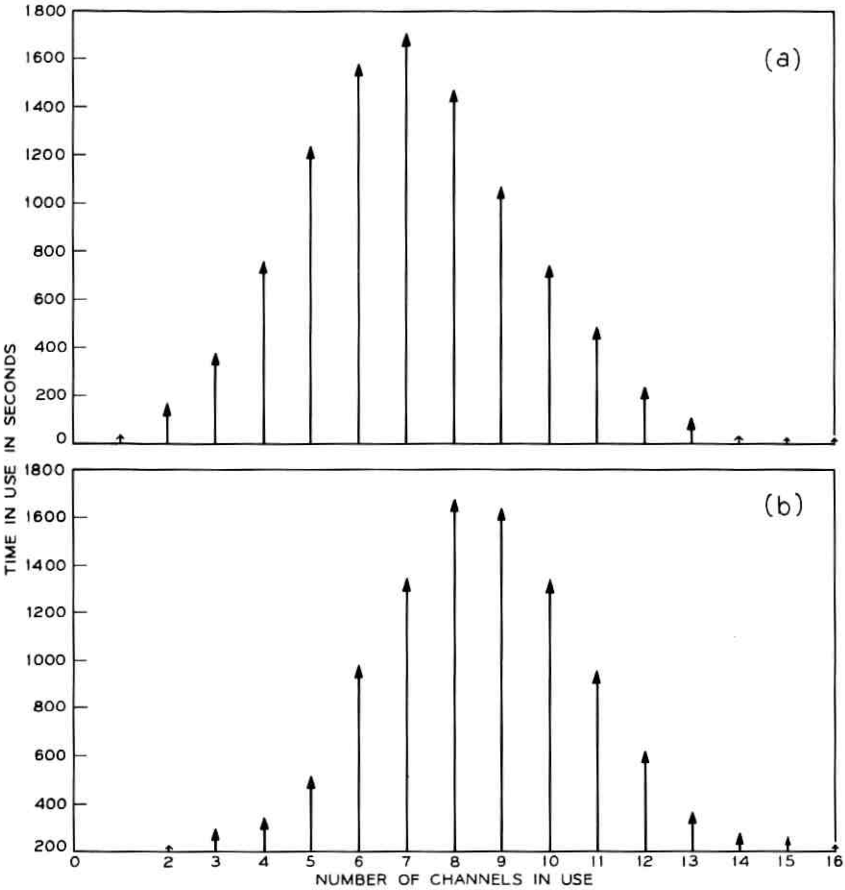


Fig. 17a—Histogram of channel usage per base station for dynamic channel assignment system with a calls-on average of 7.1.

Fig. 17b—Histogram of channel usage per base station for dynamic channel assignment system with a calls-on average of 8.66.

used at a base station seldom exceeds that value. The design of the FIXSYS allows for a maximum of 10 transmitter channels at any base station. Similar histograms of the calls-on distribution for the FIXSYS are presented in Fig. 18a and b.

7.2 *Nonuniform Call-Attempts*

The data presented to this point were produced under the assumption that the spatial distribution of call-attempts is uniform. It is reasonable to expect that variations in call-attempt rates might occur

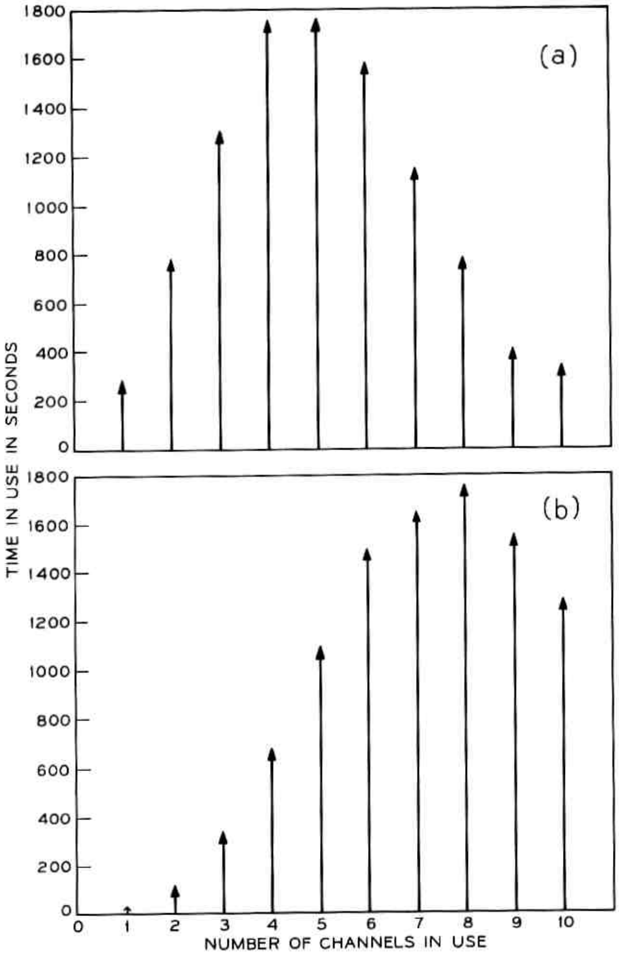
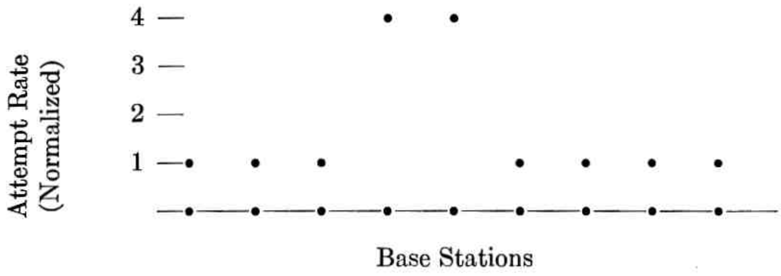


Fig. 18a—Histogram of channel usage per base station for fixed-channel assignment system with a calls-on average of 5.1.

Fig. 18b—Histogram of channel usage per base station for fixed-channel assignment with a calls-on average of 7.08.

during rush hours and there certainly will exist random occurrences of highly localized demands.

The following example is presented as a plausible situation to demonstrate the reactions of the two systems. Assume that the call-attempt density within the coverage area of two adjacent base stations is four times higher than the other base stations in the system as illustrated below:



Such a situation could occur when traffic converges at the center of a city during the morning rush hours.

One can calculate the blocking versus calls-on for the FIXSYS using the trunking formulas referenced earlier. The comparative performance of the DYNSYS was obtained by simulation. The values plotted in Fig. 19 are averages taken over one reuse interval (four base stations) centered on the two base stations with higher call-attempt rate. Since the same 1,000 system cycles were used in the simulation and the averages could not be taken over all base stations, the accuracy of the DYNSYS points is degraded compared to those presented

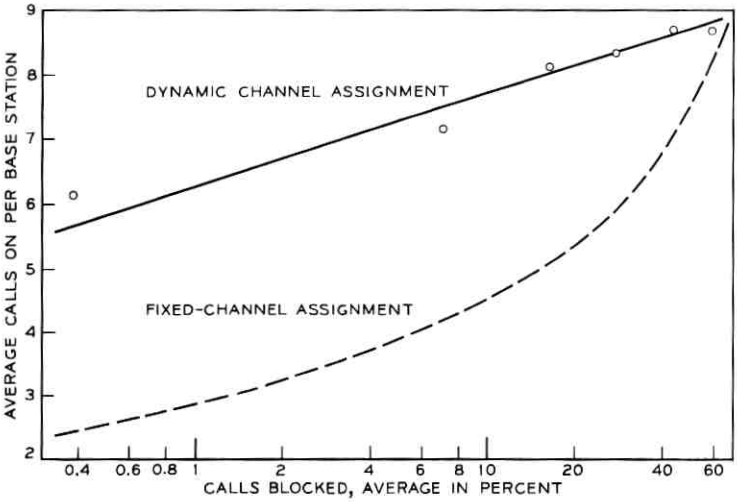


Fig. 19—Channel usage (peaked loading).

in the flat loading examples in Fig. 15. However, since there is such a marked difference in performance between the two systems more precise calculations of the points were not felt necessary. In this example, we see that the FIXSYS can handle only 40 percent of its average capacity of 10 calls "on" per base station, while maintaining 10 percent blockage, whereas the DYNSYS can work up to 70 percent capacity at the same 10-percent blockage level.

VIII. SUMMARY

The simulations described readily produce system performance parameters which are difficult, if not impossible, to obtain analytically. The system driving routine provides flexibility in parameter variation such as in demand profiles and distributions of call durations. Many channel assignment philosophies and system configurations can be readily explored by the versatile system operating routines. The graphical techniques used to display system operation for each system cycle provide a vivid picture of system dynamics.

The simulated points for uniform spatial distributions of attempts were averaged over a time sufficient to minimize statistical fluctuations. This is illustrated by the close fit of the points to the smooth curves. The fact that the simulated data for the fixed-channel assignment system agrees closely with calculations based on telephone trunking formulas lends credibility to the simulations and to the use of these formulas for determining the performance of such systems.

Portions of the system control routines such as the channel assignment procedures and the method of coding vehicle channel and base station assignments could be used essentially without modification for controlling an actual high-capacity mobile radio system.

When operating with a uniform spatial distribution of attempts and with the fixed-channel assignment system optimally matched to this distribution, the dynamic channel assignment system out-performs the fixed system for blocking rates less than 10 percent. In fact, at a blocking rate of 3 percent, the dynamic system handles 20 percent more calls per base station than the fixed system.

IX. ACKNOWLEDGMENT

The authors would like to express their appreciation to Mrs. Diane Vitello for running the computer programs and helping with program debugging.

REFERENCES

1. Schulte, H., Jr., and Cornell, W. A., "Multiarea Mobile Telephone System," IRE Trans. Veh. Commun. *9* (1960), pp. 49–53.
2. Schiff, L., "Traffic Capacity of Three Types of Common-User Mobile Radio Communication Systems," IEEE Trans. Commun. Tech., *COM-18,* No. 1 (February 1970), pp. 12–21.
3. Frenkiel, R. H., "A High-Capacity Mobile Radiotelephone System Model Using A Coordinated Small-Zone Approach," IEEE Trans. Veh. Tech., *VT-19,* No. 2 (May 1970), pp. 173–177.
4. Cox, D. C., and Reudink, D. O., "Floating Cells—A Dynamic Approach to High-Capacity Mobile Radio Systems," unpublished memorandum.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Antenna Spacing Requirement for a Mobile Radio Base-Station Diversity

By W. C. -Y. LEE

(Manuscript received February 8, 1971)

The minimum required antenna spacing between two base-station antennas in order to take advantage of spatial diversity technique was investigated. The measurements were made for two cases: (i) the incoming radio signal was perpendicular to the axis of two base-station antennas (the broadside case), and (ii) the signal was in-line with the axis of two base-station antennas (the in-line case). The correlation of signals received from two separated antennas at the base station was found to be much higher for the in-line case than for the broadside case with any given antenna spacing. For correlation up to 0.7, from which most of the advantage of two-branch diversity can still be obtained, we found the minimum required antenna spacing is around 70λ–80λ for the in-line case and 15λ–20λ for the broadside case. In order to achieve a correlation always less than 0.7 between two base-station signals regardless of the arrival direction of the incoming signal, a triangular configuration with a three-antenna array used with a three-branch diversity receiver is proposed, requiring less antenna spacing in the array than for a two-antenna setup.

I. INTRODUCTION

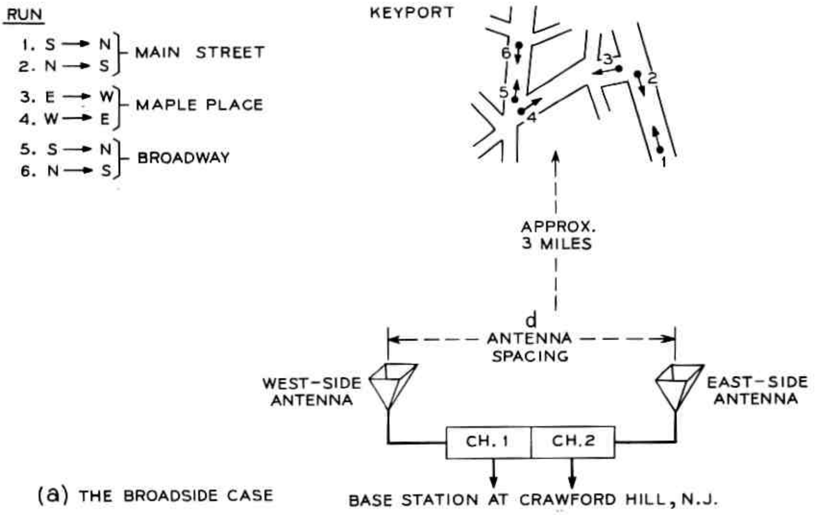
It has been shown previously that diversity reception techniques at the mobile unit often help to reduce the fading rate of a mobile radio signal.[1–3] Here we try to determine the limitations on using space diversity at the base station in order to reduce the signal fading. In particular, we would like to know how large the antenna spacing should be between two base-station antennas in order to take advantage of the diversity technique. If this turns out to be practical, then we might prefer to build a diversity system, even a complicated one, from the economical point of view at the base station, and let the transmitter and receiver in the mobile unit be as simple as possible.[4,5] In the experiments we varied the antenna spacing between two base-station antennas and calculated the cross-correlation of the envelopes of the signals received from these antennas for a number of different antenna spacings. We also determined the theoretical relationship between the cross-correlation and the cumulative distribution curve of the combined signal. If we select a particular cumulative distribution curve as acceptable for diversity operation, the corresponding correlation then indicates the required antenna spacing.

Using the experimental correlation data of both the in-line propagation case and the broadside propagation case as a guide, we will try to reduce the required antenna spacings by using an array of three base-station antennas.

II. EXPERIMENTS

A mobile radio transmitter was set up at 836 MHz in a station wagon with a $\lambda/2$ dipole mounted vertically. This transmitted to two receiving horns having 24 degrees beam width in the horizontal plane and located at the north end of Crawford Hill in Holmdel, New Jersey. During the measurements the station wagon was driven at a constant speed of 15 mph on three selected streets in the Keyport area, as shown in Fig. 1a and b. The horn antennas were used to reduce the local scattering at the base, thus simulating the conditions at a typical installation where the basic antenna is mounted well above nearby objects. Two of the streets chosen were in line with the radius vector of the base station, and the other was perpendicular to the radius vector. (The symbols •→ and ←• in Fig. 1a and b indicate different runs on the same streets. The dot indicates that the data have been received around this point on the street, and the arrow indicates the motion of the transmitter.) Figure 1a shows the experimental set up for the case where the incoming radio signal is perpendicular to the axis of two base-station antennas (the broadside case). In this case, the S → N runs were closer to the base station than the N → S runs on Main Street and on Broadway, as shown in Fig. 1a. The distance from the mobile transmitter to the base-station receivers was about three miles. The separation between the two base-station receiving antennas was variable from 25λ to 70λ in 10λ steps starting from 30λ, i.e., 25λ, 30λ, 40λ, $\cdots$.

Figure 1b shows the experimental setup for the case where the incoming radio signal is parallel to the axis of two base-station antennas (the in-line case). In this case, the S → N runs and N → S runs were made approximately in the same sections of the streets. The distance



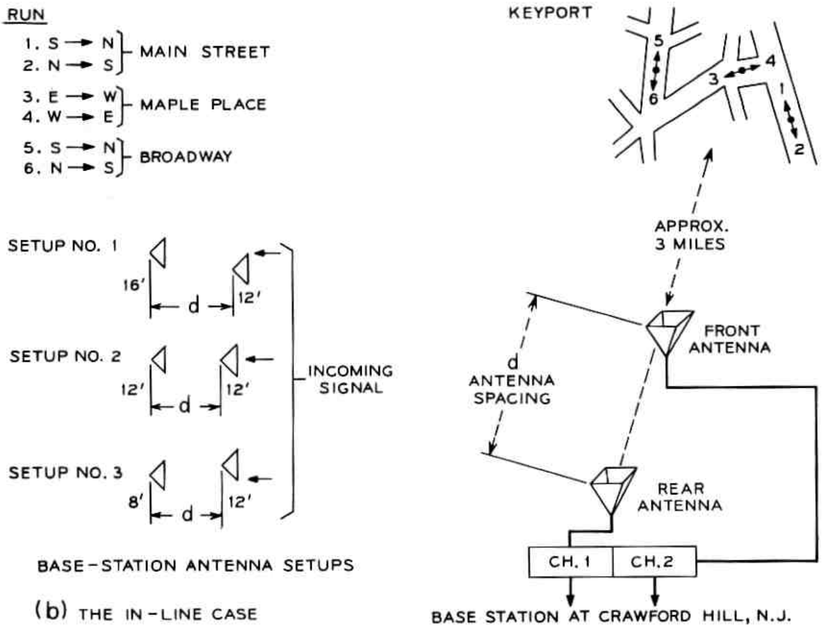


Fig. 1a—Map of testing area (Keyport, N. J.), the broadside case.

Fig. 1b—Map of testing area (Keyport, N. J.), the in-line case.

from mobile transmitter to the base-station receivers was also about three miles. The separation between the two base-station receiving antennas was variable from 30λ to 100λ in 10λ steps. The height of the front base-station antenna was fixed at 12 feet. The height of the rear antenna was variable in three steps, 16 feet, 12 feet, and 8 feet.

For both cases, each signal received from each antenna was fed to a separate receiver, which was calibrated from a common source. The two signals were recorded simultaneously on a magnetic tape recorder.

III. EXPERIMENTAL CROSS-CORRELATIONS

All data obtained from the two-channel base-station receiver were digitized at a 500-Hz sampling rate, which is fast enough to sample the data adequately. There are several questions which need to be answered before calculating the cross-correlation of two received signals. First, are the mean values of a signal in different time intervals (local means) different, and, if so, how should we compensate for this? Second, how many sample points should be taken for each of two received signals to calculate their cross-correlation? Third, are the processes of the signals we received stationary in the wide sense?[6] These can tell us that the correlations we have obtained are independent of time. Statistically analyzing our data, we have had to justify several statistical properties of individual pieces of data before calculating the cross-correlations among them:

(*i*) *Treatment of Local Means*–We have found that the mean values of a signal in different time intervals (local means) are different. The variations in the local means of a received signal are due to the different contours of land between the transmitter and the receiver at different time intervals. These variations affect the auto-correlation function of a received signal. Hence, we normalized the local means of the received signal, before calculating its auto-correlation, as follows: We broke a piece of data into many time intervals, then each sample was divided by the mean of the signal in that time interval. In our data, we used a time interval a half-second long, i.e., 250 samples.

(*ii*) *The Number of Sample Points to be Taken*–We have found from the auto-correlation curves of most data that their first nulls occur at a delay, τ, of 20 samples. Then we may say that there is no correlation between two pieces of data separated by a delay of 20 samples. Since these pieces of data are Rayleigh distributed,[4] if they are uncorrelated, they are also independent. Now we look at this a different way: how long will a piece of data be which can from an ensemble of , say, 100 independent subsets of data? Since every subset of data with a delay of at least 20 samples from each other is an independent subset, a piece of data containing over 2000 samples is required to provide an ensemble of 100 independent subsets. We know that 100 independent subsets of data are enough to calculate the ensemble average. Therefore, from the concept of ergodicity, we need a piece of data containing over 2000 samples in order to meet the requirement "time average equals ensemble average". However, most of the local means of all runs vary rapidly after 2500 samples. Therefore, we chose 2500 sample points as a unit piece of data.

(*iii*) *Verification of Stationarity*–If a piece of data is stationary, the auto-correlation of a piece of data $x(t)$ should be independent of time t. This means that if we pick any arbitrary starting point, say one second after the first sample point, and calculate the auto-correlation, the result should be the same as if we had chosen the first sample point as the starting point. This evidence was observed in our data. Hence we have shown that our data are stationary.

In summary, we have found that the local means of two signals should be properly factored out, and that 2500 sample points are sufficient for calculating the cross-correlation. Under these two requirements, the processes of the signals we received were verified to be stationary in the wide sense.

Assuming two signals received from two base-station antennas, separated by distance d, are $x(t)$ and $y(t)$, the cross-correlation coefficient of $x(t)$ and $y(t)$ delayed by time τ can be expressed as

$$\rho(t, \tau, d) = \frac{\overline{x(t)y(t+\tau)} - m_x m_y}{\sqrt{\overline{x^2(t)} - m_x^2}\sqrt{\overline{y^2(t+\tau)} - m_y^2}} \tag{1}$$

where m_x and m_y are mean values of $x(t)$ and $y(t+\tau)$, respectively. Since we have shown that our data are stationary, and also we let $\tau = 0$, equation (1) becomes

$$\rho(d) = \frac{\overline{x(0)y(0)} - m_x m_y}{\sqrt{\overline{x^2(0)} - m_x^2}\sqrt{\overline{y^2(0)} - m_y^2}}. \tag{2}$$

After normalizing the local means of the two signals from the two channels (receivers), we used 2500 sample points and calculated the cross-correlation as a function of antenna spacing for two cases.

3.1 *The Broadside Case*

The cross-correlation as a function of base-station antenna spacing was determined for runs on three streets (Main Street on Fig. 2, Maple

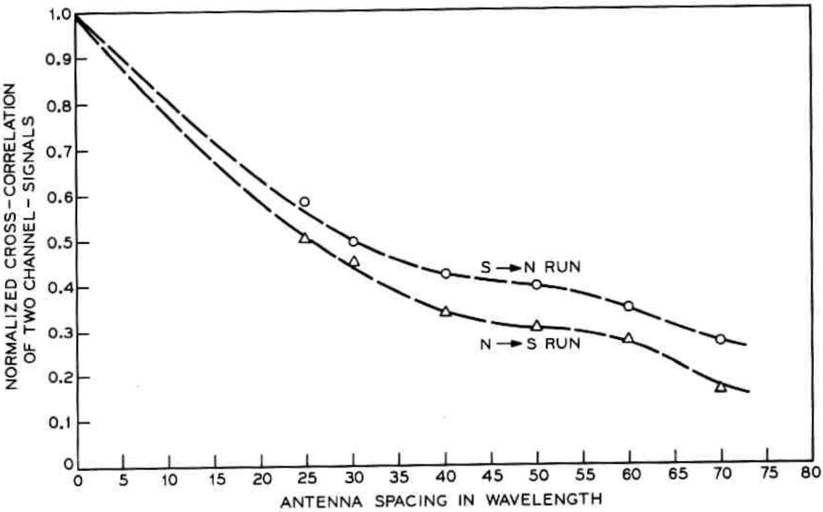


Fig. 2—The correlation coefficient of two base-station antennas broadside with the incoming wave vs antenna spacing for runs on Main St., Keyport, N. J.

Place on Fig. 3, and Broadway on Fig. 4). From these three figures, there are several major points to be disclosed:

(*i*) The cross-correlation decreases as the base-station antenna spacing increases, as one would expect.

(*ii*) In general, the cross-correlation is somewhat higher when the mobile transmitter is nearer the base station, such as the S → N runs shown in Figs. 2 and 4.

(*iii*) The tops of the trees at the northwest boundary of Crawford Hill were high and thus partially scattered the incoming signal from the northwest direction. Hence, the signals received from Broadway and the west side of Maple Place were lower. The cross-correlations obtained from these locations were also low, due to this local scattering.

(*iv*) To measure the cross-correlation for an antenna spacing of 70λ, we moved the west-side receiving antenna farther left by 10λ. Therefore, the tops of trees more affected the incoming signal received by Channel 1 from Broadway. Hence, the two points for this antenna spacing shown on Fig. 4 have very low values.

(*v*) The highest cross-correlation for the six runs for an antenna spacing of 25λ was 0.65, and for an antenna spacing of 70λ was about 0.27.

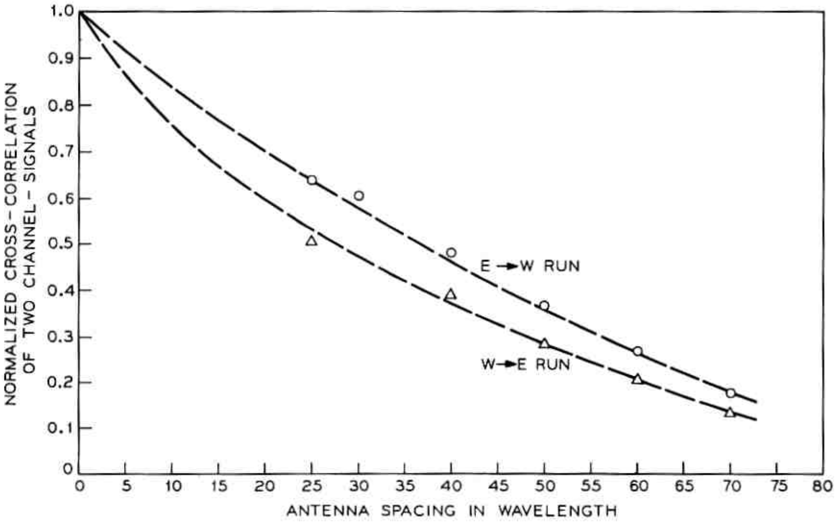


Fig. 3—The correlation coefficient of two base-station antennas broadside with the incoming wave vs antenna spacing for runs on Maple Place, Keyport, N. J.

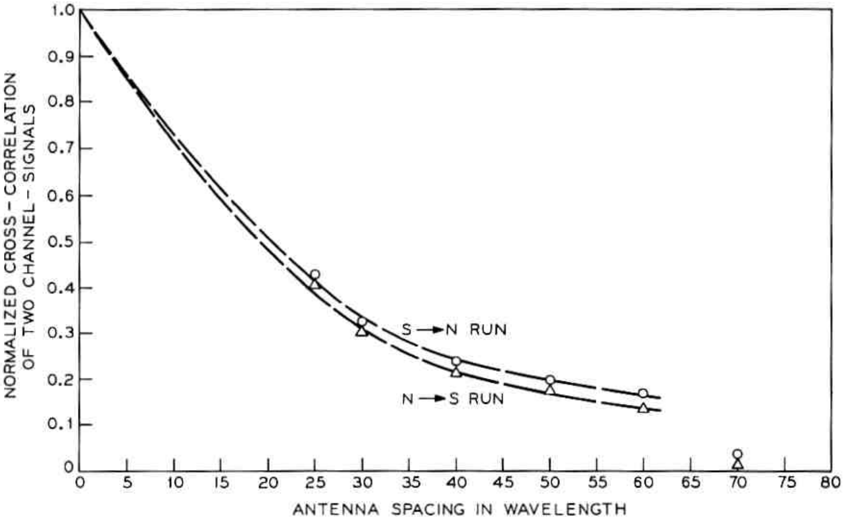


Fig. 4—The correlation coefficient of two base-station antennas broadside with the incoming wave vs antenna spacing for runs on Broadway, Keyport, N. J.

3.2 *The In-Line Case*

The cross-correlation as a function of base-station antenna spacing was determined for runs on three streets (Main Street, Maple Place, and Broadway) for three values of base-station antenna height (see Fig. 1b), and the results shown in Figs. 5 through 13. Each figure shows the correlation for a particular antenna height for runs on a particular street. Assuming that the correlation at zero antenna spacing is one, we can draw a curve for the best fit to our measured correlation points for spacing from 0 to 100λ for each run. From these nine figures, there are several major points to be disclosed:

(*i*) Between the broadside case and the in-line case: The correlation coefficient of the in-line propagation case is much higher than that of the broadside propagation case for a given antenna spacing. For antenna spacing = 70λ, the correlation coefficient is about 0.2 for the broadside propagation case and about 0.7 for the in-line propagation case. Alternatively, to get a correlation of 0.7, the antenna spacing merely needs to be 25λ in the broadside case but 70λ in the in-line case.

(*ii*) Direction of runs: From Figs. 5–7 (Main Street) and Figs. 11–13 (Broadway), we find that the correlations are higher for those runs (S → N) when the mobile radio transmitter was traveling away from the base station than those (N → S) when traveling toward the base station. Figures 8–10 (Maple Place), for runs perpendicular to

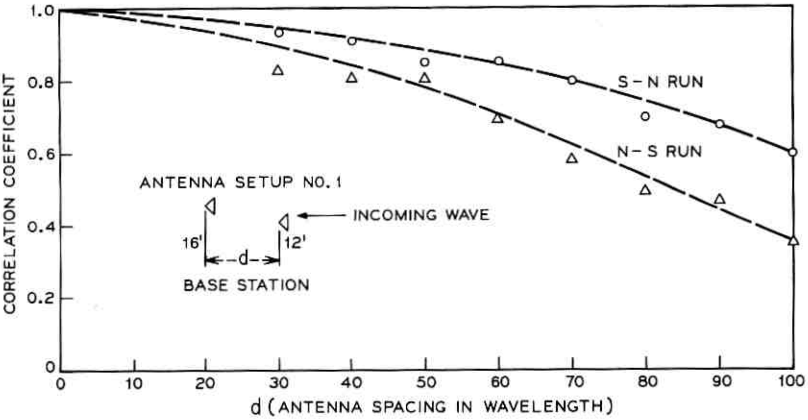


Fig. 5—The correlation coefficient of two base-station antennas (setup no. 1) in line with the incoming wave vs antenna spacing for runs on Main St., Keyport, N. J.

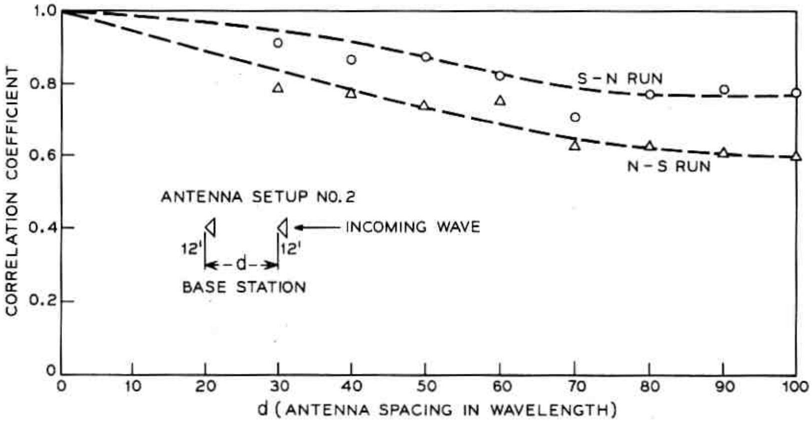


Fig. 6—The correlation coefficient of two base-station antennas (setup no. 2) in line with the incoming wave vs antenna spacing for runs on Main St., Keyport, N. J.

the radius vector of propagation, show that the signal correlations from runs in different directions, E → W or W → E, are not noticeably different.

(*iii*) Antenna height: The variation of height of the base-station antennas apparently has minor effects on the correlations of the received signals. (Figs. 5–13).

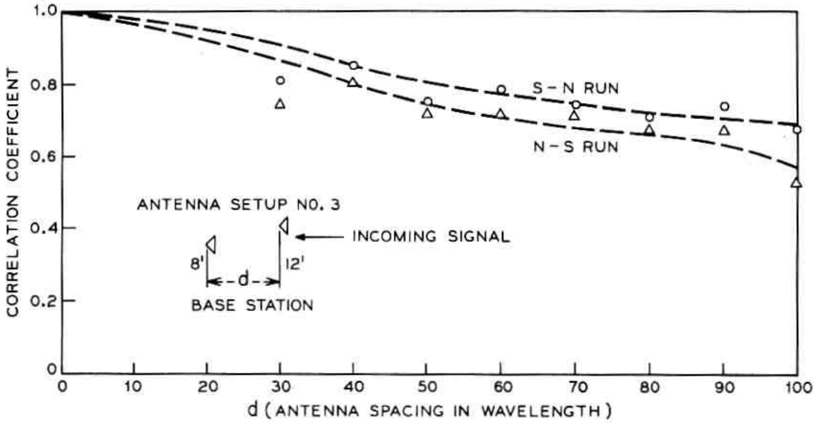


Fig. 7—The correlation coefficient of two base-station antennas (setup no. 3) in line with the incoming wave vs antenna spacing for runs on Main St., Keyport, N. J.

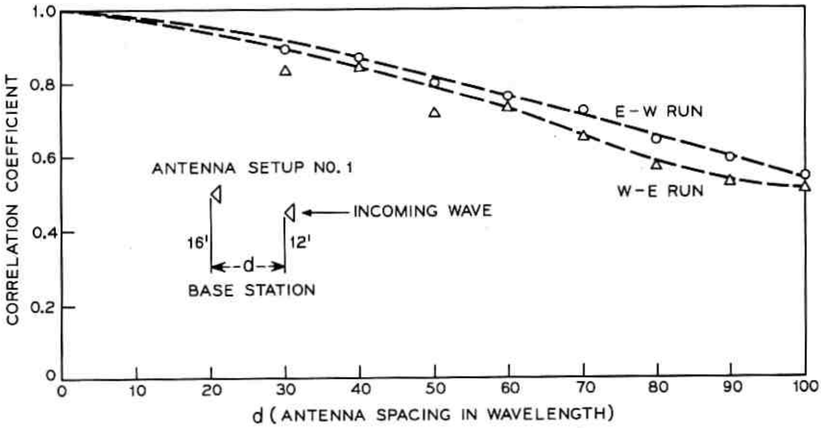


Fig. 8—The correlation coefficient of two base-station antennas (setup no. 1) in line with the incoming wave vs antenna spacing for runs on Maple Place, Keyport, N. J.

(*iv*) Different streets: On Main Street the buildings are taller and spaced closer together than on the other two streets. The buildings on Maple Place are again taller and spaced closer than those on Broadway. This may be the reason that the average correlation drops slightly faster on Main Street and more slowly on Broadway as the antenna spacing increases.

(*v*) Local scattering at the base station: Most of the correlation

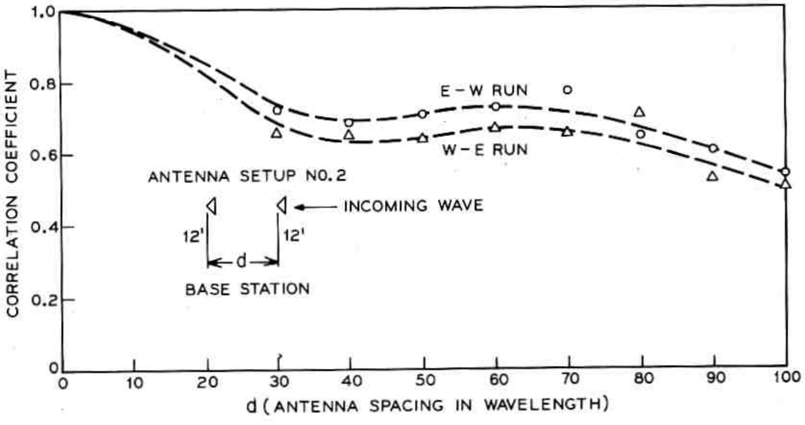


Fig. 9—The correlation coefficient of two base-station antennas (setup no. 2) in line with the incoming wave vs antenna spacing for runs on Maple Place, Keyport, N. J.

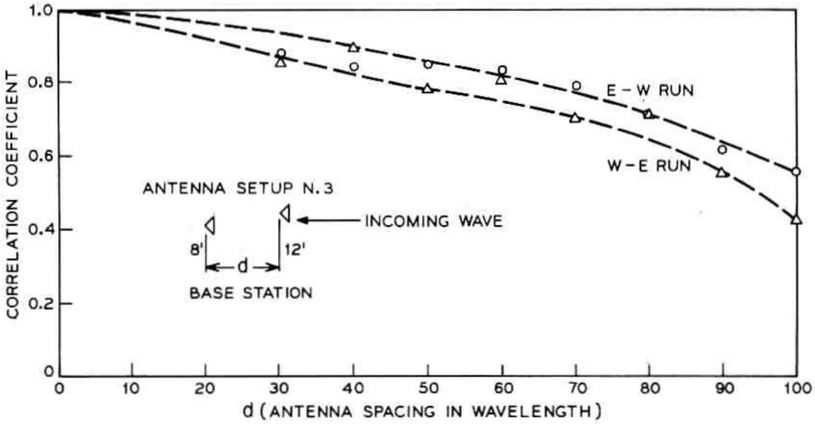


Fig. 10—The correlation coefficient of two base-station antennas (setup no. 3) in line with the incoming wave vs antenna spacing for runs on Maple Place, Keyport, N. J.

curves in Figs. 5–13, are monotonically decreasing as the antenna spacing increases. A few curves have dips; for example, see Fig. 12. These are perhaps caused by the local scattering at the base station.

(*vi*) Upper bound of the correlation data: Figure 14 shows the upper bound of the correlation data for the three antenna setups versus the antenna spacing d. When $d = 70\lambda$, the highest correlation is 0.85.

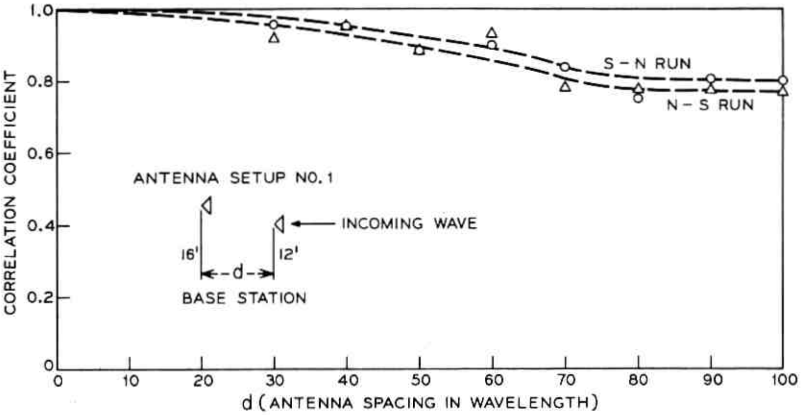


Fig. 11—The correlation coefficient of two base-station antennas (setup no. 1) in line with the incoming wave vs antenna spacing for runs on Broadway, Keyport, N. J.

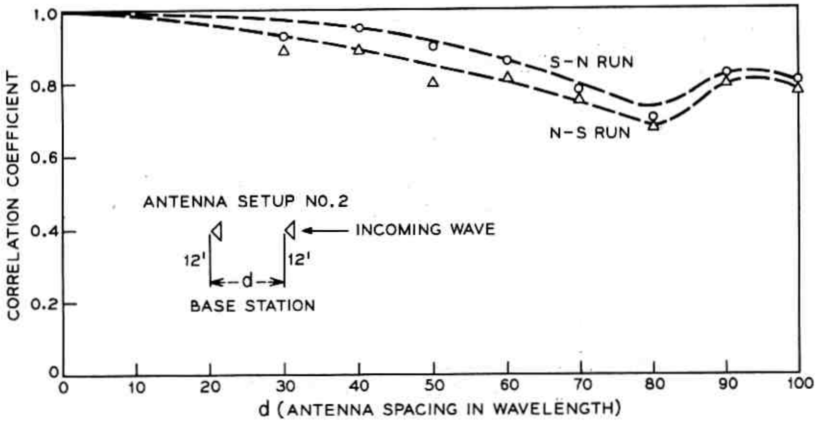


Fig. 12—The correlation coefficient of two base-station antennas (setup no. 2) in line with the incoming wave vs antenna spacing for runs on Broadway, Keyport, N. J.

IV. EFFECT OF CORRELATION ON DIVERSITY

Having obtained the cross-correlation as a function of antenna spacing, we would now like to ask how great a cross-correlation coefficient between two received signals we can tolerate and still realize a signal improvement in a two-branch predetection diversity receiver.

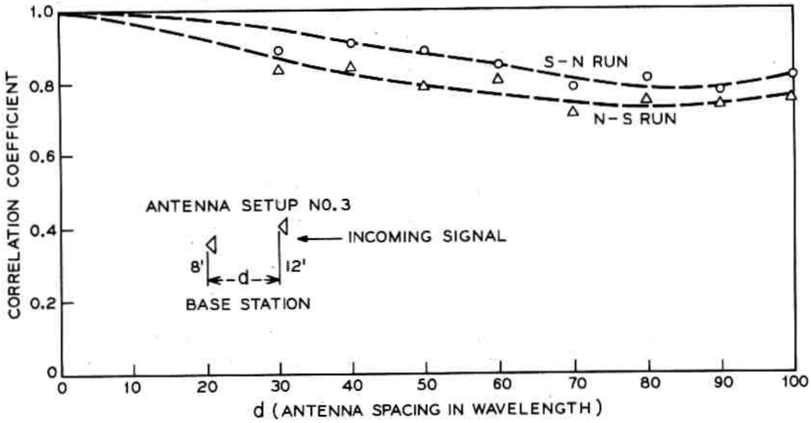


Fig. 13—The correlation coefficient of two base-station antennas (setup no. 3) in line with the incoming wave vs antenna spacing for runs on Broadway, Keyport, N. J.

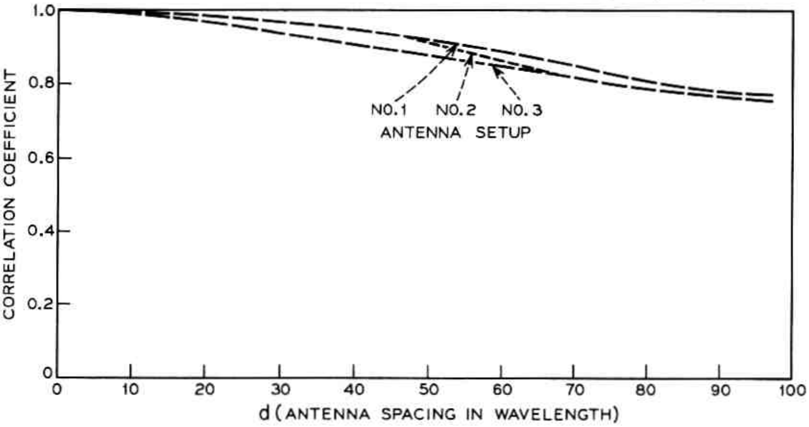


Fig. 14—Upper bound of correlation coefficients from different antenna setups at the base station (in-line case) with a mobile radio transmitting from three different streets in Keyport, N. J.

To do this we will find the cumulative distribution of a two-branch diversity combined signal as a function of the cross-correlation of the two incoming signals.

Assume that the two signals received by a two-branch diversity receiver are complex gaussian random variables[4] in a short-term fading (a piece of data a few hundred wavelengths long). If these two signals are independent, i.e., no cross-correlation, the cumulative distribution of a two-branch maximum ratio diversity* signal can be expressed as[7]

$$P(\gamma < x) = 1 - \left(1 + \frac{x}{\Gamma}\right) \exp\left(-\frac{x}{\Gamma}\right) \tag{3}$$

where γ is the sum of the instantaneous SNR's on the individual branches, Γ is the average SNR of a single channel, and x is a value greater than γ. For the case that the two received signals are not independent, we can obtain the cumulative distribution from Ref. 9 by inserting equation (10-10-21) into equation (10-10-22) for a two-branch maximum-ratio diversity. After some manipulation, we get

* The cumulative distributions of a two-branch maximum-ratio diversity combining signal and a two-branch equal-gain diversity combining signal are very similar, differing only by 0.49 dB[8] if the two branches are independent.

$$P(\gamma < x) = 1 - \left\{ \frac{(1+\sqrt{\rho}) \exp\left[-\frac{x}{(1+\sqrt{\rho})\Gamma} \right] - (1-\sqrt{\rho}) \exp\left[-\frac{x}{(1-\sqrt{\rho})\Gamma} \right]}{2\sqrt{\rho}} \right\} \tag{4}$$

where γ, x, and Γ are as previously described and ρ is the cross-correlation coefficient between the envelopes of two-branch signals. Equations (3) and (4) are plotted in Fig. 15. The abscissa is the SNR with respect to the mean value of the SNR of a single channel, in dB. We see that the curve for $\rho = 0.7$ is very close to that for $\rho = 0$. This means that for correlations between the two branches up to 0.7, the advantage of using diversity technique is still good. Figure 15 also shows that 99.9 percent of the time the SNR for $\rho = 0$ is above -13.3 dB, while at the same percentage, the SNR for $\rho = 0.7$ is above -15.8 dB. The difference is 2.5 dB. Comparing the two-branch signal for $\rho = 0$ with a single-channel signal at the 99.9 percent level, the difference is 17 dB. For a probability of 99.9 percent, the difference between two two-branch signals for $\rho = 0$ and $\rho = 0.7$ is still 2.5 dB; however, the difference between a two-branch signal with $\rho = 0$ and a single-channel signal is 22 dB. Hence, we could say that, if a two-branch signal for $\rho = 0.7$ is taken, the improvement over a single channel at the 99.99 percent level is greater than at the 99.9 percent level. Figure 16 gives a clear view of the two-branch diversity improvement over a single-channel signal for different correlations. The improvement becomes less as the correlation increases. Also the improvement becomes greater at the higher percentage signal level than at the lower percentage signal level, as mentioned previously. For ρ between 0 and 0.7, the diversity advantage changes very little. Hence, we may say that $\rho = 0.7$ is a reasonable value to pick for the maximum cross-correlation which can be tolerated in order to take good advantage of the diversity technique. The antenna spacing for a cross-correlation of 0.7 is around 15λ–25λ for the broadside case (see Figs. 2–4) and is 70λ–80λ for the in-line case (see Figs. 5–13).

V. CONSIDERATION OF A THREE-BRANCH DIVERSITY SYSTEM

As we have seen, to get a correlation of 0.7 between two base-station signals in the in-line case requires antenna spacings of 70λ to 80λ, and in the broadside case requires antenna spacings of 15λ to

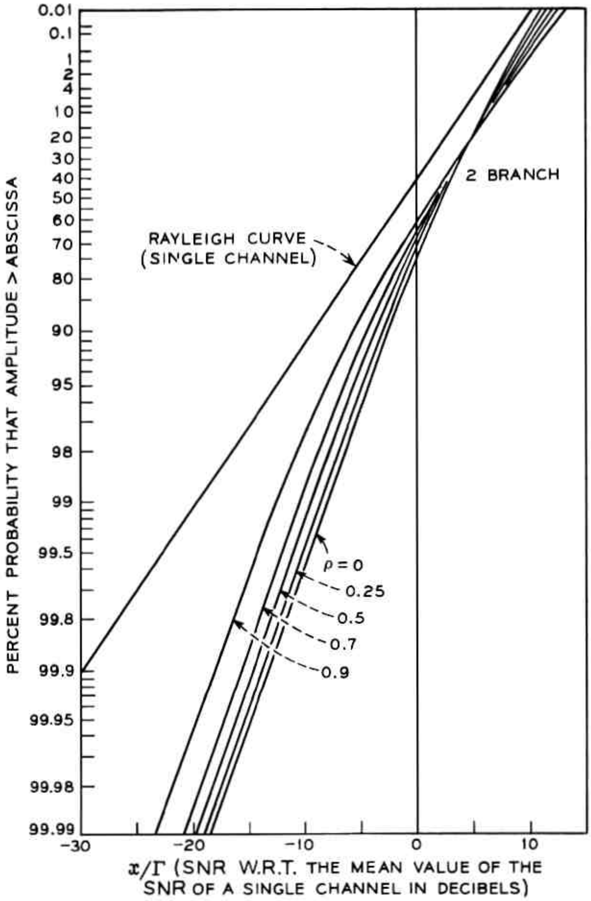


Fig. 15—The cumulative distribution curves vs correlation for a two-branch maximum-ratio diversity receiver.

25λ. Hence the broadside case is better than the in-line case as far as saving the antenna space is concerned. Since the base-station antennas are set up not only for one mobile unit in one particular direction but rather for all the mobile units under its radio coverage, some mobile units may well be in the in-line case to the base station if only two base-station antennas are considered. If we try to reduce the antenna spacing and still meet the same correlation requirement of 0.7 or less regardless of the direction of the incoming signal arrival, a triangular antenna array may provide a solution. In a triangular

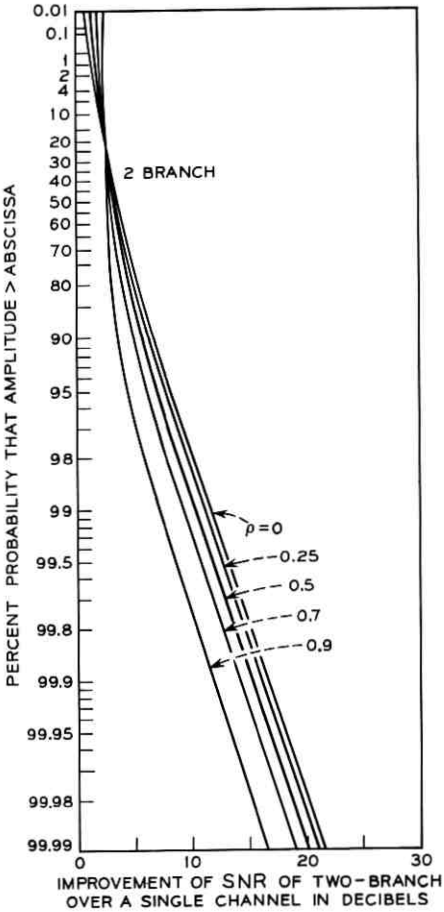


Fig. 16—Improvement of signal-to-noise ratio of a two-branch signal over a single-channel signal.

array, we have three cross-correlations between three antennas. When an incoming mobile radio signal is in line with two antennas of the array, then the correlation with these two antennas is very high but the other two correlations are lower, since they are more nearly broadside to the incoming signals. Because two of the three antennas are always approximately broadside to an incoming signal in a triangular array, we may always get at least the advantage of two-branch diversity but never more than three-branch.[10] Furthermore, by making more correlation measurements for arrival angles between in-line and broadside cases in the future, the exact performance of the triangular array could be calculated.

VI. CONCLUSION

The measurements reported here show that the correlation of the in-line propagation case is much higher than that of the broadside propagation case for any given antenna spacing. To get a correlation of 0.7, the antenna spacing merely needs to be 25λ in the broadside case but 70λ–80λ in the in-line case. Direction of runs and variation of height of the base-station antennas have minor or insignificant effects on the correlation. Local scattering at the base station may not have been entirely eliminated in our test; this, if present, would reduce the apparent correlations. In this case, larger separations might be required in actual situations, depending on the amount of local scattering that existed. Further work is needed to resolve this point.

The theoretical analysis has pointed out that the advantage of a two-branch diversity can be obtained when the cross-correlation between branches is less than 0.7.

In order to achieve a condition on the cumulative distribution curve equivalent to a correlation always less than 0.7 between two base-station antennas regardless of the direction of the incoming signal arrival, yet with antenna spacing less than the in-line case, a triangular configuration with a three-antenna array used with a three-branch diversity receiver is proposed. We estimate that, in a triangular array, the antenna spacing between any two of three antennas could be around 40λ for its cumulative curve to be better than that of a two-branch receiver with correlation of 0.7 regardless of the direction of the incoming signal onto the triangular array. The idea of applying the diversity scheme at the base station could then be realized.

VII. ACKNOWLEDGMENTS

Discussion with W. C. Jakes, M. J. Gans, and D. C. Cox have been very helpful. R. H. Brandt helped with experimental data.

REFERENCES

1. Rustako, A. J., "Evaluation of a Mobile Radio Multiple Channel Diversity Receiver Using Pre-Detection Combining," IEEE Trans. on Vehicular Tech., *16,* No. 1 (October 1967), pp. 46–57.

2. Lee, W. C.-Y., "Comparison of an Energy Density Antenna System with Pre-Detection Combining System for Mobile Radio," IEEE Trans. on Commun. Tech., *17,* No. 2 (April 1969), pp. 277–284.
3. Reudink, D. O., and Rustako, A. J., "Performance of a Four-Branch Pre-Detection FM Mobile Microwave Receiver in an Urban Area," International Conference on Communications, Boulder, Colorado, June 9–11, 1969.
4. Jakes, W. C., Jr., "A Review of Space Diversity Techniques for Reduction of Fast Fading in UHF Mobile Radio Systems," unpublished work.
5. Yeh, S. Y., "An Analysis of Adaptive Retransmission Arrays in a Fading Environment," B.S.T.J., *49,* No. 8 (October 1970), pp. 1811–1825.
6. Papoulis, A., *Probqbility, Random Variables and Stochastic Processes,* New York: McGraw-Hill, 1958, p. 302.
7. Brennan, D. G., "Linear Diversity Combining Techniques," IEEE Proc., *47* (June 1959), pp. 1075–1102.
8. Ibid., p. 1086, Table I.
9. Schwartz, M., Bennett, W. R., and Stein, S., *Communication Systems and Techniques,* New York: McGraw-Hill, 1966, p. 443.
10. Lee, W. C.-Y., "A Study of the Antenna Array Configuration of an M-Branch Diversity Combining Mobile Radio Receiver," IEEE Trans. on Vehicular Technology (to be published).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

A Simple Interframe Coder For Video Telephony

By J. O. LIMB and R. F. W. PEASE

(Manuscript received February 9, 1971)

The technique of exchanging resolution according to the amount of movement in a picture has been previously described; in stationary parts of the picture the temporal resolution is reduced while in moving parts of the picture the spatial resolution is reduced. Here, we describe a method of applying resolution exchange to a differentially quantized (DPCM) signal. The resulting channel capacity required for the subjectively satisfactory transmission of the differential signal is halved. The coder is simpler than most interframe coders and should not increase the sensitivity of the system to channel errors.

I. INTRODUCTION

In a previous paper we described a way to halve the channel capacity required for the subjectively satisfactory transmission of an 8-bit PCM television signal by exchanging spatial and temporal resolution according to the amount of movement in the local part of the picture[1] Every second picture element ("pel") is sampled and in stationary areas of the picture the values of the unsampled pels are interpolated from adjacent temporal samples (reduced temporal resolution); in the moving areas the values of the unsampled elements are interpolated from neighboring sampled elements in the same line (reduced spatial resolution).

We would like to apply this technique to a signal whose bit rate has already been reduced by an element-to-element differential quantizer (EDQ), e.g., the *Picturephone*® codec. Unfortunately, halving the horizontal sampling rate, as in Ref. 1, increases the amplitude of the sample-to-sample difference signal which, in turn, requires a larger number of quantizing levels for adequate representation. There are two ways around this problem:

(*i*) Use the vertically adjacent elements as a prediction of the current pel and quantize the resulting difference. Because vertically adjacent pels are in the previous field, such a technique can be called field difference quantization and is the subject of another study.

(*ii*) Reduce the vertical resolution rather than the horizontal resolution so that the full horizontal sampling frequency is retained; every transmitted line is left completely intact. This is the method described here.

II. PRINCIPLE AND IMPLEMENTATION

In Fig. 1 we show an outline of a system which uses resolution exchange in conjunction with differential quantization of the signal. The output of the television camera is differentially quantized and fed to a movement detector which usually consists of a frame delay circuit, a difference and threshold circuit, and associated logic. When movement is not detected, that part of the picture being coded enjoys the full spatial sampling frequency but each line is only transmitted every second frame, i.e., the temporal sampling frequency is halved. When movement is detected, then the vertical sampling frequency is halved by transmitting only every second line of the picture. For a

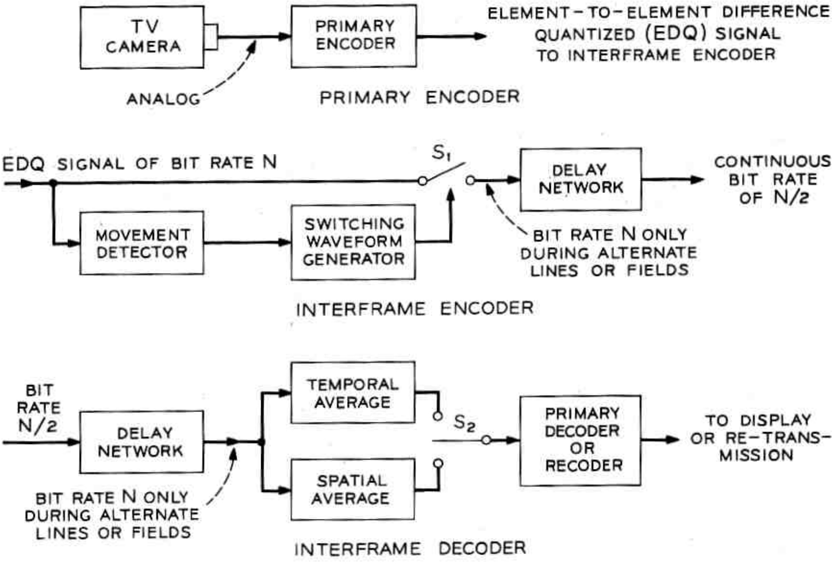

TV CAMERA
PRIMARY ENCODER
ELEMENT-TO-ELEMENT DIFFERENCE QUANTIZED (EDQ) SIGNAL TO INTERFRAME ENCODER
ANALOG
PRIMARY ENCODER
EDQ SIGNAL OF BIT RATE N
S_1
DELAY NETWORK
CONTINUOUS BIT RATE OF N/2
MOVEMENT DETECTOR
SWITCHING WAVEFORM GENERATOR
BIT RATE N ONLY DURING ALTERNATE LINES OR FIELDS
INTERFRAME ENCODER
BIT RATE N/2
DELAY NETWORK
TEMPORAL AVERAGE
S_2
PRIMARY DECODER OR RECODER
TO DISPLAY OR RE-TRANSMISSION
SPATIAL AVERAGE
BIT RATE N ONLY DURING ALTERNATE LINES OR FIELDS
INTERFRAME DECODER

Fig. 1—Schematic of communication system using simple interframe coder.

2:1 interlaced scan this means transmitting every other field and so the temporal sampling frequency is raised to 30 Hz.

At the receiver, when movement is not detected, alternate (sampled) lines in each field are decoded and displayed. In place of the unsampled lines in the field a temporal average of corresponding lines from neighboring sampled frames is formed and displayed.*

To describe this "stationary mode" we assign to each line coordinates (y, t) which refer respectively to the vertical position in each frame (or line number) and the temporal position (or field number). Figure 2 shows diagrams of lines for a 2:1 interlaced scan format. Each sampled line is marked with a full dot and the unsampled lines are shown as circles; the averaging is denoted with arrows. In the stationary mode we receive and display alternate lines in each field, i.e., lines 2, 3, 6, 7, 10, 11, $\cdots$ in fields 2 and 3 and lines 0, 1, 4, 5, 8, 9, $\cdots$ in fields 4 and 5. The averaging is done only along the time (t) axis and so stationary pictures are displayed with the full resolution of the fully sampled picture.

The results of preliminary experiments with this mode of operation continually applied to the whole picture show better motion rendition than does frame repeating but are still unsatisfactory for moderate and fast movement of subjects of normal contrast (the degradation becomes annoying at speeds of about 2 pels per frame interval). In the case of still pictures the quality is actually improved over that of normal (3- or 4-bit) EDQ because the temporal (frame-to-frame) averaging reduces the visibility of granular noise.

When movement is detected, pels from alternate sampled fields are received, decoded, and displayed. In place of pels from the unsampled fields, an average is formed from the four nearest neighbors in the $y - t$ diagram, as shown in Fig. 2b. For example, pels in lines $(y, 1)$ are replaced with the average value of pels in lines $(y - 1, 0)$ $(y + 1, 0)$ $(y - 1, 2)$ and $(y + 1, 2)$. Thus both spatial and temporal interpolation is used to form the new value in the unsampled field.

Preliminary experiments with this mode applied continually to the whole picture showed adequate motion rendition for most head and shoulders views of a person talking; the loss in vertical resolution is barely noticeable, but in some regions (especially dark regions) of high vertical detail, some aliasing patterns can be seen. Under normal viewing conditions (see Section III), fast movement (4 pels per frame interval and above) of a contrasty edge appears slightly jerky.

* Some other work on reducing the temporal resolution is given in Refs. 2 and 3.

There are a number of ways in which the sampling mode can be controlled. The subjectively ideal, but probably most difficult, method is to change to the appropriate mode as each element is encountered. Another method is to code the whole of one line in the same mode, while a further method is to code the whole field in the same mode. The method adopted in most of our experiments is to use one mode throughout each field. The transitions from stationary mode to moving mode can be made at the beginning of any field but the reverse transition is only permitted at the beginning of any even field, counting

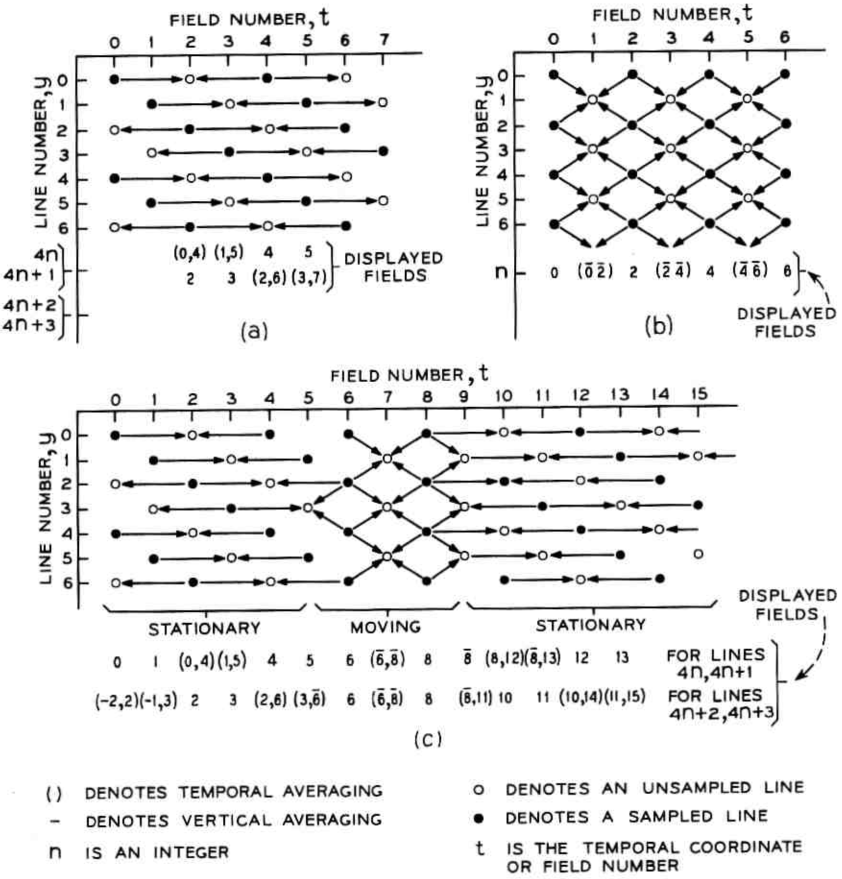


Fig. 2—Sampling patterns of lines and sequences of fields in the displayed picture for the different modes: (a) Stationary mode. (b) Moving mode. (c) A combination of modes (except near the mode changes, the fields represented in the output correspond with the temporal coordinate t).

from after the change from stationary to moving mode. This is done to prevent data being generated at a greater rate during two consecutive fields than can be handled by the channel. The mode changing information is negligible as only one extra bit of information is required after every field.

The delay network at the transmitter (Fig. 1) converts the effectively halved bit rate to a continuous bit stream of half the original rate; a similar network at the receiver reconverts the continuous bit stream back to alternate fully sampled fields or frames. The operation of these networks is described more fully in Section IV.

Fig. 2c illustrates the line sampling pattern and the interpolation used when the system starts in the stationary mode (fields 0, 1, 2, 3, 4, 5) and switches to the moving mode (fields 6, 7, 8, 9) and then switches back again. Also shown are the sequences of fields represented in the displayed picture. In both modes the lines displayed correspond either directly to the temporal coordinate or indirectly by temporally averaging from fields temporally equidistant from the current field. Near the changeovers some lines are misplaced temporally; for instance the unsampled lines of field 5 are replaced with lines from field 3 and 6 whose mean temporal position is $4\frac{1}{2}$. Different weights could be assigned to the signals of fields 3 and 6, but for simplicity we decided to subjectively test the coder using equal weight for all averaging.

To experimentally test the coder we simulated Fig. 1 with the arrangement of Fig. 3. The picture format, scenes, and viewing conditions were the same as those used in the previous resolution exchange experiments,[1] i.e., there were 271 lines per frame with a 2:1 line interlace and the frame frequency was 30 Hz; the sampling frequency was 2 MHz. The scenes were head and shoulders views of a variety of subjects engaged in conversation varying from quiet to violent. The display raster was approximately $5\frac{1}{2}$ inches horizontally by 5 inches vertically and was viewed at 3 feet under ambient illumination typical of a well-lit office (about 70 footcandles).

The primary coder is a 4-bit digital differential quantizer* whose quantizing levels are shown in Table I. The accumulated value, with a peak value of 127 levels, is fed to the frame memory via switch S_1 (Fig. 3a). In the stationary mode S_1 and S_3 are switched at the line rate and S_2 is held at 0. The waveform L, used for switching at line

* Although the quantizer has 17 levels and, strictly speaking, could not be transmitted as a 4-bit signal, certain combinations of levels were deleted so as to permit 4-bit transmission.[4]

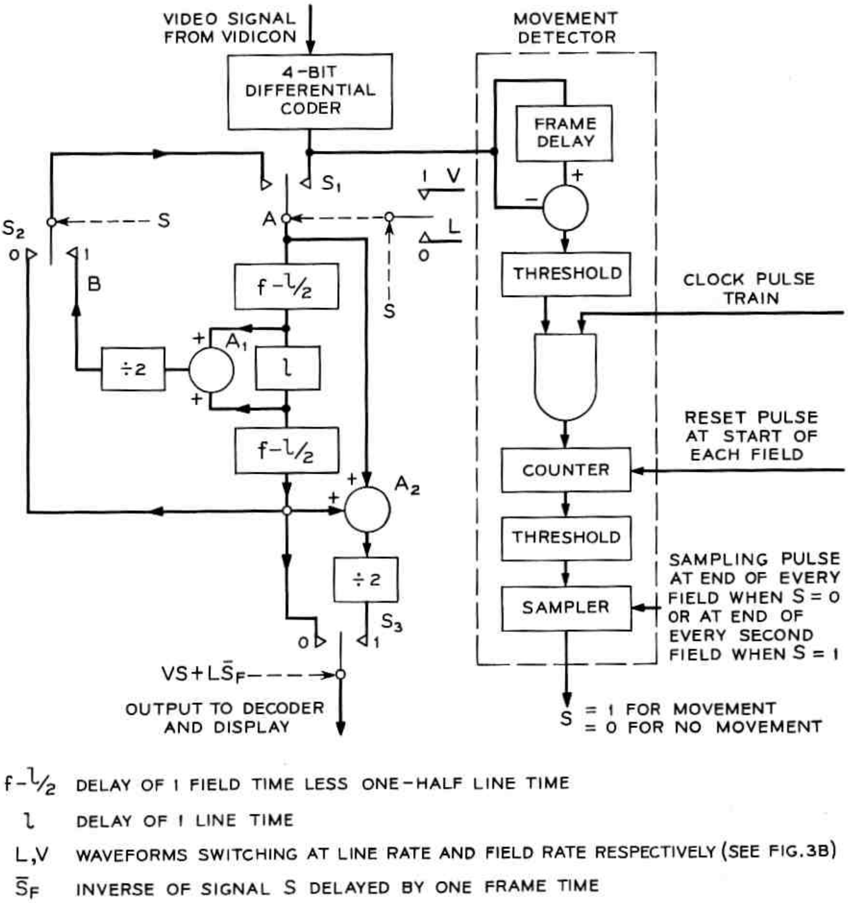
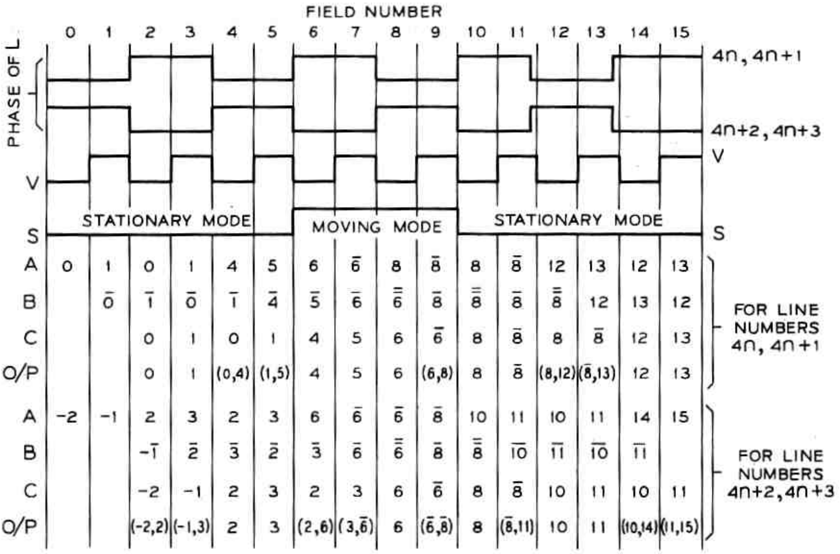

VIDEO SIGNAL FROM VIDICON
4-BIT DIFFERENTIAL CODER
MOVEMENT DETECTOR
FRAME DELAY
+
−
THRESHOLD
CLOCK PULSE TRAIN
RESET PULSE AT START OF EACH FIELD
COUNTER
THRESHOLD
SAMPLING PULSE AT END OF EVERY FIELD WHEN S = 0 OR AT END OF EVERY SECOND FIELD WHEN S = 1
SAMPLER
S = 1 FOR MOVEMENT
= 0 FOR NO MOVEMENT
S₁
I V
0 L
S₂
0
1
S
A
B
S
f−l/2
A₁
+
+
÷2
l
f−l/2
A₂
+
+
÷2
S₃
0
1
VS+LS̄F
OUTPUT TO DECODER AND DISPLAY
f−l/2 DELAY OF 1 FIELD TIME LESS ONE-HALF LINE TIME
l DELAY OF 1 LINE TIME
L,V WAVEFORMS SWITCHING AT LINE RATE AND FIELD RATE RESPECTIVELY (SEE FIG.3B)
S̄F INVERSE OF SIGNAL S DELAYED BY ONE FRAME TIME

Fig. 3a—Experimental arrangement for testing simple interframe coder. Alternate fields or lines are fed to the frame memory via switch S_1. Switch S_2 selects either the previous field or lines from the previous frame for recirculation in the (delay line) frame memory. Adder A_1 performs the vertical averaging and adder A_2 performs the temporal averaging. Switch S_3 selects the required output.

rate, is shown in Fig. 3b as two waveforms of opposite phase. One phase applies to line numbers 1, 2, 5, 6, 9, 10, $\cdots$, or $4n$ and $4n + 1$, where n is an integer, and the opposing phase applies to line numbers 3, 4, 7, 8, $\cdots$, or $4n + 2$ and $4n + 3$. Thus, during one field the value of L changes each successive line, and for a given line the waveform, L, changes polarity for each frame. Figure 3b also shows which fields are present at points A, B, C (Fig. 3a) for each of the two sets of



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
A	0	1	0	1	4	5	6	$\bar{6}$	8	$\bar{8}$	8	$\bar{8}$	12	13	12	13	FOR LINE NUMBERS 4n, 4n+1
B		$\bar{0}$	$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{\bar{6}}$	$\bar{8}$	$\bar{\bar{8}}$	$\bar{8}$	$\bar{\bar{8}}$	12	13	12	
C			0	1	0	1	4	5	6	$\bar{6}$	8	$\bar{8}$	8	$\bar{8}$	12	13	
O/P			0	1	(0,4)	(1,5)	4	5	6	(6,8)	8	$\bar{8}$	(8,12)	($\bar{8}$,13)	12	13	
A	−2	−1	2	3	2	3	6	$\bar{6}$	$\bar{6}$	$\bar{8}$	10	11	10	11	14	15	FOR LINE NUMBERS 4n+2, 4n+3
B			$-\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{2}$	$\bar{3}$	$\bar{6}$	$\bar{\bar{6}}$	$\bar{8}$	$\bar{\bar{8}}$	$\overline{10}$	$\overline{11}$	$\overline{10}$	$\overline{11}$		
C			−2	−1	2	3	2	3	6	$\bar{6}$	8	$\bar{8}$	10	11	10	11	
O/P			(−2,2)	(−1,3)	2	3	(2,6)	(3,$\bar{6}$)	6	($\bar{6}$,$\bar{8}$)	8	($\bar{8}$,11)	10	11	(10,14)	(11,15)	

Fig. 3b—Waveforms L, V, and S and fields present at points A, B, C (Fig. 3a) and at the output for both sets of lines. In real time, waveform L switches at the line rate but is shown here as two waveforms, one for each set of lines switching at the frame rate.

lines, for the same sequence of fields and modes as described in the previous section and shown in Fig. 2. In the stationary mode ($S = 0$) the required output can be obtained either by continuously averaging the signals present at A and C or by switching S_3 at the line rate; the latter method is used because it allows greater simplicity when changing modes.

In the moving mode S_1 and S_3 are switched at the field rate (waveform V) and Fig. 3b again shows which fields are present at points

TABLE I—QUANTIZING LEVELS

Level	Decision Level	Representative Level
0	1/2	0
±1	1-1/2	1
±2	3	2
±3	5	4
±4	7-1/2	6
±5	11-1/2	9
±6	16-1/2	14
±7	22	19
±8		25

A, B, C of Fig 3a; the output is formed by switching between C and (A + C)/2 at the field rate. To bring about the same sequence of outputs near the mode changes, as shown in Fig. 2, the waveform-controlling switches S_1 and S_2 are changed as soon as S changes, but the change of waveform-controlling switch S_3 is delayed by 1 frame. The sequences of fields present in the output in the experimental system (Fig. 3a) are also shown in Fig. 3b and correspond to the sequences shown in Fig. 2c.

In the experimental system, movement is deemed to be present if, during any field, 512 or more pels exhibit a frame difference amplitude greater than 15 levels (out of 255). To return to the stationary mode, less than 512 picture elements must have a frame difference amplitude greater than 15 levels during the even field where the field number is counted from the field in which the transition is made to the moving mode.

III. RESULT

The picture quality was consistent with the results of the preliminary experiments conducted on the separate modes as described above. The two faults of the moving mode, the aliasing patterns in areas of strong vertical detail and the slight jerkiness of fast-moving contrasty subjects, were still visible. Degradation due to the switching of modes was seldom visible and the effect of switching the whole picture instead of just the moving areas was not troublesome even when the plane of sharpest focus was midway between the subject's head and the curtains in the background.

In some related experiments the switching of modes was confined to the moving area. The moving area detector examined a sequence of eight frame differences to decide whether the current picture element belonged in a moving area.[1] The resulting pictures of similar scenes viewed under the same conditions were no more pleasing and the movement detector setting was more critical; i.e., with a poor setting slowly moving sharp edges tended to break up due to intermittent mode switching.

IV. DISCUSSION

4.1 *Delay Requirements*

At the transmitter every second pel is delayed by one line period, or one field period (according to the mode), to bring about a constant bit rate. This delay can also be used by the movement detector for generating the required frame differences. Of course, the movement detector must now work on one-quarter as many points as before. But because of the global nature of the decision and the large number of supra-threshold points required to change to the moving mode (512), reliable detection of movement could probably be performed with even less than this number of points. Thus, if the primary (EDQ) coder has an output bit rate of 6 Mb/s (or 200,000 bits per frame) a delay store of 50,000 bits (half a field) is required at the transmitter.

At the receiver a similar delay is needed to convert the incoming constant bit rate to the required 6 Mb/s rate during alternate lines or fields and a frame memory of 200,000 bits is needed to store the pels required for display.

It is tempting to devise schemes in which the alternate fields or lines of data from different sources are multiplexed so that the 50 k-bit delay stores at the transmitter and receiver are unnecessary. However, such schemes have so far been less practicable than using the extra storage; the main difficulty lies in synchronizing any two cameras in an unsynchronized system.

4.2 *Recoding of Output*

When the primary decoder is located remotely from the interframe decoder, the unsampled fields or frames (depending on whether there is movement or not) must be recoded since, when different quantized differences (representative values) are averaged, the resulting difference will not necessarily belong to the set of representative values allowed by the in-frame decoder. A circuit to achieve this is shown in Fig. 4.* The averaged difference signal is added to the error term from the previously quantized level: the output of the quantizer is subtracted from the input to form the new error term. The advantage, previously mentioned, of reducing granular noise in the stationary mode is lost in the process of recoding. The effect of recoding was tested experimentally by differentially quantizing the output of switch S_3 (Fig. 3a) before decoding and displaying the signal. There was no appreciable increase in noise (over the primary coded signal) due to recoding either in the stationary mode or the moving mode.

In a switching system it may be desirable to convert and reconvert the signal between the full rate and the half rate many times. This

* This is Cutler's error feedback coder.[5]

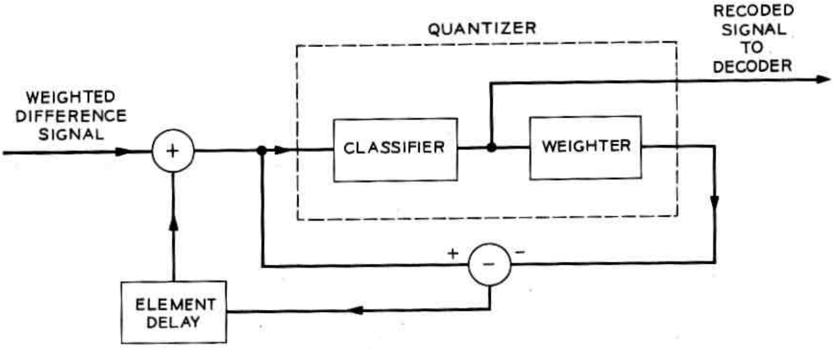


Fig. 4—Recoder to convert averaged weights to standard weights.

can be achieved with no loss of quality other than that introduced in the initial conversions by having the full-rate signal contain one code word in each field to indicate the sampling mode; thus no further movement detection need be employed and subsequent interframe encodings are simpler than the initial one. Decoding will remain unchanged.

4.3 *Comparison*

Compared with other interframe coders[6,7] the simple interframe coder has a higher bit rate and, under certain conditions, a poorer picture quality. However, there are certain offsetting advantages which are described below.

One advantage is the relatively low memory requirement (300,000 bits for the transmitter and receiver combined). Other existing interframe coders require 530,000 bits of storage for the frame memory at the transmitter and the same again at the receiver, and sophisticated buffers are also required because of the randomness in the generated bit rate of these coders. It is possible to use smaller stores in such coders by requantizing the input signal. This operation produces a small loss in picture quality and also it is not yet clear what effect it will have on recoding the signal.[8]

The simple interframe coder can be used with any primary coder which uses only previous pels along the line for prediction (see, for example, Refs. 4, 9, 10). The primary encoding stage may well be located at the first level of switching while the frame-to-frame coding section may be located at a higher level in the switching hierarchy. With such an arrangement, the secondary encoding stage can be bypassed altogether when less than half the *Picturephone* trunk connections in a group of trunks is not in use.

The effect of channel errors on the received picture is probably about the same as in a standard DPCM system. If element-to-element differential encoding were used in the primary encoder, errors would be confined to a single line (in the received signal) as the accumulators at the transmitter and receiver are reset at the end of each line.[4] If the movement detector is in the stationary mode, the effect of an error in one frame will also be displayed at half-amplitude in neighboring frames; but on the other hand, no errors will originate in the neighboring frames. Similarly, if an error occurs in the moving mode, the effect will appear at full-amplitude in the sampled field and at quarter-amplitude in the two adjacent lines in each neighboring field, but the neighboring fields will contain no indigenous errors. Most error detection techniques that can be applied to the primary encoding section can be used without modification when the frame-to-frame coding section is added (e.g., the check-summing method of error detection suggested for the differential quantizer).[4]

Apart from one mode-bit transmitted every second field the coder has no change-of-mode words, addresses, start-of-line words, or start-of-frame words. In a coder using these special words, an error in the received data, if it occurs in a special word, can be especially troublesome.

4.4 *Encoding Using the Previous Line*

There may be occasions when the primary coder uses pels from the previous line (in the same field) as well as from the present line for predicting the value of the current pel (see for example, Ref. 11). The stationary mode described above is now unsatisfactory because only every other line in each field is available at the receiver. We therefore modified the apparatus of Fig. 3 to evaluate a coder in which alternate frames are transmitted in the stationary mode and temporal interpolation is used to replace the unsampled frames. The moving mode is unchanged and whole fields are now left intact in both modes. The change from a stationary to a moving mode can now be made only at the end of a frame rather than at the end of a field, as before. In addition, the delay stores for converting to and from a constant 3Mb/s rate are twice as large as required before, but the total required memory is still less than half that of other existing interframe coders. Experimental results with this coder gave pictures which were as satisfactory as those already described.

V. ACKNOWLEDGMENTS

We would like to acknowledge the many discussions with A. J. Seyler of the Research Laboratories of the Australian Post Office and with our colleagues, J. C. Candy, D. J. Connor, B. G. Haskell, and F. W. Mounts. We are also indebted to W. G. Scholes for his technical assistance.

REFERENCES

1. Limb, J. O., and Pease, R. F. W., "Exchange of Spatial and Temporal Resolution in Television Coding," paper presented at S.M.P.T.E. Meeting, Los Angeles, October 1969; also B.S.T.J., *50,* No. 1 (January 1971), pp. 191–200.
2. Cunningham, J. E., "Frame Correction Coding," Symposium on Picture Bandwidth Compression, MIT, Cambridge, Mass., April 1969.
3. Brainard, R. C., Mounts, F. W., and Prasada, B., "Low-Resolution TV: Subjective Effects of Frame Repetition and Picture Replenishment," B.S.T.J., *46,* No. 1 (January 1967), pp. 261–271.
4. Limb, J. O., and Mounts, F. W., "Digital Differential Quantizer for Television," B.S.T.J., *48,* No. 7 (September 1969), pp. 2583–2599.
5. Cutler, C. C., "Transmission Systems Employing Quantization," U. S. Patent No. 2,927,962, March 8, 1960.
6. Mounts, F. W., "A Video Encoding System Using Conditional Picture-Element Replenishment," B.S.T.J., *48,* No. 7 (September 1969), pp. 2545–2554.
7. Candy, J. C., Franke, Mrs. M. A., Haskell, B. G., and Mounts, F. W., "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., this issue, pp. 1889–1917.
8. Candy, J. C., and Mounts, F. W., private discussion.
9. Brown, E. F., "A Sliding-Scale Direct-Feedback PCM Coder for Television," B.S.T.J., *48,* No. 5 (May–June 1969), pp. 1537–1553.
10. Bosworth, R. H., and Candy, J. C., "A Companded One-Bit Coder for Television Transmission," B.S.T.J., *48,* No. 5 (May-June 1969), pp. 1459–1479.
11. Connor, D. J., Pease, R. F. W., and Scholes, W. G., "Television Coding Using Two-Dimensional Spatial Prediction," B.S.T.J, *50,* No. 3 (March, 1971), pp. 1049–1061.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Transmitting Television as Clusters of Frame-to-Frame Differences

By J. C. CANDY, MRS. M. A. FRANKE, B. G. HASKELL, and F. W. MOUNTS

(Manuscript received February 16, 1971)

A number of redundancy reduction techniques are used in a coder that is about eight times more efficient than simple PCM. The coder is capable of transmitting Picturephone® *signals at an average rate of one bit per picture-element (2 Megabits per second). When there is movement in the scene, most transmission time is devoted to the parts of the picture that change significantly. The data are generated irregularly but the data flow is smoothed prior to transmission in a buffer that holds about one frame of data. The redundancy reduction techniques used and the behavior of the coder are discussed both from an intuitive and from a statistical viewpoint.*

The positions of elements that change are signaled by addressing the first element of a run of changes and marking the end of the run with a special code word. The changes of luminance are transmitted as frame-to-frame differences using variable-length code words. When rapid motion makes the buffer more than a quarter full, only differences for every second element are transmitted, the values of the intervening changed elements being set equal to the average of their neighbors. If the buffer continues to fill, the threshold that determines which changes are significant is raised from 4/256 to 7/256 of the maximum signal value. When violent motion causes the buffer to fill completely, replenishment is stopped for about one frame while the buffer empties.

Subsampling and raising the threshold are not objectionable because viewers rarely detect the small impairments introduced in moving images. Observers are critical, however, of small impairments in stationary scenes. Thus, to maintain high quality in stationary areas, the entire picture is forcibly updated every three seconds by transmitting 8-bit luminance values for three lines of every frame.

A record of the coder's behavior is available as a 16-millimeter movie film.

I. INTRODUCTION

The introduction of *Picturephone* service has stimulated a great deal of interest in techniques for transmitting video signals more effectively than the way normally used for broadcast television. Since digital transmission over large networks has been found to be more economical for video signals than analog transmission, digital coding is presently receiving most of the attention. One of the best known improvements on simple PCM is differential coding, which takes advantage of the similarity between consecutive samples along a scanning line. Differential coding is simple and yields a bit-rate savings of around 50 percent depending on the picture quality desired.

This paper describes a method that makes use of the similarity of pictures in successive frames as well as the similarity of adjacent samples. Until recently, using the correlation between frames to improve efficiency required complex and costly equipment; but now, with low-cost memories and integrated circuits, it is economically attractive and seems to be particularly suitable for the visual telephone application.

All techniques that reduce the required channel capacity by anticipating a redundancy in the signal have a limitation, in that signals which do not contain the expected redundancy cannot be transmitted unless some alternative process is provided for them. This limitation is possibly the reason why redundancy-reducing techniques have not been used extensively for broadcast television, where there is need to display unusual scenes to attract viewers and entertain them. With visual telephone, however, there is a great need for economical transmission because each signal will not have an audience of millions to share its costs. Indeed it is unlikely that users will want to pay for transmission capabilities that are rarely needed.

Recent work[1–5] has demonstrated how differential coding enables *Picturephone* signals to be transmitted using three or four bits per picture-element. This is about three bits less than simple PCM requires to give comparable quality. Inherent in this technique is a restriction on the amount of detail in the scene that can be reproduced faithfully. A much greater saving has been obtained by F. W. Mounts,[6] using coders that signal only the changes between successive frames. These coders require only one bit of channel capacity per picture-element, the restriction in this case being on the amount of movement that can be accepted in the scene.

It is anticipated that users of *Picturephone* service will want to display scenes containing fine detail much more often than scenes containing rapid motion; and so they would tolerate blurring of moving objects more readily than a similar loss of spatial detail in stationary scenes.[7] Consequently, ways are used for exploiting the correlation between successive frames as well as that between successive elements in order to effectively encode *Picturephone* video signals.

The work to be described here, which is a continuation of F. W. Mounts' original work, improves the coder so that more motion can be tolerated while using a much smaller buffer than was originally required. (A coder that uses no buffer but requires a data rate of 1.5 bits per element is described in Ref. 8). The techniques used here rely primarily on two phenomena:

(*i*) Large areas of the average *Picturephone* scene change very little or not at all between successive frames. Thus, in each frame only a small amount of new information is required to specify these areas to the receiver.

(*ii*) More information is needed for specifying areas of the picture that change. But they need not be reproduced at the receiver with as much spatial resolution as stationary areas because the storage properties of the camera tube tend to blur movements, and viewers cannot accurately resolve fine details that are in motion.

The coder transmits fresh information describing the stationary parts of the picture at least once in three seconds. At the receiver, this information is retained in a frame memory from which the display is derived. Portions of the picture that change significantly are updated as soon as they are detected, but sometimes with slightly lower resolution than in the stationary areas.

There follows a description of the proposed coding strategy and an account of a simulation of the coder that has been demonstrated. *

II. CONDITIONAL REPLENISHMENT

In Ref. 6 F. W. Mounts called his method "Conditional Replenishment." He transmitted only the information necessary to describe the

* A demonstration of the system processing a live scene was shown during the Keynote Session of the 1970 IEEE International Convention.

intensity of elements that changed between frames. The implementation used two frame memories, one at the receiver and another at the transmitter, to store the required reference picture. The incoming signal was compared element by element with the reference picture. When the magnitude of the frame-to-frame difference was greater than a certain threshold, it was regarded as significant,[9] and the new signal value replaced the reference value in the memory. The new value was also transmitted to the receiver, where it updated the contents of the receiver frame memory. In this way the receiver memory was made to track the transmitter memory, thereby providing a signal for display.

Every time an element value changed, two numbers were transmitted, one describing the new amplitude and the other describing the location of the element. In addition, synchronizing words were sent at the start of every frame and every line, making it necessary to address only the horizontal position of changed elements.

For most television pictures, the technique of sending only data relating to elements that change gives a considerable saving of transmission cost. There is a saving when a large fraction of the scene is stationary or when moving objects contain large areas that are uniformly bright; in either case little transmission is required. Part of the saving is offset by the need to define the position of changed elements in addition to their new values. Nevertheless, F. W. Mounts found that, by using a large buffer to smooth the highly irregular data flow, visual telephone signals could be transmitted using, on the average, only one bit per picture-element.

III. THE PICTURE FORMAT AND TYPICAL STATISTICS

The video signal, which is very similar to the *Picturephone* video signal, is derived from a picture scanned with 271 interlaced lines at 30 frames per second. An example is shown in Fig. 1. The bandwidth is nominally 1 MHz, and elements are sampled at about 2 MHz to give 248 samples in a line period. The visible portion is about 13 cm high and about 14 cm wide; it contains 255 lines each with 207 elements. For practical convenience the signal is coded as 8-bit PCM so that all processing can be performed digitally in real time. The positions of elements in the frame are described by signaling the start of every line and addressing their positions only within the line. If an 8-bit address word is used, 206 of its 256 values are needed to address the visible elements on a scanning line. In the simplest case, only three

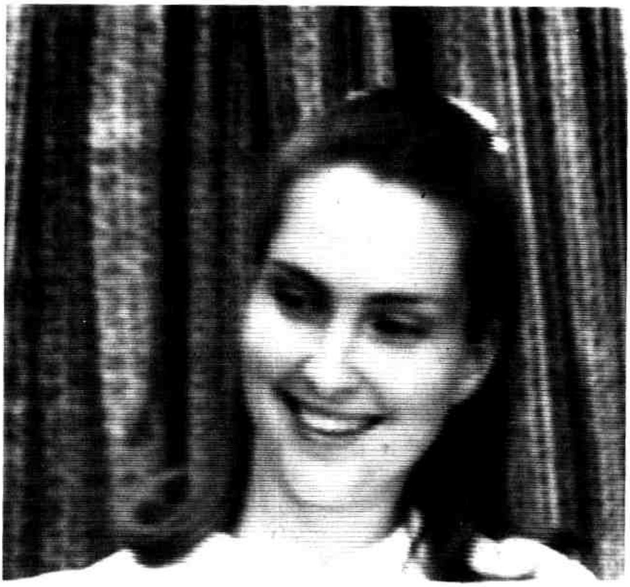


Fig. 1—A typical picture.

others are needed: one to mark the start of an even field, one to mark an odd field, and one to mark the start of a new line. In an operational system, however, all of the remaining 50 values would probably be used to help protect the synchronization from error and to signal special modes.

The frame contains a total of 67,208 elements of which 52,785 are visible. Usually only a small fraction of these elements change between any two frames.[10,11] The ordinate in Fig. 2 shows the number of elements that change in a frame-time by more than 3/256 of the maximum signal amplitude. This graph shows 18 seconds of signal measured when there was an unusually large variety of motion in the scene. For convenience the scale is classified into five ranges corresponding to the viewer's subjective judgment of the activity. Figure 3 is a probability density function of tne number of changes in a frame. It shows how often the various kinds of activity can be expected in typical *Picturephone* scenes. The data were collected from 75 minutes of picture material which included head-and-shoulder views of one and occasionally two persons, camera panning, subjects walking through the field of view, and pictures of printed text. On the average, less than 4 percent of the elements change in a frame-time.

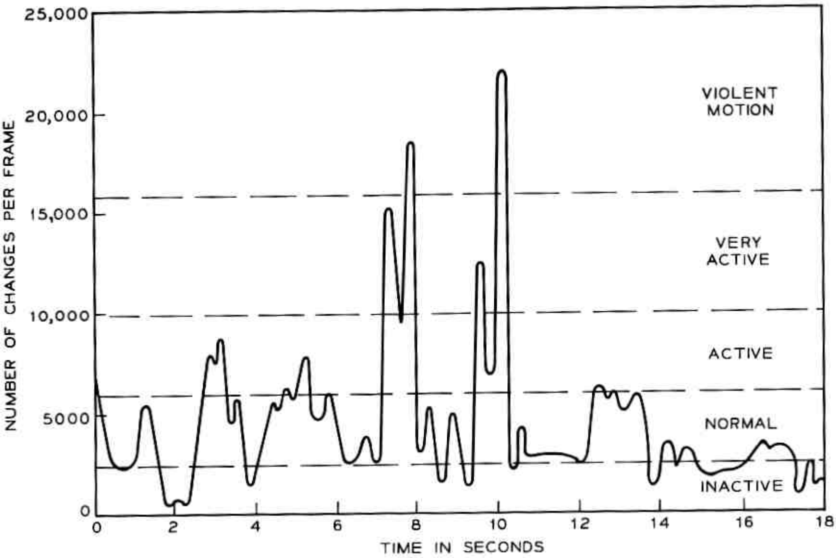


Fig. 2—The number of elements in a frame that change by more than 3/256 of maximum amplitude, plotted over a period of 18 seconds when there was a variety of motion. Each frame contains 67,000 elements of which 53,000 are visible.

IV. THE BUFFER REQUIREMENT

When only 4 percent of the elements change, it is fairly easy to design a coder whose data-rate is less than one bit per element; there are times, however, when the rate is much higher. Thus, a buffer must be used at the transmitter to smooth the highly irregular flow of data from the coder and feed it to the channel at a constant rate. Another buffer at the receiver performs the inverse operation.

It is evident in Fig. 2 that peaks of high activity last for an appreciable part of a second; thus, a buffer to smooth them out would introduce an intolerable delay into the signal path. To prevent delay and echoes in the accompanying voice channel from being a distraction to users of *Picturephone* service, the total delay between talker and viewer should be less than a third of a second;[12] i.e., ten frames. This sets an upper limit on the size of the buffer.

Not much is gained, however, by using a buffer this large. The buffer should be at least large enough to smooth the data over one field-time (two interlaced fields per frame are used throughout), since the changing elements are usually very unevenly distributed within a picture. This is illustrated in Fig. 4 by bright dots placed in the position of elements that have changed in the picture. There seems to be little advantage, however, in carrying data over from one field to another; unless it can be accumulated for many frames, the activity is usually very similar in successive fields. To illustrate this similarity, Fig. 5 gives the probability that the variation in the number of significant changes in successive fields exceeds a stated amount: the number of changes in a field differs from the number in the previous field by more than 500 in only 5 percent of the fields measured. Another study[13] has also confirmed that the buffer should indeed be capable of storing data over a field-time and not very much larger.

It is proposed that the buffer be capable of storing the amount of data that is transmitted in two field-times. This is more than enough to smooth the irregularities caused by spatial distribution of activity and also allow latitude for controlling the coder. Use of this buffer, which is about one tenth the size of the one used by F. W. Mounts,[6] requires that the data fed into it in a frame-time be approximately

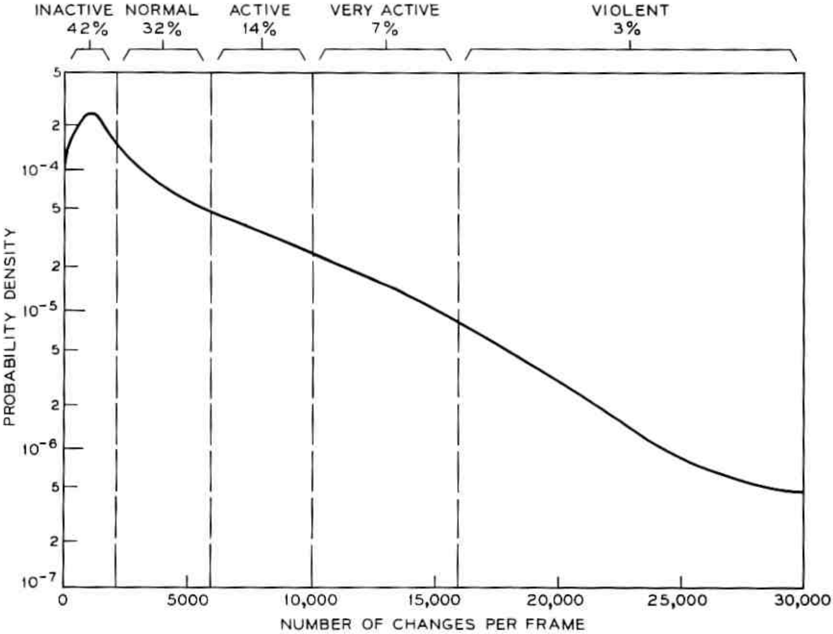


Fig. 3—The probability density function of the number of changes in a frame for 75 minutes of signal. The degrees of motion in the scene are indicated.

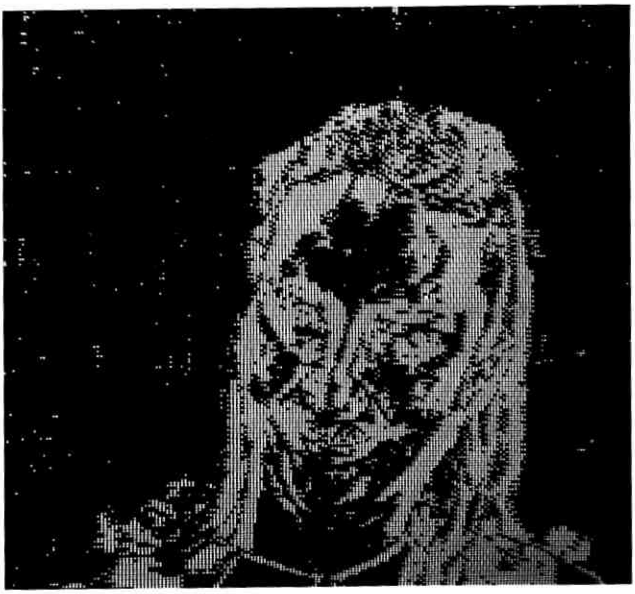


Fig. 4—A display of dots marking elements that have changed in a frame during very active motion: about 12,000 changes occurred in this frame. The picture is that shown in Fig. 1.

equal to what can be transmitted by the channel in a frame-time. To satisfy this constraint on the coder, the average number of bits used to signal a change in an element value must decrease as the activity in the scene increases. For example, Table I gives an estimate of the number of bits available to code each change for various amounts of activity if the transmission rate is 67,000 bits per frame (one bit per element for *Picturephone* use):

From Fig. 3 and Table I, it is seen that conditional replenishment accommodates pictures of normal activity, i.e., 70 percent of the scenes, when transmitting two 8-bit words (address and amplitude) for every change. It will be shown how to extend the range by encoding with fewer bits. Elements will be addressed in clusters instead of individually, and their changed amplitudes will be transmitted as frame-to-frame differences using four bits on the average instead of the eight bits needed to define a completely new amplitude.

V. ADDRESSING ELEMENTS IN CLUSTERS

In the original conditional replenishment system, about half of the transmitted data were used for addressing the positions of changed

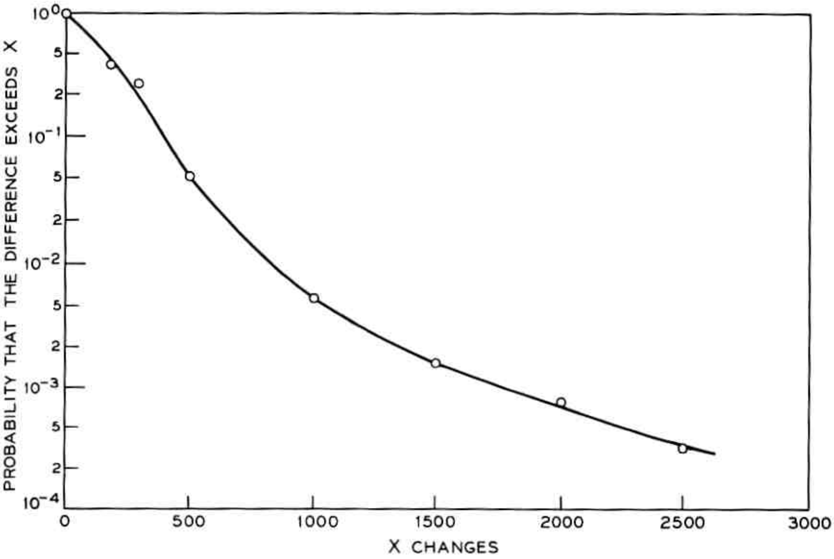


Fig. 5—The probability that the difference between the number of changes in one field and those in the next exceeds X.

elements. It is observed in Fig. 4 that significant changes usually occur in clusters crowded near brightness edges of moving objects. Thus, it is profitable to describe the position of changed elements in groups rather than individually. To explore such methods, the magnitudes of frame-to-frame differences for one whole frame were stored in a memory during a period of active movement. Figure 6 shows an example of how changes are distributed among runs of various lengths. In this example 10,547 elements changed requiring 84,376 bits to address them individually each with eight bits. Using cluster coding, the start of each run is addressed with an 8-bit word, and the end of

TABLE I—NUMBER OF BITS PER CHANGE AVAILABLE AT VARIOUS DEGREES OF ACTIVITY

Description of the Activity	No. of Changes in a Frame of 67,000 Elements	Average Number of Bits Available per Change
Inactive	< 2,000	>30
Normal	≃ 4,000	≃16
Active	≃ 8,000	≃ 8
Very active	≃13,000	≃ 5
Violent	>16,000	< 4

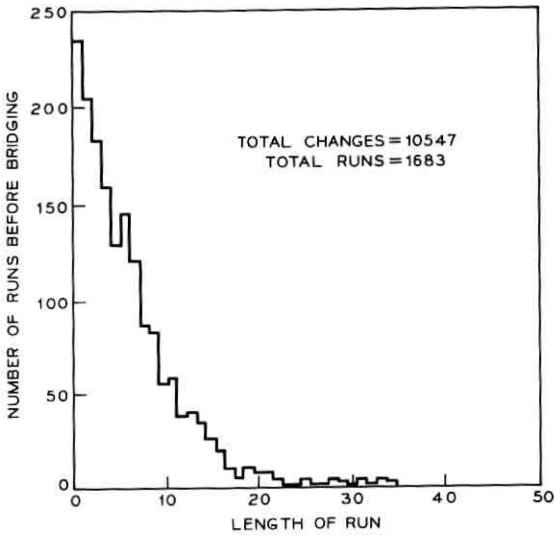


Fig. 6—A histogram of the number of contiguous runs of various lengths in a frame. The motion was very active.

the run is signaled with a special 4-bit word. With this method the 1683 runs of changed elements in Fig. 6 could be positioned using only 20,196 bits; a saving of 64,180 bits, (about 75 percent). It will be shown that an additional saving is possible.

Addressing runs is inefficient when applied to isolated changes in the picture, because 12 bits are used to fix their positions. However, it was observed that most isolated changes can be ignored without spoiling the picture quality. The following criterion provided acceptable quality: *any change in an element, no matter how large, was regarded as insignificant if it were immediately preceded and followed by two elements that changed by less than the threshold of significance.* For the data in Fig. 6, 137 changes could be ignored under this criterion.

Another improvement is obtained by coalescing runs that are separated by a small number of unchanged elements. For example, when using the cluster-coding just described and 4-bit words to signal changes of amplitude, it is preferable to continue a run that is interrupted by less than four unchanged elements, since it requires 12 bits to end a run and restart a new one, but only 4, 8, or 12 bits to continue coding the insignificant changes between runs. The technique of coalescing runs and ignoring isolated changes will be referred to as "bridging." Bridging often improves quality as well as efficiency because some elements, changing less than the threshold, are updated and isolated noise impulses are suppressed. It should be noted that the suppression of isolated changes takes place before bridging. Performing the operations in the reverse order would be somewhat less efficient.

Figure 7 shows the bright dots covering elements that would be coded using the techniques described above and Fig. 8 shows dots at the start of every bridged-run (cluster). Figure 9 shows the distribution of the cluster lengths for the data in Fig. 6. (Notice how bridging significantly lengthens the runs, reducing the number of runs from 1,683 to 736 while increasing the number of elements in the runs only from 10,410 to 11,886.) To transmit these data, about 47,544 bits would signal 4-bit amplitude changes and 8,832 bits would signal addresses. Figure 10 is plotted the same as Figs. 6 and 9 but with the threshold raised to 7 parts out of 256. Here the changed elements are grouped more closely together, and there are fewer of them.

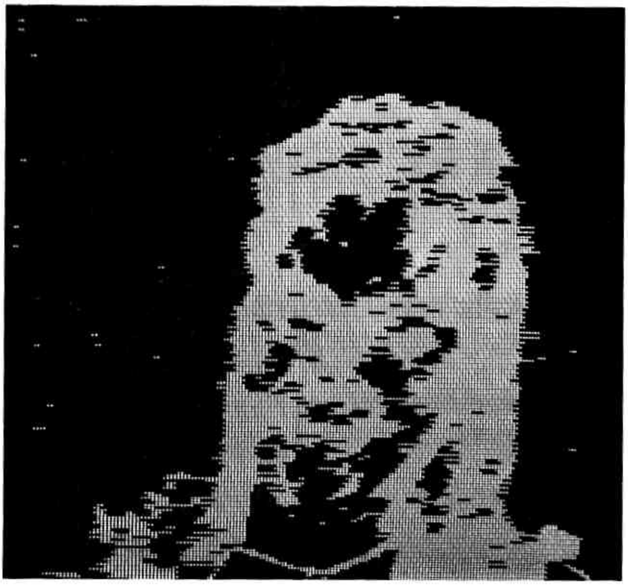


Fig. 7—A display of elements that would be transmitted using cluster coding for the changes in Fig. 4. Isolated changes are ignored and gaps of three or less are bridged.

Fig. 8—The dots show the start of every cluster.

This frame requires 32,516 bits for amplitudes and 9,132 bits for addresses.

In general, cluster coding reduces the number of address-bits significantly. Application of cluster coding to the data used in Fig. 3 is illustrated in Fig. 11. This graph shows that the number of clusters increases much slower than does the number of changes in the picture. Indeed when there are more than 10,000 changes, the number of clusters is relatively constant; therefore, cluster coding is very effective for smoothing the generation of address bits. The next two sections describe methods used to reduce and regulate the generation of bits needed to signal amplitude-changes.

VI. FRAME-TO-FRAME DIFFERENCES

6.1 *Amplitude Distribution of Difference Signals*

While discussing cluster coding, we allowed 4-bit words for describing frame-to-frame differences of the video signal. This difference signal is generated at the transmitter when the reference is subtracted from the input. Its amplitude can range between ±255 units but usually it is small, as Fig. 12 demonstrates. Figure 12 shows how the magnitudes of the significant frame differences are distributed statistically for typical *Picturephone* scenes. The motion in the scene was normal for curve (a) and very active for (b). The circuit used for generating the frame differences is shown in Fig. 13. The quantizer was bypassed while the data was being taken.

The frame-to-frame difference signal may be quantized and transmitted with much less than the eight bits needed to send the absolute amplitudes. For example, when the quantizer is used in the circuit of Fig. 13 (output levels correspond to the eight divisions of the abscissa of Fig. 12), the displayed picture has good quality for still and slowly moving scenes. However, there is noticeable distortion of rapidly moving edges: a sharp, moving edge appears as a cascade of smaller steps as each increase of luminance corresponds to the magnitude of the outer quantization level (43 units). This distortion is eliminated by providing more quantization levels, for example 64 levels can give excellent reproduction for all types of activity in the scene.

Another important reason for using more than 16 levels is to reduce

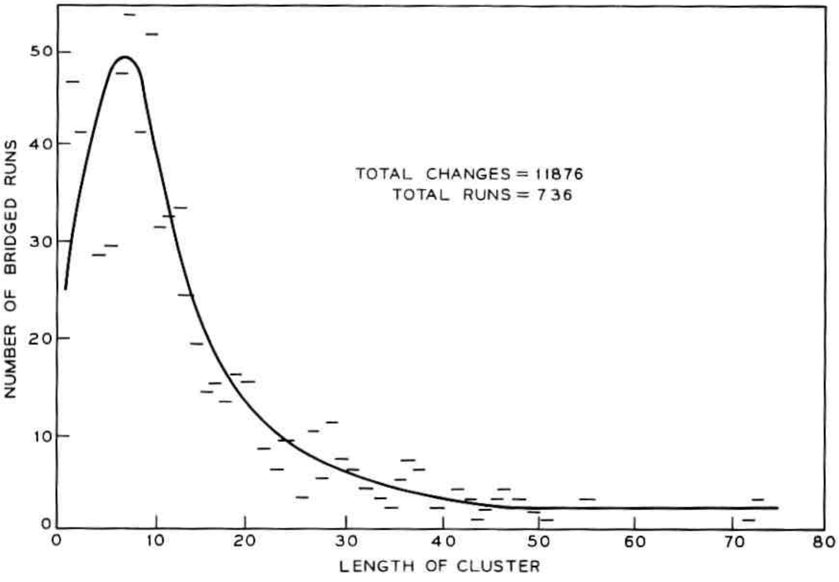


Fig. 9—A histogram of the number of clusters of various lengths for the runs shown in Fig. 6. Isolated changes are ignored and gaps of three or less are bridged.

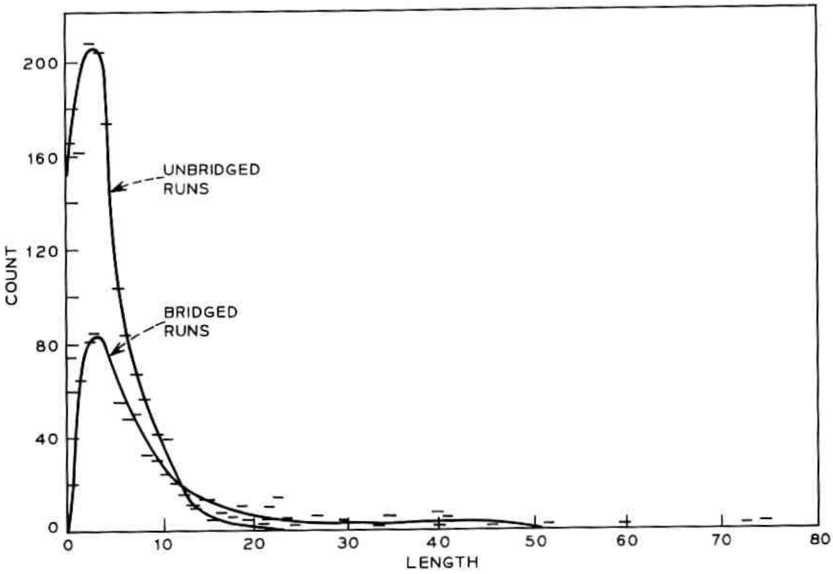


Fig. 10—The effect of raising the threshold from 4/256 to 7/256 on the runs and bridged-runs given in Figs. 7 and 9. Changes that equal or exceed the threshold are counted.

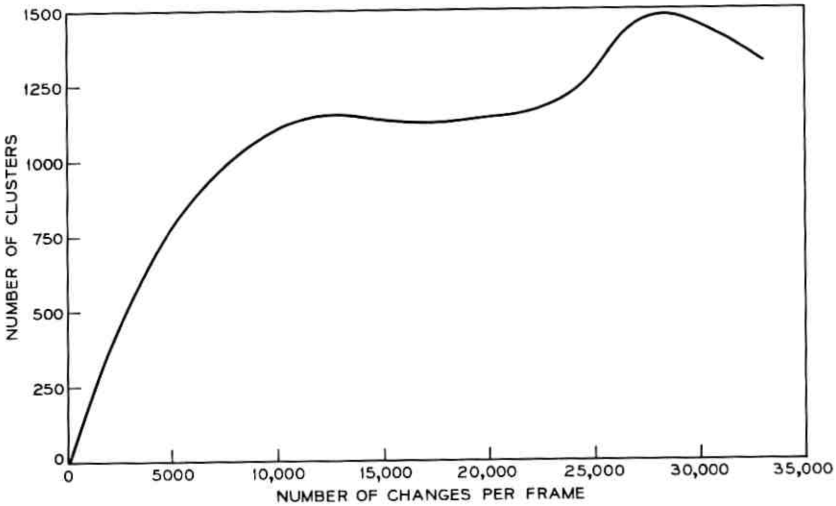


Fig. 11—The average number of clusters in a frame plotted against number of changes. The threshold was 4/256.

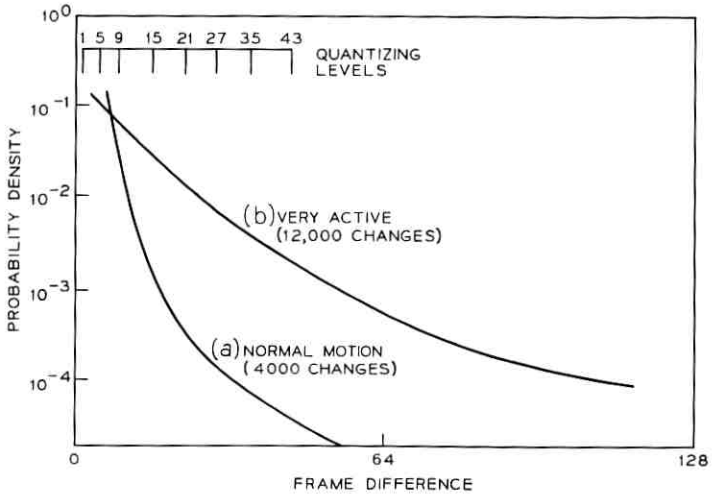


Fig. 12—The probability distribution for the amplitude of frame-to-frame difference signals during normal and active motion. Its maximum is 255 units.

the residual discrepancy between the reference and the input. The stored reference is updated by addition of the quantized difference signal; therefore, its updated value will differ from the input by an amount dependent upon the quantizing scale even after movement has ceased. If, in the next frame time the discrepancy exceeds the threshold, transmission capacity will be used to correct it even though the input signal remains unchanged. This need for additional transmission, after large changes have ceased, makes the system inefficient for certain types of input. Therefore, the coder uses a quantizer that can represent any change to within ±4 units of its true value. The quantizing scale

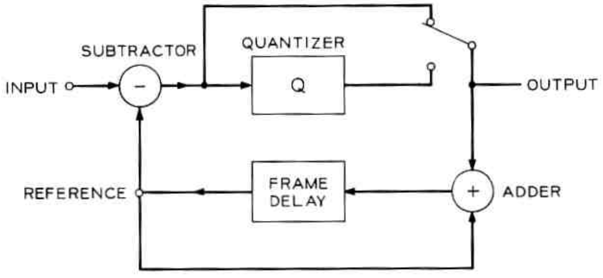


Fig. 13—The circuit used for measuring frame-to-frame difference.

has 64 levels distributed thusly:

$$\pm 1,\ \pm 5,\ \pm 10,\ \pm 15,\ \pm 20,\ \pm 27,\ \pm 35, \cdots$$

increasing in increments of ±8 up to ±235.

A signal quantized to 64 levels can be signaled with 6-bit words, but since the outer levels are rarely used, we use only a 4-bit word to signal the innermost 14 levels. The 4-bit word has sixteen distinct values. One of the remaining two values is used to specify the end of a cluster. Whenever the magnitude of the next frame difference exceeds 39 units the remaining 4-bit word is used to tell the receiver that the next 12 bits should be decoded as two 6-bit words rather than three 4-bit words. These 6-bit words specify the frame differences using the full set of 64 levels. Six-bit words continue to be transmitted in pairs until the last word represents a level lying in the innermost 14 levels; then we revert to 4-bit words.

Figure 14 shows a picture with bright spots marking the elements that are updated with 6-bit words; they occur in groups near large rapidly moving edges in the picture. On the average, 6-bit words are used for less than 10 percent of the updated elements. We find that

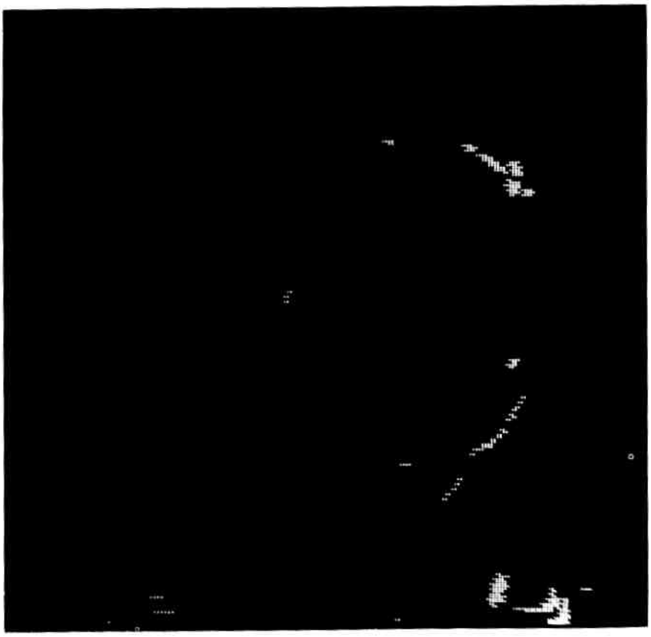


Fig. 14—The changes that are signaled with 6-bit words for the data in Fig. 4.

this method of transmitting quantized frame differences requires only about half the transmission capacity needed for sending the true picture-element brightnesses as 8-bit PCM. However, when differences are transmitted, a mechanism is needed for preventing transmission faults from introducing permanent errors into the signal. This mechanism is described in section 8.2, 'Forced Updating.'

6.2 *Subsampling*

The techniques described in the previous sections allow pictures to be transmitted using, on the average, less than six bits per changed element. Thus, pictures with active motion can be accommodated. However, to be useful, the range of the system must be extended even further. This extension is obtained by making use of the fact that in moving parts of the picture the sampling frequency can be halved, yet not cause noticeable degradation.[7]

A measure of the activity in the scene is obtained by monitoring the number of bits held in the buffer. When it exceeds a prescribed amount, indicating active motion, we switch to a subsampling mode of operation. In this mode only the changes in every other element in a cluster are transmitted and used to update the stored reference pictures. The amplitudes of the intervening elements in the cluster are set equal to the average of the two neighboring values. Table II shows the states of a sequence of sixteen elements having, initially, reference values A_1, A_2, $\cdots$, A_{16}. The new input values are assumed to be those listed on line (b) where, in every case, the change from value A to value B exceeds the threshold. Line (c) shows which difference signals are transmitted in the normal mode of operation. The change A_3 to B_3, an isolated change, is not transmitted, but the stationary element value A_{10} is transmitted in order to bridge the cluster. Line (d) shows the new reference, the apostrophe signifying that a quantization has occurred in the value. If the coder were operating in the subsampling mode, the transmitted differences would be those marked on line (e), and the new reference would be those values on line (f). Here, the symbol C represents the average of values on either side of the element; for example,

$$C_9 = \frac{B_8' + B_{10}'}{2}$$

When subsampling, the set of elements whose values are considered for transmission are arranged in a fixed checkerboard pattern in the raster. This pattern is unchanged from frame to frame in order that discrepancies between the reference and the input caused by the inter-

TABLE II—SELECTION OF ELEMENTS FOR TRANSMISSION IN BOTH THE NORMAL AND THE SUBSAMPLE MODE

A_n represents the original reference amplitude of the nth element
B_n represents a new value of an amplitude
B_n' represents a quantized amplitude
C_n represents an interpolated amplitude

	(a) Reference	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
	(b) Input	A_1	A_2	B_3	A_4	A_5	A_6	B_7	B_8	B_9	A_{10}	B_{11}	B_{12}	A_{13}	A_{14}	A_{15}	A_{16}
Normal Mode	(c) Elements selected for transmission							X	X	X	X	X	X				
	(d) New reference	A_1	A_2	A_3	A_4	A_5	A_6	B_7'	B_8'	B_9'	B_{10}'	B_{11}'	B_{12}'	A_{13}	A_{14}	A_{15}	A_{16}
Sub-sample Mode	(e) Elements selected for transmission								X		X		X				
	(f) New reference	A_1	A_2	A_3	A_4	A_5	A_6	C_7	B_8'	C_9	B_{10}'	C_{11}	B_{12}'	C_{13}	A_{14}	A_{15}	A_{16}

polation shall not instigate a transmission; moreover, a changing pattern is subjectively unpleasant.

Using the subsampling technique, very active motion can be accommodated in the scene, because, on the average, less than four bits are needed to signal each significant change. Although the spatial resolution is effectively reduced by a factor of two in regions of motion, it is almost impossible to see. Indeed, additional loss of resolution can be tolerated to further extend the range of the coder. This extension is obtained by raising the threshold[6] that determines which changes are significant.

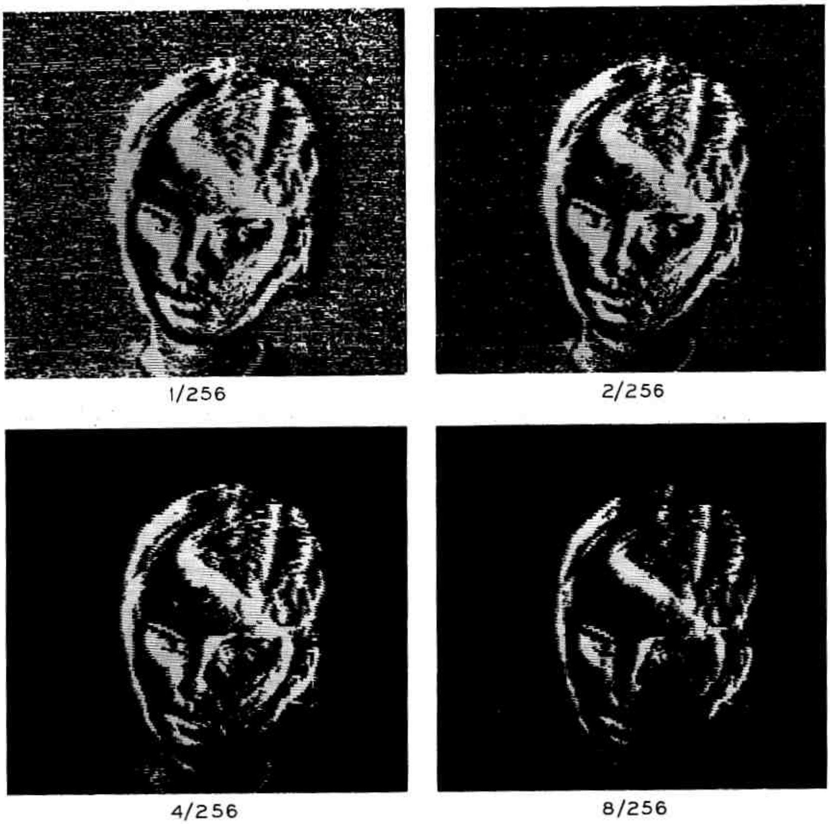
6.3 *The Threshold*

Normally the threshold is set to accept changes greater than or equal to 4 parts out of 256; this gives excellent reproduction of motion, yet prevents small amounts of noise from needlessly using transmission capacity. When there is active motion in the picture and the buffer starts filling in spite of subsampling, the threshold is raised one unit at a time as the buffer fills up, finally reaching a maximum threshold of seven units.

When the threshold is at a high value the unreplenished changes appear as a stationary noise, which has the appearance of a dirty windowpane placed before the scene. With threshold values in excess of eight this noise is clearly visible in dark, moving areas of the picture. When the threshold is less than eight the picture impairment is small, and with thresholds less than five the noise is unnoticeable. Figures 15 and 16 show how raising the threshold reduces the number of changes that are counted as significant. Besides reducing the transmission requirement, raising the threshold also makes the coder more tolerant of noise in the input video. In fact, the coder is capable of processing signals that have been contaminated with noise whose root-mean-square value is 35 dB less than the peak signal value. Such high noise levels would be unacceptable for commercial service; we expect noise to be 45 dB less than the signal.

VII. BUFFER OVERLOAD

The techniques we have described so far allow the majority of *Picturephone* scenes to be encoded and signaled at an average rate of one bit per picture-element. Only occasionally does the motion become so violent that the system is congested. Such situations occur when the camera is panned over a very detailed scene or when the



1/256

2/256

4/256

8/256

Fig. 15—A display of the elements whose changes equal or exceed thresholds of 1/256, 2/256, 4/256, 8/256.

subject suddenly stands up close to the camera; these instances are usually of short duration. To prevent the buffer from overflowing at these times, we simply inhibit all replenishment of memories at the end of the cluster during which the buffer becomes 99 percent full. Replenishment is inhibited until the buffer content falls to about 2,500 bits, which takes about one frame-time. Afterward, coding resumes with the constraint that it remain in the subsampling mode with a threshold of seven for at least one entire field.

While replenishment is inhibited, the data previously stored in the memory are displayed again, and the motion in the picture becomes somewhat jerky.[14] Viewers have not found the operation to be very objectionable, and some people have grown accustomed to seeing jerkiness in vigorous movement. However, this is one aspect of the coder that can probably be improved by using the vertical correlations in the picture[8] in much the same way as subsampling uses the horizonal correlations.

VIII. FIXED BACKGROUND TRANSMISSION

8.1 *Start-of-Line Codes*

In previous sections we have concerned ourselves with describing techniques for signaling changes which occur irregularly in the picture. Signaling addresses and amplitude changes use about 90 percent of the transmission capacity; the remaining 10 percent represents the synchronization and forced updating. These data are transmitted independently of picture content unless the buffer fills; then the forced updating is interrupted for a frame-time but the synchronization is always transmitted in order to keep the transmitter and receiver in step.

We have already mentioned that the start of each new line of the raster is signaled with a selected value of the 8-bit address words. (A

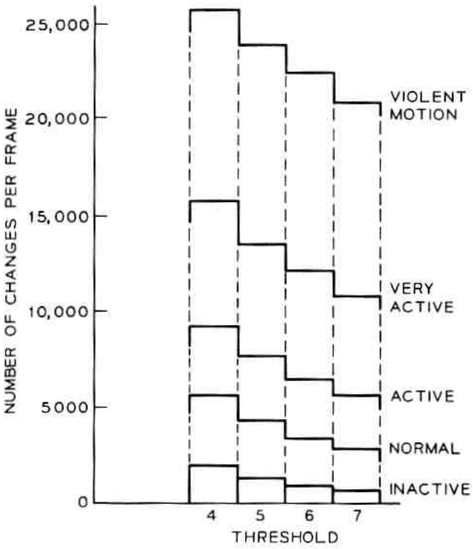


Fig. 16—A histogram of the number of changes in a frame that exceeds various thresholds for typical scenes.

suitable choice of the code will not be discussed in this paper, but notice that 50 code words are available after 206 have been reserved for addressing the positions of elements on the line.) Changes from the normal to the subsampling mode will be signaled by a change in the start-of-line code. A special code is also used at the start of the first line in each field and at the start of lines that are to be forcibly updated.

Using 8-bit words for marking starts of lines requires about 2,000 bits in each frame-time, which is about 3 percent of the intended transmission capacity.

8.2 *Forced Updating*

Forced updating is a process for periodically rewriting the entire picture with full amplitude values. It serves two purposes: To insure a very high quality reproduction of stationary parts of a scene and to correct errors introduced during transmission by periodically aligning the contents of the transmitting and receiving memories.

This forced updating is accomplished by transmitting in each frame-time, three lines of the picture as 8-bit PCM. The three lines are evenly spaced in the raster and are moved up each frame-time so that the entire picture is updated in about three seconds: Thus, less than three seconds after a portion of the picture became stationary it acquires the quality obtainable with 8-bit-PCM transmission. Nearly perfect reproduction can be obtained for graphics and similar scenes that display spatial detail rather than movement.

Forced updating of three lines in each frame uses about 5,000 bits or about 8 percent of the transmission capacity. Thus, about 11 percent of the transmitted information is fixed.

IX. PREVENTING THE BUFFER FROM EMPTYING

We have described ways for preventing buffer overflow by reducing the rate at which data are generated, always taking care to maintain a satisfactory quality in the transmitted picture. The problem of preventing the buffer from becoming empty is much easier to solve because there are many ways of acquiring data in order to improve the quality of the picture. We have chosen to use the forced updating mechanism for this purpose. The next line of input data is "forcibly updated" whenever the buffer content falls below 2,500 bits. When there is no motion, the entire picture is updated, within a quarter of a second, by this technique. This helps to obtain a high-quality reproduction of stationary scenes, such as graphic material, and always provides data for the buffer even during the vertical fly-back interval.

X. THE SIMULATION TRANSMISSION EXPERIMENT

Having described the coding strategy, we will now describe a laboratory simulation of the system.

The simulator used produces a received picture which has all the characteristics of one processed by a real system except that effects of transmission delay and digital transmission error are not included.

The simulator is a more complex version of the test circuit shown in Fig. 13; its main features are given in Fig. 17. The video signal enters at the upper left where it is sampled and coded as 8-bit PCM. This digital coding is required because it is impractical to build analog memories and analog processing circuits that are sufficiently accurate to perform the operations previously described. The coded input signal is compared with the reference signal emerging from the frame memory. The difference is then fed to a quantizer and to a threshold circuit that determines whether or not the change is significant. If the difference is insignificant, the threshold circuit tells the control

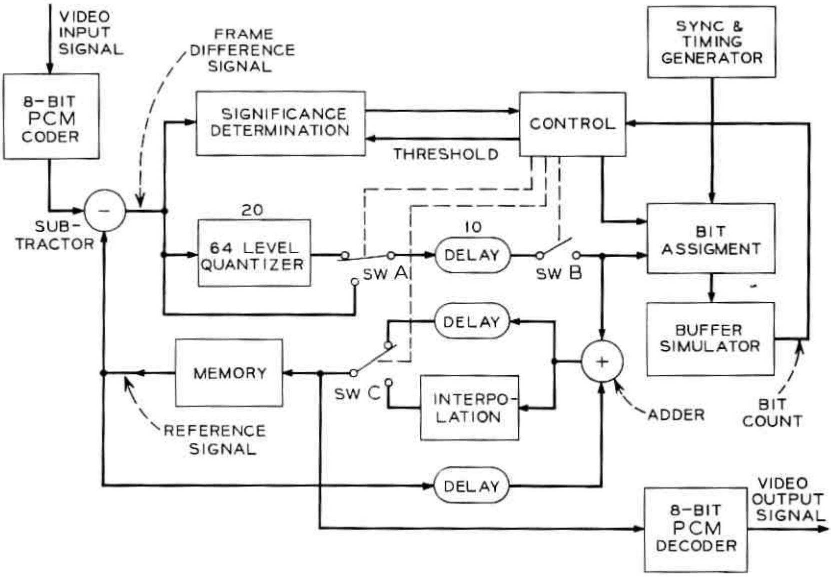


Fig. 17—The laboratory system used for simulating the behavior of the coder. It operates directly on real video signals.

to open switch B to prevent the quantized signal from being fed to the buffer or to the memory. When the first significant difference is encountered, starting a cluster, the control closes switch B so that the quantized difference can be fed toward the buffer and to the adder. The adder combines the difference with the reference that is being circulated from the output of the frame memory. Thus, for significant changes the reference picture is brought up-to-date by adding the quantized difference to it. For insignificant changes, it remains unchanged in the memory since zero is added to it.

The purpose of the delay included before switch B is to give the control circuit enough time to decide whether to start or end a cluster. Recall that changes not accompanied by other changes are ignored, and clusters are ended only when the next four picture-elements change less than the prescribed threshold. The delay in the memory feedback path insures that the reference is added to the corresponding quantized difference.

In a real system the significant frame-differences would be coded and transmitted to the receiver to update the contents of its memory as is done at the transmitter. But in the simulator we avoid the need for a transmission link and a receiver by taking the display signal from the input of the frame memory. Although we have avoided the need for transmission, we still need an indication of the fullness of the transmitting buffer in order to control the behavior of the coder as the buffer fills up. This is accomplished by counting the number of bits that would be sent to a buffer in a real system, and subtracting the amount that would be transmitted. The circuit labeled "bit assignment" determines the number of bits that should be fed to the counter in order to simulate a buffer. It receives the quantized differences, and determines the number of bits needed to code them. It also receives signals from the control circuit which mark the start and end of each cluster of changing elements and a signal to indicate when forced updating occurs. From the address generator it receives signals that mark the start of each scanning line.

A count of the number of bits in the buffer is fed to the control circuit where it is used to determine the operating mode of the coder. The diagram in Fig. 18 illustrates this control function; it shows that the threshold is raised from four to five, and then to six, and finally to seven as the count increases from 10,000 to 20,000, to 35,000, and to 50,000. At a count of 65,000 all replenishment is stopped until the count falls below 2,500. The system starts subsampling when the count exceeds 20,000 and does not return to full sampling until the count

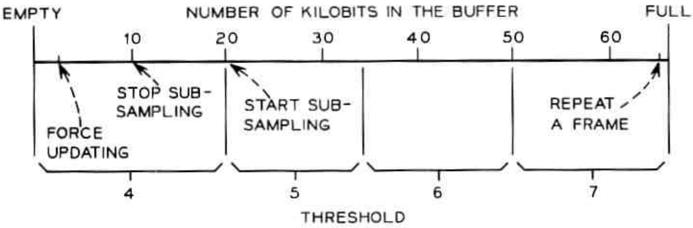


Fig. 18—A representation of the control functions associated with the buffer.

falls to 10,000.* The subsampling is accomplished by simultaneously opening switch B and moving switch C to its lower position for every other element in the cluster. The signal fed to the memory and to the output is then an interpolation of adjacent values.

Switch A is closed to bypass the quantizer and switch C removes the interpolation when lines of the picture are "force updated" with their true values.

XI. BEHAVIOR OF THE CODER

11.1 *Qualitative Behavior*

When the simulator processes the signal in a *Picturephone* transmission, very little impairment can be seen in the received picture. Indeed the received picture appears to be the same as the transmitted one except for two conditions which rarely occur for normal scenes. One is the jerkiness introduced into scenes that change violently over a large area; the other is a moiré pattern that can be seen when a graticule of vertical bars moves enough to require the subsampling mode. The moiré pattern, or aliasing, is a transient that vanishes as soon as the graticule becomes stationary and full sampling resumes.

11.2 *Quantitative Behavior*

Figure 19a shows the probability of buffer content exceeding Q bits with a subject moving vigorously. Figure 19b shows the corresponding probability density function. We see that, most of the time, the buffer is almost empty; only two percent of the time is it more than three quarters full. This indicates that there is considerable advantage in sharing buffers and transmission channels amongst several coders. Figure 20 shows how many bits are used, on the average,

* Choice of these parameters is not critical.

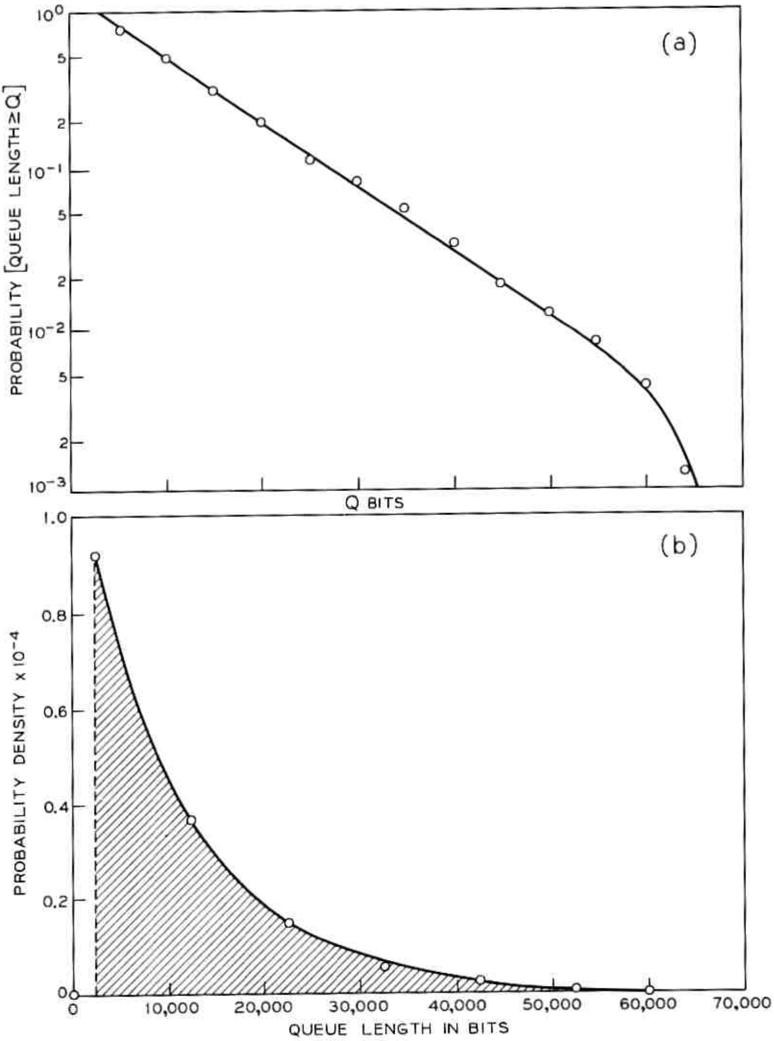


Fig. 19a—The probability of the buffer content exceeding Q bits.

Fig. 19b—The corresponding probability density function.

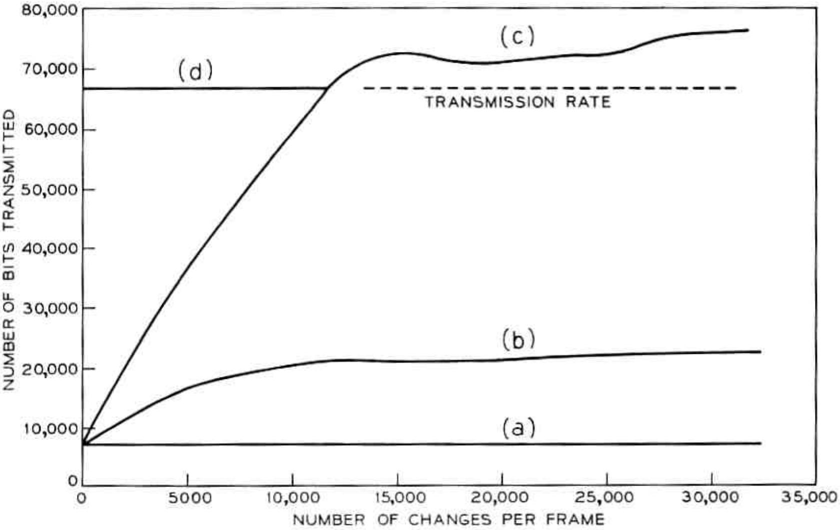


Fig. 20—The average number of bits transmitted plotted against the number of significant changes. The probability of the changes occurring is given in Fig. 2. The ordinate of (a) represents the fixed transmission. The separation (a)-(b) represents the addresses, c.f. Fig. 11; (b)-(c) represents the amplitude difference; and (c)-(d) represents the forced transmission.

for the various functions of the coder at different levels of activity in the scene. The ordinate of the straight line (a) represents the fixed transmission. The distance between curves (a) and (b) represents the bits used for addressing clusters of changed elements, and the distance between (b) and (c) represents the bits used for signaling their changes of amplitude. The change in slope of (c) near 12 kilochanges per frame and the subsequent lower rate of rise of the curve is caused by the subsampling mechanism. The distance between curves (c) and (d) represents the bits that are generated by forcibly updating lines of the picture in order to prevent the buffer from emptying. On this graph we see that the data generated per frame is less than the transmission rate, provided no more than 12,000 elements change in a frame time; i.e., 90 percent of the time. When more than 12,000 elements change, the extra bits accumulate in the buffer. Only when more than 20,000 elements change for several consecutive frames, i.e., for very violent motion covering a large area of the picture, is the overload mode needed.

XII. CONCLUSION

Taking 8-bit PCM as the standard of coding quality, we have described a method of improving transmission efficiency by a factor of nearly eight, permitting pictures to be transmitted at an average rate of one bit per element. Most viewers agree that there is little visible difference between the processed picture and the original. The quality is certainly better than has been obtained using 16-level differential coding.

The improvement in efficiency is largely obtained by avoiding the use of transmission capacity for elements that do not change from frame to frame. Additional saving is obtained by taking advantage of the fact that reduced resolution is tolerable in changing scenes. Thus, frame-to-frame differences are quantized in amplitude, and sampled at half of the Nyquist rate. In this way the number of bits generated in every field is kept approximately equal to that which can be transmitted in a field time. A buffer is used to smooth the data flow over a field-time, its fullness serving as an indication of the rate at which the data is being generated. This measure is used for controlling the behavior of the coder.

Important subjects which have not been described here are the possibilities of using line-to-line correlation in the coder, the advantages of mixing the data from several coders in order to obtain a more even data flow, and the effect of having an input signal that has already been modulated or coded in a prior transmission.

XIII. ACKNOWLEDGMENT

The authors thank C. C. Cutler and L. H. Enloe for direction and encouragement, D. J. Connor, J. O. Limb, and R. F. W. Pease for their cooperation, R. L. Eilenberger for providing the camera system, W. G. Scholes for the frame memories, and W. Kaminski for the master clock generator.

REFERENCES

1. Maunsell, H. I., and Millard, J. B., "Digital Coding of the Video Signal," B.S.T.J., *50,* No. 2 (February 1971), pp. 459–479.
2. Limb, J. O., and Mounts, F. W., "Digital Differential Quantizer for Television,'" B.S.T.J., *48,* No. 7 (September 1969), pp. 2583–2599.
3. Brown, E. F., "A Sliding-Scale Direct-Feedback PCM Coder for Television," B.S.T.J., *48,* No. 5 (May-June 1969), pp. 1537–1553.
4. Bosworth, R. H., and Candy, J. C., "A Companded One-Bit Coder for Television Transmission," B.S.T.J., *48,* No. 5 (May-June 1969), pp. 1459–1479.

5. Connor, D. J., Pease, R. F. W., and Scholes, W. G., "Television Coding Using Two-Dimensional Spatial Prediction," B.S.T.J., *50,* No. 3 (March 1971), pp. 1049–1061.
6. Mounts, F. W., "A Video Encoding System With Conditional Picture-Element Replenishment," B.S.T.J., 48, No. 7 (September 1969), pp. 2545–2554.
7. Pease, R. F. W., and Limb, J. O., "Exchange of Spatial and Temporal Resolution in Television Coding," B.S.T.J., *50,* No. 1 (January 1971), pp. 191–200.
8. Limb, J. O., and Pease, R. F. W., "A Simple Interframe Coder for Video Telephony," B.S.T.J., this issue, pp. 1877–1888.
9. Teer, K., "Some Investigations on Redundancy and Possible Bandwidth Compression in Television Transmission," Thesis Technische Hogeschool Delft, September 1959.
10. Seyler, A. J., "Probability Distribution of TV Frame Differences," SMIREE Australia Proc., November 1965, p. 355.
11. Pease, R. F. W., "Frame Difference Signals for a Personal Communication Television System of 271 Lines," Private Communication.
12. Brainard, R. C., Cutler, C. C., and Rowlinson, D. E., "*Picturephone* Communication with Delay," Private Communication.
13. Limb, J. O., and Haskell, B. G., "Buffering of Data Generated by the Coding of Moving Images," Private Communication.
14. Mounts, F. W., and Pearson, D. E., "Apparent Increase in Noise Level When Television Pictures are Frame-Repeated," B.S.T.J., *48,* No. 3 (March 1969), pp. 527–539.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Low-Loss Microstrip Filters Developed by Frequency Scaling

By W. W. SNELL, JR.

(Manuscript received January 21, 1971)

Low-pass and band-pass microstrip filters up to 30 GHz have been built on silica substrates by photolithographic reduction of optimized low-frequency models. Scaling accuracies of 1.8 percent or better have been achieved. An insertion loss of 0.28 dB has been measured for a five-section low-pass filter with a cutoff frequency of 5.7 GHz. The insertion loss of a 12-percent band-pass filter at 8.18 GHz is 0.16 dB. Good agreement between measured and calculated losses has been obtained. Circuit dimensions are given to facilitate scaling of the low-pass and band-pass microstrip filters to other frequencies.

I. INTRODUCTION

Microstrip filters are used in frequency converters, frequency multipliers, and branching networks for RF systems. Microstrip filters have been built in the past with photo-etched conductor patterns on ceramic substrates such as alumina, sapphire, and beryllia.[1,2] These materials are useful in microwave integrated circuits at a few gigahertz where a desirable size reduction relative to free-space wavelength is achieved because of the high dielectric constant of the substrate. At higher frequencies, size reduction is a disadvantage because the circuit elements are approaching the limitations of photolithographic reproduction. A suitable substrate material for use at higher frequencies is clear fused silica. Its relative dielectric constant over a wide frequency range is 3.82 and its dielectric Q is 3,000 or higher at all frequencies up to 50 GHz.

The purpose of this paper is to show that optimized microstrip filters at microwave and millimeter-wave frequencies can be built by linear reduction of a low-frequency model. The flow chart of this procedure is shown in Fig. 1. The oversize microstrip filter is designed, built, and tested at a few hundred megahertz. All circuit dimensions

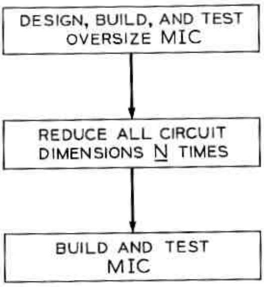

DESIGN, BUILD, AND TEST OVERSIZE MIC
REDUCE ALL CIRCUIT DIMENSIONS N TIMES
BUILD AND TEST MIC

Fig. 1—MIC flow chart (general).

are then reduced N times in order to obtain the optimized microstrip filter at microwave or millimeter-wave frequencies.

The initial design and testing of the microstrip filters described in this paper are made at 285 MHz for the low-pass filter and 818 MHz for the band-pass filter. The filters are scaled by factors of 20 and 10 times, respectively, to produce 5.7-GHz and 8.18-GHz circuits. A scaling factor of 37.5 times is also used to obtain low-pass input and band-pass output filters for a 10.1-GHz to 30.3-GHz frequency multiplier which is described in another paper.[3]

II. MICROSTRIP SUBSTRATE

Fused silica, SiO_2, is selected as the microstrip substrate for its desirable properties. These properties are as follows:

(*i*) The dielectric constant and loss tangent are low.

(*ii*) The dielectric constant is independent of frequency up to 50 GHz.

(*iii*) The surface can be highly polished.

(*iv*) It is commercially available in a wide range of sizes and thicknesses.

(*v*) It has a low thermal expansion coefficient.

Table I lists the electrical and physical properties of fused silica and other common dielectric materials. Soda-lime glass, although not normally used as a microstrip substrate, is included as a low-cost material for price comparison. Detailed data on dielectric constant, loss tangent, and volume resistivity of high-purity synthetic fused silica are given in the appendix. Since silica has a low thermal conductivity of 1.4×10^{-2} watts/cm°K it is recommended for low- or medium-power applications.

III. SCALING OF MICROSTRIP CIRCUITS

The laws governing scaling of electromagnetic circuits are described by Stratton.[4] In order to obtain circuits with similar characteristics at different frequencies, each with a characteristic length l, an RF frequency ω, a conductivity σ, a permeability μ, and a permittivity ϵ, it is necessary and sufficient that:

$$\mu\epsilon(\omega l)^2 = \mathrm{K}_1\ , \tag{1}$$

$$\mu\epsilon\sigma\omega l^2 = \mathrm{K}_2\ , \tag{2}$$

where K_1 and K_2 are constants. Scaling has not been widely used for building microstrip circuits because high dielectric constant materials are expensive and difficult to obtain in large sizes and also because air gaps between these materials and metal conductors have a significant effect on the circuit performance and are difficult to scale. These problems are substantially reduced by using a low dielectric constant material such as silica.

A detailed flow chart of the scaling process is shown in Fig. 2. All circuit dimensions are reduced by the same factor by which the frequency is increased, thus satisfying equation (1). Equation (2) is only partially satisfied because brass conductors have been used in the over-

TABLE I—PROPERTIES OF MIC SUBSTRATES

Property	Units	Silica SiO_2	Alumina Al_2O_3 (99.5%)	Beryllia BeO(99.5%)	Soda-Lime Glass Corning #0080
Dielectric Constant at 10 GHz, 25°C		3.82	9.8	6.8	6.71
Loss Tangent 10 GHz	$\times 10^{-4}$	1	4	3	170
Thermal Expansion	PPM/°C	0.35	6.6	7.5	9.2
Relative Cost Factor*		6	15	125	1
Surface Condition		Polished	Polished	Polished	Drawn

* Relative cost factor of substrate material compared to drawn soda-lime glass = 1 (Data from *Handbook of Thin Film Technology*, edited by L. I. Maissel and R. Glang, McGraw-Hill, 1970, Section 6–43).

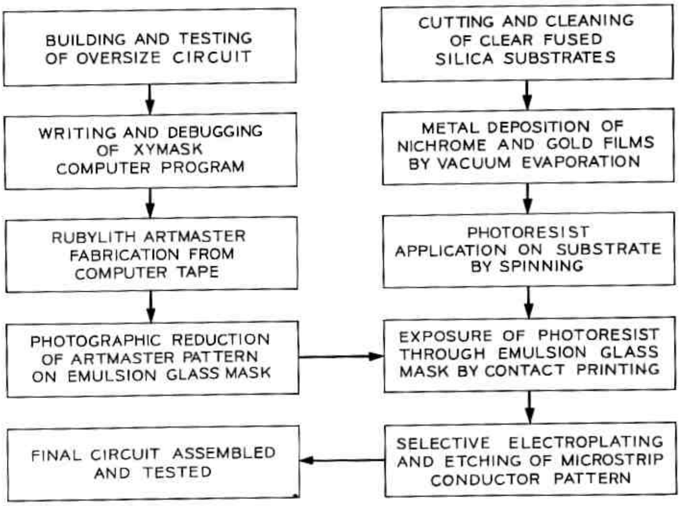

BUILDING AND TESTING OF OVERSIZE CIRCUIT
WRITING AND DEBUGGING OF XYMASK COMPUTER PROGRAM
RUBYLITH ARTMASTER FABRICATION FROM COMPUTER TAPE
PHOTOGRAPHIC REDUCTION OF ARTMASTER PATTERN ON EMULSION GLASS MASK
FINAL CIRCUIT ASSEMBLED AND TESTED
CUTTING AND CLEANING OF CLEAR FUSED SILICA SUBSTRATES
METAL DEPOSITION OF NICHROME AND GOLD FILMS BY VACUUM EVAPORATION
PHOTORESIST APPLICATION ON SUBSTRATE BY SPINNING
EXPOSURE OF PHOTORESIST THROUGH EMULSION GLASS MASK BY CONTACT PRINTING
SELECTIVE ELECTROPLATING AND ETCHING OF MICROSTRIP CONDUCTOR PATTERN

Fig. 2—MIC flow chart (detailed).

size model and gold conductors are used in the reduced size circuits. This leads to a slight increase in the scaled circuit losses because the skin effect loss increases with the square root of the frequency.

IV. LOW-PASS FILTER

The low-pass filter is a five-element network of three shunt capacitors, C_1, C_3, and C_5, and two series inductors, L_2 and L_4, designed for a 0.2-dB Chebyshev frequency response. The microstrip layout of the filter is shown in Fig. 3. The total filter length is slightly less than $\lambda_o/5$; λ_o is the free-space wavelength at the 0.2-dB cutoff frequency. The length of each element is less than one-eighth of a microstrip wavelength, which allows the initial design of the filter to be based on a lumped circuit concept. The filter parameters C_n and L_n are computed from normalized prototype parameters g_n, the input impedance $R = 50\Omega$, and the cutoff frequency $\omega_c = 2\pi f_c$:

$$C_n = \frac{1}{\omega_c R} g_n , \tag{3}$$

$$L_n = \frac{R}{\omega_c} g_n . \tag{4}$$

Table II lists the parameters of the filter. The 0.2-dB cutoff frequency is 285 MHz. The thickness of the dielectric substrate is 0.5

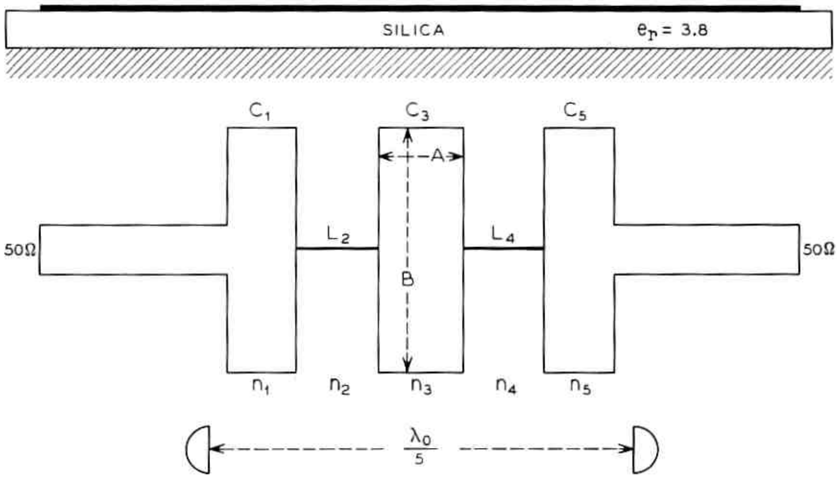


Fig. 3—Low-pass filter (circuit).

inch and its relative dielectric constant is 3.82. The computed filter area is based on the choice of a specific cutoff frequency and does not take into account distributed line effects at higher frequencies. The experimental filter has been modified in order to optimize the performance up to $4\omega_c$.

The computed dimensions of the capacitive plates shown in Fig. 3 are found as follows:

TABLE II—LOW-PASS FILTER PARAMETERS

Design Parameters				Computed Filter Dimensions		Experimental Filter Dimensions		
n	g_n	C_n pF	L_n nH	Aspect Ratio a/b	Capacitor Area A in.^{2} or Inductance Length l_n in.	a in.	b in.	Area in.^{2}
1	1.339	15		3.52	$A_1 = 5.67$	1.42	5.0	7.1
2	1.337		37.4		$l_2 = 1.82$	0.70	1.62	
3	2.166	24.2		2.91	$A_3 = 10.23$	1.72	5.0	8.6
4	1.337		37.4		$l_4 = 1.82$	0.70	1.62	
5	1.339	15		3.52	$A_5 = 5.67$	1.42	5.0	7.1

(*i*) An effective area A_{eff} is defined as the area of a capacitor plate which has no fringing fields. This area is given by

$$A_{eff} = \frac{C_n t}{\epsilon_o \epsilon_r}, \tag{5}$$

where t = dielectric substrate thickness =0.5 inch,
ϵ_r = relative dielectric constant, and
ϵ_o = 8.85 pF/m.

(*ii*) The actual area A of the capacitor plate is smaller because of fringe field effects. If a and b are the width and length of the capacitor plate shown in Fig. 3, then A_{eff} is given by

$$A_{eff} = (a + \alpha t)(b + \alpha t), \tag{6}$$

where α is a normalized factor which describes the effect of the fringe field.

(*iii*) The aspect ratio of the capacitor plate shown in Fig. 3 is defined as $r = a/b$ and the factor β by

$$\beta = \sqrt{r} + \frac{1}{\sqrt{r}}. \tag{7}$$

(*iv*) The area A is obtained from equations (5), (6), and (7):

$$A = \left\{\left[\frac{C_n t}{\epsilon_o \epsilon_r} + \frac{(\beta^2 - 4)\alpha^2 t^2}{4}\right]^{\frac{1}{2}} - \frac{\alpha \beta t}{2}\right\}^2. \tag{8}$$

One finds by empirical methods that the effective increase in each plate dimension due to the fringe field is approximately equal to the substrate thickness t; thus, $\alpha = 1$ and equation (8) is simplified to

$$A = \frac{t^2}{4}\left\{\left[\frac{4C_n}{\epsilon_o \epsilon_r t} + \beta^2 - 4\right]^{\frac{1}{2}} - \beta\right\}^2. \tag{9}$$

V. MICROSTRIP INDUCTORS

The inductance L per unit length for a narrow microstrip conductor with a width w and a substrate thickness h is

$$L = \frac{60 \ln\left(\frac{8h}{w} + \frac{w}{4h}\right)}{v_o}, \tag{10}$$

where v_o is the velocity of light in free space, $v_o = 3 \times 10^{10}$ cm/s = 1.18×10^{10} ips. The length l_n of the inductor element is given by the ratio L_n/L where L_n is the required inductance given in Table II,

$$l_n = \frac{L_n}{L} = \frac{v_o R}{60\omega_c \ln\left(\dfrac{8h}{w} + \dfrac{w}{4h}\right)} \cdot g_n \,. \tag{11}$$

The computed length for the inductive elements is 1.82 inches. The actual length of the final inductive elements in the filter is 1.61 inches because the current constrictions in the capacitor plates increase the effective length of each element.

Calculated and measured transmission loss for a low-pass filter with a cutoff frequency of 285 MHz is shown in Fig. 4. The transmission is plotted as a function of normalized frequency ω/ω_c. Figure 5 shows the frequency response after scaling the filter by a factor of 20X. The new cutoff frequency is 5.6 GHz and the scaling accuracy is 1.8 percent.

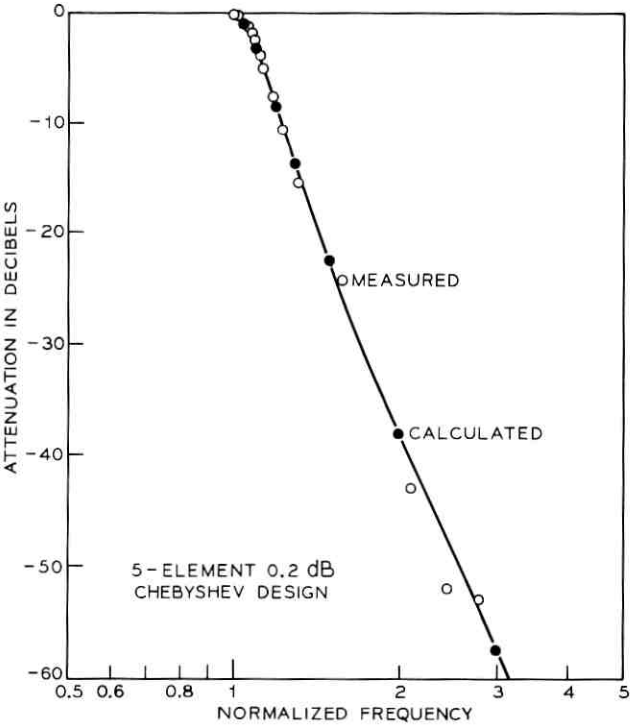


Fig. 4—Low-pass filter (comparison of calculated and measured data points).

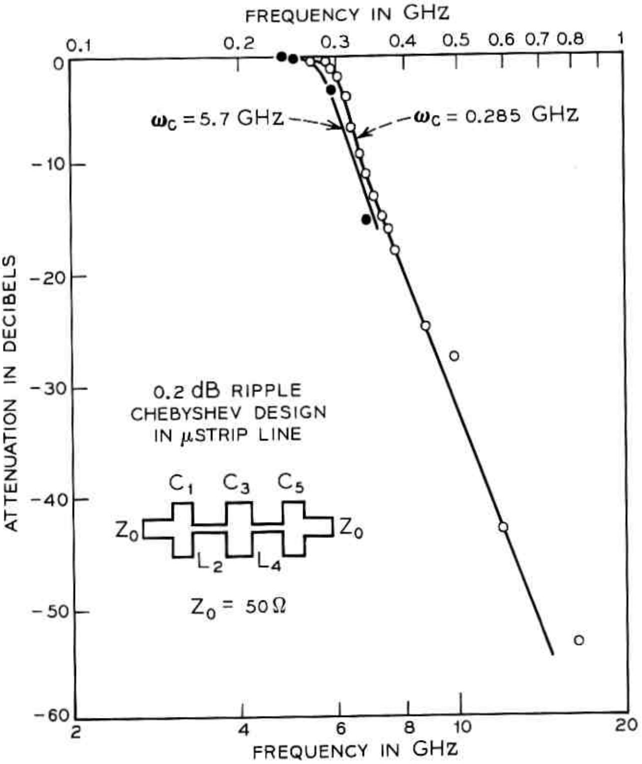


Fig. 5—5-element low-pass filter (results of scaling from 285 MHz to 5.7 GHz).

VI. BAND-PASS MICROSTRIP FILTER

Band-pass filters utilizing two parallel coupled $\lambda/2$ resonators have been designed, built, and optimized at 820 MHz and scaled to 8.2 GHz and 30 GHz. Figure 6 shows the microstrip lay-out of the conductor pattern on the silica substrate. The filter has a 0.2-dB Chebyshev ripple and 20 dB attenuation at the ±33-percent bandwidth points. The calculated and measured filter responses are shown in Fig. 7 and the filter characteristics, after scaling by a factor of 10X, are plotted in Fig. 8. The scaling accuracy for the midband frequency is better than 1 percent. The 7-percent decrease in bandwidth is due to undercutting of the conductor elements during processing of the microstrip pattern. Undercutting is etching of the conductor pattern beneath the edge of the protective photoresist layer. It decreases

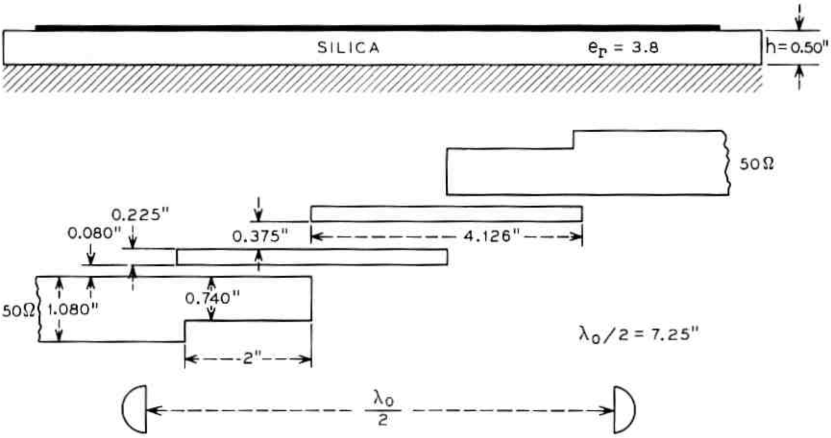


Fig. 6—Band-pass filter (circuit).

all conductor dimensions by a small distance. The effect of undercutting is most noticed on small critical circuit dimensions such as the 0.008-inch-wide coupling gap of the $\lambda/2$ resonator. Undercutting can be controlled by establishing the undercut dimension and adding this to all circuit elements on the original artmaster.

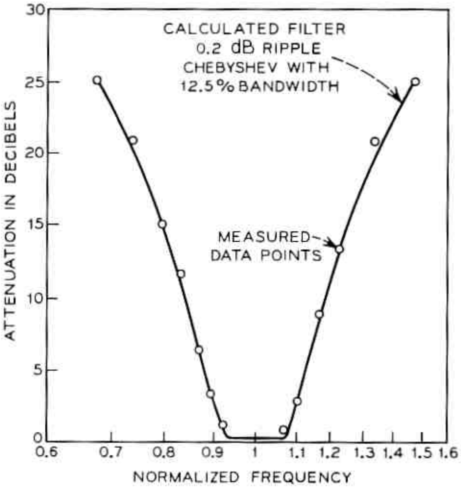


Fig. 7—Experimental and theoretical attenuation of parallel coupled band-pass filter.

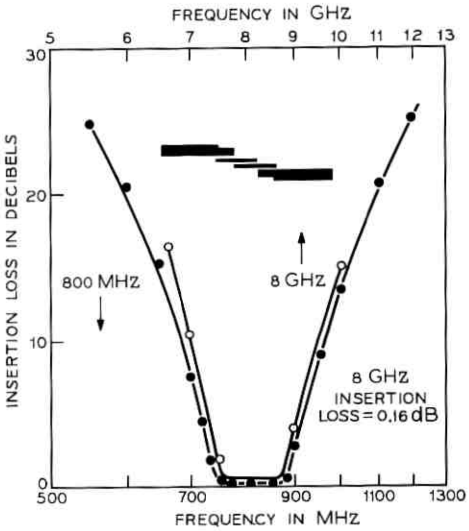


Fig. 8—Microstrip band-pass filter (results of scaling from 800 MHz to 8 GHz).

VII. ATTENUATION OF MICROSTRIP FILTERS

The dissipative loss in the pass band of a multiple-resonator filter is[5]

$$L_o = \frac{10}{w' \ln 10} \sum_{i=1}^{n} \frac{g_i}{Q_i} \, \text{dB}, \tag{12}$$

where w' is the fractional bandwidth of the filter, g_i is the normalized prototype filter parameter, and Q_i is the unloaded Q of each filter element.

The unloaded Q_i of a microstrip element is

$$Q_i = \frac{20\pi}{\ln 10} \cdot \frac{1}{\alpha_o \lambda_o}, \tag{13}$$

where λ_o is the free-space wavelength, and α_o is the free-space attenuation in dB per unit length as derived by M. V. Schneider,[6] where

$$\alpha_o = \frac{10R_s}{\pi h \ln 10} \, \frac{\left(\frac{8h}{w} - \frac{w}{4h}\right)\left(1 + \frac{h}{w} + \frac{h}{w}\frac{\partial w}{\partial t}\right)}{Z_o \exp\left(\frac{Z_o}{60}\right)} \qquad w/h \leqq 1 \tag{14}$$

or

$$\alpha_o = \frac{Z_o R_s}{720\pi^2 h \ln 10}\left[1 + \frac{0.44h^2}{w^2} + \frac{6h^2}{w^2}\left(1 - \frac{h}{w}\right)^5\right]\left(1 + \frac{w}{h} + \frac{\partial w}{\partial t}\right) \qquad w/h \leqq 1. \tag{15}$$

Z_o is the characteristic impedance of the microstrip line:

$$Z_o = 60 \ln\left(\frac{8h}{w} + \frac{w}{4h}\right) \qquad w/h \leqq 1 \tag{16}$$

$$Z_o = \frac{240\pi}{\dfrac{2w}{r} + 5 - \dfrac{h}{w}} \qquad w/h \geqq 1. \tag{17}$$

R_s is the skin resistance of the metal conductor as shown in Fig. 9, h is the conductor height, and w is the conductor width. The partial derivative $\partial w/\partial t$ is given by:

$$\frac{\partial w}{\partial t} = \frac{1}{\pi} \ln \frac{4\pi w}{t} \qquad w/h \leqq \frac{1}{2\pi} \tag{18}$$

$$\frac{\partial w}{\partial t} = \frac{1}{\pi} \ln \frac{2h}{t} \qquad w/h \geqq \frac{1}{2\pi} \tag{19}$$

where t is the conductor thickness. The calculated and measured dissipative losses for the low-pass filter and for the band-pass filter are shown in Table III.

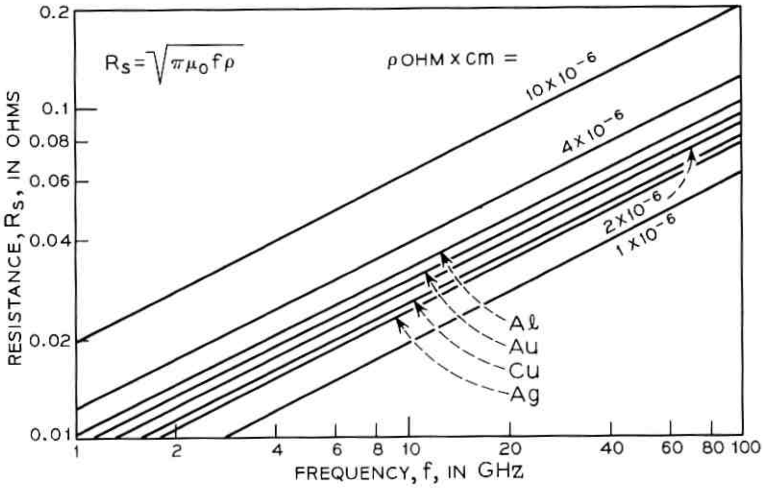


Fig. 9—Skin resistance of metals.

TABLE III—INSERTION LOSS MEASUREMENTS

Filter	Insertion Loss in dB	
	Calculated	Measured
Loss-Pass $f_c = 5.7$ GHz 5 sections	0.13	0.28
Band-Pass $f_0 = 8.2$ GHz $w' = 12\%$ 2 sections	0.11	0.16

Unloaded Q of Single Band-Pass Resonator	
Calculated Q	546
Measured Q	375

The calculated loss and the calculated Q factor are based on the assumption that the conductor resistivity is independent of frequency and equal to the dc resistivity of the metal.

VIII. CONCLUSION

It is shown that scaling is a powerful tool for building and optimizing microwave integrated circuits. A five-element low-pass filter and a two-element band-pass filter have been built and measured in the 200- to 800-MHz frequency band. The low-pass filter has been scaled to 5.7 GHz and 10 GHz, and the band-pass filter has been scaled to 8.2 GHz and 30 GHz. Good agreement between measured and calculated losses has been obtained. Unloaded Q factors of 375 at 8.2 GHz are reported.

IX. ACKNOWLEDGMENTS

The author wishes to thank M. V. Schneider for his helpful suggestions throughout this project, and to thank S. R. Shah for building the accurate, scaled microcircuit filters on fused silica.

APPENDIX

Properties of High-Purity Synthetic Fused Silica

Table IV gives detailed data on the dielectric constant, loss tangent, and volume resistivity of high-purity synthetic fused silica produced by chemical vapor deposition (Dynasil No. 4000).

Table IV—MIT Measurements of Dynasil No. 4000 (July 1970)

Dielectric Constant	100 Hz	10 kHz	1 MHz	10 MHz	8.60 GHz	24.0 GHz
25°C	3.826	3.826	3.826	3.826	3.824	3.82
100°C	3.834	3.834	3.834	3.834	3.832	3.83
200°C	3.843	3.843	3.843	3.843	3.841	3.84
300°C	3.861	3.856	3.854	3.854	3.852	3.85
400°C	4.043	3.870	3.865	3.865	3.863	3.86
500°C	4.795	3.886	3.878	3.878	3.876	3.87
Loss Tangent						
25°C	<.000004	<.000005	.000015	.00002	.00012	.00033
100°C	<.000004	<.00002	<.00005	<.0001	.00012	.00032
200°C	.00137	.00003	<.00005	<.0001	.00012	.00026
300°C	.0930	.00093	<.00005	<.0001	.00012	.00025
400°C	2.03	.0207	.00021	<.0001	.00012	.00022
500°C	11.36	.140	.00140	.00014	.00012	.00020
Volume Resistivity	$\log_{10} \rho$ ohm-cm					
25°C	>15.0	>12.9	10.5	9.34	5.65	
100°C	>15.0	>12.4	>9.95	8.65	5.65	
200°C	12.5	12.2	>9.95	>8.65	5.65	
300°C	10.9	10.9	>9.95	>8.65	5.65	
400°C	9.34	9.34	9.34	>8.65	5.65	
500°C	8.52	8.52	8.52	8.52	5.65	

Source: A. R. von Hippel and W. Westphal, Laboratory of Insulation Research, MIT.

REFERENCES

1. Chinchillo, A. R., and Perry, R. W., "Microstrip Filter Loss Reduction Techniques," NEREM 70 Record, *12* (November 1970), pp. 72–73.
2. Dell-Imagine, R. A., "A Parallel Coupled Microstrip Filter Design Procedure," G-MTT 1970 Int. Microwave Symp. Digest, May 1970, pp. 29–32.
3. Schneider, M. V., and Snell, W. W., "A Scaled Hybrid Integrated Multiplier from 10 to 30 GHz," B.S.T.J., this issue, pp. 1933–1942.
4. Stratton, J. A., *Electromagnetic Theory,* New York: McGraw-Hill, 1941, pp. 488–490.
5. Cohn, S. B., "Dissipation Loss in Multiple-Coupled-Resonator Filters," Proc. IEEE, *47,* No. 8 (August 1959), pp. 1342–1348.
6. Schneider, M. V., "Microstrip Lines for Microwave Integrated Circuits," B.S.T.J., *48,* No. 5 (May–June 1969), pp. 1421–1444.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

A Scaled Hybrid Integrated Multiplier from 10 to 30 GHz

By M. V. SCHNEIDER and W. W. SNELL, JR.

(Manuscript received February 18, 1971)

Frequency multipliers which are built on a dielectric substrate have many useful applications in radio systems. The multiplier circuits which include input and output filters, idler circuits, and matching networks can be produced by integrated circuit processing techniques. Varactor diodes, also made by batch processing, can be bonded to the metallized substrate.

A hybrid integrated three-times frequency multiplier from 10 to 30 GHz has been designed and built on a silica substrate by using frequency scaling and integrated circuit processing techniques. The multiplier produces a CW output power of 50 mW for a pump power of 178 mW with an overall efficiency of 29 percent. The maximum output power is 150 mW. Other harmonic generators with different output frequencies can be built by scaling the existing circuit.

I. INTRODUCTION

Frequency multipliers are used in radio systems for downconverters, for upconverters, or for locking solid state oscillators. A typical example is the multiplier chain used in the short hop radio system experiment[1] which delivers 100 mW at 10.46 GHz for the transmitter and 10 mW at 10.66 GHz for the receiver. Similar systems above 10 GHz for common carrier applications will require a few tens of milliwatts of transmitter power[2] and approximately 10 mW for the receiving downconverter. A compact and low-cost repeater package can be manufactured by using integrated circuit technology for building components and subassemblies. This can be achieved by using low-loss dielectric substrates, fully shielded circuits, and a circuit design which does not require high Q factors.

All three basic requirements are fulfilled for the frequency multiplier described in this paper. The dielectric substrate is silica, the circuit is fully shielded, and low-Q filters are used for separating the pump frequency and the output frequency. The pump frequency is compatible with the output of existing high-efficiency multipliers used in the short hop system experiment.[1] The substrate metallization as well as the planar diffused GaAs varactor diodes are fabricated by using integrated circuit processing steps. Optimum dimensions of all conductor patterns are obtained by building and testing an oversize model of the complete multiplier at a lower frequency, that is, with a pump frequency of 270 MHz and an output frequency of 810 MHz. Circuit adjustments are easily made in the oversize multiplier and no adjustments are necessary after the oversize model is reduced to its final size. Good stability and spurious-free operation is obtained for the oversize multiplier with an 80 percent efficiency and a power output of 2.5 W at 818 MHz.

II. DESIGN AND FABRICATION OF INTEGRATED MULTIPLIERS

Frequency multipliers have been built in the past with coaxial and waveguide circuits by using mesa type,[3] or planar diffused, Si or GaAs[4] varactor diodes with the highest possible figure of merit.[5] The basic design principles for building multipliers are well known[5–7] and overall efficiencies close to the theoretically expected values have been achieved up to X-band frequencies. Some effort has been made in building high-performance hybrid integrated multipliers up to X-band, but the design of good filters on commonly used alumina substrates is difficult at millimeter-wave frequencies and also external tuning elements are usually needed to achieve good overall efficiency. This difficulty can be resolved by using a high-quality dielectric substrate with a relatively low dielectric constant and a simple conductor pattern which can be clearly separated into a low-pass input filter, an idler circuit, and a band-pass filter. Before scaling, the filter characteristics and the performance of the complete multiplier are tested in an oversize model. The model is assembled with clear silica plates with a relative dielectric constant $\varepsilon_r = 3.8$. Small air gaps in the oversize model between the ground plane and the dielectric substrate and air gaps between the conductor pattern and the substrate are not critical because of the low dielectric constant of silica. The junction capacitance and the dimensions of the varactor diode are also scaled by the same reduction factor which is used for scaling the microstrip circuit. The parameters which are more difficult to scale are the cutoff frequency of the diode and the skin depth of the microstrip conductors. This problem has been reduced by using a low-cutoff silicon diode for the low-frequency model and a high-cutoff gallium arsenide diode in the final circuit at 30 GHz. The skin depth is partially scaled by using brass conductors for the low-frequency model and gold conductors for the final circuit.

2.1 *Multiplier Conductor Pattern*

The complete conductor pattern of the multiplier deposited on a silica substrate is shown in Fig. 1. The pattern consists of three sections, a low-pass input filter, an idler section, and a band-pass output filter. The input and the output is a microstrip line with a characteristic impedance of 50 ohms. The substrate is inserted into a channel in a brass block, shown in Fig. 2, with a coaxial-to-microstrip transition at the input and a microstrip-to-waveguide transition at the output. A planar diffused GaAs varactor diode is mounted between the substrate metallization and a metal tab which is in contact with one sidewall of the channel shown in Fig. 2. The multiplier is completely shielded by a metal plate which covers the brass channel.

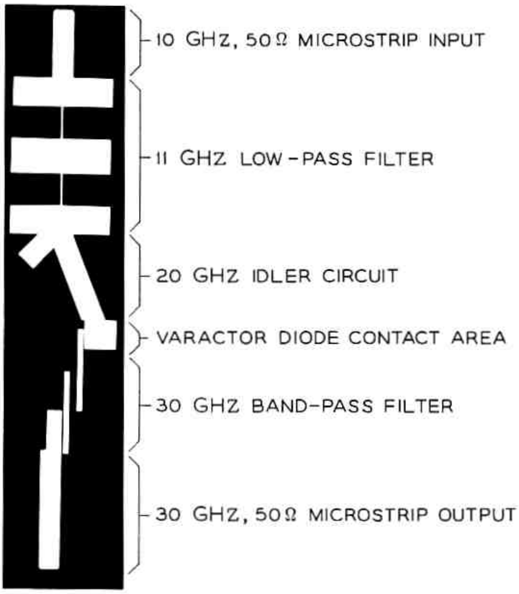

10 GHz, 50Ω MICROSTRIP INPUT
11 GHZ LOW-PASS FILTER
20 GHZ IDLER CIRCUIT
VARACTOR DIODE CONTACT AREA
30 GHZ BAND-PASS FILTER
30 GHZ, 50Ω MICROSTRIP OUTPUT

Fig. 1—Conductor pattern of hybrid integrated multiplier on silica substrate.

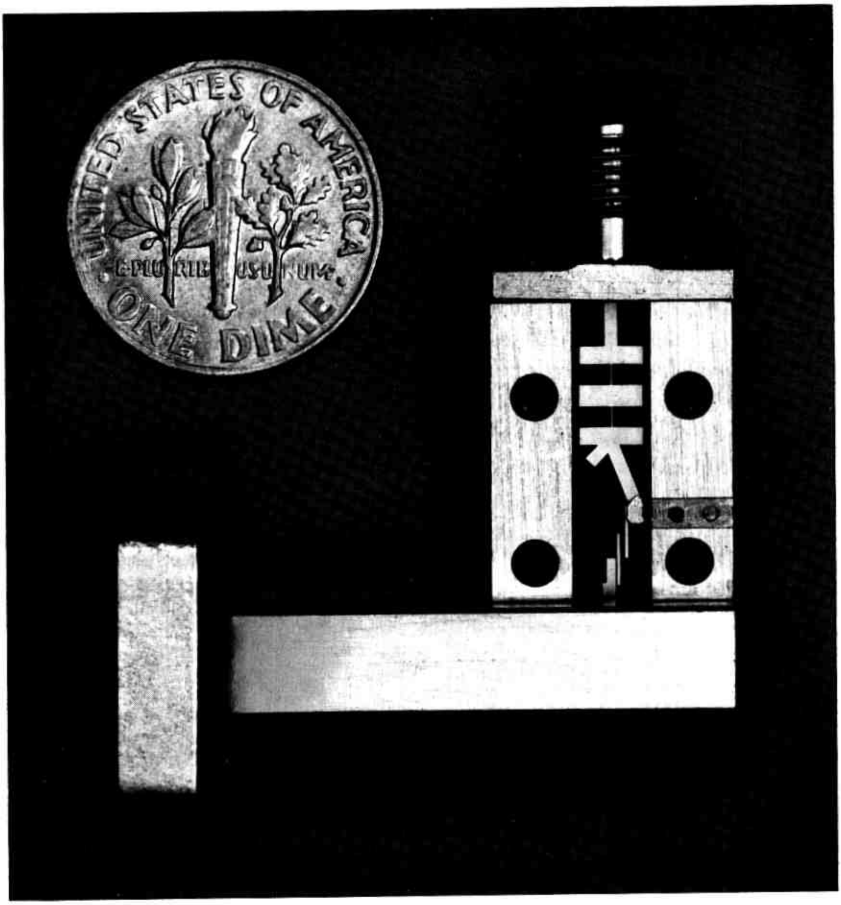


Fig. 2—Hybrid integrated multiplier from 10 GHz to 30 GHz with transitions from coaxial transmission line to microstrip, and microstrip to RG-96/U waveguide.

2.2 *Low-Pass and Band-Pass Filter Design*

The low-pass filter used in the multiplier is a five element semi-lumped structure consisting of three capacitive and two inductive sections. The element values for a Chebyshev filter with a 0.2-dB ripple are computed from tables given by G. L. Matthaei.[8] The cutoff frequency is 300 MHz for the oversize model and 11.2 GHz for the scaled filter. The attenuation of the filter at the second harmonic of the pump frequency, which is the idler frequency of the multiplier, is 35 dB.

The band-pass filter consists of two parallel-coupled microstrip resonators. The 3-dB bandwidth of the oversize filter with a center frequency of 810 MHz is 160 MHz. The 3-dB bandwidth of the reduced filter with a center frequency of 30.4 GHz is 6.0 GHz. The attenuation of the filter at the second harmonic of the pump frequency is 25 dB. The scaling of microstrip circuits, the design of the filters, and the measured characteristics of low-pass and band-pass microstrip filters on silica are described in more detail in a separate paper.[9]

The input and output impedance for each filter is 50 ohms. An additional matching network or a modification of each filter is necessary to match the input impedance Z_{in} and the output impedance Z_{out} of the multiplier. Z_{in} and Z_{out} are functions of the pump frequency ω_o, the nonlinearity coefficient γ, the breakdown capacitance C_B, and the drive D of the varactor diode. For a tripler with $f_o = \omega_o/2\pi = 10$ GHz, a nonlinearity coefficient $\gamma = 0.33$, a breakdown capacitance $C_B = 0.090$ pF, and a drive $D = 1.6$, one obtains

$$Z_{in} = \frac{F_1(\gamma, D)}{\omega_o C_B} = \frac{0.214}{\omega_o C_B} = 38\ \Omega, \tag{1}$$

$$Z_{out} = \frac{F_2(\gamma, D)}{\omega_o C_B} = \frac{0.087}{\omega_o C_B} = 15.5\ \Omega, \tag{2}$$

where the values of the functions $F_1(\gamma, D)$ and $F_2(\gamma, D)$ are obtained from tables given by C. B. Burckhardt.[6] The transformation from 38 ohms to 50 ohms at the input is achieved by a quarter-wave microstrip section with an impedance of $Z = \sqrt{50Z_{in}} = 43.5\ \Omega$. This section is also part of the idler circuit shown in Fig. 1. Its electrical length at the idler frequency is slightly longer than half a wavelength and its inductive reactance seen across the varactor terminal resonates with the average diode capacitance at the idler frequency. A short additional line section is required to tune the reactive part of the diode package.

The transformation from 15.5 ohms to 50 ohms is obtained by adjusting the coupling between the contact area of the varactor diode and the adjacent half-wavelength microstrip resonator. The spacing between the conductor edges is optimized in the oversize model of the multiplier and no adjustment is necessary after the model is reduced to its final size.

2.3 *Diode Fabrication and Packaging*

The planar diode used for the integrated multiplier is a zinc-diffused junction in n-type epitaxial gallium arsenide with a doping level of

2×10^{17} carriers/cc and an epitaxial layer thickness of 1.2 μm. The processing is done by the photoresist, etching, and zinc-diffusion process developed by C. A. Burrus[10] for fabricating planar millimeter-wave varactor diodes with zero-bias cutoff frequencies above 500 GHz for breakdown voltages between 10 and 20 volts. Contact to the gold-plated junction with a diameter of 20 μm is made with a gold wire with a diameter of 75 μm and a length of 150 μm. The gold wire is embedded in epoxy and the diode is mounted in the circuit between the gold-plated contact area on the silica substrate and a spring-loaded ground contact made with a gold-plated phosphor bronze tab which is shown in Fig. 2. The diodes can be easily replaced and the mount has the smallest possible series inductance which can be achieved without making a hole in the silica substrate.

III. PERFORMANCE OF HYBRID INTEGRATED MULTIPLIER

The microstrip and diode properties used for building the hybrid integrated multiplier are listed in Table I. The output power of this multiplier is 50 mW at 30.3 GHz for a pump power of 178 mW at 10.1 GHz. The available output power at burnout is 150 mW. A plot of the output power as a function of pump power is shown in Fig. 3 and specific performance data are listed in Table II.

The output is measured in RG-96/U rectangular waveguide by using a microstrip-to-waveguide transition which is similar to the side launcher developed by K. H. Knerr.[11] The only major difference is that the Knerr launcher is used for balanced microstrip line while the pres-

TABLE I—SUBSTRATE AND DIODE PROPERTIES

Substrate Material	Silica
Dielectric Constant	$\epsilon_r = 3.828 \pm 0.003$ at 8.5 GHz
Loss Tangent	$10^4 \cdot \tan \delta = 1.29 \pm 0.2$ at 8.5 GHz
Substrate Thickness	0.34 mm
Substrate Dimensions	4.06 mm × 20.4 mm
Substrate Metallization	100 Å Nichrome and 1000 Å Gold evaporated, 3 μm Gold electroplated
Weight of Gold	4.7 milligrams (0.57 cent)
Varactor Diode	Planar zinc-diffused junction into epitaxial GaAs N/N^+
Epitaxial Layer Thickness	1.2 μm
Doping Level	1.9×10^{17} carriers/cc
Mobility and Resistivity	4290 cm^2/Vsec, 0.0076 Ωcm
Properties of N^+ Material	Orientation $\langle 100 \rangle$, Te doped, 1×10^{-3} Ωcm
Junction Diameter	20 μm
Wafer Thickness	0.20 mm
Wafer Dimensions	0.35 × 0.35 mm

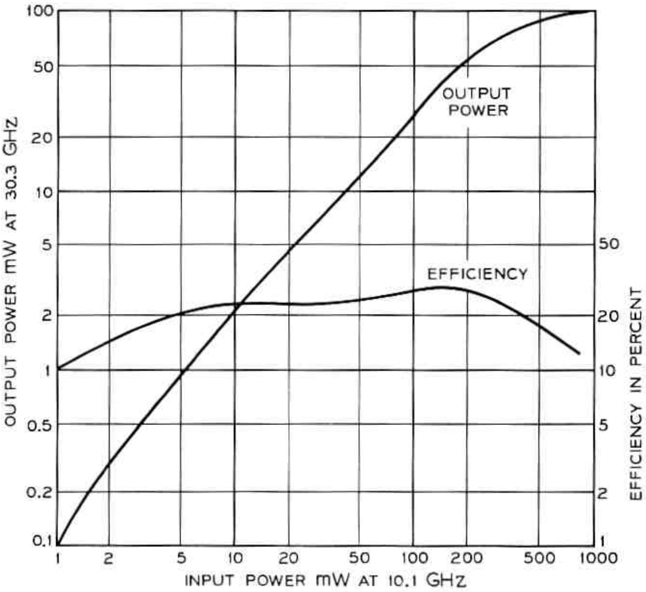


Fig. 3—Output power of multiplier at 30.3 GHz as a function of pump power at 10.1 GHz.

TABLE II—INTEGRATED FREQUENCY MULTIPLIER PERFORMANCE

Input Frequency	10.1 GHz
Input Power	178 mW
Output Frequency	30.3 GHz
Output Power	50 mW
Efficiency	29%
DC-Bias	Self-bias, 6 mA at 3V
Capacitance at Zero Bias	0.15 pF
Capacitance at Breakdown	0.09 pF
Breakdown Voltage	15 volts
Input Microstrip Filter	Five element semi-lumped low-pass filter with three capacitive and two inductive sections, cutoff frequency 11.2 GHz
Output Microstrip Filter	Band-pass filter with two parallel-coupled microstrip resonators, center frequency 30.4 GHz, 3-dB bandwidth 6 GHz
Input of Multiplier	Coaxial-to-microstrip transition, 50-ohm coax to 50-ohm microstrip
Output of Multiplier	Microstrip-to-waveguide transition, 50-ohm microstrip to RG-96/U waveguide

ent launcher is optimized for a standard microstrip line on a silica substrate.

The tuned output power versus frequency for a pump power of 137 mW is plotted in Fig. 4. The 3-dB bandwidth is 2.5 GHz or 8.5 percent and has been achieved by optimizing both diode bias voltage and plunger position of the microstrip-to-waveguide transisition at each frequency. The maximum output power is obtained at 30.3 GHz and is related to the fixed frequency resonance of the idler which occurs at 20.2 GHz. The corresponding oversize multiplier is optimized for an input frequency of 272.7 MHz and an output at 818 MHz. The scaling factor used for building the multiplier shown in Fig. 2 is 37.5 times and the optimized output after reduction of the circuit should be reached at an output frequency of 30.7 GHz. This means that the total error in scaling is 1.3 percent. The error is very small because the photolithographic technique is one of the most accurate methods to reduce the physical dimensions of a microstrip circuit and also because the dielectric constant of the substrate does not change with frequency.[12]

Figure 5 shows the frequency spectrum of the multiplier from 29.3 to 31.3 GHz. The output is stable and free of spurious signals to the maximum sensitivity of the spectrum analyzer which is 40 dB below the output of the multiplier. Good stability and spurious-free operation are also obtained for the oversize multiplier with an 80-percent efficiency and a power output of 2.5 W at 818 MHz.

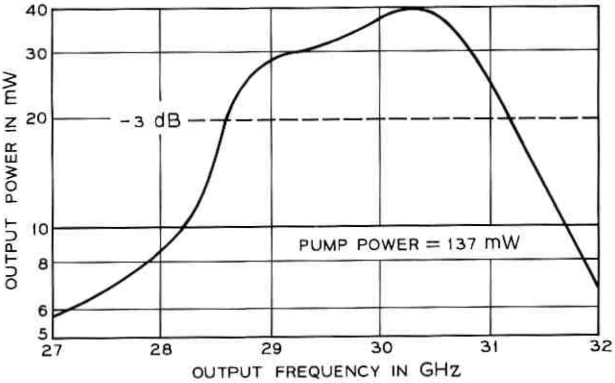


Fig. 4—Tuned output power versus frequency of multiplier for fixed pump power of 137 mW.

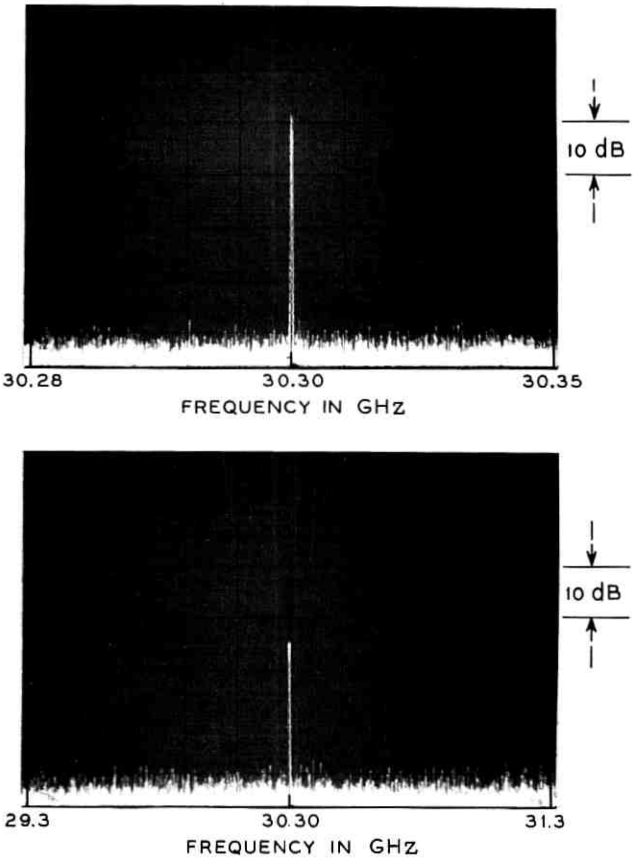


Fig. 5—Output spectrum of multiplier from 29.3 to 31.3 GHz.

IV. CONCLUSIONS

Hybrid integrated multipliers on a dielectric substrate can now be built with output frequencies up to 30 GHz. The power levels and the overall efficiencies which are obtained with a single planar diffused gallium arsenide diode are useful for many future millimeter-wave systems applications.[2,13] Scaling of oversize circuits is a powerful technique for building optimized integrated millimeter-wave circuits.

REFERENCES

1. Ruthroff, C. L., Osborne, T. L., and Bodtmann, W. F., "Short Hop Radio System Experiment," B.S.T.J., *48,* No. 6 (July–August 1969), pp. 1577–1604.
2. Tillotson, L. C., "Use of Frequencies above 10 GHz for Common Carrier Applications," B.S.T.J., *48,* No. 6 (July–August 1969), pp. 1563–1576.
3. Swan, C. B., "Design and Evaluation of a Microwave Varactor Tripler," International Solid State Circuit Conference Digest, *8* (February 1965), pp. 106–107.
4. Lee, T. P., and Burrus, C. A., "A Millimeter-Wave Quadrupler and an Up-Converter Using Planar-Diffused Gallium Arsenide Varacter Diodes," IEEE Trans. on Microwave Theory and Techniques, *MTT-16,* No. 5 (May 1969), pp. 287–296.
5. Gewartowski, J. W., and Ueonohara, M., "Varactor Applications," Chapter 8 in Watson, H. A., ed., *Microwave Semiconductor Devices and Their Circuit Applications,* New York: McGraw-Hill, 1969, pp. 194–270.
6. Burckhardt, C. B., "Analysis of Varactor Frequency Multipliers for Arbitrary Capacitance Variation and Drive Level," B.S.T.J., *44,* No. 4 (April 1965), pp. 675–692.
7. Penfield, P., Jr., and Rafuse, R. P., *Varactor Applications,* Cambridge, Mass.: M.I.T. Press, 1962, pp. 297–435.
8. Matthaei, G. L., Yound, L., and Jones, E. M. T., *Microwave Filters, Impedance-Matching Networks, and Coupling Structures,* New York: McGraw-Hill, 1964, pp. 100–101.
9. W. W. Snell, Jr., "Low-Loss Microstrip Filters Developed by Frequency Scaling," B.S.T.J., this issue, pp. 1919–1931.
10. Burrus, C. A., "Planar Diffused Gallium Arsenide Millimeter-Wave Varactor Diodes," Proc. IEEE, *55,* No. 6 (June 1967), pp. 1104–1105.
11. Knerr, K. H., "A New Type of Waveguide-to-Stripline Transition," IEEE Trans. on Microwave Theory and Techniques, *MTT-16,* No. 3 (March 1969), pp. 192–194.
12. Von Hippel, R., *Dielectric Materials and Applications,* New York: Chapman and Hall, 1954, pp. 310–311.
13. Tillotson, L. C., "Millimeter Wave Radio," Science, *170,* No. 3953 (October 2, 1970), pp. 31–36.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Some Analytical Investigations of Phase Contrast Imaging

By C. M. NAGEL, JR.

(Manuscript received December 1, 1970)

An analytical study of phase contrast imaging is performed. It is shown that the complex disturbance in the image plane can be represented as a convolution of the object disturbance and the Fourier transform of the transmission function of the phase plate. The simple theory of the phase contrast microscope is then derived as a limiting case of this more general result that is applicable when the size of the phase object is small compared to the area of the entrance aperture of the system. The response of a general system to several simple large phase objects is also examined, and it is shown that qualitative information about these objects can be obtained from the intensity pattern when the phase perturbations are small, providing the background is sufficiently uniform, and the size of the phase spot on the phase plate is carefully chosen. The study provides insight into the type of performance that can be achieved, for example, if phase contrast imaging is used to extract phase information from an optical memory or if it is used as an experimental tool to study the qualitative behavior of such phenomena as clear air turbulence.

I. INTRODUCTION

One of the most popular techniques for converting a spatial phase variation into a spatial intensity variation is the phase contrast imaging scheme originally proposed by F. Zernike[1-3] in 1935. Proceeding mainly from a heuristic point of view, Zernike recognized that, with a conventional imaging system, most of the direct light passes through a small region R in the focal plane while, in many instances, most of the diffracted light is scattered away from this region. If a plate whose transmission function is given by

$$T = \begin{cases} ae^{j\alpha} & \text{in } R \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

is placed in the focal plane, the direct light can be modified and made to interfere with the diffracted light to produce an intensity pattern in the image plane. Moreover, if the phase perturbations ϕ that are introduced by the object are small enough so that the approximation

$$e^{j\phi} \approx 1 + j\phi$$

holds, the intensity in the image plane and the phase of the object may be linearly related. These concepts led to the design of the phase contrast microscope.

The key assumption in the theory is that the effect of the phase plate on the diffracted light is negligible. In 1953, H. H. Hopkins[4] rigorously demonstrated that, if this assumption holds, a phase object will produce an intensity pattern given by

$$I = \frac{1}{M^2} \mid ae^{j\alpha} - 1 + e^{j\phi} \mid^2. \tag{2}$$

In this expression, M is the magnification of the system, and it is assumed in the derivation that the object is illuminated by a unit amplitude monochromatic plane wave. When ϕ is small, a linear relationship,

$$I = \frac{1}{M^2} (a^2 + 2a\phi \sin \alpha), \tag{3}$$

results, as predicted by Zernike.†

There are many potential applications of phase contrast imaging in which the basic assumption employed in Refs. 1–5 will not be valid. Therefore, it is interesting to examine what results will follow if these assumptions are not made. In this regard, Hopkins[6] rigorously studied the diffraction images of circular disks under coherent illumination, and M. De and S. C. Som,[7] and De and P. K. Mondal,[8] investigated the images produced by similar objects under incoherent illumination. In this paper, we will derive the theory from the point of view of Fourier optics and demonstrate that, in general, the complex disturbance over the image plane can be represented by a convolution of the object disturbance and the Fourier transform of the transmission function of the phase plate. The approximate theory, equations (2) and (3), is then shown to be a limiting case of the general theory, applicable if the size of the phase object is small compared to the entrance aperture of the system. Such is usually the case when the

† A related analysis is also given in Born and Wolf,[5] Chapter 8.

magnification of the system is large. For larger objects, the convolution integral yields a complicated relationship between the phase of the object and the intensity of the image.

The response of the system to several simple large phase objects is then studied, and it is shown that qualitative information about the phase distributions of these objects can be obtained when the phase perturbations are small if an appropriate phase plate is employed.

The work that is discussed in this paper complements the experimental investigations carried on at Bell Telephone Laboratories by N. J. Kolettis.

II. GENERAL THEORY

The optical system that we will consider is shown in Fig. 1. A circular entrance aperture with radius R_o is located a distance d_o in front of the lens, and a phase plate with a transmission function

$$T_f(r_f) = \begin{cases} ae^{i\alpha} & r_f \leqq R_f^o \\ 1 & R_f^o < r_f \leqq R_{fo} \\ 0 & R_{fo} < r_f \end{cases} \tag{4}$$

is centered in the focal plane. The phase object is assumed to occupy some portion of the entrance aperture, and the illuminating light is taken to be a normally incident unit amplitude monochromatic plane wave.

Let the complex disturbance in the object plane be denoted by $U_o(x_o, y_o)$ where U_o is assumed to be zero at points outside the entrance aperture. Then, from the Fresnel approximation,[9] it follows

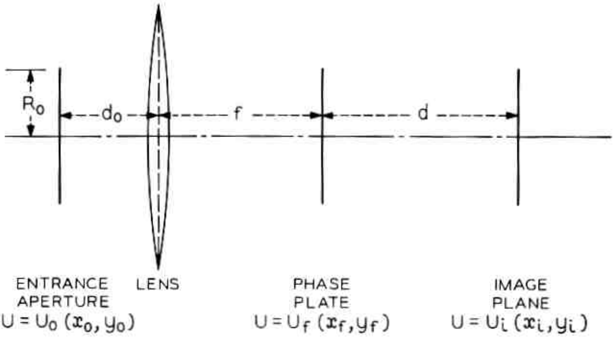


Fig. 1—System geometry.

that the complex disturbance in the focal plane to the left of the phase plate is

$$U_f(x_f\,,\,y_f) = \frac{\exp\,[jk(d_o + f + n\,\Delta_o)]}{j\lambda f}\exp\left[\frac{jk}{2f}\left(1 - \frac{d_o}{f}\right)(x_f^2 + y_f^2)\right]$$
$$\cdot \iint_{-\infty}^{\infty} U_o(x_o\,,\,y_o)P_L\left(x_o + \frac{d_o}{f}\,x_f\,,\,y_o + \frac{d_o}{f}\,y_f\right)$$
$$\cdot \exp\left[\frac{-jk}{f}\,(x_ox_f + y_oy_f)\right]dx_o\,dy_o\,. \qquad (5)$$

In this expression f is the focal length of the lens, λ is the wavelength of the illuminating light, k is the wavenumber $2\pi/\lambda$, and $kn\Delta_o$ is the phase shift introduced by the lens. P_L is the pupil function of the lens which assumes the value of unity over the region covered by the lens, and is zero elsewhere.

Now, the complex disturbance on the right side of the phase plate is simply U_fT_f, and, therefore, if the Fresnel approximation is also used to describe the propagation from this plane to the image plane, the complex disturbance in the image plane becomes

$$U_i(x_i\,,\,y_i)$$
$$= \frac{K}{\lambda^2 fd}\iint_{-\infty}^{\infty}\iint_{-\infty}^{\infty} U_o(x_o\,,\,y_o)P_L\left(x_o + \frac{d_o}{f}\,x_f\,,\,y_o + \frac{d_o}{f}\,y_f\right)T_f(x_f\,,\,y_f)$$
$$\cdot \exp\left\{-\frac{j2\pi}{\lambda d}\,[(Mx_o + x_i)x_f + (My_o + y_i)y_f]\right\}dx_o\,dy_o\,dx_f\,dy_f\,, \qquad (6)$$

where

$$K = -\exp\,[jk(d_o + f + d + n\,\Delta_o)]\,\exp\left[\frac{jk}{2d}\,(x_i^2 + y_i^2)\right].$$

Here we are assuming that the distances d_o, d, and f obey the lens law,

$$\frac{1}{f}\left(1 - \frac{d_o}{f}\right) + \frac{1}{d} = 0,$$

and we have denoted the magnification of the system, d/f, by M.

The integral in (6) is complicated by the coupling of the coordinates x_o, x_f and y_o, y_f via the pupil function of the lens. This phenomenon, known as the vignetting effect,[9] can be ignored if it is assumed that the lens is large enough so that the sum of the squares of the arguments of P_L is always less than the square of the radius of the lens

for all values of x_o, x_f, y_o, and y_f for which U_o and T_f are nonzero.†
With this assumption P_L may be taken as one.

To further simplify equation (6) we introduce the notation

$$\tilde{x}_i = \frac{x_i}{M}, \qquad \tilde{y}_i = \frac{y_i}{M}, \qquad x = x_o + \tilde{x}_i , \qquad y = y_o + \tilde{y}_i .$$

Then, neglecting the vignetting effect,

$$U_i(x_i , y_i) = \frac{K}{\lambda^2 fd} \iint_{-\infty}^{\infty} U_o(x - \tilde{x}_i , y - \tilde{y}_i) F[T_f] \, dx \, dy$$

where

$$F[T_f] = \iint_{-\infty}^{\infty} T_f(x_f , y_f) \exp\left[-\frac{j2\pi M}{\lambda d} (xx_f + yy_f) \right] dx_f \, dy_f . \quad (7)$$

We observe that the disturbance in the image plane can be written as a convolution of the disturbance in the object plane and the Fourier transform of the transmission function of the phase plate evaluated at the spatial frequencies $Mx/\lambda d$, $My/\lambda d$.

Let us now examine the form of $F[T_f]$ more carefully. The transmission function of the phase plate, from (4), is given by

$$T_f(r_f) = (ae^{i\alpha} - 1) \operatorname{circ} (r_f/R_f^o) + \operatorname{circ} (r_f/R_{fo})$$

with

$$r_f = (x_f^2 + y_f^2)^{\frac{1}{2}}$$

and

$$\operatorname{circ} (r_f/R) = \begin{cases} 1 & r_f \leqq R \\ 0 & r_f > R. \end{cases}$$

It follows then that

$$F[T_f] = \frac{\lambda^2 d^2}{M^2} \left[(ae^{i\alpha} - 1) \frac{MR_f^o}{\lambda d} \frac{J_1\left(\frac{2\pi MR_f^o}{\lambda d} r\right)}{r} + \frac{MR_{fo}}{\lambda d} \frac{J_1\left(\frac{2\pi MR_{fo}}{\lambda d} r\right)}{r} \right] \quad (8)$$

where

$$r = (x^2 + y^2)^{\frac{1}{2}},$$

and J_1 is the Bessel function of the first kind of order one.

† Recall $U_o = 0$ for $x_o^2 + y_o^2 > R_o$, and $T_f = 0$ for $x_f^2 + y_f^2 > R_{fo}$.

Now, at optical wavelengths, R_{fo}/λ is exceedingly large, and for all practical purposes, the second term in (8) can be approximated by a two-dimensional delta function with little error.[†] On the other hand, R_f^o is typically of the order of the radius of the Airy disk of the diffraction pattern of the entrance aperture so that the first term cannot be so simplified. To emphasize this fact, R_f^o will be written as

$$R_f^o = \gamma \frac{\lambda f}{R_o}$$

where γ is a constant of the order of unity. ($\gamma \approx 0.61$ if R_f^o is the radius of the Airy disk.)

With this notation and the definition of M, equation (8) becomes

$$F[T_f] \approx \frac{\lambda^2 d^2}{M^2}\left[(ae^{i\alpha} - 1)\frac{\gamma}{R_o}\frac{J_1\left(\frac{2\pi\gamma}{R_o} r\right)}{r} + \delta(r)\right]. \tag{9}$$

Returning to the expression for the complex disturbance in the image plane (7), we conclude that

$$U_i(x_i\,,\, y_i) \approx \frac{K}{M}\left[(ae^{i\alpha} - 1)\frac{\gamma}{R_o}\iint_{-\infty}^{\infty} U_o\left(x - \frac{x_i}{M}\,,\, y - \frac{y_i}{M}\right)\right.$$
$$\left.\cdot\frac{J_1\left[\frac{2\pi\gamma}{R_o}(x^2 + y^2)^{\frac{1}{2}}\right]}{(x^2 + y^2)^{\frac{1}{2}}}\,dx\,dy + U_o\left(-\frac{x_i}{M}\,,\, -\frac{y_i}{M}\right)\right]. \tag{10}$$

Equation (10) is the fundamental result that will be employed throughout the remainder of this paper.

III. SMALL PHASE OBJECTS—THE PHASE CONTRAST MICROSCOPE

The form of equation (10) indicates that, in general, a complicated relationship exists between the object and the image. We will now demonstrate, however, that the simple theory, equations (2) and (3), may be employed *if the ratio of the area of the phase object to the area of the entrance aperture is sufficiently small.* This is usually the case for a microscopic system.

We will assume that the phase object is centered in the entrance

[†] We shall observe later that this approximation yields sharp jumps in intensity when the phase object possesses phase discontinuities. In practice some smoothing of the discontinuities in the intensity pattern will always result. (See Ref. 10 for example.) We are neglecting these lower order effects for simplicity.

aperture and write $U_o(x_o, y_o)$ as

$$U_o(x_o, y_o) = \text{circ}\left[\frac{(x_o^2 + y_o^2)^{\frac{1}{2}}}{R_o}\right] + (e^{i\phi} - 1)P_o ,$$

where P_o has a value of one over the region covered by the phase object and a value of zero elsewhere. In the limit of a very small phase object,† the contribution of the second term in U_o to the *integral* in (10) approaches zero, and the integral can be approximated by

$$\text{I} = \frac{\gamma}{R_o} \iint_{-\infty}^{\infty} \text{circ}\left[\frac{((x - \tilde{x}_i)^2 + (y - \tilde{y}_i)^2)^{\frac{1}{2}}}{R_o}\right] \frac{J_1\left[\frac{2\pi\gamma}{R_o}(x^2 + y^2)^{\frac{1}{2}}\right]}{(x^2 + y^2)^{\frac{1}{2}}}\, dx\, dy.$$

The convolution is most easily evaluated by first taking the Fourier transform of I and then taking the inverse transform of the results. These two operations yield

$$\text{I} = \int_0^{2\pi\gamma} J_1(t) J_0\left(\frac{r_i}{MR_o}\, t\right) dt \tag{11}$$

where r_i is the radial coordinate in the image plane.

If we now restrict the coordinates in the image plane to the region corresponding to the image of the phase object, the ratio r_i/MR_o will be small, and I can be approximated by

$$\text{I} \approx \int_0^{2\pi\gamma} J_1(t)\, dt = 1 - J_0(2\pi\gamma).$$

Finally, with an appropriate choice of the radius of the central spot of the phase plate (and thus γ), the contribution from the Bessel function $J_0(2\pi\gamma)$ can be made to vanish, and we conclude that, for a sufficiently small phase object and a suitable γ,

$$U_i(x_i, y_i) \approx \frac{K}{M}\left[(ae^{i\alpha} - 1) + \exp\left[j\phi\left(-\frac{x_i}{M}, -\frac{y_i}{M}\right)\right]\right].$$

The intensity of the image is then

$$I(x_i, y_i) \equiv |U_i|^2 \approx \frac{1}{M^2}\left|(ae^{i\alpha} - 1) + \exp\left[j\phi\left(-\frac{x_i}{M}, -\frac{y_i}{M}\right)\right]\right|^2$$

which agrees with equation (3). The simple theory of the phase contrast microscope is, therefore, a limiting case of the more general approach.

† Consider a single cell under a microscope, for example.

IV. MACROSCOPIC SYSTEMS

In many potential applications of phase contrast imaging the magnification of the system may be of the order of unity, and the area of the phase object may be comparable to the area of the entrance aperture. In these cases, the approximations that were employed in the last section cannot be used, and the complex disturbance in the image plane can, in general, only be determined via a direct application of equation (10). In order to provide some representative results, we have chosen a few simple, yet important, examples for which the integral in (10) can be easily evaluated.

4.1 *Background*

The purpose of any phase contrast system is to give a visual representation of the phase distribution in the entrance aperture. If a system is to yield meaningful results, therefore, the intensity of the image of the aperture should be essentially constant when no phase objects are present. In this case

$$U_o(x_o\,,\,y_o) = \text{circ}\left[\frac{(x_o^2 + y_o^2)^{\frac{1}{2}}}{R_o}\right]$$

and from (10) and (11) it follows that

$$I(r_i) = \frac{1}{M^2}\left|\,(ae^{i\alpha} - 1)\int_0^z J_1(t)J_0\!\left(\frac{r_i}{MR_o}\,t\right)dt + 1\,\right|^2 \qquad (r_i \leqq MR_o) \tag{12}$$

where the notation $z = 2\pi\gamma$ has been introduced.

For our later purposes, it is convenient to define a normalized intensity, $\tilde{I} = M^2 I$. Plots of this function are shown in Figs. 2–5 for various values of z. Clearly, the response of the system to the illuminating light is critically dependent upon the size of the central spot of the phase plate. Where $a = 1$ and $\alpha = \pi/2$, a typical case that is employed in phase contrast imaging, near uniform backgrounds result for values of z corresponding to spot sizes on the order of one-half the Airy disk or smaller.† For larger z, the background varies considerably until several rings in the diffraction pattern of the entrance aperture are covered by the phase spot.‡ We shall see later that these variations produce distortions in the intensity pattern of phase objects when they

† The task of making a phase plate with a spot size this small may be difficult.

‡ The phase contrast system will begin to behave like a direct imaging system for spot sizes larger than those shown, and thus large z values are not practical.

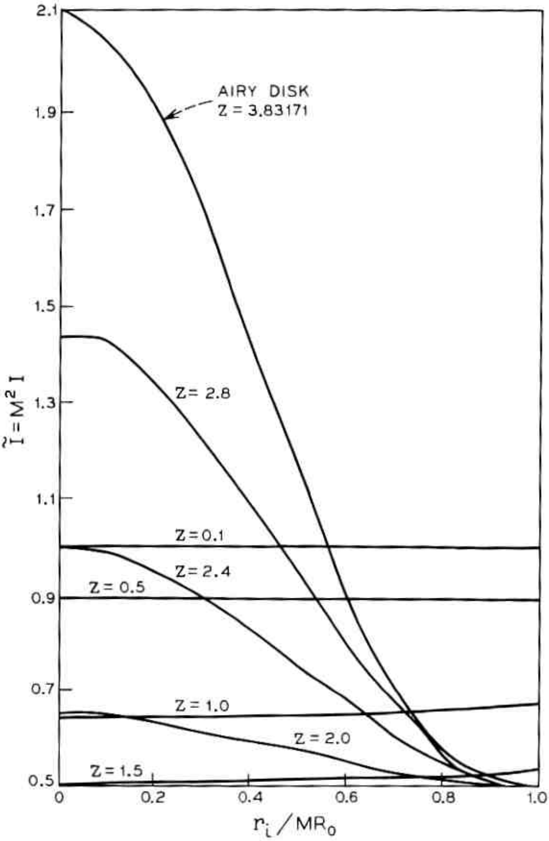


Fig. 2—Background intensity distributions for spot sizes less than or equal to the Airy disk ($a = 1$, $\alpha = \pi/2$).

are placed in the aperture. The case $a = 1$, $\alpha = \pi/2$, $z = 1.5$ will be of special interest to us, since in this case the integral in (12) is approximately 1/2 except when r_i is near the edge of the image of the aperture, the background is essentially constant, and several interesting results can be obtained.

Frequently, a phase contrast system will employ a partially absorbing phase plate ($a < 1$) to improve the contrast between the image of the phase object and the background [see equation (3)]. Background intensities for $a = 0.1$, $\alpha = \pi/2$ and various values of z are shown in Figs. 4 and 5. There it may be observed that, in general, the background intensity is reduced, although a bright ring appears in the

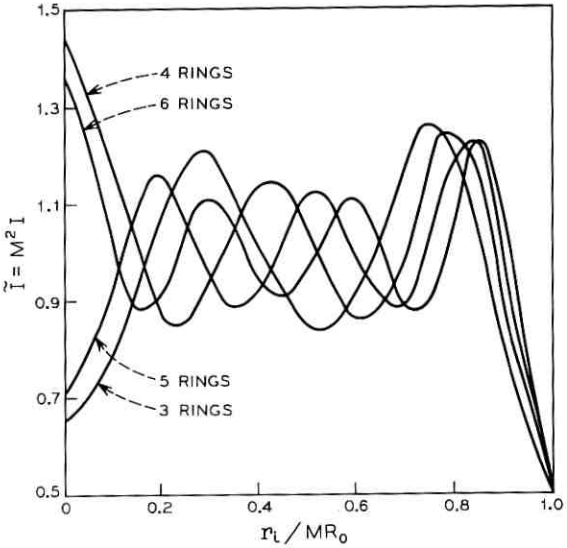


Fig. 3—Background intensity distributions for spot sizes larger than the Airy disk ($a = 1$, $\alpha = \pi/2$).

region close to the edge of the image of the aperture. (This ring has been observed experimentally by Kolettis.) If this edge effect is ignored, near uniform backgrounds are obtained for values of z that include five or six Airy rings.

Taken together, the above results indicate that care must be exercised in designing a phase contrast system if near uniform backgrounds are to be achieved.

4.2 *Phase Disk*

Consider the "phase disk" defined by

$$U_o(x_o\,,\,y_o) = (e^{i\phi} - 1)\ \mathrm{circ}\left[\frac{(x_o^2 + y_o^2)^{\frac{1}{2}}}{R_o^{(1)}}\right] + \mathrm{circ}\left[\frac{(x_o^2 + y_o^2)^{\frac{1}{2}}}{R_o}\right]$$

$$(0 \leqq R_o^{(1)} \leqq R_o) \qquad (13)$$

where ϕ is a constant. The geometry corresponds to a disk with radius $R_o^{(1)}$ centered in the entrance aperture which introduces a constant phase shift ϕ in the incident wave. Substituting in equation (10) and using the technique that was employed to derive equation (11), we obtain the intensity distribution

$$\tilde{I}(r_i) = \begin{cases} |\,(ae^{i\alpha} - 1)[(e^{i\phi} - 1)\mathrm{I}^{(1)}(r_i) + \mathrm{I}(r_i)] + e^{i\phi}\,|^2, \\ \hfill r_i \leqq MR_o^{(1)} \\ |\,(ae^{i\alpha} - 1)[(e^{i\phi} - 1)\mathrm{I}^{(1)}(r_i) + \mathrm{I}(r_i)] + 1\,|^2, \\ \hfill MR_o^{(1)} < r_i \leqq MR_o\,. \end{cases} \tag{14}$$

In this expression $\mathrm{I}(r_i)$ is the integral given in (11) and

$$\mathrm{I}^{(1)}(r_i) = \frac{R_o^{(1)}}{R_o} \int_0^{2\pi\gamma} J_1\left(\frac{R_o^{(1)}t}{R_o}\right) J_0\left(\frac{r_i}{MR_o}\,t\right) dt. \tag{15}$$

Note that the integral I only takes into account the finite extent of the entrance aperture, while $\mathrm{I}^{(1)}$ includes the size of the phase object as well.

Plots of (14) are shown in Figs. 6–8 for $a = 1$, $\alpha = \pi/2$ and $z = 1.5$. The value of z has been chosen to correspond to a near uniform back-

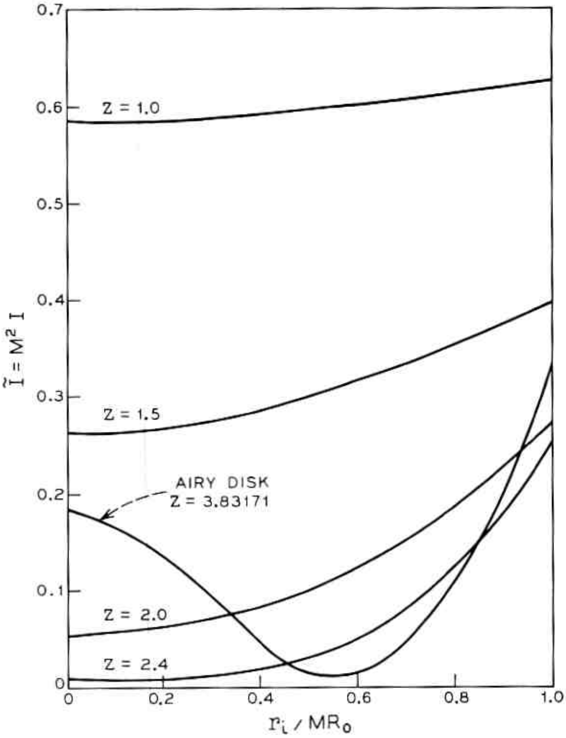


Fig. 4—Background intensity distributions for spot sizes less than or equal to the Airy disk ($a = 0.1$, $\alpha = \pi/2$).

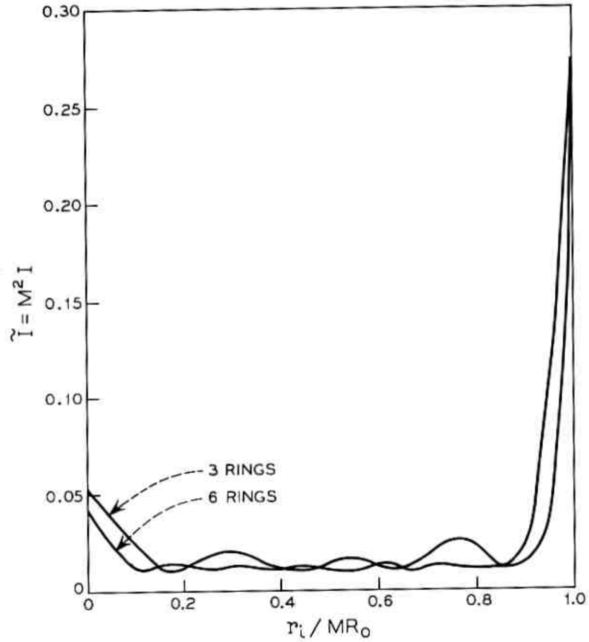


Fig. 5—Background intensity distributions for spot sizes larger than the Airy disk ($a = 0.1$, $\alpha = \pi/2$).

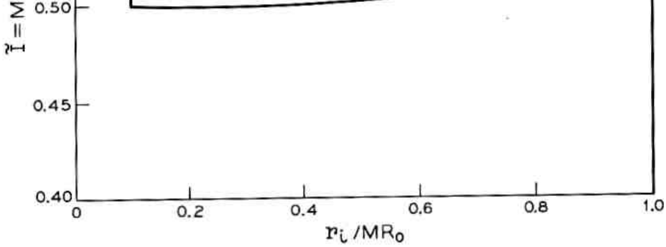


Fig. 6—Response to a phase disk ($\phi = 0.1$, $R_o^{(1)}/R_o = 0.1$).

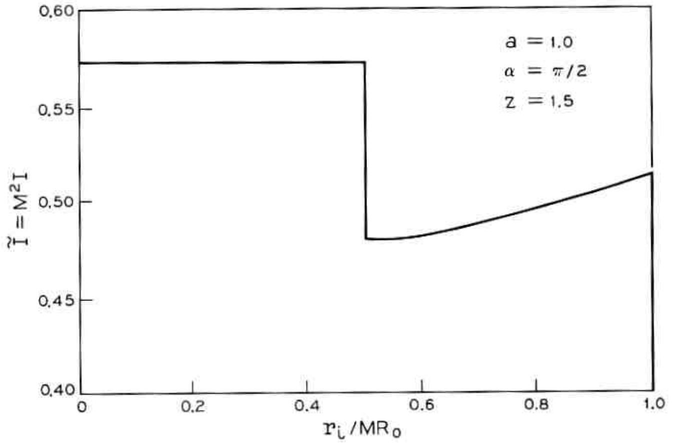


Fig. 7—Response to a phase disk ($\phi = 0.1, R_o{}^{(1)}/R_o = 0.5$).

ground when $R_o^{(1)} = R_o$ or when $R_o^{(1)} = 0$. It is clear from the figures that, for the small value of ϕ chosen, the qualitative behavior of the phase distribution in the entrance aperture is reproduced quite nicely. The phase disk can be readily observed and the magnitude of the discontinuity in $\tilde{I}$ is on the order of ϕ.†

In Fig. 9 we have indicated the response of the same system with the exception that ϕ has now been assigned a value of π. Note that in this case, the jump in $\tilde{I}$ depends critically upon the size of the phase object. Moreover, it appears that a value of $R_o^{(1)}$ exists for which no contrast results and beyond which the contrast is reversed. This suggests that for larger ϕ values, the system performs quite poorly.

To explain this behavior, we need only to return to equation (14) from which it immediately follows that the magnitude of the discontinuity in $\tilde{I}$ is

$$\begin{aligned}\Delta\tilde{I} &= 2\ \mathrm{Re}\ \{(ae^{i\alpha} - 1)(e^{i\phi} - 1)[(e^{i\phi} - 1)\mathrm{I}^{(1)}(MR_o^{(1)}) + \mathrm{I}(MR_o^{(1)})]\}\\ &= 2(a\cos\alpha - 1)(1 - \cos\phi)(2\mathrm{I}^{(1)} - \mathrm{I}) + 2a\mathrm{I}\sin\alpha\sin\phi. \qquad (16)\end{aligned}$$

For small ϕ,

$$\Delta\tilde{I} \approx (2a\mathrm{I}\sin\alpha)\phi,$$

while for larger ϕ, the full equation (16) must be employed. Recalling, that for $z = 1.5$, $\mathrm{I} \approx 0.5$ for most values of r_i, we conclude that, for

† The increase in intensity in the outer portion of the image is due to the slight nonuniformity of the background in this region.

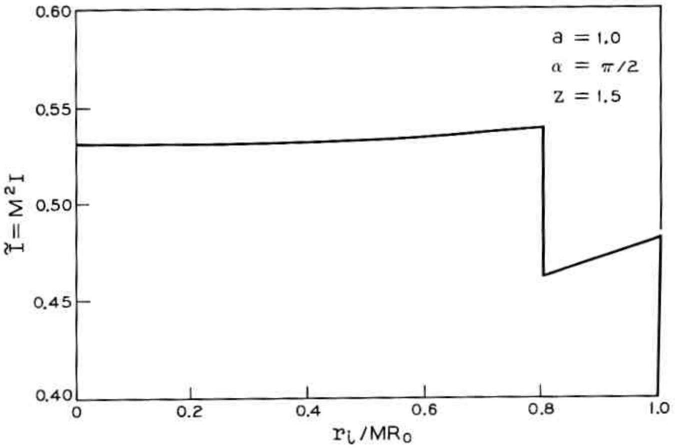


Fig. 8—Response to a phase disk ($\phi = 0.1$, $R_o^{(1)}/R_o = 0.8$).

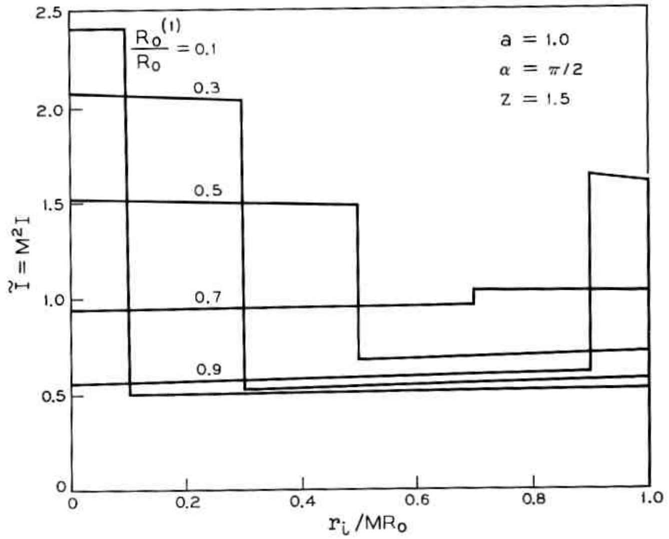


Fig. 9—Response to phase disks of various sizes ($\phi = \pi$).

$a = 1$, $\alpha = \pi/2$, $z = 1.5$, and ϕ small,

$$\Delta\tilde{I} \approx \phi$$

as observed.

When $\phi = \pi$, however, equation (16) reduces to

$$\Delta\tilde{I} = 4(a \cos \alpha - 1)[2\mathrm{I}^{(1)}(MR_o^{(1)}) - \mathrm{I}(MR_o^{(1)})].$$

From the definitions of the integrals (11) and (15), it is clear that for $\phi = \pi$ there exists a value of $R_o^{(1)}$ for which $\Delta\tilde{I} = 0$ and beyond which the contrast is reversed.

In the preceding section, we observed that nonuniform backgrounds can result when the size of the central spot of the phase plate is not carefully chosen. Figure 10 demonstrates the response of a phase contrast system to a phase disk when this is the case. The value of z has been chosen so that the spot size corresponds to the Airy disk of the diffraction pattern of the entrance aperture. The background corresponding to this value of z is shown in Fig. 2. Note that the phase disk simply perturbs the background pattern yielding an intensity profile that does not have the desired step function shape.

4.3 *Phase Rings*

The analytical results presented in Section 4.2 indicate that a discontinuity in the radial coordinate of the phase distribution of the entrance aperture can be qualitatively reproduced in the intensity pattern providing a sufficiently uniform background is chosen and the magnitude of the discontinuity is small. In other cases, such as a large discontinuity or a nonuniform background, rather poor results are obtained. We will now generalize these results to a more complicated radial phase distribution.

Suppose the phase distribution in the entrance aperture consists of a central disk and a sequence of annular rings in each of which the phase is some constant ϕ_k. Let the radii defining the location of the phase discontinuities be $R_o^{(k)}$, $k = 1, \cdots, n - 1$, and let $R_o^{(n)} = R_o$. Then U_o can be written as

$$U_o(x_o, y_o) = \sum_{k=1}^{n} U_o^{(k)} \operatorname{circ}\left[\frac{(x_o^2 + y_o^2)^{\frac{1}{2}}}{R_o^{(k)}}\right] \tag{17}$$

where

$$U_o^{(k)} = \begin{cases} \exp[j\phi_k] - \exp[j\phi_{k+1}] & k = 1, \cdots, n-1 \\ \exp[j\phi_n] & k = n. \end{cases}$$

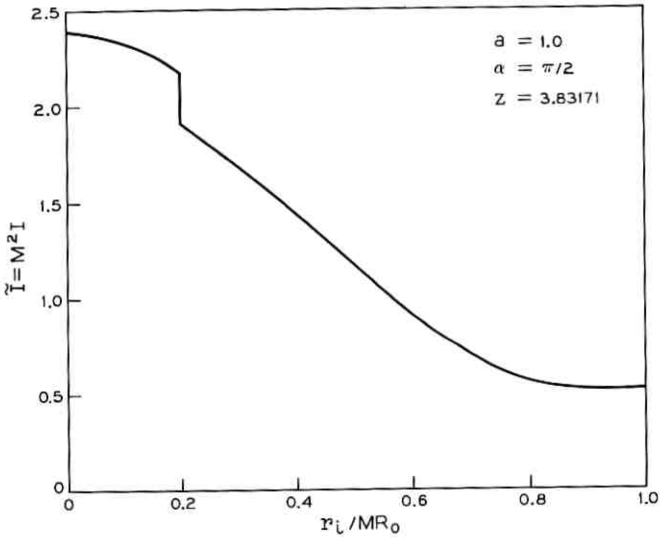


Fig. 10—Response to a phase disk when the phase plate spot size equals the Airy disk of the entrance aperture ($\phi = 0.1$, $R_o^{(1)}/R_o = 0.2$).

The normalized intensity distribution immediately follows as

$$\tilde{I}(r_i) = \begin{cases} \left| (ae^{j\alpha} - 1) \sum_{k=1}^{n} U_o^{(k)} \mathrm{I}^{(k)}(r_i) + \exp[j\phi_1] \right|^2 \\ \qquad\qquad 0 \leqq r_i \leqq MR_o^{(1)} \\ \left| (ae^{j\alpha} - 1) \sum_{k=1}^{n} U_o^{(k)} \mathrm{I}^{(k)}(r_i) + \exp[j\phi_{l+1}] \right|^2 \\ \quad MR_o^{(l)} < r_i \leqq MR_o^{(l+1)}, \qquad l = 1, \cdots, n-1. \end{cases} \tag{18}$$

In this expression $\mathrm{I}^{(k)}(r_i)$ is the generalization of (15) which is given by

$$\mathrm{I}^{(k)}(r_i) = \frac{R_o^{(k)}}{R_o} \int_0^z J_1\left(\frac{R_o^{(k)}}{R_o}\, t\right) J_0\left(\frac{r_i}{MR_o}\, t\right) dt. \tag{19}$$

Plots of (18) are shown in Figs. 11–13 for $a = 1$, $\alpha = \pi/2$, and $z = 1.5$. The assumed phase distributions in the entrance aperture are given at the top of the figures and the resulting intensity profiles are shown below. A qualitative reproduction of the phase distributions results in each case.

The behavior of (18) in the limit of small phases, ϕ_k, explains some of the success that was achieved in the figures. From the definition of the coefficients $U_o^{(k)}$ it follows that, if all the ϕ_k are small and, in particular, if $\phi_n = 0$,

$$U_o^{(k)} \approx \begin{cases} j(\phi_k - \phi_{k+1}) & k = 1, \cdots, n-1 \\ 1 & k = n. \end{cases}$$

Therefore, the jumps in intensity at the points $r_i = MR_o^{(k)}$, to first order in the phase jumps, is just

$$\begin{aligned} \Delta\tilde{I}(MR_o^{(k)}) &\approx 2 \operatorname{Re} \{(ae^{j\alpha} - 1)j(\phi_{k+1} - \phi_k)\mathrm{I}^{(n)}(MR_o^{(k)})\} \\ &= 2a \sin \alpha(\phi_k - \phi_{k+1})\mathrm{I}^{(n)}(MR_o^{(k)}). \end{aligned} \quad (20)$$

Now, in the cases shown in the figures, $a = 1$, $\alpha = \pi/2$, and $z = 1.5$. Recalling that for $z = 1.5$, $\mathrm{I}^{(n)} \approx 0.5$ we conclude that

$$\Delta\tilde{I}(MR_o^{(k)}) \approx \phi_k - \phi_{k+1}\,.$$

Thus, if the ϕ_k are sufficiently small, and an appropriate value of z

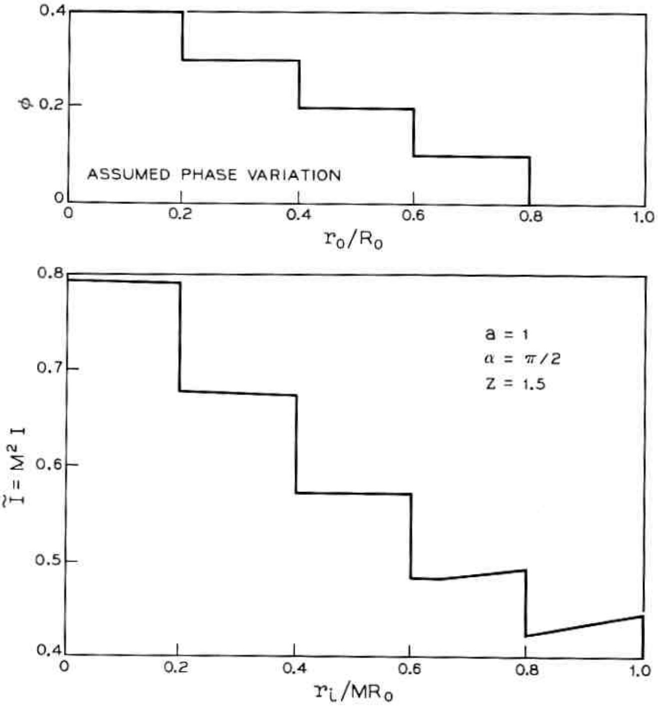


Fig. 11—Response to phase rings with decreasing phase.

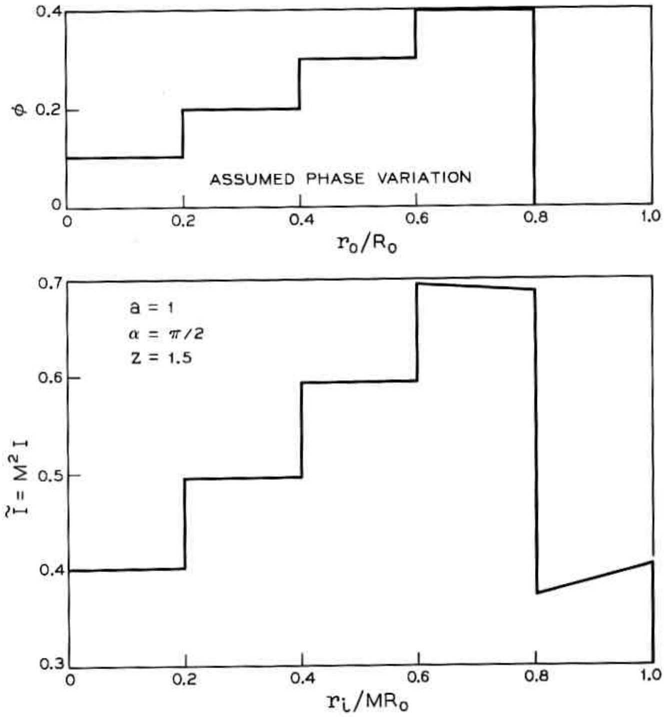


Fig. 12—Response to phase rings with increasing phase.

is chosen, the jumps in intensity are of the order of the jumps in phase.

Figures 11–13 also indicate that, for the values of the parameters chosen, the piecewise constant behavior of the phase distributions is essentially reproduced in the intensity profile. The success here is mainly due to the fact that not only was z chosen to yield a nearly uniform background, but also the z value selected was relatively small. Under these conditions the dependence of the integrals $I^{(k)}(r_i)$ is suppressed,[†] and the effect of the phase object is to introduce essentially piecewise constant perturbations to a nearly uniform background.

To demonstrate what effects can occur for a larger z value, we have plotted in Fig. 14 the response of a phase contrast system to a sequence of phase rings when $a = 0.1$, $\alpha = \pi/2$, and $z = 21.21163$.[‡] This is a practical case to consider since small values of a are fre-

[†] Consider equation (19) in the limit of small z.

[‡] This value of z corresponds to a phase plate spot size that includes about five and one-half Airy rings.

quently employed in phase constant microscopy to improve the contrast between the phase object and the background. (For our purposes, large z values are needed to give a fairly uniform background over a large portion of the image of the entrance aperture when a is small.) Unfortunately, the results that are observed are then quite poor. The problem here is that the integrals $I^{(k)}(r_i)$ are now highly nonlinear functions of r_i, and, therefore, nonlinear perturbations of the order of the background are introduced.[†]

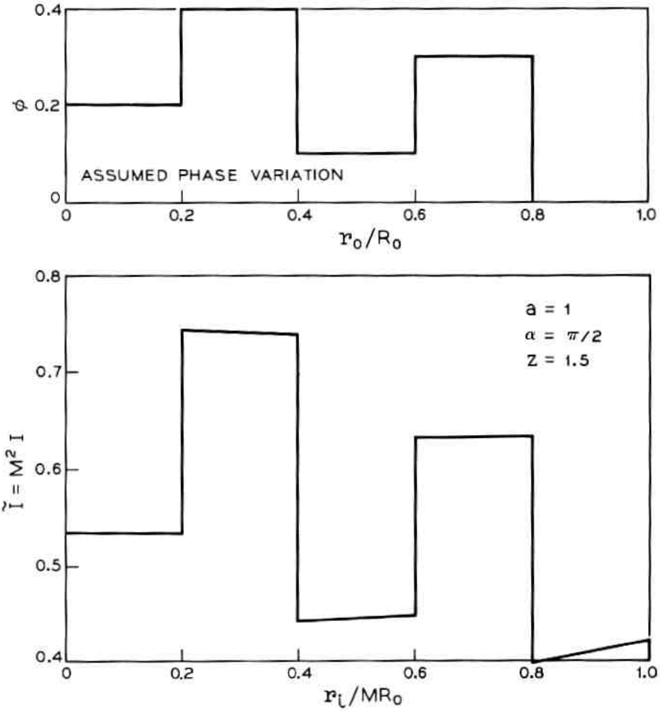


Fig. 13—Response to phase rings with alternating phase.

4.4 *Semi-circular Phase Disks*

The results of the last section demonstrate that with an appropriate choice of the parameters a, α, and z, circularly symmetric phase

[†] Figure 5 demonstrates that for the values of the parameters chosen, the background is of the order of $a^2 = 0.01$. However, from equation (20), we observe that the jumps in $\tilde{I}$, which we may take as indications of the magnitudes of the perturbations of the background, are of the same order when $\phi_k - \phi_{k+1}$ is small.

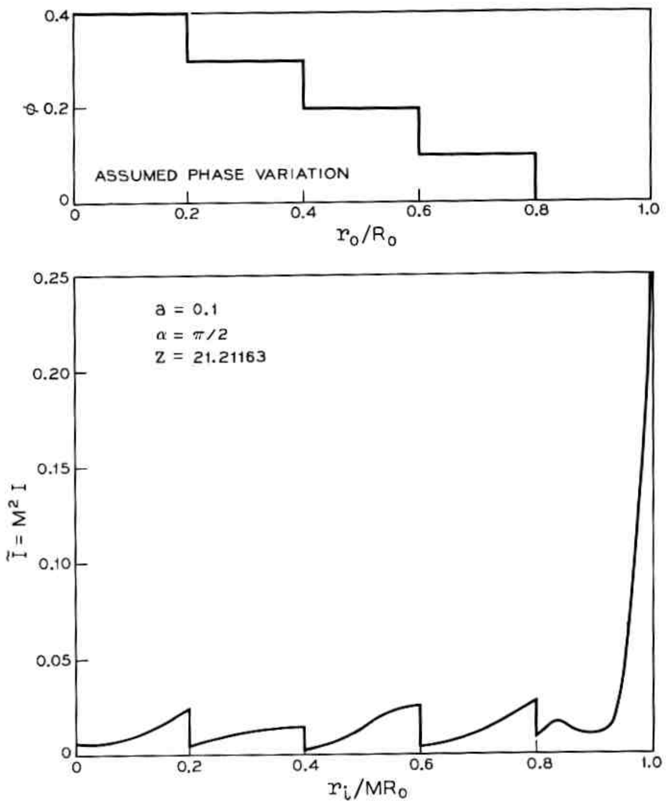


Fig. 14—Response to phase rings ($a = 0.1$, $z = 21.21163$).

distributions possessing discontinuities in the radial coordinate can be qualitatively reproduced in the intensity patterns. As a final example, we now examine the response of a phase contrast system to a simple angular discontinuity.

Let the coordinates in the object plane be r_o and θ_o and consider the complex disturbance

$$U_o(r_o , \theta_o) = \text{circ } (r_o/R_o) \exp [j\Theta(\theta_o)], \tag{21}$$

where

$$\Theta(\theta_o) = \begin{cases} \gamma_1 & 0 \leqq \theta_o \leqq \pi \\ \gamma_2 & \pi < \theta_o < 2\pi \end{cases}$$

with γ_1 and γ_2 assumed constant. With this phase object, the entrance aperture is divided into two semi-circular regions in each of which the phase is constant.

In order to calculate the intensity pattern that is produced by (21), we must compute the convolution integral (see equation (10)):

$$\mathrm{I}_s = U_o * \frac{\gamma}{R_o} \frac{J_1(2\pi\gamma r/R_o)}{r}. \tag{22}$$

As in our previous analysis, we will accomplish this task by first taking the Fourier transform of I_s and then taking the inverse transform of the result.

Now, it is a well-known result in the theory of Fourier transforms[9] that if a function $g(r, \theta)$ is separable in r and θ, i.e., $g(r, \theta) = g_r(r) g_\theta(\theta)$, then the Fourier transform of g can be expressed by the following infinite series of Hankel transforms:

$$F[g] = \sum_{k=-\infty}^{\infty} c_k(-j)^k e^{ik\phi} \mathrm{H}_k[g_r(r)]$$

where

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} g_\theta(\theta) e^{-ik\theta}\, d\theta$$

and

$$\mathrm{H}_k[g_r(r)] = 2\pi \int_0^\infty r g_r(r) J_k(2\pi r\rho)\, dr.$$

In these expressions ρ and ϕ are the coordinates in the transform space, and J_k is the Bessel function of the first kind of order k.

Applying this result to (22) yields

$$F[\mathrm{I}_s] = \operatorname{circ}\left(\frac{\rho R_o}{\gamma}\right) 2\pi \sum_{k=-\infty}^{\infty} c_k(-j)^k e^{ik\phi} \int_0^{R_o} r J_k(2\pi r\rho)\, dr \tag{23}$$ †

where

$$c_k = \begin{cases} \frac{1}{2}(\exp[j\gamma_1] + \exp[j\gamma_2]), & k = 0 \\ 0 & k \text{ even.} \\ \frac{1}{\pi jk}(\exp[j\gamma_2] - \exp[j\gamma_1]), & k \text{ odd} \end{cases} \tag{24}$$

† Here we have used the easily established result that

$$F\left[\frac{\gamma}{rR_o} J_1(2\pi\gamma r/R_o)\right] = \operatorname{circ}(\rho R_o/\gamma).$$

The appropriately normalized inverse transform of (23) is

$$\begin{aligned}\mathrm{I}_s &= \int_0^\infty \int_0^{2\pi} F[\mathrm{I}_s]\rho \exp\,[j2\pi\tilde{r}_i\rho \cos\,(\theta_i - \phi)]\, d\rho\, d\phi \\ &= 4\pi^2 \sum_{k=-\infty}^{\infty} c_k \exp\,[jk\theta_i] \int_0^{\gamma/R_o} \rho \int_0^{R_o} rJ_k(2\pi r\rho)\, drJ_k(2\pi\tilde{r}_i\rho)\, d\rho .\end{aligned}$$

Here θ_i is the angular coordinate in the image plane, and we have already introduced $\tilde{r}_i = r_i/M$ for later simplicity. Now, letting $t = 2\pi R_o\rho$, $s = r/R_o$, and substituting expressions (24) for the coefficients c_k finally gives

$$\mathrm{I}_s = \frac{1}{2}(\exp\,[j\gamma_1] + \exp\,[j\gamma_2])h_o(r_i) + \frac{2}{\pi}\,(\exp\,[j\gamma_2] - \exp\,[j\gamma_1]) \cdot \sum_{k\text{ odd}} h_k(r_i)\,\frac{\sin k\theta_i}{k} \tag{25}$$

where

$$h_k(r_i) = \int_0^{2\pi\gamma} t \int_0^1 sJ_k(st)\, dsJ_k\!\left(\frac{r_i t}{MR_o}\right) dt. \tag{26}$$

Note that the integral $h_o(r_i)$ is just the integral given in (11), and that the integrals $h_k(r_i)$ are higher order generalizations of it.

The intensity in the image plane is obtained from equations (10) and (25) as

$$\tilde{I}(r_i\,,\,\theta_i) = \begin{cases} |\,(ae^{j\alpha} - 1)\mathrm{I}_s + \exp\,[j\gamma_2]\,|^2, & 0 \leqq \theta_i \leqq \pi \\ & 0 \leqq r_i \leqq MR_o\,. \\ |\,(ae^{j\alpha} - 1)\mathrm{I}_s + \exp\,[j\gamma_1]\,|^2, & \pi < \theta_i < 2\pi \end{cases} \tag{27}$$

Clearly $\tilde{I}$ depends upon both the radial and the angular coordinates in the image plane while the phase distribution over the entrance aperture depends only upon θ. Therefore, one would not expect, in general, that the semicircles would be visible in the intensity pattern. We will now demonstrate, however, that if a small value of $z = 2\pi\gamma$ is used for which the background is essentially uniform, and if $\gamma_2 - \gamma_1$ is small, the semicircles can be reproduced, and the jump in the normalized intensity will be proportional to $\gamma_2 - \gamma_1$, at least to first order. Moreover, as in our previous analysis, if $a = 1$, $\alpha = \pi/2$, and $z = 1.5$, the first condition will be satisfied, and the proportionality factor will be approximately unity.

To observe these facts we need only to substitute the first term of the power series for the Bessel functions in (26) and conclude that for z sufficiently small

$$h_k(r_i) \approx \left(\frac{r_i}{MR_o}\right)^k \frac{z^{2k+2}}{(2^k k!)^2(k+2)(2k+2)}.$$

Thus for small z, the coefficients $h_k(r_i)$ are small, and they decrease extremely rapidly as k increases. Hence, if $\gamma_2 - \gamma_1$ is sufficiently small, the first term in (25) will dominate the second and I_s can be approximated by

$$\mathrm{I}_s \approx \tfrac{1}{2}(\exp[j\gamma_1] + \exp[j\gamma_2])h_o(r_i).$$

With this approximation the normalized intensity is approximately

$$\tilde{I}(r_i\,,\,\theta_i) \approx \begin{cases} \left|\,\tfrac{1}{2}(ae^{j\alpha} - 1)(\exp[j\gamma_1] + \exp[j\gamma_2])h_o(r_i) + \exp[j\gamma_2]\,\right|^2, \\ \qquad\qquad\qquad 0 \leqq \theta_i \leqq \pi \\ \qquad\qquad\qquad 0 \leqq r_i \leqq MR_o\,. \\ \left|\,\tfrac{1}{2}(ae^{j\alpha} - 1)(\exp[j\gamma_1] + \exp[j\gamma_2])h_o(r_i) + \exp[j\gamma_1]\,\right|^2, \\ \qquad\qquad\qquad \pi < \theta_i < 2\,\pi. \end{cases}$$

Now, if the background is essentially uniform, the integral $h_o(r_i)$ in the above expression is essentially constant, and we observe that, for all practical purposes, $\tilde{I}$ is constant in each of the semicircular regions $0 \leqq \theta_i \leqq \pi$, $\pi < \theta_i < 2\pi$. Furthermore, the jump in $\tilde{I}$ between these two regions is just

$$\begin{aligned} \Delta\tilde{I} &\approx 2\,\mathrm{Re}\,\{\tfrac{1}{2}(ae^{j\alpha} - 1)(\exp[-j\gamma_2] - \exp[-j\gamma_1])(\exp[j\gamma_1] \\ &\qquad + \exp[j\gamma_2])h_o(r_i)\} \\ &= 2a \sin\alpha\, h_o(r_i) \sin(\gamma_2 - \gamma_1) \\ &\approx 2a \sin\alpha\, h_o(r_i)(\gamma_2 - \gamma_1). \end{aligned}$$

From our previous work, we may recall that, for $z = 1.5$, $h_o(r_i) \approx 0.5$, and thus, for $a = 1$, $\alpha = \pi/2$, and $z = 1.5$,

$$\Delta\tilde{I} \approx (\gamma_2 - \gamma_1).$$

The same conditions that permitted radial discontinuities in the phase distribution to be qualitatively observed in the intensity pattern appear to permit a simple angular discontinuity such as (21) also to be observed. The representative results of numerical computations of the exact expressions (25), (26), (27) shown in Fig. 15 bear this out.

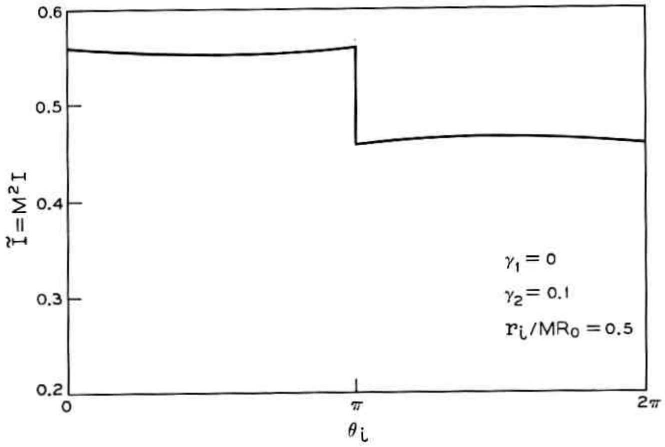


Fig. 15—Response to a semi-circular phase disk.

V. CONCLUSIONS

A general expression for the intensity distribution that is produced by a phase contrast imaging system with a circular phase plate was derived. It was shown that the results reduced to the well known expression for the phase contrast microscope when the size of the phase object is small compared to the area of the entrance aperture of the system. To obtain the intensity distribution for larger phase objects, a convolution integral must be evaluated by analytical and/or numerical techniques.

The intensity patterns that are produced by phase disks, phase rings, and semicircular phase distributions were derived, and the results of numerical computations that were based on these derivations were studied. In general, it appeared that qualitative reproductions of these simple phase distributions could be observed in the intensity patterns if the size of the phase spot on the phase plate was chosen to yield a uniform background, if the resulting parameter, z, was small, and if the magnitudes of the phase perturbations were small enough so that the approximation $e^{j\phi} \approx 1 + j\phi$ held. Moreover, it appeared that, if these conditions were not met, rather poor reproductions of even these simple phase distributions could result.

VI. ACKNOWLEDGMENTS

The author would like to thank J. Alan Cochran, E. R. Nagelberg, and N. J. Kolettis for helpful discussions on this research.

REFERENCES

1. Zernike, F., Z. Tech. Physik, *16* (1935), p. 454 ff.
2. Zernike, F., Physica *IX*, No. 7 (1942), p. 686 ff.
3. Zernike, F., Physica *IX*, No. 10 (1942), p. 974 ff.
4. Hopkins, H. H., Proc. Phys. Soc. B, *66* (1953), p. 331 ff.
5. Born, M., and Wolf, E., *Principles of Optics,* New York: Pergamon Press, 1965.
6. Hopkins, H. H., Rev. Opt., *31* (1952), p. 142 ff.
7. De, M., and Som, S. C., J. Opt. Soc. Amer., *53* (1963), p. 779 ff.
8. De, M., and Mondal, P. K., Optica Acta, *17* (1970), p. 397 ff.
9. Goodman, J. W., *Introduction to Fourier Optics,* New York: McGraw-Hill, 1959.
10. Mondal, P. K., and Slansky, S., Optica Acta, *17* (1970), p. 425 ff.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

A New Equalizer Structure for Fast Start-Up Digital Communication

By ROBERT W. CHANG

(Manuscript received November 18, 1970)

The use of transversal filters for automatic equalization has made possible high-speed data communication over voiceband telephone channels. Recently much attention has been focused on the possible use of the high-speed data sets in private line multiparty polling systems. However, for such applications, it is necessary to reduce the start-up time of the present automatic equalizer drastically. This paper examines the start-up time (settling time) of the transversal filter equalizer for two important classes of data communication systems: Class IV partial-response systems and single-sideband Nyquist systems. (The latter represents the limiting case of vestigial-sideband systems with small roll-off bandwidth.) It is shown that in single-sideband Nyquist systems the input signals to the gain controls of the transversal equalizer may be nearly orthonormal. Consequently the equalizers may have a short settling time. It is also shown that the equalizer settling time is much longer in Class IV partial-response systems, because such systems use controlled intersymbol interference and the input signals to the gain controls are highly correlated.

The possibility of reducing the settling time of the automatic equalizers is examined. A new equalizer structure is developed based on the following principles: (i) Equalizer settling time can be minimized by making the input signals to the gain controls orthonormal, and (ii) Such a minimization does not change the noise power, the mean-square equalization error, the convexity of the gain control adjustment, and the feedback control loops in the equalizer. These principles are general in that they apply regardless of the type of modulation—single- , vestigial- , or double-sideband (SSB, VSB, or DSB)—or the signaling scheme (Nyquist or partial-response). Application of these principles to Class IV partial-response systems is considered. For private line systems and systems where amplitude *distortions in the communication channels are not severe (delay distortions can be arbitrary), the new equalizer can be implemented by simply adding a prefixed weighting matrix to the conventional transversal equalizer. Analysis and computer simulation show that the use of such a new equalizer can result in a significant reduction in the system's start-up time.*

I. INTRODUCTION

In order to meet the needs of the rapidly growing computer and data processing industries, a number of high-speed data sets have been developed in recent years for voiceband telephone channels. Most of these data sets use transversal filters[1] for precise automatic equalization. The transversal equalizer consists of a tapped delay line with variable tap gains. During a start-up period prior to data transmission, the tap gains are adjusted automatically to minimize the peak distortion[1] or the mean-square error[2,3,4] of the received pulses. The time required to adjust the tap gains to nearly their optimum settings is usually called the settling time of the equalizer. Most automatic equalizers have settling times of a few seconds. The start-up time of the system can be longer because there are usually other operations to be performed in the start-up period (operations such as synchronization, carrier recovery, and so forth).

Recently, much attention has been focused on the possible use of the high-speed data sets in private-line multiparty polling systems (such as airline reservation systems, on-line banking systems, and so forth). Such systems are generally real-time information retrieval systems where the inquiry and response are short (the message lengths are usually less than 1000 bits[5]). With a 4800 b/s data set, it takes only 0.208 second to transmit a message of 1000 bits. Consequently the actual transmission time can be much less than the start-up time of the system (which can be five seconds). In order to allow additional stations to be served or to reduce the response time of the system (response time is important in real-time systems), it is necessary[5] to reduce the start-up time of the high-speed data sets drastically. This means that the settling time of the automatic equalizer must be reduced drastically (for example, from a few seconds to tens of milliseconds).

Such a drastic reduction in the start-up time of the high-speed data sets raises a number of theoretical questions. For example, it is no longer sufficient to just prove the convergence of the equalizer adjustment; it must be shown that the convergence is sufficiently fast to meet the requirement. It becomes necessary to examine the dependence of equalizer settling time on modulation and signaling schemes. Instead of considering synchronization, carrier recovery, and automatic equalization separately, one has to consider the mutual dependence of these adjustments and the possibility of minimizing the overall adjustment time. The start-up time requirement also provides a strong motivation to search for a new equalizer that has a settling time shorter than that of the conventional transversal equalizer. Answers to some of these problems are presented in this paper. Because the paper is lengthy, the contents of each section are outlined here; the results are summarized in Section VIII (the reader may read Section VIII first).

Section II includes a description of the mathematical model and reviews some of the fundamental works on automatic equalization (particularly those by Lucky and Gersho). Section III examines the equalizer settling time for two important classes of digital communication systems: the Class IV partial-response system and the SSB Nyquist system. (The latter represents the limiting case of VSB systems with small roll-off bandwidth.) Surprisingly, the results show that the equalizer settling times of these two systems can be very different. The reason for this difference (eigenvalue spread) is explained so that the method of analysis can be extended to other systems. Sections IV and V consider a general data communication system and develop a new equalizer structure for fast start-up purpose. These sections stress the underlying principle (condition of orthogonality) and analyze the various properties of the new equalizer (including convergence rate of equalizer adjustment, residual noise power, minimum mean-square error, and convexity of the adjustment). In Sections VI and VII, application of the new equalizer to the Class IV partial-response system is considered. Most importantly, it is shown that the new equalizer can be implemented by simply adding a prefixed weighting matrix to a conventional transversal equalizer. The related analytical studies and computer simulation are described; fast convergence of the equalizer adjustments is demonstrated. Section VIII is a summary of the results.

II. REVIEW OF FUNDAMENTALS

An amplitude modulation data communication system utilizing a transversal equalizer is depicted in Fig. 1. The equalizer consists of

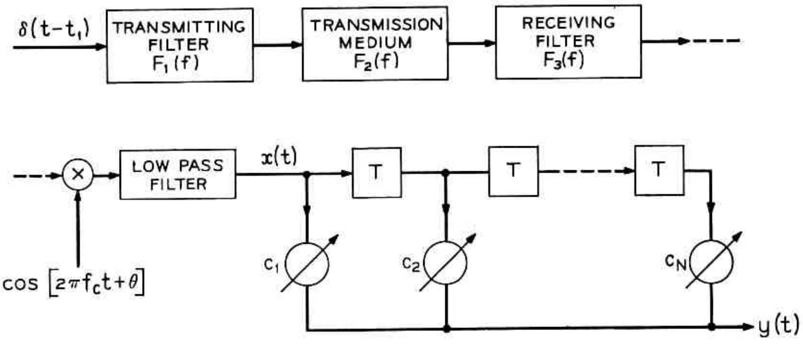


Fig. 1—Block diagram of an amplitude modulation data communication system with a conventional transversal equalizer.

a delay line tapped at T-second intervals, where T is the signaling interval of the system. The ith tap, $i = 1$ to N, is connected through a variable gain control c_i to a summing bus. During data transmission, the transmitter transmits the information digits sequentially at time-instants $t = \cdots, t_1 - T, t_1, t_1 + T, t_1 + 2T, \cdots$. The equalizer output is sampled sequentially at time-instants $t = \cdots, t_2 - T, t_2, t_2 + T, t_2 + 2T, \cdots$, and the time-samples are used to recover the information digits. To simplify the notations, we shall shift the origin of the time-axis to make $t_2 = 0$.

When an impulse $\delta(t - t_1)$ is applied at the transmitter input, the equalizer input and output are, respectively, $x(t)$ and $y(t)$. Since we consider only linear systems, $y(t)$ is the overall impulse response. The desired overall impulse response of the system is $d(t)$. In this study, we adopt the familiar mean-square error criterion[3,4,6] and adjust the gain controls of the equalizer to minimize the mean-square error between $y(t)$ and $d(t)$ at the receiver sampling instants $t = \cdots, -T, 0, T, 2T, \cdots$. The mean-square error can be written as

$$\epsilon = \sum_{i=-\infty}^{\infty} [y(iT) - d(iT)]^2. \tag{1}$$

It can be seen from Fig. 1 that

$$y(t) = \sum_{k=1}^{N} c_k x[t - (k - 1)T]. \tag{2}$$

For the sake of simplicity, we shall use the abbreviations $y_i = y(iT)$, $d_i = d(iT)$, and $x_i = x(iT)$. It can be seen from (2) that (1) can be written in the following matrix form

$$\epsilon = \mathbf{c}'\mathbf{A}\mathbf{c} - 2\mathbf{c}'\mathbf{v} + \sum_{i=-\infty}^{\infty} d_i^2 , \tag{3}$$

where

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix}, \tag{4}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}, \tag{5}$$

$$a_{ij} = \sum_{l=-\infty}^{\infty} x_{l-i+1}x_{l-j+1} , \quad \text{all} \quad i, j, \tag{6}$$

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}, \tag{7}$$

$$v_k = \sum_{i=-\infty}^{\infty} x_{i-k+1}d_i , \quad \text{all} \quad k. \tag{8}$$

It can be shown that **A** is positive definite.

Let $\partial\epsilon/\partial c_i$ be the partial derivative of ϵ with respect to c_i, $i = 1$ to N, and let $\partial\epsilon/\partial\mathbf{c}$ represent an $N \times 1$ column vector whose ith element is $\partial\epsilon/\partial c_i$; i.e.,

$$\frac{\partial\epsilon}{\partial\mathbf{c}} = \begin{bmatrix} \dfrac{\partial\epsilon}{\partial c_1} \\ \dfrac{\partial\epsilon}{\partial c_2} \\ \vdots \\ \dfrac{\partial\epsilon}{\partial c_N} \end{bmatrix}.$$

From (1), we obtain

$$\frac{\partial \epsilon}{\partial \mathbf{c}} = 2\mathbf{A}\mathbf{c} - 2\mathbf{v}. \tag{9}$$

The measured $\partial\epsilon/\partial\mathbf{c}$ in actual data sets will deviate[4] from that in (9) due to noise in the receiver. However this deviation is negligible in high signal-to-noise ratio systems. For example, it has been found (Section 7.3) that at a 30-dB signal-to-noise ratio this deviation has only a minor effect on the settling time of the equalizer. In this paper we consider high signal-to-noise ratio systems and neglect such a deviation.

The optimum value of $\mathbf{c}$ that minimizes the mean-square error ϵ will be called $\mathbf{c}_{\text{opt}}$. It is clear that $\mathbf{c}$ minimizes ϵ if and only if $\partial\epsilon/\partial\mathbf{c} = \mathbf{0}$; therefore, from (10) the optimum $\mathbf{c}$ is

$$\mathbf{c}_{\text{opt}} = \mathbf{A}^{-1}\mathbf{v}. \tag{10}$$

The difference between $\mathbf{c}$ and $\mathbf{c}_{\text{opt}}$ is denoted by $\mathbf{e}$: i.e.,

$$\mathbf{e} = \mathbf{c} - \mathbf{A}^{-1}\mathbf{v}. \tag{11}$$

Let $\epsilon_{\min}$ be the minimum value of ϵ when $\mathbf{c} = \mathbf{c}_{\text{opt}}$. From (3) and (10)

$$\epsilon_{\min} = \sum_{i=-\infty}^{\infty} d_i^2 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{v}. \tag{12}$$

Now consider the adjustment of the equalizer. As is well known,[3,7] the equalizer can be adjusted in the training period prior to data transmission by transmitting either a succession of isolated test pulses or a sequence of Pseudo-random numbers. Since these two methods differ considerably, we shall consider only the first method (isolated test pulses) in this paper.

In the training period, isolated impulses are applied to the transmitter input. For instance, $\delta(t - t_1)$ in Fig. 1 may be the kth such impulse. The transmission of $\delta(t - t_1)$ produces the test pulse $x(t)$ at the equalizer input. From $x(t)$ the partial derivatives $\partial\epsilon/\partial c_i$, $i = 1$ to N, are computed[3,4,6] and the gain control c_i is changed by an amount proportional to $\partial\epsilon/\partial c_i$. This process is then repeated for the next test pulse.

The adjustment made after the kth test pulse, $k = 1, 2, 3, \cdots$, will be referred to as the kth adjustment. The initial values of $\mathbf{c}$, $\mathbf{e}$, and ϵ (i.e., their values prior to the first adjustment) will be denoted, respectively, by $\mathbf{c}_0$, $\mathbf{e}_0$, and ϵ_0. The values of $\mathbf{c}$, $\mathbf{e}$, and ϵ after the

kth adjustment are denoted, respectively, by $\mathbf{c}_k$, $\mathbf{e}_k$, and ϵ_k. From (11)

$$\mathbf{e}_k = \mathbf{c}_k - \mathbf{A}^{-1}\mathbf{v} \qquad k = 0, 1, 2 \cdots . \tag{13}$$

It can be easily shown from (3), (12), and (13) that

$$\epsilon_k = \epsilon_{\min} + \mathbf{e}_k'\mathbf{A}\mathbf{e}_k, \quad k = 0, 1, 2, \cdots . \tag{14}$$

The kth adjustment is made according to the equation

$$\mathbf{c}_k = \mathbf{c}_{k-1} - \frac{1}{2}\alpha_k\left[\frac{\partial\epsilon}{\partial\mathbf{c}}\right]_k, \tag{15}$$

where α_k, $k = 1, 2, 3, \cdots$, are suitably chosen constants.[4]

The subscript k of $\partial\epsilon/\partial\mathbf{c}$ indicates that the $\partial\epsilon/\partial\mathbf{c}$ is computed from the kth test pulse. It can be shown from (9), (15), and (13) that

$$\mathbf{e}_k = (\mathbf{I} - \alpha_k\mathbf{A})\mathbf{e}_{k-1}, \quad k = 1, 2, 3, \cdots . \tag{16}$$

We now proceed to study the convergence of the mean-square error for several data communication systems.

III. A STUDY OF CONVERGENCE FOR PARTIAL-RESPONSE AND NYQUIST SYSTEMS

We have defined ϵ_k as the value of ϵ after the kth adjustment. Note from (14) that ϵ_k consists of two terms. The first term $\epsilon_{\min}$ is the irreducible value of ϵ. Only the second term $\mathbf{e}_k'\mathbf{A}\mathbf{e}_k$ depends on $\mathbf{c}$ and the adjustments. Thus, we shall study the convergence of the second term in this section.

It can be seen from (6) that $\mathbf{A}$ is a symmetric matrix. Let the eigenvalues of $\mathbf{A}$ be denoted, in the order of increasing magnitude, by λ_i, $i = 1$ to N, so that

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_N, \tag{17}$$

and let $\mathbf{u}_i$, $i = 1$ to N, be a set of orthonormal eigenvectors of $\mathbf{A}$ ($\mathbf{u}_i$ is the eigenvector corresponding to λ_i). It is well known that $\mathbf{A}$ can be represented in the form

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}', \tag{18}$$

where $\mathbf{D}$ is an $N \times N$ diagonal matrix whose ith diagonal element is λ_i, and $\mathbf{Q}$ is an $N \times N$ matrix whose ith column is the eigenvector $\mathbf{u}_i$. It is also well known that Q is an orthogonal matrix; i.e.,

$$\mathbf{Q}' = \mathbf{Q}^{-1}. \tag{19}$$

From (18) and (19),

$$\mathbf{I} - \alpha_k \mathbf{A} = \mathbf{Q}[\mathbf{I} - \alpha_k \mathbf{D}]\mathbf{Q}'. \tag{20}$$

By repeated application of (16), one obtains

$$\begin{aligned}\mathbf{e}_k &= (\mathbf{I} - \alpha_k \mathbf{A})(\mathbf{I} - \alpha_{k-1}\mathbf{A}) \cdots (\mathbf{I} - \alpha_2 \mathbf{A})(\mathbf{I} - \alpha_1 \mathbf{A})\mathbf{e}_0 \\ &= \prod_{n=1}^{k} (\mathbf{I} - \alpha_n \mathbf{A})\mathbf{e}_0 , \qquad k = 1, 2, 3, \cdots . \end{aligned} \tag{21}$$

Substituting (20) into (21) and noting that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ gives

$$\mathbf{e}_k = \mathbf{Q}\left[\prod_{n=1}^{k} (\mathbf{I} - \alpha_n \mathbf{D})\right]\mathbf{Q}'\mathbf{e}_0 , \qquad k = 1, 2, 3, \cdots . \tag{22}$$

From (22) and (18),

$$\mathbf{e}_k'\mathbf{A}\mathbf{e}_k = \mathbf{e}_0'\mathbf{Q}\left[\prod_{n=1}^{k} (\mathbf{I} - \alpha_n \mathbf{D})\right]\mathbf{D}\left[\prod_{n=1}^{k} (\mathbf{I} - \alpha_n \mathbf{D})\right]\mathbf{Q}'\mathbf{e}_0 , \qquad k = 1, 2, 3, \cdots . \tag{23}$$

Using the properties of **D** and **Q** described one can carry out the matrix multiplications in (23) and obtain

$$\mathbf{e}_k'\mathbf{A}\mathbf{e}_k = \sum_{i=1}^{N} \xi_i(k), \qquad k = 1, 2, 3, \cdots . \tag{24}$$

where

$$\xi_i(k) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i\left[\prod_{n=1}^{k} (1 - \alpha_n\lambda_i)^2\right], \qquad i = 1 \text{ to } N. \tag{25}$$

In a similar manner, we find that

$$\mathbf{e}_0'\mathbf{A}\mathbf{e}_0 = \sum_{i=1}^{N} \xi_i(0), \tag{26}$$

where

$$\xi_i(0) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i , \qquad i = 1 \text{ to } N. \tag{27}$$

Since **A** is positive definite, $\lambda_i > 0$ for all i. Thus, $\xi_i(k) \geqq 0$ for all i. Consequently, $\mathbf{e}_k'\mathbf{A}\mathbf{e}_k$ converges to zero if and only if $\xi_i(k)$, $i = 1$ to N, all converge to zero (see (24)). For this reason, $\xi_i(k)$ will be called the ith error component. If the error components all converge rapidly to zero as k increases, $\mathbf{e}_k'\mathbf{A}\mathbf{e}_k$ converges rapidly to zero.

It is clear from (25) that $\xi_i(k)$ converges to zero if and only if the factor

$$\left[\prod_{n=1}^{k} (1 - \alpha_n \lambda_i)^2\right]$$

converges to zero. Thus, the convergence of $\xi_1(k)$ *to* $\xi_N(k)$ depend only on two sets of parameters: $\alpha_1, \cdots, \alpha_k$ and $\lambda_1, \cdots, \lambda_N$. The first set of parameters corresponds to the magnitudes of the gain-control adjustments [As can be seen from (15), α_k determines the magnitude of the kth adjustment.]. In the following subsections, we show that the second set of parameters, $\lambda_1, \cdots, \lambda_N$, depend on the modulation scheme and channel characteristics. In some systems, $\lambda_1, \cdots, \lambda_N$ differ only slightly in value. Consequently α_k can be selected such that each adjustment reduces each of the error components by a large factor (such as 100). However, this is not possible in some other systems.

3.1 *Class IV Partial-Response System*

There are several classes of partial-response systems.[8] We shall consider the most important one: Class IV partial response system.[9,10,11,12,13] The results can be easily extended to other classes.

As depicted in Fig. 1, the transfer functions of the transmitting filter, transmission medium, and receiving filter are, respectively, $F_1(f)$, $F_2(f)$, and $F_3(f)$. The amplitude and phase characteristics of $F_i(f)$ will be denoted, respectively, by $|F_i(f)|$ and $\beta_i(f)$, i.e.,

$$F_i(f) = |F_i(f)| e^{J\beta_i(f)}, \qquad i = 1, 2, 3. \tag{28}$$

Note that in this paper J will be used to denote the imaginary number $\sqrt{-1}$ (j is used as an index).

In a Class IV partial-response system, the transmitting and the receiving filters are band-limited; that is,

$$|F_1(f)F_3(f)| = 0,$$

when

$$|f| \leqq f_1 \quad \text{and} \quad |f| \geqq f_2, \tag{29}$$

where f_1 and f_2 are, respectively, the lower and the upper cutoff frequencies. The demodulating carrier frequency, f_c, is usually equal to f_2. This implies that the system is SSB. According to the previous definition, when an impulse $\delta(t - t_1)$ is applied at the transmitter input, the equalizer input is $x(t)$. Let $X(f)$ denote the Fourier transform of $x(t)$. It can be shown from (28), (29), and the demodulation process that

$$
\begin{aligned}
X(f) &= \tfrac{1}{2}\,|F_1(f - f_c)F_2(f - f_c)F_3(f - f_c)| \\
&\quad \cdot \exp\{J[\beta_1(f - f_c) + \beta_2(f - f_c) + \beta_3(f - f_c) - 2\pi(f - f_c)t_1 + \theta]\}, \\
&\qquad\qquad 0 \leqq f \leqq f_2 - f_1 \\
&= \tfrac{1}{2}\,|F_1(f + f_c)F_2(f + f_c)F_3(f + f_c)| \\
&\quad \cdot \exp\{J[\beta_1(f + f_c) + \beta_2(f + f_c) + \beta_3(f + f_c) - 2\pi(f + f_c)t_1 - \theta]\}, \\
&\qquad\qquad -(f_2 - f_1) \leqq f \leqq 0 \\
&= 0, \quad \text{all other frequencies.}
\end{aligned} \tag{30}
$$

Now we can determine the **A** matrix. From (6), the elements of **A** are

$$a_{ij} = \sum_{l=-\infty}^{\infty} x_{l-i+1}x_{l-j+1} \,. \tag{6}$$

In a Class IV partial-response system the signaling interval is

$$T = \frac{1}{2(f_2 - f_1)}. \tag{31}$$

It is clear from (30) that $x(t)$ is band-limited from 0 to $(f_2 - f_1)$ Hz. Thus, the time-samples $x(kT)$, $k = \cdots, 0, 1, 2, \cdots$, are taken at the Nyquist rate. Therefore, according to the sampling theorem

$$\int_{-\infty}^{\infty} x[t - iT + T]x[t - jT + T]\,dt = \frac{1}{2(f_2 - f_1)} \sum_{l=-\infty}^{\infty} x_{l-i+1}x_{l-j+1} \,. \tag{32}$$

Comparing (6) and (32) shows that

$$a_{ij} = 2(f_2 - f_1)\int_{-\infty}^{\infty} x[t - iT + T]x[t - jT + T]\,dt. \tag{33}$$

From Parseval's theorem and (30), one can rewrite (33) as

$$
\begin{aligned}
a_{ij} = (f_2 - f_1)\int_0^{f_2-f_1} &[\cos 2\pi f(i - j)T] \\
&\cdot [|F_1(f - f_c)F_2(f - f_c)F_3(f - f_c)|]^2\,df.
\end{aligned} \tag{34}
$$

It can be seen from (34) that a_{ij} is independent of the following parameters: demodulating carrier phase θ, system timing t_1, phase characteristics of the transmitting and receiving filters, and phase characteristic of the transmission medium. To evaluate a_{ij}, we need to specify only the amplitude characteristics $|F_1(f - f_c)|$, $|F_2(f - f_c)|$, and $|F_3(f - f_c)|$. In a Class IV partial-response system, the transmitting and receiving filters are designed such that

$$
\begin{aligned}
|F_1(f)|\,|F_3(f)| &= \sin \pi\left[\frac{f-f_1}{f_2-f_1}\right], \qquad f_1 \leqq f \leqq f_2 \\
&= \sin \pi\left[\frac{-f-f_1}{f_2-f_1}\right], \qquad -f_2 \leqq f \leqq -f_1\,. \qquad (35)
\end{aligned}
$$

Substituting (35) into (34) gives

$$
\begin{aligned}
a_{ij} = (f_2 - f_1) \int_0^{f_2-f_1} & [\cos 2\pi f(i-j)T] \\
& \cdot\left[\left(\sin \pi \frac{f}{f_2-f_1}\right) |F_2(f-f_c)|\right]^2 df. \qquad (36)
\end{aligned}
$$

We first consider the case where the transmission medium has a constant amplitude characteristic

$$|F_2(f)| = 1 \qquad (37)$$

in the pass band $f_1 \leqq f \leqq f_2$. Substituting (37) into (36) and evaluating the integral, we obtain

$$
\begin{aligned}
a_{ij} &= -\frac{(f_2-f_1)^2}{4}, \qquad i-j = -2 \quad \text{or} \quad 2 \\
&= \frac{(f_2-f_1)^2}{2}, \qquad i-j = 0 \\
&= 0, \qquad \text{all other} \quad i-j. \qquad (38)
\end{aligned}
$$

The constant term $(f_2 - f_1)^2/2$ in a_{ii} may be dropped (this corresponds to introducing a gain of $\sqrt{2}/(f_2 - f_1)$ in the channel). Then (38) becomes

$$
\begin{aligned}
a_{ij} &= -\tfrac{1}{2}, \qquad i-j = -2 \quad \text{or} \quad 2 \\
&= 1, \qquad i-j = 0 \\
&= 0, \qquad \text{all other} \quad i-j. \qquad (39)
\end{aligned}
$$

Thus, the matrix **A** is of the following form

$$
\mathbf{A} = \begin{bmatrix}
1 & 0 & -\frac{1}{2} & 0 & \cdots & 0 \\
0 & 1 & 0 & -\frac{1}{2} & \cdots & 0 \\
-\frac{1}{2} & 0 & 1 & 0 & \cdots & 0 \\
0 & -\frac{1}{2} & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1
\end{bmatrix}. \qquad (40)
$$

In particular the main diagonal elements are 1, the second off-diagonal elements are $-1/2$, and the other elements are all 0 (notice that **A** belongs to a class of matrixes known as the "Toeplitz"[14]). We have defined λ_1 to λ_N as the eigenvalues of **A**, and $\mathbf{u}_1$ to $\mathbf{u}_N$ as a set of orthonormal eigenvectors of **A**. These eigenvalues and eigenvectors are determined in Appendix A. Since transversal equalizers usually have an odd number of taps, we shall assume that N is odd in the following discussion. From Appendix A, the N eigenvalues of **A** are

$$1 - \cos \frac{k\pi}{\frac{N+1}{2}+1}, \qquad k = 1, 2, \cdots, \frac{N+1}{2}, \tag{41}$$

and

$$1 - \cos \frac{k\pi}{\frac{N-1}{2}+1}, \qquad k = 1, 2, \cdots, \frac{N-1}{2}. \tag{42}$$

Thus, the eigenvalues of **A** are points on the curve $1 - \cos \theta$, $0 < \theta < \pi$ (see Fig. 2). The minimum eigenvalue

$$1 - \cos \frac{\pi}{\frac{N+1}{2}+1}$$

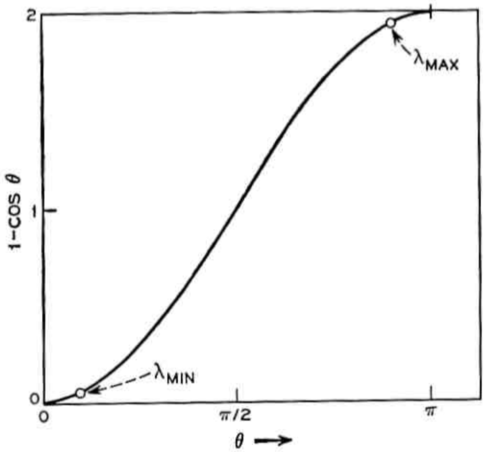


Fig. 2—Distribution of the eigenvalues of $\mathbf{A}(\lambda_{\min}$ and $\lambda_{\max}$ shown are for $N = 17)$.

and the maximum eigenvalue

$$1 - \cos \frac{\frac{N+1}{2}\pi}{\frac{N+1}{2} + 1}$$

are also shown in Fig. 2. Other eigenvalues lie between these two points. It can be seen from this figure that, for typical values of N (13, 17, etc.), the minimum eigenvalue is very close to 0 and the maximum eigenvalue is very close to 2. For example, consider a transversal equalizer with 17 taps ($N = 17$). From (41) and (42), the 17 eigenvalues of $\mathbf{A}$ are 0.0489, 0.0603, 0.1910, 0.2340, 0.4122, 0.5, 0.6910, 0.8263, 1, 1.174, 1.309, 1.5, 1.588, 1.766, 1.809, 1.940, and 1.951.

When the amplitude characteristic of the transmission medium is not exactly a constant in the passband $f_1 \leqq f \leqq f_2$, the calculations from (38) to (42) will change slightly. Consequently the eigenvalues will differ slightly from those above.

Having determined the distribution of the eigenvalues, we now consider the convergence of the error components $\xi_i(k)$, $i = 1$ to N. Prior to the kth adjustment, the ith error component is

$$\xi_i(k-1) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i\left[\prod_{n=1}^{k-1}(1 - \alpha_n\lambda_i)^2\right]. \tag{43}$$

After the kth adjustment, the ith error component is

$$\xi_i(k) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i\left[\prod_{n=1}^{k}(1 - \alpha_n\lambda_i)^2\right].$$

Substituting (43) into tne above equation gives

$$\xi_i(k) = (1 - \alpha_k\lambda_i)^4\,\xi_i(k-1), \quad k = 1, 2, 3, \cdots. \tag{44}$$

It is easily seen from (44) that the kth adjustment will reduce each of the error components by a large factor if

$$(1 - \alpha_k\lambda_i)^2 \ll 1, \qquad \text{for all } i$$

or if

$$\alpha_k \cong \frac{1}{\lambda_i}, \qquad \text{for all } \quad i. \tag{45}$$

Unfortunately, α_k cannot satisfy (45) because $\lambda_1, \cdots, \lambda_N$ are very different in value. Therefore, the kth adjustment cannot reduce each of the error-components by a large factor. To illustrate this quantitatively, consider again the 17-tap transversal equalizer ($\lambda_1 = 0.0489$ and $\lambda_{17} = 1.951$). In order to reduce the first error-component $\xi_1(k)$, we set $(1 - \alpha_k\lambda_1)^2 \ll 1$ or $\alpha_k \cong 1/\lambda_1 = 20.4$. However, this α_k will cause a large increase in some other error-components. For example, the 17th error-component will increase about 1505 times because $(1 - \alpha_k\lambda_{17})^2 = (1 - 20.4 \times 1.951)^2 = 1505$. On the other hand, if we wish to reduce the 17th error-component, we should set $(1 - \alpha_k\lambda_{17})^2 \ll 1$, or $\alpha_k \cong 1/\lambda_{17} = 0.512$. However, this α_k is too small to reduce some other error-components rapidly. For instance, the first error-component will reduce only five percent because $(1 - \alpha_k\lambda_1)^2 = (1 - 0.512 \times 0.0489)^2 = 0.95$. These examples clearly demonstrate that the kth adjustment cannot reduce each of the error-components by a large factor.

We now summarize the results in this subsection:

(*i*) In a Class IV partial-response system, the eigenvalues λ_1 to λ_N of the **A** matrix depend only on the amplitude characteristic of the channel, but not on carrier phase, system timing, and phase characteristic of the channel.

(*ii*) These eigenvalues are very different in value. For typical values of N (13, 17, etc.), the minimum eigenvalue λ_1 is very close to 0, while the maximum eigenvalue λ_N is very close to 2.

(*iii*) Because of the large differences in the eigenvalues, each adjustment of the gain controls cannot reduce each of the error-components by a large factor.

3.2 *SSB Nyquist Systems*

In this section, we study SSB data communication systems which transmit at the Nyquist rate with sin x/x pulses (hereafter referred to as SSB Nyquist systems). Such a system has the same configuration as the Class IV partial-response system except that its transmitting and receiving filters have constant amplitude characteristics in the passband. The main advantage of an SSB Nyquist system is that it transmits at the maximum possible bauds (Nyquist rate) without noise penalty. Saltzberg[15] has shown that this signaling method is not as sensitive to timing error as is commonly believed. In fact we show in this section that this scheme exhibits fast start-up advantages.

The transmitting and receiving filters are specified by

$$\begin{aligned} |F_1(f)| &= 1, \qquad f_1 \leqq |f| \leqq f_2 \\ &= 0, \qquad \text{other } f. \end{aligned} \tag{46}$$

and

$$|F_3(f)| = 1, \qquad f_1 \leqq |f| \leqq f_2$$
$$= 0, \qquad \text{other} \quad f. \tag{47}$$

The signaling interval T is again given by (31), and a_{ij} by (34). Substituting (46) and (47) into (34), we have

$$a_{ij} = (f_2 - f_1) \int_0^{f_2 - f_1} [\cos 2\pi f(i - j)T] \, |F_2(f - f_c)|^2 \, df. \tag{48}$$

Note again that a_{ij} (and hence the eigenvalues λ_1 to λ_N) depends only on the amplitude characteristics of the transmission medium and the filters. When the transmission medium has a constant amplitude characteristic

$$| F_2(f) | = 1$$

in the passband $f_1 \leqq f \leqq f_2$, (48) yields

$$a_{ij} = (f_2 - f_1)^2, \qquad i - j = 0$$
$$= 0, \qquad i - j \neq 0. \tag{49}$$

Neglecting the constant $(f_2 - f_1)^2$ above, we see that $\mathbf{A}$ is simply the identifying matrix and the eigenvalues of $\mathbf{A}$ are

$$\lambda_i = 1, \qquad i = 1 \text{ to } N. \tag{50}$$

Now consider the convergence of the error components $\xi_i(k)$, $i = 1$ to N. Prior to the first adjustment, the ith error component is

$$\xi_i(0) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i \,. \tag{51}$$

After the first adjustment, the ith error component is

$$\xi_i(1) = (\mathbf{e}_0'\mathbf{u}_i)^2\lambda_i(1 - \alpha_1\lambda_i)^2. \tag{52}$$

Substituting (50) and (51) into (52) yields

$$\xi_i(1) = (1 - \alpha_1)^2\xi_i(0). \tag{53}$$

If we set $\alpha_1 = 1$, $\xi_i(1) = 0$ for all i. In other words, when $\alpha_1 = 1$, the first adjustment reduces all the error-components to zero. Consequently the mean-square error ϵ is reduced to its irreducible value $\epsilon_{\min}$ after only one adjustment. This fastest possible convergence is obtained regardless of the initial equalizer settings, the carrier phase and system timing, the phase characteristic of the transmission medium, and the phase characteristics of the transmitting and receiving filters (because the eigenvalues λ_1 to λ_N are independent of all these parameters).

From (48), a_{ij} can be computed for any given $|F_2(f)|$. When $|F_2(f)|$ is not exactly a constant in the passband, **A** will not be an identity matrix and the eigenvalues will not be all equal. However, fast convergence is obtainable as long as the differences between the eigenvalues are small. For example, consider the case where the minimum eigenvalue is 0.9 and the maximum eigenvalue is 1.1. If we use $\alpha_1 = 1$, the maximum value of $(1 - \alpha_1\lambda_i)^2$ will be $(1 - 0.9)^2 = 0.01$. Consequently the first adjustment will reduce each of the error components by at least a factor of 100 (as will each following adjustment). The equalizer adjustment therefore will be completed after only a few adjustments.

IV. A NEW EQUALIZER STRUCTURE

In the preceding sections, we have clearly shown that, when the eigenvalues of **A** have close magnitudes, each adjustment of the equalizer reduces each of the error-components by a large factor (such as 100). We have also shown that such a fast convergence is not possible when the eigenvalues are very different in magnitude (such as in the case of a Class IV partial-response system). These results suggest that in order to improve the convergence rate we should attempt to reduce the differences between the eigenvalues.

Now we ask: *What causes the eigenvalues to be different? Is it possible to reduce such differences (and hence increase the convergence rate) by changing the equalizer structure? Does the use of the new equalizer structure alter the system's performance otherwise?* We shall consider the first two questions in this section, and derive a new equalizer structure. The last question will be considered in the next section. Application of the results to a Class IV partial-response system and the various related problems will be studied in Sections VI and VII.

It may appear that the new equalizer structure derived in this section requires complicated mathematical operations (computation of eigenvalues and eigenvectors of a matrix). However, it will be shown in Section VI that such mathematical operations can be completely eliminated for the systems of interest here.

Now consider the first question: What causes the eigenvalues to be different? From the definition in (6)

$$a_{ij} = \sum_{l=-\infty}^{\infty} x[(l-i+1)T]x[(l-j+1)T].$$

Since the input to the ith gain control is $x[t-(i-1)T]$, $x[(l-i+1)T]$ are time-samples of the input to the ith gain control, while $x[(l-j+1)T]$ are time-samples of the input to the jth gain control. Therefore, a_{ij} is simply the cross-correlation between the inputs to the ith and the jth gain control, and $\mathbf{A}$ has the interpretation of a correlation matrix. Therefore, we might state that it is the correlations between the inputs to the gain controls that result in the differences between the eigenvalues. When the inputs to the gain controls are orthonormal, $\mathbf{A}$ is an identity matrix and the eigenvalues are equal. When the inputs to the gain controls are correlated, the eigenvalues are different.

The above discussion suggest that, if we could construct a new equalizer such that the inputs to the gain controls are orthonormal, the eigenvalues of the correlation matrix of the new equalizer might be all equal and it might be possible to minimize the mean-square error in a single adjustment. Following this line of thinking, we obtain the generalized equalizer structure depicted in Fig. 3. As shown in Fig. 3, the equalizer input $x(t)$ is connected to a bank of filters. The output of the ith filter is connected through a variable gain control c_i to the summing bus. The equalizer output is

$$y(t) = \sum_{i=1}^{N} c_i z_i(t), \tag{54}$$

where $z_i(t)$ is the input to the ith gain control c_i. These signals $z_i(t)$, $i = 1$ to N, are orthonormal when

$$\sum_{l=-\infty}^{\infty} z_i(lT)z_j(lT) = \delta_{ij} . \tag{55}$$

In the following, we show that, for any given channel, the filters in Fig. 3 can be designed to satisfy (55). We also show that, when (55) is satisfied, the mean-square error can be minimized by a single adjustment of the gain controls.

Because of the press toward digitalization and integrated circuits, we shall specify the filters in Fig. 3 directly in digital filter form (instead of specifying them in analog filter form and then approximating them with digital filters). We shall use nonrecursive digital

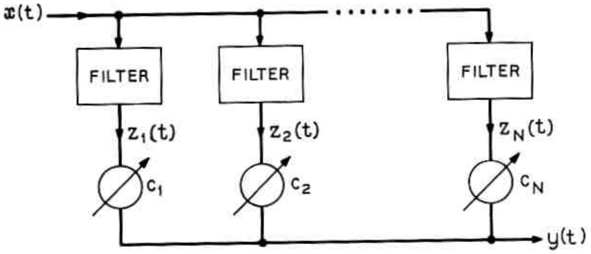


Fig. 3—Generalized equalizer structure.

filters[16] because they do not have stability problems and can be implemented most easily.

A realization of the generalized equalizer using nonrecursive digital filters is depicted in Fig. 4. The N digital filters share the same tapped delay line. The ith digital filter, $i = 1$ to N, consists of the tapped delay line and the N coefficients P_{i1} to P_{iN}. Its output is

$$z_i(t) = \sum_{j=1}^{N} P_{ij}x[t - (j - 1)T], \qquad i = 1 \quad \text{to} \quad N. \tag{56}$$

The gain controls c_1 to C_N are again adjusted to minimize the mean-square error ϵ defined in (1). From (54) and (56), we can rewrite (1) into the following matrix form

$$\epsilon = \mathbf{c}'\mathbf{PAP}'\mathbf{c} - 2\mathbf{c}'\mathbf{Pv} + \sum_{k=-\infty}^{\infty} d_k^2 \tag{57}$$

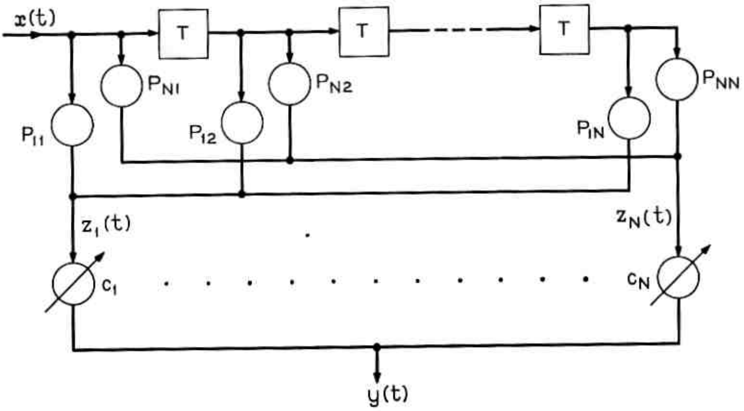


Fig. 4—A new equalizer structure (a realization of the generalized equalizer structure using nonrecursive digital filters).

where $\mathbf{c}$, $\mathbf{A}$, and $\mathbf{v}$ have been defined previously [see (4), (5), and (7)]. The matrix $\mathbf{P}$ is defined by

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{21} & P_{22} & \cdots & P_{2N} \\ \vdots & \vdots & & \vdots \\ P_{N1} & P_{N2} & \cdots & P_{NN} \end{bmatrix}. \tag{58}$$

As in Section II, we define $\partial\epsilon/\partial\mathbf{c}$ as the $N \times 1$ column vector whose ith element is $\partial\epsilon/\partial c_i$. It can be shown from (57) that

$$\frac{\partial\epsilon}{\partial\mathbf{c}} = 2\mathbf{PAP'c} - 2\mathbf{Pv}. \tag{59}$$

It can also be shown from (57) that $\mathbf{c}$ minimizes ϵ if and only if $\partial\epsilon/\partial\mathbf{c} = \mathbf{0}$; therefore, from (59) the optimum value of $\mathbf{c}$, $\mathbf{c}_{\text{opt}}$, is

$$\mathbf{c}_{\text{opt}} = (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v}. \tag{60}$$

We again define $\mathbf{e}$ as the difference between $\mathbf{c}$ and $\mathbf{c}_{\text{opt}}$; hence, from (60)

$$\mathbf{e} = \mathbf{c} - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v}. \tag{61}$$

As in Section II, the initial values of $\mathbf{c}$, $\mathbf{e}$, and ϵ are denoted, respectively, by $\mathbf{c}_0$, $\mathbf{e}_0$, and ϵ_0. The values of $\mathbf{c}$, $\mathbf{e}$, and ϵ after the kth adjustment are denoted, respectively, by $\mathbf{c}_k$, $\mathbf{e}_k$, and ϵ_k. From (61)

$$\mathbf{e}_k = \mathbf{c}_k - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v}, \qquad k = 0, 1, 2, \cdots. \tag{62}$$

It can be shown from (57) and (62) that

$$\epsilon_k = \sum_{k=-\infty}^{\infty} d_k^2 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{v} + \mathbf{e}_k'\mathbf{PAP'e}_k \qquad k = 0, 1, 2, \cdots. \tag{63}$$

The kth adjustment, $k = 1, 2, 3, \cdots$, of the gain control is again made according to (15). From (62), (15), and (59),

$$\begin{aligned}\mathbf{e}_k &= \mathbf{c}_{k-1} - \frac{1}{2}\alpha_k\left[\frac{\partial\epsilon}{\partial\mathbf{c}}\right]_k - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v} \\ &= \mathbf{c}_{k-1} - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v} - \alpha_k[\mathbf{PAP'c}_{k-1} - \mathbf{Pv}].\end{aligned} \tag{64}$$

From (62), we have

$$\mathbf{e}_{k-1} = \mathbf{c}_{k-1} - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v}. \tag{65}$$

Using (65), we can rewrite (64) as

$$\begin{aligned}\mathbf{e}_k &= \mathbf{e}_{k-1} - \alpha_k[\mathbf{PAP}'\mathbf{c}_{k-1} - \mathbf{Pv}] \\ &= \mathbf{e}_{k-1} - \alpha_k\mathbf{PAP}'[\mathbf{c}_{k-1} - (\mathbf{P}')^{-1}\mathbf{A}^{-1}\mathbf{v}] \\ &= [\mathbf{I} - \alpha_k\mathbf{PAP}']\mathbf{e}_{k-1}\ , \\ &\qquad\qquad k = 1, 2, \cdots . \end{aligned} \tag{66}$$

Now we can prove the first statement after (55) [that is, for any given channel, **P** can be chosen to satisfy (55)]. It is easily shown from (58), (5), (6), and (56) that **PAP′** can be written in the form

$$\mathbf{PAP}' = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1N} \\ b_{21} & b_{22} & \cdots & b_{2N} \\ \vdots & \vdots & & \vdots \\ b_{N1} & b_{N2} & \cdots & b_{NN} \end{bmatrix}. \tag{67}$$

where

$$b_{ij} = \sum_{l=-\infty}^{\infty} z_i(lT)z_j(lT). \tag{68}$$

Equation (55) is therefore equivalent to $b_{ij} = \delta_{ij}$, or

$$\mathbf{PAP}' = \mathbf{I}. \tag{69}$$

From (18),

$$\mathbf{A} = \mathbf{QDQ}'. \tag{18}$$

Let **H** be an $N \times N$ diagonal matrix whose ith diagonal element is $\sqrt{\lambda_i}$, that is,

$$\mathbf{H} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_N} \end{bmatrix}. \tag{70}$$

Then $\mathbf{D} = \mathbf{HH}'$ and (18) becomes

$$\mathbf{A} = \mathbf{QHH}'\mathbf{Q}'. \tag{71}$$

Substituting (71) into (69) yields

$$\mathbf{PQH}(\mathbf{PQH})' = \mathbf{I}. \tag{72}$$

Equation (72) holds if and only if **PQH** is an orthogonal matrix **G**.

From $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, we can rewrite

$$\mathbf{PQH} = \mathbf{G} \tag{73}$$

as

$$\mathbf{P} = \mathbf{GH}^{-1}\mathbf{Q}'. \tag{74}$$

For any given channel [i.e., for any given $x(t)$], we can compute a_{ij} from (6) and determine the eigenvalues λ_1 to λ_N and the eigenvectors $\mathbf{u}_1$ to $\mathbf{u}_N$. This determines $\mathbf{D}$, $\mathbf{Q}$, and $\mathbf{H}$. There are an infinite number of $N \times N$ orthogonal matrices. Any of them can be used as $\mathbf{G}$ (the choice of $\mathbf{G}$ will be discussed later). Then from (74) $\mathbf{P}$ can be determined, and this $\mathbf{P}$ satisfies (72), (69), and (55). This proves that, for any given channel, the digital filters in Fig. 4 can be designed to satisfy (55).

Next we prove the second statement after (55). As in Section II, we use $\epsilon_{\min}$ to denote the minimum value of ϵ when $\mathbf{c} = \mathbf{c}_{\text{opt}}$. Substituting $\mathbf{c}_{\text{opt}}$ in (60) into (57) yields

$$\epsilon_{\min} = \sum_{k=-\infty}^{\infty} d_k^2 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{v}. \tag{75}$$

Combining (63) and (75) gives

$$\epsilon_k = \epsilon_{\min} + \mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k \qquad k = 0, 1, 2, \cdots. \tag{76}$$

From (76), the initial value of ϵ is

$$\epsilon_0 = \epsilon_{\min} + \mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0, \tag{77}$$

and the value of ϵ after the first adjustment is

$$\epsilon_1 = \epsilon_{\min} + \mathbf{e}_1'\mathbf{PAP}'\mathbf{e}_1. \tag{78}$$

From (66), $\mathbf{e}_1 = [\mathbf{I} - \alpha_1\mathbf{PAP}']\mathbf{e}_0$. Substituting this into (78), we have

$$\epsilon_1 = \epsilon_{\min} + \mathbf{e}_0'[\mathbf{I} - \alpha_1\mathbf{PAP}']\mathbf{PAP}'[\mathbf{I} - \alpha_1\mathbf{PAP}']\mathbf{e}_0. \tag{79}$$

when $\mathbf{P}$ satisfies (74), (69) holds; i.e., $\mathbf{PAP}' = \mathbf{I}$. If we set $\alpha_1 = 1$, $\mathbf{I} - \alpha_1\mathbf{PAP}' = 0$. Then from (79) $\epsilon_1 = \epsilon_{\min}$. This proves that, when $\mathbf{P}$ satisfies (74) and $\alpha_1 = 1$, the mean-square error ϵ reduces to its minimum value $\epsilon_{\min}$ after only one adjustment of the gain controls.

In general it can be seen from (66) that, when $\mathbf{P}$ satisfies (74) and $\alpha_k = 1$, $\mathbf{e}_k = [\mathbf{I} - \alpha_k\mathbf{PAP}']\mathbf{e}_{k-1} = 0$, regardless of the value of $\mathbf{e}_{k-1}$. Consequently $\epsilon_k = \epsilon_{\min}$, regardless of the value of ϵ_{k-1}. This means that each of the adjustments is capable of reducing the mean-square error to its minimum value $\epsilon_{\min}$, regardless of what is the value of ϵ prior to that adjustment.

Now we summarize this section. It is emphasized at the beginning of this section that it is the correlations between the inputs to the gain controls that determine the differences between the eigenvalues and consequently the rate of convergence of ϵ. When the inputs to the gain controls are orthonormal, ϵ can be minimized in one adjustment. The inputs to the gain controls can be made orthonormal by using the generalized equalizer structure depicted in Fig. 3. The generalized equalizer can be realized with digital or analog filters. A realization using nonrecursive digital filters is depicted in Fig. 4, and is analyzed in detail. It is shown that, for any given channel, one can design the digital filters from the simple equation (74) to orthonormalize the inputs to the gain controls. When the digital filters are so designed, each adjustment of the gain controls can reduce ϵ to its minimum value, regardless of the value of ϵ prior to that adjustment. Therefore, ϵ can be minimized by only one adjustment of the gain controls.

V. FURTHER PROPERTIES OF THE NEW EQUALIZER

We have seen in the preceding section that it is possible to minimize ϵ in one adjustment. In this section we show that such an improvement in convergence rate is obtained without changing the residue noise power, the minimum mean-square error, and the convexity of the adjustments, and without complicating the gain-control adjustment loop.

5.1 *Residual Noise Power and Minimum Mean-Square Error*

As described in Section II, during data transmission the equalizer output is sampled sequentially at time-instants $t = \cdots, t_2 - T, t_2, t_2 + T, t_2 + 2T, \cdots$, and the time-samples are used to recover the transmitted information digits. Each of these time-samples consists of a signal and a noise component. The variance of this noise component (that is, the residue noise power) can be determined for both the new and the conventional equalizer.

Consider first the new equalizer (Fig. 4). Let the noise at the input of the equalizer be denoted by $n(t)$. The resulting noise at the input of c_i is denoted by $\mu_i(t)$, while the resulting noise at the equalizer output is denoted by $\nu(t)$. Clearly

$$\mu_i(t) = \sum_{j=1}^{N} P_{ij} n[t - (j-1)T], \qquad i = 1 \text{ to } N \tag{80}$$

and

$$\nu(t) = \sum_{i=1}^{N} c_i \mu_i(t). \tag{81}$$

We assume that the noise $n(t)$ is a zero-mean stationary Gaussian process. Then $\mu_i(t)$, $i = 1$ to N, and $\nu(t)$ are also zero-mean stationary Gaussian processes. Consequently, the variance of $\nu(t_k)$ is the same for all t_k. It is therefore sufficient to consider a single t_k (for example, $t_k = 0$). From (80) and (81)

$$\nu(0) = \sum_{i=1}^{N} c_i \sum_{j=1}^{N} P_{ij} n[(1 - j)T]. \tag{82}$$

Let $\mathbf{n}$ be an $N \times 1$ vector whose ith element is $n[(1 - i)T]$. It can be easily shown from (82) that

$$\nu(0) = \mathbf{c}'\mathbf{P}\mathbf{n}. \tag{83}$$

The noise vector $\mathbf{n}$ is distributed normally with zero mean. Let the covariance matrix of $\mathbf{n}$ be denoted by $\mathbf{\Lambda}$. Then $\nu(0)$ is distributed as the normal distribution with zero mean and variance $\mathbf{c}'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{c}$. Since $\mathbf{c}$ can be optimized in the training period, we assume that during data transmission $\mathbf{c} = \mathbf{c}_{\text{opt}}$. Thus, during data transmission the variance of $\nu(0)$ is $\text{Var}\,[\nu(0)] = \mathbf{c}'_{\text{opt}}\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{c}_{\text{opt}}$. Substituting (60) into above gives

$$\text{Var}\,[\nu(0)] = \mathbf{v}'\mathbf{A}^{-1}\mathbf{\Lambda}\mathbf{A}^{-1}\mathbf{v}. \tag{84}$$

Note from (84) that $\text{Var}[\nu(0)]$ is independent of $\mathbf{P}$.

Next consider the conventional equalizer. It can be easily seen that when $\mathbf{P} = \mathbf{I}$ the new equalizer in Fig. 4 is reduced to the conventional equalizer in Fig. 1. Since $\text{Var}[\nu(0)]$ in (84) remains unchanged when $\mathbf{P}$ is set to $\mathbf{I}$, $\text{Var}[\nu(0)]$ of the new equalizer is equal to $\text{Var}[\nu(0)]$ of the conventional equalizer. In other words, the use of the new equalizer does not change the residue noise power of the system.

The minimum mean-square error, $\epsilon_{\min}$, has already been determined in the preceding sections. Comparing (12) with (75) shows that the minimum mean-square error does not change when the new equalizer is used instead of the conventional equalizer.

5.2 *Convexity of the Adjustment*

It has been shown[3,4] that, when the conventional equalizer is used, ϵ is a strict convex function of the gain controls $\mathbf{c}$. This ensures that ϵ converges for all initial settings of $\mathbf{c}$, and that ϵ converges when the gradient method [equation (15)] is approximated by certain other iterative techniques. In this subsection we show that ϵ remains as a strict convex function of $\mathbf{c}$ when the new equalizer is used.

Let $\mathcal{K}$ denote the open convex set $\{\mathbf{c}: -\infty < c_i < \infty, i = 1 \text{ to } N\}$. Let $\mathbf{h}$ be an $N \times 1$ vector whose ith element is h_i. Clearly,[17] ϵ is a function of class $c^{(2)}$ on the set $\mathcal{K}$. If we can show that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} h_i h_j \frac{\partial^2 \epsilon}{\partial c_i \, \partial c_j} > 0 \tag{85}$$

for every $\mathbf{c} \, \varepsilon \, \mathcal{K}$ and every $\mathbf{h} \neq \mathbf{0}$, then [17] ϵ is strictly convex on $\mathcal{K}$.

It can be shown from (57) that

$$\frac{\partial^2 \epsilon}{\partial c_i \, \partial c_j} = 2b_{ij}, \quad \text{all} \quad i \quad \text{and} \quad j \tag{86}$$

where b_{ij} has been specified in (68). From (86) and (67), we have

$$\sum_{i=1}^{N} \sum_{j=1}^{N} h_i h_j \frac{\partial^2 \epsilon}{\partial c_i \, \partial c_j} = 2\mathbf{h}'\mathbf{PAP}'\mathbf{h}. \tag{87}$$

It can be seen from (87) that (85) holds for every $\mathbf{c} \, \varepsilon \, \mathcal{K}$ and every $\mathbf{h} \neq \mathbf{0}$ if and only if $\mathbf{PAP}'$ is positive definite. Since $\mathbf{P}$ is a nonsingular matrix [see (74)], and $\mathbf{A}$ is positive definite, $\mathbf{PAP}'$ is positive definite. Therefore, (85) holds for every $\mathbf{c} \, \varepsilon \, \mathcal{K}$ and every $\mathbf{h} \neq \mathbf{0}$, and consequently ϵ is strictly convex on $\mathcal{K}$. This proves that ϵ remains as a strict convex function of $\mathbf{c}$ when the new equalizer is used.

5.3 *Gain-Control Adjustment Loop*

In the case of the conventional equalizer, the gain controls $\mathbf{c}$ are adjusted according to (15). The partial derivative $\partial\epsilon/\partial c_i$ used in (15) is obtained[3] by correlating the time-samples of the tap signal and the time-samples of the error signal $y(t) - d(t)$.

In the case of the new equalizer, the gain controls $\mathbf{c}$ are again adjusted according to (15). Since

$$y(t) = \sum_{i=1}^{N} c_i z_i(t),$$

$$\frac{\partial \epsilon}{\partial c_i} = \frac{\partial}{\partial c_i} \left\{ \sum_{k=-\infty}^{\infty} [y(kT) - d(kT)]^2 \right\}$$

$$= 2 \sum_{k=-\infty}^{\infty} [z_i(kT)][y(kT) - d(kT)],$$

$$i = 1 \quad \text{to} \quad N. \tag{88}$$

It is seen from (88) that $\partial\epsilon/\partial c_i$ can be obtained similarly by correlating the time-samples of $z_i(t)$ and the time-samples of $y(t) - d(t)$. Therefore, the gain-control adjustment loop of the new equalizer is essentially the same as that of the conventional equalizer.

VI. APPLICATION OF THE NEW EQUALIZER STRUCTURE

In this and the next sections, we consider how to use the new equalizer structure for a Class IV partial-response system. It will be shown, both analytically and by computer simulation, that when the amplitude characteristic of the transmission medium does not vary appreciably from channel to channel (such as in the case of multiparty private line polling systems), we can use a fixed **P** matrix for the new equalizer (that is, P_{ij} need not be adjusted in each training period). Such a simplification is possible even though the system timing, the demodulating carrier phase, and the phase characteristic of the transmission medium change from channel to channel.

It has been shown in Section IV that ϵ can be minimized in one adjustment if $\alpha_k = 1$ and **P** satisfies the equation

$$\mathbf{PAP}' = \mathbf{I}. \tag{69}$$

As pointed out in Section 3.1, the elements a_{ij} of the **A** matrix are independent of the demodulating carrier phase, the system timing, the phase characteristics of the transmitting and receiving filters, and the phase characteristic of the transmission medium. Therefore, from (69), **P** is independent of all these parameters. It can also be seen from (36) in Section 3.1 that a_{ij} depends only on the amplitude characteristic $|F_2(f)|$ of the transmission medium (amplitude distortions of the transmitting and receiving filters are usually negligible). In some systems, such as private-line polling systems, $|F_2(f)|$ does not vary appreciably from channel to channel. For such systems, a fixed **P** matrix may approximately satisfy (69) for all channels. Therefore, in designing such systems, we may estimate $|F_2(f)|$ and compute from this an estimate of **A** (this estimate of **A** will be denoted by **S**). We can then estimate **P** from the equation

$$\mathbf{PSP}' = \mathbf{I}. \tag{89}$$

and use this estimated **P** for all channels. When the estimated $|F_2(f)|$ agrees with the actual $|F_2(f)|$, **S** agrees with **A**. Consequently the estimated **P** satisfies (69), and ϵ is minimized in one adjustment. When the estimated $|F_2(f)|$ differs from the actual $|F_2(f)|$, the estimated **P** will not satisfy (69) and the convergence rate of ϵ will be reduced. In order to see whether the reduced convergence rate is satisfactory, we need formulas to relate the convergence rate to $|F_2(f)|$. Such formulas will be derived in this section, and will be used in the next section for a hypothetical data communication system.

Let the difference between the estimate of **A** and **A** be denoted by **R**, that is,

$$\mathbf{R} = \mathbf{S} - \mathbf{A}. \tag{90}$$

Since **S** is an correlation matrix, it is symmetric. Therefore, **R** is also a symmetric matrix.

From (66) and (90), we have

$$\mathbf{e}_k = [\mathbf{I} - \alpha_k\mathbf{PSP}' + \alpha_k\mathbf{PRP}']\mathbf{e}_{k-1}, \qquad k = 1, 2, \cdots . \tag{91}$$

As in Section IV, we set

$$\alpha_k = 1, \qquad \text{all } k. \tag{92}$$

Since the estimated **P** is computed from (89), we can substitute (89) and (92) into (91) to obtain

$$\mathbf{e}_k = \mathbf{PRP}'\mathbf{e}_{k-1}, \qquad k = 1, 2, \cdots . \tag{93}$$

It can be easily seen from the recursive equation (93) that

$$\mathbf{e}_k = (\mathbf{PRP}')^k\mathbf{e}_0, \qquad k = 1, 2, \cdots . \tag{94}$$

From (76), the mean-square error after the kth adjustment is

$$\epsilon_k = \epsilon_{\min} + \mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k . \tag{76}$$

We now evaluate the last term $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ in the equation above. It can be shown from (89) that

$$\mathbf{P}'\mathbf{P} = \mathbf{S}^{-1}. \tag{95}$$

From (95), (94) can be rewritten as

$$\mathbf{e}_k = \mathbf{PR}(\mathbf{S}^{-1}\mathbf{R})^{k-1}\mathbf{P}'\mathbf{e}_0, \qquad k = 1, 2 \cdots . \tag{96}$$

From (96) and (95), we obtain

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k = \mathbf{e}_0'\mathbf{P}(\mathbf{RS}^{-1})^k\mathbf{A}(\mathbf{S}^{-1}\mathbf{R})^k\mathbf{P}'\mathbf{e}_0, \quad k = 1, 2, \cdots . \tag{97}$$

Using $\mathbf{A} = \mathbf{S} - \mathbf{R}$ and (89), we can rearrange (97) into the following form

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k = \mathbf{e}_0'\mathbf{P}(\mathbf{RS}^{-1})^{2k}(\mathbf{I} - \mathbf{RS}^{-1})\mathbf{P}^{-1}\mathbf{e}_0, \qquad k = 1, 2, \cdots . \tag{98}$$

From (77), the initial value of mean-square error is

$$\epsilon_0 = \epsilon_{\min} + \mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0 \, . \tag{77}$$

It can be easily shown from (90) and (89) that

$$\mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0 = \mathbf{e}_0'\mathbf{P}[\mathbf{I} - \mathbf{RS}^{-1}]\mathbf{P}^{-1}\mathbf{e}_0 \, . \tag{99}$$

Since $\mathbf{A}$ is symmetric, $\mathbf{PAP}'$ is symmetric. Let $\zeta_1 , \zeta_2 , \cdots , \zeta_N$ be the eigenvalues of $\mathbf{PAP}'$, and let $\mathbf{w}_1 , \mathbf{w}_2 , \cdots \mathbf{w}_N$ be a set of orthonormal eigenvectors of $\mathbf{PAP}'$ ($\mathbf{w}_i$ corresponds to ζ_i). From

$$\mathbf{PAP}'\mathbf{w}_i = \zeta_i\mathbf{w}_i \tag{100}$$

and [see (99)]

$$\mathbf{PAP}' = \mathbf{P}[\mathbf{I} - \mathbf{RS}^{-1}]\mathbf{P}^{-1} \tag{101}$$

we have

$$\mathbf{RS}^{-1}\mathbf{P}^{-1}\mathbf{w}_i = (1 - \zeta_i)\mathbf{P}^{-1}\mathbf{w}_i \, . \tag{102}$$

It is seen from (102) that the ith eigenvalue of $\mathbf{RS}^{-1}$ is $(1 - \zeta_i)$, and the ith eigenvector of $\mathbf{RS}^{-1}$ is $\mathbf{P}^{-1}\mathbf{w}_i$. We shall denote the ith eigenvalue of $\mathbf{RS}^{-1}$ by $\hat{\lambda}_i$ so

$$\hat{\lambda}_i = 1 - \zeta_i \, , \qquad i = 1 \text{ to } N. \tag{103}$$

Now consider the matrix product $\mathbf{P}(\mathbf{RS}^{-1})^{2k}(\mathbf{I} - \mathbf{RS}^{-1})\mathbf{P}^{-1}$ in (98). To facilitate writing, we shall denote this product by $\mathbf{\Phi}$, that is

$$\mathbf{\Phi} = \mathbf{P}(\mathbf{RS}^{-1})^{2k}(\mathbf{I} - \mathbf{RS}^{-1})\mathbf{P}^{-1}. \tag{104}$$

From the eigenvalue and eigenvectors of $\mathbf{RS}^{-1}$, we see that the ith eigenvalue of $\mathbf{\Phi}$ is $\hat{\lambda}_i^{2k}(1 - \hat{\lambda}_i)$, and the ith eigenvector of $\mathbf{\Phi}$ is $\mathbf{w}_i$. We shall denote the ith eigenvalue of $\mathbf{\Phi}$ by δ_i , so

$$\delta_i = \hat{\lambda}_i^{2k}(1 - \hat{\lambda}_i). \tag{105}$$

We have shown in the above that the orthonormal eigenvectors $\mathbf{w}_1$ to $\mathbf{w}_N$ of $\mathbf{PAP}'$ are also eigenvectors of $\mathbf{\Phi}$. Hence, $\mathbf{PAP}'$ and $\mathbf{\Phi}$ can be simultaneously diagonalized by $\mathbf{w}_1$ to $\mathbf{w}_N$. By such diagonalizations we can reduce (99) to the form

$$\mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0 = \sum_{i=1}^{N} (1 - \hat{\lambda}_i)(\mathbf{e}_0'\mathbf{w}_i)^2 \tag{106}$$

and reduce (98) to the form

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k = \sum_{i=1}^{N} \hat{\lambda}_i^{2k}(1 - \hat{\lambda}_i)(\mathbf{e}_0'\mathbf{w}_i)^2, \qquad k = 1, 2, \cdots . \tag{107}$$

It can be seen from (106) and (107) that the convergence rate of ϵ depends on $\hat{\lambda}_1$ to $\hat{\lambda}_N$. Let $|\hat{\lambda}_i|_{\max}$ denote the largest $|\hat{\lambda}_i|$ among $|\hat{\lambda}_1|$ to $|\hat{\lambda}_N|$, that is,

$$|\hat{\lambda}_i|_{\max} \geqq |\hat{\lambda}_i|, \qquad i = 1 \text{ to } N. \tag{108}$$

Since $\mathbf{PAP}'$ is positive definite, its eigenvalues must all be positive. Hence,

$$\zeta_i = 1 - \hat{\lambda}_i > 0, \qquad i = 1 \text{ to } N. \tag{109}$$

From (106) to (109), the following bound is obtained

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k \leqq [\,|\hat{\lambda}_i|_{\max}]^{2k}\mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0\,, \quad k = 1, 2, \cdots. \tag{110}$$

Based on (110), we have the first method of estimating the convergence rate of $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$:

First Method. From each $|F_2(f)|$, compute $\hat{\lambda}_1$ to $\hat{\lambda}_N$. This gives $|\hat{\lambda}_i|_{\max}$. According to (110), $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ reduces by at least a factor of $[\,|\hat{\lambda}_i|_{\max}]^{-2}$ after each adjustment. This completes the estimation.

This method will be illustrated in the next section. In certain applications, measurements of $|F_2(f)|$ may not be available. It may only be specified that $|F_2(f)|$ varies within certain bounds. In such cases, $|\hat{\lambda}_i|_{\max}$ cannot be determined. However, it can be bounded from the bounds of $|F_2(f)|$. It is shown in Appendix B that

$$1 - [\eta(F)]_{\max} \leqq \hat{\lambda}_i \leqq 1 - [\eta(f)]_{\min}\,, \qquad i = 1 \text{ to } N. \tag{111}$$

where

$$\eta(f) = \frac{[|F_2(f - f_c)|_{\text{act}}]^2}{[|F_2(f - f_c)|_{\text{est}}]^2}\,, \qquad 0 \leqq f \leqq f_2 - f_1\,. \tag{112}$$

As explained in Appendix B, $|F_2(f - f_c)|_{\text{act}}$ is the actual value of $|F_2(f - f_c)|$, $|F_2(f - f_c)|_{\text{est}}$ is the estimated value of $|F_2(f - f_c)|$, $[\eta(f)]_{\max}$ is the maximum value of $\eta(f)$ in the frequency range $0 \leqq f \leqq f_2 - f_1$, *and* $[\eta(f)]_{\min}$ is the minimum value of $\eta(f)$ in the frequency range of $0 \leqq f \leqq f_2 - f_1$.

Based on (110) to (112), we have the second method of estimating the convergence rate of $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$:

Second Method. From the bounds of $|F_2(f)|$, determine $[\eta(f)]_{\max}$ and $[\eta(f)]_{\min}$ from (112). If

$$|1 - [\eta(f)]_{\min}| > |1 - [\eta(f)]_{\max}|,$$

use $|1 - [\eta(f)]_{\min}|$ as the upper bound of $|\hat{\lambda}_i|_{\max}$. If

$$|1 - [\eta(f)]_{\max}| > |1 - [\eta(f)]_{\min}|,$$

use $|1 - [\eta(f)]_{\max}|$ as the upper bound of $|\hat{\lambda}_i|_{\max}$. Use the upper bound of $|\hat{\lambda}_i|_{\max}$ as $|\hat{\lambda}_i|_{\max}$ to obtain from (110) an estimate of the minimum convergence rate of $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$.

As illustrated in the next section, the second method requires only some simple hand-calculations, and hence may be applied first even if measurements of $|F_2(f)|$ are available. However, it should be warned that the second method is usually too pessimistic because the upper bound of $|\hat{\lambda}_i|_{\max}$, obtained in the above fashion, can be several times larger than the actual $|\hat{\lambda}_i|_{\max}$. (Consequently, the convergence rate may appear to be unsatisfactory when it is actually satisfactory.) Therefore, for borderline cases, the first method should be used.

We illustrate these two methods in the next section.

VII. NUMERICAL EXAMPLES AND COMPUTER SIMULATION

In order to illustrate the methods we consider a hypothetical private-line data communication system (hereafter referred to as the hypothetical system). We assume that the system uses a Class IV partial-response signal and transmits over voiceband at a baud rate of 4800 bauds. The cutoff frequencies f_1 and f_2 defined in Section 3.1 will be 400 and 2800 Hz respectively. The demodulating carrier frequency f_c is equal to f_2. It is assumed that the new equalizer in Fig. 4 is used with $N = 13$ and a prefixed $\mathbf{P}$ matrix computed from (89) (see footnote in Section 7.1). The two methods will be used in Sections 7.1 and 7.2, which follow, to estimate the convergence rate of $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ {the convergence rate of the mean-square error ϵ is identical to the convergence rate of $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ [see (76) in Section IV]}. A third and very elaborate method, computer simulation of the data communication system, will be used in Section 7.3 to demonstrate the convergence rate.

7.1 *Estimation by the First Method*

Since $|F_2(f)|$ of private lines does not deviate considerably from a constant in the frequency band 400 to 2800 Hz, we shall simply use a constant as the estimate of $|F_2(f)|$ in this frequency band, and compute a $\mathbf{P}$ matrix accordingly from (89). This prefixed $\mathbf{P}$ matrix will be used for all lines.

Now consider what happens to the convergence rate when the private line has a $|F_2(f)|$ as depicted in Fig. 5. It can be seen that this $|F_2(f)|$ varies from -4.8 dB to 0.4 dB, and then to -6.3 dB in the frequency band 400 to 2800 Hz. Such a variation is large for private lines. However, we show that even for this large variation the prefixed $\mathbf{P}$ matrix

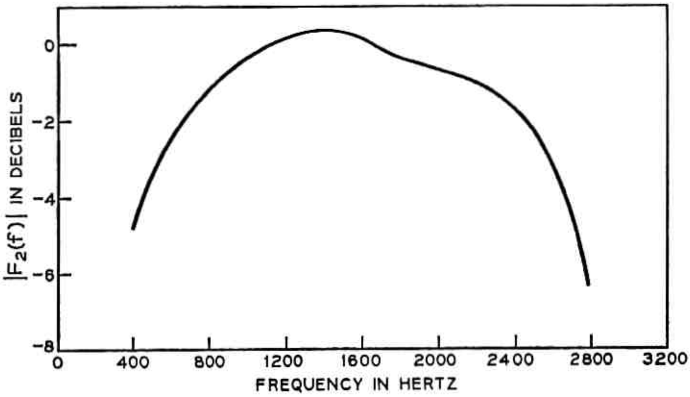


Fig. 5—The $|F_2(f)|$ used in Section 7.1.

can be used and the convergence rate of the mean-square error is satisfactory.

In order to use (110), we compute the eigenvalues $\hat{\lambda}_1$ to $\hat{\lambda}_N$ in the following steps:

Step 1. Compute the elements a_{ij} of the **A** matrix from (36) and the $| F_2(f) |$ in Fig. 5. It is sufficient to compute only the first row of **A**. To see this, note from (6) that a_{ij} satisfies the condition

$$a_{ij} = a_{ji} \tag{113}$$

and the condition

$$a_{ij} = a_{(i+h)(j+h)} \, . \tag{114}$$

From (113) and (114), we see that **A** can be written in the form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1N} \\ a_{12} & a_{11} & a_{12} & a_{13} & \cdots & a_{1(N-1)} \\ a_{13} & a_{12} & a_{11} & a_{12} & \cdots & a_{1(N-2)} \\ a_{14} & a_{13} & a_{12} & a_{11} & \cdots & a_{1(N-3)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{1N} & a_{1(N-1)} & a_{1(N-2)} & a_{1(N-3)} & \cdots & a_{11} \end{bmatrix}, \tag{115}$$

where the elements on the diagonal line are all equal to a_{11}, and the elements on each off-diagonal line are equal (for example, the elements on the second off-diagonal line are all equal to a_{13}). Thus, it is sufficient to compute only the first row of **A**.

Step 2. Compute the **S** matrix. It can be seen from Section 3.1 (from (37) to (40)) that, when

$$|F_2(f)| = \text{a constant}, \qquad f_1 \leqq f \leqq f_2, \tag{116}$$

we have

$$\mathbf{A} = a_{11}\mathbf{\Delta}, \tag{117}$$

where

$$\mathbf{\Delta} = \begin{bmatrix} 1 & 0 & -\frac{1}{2} & 0 & \cdots & 0 \\ 0 & 1 & 0 & -\frac{1}{2} & \cdots & 0 \\ -\frac{1}{2} & 0 & 1 & 0 & \cdots & 0 \\ 0 & -\frac{1}{2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{118}$$

Since we have used a constant as the estimate of $|F_2(f)|$, we have

$$\mathbf{S} = a_{11}\mathbf{\Delta}. \tag{119)*}$$

Step 3. Compute $\mathbf{R} = \mathbf{S} - \mathbf{A}$, and compute the eigenvalues $\hat{\lambda}_1$ to $\hat{\lambda}_N$ of $\mathbf{RS}^{-1}$.

The computations in the above three steps can best be carried out by a computer program. Such a program has been written. For the $|F_2(f)|$ in Fig. 5, the eigenvalues of $\mathbf{RS}^{-1}$ are found to be: -0.1573, -0.1434, -0.0902, -0.0615, -0.0034, 0.0242, 0.0716, 0.1084, 0.1834, 0.1945, 0.2954, 0.3797, and 0.4651. The largest magnitude of these eigenvalues is 0.4651; therefore,

$$|\hat{\lambda}_i|_{\max} = 0.4651. \tag{120}$$

Substituting (120) into (110) gives

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k \leqq \left[\frac{1}{4.62}\right]^k \mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0 \, . \tag{121}$$

It can be seen from (121) that $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ reduces at least 4.62 times

* From (119) and (89), we may determine the **P** matrix. Since **P** is not needed for the following computation, we do not carry out such calculations here. However, it should be pointed out that the method from (69) to (74) may be used for such calculations. It should also be noted that by setting the **G** matrix in (74) to **I**, we can reduce most of the elements in **P** to zero. For example, when N is a multiple of four, it is sufficient to implement only $N^2/8$ of the P_{ij} s in Fig. 4. This greatly simplifies the implementation.

after each adjustment. In other words, each adjustment reduces the mean-square error by at least 6.65 dB. This convergence rate is satisfactory. For example, let $\epsilon_{\min}$ be 0.01 and let the initial mean-square error ϵ_0 be as large as 4 (see computer simulation in Section 7.3). From (76), (77), and (121), we see that after four adjustments (approximately 16 milliseconds) the mean-square error ϵ reduces to less than 0.0187. Clearly, this convergence rate is satisfactory.

7.2 *Estimation by the Second Method*

When measurements of $|F_2(F)|$ are not available, we may estimate the convergence rate from the bounds of $|F_2(f)|$. For illustrative purposes, let us assume that $|F_2(f)|$ does not deviate from unity by more than -2.5 dB or 1.5 dB, that is,

$$-2.5 \leqq 20 \log_{10} |F_2(f)| \leqq 1.5, \qquad 400 \leqq f \leqq 2800.$$

Notice that $|F_2(f)|$ can vary in any manner within these bounds. We simply use unity as the estimate of $|F_2(f)|$. Then from (112), $[\eta(f)]_{\max} = 1.414$ and $[\eta(f)]_{\min} = 0.562$. Since $|1 - [\eta(f)]_{\min}| > |1 - [\eta(f)]_{\max}|$, we use $|1 - [\eta(f)]_{\min}|$ as the upper bound of $|\hat{\lambda}_i|_{\max}$, that is, $|\hat{\lambda}_i|_{\max} \leqq 0.438$. Using $|\hat{\lambda}_i|_{\max} = 0.438$, we obtain from (110)

$$\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k \leqq \left[\frac{1}{5.2}\right]^k \mathbf{e}_0'\mathbf{PAP}'\mathbf{e}_0 .$$

Thus, $\mathbf{e}_k'\mathbf{PAP}'\mathbf{e}_k$ reduces at least 5.2 times after each adjustment (a 7.16 dB reduction per adjustment). Notice that since the actual value of $|\hat{\lambda}_i|_{\max}$ may be much less than 0.438, the actual convergence rate can be much faster.

7.3 *Computer Simulation of the Hypothetical System*

The analytical methods in the preceding subsections yield bounds of the convergence rate. In order to demonstrate the actual convergence rate, we simulate the hypothetical system on the computer, using a number of private-line characteristics obtained from field measurement. In order to conserve space we describe here only the most important results. Fig. 6 shows the convergence of the mean-square error for a private line which has severe amplitude and phase distortions (see Table I). Other lines simulated have less severe distortions and hence faster convergence rates. A 30-dB signal-to-noise ratio is assumed at the equalizer input. The bottom curve shows the convergence of ϵ when the new equalizer is used (with a prefixed $\mathbf{P}$ matrix computed from (119) and (89)). It can be seen that, when the

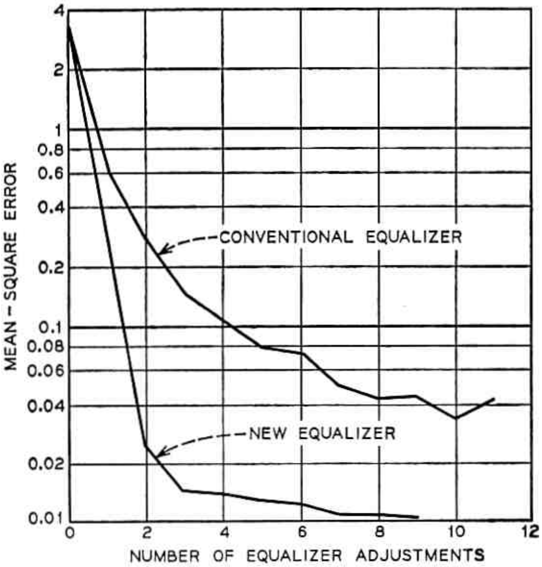


Fig. 6—Convergence of the mean-square error.

Table I—Relative Envelope Delay and Relative Loss Characteristics of Private Line used in Section 7.3.

Frequency (Hz)	Relative Envelope Delay (μs)	Relative Loss (dB)
300	5500	6.3
500	2830	2.4
600	2060	1.9
800	1040	0.6
1000	590	0
1200	390	−0.9
1400	280	−1.3
1600	150	−0.8
1800	0	0
2000	−100	0.7
2200	− 80	1.0
2400	15	2.2
2600	270	2.7
2800	260	4.5
3000	1500	7.1

new equalizer is used, ϵ converges rapidly to its minimum value, and the training period may be terminated after the third adjustment (three adjustments require approximately 12 milliseconds). These simulation results are in full agreement with the theoretical results in the preceding subsections.

The top curve in Fig. 6 shows the convergence of ϵ when the conventional equalizer in Fig. 1 is used (also with $N = 13$). It can be seen that the convergence of ϵ is rather slow, and that it takes more than 11 adjustments (approximately 44 milliseconds) to reduce ϵ to close to its minimum value.

Figure 6 is obtained with a specific setting of system timing and demodulating carrier phase. The computer program has also been executed for a sufficiently large number of other timing and carrier phase settings. Curves similar to those in Fig. 6 have been obtained. The results may be summarized in Fig. 7. The horizontal axis in Fig. 7

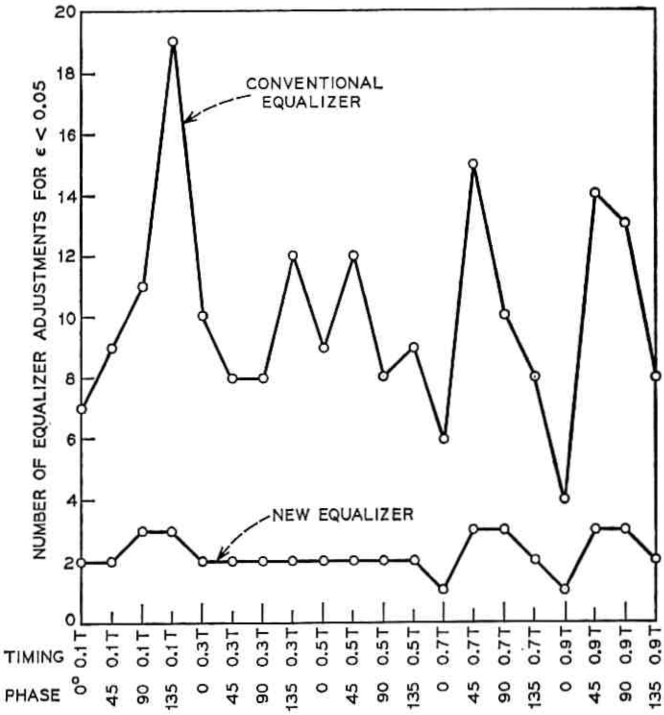


Fig. 7—Number of equalizer adjustments required to reduce the mean-square error to less than 0.05 for different timing and carrier-phase settings.

shows the setting of the timing and the carrier phase, while the vertical axis shows the number of adjustments required to reduce the mean-square error ϵ to less than 0.05. For example, when timing is set to $0.1T$ and the carrier phase is set to 90 degrees, three adjustments are required to reduce ϵ to less than 0.05 when the new equalizer is used, while 11 adjustments are required when the conventional equalizer is used. It can be seen from the top curve in Fig. 7 that when the conventional equalizer is used, the number of required adjustments varies a great deal with timing and carrier phase, and as many as 19 adjustments can be required. On the other hand, when the new equalizer is used (bottom curve), the number of required adjustments remains small for all time and carrier-phase settings. (This is due to the first convergence of the new equalizer.) Thus, when the new equalizer is used, it is not necessary to make fine adjustments of timing and carrier phase at the beginning of the training period. We may simply set these parameters to some reasonable values, and then use the new equalizer to quickly reduce the mean-square error. It is possible that the whole process of adjusting the timing, the carrier phase, and the equalizer can be completed in a brief training period.

VIII. SUMMARY AND CONCLUSIONS

Section II considers amplitude modulation data communication systems that use transversal filters for automatic equalization. The gain controls of the transversal filters are adjusted by the gradient method [equation (15)] to minimize the mean-square error between the received and the desired pulses. After the fundamentals are reviewed, the mean-square error is decomposed into N error-components (N is the number of taps on the transversal filter). The error-components depend on the eigenvalues $\lambda_1, \cdots, \lambda_N$ of the correlation matrix $\mathbf{A}$. These eigenvalues and the convergence of the error-components depend on the signaling scheme and the channel characteristics. Two important classes of digital communication systems are examined: the class IV partial-response system and the SSB Nyquist system. It is shown that for both systems, the engenvalues $\lambda_1, \cdots, \lambda_N$ are independent of the demodulation carrier phase, the system timing, the phase characteristic of the transmission medium, and the phase characteristics of the transmitting and receiving filters. Consequently, the eigenvalues depend on only the amplitude characteristics of the transmission medium and the filters. Since amplitude characteristics of the filters depend on the signaling scheme, it might be said that the eigenvalues depend on only the signaling scheme and the amplitude characteristic $|F_2(f)|$ of the transmission medium. For applications where $|F_2(f)|$ does not vary considerably from channel to channel (such as private-line systems), the eigenvalues depend primarily on the signaling scheme. For the SSB Nyquist system, the eigenvalues $\lambda_1, \cdots, \lambda_N$ are nearly equal. Consequently, as shown in Section 3.2, each adjustment reduces each of the error-components by a large factor, and the mean-square error can be minimized after only a few adjustments (for example, two adjustments). For the class IV partial-response system, the eigenvalues $\lambda_1, \cdots, \lambda_N$ are very different in magnitude. (The ratio of the maximum eigenvalue to the minimum eigenvalue is very large; e.g., 40 or larger.) Therefore, each adjustment cannot reduce each error-component by a large factor, and, as can be seen from Figs. 6 and 7, the convergence rate may not be satisfactory for fast start-up purposes.

The results in Section III suggest that in order to improve the convergence rate an attempt should be made to reduce the differences between the eignevalues. As emphasized in Section IV, differences in eigenvalues are caused by the correlation between the inputs to the gain controls. When the inputs to the gain controls are orthonormal, the eigenvalues are all equal and the mean-square error can be minimized in only one adjustment. The inputs to the gain controls can be made orthonormal by using the generalized equalizer structure in Fig. 3. The generalized equalizer can be realized with digital or analog filters. Because of the press toward digitalization and integrated circuits, digital filters are used. A realization using nonrecursive digital filters is depicted in Fig. 4. This equalizer structure, referred to as the new equalizer structure, is analyzed in detail. It is shown that for any given channel the coefficients of the digital filters can be set according to (69) or (74) to make the inputs to the gain controls orthonormal. Then each adjustment of the gain controls can reduce the mean-square error ϵ to its minimum value, regardless of the magnitude of ϵ prior to that adjustment. In Section V it is shown that such a fast convergence (one-step convergence) is obtained without changing the residual noise power, the minimum mean-square error, and the convexity of the adjustments, and without complicating the gain-control adjustment loops.

The theory in Sections IV and V is general in that it applies regardless of the type of modulation (SSB, VSB, or DSB) or the signaling scheme (partial-response or Nyquist). Sections VI and VII consider the application of the theory to a Class IV partial-response system. Most importantly, it is shown that the coefficients P_{ij} of the digital filters in the new equalizer can be prefixed. It is first observed that for Class IV partial-response systems the coefficients P_{ij} in the new equalizer depend only on the amplitude characteristic $|F_2(f)|$ of the transmission medium. Consequently, for private-line systems and systems where $|F_2(f)|$ does not vary considerably from channel to channel (delay distortion can vary arbitrarily), we may compute the coefficients P_{ij} from an estimated (typical) $|F_2(f)|$ and use these prefixed P_{ij} for all channels. One-step convergence is obtained when actual $|F_2(f)|$ agrees with the estimated $|F_2(f)|$. When actual $|F_2(f)|$ differs from the estimated $|F_2(f)|$, the convergence rate of the mean-square error is reduced. In order to determine if the reduced convergence rate is satisfactory, two analytical methods are developed to relate the convergence rate to $|F_2(f)|$. These methods are first presented in Section VI, and are used in Section VII for a hypothetical data communication system. This hypothetical system uses a Class IV partial-response signal and transmits over private voice lines at a data rate of 4800 bauds. For private lines, $|F_2(f)|$ does not vary considerably in the passband 400 to 2800 Hz; therefore, a constant is used as the estimate of $|F_2(f)|$ and the prefixed P_{ij} are computed from (89) and (119). It is shown analytically in Section 7.1 that the use of the prefixed P_{ij} yields a satisfactory convergence rate even when the line has severe amplitude distortion. To further demonstrate the convergence rate, the hypothetical system is simulated on a digital computer, using the prefixed P_{ij} and a number of private-line characteristics obtained from field measurements. A 30-dB signal-to-noise ratio at the equalizer input is assumed. Figs. 6 and 7 illustrate the convergence of the mean-square error for a private line that has severe amplitude and phase distortion. (The convergence is faster for other lines.) It can be seen from these figures that, when the new equalizer is used, the equalizer settles after three adjustments (approximately 12 milliseconds). Furthermore, because of the fast convergence, the settling time remains small for all settings of timing and demodulating carrier phase. Thus, when the new equalizer is used, it is not r ecessary to make fine adjustments of timing and demodulating carrier phase in the start-up period. These parameters can simply be set to some reasonable values at the beginning of the start-up period. This can further reduce the overall start-up time of the system.

IX. ACKNOWLEDGMENT

The author wishes to thank Mrs. B. Butler for writing the computer programs used in Sections 7.1 and 7.3, and J. Salz, D. A. Spaulding, K. H. Mueller, and E. Y. Ho for helpful discussions.

APPENDIX A

A.1 *Determination of Eigenvalues and Eigenvectors*

In this appendix we determine the eigenvalues and eigenvectors of the $N \times N$ matrix $\mathbf{A}$ in equation (40). The two cases of odd N and even N are considered separately.

Case 1. Odd N.

Let λ be an eigenvalue of $\mathbf{A}$, and let

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_N \end{bmatrix} \tag{122}$$

be the corresponding eigenvector. We can rewrite the equation $\mathbf{A}\ \mathbf{u} = \lambda\ \mathbf{u}$ as

$$\sum_{j=1}^{N} a_{ij}u_j = \lambda u_i , \qquad i = 1 \quad \text{to} \quad N. \tag{123}$$

Equation (123) can be split into the following two equations:

$$\sum_{j=1}^{N} a_{ij}u_j = \lambda u_i , \qquad i = \text{odd} \tag{124}$$

and

$$\sum_{j=1}^{N} a_{ij}u_j = \lambda u_i , \qquad i = \text{even}. \tag{125}$$

From (39), $a_{ij} = 0$ when $i - j$ is odd. Thus (124) can be reduced to

$$\sum_{j=1,3,\cdots,N} a_{ij}u_j = \lambda u_i , \qquad i = 1, 3, \cdots , N.$$

Using (39), we can write the above equation as

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 & \cdots & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \cdots & 0 \\ 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & \cdots & 0 \\ 0 & 0 & -\frac{1}{2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_3 \\ u_5 \\ u_7 \\ \vdots \\ u_N \end{bmatrix} = \lambda \begin{bmatrix} u_1 \\ u_3 \\ u_5 \\ u_7 \\ \vdots \\ u_N \end{bmatrix}. \tag{126}$$

The first matrix on the left-hand side of (126) is an $(N + 1)/2 \times (N + 1)/2$ tridiagonal matrix. In addition to the trivial solution

$$u_1 = u_3 = u_5 = \cdots = u_N = 0,$$

there are $(N + 1)/2$ nonzero solutions to (126). The kth such solution, $k = 1, 2, \cdots, (N + 1)/2$, is[14]

$$\lambda = 1 - \cos \frac{k\pi}{\dfrac{N+1}{2} + 1} \tag{127}$$

and

$$\begin{bmatrix} u_1 \\ u_3 \\ \vdots \\ u_N \end{bmatrix} = \left[\frac{2}{\dfrac{N+1}{2} + 1} \right]^{\frac{1}{2}} \begin{bmatrix} \sin \dfrac{k\pi}{\dfrac{N+1}{2} + 1} \\ \sin \dfrac{k2\pi}{\dfrac{N+1}{2} + 1} \\ \vdots \\ \sin \dfrac{k \dfrac{N+1}{2} \pi}{\dfrac{N+1}{2} + 1} \end{bmatrix}. \tag{128}$$

From (39), (125) can be written as

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 & \cdots & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & \cdots & 0 \\ 0 & -\frac{1}{2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_2 \\ u_4 \\ u_6 \\ \vdots \\ u_{(N-1)} \end{bmatrix} = \lambda \begin{bmatrix} u_2 \\ u_4 \\ u_6 \\ \vdots \\ u_{(N-1)} \end{bmatrix}. \tag{129}$$

where the first matrix on the left-hand side is an $(N - 1)/2 \times (N - 1)/2$ tridiagonal matrix. In addition to the trivial solution

$$u_2 = u_4 = u_6 = \cdots = u_{(N-1)} = 0,$$

there are $(N - 1)/2$ nonzero solutions to (129). The kth such solution, $k = 1, 2, \cdots, (N - 1)/2$, is

$$\lambda = 1 - \cos \frac{k\pi}{\frac{N-1}{2} + 1} \tag{130}$$

and

$$\begin{bmatrix} u_2 \\ u_4 \\ \vdots \\ u_{(N-1)} \end{bmatrix} = \left[\frac{2}{\frac{N-1}{2} + 1}\right]^{\frac{1}{2}} \begin{bmatrix} \sin \frac{k\pi}{\frac{N-1}{2} + 1} \\ \sin \frac{k2\pi}{\frac{N-1}{2} + 1} \\ \vdots \\ \sin \frac{k\frac{N-1}{2}\pi}{\frac{N-1}{2} + 1} \end{bmatrix}. \tag{131}$$

It can be shown that the λ given by (127) is not equal to that given by (130). Therefore, (126) and (129) do not have the same nonzero solution.

We have split (123) into (124) and (125), and rewritten (124) and (125), respectively, as (126) and (129). Thus, λ and $\mathbf{u}$ satisfy (123) if and only if they satisfy both (126) and (129). Since (126) and (129) do not have the same nonzero solution, λ and $\mathbf{u}$ are given either by the nonzero solution of (126) plus the trivial solution of (129), or by the nonzero solution of (129) plus the trivial solution of (126).

The above may be summarized in the form of a lemma:

Lemma A-1: For odd N, the eigenvalues and eigenvectors of $\mathbf{A}$ *may be divided into two groups. The kth eigenvalue and eigenvector, $k = 1, 2, \cdots, (N + 1)/2$, in the first group are given by* (127), (128), *and*

$$u_2 = u_4 = u_6 = \cdots = u_{(N-1)} = 0.$$

The kth eigenvalue and eigenvector, $k = 1, 2, \cdots, (N-1)/2$, in the second group are given by (130), (131), *and*

$$u_1 = u_3 = u_5 = \cdots = u_N = 0.$$

It can be shown that the eigenvectors specified in Lemma A-1 form a set of N orthonormal eigenvectors.

Case 2. Even N

Using the same method as in Case 1, one can verify the following lemma:

Lemma A-2: For even N, the eigenvalues and eigenvectors of **A** *may be divided into two groups. The kth eigenvalue and eigenvector in the first group, $k = 1, 2, \cdots, N/2$, are given by*

$$\lambda = 1 - \cos \frac{k\pi}{\frac{N}{2}+1} \tag{132}$$

and

$$\begin{bmatrix} u_1 \\ u_3 \\ \vdots \\ u_{(N-1)} \end{bmatrix} = \left(\frac{2}{\frac{N}{2}+1}\right)^{\frac{1}{2}} \begin{bmatrix} \sin \dfrac{k\pi}{\frac{N}{2}+1} \\ \sin \dfrac{k2\pi}{\frac{N}{2}+1} \\ \vdots \\ \sin \dfrac{k\frac{N}{2}\pi}{\frac{N}{2}+1} \end{bmatrix} \tag{133}$$

and

$$u_2 = u_4 = u_6 = \cdots = u_N = 0. \tag{134}$$

The kth eigenvalue and eigenvector in the second group, $k = 1, 2, \cdots, N/2$, are given by

$$\lambda = 1 - \cos \frac{k\pi}{\frac{N}{2}+1} \tag{135}$$

and

$$\begin{bmatrix} u_2 \\ \\ u_4 \\ \\ \vdots \\ \\ u_N \end{bmatrix} = \left(\frac{2}{\frac{N}{2}+1}\right)^{\frac{1}{2}} \begin{bmatrix} \sin \dfrac{k\pi}{\frac{N}{2}+1} \\ \sin \dfrac{k2\pi}{\frac{N}{2}+1} \\ \vdots \\ \sin \dfrac{k\frac{N}{2}\pi}{\frac{N}{2}+1} \end{bmatrix} \tag{136}$$

and

$$u_1 = u_3 = u_5 = \cdots = u_{(N-1)} = 0. \tag{137}$$

The eigenvectors given by (133) and (134) together with the eigenvectors given by (136) and (137) form a set of N orthonormal eigenvectors.

APPENDIX B

B.1 *Bounds on Eigenvalues*

It has been defined in Section VI that $\hat{\lambda}_i$, $i = 1$ to N, are the eigenvalues of $\mathbf{RS}^{-1}$. In this appendix we study the relation between $\hat{\lambda}_i$ and the amplitude characteristic $|F_2(f)|$ of the transmission medium. It is easier to first study the relation between $|F_2(f)|$ and the eigenvalues of $\mathbf{AS}^{-1}$. Let $\tilde{\lambda}$ denote the eigenvalue of $\mathbf{AS}^{-1}$. Since $\mathbf{AS}^{-1}$ and $\mathbf{S}^{-1}\mathbf{A}$ have the same eigenvalues, $\tilde{\lambda}$ is also the eigenvalue of $\mathbf{S}^{-1}\mathbf{A}$. Thus

$$\mathbf{S}^{-1}\mathbf{A}\tilde{\mathbf{x}} = \tilde{\lambda}\tilde{\mathbf{x}}, \tag{138}$$

where $\tilde{\mathbf{x}}$ is the eigenvector of $\mathbf{S}^{-1}\mathbf{A}$. Premultiplying both sides of (138) with $\tilde{\mathbf{x}}'\mathbf{S}$ gives

$$\tilde{\mathbf{x}}'\mathbf{A}\tilde{\mathbf{x}} = \tilde{\lambda}\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}}. \tag{139}$$

Since $\mathbf{S}$ is a correlation matrix like $\mathbf{A}$, $\mathbf{S}$ is positive definite. Consequently $\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}} > 0$ for every $\tilde{\mathbf{x}}$ except $\tilde{\mathbf{x}} = \mathbf{0}$. Since $\tilde{\mathbf{x}}$ is an eigenvector, $\tilde{\mathbf{x}} \neq \mathbf{0}$. Therefore

$$\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}} > 0.$$

From (139), we have

$$\tilde{\lambda} = \frac{\tilde{\mathbf{x}}'\mathbf{A}\tilde{\mathbf{x}}}{\tilde{\mathbf{x}}'\mathbf{S}_{\blacktriangle}}. \tag{140}$$

In this appendix, $\mathbf{c}$ is the vector defined in (4) for the conventional equalizer. Let $[(\mathbf{c}'\mathbf{A}\mathbf{c})/(\mathbf{c}'\mathbf{S}\mathbf{c})]_{\min}$ denote the minimum value of $(\mathbf{c}'\mathbf{A}\mathbf{c})/(\mathbf{c}'\mathbf{S}\mathbf{c})$ over all $\mathbf{c} \neq \mathbf{0}$, and let $[(\mathbf{c}'\mathbf{A}\mathbf{c})/(\mathbf{c}'\mathbf{S}\mathbf{c})]_{\max}$ denote the maximum value of $(\mathbf{c}'\mathbf{A}\mathbf{c})/(\mathbf{c}'\mathbf{S}\mathbf{c})$ over all $\mathbf{c} \neq \mathbf{0}$. Then since the eigenvectors $\tilde{\mathbf{x}}$ are elements of the set $\{\mathbf{c} \mid \mathbf{c} \neq \mathbf{0}\}$, we have

$$\left[\frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{S}\mathbf{c}}\right]_{\min} \leqq \frac{\tilde{\mathbf{x}}\ \mathbf{A}\tilde{\mathbf{x}}'}{\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}}} \leqq \left[\frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{S}\mathbf{c}}\right]_{\max}. \tag{141}$$

Now we evaluate the two bounds in (141). Using the notations in Section II, we can easily show that

$$\mathbf{c}'\mathbf{A}\mathbf{c} = \sum_{i=-\infty}^{\infty} y_i^2 . \tag{142}$$

Since y_i are time samples of $y(t)$ taken at the Nyquist rate, we have by sampling theorem

$$\sum_{i=-\infty}^{\infty} y_i^2 = 2(f_2 - f_1) \int_{-\infty}^{\infty} [y(t)]^2\, dt. \tag{143}$$

According to Parseval's theorem

$$\int_{-\infty}^{\infty} [y(t)]^2\, dt = \int_{-\infty}^{\infty} Y(f)\, Y^*(f)\, df, \tag{144}$$

where $Y(f)$ is the Fourier transform of $y(t)$, and $Y^*(f)$ is the complex conjugate of $Y(f)$. From (2), we obtain

$$Y(f) = X(f)e^{J2\pi fT} \sum_{k=1}^{N} c_k e^{-J2\pi fkT}. \tag{145}$$

Substituting (145) into (144) gives

$$\int_{-\infty}^{\infty} [y(t)]^2\, dt = \int_{-\infty}^{\infty} |X(f)|^2 \left[\sum_{k=1}^{N} c_k e^{-J2\pi fkT}\right]\left[\sum_{k=1}^{N} c_k e^{J2\pi fkT}\right] df. \tag{146}$$

One can rearrange (146) into the following form

$$\int_{-\infty}^{\infty} [y(t)]^2\, dt = 2 \int_{0}^{\infty} |X(f)|^2\, q(f)\, df, \tag{147}$$

where

$$q(f) = \left[\sum_{k=1}^{N} c_k \cos 2\pi fkT \right]^2 + \left[\sum_{k=1}^{N} c_k \sin 2\pi fkT \right]^2. \tag{148}$$

Since we are considering a Class IV partial-response system, we can use (30) in Section 3.1 to determine $|X(f)|$, and use (35) for the product $|F_1(f - f_c)F_3(f - f_c)|$ in $|X(f)|$. From these equations, (147) can be reduced to the following form

$$\int_{-\infty}^{\infty} [y(t)]^2 \, dt = \frac{1}{2} \int_0^{f_2 - f_1} |F_2(f - f_c)|^2 \left[\sin \frac{\pi f}{f_2 - f_1} \right]^2 q(f) \, df. \tag{149}$$

Combining (142), (143), and (149) gives

$$\mathbf{c}'\mathbf{A}\mathbf{c} = (f_2 - f_1) \int_0^{f_2 - f_1} |F_2(f - f_c)|^2 \left[\sin \frac{\pi f}{f_2 - f_1} \right]^2 q(f) \, df. \tag{150}$$

In the following discussion, we have to distinguish between the actual $|F_2(f - f_c)|$ (denoted $|F_2(f - f_c)|_{\text{act}}$) and the estimated $|F_2(f - f_c)|$ (denoted $|F_2(f - f_c)|_{\text{est}}$). In order to emphasize that the $|F_2(f - f_c)|$ in (150) is the actual $|F_2(f - f_c)|$, we rewrite (150) as

$$\mathbf{c}'\mathbf{A}\mathbf{c} = (f_2 - f_1) \int_0^{f_2 - f_1} [|F_2(f - f_c)|_{\text{act}}]^2 \left[\sin \frac{\pi f}{f_2 - f_1} \right]^2 q(f) \, df. \tag{151}$$

When the actual $|F_2(f - f_c)|$ is replaced by the estimated $|F_2(f - f_c)|$, $\mathbf{A}$ is replaced by $\mathbf{S}$, and (151) becomes

$$\mathbf{c}'\mathbf{S}\mathbf{c} = (f_2 - f_1) \int_0^{f_2 - f_1} [|F_2(f - f_c)|_{\text{est}}]^2 \left[\sin \frac{\pi f}{f_2 - f_1} \right]^2 q(f) \, df. \tag{152}$$

Let $\eta(f)$ be defined as

$$\eta(f) = \frac{[|F_2(f - f_c)|_{\text{act}}]^2}{[|F_2(f - f_c)|_{\text{est}}]^2}, \qquad 0 \leqq f \leqq f_2 - f_1 . \tag{153}$$

From (151), (152), and (153), we obtain

$$\frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{S}\mathbf{c}} = \frac{\displaystyle\int_0^{f_2 - f_1} \eta(f)\Omega(f) \, df}{\displaystyle\int_0^{f_2 - f_1} \Omega(f) \, df} \tag{154}$$

where

$$\Omega(f) = [|F_2(f - f_c)|_{\text{est}}]^2 \left[\sin \frac{\pi f}{f_2 - f_1} \right]^2 \cdot q(f). \tag{155}$$

It can be easily seen from (155) and (148) that

$$\Omega(f) \geqq 0, \qquad 0 \leqq f \leqq f_2 - f_1 . \tag{156}$$

Furthermore, for $\mathbf{c} \neq 0$, $\Omega(f)$ cannot be identically 0 in the frequency range $0 \leqq f \leqq f_2 - f_1$ (for if so $\mathbf{c}'\mathbf{S}\mathbf{c}$ would be 0, contradicting the fact that $\mathbf{S}$ is positive definite). It can be seen from (153) that

$$\eta(f) \geqq 0, \qquad 0 \leqq f \leqq f_2 - f_1 . \tag{157}$$

From these properties of $\Omega(f)$ and $\eta(f)$, one can see that

$$\frac{\displaystyle\int_0^{f_2-f_1} \eta(f)\Omega(f)\, df}{\displaystyle\int_0^{f_2-f_1} \Omega(f)\, df} \leqq [\eta(f)]_{\max} , \tag{158}$$

where $[\eta(f)]_{\max}$ is the maximum value of $\eta(f)$ in the frequency range $0 \leqq f \leqq f_2 - f_1$. It can alse be seen that

$$\frac{\displaystyle\int_0^{f_2-f_1} \eta(f)\Omega(f)\, df}{\displaystyle\int_0^{f_2-f_1} \Omega(f)\, df} \geqq [\eta(f)]_{\min} , \tag{159}$$

where $[\eta(f)]_{\min}$ is the minimum value of $\eta(f)$ in the frequency range $0 \leqq f \leqq f_2 - f_1$. Now we tie the results together. From (140), (141), (154), and (158), we have

$$\tilde{\lambda} = \frac{\tilde{\mathbf{x}}'\mathbf{A}\tilde{\mathbf{x}}}{\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}}} \leqq \left[\frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{S}\mathbf{c}}\right]_{\max} \leqq [\eta(f)]_{\max} . \tag{160}$$

Similarly, from (140), (141), (154), and (159),

$$\tilde{\lambda} = \frac{\tilde{\mathbf{x}}'\mathbf{A}\tilde{\mathbf{x}}}{\tilde{\mathbf{x}}'\mathbf{S}\tilde{\mathbf{x}}} \geqq \left[\frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{S}\mathbf{c}}\right]_{\min} \geqq [\eta(f)]_{\min} . \tag{161}$$

We may combine (160) and (161) as

$$[\eta(f)]_{\min} \leqq \tilde{\lambda} \leqq [\eta(f)]_{\max} . \tag{162}$$

It can be easily shown that the eigenvalue $\hat{\lambda}_i$ of $\mathbf{R}\mathbf{S}^{-1}$ is related to the eigenvalue $\tilde{\lambda}$ of $\mathbf{A}\mathbf{S}^{-1}$ by

$$1 - \hat{\lambda}_i = \tilde{\lambda}. \tag{163}$$

From (163) and (162), we obtain the final result

$$1 - [\eta(f)]_{\max} \leqq \hat{\lambda}_i \leqq 1 - [\eta(f)]_{\min} , \qquad i = 1 \text{ to } N. \tag{164}$$

Equations (164) and (153) are, respectively, (111) and (112) in Section VI.

REFERENCES

1. Lucky, R. W., "Automatic Equalization for Digital Communication," B.S.T.J., *44,* No. 4, part 1 (April 1965), pp. 547–588.
2. DiToro, M. J., "Communication in Time-Frequency Spread Media Using Adaptive Equalization," Proc. IEEE *56,* (October 1968), pp. 1653–1679.
3. Lucky, R. W., and Rudin, H. R., "An Automatic Equalizer for General-Purpose Communication Channels," B.S.T.J., *46,* No. 9, part 2 (November 1967), pp. 2179–2208.
4. Gersho, A., "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48,* No. 1, part 1 (January 1969), pp. 55–70.
5. Koman, C. W., unpublished work.
6. Chang, R. W., "Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication," B.S.T.J., *49,* No. 6 (July-August 1970), pp. 1069–1104.
7. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communication,* New York: McGraw-Hill Book Company, 1968.
8. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. Comm. Tech., *COM-14,* (February 1966), pp. 67–68.
9. Becker, F. K., Kretzmer, E. R., and Sheehan, J. R., "A New Signal Format for Efficient Data Transmission," B.S.T.J., *45,* No. 5, part 1 (May–June 1966), pp. 755–758.
10. Gerrish, A. M., and Lawless, W. J., "A New Wideband Partial-Response Data Set," *Conference Record,* IEEE Int. Conf. Commun., San Francisco, June 8–10, 1970.
11. Gunn, J. F., and Weller, D. C., "A Digital Mastergroup Channel for Modern Coaxial Carrier Systems," *Conference Record,* IEEE Int. Conf. Commun., San Francisco, June 8–10, 1970.
12. Lender, A., "Decision-Directed Digital Adaptive Equalization Technique for High-Speed Data Transmission," *Conference Record,* IEEE Int. Conf. Commun., San Francisco, June 8–10, 1970.
13. Forney, G. D., "Error Correction for Partial Response Modems," Int. Symp. Inform. Theory, Noordwijk, the Netherlands, June 1970.
14. Grenander, U., and Szegö, G., *Toeplitz Forms and their Applications,* Berkeley, California: University of California Press, 1958.
15. Saltzberg, B. R., "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," IEEE Trans. Inform. Theory, *IT-14,* (July 1968), pp. 563–568.
16. Kuo, F. F., and Kaiser, J. F., *System Analysis by Digital Computer,* New York: John Wiley and Sons, 1966.
17. Fleming, W. H., *Functions of Several Variables,* Reading, Mass.: Addison-Wesley Publishing Company, 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Combining Correlated Streams of Nonrandom Traffic

By SCOTTY NEAL

(Manuscript received October 6, 1970)

The Equivalent Random method is used for engineering many of the telephone overflow-networks in the Bell System. But since this method is not directly applicable to the analysis of graded-multiple trunk-groups which carry overflow traffic, we extend the method to cover such arrangements. The key to this extension is a technique for taking correlation into account when combining dependent streams of traffic which are themselves more variable than Poisson. In principle, the technique is applicable wherever a stream of overflow traffic is divided, submitted to independent trunk groups, and then recombined.

The extended Equivalent Random method provides adequate estimates of load-service relations for graded multiples which carry overflow traffic, provided the grading capacity is not substantially influenced by the network that precedes the grading.

I. INTRODUCTION

The Equivalent Random method[1,2] is used for engineering many of the telephone overflow-networks in the Bell System. However, the method is not directly applicable to the analysis of graded-multiple trunk-groups[†] which carry overflow traffic; e.g., gradings used as alternate routes in step-by-step switching systems having common control and alternate-routing capability. Apparently, Lotze[5,6] is the only author with results for estimating load-loss relations for gradings

[†] The reader should have some knowledge of graded multiples and the methods associated with the engineering of telephone overflow-networks. Some familiarity with the step-by-step switching system would also be helpful. Those not acquainted with these concepts may find it worthwhile to consult Ref. 1 for a discussion of telephone overflow-networks. An introduction to graded multiples is given in Ref. 3. Reference 4 contains a description of the pertinent aspects of the step-by-step system.

which carry overflow traffic. Unfortunately, his method requires data which cannot be obtained in a step-by-step system at reasonable cost.

We extend the Equivalent Random method to cover such applications. The key to the extension is a technique for taking correlation into account when combining dependent streams of overflow traffic. (In principle, the technique is applicable wherever a stream of overflow traffic is divided, submitted to independent trunk groups, and then recombined.) In Section II, we derive the appropriate covariance function. In order to describe how the covariance function is used to extend the Equivalent Random method,[†] we begin with an example of an application of the extended method for the analysis of a step-by-step graded multiple.

Figure 1 represents schematically a step-by-step grading which might be used as a final route. Each horizontal bar denotes one trunk (server). The traffic offered to the grading[‡] is an overflow stream from a subordinate network. The traffic is represented by the mean α and variance v of the number of simultaneous calls that would be in progress if this traffic were carried on a full-access group without blocking. The diagram is designed to indicate that an arriving call is first directed at random to one of the four first-choice subgroups; i.e., an arrival is directed to the ith subgroup with probability p_i. After reaching a particular subgroup, the call hunts vertically upward for an idle trunk. If all three trunks in the subgroup are busy, the call overflows into the second major level of the grading. The call then seizes the lowest idle trunk in the second-level subgroup. If both trunks in the second level are busy, the call overflows to the third major level containing five trunks (usually called finals). If all five finals are busy, the call leaves the system and does not return; i.e., the call is blocked and cleared. The symbol α_0 and v_0 denote respectively the mean and variance of the overflow from the grading.

In a step-by-step system, the arrangement of line-finders and selectors through which calls reach a grading causes an inherent load-balancing[4] over the first-choice subgroups; there is positive correlation between the numbers of occupied trunks in the individual subgroups. Attempting to introduce the correlation into our model, we assume

[†] The Equivalent Random method is known to be adequate for estimating load-service relation for overflow-networks having Poisson input.[7,8]

[‡] For the present study, we assume that all offered loads are constant. Hence, we are estimating single-hour capacities of gradings. Utilization of our results for normal engineering involving average busy-hour loads, would require peripheral operations to adjust the load-service relationship to reflect the effects of low, medium, or high day-to-day variations in the load.[9]

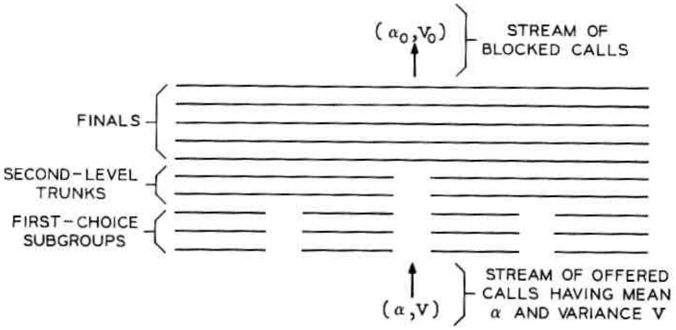


Fig. 1—Schematic representation of a step-by-step graded multiple containing 21 trunks.

the configuration given in Fig. 2. The four arrows above the full-access group denote (correlated) overflow streams caused by the corresponding input streams. The intensity of the ith input stream is $a_i = p_i a$.

At this stage, we have modeled an arrival process having mean α and variance v. The individual substreams to the first-choice subgroups of the grading are certainly correlated. How well the correla-

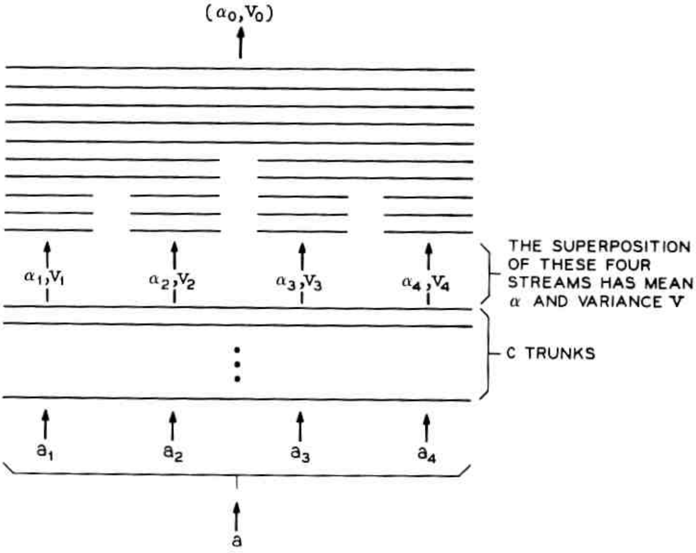


Fig. 2—A model for correlated input streams.

tion approximates the correlation that exists in an actual step-by-step system is a point which must be tested. In Section III, we show that the approximation works quite well when a large group of selectors is connected to the grading. In any event, we can view the above figure as a first approximation for the step-by-step problem. Moreover, the general features of the above configuration are not restricted to step-by-step systems.

Figure 2 also illustrates the two basic features of the Equivalent Random method. First, the method assumes that, for engineering purposes, only the first two moments of an overflow process are required. Second, the method assumes that any overflow process having mean α and variance v is adequately approximated by the overflow from a unique "equivalent system" consisting of a full-access trunk-group with Poisson input. Whenever α and v are known, standard techniques are available[1] to obtain the equivalent system of c trunks and intensity $a = a_1 + a_2 + a_3 + a_4$ of the Poisson input.

The next step consists of using results obtained independently by Descloux[10] and Lotze[6] to determine the mean α_i and the variance v_i of the overflow (due to a_i) which is submitted to the ith first-choice subgroup of our grading. This "splitting" is determined by[10]

$$\alpha_i = p_i \alpha \tag{1}$$

and

$$\frac{v_i}{\alpha_i} - 1 = p_i\left(\frac{v}{\alpha} - 1\right). \tag{2}$$

The covariance c_{ij} between the ith and jth "split" streams is given by[10]

$$c_{ij} = p_i p_j (v - \alpha). \tag{3}$$

After splitting, our system is represented as

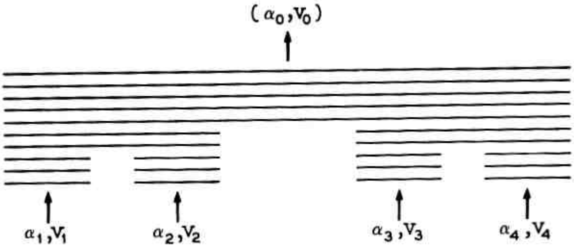


Fig. 3—Graded multiple with correlated input streams.

where it is understood that the parameters α_i and v_i of the individual substreams satisfy equations (1) through (3). Using the Equivalent Random method, we now approximate the individual substreams as overflows from full-access trunk groups. Moreover, to determine the total overflow which goes to the second major level of the grading, we view the entire system in the following manner:

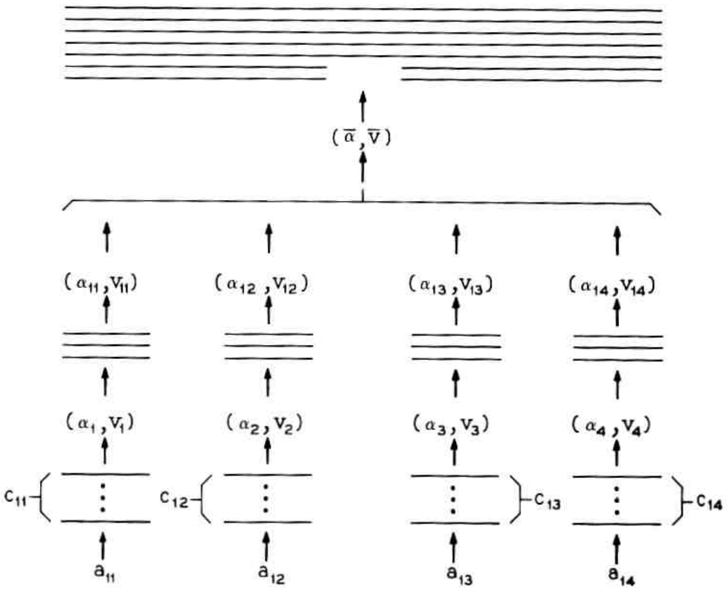


Fig. 4—First cycle of application of the extended equivalent random method.

Figure 4 indicates that an overflow of mean α_i and variance v_i results from a Poisson stream of intensity a_{1i} being offered to a full-access group of c_{1i} trunks, $i = 1, \cdots, 4$. Furthermore, the ith substream is offered to the three trunks in the ith first-choice subgroup of the grading, and causes an overflow of mean α_{1i} and variance v_{1i} to be submitted to the second major level of our grading.

One can see the main reason for looking at the grading in the manner described above: the parameters α_{1i} and v_{1i} are easily computed by[1]

$$\alpha_{1i} = a_{1i}E_{1,c_{1i}+3}(a_{1i})$$

and

$$v_{1i} = \alpha_{1i}\left[1 - \alpha_{1i} + \frac{a_{1i}}{(c_{1i} + 3) + \alpha_{1i} - a_{1i} + 1}\right],$$

where $E_{1,s}(a)$ denotes the first Erlang loss-function (Erlang-B blocking probability).

To complete the first cycle of computation, we need to obtain the mean $\bar{\alpha}$ and variance $\bar{v}$ of the total overflow submitted to the second major level of our grading. The mean $\bar{\alpha}$ is given by $\bar{\alpha} = \sum_{i=1}^{4} \alpha_{1i}$. Unfortunately, the variance $\bar{v}$ is more difficult to obtain, since the individual substreams are correlated.

To obtain $\bar{v}$ we need the covariance, cov (i, j), between the ith and jth overflow streams which are offered to the second level. A reasonable method for computing cov (i, j) has not been available in the past. Our extension of the Equivalent Random method consists of an algorithm for computing cov (i, j) in a fairly efficient fashion for many configurations of interest. A derivation of the algorithm is given in Section II.

After cov (i, j) has been determined, $\bar{v}$ is obtained from

$$\bar{v} = \sum_{i=1}^{4} v_{1i} + \sum_{\substack{i,j=1 \\ i \neq j}}^{4} \operatorname{cov}(i, j). \tag{4}$$

Having $\bar{\alpha}$ and $\bar{v}$, we reduce the system configuration to that shown in Fig. 5.

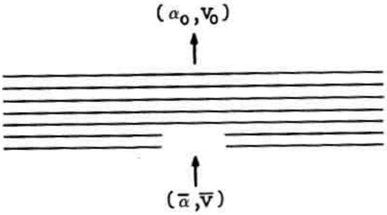


Fig. 5—Starting point for the second cycle of the extended equivalent random method.

The proportion of the traffic $(\bar{\alpha}, \bar{v})$ offered to the first subgroup of the second level is $\bar{p}_1 = (\alpha_{11} + \alpha_{12})/\bar{\alpha}$, and $\bar{p}_2 = (\alpha_{13} + \alpha_{14})/\bar{\alpha}$ is the proportion offered to the second subgroup. Consequently, one cycle of computation is completed.

Repetition of the logic described above will yield estimates of the overflow mean α_0 and variance v_0 . Moreover, the cyclic nature of the procedure allows the logic to be programmed on a digital computer so that load-service tables and other relevant information can be generated in a straightforward manner.

II. MATHEMATICAL MODEL

A service system S is composed of a collection $\mathfrak{M}$ of c first-choice servers, two groups $\mathfrak{N}_1$, $\mathfrak{N}_2$ containing d_1, d_2 second-choice servers respectively, and two last-choice groups $\mathfrak{L}_1$, $\mathfrak{L}_2$ each containing an infinite number of servers (see Fig. 6). The arrivals into the system are generated by ν independent groups of customers $G_1, \cdots, G_\nu$. The arrivals from group G_i occur according to a Poisson process with intensity a_i. All service times throughout the system are independent and have a negative-exponential distribution with unit mean.

A customer, arriving to find an idle first-choice server, selects an idle server from $\mathfrak{M}$, and service commences immediately. If an arrival from group G_i, $i = 1, 2$, occurs when all c of the first-choice servers are busy, but at least one of the d_i servers from the second-choice group $\mathfrak{N}_i$ is idle, a server is selected from $\mathfrak{N}_i$, and service commences at once. If a customer from group G_i arrives to find all the servers busy in both $\mathfrak{M}$ and $\mathfrak{N}_i$, then he is served by one of the servers from the group $\mathfrak{L}_i$. If $k \neq 1$ and $k \neq 2$, requests for service from group G_k, which occur when all c servers in $\mathfrak{M}$ are busy, are dismissed and do not return.

We assume that the system is in statistical equilibrium and define M, N_i, L_i to be the number of busy servers in $\mathfrak{M}$, $\mathfrak{N}_i$ and $\mathfrak{L}_i$ respectively (at a random instant of time) for $i = 1, 2$. Define the state of the system to be (M, N_1, N_2, L_1, L_2) with joint probability density function $f(m, n_1, n_2, l_1, l_2) =$

$$P\{M = m, N_1 = n_1, N_2 = n_2, L_1 = l_1, L_2 = l_2\}.$$

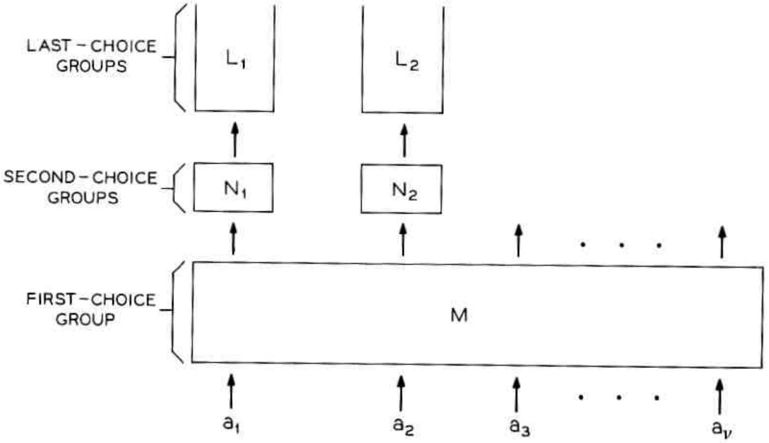


Fig. 6—System configuration.

Setting $a = a_1 + a_2 + \cdots + a_\nu$, it follows that f must satisfy relations of the following form:

For $0 \leqq m \leqq c - 1, 0 \leqq n_i \leqq d_i$, and $l_i \geqq 0$,

$$\begin{aligned}(a + m + n_1 + n_2 + l_1 + l_2)f(m, n_1, n_2, l_1, l_2) &= af(m - 1, n_1, n_2, l_1, l_2) \\ &\quad + (m + 1)f(m + 1, n_1, n_2, l_1, l_2) \\ &\quad + (n_1 + 1)f(m, n_1 + 1, n_2, l_1, l_2) \\ &\quad + (n_2 + 1)f(m, n_1, n_2 + 1, l_1, l_2) \\ &\quad + (l_1 + 1)f(m, n_1, n_2, l_1 + 1, l_2) \\ &\quad + (l_2 + 1)f(m, n_1, n_2, l_1, l_2 + 1). \qquad (5)\end{aligned}$$

Similar relations hold on the boundary of the state space $\{(m, n_1, n_2, l_1, l_2) : 0 \leqq m \leqq c, 0 \leqq n_i \leqq d_i, l_i \geqq 0\}$. We define f to be zero at all points not in the state space.

The preceding infinite set of equations is quite difficult to solve. However, when it suffices to know the various moments of the random variables L_1 and L_2, the problem can be simplified by introducing a two-dimensional binomial-moment generating-function. This function is defined by

$$\begin{aligned}B(m, n_1, n_2 ; x_1, x_2) &= \sum_{l_1=0}^{\infty} \sum_{l_2=0}^{\infty} f(m, n_1, n_2, l_1, l_2)(1 + x_1)^{l_1}(1 + x_2)^{l_2} \qquad (6)\end{aligned}$$

for $-1 \leqq x_i \leqq 0, 0 \leqq m \leqq c$, and $0 \leqq n_i \leqq d_i$. Assuming that the binomial moments

$$B_{l_1, l_2}(m, n_1, n_2) = \sum_{k_1=l_1}^{\infty} \sum_{k_2=l_2}^{\infty} \binom{k_1}{l_1}\binom{k_2}{l_2} f(m, n_1, n_2, k_1, k_2) \qquad (7)$$

exist, it follows that[†]

$$B(m, n_1, n_2 ; x_1, x_2) = \sum_{l_1=0}^{\infty} \sum_{l_2=0}^{\infty} B_{l_1, l_2}(m, n_1, n_2) x_1^{l_1} x_2^{l_2}.$$

Of course, the binomial moments $B_{l_1, l_2}(m, n_1, n_2)$ are the entities of interest since

[†] Various manipulations of these double series will be carried out in the sequel. The mathematical justification for the validity of the manipulations can be obtained from Ref. 11, Sections 5.3 through 5.5.

$$B_{(l_1,l_2)} \overset{\mathrm{d}}{=} \sum_{m=0}^{c} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} B_{l_1,l_2}(m, n_1, n_2) = E\left[\binom{L_1}{l_1}\binom{L_2}{l_2}\right]. \tag{8}$$

In particular, $B_{(1,0)} = E(L_1)$, $B_{(0,1)} = E(L_2)$, and $B_{(1,1)} = E(L_1L_2)$ so that $\operatorname{cov}(L_1, L_2) = B_{(1,1)} - B_{(1,0)}\, B_{(0,1)}$.

Relations for the binomial moments are obtained by multiplying both sides of Equation (5) (and the boundary equations which were not given) by $(1 + x_1)^{l_1}(1 + x_2)^{l_2}$ and summing on l_1 and l_2. Equating like powers of x_1 and x_2 yields the following finite system of equations:

If $0 \leqq m \leqq c - 1$ and $0 \leqq n_i \leqq d_i$,

$$\begin{aligned}(a + m + n_1 + n_2 + l_1 + l_2)&B_{l_1,l_2}(m, n_1, n_2)\\ &= aB_{l_1,l_2}(m - 1, n_1, n_2) + (m + 1)B_{l_1,l_2}(m + 1, n_1, n_2)\\ &\quad + (n_1 + 1)B_{l_1,l_2}(m, n_1 + 1, n_2)\\ &\quad + (n_2 + 1)B_{l_1,l_2}(m, n_1, n_2 + 1).\end{aligned} \tag{9}$$

For $0 \leqq n_i \leqq d_i - 1$,

$$\begin{aligned}(a_1 + a_2 + c + n_1 + n_2 + l_1 + l_2)&B_{l_1,l_2}(c, n_1, n_2)\\ &= aB_{l_1,l_2}(c - 1, n_1, n_2)\\ &\quad + a_1B_{l_1,l_2}(c, n_1 - 1, n_2) + a_2B_{l_1,l_2}(c, n_1, n_2 - 1)\\ &\quad + (n_1 + 1)B_{l_1,l_2}(c, n_1 + 1, n_2)\\ &\quad + (n_2 + 1)B_{l_1,l_2}(c, n_1, n_2 + 1).\end{aligned} \tag{10}$$

Whenever $0 \leqq n_1 \leqq d_1 - 1$,

$$\begin{aligned}(a_1 + c + n_1 + d_2 + l_1 + l_2)&B_{l_1,l_2}(c, n_1, d_2)\\ &= aB_{l_1,l_2}(c - 1, n_1, d_2) + a_1B_{l_1,l_2}(c, n_1 - 1, d_2)\\ &\quad + a_2B_{l_1,l_2}(c, n_1, d_2 - 1)\\ &\quad + (n_1 + 1)B_{l_1,l_2}(c, n_1 + 1, d_2) + a_2B_{l_1,l_2-1}(c, n_1, d_2).\end{aligned} \tag{11}$$

A similar result holds for $0 \leqq n_2 \leqq d_2 - 1$. At the extreme boundary point (c, d_1, d_2),

$$\begin{aligned}(c + d_1 + d_2 + l_1 + l_2)&B_{l_1,l_2}(c, d_1, d_2)\\ &= aB_{l_1,l_2}(c - 1, d_1, d_2) + a_1B_{l_1,l_2}(c, d_1 - 1, d_2)\\ &\quad + a_2B_{l_1,l_2}(c, d_1, d_2 - 1)\\ &\quad + a_1B_{l_1-1,l_2}(c, d_1, d_2) + a_2B_{l_1,l_2-1}(c, d_1, d_2).\end{aligned} \tag{12}$$

Since $B_{0,0}\ (m, n_1, n_2) = P\{M = m, N_1 = n_1, N_2 = n_2\}$, it follows that

$$\sum_{m=0}^{c} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} B_{0,0}(m, n_1, n_2) = 1. \tag{13}$$

We set $B_{l_1,l_2}(m, n_1, n_2) = 0$ for any point (l_1, l_2, m, n_1, n_2) not in the set

$$\{(l_1, l_2, m, n_1, n_2):\ l_i \geqq 0,\ \ 0 \leqq m \leqq c,\ \ 0 \leqq n_i \leqq d_i\}.$$

A very useful relation can be obtained by summing all of equations (9) through (12) to obtain

$$(l_1 + l_2)B_{(l_1,l_2)} = a_1 \sum_{n_2=0}^{d_2} B_{l_1-1,l_2}(c, d_1, n_2) + a_2 \sum_{n_1=0}^{d_1} B_{l_1,l_2-1}(c, n_1, d_2). \tag{14}$$

Consequently, $B_{(1,1)} = E\{L_1L_2\}$ can be obtained from

$$q_1(m, n_2) \overset{\mathrm{d}}{=} \sum_{n_1=0}^{d_1} B_{1,0}(m, n_1, n_2) \tag{15}$$

and

$$q_2(m, n_1) \overset{\mathrm{d}}{=} \sum_{n_2=0}^{d_2} B_{0,1}(m, n_1, n_2). \tag{16}$$

That is,

$$2B_{(1,1)} = a_1q_2(c, d_1) + a_2q_1(c, d_2). \tag{17}$$

Relations for q_1 and q_2 are obtained directly from equations (9) through (12) (see Appendix B). However, the relations require $B_{0,0}(c, n_1, d_2)$ and $B_{0,0}(c, d_1, n_2)$ for $0 \leqq n_i \leqq d_i$. (See Refs. 12 and 13 for a related problem.) Setting $l_1 = l_2 = 0$ in equations (9) through (13) yields, with equation (14), a system of $(c + 1)\ (d_1 + 1)\ (d_2 + 1)$ independent linear equations for $B_{0,0}$ which in principle can be solved numerically. Unfortunately, for the step-by-step applications described in the introduction, c can be quite large (20 or more) although d_1 and d_2 normally do not exceed 5. Consequently, systems of 500 and more equations would not be uncommon. Since a solution might be required for several sets of parameters in any particular network, a direct numerical solution is not attractive.

In Appendix A, we obtain a closed-form expression for $B_{0,0}\ (m, n_1, n_2)$ in terms of the $(d_1 + 1)\ (d_2 + 1)$ constants

$$\beta(n_1\,,n_2) = E\left[\binom{M}{c}\binom{N_1}{n_1}\binom{N_2}{n_2}\right], \qquad 0 \leqq n_i \leqq d_i\,. \tag{18}$$

Furthermore, these constants satisfy $(d_1 + 1)\ (d_2 + 1)$ independent linear relations, so that a numerical solution is quite practical. In fact, the *maximum* size of the system of equations requiring solution for step-by-step systems is reduced from more than 500 to 36.

A closed-form expression for $q_1(c,\, d_2)$ and $q_2(c,\, d_1)$ is derived in Appendix B. These results combine into the following computational algorithm for $\operatorname{cov}(L_1\,, L_2)$: First of all,

$$\beta(0,\,0) = E_{1,c}(a) \tag{19}$$

(the Erlang-B blocking probability for the first-choice group), and for $0 \leqq n_i \leqq d_i\,,\ n_1 + n_2 > 0$,

$$\begin{aligned}(n_1 + n_2)\nu_{n_1+n_2}\beta(n_1\,,n_2)&\\ = a_1\beta(n_1 - 1,\,n_2) &+ a_2\beta(n_1\,,n_2 - 1)\\ - a_1\binom{d_1}{n_1 - 1}\beta(d_1\,,n_2) &- a_2\binom{d_2}{n_2 - 1}\beta(n_1\,,d_2).\end{aligned} \tag{20}$$

By definition, $\beta(n_1\,,-\,1) = \beta(-\,1,\,n_2) = 0$ for $0 \leqq n_i \leqq d_i$. The numbers ν_n are intimately related to various aspects of overflow systems[1] and satisfy the following recurrence relation:

$$\frac{1}{\nu_0} = E_{1,c}(a) \tag{21}$$

and

$$\nu_n = \frac{a}{n\nu_{n-1}} + 1 + \frac{c-a}{n} \quad \text{for} \quad n \geqq 1. \tag{22}$$

Then, (from Appendix B)

$$q_1(c,\,d_2) = \frac{\dfrac{a_1}{a_2}\displaystyle\sum_{j=0}^{d_2}\left[\beta(d_1\,,j)\prod_{k=j+1}^{d_2+1}\frac{a_2}{k\nu_k}\right]}{1 + \displaystyle\sum_{j=1}^{d_2}\left[\binom{d_2}{j-1}\prod_{k=j+1}^{d_2+1}\frac{a_2}{k\nu_k}\right]} \tag{23}$$

and

$$q_2(c,\,d_1) = \frac{\dfrac{a_2}{a_1}\displaystyle\sum_{j=0}^{d_1}\left[\beta(j,\,d_2)\prod_{k=j+1}^{d_1+1}\frac{a_1}{k\nu_k}\right]}{1 + \displaystyle\sum_{j=1}^{d_1}\left[\binom{d_1}{j-1}\prod_{k=j+1}^{d_1+1}\frac{a_1}{k\nu_k}\right]}. \tag{24}$$

From Appendix A,

$$E\{L_1\} = a_1\beta(d_1, 0) \quad \text{and} \quad E\{L_2\} = a_2\beta(0, d_2), \tag{25}$$

so that

$$2 \operatorname{cov}(L_1, L_2) = a_1q_2(c, d_1) + a_2q_1(c, d_2) - 2a_1a_2\beta(d_1, 0)\beta(0, d_2). \tag{26}$$

Equations (19) through (26) constitute an algorithm for the computation of $\operatorname{cov}(L_1, L_2)$.

Whenever

$$a_1 = a_2 \quad \text{and} \quad d_1 = d_2$$

the symmetry of the problem (see Fig. 1 and equation (18)) implies that

$$\beta(n_1, n_2) = \beta(n_2, n_1). \tag{27}$$

The symmetry required by (27) would prevail for most step-by-step graded multiples. In such cases (27) can be used to reduce the number of equations to $\frac{1}{2}(d_1 + 1)(d_1 + 2)$. Consequently, the dimensions of the systems of equations needed for an analysis of most step-by-step graded multiples would not exceed 21. Such systems can be solved very efficiently by numerical matrix inversion.

III. NUMERICAL RESULTS

In order to establish a base for comparison, we used a simulation to obtain load-service relations for a 25-trunk and a 45-trunk step-by-step graded multiple. We obtained results for several values of variance-to-mean ratio $z = v/\alpha$ and several selector configurations. We found that the extended Equivalent Random method furnished adequate estimates of blocking probability in each case where the maximum number of selectors was used. However, as the number of selectors was reduced for a particular grading, the inherent load-balancing[4] caused actual grading capacity to be higher than indicated by the extended Equivalent Random method. Consequently, we conclude that *the extended Equivalent Random method provides adequate estimates of load-service relations for graded multiples which carry overflow traffic, provided that the network through which calls reach the grading does not significantly influence grading capacity.*

In view of the effort required to obtain the covariance $\operatorname{cov}(L_i, L_j)$, one naturally questions the necessity of accounting for the correlation which occurs in our problem. In fact, on several occasions, we encountered the following question: What sort of results would be obtained if both mean and variance were split by proportion (i.e., $\alpha_i = p_i\alpha$ and $v_i = p_iv$) when splitting is required, and $\text{cov}(L_i, L_j)$ were assumed to be zero whenever a variance recombination is required?

In order to consider the preceding question, as well as to obtain a better understanding of the behavior of $\text{cov}(L_i, L_j)$, three computer programs were written to generate load-loss relations for step-by-step graded multiples. These programs were based respectively on the following assumptions:

(*i*) The input traffic is completely random, i.e., Poisson.[†]

(*ii*) The input traffic is nonrandom. Mean and variance are split by proportion when required and cov (L_i, L_j) is assumed to be zero.[‡]

(*iii*) The input traffic is nonrandom, and the gradings are analyzed by using the extended Equivalent Random method.

Throughout the study, the offered traffic was assumed to be balanced over the subgroups of any grading under consideration (i.e., $p_i = p_j$ for all i, j).

For comparison, we used each of the three assumptions to compute load-service relations for a 25-trunk and a 45-trunk graded multiple. The results are displayed in Fig. 7. For these examples, the variance-to-mean ratio of the nonrandom offered traffic was held constant at 2.25.

The different results arising from assumptions (*ii*) and (*iii*) were surprising. It was originally felt by some that assumption (*ii*) would cause over-trunking, but not by the amounts indicated. For example, notice that the 45-trunk grading yields a B.01 blocking probability for an offered load of 330 ccs (9.17 erlangs) with a variance-to-mean ratio of 2.25 when the correlation is neglected as outlined in assumption (*ii*). However, Fig. 7 indicates that the same traffic can actually be handled at B.01 on the 25-trunk grading when the correlation is taken into account. Hence, assumption (*ii*) leads to at least 80 percent overtrunking for the example. An examination of other portions of the curves yields similar results. Consequently assumption (*ii*) must be discarded.

Lower bounds to the load-loss relations for nonrandom traffic

[†] Assumption (*i*) is known to cause an underprovision of trunks.[1]

[‡] Assumption (*ii*) was considered by many to be the most natural approach to improving on assumption (*i*).

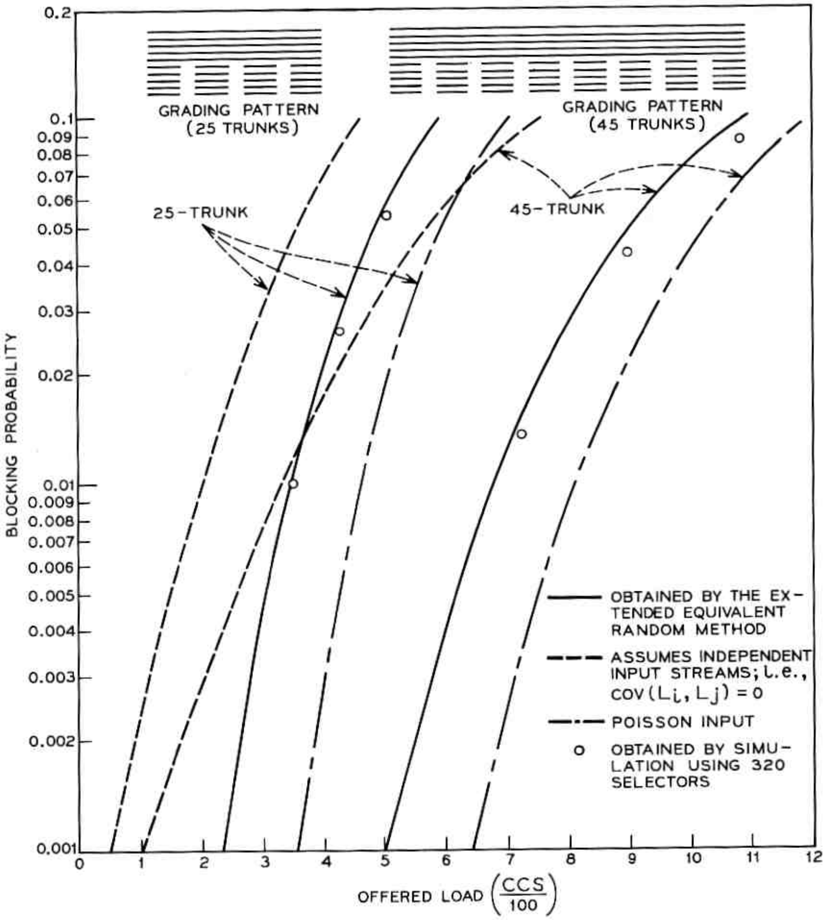


Fig. 7—Load-loss relations for the 25-trunk and the 45-trunk graded multiples. Variance-to-mean ratio of the offered load is $z = 2.25$.

result from assumption (*i*) as illustrated in Fig. 7. For these examples, undertrunking by 18 to 25 percent results from approximation (*i*). Since the disparity will increase for larger variance-to-mean ratios, assumption (*i*) does not seem applicable either.

Two other approximations were also tried but the results were very poor. The first used the correct splitting equations (1) through (3) but assumed $\text{cov}(L_i, L_j) = 0$. As a result, some of the load-service curves actually intersected each other. The second also used the correct splitting formulas but assumed the correlation coefficient.

$$\rho(L_i\ ,\ L_j) = \frac{\operatorname{cov}\ (L_i\ ,\ L_j)}{[\operatorname{var}\ (L_i)\ \operatorname{var}\ (L_j)]^{\frac{1}{2}}}$$

to be constant at recombination points and equal to the correlation coefficient for the split (nonrandom) offered traffic. The predicted blocking resulting from the last approximation actually *decreased* as the intensity of the offered traffic increased.

IV. CONCLUSIONS

We have presented a technique for taking correlation into account when combining certain dependent streams of overflow traffic. The result was used to define an extension of the Equivalent Random method; an engineering approximation for estimating the capacities of overflow networks.

The extended Equivalent Random method yields good estimates of load-service relations for graded multiples which carry overflow traffic provided the network through which calls reach the grading does not significantly affect grading capacity. Consequently, for application to step-by-step gradings, the extended method is restricted to gradings which are connected to large groups of selectors. We are currently investigating techniques which would allow us to consider systems which do influence grading capacity.

V. ACKNOWLEDGMENTS

I am grateful to Mrs. H. J. Tauson for writing the computer programs which were necessary for this study. My understanding of basic aspects of step-by-step switching systems is primarily due to the patience of M. F. Morse and R. I. Wilkinson. Discussions with P. J. Burke were very helpful in establishing a proper mathematical model. Burke also pointed out the original work of Kosten[14] which was the key to the solutions presented in Appendixes A and B.

APPENDIX A

System State Probabilities

In this section, we obtain the solution for the system of equations (9) through (13). We use a generalization of a technique employed by Kosten.[14]

For notational simplicity, let $p(m, n_1, n_2) = B_{0,0}(m, n_1, n_2)$, and let $p_1(m, n_1, n_2)$ denote any function which belongs to the family $\mathcal{a}$ of functions satisfying equation (9) for $0 \leqq m < \infty$ and $0 \leqq n_i \leqq d_i$. Define

$$P(t, q_1, q_2) = \sum_{m=0}^{\infty} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} p_1(m, n_1, n_2) t^m q_1^{n_1} q_2^{n_2}. \tag{28}$$

It follows from equation (9) that

$$(1 - t)\frac{\partial P}{\partial t} + (1 - q_1)\frac{\partial P}{\partial q_1} + (1 - q_2)\frac{\partial P}{\partial q_2} = a(1 - t)P. \tag{29}$$

This linear first-order partial differential equation can be solved by (Lagrange's) method of characteristics (see Ref. 15, Chap. 2). The characteristic equations are

$$\frac{dt}{1 - t} = \frac{dq_1}{1 - q_1} = \frac{dq_2}{1 - q_2} = \frac{dP}{a(1 - t)P}$$

with solutions

$$\frac{1 - q_i}{1 - t} = k_i \quad \text{and} \quad k = e^{-at}P$$

where k and k_i are arbitrary constants. Hence, the solution to equation (29) is given by

$$P(t, q_1, q_2) = e^{at} H\left(\frac{1 - q_1}{1 - t}, \frac{1 - q_2}{1 - t}\right) \tag{30}$$

where $H(z_1, z_2)$ denotes any analytic function of the arguments z_1, z_2. From (28), it follows that the Taylor series for H must be finite, and so

$$P(t, q_1, q_2) = e^{at} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1, n_2)\left(\frac{1 - q_1}{1 - t}\right)^{n_1}\left(\frac{1 - q_2}{1 - t}\right)^{n_2}. \tag{31}$$

Following the notation of Riordan (Ref. 1, p. 89), let $\{\sigma_k(m): m = 0, 1, \cdots\}$ be the sequence with generating function $(1 - t)^{-k} \exp(at)$, that is,

$$\sum_{m=0}^{\infty} \sigma_k(m) t^m = \frac{e^{at}}{(1 - t)^k}. \tag{32}$$

The variables $\sigma_k(m)$ satisfy the following recurrence relations.[1,14] For $m \geqq 0$ and $k \geqq 0$,

$$m\sigma_k(m) = a\sigma_k(m - 1) + k\sigma_{k+1}(m - 1), \tag{33}$$

$$k\sigma_{k+1}(m) = (k + m - a)\sigma_k(m) + a\sigma_{k-1}(m), \tag{34}$$

and

$$\sum_{m=0}^{j} \sigma_k(m) = \sigma_{k+1}(j). \tag{35}$$

It is convenient to define

$$\sigma_{-1}(m) = \sigma_k(-1) = 0. \tag{36}$$

Hence, from Equations (31) and (32),

$$P(t, q_1, q_2)$$

$$= \sum_{m=0}^{\infty} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1, n_2)\sigma_{n_1+n_2}(m)t^m(1 - q_1)^{n_1}(1 - q_2)^{n_2} \tag{37}$$

$$= \sum_{m=0}^{\infty} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \Bigg[(-1)^{n_1+n_2} \sum_{k_1=n_1}^{d_1} \sum_{k_2=n_2}^{d_2} \begin{bmatrix}k_1\\ n_1\end{bmatrix}\begin{bmatrix}k_2\\ n_2\end{bmatrix}\alpha(k_1, k_2)$$

$$\cdot\sigma_{k_1+k_2}(m)\Bigg]t^m q_1^{n_1} q_2^{n_2}. \tag{38}$$

Comparing equations (29) and (38), one can see that the functions p_1 in $\mathcal{A}$ are of the form

$$p_1(m, n_1, n_2) = (-1)^{n_1+n_2} \sum_{k_1=n_1}^{d_1} \sum_{k_2=n_2}^{d_2} \begin{pmatrix}k_1\\ n_1\end{pmatrix}\begin{pmatrix}k_2\\ n_2\end{pmatrix}\alpha(k_1, k_2)\sigma_{k_1+k_2}(m), \tag{39}$$

and are determined up to the $(d_1 + 1)$ $(d_2 + 1)$ constants $\{\alpha(n_1, n_2)$: $0 \leqq n_i \leqq d_i\}$. There are $(d_1 + 1)$ $(d_2 + 1)$ independent linear equations among equations (10) through (13), and so it follows that there is exactly one function, p^*, in $\mathcal{A}$ which satisfies equations (9) through (13). The appropriate restriction of p^* must be p. The remainder of this section is devoted to a derivation of the relations which the constants $\{\alpha(n_1, n_2)\}$ must satisfy in order to obtain the solution.

An equivalent but less complex set of boundary conditions is obtained by putting $m = c$ in (9) and subtracting equations (2) through (13) respectively. Hence, if $0 \leqq n_i \leqq d_i - 1$,

$$(a - a_1 - a_2)p(c, n_1, n_2) + a_1p(c, n_1 - 1, n_2) + a_2p(c, n_1, n_2 - 1)$$

$$= (c + 1)p(c + 1, n_1, n_2), \tag{40}$$

for $0 \leqq n_1 \leqq d_1 - 1$,

$$(a - a_1)p(c, n_1 , d_2) + a_1 p(c, n_1 - 1, d_2) + a_2 p(c, n_1 , d_2 - 1) = (c + 1)p(c + 1, n_1 , d_2), \tag{41}$$

and a similar relation holds for $0 \leqq n_2 \leqq d_2 - 1$. Finally,

$$ap(c, d_1 , d_2) + a_1 p(c, d_1 - 1, d_2) + a_2 p(c, d_1 , d_2 - 1) = (c + 1)p(c + 1, d_1 , d_2). \tag{42}$$

Now, define

$$P_m(q_1 , q_2) = \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} p(m, n_1 , n_2) q_1^{n_1} q_2^{n_2}, \tag{43}$$

$$G_{m,n_1}(q_2) = \sum_{n_2=0}^{d_2} p(m, n_1 , n_2) q_2^{n_2}, \tag{44}$$

and

$$H_{m,n_2}(q_1) = \sum_{n_1=0}^{d_1} p(m, n_1 , n_2) q_1^{n_1}. \tag{45}$$

It follows from equations (40) through (42) that

$$[a - a_1(1 - q_1) - a_2(1 - q_2)]P_c(q_1 , q_2) + a_1(1 - q_1)q_1^{d_1} G_{c,d_1}(q_2) + a_2(1 - q_2)q_2^{d_2} H_{c,d_2}(q_1) = (c + 1)P_{c+1}(q_1 , q_2). \tag{46}$$

Equations (39) and (43) through (45) imply that

$$P_m(q_1 , q_2) = \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1 , n_2)\sigma_{n_1+n_2}(m)(1 - q_1)^{n_1}(1 - q_2)^{n_2}, \tag{47}$$

$$G_{c,d_1}(q_2) = (-1)^{d_1} \sum_{n_2=0}^{d_2} \alpha(d_1 , n_2)\sigma_{d_1+n_2}(c)(1 - q_2)^{n_2}, \tag{48}$$

and

$$H_{c,d_2}(q_1) = (-1)^{d_2} \sum_{n_1=0}^{d_1} \alpha(n_1 , d_2)\sigma_{n_1+d_2}(c)(1 - q_1)^{n_1}. \tag{49}$$

Using the identity $q_i = 1 - (1 - q_i)$ and the relations (47) through (49) in (46) obtains

$$a \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1 , n_2)\sigma_{n_1+n_2}(c)(1 - q_1)^{n_1}(1 - q_2)^{n_2}$$
$$- a_1 \sum_{n_1=1}^{d_1+1} \sum_{n_2=0}^{d_2} \alpha(n_1 - 1, n_2)\sigma_{n_1+n_2-1}(c)(1 - q_1)^{n_1}(1 - q_2)^{n_2}$$

$$- a_2 \sum_{n_1=0}^{d_1} \sum_{n_2=1}^{d_2+1} \alpha(n_1 , n_2 - 1)\sigma_{n_1+n_2-1}(c)(1 - q_1)^{n_1}(1 - q_2)^{n_2}$$

$$- (-1)^{d_1} a_1 \sum_{n_1=1}^{d_1+1} \sum_{n_2=0}^{d_2} (-1)^{n_1} \left[\begin{matrix} d_1 \\ n_1 - 1 \end{matrix}\right]$$

$$\cdot \alpha(d_1 , n_2)\sigma_{d_1+n_2}(c)(1 - q_1)^{n_1}(1 - q_2)^{n_2}$$

$$- (-1)^{d_2} a_2 \sum_{n_1=0}^{d_1} \sum_{n_2=1}^{d_2+1} (-1)^{n_2} \left[\begin{matrix} d_2 \\ n_1 - 1 \end{matrix}\right]$$

$$\cdot \alpha(n_1 , d_2)\sigma_{n_1+d_2}(c)(1 - q_1)^{n_1}(1 - q_2)^{n_2}$$

$$= (c + 1) \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1 , n_2)\sigma_{n_1+n_2}(c + 1)(1 - q_1)^{n_1}(1 - q_2)^{n_2}.$$

Equating the coefficients of $(1 - q_1)^{n_1}(1 - q_2)^{n_2}$ yields

$$[a\sigma_{n_1+n_2}(c) - (c + 1)\sigma_{n_1+n_2}(c + 1)]\alpha(n_1 , n_2)$$

$$= [a_1\alpha(n_1 - 1, n_2) + a_2\alpha(n_1 , n_2 - 1) + a_2\alpha(n_1 , n_2 - 1]\sigma_{n_1+n_2-1}(c)$$

$$+ (-1)^{d_1+n_1} a_1 \left[\begin{matrix} d_1 \\ n_1 - 1 \end{matrix}\right] \alpha(d_1 , n_2)\sigma_{d_1+n_2}(c)$$

$$+ (-1)^{n_2+d_2} a_2 \left[\begin{matrix} d_2 \\ n_2 - 1 \end{matrix}\right] \alpha(n_1 , d_2)\sigma_{n_1+d_2}(c). \tag{50}$$

Using (33), we see that

$$a\sigma_{n_1+n_2}(c) - (c + 1)\sigma_{n_1+n_2}(c + 1) = -(n_1 + n_2)\sigma_{n_1+n_2+1}(c). \tag{51}$$

It is worthwhile to define

$$\beta(n_1 , n_2) = (-1)^{n_1+n_2}\alpha(n_1 , n_2)\sigma_{n_1+n_2}(c), \quad \text{and} \quad n = n_1 + n_2 . \tag{52}$$

Now, substitute (51) and (52) into (50) to obtain

$$n \frac{\sigma_{n+1}(c)}{\sigma_n(c)} \beta(n_1 , n_2) = a_1\beta(n_1 - 1, n_2) + a_2\beta(n_1 , n_2 - 1)$$

$$- a_1 \left[\begin{matrix} d_1 \\ n_1 - 1 \end{matrix}\right] \beta(d_1 , n_2)$$

$$- a_2 \left[\begin{matrix} d_2 \\ n_2 - 1 \end{matrix}\right] \beta(n_1 , d_2) \tag{53}$$

for $0 \leqq n_i \leqq d_i$ and $n > 0$. (Both sides vanish for $n_1 = n_2 = 0$.)

One additional relation is required. Using (39),

$$\sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} p(m, n_1, n_2) = \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} \alpha(n_1, n_2)\sigma_{n_1+n_2}(m) \cdot \left[\sum_{k_1=0}^{n_1} \binom{n_1}{k_1}(-1)^{k_1}\right] \left[\sum_{k_2=0}^{n_2} \binom{n_2}{k_2}(-1)^{k_2}\right] = \alpha(0, 0)\sigma_0(m).$$

Consequently, noting relation (36), one obtains

$$1 = \sum_{m=0}^{c} \sum_{n_1=0}^{d_1} \sum_{n_2=0}^{d_2} p(m, n_1, n_2) = \alpha(0, 0) \sum_{m=0}^{c} \sigma_0(m) = \alpha(0, 0)\sigma_1(c).$$

Thus,

$$\alpha(0, 0) = \frac{1}{\sigma_1(c)}$$

and

$$\beta(0, 0) = \frac{\sigma_0(c)}{\sigma_1(c)} = E_{1,c}(a); \tag{54}$$

i.e., $\beta_{(0,0)}$ is the Erlang-B blocking probability (also known as the first Erlang loss-function) for the first-choice trunk group.

Equations (53) and (54) completely determine $\beta(n_1, n_2)$ for $0 \leqq n_i \leqq d_i$. Using (52) and (39), we see that the state probabilities are given by

$$p(m, n_1, n_2) = \sum_{k_1=n_1}^{d_1} \sum_{k_2=n_2}^{d_2} (-1)^{k_1-n_1}(-1)^{k_2-n_2}\binom{k_1}{n_1}\binom{k_2}{n_2}\beta(k_1, k_2) \frac{\sigma_{k_1+k_2}(m)}{\sigma_{k_1+k_2}(c)}. \tag{55}$$

Using the relation

$$\binom{k}{m}\binom{m}{n} = \binom{k}{n}\binom{k-n}{m-n}$$

it is straightforward to show that if

$$\Lambda_{(m,n_1,n_2)} = \sum_{i=m}^{c} \sum_{k_1=n_1}^{d_1} \sum_{k_2=n_2}^{d_2} \binom{i}{m}\binom{k_1}{n_1}\binom{k_1}{n_1} p(i, k_1, k_2)$$

for $0 \leqq m \leqq c$ and $0 \leqq n_i \leqq d_i$, then an inverse relation is given by

$$p(m, n_1, n_2) = \sum_{i=m}^{c} \sum_{k_1=n_1}^{d_1} \sum_{k_2=n_2}^{d_2} (-1)^{i-m}(-1)^{k_1-n_1}(-1)^{k_2-n_2} \binom{i}{m} \binom{k_1}{n_1} \binom{k_2}{n_2} \cdot \Lambda_{(i,k_1,k_2)} \tag{56}$$

for $0 \leqq m \leqq c$ and $0 \leqq n_i \leqq d_i$.

Consequently, comparing (55) and (56) we see that

$$\beta(n_1, n_2) = \Lambda_{(c,n_1,n_2)} = E\left\{\binom{M}{c}\binom{N_1}{n_1}\binom{N_2}{n_2}\right\}. \tag{57}$$

Equations (14) and (55), imply that

$$E\{L_1\} = a_1\beta(d_1, 0), \tag{58}$$

and

$$E\{L_2\} = a_2\beta(0, d_2). \tag{59}$$

For the computation of $\beta(n_1, n_2)$ using (53), the ratio

$$\nu_n = \frac{\sigma_{n+1}(c)}{\sigma_n(c)}$$

is required for $n \geqq 1$. A recursive relation for the ratio is obtained from (32) via

$$\frac{\sigma_{n+1}(c)}{\sigma_n(c)} = \left[1 + \frac{c-a}{n}\right] + \frac{a}{n}\frac{\sigma_{n-1}(c)}{\sigma_n(c)};$$

that is,

$$\nu_n = \frac{a}{n}\left(\frac{1}{\nu_{n-1}}\right) + \frac{c-a}{n} + 1 \quad \text{for} \quad n \geqq 1. \tag{60}$$

The first-order (nonlinear) algorithm is initiated with

$$\frac{1}{\nu_0} = \frac{\sigma_0(c)}{\sigma_1(c)} = E_{1,c}(a); \tag{61}$$

i.e., the Erlang-B blocking probability for the first-choice trunk group.

APPENDIX B ·

A Conditional Mean

In this appendix, a formula is obtained for the computation of[†]

$$q\,(m, n) = \sum_{n_1=0}^{d_1} B_{1,0}(m, n_1\ , n)$$
$$= E\{L_1 \mid M = m, N_2 = n\}P\{M = m, N_2 = n\}. \qquad (62)$$

From equations (9) through (12), it follows that for $0 \leqq m \leqq c - 1$ and $0 \leqq n \leqq d_2$,

$$(a + m + n + 1)q(m, n) = aq(m - 1, n) + (m + 1)q(m + 1, n) + (n + 1)q(m, n + 1) \qquad (63)$$

and for $0 \leqq n \leqq d_2 - 1$,

$$(a_2 + c + n + 1)q(c, n) = aq(c - 1, n) + a_2q(c, n - 1) + (n + 1)q(c, n + 1) + a_1B_{0,0}(c, d_1\ , n). \qquad (64)$$

Also,

$$(c + d_2 + 1)q(c, d_2) = aq(c - 1, d_2) + a_2q(c, d_2 - 1) + a_1B_{0,0}(c, d_1\ , d_2). \qquad (65)$$

Using the methods and results presented in Appendix A, we can show that

$$q(m, n) = \sum_{k=n}^{d_2} (-1)^{n-k}\omega_k \frac{\sigma_{k+1}(m)}{\sigma_{k+1}(c)} \quad \text{for} \quad 0 \leqq m \leqq c \quad \text{and} \quad 0 \leqq n \leqq d_2\ , \qquad (66)$$

where $\omega_{-1} = 0$, and

$$(n + 1)\nu_{n+1}\omega_n = a_2\omega_{n-1} - a_2 \begin{pmatrix} d_2 \\ n - 1 \end{pmatrix} \omega_{d_2} + a_1\beta(d_1\ , n) \quad \text{for} \quad 0 \leqq n \leqq d_2\ . \qquad (67)$$

In particular, $q(c, d_2) = \omega_{d_2}$, and can be obtained from (67) in the following closed form:

[†] Throughout this appendix, the subscript 1 on q_1 has been omitted for notational simplicity.

$$q(c, d_2) = \frac{\dfrac{a_1}{a_2} \displaystyle\sum_{j=0}^{d_2} \left[\beta(d_1\, , j) \prod_{k=j+1}^{d_2+1} \frac{a_2}{k\nu_k} \right]}{1 + \displaystyle\sum_{j=1}^{d_2} \left[\binom{d_2}{j-1} \prod_{k=j+1}^{d_2+1} \frac{a_2}{k\nu_k} \right]}. \tag{68}$$

A similar expression is valid for $q_2(c, d_1)$.

REFERENCES

1. Wilkinson, R. I., (Appendix, by Riordan, J.), "Theories for Toll Traffic Engineering in the U.S.A." B.S.T.J., *35,* No. 2 (March 1956), pp. 421–514.
2. Bretschneider, G., "Die Berechnung von Leitungsgruppen für überfliessenden Verkehr in Fernsprechwählanlagen," Nachr. Techn. Zeitschr. *9* (1956), pp. 533–540.
3. Wilkinson, R. I., "The Interconnection of Telephone Systems—Graded Multiples," B.S.T.J., *10,* No. 4 (October 1931), pp. 531–564.
4. Buchner, M. M., Jr., and Neal, S. R., "Inherent Load-Balancing in Step-by-Step Switching Systems," B.S.T.J., *50,* No. 1 (January 1971), pp. 135–165.
5. Lotze, A., "History and Development of Grading Theory," Proceedings of the Fifth International Teletraffic Congress, Rockefeller Univ., New York, June 14–20, 1967, pp. 148–161.
6. Lotze, A., "A Traffic Variance Method for Gradings of Arbitrary Type," Proc. Fourth Int. Teletraffic Congress, Document 80, London, July 15–21, 1964.
7. Wilkinson, R. I., "Notes on Comparison of British Post Office Simulation Tests on Graded Multiples, with Equivalent Random Theory," unpublished work.
8. McGrath, H. T., "The Electronic Analyzer—a Survey of its Use and Limitations," Post Office E. E. Journal, *52,* Part 1 (April 1959), pp. 19–25.
9. Wilkinson, R. I., "A Study of Load and Service Variations in Toll Alternate Route Systems," Proc. Second Int. Teletraffic Congress, *3,* Paper 29, The Hague, July 7–11, 1958.
10. Descloux, A., "On the Components of Overflow Traffic," unpublished work.
11. Hille, E., *Analytic Function Theory,* Vol. 1, New York: Ginn, 1959.
12. Brockmeyer, E., "The Simple Overflow Problem in the Theory of Telephone Traffic," Teleteknik *1,* No. 1, 1957, pp. 92–105.
13. Wallström, B., "Congestion Studies in Telephone Systems with Overflow Facilities," Ericsson Technics, *3,* 1966, pp. 190–351.
14. Kosten, L., "Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen," Elktr. Nachr.-Techn., *14,* No. 1, 1937, pp. 5–12.
15. Garabedian, P. R., *Partial Differential Equations,* New York: Wiley, 1964, pp. 18–22.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

Phase and Amplitude Variations in Multipath Fading of Microwave Signals

By K. BULLINGTON

(Manuscript received November 10, 1970)

Experimental data on the duration of individual fades in a 4-GHz radio signal are used to derive the rates of change in amplitude and phase of a signal propagated through a 30-mile multipath medium. A by-product is a relation between the rate of change in phase and the effective vertical velocity of the atmospheric irregularities that can be considered to be the cause of the multipath phenomena. The derived distribution of velocities is then used to predict the variations in fade duration over a wide range of frequency and path length.

I. INTRODUCTION

Multipath fading on line-of-sight paths causes substantial variations in the amplitude and phase of the received signal. Theoretically, if the response to an impulse function were known *with sufficient accuracy*, all magnitudes and rates of change could be computed. In practice, this possibility is somewhat of an illusion because many different impulse responses occur on any given path, with significant diurnal and seasonal variations. Moreover, the desired accuracy is generally not achieved, because either the required impulse function needs to be only one cycle of the radio frequency or extensive correlation techniques are needed to achieve a comparable result.

The quantitative estimates given below are based on elementary theory and on the measured distribution of fade durations. The results agree with (or at least do not contradict) the available experimental data.

II. GENERAL CONSIDERATIONS

The signal received through a multipath medium can be represented by the sum of the principal wave plus the delayed components $(a_m e^{i\Delta t_m})$:

$$
\begin{aligned}
E &= Ae^{i(\omega+\Delta\omega)t}\left(1 + \sum_{m=1}^{M} a_m e^{i(\omega+\Delta\omega)\Delta t_m}\right) \\
&= Ave^{i(\omega+\Delta\omega)t}
\end{aligned} \tag{1}
$$

where a_m and Δt_m represent the magnitude and relative time delay of the mth "echo."

The effect of frequency deviation, $\Delta\omega/2\pi$, is discussed in a later section on the variation in net loss within the modulation bandwidth. For the single-frequency case ($\Delta\omega = 0$), it is convenient to define R, L, ϕ, and γ as follows:

$$
v = 1 + Re^{i[\phi+(n-1)\pi]} = Le^{-i\gamma}, \tag{2}
$$

$$
\tan\gamma = \frac{R\sin\phi}{1 - R\cos\phi}
$$

where R and ϕ, respectively, are the instantaneous amplitude and residual phase of the "composite echo" given by the summation term in (1) with $\Delta\omega = 0$. For the deep fades $L \ll 1$, $| 1 - R | \leqq L$, $| \phi | \leqq L$, and n is an even integer.

It is shown in Appendix A that

$$
\frac{dL}{dt} = \frac{d\phi}{dt}\left[\sin\gamma - (\cos\gamma - L)\frac{dR}{R\,d\phi}\right], \tag{3}
$$

$$
\frac{L\,d\gamma}{dt} = \frac{d\phi}{dt}\left[(\cos\gamma - L) + \sin\gamma\frac{dR}{R\,d\phi}\right], \tag{4}
$$

and

$$
\frac{dL}{L\,d\gamma} = \tan(\alpha - \beta)
$$

where

$$
\begin{aligned}
\alpha &= \tan^{-1}\left(\frac{\sin\gamma}{\cos\gamma - L}\right) = \tan^{-1}\left(\frac{\sin\phi}{\cos\phi - R}\right) \\
&= \tan^{-1}\left(\frac{\tan\gamma}{1 - L/\cos\gamma}\right) \approx \gamma, \\
\beta &= \tan^{-1}\frac{dR}{R\,d\phi}.
\end{aligned}
$$

At the bottom of a fade $dL/dt = 0$ and $L\,d\gamma/dt$ has a maximum value of

$$
\frac{d\phi}{dt}\frac{R^2}{(\cos\gamma - L)};
$$

conversely, dL/dt has a maximum value of

$$\frac{d\phi}{dt}\frac{R^2}{\sin\gamma} \quad \text{when} \quad d\gamma = 0.$$

Frequency selective fading requires that $d\phi/dt \gg dR/dt(\beta \ll 1)$, but dR/dt may not be entirely negligible at the very deep fade levels. This observation is consistent with physical models of "reflecting" layers that are slowly rising or falling with time. The physical model suggests that during any particular fade $d\phi/dt$ and dR/dt can be constant, or at least they vary much more slowly with time than the "non-linear" parameters L, γ, dL/dt, and $d\gamma/dt$. This information is useful in examining the detailed wave interference characteristics during a few seconds or minutes in time. On the other hand, the process is nonstationary and the *average* fading characteristics taken over many fades depend on the long term *average* values of $d\phi/dt$ and dR/dt, and hence on meteorological conditions.

Since the variations in signal level are caused by variable meterological conditions along the path, experimental data are needed to obtain quantitative results. In principle, it should be possible to predict the radio results from meteorological theory, if the meterological parameters were known with sufficient accuracy. In practice, this indirect approach is not feasible and direct radio measurements are required. Data on the number and durations of deep fades will be used to estimate the probability distribution of cos γ and $d\phi/dt$ for use in (3) and (4).

III. NUMBER OF FADES VS DEPTH OF FADE

Experimental data show that 20-dB fades occur about ten times more often than 40-dB fades and on the average last about ten times longer. In other words, both the number of fades and their average duration are proportional to L.[1] Of all the fades reaching a given level L_1, 30 percent will "bottom out" at or before $0.7L_1$, 50 percent will not go deeper than $0.5L_1$, and only 10 percent will reach or go deeper than $0.1L_1$. Thus the probability can be written as

$$\text{Prob}\left(\frac{L_{\min}}{L_1} \leqq X\right) = X$$

where X varies uniformly from $0 \leqq X \leqq 1$.

It can be seen from (2) that

$$L = (1 - R\cos\phi)\sqrt{1 + \tan^2\gamma} = \frac{1 - R\cos\phi}{\cos\gamma} = \frac{R\sin\phi}{\sin\gamma}$$

and $\phi \leqq L$. At the bottom of a fade, $\phi \rightarrow 0$, $dL = 0$, $\gamma = 0$, and

$$L_{\min} \approx 1 - R_0 .$$

It follows that at any level $L_1 < 0.1$ the ratio

$$\frac{L_{\min}}{L_1} = \left(\frac{1 - R_0}{1 - R_1 \cos \phi_1}\right) \cos \gamma_1$$

$$\approx \cos \gamma_1 \quad \text{for} \quad R_1 \approx R_0 . \tag{5}$$

Thus the experimentally determined cumulative distribution for $L_{\min}/L_1$ also applies to $\cos \gamma$, when $dR/dt \ll d\phi/dt$; that is

$$\text{Prob}\ (\cos \gamma \leqq X) = X$$

and the average values are

$$\overline{\frac{L_{\min}}{L_1}} \approx \overline{\cos \gamma} = \int_0^1 X\, dX = \tfrac{1}{2}.$$

The corresponding average value of $\overline{\sin \gamma}$, needed in a later section, is given by

$$\int_0^1 \sqrt{1 - X^2}\, dX = \frac{\pi}{4}.$$

IV. DURATION OF FADES

The average duration of a 40-dB fade ($L = 0.01$) at 4 GHz on a 30-mile path is about 3 or 4 seconds, and the average duration of a 20-dB fade is about 10 times longer. The results of an extensive propagation experiment have shown that the distribution of fade duration τ follows a log normal distribution with a standard deviation of about $10 \log(\tau_{16}/\tau_{50}) = 5.6$ or $\ln(\tau_{16}/\tau_{50}) = 1.3$. Quantitatively, these results are summarized as follows for 4 GHz on a 30-mile line-of-sight path:

99 percent of the fades are shorter than 4000 L seconds,
70 percent of the fades are shorter than 400 L (average duration),
50 percent of the fades are shorter than 200 L (median duration).

These data are plotted on Fig. 1 and it is assumed that fade durations shorter than the median are given by the extension of this log normal distribution with the same standard deviation.*

* Recent unpublished data on the duration of rapid fades indicate that the standard deviation may be somewhat larger than the "5.6 dB" used in this paper, which means that the lines on Fig. 1 might be more nearly vertical than shown.

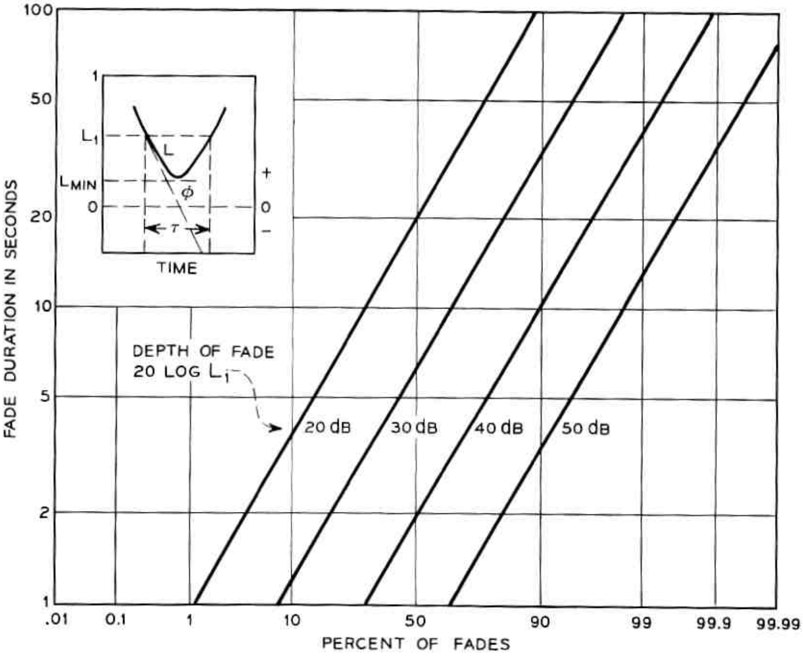


Fig. 1—Duration of fading at 4 GHz on a 30-mile path.

The voltage-time diagram shown on Fig. 1 defines L, $L_{\min}$, ϕ, and the fade duration τ for a fixed frequency. From the diagram it can be seen that for $dR/dt = 0$ the fade duration τ is given by

$$\tau = \frac{2}{\dfrac{d\phi}{dt}} \int_0^{\sin^{-1}(L_1 \sin \gamma_1/R)} d\phi \approx \frac{2}{\dfrac{d\phi}{dt}} \frac{L_1 \sin \gamma_1}{R} \tag{6a}$$

and since the average value

$$\overline{\sin \gamma_1} = \frac{\pi}{4},$$

$$\overline{\frac{d\phi}{dt}} = \frac{\pi L_1}{2\tau R}. \tag{6b}$$

Substituting into (3) and (4), the average values are

$$\overline{\frac{dL}{L\,dt}} = \overline{\frac{d(\ln L)}{dt}} = \frac{\pi^2}{8\tau R}, \tag{7a}$$

$$\overline{\frac{d\gamma}{dt}} = \frac{\pi}{2\tau R}\,\frac{(1-2L)}{2} \approx \frac{\pi}{4\tau R} \quad \text{for} \quad L \ll 1. \tag{7b}$$

It follows that, since 20 log L = 8.7 ln L dB,

$$\overline{\frac{d\gamma}{dt}} \approx \frac{2}{\pi}\,\frac{\overline{d(\ln L)}}{dt} = \frac{2\,\overline{\text{dB/s}}}{8.7\pi} = \frac{\overline{\text{dB/s}}}{13.6}. \tag{8}$$

Figure 2 shows the distribution of rate of change of phase $d\gamma/dt$ and the rate of change of L in dB/second, based on (7a) and (7b) and on the distribution of fade durations shown on Fig. 1.

It will be noted that if γ were uniformly distributed ($\overline{\sin\gamma} = \overline{\cos\gamma}$) it would be expected that

$$\overline{\text{dB/s}} = 8.7\,\overline{\frac{d\gamma}{dt}},$$

but this assumption is not consistent with the experimental result that the number of fades is proportional to L.

In a similar manner, when $d\phi/dt = 0$ ($|\phi|$ is a constant $< L_{\min}$) the fade duration might be represented by

$$\tau = \frac{2}{\frac{dR}{dt}} \int_{R_0}^{R_1} dR = \frac{2}{\frac{dR}{dt}}\,(R_1 - R_0)$$

or

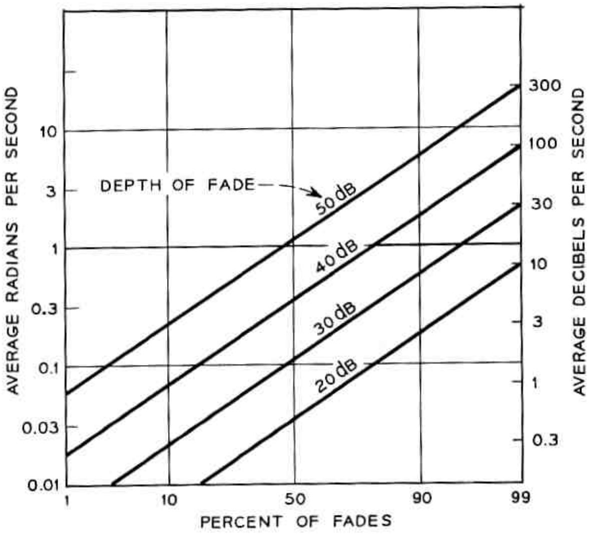


Fig. 2—Rates of change in amplitude and phase during fading at 4 GHz on a 30-mile path.

$$\frac{dR}{dt} = \frac{2L}{\tau}\left[\frac{R_1 - R_0}{L\cos\phi}\right]. \tag{9}$$

There is no obvious way to relate a change in R to a change in meteorological conditions and there is doubt that it is important to do so. The frequency selective nature of fading requires that $d\phi/dt$ be controlling most of the time. In addition, essentially all fast "flat" fading can be explained as a phase interference effect with a delay of one or three half-cycles as well as by a supposed change in R. Moreover, the experimental results that the number of fades and their average durations are both proportional to L are characteristic of a Rayleigh type distribution, which is usually explained on the basis of near-random phase. Consequently, it is assumed in the remainder of the paper that

$$\overline{\frac{d\phi}{dt}} \gg \overline{\frac{dR}{dt}}.$$

V. VERTICAL VELOCITY OF ATMOSPHERIC IRREGULARITIES

Although dL/dt and $d\gamma/dt$ are important parameters in the design of communication systems, the search for a better physical understanding of the fading mechanism needs to begin with the less complicated parameters R, ϕ, and $d\phi/dt$ of the composite echo given by the bracketed term in (1). It may be seen from (4) that for $dR/dt = 0$,

$$\frac{d\phi}{dt} = \frac{L\,d\gamma}{dt}\,\frac{1}{(\cos\gamma - L)}$$

which on the average is

$$\overline{\frac{d\phi}{dt}} \approx \overline{\frac{2L\,d\gamma}{dt}}. \tag{10}$$

In principle, the rate of change in phase $d\phi/dt$ might also be obtained by either: (*i*) direct measurement of phase, or (*ii*) computation from known or assumed values for vertical motion in the atmosphere. For example, it is reasonable to expect that

$$\overline{\frac{d\phi}{dt}} = \frac{\pi\bar{v}}{H_n - H_{n-1}} \tag{11}$$

or conversely,

$$\bar{v} = \left(\frac{H_n - H_{n-1}}{\pi}\right)\overline{\frac{d\phi}{dt}}$$

where $d\phi/dt$ is the rate of change in phase caused by an echo source moving vertically through a thickness $(H_n - H_{n-1})$ at a velocity v. H_n is the height of the nth Fresnel zone and $H_n - H_{n-1}$ is the thickness of the nth Fresnel zone. The relation between $d\phi/dt$ and the effective velocity v is shown on Fig. 3 and the probability of exceeding specified values of $d\phi/dt$ has been obtained from (10) and from Fig. 2. It appears that a vertical motion of only a few centimeters per second (a few feet per minute) is sufficient to explain the fade duration measurements and the values of $d\gamma/dt$ and dB/s derived therefrom. An alternative derivation of (11) is given in Appendix B.

VI. FADE DURATION AT OTHER FREQUENCIES AND PATH LENGTHS

The distribution of velocities obtained in the preceding paragraph is expected to be a meteorological phenomena independent of radio frequency. It should be possible to use this information to predict the fade duration at other frequencies and path lengths.

Substituting (6b) into (11), and using equation (17), it follows that the

$$\text{Average Fade Duration} = \frac{1}{2}\frac{L}{\bar{v}}(H_n - H_{n-1}) = \frac{L}{8\bar{v}}\sqrt{\frac{\lambda D}{n}} \qquad (12)$$

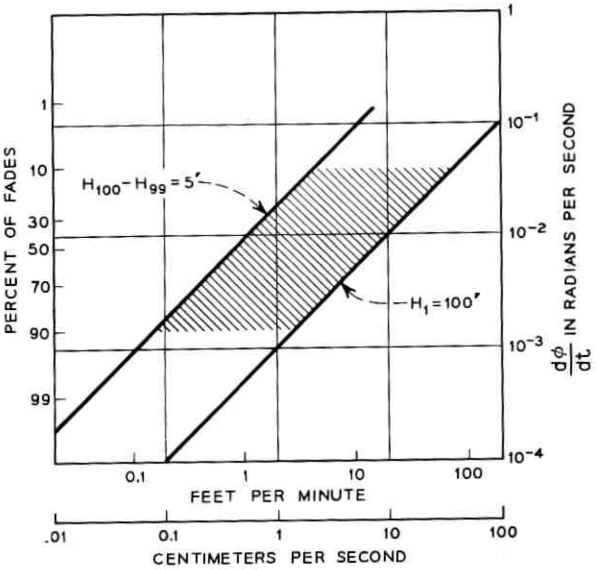


Fig. 3—Change in phase at 4 GHz vs vertical velocity on a 30-mile path.

where D is the path length and λ is the wavelength. Equation (12) indicates that the fade durations vary as the $\sqrt{\lambda\, D}$, *if* the effective value of n is essentially independent of frequency for the average fade.

The preceding information and assumptions lead to the following estimates for the average fade duration at various frequencies, when the path length D is measured in km.

Frequency	Average Fade Duration in Seconds
160 MHz	$2000\, L \sqrt{\frac{D}{50}}$
4 GHz	$400\, L \sqrt{\frac{D}{50}}$
6 GHz	$320\, L \sqrt{\frac{D}{50}}$
6×10^5 GHz (visible light)	$L \sqrt{\frac{D}{50}}$

The experimental results at 4 and 6 GHz are not accurate enough to establish the $\lambda^{1/2}$ variation over this narrow range, but at lower frequencies the fades, when they occur, last longer than at 4 and 6 GHz so the fading rate is not independent of λ.

An extrapolation of the radio data to optical frequencies indicates that the *average* fading rate on a 5-km path through the atmosphere would be of the order of one or two cycles per second with much more rapid fluctuations for a small percentage of the time. In principle, scintillations are rapid multipath phenomena. Although differences, such as the effect of water vapor, antenna beamwidth, etc., need to be included, the fading rate must vary with the wavelength as λ^p, where p is a positive fraction less than one. A variation of as much as λ^1 is unrealistic because it would predict that the twinkling of the stars would ordinarily be above the flicker rate acceptable by the human eye. A consideration of both radio and optical phenomena requires a variation in fade duration with frequency close to $\lambda^{1/2}$ (at least in the range between $\lambda^{1/3}$ and $\lambda^{2/3}$), so the elementary result shown in (12) is a reasonable working hypothesis. At optical frequencies the Fresnel zones are more nearly comparable in size to the atmospheric irregularities and both horizontal and vertical motion of atmospheric irregularities produce significant scintillations.

VII. VARIATIONS IN NET LOSS WITH FREQUENCY

The principal interest thus far has been in the characteristics of fading at a single fixed frequency. Even if an automatic equalizer could be built to compensate perfectly for the effects of fading at any reference frequency, the resulting equalization would be less than perfect at frequencies other than the reference frequency.[2–4] This type of distortion results in variations in net loss (deviations from a constant output level) over the modulation bandwidth. By the use of the concept of Fresnel zones and some simplifying assumptions, it is shown in Appendix C that the variations in net loss relative to a fixed level at the carrier or pilot frequency is given approximately by

$$20 \log \left| 1 - i \frac{\delta_{\text{eff}}}{L} \right| , \tag{13}$$

$$10 \log \left(1 + \left(\frac{\delta_{\text{eff}}}{L} \right)^2 \right) , \tag{14}$$

where

$$\delta_{\text{eff}} = \frac{\Delta f}{f} \pi n_{\text{eff}} .$$

20 log L is the depth of fade in decibels ($L \ll 1$) and n_{eff} is the effective number of half-wavelengths in the relative delay of the strongest delayed component (or composite "echo") that is causing the fade at the reference frequency. The corresponding variations in net loss are shown in Table I and on Fig. 4. Table I indicates that most of the deep fades are caused by "echo" delays of 1, 3, or 5 half-wavelengths ($n_{\text{eff}} \leqq 5$), because experience has shown that good diversity requires a frequency separation greater than 1 percent and that acceptable transmission quality can almost always be obtained with bandwidths of 0.1 percent or less. The corresponding values of $\Delta f/f$ for a 20-dB fade are ten times larger than shown in Table I.

A comparison between theory and experiment is shown on Fig. 4 for the case of 40-dB fades. This figure also suggests that the median 40-dB fade corresponds to $1 \leqq n_{\text{eff}} \leqq 5$, but that 10 percent of the 40-dB fades correspond to delays of many half-wavelengths. The curve for $n = 100$ represents the approximate upper boundary set by 40-dB antenna patterns at 4 GHz on a 30-mile path. The corresponding theoretical curves for a 20-dB fade would be shifted one decade to the right. The crossover of the experimental 40-dB fade data and the $n = 1$ curve shown on the figure is not possible with a single echo,

Table I—Variations in Net Loss

$\frac{\delta_{eff}}{L}$	Net Loss Variation (dB)	Quality	$\frac{\Delta f}{f}$ for 40-dB fade		
			$n_{eff} = 1$ (%)	$n_{eff} = 3$ (%)	$n_{eff} = 5$ (%)
0.1	0.04	Excellent transmission	0.03	0.01	0.006
0.3	0.4	Acceptable transmission	0.1	0.03	0.02
1.0	3.0	Poor transmission	0.3	0.1	0.06
3.0	10.0	Fair diversity	1.0	0.3	0.2
10.0	20.0	Good diversity	3.0	1.0	0.6

and is an indication that more than two components are present much of the time. The "experimental 40-dB fade" frequently represents the combination of two or more echoes whose effects add to 40 dB; in this case, the limiting theoretical curve for $n = 1$ is shifted to the right because the principal component of the composite echo, taken by itself, represents a fade of less than 40 dB.

The agreement between theory and experiment is considered satisfactory and as evidence that most of the fades are caused by strong components delayed by only a few half-wavelengths. It has been suggested that a better fit with the shape of the experimental data exceeded 10 percent and 50 percent of the time can be obtained if the $i = \sqrt{-1}$ in (13) could be omitted; such an assumption is rejected

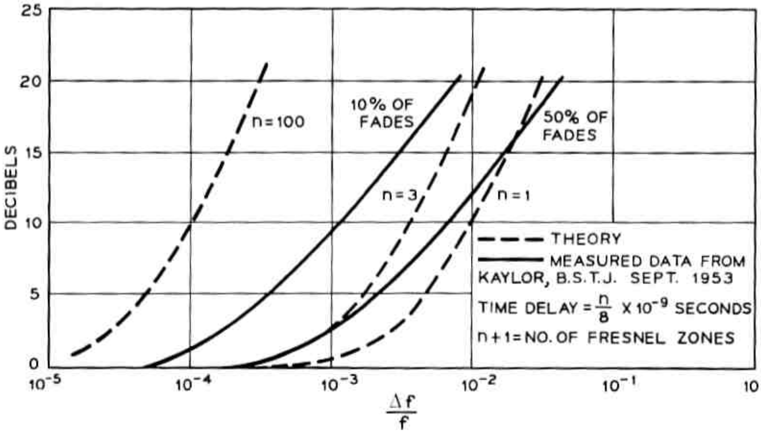


Fig. 4—Variations in net loss during 40-dB fades at 4 GHz on a 30-mile path.

because it seems contrary to logic and because it would create difficulties in explaining the data exceeded by, say, 90 percent of the fades.

VIII. CONCLUSION

Experimental data on the number and duration of fading have been used to estimate the rate of change in phase and the rate of change in amplitude of an electromagnetic wave received through a multipath medium. The elementary theory used herein achieves quantitative results, primarily because it explicitly includes the first-order effects of phase associated with the use of Fresnel zones. Alternative methods starting from the more fundamental wave equations are more sophisticated and may be more exact in a mathematical sense, but their inherent generality does not ordinarily lead to results that can be compared with experiment, unless somewhat arbitrary "constants" or correlation functions are assumed.

APPENDIX A

The relative magnitude and phase of a multipath signal can, by definition, be represented as

$$v = 1 + Re^{i[\phi+(n-1)\pi]} = Le^{-i\gamma}$$

where n is an even integer and R and ϕ are, respectively, the instantaneous magnitude and residual phase of the sum of all the delayed components. It follows that

$$\begin{aligned} L &= \sqrt{(1 - R\cos\phi)^2 + (R\sin\phi)^2} \\ &= (1 - R\cos\phi)\sqrt{1 + \tan^2\gamma} \\ &= \sqrt{(1 - R)^2 + 2R(1 - \cos\phi)} \\ &= \frac{1 - R\cos\phi}{\cos\gamma} = \frac{R\sin\phi}{\sin\gamma}. \end{aligned}$$

By differentiation,

$$\begin{aligned} dL &= \frac{R\sin\phi\, d\phi - (\cos\phi - R)\, dR}{L} \\ &= \sin\gamma\, d\phi - \frac{(R\cos\phi - R^2 + 1 - 2R\cos\phi - 1 + 2R\cos\phi)}{L}\frac{dR}{R} \\ &= \sin\gamma\, d\phi - \left(\frac{1 - R\cos\phi - L^2}{L}\right)\frac{dR}{R} \end{aligned}$$

$$= \sin \gamma \, d\phi - (\cos \gamma - L) \frac{dR}{R}.$$

In a similar manner,

$$L \, d\gamma = (\cos \gamma - L) \, d\phi + \sin \gamma \frac{dR}{R}.$$

APPENDIX B

As shown in several textbooks, the height of the nth Fresnel zone at the middle of path length D is given by

$$H_n = y_n = \frac{\sqrt{\lambda \, Dn}}{2}. \tag{15}$$

The corresponding phase delay is $\theta_n = n\pi$, by definition of Fresnel zones.

In order to restrict the parameter n to integral values, we can generalize as follows:

$$\theta = n\pi - \phi; \qquad \frac{d\theta}{dt} = -\frac{d\phi}{dt}$$

and

$$y = \frac{1}{2} \sqrt{\frac{\lambda \, D(\theta + \phi)}{\pi}}.$$

It follows that

$$v = \frac{dy}{dt} = \frac{1}{4\pi} \sqrt{\frac{\lambda \, D}{n}} \frac{d\phi}{dt} \tag{16}$$

where v is the effective velocity in a plane perpendicular to the direction of propagation.

From (15) we have

$$H_n - H_{n-1} = \tfrac{1}{2} \sqrt{\lambda \, Dn} \left(1 - \sqrt{\frac{n-1}{n}} \right)$$

$$\approx \frac{1}{4} \sqrt{\frac{\lambda \, D}{n}} \quad \text{for} \quad n > 3. \tag{17}$$

By substituting (17) into (16), we have

$$\frac{d\phi}{dt} = \frac{\pi v}{H_n - H_{n-1}},$$

which is the same as equation (11).

APPENDIX C

As noted in (1) the signal received through a multipath medium can be represented by the sum of the principal wave plus the delayed components ($a_m e^{i\Delta t_m}$):

$$E = Ae^{i(\omega+\Delta\omega)t}\left(1 + \sum_{m=1}^{M} a_m e^{i(\omega+\Delta\omega)\Delta t_m}\right) \tag{18}$$

where a_m and Δt_m represent the magnitude and relative time delay of the mth "echo." The term in brackets can also be resolved into Fresnel zones so that

$$v = 1 + \sum_{n=1}^{N} a_n e^{i(\omega+\Delta\omega)(n\pi/\omega)} \tag{19}$$

where n is an integer and a_n is the magnitude of the signal in the $(n + 1)$th Fresnel zone. The upper limit N is the maximum number of Fresnel zones enclosed by the antenna beamwidth; components with longer delays, if present, are relatively unimportant because their magnitudes are reduced by the off-beam antenna patterns. The expression for v can be separated into the following three terms:

$$v = 1 + \sum_{n=1}^{N} a_n e^{in\pi} + \sum_{n=1}^{N} a_n e^{in\pi}(e^{i\delta} - 1) \tag{20}$$

where $\delta = (\Delta f/f)n\pi$. The first two terms in (20) show the fading characteristic at frequency f, which in principle can be corrected by a perfect automatic equalizer based on a carrier frequency or pilot tone. The third term on the right in (20) shows the relative variation (distortion) at frequency Δf.

By definition, let the first two terms be

$$1 + \sum_{n=1}^{N} a_n e^{in\pi} = Le^{-i\gamma}. \tag{21}$$

Substituting into (20) and introducing an effective value of δ, we have

$$\begin{aligned} v &= Le^{-i\gamma} + i\delta_{\text{eff}} \sum a_n e^{in\pi}\left(\frac{e^{i\delta} - 1}{i\delta_{\text{eff}}}\right) \\ &= Le^{-i\gamma} + i\delta_{\text{eff}}\left[Le^{-i\gamma} - 1 - \sum_{n=1}^{N} a_n e^{in\pi} + \sum_{n=1}^{N} a_n e^{in\pi}\left(\frac{e^{i\delta} - 1}{i\delta_{\text{eff}}}\right)\right] \\ &= Le^{-i\gamma} - i\delta_{\text{eff}}\left[1 - Le^{-i\gamma} - \sum_{n=1}^{N} a_n e^{in\pi}\left(\frac{e^{i\delta} - 1}{i\delta_{\text{eff}}} - 1\right)\right]. \end{aligned} \tag{22}$$

By this series approximation and a suitable choice for the effective value of δ (δ_{eff}), the last term on the right can be made negligible during fading conditions ($L \ll 1$),* so that

$$v \approx Le^{-i\gamma}\left(1 - \frac{i\delta_{\text{eff}}e^{i\gamma}}{L}\right). \tag{23}$$

A "perfect" equalizer (at any given fixed frequency) would leave a residual distortion of

$$v' = 1 - \frac{i\delta_{\text{eff}}e^{i\gamma}}{L}. \tag{24}$$

The phase angle γ is zero at the bottom of the fade, is ±45 degrees at 3 dB above the bottom, and varies over the range $-\pi/2 < \gamma < \pi/2$. In the region of good transmission ($\delta_{\text{eff}}/L \ll 1$) it is sufficiently accurate to approximate the net loss variations as

$$20 \log v' = 20 \log \left| 1 - \frac{i\delta_{\text{eff}}}{L} \right| = 10 \log \left[1 + \left(\frac{\delta_{\text{eff}}}{L}\right)^2\right]. \tag{25}$$

While both δ_{eff} and L are individually less than unity (and usually <0.1) their ratio can be greater than unity, and in that case the net loss variations calculated from (25) need to be interpreted as the median value of a distribution of net loss variations.

REFERENCES

1. Vigants, A., "Number and Duration of Fades at 6 and 4 GHz," B.S.T.J., *50,* No. 3 (March 1971), pp. 815–842.
2. De Lange, O. E., "Propagation Studies at Microwave Frequencies by Means of Very Short Pulses," B.S.T.J., *31,* No. 1 (January 1952), pp. 91–103.
3. Crawford, A. B., and Jakes, W. C., Jr., "Selective Fading of Microwaves," B.S.T.J., *31,* No. 1 (January 1952), pp. 68–90.
4. Kaylor, R. L., "A Statistical Study of Selective Fading of Super-High Frequency Radio Signals," B.S.T.J., *32,* No. 5 (September 1953), pp. 1187–1202.

* An associated assumption is that the maximum value of δ is restricted to

$$\delta_{\max} = \frac{\Delta f}{f} N\pi < 1.$$

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 6, July-August, 1971
Printed in U.S.A.

The Computation of Error Probability for Digital Transmission

By F. S. HILL, JR.

(Manuscript received October 19, 1970)

A method is presented for calculating the probability of error for a digital signal contaminated by intersymbol interference and additive Gaussian noise. The method constructs a close approximation to the probability density function of the intersymbol interference, circumventing the heretofore formidable computational problems by decomposing the calculation into a sequence of simple calculations. This method is applicable to binary transmissions, as well as 4-, 8-, 16-, . . . level transmission. The method is rapid enough for use on time-sharing facilities. As part of the method, a new and rather simple scheme is presented for including the effects of partial response source coding. Several interesting examples which use and give insight into the method are included.

I. INTRODUCTION

In a large class of digital transmission systems, a succession of amplitude modulated pulses is sent over a channel to the receiver. The signal suffers two main types of distortion: additive noise and intersymbol interference (ISI). The latter arises from dispersion in the channel, and is a noise-like process due to the overlapping of many neighboring pulses.[1] The number of these neighboring pulses, or "interferers," depends on the system impulse response, and can be rather large for systems with restricted bandwidth.

To properly analyze a system one must compute its probability of error due to the noise and intersymbol interference. The effect of the additive noise is simple to analyze, but that of the ISI is extremely difficult to find due to the complicated nature of its probability density function (pdf). The complexity of the pdf grows exponentially with the number of interferers, and for more than about 12 interferers the computation of error probability has long seemed impossible.

Until recently researchers have analyzed such systems by one of three main routes: (*i*) they have used more tractable—although less meaningful—performance measures such as peak distortion[2] or mean squared error,[1,3] (*ii*) they have found upper bounds on the probability of error,[4] or (*iii*) they have found the probability of error using a truncated impulse response, for which only a few interferers were assumed to be significant.[5,6] Unfortunately, in many restricted-bandwidth systems as many as 50 or 60 interferers can be important, so the use of a truncated impulse response is inadequate for a large class of systems. The goal is therefore to find methods which will treat large numbers of interferers while avoiding the exponential growth of computational complexity. This paper describes such a method.

Recently two other schemes have been presented which also provide accurate estimates of the error probability for large numbers of interferers.[7,8] These methods do not provide the pdf of ISI, but calculate instead the terms in a series expansion of the expression for the error probability. They were devised primarily for the binary transmission case, and the extension to multilevel transmission would be rather awkward.

The new technique presented here constructs a close approximation to the pdf of intersymbol interference, and uses it to find the error probability. Knowledge of the pdf can provide useful information about the intersymbol interference process, illuminating for the system designer the relationship between certain types of channel characteristics and error performance. The technique is applicable to binary transmission, as well as to 4-, 8-, 16-, $\cdots$ level transmission. The method also allows simple inclusion of certain kinds of source coding such as partial response.[9] The ability to treat the multilevel case and different kinds of source coding is very important. In feasibility studies a system designer might know the general type of channel characteristics the signal will encounter, but he does not know how sensitive the signal will be to various combinations of numbers of source levels and source coding. The method given here is envisioned as providing a valuable tool in such studies.

In Section II the probability of error is related to a single random variable, the error voltage z, which then characterizes system performance. It is this random variable whose pdf is sought. Examples taken from the applicable class of source coders are presented and it is shown that the effects of coding can be transferred from the source statistics to the channel description. Alternative performance measures are noted for later use. In Section III the details of the method are described with the computer programmer in mind, and various empirical rules-of-thumb are suggested. In Section IV several simple examples are presented which illuminate some of the potential uses of the method. In Section V the accuracy of the method is considered, and guidelines for choosing some parameter values in the method are given. The method presented here is an extended and refined version of a technique developed by W. E. Norris.[10]

II. ANALYSIS

Figure 1 indicates the system under consideration. The source emits a symbol every T seconds. The symbols are statistically independent, and form the sequence $\{a_k\}$ where each a_k takes on one of the N equally likely values $m(i)$

$$m(i) = (2i - N - 1)V/(N - 1), \quad i = 1, 2, \cdots, N. \tag{1}$$

The outermost levels are $\pm V$ volts. The sequence is then passed through a coder to form the sequence $\{c_k\}$. Only two kinds of coding are treated in the main body of the paper, although the method is extended to more general coders in the Appendix. The two special types treated here are the uncoded case and the class IV partial response case,[9] where

$$\begin{aligned} \text{uncoded: } c_k &= a_k , \\ \text{partial response IV: } c_k &= a_k - a_{k-2} . \end{aligned} \tag{2}$$

Partial response IV coding provides useful shaping of the signal spectrum, and precoding of the source sequence can be used to prevent error propagation.[9] It is assumed that the sequence $\{a_k\}$ has already been precoded appropriately.

The sequence $\{c_k\}$ modulates the impulse response $r(t)$ of the channel. White Gaussian noise is added, and the received signal $y(t)$ is sampled every T seconds. The decision device (typically an A/D converter), determines which of the possible transmitted levels each

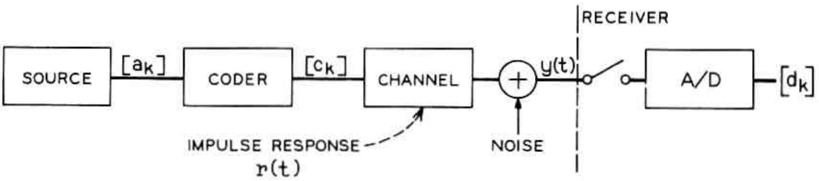


Fig. 1—System configuration.

sample is nearest, and outputs the sequence $\{d_k\}$. (The receiver can base its decision simply on the present sample because of the precoding.[9]) Ideally, each $d_k = c_k$, but because of the noise and ISI they will occasionally differ, and an error is made. The method discussed here calculates the probability of occurence of such an event.

The sample at $t = 0$ has the form

$$y(0) = \sum_{k=-\infty}^{\infty} c_k r(-kT) + n, \tag{3}$$

where n is a realization of a zero mean Gaussian random variable with variance σ^2. Because the sequence $\{a_k\}$ and the noise are stationary in a statistical sense, studying the error probability for the single sample $y(0)$ is equivalent to studying it for the entire ensemble of samples $y(mT)$.

Ideally $y(0)$ will equal c_0. The error voltage z is therefore given by

$$z = \sum_{k=-\infty}^{\infty} c_k x_k + n, \tag{4}$$

where

$$x_k = r(-kT) - \delta_k \tag*{(5)*}$$

are the "error samples" of the system. It is very convenient to incorporate the effects of any source coding in the error samples. In the case of partial response coding:

$$\begin{aligned} z &= \sum_{k=-\infty}^{\infty} (a_k - a_{k-2})x_k + n \\ &= \sum_{k=-\infty}^{\infty} a_k e_k + n, \end{aligned} \tag{6}$$

where

$$e_k = x_k - x_{(k+2)} \tag{7}$$

are the "coded error samples." Thus z of (6) depends on the simple statistics of the uncoded source symbols a_k. (If no coding is used, then $e_k = x_k$.)

The error voltage z is the sum of a large number of independent random variables. Each a_k has the same discrete probability density function: A set of N equidistant Dirac delta-functions with weights

* δ_k = Kronecker delta function: $\delta_o = 1$; $\delta_k = 0$, $k \neq 0$.

$1/N$. Each a_k is scaled by the coded error sample e_k. The n is Gaussian, and assumed independent of the a_k's. The pdf $p_z(\cdot)$ of z is found by convolving the individual density functions of the constituent random variables. Because each of these constituents has a symmetrical pdf about $z = 0$, $p_z(\cdot)$ is also symmetrical. Once $p_z(\cdot)$ has been found, the error probability for the sample $y(0)$ is simply the probability that $y(0)$ is further from c_0 than from some other level. When c_0 is an inner level this is just the probability that $|z| > V/(N - 1)$. When c_0 is an outer level only one polarity of z causes an error, and an error occurs with the probability that $z > V/(N - 1)$. For uncoded symbols outer levels occur with probability $1/N$, while for partial response IV coding they occur with probability $1/N^2$. Hence the error probability is

$$P_e = 2(1 - N^{-h}) \int_{V/(N-1)}^{\infty} p_z(e)\, de, \tag{8}$$

where $h = 1$ for an uncoded source, and $h = 2$ for a partial response IV coded source.

Clearly, the real task in finding P_e lies in computing $p_z(\cdot)$, since so many random variables are involved. The method used to achieve this is described in the next section, after two simpler measures of system distortion have been introduced. These measures have been widely used in the past, and will provide useful definitions in the discussion to follow. They are both defined in the noise-free case (i.e., $n = 0$) here in order to isolate the effect of the intersymbol interference.

2.1 *Mean Squared Error*[1,3]

The variance of the error voltage z is given by

$$E[y(0) - c_0]^2 = E[\sum_k a_k e_k]^2 = E[a^2] \sum_k e_k^2, \tag{9}$$

where $E[a^2]$ is the variance of the source. A simple calculation based on equation (1) shows that $E[a^2] = \frac{1}{3}V^2(N + 1)/N - 1)$. A normalized version of the mean squared error then depends only on the coded error samples, and will be called MSE:

$$\text{MSE} = \sum_k e_k^2. \tag{10}$$

2.2 *Peak Distortion*[2]

The maximum value that z of equation (6) can have when $n = 0$ is

$$\max_{a_k} z = V \sum_k | e_k | , \tag{11}$$

which occurs when all of the a_k symbols have maximum amplitude V and such signs that the interfering symbols add constructively. Such a sequence occurs with very low probability. Peak distortion here is defined as D:

$$D = \sum_k | e_k | . \tag{12}$$

III. THE COMPUTATIONAL METHOD

The starting point for computing $p_z(\cdot)$ is the set of coded error samples $\{e_k\}$. The only other relevant parameters are V, N and σ^2. Because each random variable a_k has the same pdf, and this function is symmetrical, one need only consider the magnitudes $|e_k|$, and these can be relabelled to rank order the samples for convenience.

3.1 *Obtaining the Error Samples*

Channel characteristics are typically measured or calculated in the frequency domain, although in some simple cases an analytical expression for the channel impulse response is available. If the impulse response is indeed available, it is simply sampled as in equation (5) to obtain the error samples. Otherwise the error samples are computed from $R(f)$ as follows.* Because by inspection

$$r(kT) = \int_{-\infty}^{\infty} R(f)e^{i2\pi fkT}\, df = \int_{-1/2T}^{1/2T} e^{i2\pi fkT} \sum_{m=-\infty}^{\infty} R(f - m/T)\, df \tag{13}$$

the Fourier coefficients of $\hat{R}(f) = \sum_m R(f - m/T)$ are desired. A Discrete Fourier Transform (DFT)—perhaps using a Fast Fourier Transform algorithm—of N_s samples of $\hat{R}(f)/N_sT$, $f_i = (i - 1)/N_sT$, $i = 1, 2, \cdots , N_s$, yields N_s samples of the aliased version $\hat{r}(t) = \sum_m r(t - mT)$, $t_k = (k - 1)T$, $k = 1, 2, \cdots , N_s$.[11] Although the samples of $\hat{r}(t)$ are not identical to those of $r(t)$, for well-behaved channels and sufficiently large N_s the two are indistinguishable.

The error samples of equation (5) are formed by removing unity from r_o. The coded error samples (for the partial response IV case) are then formed by subtracting from each sample the value of the error sample two places to the right in accordance with (7). The last two error samples, e_{N_s-1} and e_{N_s} are formed making use of the periodic nature of $r(t)$:[11] e.g., because $x_{N_s+2} = x_2$ we have $e_{N_s} = x_{N_s} - x_2$. Finally, only the magnitudes are retained, and these are rank ordered.

* Upper and lower case letters indicate Fourier Transform pairs.

The result is the vector $\mathbf{e}$:

$$\mathbf{e} = (e_1, e_2, \cdots, e_{N_s}), e_1 \geqq e_2 \geqq \cdots \geqq e_{N_s} \geqq 0. \tag{14}$$

(Any $e_k = 0$ could be discarded.)

3.2 *Formation of the Pdf for Binary Transmission*

The number of levels N is normally a power of two. In such cases the pdf $p_z(\cdot)$ for N levels can be built up from the pdf for the binary case as discussed below. Thus the binary case is of fundamental importance. Ignoring the Gaussian noise for the moment, the error voltage z of equation (6) is

$$z = \sum_{k=1}^{N_s} a_k e_k, \tag{15}$$

where the e_k are elements of $\mathbf{e}$ and for the binary case each a_k takes on the values ±V with equal probability. There are 2^{N_s} possible realizations of the sequence $(a_1, a_2, \cdots, a_{N_s})$, each occuring with probability 2^{-N_s}. Each realization yields a value for z, (not necessarily all different), which contributes a delta function of weight 2^{-N_s} to the pdf of z. Since N_s can be 100 or more, an attempt to form z for each possible sequence of a_k would be futile ($2^{100} = 10^{30}$). Instead, decomposition and quantization are used to reduce this formidable task to a sequence of rather simple ones.

3.2.1 *Decomposition*

The solution used here to circumvent the overwhelming computational problem is to decompose the N_s samples into blocks of K samples, and to find the pdf of ISI due to the samples for each block.* The resulting pdf's are convolved together to form the final binary pdf of z. Using specific numbers for concreteness, if $N_s = 54$ and one chooses $K = 9$, then z can be written as a sum of six random variables $A_1, \cdots, A_6$, where

$$A_i = \sum_{k=9i-8}^{9i} a_k e_k, \qquad i = 1, \cdots, 6. \tag{16}$$

Each A_i is a discrete random variable, being the sum of nine discrete random variables. As the pdf of each A_i is formed, it is convolved with previous ones until all 54 samples have been included.

* Assume for the moment that N_s is an integer multiple of K. A rule-of-thumb for discarding some of the small error samples in order to reduce the computation time is described in the following text.

Consider the formation of A_1. Since a_k can take on values $\pm V$, there are 2^9 possible values for A_1. The largest value, denoted as A_1^*, occurs when all $a_k = V$, and has value $A_1^* = V(e_1 + e_2 + \cdots + e_9)$. The 2^9 delta functions that make up the exact pdf of A_1 will therefore lie in the interval $[-A_1^*, A_1^*]$. Since the pdf is symmetrical it is sufficient to compute it for nonnegative values of A_1 only. Every combination of $\pm V$ values for $a_1, \cdots, a_9$ is tried. For each the corresponding value of A_1 is found according to equation (16), and is discarded if $A_1 < 0$.

The following simple technique insures that all combinations are included. A 9-digit binary number B is initialized to zero, and then in turn incremented by one until all digits are one (2^9 iterations in all). For each value of B, the value of a_k is V if the kth digit of B is 1, and $-V$ otherwise.

3.2.2 *Quantization*

Decomposition alone would not significantly reduce the size of the computation, since the total number of delta functions in the final binary pdf of z would still be on the order of 2^{54}. Quantization of the voltage axis is used to keep the number of possible levels of z within reasonable bounds.[10]

The interval $[0, A_1^*]$ is marked off with $2N_1 + 1$ "location" points t_i, $i = 0, 1, \cdots, 2N_1$. Associated with each t_i is a probability p_i. Instead of recording the exact value of A_1, it is tested to see which t_i it is nearest, and the corresponding p_i is incremented by 2^{-9}. There are many roughly equivalent ways of assigning the values t_i. The following scheme was adopted for its computational simplicity. The interval $[0, A_1^*]$ is broken into two parts at some level kA_1^*, and then each part is uniformly quantized into N_1 levels. Hence, the t_i are given by

$$t_i = \begin{cases} kA_1^* i/N_1, & i = 0, 1, \cdots, N_1, \\ kA_1^* + (1-k)A_1^*(i/N_1 - 1), & i = N_1 + 1, \cdots, 2N_1. \end{cases} \tag{17}$$

For $k = 1/2$, the quantizing is therefore uniform over the whole interval. For $k > 1/2$ the quantizing is finer for larger values of A_1, so that the most important levels of error voltage are most accurately represented. An efficient choice for k is 0.6, as shown in Section V. (However, uniform quantizing may have other advantages since convolution could be performed using a Fast Fourier Transform algorithm.[11]

The pdf's of $A_2, \cdots, A_9$ are computed in the same way. After each has been found, it is convolved with the resultant pdf for the previous ones. Quantization is also used in the convolution process.

3.3 *Formation of the N-Level Pdf*

Once the binary pdf for z has been obtained, it is a straightforward matter to find the pdf for 4-, 8-, 16-, $\cdots$ level sources. The key to this is to envision N level transmission as the sum of $\log_2 N$ independent binary transmission systems.[10] For instance, an 8-level source with voltage range from $-V$ to V can be decomposed into three binary sources with symbols f_k, g_k, and h_k having only values $\pm V$. Each 8-level symbol a_k can be written as:

$$a_k = \frac{4f_k + 2g_h + h_k}{7} \tag{18}$$

and the error voltage for eight levels can be written in the noise-free case as

$$\begin{aligned} z &= \tfrac{4}{7} \sum f_k e_k + \tfrac{2}{7} \sum g_k e_k + \tfrac{1}{7} \sum h_k e_k \\ &= \tfrac{4}{7} E_1 + \tfrac{2}{7} E_2 + \tfrac{1}{7} E_1, \end{aligned} \tag{19}$$

where E_1, E_2, and E_3 are independent random variables, all having the same pdf as z in the binary case. (The generalization to any power of 2 is straightforward.) Since z is the sum of three random variables, and the pdf of each is known, it is a simple matter to convolve scaled versions of the binary pdf to obtain the 8-level pdf. Quantization is used here also, employing the scheme described above. The final pdf is thus approximated over the interval $[0, D \cdot V]$ by a set of $2N_1 + 1$ delta-functions having locations t_i similar to those in equation (17), and probabilities p_i. By symmetry there is the same pattern of delta-functions at location $-t_i$ with probabilities p_i.

3.4 *Calculation of the Error Probability*

The Gaussian random variable n of equation (6) is reinserted by convolving its pdf with that found above. The result is simply

$$P_z(x) = \frac{1}{\sqrt{2\pi\sigma}} \sum_{i=0}^{2N_1} p_i \{e^{-(x-t_i)^2/2\sigma^2} + e^{-(x+t_i)^2/2\sigma^2}\}. \tag{20}$$

From equation (8) the error probability is then

$$P_e = 2(1 - N^{-h}) \sum_{i=0}^{2N_1} p_i \left[Q\left(\frac{V - t_i}{\sigma(N-1)}\right) + Q\left(\frac{V + t_i}{\sigma(N-1)}\right)\right] \tag{21}$$

where $Q(x)$ is the normal probability integral

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt. \tag{22}$$

The accuracy of the method is examined in Section V, following a set of examples of the method applied to typical channel distortions.

3.5 *Reducing the Computation Time by Discarding Samples*

In the preceding discussion all N_s error samples of equation (14) were involved in the computation of the pdf of ISI. In most practical cases, however, a sizable number of the error samples are very small, and consequently their effect on the nature of the pdf as calculated is negligible. In order to save some computation time it behooves one to truncate the rank-ordered vector **e** of (14) at some index M, and to lump the other samples together in some fashion. This is an optional procedure, since the computation time grows only approximately linearly with the number of blocks of K samples processed, but it can indeed improve the efficiency of the method.

One way to approximate the effect of the remaining samples is to treat the random variable

$$A_r = \sum_{K=M+1}^{N_s} a_k e_k \tag{23}$$

as a Gaussian random variable with variance[4]

$$\sigma_r^2 = E[a^2] \sum_{K=M+1}^{N_s} e_k^2. \tag{24}$$

The effect of these samples is thus a contribution to the Gaussian noise n, replacing the previous variance σ^2 in equation (21) with $\sigma_1^2 = \sigma^2 + \sigma_r^2$.

The question of choosing M still remains. One rule of thumb is that the sum of the samples lumped together should not exceed 1 percent of the total sum D. This was found to be very useful empirically: for all choices of channel studied P_e was the same whether all N_s samples were used, or only M of them, while if only M were used the saving in computation time was significant.

There are some pathological cases of mainly academic interest where D is theoretically infinite. This is true, for instance, for an ideal band-limited signal with a timing error. Saltzberg has shown, using an upper bound method, that the error probability can still be small in such cases.[4] The present method will not work, however, since the value of D is a crucial landmark in the computational procedure, and it must be finite. Such a restriction does not significantly reduce the power of the method.

IV. EXAMPLES OF ERROR RATE CALCULATION

In each of the examples the following definitions are used:

$$\text{Error Rate} = \frac{P_e}{\log_2 N}, \qquad \text{SNR} = \frac{E[c^2]}{\sigma^2} = \frac{h_s V^2(N+1)}{3\sigma^2(N-1)} \tag{25}$$

P_e is normalized by the number of information bits in each symbol so that comparisons between different values of N will be more meaningful. Signal-to-noise ratio is given by the ratio of the variance $E[c^2]$ of transmitted levels to average noise power. Since twice as much power is transmitted (if V is unchanged) when partial response IV coding is used, $h_s = 1$ for uncoded sources, and $h_s = 2$ for partial response IV sources.

(*i*) *Error in Sampling Time*:

A convenient example of an ideal signal shaping characteristic $S(f)$ is shown in Fig. 2. $S(f)$ satisfies the Nyquist criterion and consequently if $R(f)$ of Fig. 1 were equal to $S(f)$ no intersymbol interference would be present at the receiver. Instead it is assumed that the sampling time is in error by ρT seconds, so that $R(f)$ is given by

$$R(f) = S(f)e^{-j2\pi f\rho T}. \tag{26}$$

Figure 3 shows the vector of error samples $\mathbf{e}$ of equation (14)

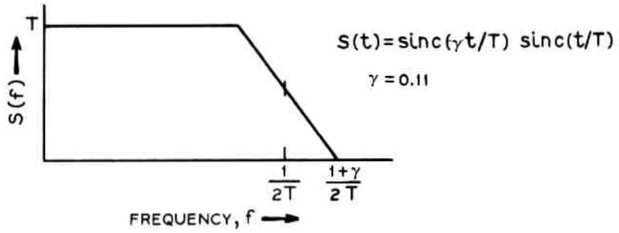


Fig. 2—Ideal signal shaping characteristic.

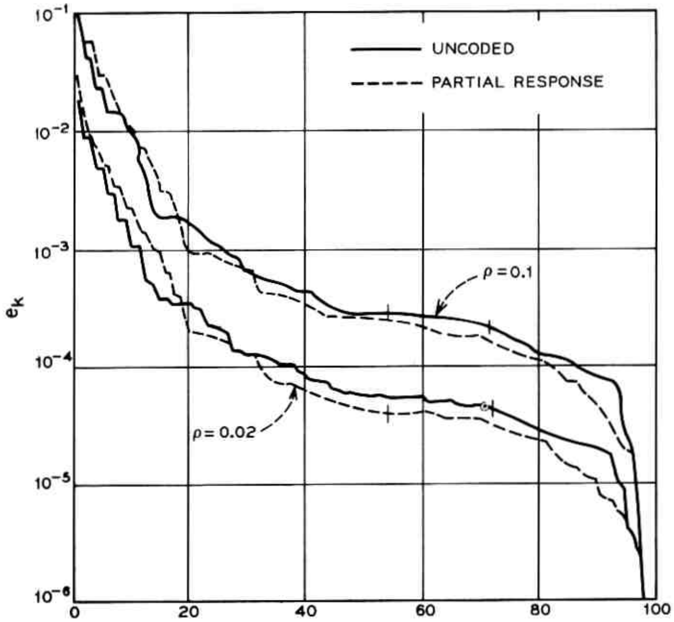


Fig. 3—Error samples e_k for timing error.

for $\rho = 0.02$ and $\rho = 0.1$, and for coded and uncoded sources. The samples fall off rapidly in each case, and the significant samples are somewhat larger for partial response IV coding. In order to include 99 percent of D, six blocks of nine samples must be used in the partial response IV case, and eight blocks of nine in the uncoded case. Figure 4 shows the resulting probability density functions for an uncoded source with timing error $\rho = 0.1$. (In the quantization process $N_1 = 40$ so that 81 locations were used in the interval $[0, V \cdot D]$.) The binary and 8-level pdf's are shown along with a Gaussian pdf having the same value of MSE. The Gaussian function is a reasonably good approximation to the binary case near the origin, but is of course very poor near $z = D \cdot V$, since the pdf's of intersymbol interference are strictly zero beyond this point. Figures 5 and 6 show curves of error-rate versus SNR for several values of timing error. An 8-level source is of course much more sensitive to timing error than is a binary source. In the absence of intersymbol interference a partial response IV signal requires about 3 dB more signal-to-noise ratio than an uncoded signal to achieve the same error-rate, under a constraint of equal average signal power. It is clear from the figures that this "SNR cost" need not be maintained when intersymbol interference is present. The partial response IV signal is more sensitive to timing error than is the uncoded signal, such that at $\rho = 0.15$ the cost is around 9 dB. However, for types of distortion other than timing error, this situation can be markedly reversed, since a partial response IV signal is rather impervious to distortions near dc or the Nyquist frequency.[1]

(*ii*) *Single Echo in the Channel:*

Echoes frequently occur in transmission channels, generated for instance by mismatched terminations. A sizable amount of analysis has been done on echo theory in relation to system distortion.[1,12] Here a single echo of amplitude g and relative delay βT is assumed to occur in the ideal channel of Fig. 2, so that

$$R(f) = S(f)(1 + ge^{-i2\pi f\beta T}). \tag{27}$$

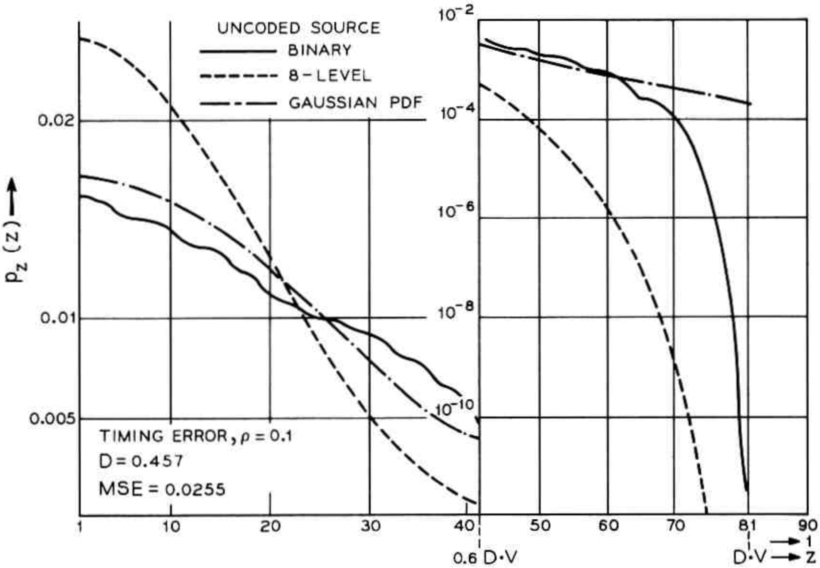


Fig. 4—PDF of intersymbol interference.

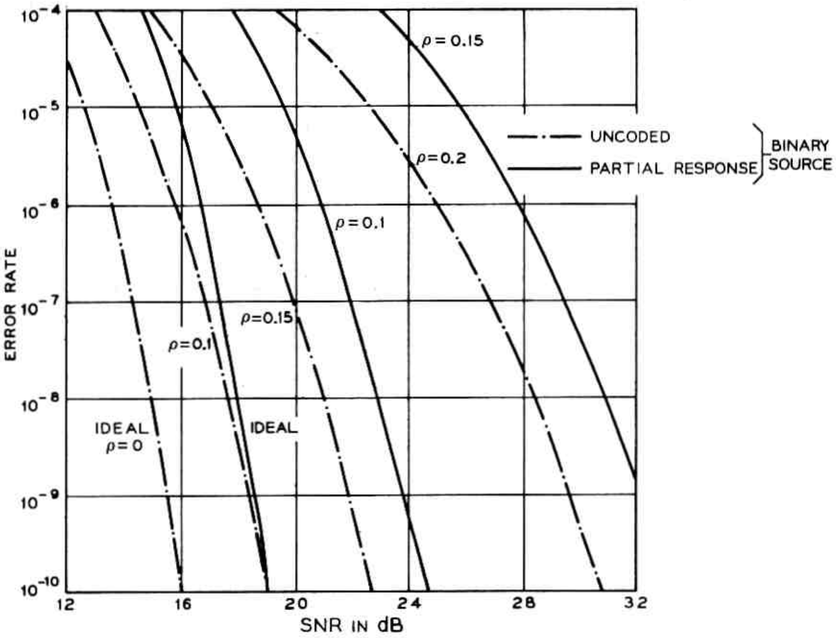


Fig. 5—Error rate vs SNR for timing error.

Because $s(t)$ is symmetrical and introduces no intersymbol interference, an echo at βT is identical in its effect to an echo at $(\pm \beta \pm n)T$ for any integer n. For instance, echoes at $\beta = \pm 0.4, \pm 3.4 \pm 0.6$ all have identical error sample vectors. Consequently one need only consider values of β between 0 and 0.5.

Figure 7 shows curves of error rate versus SNR for echoes of various strengths, and for an 8-level partial response IV coded source. Direct calculation shows that, because of the narrow excess bandwidth γ of $S(f)$ in Fig. 2, the sample error variance MSE of equation (10) is essentially independent of echo position. Echo strengths were chosen to yield convenient values of MSE = 0, 0.001, 0.003, and 0.005. It is clear that echo location has a strong effect on the error rate curves. Echoes at $\beta = .5$ yield the largest error rates. The error rate does not necessarily vanish as SNR $\rightarrow \infty$, as indicated by the asymptote labelled "isi alone."

(*iii*) *Equal Error Samples*:

Consider the error sample vector $\mathbf{e}$ consisting of n identical elements, $e_k = u$, $i = 1, \cdots, n$. A considerable amount of insight is gained by examining this case, although it is unlikely that any physical system would give rise to such a vector of error samples. In addition, one can derive the pdf analytically for the binary case providing a useful check on the computer method.

The error voltage z of equation (15) clearly has the value $uV(n - 2r)$ if and only if r of the symbols a_k equal $-V$, and the rest equal V. Such an event occurs with probability $\binom{n}{r} 2^{-n}$ so that z has the pdf

$$p_z(e) = \sum_{r=0}^{n} \binom{n}{r} 2^{-n} \, \delta(e - uV(n - 2r)), \tag{28}$$

and the error probability follows immediately using the method

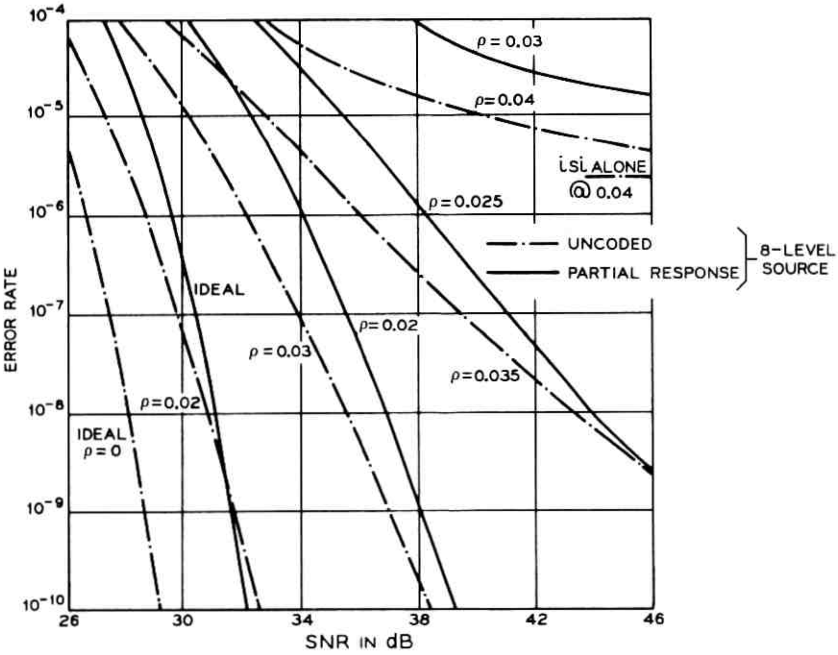


Fig. 6—Error rate vs SNR for timing error.

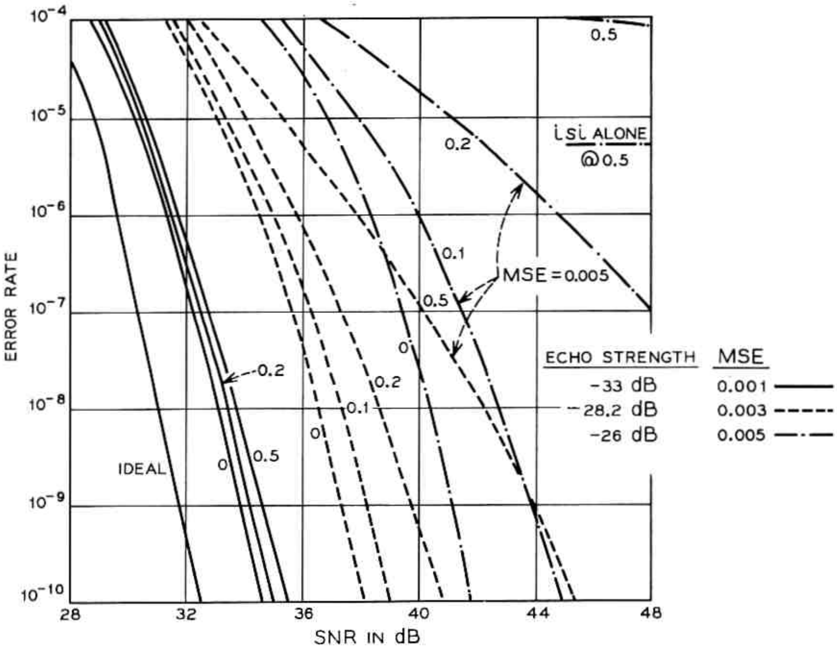


Fig. 7—Error rate vs SNR for single echo.

of Section 3.4:

$$P_e = \sum_{r=0}^{n} \binom{n}{r} 2^{-n} Q[\sqrt{\text{SNR}}\,[1 - u(n - 2r)]]. \tag{29}$$

The case $n = 1$ is the same as an echo positioned at $\beta T = 0$. Figure 8 shows curves of error rate versus SNR as a function of n. (Curves formed using equation (29) and the method just given were indistinguishable.) The samples have size $u = 0.5/n$ so that $d = 0.5$ for all n, while $\text{MSE} = 0.25/n$. From the curves it is clear that a few large error samples are more degrading than many small ones for the same value of D. This is intuitively reasonable since only very special source sequences $\{a_k\}$ will cause the error voltage z to be large when $\mathbf{e}$ consists of many small error samples, and such sequences occur with very small probability. Figure 9 shows similar curves, where now the samples have sizes $0.1/\sqrt{n}$ or $0.2/\sqrt{n}$, so that MSE is constant at 0.01 or 0.04. For $\text{MSE} = 0.01$ error rate is somewhat insensitive to the number of error samples, because the increasing value of $D = 0.1\sqrt{n}$ with n is offset by the decreasing probability that sequences $\{a_k\}$ yielding large errors will occur. However, as n increases for MSE $= 0.04$, D approaches 1, the value at which errors occur due to intersymbol interference alone, ($D = 1$ for $n = 25$). The offset in probability is not strong enough here, and error rate is very sensitive to n.

V. ACCURACY OF THE METHOD AND THE SALTZBERG BOUND

The applications of quantization at various steps in the computation shift the locations of some of the delta-functions slightly in one direction or the other, introducing some error in the t_i locations of equation (17). It would be extremely difficult to estimate analytically the error resulting in the value of P_e. Instead an empirical check was made over several channel characteristics. Figure 10 shows how error rate converges with increasing N_1 for a typical example of channel distortion. Convergence was typically most rapid for $k = 0.6$. For high values of SNR uniform quantization ($k = 0.5$) was noticeably

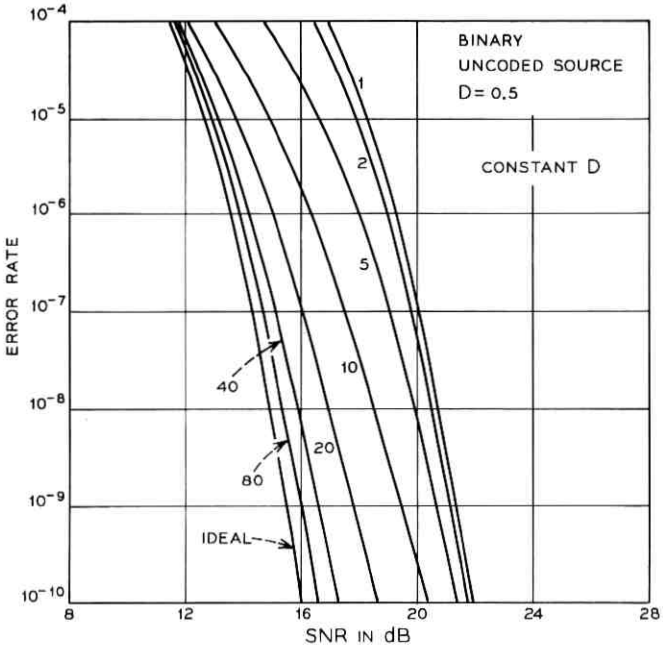


Fig. 8—Error rate for equal error samples.

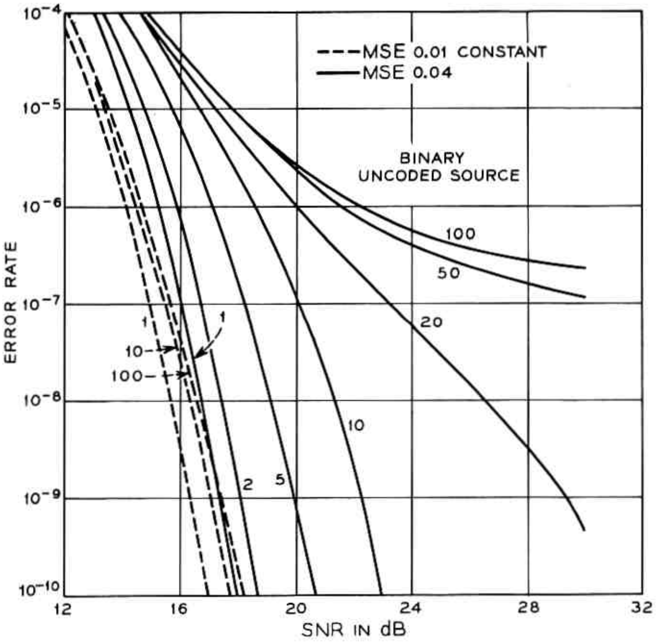


Fig. 9—Error rate for equal error samples.

inferior. Because computation time grows rapidly with N_1, values of $N_1 = 40$ and $k = 0.6$ were used in the examples of Section IV. With these values the algorithm is acceptably rapid, even on time-sharing facilities.

A much faster and simpler scheme was proposed by Saltzberg.[4] This technique found an upper bound for the error probability, but it was not clear from the theory how tight a bound was being computed. Figures 11 and 12 compare the error rate using the method given above with the upper bound found using this technique. For binary transmission the bound is between one and three orders of magnitude above the true value. For 8-level transmission the bound is less tight. The discontinuities in the bound versus signal-to-noise ratio occur due to the optimum partitioning of the error samples into two sets in the upper-bound method.

(*i*) *Choice of the Block Size in the Method*:

Given M error samples (the M "significant" samples from the original set of N_s), it was desired to find the binary pdf of z in equation (15). To do this, the M samples were partitioned into M/K blocks of K samples each. Consequently one must compute M/K pdf's based on sets of K samples, and perform $M/K - 1$ convolutions of these pdf's to obtain the binary pdf of z. The speed of the convolution process depends on $2N_1 + 1$, the number of quantization levels. By direct examination of the algorithm one finds: (*i*) that to form a pdf trying all combinations of K samples requires 2^K multiplications; (*ii*) that each convolution requires $5(2N_1 + 1)^2$ multiplications. The total number of multiplications is therefore

$$N_{\text{mult.}} = 2^K M/K + (M/K - 1)5(2N_1 + 1)^2. \tag{30}$$

For example, if $N_1 = 40$ and $M = 50$, a search shows that $N_{\text{mult.}}$ has a minimum with respect to K of 1.2×10^5 at $K = 12$. The saving in computation effort is dramatic: if no decomposition were used ($K = M$), $N_{\text{mult.}}$ would be 1.1×10^{15}. More generally, examination of equation (30) reveals that if $M/K \gg 1$ then $N_{\text{mult.}}$ varies linearly with M. Therefore $N_{\text{mult.}}$ has a minimum

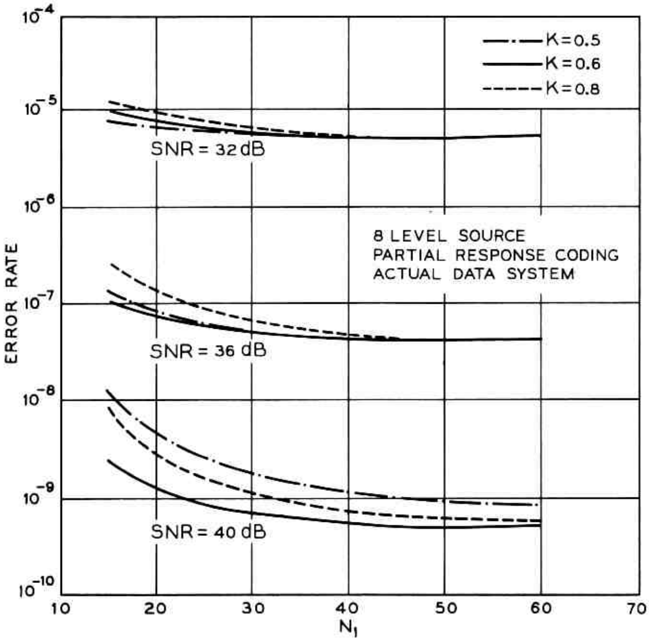


Fig. 10—Convergence of error-rate values.

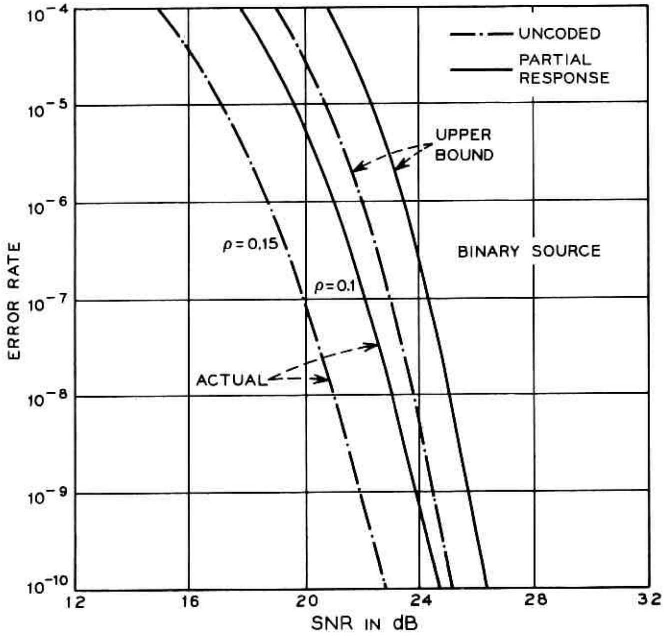


Fig. 11—Comparison with upper bound.

with respect to K that depends only on N_1. This minimum is $K = 12$ for $N_1 = 40$, and $K = 13$ for $N_1 = 60$, although $N_{\text{mult.}}$ is rather insensitive to K for $K \epsilon$ (8, 15).

VI. CONCLUSIONS

A method has been developed that accurately computes the probability of error for a multilevel digital signal contaminated by intersymbol interference and noise. The method constructs a close approximation to the probability density function for the intersymbol interference, which can provide useful insight into the nature of the interference. The method is rapid enough for use on time-sharing facilities, and can provide a useful tool to aid the analysis and design of digital communication systems.

Starting with the transfer function or impulse response of the channel, a set of error samples is found, and the method operates on these to form the probability density function of intersymbol interference. Partial response coding of the source is easily included if desired. The formidable size of the computation that has heretofore blocked direct calculation of this probability density function is circumvented by partitioning the problem into a sequence of easily handled computations. Quantization keeps the total number of elements in the probability density function within reasonable limits, and the density function for binary transmission is used repeatedly to form the density function for 4-, 8-, 16-, ··· level transmission.

VII. ACKNOWLEDGMENT

The author wishes to acknowledge the large contribution of W. E. Norris who developed several of the key ideas in the method described here. Also, the aid and encouragement of John F. Gunn is greatly appreciated.

APPENDIX

Extension to Other Coding Types

The method described above considers only two coding types: the uncoded and partial response IV cases. With slight adjustments in the method one can also treat other interesting cases. The class of coders

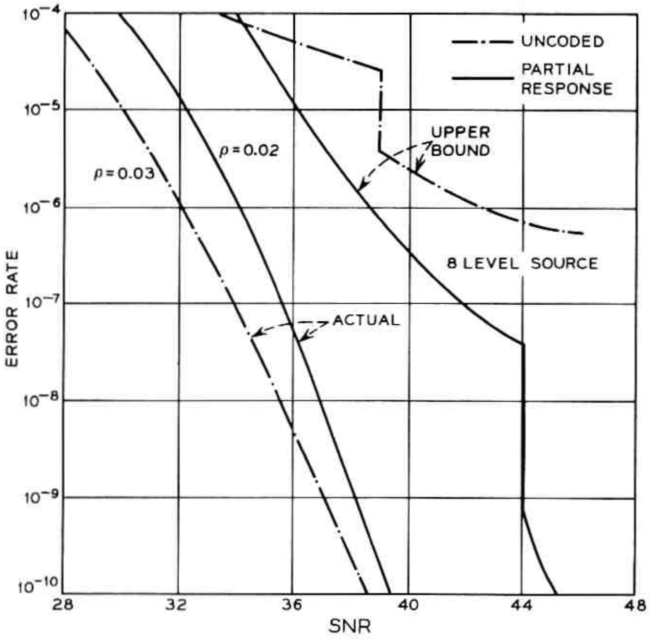


Fig. 12—Comparison with upper bound.

considered is that of Kretzmer[9,13]: each symbol c_k is related to the input symbols $\{a_k\}$ by

$$c_k = b_0 a_k + b_1 a_{k-1} + b_2 a_{k-2} + \cdots + b_v a_{k-v} , \tag{31}$$

where the elements b_i of the vector $\mathbf{b}$ are integers. (The uncoded case is then $b_0 = 1$, all other $b_i = 0$). Kretzmer tabulates five partial response classes which have proven most interesting: (I) $\mathbf{b} = (1, 1)$; (II) $\mathbf{b} = (1, 2, 1)$; (III) $\mathbf{b} = (2, 1, -1)$; (IV) $\mathbf{b} = (1, 0, -1)$; and (V) $\mathbf{b} = (-1, 0, 2, 0, -1)$. Other simple types exist as well, such as $\mathbf{b} = (1, 0, 0, 0, 0, 0, 0, 0, 0, -1)$, a generalization of class IV partial response that has five "lobes" in the signal spectrum.[9]

To see how codes of the type in equation (31) could be included in the method, note that there are only three places where the coding scheme used makes a difference: (*i*) in forming the coded error samples, as in equation (7), (*ii*) in finding P_e from the pdf, as in equation (8), and (*iii*) in evaluating signal-to-noise ratio, as in equation (25).

(*i*) The coded error samples are easily found for any vector $\mathbf{b}$. Using equation (30) in equation (4), and manipulating

$$\begin{aligned} z &= \sum_k \sum_{m=0}^{v} b_m a_{k-m} x_k + n \\ &= \sum_{m=0}^{v} b_m \sum_j a_j x_{j+m} + n \\ &= \sum_j a_j e_j + n, \end{aligned}$$

where

$$e_j = \sum_{m=0}^{v} b_m x_{j+m} \tag{32}$$

are the coded error samples.

(*ii*) Once the pdf of isi has been found from the coded error samples, the error probability follows as in equation (8). To use signal power efficiently a coder will be chosen so that the transmitted voltage levels are equally spaced. Thus one need only evaluate the probability of an outer level in order to adapt equation (8) to this larger class of coders. An outer level of c_k occurs only if all relevant a_k have their maximum size and correct polarity. For N-level a_k symbols then, if M of the coefficients b_i in equation (31) are nonzero, each outer level has probability N^{-M}. Thus M replaces h in equation (8).

(*iii*) Finally, since the symbols a_k are statistically independent and have mean square value $E[a^2]$ as in equation (25), it is a simple matter to show that the symbols c_k have mean square value $E[c^2] = h_s E[a^2]$, where

$$h_s = \sum_{m=0}^{v} b_m^2 . \tag{32}$$

Therefore h_s replaces h_s in equation (25) in the evaluation of signal-to-noise ratio.

REFERENCES

1. Lucky, R. W., Salz, J., Weldon, E. J., Jr., "Principles of Data Communications," New York: McGraw Hill Book Co., 1968.
2. Lucky, R. W., "Automatic Equalization for Digital Communication," B.S.T.J., *44,* No. 4, part 1 (April 1965), pp. 547–588.
3. Gersho, A., "Adaptive Equalization of Highly Dispersive Channels for Data Transmission", B.S.T.J. *48,* No. 1, part 1 (January 1969), pp. 55–70.
4. Saltzberg, B. R., "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling", IEEE Trans. Inform. Theory *IT-14* (July 1968), pp. 563–568.
5. Aein, J. M., and Hancock, J. C., "Reducing the Effects of Intersymbol Interference with Correlation Receivers," IEEE Trans. Inform. Theory, *IT-9,* No. 3 (July 1963), pp. 167–175.
6. Aaron, M. R., and Tufts, D. W., "Intersymbol Interference and Error Probability," IEEE Trans. Inform. Theory, *IT-12,* No. 1 (January 1966), pp. 26–34.
7. Ho, E. Y., and Yeh, Y. S., "A New Approach for Evaluating the Error Probability in the Presence of Intersymbol Interference and Additive Gaussian Noise," B.S.T.J., *49,* No. 9, part 3 (November 1970), pp. 2249–2265.
8. Celebiler, M. I., and Shimbo, O., "Intersymbol Interference Considerations in Digital Communications," Int. Conf. on Commun., *ICC 1970,* San Francisco (June 8, 1970), pp. 8.1–8.10.
9. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. on Commun. Technology, *COM-14,* No. 1 (February 1966), pp. 67–68.
10. Norris, W. E., unpublished work, Bell Telephone Laboratories.
11. Cooley, J. W., Lewis, P. A., Welch, P. D., "Application of the Fast Fourier Transform to Computation of Fourier Integrals, Fourier Series, and Convolution Integrals," IEEE Trans. Audio and Electroactoustics, *AV-15* (June 1967), pp. 79–84.
12. Sunde, E. D., "Theoretical Fundamentals of Pulse Transmission," I and II, B.S.T.J., *33,* Nos. 3 and 4, parts 1 and 2 (May, July, 1954), pp. 721–88, 987–1010.
13. Gerrish, A. M., Howson, R. D., "Multilevel Partial-Response Signaling," IEEE Int. Conf. on Commun., Minneapolis, Minnesota, June 12, 1967, p. 186.

Contributors to This Issue

Walter J. Bertram, Jr., B.S., 1956, M.S., 1957, Ph.D. (Physics), 1961, Carnegie Inst. of Tech.; Bell Telephone Laboratories, 1960—. Mr. Bertram's early work at Bell Laboratories was on the development of low-noise traveling-wave tubes. From 1962 through 1966 he supervised a group developing low-noise wide-band parametric amplifiers, tunnel diode amplifiers, and down-converters. He then supervised a group involved in the development of electroluminescent diodes and in the development of acousto-optic modulators. Since 1970 he has supervised the development of imaging devices using the charge coupling principle. Member, American Physical Society, Institute of Electrical and Electronics Engineers, American Association for the Advancement of Science.

Kenneth Bullington, B.S., 1936, University of New Mexico; S.M., 1937, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1937—. Mr. Bullington has worked on transmission engineering problems on wire, radio, and submarine cable systems. He is now Radio Consultant. In 1956 he received the Morris Liebmann Memorial Prize of the Institute of Radio Engineers and the Franklin Institute's Stuart Ballantine Medal for contributions in tropospheric transmission and its application to practical communications systems. Fellow, IEEE. Member, Phi Kappa Phi, Sigma Tau, Kappa Mu Epsilon.

J. C. Candy, B.Sc., 1951, and Ph.D., 1954, University of Wales, Bangor, Wales; Research Associate at the University of Minnesota, 1959–60; Bell Telephone Laboratories, 1960—. Mr. Candy has worked on digital circuits and pulse transmission systems. He is studying methods of processing video signals, and is designing digital coders. Member, IEEE.

Robert W. Chang, B.S.E.E., 1955, National Taiwan University; M.S.E.E., 1960, North Carolina State University; Ph.D., 1965, Purdue University; Bendix Corporation, 1960–1963; Bell Telephone Laboratories, 1965—. Mr. Chang has worked on a variety of problems in data transmission and communication system theory. Member, Phi Kappa Phi, Eta Kappa Nu, Sigma Xi, Association for Computing Machinery, IEEE.

Donald C. Cox, B.S. (E.E.), 1959, and M.S. (E.E.), 1960, University of Nebraska; Ph.D. (E.E.), 1968, Stanford University; U. S. Air Force Research and Development Officer, Wright-Patterson AFB, Ohio, 1960–1963; Bell Telephone Laboratories, 1968—. Since coming to Bell Laboratories from Stanford, where he was engaged in microwave trans-horizon propagation research, Mr. Cox has been engaged in microwave propagation research in mobile radio environments and in high-capacity mobile radio systems studies. Member, IEEE, Commission II of USNC/URSI, Sigma Xi, Sigma Tau, Eta Kappa Nu, Pi Mu Epsilon.

Mrs. Marie A. Franke, A.A.S., 1967, Fayetteville Technical Institute; Bell Telephone Laboratories, 1967—. Since joining Bell Telephone Laboratories, Mrs. Franke has been engaged in developing digital techniques to reduce the frame-to-frame redundancy in television pictures for visual telephone applications.

Arthur D. Friedman, B.A., 1961, B.S., 1962, M.S., 1963, and Ph.D., 1965, Columbia University; Bell Telephone Laboratories 1965—. As a member of the Computing Techniques Research Department, Mr. Friedman has been interested in various switching theory problems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Barry G. Haskell, B.S., 1964, M.S., 1965, and Ph.D. (Electrical Engineering), 1968, University of California; Research Assistant, University of California, 1965–68; Bell Telephone Laboratories, 1968—. Mr. Haskell is engaged in TV picture processing studies. Member, Phi Beta Kappa, Sigma Xi.

Francis S. Hill, Jr., B.E., 1962, M.E., 1964, and Ph.D., 1968, Yale University; Bell Telephone Laboratories, 1967–1970. Mr. Hill was concerned with various analysis problems in the area of digital data communications. He is presently an Assistant Professor at the University of Massachusetts, Amherst, Mass. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

William B. Joyce, B.E.P., 1955, Cornell University; Ph.D., 1966, Ohio State University; Bell Telephone Laboratories, 1966—. Mr. Joyce has worked on various mathematical aspects of semiconductor device development. Member, APS.

WILLIAM C.-Y. LEE, B.S.C. in Engineering, 1954, Chinese Naval Academy; M.Sc. in E.E., 1960, and Ph.D. in E.E., 1963, Ohio State University; Bell Telephone Laboratories, 1964—. Mr. Lee has been concerned with the study of wave propagation in anisotropic medium and antenna theory. His present work has included studies of mobile radio antennas, signal fading problems, and communication systems, particularly those relating to the UHF and X-band regions. Member, Sigma Xi, IEEE.

JOHN O. LIMB, B.E.E., 1963, and Ph.D., 1967, University of Western Australia; Research Laboratories, Australian Post Office, 1966–1967; Bell Telephone Laboratories, 1967—. Mr. Limb has worked on the coding of picture signals to reduce channel capacity requirements. More recently this has included the coding of color pictures. He is now working on methods of reducing redundancy in moving pictures for *Picturephone*® visual telephone applications.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Telephone Laboratories at the University of Utah where he wrote a book on quantum electronics. He is presently working on the transmission aspect of a light communications system. Member, IEEE, Optical Society of America.

D. MAYDAN, B.Sc., 1957, and M.Sc., 1962, Technion, Israel Institute of Technology, Israel; Ph.D., 1965, Edinburgh University, Scotland; Israel Atomic Energy Commission, 1957–1962 and 1965–1967; Bell Telephone Laboratories, 1967—. Mr. Maydan is a member of the Active Optical Device Department and is working on acoustooptical devices and laser machining for fast image recording.

PREMACHANDRAN R. MENON, B.Sc. 1953, Banaras Hindu University, India; Ph.D., 1962, University of Washington; Bell Telephone Laboratories, 1963—. Mr. Menon is a member of the Computing Techniques Research Department, and has been engaged in research in switching theory. Member, IEEE, Sigma Xi.

F. W. Mounts, E.E., 1953, and M.S., 1956, University of Cincinnati; Bell Telephone Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, Eta Kappa Nu.

Charles M. Nagel, Jr., B.S., 1964, and M.S., 1967, Stevens Institute of Technology; Bell Telephone Laboratories, 1965–1971. At the time of his death, Mr. Nagel was a member of the Applied Mathematics and Statistics Department, working in the area of optical communications theory. Previously, he had been engaged in research into the numerical aspects of various waveguide propagation problems, Faraday rotation in the ionosphere, and electromagnetic scattering. Mr. Nagel was a member of Tau Beta Pi, Pi Delta Epsilon, and the American Association for the Advancement of Science.

Scotty R. Neal, B.A. (Mathematics), 1961, M.A. (Mathematics), 1963, and Ph.D. (Mathematics), 1965, University of California, Riverside; Research Mathematician, Naval Weapons Center, China Lake, California, 1964–1967; Bell Telephone Laboratories, 1967—. Since coming to Bell Laboratories, Mr. Neal has been primarily concerned with the analysis of various aspects of telephone traffic systems. He has also worked on applications of optimal linear estimation theory and certain aspects of communication theory. Member, American Math Society, SIAM, Sigma Xi.

R. F. W. Pease, B.A., 1960, M.A. and Ph.D., 1964, University of Cambridge; Bell Telephone Laboratories, 1967—. Mr. Pease held a faculty appointment at the University of California at Berkeley prior to joining Bell Laboratories and worked on electron microscopy. He is now trying to efficiently encode moving and still pictures.

Douglas O. Reudink, B.A., 1961, Linfield College; Ph.D. (Mathematics), 1965, Oregon State University; Bell Telephone Laboratories, 1964—. Since joining Bell Laboratories, Mr. Reudink has been engaged in electronic systems research with particular emphasis in the field of mobile communications. His recent work has been concerned with fundamentals of mobile radio propagation, diversity techniques, and the configuration and control of mobile systems. Member, Sigma Pi Sigma, American Mathematical Society, Pi Mu Epsilon, IEEE.

M. V. SCHNEIDER, M.S., 1956, and Ph.D., 1959, Swiss Federal Institute of Technology, Zurich, Switzerland; Bell Telephone Laboratories, 1961—. Mr. Schneider has been engaged in experimental work on thin film solid-state devices, Schottky barrier photodetectors, and thin film mode selection filters. He has published contributions on microstrip and thin film technology, on the design of microstrip circuits, and on the reduction of conductor and dielectric losses in microstrip transmission lines. He is presently working on hybrid integrated circuits for systems applications at microwave and millimeter-wave frequencies. Member, American Physical Society, American Vacuum Society, IEEE.

N. L. SCHRYER, B.S., 1965, M.S., 1966, Ph.D., 1969, University of Michigan; Bell Telephone Laboratories, 1969—. Mr. Schryer has worked on the numerical solution of parabolic and elliptic partial differential equations. He is currently studying problems of this type which arise in semiconductor device theory.

WILLIAM W. SNELL, JR., Williamsport Technical Institute, 1951; Bell Telephone Laboratories, 1955—. In his first years with Bell Laboratories Mr. Snell was concerned with the design of waveguide components for use in the 4-, 6-, and 11-GHz common carrier band. He later participated in the Shotput and Project Echo satellite communications experiments where he designed several components of the Holmdel Space Communication Receiver. He is presently concerned with the design and fabrication of high-order varactor frequency multipliers, high-quality varactor diodes, and low-loss microstrip filters for use in hybrid integrated circuits at frequencies above 10 GHz.

R. J. STRAIN, B.S.E.E., 1958, M.S., 1959, Ph.D., 1963, University of Illinois; Standard Telecommunication Laboratories, Harlow, England; Bell Telephone Laboratories, 1965—. Mr. Strain's initial activities in the Electron Device Laboratory were concerned with electroluminescent display devices, particularly GaP diodes. In 1968 he joined the Semiconductor Device Laboratory as a supervisor, and he is now in charge of the Functional Interface Device Group. Member, APS, AAAS, ECS, IEEE, Sigma Xi.