Practical Design Rules for Optimum Finite Impulse Response Low-Pass Digital Filters

By O. HERRMANN,* L. R. RABINER, and D. S. K. CHAN

(Manuscript received November 22, 1972)

Although a great deal is known about design techniques for optimum (in a minimax error sense) finite impulse response (FIR) low-pass digital filters, there have not been established any practical design rules for such filters. Thus, a user is unable to easily decide on the (approximate or exact) filter order required to meet his design specifications and must resort to tables or trial and error procedures. In this paper, such a set of design rules is given. In the case of very narrow bandwidth or very wide bandwidth filters, analytic relations between the filter parameters can be readily obtained. In all other cases, exceedingly good linear and nonlinear fits to the data can be obtained over somewhat restricted ranges of the parameters. These fitting procedures lead to a practical set of simple design rules for estimating filter order from the desired specifications.

I. INTRODUCTION

The problem of designing an optimal (in the minimax sense) low-pass FIR digital filter to meet design specifications has been thoroughly investigated[1-5] and may be considered to be solved. Thus, given a

---

* University of Erlangen—Nuremberg, Germany.
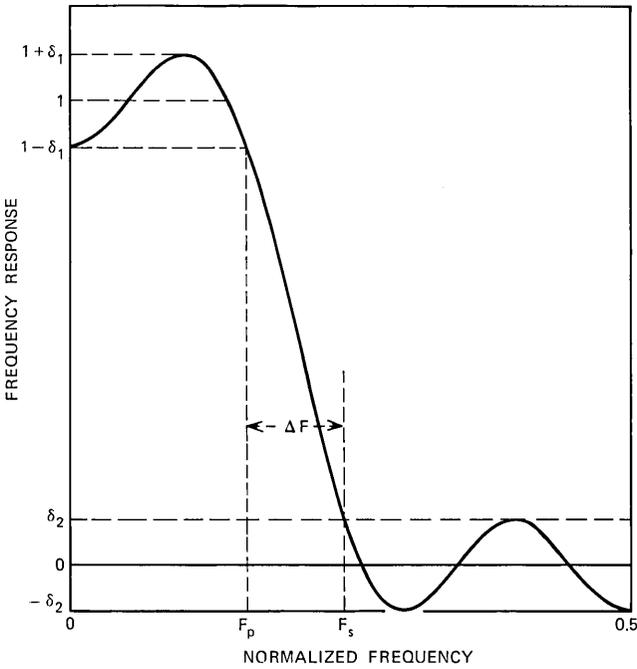
Fig. 1—Definition of low-pass filter parameters.

specified impulse response duration of N samples, a passband cutoff frequency F_p (see Fig. 1), a stopband cutoff frequency F_s, and a ratio of tolerances $K = \delta_1/\delta_2$ (where δ_1 = passband tolerance and δ_2 = stopband tolerance), an optimal approximation to these specifications can be designed. The approximation is optimal in the sense that, for given values of N, F_p, F_s, and K, δ_1 (or equivalently δ_2) is minimum. The nature of the solution is such that there are three distinct classes into which it may belong, depending on the specific design parameters. These classes have been called extraripple filters,[4,6] scaled extraripple filters,[7] and equiripple filters with one less than the maximum possible number of ripples. The differences between these classes lie in the number and amplitude of the ripples in the weighted error curve. The weighted error curve is defined as:

$$E(e^{j2\pi f}) = \begin{cases} \dfrac{1 - H(e^{j2\pi f})}{K} & 0 \leq f \leq F_p \\ -H(e^{j2\pi f}) & F_s \leq f \leq 0.5, \end{cases} \qquad (1)$$

where $H(e^{j2\pi f})$ is the frequency response of the optimal filter. Extra-

ripple filters have $(N + 5)/2$ equal magnitude extrema in their error curves. Scaled extraripple filters also have $(N + 5)/2$ extrema, all but one of which are of equal magnitude. The third class of solution has $(N + 3)/2$ equal amplitude extrema in its error curves. Figure 2 shows plots of curves of transition bandwidth, $\Delta F = F_s - F_p$, versus passband cutoff frequency, F_p, for two sets of conditions.[6] The data in Fig. 2a show the curves for $K = 1$, $\delta_1 = \delta_2 = 0.1$, $N = 9$, and $N = 11$; whereas Fig. 2b shows the curves for $K = 100$, $\delta_1 = 0.01$ ($\delta_2 = 0.0001$), $N = 19$, and $N = 21$. The minima along each curve are the extraripple filters. (For fixed values of δ_1 and δ_2, as in this figure, there are only $(N - 1)/2$ distinct extraripple filters.) The local regions around the minima are the scaled extraripple filters[7] and the remainder of the curve represents equiripple solutions with one less than the maximum number of ripples. As will be shown in the next section, the first and last extraripple filters can be obtained analytically because they are simply related to the Chebyshev polynomial of appropriate degree.[8,9]

As seen from Fig. 2, as F_p varies from 0 to its maximum possible value, the transition width goes through a sequence of minima and maxima. The variation in the transition widths of the minima and maxima decreases as N increases. Furthermore, the variation in transition width between adjacent maxima and minima also decreases as N increases. In fact, except for a narrow region at the beginning and end of the curve, the curve of transition width versus passband cutoff frequency is relatively flat over a wide range of values of F_p, δ_1, and K. Figure 3 shows a sequence of three plots of transition width versus passband cutoff frequency for extraripple filters (i.e., only the minima of the curve are plotted) of length $N = 101$ for various values of δ_1 and K. (The entire curves are not plotted because the amount of computation required for a smooth curve of such high order is inordinately high.) Figure 3a shows a sequence of four curves for $K = 1$, $\delta_1 = 0.1$, 0.01, 0.001, 0.0001, whereas Fig. 3b shows the same sequence for $K = 10$, and Fig. 3c shows the sequence for $K = 100$. Several observations can be made from this figure.

(*i*) For a wide range of values of F_p, and fixed δ_1 and δ_2, the transition width of the extraripple filters is relatively insensitive to F_p. The larger the value of K, or the smaller the value of δ_1, the worse this approximation becomes.

(*ii*) In the regions of either very small or very large values of F_p, the transition width generally decreases.

Based on these curves and on previous results with window designs,[10]
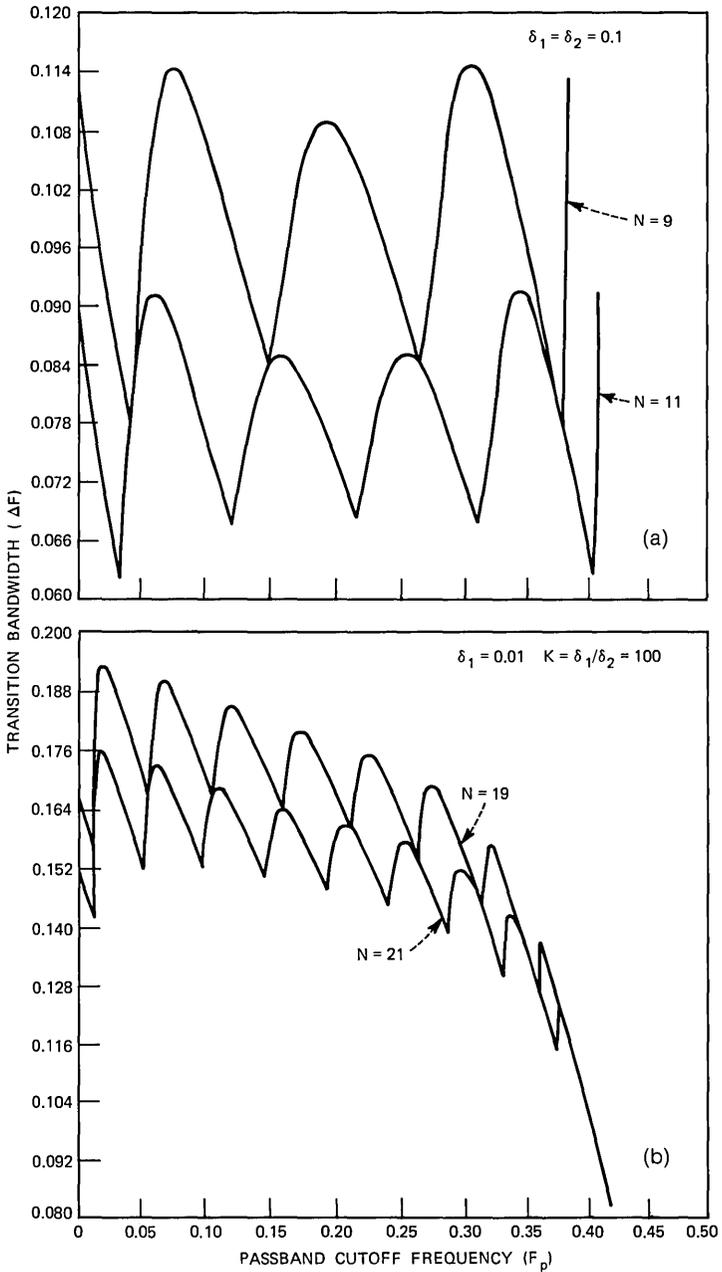
Fig. 2—The transition width as a function of passband cutoff frequency for two sets of filter parameters.
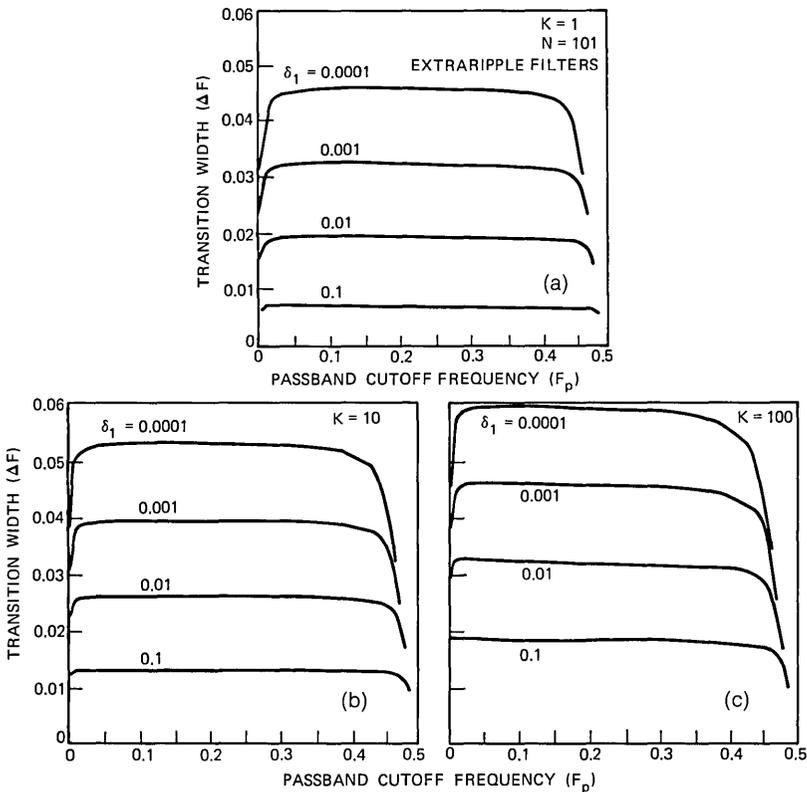
Fig. 3—The transition width as a function of passband cutoff frequency for $N = 101$-point extraripple filters and $K = 1$, 10, and 100.

it seems fairly reasonable to expect some simple design relationships to exist between the five basic filter parameters, N, F_p, F_s, δ_1, and δ_2 (or K), at least in the extreme case of the Chebyshev solution, and also for a reasonably large region near $F_p = 0.25$. Experience in practical situations has shown that the number of terms needed in the optimum FIR low-pass filter to meet design specifications is significantly less than the number of terms estimated by known relationships on windows.[10] Therefore, the goals of this paper are to obtain approximations to the actual design relationships between linear-phase, low-pass filter design parameters and to illustrate their use in actual design examples.

The organization of this paper is as follows. First, the design relationships for the Chebyshev solution are derived, and approximate formulas for the transition width in the limit of large values of N are obtained. Then the results of measurements on a wide range of filters

where the number of passband and stopband ripples are equal are shown. Then minimum mean-square relative error fits to the observed data (for large N) assuming both linear and nonlinear dependency on the basic filter parameters are computed. To apply the design relationships for all values of N, a correction formula is derived, based on consideration of the appropriate transition width of the filter. Finally, a set of rules is presented for going from a set of desired filter parameters to an estimate of the lowest-order filter which meets these specifications.

1.1 Summary of Design Relationships

Given the low-pass filter parameters F_p, F_s, δ_1, and δ_2 (or, equivalently, $K = \delta_1/\delta_2$), the minimum filter impulse response duration, N required to meet the above specifications can be estimated from the relation

$$N = \frac{D_\infty(\delta_1, \delta_2) - f(K)(F_s - F_p)^2}{(F_s - F_p)} + 1,$$

where

$$D_\infty(\delta_1, \delta_2) = [5.309 \times 10^{-3} (\log_{10} \delta_1)^2 + 7.114 \times 10^{-2} (\log_{10} \delta_1)$$
$$- 0.4761] \log_{10} \delta_2 - [2.66 \times 10^{-3} (\log_{10} \delta_1)^2$$
$$+ 0.5941 (\log_{10} \delta_1) + 0.4278]$$

and

$$f(K) = 0.51244 \log_{10} K + 11.01217.$$

The above relations are valid to within 1.3 percent relative error in N if $\delta_1 \leqq 0.1$ and $\delta_2 \leqq 0.1$.

II. CHEBYSHEV SOLUTIONS

Let $\{h(n), n = -(N - 1)/2, \cdots, (N - 1)/2\}$ be the impulse response of the desired digital filter. (N is assumed to be odd throughout this paper.) The impulse response satisfies the symmetry condition $h(n) = h(-n)$ to give the desired linear phase. The frequency response of the filter is given by

$$H(e^{j2\pi f}) = h(0) + \sum_{n=1}^{(N-1)/2} 2h(n) \cos (2\pi fn). \tag{2}$$

By making the substitution

$$\cos (2\pi f) = \frac{x - \left(\dfrac{X_0 - 1}{2}\right)}{\left(\dfrac{X_0 + 1}{2}\right)} \tag{3a}$$

or

$$x = \left(\frac{X_0 + 1}{2} \right) \cos (2\pi f) + \left(\frac{X_0 - 1}{2} \right) \tag{3b}$$

the interval $0 \leqq f \leqq 0.5$ is mapped to the interval $X_0 \geqq x \geqq -1$. It is easily shown[3,7] that the mapping transforms the trigonometric polynomial of eq. (1) to an algebraic polynomial in x of the form

$$G(x) = \sum_{n=0}^{(N-1)/2} b(n)x^n, \tag{4}$$

where the $\{b(n)\}$ are straightforwardly related to the $\{h(n)\}$.

The basic filter design problem is to find coefficients $b(n)$ [or $h(n)$] such that the weighted error of approximation is equiripple in both the passband and stopband. In the case when either the passband or the stopband has only one ripple, the solution to the filter design problem may be obtained analytically based on the theory of Chebyshev polynomials. In all other cases, alternative techniques must be used to arrive at the appropriate solution.

Consider the Mth degree Chebyshev polynomial, $T_M(x)$, as shown in Fig. 4 for $M = 4$. The standard representation for $T_M(x)$ is

$$T_M(x) = \cos [M \cos^{-1}x], \qquad |x| \leqq 1 \tag{5}$$

and

$$T_M(x) = \cosh [M \cosh^{-1} x], \qquad |x| > 1. \tag{6}$$

In the interval $-1 \leqq x \leqq 1$, $T_M(x)$ oscillates between ± 1 and, beyond the value $x = +1$, $T_M(x)$ grows approximately as x^M. If we define X_0 (see Fig. 4) as the point where $T_M(X_0) = (1 + \delta_1)/\delta_2$, and X_p as the point where $T_M(X_p) = (1 - \delta_1)/\delta_2$, it is readily seen that $\delta_2 \cdot T_M(x)$ is a polynomial of the form of eq. (4) [with $M = (N - 1)/2$] which is equiripple in both the passband $X_p \leqq x \leqq +X_0$, and the stopband $-1 \leqq x \leqq +1$, and hence is an optimal solution to the filter design problem. Of course, this solution is a special case of the general solution in that there is only one ripple in the passband, but at least it is an analytically tractable case from which a great deal can be learned about the relationships between the filter parameters.

It is straightforward to solve for the points X_p and X_0 in terms of δ_1, δ_2, and N. If we set $M = (N - 1)/2$, then at $x = X_0 > 1.0$, we get the relation

$$T_M(X_0) = \frac{1 + \delta_1}{\delta_2} = \cosh [M \cosh^{-1} X_0] \tag{7}$$

or

$$X_0 = \cosh \left[\frac{1}{M} \cosh^{-1} \left(\frac{1 + \delta_1}{\delta_2} \right) \right]. \tag{8}$$
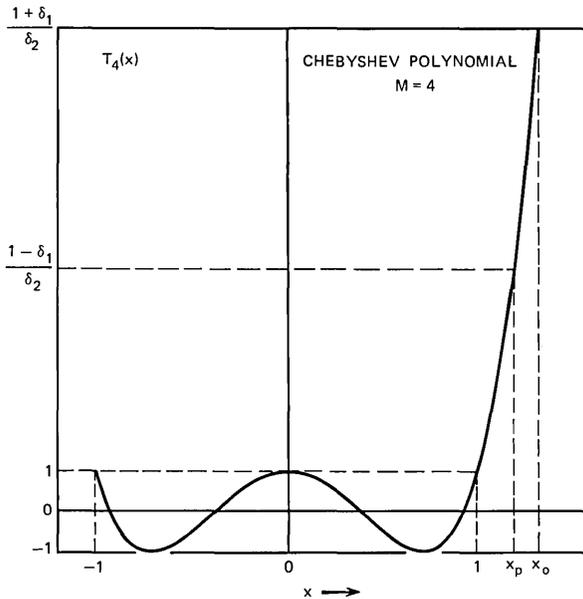
Fig. 4—A fourth-degree Chebyshev polynomial defining the points X_0 and X_p.

At $x = X_p > 1.0$, we get the relation

$$T_M(X_p) = \frac{1 - \delta_1}{\delta_2} = \cosh \left[M \cosh^{-1} X_p \right] \qquad (9)$$

or

$$X_p = \cosh \left[\frac{1}{M} \cosh^{-1} \left(\frac{1 - \delta_1}{\delta_2} \right) \right]. \qquad (10)$$

The inverse mapping of eq. (3) can be used to determine the filter cutoff frequencies by the relation

$$f = \frac{1}{2\pi} \cos^{-1} \left[\frac{2x - X_0 + 1}{X_0 + 1} \right], \qquad |x| \le X_0. \qquad (11)$$

Thus, F_p (corresponding to X_p) and F_s (corresponding to $x = +1.0$) can be readily obtained using eq. (11) with the appropriate values for x. In this case, the transition width $(F_s - F_p)$ can be analytically determined for all values of N, δ_1, and δ_2 as:

$$\Delta F = F_s - F_p = \frac{1}{2\pi} \left[\cos^{-1} \left(\frac{3 - X_0}{1 + X_0} \right) \right.$$
$$\left. - \cos^{-1} \left(\frac{2X_p - X_0 + 1}{1 + X_0} \right) \right]. \qquad (12)$$

In the limit of large values of N (or, equivalently, M), eq. (12) can be simplified by the following approximations. First, for sufficiently small values of α,

$$\cosh \alpha \approx 1 + \frac{\alpha^2}{2}, \qquad \alpha \ll 1. \tag{13}$$

Thus, eqs. (10) and (8) can be simplified to the form [replacing M by $(N-1)/2$]

$$X_p \approx 1 + \frac{\left[\left(\frac{2}{N-1}\right)\cosh^{-1}\left(\frac{1-\delta_1}{\delta_2}\right)\right]^2}{2}. \tag{14}$$

$$X_0 \approx 1 + \frac{\left[\left(\frac{2}{N-1}\right)\cosh^{-1}\left(\frac{1+\delta_1}{\delta_2}\right)\right]^2}{2}. \tag{15}$$

The approximation is then made that for sufficiently small ϵ,

$$\cos^{-1}(1-\epsilon) \approx \sqrt{2\epsilon}. \tag{16}$$

Thus, F_s and F_p are well approximated as

$$F_s \approx \frac{1}{2\pi}\left[\frac{2}{N-1}\cosh^{-1}\left(\frac{1+\delta_1}{\delta_2}\right)\right] \tag{17}$$

$$F_p \approx \frac{1}{2\pi}\left(\frac{2}{N-1}\right)\left[\left[\cosh^{-1}\left(\frac{1+\delta_1}{\delta_2}\right)\right]^2 - \left[\cosh^{-1}\left(\frac{1-\delta_1}{\delta_2}\right)\right]^2\right]^{\frac{1}{2}}. \tag{18}$$

Thus, for large N ($N \gg 1$) the approximation to the transition width curve is given by

$$\Delta F = F_s - F_p \approx \frac{1}{\pi(N-1)}\left[\cosh^{-1}\left(\frac{1+\delta_1}{\delta_2}\right)\right.$$
$$\left. - \left[\left[\cosh^{-1}\left(\frac{1+\delta_1}{\delta_2}\right)\right]^2 - \left[\cosh^{-1}\left(\frac{1-\delta_1}{\delta_2}\right)\right]^2\right]^{\frac{1}{2}}\right. \tag{19}$$

(This approximation is valid to within 1 percent for $N \geqq 51$.) Equation (19) shows ΔF to be inversely proportional to $(N-1)$. This identical inverse behavior has been noted previously for filters designed by windowing.[10] These and other considerations lead one to consider as a performance measure of a low-pass digital filter the quantity D defined as

$$D = (N-1)\Delta F = (N-1)(F_s - F_p) \tag{20}$$

which, in many cases, depends only on δ_1 and δ_2.

The curves in Fig. 5 (for $N = 127$) show plots of performance D versus $20 \log_{10} (\delta_2)$ both for values of $K = \delta_1/\delta_2$ from $K = 1$ to $K = 1000$ and for values of δ_1 from 0.5 to 0.0001. The behavior of these curves is as predicted from eq. (19) and from intuition about the behavior of D as δ_1 and δ_2 get large. It is clear that when $\delta_1 + \delta_2 = 1.0$ [i.e., $\delta_1 = K/(K + 1)$, $\delta_2 = 1/(K + 1)$], then $D = 0$ since there is no transition band. In this case, the term $\cosh^{-1} (1 - \delta_1)/\delta_2$ vanishes, and the first and second terms in eq. (19) cancel exactly. In the limit of small values of δ_1, the second and third terms in eq. (19) are approximately equal and cancel. Thus, D is approximately of the form

$$D \approx \frac{1}{\pi} \cosh^{-1} \left(\frac{1 + \delta_1}{\delta_2} \right). \tag{21}$$

Since

$$\cosh^{-1} (x) = \ln (x + \sqrt{x^2 - 1}), \tag{22}$$

eq. (21) can be rewritten as (assuming δ_1 negligible)

$$D \approx \frac{1}{\pi} (\ln 2 - \ln \delta_2) \tag{23}$$

which is independent of δ_1. Thus, as seen in Fig. 5b, in the case of small δ_1, D is essentially independent of δ_1.

The curves of Fig. 6 show the behavior of D for various values of $N \leq 127$, for $K = 1, 10,$ and 100. The values of N used for these plots were $N = 3, 7, 11, 19, 51,$ and 127. The differences between the data for $N = 127$ and the data for $N = 51$ are relatively small. These curves also exhibit another interesting phenomenon. The curve for $N = 3$ saturates at a value of $D = 1$. This is due to the limitation that the transition width, ΔF, must be less than or equal to 0.5. Thus, the saturation value of D is $(N - 1)/2$, or 1.0 for $N = 3$. The larger the value of K, the larger the value of δ_2 beyond which the curve for $N = 3$ saturates.

In the case of the Chebyshev solution to the optimal filter design problem, a formula can be derived for the impulse response duration N of a filter whose response meets specified values of δ_1, δ_2, and F_s. Since F_p is not specified, this result is useful only as a first guess of a value of N which meets specifications on all four filter parameters. From the discussion given earlier in this section, $f = F_s$ when $x = 1.0$. Thus, eq. (11) can be used to solve for X_0 as

$$X_0 = \frac{3 - \cos (2\pi F_s)}{1 + \cos (2\pi F_s)}. \tag{24}$$
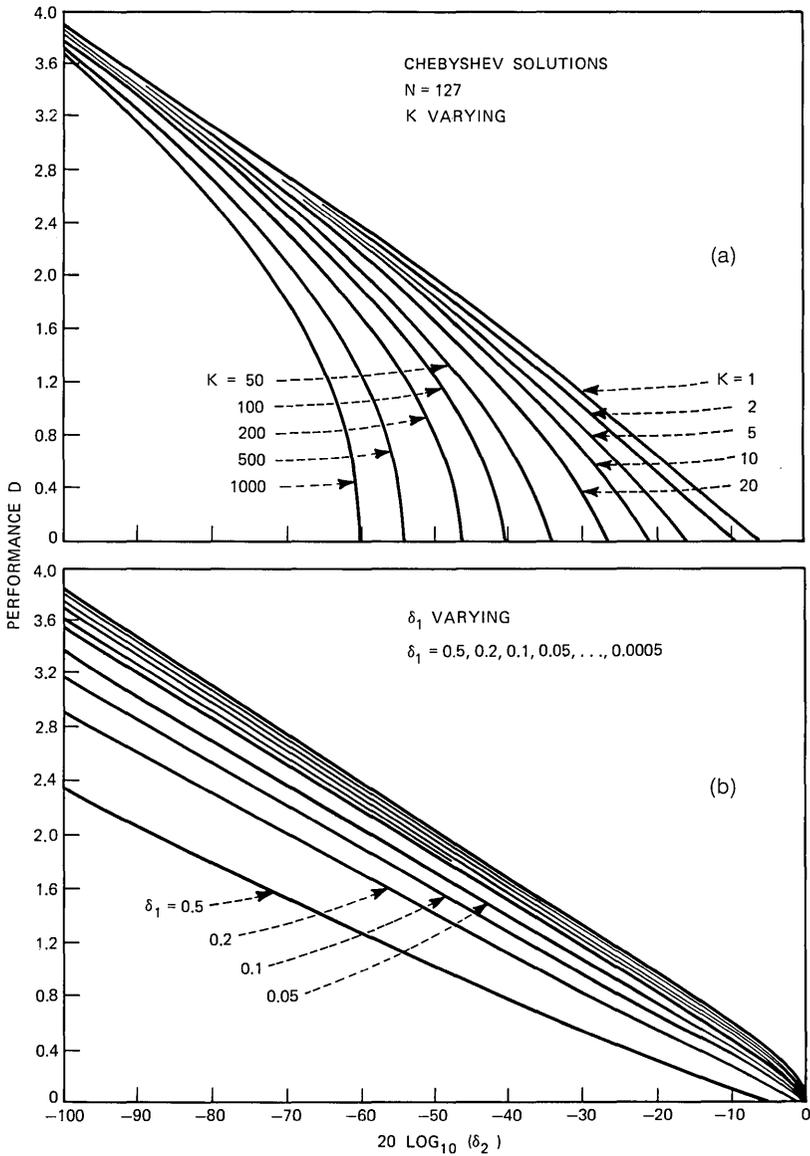
Fig. 5—Plots of D versus $\log_{10} \delta_2$ as a function of K and δ_1 for the $N = 127$ Chebyshev solutions.
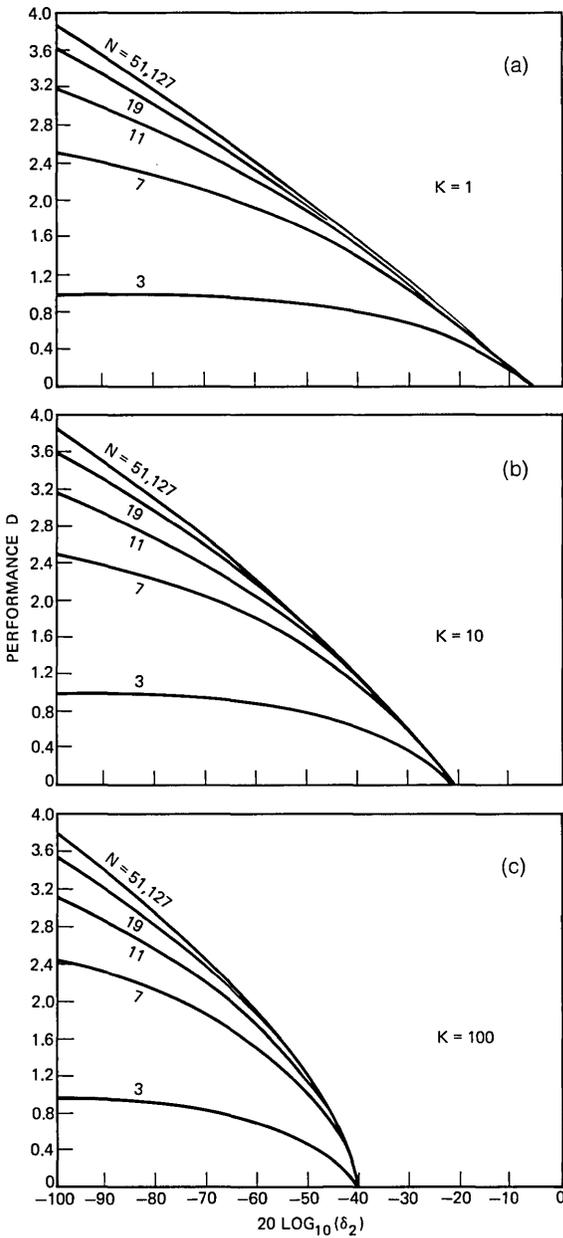
Fig. 6—Plots of D versus $\log_{10} \delta_2$ as a function of N for $K = 1$, 10, and 100, for the Chebyshev solutions.

Equation (8) may now be used to solve for N as

$$N = 1 + \frac{2 \cosh^{-1}\left(\dfrac{1 + \delta_1}{\delta_2}\right)}{\cosh^{-1}\left[\dfrac{3 - \cos(2\pi F_s)}{1 + \cos(2\pi F_s)}\right]}. \tag{25}$$

Using trigonometric identities, eq. (25) can be simplified to the form

$$N = 1 + \frac{\cosh^{-1}\left(\dfrac{1 + \delta_1}{\delta_2}\right)}{\cosh^{-1}\left(\dfrac{1}{\cos(\pi F_s)}\right)}. \tag{26}$$

For these values of X_0 and N, F_p may be obtained from eqs. (10) and (11). If the transition width obtained is too large (too small), N is then decreased (increased) until the desired specifications are approximately achieved. (Since there is only one Chebyshev solution for fixed δ_1, δ_2, and N, exact values of both F_p and ΔF cannot usually be obtained.) As will be seen later, eq. (26) forms the basis for estimating a lower bound on the filter order required to meet given design specifications on ΔF, δ_1, and δ_2.

As mentioned earlier, Chebyshev polynomials can be used as the optimal solution in the case when there is one ripple in either the pass-band or the stopband. To see how this second case can be handled, consider the Chebyshev polynomial $T_M(x)$ where $T_M(X_0) = (1 + \hat{\delta}_2)/\hat{\delta}_1$ and $M = (N - 1)/2$. As shown earlier, this polynomial represents an optimal filter with passband ripple $\hat{\delta}_2$, stopband ripple $\hat{\delta}_1$, passband cutoff frequency F_p, and stopband cutoff frequency F_s. Consider the function $R_M(x)$ defined as

$$R_M(x) = 1 - \hat{\delta}_1 T_M(-x + X_0 - 1), \tag{27}$$

where $-1 \leqq x \leqq X_0$. An examination of the properties of $R_M(x)$ shows

(i) $R_M(x)$ is a polynomial in x of degree M.
(ii) In the interval $X_0 - 2 \leqq x \leqq X_0$, $R_M(x)$ oscillates between the values $1 - \hat{\delta}_1$, and $1 + \hat{\delta}_1$.
(iii) In the interval $-1 \leqq x \leqq X_0 - 1 - X_p$, $R_M(x)$ goes from $-\hat{\delta}_2$ to $\hat{\delta}_2$.

If we define $\hat{F}_p$ and $\hat{F}_s$ as the equivalent filter cutoff frequencies, then

it is readily shown that

$$\hat{F}_p = 0.5 - F_s$$
$$\hat{F}_s = 0.5 - F_p. \tag{28}$$

Thus $R_M(x)$ is a polynomial with only one stopband ripple which satisfies the filter optimality criterion. In summary, to design Chebyshev approximations with only one *stopband* ripple, one merely makes the substitution

$$\delta_1 = \hat{\delta}_2$$
$$\delta_2 = \hat{\delta}_1$$
$$F_p = 0.5 - \hat{F}_s$$
$$F_s = 0.5 - \hat{F}_p \tag{29}$$

and solves for an equivalent filter with one *passband* ripple using the formulas of this section.

The data on the Chebyshev solutions provide valuable insights into the behavior of the design relationships between filter parameters in more general cases. These data and their design relationships are presented in the next sections.

III. MEASUREMENTS OF D

Earlier, it was shown that the transition width or, equivalently, the performance measure D for the Chebyshev solutions (i.e., either small F_p or large F_s) was generally significantly smaller than for most of the range of values of passband cutoff frequency. To obtain data on a more realistic set of filters, the value of D was measured for a large number of extraripple filters designed with the constraint that the number of passband and stopband ripples were the same. In this manner, the passband and stopband widths were almost equal and, thus, the measured data would characterize the parameter set over as wide a range of values of δ_1, δ_2, and F_p as possible. Six different values of N were used including $N = 3$, 7, 11, 19, 51, and 127. Over 1500 filters were designed to cover the parameter range $0.00001 \leq \delta_2 \leq 0.5$, and $1 \leq K \leq 500$ ($K = \delta_1/\delta_2$). Figures 7a through 7i show plots of D versus $20 \log_{10}(\delta_2)$ for the nine values of K and the chosen values of N. All these data are presented since they are fairly general and may be useful in a wide variety of contexts other than this paper. (Also, their measurement required almost 3 hours of computer time on a fairly fast processor.)

It is remarkable how similar the plots of D versus $\log \delta_2$ for the more general case of Fig. 7 are to the identical plots for the Chebyshev solutions. Similar behavior for small D is expected, since D tends to 0 as

$\delta_1 + \delta_2$ tends to 1.0 independent of N, F_p, and F_s. However, the approximately linear behavior of D as a function of log δ_2 and log K (for large N) is unexpected. Another similarity between the two cases is the independence of D of $(N - 1)$ for large N. The tendency of D to saturate for small values of N is yet another similarity between the curves. The main difference between the sets of curves is that, in the Chebyshev case, D is approximately independent of δ_1 for small δ_1. This behavior is not observed for the extraripple filters of Fig. 7.

A summary of the behavior of D for $N = 127$, as a function of log δ_2 for various values of either K or δ_1, is presented in Fig. 8. The values of δ_1 used were 0.5, 0.2, 0.1, $\cdots$, 0.00002, whereas K ranged from 1 to 500 as in Figs. 7a through 7i. In some sense, this figure represents a set of design curves for high-degree low-pass filters. In the next section, we show how the data of Fig. 8 can be approximated by linear and non-linear fits, and how simple modifications can be made to correct the results for values of N less than 127.

IV. DATA-FITTING PROCEDURES

In order to make most efficient use of the data of the previous section in a practical design problem, it is useful to express the relationships between the filter parameters in a simple manner. Since we know of no way of deriving exact analytical formulas, as in the Chebyshev case, a minimum mean-square relative error fit to the data over a restricted but reasonable range was sought. Both a linear and a nonlinear fit to the data of Fig. 8 (N large) were obtained. Corrections for smaller values of N were then obtained giving a complete set of design rules.

The data of Fig. 8 suggest that except in the region $D = 0$ (large values of δ_1) a simple linear fit can be obtained. The curves of Fig. 8 were assumed to be of the form

$$
\begin{aligned}
D_{\text{LIN}} &= a + b \log_{10}(\delta_2) + c \log_{10} K, \\
a &= -0.803, \\
b &= -1.359, \\
c &= -0.737.
\end{aligned}
\tag{30}
$$

where D_{LIN} is the predicted value of D. The values for a, b, and c were chosen to minimize the sum of the squares of the relative differences between D_{LIN} and D for the $N = 127$ data for values of δ_1 in the range $0.01 \geqq \delta_1$; $0.01 \geqq \delta_2$. The reason relative rather than absolute errors were considered is that a fixed percentage error in D, δD, approximately gives a fixed percentage error in N, δN, when transition width is held
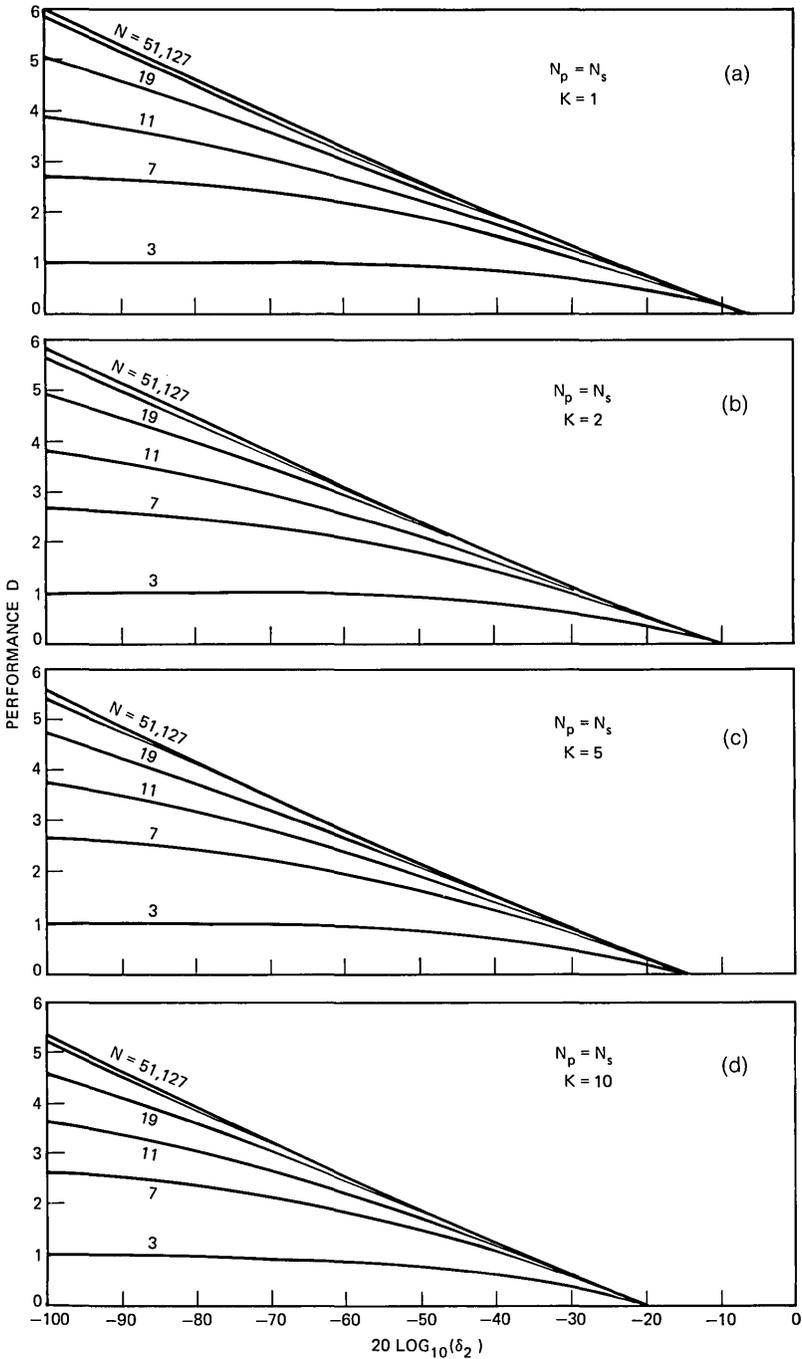
Fig. 7—Plots of D versus $\log_{10} \delta_2$ as a function of N for $K = 1, 2, 5, 10, 20, 50, 100, 200, 500$ for the case of extraripple filters with the same number of passband and stopband ripples.
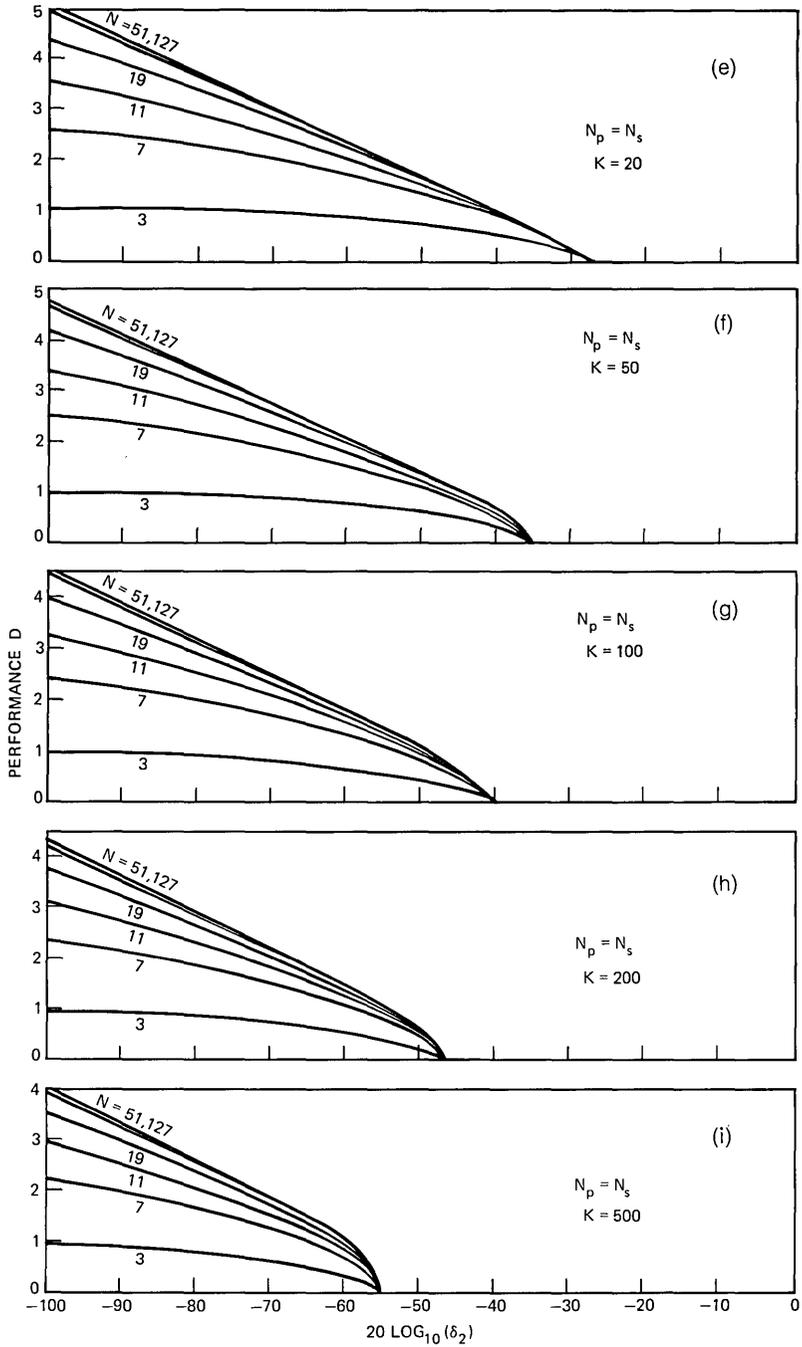
Fig. 7 (continued).
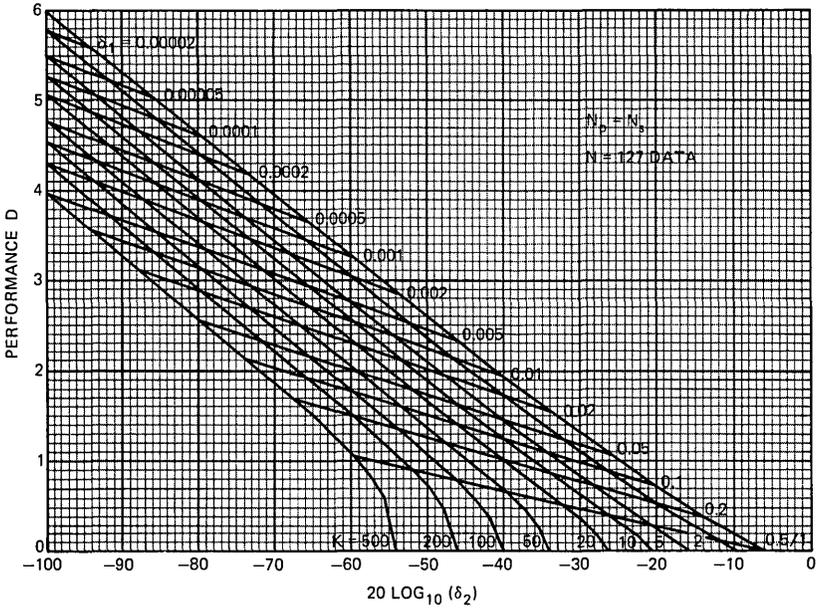
Fig. 8—Plot of D versus $\log_{10} \delta_2$ for the $N = 127$ data of Fig. 7 showing curves for various values of K and δ_1.

fixed, as in most design problems, i.e.,

$$\frac{\delta D}{D} = \frac{\delta(N-1)}{N-1} = \frac{\delta N}{N-1}. \tag{31}$$

Figure 9 shows a plot of the relative error of the predicted values of D versus $\log \delta_2$ for $\delta_1 = 0.01, 0.005, 0.002, \cdots, 0.00002$. Except for a small region on the curve $\delta_1 = 0.01$, the relative error is less than 1.0 percent for the entire range of δ_2 and δ_1 considered. Based on a value of $N = 127$, a relative error of 1.0 percent in D is equivalent to an error of 1.26 samples in N, or approximately one-half a filter order off from the correct order. Errors of this magnitude are generally considered to be quite small, i.e., the prediction is reasonably good.

In an effort to improve the fit and extend the range of applicability of the approximation, a nonlinear formula was chosen for D. Based on the data of Fig. 8, it was observed that the slope of the curve of D versus $\log_{10} \delta_2$ changes nonlinearly with $\log \delta_1$. The simplest approximation was to try a fit which was linear with respect to $\log \delta_2$ and

quadratic in $\log \delta_1$. Such a fit is of the form

$$D_{NL} = [a_1(\log_{10} \delta_1)^2 + a_2 \log_{10} \delta_1 + a_3] \log_{10} \delta_2$$
$$+ [a_4(\log_{10} \delta_1)^2 + a_5 \log_{10} \delta_1 + a_6]. \quad (32)$$

The constants a_1 to a_6 were chosen to minimize the mean-square relative error for the $N = 127$ data over the range $0.1 \geq \delta_1 \geq 0.000001$, $0.1 \geq \delta_2 \geq 0.000001$, and turned out to be

$$a_1 = 5.309 \times 10^{-3}$$
$$a_2 = 7.114 \times 10^{-2}$$
$$a_3 = -4.761 \times 10^{-1}$$
$$a_4 = -2.660 \times 10^{-3}$$
$$a_5 = -5.941 \times 10^{-1}$$
$$a_6 = -4.278 \times 10^{-1}.$$

Figure 10 shows the relative error of the predicted value of D as a function of $\log \delta_2$ for values of δ_1 from 0.1 to 0.00002. The peak percentage error is 1.3 percent, and over most of the range the percentage error is
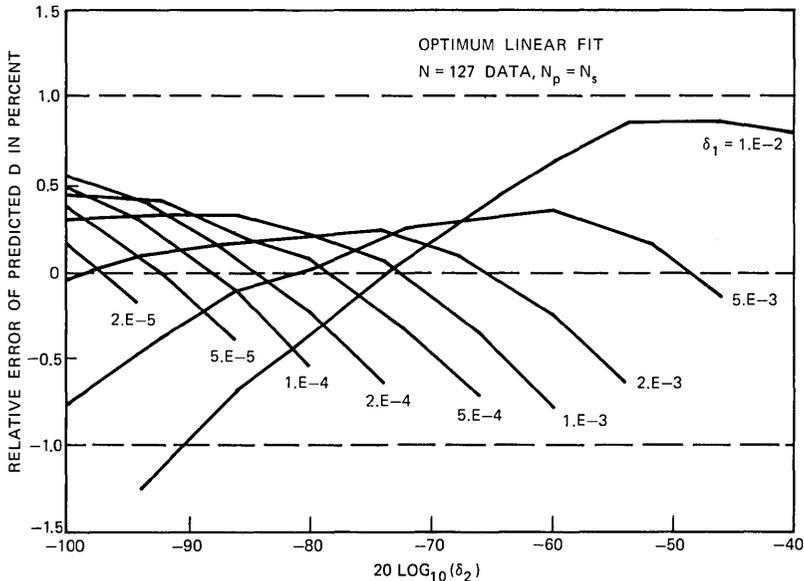
Fig. 9—The relative errors in fitting the data of Fig. 8 with a linear curve over the range $\delta_1 \leq 0.01$, for various values of δ_1 $(N = 127)$.
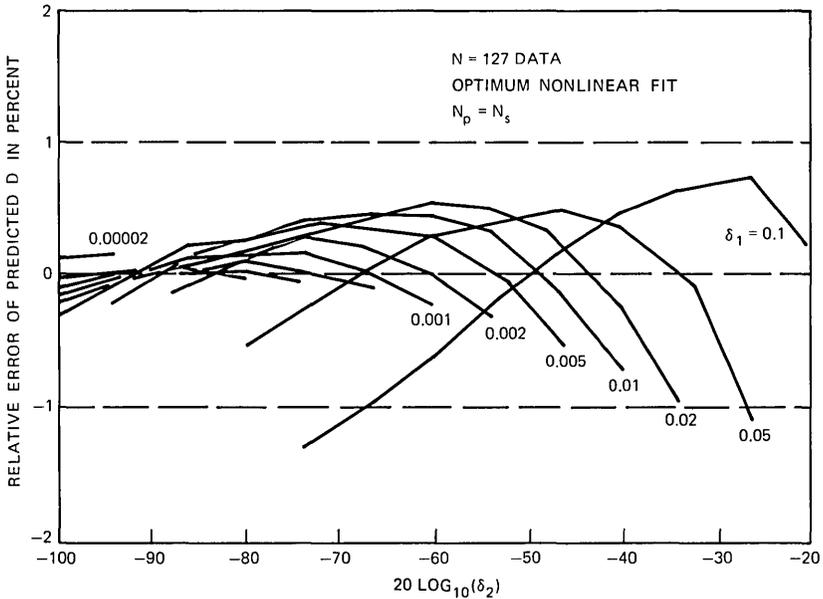
Fig. 10—The relative errors in fitting the data of Fig. 8 with a nonlinear curve over the range $\delta_1 \leqq 0.1$, for various values of δ_1 ($N = 127$).

much smaller. Clearly, this prediction formula is acceptable for almost any design application.

Figure 11 shows a summary of the predicted values of D as a function of stopband attenuation for a wide range of values of passband ripple. In this case, passband ripple in dB is defined as

$$\text{Passband ripple} = 20 \log_{10} \left(\frac{1}{1 - \delta_p} \right), \qquad (33)$$

$$\text{Stopband attenuation} = -20 \log_{10} (\delta_s), \qquad (34)$$

where

$$\delta_p = \frac{2\delta_1}{1 + \delta_1} \qquad (35)$$

and

$$\delta_s = \frac{\delta_2}{1 + \delta_1}. \qquad (36)$$

These data correspond to standard design data for continuous-time filters where the frequency response magnitude is constrained to be less than or equal to 1.0.
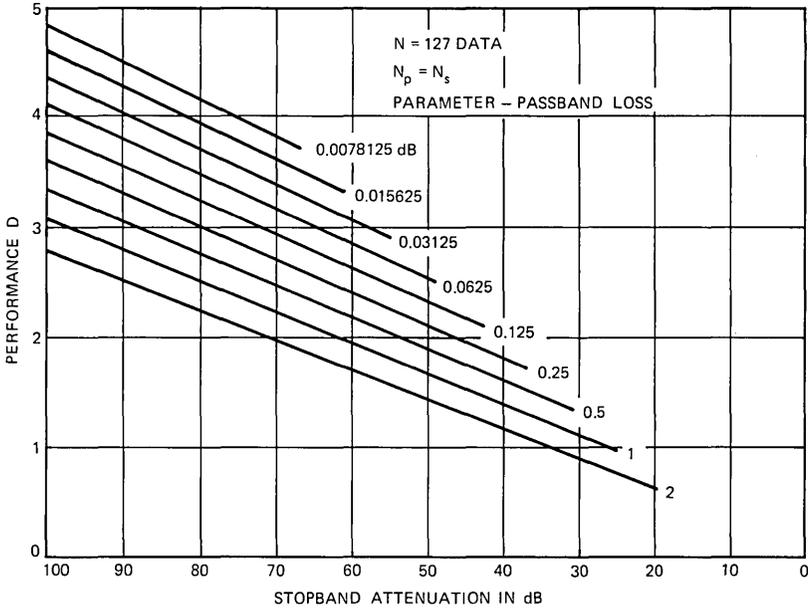
Fig. 11—Plots of D versus $20 \log_{10} \delta_2$ for various values of the parameter $20 \log_{10} (1/1 - \delta_p)$, where $\delta_p = 2\delta_1/(1 + \delta_1)$, as calculated from the optimum non-linear fits to the data of Fig. 8 ($N = 127$).

V. CORRECTIONS FOR SMALL VALUES OF N

The formulas in the previous section are accurate for predicting D (or, equivalently, N) for values of N greater than about 51. As seen from the curves in Fig. 7, as N decreases, D decreases for fixed values of δ_2 and K. It is also seen from Fig. 7 that the differences increase with decreasing δ_2 or, equivalently, increasing transition width. An examination of the relative deviation of D_{NL}, the predicted value of D, from its true value as a function of transition bandwidth showed that, independent of N, the deviations could be simply approximated by a curve of the form

$$D_{\text{cor}} = f(K)(\Delta F)^2, \tag{37}$$

where D_{cor} is the correction term and $f(K)$ is of the form

$$f(K) = (0.51244 \log_{10} K + 11.01217). \tag{38}$$

(The constants in eq. (38) were again obtained by a minimum mean-square relative error data-fitting procedure.) Thus, using eqs. (37) and (38), a formula for D which depends on N, K, and δ_2 can be obtained.

Adopting the notation

   (*i*) $D_\infty(\delta_2, K) = D_{NL}$ of eq. (32)*

   (*ii*) $\hat{D}(\delta_2, K, N)$ = predicted value of D as a function of N as well as δ_2 and K

   (*iii*) $D(\delta_2, K, N)$ = true value of D,

we obtain the relation

$$D_\infty(\delta_2, K) - \hat{D}(\delta_2, K, N) = f(K)(\Delta F)^2. \tag{39}$$

Thus, in a design case where δ_2, K (or δ_1), F_p, and F_s (i.e., ΔF) are specified and the problem is to estimate N, the impulse response duration required to meet these specifications, eq. (39) can be used directly since

$$\hat{D}(\delta_2, K, N) = (N - 1)\Delta F. \tag{40}$$

Thus, combining eqs. (37) and (38) and solving for N gives

$$N = \frac{D_\infty(\delta_2, K) - f(K)(\Delta F)^2}{\Delta F} + 1. \tag{41}$$

A closed-form expression for $\hat{D}(\delta_2, K, N)$ may be obtained by substituting eq. (40) for $\hat{D}(\delta_2, K, N)$ in eq. (39) and solving the quadratic equation for ΔF. Equation (40) is then used to give $\hat{D}(\delta_2, K, N)$. Thus, we get

$$f(K)(\Delta F)^2 + (N - 1)\Delta F - D_\infty(\delta_2, K) = 0 \tag{42}$$

$$\Delta F = \frac{(N - 1)}{2f(K)} \left(\sqrt{1 + \frac{4f(K)D_\infty(\delta_2, K)}{(N - 1)^2}} - 1 \right) \tag{43}$$

$$\hat{D}(\delta_2, K, N) = \frac{(N - 1)^2}{2f(K)} \left(\sqrt{1 + \frac{4f(K)D_\infty(\delta_2, K)}{(N - 1)^2}} - 1 \right). \tag{44}$$

In the limit, as N tends to infinity, eq. (44) shows $\hat{D}(\delta_2, K, N)$ tends to $D_\infty(\delta_2, K)$ as expected.

Using eq. (44), the relative error of the predicted value of D from the true value was measured for the data for $N = 3, 7, 11, 19, 51,$ and 127 with $\delta_1 \leqq 0.1$. The relative errors for values of K from 1 to 500 are plotted in Figs. 12a through 12f. In all cases, the worst relative error in D is sufficiently small that the equivalent error in N is less than *one sample*. Thus, for all practical purposes, the design equations above serve as a useful guide for estimating the order of the filter required to meet design specifications.

---

* Eq. (32) gives D_{NL} as a function of δ_1 and δ_2 but since $\delta_1 = K\delta_2$, it is also implicitly a function of δ_2 and K.

VI. SUMMARY OF DESIGN PROCEDURES AND EXAMPLES

Based on the results presented in this paper, it is now possible to give a set of rules for estimating the filter impulse response duration required to meet given design specifications. These rules are as follows:

(i) Check if either δ_1 or δ_2 is greater than 0.1, in which case the graphical data of Fig. 7 are used directly to estimate N.

(ii) Calculate a value of N, call this N_1, corresponding to the extra-ripple case where $N_p = N_s$, from the equation

$$N_1 = \frac{D_\infty(\delta_1, \delta_2) - f(K)(\Delta F)^2}{\Delta F} + 1,$$

where $D_\infty(\delta_1, \delta_2)$ is the optimum nonlinear fit to the data for large N and is given by:

$$D_\infty(\delta_1, \delta_2) = [5.309 \times 10^{-3} [\log_{10} \delta_1]^2 + 7.114 \times 10^{-2} \log_{10} \delta_1$$
$$- 0.4761] \log_{10} \delta_2 - [2.66 \times 10^{-3} (\log_{10} \delta_1)^2$$
$$+ 0.5941 \log_{10} \delta_1 + 0.4278]$$

and $f(K)$ is given by

$$f(K) = 0.51244 \log_{10}(K) + 11.01217.$$

(iii) If the desired value of F_p is less than or equal to 0.04 (let us call this case 1), or if the desired value of F_s is greater than or equal to 0.46 (case 2), then the estimate of N is obtained in rule (iv). Otherwise, the value N_1 of rule (ii) is used as the estimate of N.

(iv) To obtain the value of N_c for the Chebyshev solution for case 1, eq. (26) is used to get a first approximation to N_c. N_c is then systematically varied until a Chebyshev solution is obtained which meets specifications on δ_1, δ_2, and ΔF. (As discussed earlier, it is not generally possible to find a Chebyshev solution which meets specifications exactly on all four filter parameters.) For case 2, δ_1 and δ_2 are interchanged, and F_s is replaced by $0.5 - F_p$, in order to solve for N_c from eq. (26). In cases where this rule is applied, the value N_c obtained is a lower bound to the true value of N_1. No upper bound may be given in this case. A discussion of this problem is given below.

Several comments are necessary about these rules before proceeding to some examples. The discussion in this paper has concentrated on two regions of the curve of ΔF versus F_p—the Chebyshev solutions and the case of extraripple filters with an equal number of passband and stop-
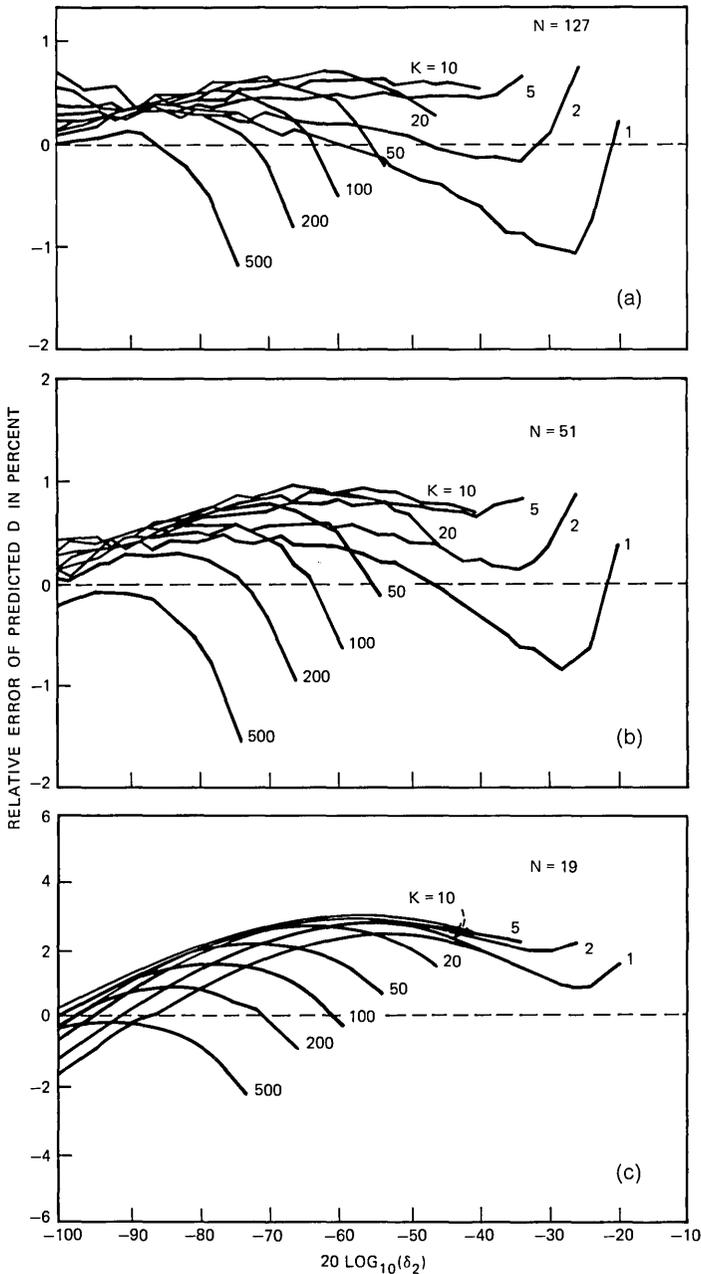
Fig. 12—Plots of the relative errors in fitting the data of Fig. 7 using the corrected values of D for $N = 3, 7, 11, 19, 51,$ and 127.
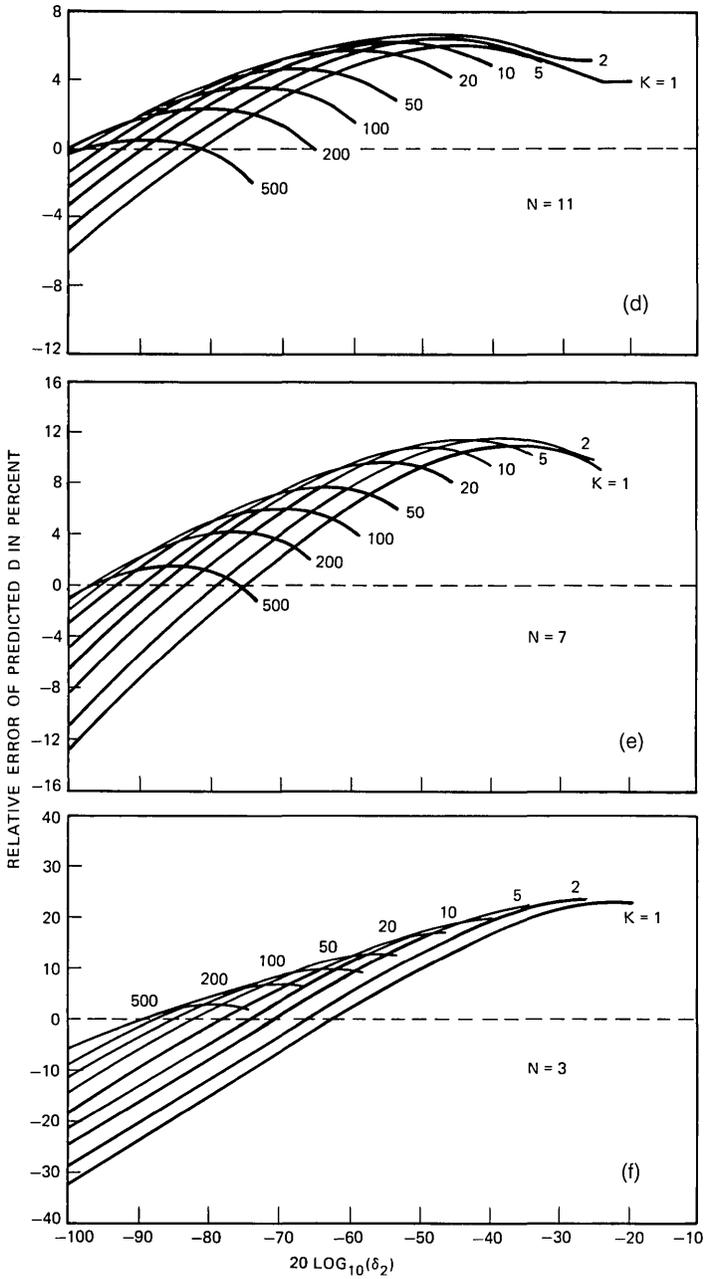
Fig. 12 (continued).

band ripples. The justification for such emphasis was the results shown in Figs. 2 and 3 which indicated that the extraripple solution for $N_p = N_s$ was fairly representative of a large region of the curve of ΔF versus F_p, except in the case where F_p was either very small or very large, in which case the Chebyshev solution became important. Also, as seen from Fig. 2, between extraripple solutions, the curve of ΔF versus F_p peaks up. However, it is seen that in many cases one can "approximately" bound the maximum between extraripple solutions by the next lower-degree extraripple solution. Since the design equations can predict ΔF for this case (the next lower-degree extraripple solution), a good bound on N can be obtained for a reasonably large region of the curve of ΔF versus F_p. In these cases, the value given by rule (ii) is good to within ± 4 in the worst cases, i.e., large values of K, and generally to within ± 2.
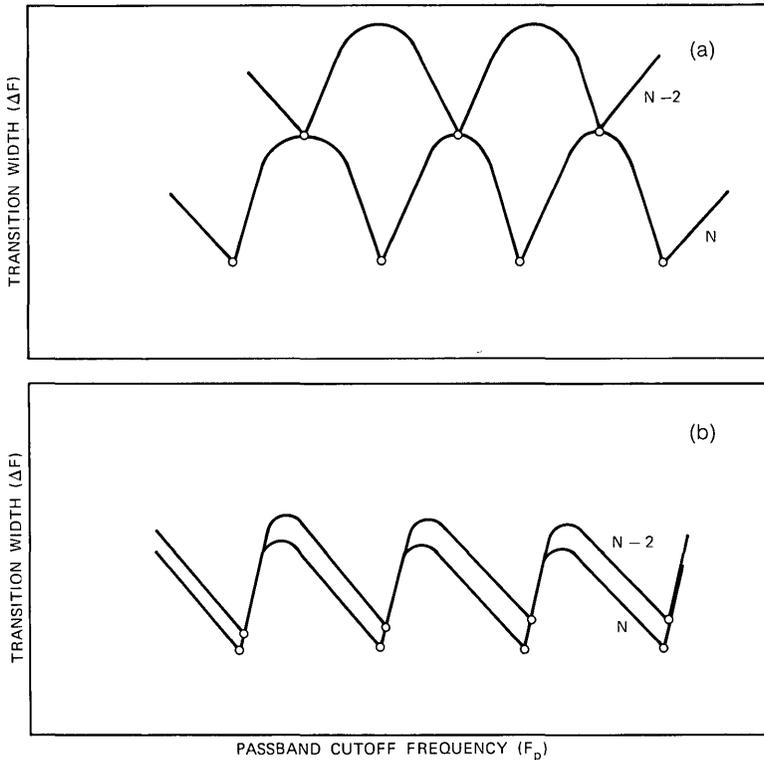
Fig. 13—Explanation of the types of behavior of curves of transition width versus passband cutoff frequency.

In the region of the Chebyshev solutions, however, such bounding procedures no longer are valid. An explanation of the difficulties which may be encountered is given in Fig. 13, which shows two types of curves of ΔF versus F_p. Fig. 13a shows the case, discussed above, where the extraripple solutions of impulse response duration $N - 2$ are approximately midway between extraripple solutions of impulse response duration N. In this case, the next degree solution bounds the maximum ΔF between extraripple solutions. Figure 13b shows the case where the extraripple solutions of impulse response duration $N - 2$ have approximately the same values for ΔF and F_p as the extraripple solutions of impulse response duration N. In these cases, there is no good bound on the maximum value of ΔF between adjacent extraripple solutions. The case of Fig. 13b corresponds to regions of F_p near 0.0 and F_s near 0.5, i.e., in the regions of the Chebyshev solutions. In these cases, as discussed in rule (*iv*), there is only an underbound on N, and no overbound.

The choice of a value of 0.04 in rule (*iii*) as the width of the region during which the behavior of N can only be underbounded was obtained from the data of Fig. 3 which shows that, beyond this region, the variation in the values of ΔF for extraripple solutions is small.

Figures 14 through 16 illustrate typical behavior of the curve of minimum value of impulse response duration N, to meet given specification on δ_1, δ_2, and ΔF as a function of F_p. Figure 14 shows data for the case $\delta_1 = 0.01$, $\delta_2 = 0.0001$, $\Delta F = 0.158$. The value of N_1 from
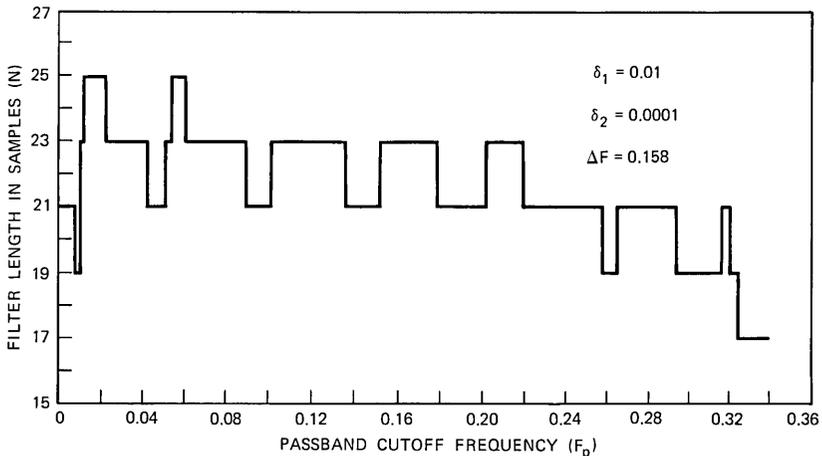
Fig. 14—Optimum values of N to meet given design specifications on δ_1, δ_2, and ΔF, as a function of F_p.
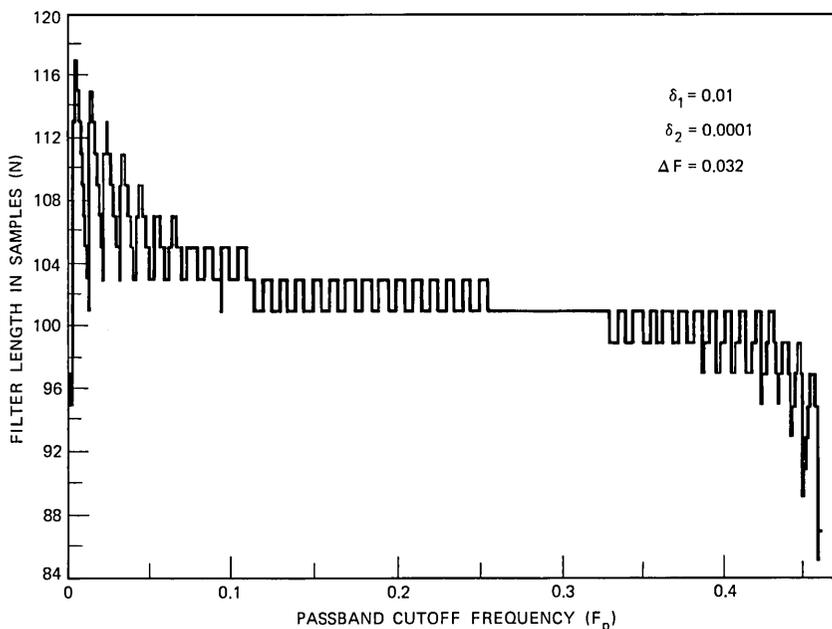
Fig. 15—Optimum values of N to meet given design specifications on δ_1, δ_2, and ΔF, as a function of F_p.

rule (ii) is 21, and the values of N_c are 19 (for small F_p) and 13 (for large F_p). Over the region $0.06 \leqq F_p \leqq 0.32$, N is within 2 of the nominal value of $N_1 = 21$. Over the entire range of F_p, N is within 4 of the value $N_1 = 21$. The data of Fig. 14 correspond to the case of Fig. 13a.

Figure 15 shows data for the case $\delta_1 = 0.01$, $\delta_2 = 0.0001$, $\Delta F = 0.032$. The value of N_1 from rule (ii) is 101, and the values of N_c are 95 (for small F_p) and 55 (for large F_p). Over the region $0.1 \leqq F_p \leqq 0.38$, N is within 2 of the nominal value of $N_1 = 101$. However, in the region $0 \leqq F_p \leqq 0.036$, the value of N fluctuates from a minimum of 95 (the Chebyshev lower bound) to a maximum of 117. The explanation of this erratic behavior of N is seen in Fig. 16. Figure 16a shows the data of Fig. 15 on an expanded horizontal scale, and Fig. 16b shows a plot of the approximate curves of ΔF versus F_p for all values of N from 95 to 117. The heavily traced parts of these curves show the lowest-order solution which just meets specifications on ΔF. From Fig. 16b, it is clear that in the vicinity of the first few extraripple solutions, the curves of ΔF versus F_p are exceedingly steep. Hence, a slight change in F_p greatly increases the required order solution to meet specifications. In
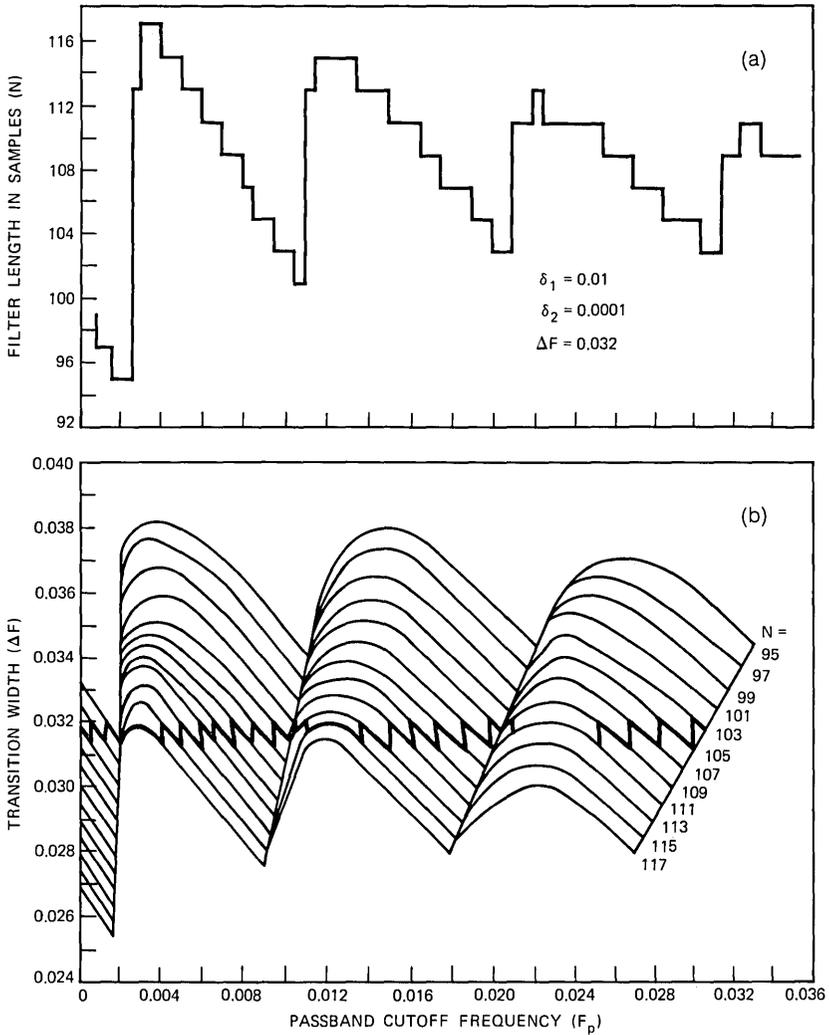
Fig. 16—Explanation of the behavior of the data of Fig. 15 in the region $0 \leqq F_p \leqq 0.036$.

these cases, it is impossible to estimate the exact filter impulse response duration which is required. Instead, only a lower bound can be given. Fortunately, as seen in Fig. 15, the regions in which this erratic behavior can occur are limited.

We conclude this section with a set of examples which illustrate the use of the design rules.

Example 1: Find the minimum value of N required to meet the specifications $\delta_1 = 0.05$, $\delta_2 = 0.0001$, $F_p = 0.19$, $F_s = 0.21$.

From rule *(ii)* we get $N_1 = 129.7$ which is rounded to 129. (Herein, all values of N will be rounded to the nearest odd integer.) Since F_p is far from the range for the Chebyshev solutions, the value of 129 is used as the appropriate estimate. The actual filter impulse response duration required is $N = 131$, although the $N = 129$ filter just missed meeting specifications.

Example 2: Find the minimum value of N required to meet the specifications $\delta_1 = 0.01$, $\delta_2 = 0.0001$, $F_p = 0.213$, $F_s = 0.373$.

From rule *(ii)* we get $N_1 = 19$. Since F_p is again out of the range of the Chebyshev solutions, the value 19 is used as the estimate of N. The actual filter impulse response duration required is $N = 19$.

Example 3: Find the minimum value of N required to meet the specifications $\delta_1 = 0.1$, $\delta_2 = 0.1$, $F_p = 0.12$, $F_s = 0.19$.

From rule *(ii)* we get $N_1 = 11$. Since F_p is out of the range of the Chebyshev solutions, the value 11 is used as the estimate of N. The actual value of N is 11. In this case, it is interesting to note that the value of N_c for the Chebyshev solution is also 11. This example points out that, for filter specifications leading to small values of N, there is very little variation in the actual value of N as F_p varies. This observation has been made earlier with respect to Fig. 14.

Example 4: Find the minimum value of N required to meet the specifications $\delta_1 = 0.01$, $\delta_2 = 0.0001$, $F_p = 0.36$, $F_s = 0.497$.

From rule *(ii)* we get $N_1 = 23$. Since F_s is within the bounds of rule *(iii)* (case 2), N_c is computed from rule *(iv)* as 13. The actual value of N is 19. In this case, the lower bound is within three filter orders of the true solution.

VII. SUMMARY

This paper has presented a wide variety of data on the relationships between design parameters for optimal low-pass FIR linear-phase digital filters. Analytical formulas were derived for the Chebyshev solutions, i.e., when there was only one passband or stopband ripple. Approximate fits to the data using nonlinear relationships between ΔF, N, δ_1, and δ_2 were given in the case where the number of passband and stopband ripples was equal, and it was argued that these relationships

were valid over a wide range of values of the filter parameters. Finally, a simple set of rules for estimating the minimum value of N, which meets given specifications on F_p, F_s, δ_1, and δ_2, was discussed. Examples were given to illustrate the application of the rules.

VIII. ACKNOWLEDGMENT

The authors would like to acknowledge the most helpful comments and criticisms provided by James Kaiser of Bell Laboratories.

REFERENCES

1. Rabiner, L. R., "Techniques for Designing Finite-Duration Impulse-Response Digital Filters," IEEE Trans. Commun. Tech., *COM-19* (April 1971), pp. 188–195.
2. Herrmann, O., "On the Design of Nonrecursive Digital Filters with Linear Phase," Elec. Letters, *6*, No. 11, 1970, pp. 328–329.
3. Hofstetter, E., Oppenheim, A. V., and Siegel, J., "A New Technique for the Design of Nonrecursive Digital Filters," Proc. Fifth Annual Princeton Conf. Inform. Sci. Syst., 1971, pp. 64–72.
4. Parks, T. W., and McClellan, J. H., "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," IEEE Trans. Circuit Theory, *DT-19* (March 1972), pp. 89–194.
5. Rabiner, L. R., "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1177–1198.
6. Parks, T. W., Rabiner, L. R., and McClellan, J. H., "On the Transition Width of Finite Impulse Response Digital Filters," IEEE Trans. Audio and Electroacoustics, *21*, No. 1 (February 1973), pp. 1–4.
7. Rabiner, L. R., and Herrmann, O., "The Predictability of Certain Optimum Finite Impulse Response Digital Filters," IEEE Trans. on Circuit Theory, *20*, No. 4 (July 1973), pp. 401–408.
8. Helms, H. D., "Nonrecursive Digital Filters: Design Methods for Achieving Specifications on Frequency Response," IEEE Trans. Audio and Electroacoustics, *16*, No. 3 (September 1968), pp. 336–342.
9. Heyliger, G. E., "Simple Design Parameters for Chebyshev Arrays and Filters," IEEE Trans. Audio and Electroacoustics, *18*, No. 4 (December 1970), pp. 502–503.
10. Kaiser, J., "Digital Filters," Chapter 7 in *System Analysis by Digital Computer*, edited by F. F. Kuo and J. F. Kaiser, New York: J. Wiley & Sons, 1966, pp. 230–238.

Impulse Response of Clad Optical Multimode Fibers

By D. GLOGE

Loss, coupling, and delay differences among the modes of multimode fibers influence their response to intensity-modulated optical signals. This "baseband" response is derived here from a time-dependent continuous description of the power flow in the fiber. Particular attention is given to the output as a function of angle and to the impulse response, its width and symmetry. We find that coupling narrows the impulse response but, at the same time, causes additional loss. Under practical conditions, this loss may limit the usefulness of coupling for the purpose of reducing the mode dispersion. We calculate a possible data rate of 12 Mb/s for a 10-km repeater spacing and an effective numerical aperture of 0.1, but we show that further improvements can be gained from an optimization of the coupling characteristic and of other parameters.

I. INTRODUCTION

Although single-mode operation of clad optical fibers is possible and, in general, offers very good transmission characteristics, multimode fibers have two advantages: They impose less stringent requirements on the optical carrier (they transmit even the incoherent light from a luminescent diode) and their larger dimensions alleviate splicing problems or at least relax the tolerances required for connection. Typically, the core diameter is of the order of a hundred wavelengths and the fiber therefore transmits thousands of modes, even if the index difference between core and cladding is only a few percent (corresponding to a numerical aperture of 0.2 to 0.3).

The usefulness of such fibers depends on their dispersion characteristics. Delay differences among the many modes[1] distort the signal and certainly produce a signal response inferior to that of the single-mode fiber. For certain systems, on the other hand, overall system economy may place the desirable information rate of individual fiber channels

in a range where the characteristic signal response of multimode fibers is adequate (below 100 megabits per second, say). In such systems, multimode operation would surpass single-mode operation because of the advantages mentioned earlier.

Effective use of the multimode fiber would presuppose the excitation of a large number of modes right at the input with the objective of transmitting *all* of these modes to the receiver. Experiments that have approximated these conditions have revealed a rather intricate response to short input pulses both in liquid- and solid-core multimode fibers.[2,3] For example, the width of the output pulse did not increase linearly with fiber length, but showed a less-than-proportional increase for long fibers. Coupling among the modes and a dependence of loss on mode number seemed to play a part.[4] In some fibers, this resulted in an optimal mode distribution (causing lowest overall loss) which comprised only a fraction of the modes capable of propagating. Measurements of the coupling strength showed that a total exchange of power between two modes was likely to occur within less than a meter of fiber.[5] This result made it clear that a perturbation theory depending on small coupling rates was not applicable. A closed and unrestricted description was achieved by assuming a modal continuum rather than thousands of individual modes. In this theory, mode coupling took the form of a diffusion process not limited to small coupling amplitudes.

The work discussed here extends the approach outlined in Ref. 5 by taking the velocity differences among the modes into account. We first consider fibers in which the optimal (steady-state) mode distribution does not include modes close to cutoff. The power in the fiber is calculated as a function of time and output angle (mode number) for the case of a short input pulse. Particular attention is given to the "fiber impulse response" obtained by integrating over all angles at the output. A simple formula relates the output pulse width to the fiber length and to the attenuation and coupling parameters. The latter can be measured in short samples (a few meters in length) permitting the immediate computation of the pulse broadening in a long fiber and, hence, of the obtainable data rate for a given fiber length.

The results may also shed some light on the prospects of mode coupling introduced artificially as a means of equalization: It has been predicted that, under certain circumstances, increased mode coupling reduces the signal distortion (ultimately forcing all energy to propagate at an average velocity).[6,7] The objective of this paper is to outline an analytic approach which can answer these and other questions.

II. TIME-DEPENDENT POWER FLOW EQUATION

The differential equation obtained in Ref. 5 for the power flow in multimode fibers was originally derived from the mode characteristics by assuming certain statistics for the modal field coupling. By making some approximations acceptable for high-order modes, a description results which admits a simple ray-optics interpretation. Each high-order mode can be represented by a characteristic ray propagating inside the core along a meridional zigzag path. Internal reflection guides the rays at the core-cladding interface and limits the range of angles which can be formed with the guide axis. If n and n_c are the indices of core and cladding, respectively, and

$$\Delta = 1 - \frac{n_c}{n} \tag{1}$$

is a small difference, the maximum angle is given by the condition of critical internal reflection which is approximately

$$\theta_{\max} = \sqrt{2\Delta}. \tag{2}$$

The rays form a uniform distribution within the cone of apex angle $\theta_{\max}$.

If the core cross section permits many modes to propagate, the rays are so densely spaced that their distribution can be considered as continuous. The state of the fiber at a point z and at time t can then be described by a distribution $P(\theta, z, t)$ where θ is a continuous variable.* Reference 5 expresses the incremental change dP in the power P as a sum of two terms:

(i) A loss $-A\theta^2 P dz$; this term comprises attenuation effects in the cladding and the core-cladding interface and increases as the square of the characteristic angle θ. The coefficient A is measured in $m^{-1} rad^{-2}$. θ-independent loss is omitted, but can easily be incorporated later in the final solution.

(ii) Mode coupling; in practical multimode fibers, coupling was found to occur essentially only between closely adjacent modes and, for this reason, takes the form of a diffusion process in the ray picture. The incremental increase in $P(\theta, z)$ as a result of diffusion is $(1/\theta)(\partial/\partial\theta)(\theta D \partial P/\partial\theta)dz$, a term typical for radial diffusion in cylindrical configurations. D is a coupling coefficient

---

* θ is related to the transverse wave number u of the corresponding mode by $u = nk\theta$ where k is the vacuum wave number.

which, for most of the following discussion, is assumed to be independent of θ.

The total variation in P thus becomes

$$dP = -A\theta^2 P dz + \frac{1}{\theta} \frac{\partial}{\partial \theta} \left(\theta D \frac{\partial P}{\partial \theta} \right) dz. \tag{3}$$

If P is a function of time t, we can also write

$$dP = \frac{\partial P}{\partial z} dz + \frac{\partial P}{\partial t} dt. \tag{4}$$

Equating (3) and (4) and dividing by dz results in the equation

$$\frac{\partial P}{\partial z} + \frac{dt}{dz} \frac{\partial P}{\partial t} = -A\theta^2 P + \frac{1}{\theta} \frac{\partial}{\partial \theta} \left(\theta D \frac{\partial P}{\partial \theta} \right). \tag{5}$$

The derivative dz/dt is the velocity of the power $P(\theta)$ or, equivalently, the group velocity of a mode with characteristic angle θ. By using the relation between θ and the transverse wave number u, we can calculate this velocity from eq. (25) of Ref. 8. Except for the few modes close to cutoff, we obtain the simple relation

$$\frac{dz}{dt} = \frac{c}{n(1 + \theta^2/2)}. \tag{6}$$

It relates the mode velocity to the vacuum light velocity, c, reduced by n because of the material retardation and by a factor $1 + \theta^2/2$ which accounts for the increased path length as a result of the zigzag propagation. The derivative dt/dz required in (5) is the inverse of (6) and has the meaning of a delay per unit length. If we ignore the delay n/c common to all modes (it can be added later if necessary), we obtain from (5) and (6)

$$\frac{\partial P}{\partial z} = -A\theta^2 P - \frac{n}{2c} \theta^2 \frac{\partial P}{\partial t} + \frac{1}{\theta} \frac{\partial}{\partial \theta} \left(\theta D \frac{\partial P}{\partial \theta} \right). \tag{7}$$

With the help of the Laplace transformation

$$p(\theta, z, s) = \int_0^\infty e^{-st} P(\theta, z, t) dt, \tag{8}$$

we can write (7) in the form

$$\frac{\partial p}{\partial z} = -A\sigma^2\theta^2 p + \frac{1}{\theta} \frac{\partial}{\partial \theta} \left(\theta D \frac{\partial p}{\partial \theta} \right) \tag{9}$$

where

$$\sigma = (1 + ns/2cA)^{\frac{1}{2}}. \tag{10}$$

Except for the factor σ^2, (9) agrees with (22) of Ref. 5; we can therefore use the solution derived there if we replace A by $A\sigma^2$. For the Gaussian input distribution

$$p_{in} = f(0, s) \exp(-\theta^2/\Theta_o^2), \tag{11}$$

we obtain

$$p(\theta, z, s) = f(z, s) \exp[-\theta^2/\Theta^2(z, s)] \tag{12}$$

where

$$\Theta^2(z, s) = \frac{\Theta_\infty^2}{\sigma} \frac{\sigma\Theta_o^2 + \Theta_\infty^2 \tanh \sigma\gamma_\infty z}{\Theta_\infty^2 + \sigma\Theta_o^2 \tanh \sigma\gamma_\infty z} \tag{13}$$

and

$$f(z, s) = \frac{f(0, s)\sigma\Theta_o^2}{\Theta_\infty^2 \sinh \sigma\gamma_\infty z + \sigma\Theta_o^2 \cosh \sigma\gamma_\infty z} \tag{14}$$

with

$$\Theta_\infty = (4D/A)^{\frac{1}{4}} \tag{15}$$

and

$$\gamma_\infty = (4DA)^{\frac{1}{2}}. \tag{16}$$

For cw excitation ($s = 0$), the angular width $\Theta(z, 0)$ changes monotonically from Θ_o to Θ_∞ as z increases. The width Θ_∞ characterizes a distribution which propagates unchanged (at steady state) and with the minimum overall loss coefficient γ_∞. It seems practical to excite this distribution right from the beginning. The condition $\Theta_o = \Theta_\infty$ will therefore receive particular attention in the following. The solutions (12) through (16) assume that Θ_o and Θ_∞ are so small compared to $\theta_{max} = \sqrt{2\Delta}$ that practically no light propagates at angles close to the critical one. In other words, modes close to cutoff do not take part in the transmission process. Experiments have shown that the steady state in certain liquid-core fibers ($C_2C\ell_4$ in quartz, for example) is of that type.

Closed-form Laplace transformations of (12) exist only for the approximations given in the limits $z \ll 1/\gamma_\infty$ and $z \gg 1/\gamma_\infty$ and these two cases are discussed in Section III. Certain important characteristics of $P(\theta, z, t)$, however, can be derived for all z, as we shall see in Section IV.

III. CLOSED-FORM SOLUTIONS FOR THE IMPULSE RESPONSE

In a practical communication system, the multimode fiber is likely to be fed with a pulse $F(0, t)$ whose width is typically of the same order as the broadening expected in the fiber. Its Laplace transform $f(0, s)$, appearing in (12) and (14), and its dependence on s therefore cannot be ignored. We assume, however, that the input is simultaneous in all modes being excited, in which case Θ_o is independent of s. Sacrificing

generality for clarity, we restrict this discussion to the practical input condition $\Theta_o = \Theta_\infty$; the general case can be treated in exactly the same way, but leads to more complicated results.

In the case of a short fiber, we replace $\sinh \sigma\gamma_\infty z$ and $\tanh \sigma\gamma_\infty z$ in (13) and (14) by the argument $\sigma\gamma_\infty z$ and set $\cosh \sigma\gamma_\infty z = 1$. With the help of (10), (15), and (16), (12) then becomes

$$p(\theta, z, s) = \frac{f(0, s)}{1 + \gamma_\infty z} \exp\left[-\theta^2 \left(\frac{1}{\Theta_o^2} + \frac{nz}{2c} s \right) \right] \qquad (17)$$

which has the Laplace transform

$$P(\theta, z, t) = \frac{\exp\left(-\theta^2/\Theta_o^2\right)}{1 + \gamma_\infty z} F(0, t - n\theta^2 z/2c). \qquad (18)$$

The denominator $1 + \gamma_\infty z$ expresses the loss in the short distance z; $\exp\left(-\theta^2/\Theta_o^2\right)$ indicates that the input condition has been conserved, and $F(0, t - n\theta^2 z/2c)$ shows that the portion of the input pulse $F(0, t)$ which propagated at an angle θ was delayed by $n\theta^2 z/2c$. Clearly, coupling has not affected the propagation at this distance.

The total output is obtained from the integration[5]

$$q(z, s) = 2\pi \int_0^\infty p(\theta, z, s)\theta d\theta. \qquad (19)$$

For $z \ll 1/\gamma_\infty$, we obtain with (17)

$$q = \frac{\pi f(0, s) \Theta_o^2}{(1 + \gamma_\infty z)(1 + n \Theta_o^2 zs/2c)}. \qquad (20)$$

If we now set $f(0, s) = 1$, which corresponds to an infinitesimally short input pulse of energy 1, the Laplace transformation of (20) yields the impulse response of the fiber:

$$Q(z, t) = \frac{2c\pi}{nz(1 + \gamma_\infty z)} \exp\left(-2ct/n\Theta_o^2 z\right). \qquad (21)$$

The assumption of a mode continuum has the consequence that the impulse response is a continuous and well-behaved function, in spite of the somewhat artificial condition of an infinitesimally narrow input pulse. That Q extends mathematically to infinity results from the assumption of the unbounded distributions (11) and (12). Remember that this assumption was acceptable since $\Theta_o = \Theta_\infty < \theta_{\max}$. The same condition limits $Q(t)$ practically to a time interval narrower than $n\theta_{\max}^2 z/2c$, which is the delay between the fastest and the slowest mode. Since (18) and (21) neglect mode coupling, they could have been ob-
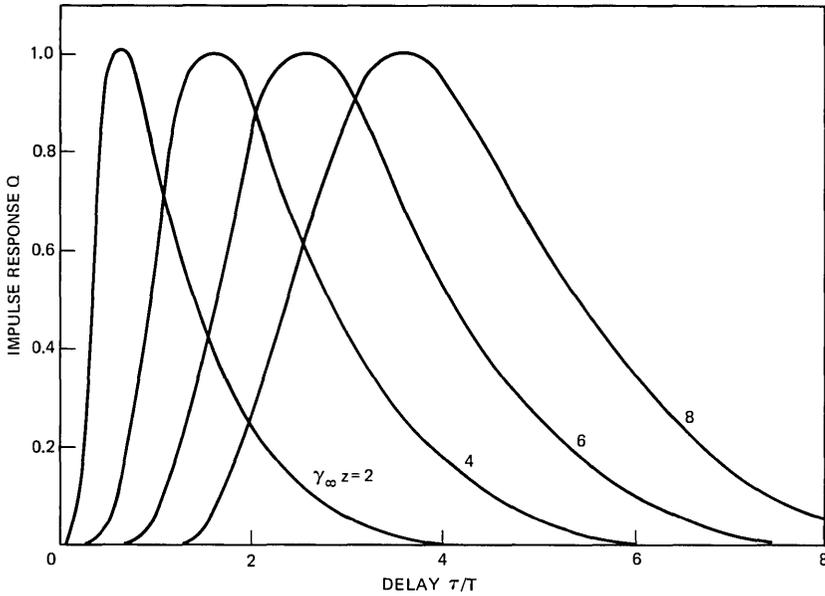
Fig. 1—Impulse response according to (24) normalized for equal peak values and plotted versus normalized time for different fiber lengths.

tained without the help of the power flow equation. They were derived here merely for a better understanding of the physical implications involved.

In the case of a very long fiber, we may use the approximation $\tanh \sigma\gamma_\infty z = 1$ and $\sinh \sigma\gamma_\infty z = \cosh \sigma\gamma_\infty z = \frac{1}{2} \exp \sigma\gamma_\infty z$ in (13) and (14). Equation (12) thus assumes the form

$$p = \frac{2\sigma}{1+\sigma} \exp\left[-\sigma(\theta^2/\Theta_o^2 + \gamma_\infty z)\right] \qquad (22)$$

which leads to

$$q = \frac{2\pi\Theta_o^2}{1+\sigma} \exp\left(-\sigma\gamma_\infty z\right) \qquad (23)$$

where p is integrated over all angles θ with the help of (19). After introducing (10) for σ into (23) we can form the Laplace transform of $q(s)$. By using the condition $\gamma_\infty z > 1$, we arrive at

$$Q(z, t) = \Theta_o^2 \sqrt{\frac{\pi}{Tt}} \left(\frac{t}{\gamma_\infty z T} + \frac{1}{2}\right)^{-1} \exp\left(-\frac{\gamma_\infty^2 z^2 T}{4t} - \frac{t}{T}\right) \qquad (24)$$

where

$$T = \frac{n}{2cA} = \frac{n}{2c}\frac{\Theta_o^2}{\gamma_\infty}. \qquad (25)$$

An evaluation of (24) is shown in Fig. 1 for various normalized lengths $\gamma_\infty z$. The plotted impulse responses are normalized for equal peak values.

The normalizing distance $1/\gamma_\infty$ is the distance within which a 1-neper loss is incurred as a result of the θ-dependent loss characteristic. Note that additional θ-independent loss can be present. As indicated by (21), T is the $1/e$-width, which the impulse response Q would assume at the distance $1/\gamma_\infty$ if no coupling were present. Closed-form solutions for Q are available only in the two cases discussed here, but another very practical characterization of the fiber output distribution can be obtained without performing the Laplace transformation.

IV. PULSE DELAY AND PULSE WIDTH

Because of a general relation between $P(t)$ and its Laplace transform $p(s)$, we obtain the mth moment of $P(t)$ from

$$(-1)^m \frac{\partial^m p}{\partial s^m}\bigg|_{s=0} = \int_0^\infty t^m P dt. \tag{26}$$

To achieve a suitable normalization we set $m = 0$ which yields

$$p(s = 0) = \int_0^\infty P dt \tag{27}$$

and divide (26) through (27). This leads to

$$(-1)^m \frac{\partial^m \ell n p}{\partial s^m}\bigg|_{s=0} = \frac{\int_0^\infty (t - \delta)^m P dt}{\int_0^\infty P dt}, \tag{28}$$

where

$$\delta = \frac{\int_0^\infty (-t) P dt}{\int_0^\infty P dt} = -\frac{\partial \ell n p}{\partial s}\bigg|_{s=0}. \tag{29}$$

The second derivative represents the variance of $P(t)$ and, hence, a measure of the width of $P(t)$:

$$\tau^2 = \frac{\int_0^\infty (t - \delta)^2 P dt}{\int_0^\infty P dt} = \frac{\partial^2 \ell n p}{\partial s^2}\bigg|_{s=0}. \tag{30}$$

The third derivative

$$\eta^3 = \frac{\int_0^\infty (t - \delta)^3 P dt}{\int_0^\infty P dt} = \frac{\partial^3 \ln p}{\partial s^3}\bigg|_{s=0} \tag{31}$$

is generally called the "skewness" of the distribution $P(t)$. The ratio η/τ is a measure of the asymmetry of P and, as we shall see later, permits an immediate estimate of the value of $\gamma_\infty z$ without the knowledge of any other fiber parameters.

The fact that the Laplace term $ns/2c$ appears in (9) as part of the sum $A\sigma^2 = A + ns/2c$ permits us to use the relation

$$\frac{\partial^m p(\theta, z, s)}{\partial s^m}\bigg|_{s=0} = \left(\frac{n}{2c}\right)^m \frac{\partial^m p(\theta, z, 0)}{\partial A^m}, \tag{32}$$

which will greatly simplify the following calculations.

Let us now apply (30) and (31) to the general solution (12) assuming again the special but practical condition $\Theta_o = \Theta_\infty$. We obtain

$$\delta_P(\theta) = \frac{T}{2}\left[\gamma_\infty z + \left(\frac{\theta^2}{\Theta_o^2} - \frac{1}{2}\right)(1 - e^{-2\gamma_\infty z})\right] \tag{33}$$

for the mean delay (in addition to the overall delay nz/c) and

$$\tau_P(\theta) = \frac{T}{2}\left[\gamma_\infty z + \left(\frac{\theta^2}{\Theta_o^2} - \frac{5}{4}\right) - 2\gamma_\infty z\left(2\frac{\theta^2}{\Theta_o^2} - 1\right)e^{-2\gamma_\infty z}\right.$$
$$\left. + e^{-2\gamma_\infty z} - \left(\frac{\theta^2}{\Theta_o^2} - \frac{1}{4}\right)e^{-4\gamma_\infty z}\right]^{\frac{1}{2}} \tag{34}$$

for the half-width of the pulse. Figure 2 shows δ_P and τ_P plotted as a function of $\gamma_\infty z$ for $\theta = 0$ and $\theta = \Theta_o$. At first, a replica of the input pulse ($\tau = 0$) propagates in every mode without broadening and merely suffers a mode-dependent delay $n\theta_z^2/2c$, as we learned already from (18). Very soon, however, the pulses in the individual modes widen; they begin to overlap even before the length $1/\gamma_\infty$ is reached. Once $1/\gamma_\infty$ is passed, the pulse width in all modes increases mainly as $T(\gamma_\infty z)^{\frac{1}{2}}$. Compared to this increase, delay and pulse width differences in different modes become negligible since they cease to increase for large z. Specifically,

$$\delta_P(\theta) = \frac{T}{2}\left(\gamma_\infty z - \frac{1}{2} + \frac{\theta^2}{\Theta_o^2}\right), \qquad \text{for } \gamma_\infty z \gg 1 \tag{35}$$

and

$$\tau_P(\theta) = \frac{T}{2}\left(\gamma_\infty z - \frac{5}{4} + \frac{\theta^2}{\Theta_o^2}\right)^{\frac{1}{2}} \qquad \text{for } \gamma_\infty z \gg 1. \tag{36}$$
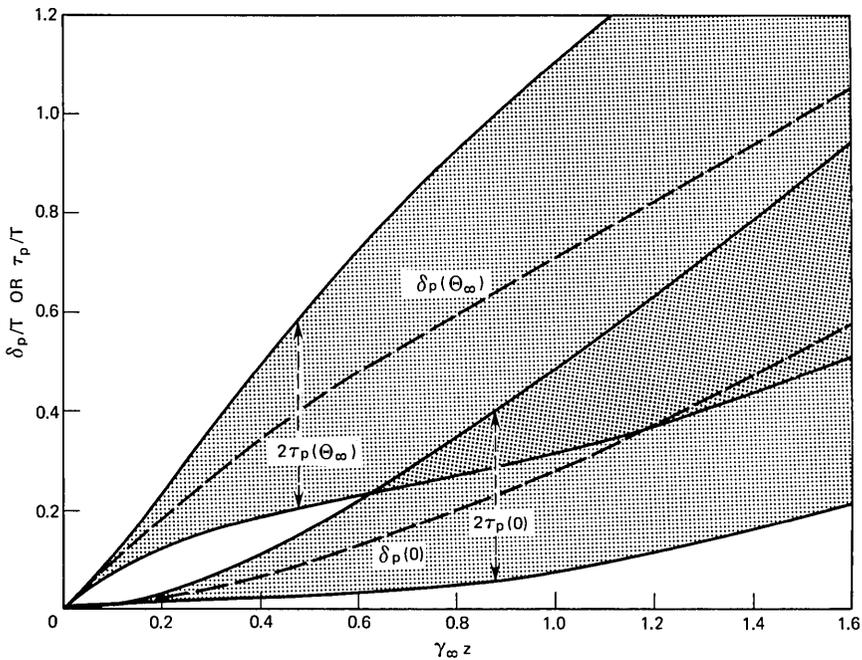
Fig. 2—Delay and time spread of the fiber output on axis and at an angle $\theta = \Theta_\infty$.

To calculate delay and width of the impulse response $Q(t)$, we must first apply the integration (19) to the general solution (12). This yields

$$q = \pi\Theta_o^2(\sigma \sinh \sigma\gamma_\infty z + \cosh \sigma\gamma_\infty z)^{-1}. \tag{37}$$

Now, by forming the first and second derivative of q according to (27) and (28), we obtain

$$\delta_Q = \frac{T}{2}\left[\gamma_\infty z + \frac{1}{2}(1 - e^{-2\gamma_\infty z})\right] \tag{38}$$

and

$$\tau_Q = \frac{T}{2}\left[\gamma_\infty z(1 - 2e^{-2\gamma_\infty z}) + \frac{3}{4} - e^{-2\gamma_\infty z} + \frac{1}{4}e^{-4\gamma_\infty z}\right]^{\frac{1}{2}}. \tag{39}$$

The ratio τ/T is shown in Fig. 3 plotted versus the normalized fiber length $\gamma_\infty z$. For $z \ll 1/\gamma_\infty$, the width τ approaches $T\gamma_\infty z$, as expected for negligible coupling. At $z = 1/4\gamma_\infty$, τ begins to follow a new asymptote

$$\tau = (T/2)(\gamma_\infty z)^{\frac{1}{2}}. \tag{40}$$

The quality of the approximation (40) is amazingly good even for small z.
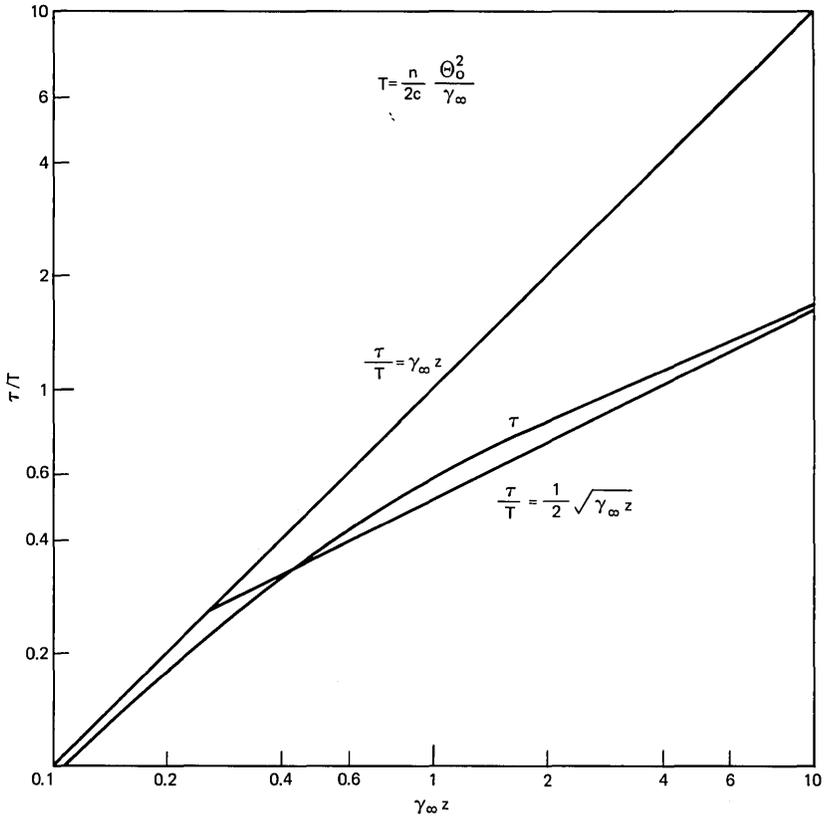
Fig. 3—Relative width of the impulse response plotted versus the normalized fiber length. The two straight lines show the asymptotic behavior for very short and very long fibers.

The amount by which τ deviates from $T\gamma_\infty z$ indicates the (desirable) effect of coupling: The width of the impulse response increases less with coupling than without,[6] the increase being proportional to $z^{\frac{1}{2}}$ rather than z.

V. SOME GENERAL RESULTS FOR LONG FIBERS

The simple approximation (40) can be obtained directly from (23) if σ in the denominator of that equation is set equal to unity. In this case,

$$q = \pi\Theta_\infty^2 \exp\left(-\sigma\gamma_\infty z\right) \qquad (41)$$

independent of the input condition. By applying (32) to the approxi-

mation (41), we obtain

$$\frac{\int_0^\infty t^m Q dt}{\int_0^\infty Q dt} = - \left(\frac{n}{2c}\right)^m \frac{\partial^m}{\partial A^m} (-\gamma_\infty z) \tag{42}$$

where Q is the inverse Laplace transform of $q(s)$ and hence the impulse response as before. Equation (42) is an important and powerful relation which permits us to calculate all moments of the impulse response from the steady-state loss coefficient of the *time-independent* power flow equation.

If we let $m = 2$ in (42), we obtain (40) as expected. For $m = 3$ we have

$$\eta = \frac{T}{2} (3z\gamma_\infty)^{\frac{1}{3}} \tag{43}$$

which according to its definition (31) describes the "skewness" of $Q(t)$. The ratio

$$\eta/\tau = (9/\gamma_\infty z)^{1/6} \tag{44}$$

is a measure of the asymmetry of the impulse response. For large z, η/τ approaches zero and, hence, $Q(t)$ becomes a symmetric function. We can compute this function by introducing $t = t' + T\gamma_\infty z/2$ with $t' \ll T\gamma_\infty z/2$ in (24). The impulse response

$$Q(t) = \Theta_\infty^2 (2\pi/\gamma_\infty z)^{\frac{1}{2}} \exp (-\gamma_\infty z - 2t'^2/T^2\gamma_\infty z) \tag{45}$$

is then Gaussian in time with the variance τ of (40).

The asymmetry parameter (44) can be used to determine γ_∞. Particularly if merely the order of magnitude of γ_∞ is of interest, this can be obtained, with some experience, from a quick look at the asymmetry of the impulse response.

Another conveniently measurable fiber characteristic is the angular width Θ_∞ of the steady-state mode distribution. It can be obtained from a scan of the (angular) far-field distribution at the end of a long fiber ($z > 1/\gamma_\infty$). If we define the effective numerical aperture NA of the fiber as the sine of the apex angle of this cone of radiation (measured at the $1/e$-points of the intensity), then

$$NA = n \sin \Theta_\infty \approx n\Theta_\infty. \tag{46}$$

Using (15), (16), (25), and (40), we can now write the width of the impulse response as

$$\tau = \frac{(NA)^2}{2nc} (z/4\gamma_\infty)^{\frac{1}{2}}. \tag{47}$$

This formula clearly shows the improvement, and the penalty, that results from coupling. Uncoupled, uniformly attenuated modes cause the impulse response to broaden to an effective width of $z(\text{NA})^2/2nc$ in z km of fiber. This width can be reduced by a factor $(4\gamma_\infty z)^{\frac{1}{2}}$ in exchange for an increase in the overall attenuation by $4.35\gamma_\infty$ dB/km.

The physical contents of these results can best be summarized if we define a "coupling length" $L = 1/4\gamma_\infty$. As shown in Fig. 3, this length marks the point at which the width of the impulse response changes from a linear to a square-root dependence on length. Together with this change, the impulse response undergoes a transition from the exponential shape (21) to the Gaussian shape (45). The inverse of the coupling length (in km) is very nearly equal to the excess loss (in dB/km) incurred because of the coupling phenomenon.

Equation (47) as well as the previous results are limited to fibers in which the coupling coefficient D is independent of θ. The study of liquid-core fibers, on the other hand, has given us reason to believe that, in some fibers, D decreases with increasing θ. This characteristic could have a desirable effect on the impulse response and the excess loss, since it reduces the power flow toward the lossy modes (large angles) while, at the same time, enhancing the coupling among all other modes. We therefore studied the general case

$$D(\theta) = D_o\, \theta^{-\nu}, \qquad \nu = 0, 1, 2, \ldots, \tag{48}$$

in some detail.

The steady-state parameters Θ_∞ and γ_∞ can be obtained as general functions of A, D_o, and ν by using the Rayleigh-Ritz procedure.[9] Due to (42), twofold derivation of γ_∞ with respect to A then yields directly the effective width of the impulse response. This calculation leads to

$$\tau = \frac{(\text{NA})^2}{2nc} \left[z/(4 + \nu)\gamma_\infty\right]^{\frac{1}{2}} \tag{49}$$

where $\text{NA} = n\Theta_\infty$ denotes the effective width of the output radiation as before. For $\nu = 0$, (49) reduces to (47). An exponent $\nu > 0$ indeed narrows the impulse response, although not very significantly. For $D = D_o/\Theta^4$, for example (a strong angular dependence indeed), τ is only 0.7 times narrower than in the case $D = \text{const}$.

Another limitation of these results is the requirement that $\Theta_\infty \ll (2\Delta)^{\frac{1}{2}}$. If this is not the case, the steady-state distribution is generally determined by a sharply rising loss term at $\theta_{\max} = (2\Delta)^{\frac{1}{2}}$ rather than by the quadratic term $A\theta^2$. Under these conditions we find that the functional relation (47) still holds, although with different

coefficients, so that for practical distributions $D(\theta)$ the width τ can be up to three times smaller than indicated by (47) or (36).

As a typical example, we shall use (36) to compute the data rates achievable. Let us assume that we had means to design and manufacture a coupling structure in the fiber which produced the desired excess loss γ_∞ and the desired numerical aperture. For simplicity we assume the input pulse to be somewhat narrower than the fiber impulse response so that (36) gives a good measure of the half-width of the expected output pulse. We then choose a data rate

$$B = 1/2\tau. \tag{50}$$

We assume the core loss common to all modes to be 4 dB/km and allow 50 dB of loss between repeaters.

Using (36) we can then calculate the excess loss and the repeater spacing necessary for a desired data rate. These results are plotted in Fig. 4. A 1-dB/km excess loss decreases the possible repeater spacing by only 25 percent but, at the same time, triples the data rate. An attempt to further increase the data rate by even more coupling is costly:
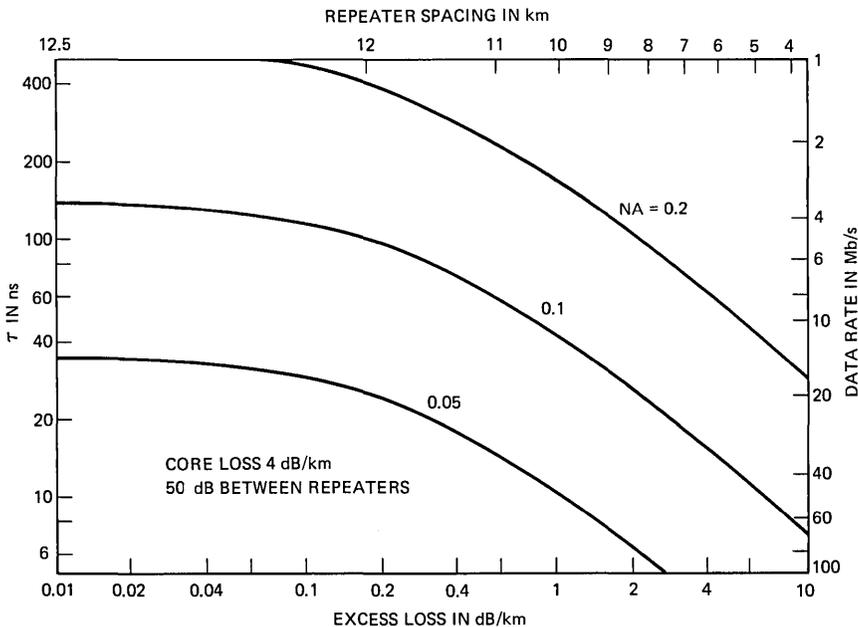
Fig. 4—Width of the impulse response plotted versus the excess loss incurred because of coupling for 4 dB/km core attenuation and 50 dB loss between repeaters. Right side shows equivalent data rates. Repeater spacing is shown at the top.

Another threefold increase in the data rate requires 4 dB/km excess loss, and divides the repeater spacing in half. The results of Fig. 4, of course, are based on a uniform coupling distribution and an excess loss increasing as θ^2. As mentioned earlier, more favorable distributions could result in an impulse response permitting a higher data rate, although there may be practical limitations to the extent of this improvement.

VI. CONCLUSIONS

The description of fiber modes by a continuum results in a partial differential equation whose solution yields the response function of the fiber. We find a characteristic length indicating the region in which the impulse response changes from an exponential to a Gaussian shape. Beyond this length, the width of the impulse response increases only as the square root of the fiber length. In practical fibers, the inverse of this length turns out to be proportional to the excess loss incurred because of the coupling phenomenon. The latter may represent a practical limit to the improvement that can be gained from coupling. As an example, we find a data rate of 12 Mb/s achievable for 10 km repeater spacing and an effective numerical aperture of 0.1. The data rate is inversely proportional to the square of the numerical aperture. Thus half the numerical aperture permits a fourfold increase of the data rate. Another increase of the data rate without a penalty in loss or numerical aperture is theoretically possible by artificially creating a more suitable coupling characteristic in the fiber, but it seems that the technological requirements for doubling or tripling the data rate in this way are high.

VII. ACKNOWLEDGMENTS

REFERENCES

1. Dyott, R. B., and Stern, J. R., "Group Delay in Glass Fiber Waveguide," Elec. Lett., 7, 1971, pp. 82–84.
2. Gloge, D., Chinnock, E. L., and Lee, T. P., "Self-Pulsing GaAs Laser for Fiber Dispersion Measurements," IEEE J. Quantum Elec., QE-8, 1972, pp. 844–846.
3. Gloge, D., Chinnock, E. L., Standley, R. D., and Holden, W. S., "Dispersion in a Low-Loss Multimode Fiber Measured at Three Wavelengths," Elec. Lett., 8, 1972, pp. 527–529.
4. Gloge, D., Tynes, A. R., Duguay, M. A., and Hansen, J. W., "Picosecond Pulse Distortion in Optical Fibers," IEEE J. Quantum Elec., QE-8, 1972, pp. 217–221.

5. Gloge, D., "Optical Power Flow in Multimode Fibers," B.S.T.J., *51*, No. 8 (October 1972), pp. 1767–1783.
6. Personick, S. D., "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
7. Marcuse, D., "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1199–1232.
8. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, 1971, pp. 2252–2258.
9. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics*, vol. II, New York: McGraw-Hill, 1953, p. 1115.

Coupled Mode Theory of Round Optical Fibers

By D. MARCUSE

This paper presents a comprehensive theory of mode coupling in optical fibers with imperfections. The paper begins with the derivation of a general coupled wave theory based on the modes of the ideal fiber. The general theory is applied to a simplified description of guided and radiation modes of the fiber that is valid for small core-cladding index differences. The simplified theory results in expressions for the coupling coefficients that are nearly as simple as those of the slab waveguide. As an example, the theory is applied to the calculation of radiation losses caused by pure core diameter changes and by elliptical deformations of the fiber core.

I. INTRODUCTION

Dielectric optical waveguides support a finite number of guided modes and an infinite number of radiation modes.[1] Even if the number of guided modes that can be supported by the waveguide is reduced to one, the presence of the infinite number of radiation modes forces us to be concerned about mode conversion phenomena. Coupling among the guided modes of a multimode optical waveguide (multimode waveguide refers to the guided modes) is caused by imperfections in the refractive index distribution or the geometry of the optical waveguide. Its effects may be beneficial for reducing the delay distortion that results from uncoupled multimode operation.[2,3] Coupling between guided modes and the continuum of radiation modes is usually not desired unless the waveguide is intended to serve as an antenna. However, a certain amount of coupling is unavoidable and results in scattering losses.[1,4,5]

E. G. Rawson has calculated light scattering from fiber waveguides with irregular core surfaces by an approximate technique.[6] However, his method is not suitable to calculate coupling between guide modes.

A theory of radiation losses in round optical fibers has been presented in Ref. 7. This theory was based on a field expansion in terms of the exact modes of the fiber. Because of the complicated mode fields, a theory based on the exact modes of the guide is very tedious and results in equations whose numerical evaluation is difficult and costly. However, Snyder[8] and Gloge[9] have shown that the description of the modes of dielectric waveguides can be greatly simplified if it is assumed that the difference of the refractive indices of core and cladding is only very slight. This assumption allows an approximate treatment of the mode problem, resulting in very much simpler field expressions. The coupled mode theory based on approximate modes yields expressions for the coupling coefficients that are almost as simple as those for the slab waveguide. The entire coupled mode theory thus becomes simpler and its numerical evaluation becomes cheaper.

This paper starts out with a derivation of the coupled wave equations in terms of modes of the ideal guide. This mode description is somewhat different from the coupled mode theory in terms of local normal modes used by Snyder.[10,11] It results in simpler expressions for the coupling coefficients. The exact coupled mode theory is then applied to the problem of coupling between the simplified waveguide modes. We limit the discussion to coupling caused by changes or imperfections in the waveguide geometry. Coupling caused by refractive index inhomogeneities, which could be handled in a similar fashion with the use of the exact expressions for the coupling coefficients and the approximate mode description, is not discussed in this paper.

Finally, we apply our results to the problem of scattering losses of guided modes caused by diameter changes and elliptical deformations of the waveguide core. We also derive simplified expressions for the coupling coefficients between guided modes far from cutoff and discuss coupling between guided modes caused by deformations of the fiber core and by curvature of the waveguide axis.

II. EXACT COUPLED MODE THEORY

The dielectric optical waveguide with imperfections is defined by a certain refractive index distribution $n = n(x, y, z)$ that enters Maxwell's equations:

$$\nabla \times \mathbf{H} = i\omega\epsilon_o n^2 \mathbf{E} \tag{1}$$

$$\nabla \times \mathbf{E} = -i\omega\mu_o \mathbf{H}. \tag{2}$$

E and **H** are the electric and magnetic field vectors of a general field distribution in the waveguide. The fields are assumed to have the time

dependence $e^{i\omega t}$, with radian frequency ω; ϵ_o and μ_o are the vacuum values of the electric permittivity and magnetic permeability. In addition to the refractive index distribution n of the real waveguide, we consider the index distribution n_o that defines an ideal guide from which the real guide deviates in some way.

We decompose the fields into their transverse and longitudinal parts. The electric field is thus represented by the equation

$$\mathbf{E} = \mathbf{E}_t + \mathbf{E}_z \tag{3}$$

and the magnetic field is given by

$$\mathbf{H} = \mathbf{H}_t + \mathbf{H}_z. \tag{4}$$

By using a similar decomposition of the ∇ operator ($\mathbf{e}_z$ is a unit vector in the z-direction),

$$\nabla = \nabla_t + \mathbf{e}_z \frac{\partial}{\partial z}, \tag{5}$$

Maxwell's equations can be written in the form

$$\nabla_t \times \mathbf{H}_z + \left(\mathbf{e}_z \times \frac{\partial \mathbf{H}_t}{\partial z} \right) = i\omega\epsilon_o n^2 \mathbf{E}_t \tag{6}$$

and

$$\nabla_t \times \mathbf{E}_z + \left(\mathbf{e}_z \times \frac{\partial \mathbf{E}_t}{\partial z} \right) = -i\omega\mu_o \mathbf{H}_t. \tag{7}$$

The longitudinal field components are expressed in terms of the transverse field components

$$\mathbf{E}_z = \frac{1}{i\omega\epsilon_o n^2} \nabla_t \times \mathbf{H}_t \tag{8}$$

and

$$\mathbf{H}_z = - \frac{1}{i\omega\mu_o} \nabla_t \times \mathbf{E}_t. \tag{9}$$

The modes of the ideal waveguide with index distribution $n_o(x, y)$ are defined as solutions of the equations

$$\nabla_t \times \mathfrak{3C}_{\nu z} - i\beta_\nu(\mathbf{e}_z \times \mathfrak{3C}_{\nu t}) = i\omega\epsilon_o n_o^2 \mathcal{E}_{\nu t} \tag{10}$$

and

$$\nabla_t \times \mathcal{E}_{\nu z} - i\beta_\nu(\mathbf{e}_z \times \mathcal{E}_{\nu t}) = -i\omega\mu_o \mathfrak{3C}_{\nu t}. \tag{11}$$

The index ν is a mode label and β_ν is the propagation constant of the νth mode. The longitudinal components of the mode fields are similarly expressed as

$$\mathcal{E}_{\nu z} = \frac{1}{i\omega\epsilon_o n_o^2} \nabla_t \times \mathfrak{3C}_{\nu t} \tag{12}$$

and

$$\mathcal{3C}_{\nu z} = - \frac{1}{i\omega\mu_o} \nabla_t \times \mathcal{E}_{\nu t}. \tag{13}$$

The transverse field of the waveguide with the index distribution $n(x, y, z)$ is now expressed as a superposition of modes of the ideal waveguide.

$$\mathbf{E}_t = \sum_\nu a_\nu \mathcal{E}_{\nu t} \tag{14}$$

and

$$\mathcal{3C}_t = \sum_\nu b_\nu \mathcal{3C}_{\nu t}. \tag{15}$$

The summation symbol in (14) and (15) indicates summation over guided modes and integration over radiation modes. Indicating the index ν by the symbol ρ in case that it belongs to the continuum of radiation modes, we have to replace

$$\sum_\nu \rightarrow \sum \int_0^\infty d\rho. \tag{16}$$

The sum in front of the integral on the right-hand side of (16) indicates a summation over the various types of radiation modes.

For the derivation of coupled differential equations for the expansion coefficients a_ν and b_ν we need the orthogonality relations of the modes of the ideal waveguide.

$$\int_{-\infty}^\infty \int_{-\infty}^\infty \mathbf{e}_z \cdot (\mathcal{E}_{\nu t} \times \mathcal{3C}_{\mu t}^*) dx dy = 2 \frac{\beta_\mu^*}{|\beta_\mu|} P \delta_{\nu\mu}. \tag{17}$$

The asterisks indicate complex conjugation. The symbol $\delta_{\nu\mu}$ indicates Kronecker's delta for discrete values of ν and μ; it is zero if one of the indices labels a guided mode while the other labels a radiation mode, and it becomes Dirac's delta function if both indices label radiation modes.

The series expansions (14) and (15) are now substituted into the equations (6) through (9). Making use of the fact that the mode fields satisfy the equations (10) through (13), we obtain

$$\sum_\nu \left\{ \left(\frac{db_\nu}{dz} + i\beta_\nu a_\nu \right) (\mathbf{e}_z \times \mathcal{3C}_{\nu t}) - i\omega\epsilon_o(n^2 - n_o^2)a_\nu \mathcal{E}_{\nu t} \right\} = 0 \tag{18}$$

and

$$\sum_\nu \left\{ \left(\frac{da_\nu}{dz} + i\beta_\nu b_\nu \right) (\mathbf{e}_z \times \mathcal{E}_{\nu t}) + \frac{1}{i\omega\epsilon_o} b_\nu \nabla_t \right.$$
$$\left. \times \left[\left(\frac{1}{n^2} - \frac{1}{n_o^2} \right) (\nabla_t \times \mathcal{3C}_{\nu t}) \right] \right\} = 0. \tag{19}$$

We take the scalar product of (18) with $\mathcal{E}_{\mu t}^*$ and of (19) with $\mathcal{K}_{\mu t}^*$. After integration over the infinite cross section, we obtain with the help of the orthogonality relation (17)

$$\frac{db_\mu}{dz} + i\beta_\mu a_\mu = 2 \sum_\nu \bar{K}_{\mu\nu} a_\nu \tag{20}$$

and

$$\frac{da_\mu}{dz} + i\beta_\mu b_\mu = 2 \sum_\nu \bar{k}_{\mu\nu} b_\nu \tag{21}$$

with the coupling coefficients

$$\bar{K}_{\mu\nu} = \frac{\omega\epsilon_o}{4iP} \frac{|\beta_\mu|}{\beta_\mu} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (n^2 - n_o^2) \mathcal{E}_{\mu t}^* \cdot \mathcal{E}_{\nu t} dx dy \tag{22}$$

and

$$\bar{k}_{\mu\nu} = \frac{-1}{4i\omega\epsilon_o P} \frac{|\beta_\mu|}{\beta_\mu^*} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{K}_{\mu t}^* \cdot \nabla_t$$
$$\times \left[\left(\frac{1}{n^2} - \frac{1}{n_o^2} \right) \nabla_t \times \mathcal{K}_{\nu t} \right] dx dy. \tag{23}$$

Equation (23) can be brought into a simpler form with the help of (12) and by performing a partial integration

$$\bar{k}_{\mu\nu} = \frac{\omega\epsilon_o}{4iP} \frac{|\beta_\mu|}{\beta_\mu^*} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{n_o^2}{n^2} (n^2 - n_o^2) \mathcal{E}_{\mu z}^* \cdot \mathcal{E}_{\nu z} dx dy. \tag{24}$$

Finally, we introduce the amplitudes $c_\mu^{(+)}$ and $c_\mu^{(-)}$ of forward and backward traveling waves by means of the transformation

$$a_\mu = c_\mu^{(+)} e^{-i\beta_\mu z} + c_\mu^{(-)} e^{i\beta_\mu z} \tag{25}$$

and

$$b_\mu = c_\mu^{(+)} e^{-i\beta_\mu z} - c_\mu^{(-)} e^{i\beta_\mu z}. \tag{26}$$

Substitution into (20) and (21), addition and subtraction of the resulting equations, and regrouping of terms results in the desired coupled wave equations

$$\frac{dc_\mu^{(+)}}{dz} = \sum_\nu \{ K_{\mu\nu}^{(+,+)} c_\nu^{(+)} e^{i(\beta_\mu - \beta_\nu)z} + K_{\mu\nu}^{(+,-)} c_\nu^{(-)} e^{i(\beta_\mu + \beta_\nu)z} \} \tag{27}$$

$$\frac{dc_\mu^{(-)}}{dz} = \sum_\nu \{ K_{\mu\nu}^{(-,+)} c_\nu^{(+)} e^{-i(\beta_\mu + \beta_\nu)z} + K_{\mu\nu}^{(-,-)} c_\nu^{(-)} e^{-i(\beta_\mu - \beta_\nu)z} \}. \tag{28}$$

The coupling coefficients are defined as

$$K_{\mu\nu}^{(p,q)} = \frac{\omega\epsilon_o}{4iP} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (n^2 - n_o^2) \left\{ p \frac{|\beta_\mu|}{\beta_\mu} \mathcal{E}_{\mu t}^* \cdot \mathcal{E}_{\nu t} \right.$$
$$\left. + q \frac{|\beta_\mu|}{\beta_\mu^*} \frac{n_o^2}{n^2} \mathcal{E}_{\mu z}^* \cdot \mathcal{E}_{\nu z} \right\} dx dy. \tag{29}$$

The factors and superscripts p and q indicate the symbols $(+)$ or $(-)$ or the corresponding factors $+1$ and -1. The propagation constants β_μ are positive quantities. The coupled wave equations (27) and (28) provide an exact description of the imperfect waveguide in terms of normal modes of the perfect guide. The use of normal modes of the perfect guide results in the simple general form (29) for the coupling coefficients.

Our coupled mode theory can be applied to any type of waveguide problem such as waveguides with refractive index inhomogeneities, tapers, or bends. It may be that the expansion in terms of ideal modes of the waveguide does not provide the most convenient basis for some problems. For the description of tapers, for example, we face the following situation. Consider a waveguide which is perfectly straight and uniform up to a point where its cross section begins to increase. After some distance, the taper connects to a uniform waveguide of constant cross section. If we use the modes of the smaller guide for our mode expansion, we see immediately that the coupling coefficients have nonzero values not only on the taper itself but also throughout the entire waveguide of larger cross section. Even though our description is precise and yields the right answers, it is inconvenient for the problem at hand. It would be far better to use so-called local normal modes that do not themselves describe wave forms in any real waveguide but correspond at each point z along the nonuniform guide to the modes of a hypothetical uniform guide whose cross section coincides locally with that of the waveguide under study. Using local normal modes results in coupled wave equations of the form (27) and (28) but with different coupling coefficients. In case of the taper, these coupling coefficients would be nonzero only on the taper itself but would vanish on the uniform waveguide sections. Local normal modes are obviously better suited for the description of tapers. For the description of waves in bent waveguides it would be most convenient to use modes that locally correspond to a straight waveguide whose axis is tangential to that of the actual guide. In addition to these problems of convenience, there exist problems of convergence of the series expansions (14) and (15). A series expansion in terms of ideal modes may converge more slowly than an expansion in terms of local normal modes. However, we shall see that we can use the series expansion in terms of ideal modes to treat most problems of waveguides with only slight refractive index differences between the core and cladding materials. In addition, it is usually possible to guess the form of the coupling coefficients of a particular expansion from the coupling coefficients for the ideal mode expansion.

For the purposes of this paper, we restrict ourselves to the description of waveguides with piecewise constant refractive index distributions and allow only deformations of the cross section of the waveguide core. Figure 1 shows a typical waveguide imperfection. The refractive index distributions n and n_o coincide inside of the core and in the cladding region. They differ only near the core-cladding boundary. If the boundary has moved outward from its ideal position, the index difference $n^2 - n_o^2$ equals $n_1^2 - n_2^2$ in the region where the actual core overlaps the ideal cladding; it vanishes everywhere else. If the core boundary has moved inwards, we have $n^2 - n_o^2 = -(n_1^2 - n_2^2)$ in the region where the actual cladding overlaps the ideal core and zero values everywhere else. The field components that multiply the refractive index difference term in (29) are either continuous, if they are tangential to the boundary of the ideal waveguide, or they jump by a factor $(n_1/n_2)^2$ if they are normal to the ideal core boundary. If we restrict the discussion to core boundary displacements that are so slight that the fields can be considered constant over the region of the displacement and to weakly guiding fibers with $(n_1/n_2 - 1) \ll 1$, we obtain from (29)

$$K_{\mu\nu}^{(p,q)} = -\frac{i\omega\epsilon_o(n_1^2 - n_2^2)a}{4P} \int_0^{2\pi} [r(x, y, z) - a]$$
$$\times [p\,\boldsymbol{\mathcal{E}}_{\mu t}^* \cdot \boldsymbol{\mathcal{E}}_{\nu t} + q\,\boldsymbol{\mathcal{E}}_{\mu z}^* \cdot \boldsymbol{\mathcal{E}}_{\nu z}]\}d\phi. \quad (30)$$

It is noteworthy that the approximate coupling coefficient (30) is identical to the coupling coefficient (13) of Ref. 12 for the local normal mode expansion. The only difference in the appearance of these two coupling coefficients consists in the fact that the derivative of the boundary function instead of the function itself appears in Ref. 12 and that the entire expression is divided by $\beta_\nu - \beta_\mu$. It has been explained
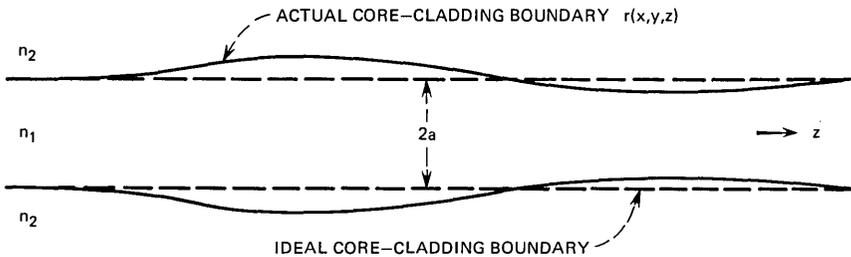

Fig. 1—Sketch of a fiber with distorted core-cladding interface.

in a different place[13] that it is the Fourier component of the boundary function at the spatial frequency $\beta_\nu - \beta_\mu$ that determines the coupling behavior of the modes. This fact makes it clear that (we write f instead of $r - a$ for simplicity)

$$\frac{1}{i(\beta_\nu - \beta_\mu)} \frac{df}{dz} \tag{31}$$

is fully equivalent to the function f itself as far as its effect on mode coupling is concerned. We can carry the argument one step further and replace f with

$$- \frac{1}{(\beta_\nu - \beta_\mu)^2} \frac{d^2 f}{dz^2}. \tag{32}$$

If we replace $f = r - a$ in (30) with (32), we obtain a coupling coefficient that vanishes in straight, uniform waveguide sections. It is not hard to guess that a coupling coefficient of this type belongs to an expansion in terms of local normal modes of a hypothetical waveguide that is tangential to the curved axis of the actual guide. The modification of (30) that is indicated by (32) is thus particularly suitable for the description of mode coupling caused by bends of the waveguide axis. A description in terms of ideal modes or even in terms of local normal modes of a hypothetical guide with straight axis is unsuitable for a description of waveguide bends since it leads to coupling coefficients that do not vanish on the straight waveguide section behind the bend. This brief discussion shows that it is not hard to modify the coupling coefficients of the ideal mode coupling theory to extend it to the case of local normal mode expansions of different types.

III. SIMPLIFIED DESCRIPTION OF GUIDED MODES OF THE FIBER

A. W. Snyder[8] realized that the modes of round fibers and their eigenvalue equations simplify considerably if use is made of the fact that $(n_1/n_2 - 1) \ll 1$ applies to most fibers of practical interest. D. Gloge[9] went one step further and showed that the mode fields become simple in appearance if they are expressed in Cartesian instead of the more conventional description in cylindrical coordinates. Gloge's technique is useful for even more complicated waveguide structures such as tubes.[14]

We write down the field expressions for the guided modes of the round fiber without derivation.[9] The mode fields can be polarized in two mutually orthogonal directions. We have for one polarization in the core region for $r < a$

$$\mathcal{E}_z = \frac{iA\kappa}{2\beta_\nu} \left[J_{\nu+1}(\kappa r) \left\{ \begin{array}{c} \sin(\nu+1)\phi \\ -\cos(\nu+1)\phi \end{array} \right\} \right.$$
$$\left. + J_{\nu-1}(\kappa r) \left\{ \begin{array}{c} \sin(\nu-1)\phi \\ -\cos(\nu-1)\phi \end{array} \right\} \right] \quad (33)$$

$$\mathcal{E}_y = AJ_\nu(\kappa r) \left\{ \begin{array}{c} \cos \nu\phi \\ \sin \nu\phi \end{array} \right\} \quad (34)$$

$$\mathcal{H}_z = -\frac{iA\kappa}{2k} \sqrt{\frac{\epsilon_o}{\mu_o}} \left[J_{\nu+1}(\kappa r) \left\{ \begin{array}{c} \cos(\nu+1)\phi \\ \sin(\nu+1)\phi \end{array} \right\} \right.$$
$$\left. - J_{\nu-1}(\kappa r) \left\{ \begin{array}{c} \cos(\nu-1)\phi \\ \sin(\nu-1)\phi \end{array} \right\} \right] \quad (35)$$

$$\mathcal{H}_x = -nA \frac{\beta_\nu}{|\beta_\nu|} \sqrt{\frac{\epsilon_o}{\mu_o}} J_\nu(\kappa r) \left\{ \begin{array}{c} \cos \nu\phi \\ \sin \nu\phi \end{array} \right\}. \quad (36)$$

The field in the (infinite) cladding region for $r > a$ is obtained from the field expressions (33) through (36) by replacing the amplitude constant A with $[J_\nu(\kappa a)/H_\nu^{(1)}(i\gamma a)]A$. In addition, we replace κ with $i\gamma$ and the Bessel function $J_\nu(\kappa r)$ with the Hankel function of the first kind $H_\nu^{(1)}(i\gamma r)$. The parameters κ and γ are defined as ($k^2 = \omega^2\epsilon_o\mu_o$)

$$\kappa = (n_1^2 k^2 - \beta_\nu^2)^{\frac{1}{2}} \quad (37)$$

and

$$\gamma = (\beta_\nu^2 - n_2^2 k^2)^{\frac{1}{2}}. \quad (38)$$

The remaining field components vanish, $\mathcal{E}_x = 0$ and $\mathcal{H}_y = 0$. The two sets of circular functions that are shown in the field equations are necessary to obtain a complete set of orthogonal modes. The functions in the upper as well as those in the lower position belong together. We have used n to indicate $n \approx n_1 \approx n_2$.

The set of guided modes is still not complete unless we also include the orthogonal polarization. We have again for $r < a (\mathcal{E}_y = \mathcal{H}_x = 0)$

$$\mathcal{E}_z = \frac{iA\kappa}{2\beta_\nu} \left[J_{\nu+1}(\kappa r) \left\{ \begin{array}{c} \cos(\nu+1)\phi \\ \sin(\nu+1)\phi \end{array} \right\} - J_{\nu-1}(\kappa r) \left\{ \begin{array}{c} \cos(\nu-1)\phi \\ \sin(\nu-1)\phi \end{array} \right\} \right] \quad (39)$$

$$\mathcal{E}_x = AJ_\nu(\kappa r) \left\{ \begin{array}{c} \cos \nu\phi \\ \sin \nu\phi \end{array} \right\} \quad (40)$$

$$\mathcal{H}_z = \frac{iA\kappa}{2k} \sqrt{\frac{\epsilon_o}{\mu_o}} \left[J_{\nu+1}(\kappa r) \left\{ \begin{array}{c} \sin(\nu+1)\phi \\ -\cos(\nu+1)\phi \end{array} \right\} \right.$$
$$\left. + J_{\nu-1}(\kappa r) \left\{ \begin{array}{c} \sin(\nu-1)\phi \\ -\cos(\nu-1)\phi \end{array} \right\} \right] \quad (41)$$

$$\mathcal{H}_y = nA \frac{\beta_\nu}{|\beta_\nu|} \sqrt{\frac{\epsilon_o}{\mu_o}} J_\nu(\kappa r) \left\{ \begin{array}{c} \cos \nu\phi \\ \sin \nu\phi \end{array} \right\}. \quad (42)$$

The field in the cladding is again obtained by the replacements mentioned above.

The mode amplitudes must be related to the power P (that is the same for all the modes). We have for the fields (33) through (36) and for the orthogonally polarized field (39) through (42)

$$A = \left\{ \frac{4\sqrt{\frac{\mu_o}{\epsilon_o}}\,\gamma^2 P}{e_\nu \pi a^2 n(n_1^2 - n_2^2)k^2 |J_{\nu-1}(\kappa a)J_{\nu+1}(\kappa a)|} \right\}^{\frac{1}{2}} \tag{43}$$

with

$$e_\nu = \begin{cases} 2 & \text{for} \quad \nu = 0 \\ 1 & \text{for} \quad \nu \neq 0 \end{cases}. \tag{44}$$

The eigenvalue equation of these simplified fiber modes is

$$\kappa \frac{J_{\nu-1}(\kappa a)}{J_\nu(\kappa a)} = i\gamma \frac{H_{\nu-1}^{(1)}(i\gamma a)}{H_\nu^{(1)}(i\gamma a)}. \tag{45}$$

With the help of the functional relations of the cylinder functions, it is easy to show that (45) is also valid if $\nu - 1$ is replaced by $\nu + 1$.

The simplified guided modes listed here are not the same as the usual HE and EH modes of the round fiber. It can be shown[9,15] that the simplified modes listed in this paper result from the usual fiber modes as superpositions of an HE and an EH mode. The $HE_{\nu+1,\mu}$ and $EH_{\nu-1,\mu}$ modes have very nearly the same propagation constant, they are almost degenerate. Since this degeneracy is not perfect, our simplified modes are not modes in the true sense of the word. They decompose into the $HE_{\nu+1,\mu}$ and $EH_{\nu-1,\mu}$ modes of the round fiber as they travel along the waveguide thus changing their shape. A true mode is defined by the fact that only its phase changes (in the lossless case) as it travels down the guide. However, the approximate modes do form a complete orthogonal set of modes and can thus be used to express any field that can exist in the fiber. Even after one of the approximate modes has decomposed into HE and EH modes, it can again be expressed in terms of approximate modes at this point. The fact that the approximate modes are not true modes in the usual sense does not limit their usefulness for studying mode conversion and radiation problems.

The important fundamental HE_{11} mode of the fiber corresponds to the lowest-order approximate mode with $\nu = 0$. This is a true mode that does not decompose as it travels along the waveguide.

IV. SIMPLIFIED RADIATION MODES OF THE FIBER

The radiation modes of the fiber can again be simplified by using $(n_1/n_2 - 1) \ll 1$.[16] There is a slight complication, however. The simpli-

fied description of the guided modes was made possible by the fact that they are very nearly transverse modes, their transverse components being much larger than their longitudinal components. The radiation modes are nearly transverse only if their propagation constants are nearly $\beta = n_2 k$. Since the continuous spectrum of radiation modes extends from $\beta = -n_2 k$ to $\beta = n_2 k$, only the modes in the immediate vicinity of the two end points of this interval are also nearly transverse modes. Throughout most of the spectral region of β values, the approximation corresponding to that for the guided modes does not work. However, we can still use the fact that the refractive indices of core and cladding are nearly identical and use the radiation modes of free space in the region where our mode approximation technique fails. A simplified treatment of all the radiation modes is thus also possible. The two approximations complement each other. In the region near $\beta = \pm n_2 k$, where we use the approximate radiation modes of the guide, the free-space radiation modes do not work very well because reflections at the core-cladding interface at grazing angles are important. Inside of the β range, where the waveguide mode approximation method fails, we can use the free-space radiation modes with confidence since the interface does not cause much reflection for waves passing through it at reasonably steep angles.

We begin by listing the approximate radiation modes of the fiber for β near $\pm n_2 k$. The field equations are very similar to those of the guided modes. In the fiber core at $r < a$ we have

$$\mathscr{E}_z = \frac{iB\sigma}{2\beta} \left[J_{\nu+1}(\sigma r) \begin{Bmatrix} \sin(\nu+1)\phi \\ -\cos(\nu+1)\phi \end{Bmatrix} \right.$$
$$\left. + J_{\nu-1}(\sigma r) \begin{Bmatrix} \sin(\nu-1)\phi \\ -\cos(\nu-1)\phi \end{Bmatrix} \right] \quad (46)$$

$$\mathscr{E}_y = B J_\nu(\sigma r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \quad (47)$$

$$\mathscr{H}_z = -\frac{iB\sigma}{2k} \sqrt{\frac{\epsilon_o}{\mu_o}} \left[J_{\nu+1}(\sigma r) \begin{Bmatrix} \cos(\nu+1)\phi \\ \sin(\nu+1)\phi \end{Bmatrix} \right.$$
$$\left. - J_{\nu-1}(\sigma r) \begin{Bmatrix} \cos(\nu-1)\phi \\ \sin(\nu-1)\phi \end{Bmatrix} \right] \quad (48)$$

$$\mathscr{H}_x = -nB \frac{\beta}{|\beta|} \sqrt{\frac{\epsilon_o}{\mu_o}} J_\nu(\sigma r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix}. \quad (49)$$

The remaining field components vanish. The propagation constant β is a continuous variable for radiation modes unrestricted by an eigenvalue

equation. The parameter σ is defined as

$$\sigma = (n_1^2 k^2 - \beta^2)^{\frac{1}{2}} \tag{50}$$

Instead of specifying the modifications that are required to transform the expression of the field inside of the core into the expression for the cladding field, we state the field in the region $r > a$ in detail.

$$\mathscr{E}_z = \frac{iC\rho}{2\beta} \left[(H_{\nu+1}^{(1)}(\rho r) + D H_{\nu+1}^{(2)}(\rho r)) \begin{Bmatrix} \sin (\nu + 1)\phi \\ -\cos (\nu + 1)\phi \end{Bmatrix} \right.$$
$$\left. + (H_{\nu-1}^{(1)}(\rho r) + D H_{\nu-1}^{(2)}(\rho r)) \begin{Bmatrix} \sin (\nu - 1)\phi \\ -\cos (\nu - 1)\phi \end{Bmatrix} \right] \tag{51}$$

$$\mathscr{E}_y = C(H_\nu^{(1)}(\rho r) + D H_\nu^{(2)}(\rho r)) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \tag{52}$$

$$\mathscr{H}_z = -\frac{iC\rho}{2k} \sqrt{\frac{\epsilon_o}{\mu_o}} \left[(H_{\nu+1}^{(1)}(\rho r) + D H_{\nu+1}^{(2)}(\rho r)) \begin{Bmatrix} \cos (\nu + 1)\phi \\ \sin (\nu + 1)\phi \end{Bmatrix} \right.$$
$$\left. - (H_{\nu-1}^{(1)}(\rho r) + D H_{\nu-1}^{(2)}(\rho r)) \begin{Bmatrix} \cos (\nu - 1)\phi \\ \sin (\nu - 1)\phi \end{Bmatrix} \right] \tag{53}$$

$$\mathscr{H}_x = -nC \frac{\beta}{|\beta|} \sqrt{\frac{\epsilon_o}{\mu_o}} (H_\nu^{(1)}(\rho r) + D H_\nu^{(2)}(\rho r)) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix}. \tag{54}$$

$H_\nu^{(1)}$ and $H_\nu^{(2)}$ are the Hankel functions of the first and second kind. The parameter ρ is defined as

$$\rho = (n_2^2 k^2 - \beta^2)^{\frac{1}{2}}. \tag{55}$$

The amplitude coefficients are

$$C = \frac{i\pi a}{4} \left[\sigma J_{\nu+1}(\sigma a) H_\nu^{(2)}(\rho a) - \rho J_\nu(\sigma a) H_{\nu+1}^{(2)}(\rho a) \right] B \tag{56}$$

and

$$D = -\frac{\sigma J_{\nu+1}(\sigma a) H_\nu^{(1)}(\rho a) - \rho J_\nu(\sigma a) H_{\nu+1}^{(1)}(\rho a)}{\sigma J_{\nu+1}(\sigma a) H_\nu^{(2)}(\rho a) - \rho J_\nu(\sigma a) H_{\nu+1}^{(2)}(\rho a)}. \tag{57}$$

For the field with the orthogonal polarization, we simply state the field expressions inside of the core. It should be apparent from inspection of (46) through (57) how the field expression in the cladding is obtained from that of the core. The relations between the amplitude coefficient are the same in either case. We have for $r < a$

$$\mathscr{E}_z = \frac{iB\sigma}{2\beta} \left[J_{\nu+1}(\sigma r) \begin{Bmatrix} \cos (\nu + 1)\phi \\ \sin (\nu + 1)\phi \end{Bmatrix} \right.$$
$$\left. - J_{\nu-1}(\sigma r) \begin{Bmatrix} \cos (\nu - 1)\phi \\ \sin (\nu - 1)\phi \end{Bmatrix} \right] \tag{58}$$

$$\mathcal{E}_x = BJ_\nu(\sigma r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \tag{59}$$

$$\mathcal{H}_z = \frac{iB\sigma}{2k} \sqrt{\frac{\epsilon_o}{\mu_o}} \left\{ J_{\nu+1}(\sigma r) \begin{bmatrix} \sin (\nu + 1)\phi \\ -\cos (\nu + 1)\phi \end{bmatrix} \right.$$
$$\left. + J_{\nu-1}(\sigma r) \begin{bmatrix} \sin (\nu - 1)\phi \\ -\cos (\nu - 1)\phi \end{bmatrix} \right\} \tag{60}$$

$$\mathcal{H}_y = nB \frac{\beta}{|\beta|} \sqrt{\frac{\epsilon_o}{\mu_o}} J_\nu(\sigma r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix}. \tag{61}$$

It remains to relate the amplitude coefficient to the power P. Because of the continuous mode spectrum, it is not possible to normalize the radiation modes with a finite amount of power. The parameter P is defined by the relation

$$\frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{E}_\rho \times \mathcal{H}_{\rho'}^*) \cdot e_z dx dy = \frac{\beta^*}{|\beta|} P\delta(\rho - \rho'). \tag{62}$$

The amplitude coefficient B is thus [e_ν is defined by (36)]

$$B = \frac{\left(\dfrac{\mu_o}{\epsilon_o}\right)^{\frac{1}{4}}(8\rho P)^{\frac{1}{2}}}{\sqrt{e_\nu n}a\pi^{\frac{1}{2}}|\sigma J_{\nu-1}(\sigma a)H_\nu^{(1)}(\rho a) - \rho J_\nu(\sigma a)H_{\nu-1}^{(1)}(\rho a)|}. \tag{63}$$

It is important to remember that the radiation modes listed so far are valid only in the immediate vicinity of $\beta = \pm n_2 k$. Inside the β range, we use the radiation modes of homogeneous space with refractive index n_2. These modes can be expressed in a number of different ways. The simplest expressions would result from a plane-wave representation. However, for our present purposes, it seems advisable to use field expressions that resemble most closely the radiation modes (46) through (54) and (58) through (61) in order to achieve continuity of the field expressions throughout the entire β range. The modes of the homogeneous medium (in vacuum we would say free-space modes) are simpler than the radiation modes of the fiber, since one expression applies throughout all of space. There is no need to treat the fields inside and outside of the core separately. The field expressions that satisfy our requirements are [ρ is defined by (55), we use $n = n_2$]

$$\mathcal{E}_z = \frac{iC\rho}{2\beta} \left[J_{\nu+1}(\rho r) \begin{Bmatrix} \sin (\nu + 1)\phi \\ -\cos (\nu + 1)\phi \end{Bmatrix} \right.$$
$$\left. + J_{\nu-1}(\rho r) \begin{Bmatrix} \sin (\nu - 1)\phi \\ -\cos (\nu - 1)\phi \end{Bmatrix} \right] \tag{64}$$

$$\mathcal{E}_x = 0 \tag{65}$$

$$\mathscr{E}_y = C J_\nu(\rho r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \tag{66}$$

$$\mathscr{H}_z = -\frac{iC\rho}{2k}\sqrt{\frac{\epsilon_o}{\mu_o}}\left[J_{\nu+1}(\rho r) \begin{Bmatrix} \cos (\nu+1)\phi \\ \sin (\nu+1)\phi \end{Bmatrix} \right.$$
$$\left. - J_{\nu-1}(\rho r) \begin{Bmatrix} \cos (\nu-1)\phi \\ \sin (\nu-1)\phi \end{Bmatrix} \right] \tag{67}$$

$$\mathscr{H}_x = -\frac{C}{4}\sqrt{\frac{\epsilon_o}{\mu_o}}\left[2\left(\frac{\beta}{k}+\frac{n^2 k}{\beta}\right) J_\nu(\rho r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \right.$$
$$+ \left(\frac{n^2 k}{\beta}-\frac{\beta}{k}\right)\left(J_{\nu+2}(\rho r) \begin{Bmatrix} \cos (\nu+2)\phi \\ \sin (\nu+2)\phi \end{Bmatrix} \right.$$
$$\left.\left. + J_{\nu-2}(\rho r) \begin{Bmatrix} \cos (\nu-2)\phi \\ \sin (\nu-2)\phi \end{Bmatrix} \right) \right] \tag{68}$$

$$\mathscr{H}_y = \frac{C}{4}\sqrt{\frac{\epsilon_o}{\mu_o}}\left(\frac{\beta}{k}-\frac{n^2 k}{\beta}\right)\left[J_{\nu+2}(\rho r) \begin{Bmatrix} \sin (\nu+2)\phi \\ -\cos (\nu+2)\phi \end{Bmatrix} \right.$$
$$\left. - J_{\nu-2}(\rho r) \begin{Bmatrix} \sin (\nu-2)\phi \\ -\cos (\nu-2)\phi \end{Bmatrix} \right]. \tag{69}$$

The orthogonally polarized modes are

$$\mathscr{E}_z = \frac{iC\rho}{2\beta}\left[J_{\nu+1}(\rho r) \begin{Bmatrix} \cos (\nu+1)\phi \\ \sin (\nu+1)\phi \end{Bmatrix} - J_{\nu-1}(\rho r) \begin{Bmatrix} \cos (\nu-1)\phi \\ \sin (\nu-1)\phi \end{Bmatrix} \right] \tag{70}$$

$$\mathscr{E}_x = C J_\nu(\rho r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \tag{71}$$

$$\mathscr{E}_y = 0 \tag{72}$$

$$\mathscr{H}_z = \frac{iC\rho}{2k}\sqrt{\frac{\epsilon_o}{\mu_o}}\left[J_{\nu+1}(\rho r) \begin{Bmatrix} \sin (\nu+1)\phi \\ -\cos (\nu+1)\phi \end{Bmatrix} \right.$$
$$\left. + J_{\nu-1}(\rho r) \begin{Bmatrix} \sin (\nu-1)\phi \\ -\cos (\nu-1)\phi \end{Bmatrix} \right] \tag{73}$$

$$\mathscr{H}_x = -\frac{C}{4}\sqrt{\frac{\epsilon_o}{\mu_o}}\left(\frac{\beta}{k}-\frac{n^2 k}{\beta}\right)\left[J_{\nu+2}(\rho r) \begin{Bmatrix} \sin (\nu+2)\phi \\ -\cos (\nu+2)\phi \end{Bmatrix} \right.$$
$$\left. - J_{\nu-2}(\rho r) \begin{Bmatrix} \sin (\nu-2)\phi \\ -\cos (\nu-2)\phi \end{Bmatrix} \right] \tag{74}$$

$$\mathscr{H}_y = \frac{C}{4}\sqrt{\frac{\epsilon_o}{\mu_o}}\left[2\left(\frac{\beta}{k}+\frac{n^2 k}{\beta}\right) J_\nu(\rho r) \begin{Bmatrix} \cos \nu\phi \\ \sin \nu\phi \end{Bmatrix} \right.$$
$$+ \left(\frac{\beta}{k}-\frac{n^2 k}{\beta}\right)\left(J_{\nu+2}(\rho r) \begin{Bmatrix} \cos (\nu+2)\phi \\ \sin (\nu+2)\phi \end{Bmatrix} \right.$$
$$\left.\left. + J_{\nu-2}(\rho r) \begin{Bmatrix} \cos (\nu-2)\phi \\ \sin (\nu-2)\phi \end{Bmatrix} \right) \right]. \tag{75}$$

The amplitude coefficient C is related to P

$$C = \left\{ \frac{4 \sqrt{\frac{\mu_o}{\epsilon_o}} k\rho\beta P}{e_\nu \pi (\beta^2 + n^2 k^2)} \right\}^{\frac{1}{2}}. \tag{76}$$

The modes of the homogeneous medium, like all the other modes, are mutually orthogonal among each other. With the help of the relations

$$J_\nu(x) H^{(1)}_{\nu+1}(x) - J_{\nu+1}(x) H^{(1)}_\nu(x) = \frac{2}{i\pi x} \tag{77a}$$

and

$$J_\nu(x) H^{(2)}_{\nu+1}(x) - J_{\nu+1}(x) H^{(2)}_\nu(x) = -\frac{2}{i\pi x} \tag{77b}$$

it is easy to show that the radiation modes of the homogeneous medium and the radiation modes of the round fiber reduce to the same expressions in the limit $|\beta| = n_2 k$, $n_1 = n_2 = n$.

V. COUPLING COEFFICIENTS FOR CORE-CLADDING IMPERFECTIONS

We consider a fiber whose core-cladding interface is described by the function

$$r(x, y, z) = a + f(z) \cos (m\phi + \psi). \tag{78}$$

If we choose a different function $f(z)$ and different phase ψ for each integer m and sum over the second term on the right side, we generate a Fourier series which allows us to describe core-cladding imperfections of the most general kind.

We assume that a given mode labeled ν is traveling in the waveguide and ask for the coupling from this mode to all other guided and radiation modes. The function $f(z)$ can be separated out from the coupling coefficient by defining

$$K_{\mu\nu} = \bar{K}_{\mu\nu} f(z). \tag{79}$$

We have mentioned earlier that the longitudinal field components of the guided modes are much smaller than their transverse components. The same statement is true for the radiation modes only if their propagation constant β is very nearly equal to $\pm n_2 k$. For $|\beta|$ values much smaller than $n_2 k$, the longitudinal field components of the radiation modes can be as large as or larger than the transverse components. The coupling coefficients contain scalar products of the two fields that are coupled together. Coupling coefficients that involve at least one guided mode are thus determined primarily by the transverse component of both fields, since the product of the longitudinal components is small over most of the range of β values. We thus neglect the contribu-

tion of the longitudinal components and gain the advantage of much simpler expressions for the coupling coefficients. The only region of the β range where the longitudinal components could make a significant contribution to the coupling process is near $\beta = 0$. Outside of a small region near $\beta = 0$, our approximation is reasonably accurate. To this approximation, modes with orthogonal (transverse) polarization are not coupled by core-cladding imperfections of the form (78).

For coupling between two guided modes ν and μ we obtain from (30) and (79) with the help of the field expressions

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{e_{\mu\nu m}}{\sqrt{e_\mu e_\nu}} \frac{p\gamma_\nu\gamma_\mu J_\nu(\kappa_\nu a)J_\mu(\kappa_\mu a)}{2iank[\,|\,J_{\nu-1}(\kappa_\nu a)J_{\nu+1}(\kappa_\nu a)J_{\mu-1}(\kappa_\mu a)J_{\mu+1}(\kappa_\mu a)\,|\,]^{\frac{1}{2}}}. \tag{80}$$

The factor $e_{\mu\nu m}$ can assume the values 4, 2, 1, or 0. It is zero unless $\mu = \nu \pm m$. Table I shows the values of $e_{\mu\nu m}$ for all possible cases. The factor e_ν is defined by (44), κ_ν and γ_ν are determined by (37) and (38).

Coupling between the guided mode ν and radiation modes μ must be described by two different coupling coefficients depending on the value of the propagation constant β of the radiation mode. For values of $|\beta|$ close to n_2k, we use the radiation modes of the fiber and obtain

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{e_{\mu\nu m}}{\sqrt{e_\nu e_\mu}}$$

$$\times \frac{p\left(\dfrac{n_1}{n_2} - 1\right)^{\frac{1}{2}}\gamma_\nu\sqrt{\rho}J_\nu(\kappa_\nu a)J_\mu(\sigma a)}{i\pi a\,|\,J_{\nu-1}(\kappa_\nu a)J_{\nu+1}(\kappa_\nu a)\,|^{\frac{1}{2}}\,|\,\sigma J_{\mu-1}(\sigma a)H_\mu^{(1)}(\rho a) - \rho J_\mu(\sigma a)H_{\mu-1}^{(1)}(\rho a)\,|}. \tag{81}$$

For β values inside the range $-n_2k < \beta < n_2k$ excluding the end points, we use the radiation modes of homogeneous space and find

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{e_{\mu\nu m}}{\sqrt{e_\nu e_\mu}} \frac{p\left[n\left(\dfrac{n_1}{n_2} - 1\right)k\rho\beta\right]^{\frac{1}{2}}\gamma_\nu J_\nu(\kappa_\nu a)J_\mu(\rho a)}{i[2(\beta^2 + n^2k^2)\,|\,J_{\nu-1}(\kappa_\nu a)J_{\nu+1}(\kappa_\nu a)\,|\,]^{\frac{1}{2}}}. \tag{82}$$

Use of (77) allows us again to see that (81) and (82) become identical in the limit $n_1 = n_2 = n$, $|\beta| = n_2k$.

The coupling coefficients for the approximate guided and radiation modes of the round optical fiber allow us to solve a large number of problems involving fibers with core-cladding interface irregularities.

VI. FAR-FROM-CUTOFF APPROXIMATIONS

For purposes of multimode operation it is often desirable to have simple expressions for the coupling coefficients which are valid far from

TABLE I—TABULATION OF THE FACTOR $e_{\mu\nu m}$ FOR ALL POSSIBLE COMBINATIONS OF ANGULAR FIELD DEPENDENCE OF THE MODES AND THE CORE-CLADDING INTERFACE DISTORTION

$e_{\mu\nu m} = 0$ unless specified otherwise		
incident mode $\cos \nu\phi$	spurious mode $\cos \mu\phi$	distortion $\cos m\phi$
$e_{\mu\nu m} = \begin{cases} 4 & \nu = \mu = m = 0 \\ 2 & \begin{cases} \nu = 0, & \mu = m \\ \mu = 0, & \nu = m \end{cases} \\ 1 & \begin{cases} 0 < \mu = \nu \pm m \\ 0 < \mu = m - \nu \end{cases} \end{cases}$		
incident mode $\sin \nu\phi$	spurious mode $\sin \mu\phi$	distortion $\cos m\phi$
$e_{\mu\nu m} = \begin{cases} 0 & \nu \text{ or } \mu = 0 \\ +1 & 0 < \mu = \nu \pm m \\ -1 & 0 < \mu = m - \nu \end{cases}$		
incident mode $\cos \nu\phi$	spurious mode $\sin \mu\phi$	distortion $\sin m\phi$
$e_{\mu\nu m} = \begin{cases} 0 & \mu = 0 \\ 2 & \nu = 0, \quad \mu = m \\ 1 & \mu = \nu + m \\ -1 & 0 < \mu = \nu - m \\ 1 & 0 < \mu = m - \nu \end{cases}$		
incident mode $\cos \nu\phi$	spurious mode $\sin \mu\phi$ $e_{\mu\nu m} = 0$	distortion $\cos m\phi$
incident mode $\cos \nu\phi$	spurious mode $\cos \mu\phi$ $e_{\mu\nu m} = 0$	distortion $\sin m\phi$
incident mode $\sin \nu\phi$	spurious mode $\sin \mu\phi$ $e_{\mu\nu m} = 0$	distortion $\sin m\phi$

cutoff. We obtain such approximations by using the approximation for large argument of the Hankel function ($\gamma a \gg 1$)

$$H_\nu^{(1)}(i\gamma a) \approx \sqrt{\frac{2}{i\pi\gamma a}} \, e^{-i[\pi/4 + \nu(\pi/2)]} e^{-\gamma a}. \tag{83}$$

With (83), we obtain from (45) for $\gamma a \gg 1$

$$\gamma J_\nu(\kappa a) = -\kappa J_{\nu-1}(\kappa a). \tag{84}$$

We remarked earlier that (45) is also valid if $\nu - 1$ is replaced with $\nu + 1$. We thus also have

$$\gamma J_\nu(\kappa a) = \kappa J_{\nu+1}(\kappa a). \tag{85}$$

Multiplying (84) with (85) and taking the square root results in

$$\gamma J_\nu(\kappa a) = \kappa \big[-J_{\nu-1}(\kappa a) J_{\nu+1}(\kappa a) \big]^{\frac{1}{2}}$$
$$= \kappa \big[\, | J_{\nu-1}(\kappa a) J_{\nu+1}(\kappa a) | \, \big]^{\frac{1}{2}}. \tag{86}$$

Far from cutoff, (86) allows us to write the coupling coefficient (80) between guided modes in the simple form

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{p e_{\mu\nu m}}{\sqrt{e_\nu e_\mu}} \frac{\kappa_\mu \kappa_\nu}{2 i a n k}. \tag{87}$$

Very far from cutoff, $\gamma a \to \infty$, we obtain the approximate eigenvalue equation

$$J_\nu(\kappa_\nu a) = 0 \tag{88}$$

from (84) or (85). The roots of the Bessel function $J_\nu(\kappa_\nu a)$ thus determine the values of $\kappa_\nu a$ that appear in (87). For higher-order modes $\kappa_\nu a$ becomes large so that we can approximate

$$J_\nu(\kappa_\nu a) \approx \sqrt{\frac{2}{\pi \kappa_\nu a}} \cos \left[\kappa_\nu a - \left(\nu + \frac{1}{2} \right) \frac{\pi}{2} \right]. \tag{89}$$

Equation (88) requires that the argument of the cosine function equal $(2N + 1)\pi/2$. This leads to a direct determination of

$$\kappa_\nu a \approx \left(\nu + 2N + \frac{3}{2} \right) \frac{\pi}{2} \tag{90}$$

with $N = 0, 1, 2 \cdots$. The equations (87) and (90) provide us with an approximate determination of the coupling coefficient between two guided modes without the need for solving a transcendental eigenvalue equation. In a strict sense, we would have to label κ and the coupling coefficient with N as well as ν. We refrain from burdening the symbols with too many indices.

The coupling coefficients between guided and radiation modes can similarly be simplified far from cutoff of the guided modes. The far-from-cutoff approximation of (81) is

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{e_{\mu\nu m}}{\sqrt{e_\nu e_\mu}} \frac{p \left(\dfrac{n_1}{n_2} - 1 \right)^{\frac{1}{2}} \sqrt{\rho} \kappa_\nu J_\mu(\sigma a)}{i \pi a \, | \, \sigma J_{\mu-1}(\sigma a) H_\mu^{(1)}(\rho a) - \rho J_\mu(\sigma a) H_{\mu-1}^{(1)}(\rho a) \, |}. \tag{91}$$

This equation is valid only for $|\beta|$ values very close to $n_2 k$. The coefficient that describes coupling between guided modes and the radiation modes of homogeneous space, eq. (82), leads to the far-from-cutoff

approximation

$$\bar{K}_{\mu\nu}^{(p,q)} = \frac{e_{\mu\nu m}}{\sqrt{e_\nu e_\mu}} \frac{p\left[n\left(\frac{n_1}{n_2} - 1\right)k\rho\beta\right]^{\frac{1}{2}}\kappa_\nu J_\mu(\rho a)}{i[2(\beta^2 + n^2k^2)]^{\frac{1}{2}}}. \tag{92}$$

This expression is valid inside the range $-n_2k < \beta < n_2k$ but not near $|\beta| = n_2k$. For most scattering problems, it is sufficient to use the coupling coefficients to the radiation modes of the homogeneous medium. Only if the scattering is sharply forward or backward directed do we have to use the slightly more complicated expression (91) of the coupling coefficients to the true radiation modes of the round fiber.

In the remainder of the paper, we apply our results to special cases.

VII. COUPLING CAUSED BY WAVEGUIDE BENDS

We consider the case of a straight fiber that is connected to a fiber section that is bent with a constant radius of curvature and finally continues in a straight section. If the curved piece of waveguide causes considerable mode conversion, the system of coupled equations (27) and (28) must be solved. However, for slight mode conversion, we can use the approximation that the incident mode does not change very much while power builds up in some of the spurious modes. In this case, we obtain the following approximate solution from (27):

$$c_\mu(L) = c_\nu(0)\int_0^L K_{\mu\nu}(z)e^{i(\beta_\mu - \beta_\nu)z}dz. \tag{93}$$

We assume $z = 0$ at the beginning of the curved section of length L. The description (32) is most appropriate in this case. The second derivative assumes the constant value $1/R$, with R being the radius of curvature of the circular bend. We thus obtain from (79) [with f replaced by (32)] and (93)

$$\left|\frac{c_\mu(L)}{c_\nu(0)}\right|^2 = \frac{1}{R^2}\frac{4|\bar{K}_{\mu\nu}|^2}{(\beta_\mu - \beta_\nu)^6}\sin^2(\beta_\mu - \beta_\nu)\frac{L}{2}. \tag{94}$$

We obtain $\bar{K}_{\mu\nu}$ from (80) or in its "far-from-cutoff" approximation from (87). The integer m appearing in (78) must be set $m = 1$ in this case, since we want to describe a continuous offset of the fiber which results in a bend. It is apparent that the amount of power transfer between the incident and the spurious mode depends critically on the separation between the two propagation constants. The sine factor in (94) describes the phasing between the two modes. If the incident and spurious modes travel with equal phase velocity, power would be

transferred from the incident mode to the spurious mode in proper phase so that all the power could be exchanged between the modes. If both modes have different phase velocities (and our formula holds only in this case), the two modes get out of step so that after some distance the power that is fed from the incident mode to the spurious mode tends to interfere destructively and destroys the power that has already been transferred. The power in the spurious mode thus builds up and decays. This process does not involve reconversion of power from the spurious to the incident mode, since this process is not included in our perturbation solution.

It can be shown that a guided mode of the type (34) produces a ring-shaped far-field pattern if it is allowed to radiate out of the end of the fiber. The maximum of the ring as seen from the end of the fiber appears at an angle

$$\theta_\nu = \frac{\kappa_\nu}{k}. \tag{95}$$

The circular far-field pattern on a screen is broken up into 2ν bright dots corresponding to the angular intensity maxima of the field distribution in the fiber. The angle θ_ν is useful to distinguish guided modes experimentally. It may thus be of interest to express (94) in terms of this mode angle. Using (87), (95), and

$$\beta_\nu \approx n_1 k - \frac{\kappa_\nu^2}{2 n_1 k} = n_1 k - \frac{k}{2 n_1} \theta_\nu^2, \tag{96}$$

we can write (94) in the following form:

$$\left| \frac{c_\mu}{c_\nu} \right|^2 = \frac{e_{\mu\nu 1}^2}{e_\nu e_\mu} \frac{2^6 n^4 \theta_\nu^2 \theta_\mu^2}{k^4 a^2 R^2 (\theta_\nu^2 - \theta_\mu^2)^6} \sin^2(\beta_\mu - \beta_\nu) \frac{L}{2}. \tag{97}$$

For all practical applications, the separation between the angles θ_ν and θ_μ is so small that we can replace them with one angle θ. According to (90) we get for the difference

$$\theta_\nu - \theta_\mu = \Delta\theta = \pm \frac{\pi}{2ka} \tag{98}$$

if $\mu = \nu \pm 1$. We restrict the discussion to coupling between modes with the same value of N but adjacent ν values. If the results of this discussion are implemented, (97) assumes the form

$$\left| \frac{c_\mu}{c_\nu} \right|^2 = \frac{e_{\mu\nu 1}^2}{e_\nu e_\mu} \frac{2^6 k^2 a^4 n^4}{\pi^6 R^6 \theta^2} \sin^2(\beta_\mu - \beta_\nu) \frac{L}{2}. \tag{99}$$

This equation shows that the amount of power transfer caused by waveguide bends decreases with increasing mode angle.

Next we consider random bends. It has been shown in Refs. 3 and 17 that the exchange of power among randomly coupled modes can be described by coupled power equations. The power coupling coefficient is given by

$$h_{\mu\nu} = |\bar{K}_{\mu\nu}|^2 F(\beta_\mu - \beta_\nu). \tag{100}$$

$F(\beta_\mu - \beta_\nu)$ is the power spectrum of the coupling function $f(z)$ or of its equivalents (31) or (32). Using (32) and writing $1/R$ instead of $d^2 f/dz^2$, we obtain for the power coupling coefficient for random bends

$$h_{\mu\nu} = \frac{e_{\mu\nu1}^2}{e_\nu e_\mu} \frac{\kappa_\nu^2 \kappa_\mu^2}{4a^2 n^2 k^2 (\beta_\mu - \beta_\nu)^4} C\left(\frac{1}{R}\right) \tag{101}$$

with the power spectrum of the curvature function

$$C\left(\frac{1}{R}\right) = \left\langle \left| \frac{1}{\sqrt{L}} \int_0^L \frac{1}{R(z)} e^{i(\beta_\mu - \beta_\nu)z} dz \right|^2 \right\rangle. \tag{102}$$

The symbol $\langle \ \rangle$ indicates an ensemble average.

The guided modes suffer radiation losses even in uniformly bent waveguide sections, $R(z) = \text{const.}$[18,19] These curvature losses cannot be obtained by perturbation theory and thus are not included in our discussion. However, our perturbation theory includes mode conversion losses between guided modes and radiation losses caused by changes in the waveguide curvature.

In terms of the mode angle θ, (102) can be written in the form

$$h_{\mu\nu} = \frac{e_{\mu\nu1}^2}{e_\nu e_\mu} \frac{4n^2 k^2 a^2}{\pi^4} C\left(\frac{1}{R}\right). \tag{103}$$

The dependence on the mode angle is contained only in the power spectrum of the curvature function.

With the help of the coupling coefficients (91) and (92), radiation losses caused by random bends can be calculated. However, the explicit expression will not be given here.

VIII. MODE CONVERSION AND LOSSES DUE TO DISTORTED CORE-CLADDING INTERFACES

Instead of writing down the general formula for the power coupling coefficient between guided modes based on (80), we restrict ourselves to the far-from-cutoff approximation. From (87), (95), and (100) we

obtain

$$h_{\mu\nu} = \frac{e_{\mu\nu m}^2}{e_\nu e_\mu} \frac{k^2 \theta_\mu^2 \theta_\nu^2}{4a^2 n^2} F(\beta_\mu - \beta_\nu) \tag{104}$$

with

$$F(\beta_\mu - \beta_\nu) = \left\langle \left| \frac{1}{\sqrt{L}} \int_0^L f(z) e^{i(\beta_\mu - \beta_\nu)z} dz \right|^2 \right\rangle. \tag{105}$$

$f(z)$ is the function that appears in (78). Equation (104) is remarkably similar to eq. (37) of Ref. 3 which was derived for the slab waveguide model. For comparison of the two equations, it is necessary to remember that we have assumed that the angles θ_ν and θ_μ are much smaller than unity and that $\gamma d \gg 1$ according to our far-from-cutoff approximation. The difference in the position of n in the two equations is attributable to the different definitions of the mode angles. In our present discussion, we consider θ_ν as the angle of the far-field radiation cone outside of the fiber, while this angle was defined as the angle of the plane waves inside of the fiber core in Ref. 3. The correspondence between the two coupling coefficients requires us to consider the case of pure diameter changes, $m = 0$, and assume that both modes have no circumferential variation, $\nu = 0$ and $\mu = 0$. In this case, we have $e_{ooo}^2/(4e_o e_o) = 1$ instead of the factor $1/2$ appearing in (37) of Ref. 3. The difference is accounted for if we remember that the round fiber corresponds to a slab in which both interfaces have irregularities which are perfectly correlated. The slab waveguide theory of Ref. 3 assumed, on the other hand, that the two interfaces had uncorrelated irregularities. This comparison shows that the results of the slab waveguide theory and the round fiber are in very good agreement. Our present formula (104) holds for core-cladding interface irregularities of a much more general kind. Not only pure diameter changes but elliptical deformations and deformations of even more general shapes are included.

Next, we turn to the problem of radiation losses. The power loss coefficient is defined by [compare (9.3–14) and (9.3–42) of Ref. 1]

$$a_\nu = \sum_\mu \int_{-n_2 k}^{n_2 k} |\bar{K}_{\mu\nu}|^2 F(\beta - \beta_\nu) \frac{|\beta|}{\rho} d\beta \tag{106}$$

with $\bar{K}_{\mu\nu}$ being the coefficient for coupling between a guided mode ν and a radiation mode with angular symmetry μ and propagation constant β. The power spectrum F is defined by (105) with β_μ replaced by β. Not much can be gained from substituting (81) and (82) or their far-from-cutoff approximations (91) and (92) into (106). The integral

in (106) is hard to solve and useful approximations covering the whole range of correlation lengths have not yet been found. However, using, for example, (81) and (82) in (106) simplifies the numerical integration compared to the problem discussed in Ref. 7. The radiation losses can be calculated from (81), (82), and (106) with much simpler computer programs and at considerable savings compared to the theory of Ref. 7. We use an exponential correlation function

$$R(u) = \langle f(z)f(z + u) \rangle = \bar{\sigma}^2 \exp\left(-|u|/B\right) \tag{107}$$

and obtain (Ref. 1, p. 371)

$$F(\beta - \beta_\nu) = \frac{2\bar{\sigma}^2}{B\left[(\beta - \beta_\nu)^2 + \dfrac{1}{B^2}\right]}. \tag{108}$$

The resulting radiation losses for pure diameter changes, $m = 0$, are plotted for the HE_{11} mode, $\nu = 0$, in Fig. 2 as functions of the ratio of correlation length B over core radius a for $n_1/n_2 = 1.01$. The curves were obtained by numerical integration of (106) with $\bar{K}_{\mu\nu}$ of (81) in the range $0.95n_2k \leqq |\beta| \leqq n_2k$ and with $\bar{K}_{\mu\nu}$ of (82) in the range $-0.95n_2k < \beta < 0.95n_2k$. For small index differences between core
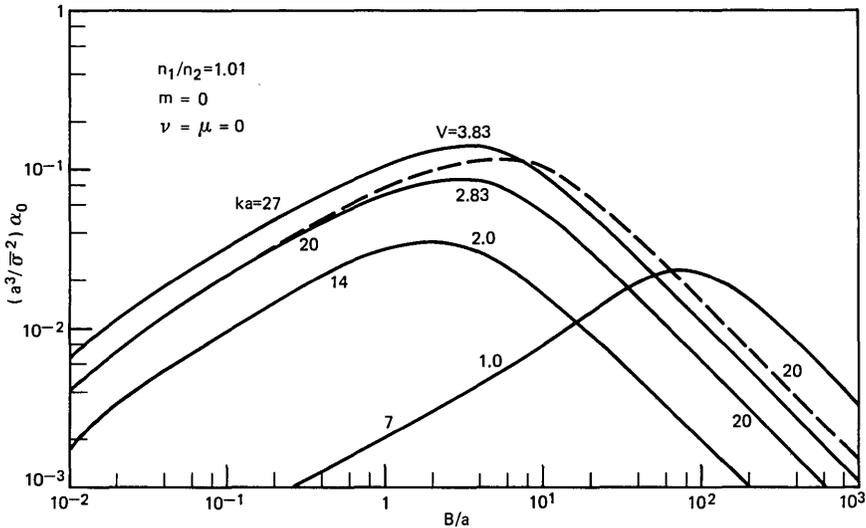

Fig. 2—Normalized radiation losses of the HE_{11} mode, $\nu = 0$, as functions of the ratio of correlation length B to core radius a for different values of $ka = 2\pi a/\lambda$ for pure diameter changes, $m = 0$. $n_1/n_2 = 1.01$. The dotted line results from using only "free space" radiation modes.

and cladding, the losses calculated from our simplified theory are in perfect agreement with the theory of Ref. 7. For larger index ratios, our approximate theory begins to fail. For $n_1/n_2 = 1.43$, the error caused by our approximation is in the order of 60 percent. The simplification gained from using (81), (82), and (106) is apparent by glancing at the complex formulas of Ref. 7.

The dotted line in Fig. 2 was computed by using the radiation modes of homogeneous space alone so that (82) instead of a combination of (81) and (82) was used in (106). It is apparent that the radiation modes of homogeneous space are not suitable to calculate the radiation losses for large values of B/a. It was pointed out in Ref. 5 that large B/a ratios lead to forward scattering. The radiation makes small angles with the core-cladding interface so that reflection at this interface becomes important. It is thus necessary to use the radiation modes of the fiber for β values near $n_2 k$.

Our theory allows us to calculate radiation losses for more general core-cladding interface distortions. As a second example, we consider elliptical deformation, $m = 2$, and plot the result of the numerical integration of (106) in Fig. 3. The power spectrum (108) of the function $f(z)$ [defined by (78)] was used again. We also used a combination of radiation modes of the fiber and of free-space radiation modes in the same way as indicated before. The radiation losses caused by elliptical
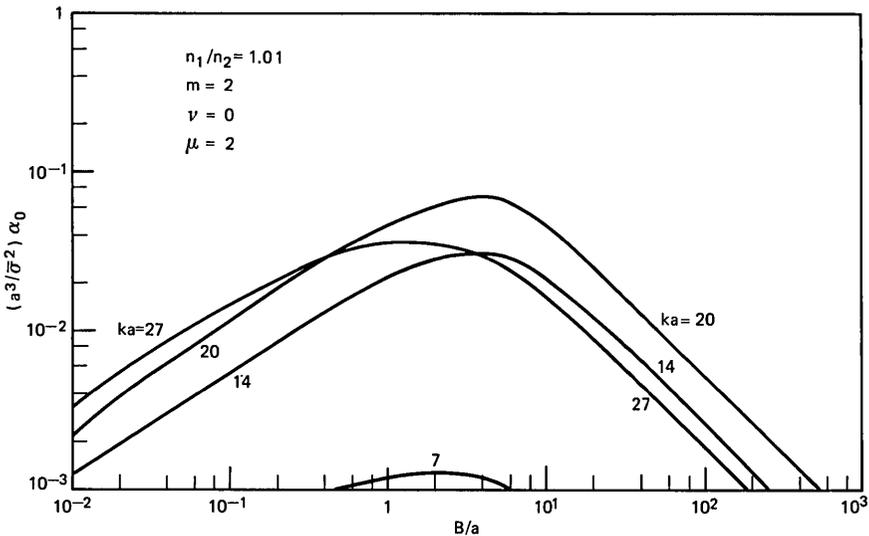

Fig. 3—HE_{11} mode radiation losses caused by elliptical deformations, $m = 2$, of the core-cladding boundary. $n_1/n_2 = 1.01$.

core-cladding interface irregularities are smaller than those of pure diameter changes.

Actual numerical values of radiation losses obtained from curves like Fig. 2 and Fig. 3 were discussed in previous publications.[1,5,7]

IX. CONCLUSIONS

We have presented a simplified theory of mode coupling in imperfect round optical fibers. The simplification was a result of restricting the discussion to fibers with small values of $n_1/n_2 - 1$. The simplified theory results in much simpler expressions for the guided and radiation modes of the fiber and consequently leads to simple expressions for the coupling coefficients. For small core-cladding index differences, the simplified theory is in excellent agreement with more general theories.

The principal contribution of this paper is a tabulation of coupling coefficients for coupling between guided and radiation modes that are necessary for solving mode coupling problems caused by general core-cladding interface imperfections. A general coupling theory based on the modes of the ideal fiber is also presented.[20]

REFERENCES

1. Marcuse, D., *Light Transmission Optics*, New York: Van Nostrand Reinhold Company, 1972.
2. Personick, S. D., "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
3. Marcuse, D., "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1199–1232.
4. Snyder, A. W., "Excitation and Scattering of Modes on a Dielectric or Optical Fiber," IEEE Trans. Microwave Theory and Techniques, *MTT-17*, No. 12 (December 1969), pp. 1138–1144.
5. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3187–3215.
6. Rawson, E. G., "Analysis of Scattering from Fiber Waveguides with Irregular Core Surfaces," 1972 Annual Meeting of the Optical Society of America, 17–20 October, 1972, San Francisco.
7. Marcuse, D., "Radiation Losses of the Dominant Mode in Round Dielectric Waveguides," B.S.T.J., *49*, No. 8 (October 1970), pp. 1665–1693.
8. Snyder, A. W., "Asymptotic Expressions for Eigenfunctions and Eigenvalues of a Dielectric or Optical Waveguide," IEEE Trans. Microwave Theory and Techniques, *MTT-17*, No. 12 (December 1969), pp. 1130–1138.
9. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, No. 10 (October 1971), pp. 2252–2258.
10. Snyder, A. W., "Coupling of Modes on a Tapered Dielectric Cylinder," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 7 (July 1970), pp. 383–392.
11. Snyder, A. W., "Mode Propagation in a Nonuniform Cylindrical Medium," IEEE Trans. Microwave Theory and Techniques, *MTT-19*, No. 4 (April 1971), pp. 402–403.
12. Marcuse, D., "Coupling Coefficients for Imperfect Asymmetric Slab Waveguides," B.S.T.J., *52*, No. 1 (January 1973), pp. 63–82.
13. Ref. 1, p. 348.

14. Marcuse, D., and Mammel, W. L., "Tube Waveguide for Optical Transmission," B.S.T.J., *52*, No. 3 (March 1973), pp. 423–435.
15. Ref. 1, p. 338.
16. Snyder, A. W., "Continuous Mode Spectrum of a Circular Dielectric Rod," IEEE Trans. Microwave Theory and Techniques, *MTT-19*, No. 8 (August 1971), pp. 720–727.
17. Marcuse, D., "Derivation of Coupled Power Equations," B.S.T.J., *51*, No. 1 (January 1972), pp. 229–237.
18. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2103–2132.
19. Marcuse, D., "Bending Losses of the Asymmetric Slab Waveguide," B.S.T.J., *50*, No. 8 (October 1971), pp. 2551–2563.
20. Snyder, A. W., "Coupled Mode Theory for Optical Fibers," J. Opt. Soc. Am., *62*, No. 11 (November, 1972) pp. 1267–1277.

Receiver Design for Digital Fiber Optic Communication Systems, I

By S. D. PERSONICK

(Manuscript received January 15, 1973)

This paper is concerned with a systematic approach to the design of the "linear channel" of a repeater for a digital fiber optic communication system. In particular, it is concerned with how one properly chooses the front-end preamplifier and biasing circuitry for the photodetector; and how the required power to achieve a desired error rate varies with the bit rate, the received optical pulse shape, and the desired baseband-equalized output pulse shape.

It is shown that a proper front-end design incorporates a high-impedance preamplifier which tends to integrate the detector output. This must be followed by proper equalization in the later stages of the linear channel. The baseband signal-to-noise ratio is calculated as a function of the preamplifier parameters. Such a design provides significant reduction in the required optical power and/or required avalanche gain when compared to a design which does not integrate initially.

It is shown that, when the received optical pulses overlap and when the optical channel is behaving linearly in power,[1] baseband equalization can be used to separate the pulses with a practical but significant increase in required optical power. This required power penalty is calculated as a function of the input and equalized pulse shapes.

I. INTRODUCTION

The purpose of this paper is to provide insight into a systematic approach to designing the "linear channel" of a repeater for a digital fiber optic communication system.

In particular, we are interested in how one properly chooses the biasing circuitry for the photodetector; and how the required power to achieve a desired error rate varies with the bit rate, the received optical pulse shape, and desired baseband output pulse shape.

Throughout this paper, performance will be measured in terms of signal-to-noise ratios. Efforts to calculate exact error rates and bounds

to error rates are difficult to carry out, and, in the past, the results of such efforts have shown little deviation (for practical design purposes) from calculations of error rates using the signal-to-noise ratio (Gaussian approximation) approach. (See Refs. 2 through 5 and Appendix A.)

II. INPUT–OUTPUT RELATIONSHIPS FOR AN AVALANCHE DETECTOR

An avalanche photodiode is the device of interest in fiber applications for converting optical power into current for amplification and equalization, ultimately to produce a baseband voltage for regeneration.

In order to appreciate its performance in practical optical systems, we have to characterize the avalanche photodiode from three points of view: the physical viewpoint, the circuit viewpoint, and the statistical viewpoint.

When we study the device from the physical viewpoint, we ask how does it operate, how do we develop circuit and statistical models of its operation, and what are the limitations of the models.

From the circuit viewpoint, we investigate how to design a piece of equipment in which the device will perform some function.

From the statistical viewpoint, we investigate the probabilistic behavior of the device to allow us to quantify its performance in a circuit.

2.1 The Physical Viewpoint

The avalanche photodiode is a semiconductor device which is normally operated in a backbiased manner–producing a region within the device where there is a high field (see Fig. 1). Due to thermal agitation and/or the presence of incident optical power, pairs of holes and electrons can be generated at various points within the diode. These carriers drift toward opposite ends of the device under the influence of the applied field. When a carrier passes through the high-field region, it may gain sufficient energy to generate one or more new pairs of holes and electrons through collision ionization. These new pairs can in turn generate additional pairs by the same mechanism. Carriers accumulate at opposite ends of the diode, thereby reducing the potential across the device until they are removed by the biasing and other circuitry in parallel with the diode (see Fig. 2). The chances that a carrier will generate a new pair when passing through the high-field region depends upon the type of carrier (hole or electron), the material out of which the diode is constructed, and the voltage across the device. To the extent that carriers do not accumulate to significantly modulate the
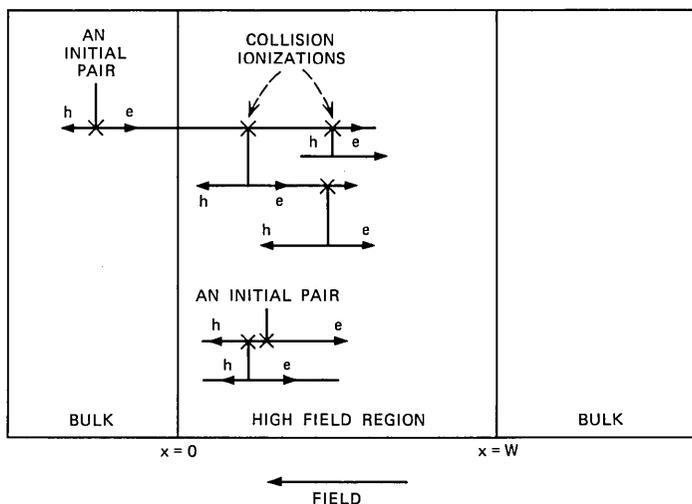
Fig. 1—Avalanche detector.

voltage across the device, it can be assumed that all ionizing collisions are statistically independent. This assumption also requires that the mean time between ionizing collisions be large compared to the time it takes for a carrier in the high-field region to randomize its momentum.

2.2 The Circuit Viewpoint

From the discussion above, and of course more detailed investigations,[3,6-8] it has been concluded that a reasonable small-signal model of an avalanche photodiode with a biasing circuit shown in Fig. 2 is the equivalent circuit of Fig. 3. In Fig. 3, C_d is the junction capacitance of the diode[†] across which voltage accumulates when charges produced within the device separate under the influence of the bias field. The current generator $i(t)$ represents the production of charges (holes and electrons) by optical and thermal generation and collision ionization in the diode high-field region. In order to use the photodiode efficiently,
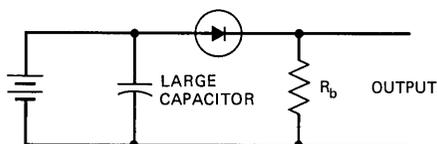
Fig. 2—Detector biasing circuit.

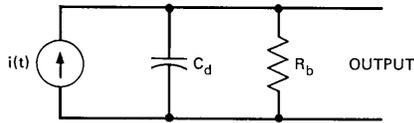[†] Not to be confused with the large power supply bypass capacitor of Fig. 2.

Fig. 3—Equivalent circuit of biased detector.

we must design a circuit which will respond to the current $i(t)$ with as little distortion and added noise as possible.

In order to derive information from the circuit responding to $i(t)$, we must understand the statistical relationship between $i(t)$ (the equivalent current generator) and the incident optical power $p(t)$.

2.3 The Statistical Viewpoint

In Fig. 3, the current source $i(t)$ can be considered to be a sequence of impulses corresponding to electrons generated within the photodiode due to optical or thermal excitation or collision ionization. We shall now specify, in a statistical way, how many electrons are produced and when they are produced.

From various physical studies,[3,7,9] it has been concluded that for cases of current interest the electron production process can be modeled as shown in Fig. 4.

Let the optical power falling upon the photon counter be $p(t)$.[†] In response to this power and due to thermal effects, the photon counter of Fig. 4 produces electrons at average rate $\lambda(t)$ per second where

$$\lambda(t) = [(\eta/\hbar\Omega)p(t)] + \lambda_0, \tag{1}$$

where

$$\eta = \text{photon counter quantum efficiency}$$
$$\hbar\Omega = \text{energy of a photon}$$
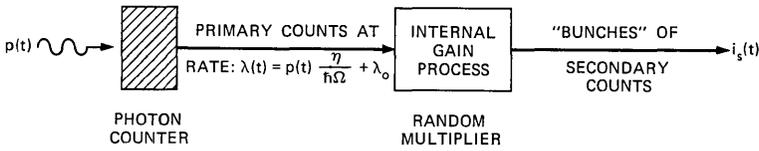$$\lambda_0 = \text{dark current ``counts'' per second.}$$

$\lambda(t)$ is only the average rate at which electrons are produced. In any interval T seconds long, the probability that exactly N counts are produced is given by

$$P[N, (t_0, t_0 + T)] = \frac{\Lambda^N e^{-\Lambda}}{N!},$$

where

$$\Lambda = \int_{t_0}^{t_0+T} \lambda(t)dt. \tag{2}$$

[†] The reader is cautioned not to confuse $p(t)$, the optical power, with the probability densities (e.g., $P[N, \{t_k\}]$) in this paper.

Fig. 4—Model of $i_s(t)$ generation process.

Given $p(t)$, the number of electrons produced in any interval is statistically independent of the number produced in any other disjoint interval.

A process of impulse (electron) production satisfying (2) and the above independent increments condition is said to be a "Poisson impulse process" with arrival rate $\lambda(t)$.[10]

A useful equivalent description of the above process follows.

If T is an interval, the probability that exactly N electrons will be produced at the (approximate) times $t_1 \pm \frac{1}{2}\Delta$, $t_2 \pm \frac{1}{2}\Delta$, $\cdots$ $t_N \pm \frac{1}{2}\Delta$ where the widths Δ are very small is

$$P[N, \{t_k\}] = \{e^{-\Lambda} \prod_1^N [\lambda(t_k)\Delta]/N!\} + o(\Delta), \qquad (3)$$

where Λ is defined in (2) and $o(\Delta)$ is a term such that

$$\lim_{\Delta \to 0} \frac{o(\Delta)}{\Delta} = 0.$$

It is *important* to note that in (3) the times $\{t_k\}$ are *not* in order, that is, in (3) it is *not* necessarily true that $t_1 < t_2$, etc.

Each of the "primary" impulses (electrons) produced by the photon counter enters a random multiplier where, corresponding to collision ionization, it is replaced by g contiguous "secondary" impulses (electrons). The number g is governed by the statistics of the internal gain mechanism of the photodiode. Each primary impulse (electron) is "multiplied" in this manner by a value g which is statistically independent of the value g assigned to other primaries.

Thus the current leaving the photodiode consists of "bunches" of electrons, the number of electrons in the bunch being a random quantity having statistics to be described below. For applications of interest here, it will be assumed that all electrons in a bunch exit the photodiode at the time when the primary is produced. This implies that the duration of the photodiode response to a single primary hole-electron pair is very short compared to the response times of circuitry to be used with the photodiode.
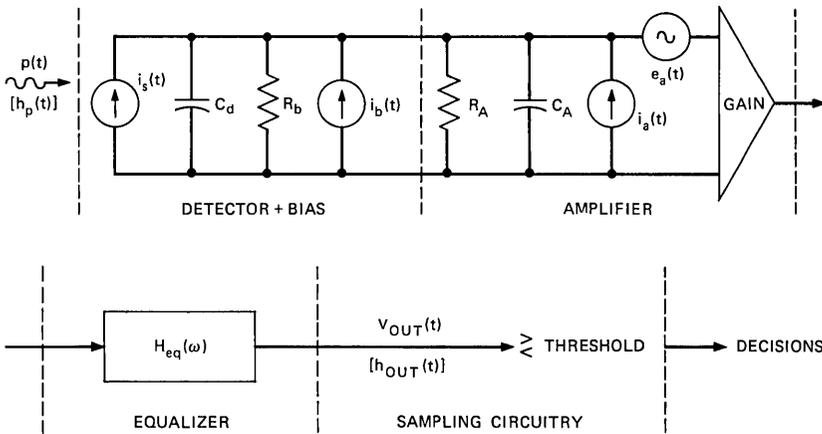
Fig. 5—Receiver.

Different avalanche photodiodes have different statistics governing the number of electrons in a bunch, i.e., the gain. For applications below, we will only need to know the mean gain $\langle g \rangle$ and the mean square gain $\langle g^2 \rangle$. For a large class of avalanche photodiodes of interest, it has been found that[3,7]

$$\langle g^2 \rangle \cong \langle g \rangle^{2+x}, \tag{4}$$

where $\langle g \rangle$ is determined by the applied bias voltage and x, a number usually between 0 and 1, depends upon the materials out of which the diode is constructed. For germanium photodiodes, $x \cong 1$; for well-designed silicon photodiodes, $x \approx 0.5$.

III. AN OPTICAL RECEIVER

Figure 5 shows a fairly typical receiver, in schematic form, consisting of an avalanche photodiode, an amplifier, and an equalizer.

The amplifier is modeled as an ideal high-gain infinite-impedance amplifier with an equivalent shunt capacitance and resistance at the input and with two noise sources referred to the input. For the purposes of this paper, the noise sources will be assumed to be white, Gaussian, and uncorrelated. Extensions to other amplifier models will be straightforward when the techniques of this paper are understood.[†]

It is assumed that the amplifier gain is sufficiently high so that noises introduced by the equalizer are negligible.

[†] With this model, the noise sources of the amplifier do not change when the input and output load circuitry changes.

The power falling upon the detector will be assumed to be of the form of a digital pulse stream

$$p(t) = \sum_{-\infty}^{\infty} b_k h_p(t - kT),\tag{5}$$

where b_k takes on one of two values for each integer value of k, $T =$ the pulse spacing, $h_p(t - kT) =$ pulse shape and is positive for all t. We shall assume $\int_{-\infty}^{\infty} h_p(t - kT)dt = 1$, therefore b_k is the energy in pulse k. The assumption that the received power will be in the form (5) appears reasonable for intensity modulation and fiber systems of interest.[1]

From (1) we have the average detector output current $\langle i_s(t) \rangle$ given by

$$\langle i_s(t) \rangle = \frac{\eta \langle g \rangle e p(t)}{\hbar \Omega} + \langle g \rangle e \lambda_0,$$

where

$\langle g \rangle =$ average detector internal gain

$e =$ electron charge

$\lambda_0 =$ dark current electrons per second

$\frac{\eta}{\hbar \Omega} p(t) =$ average optical primary electrons per second.

Therefore, the average voltage (neglecting dc components) at the equalizer output is

$$\langle v_{\text{out}}(t) \rangle = \frac{A \eta \langle g \rangle e p(t)}{\hbar \Omega} * h_{\text{fe}}(t) * h_{\text{eq}}(t),\tag{6}$$

where "*" indicates convolution and A is an arbitrary constant.

$$h_{\text{fe}}(t) = F \left\{ \frac{1}{\frac{1}{R_T} + j\omega(C_d + C_A)} \right\}$$

$$= \text{amplifier input circuit current impulse response,}$$

$$R_T = \left[\frac{1}{R_b} + \frac{1}{R_A} \right]^{-1} = \text{total detector parallel load resistance,}$$

and $h_{\text{eq}}(t) =$ equalizer impulse response.

Clearly, $\langle v_{\text{out}}(t) \rangle$ is of the form

$$\langle v_{\text{out}}(t) \rangle = \sum_{-\infty}^{\infty} b_k h_{\text{out}}(t - kT)\tag{7}$$

and $v_{out}(t)$ is of the form

$$v_{out}(t) = \sum_{-\infty}^{\infty} b_k h_{out}(t - kT) + n(t),$$

where $n(t)$ represents *deviations* (or noises) of $v_{out}(t)$ from its average.

The fundamental task ahead is to pick R_b (the bias circuit resistor) and $h_{eq}(t)$ so that a system which samples $v_{out}(t)$ at the times $\{kT\}$ can make decisions as to which value b_k has assumed (by comparing the sample to a threshold) with minimum chance of error.

IV. CALCULATING SIGNAL–TO–NOISE RATIO IN TERMS OF THE EQUALIZED PULSE SHAPE

Having defined the receiver and its statistics in the above sections, we can now calculate the variance of $n(t)$, the noise portion of the output $v_{out}(t)$ of the system of Fig. 5, defined as follows:

$$N = \langle (n(t))^2 \rangle = \langle v_{out}^2(t) \rangle - \langle v_{out}(t) \rangle^2. \tag{8}$$

The noise, N, of (8) above depends upon the coefficients $\{b_k\}$ defined in (5) and upon the time t.

We shall first of all restrict consideration to the set of times $t = \{kT\}$ when a decision as to the values $\{b_k\}$ will be made by sampling $v_{out}(t)$. We shall next assume that the equalized pulses satisfy

$$\begin{aligned} h_{out}(0) &= 1 \\ h_{out}(t) &= 0 \quad \text{for} \quad t = kT, \quad k \neq 0. \end{aligned} \tag{9}$$

That is, we shall assume that the equalized pulse stream has no inter-symbol interference at the sampling times kT.[†] Therefore,

$$v_{out}(kT) = b_k + n(kT). \tag{10}$$

In eq. (10) the noise, $n(t)$, still depends upon all the $\{b_k\}$ and the time t. This is a property which distinguishes fiber optic systems from many other systems where the noise is signal-independent and stationary (not time-dependent). Consider, without loss of generality, the output, $v(t)$, at $t = 0$. We define the worst-case noise, $NW(b_0)$, for each of the two possible values of b_0 as follows:

$$NW(b_0) = \max_{\{b_k\}, k \neq 0} [\langle v_{out}^2(0) \rangle - \langle v_{out}(0) \rangle^2], \tag{11}$$

where in (11) the maximization is over all possible sets $\{b_k\}$ for $k \neq 0$, and where b_0 can take on either of two values as previously stated. The

---

[†] The limitations imposed by this assumption are discussed in Section VII.

quantity $NW(b_0)$ shows, for the two possible values of b_0, what the noise for the worst combination of the other symbols is.

We shall next calculate $\langle v_{\text{out}}^2(t) \rangle - \langle v_{\text{out}}(t) \rangle^2$ as a function of the set $\{b_k\}$.

Examine Fig. 5. We shall define the two-sided spectral density of the amplifier-current noise source $i_a(t)$ as S_I and the two-sided spectral height of the amplifier-voltage noise source $e_a(t)$ as S_E. The two-sided spectral density of the Johnson-current noise source $i_b(t)$ associated with R_b is $2k\theta/R_b$, where k is Boltzmann's constant and θ is the absolute temperature.

We can write the output noise as follows:

$$v_{\text{out}}(t) - \langle v_{\text{out}}(t) \rangle = n_S(t) + n_R(t) + n_I(t) + n_E(t), \qquad (12)$$

where

$n_S(t)$ is the output noise due to the random multiplied Poisson process nature of the current $i_s(t)$ produced by the detector,

$n_R(t)$ is the output noise due to the Johnson noise current source of the resistor R_b,

$n_I(t)$ is the output noise due to the amplifier input current noise source $i_a(t)$, and

$n_E(t)$ is the output noise due to the amplifier input voltage noise source $e_a(t)$.

We have

$$\langle v_{\text{out}}^2(t) \rangle - \langle v_{\text{out}}(t) \rangle^2$$
$$= \langle (v_{\text{out}}(t) - \langle v_{\text{out}}(t) \rangle)^2 \rangle$$
$$= \langle n_S^2(t) \rangle + \langle n_R^2(t) \rangle + \langle n_I^2(t) \rangle + \langle n_E^2(t) \rangle$$
$$= \langle n_S^2(t) \rangle + (2k\theta/R_b) \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| H_{\text{eq}}(\omega) \frac{1}{\dfrac{1}{R_b} + \dfrac{1}{R_A} + j\omega(C_d + C_A)} \right|^2 d\omega$$
$$+ (S_I) \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| H_{\text{eq}}(\omega) \frac{1}{\dfrac{1}{R_b} + \dfrac{1}{R_A} + j\omega(C_d + C_A)} \right|^2 d\omega$$
$$+ (S_E) \frac{1}{2\pi} \int_{-\infty}^{\infty} |H_{\text{eq}}(\omega)|^2 d\omega. \qquad (13)$$

In (13), the last three terms were evaluated using the well-known formula for the average-squared output of a filter driven by white noise. We must now calculate the "shot noise" term $\langle n_S^2(t) \rangle$.

Recall that $i_s(t)$ consists of impulses of random charge corresponding to "bunches" of electrons with a random number g per bunch, this number being independent from bunch to bunch.

Consider a finite interval of duration L. Let g_k be the number of electrons in bunch k in the interval; where the bunches are labeled *not* in order of time but at random. Let t_k be the arrival time of bunch k. Let $h_I(t)$ be the response of the RC circuit, amplifier, equalizer combination to a current impulse from $i_s(t)$. Then the output $v_{out}(t)$ *just due to the current $i_s(t)$ in the interval L is*

$$v_{out}^L(t) = \sum_1^N e g_k h_I(t - t_k), \tag{14}$$

where N is the number of bunches.

Recall that the probability density of N bunches at the times $\{t_k\}$ is

$$p[N, \{t_k\}] = \frac{e^{-\Lambda} \prod_1^N \lambda(t_k)}{N!}, \tag{15}$$

where

$$\lambda(t) = p(t) \frac{\eta}{\hbar\Omega} + \lambda_0.$$

Thus combining (14) and (15) and leaving out some tedious algebra we obtain[10]

$$\langle v_{out}^L(t) \rangle = \int_{\text{interval } L} e\langle g \rangle (p(t')\eta/\hbar\Omega + \lambda_0) h_I(t - t')dt'. \tag{16}$$

In a similar manner, we obtain

$$\langle (v_{out}^L(t))^2 \rangle - \langle v_{out}^L(t) \rangle^2 = \int_{\text{interval } L} e^2 \langle g^2 \rangle \left(p(t') \frac{\eta}{\hbar\Omega} + \lambda_0 \right) h_I^2(t - t')dt,$$

where $\langle g \rangle$ is the mean internal gain of the detector and $\langle g^2 \rangle$ is the mean-squared internal gain.

We therefore obtain, letting $L \to \infty$, the result

$$\langle n_S^2(t) \rangle = \lim_{L \to \infty} \left[\langle [v_{out}^L(t)]^2 \rangle - \langle v_{out}^L(t) \rangle^2 \right]$$

$$= \int_{-\infty}^{\infty} e^2 \langle g^2 \rangle \left\{ [\sum b_k h_p(t' - kT)] \frac{\eta}{\hbar\Omega} + \lambda_0 \right\} h_I^2(t - t')dt'. \tag{17}$$

Further,

$$H_I(\omega) = F\{h_I(t - t')\} = H_{eq}(\omega) \frac{1}{\dfrac{1}{R_b} + \dfrac{1}{R_A} + j\omega(C_d + C_A)}. \tag{18}$$

Thus we have the remaining term in (13) in terms of the input optical pulse, the equalizer response, and the RC circuit at the amplifier input.

Converting everything to the frequency domain and recalling that we have normalized the equalized output pulse $h_{\text{out}}(t)$ to unity at $t = 0$, we obtain

$$NW(b_0) \underset{\{b_k, k \neq 0\}}{\overset{\max}{=}} \left[\left(\frac{1}{2\pi} \right)^2 \int_{-\infty}^{\infty} \frac{\langle g^2 \rangle}{\langle g \rangle^2} \frac{\hbar \Omega}{\eta} H_p(\omega) \left(\sum_{-\infty}^{\infty} b_k e^{j\omega kT} \right) \right.$$

$$\times \left(\frac{H_{\text{out}}(\omega)}{H_p(\omega)} * \frac{H_{\text{out}}(\omega)}{H_p(\omega)} \right) d\omega$$

$$+ \frac{(\hbar \Omega / \eta)^2}{2\pi \langle g \rangle^2 e^2} \left(\frac{2k\theta}{R_b} + S_I + e^2 \langle g^2 \rangle \lambda_0) \right) \int_{-\infty}^{\infty} \left| \frac{H_{\text{out}}(\omega)}{H_p(\omega)} \right|^2 d\omega$$

$$+ \frac{(\hbar \Omega / \eta)^2}{2\pi \langle g \rangle^2 e^2} S_E \int_{-\infty}^{\infty} \left| \frac{H_{\text{out}}(\omega) \left(\frac{1}{R_b} + \frac{1}{R_A} + j\omega(C_d + C_A) \right)}{H_p(\omega)} \right|^2 d\omega \left. \right], \quad (19)$$

where

$$H_p(\omega) = F\{h_p(t)\} = \text{input power pulse transform},$$
$$H_{\text{out}}(\omega) = F\{h_{\text{out}}(t)\} = \text{output pulse transform},$$
$$\text{``*''} = \text{convolution},$$
$$b_0 = \text{coefficient multiplying zeroth input pulse},$$

and

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} H_{\text{out}}(\omega) d\omega = 1. \quad (20)$$

In principle, we wish to minimize $NW(b_0)$ by choosing R_b and $H_{\text{eq}}(\omega)$ for the worst-case combination of symbols $\{b_k\}$, subject to the zero intersymbol interference condition on the ouput pulse stream $v_{\text{out}}(t)$ [recognizing that we have normalized $h_{\text{out}}(t)$ and $H_{\text{out}}(\omega)$ as given in (9) and (20) above].

4.1 Comments

    (i) One observation, which follows regardless of the choice of $H_{\text{out}}(\omega)$, is that the noise is always made smaller when R_b is increased. Therefore, subject to practical constraints and for a fixed amplifier and a fixed desired output pulse shape (which is determined by the equalizer and R_b), it is always best to make R_b, the bias circuit resistor, as large as possible.

    (ii) It is also clear, from (17) and the fact that the input pulse $h_p(t)$ is positive for all t, that the worst-case noise occurs when all the b_k (except b_0) assume the larger of the two possible

values. Recall that we are interested in the noise for both values of b_0.

  (iii) Furthermore, for a given S_E and S_I and a given output pulse shape, it is desirable that the amplifier input resistance be as large as possible and that the amplifier shunt capacitance be as small as possible.

  (iv) It is desirable that the diode shunt capacitance be as small as possible.

V. CHOOSING THE EQUALIZED PULSE SHAPE

In principle, using (19) and given $H_p(\omega)$, $\langle g \rangle$, $\langle g^2 \rangle$, S_I, S_E, R_b, R_A, C_d, and C_A one can find the equalized pulse shape $H_{\text{out}}(\omega)$ for each value of b_0 that minimizes the worst-case noise.

In practice, other considerations in addition to the noise are also of interest. In particular, it is important not only that the intersymbol interference be low at the nominal decision times kT, but that it be sufficiently small at times offset from $\{kT\}$ to allow for timing errors in the sampling process.

Therefore, rather than seeking the equalized pulse shape that minimizes the noise, we shall consider various equalized pulse shapes to see how the noise trades off against eye width.

Before proceeding, it is helpful to perform some normalizations upon (19) to reduce the number of parameters.

Make the following definitions:

$$R_T = \left(\frac{1}{R_b} + \frac{1}{R_A} \right)^{-1} = \text{total detector parallel load resistance}, \quad (21)$$

$C_T = C_d + C_A = \text{total detector parallel load capacitance},$
$b_{\max} = \text{larger value of } b_k, \; b_{\min} = \text{smaller value of } b_k,$

$$H_p'(\omega) = H_p \left(\frac{2\pi\omega}{T} \right),$$

$$H_{\text{out}}'(\omega) = \frac{1}{T} H_{\text{out}} \left(\frac{2\pi\omega}{T} \right).$$

In this normalization, the functions $H_p'(\omega)$ and $H_{\text{out}}'(\omega)$ depend only upon the *shapes* of $H_p(\omega)$ and $H_{\text{out}}(\omega)$, not upon the time slot width T. The previous normalizing conditions on $H_p(\omega)$ and $H_{\text{out}}(\omega)$ imply conditions on $H_p'(\omega)$ and $H_{\text{out}}'(\omega)$

$$H_p(0) = 1 \Rightarrow H_p'(0) = 1 \quad (22)$$

which implies

$$\int_{-\infty}^{\infty} h_p(t)dt = 1.$$

Also,

$$h_{\text{out}}(0) = 1 \Rightarrow \frac{1}{2\pi} \int_{-\infty}^{\infty} H_{\text{out}}(\omega)d\omega = 1 \Rightarrow \int_{-\infty}^{\infty} H'_{\text{out}}(f)df = 1.$$

With the above normalizations, (19) becomes

SHOT NOISES

$$NW(b_0) = \left(\frac{\hbar\Omega}{\eta}\right)^2 \left\{ \frac{\langle g^2 \rangle}{\langle g \rangle^2} \frac{\eta}{\hbar\Omega} [b_0 I_1 + b_{\max}[\textstyle\sum_1 - I_1]] \right.$$

SHOT NOISE

$$+ \frac{T}{(\langle g \rangle e)^2} \left[S_I + \frac{2k\theta}{R_b} + \langle g^2 \rangle e^2 \lambda_d + \frac{S_E}{R_T^2} \right] I_2$$

THERMAL NOISES

$$+ \frac{(2\pi C_T)^2 S_E I_3}{T(\langle g \rangle e)^2} \left. \right\}, \tag{23}$$

where

$$I_1 = \int_{-\infty}^{\infty} H'_p(f) \left[\frac{H'_{\text{out}}(f)}{H'_p(f)} * \frac{H'_{\text{out}}(f)}{H'_p(f)} \right] df$$

$$\textstyle\sum_1 = \sum_{k=-\infty}^{\infty} H'_p(k) \left[\frac{H'_{\text{out}}(k)}{H'_p(k)} * \frac{H'_{\text{out}}(k)}{H'_p(k)} \right]$$

$$I_2 = \int_{-\infty}^{\infty} \left| \frac{H'_{\text{out}}(f)}{H'_p(f)} \right|^2 df$$

$$I_3 = \int_{-\infty}^{\infty} \left| \frac{H'_{\text{out}}(f)}{H'_p(f)} \right|^2 f^2 df.$$

In (23), the first shot-noise term is due to the pulse in the time slot under decision, the second term being shot noises from the other pulses which are assumed to be all "on." From this normalized form of (19), we see that for a fixed input pulse *shape* and a fixed output pulse *shape* and with fixed R_b, R_A, C_T, S_E, and S_I, the noise decreases as the bit rate, $1/T$, increases (a consequence of the square-law detection) until the term involving I_3 dominates. After that, the noise increases with increasing bit rate (due to the shunt capacitance C_T).

$$h_p(t) = \frac{1}{\alpha T} \text{ FOR } \frac{-\alpha T}{2} < t < \frac{\alpha T}{2}, 0 \text{ OTHERWISE}$$

$$h_p(t) = \frac{1}{\sqrt{2\pi}\,\alpha T} \text{ EXP } \left\{ -t^2/[2\,(\alpha T)^2] \right\}$$

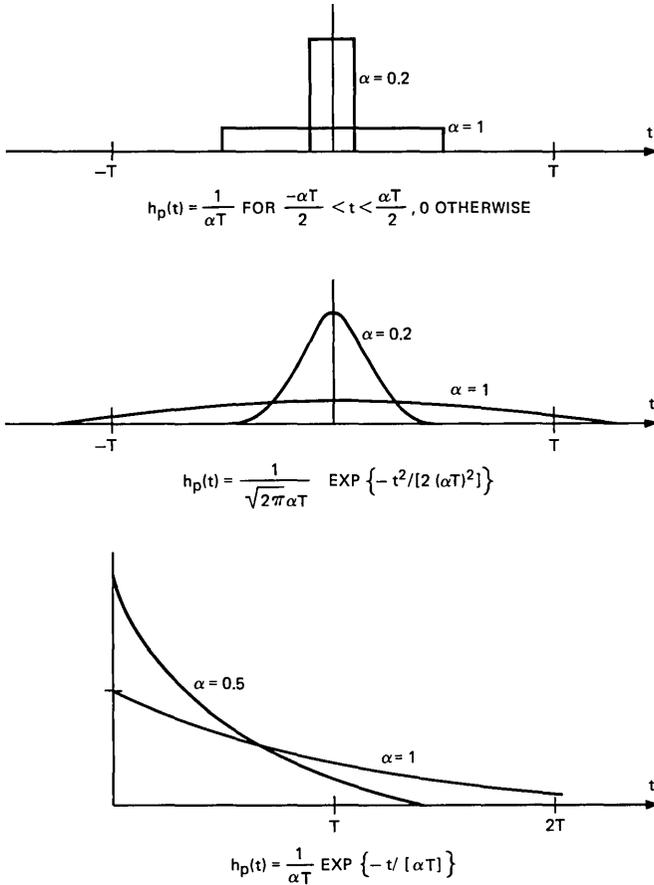$$h_p(t) = \frac{1}{\alpha T} \text{ EXP } \left\{ -t/\,[\alpha T] \right\}$$

Fig. 6a—Input pulse families.

Example of Normalization:

Suppose the input optical pulse is a rectangular pulse of unit area having width equal to one-half a time slot T; then

$$H_p(\omega) = \int_{-\frac{1}{4}T}^{\frac{1}{4}T} \frac{2}{T} e^{i\omega t} dt$$

$$= \frac{1}{i\omega} \left(\frac{2}{T} \right) (e^{i\omega t/4} - e^{-i\omega T/4})$$

$$= \sin \frac{(\omega T/4)}{\omega T/4}. \tag{24}$$

$$H'_{OUT}(f)$$

$$h_{OUT}(t) = \frac{\sin(\pi t/T)\cos(\pi\beta t/T)}{\pi t/T\,[1-(2\beta t/T)^2]}$$

$$h_{OUT}(t)$$

WORST–CASE EYE DIAGRAMS
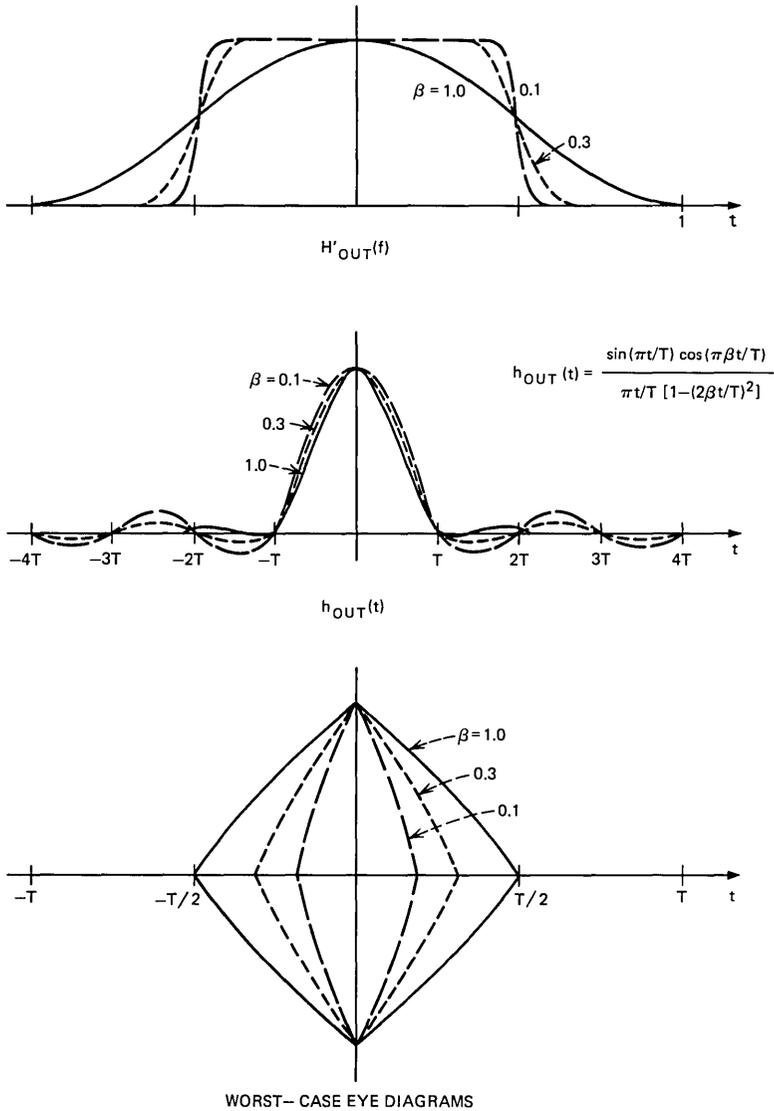
Fig. 6b—Frequency domain, time domain, and eye diagram representations of raised cosine family.

Therefore,

$$H'_p(f) = H_p\left(\frac{2\pi f}{T}\right) = \frac{\sin(\pi f/2)}{\pi f/2}.$$

As expected, the normalized pulse spectrum $H_p'(f)$ is independent of the time slot width T and merely reflects the fact that the pulse $h_p(t)$ is a rectangular pulse with width equal to half a time slot.

In order to obtain the noise for various input and output (equalized) pulse shapes, one needs to calculate the three integrals I_1, I_2, and I_3 and the sum Σ_1.

Consider the following three families of input pulse shapes (see Fig. 6a) and single family of output pulse shapes (see Fig. 6b).

(*i*) Rectangular input pulses:

$$h_p(t) = \frac{1}{\alpha T}, \qquad -\frac{\alpha T}{2} < t < \frac{\alpha T}{2}, \qquad 0 \text{ otherwise} \quad (25)$$

$$H_p'(f) = \frac{\sin(\alpha\pi f)}{\alpha\pi f}.$$

(*ii*) Gaussian input pulses:

$$h_p(t) = \frac{1}{\sqrt{2\pi}\alpha T}\, e^{-[t^2/2(\alpha T)^2]}$$

$$H_p'(f) = e^{-(2\pi\alpha f)^2/2}.$$

(*iii*) Exponential input pulses:

$$h_p(t) = \frac{1}{\alpha T}\, e^{-t/\alpha T}$$

$$H_p'(f) = \frac{1}{1 + j2\pi\alpha f}.$$

(*iv*) "Raised cosine" output pulses:

$$h_{\text{out}}(t) = \left[\sin\left(\frac{\pi t}{T}\right)\cos\left(\frac{\pi\beta t}{T}\right)\right]\left[\frac{\pi t}{T}\left(1 - \left(\frac{2\beta t}{T}\right)^2\right)\right]^{-1}$$

$$H_{\text{out}}'(f) = 1, \qquad \text{for} \qquad 0 < |f| < \frac{(1-\beta)}{2}$$

$$= \frac{1}{2}\left[1 - \sin\left(\frac{\pi f}{\beta} - \frac{\pi}{2\beta}\right)\right], \qquad \text{for } \frac{1-\beta}{2} < |f| < \frac{1+\beta}{2}$$

$$= 0 \text{ otherwise.}$$

(Time, frequency, and eye diagram representations of the raised cosine family are shown as a function of β in Fig. 6b.[11])

In Figs. 7 through 18 calculations of I_1, I_2, I_3, and I_4 are given graphically for each input pulse family as a function of α and β.
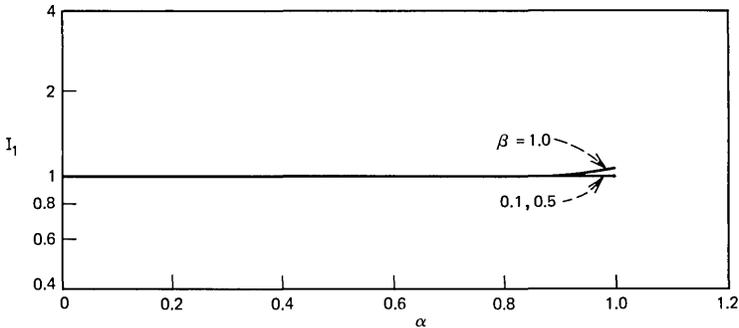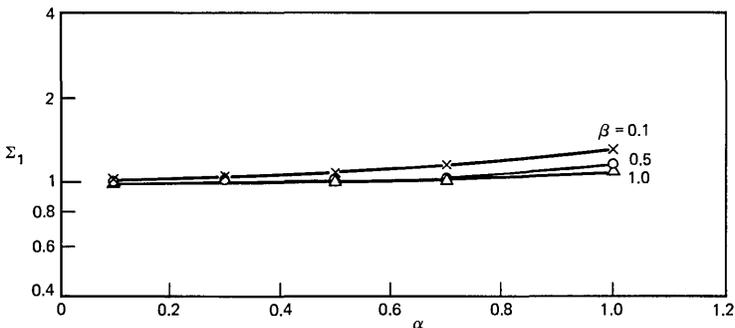
Fig. 7—Rectangular family I_1 vs α and β.

5.1 Comments on the Numerical Results

For the rectangular input pulses with widths between 0.1 and 1 time slot, I_1, $\sum_1$, I_2, and I_3 vary very little. Thus, if one expects to receive rectangular optical pulses which are fixed in energy, the required energy per pulse is insensitive to the pulse width for widths up to 1 time slot.

The curves for Gaussian-shaped input pulses imply very strong sensitivity of required energy per pulse to pulse width. This is a consequence of the rapid falloff of the spectrum of a Gaussian pulse with frequency. It is suspected that, although for certain fiber systems the received pulses may appear approximately Gaussian in the time domain, the frequency spectrum will not suffer such a rapid falloff. The results for the exponential-shaped input pulses seem much more realistic.

For exponential-shaped optical pulses we notice, from Figs. 15 and 16, that the shot noise coefficients I_1 and $\sum_1$ are sensitive to the optical

Fig. 8—Rectangular family Σ_1 vs α and β.
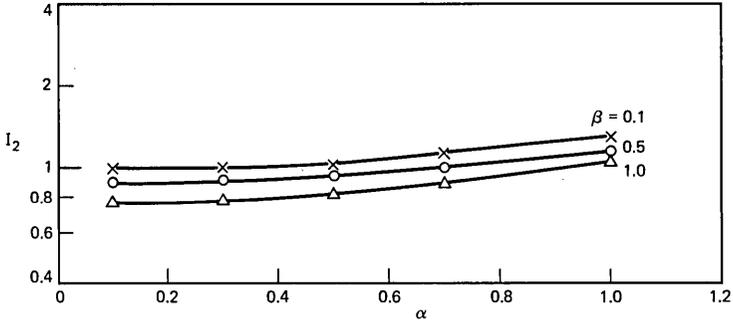
Fig. 9—Rectangular family I_2 vs α and β.
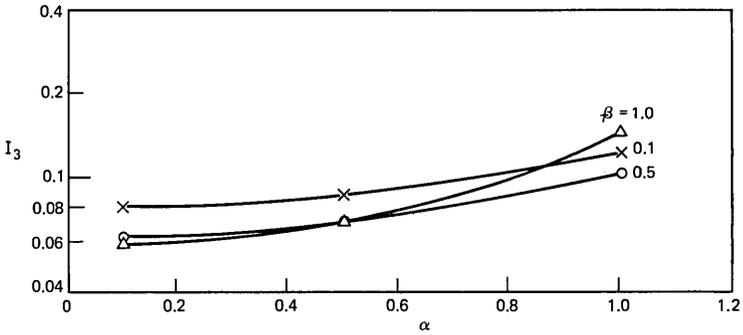
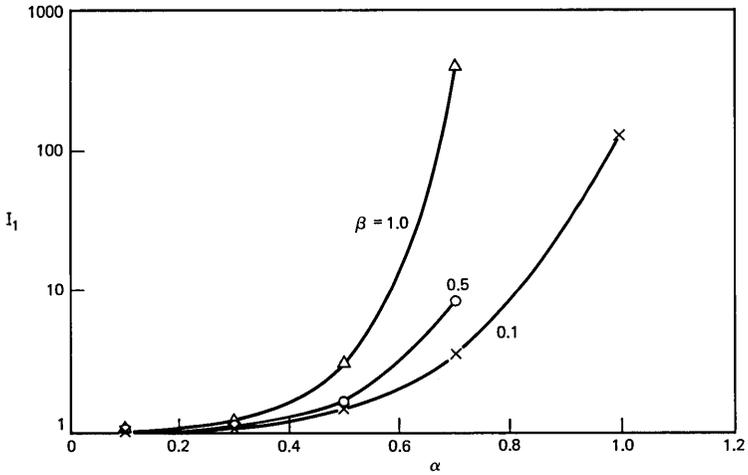Fig. 10—Rectangular family I_3 vs α and β.
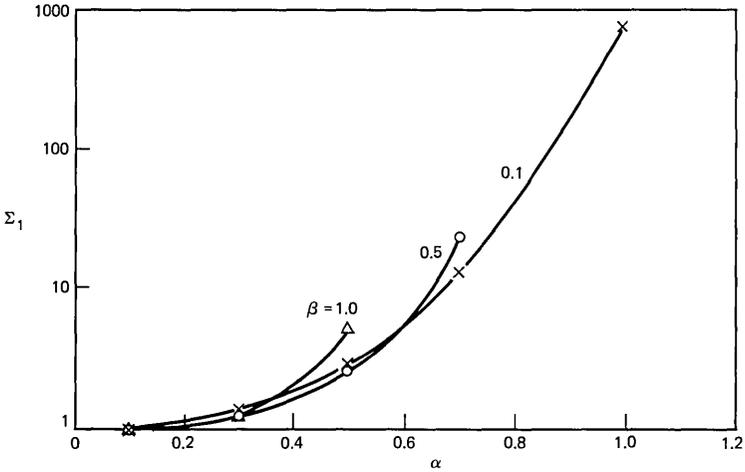

Fig. 11—Gaussian family I_1 vs α and β.

Fig. 12—Gaussian family Σ_1 vs α and β.

pulse width, but that these sensitivities imply a practically useful tradeoff in required optical power vs allowable bit rate. That is, one might take a certain power penalty to allow equalization which can substantially increase the usable bit rate on a channel having a fixed optical output pulse width. The sensitivity of I_2 and I_3 to the optical pulse width is similar to that of $\sum_1$ and less significant because in-
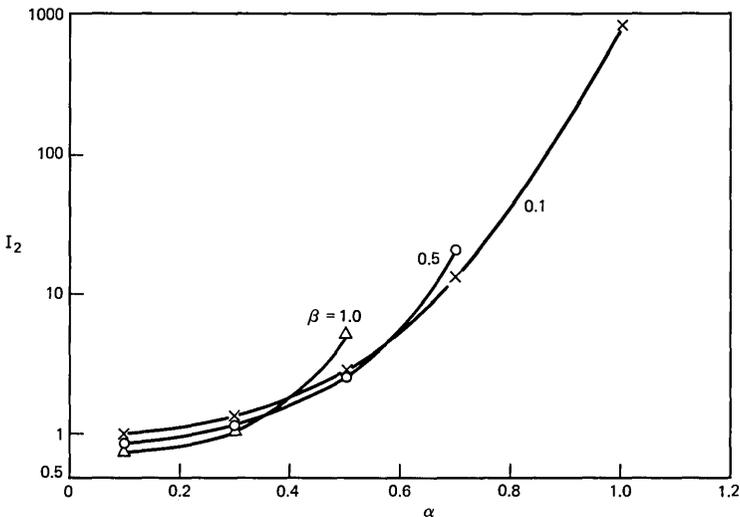
Fig. 13—Gaussian family I_2 vs α and β.
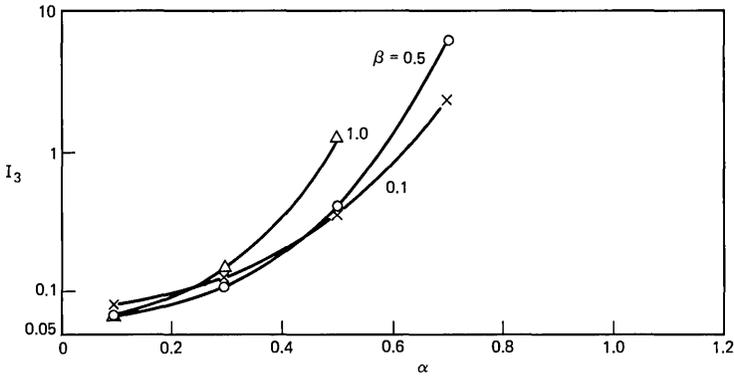
Fig. 14—Gaussian family I_3 vs α and β.

creases in the thermal noises of the receiver are for the most part compensated for by adjustment of the avalanche gain, with only a small penalty in excess shot noise. The above statements will be made quantitative in Section VI.

VI. OBTAINING THE RELATIONSHIPS FOR FIXED ERROR RATE BETWEEN THE REQUIRED ENERGY PER PULSE, OPTIMAL AVALANCHE GAIN, AND OTHER PARAMETERS

Suppose that in (23) all parameters are fixed except $\langle g \rangle$, $\langle g^2 \rangle$, b_{min}, and b_{max}.
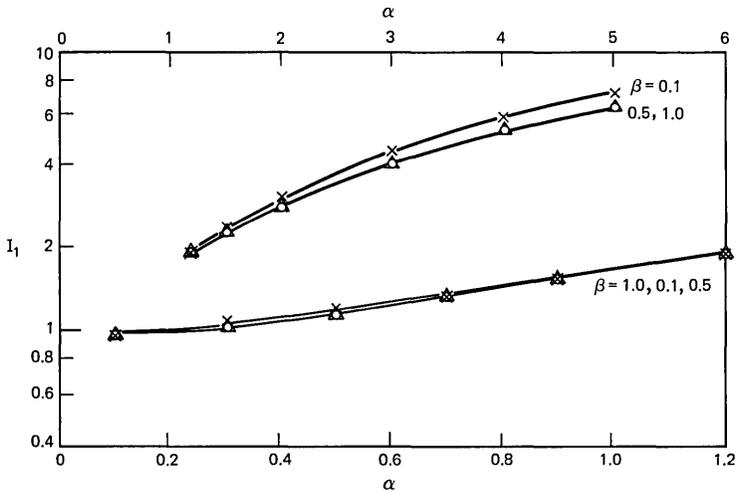

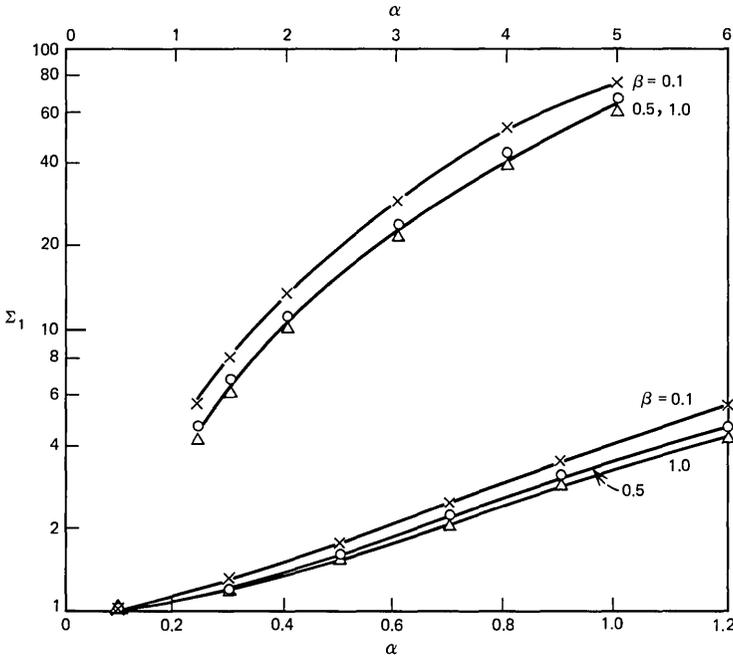
Fig. 15—Exponential family I_1 vs α and β.

Fig. 16—Exponential family Σ_1 vs α and β.

The receiver equalized output at the sampling time, due to an optical pulse of energy b_0, is b_0.

When $b_0 = b_{\min}$, we must be sure that the probability that noise drives the receiver output $v_{\text{out}}(t)$ at the sampling time above the threshold D is less than 10^{-9}.[†] Using the signal-to-noise ratio approximation,[‡] we require the noise variance, $NW(b_{\min})$, to be less than $\{\frac{1}{6}[D - b_{\min}]\}^2$.

Therefore, we require that

$$NW(b_{\min}) \leqq \frac{1}{36}[D - b_{\min}]^2. \tag{26}$$

Furthermore, when $b_0 = b_{\max}$ we must be sure that the probability that the noise drives the receiver output below the threshold is less than 10^{-9}. Therefore, we require that

$$NW(b_{\max}) \leqq \frac{1}{36}[b_{\max} - D]^2. \tag{27}$$

[†] An error rate of 10^{-9} is arbitrarily chosen here. Dependence of required optical power on error rate is discussed in Part II of this paper.
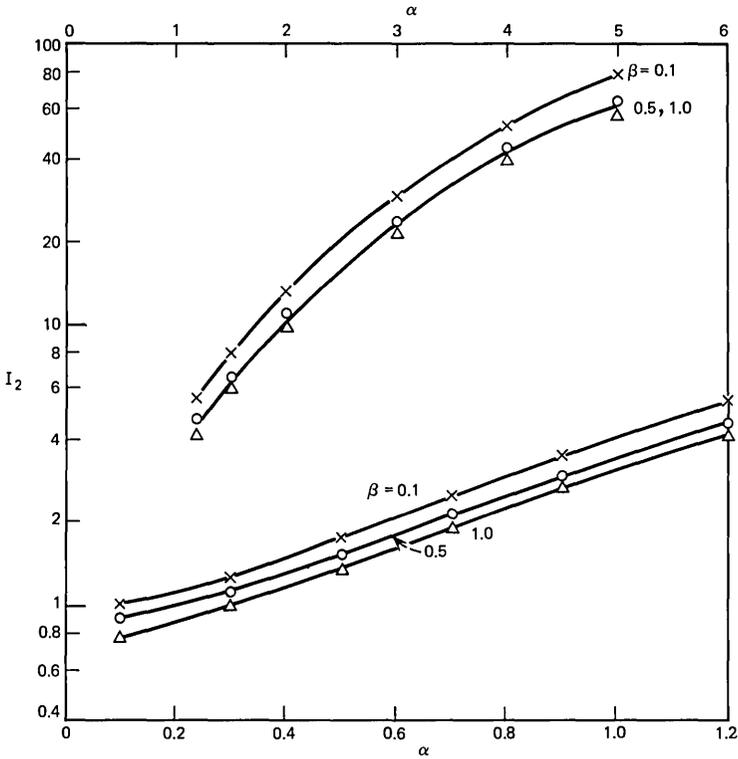
[‡] See Appendix A.

Fig. 17—Exponential family I_2 vs α and β.

Using equality in (26) and (27), we require for a 10^{-9} error rate

$$\sqrt{NW(b_{\max})} + \sqrt{NW(b_{\min})} = \tfrac{1}{6}(b_{\max} - b_{\min}).\qquad (28)$$

Very often we have a fixed ratio $(b_{\min}/b_{\max}) = \rho$.

Rearranging (28) we obtain

$$b_{\max} = \frac{6}{1 - \rho} \left[\sqrt{NW(b_{\max})} + \sqrt{NW(\rho b_{\max})}\right].\qquad (29)$$

In order to obtain numerical results, we shall make the following reasonable assumptions. Let the dark current be negligible and let $b_{\min}/b_{\max}$ be much less than unity. Therefore we shall set $\lambda_0 = 0$, $b_{\min} = 0.$[†] We obtain from (23)

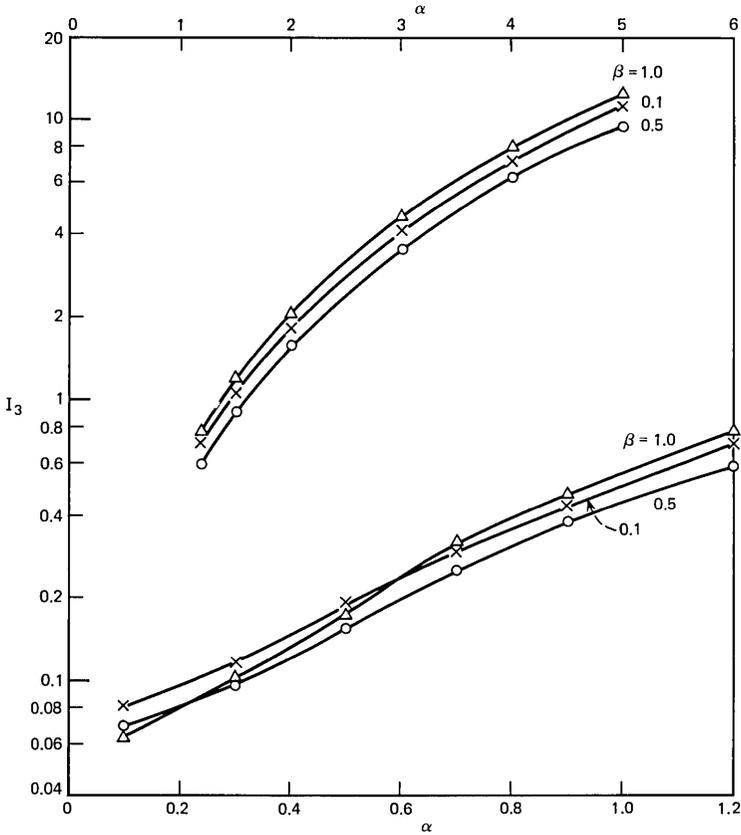[†] Quantitative discussion of the consequences of these approximations are given in Part II.

Fig. 18—Exponential family I_3 vs α and β.

$$NW(b_0) \cong \left[\frac{\hbar\Omega}{\eta} \right]^2 \left\{ \frac{\langle g^2 \rangle}{\langle g \rangle^2} \frac{\eta}{\hbar\Omega} \left[b_0 I_1 + b_{\max}(\Sigma_1 - I_1) \right] + \frac{1}{\langle g \rangle^2} [Z] \right\},$$

where

$$Z \triangleq \left\{ \frac{T}{e^2} \left[S_I + \frac{2k\theta}{R_b} + \frac{S_E}{R_T^2} \right] I_2 + \frac{(2\pi C_T)^2}{Te^2} S_E I_3 \right\}. \tag{30}$$

In (30), Z includes all the thermal noise terms of (23).

From (29), taking the limit as $\rho \to 0$ ($b_{\min} \to 0$), we obtain the conditions to achieve a 10^{-9} error rate as follows.

Case I: Thermal noise (Z) dominates (i.e., little or no avalanche gain).

$$b_{\max} = \frac{12\hbar\Omega}{\eta\langle g \rangle} Z^{\frac{1}{2}}. \tag{31}$$

Case II: Optimal gain (i.e., $\langle g \rangle$ adjusted to minimize the required optical energy in an "on" pulse b_{max}).

Let the relationship between $\langle g^2 \rangle$ and $\langle g \rangle$ be specified in the usual way:

$$\langle g^2 \rangle = \langle g \rangle^{2+x}, \tag{32}$$

where x depends upon the type of detector. We obtain the following formula for the optimal gain:

$$\langle g \rangle_{optimal} = (6)^{-1/(1+x)}(Z)^{1/(2+2x)}(\gamma_1)^{1/(2+2x)}(\gamma_2)^{-1/(1+x)}, \tag{33}$$

where defining $I_5 = \sum_1 - I_1$ [see eq. (23)]

$$\gamma_1 \triangleq \frac{-[\sum_1 + I_5] + \sqrt{(\sum_1 + I_5)^2 + \dfrac{16(1+x)}{x^2} \sum_1 I_5}}{2 \sum_1 I_5} \tag{34}$$

$$\gamma_2 \triangleq \sqrt{1/\gamma_1 + I_5} + \sqrt{1/\gamma_1 + \sum_1}.$$

We obtain the following formula for b_{max}:

$$b_{max} = \frac{\hbar\Omega}{\eta} (6)^{(2+x)/(1+x)}(Z)^{x/(2+2x)}(\gamma_1)^{x/(2+2x)}(\gamma_2)^{(2+x)/(1+x)}. \tag{35}$$

That is,

$$b_{max} \propto [Z]^{x/(2+2x)}. \tag{36}$$

We therefore see that for these assumptions and $x = 0.5$ corresponding to a silicon avalanche detector the minimum required energy per pulse varies as the one-half power of the thermal noise term, Z, without avalanche gain, and as the one-sixth power of the thermal noise term, Z, with optimal gain.

However, this does not mean that at optimal gain the value of Z is unimportant. By reducing Z (the thermal noise terms) through proper choice of biasing and amplifier circuitry, we still minimize the optimizing avalanche gain [see (33)] and obtain some reduction in the required energy per pulse (see Part II).

6.1 *Example*

From eqs. (23), (30), (34), and (35) we can calculate, for various shaped optical pulses, the effect of intersymbol interference on the required energy per "on" pulse (b_{max}) and therefore on the required average optical power needed for a 10^{-9} error rate.[†] We shall assume

---

[†] That is, if pulses are "on" half the time, the required optical power equals $b_{max}/2T$.
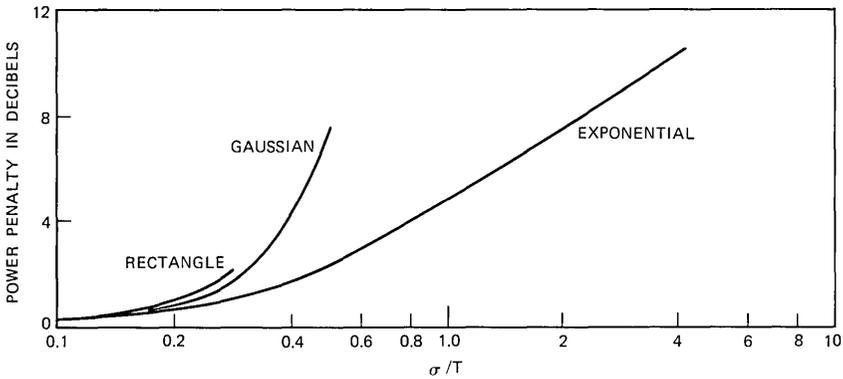
Fig. 19—No avalanche gain.

that the detector amplifier shunt resistance R_T is sufficiently large so that the term $((2\pi C_T)^2/Te^2)S_E I_3$ dominates the thermal noise in (23) and (30).

The minimal required optical power is obtained for very narrow optical input pulses.[†] For other pulse shapes, the *excess* required optical power can be defined as a *penalty* in dB for not using narrow pulses. This penalty is plotted in Figs. 19 and 20 for the case of no avalanche gain and optimal avalanche gain using the pulse shapes of (25), assuming a silicon detector ($x = 0.5$). In those figures, the abscissa is the normalized rms optical pulse width defined as follows:

$$\frac{\sigma^2}{T^2} = \frac{\left(\int t^2 h_p(t)dt\right) - \left(\int t h_p(t)dt\right)^2}{T^2} , \qquad (37)$$

where T = time slot width.

VII. CONCLUSIONS

7.1 *Conclusion on Choosing the Biasing Circuitry*

From the results of Sections IV and VI, and from (23), it is clear that, to minimize the thermal noise degradations introduced by the amplifiers following the detector, it is necessary to make the amplifier input resistance and the biasing circuit resistance sufficiently large so that the amplifier series noise source dominates the Johnson noise of these parallel resistances. When designing the amplifier, one should keep in mind that for a silicon avalanche detector the required optical
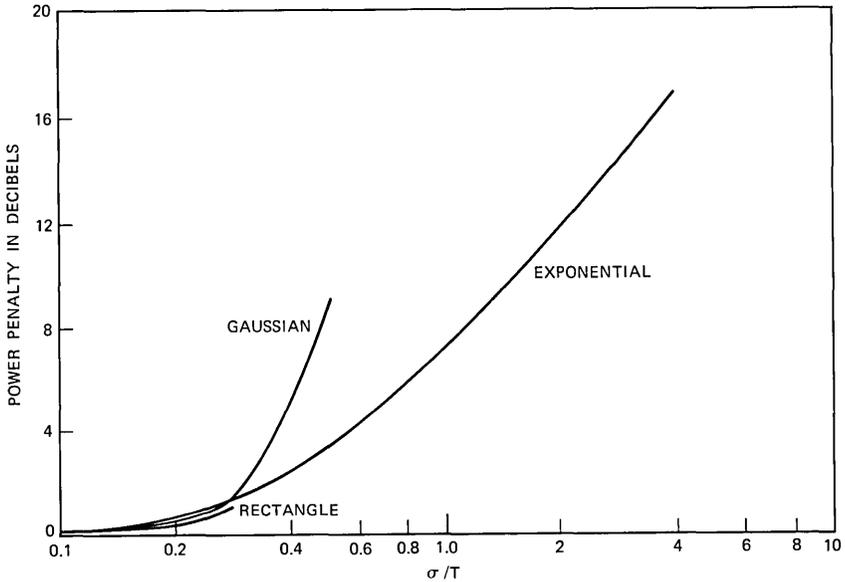
[†] See Appendix B.

Fig. 20—Optimal gain.

energy per pulse at optimal gain varies roughly as the one-sixth power of the thermal noise variance at the receiver output, and therefore it is not wise to spend too much money on thermal noise reduction. On the other hand, if one is not using avalanche gain, the required energy per pulse varies roughly as the one-half power of the thermal noise variance at the receiver output.

In order to minimize the effects of the thermal noise, the total capacitance shunting the detector should be as small as possible and the equivalent series thermal noise source of the amplifier should also be as small as possible.

7.2 *The Effect of Bit Rate on Required Energy Per Pulse*†

The effect of bit rate on the required energy per pulse is small if the received pulses remain well confined to a time slot. In (23), assume I_1, $\sum_1$, I_2, and I_3 are fixed corresponding to a fixed received pulse width relative to a time slot. Then the shot noise terms due to the signal are independent of the bit rate $1/T$, and the shot noise due to the dark current decreases with increasing bit rate. If the series noise from the amplifier dominates, then the thermal noise increases with increasing

---

† This subject will be discussed in more detail in Part II.

bit rate, but is for the most part compensated for by the avalanche gain with little penalty in required energy per pulse.

If considerable equalization is being used, then the required energy per pulse increases with the bit rate (because a higher bit rate necessitates greater equalization). For the equalization assumed above, where the equalized pulses are forced to go to zero at all sampling times except one, the required energy per pulse is a strong function of the bit rate. For example, with exponential-shaped received pulses, the required optical power at optimal avalanche gain was roughly 6 dB higher for a pulse 1 time slot wide to the $1/e$ point compared to a pulse only 0.25 time slot wide to the $1/e$ point (see Fig. 20).

On the other hand, it is clear that zero-forcing-type equalization is not optimal, particularly for received pulses whose spectra fall off rapidly with frequency. It is more likely that some compromise between eye opening and output noise variance results in minimum required energy per pulse.

For the assumed zero-forcing equalization, we still can conclude that a usable tradeoff exists between required energy per pulse and bit rate, and this will allow some extension of the usable rate on "dispersion-limited" fibers.

7.3 Comments on Previous Work

The purpose of this paper has been to illustrate the application of the "high-impedance" front-end design to optical digital repeaters, to take into account precisely the input pulse shape and the equalizer-filter shape, and to obtain explicit formulas for the required optical power to achieve a desired error rate as a function of the other parameters.

Previous authors[12,13] working in the areas of particle counting and video amplifier design have recognized that a high-impedance front end followed by proper equalization in later stages provides low noise and adequate bandwidth. However, optical communication theorists[5,6,14,15] have in the past often used the criterion "$RC \leqq T$"–loading down the front-end amplifier so as to have adequate bandwidth without equalization–therein incurring an unnecessary noise penalty. Some optical experimenters[16,17] have recognized the high-impedance design for observing isolated pulses or single frequencies, but failed to recognize the use of equalization.

Many previous authors[3,4,6,15] have used simple formulas (which usually assume isolated rectangular input pulses and a front-end bandwidth of the reciprocal pulse width) to obtain the required power in

optical communication systems for a desired signal-to-noise ratio. Often these formulas average out the signal-dependent nature of the shot noise. If modified to include the high-impedance design concept, such formulas are very useful for obtaining "ball park" estimates of optical power requirements. Such formulas are, in general, special cases of the formulas described here.

7.4 *Experimental Verification*

In work recently reported,[18] J. E. Goell has shown that, in a 6.3-Mb/s repeater operating at an error rate of 10^{-9}, agreement of experimentally determined power requirements and the above theory were within 1 dB (0.25 dB in cases without avalanche gain). In particular, using an FET front end and the "high-impedance" design, the optical power requirement without avalanche gain was 8 dB less than with the front end loaded down to the "$RC = T$" design.

APPENDIX A

Signal-to-Noise Ratio Approximation

In this paper we have calculated the mean voltage (b_{max} or b_{min}) and the average-squared deviation from the mean voltage ($NW(b_{max})$ or $NW(b_{min})$) at the receiver output at the sampling times. In order to calculate error rates, we shall assume that the output voltage is approximately a Gaussian random variable. This is the signal-to-noise ratio approximation. Thus if the threshold, to which we compare the output voltage, is D, and if the desired error probability is P_e, we have

$$\frac{1}{\sqrt{2\pi\sigma_o^2}} \int_D^\infty \exp\left[-(v - b_{min})^2/2\sigma_o^2\right]dv = P_e, \tag{38}$$

where

$$\sigma_o^2 = NW(b_{min})$$

and

$$\frac{1}{\sqrt{2\pi\sigma_1^2}} \int_{-\infty}^D \exp\left[-(v - b_{max})^2/2\sigma_1^2\right]dv = P_e,$$

where

$$\sigma_1^2 = NW(b_{max}).$$

Changing the variables of integration we obtain the following expressions, equivalent to (38):

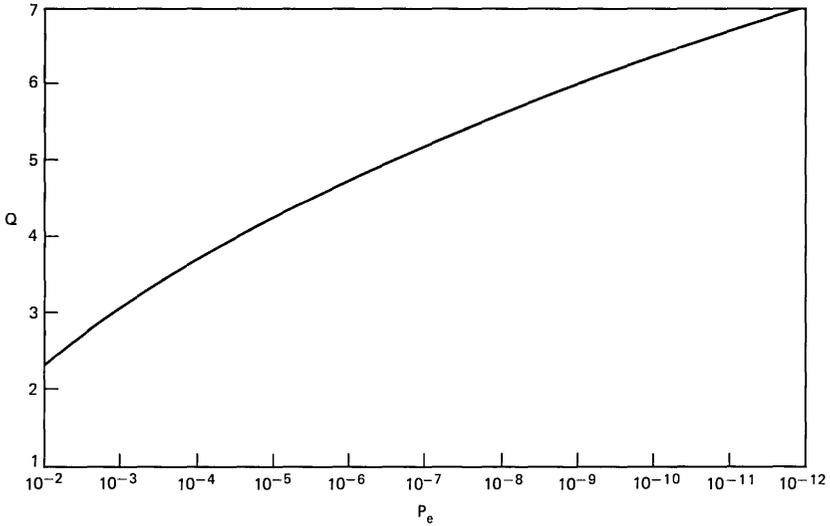$$\frac{1}{\sqrt{2\pi}} \int_Q^\infty e^{-x^2/2}dx = P_e, \tag{39}$$

Fig. 21—Q vs P_e, $P_e = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_Q^\infty e^{-x^2/2}dx.$

where
$$Q = (D - b_{min})/\sigma_o$$
and also
$$Q = (b_{max} - D)/\sigma_1.$$
Thus we must have
$$\sigma_o = \sqrt{NW(b_{min})} = (D - b_{min})/Q$$
and
$$\sigma_1 = \sqrt{NW(b_{max})} = (b_{max} - D)/Q.$$
Therefore we must also have (eliminating D)
$$\sqrt{NW(b_{max})} + \sqrt{NW(b_{min})} = (b_{max} - b_{min})/Q.$$

The value of Q is determined by the error rate through (39) above. Figure 21 shows a plot of Q vs P_e which can be obtained from standard tables.

Equation (39) states that the threshold must be Q standard deviations (of the noise at b_{min}) above b_{min}, and also must be Q standard deviations (of the noise at b_{max}) below b_{max} to insure the desired error rate. For an error rate of 10^{-9} ($P_e = 10^{-9}$) Q is roughly 6 (5.99781).

APPENDIX B

Optimal Input Pulse Shape

We now wish to show that the optimal input pulse, $h_p(t)$ shape (which minimizes the required average optical power) is ideally an impulse; and for practical purposes a pulse which is sufficiently narrow so that its Fourier transform is almost constant for all frequencies passed by the receiver. To do this, we shall show that such a narrow pulse minimizes the noises $NW(b_{max})$ and $NW(b_{min})$ defined in (23).

We begin with the already established condition that the area of $h_p(t)$ is equal to unity and that $h_p(t)$ is positive (power must, of course, be positive).

$$\int h_p(t)dt = 1; \qquad h_p(t) > 0. \tag{40}$$

These conditions imply the following weaker condition:

$$|H_p'(f)| = \left| \int h_p(t)e^{-i2\pi ft/T}dt \right|$$

$$\leq \int |h_p(t)| |e^{-i2\pi ft/T}| dt = \int h_p(t)dt = H_p'(0) = 1. \tag{41}$$

Consider first the thermal noise terms of (23) involving the integrals I_2 and I_3:

$$I_2 = \int \frac{|H_{out}'(f)|^2}{|H_p'(f)|^2} df; \qquad I_3 = \int \frac{|H_{out}'(f)|^2}{|H_p'(f)|^2} f^2 df. \tag{42}$$

Using (41) in (42) we see that these terms I_2 and I_3 are minimized for any desired output pulse $H_{out}'(f)$ by setting $|H_p'(f)| = H_p'(0) = 1$ for all frequencies, f, for which $|H_{out}'(f)| > 0$. Thus, ideally, to minimize I_2 and I_3, $h_p(t)$ is an impulse of unit area which also satisfies the conditions (40).

We must now show that the shot noise terms of (23), I_1 and $\sum_1 - I_1$, are minimized by a very narrow pulse $h_p(t)$.

First recall that $(\sum_1 - I_1)b_{max}(\hbar\Omega/\eta)\langle g^2 \rangle / \langle g \rangle^2$ is the worst-case, mean-square shot noise at the sampling time due to all other pulses except the one under decision, and assuming all of those pulses are "on" ($b_k = b_{max}$ for $k \neq 0$). Thus, from (17), we obtain

$$\sum_1 - I_1 \geq 0, \tag{43}$$

where

$$\sum_1 - I_1 = \int \left(\sum_{k \neq 0} h_p(t' - kT) \right) h_I^2(-t')dt'$$

and where $h_I(t)$ is the overall receiver impulse response relating $h_p(t)$ to $h_{out}(t)$.

Now let the optical pulse $h_p(t)$ be an impulse of unit area. Then the overall impulse response h_I must be equal to $h_{out}(t)$ and therefore using (43) and (9) we obtain

$$\Sigma_1 - I_1 = \sum_{k \neq 0} h_{out}^2(-kT) = 0 \tag{44}$$
$$\text{(for } h_p(t) = \delta(t)\text{).}$$

Because condition (9) requires zero-crossing equalization, we have shown that an impulse shape for $h_p(t)$ minimizes (removes) the shot noise from pulses other than the one under decision.

Finally, consider the shot noise from the pulse under decision given by $I_1(\hbar\Omega/\eta)b_o\langle g^2\rangle/\langle g\rangle^2$ where

$$I_1 = \int h_p(t')h_I^2(-t')dt' > 0. \tag{45}$$

We already have the condition (9)

$$h_{out}(0) = \int h_p(t')h_I(-t')dt' = 1. \tag{46}$$

We can next use the Shwarz inequality on (46)

$$(h_{out}(0))^2 = 1 = \left(\int h_p^{\frac{1}{2}}(t)h_p^{\frac{1}{2}}(t)h_I(-t)dt \right)^2$$
$$\leq \int h_p(t)dt \int h_p(t)h_I^2(-t)dt. \tag{47}$$

Since $\int h_p(t)dt = 1$, we have from (47) and (45)

$$I_1 \geqq 1. \tag{48}$$

Now set $h_p(t)$ equal to a unit area impulse. It then must follow from (46) that $h_I(0) = 1$. We finally obtain

$$\int h_p(t)h_I^2(t)dt = h_I^2(0) = 1. \tag{49}$$

From (48) and (49) we see that an impulse-shaped $h_p(t)$ makes I_1 achieve its minimum value of unity.

Summarizing, an impulse-shaped optical input pulse $h_p(t)$ (for practical purposes a sufficiently narrow pulse so that its Fourier transform is approximately constant for all frequencies passed by the receiver) minimizes all the pulse-shape-dependent coefficients (I_1, $\Sigma_1 - I_1$, I_2,

and I_3) in the noise expression (23) and thereby minimizes the required average optical power to achieve a desired error rate (using the signal-to-noise ratio approximation of Appendix A).

REFERENCES

1. Personick, S. D., "Baseband Linearity and Equalization in Fiber Optic Communication Systems," to appear in B.S.T.J., September 1973.
2. Personick, S. D., "Statistics of a General Class of Avalanche Detectors with Applications to Optical Communication," B.S.T.J., *50*, No. 10 (December 1971), pp. 3075–3096.
3. Melchior, H., et al., "Photodetectors for Optical Communication Systems," Proc. IEEE, *58*, No. 10 (October 1970), pp. 1466–1486.
4. Pratt, W. R., *Laser Communication Systems*, New York: John Wiley and Sons, 1969.
5. Hubbard, W. M., "Comparative Performance of Twin-Channel and Single-Channel Optical Frequency Receivers," IEEE Trans. Commun. *COM20*, No. 6 (December 1972), pp. 1079–1086.
6. Anderson, L. K., and McCurtry, B. J., "High Speed Photodetectors," Proc. IEEE, *54* (October 1966), pp. 1335–1349.
7. McIntyre, R. J., "Multiplication Noise in Uniform Avalanche Diodes," IEEE Trans. Electron Devices, *ED-13*, No. 1 (January 1966), pp. 164–168.
8. Melchior, H., and Lynch, W. T., "Signal and Noise Response of High Speed Germanium Photodiodes," IEEE Trans. Electron Devices, *ED-13* (December 1966), pp. 829–838.
9. Klauder, J. R., and Sudarshan, E. C. G., *Fundamentals of Quantum Optics*, New York: W. A. Benjamin, Inc., 1968, pp. 169–178.
10. Parzen, E., *Stochastic Processes*, San Francisco: Holden-Day, 1962, p. 156.
11. Figure 5b is from *Transmission Systems for Communication*, Bell Telephone Laboratories, 1970, p. 651.
12. Gillespie, A. B., *Signal, Noise, and Resolution in Nuclear Particle Counters*, New York: Pergamon Press, Inc., 1953.
13. Schade, O. H., Sr., "A Solid-State Low-Noise Preamplifier and Picture-Tube Drive Amplifier for a 60 MHz Video System," RCA Rev., *29*, No. 1 (March 1968), p. 3.
14. Chown, M., and Kao, K. C., "Some Broadband Fiber-System Design Considerations," ICC 1972 Conf. Proc., June 19–21, 1972, Philadelphia, Pa., pp. 12–1, 12–5.
15. Ross, M., *Laser Receivers*, New York: John Wiley and Sons, 1967, p. 328.
16. Edwards, B. N., "Optimization of Preamplifiers for Detection of Short Light Pulses with Photodiodes," Appl. Opt. *5*, No. 9 (September 1966), pp. 1423–1425.
17. Mathur, D. P., McIntyre, R. J., and Webb, P. P., "A New Germanium Photodiode with Extended Long Wavelength Response," Appl. Opt., *9*, No. 8 (August 1970), pp. 1842–1847.
18. Goell, J. E., work to be presented at the Conference on Laser Engineering and Applications (CLEA) Washington, D.C., May 30–June 2, 1973.

Receiver Design for Digital Fiber Optic Communication Systems, II

By S. D. PERSONICK

This paper applies the results of Part I to specific receivers in order to obtain numerical results. The general explicit formulas for the required optical average power to achieve a desired error rate are summarized. A specific receiver is considered and the optical power requirements solved for. The parameters defining this receiver (e.g., bit rate, bias resistance, dark current, etc.) are then varied, and the effects on the required optical power are plotted.

I. INTRODUCTION

This paper will apply the theory of Part I to illustrate in detail how the required received optical power in a digital fiber optic repeater varies with the parameters such as the desired error rate, the thermal noise sources, the bit rate, detector dark current, imperfect modulation, etc. We shall begin by first applying the formulas of Part I to a specific realistic example to obtain reference point. We shall then derive curves of how the required power varies around this point as we vary the system parameters.

II. REVIEW OF RESULTS OF PART I

In Part I we derived explicit formulas for the required optical power at the input of a digital fiber optic communication system repeater to achieve a desired error rate. One formula was applicable when little or no internal (avalanche) detector gain was used, so that thermal noise from the amplifier dominated. The other formula was applicable when optimal gain was being used. These formulas are repeated below:

$$p_{\text{required}} = \frac{QZ^{\frac{1}{2}}}{GT}\frac{\hbar\Omega}{\eta}, \qquad \text{(Thermal Noise Dominates)} \qquad (1)$$

where

$$Z = \left\{ \frac{T}{e^2} \left[S_I + \frac{2k\theta}{R_b} + \frac{S_E}{R_T^2} \right] I_2 + \frac{(2\pi C_T)^2 S_E I_3}{Te^2} \right\} ; \qquad (1a)$$

$$p_{\text{required}} = \frac{1}{2T} (Q)^{(2+x)/(1+x)} [(Z)^{x/(2+2x)}(\gamma_1)^{x/(2+2x)}(\gamma_2)^{(2+x)/(1+x)}] \frac{\hbar\Omega}{\eta} ,$$

(Optimal Avalanche Gain)   (2)

where

$$G_{\text{optimal}} = (Q)^{-1/(1+x)} [(Z)^{1/(2+2x)}(\gamma_1)^{1/(2+2x)}(\gamma_2)^{-1/(1+x)}],$$

where (referring to Fig. 1)

$\eta/\hbar\Omega$ = detector quantum efficiency/energy in a photon
T = interval between bits = $1/$bit rate
G = average detector internal gain
G^x = detector random internal gain excess noise factor
Q = number of noise standard deviations between signal and threshold at receiver output. $Q = 6$ for an error rate of 10^{-9}. (See Fig. 21 in Part I for a graph of error rate vs Q.)
e = electron charge
$k\theta$ = Boltzman's constant·the absolute temperature
R_T = total parallel resistance in shunt with the detector including the physical biasing resistor and the amplifier input resistance
R_b = value of physical detector biasing resistor
C_T = total shunt capacitance across the detector including the shunt capacitance of the detector and that of the amplifier
S_I = amplifier shunt noise source spectral height (two-sided) in amperes²/Hz
S_E = amplifier series noise source spectral height (two-sided) in volts²/Hz.

I_2, I_3, γ_1, and γ_2 are functions only of the shapes of the input optical pulses and the equalized repeater output pulses, where the length of a time slot has been scaled out. These functions are defined in eqs. (23) and (34) of Part I.

Formulas (1) and (2) neglect dark current and assume perfect modulation (received optical pulses completely on or off). We shall investigate deviations from these idealizations later in the paper. For silicon detectors and bit rates above a few megabits per second, these idealizations are reasonable approximations.
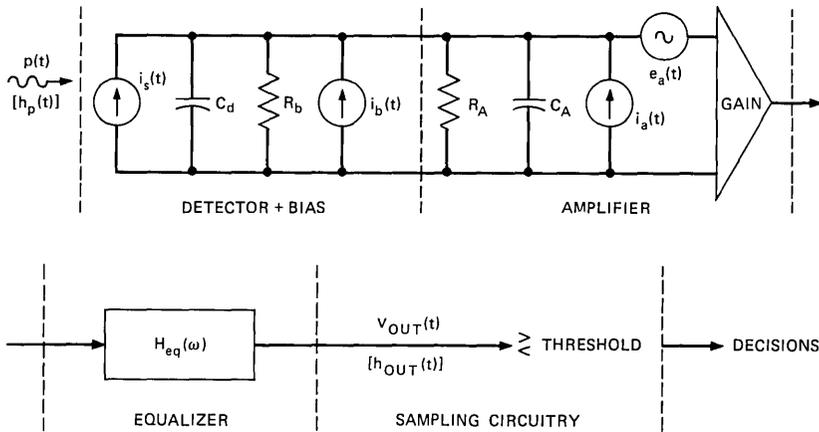
Fig. 1—Receiver.

III. A TYPICAL OPTICAL REPEATER

Consider the following practical optical repeater, operating at a bit rate of 2.5×10^7 bits per second and an error rate of 10^{-9}. The detector is a silicon device with excess noise exponent $x = 0.5$, quantum efficiency 75 percent, dark current before avalanche gain of 100 pico-amperes, and an operating wavelength of 8500 angstroms. The front-end amplifier is a field-effect transistor in a common-source configuration. The total shunt capacitance across the detector is 10 pF. The detector biasing resistor is 1 megohm. The amplifier input resistance is 1 megohm. The amplifier shunt-current noise-source spectral height is equal to the thermal noise of a 1-megohm resistor. The amplifier series-voltage noise-source spectral height is equal to the thermal noise of a conductance with a value equal to the transistor transconductance, g_m, which is 5000 micromhos. The received optical pulses are half-duty-cycle rectangular pulses. The desired equalized output pulse is a raised cosine pulse [see Part I, eq. (25)] with parameter $\beta = 1$.

We must first calculate the value of Q which depends only upon the desired error rate. From Part I, Fig. 21, we see that for an error rate of 10^{-9}, $Q = 6$.

Next we must obtain the constants I_2, I_3, γ_1, and γ_2. These depend only upon the input optical pulse shape and the equalized output pulse shape. From (23) and (34) of Part I we obtain

$$I_2 = 0.804046, \quad I_3 = 0.071966, \quad \gamma_1 = 21.4106,$$
$$\gamma_2 = 1.25424. \tag{3}$$

Using the above data we obtain the thermal noise parameter Z as follows:

$$Z = \left\{ \left[\frac{4 \times 10^{-8}}{(1.6 \times 10^{-19})^2} \right]_{\overset{\nearrow\; T}{\underset{\swarrow\; e}{}}} \left[8.28 \overset{\nearrow\; 2k\theta}{10^{-21}} \left(10^{-6} + 10^{-6} \right. \right. \right.$$

$$\left. \left. + \frac{4 \times 10^{-12}}{5 \times 10^{-3}} \right) \overset{\nearrow\; I_2}{0.804046} \right]$$

$$\left. + \frac{(2\pi \times 10^{-11})^2 \left(\dfrac{8.28 \times 10^{-21}}{5 \times 10^{-3}} \right) \overset{\nearrow\; I_3}{(0.071966)}}{(1.6 \times 10^{-19})^2 (4 \times 10^{-8})} \right\} = 4.8027 \times 10^5. \quad (4)$$

From the data we have $\hbar\Omega/\eta = 3.117 \times 10^{-19}$ joules.

We obtain from (1), at unity internal gain (no avalanche), $p_{\text{required}} = 3.25 \times 10^{-8}$ watts $= -44.89$ dBm (no gain).

We obtain from (2), at optimal avalanche gain, $p_{\text{required}} = 1.6409 \times 10^{-9}$ watts $= -57.85$ dBm, $G_{\text{optimal}} = 56.89$.

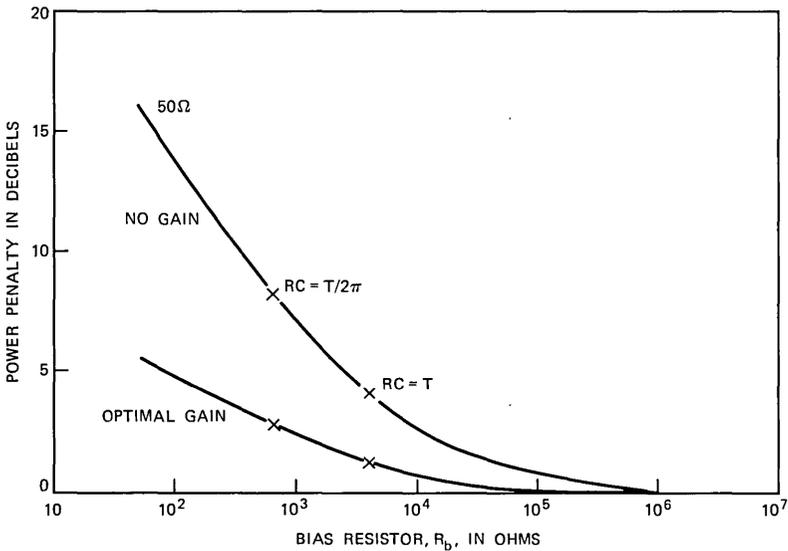

Fig. 2—Required power penalty vs R_b.

We therefore observe that optimal avalanche gain buys a 13-dB reduction in required optical power in this example. Before proceeding, we can check the validity of neglecting dark current. The average number of primary photoelectrons produced by the signal per pulse interval T is the required optical power multiplied by $\eta T / \hbar \Omega$. When shot noise is important (with avalanche gain) this number is 210 primary signal counts per interval T. The number of dark current counts per interval T is the dark current in amperes multiplied by T/e, which in this example is 25 primary dark current counts. Thus, the shot noise due to the dark current is about 10 percent of the signal shot noise. It is therefore a reasonable approximation to neglect this dark current noise. In Section VI we shall calculate precisely the effect of dark current upon the required optical power.

IV. VARYING THE PARAMETERS

In this section, we shall calculate the effect of varying parameter values used in the example of Section III.

4.1 *Biasing Resistor Value*

It was pointed out in Part I that the biasing resistor R_b should be sufficiently large so that the amplifier series noise source S_E dominates in the expression for Z of (1a). This was in fact the case in the example of Section III. We can calculate the penalty in required optical power for using a smaller biasing resistor. This penalty is plotted in Fig. 2 in dB with zero dB being the penalty associated with an infinitely large biasing resistor. The exact penalty of Fig. 2 is applicable with the other relevant parameters (which make up Z) given in the example above. However, the qualitative conclusions are that significantly more optical power is needed if one adheres to the "$RC = T$" design rather than the "large R" (high-impedance) design, in the absence of avalanche gain. Figure 3 shows how the optimal avalanche gain varies when R_b is changed. The qualitative conclusion is that the "$RC = T$" design requires significantly more avalanche gain that the "large R" (high-impedance) design. It should be pointed out that, for lower bit rates and/or a smaller capacitance C_T, the improvement associated with use of a large R_b rather than a value to keep $R_b C_T = T$ is more pronounced.

4.2 *Desired Error Rate*

As mentioned before, the error rate is coupled to the parameter Q in (1) and (2). Figures 4a and 4b show plots of the variation in the re-
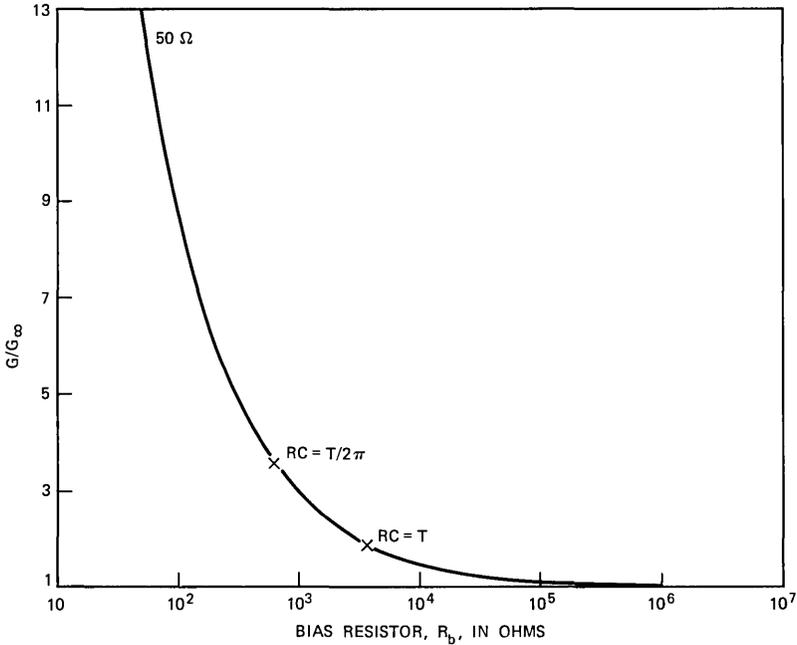
Fig. 3—Optimal gain penalty vs R_b, G_∞ = optimal gain at $R_b = \infty$.

quired power in dB with the desired error rate without gain and with optimal gain. The absolute power in dBm is only applicable to the example of Section III above. However, the difference in required power in dB between any two error rates is applicable in general, as should be apparent from (1) and (2), provided a silicon detector ($x = 0.5$) is being used.

4.3 Bit Rate (1/T)

As mentioned before, the pulse spacing T is scaled out of I_2, I_3, γ_1, and γ_2. These numbers depend only upon the input and output pulse *shapes* (e.g., half-duty-cycle rectangular input pulse, raised-cosine equalized output pulse). Therefore, the effect of the parameter T is explicitly given in (1) and (2) without any hidden dependencies. (This of course assumes that the input pulse shape is not limited by dispersion in the transmission medium and can therefore be held to a half-duty-cycle rectangle.) If we assume that the high-impedance design is being used and that this dominance of the term proportional to $1/T$ in Z of (1a) can be maintained as the bit rate is varied (becomes difficult
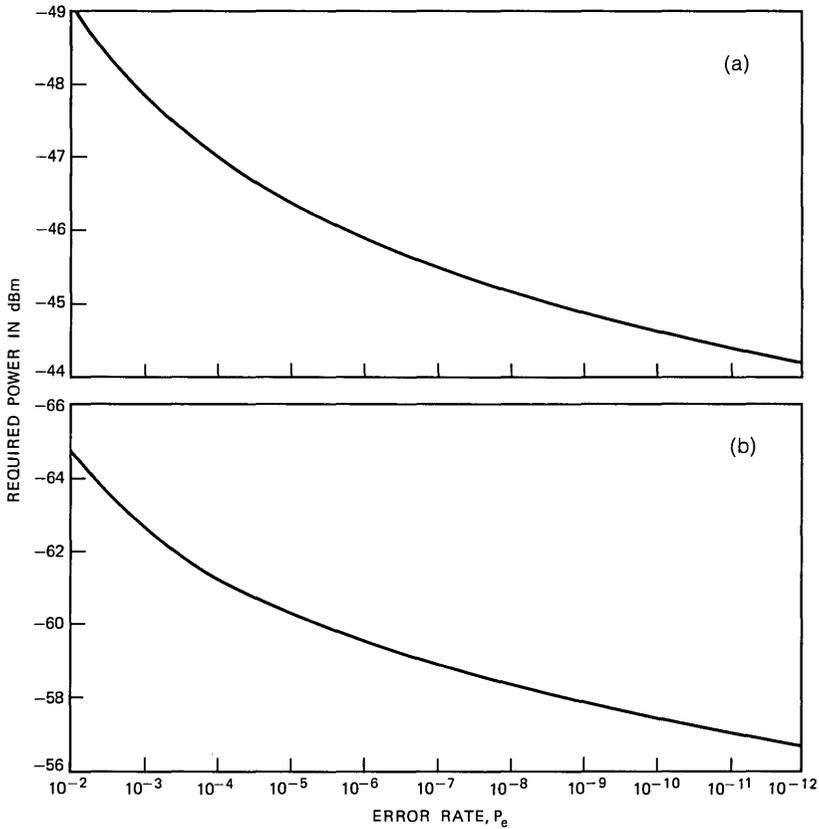
Fig. 4—(a) Required power vs error rate (no avalanche gain). (b) Required power vs error rate (optimal gain).

at low bit rates), then we have the following dependence of the required optical power upon the bit rate $1/T$ without gain and with optimal gain:

$$p_{required} \propto T^{-\frac{1}{2}} \quad \text{(no gain)}$$
$$(4.5 \text{ dB/octave of bit rate}) \tag{5a}$$

$$p_{required} \propto T^{-7/6} \quad \text{(optimal silicon gain)}$$
$$(3.5 \text{ dB/octave of bit rate}) \tag{5b}$$

$$G_{optimal} \propto T^{-\frac{1}{3}}$$
$$(1 \text{ dB/octave of bit rate}).$$

One should be careful extrapolating (5a) and (5b) to very low bit rates. First, the shot noise is no longer negligible compared to the

thermal noise at bit rates where the optimal gain is low. Thus (5a) loses validity at very low bit rates. Further, (5b) is only valid for optimal gains greater than unity. Near unity optimal gain, the silicon excess noise factor departs from $G^{.5}$. In addition, at low bit rates, dark current may not be negligible. It is reasonable to use (5a) and (5b) to extrapolate the results of the example in Section III to bit rates between 5 and 300 Mb/s.

V. THE EFFECT OF IMPERFECT MODULATION

The above formulas (1) and (2) assume that there is perfect modulation. That is, it was assumed that each optical pulse is either completely on or completely off. In this section we shall investigate two versions of imperfect modulation.

Case 1: Pulses Not Completely Extinguished

This case is illustrated in Fig. 5. In each time slot the optical pulse is either completely or partly on. The partly on pulse has the same shape as a completely on pulse, but has area EXT times the area of a completely on pulse. This may correspond to an externally modulated mode-locked laser source. Thus the ratio of the power received when a sequence of all "off" pulses is transmitted to the power received when a sequence of all "on" pulses is transmitted is EXT. Using the results of Part I eqs. (23) and (34), we obtain the following power requirements which are modifications of (1) and (2) above:

$$p_{\text{required}} = \left(\frac{1 + EXT}{1 - EXT} \right) \frac{QZ^{\frac{1}{2}}}{GT} \frac{\hbar\Omega}{\eta}, \qquad \text{(Thermal Noise Dominates)} \quad (6)$$

$$p_{\text{required}} = \frac{1 + EXT}{2T} \left(\frac{Q}{1 - EXT} \right)^{(2+x)/(1+x)}$$

$$\times \left[(Z)^{x/(2+2x)} (\gamma_1')^{x/(2+2x)} (\gamma_2')^{(2+x)/(1+x)} \right] \frac{\hbar\Omega}{\eta}, \qquad \text{(Optimal Gain)} \quad (7)$$

where defining from Part I (23) and (34)

$$I_6 = \Sigma_1 - (1 - EXT)I_1$$

we have

$$\gamma_1' = \frac{- (\Sigma_1 + I_6) + \sqrt{(\Sigma_1 + I_6)^2 + \dfrac{16(1 + x)}{x^2} \Sigma_1 I_6}}{2 \Sigma_1 I_6}$$

$$\gamma_2' = \sqrt{1/\gamma_1' + \Sigma_1} + \sqrt{1/\gamma_1' + I_6}.$$
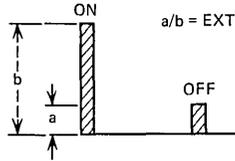
[Compare (6) and (7) to (1) and (2).]

Fig. 5—Imperfect modulation, pulses not completely extinguished.

Using (1), (2), (7), and (8), we can calculate the extra required optical power due to a nonzero value of EXT with and without avalanche gain. When avalanche gain is being used, this power penalty depends upon the input and output pulse shapes. We plot in Fig. 6 the power penalty vs EXT, assuming the pulse shapes of the example in Section III above for the avalanche gain case.

Case 2: Pulses on a Pedestal

This case is illustrated in Fig. 7. The received optical pulses arrive on a pedestal, which may correspond to inability to completely extinguish the light from a modulated source which is not in a pulsing (mode-locked) condition. We set the ratio of average received optical power when all pulses are "off" to average received optical power when all
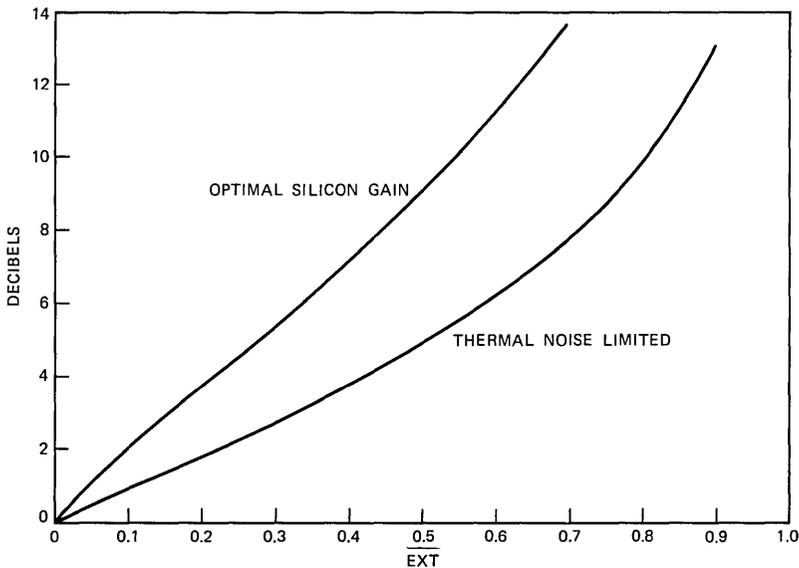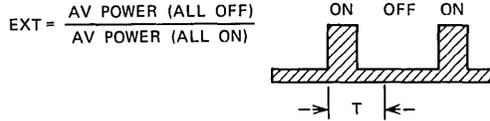

Fig. 6—EXT penalty (dB) vs EXT.

$$EXT = \frac{\text{AV POWER (ALL OFF)}}{\text{AV POWER (ALL ON)}}$$

Fig. 7—Imperfect modulation, pulses on pedestal.

pulses are "on" to be EXT in analogy to Case 1 above. This ratio will remain fixed if the pulse changes in propagation from transmitter to receiver. Using the results of Part I we obtain the following formulas for the required optical power:

$$p_{\text{required}} = \frac{1 + EXT}{1 - EXT} \frac{QZ^{\frac{1}{2}}}{GT} \frac{\hbar\Omega}{\eta}, \qquad \text{(Thermal Noise Dominates)} \qquad (8)$$

$$p_{\text{required}} = \frac{1 + EXT}{1 - EXT} \frac{(Q)^{(2+x)/(1+x)}}{2T}$$

$$\times \left[(Z)^{x/(2+2x)} (\gamma_1'')^{x/(2+2x)} (\gamma_2'')^{(2+x)/(1+x)} \right] \frac{\hbar\Omega}{\eta}, \qquad \text{(Optimal Gain)} \qquad (9)$$

where defining from Part I (23) and (34)

$$\Sigma_1' = \Sigma_1 + \left(\frac{EXT}{1 - EXT} \right) I_2$$

$$I_7 = \Sigma_1 - I_1 + \left(\frac{EXT}{1 - EXT} \right) I_2$$

we have

$$\gamma_1'' = \frac{-(\Sigma_1' + I_7) + \sqrt{(\Sigma_1' + I_7)^2 + \dfrac{16(1 + x)}{x^2} (\Sigma_1')I_7}}{2(\Sigma_1')I_7}$$

$$\gamma_2'' = \sqrt{1/\gamma_1'' + \Sigma_1'} + \sqrt{1/\gamma_1'' + I_7}.$$

Once again we can use (1), (2), (8), and (9) to calculate the penalty for nonzero extinction. This penalty is plotted in Fig. 8 vs EXT where we assume the input and output pulse shapes of the example in Section III when there is optimal avalanche gain.

VI. THE EFFECT OF DARK CURRENT

In order to allow for dark current, we must solve the following set of simultaneous equations which treat the dark current as an equivalent pedestal-type nonzero extinction. (When thermal noise dominates, dark current is either negligible or its shot noise can be added trivially

to the amplifier parallel current noise source S_I.)

$$p_{\text{required}} = \frac{Q^{(2+x)/(1+x)}}{2T} (Z)^{x/(2+2x)}(\gamma_1''')^{x/(2+2x)}(\gamma_2''')^{(2+x)/(1+x)} \frac{\hbar\Omega}{\eta} ,$$

$$\text{(At optimal avalanche gain)} \quad (10)$$

where defining from Part I (23) and (34)

$$\Sigma_1'' = \Sigma_1 + \delta I_2$$
$$I_8 = \Sigma_1 - I_1 + \delta I_2$$

we have

$$\gamma_1''' = \frac{-(\Sigma_1'' + I_8) + \sqrt{(\Sigma_1'' + I_8)^2 + \dfrac{16(1+x)}{x^2}\Sigma_1'' I_8}}{2\Sigma_1'' I_8}$$

$$\gamma_2''' = \sqrt{1/\gamma_1''' + \Sigma_1''} + \sqrt{1/\gamma_1''' + I_8}$$

$$i_d = (2p_{\text{required}})\frac{\eta e\delta}{\hbar\Omega} , \qquad (11)$$

where i_d = primary dark current in amperes.

There are various ways to solve (10) and (11) simultaneously. One way is to solve (10) first with $\delta = 0$ for p_{required}. Then one can solve
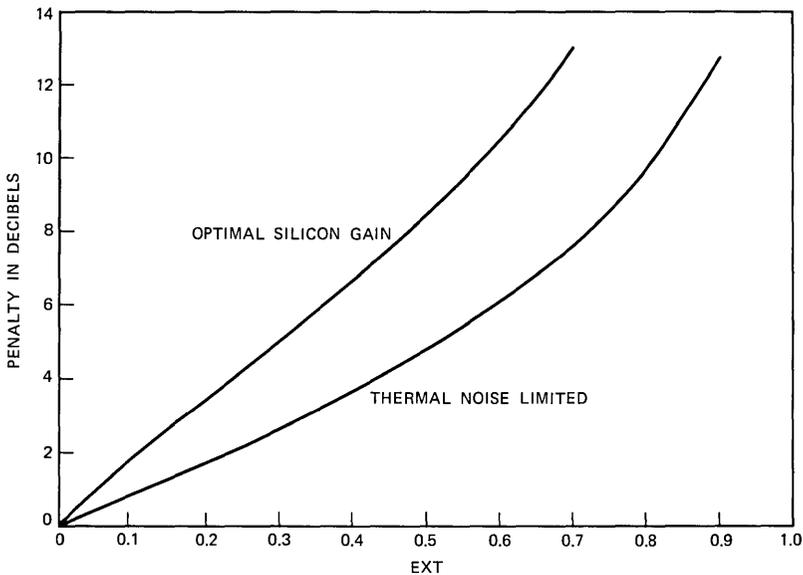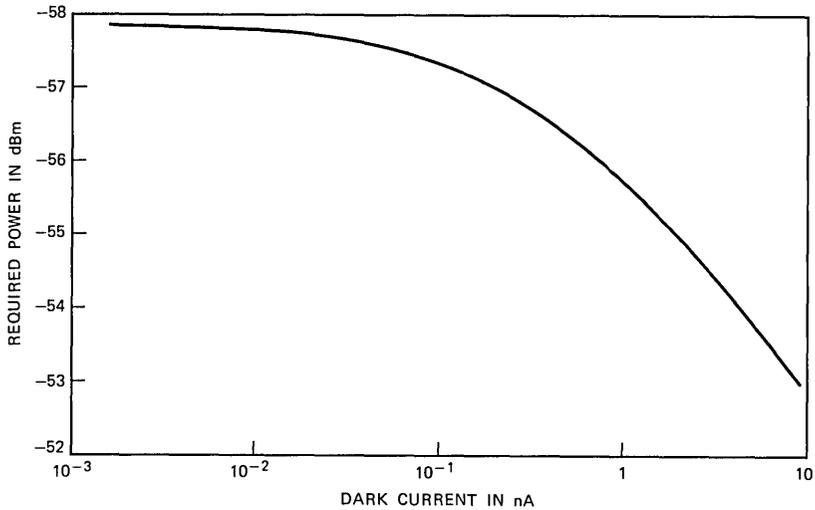

Fig. 8—Power penalty vs EXT.

Fig. 9—Required power vs dark current.

(11) for a new value of δ. One then resolves (10) and then (11), etc.–
repeating the iterations until satisfactory convergence is obtained.
Figure 9 shows a plot of the required power in dBm for the example in
Section III vs dark current in nanoamperes. We see that a dark current
of 100 picoamperes results in an optical power requirement which is
about 0.5 dB more than that which would be required with zero dark
current. Thus it was reasonable to neglect dark current when calculat-
ing the required power in Section III. Dark current will result in even
less of a penalty at higher bit rates. Although the curve of Fig. 9 is ap-
plied to the specific example of Section III, it is apparent in general
that, at bit rates above a few megabits per second and with primary
dark currents less than 0.1 nanoampere, dark current will have a small
effect upon the required optical power.

Gain-Induced Modes in Planar Structures

By W. O. SCHLOSSER

The properties of modes in a slab structure with gain in the center region and loss in the surrounding medium are investigated. The propagation constant and field distribution of the lowest-order modes are determined. The cutoff frequencies and propagation constants of the next-higher mode are given. Furthermore, the effect of a refractive index depression or increase in the center region is determined. The depression does not destroy the mode, as may be expected, but causes it to have a cutoff frequency. Comparison of these results to the experimental data shows that gain-induced modes play an important part in the lateral confinement in stripe-geometry GaAs lasers.

I. INTRODUCTION

Modes in cylindrical structures with refractive index boundaries are well known. Their existence and basic properties are generally visualized by superposition of plane waves reflected at the index boundaries. Much less is known about modes in structures with spatially nonuniform gain or attenuation.[1,2] However, it seems intuitively possible that some kind of mode should also exist in that case; indeed, Kogelnik[3] showed that a cylindrical structure with a radial gain profile can support a Gaussian beam of constant diameter even if there are no refractive index differences present. Evidently, the nonuniform transparency of the medium counteracts the natural tendency of the beam to spread.

In this paper we will consider a planar structure with stepwise discontinuities of gain or absorption. This geometry is of considerable practical interest. As will be shown in this paper, the lateral confinement of the optical field in a stripe-geometry GaAs laser is due to the gain-loss interface at the edge of the stripe. Furthermore, it is easy to create nonuniform gain distributions in planar structures either by masking and optically pumping or by nonuniform injection current distribution. The modes induced by these gain distributions can be easily influenced from the outside by changing the pumping intensity or injection current.

II. GAIN-INDUCED MODES IN A THREE-LAYER STRUCTURE

The physical configuration with which we will concern ourselves in this paper is shown in Fig. 1. All material parameters are assumed to be constant throughout the regions and differ only at the boundaries $x = \pm a$. As is well known from the theory of refractive guiding,[4] the modes can be divided into four classes: TE even and odd in x direction and TM even and odd in x. The two lowest-order even modes have no cutoff frequency, whereas the odd modes can be guided only above a certain cutoff frequency. For our application, the lowest-order even modes are of greatest interest and we will deal with them first.

Before going into details, let us explain the notation which we will use subsequently.

(i) The complex relative dielectric constant ϵ is split into refractive index n and extinction coefficient k according to

$$\epsilon = (n + jk)^2. \qquad (1)$$

(ii) u determines the x dependence of the fields inside the center layer $|x| \leq a$, i.e., they have the functional form

$$\begin{matrix} \sin \\ \cos \end{matrix} \left(u \frac{x}{a} \right).$$

(iii) w gives the dependence of the fields in the cladding $|x| \geq a$ which follow the function $e^{-w|x/a|}$. u and w are related by

$$u^2 + w^2 = (\epsilon_1 - \epsilon_2)(k_0 a)^2. \qquad (2)$$

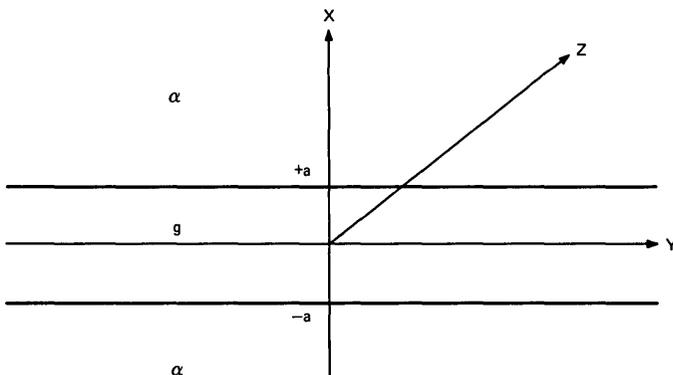With these quantities we can derive the characteristic equations (the

Fig. 1—Cross section of slab waveguide.

field components are listed in Appendix A) for the even modes.

<div align="center">

TE Mode	TM Mode

</div>

$$w = u \tan u \qquad \frac{\epsilon_1}{\epsilon_2} w = u \tan u \qquad (3)$$

They are formally the same as in the case of purely reactive modes but, since ϵ_1 and ϵ_2 are complex, the solutions will be complex.

We determine under which conditions eqs. (3) have a solution. We restrict ourselves to cases where the fields cannot increase exponentially with increasing distance from the interface nor can there be any movement of wavefronts from infinity toward the guiding structure. This limits w to the first quadrant. It is then easy to derive a necessary but not sufficient condition from eqs. (2) and (3): $\epsilon_1 - \epsilon_2$ has to stay in the first or second quadrant, i.e., Im $(\epsilon_1 - \epsilon_2) > 0$, which is equivalent to

$$n_1 k_1 - n_2 k_2 > 0. \qquad (4)$$

Let us interpret this inequality for some special cases:

  (i) $k_1 = k_2$. Condition (4) reduces to $n_1 > n_2$ which is the well-known requirement for refractive guiding.

  (ii) $n_1 = n_2$. The condition (4) now reads $k_1 > k_2$, equivalent to saying that the center region should be more transparent than the sides, which agrees with physical intuition.

  (iii) $n_1 = n_2 + \Delta n$, where Δn is small compared to n_2 but not necessarily small compared to k. In this case (4) can be expressed by

$$\frac{\Delta k}{k_1} + \frac{\Delta n}{n_2} > 0, \qquad (4a)$$

where $\Delta k = k_1 - k_2$. This inequality shows that there can be modes even if Δn or Δk is negative, as long as the other mechanism is strong enough to generate the mode. However, since (4) is only a necessary condition, this case requires further examination.

The second problem we have to address ourselves to is that of the *cutoff frequency*. For refractive guidance the lowest-order even modes have no cutoff frequency. We will now establish the conditions under which gain-induced modes have no cutoff frequency either. For small values of $(\epsilon_1 - \epsilon_2)(k_0 a)^2$, w and u will be small and the tangent in eq. (3) can be replaced by its argument. The solution of eqs. (2) and (3) is thus $w = (\epsilon_1 - \epsilon_2)(a k_0)^2$ for TE modes and $w = (\epsilon_2/\epsilon_1)$ $\times (\epsilon_1 - \epsilon_2)(a k_0)^2$ for TM modes with $a k_0$ very small. This result shows that the solution of (3) exists no matter how small $(a k_0)$ is, independent of the guide parameters, as long as $\epsilon_1 - \epsilon_2$ is in the

first quadrant. In other words, *a gain-induced mode will have no cutoff frequency as long as there is a refractive index difference to keep* $\epsilon_1 - \epsilon_2$ *in the first quadrant. If there exists a refractive index depression, i.e.,* $Re\ (\epsilon_1 - \epsilon_2) < 0$*, the modes do have a cutoff frequency.*

We will now consider the case of *gain-induced modes without a refractive index change.* We specialize the discussion to small extinction coefficients in the order of 10^{-4}, whereas n is typically greater than one. This covers typical laser applications reasonably well. The characteristic equations for TE and TM modes [eqs. (3)] differ only by a factor of $\epsilon_1/\epsilon_2 = 1 + j(2/n)(k_1 - k_2)$ which is sufficiently close to one to cause only a negligible perturbation. In the following, we will therefore neglect the difference between TE and TM modes.

To simplify the discussion we define two new quantities: A normalized frequency is given by

$$v = ak_0\sqrt{\epsilon_1 - \epsilon_2}.$$ (5)

For gain-induced modes, v reduces to the form

$$v = ak_0 n \sqrt{j2\frac{(k_1 - k_2)}{n}},$$ (6)

which means that in this particular case the phase angle of v is independent of the material parameters. It is furthermore customary[5] to use a normalized propagation constant b

$$b = \frac{(\beta_z/k_0)^2 - \epsilon_2}{\epsilon_1 - \epsilon_2} \quad \text{or} \quad \left(\frac{\beta_z}{k_0}\right)^2 = \epsilon_2 + (\epsilon_1 - \epsilon_2)b.$$ (7)

Since in our application $\epsilon_1 - \epsilon_2$ is a small quantity, $(\beta_z/k_0)^2$ is always equal to ϵ_2 plus a small perturbation, whereas the variation of $|b|$ is in the order of one, thus alleviating some computational problems. The two parameters u and w are related to v and b by $u = v\sqrt{1 - b}$ and $w = v\sqrt{b}$.

With these new quantities, the two characteristic eqs. (3) are transformed into

$$\sqrt{b} = \sqrt{1 - b} \tan (v\sqrt{1 - b}).$$ (8)

It should be noted that the normalized propagation constant b is exclusively a function of v. It is, therefore, only necessary to solve the characteristic eq. (8) once to cover all material parameters and dimensions, assuming, of course, the validity of the initial assumptions.

The solution of the characteristic equation was done on the computer. Figure 2 shows the normalized propagation constant b for the lowest-order mode as a function of $|v|$. We can interpret b more easily
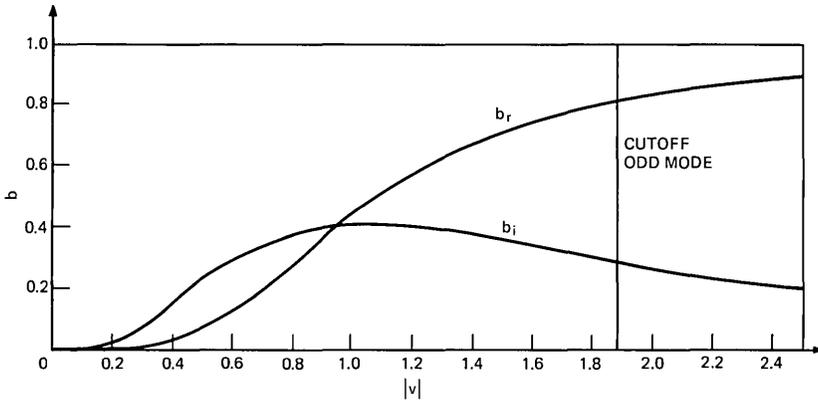
Fig. 2—$b(|v|)$ for lowest-order even mode ($n_1 = n_2$).

if we rewrite the propagation constant β_z [eq. (7)] to show its real and imaginary parts.

$$\frac{\beta_z}{k_0} \approx n - b_i(k_1 - k_2) + j(k_2(1 - b_r) + k_1 b_r). \qquad (9)$$

The attenuation or gain of the mode as represented by the imaginary part of β_z/k_0 is solely dependent on b_r, which varies between 0 and 1. For small values of b_r, the attenuation will be determined by the outer medium and only for higher values of b_r will the gain in the center region be of significance. This situation corresponds very closely to the case of refractive guidance. For further reference we have included in Fig. 3 the field distribution with $|v|$ as parameter.

For any practical application, it is important to know up to which frequency or dimension the guide is single moded. We will therefore determine the cutoff frequency of the lowest-order odd mode and of the first-order even mode. We define the cutoff value of b as the one at which the radiation condition just ceases to be fulfilled. It turns out that this is the case when w is purely imaginary, i.e., only power is radiated away from the guiding structure, but the amplitude does not decrease with increasing $|x|$. Re $(w) = 0$ is equivalent with the condition Re $(b) = 0$ [eq. (7)]. The solution of this problem has to be found on the computer. The results are

$$|v_{\text{odd}}| = 1.877 \qquad |v_{\text{even}}| = 2.759.$$

For future reference we have included $b(|v|)$ for the first-order even and the lowest-order odd mode (Figs. 4 and 5).
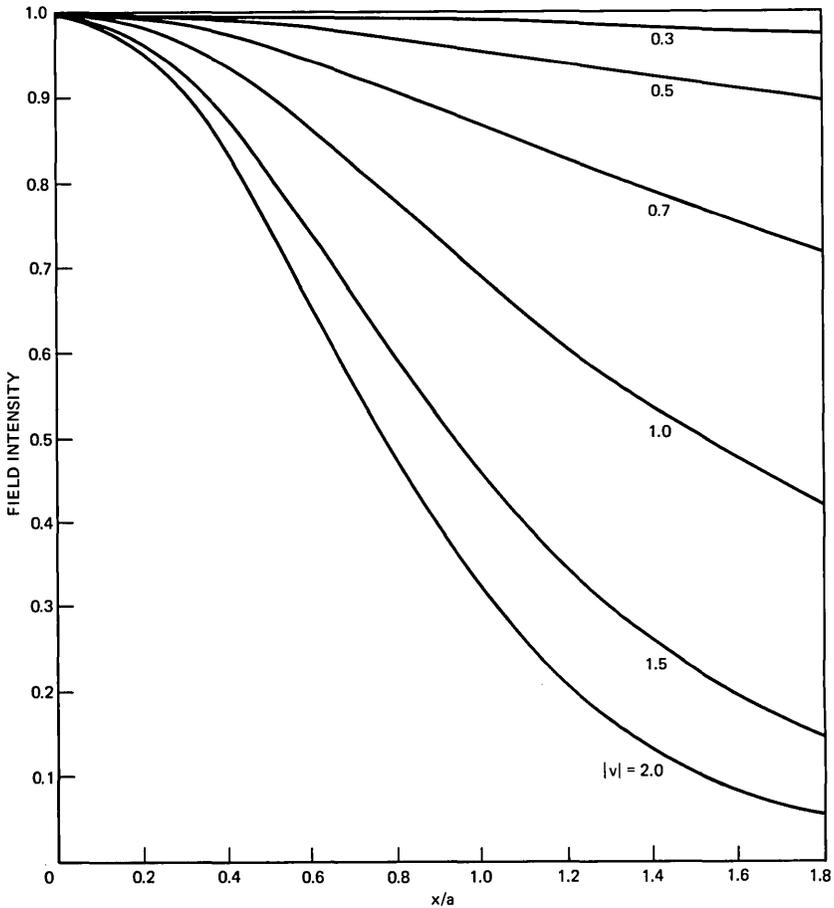
Fig. 3—Field intensity for lowest-order even mode.

III. THE INFLUENCE OF ADDITIONAL REFRACTIVE INDEX DIFFERENCES

In general, if there is a gain or loss difference between two regions, the refractive index will be different as well. Gain and refractive index are related by the Kramers-Kronig relations and thus the presence of gain can change the refractive index. Furthermore, in the case of the injection laser, the injected carriers change the refractive index due to the plasma resonance. In this particular case, the gain region can be expected to have a lower refractive index than the side regions. It is thus necessary to explore the effects of refractive index increases and depressions in the center region in conjunction with gain-loss differences.
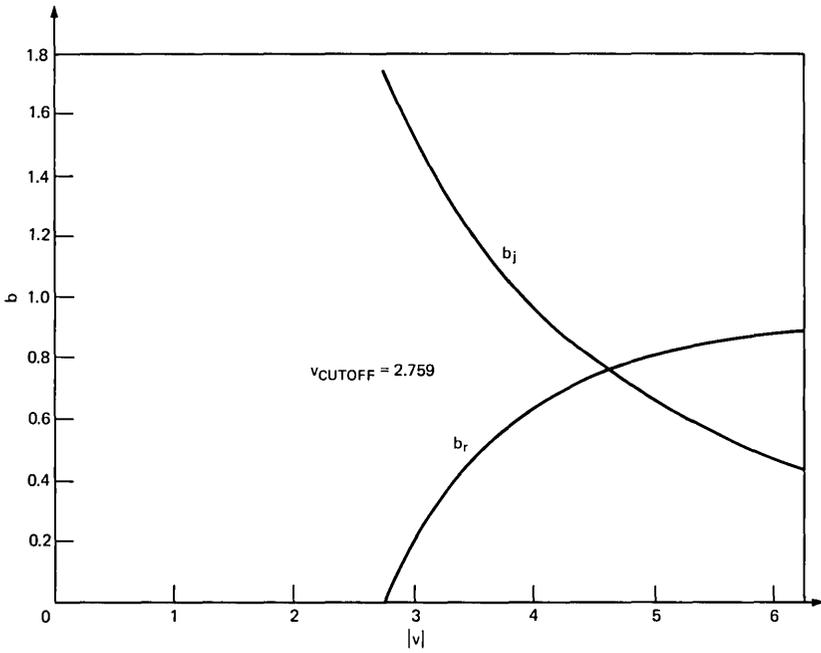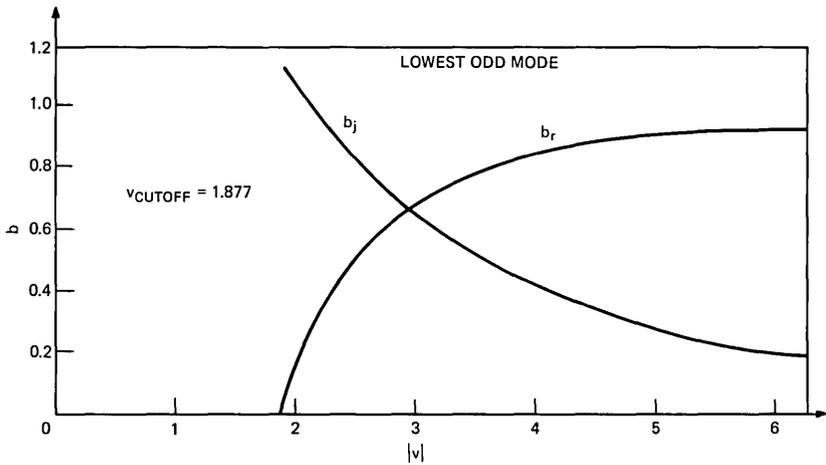
Fig. 4—$b(|v|)$ for first-order even mode.

Fig. 5—$b(|v|)$ for lowest-order odd mode.

If there is an increase in refractive index in the center, we can expect the confinement of the energy to increase with increasing index difference. However, in the case of a refractive index depression in the center, the two confining mechanisms will counteract each other.

Let us consider the effect of *increased* refractive index in detail. In contrast to the case of gain-induced modes where the phase of v was $\pi/4$ [eq. (6)] independent of material parameters, it is now dependent on the complex difference $\epsilon_1 - \epsilon_2$. This precludes a representation of the guiding parameters independently of the waveguide parameters as it was possible in the previous case. We are thus forced to determine a solution of the characteristic equation for each set of parameters. It is therefore of importance to develop approximations for the propagation constant. We treat first the case, $\Delta k/\Delta n \ll 1$. We regard the waveguide as a perturbed refractive guide. The $b(v)$ characteristic is plotted in Fig. 6. Obviously, both v and b are real. A small perturbation Δk has two effects: (*i*) the characteristic eq. (8) becomes complex and yields an imaginary part b_i, whereas the real part is to a first order unperturbed; and (*ii*) the propagation constant (9) is now approximated by

$$\frac{\beta_z}{k_0} \approx n + \Delta n b_r - b_i \Delta k + j(k_2 + b_r \Delta k + b_i \Delta n), \qquad (10)$$

i.e., there appears a second imaginary component $b_i \Delta n$, which acts like a gain since Δn and b_i are both positive. This effect is due to the improvement in guiding by the gain-loss mechanism. The calculations are straightforward and are listed in Appendix B. We note here only the result:

$$b_i = \frac{\Delta k}{\Delta n} \frac{(1 - b_0)v_0\sqrt{b_0}}{1 + v_0\sqrt{b_0}}, \qquad (11)$$

where the quantities indexed with a zero denote the unperturbed state. In particular, $v_0 = ak_0n\sqrt{2\Delta n/n}$. We notice that b_i contains a factor $1 - b_0$, which decreases with b_0 approaching unity. Inserting (11) into eq. (10) yields an expression for the loss (or gain) of the mode:

$$\alpha \approx k_0 \left[k_2 + \Delta k \left(b_0 + \frac{(1 - b_0)v_0\sqrt{b_0}}{1 + v_0\sqrt{b_0}} \right) \right]. \qquad (12)$$

We note that the function

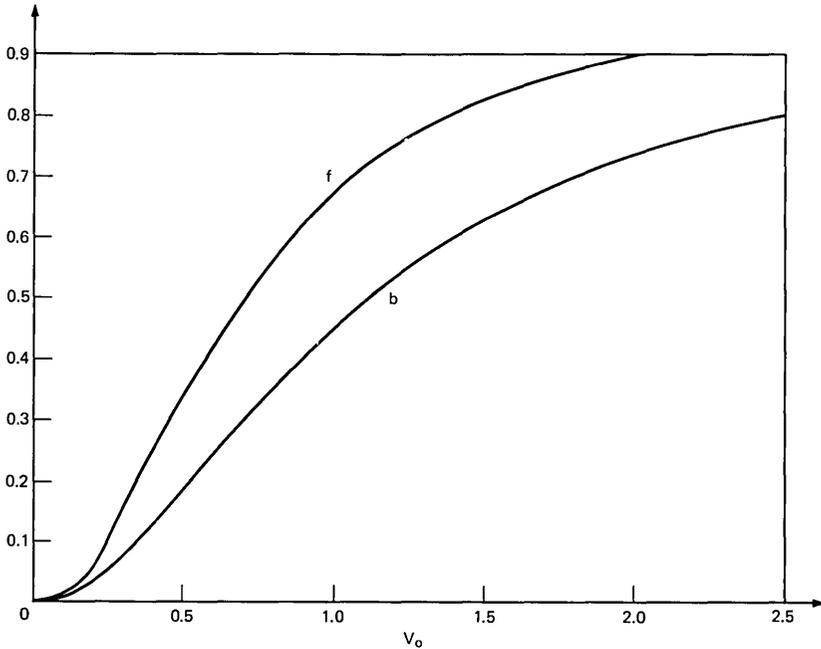$$f(v_0) = b_0 + \frac{(1 - b_0)v_0\sqrt{b_0}}{1 + v_0\sqrt{b_0}}$$

Fig. 6—$b(v)$ for lowest-order even mode in a refractive slab guide and loss-weighing function $f(v_0)$ for lowest-order even mode.

associated with Δk is only dependent on v_0 and we have thus again achieved a representation valid for all waveguide parameters. Figure 6 shows that $f(v_0)$ approaches unity much faster than $b_0(v_0)$; the loss is approaching the loss or gain in the center region indicating improved confinement due to the gain-loss difference.

We will now consider the refractive index *depression* in the center region. From our discussion of cutoff frequencies, we know that modes in this case do have a cutoff frequency if they exist at all. It can be determined from the condition that $w = v\sqrt{b}$ is purely imaginary.

In Fig. 7 real and imaginary parts of b are plotted as functions of s/λ with the parameters listed in the figure caption. Since b_r is larger than one, eq. (10) shows that the effective refractive index is smaller than that of either medium (Δn is negative). The phase velocity is therefore larger than that of a plane wave in either medium, as in metallic waveguides. Altogether, this is quite a different behavior from the ordinary dielectric waveguide. One would expect the index depression to counteract the confinement and, if the depression is strong enough, to destroy it completely. As discussed before, this is only the
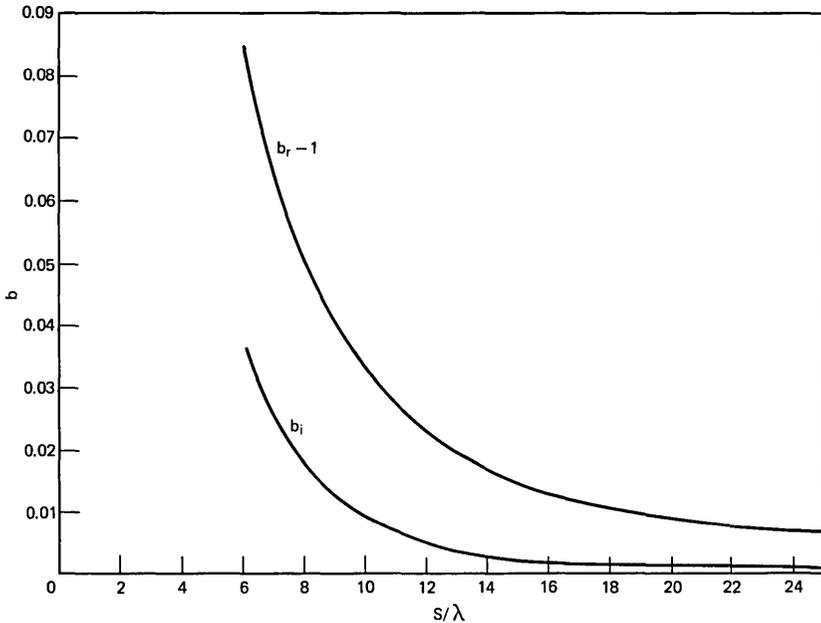
Fig. 7—$b(s/\lambda)$ for lowest-order mode for a refractive index depression $\Delta n = 10^{-2}$, $n = 3.56$, $k_1 = 2.10^{-4}$, $k_2 = 10^{-4}$.

case below the cutoff frequency. To ensure that the solution of the characteristic equation is real and not a freak of the computer program, we have derived an approximate solution for $\Delta k/\Delta n \ll 1$ in Appendix C. The normalized propagation constant is given by

$$b = 1 + \frac{\pi^2}{4v^2}\left\{1 + j\frac{2}{v}\left(1 + v\frac{\Delta k}{2|\Delta n|}\right)\right\},\tag{13}$$

where

$$v = \frac{s}{\lambda}\pi n\sqrt{\frac{2|\Delta n|}{n}}.$$

IV. GAIN-INDUCED MODES IN STRIPE-GEOMETRY GaAs LASERS

The function of the stripe geometry has been viewed as selecting a filament and preventing others from forming.[6] We will now show that, in contrast to a laser with a wide area contact, the stripe geometry does provide a confining mechanism for the optical power and is not merely "selecting" the filament.

In stripe-geometry lasers, the flow of carriers is confined in a stripe region parallel to the junction (laterally). This is done by proton bom-

barding the adjacent regions.[7] The proton bombardment and subsequent annealing alter the free-carrier density but not the optical properties of the material.[8] There are no deliberately built-in refractive index changes between the active and the passive regions. The experiments show that with stripe widths of ~ 12 μm the lasers operate generally in a single lateral mode; the origin of this confinement has not yet been explained. Figure 8a gives a linear scan of a typical nearfield distribution. However, at high current levels or wider stripe widths, numerous deviations from such a single-mode field distribution occur. Figure 8b gives an example of the nearfield in such a case. The field still fills essentially the whole stripe width but is highly nonuniform. Frequently the intensity maximum changes its position with changing current. It is clear that the simple case we have been analyzing so far cannot explain these effects. However, not enough knowledge is available at this time about the numerous parameters involved. We therefore try only to isolate the common properties of the majority of stripe-geometry lasers and explain them in terms of gain-induced modes. We had three major facts to consider:

(i) For stripe widths in the order of 12 μm and less, there is generally a single mode for current levels close to threshold.

(ii) At stripe widths above ~ 18 μm, the field distribution is very often nonuniform even at threshold.

(iii) The threshold current increases steeply with decreasing stripe width.

Since we know that there must be a gain-loss difference between the stripe and the surrounding regions, we will apply the previous results on gain-induced modes and investigate if the experimental results can be explained by this effect or if additional mechanisms have to be invoked. In doing so we must keep in mind that we can expect the agreement between experiment and this theory to be only qualitative since a number of effects are neglected. We mention the few most obvious:

(i) The loss in the layers above and below the active region contributes to the overall loss. This contribution is dependent on stripe width.

(ii) The gain depends on the field intensity and is therefore not constant.

(iii) The mirror loss may depend on the stripe width.

The mirror reflectivity R of the lowest-order mode was determined to
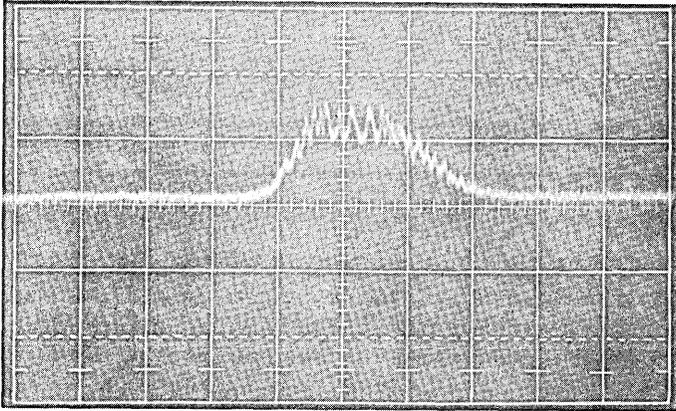
(a)
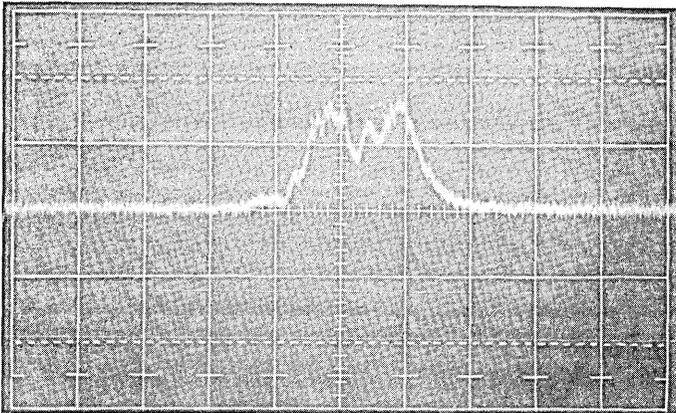
(b)

Fig. 8—Linear scan of nearfield distribution of stripe geometry lasers with stripe width of $\sim$12 μm. One division $\approx$3.7 μm. Current in both cases is $\sim$20 percent above threshold.

be $\sim$0.3 and hence the mirror loss, defined by $\alpha_M = 1/L \ln 1/R$, to be $\cong$10/cm for $L = 400$ μm sample length.[9] The attenuation of GaAs is $\sim$10/cm. We ask now how much gain would one need to reach threshold with an active region width of $\sim$12 μm. In the following, we will call gain g the excess gain over the intrinsic attenuation in the active region. To reach threshold, the effective gain of the lowest-order mode $(gb_r - \alpha(1 - b_r)$ from eq. (9)) must equal the mirror losses:
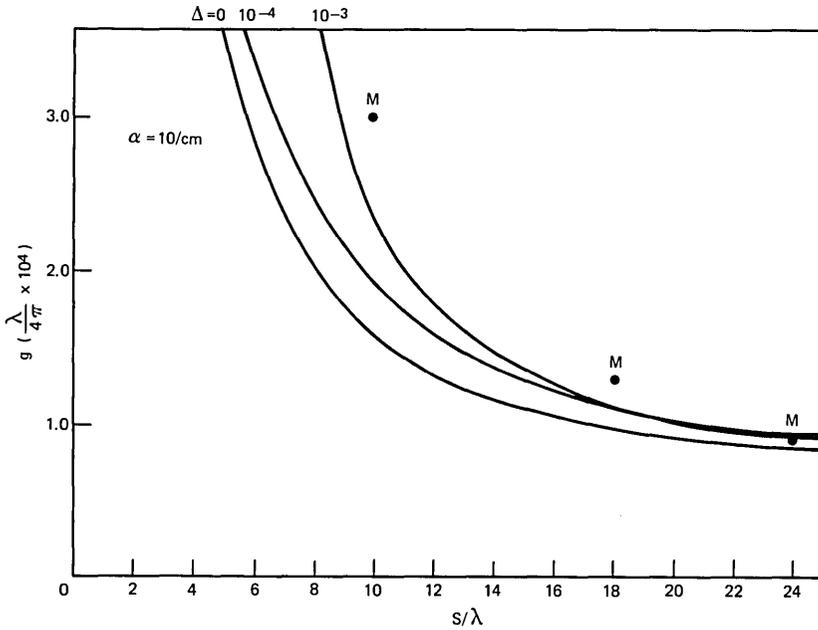
$$gb_r - \alpha(1 - b_r) = \alpha_M, \tag{14}$$

Fig. 9—Threshold gain as function of stripe width. Mirror loss $\alpha_M = 10/\text{cm}$, $n = 3.56$, $\alpha = 10/\text{cm}$, and refractive index depressions $\Delta n = 10^{-3}$, 10^{-4}.

where α is the loss in the side regions. b_r—as pointed out before–is exclusively a function of $|v|$ (Fig. 2). $g\lambda/4\pi$ is plotted as a function of stripe width in Fig. 9. We can see that the gain value corresponding to ~ 12 μm stripe width is quite realistic and we conclude that the gain-induced guiding must play at least some role in the stripe-geometry confinement. As a further piece of experimental evidence, it has been shown by Dyment[10] that the threshold current in a stripe-geometry laser increases strongly with decreasing stripe width. There is, unfortunately, some uncertainty about the relationship between current and gain. Generally, an exponential dependency is assumed ($g \sim J^q$, where q varies between 1.5 and 3). We have used $q = 2$ to insert Dyment's measured results into Fig. 9. We have furthermore introduced $g(s/\lambda)$ curves for refractive index depressions of 10^{-4} and 10^{-3} which correspond to carrier densities of 10^{18} and $10^{19}/\text{cm}$, respectively. The agreement is satisfactory considering the accuracy of the measurements. It therefore seems reasonable to conclude that the gain-induced modes provide the dominant confining mechanism in stripe-geometry lasers. In view of this conclusion we will now derive a few properties of the gain-induced modes which should be useful for predicting some

properties of the stripe-geometry lasers. Wherever the ouput of the laser is to be coupled into an optical system, it is desirable to reduce the stripe width without increase in threshold current. Equation (14) was therefore evaluated for various loss values in the side region. Figure 10 shows that–at least for zero refractive index difference–α has very little influence on the necessary gain. Evidently the increased loss in the higher α case is compensated for by improved guidance.

A refractive index depression generates a gain-width relationship of the form [eq. (13)]

$$ g = \alpha_M + \frac{k_0}{4\pi n^2} \frac{1}{(s/\lambda)^3} \frac{1}{\sqrt{\frac{2\Delta n}{n}}} . \tag{15} $$

It is quite evident that it is very difficult to compensate for the very strong s/λ dependence by proper choice of Δn. It is, however, possible to reduce the stripe width with a refractive index increase, in which case the effective gain is given by [eq. (12)]

$$ g = \frac{\alpha_M + \alpha(1 - f(v))}{f(v)} , \tag{16} $$

where $f(v)$ is plotted in Fig. 6. However, at present, the technological difficulties of this approach have not been solved.
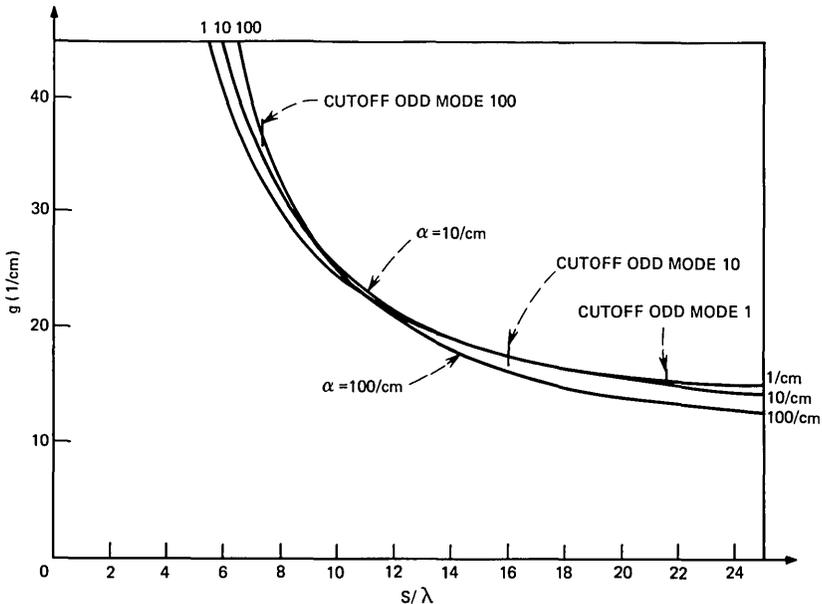

Fig. 10—Threshold gain as function of stripe widths for different attenuation α in the side regions. $\alpha_M = 12.5/\text{cm}$.

V. CONCLUSION

Modes caused by gain attenuation differences in a slab structure have been investigated. They can exist without refractive index differences, in which case the lowest-order even modes do not have a cutoff frequency. The cutoff frequencies for the next-higher modes are given. The influence of additional refractive index increase or decrease in the center region is studied. One particularly interesting result is that, in the presence of gain, a refractive index depression does not remove the mode, but it causes it to have a cutoff frequency.

An application of these results to the problem of lateral confinement in stripe-geometry lasers shows that gain-induced modes are mainly responsible for the optical confinement. Thus the stripe geometry does not just select a filament, but provides a confining mechanism for the optical power. (A similar conclusion was reached independently by F. R. Nash.[11]) The sharp increase of the threshold current with reduced stripe width is a direct consequence of this gain-induced confinement. It therefore appears impossible to reduce the stripe width significantly below $\sim 10 \ \mu\mathrm{m}$ without a substantial sacrifice in the threshold current. A deliberately built-in refractive index increase in the center region should alleviate this situation.

APPENDIX A

Field Components in Slab Guide

<table>
<tr><th>TE</th><th>TM</th></tr>
</table>

$$|x| \leqq a$$

$$E_y = A\omega\mu_o \cos u \frac{x}{a} e^{-j\beta_z z} \qquad\qquad H_y = A\omega\epsilon_0\epsilon_1 \cos u \frac{x}{a} e^{-j\beta_z z}$$

$$H_x = -A\beta_z \cos u \frac{x}{a} e^{-j\beta_z z} \qquad\qquad E_x = A\beta_z \cos u \frac{x}{a} e^{-j\beta_z z}$$

$$H_z = -jA \frac{u}{a} \sin u \frac{x}{a} e^{-j\beta_z z} \qquad\qquad E_z = jA \frac{u}{a} \sin u \frac{x}{a} e^{-j\beta_z z}$$

$$|x| \geqq a$$

$$E_y = A\omega\mu_0 \cos u e^{-[j\beta_z z + w(|x/a|-1)]} \qquad H_y = A\omega\epsilon_0\epsilon_2 \cos u e^{-[j\beta_z z + w(|x/a|-1)]}$$

$$H_x = -A\beta_z \cos u e^{-[j\beta_z z + w(|x/a|-1)]} \qquad E_x = A\beta_z \frac{\epsilon_1}{\epsilon_2} \cos u e^{-[j\beta_z z + w(|x/a|-1)]}$$

$$H_z = \mp jA \frac{w}{a} \cos u e^{-[j\beta_z z + w(|x/a|-1)]} \qquad E_z = \pm jA \frac{w}{a} \frac{\epsilon_1}{\epsilon_2}$$

$$\times \cos u e^{-[j\beta_z z + w(|x/a|-1)]}$$

The notations are explained in the list of notations at the end of the paper. The separation parameters are related to each other by

$$(\beta_z a)^2 + u^2 = \epsilon_1 (k_0 a)^2, \qquad (\beta_z a)^2 - w^2 = \epsilon_2 (k_0 a)^2,$$

and $u^2 + w^2 = (\epsilon_1 - \epsilon_2)(k_0 a)^2$.

APPENDIX B

Approximate Solution of Characteristic Equation for Refractive Index Increase in Outer Region

We assume that $\Delta k / \Delta n \ll 1$, i.e., the refractive index step is governing the guiding properties. The normalized frequency v can then be expressed by

$$v \approx \frac{s}{\lambda} \pi n \sqrt{2 \frac{\Delta n}{n}} \left(1 + j \frac{\Delta k}{2 \Delta n} \right). \tag{17}$$

We treat the characteristic equation for TE modes first [eqs. (3)]. The regular solution for $\Delta k = 0$ will yield real values for u and w. For small $\Delta k / \Delta n$, both will be complex with a small imaginary part. We list the real and imaginary part of the characteristic equation separately, neglecting second-order terms:

$$u_r \sin u_r = w_r \cos u_r \tag{18a}$$

$$u_i [(1 + w_r) \sin u_r + u_r \cos u_r] = w_i \cos u_r. \tag{18b}$$

The real part (18a) is identical to the unperturbed equation and we assume, therefore, u_r and w_r to take the values of u and w, respectively, of the $\Delta k = 0$ case. Combining (18a) and (18b) yields the following relation between u_i and w_i:

$$u_i [u_r^2 + w_r (1 + w_r)] - w_i u_r = 0. \tag{19}$$

Now we have to relate u_i, w_i to b_i, the quantity we really want to determine.

$$w = v \sqrt{b} \qquad u = v \sqrt{1 - b}. \tag{20}$$

With a small imaginary part of b we get

$$w = v_r \sqrt{b_r} \left[1 + j \left(\frac{v_i}{v_r} + \frac{b_i}{2 b_r} \right) \right] \tag{21a}$$

$$u = v_r \sqrt{1 - b_r} \left[1 + j v_r \left(\frac{v_i}{v_r} - \frac{b_i}{2(1 - b_r)} \right) \right]. \tag{21b}$$

These equations show that, if the real parts of w and u are unperturbed to the first order, the real part of b will be unperturbed also. If we intro-

duce eqs. (21) into the characteristic eq. (19), b_i can be determined

$$b_i = \frac{\Delta k}{\Delta n} \frac{(1 - b_r)v_r\sqrt{b_r}}{1 + v_r\sqrt{b_r}}.$$
(22)

The characteristic equation for the TM mode has an additional factor ϵ_1/ϵ_2. It is easy to show that the real part of b is altered by Δb from the TE value b_0

$$\frac{\Delta b_r}{b_0} = - \frac{\Delta n}{n} \frac{4(1 - b)}{1 + w}.$$
(23)

The imaginary part is to a first approximation the same as for the TE mode [eq. (22)].

APPENDIX C

Approximate Solution of Characteristic Equation for Refractive Index Depression in Center Region

Again we assume $\Delta k/|\Delta n| \ll 1$. Δn will be negative. The normalized frequency v is now expressed by

$$v = j \frac{s}{\lambda} n\pi \sqrt{\frac{2|\Delta n|}{n}} \left(1 + j \frac{\Delta k}{2|\Delta n|}\right) = jv_i\left(1 - j \frac{\Delta k}{2|\Delta n|}\right).$$
(24)

Let us assume that u is very close to $\pi/2$,

$$u = \pi/2 + \delta + j\psi,$$
(25)

where δ and ψ are small quantities. Since

$$w = v \sqrt{1 - \left(\frac{u}{v}\right)^2},$$

we use

$$w = v \left[1 - \frac{1}{2}\left(\frac{u}{v}\right)^2\right],$$
(26)

where we note that u/v is small compared to one. The characteristic equation to a first-order approximation takes the form

$$\left(\frac{\pi}{2} + \delta + j\psi\right) + jv_i(\delta + j\psi) = 0.$$
(27)

The solution is

$$\delta \approx - \frac{\pi}{2} \frac{1}{1 + v_i^2} \qquad \psi = \frac{\pi}{2} \frac{v_i}{1 + v_i^2}.$$
(28)

It is now easy to calculate b, which is related to u by

$$b = 1 - \left(\frac{u}{v}\right)^2. \tag{29}$$

The result is

$$b = 1 + \frac{\pi^2}{4v_i^2}\left\{1 + j\frac{2}{v_i}(1 + v_r)\right\} \tag{30}$$

keeping in mind that $v_r/v_i \ll 1$ and $v_i^2 \gg 1$.

NOTATIONS

$\alpha = \dfrac{4\pi k}{\lambda}$ attenuation constant

a halfwidth of slab

b normalized propagation constant [defined in eq. (7)]

$s = 2a$ width of slab

$\Delta n = n_1 - n_2$

$\Delta k = k_1 - k_2$

ϵ_0 dielectric constant of vacuum

ϵ_1, ϵ_2 relative dielectric constants (subscript 1 refers to the center and 2 to the outer region)

g gain in center region

$k_0 = \omega\sqrt{\mu_0\epsilon_0}$ free-space wave number

$k_{1,2}$ extinction coefficient (subscript 1 refers to the center and 2 to the outer region)

λ free-space wavelength

μ_0 permeability of free space

$n_{1,2}$ refractive index (subscript 1 refers to the center and 2 to the outer region)

$\omega = \dfrac{2\pi c}{\lambda}$ angular frequency

u separation parameter

$v = ak_0\sqrt{\epsilon_1 - \epsilon_2}$ normalized frequency

w separation parameter

REFERENCES

1. Yariv, A., *Introduction to Optical Electronics*, New York: Holt, Rinehart and Winston, 1971.
2. McWorther, A. L., Zeiger, H. J., and Lax, B., "Theory of Semiconductor Maser of GaAs," J. Appl. Phys., *34*, 1963, p. 235.
3. Kogelnik, H., "On the Propagation of Gaussian Beams of Light Through Lenslike Media Including Those With A Loss or Gain Variation," Appl. Opt., *4*, 1965, pp. 1562–1569.

4. Collin, R. E., *Field Theory of Guided Waves*, New York: McGraw-Hill, 1960, chapter 11.
5. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, 1971, pp. 2252–2258.
6. D'Asaro, L. A., "Advances in GaAs Junction Lasers with Stripe Geometry," to be published in J. Electroluminescence.
7. Dyment, J. C., D'Asaro, L. A., North, J. C., Miller, B. I., and Ripper, J. E., "Proton Bombardment Formation of Stripe Geometry Heterostructure Lasers for 300° K CW Operation," Proc. IEEE, *60*, 1972, pp. 726–728.
8. Dyment, J. C., North, J. C., and D'Asaro, L. A., "Optical and Electrical Properties of Proton Bombarded p-Type GaAs," J. Appl. Phys. *44*, 1973 pp. 207–213.
9. Reinhart, F. K., Hayashi, I., and Panish, M. B., "On the Mode Reflectivity and the Waveguide Properties of Double Heterostructure Injection Lasers," J. Appl. Phys., *42*, 1971, p. 4466.
10. Dyment, J. C., Hartman, A. R., and Hwang, C. J., private communication.
11. Nash, F. R., "Mode Guidance Parallel to the Junction Plane of GaAs Lasers," to be published in J. Appl. Phys.

Heat Transfer in Electronic Systems With Emphasis on Asymmetric Heating

By W. AUNG

The trend in electronic circuit design is toward increasing power dissipation density. The performance and reliability of increasing number of electronic systems are now seriously threatened by thermal effects, so that it is necessary to reappraise the relevant thermal design procedures. This paper examines natural convection cooling and concerns the prediction of maximum temperatures of electronic cabinets containing arrays of vertically oriented circuit cards with unequal power dissipation levels. By means of a vertical channel model, the effects of channel spacing, channel height, and power dissipation level are assessed with emphasis on asymmetric powering of the channel walls. Methods are indicated for rapid evaluation of maximum temperatures and optimum channel spacing with asymmetric heating. The present results show that asymmetry reduces the thermal performance of the channel. Consequently, the power dissipation on the channel walls should be made nearly equal.

I. INTRODUCTION

For the past several years, the trend in electronic equipment design has been toward ever-increasing circuit speeds. Now it is common for response times of telecommunication equipment to be specified in terms of nanoseconds while, in high-performance data processing systems, the required response times are given in the picosecond range.

The increase in circuit speed is facilitated by integration of circuit functions, circuit miniaturization, and higher-density packaging. Although the miniaturization of circuits results in decreased power dissipation per circuit, the power generation per unit volume, which is the important parameter in determining the circuit temperature, is actually increased due to the much higher packaging densities. The thermal problem is also compounded by the lower operating temperature re-

quirements of integrated circuits. This gives rise to a challenging under-
taking in thermal design. And yet a poor thermal design could pos-
sibly lead to complete failure or unreliable performance of equipment.
Thus, it is imperative that a sound thermal design be initiated, and this
at the earliest possible stage of system planning.

The thermal design of a modern electronic system is based on a ra-
tional selection of a cooling option followed by a thoughtful design con-
sideration. A design should not only be practicable, but also economi-
cal, serviceable, reliable, and compatible with other system compo-
nents. It is the purpose of this paper to present design data on natural
convection cooling of modern electronic systems. The emphasis is on
the prediction of maximum temperatures in equipment. Although na-
tural convection is the oldest method used in electronics cooling, the
method is still employed extensively in new generations of equipment
both in the communications and the data processing fields. In many
cases, this method is used in conjunction with one or more of the newer
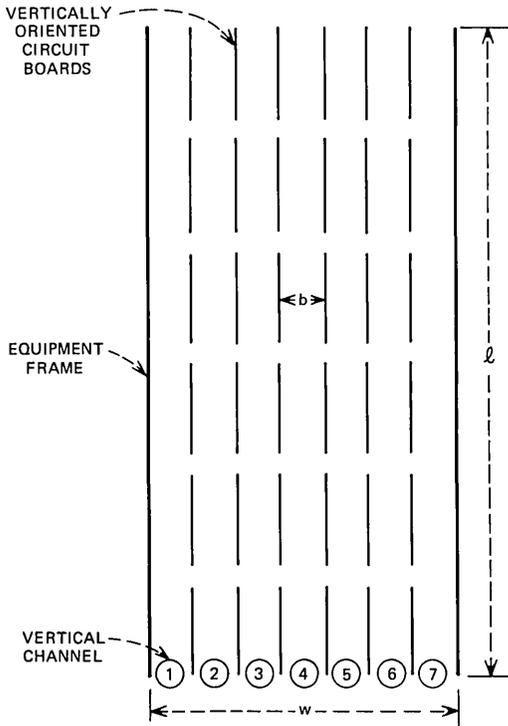

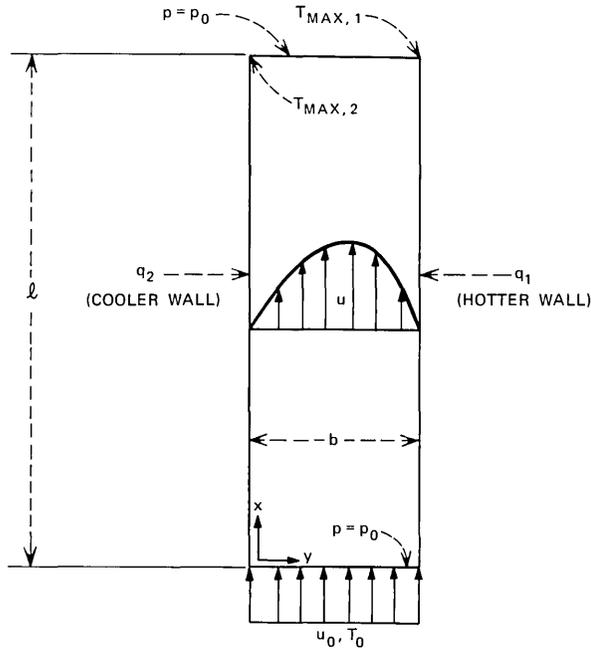Fig. 1—Schematic diagram of electronic equipment.

Fig. 2—Two-dimensional vertical channel model.

methods such as forced flow and the application of heat pipes and liquid cooling.

The present paper concerns equipment wherein circuit boards containing heat-dissipative components are mounted vertically in channel-like fashion as shown in Fig. 1. Since the circuit boards are not identically powered, the average heat fluxes from the columns are different, giving rise to asymmetrical air flow. In each channel, the heat transfer may be modeled by that in a vertical, two-dimensional, smooth-walled channel as shown in Fig. 2. The channel walls are treated as uniformly heated. However, the two wall heat fluxes need not be identical.

A number of investigations have been reported in the literature concerning free convection in vertical channels. Ostrach[1] solved the combined free (natural) and forced convection problem in a vertical channel in fully developed flow with symmetrical uniform heat flux and internal heat generation. By fully developed flow is meant a situation in which the fluid velocity is invariant in the direction of flow. Engel and Mueller[2] investigated the effect of nonisothermal channel walls by assuming constant heat fluxes. Their results, however, are limited to

symmetrical wall heating. Lauber and Welch[3] considered the combined free and forced convection between vertical flat plates which are heated at uniform but different heat fluxes. Their work is confined to fully developed flow.

Two standard references widely applied to heat transfer in electronic equipment are the experimental work of Elenbaas[4] and the numerical solution of Bodoia and Osterle.[5] Inasmuch as the above references deal with a channel whose walls are at a constant temperature, two factors are therefore ignored in their results: the effect of nonisothermal surfaces and asymmetric heating. These effects are considered in the present paper.

It is to be noted that, as in the references cited above, radiative transfer of heat is ignored in this paper. In some electronic systems, radiation may be a significant factor in heat removal. Inasmuch as temperatures of heat sources are of interest, the neglect of radiation usually results in conservative (too high) estimates of the temperatures.

II. ANALYSIS

Referring to Fig. 2 which shows a two-dimensional, straight vertical channel, let the channel height be ℓ and width be b. The channel walls
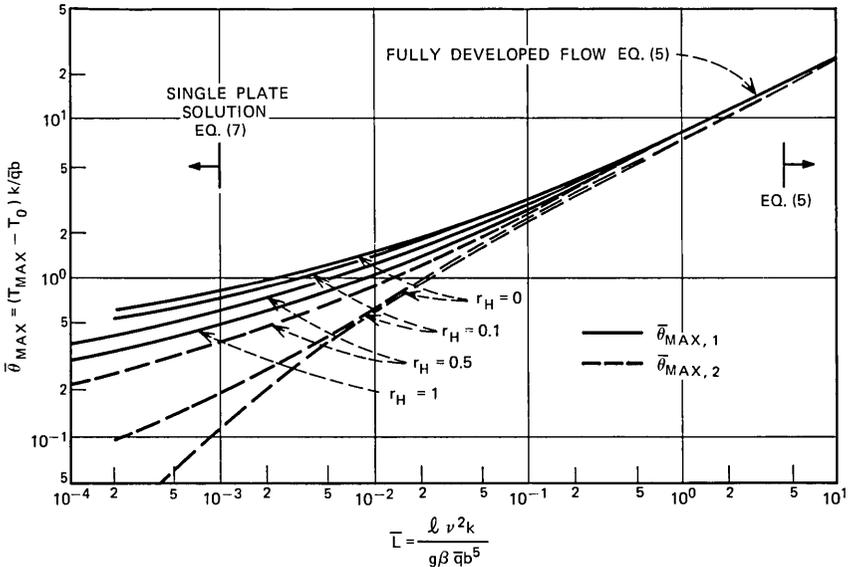

Fig. 3—Relation between dimensionless maximum wall temperature rise and the dimensionless parameter $\bar{L}$ (from Aung, et al.[6]).

are heated causing fluid to rise between them by free convection. The thermal conditions of the walls are characterized as uniform heat flux (UHF), but the individual values for the two walls need not be the same. The fluid that enters the channel at the ambient temperature T_0 is assumed to have a flat velocity profile u_0. The equations expressing conservation of mass, momentum, and energy for free convection in the constant-property (except for density in the buoyancy term) laminar flow may be written respectively as:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{1}$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \nu \frac{\partial^2 u}{\partial y^2} - \frac{g_c}{\rho} \frac{dp'}{dx} + g\beta(T - T_0), \tag{2}$$

$$\rho c_p \left(u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} \right) = k \frac{\partial^2 T}{\partial y^2}, \tag{3}$$

where p' is the pressure excess above the hydrostatic pressure.

The boundary conditions are:

For $x = 0$ and $0 < y < b$:
$$u = u_0, \qquad T = T_0;$$
For $y = 0$ and $x \geq 0$:
$$u = 0, \qquad v = 0, \qquad k \frac{\partial T}{\partial y} = -q_2;$$
For $y = b$ and $x \geq 0$:
$$u = 0, \qquad v = 0, \qquad k \frac{\partial T}{\partial y} = q_1. \tag{4}$$

At $x = 0$ and $x = \ell$: $p = p_0$, where p_0 is the hydrostatic pressure.

The system of eqs. (1) to (4) may be cast in nondimensional form and then solved by numerical integration using a digital computer. Details are contained in Refs. 6 and 7. The latter references show that the numerical results agree well with experimental data. It may be pointed out that, for a channel whose height is large compared to the spacing so that the so-called fully developed flow condition exists, the governing eqs. (1) to (3) may be simplified and the solution is then obtainable in closed form. This is indicated in Ref. 8. In the present paper, these results will be applied to indicate the effect of operating parameters on maximum wall temperatures. The cooling fluid considered is exclusively air.

III. RESULTS

When the channel walls are individually heated at uniform heat fluxes, the quantity that is of design interest is generally the maximum temperature rise on the hotter wall. This quantity may be obtained from the general results shown in Fig. 3. It is clear from the latter that the maximum temperature increase depends on the parameter $\bar{L}$ which in turn depends on the quantity $\ell/\bar{q}b^5$. When the latter quantity is large, fully developed flow is approached[6] and the desired information on the maximum temperature increase is given by a rather simple expression which follows.

3.1 Results at Large $\ell/\bar{q}b^5$

At large values of the quantity $\ell/\bar{q}b^5$, the flow in the channel exhibits interesting characteristics whereby the velocity distribution across the channel remains unaltered with axial distance. All axial temperature variations are also given by linear relations. Typical velocity and temperature distributions in this situation are given in Fig. 4. This is the so-called fully developed flow situation. Clearly, it prevails most commonly in a channel whose height ℓ is large compared to its width. In a channel with developing flow, typical velocity and temperature dis-
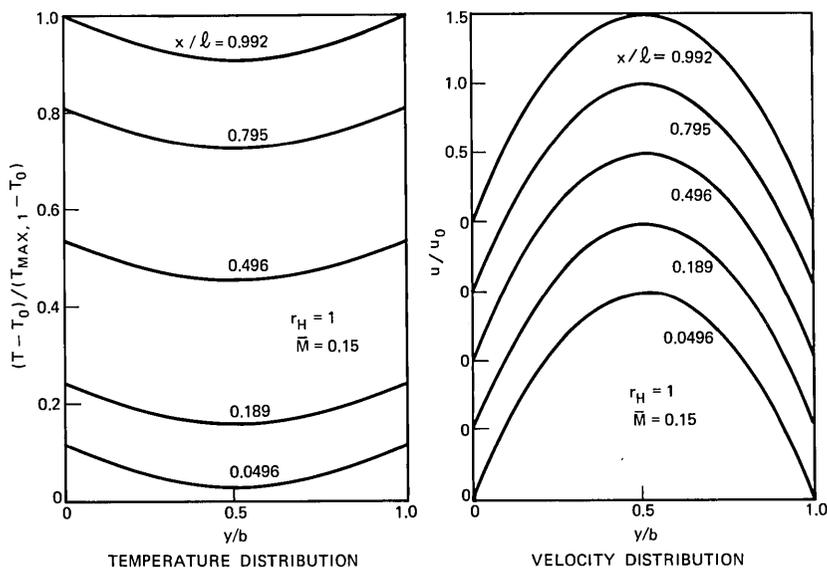

Fig. 4—Typical temperature and velocity distributions in a channel having nearly fully developed flow.
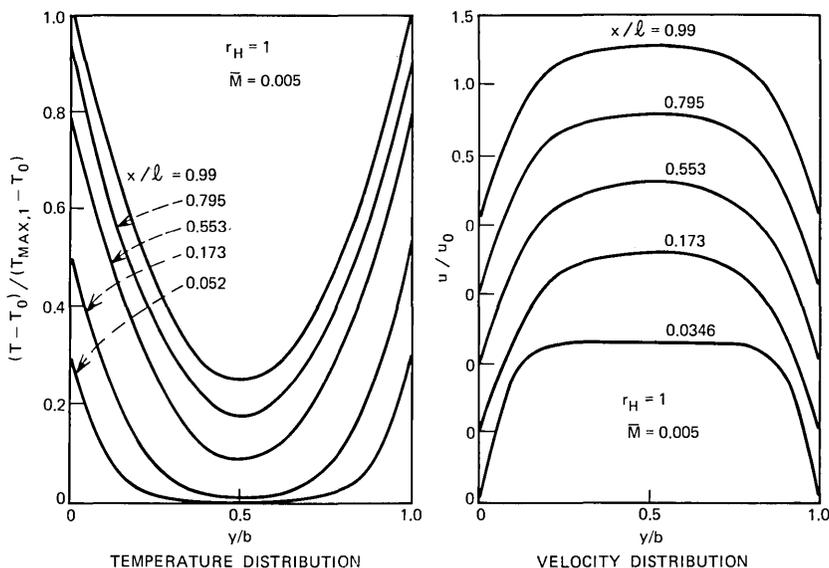
Fig. 5—Typical temperature and velocity distributions in a channel having developing flow.

tributions are as indicated in Fig. 5. In both Fig. 4 and Fig. 5, the quantity $\bar{M}$ is a dimensionless mass flow parameter which is related to $\bar{L}$ (see Refs. 6 and 8).

Figure 3 shows that, as fully developed flow is approached, all curves coalesce into one. Based on the derivations given in Ref. 8, it can be shown that, in fully developed uniform heat flux channels with asymmetric heating, the maximum temperatures on the hotter and cooler walls become practically identical and may be evaluated using the equation:

$$\left.\begin{array}{l} \bar{\theta}_{\max,1} = \bar{\theta}_{\max,2} = 6.9285(\mathrm{Pr})^{-1/2}(\bar{L})^{1/2}, \\ \bar{L} \geqq 5, \qquad \text{all } r_H \end{array}\right\}. \tag{5}$$

For air at 120°F, the Prandtl number (Pr) is about 0.7. Evaluating the thermal physical properties for air in eq. (5) at a temperature of 120°F, we may rewrite eq. (5) as:

$$\left.\begin{array}{l} T_{\max,1} - T_0 = T_{\max,2} - T_0 = 0.1610(\bar{q})^{0.5}(\ell)^{0.5}(b)^{-1.5} \\ \text{for } \dfrac{\ell}{\bar{q}b^5} \geqq 1.19 \times 10^9, \qquad \text{all } r_H \end{array}\right\}. \tag{6}$$

In eq. (6), ℓ is expressed in ft, $\bar{q}$ in W/ft², and b in ft. Note that, if $T_{\max,1}$ alone is to be evaluated, eq. (6) is applicable at $\ell/\bar{q}b^5 \geqq 4.76 \times 10^7$.

The reason is apparent from Fig. 3 where at $\bar{L} \geqq 0.2$ the deviations among curves of $\bar{\theta}_{max,1}$ at all values of r_H may be neglected.

3.2 Results at Small $\ell/\bar{q}b^5$

The counter-case to the situation just discussed is the one in which the channel height ℓ is relatively short compared to its width b. Again, in this situation the maximum temperature rise on the channel walls may be calculated by a simple relation. Note that, to use this relation, which will be given below, it is not necessary that ℓ be smaller than b. It is only necessary to fulfill the validity limit attached to the equation.

For relatively short channels, the results in Ref. 6 indicate that the maximum temperature in each wall is independent of the other wall, and may be calculated using the result for a single vertical flat plate.[9] The latter result is valid for the asymmetrically heated channel when $\bar{L} \leqq 10^{-3}$. Thus, we may write:

$$\theta_{max} = 2.05(L)^{1/5} \qquad \text{at } \bar{L} \leqq 10^{-3}. \tag{7}$$

Again, inserting thermal physical properties of air at 120°F, the above may be written:

$$\left. \begin{array}{l} T_{max} - T_0 = 8.66\ell^{1/5}q^{4/5} \\ \text{at } \dfrac{\ell}{\bar{q}b^5} \leqq 2.35 \times 10^5, \qquad \text{all } r_H \end{array} \right\}. \tag{8}$$

In eq. (8), $(T_{max} - T_0)$ is in degrees Fahrenheit, ℓ in feet, and q in watts per square foot of surface area. Equation (8), which gives the maximum temperature on either the hotter or cooler wall when q is appropriately replaced by q_1 or q_2, shows that the maximum temperature on the hotter wall, subject to the attached condition, is independent of the channel spacing which is to be expected. It may also be noted for reference that

$$\frac{\ell}{q_1 b^5} = \left(\frac{1 + r_H}{2} \right) \frac{\ell}{\bar{q}b^5}. \tag{9}$$

3.3 Results at Intermediate $\ell/\bar{q}b^5$

If the proposed design is such that neither eq. (6) nor eq. (8) may be applied, then the maximum temperature increase in the channel can be evaluated using Fig. 3. Parametric curves can be generated from Fig. 3 to display the effect of various operating variables on the maximum temperature rise on the hotter wall. Since the value of $\bar{L}$ depends on the channel spacing, the average heat flux, and the channel height, any one of these may be varied while others remain constant at typical
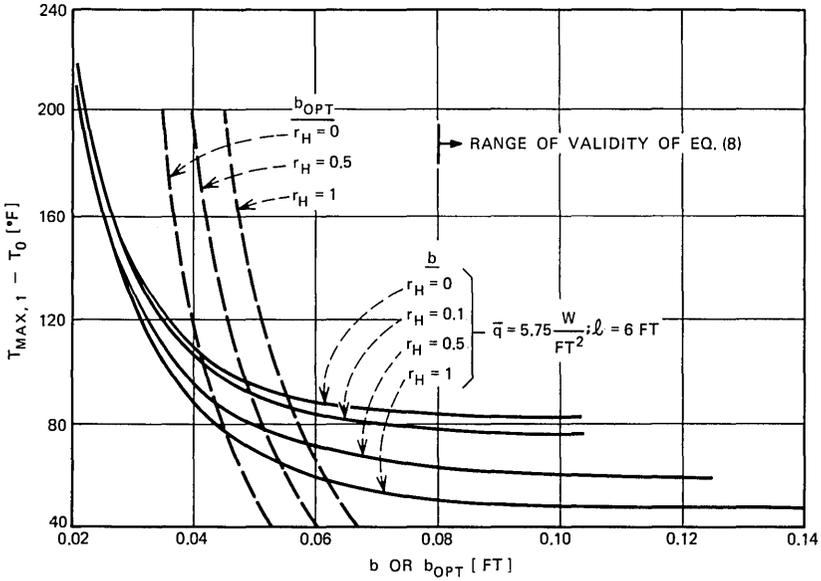
Fig. 6—Effect of channel spacing on the maximum wall temperature rise.
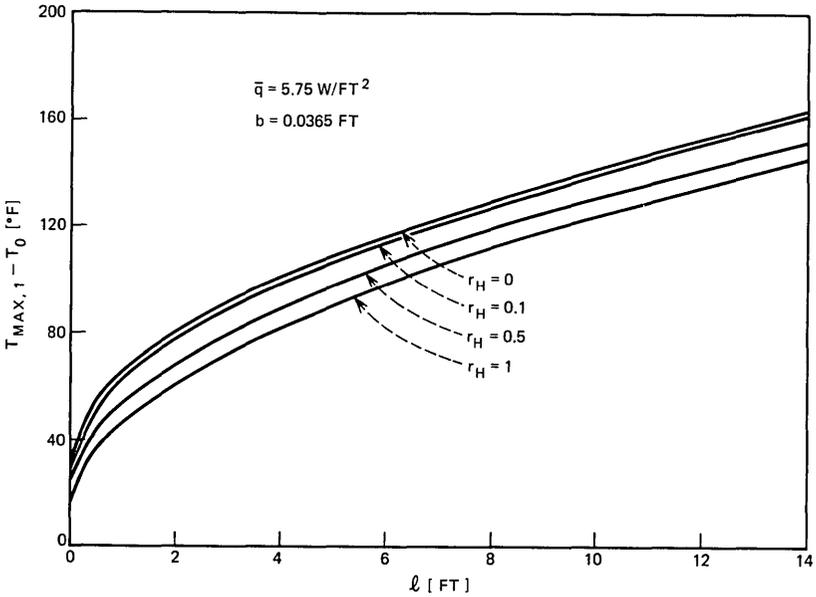

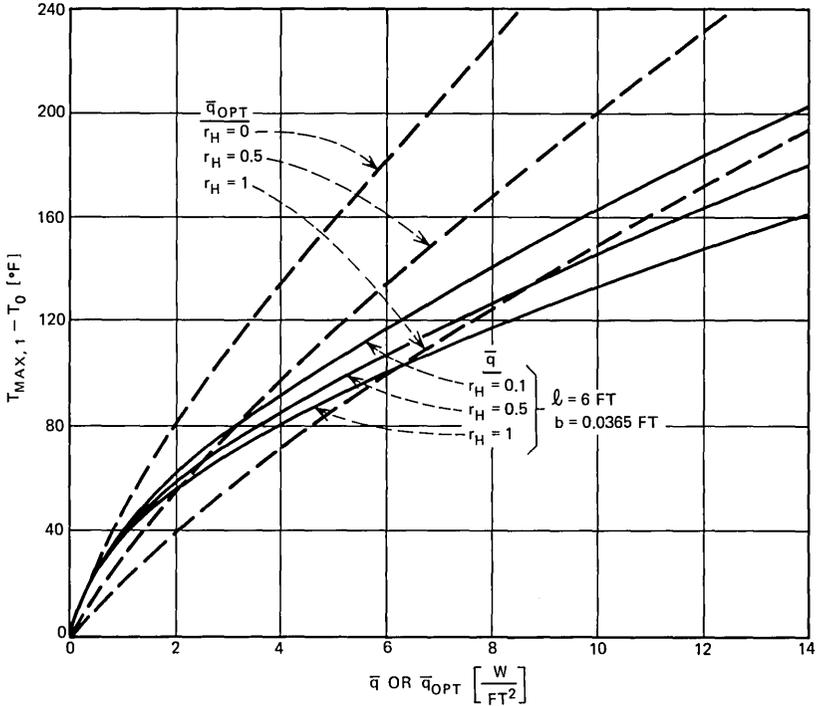Fig. 7—Effect of channel height on the maximum wall temperature rise.

Fig. 8—Effect of average heat flux on the maximum wall temperature rise.

operating values. The following conditions are chosen as an example:

$$b = 0.0365 \text{ foot (0.4375 inch)}$$
$$\ell = 6 \text{ feet} \tag{10}$$
$$\bar{q} = 5.75 \text{ watts per square foot.}$$

The resulting curves are shown in Figs. 6 through 8 which can be used by the designer. To obtain Fig. 6, ℓ and $\bar{q}$ are obtained from eqs. (10) and, corresponding to each pair of assigned values of b and r_H, $\bar{L}$ is calculated. A value of $\theta_{\max,1}$ is then obtained from Fig. 3. The maximum temperature rise can then be obtained. By varying r_H and b, the solid curves in Fig. 6 result. The range of validity of eq. (8) is also indicated. Figures 7 and 8 are obtained in the same manner. The dashed lines in Figs. 6 and 8 pertain to optimum spacing which will be discussed later. From Figs. 6 through 8, it can be seen that at any fixed maximum temperature rise the effect of asymmetry is to increase the necessary channel spacing, decrease the channel height, or decrease the heat flux when

any one of these quantities is taken as the sole variable in the equipment.

The effect of asymmetric wall heating may be more conveniently examined by replotting Fig. 3 in the manner of Fig. 9. In the latter, the maximum temperature increase on each wall of an asymmetric channel is normalized by the maximum temperature increase on a similar channel wherein the same average heat flux is distributed evenly on the two walls, so that $r_H = 1$. Clearly, as the degree of asymmetry is increased so that r_H is decreased without affecting the average heat flux, the result is that the maximum temperature on the hotter wall is raised while that on the cooler wall is depressed. Therefore, in equipment design, it is desirable to obtain equal heating on the two channel walls. This is also desirable from the standpoint of optimum spacing, as will be seen.

With the help of Fig. 9, the maximum temperature increases in an asymmetric channel may be evaluated once the symmetric heating value is known. The latter may be obtained rapidly with the aid of the nomogram shown in Fig. 10 once ℓ, b, and $\bar{q}$ are known, r_H being 1 in Fig. 10. The use of the latter has been described in Ref. 7 but is repeated here for ease of reference. To use Fig. 10, first locate Points 1, 2, and 4.
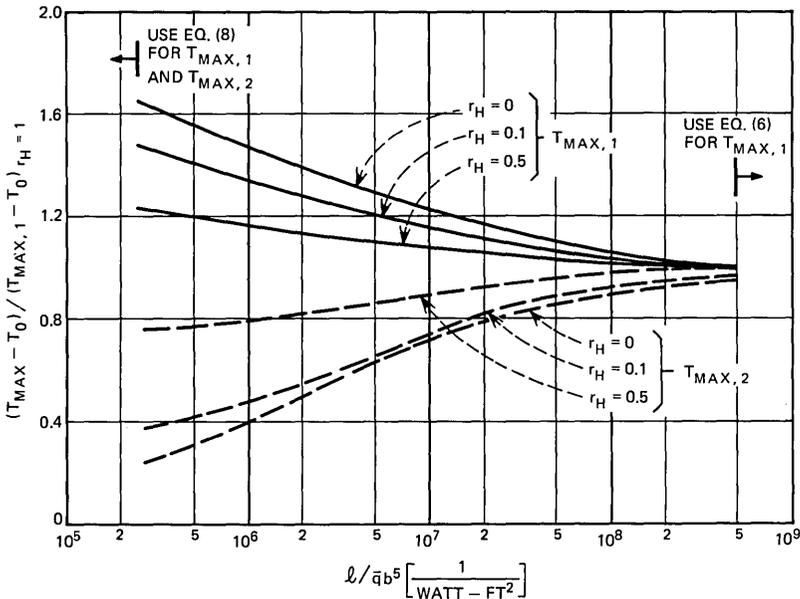

Fig. 9—Maximum wall temperature rise as a function of the quantity $\ell/\bar{q}b^5$.
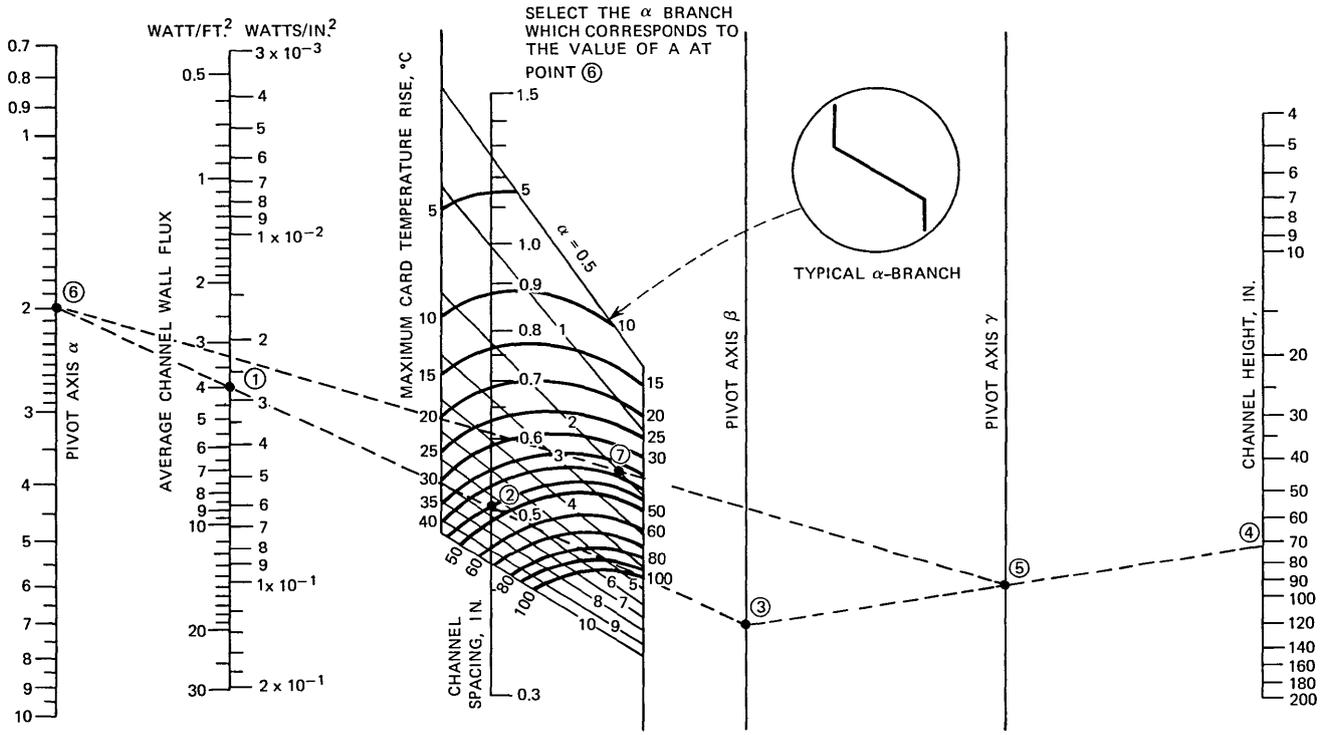
Fig. 10—Nomogram for evaluation of maximum wall temperature increase, $r_H = 1$ (from Aung, Kessler, and Beitin[7]).

Draw a line connecting Points 1 and 2. Extend this line to intersect with the Pivot Axis β to give Point 3, and with Pivot Axis α to give Point 6. Draw a line joining Points 3 and 4. The intersection of this line with Pivot Axis γ gives Point 5. Draw a line connecting Points 5 and 6. Read the α value at Point 6 which for the case illustrated is 2. The intersection of line 5–6 with the appropriate α-branch ($\alpha = 2$ in this case) in the center of the nomogram (Point 7) gives the maximum card temperature rise over room ambient, which in this example is 38°C or 68°F. It should be noted that each α-branch is made up of three segments: a left-side vertical line which corresponds to flow around a single flat plate, a right-side vertical line which corresponds to fully developed flow, and an inclined straight line joining the two vertical ones. Moreover, the asymptotes of the vertical lines coincide for all values of α. Hence, if the line joining Points 5 and 6 does not intersect the appropriate α-branch on its inclined segment, either the left or the right vertical segment should be chosen depending on whether the 5–6 line passes above the inclined segment or below it.

3.4 Optimum Spacing Including the Effect of Asymmetry

Consideration may now be given to the overall performance of the channel as a heat-removing device. In electronic equipment the maximum device junction temperature is usually specified. Since the junction temperature is related to the wall temperature, the maximum wall temperature is therefore implied. In modern equipment, it is desirable to increase the packing density and hence the total power dissipation of the entire equipment without increasing the maximum wall temperature beyond the allowable limit. As shown in the appendix, the task here is to select b so as to maximize the heat transfer. The latter value will be designated q_{opt} and the corresponding spacing is called b_{opt}. Design curves for finding b_{opt} and q_{opt} will be given below. The emphasis here is on asymmetrically heated channels. A more detailed discussion of optimum spacing in a symmetric channel may be found in Ref. 7. Note that maximum power dissipation is realized only when the equipment is strictly operating at b_{opt} and q_{opt}. If a different (smaller) spacing is used, then a different (smaller) heat flux must be employed to yield the same maximum temperature increase. For this purpose, design curves such as those in Figs. 6 through 8 can be consulted.

Following Bodoia,[10] it can be shown that the power dissipation is maximized if the channel parameters are so selected that the channel is operated at the point where the slope of a log-log plot of Nu versus Ra is one-half, where Nu is the Nusselt number and Ra the Rayleigh

number defined as:

$$\text{Nu} = \frac{\bar{q}b}{(T_{\max,1} - T_0)k} = \frac{1}{\bar{\theta}_{\max,1}},$$

$$\text{Ra} = \text{Pr} \times \frac{g\beta(T_{\max,1} - T_0)b^4}{\ell\nu^2} = 0.7 \times \bar{\theta}_{\max,1} \times \bar{L}.$$

From Fig. 3, a relation between Nu and Ra can be obtained. This is shown in Fig. 11. Calling the values of Nu and Ra for maximum power dissipation Nu_{opt} and Ra_{opt}, respectively, Table I may be constructed from Fig. 11. In Table I, E is an efficiency of heat transfer defined as

where

$$\left. \begin{aligned} E &= \frac{\bar{p}}{(\bar{p})_{r_H=1}} \\ \bar{p} &= \text{Nu}_{\text{opt}}/(\text{Ra}_{\text{opt}})^{1/2} \\ &\propto \frac{h_{\text{opt}}}{b_{\text{opt}}} \end{aligned} \right\} . \tag{11}$$

In the above, h_{opt} and b_{opt} are optimum values of $h_{\max,1}$ and b which give maximum power dissipation.
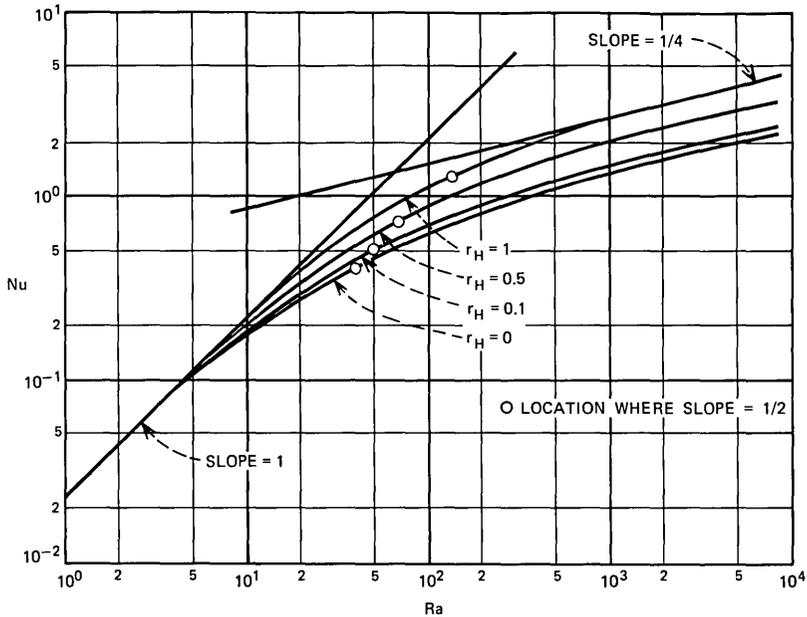

Fig. 11—Relation between Nusselt and Rayleigh numbers.

TABLE I—OPTIMUM RAYLEIGH AND NUSSELT NUMBERS

r_H	Ra_{opt}	Nu_{opt}	E
0	42	0.43	65%
0.1	51	0.51	70%
0.5	70	0.73	86%
1.0	135	1.18	100%

From Table I, it is seen that the efficiency E varies in the same manner as r_H. Asymmetry therefore decreases the efficiency of heat transfer. From the thermal standpoint all equipment should therefore be designed to yield symmetrical heating as closely as possible.

Using values of Ra_{opt} from Table I, it is possible to obtain optimum spacings for different maximum temperatures once the channel height is specified. In like manner, the necessary heat flux at optimum spacing corresponding to different prescribed maximum temperatures can also be obtained from Nu_{opt}. Results have been obtained for a 6-foot channel with $r_H = 0.0$ and $r_H = 0.5$ and compared with the case $r_H = 1$ in Figs. 6 and 8 (indicated by the dashed line). It may be ascertained from these figures that the effect of asymmetric heating is to decrease b_{opt} and q_{opt}. The net effect of asymmetry is to decrease the total power dissipation in the cabinet (see appendix).
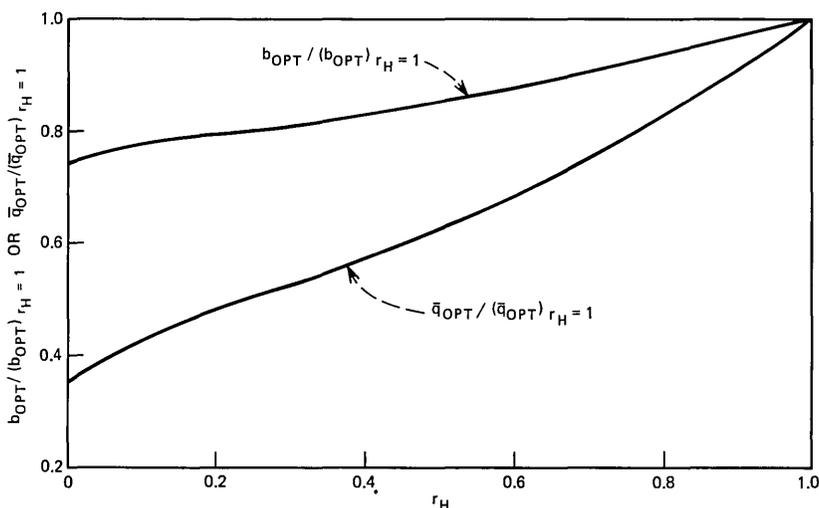

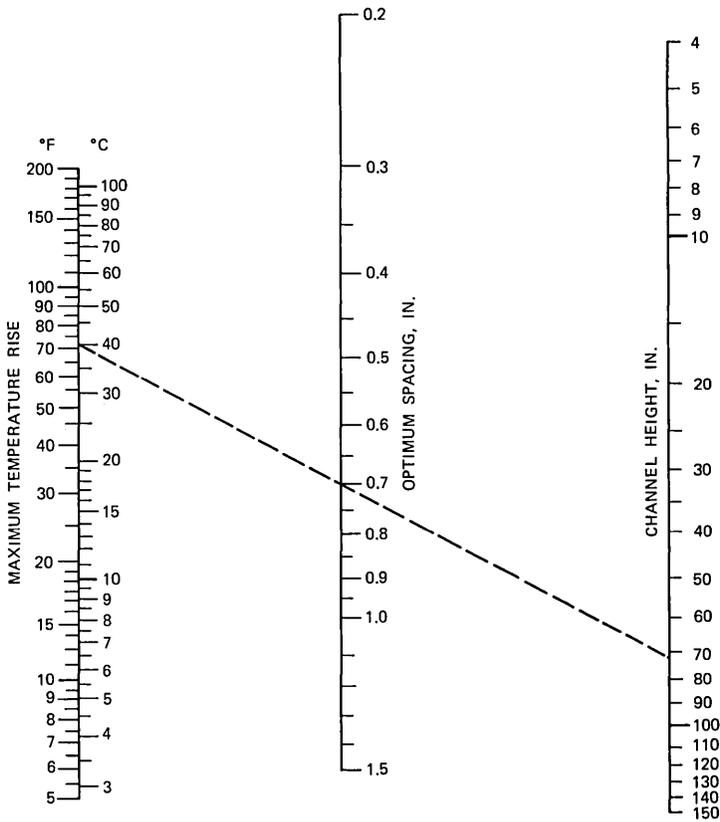Fig. 12—Optimum spacing and optimum heat flux as a function of the heat flux ratio r_H.

Fig. 13.—Nomogram for evaluation of optimum spacing, $r_H = 1$ (from Aung, Kessler, and Beitin[7]).

Using the values of Ra_{opt} and Nu_{opt} given in Table I, it is possible to obtain a general relation between optimum spacing b_{opt} or optimum heat flux q_{opt} and r_H, since

$$\frac{b_{opt}}{(b_{opt})_{r_H=1}} = \frac{(Ra_{opt})^{1/4}}{(Ra_{opt})^{1/4}_{r_H=1}}$$

and

$$\frac{\bar{q}_{opt}}{(\bar{q}_{opt})_{r_H=1}} = \frac{Nu_{opt}}{(Nu_{opt})_{r_H=1}}.$$

The results are indicated in Fig. 12. With the aid of this figure and the nomogram given in Fig. 13 which yields $(b_{opt})_{r_H=1}$, the optimum spacing at specified r_H, ℓ, and maximum hotter wall temperature rise

$(T_{max,1} - T_0)$ can be rapidly evaluated. Note that the maximum temperature rise to be entered in Fig. 13 is also the specified value of $(T_{max,1} - T_0)$.

IV. CONCLUDING REMARKS

In the present paper, consideration has been given to the free or natural convection cooling of electronic cabinets containing arrays of vertical circuit cards with asymmetric heating. It was indicated that fully developed flow prevailed in a channel whose height is large compared to its width. In this case, the maximum temperature in the channel can be calculated using the simple relation presented here. In contrast, when the channel spacing is relatively large, the maximum temperature on each wall is independent of the conditions on the other wall but can again be calculated by a simple equation.

In situations where neither of the above two approaches can be applied, the maximum temperature in the channel can be evaluated rapidly using the graphical procedure outlined in this paper. The optimum channel spacing for maximizing the power dissipation is discussed and it is emphasized that to reduce adverse thermal effects channel walls should be powered as equally as possible.

APPENDIX

The total power dissipation in the cabinet per unit depth can be written as

$$\text{Power} = 2\ell N \bar{q}$$
$$= 2\ell N h (T_{max,1} - T_0),$$

where N is the number of channels (see Fig. 1). If the thickness of the channel wall is neglected, we have

$$N = \frac{W}{b}.$$

Hence,

$$\text{Power} = 2\ell W (T_{max,1} - T_0) \frac{h}{b}$$

$$\propto \frac{h}{b}$$

for fixed $(T_{max,1} - T_0)$, W, and ℓ. Clearly, to obtain the largest possible power one needs to obtain the maximum h (that is $\bar{q}$) and the minimum b, the combination of which gives the specified $(T_{max,1} - T_0)$.

NOTATIONS

b	clear channel or card spacing in feet (ft)
c_p	specific heat of air in W-s/lb_m-F
g	acceleration due to gravity in ft/s^2
h	$\bar{q}/(T_{max,1} - T_0)$, average heat transfer coefficient, in W/ft^2-F
k	thermal conductivity of air in W/ft-F
L_1	$(\ell \nu^2 k)/(g\beta q_1 b^5)$, dimensionless
$\bar{L}$	$(\ell \nu^2 k)/(g\beta \bar{q} b^5)$, dimensionless
ℓ	channel height in ft
Pr	Prandtl number defined by $\mu c_p/k$, dimensionless
p	pressure in lb_f/ft^2
p_0	hydrostatic pressure in lb_f/ft^2
p'	$(p - p_0)$ in lb_f/ft^2
q	heat flux on channel wall per unit surface area, in W/ft^2
$\bar{q}$	$(q_1 + q_2)/2$, W/ft^2
r_H	q_2/q_1, dimensionless
T	temperature in degrees Fahrenheit (F)
T_0	ambient temperature in degrees Fahrenheit
u, v	axial and transverse velocity in ft/s
W	width of cabinet
x, y	axial and transverse coordinates in ft, see Fig. 2
β	thermal expansion coefficient of air in 1/F
μ	dynamic viscosity in lb_m/ft-s
ν	kinematic viscosity of air in ft^2/s
ρ	density of air in lb_m/ft^3
$\bar{\theta}$	dimensionless temperature rise defined in Fig. 3

Subscripts

1	refers to hotter wall
2	refers to cooler wall
max	maximum value
max, 1	maximum value on hotter wall
max, 2	maximum value on cooler wall

REFERENCES

1. Ostrach, S., "Combined Natural and Forced-Convection Laminar Flow and Heat Transfer of Fluids with and without Heat Sources in Channels with Linearly Varying Wall Temperatures," NACA TN 3141 (1954).
2. Engel, R. K., and Mueller, W. K., "An Analytical Investigation of Natural Convection in Vertical Channels," ASME Paper No. 67-HT-16, 1967.
3. Lauber, T. S., and Welch, A. U., "Natural Convection Heat Transfer Between Vertical Flat Plates with Uniform Heat Flux," Proc. Third Int. Heat Transfer Conf., Chicago, Illinois (1966), pp. 126–131.

4. Elenbaas, W., "Heat Dissipation of Parallel Plates by Free Convection," Physica, *9*, No. 1 (1942), pp. 1–28.
5. Bodoia, J. R., and Osterle, J. F., "The Development of Free Convection Between Heated Vertical Plates," J. Heat Transfer, Trans. ASME, Series C, *84*, No. 1 (1962), pp. 40–44.
6. Aung, W., Fletcher, L. S., and Sernas, V., "Developing Laminar Free Convection Between Vertical Flat Plates with Asymmetric Heating," Int. J. Heat Mass Transfer, *15*, 1972, pp. 2293–2308.
7. Aung, W., Kessler, T. J., and Beitin, K. I., "Natural Convection Cooling of Electronic Cabinets Containing Arrays of Vertical Circuit Cards," ASME Paper No. 72-WA/HT-40. See also W. Aung, T. J. Kessler, and K. I. Beitin, "Free Convection Cooling of Electronic Systems," IEEE Trans. Parts, Hybrids, and Packaging, *9*, No. 2 (June 1973), pp. 75–86.
8. Aung, W., "Fully Developed Laminar Free Convection Between Vertical Plates Heated Asymmetrically," Int. J. Heat Mass Transfer, *15*, 1972, pp. 1577–1580.
9. Sparrow, E. M., and Gregg, J. L., "Laminar Free Convection from a Vertical Plate with Uniform Surface Heat Flux," Trans. ASME, February 1956, pp. 435–440.
10. Bodoia, J. R., Ph.D. Thesis, Carnegie Institute of Technology, 1959.

On the Selection of a Two-Dimensional Signal Constellation in the Presence of Phase Jitter and Gaussian Noise

By G. J. FOSCHINI, R. D. GITLIN, and S. B. WEINSTEIN

(Manuscript received February 1, 1973)

A long-standing communications problem is the efficient coding of a block of binary data into a pair of in-phase and quadrature components. This modulation technique may be regarded as the placing of a discrete number of signal points in two dimensions. Quadrature amplitude modulation (QAM) and combined amplitude and phase modulation (AM-PM) are two familiar examples of this signaling format. Subject to a peak or average power constraint, the selection of the signal coordinates is done so as to minimize the probability of error. In the design of high-speed data communication systems this problem becomes one of great practical significance since the dense packing of signal points reduces the margin against Gaussian noise. Phase jitter, which tends to perturb the angular location of the transmitted signal point, further degrades the error rate. Previous investigations have considered the signal evaluation and design problem in the presence of Gaussian noise alone and within the framework of a particular structure, such as conventional amplitude and phase modulation. We present techniques to evaluate and optimize the choice of a signal constellation in the presence of both Gaussian noise and carrier phase jitter. The performance of a number of currently used or proposed signal constellations are compared.

The evaluation and the optimization are based upon a perturbation analysis of the probability density of the received signal given the transmitted signal. Laplace's asymptotic formula is used for the evaluation. Discretizing the signal space reduces the optimal signal design problem under a peak power constraint to a tractable mathematical programming problem.

Our results indicate that in Gaussian noise alone an improvement in signal-to-noise ratio of as much as 2 dB may be realized by using quadrature amplitude modulation instead of conventional amplitude and phase

modulation. New modulation formats are proposed which perform very well in Gaussian noise and additionally are quite insensitive to moderate amounts of phase jitter.

I. INTRODUCTION

A very attractive modulation format for coherent high-speed data transmission is the family of suppressed-carrier, two-dimensional signal constellations of which quadrature amplitude modulation (QAM) and combined amplitude and phase modulation (AM-PM) are two examples. In this paper we will consider using this more general signal format, which is equivalent to the arbitrary placement of a discrete number of signal points in the plane, subject only to a peak or average power constraint. The object will be to mitigate the major *statistical* transmission impairments encountered on the voice-grade telephone channel, such as carrier frequency offset, carrier phase jitter, and additive noise. Our attention is focused on constellations of 16 points, since this seems to be the largest constellation which is practical for the typical telephone channel. However, the techniques we develop are applicable to constellations of arbitrary size.

The placing of signal points in the plane is a long-standing problem that has received considerable attention in the past. Previous investigations[1-3] have considered the signal evaluation and design problem in the presence of Gaussian noise alone and within the framework of a particular structure such as combined amplitude and phase modulation. When Gaussian noise is the only transmission impairment, it is well known that at high signal-to-noise ratios (>25 dB) the signal points should be placed as far apart from each other as possible (the circle-packing problem). In the application to high-speed digital communication systems, the two-dimensional signal design problem becomes one of great practical significance because the dense packing of a large number of signal points markedly reduces the margin against random noise and phase jitter. The signal design problem in the presence of both phase jitter and Gaussian noise has not been solved before and is the subject of our discussion.

Coherent receiver structures have recently been proposed which employ an adaptive equalizer[4] to compensate for any linear distortion, low-frequency phase jitter, or small amounts of frequency offset introduced by the channel. A phase-locked loop[5] may be used to suppress high-frequency phase jitter and frequency offset. Of course, the output of a phase-locked loop will still deviate somewhat from the optimum demodulating phase angle. One purpose of this study is to assess the

effect of such phase errors on the system error rate and to indicate how this knowledge can be incorporated into the system design.

By system design we have in mind the selection of both a particular two-dimensional signaling format and the decision device placed at the demodulator output. In order to pursue these objectives, we discuss:

  (i) the relative immunity of various signal constellations as a function of the degree of noncoherency (i.e., the size of the phase error);

 (ii) an efficient iterative procedure for determining optimum signal formats under a peak power constraint;

(iii) system performance for the following hierarchy of decision devices: easily implementable, optimum in Gaussian noise, and a maximum-likelihood detector which uses the statistics of the phase error;

 (iv) the accuracy required in any phase-locked loop to attain a satisfactory error rate; and

  (v) the resulting error rate when no attempt is made to track the jitter.*

Our approach is to assume that intersymbol interference has been effectively eliminated by the equalizer while the phase-locked loop, if there is one, has only partially removed the phase jitter. Thus the equalizer output will be the sum of the partially coherent† transmitted signal and additive Gaussian noise. We adopt a phenomenological model which assumes that the (slowly varying) phase error has a Tikhonov density.[5] The Tikhonov density is associated with a conventional first-order phase-locked loop whose input is the sum of a sinusoid (whose phase is being tracked by the loop) and Gaussian noise. Under our assumed operating conditions of high signal-to-noise ratio, this density will closely approximate the actual phase density. When no attempt at tracking is made,‡ the jitter is modeled as being uniformly distributed in a reasonable peak-to-peak range. For each of these jitter densities, the probability density of the demodulator output, conditioned on the transmitted symbol, is used to estimate the error rate. This estimate is of the minimum distance type, where the "distance," which reflects the presence of phase jitter, is measured in

---

  * In the sequel, we will use the term jitter as a catch-all when referring to phase jitter and/or frequency offset.
  † A partially coherent signal is one whose carrier phase is jittered by a random component which is *not* uniformly distributed in the range $(-\pi, \pi)$.
  ‡ Since the passband equalizer[4] will determine the optimum static demodulating phase, the absence of a phase-locked loop does not imply the use of an arbitrary demodulating phase.

a non-Euclidean manner. The error rate is given for various signal constellations and detector structures under peak and average power constraints. By discretizing the received signal space, an iterative procedure is developed to determine (locally) optimum signal formats under a peak power constraint. This technique, which assumes a maximum-likelihood detector, makes use of an efficient search procedure developed by Kernighan and Lin.[6]

The system model and the problem formulation are presented in Section II. An asymptotic estimate and an upper bound on the error rate are developed in Section III. A comparison of the relative immunity of some popular signal constellations and detectors to phase jitter is described in Section IV. Section V discusses a technique to obtain locally optimum signal structures under a peak power constraint.[*]

II. SYSTEM MODEL AND PROBLEM FORMULATION

2.1 Preliminaries

We consider the two-dimensional synchronous data communication system shown in Fig. 1. Binary data are first grouped into blocks of M bits, and each block of M bits is then mapped into one of 2^M two-tuples (a, b). The sequences $\{a_k\}$ and $\{b_k\}$ amplitude modulate, respectively, an in-phase and quadrature carrier to generate the transmitted signal

$$m(t) = \sum_k a_k p(t - kT) \cos \omega_c t + \sum_k b_k p(t - kT) \sin \omega_c t, \qquad (1)$$

where $1/T$ is the symbol rate,[†] $p(\cdot)$ represents the transmitter pulse shaping, and ω_c is the carrier frequency. It will be assumed that the two-tuples are equiprobable. The received signal at the output of the bandpass filter is given by

$$r(t) = \left(\sum_k a_k x(t - kT) - \sum_k b_k y(t - kT)\right) \cos \left((\omega_c + \Delta)t + \theta(t)\right)$$

$$- \left(\sum_k a_k y(t - kT) + \sum_k b_k(t - kT)\right)$$

$$\times \sin \left((\omega_c + \Delta)t + \theta(t)\right) + n(t), \quad (2)$$

where $x(t)$ and $y(t)$ are the system (baseband) in-phase and quadrature impulse responses,[‡] Δ is the carrier frequency offset, $\theta(t)$ is the random

---

[*] The present authors have recently treated the two-dimensional signal design problem under an average power constraint.[7]

[†] Note that the data rate is M/T bits/second.

[‡] These pulses represent the cascade of the transmitter shaping filter, the channel, and the receiving filter.
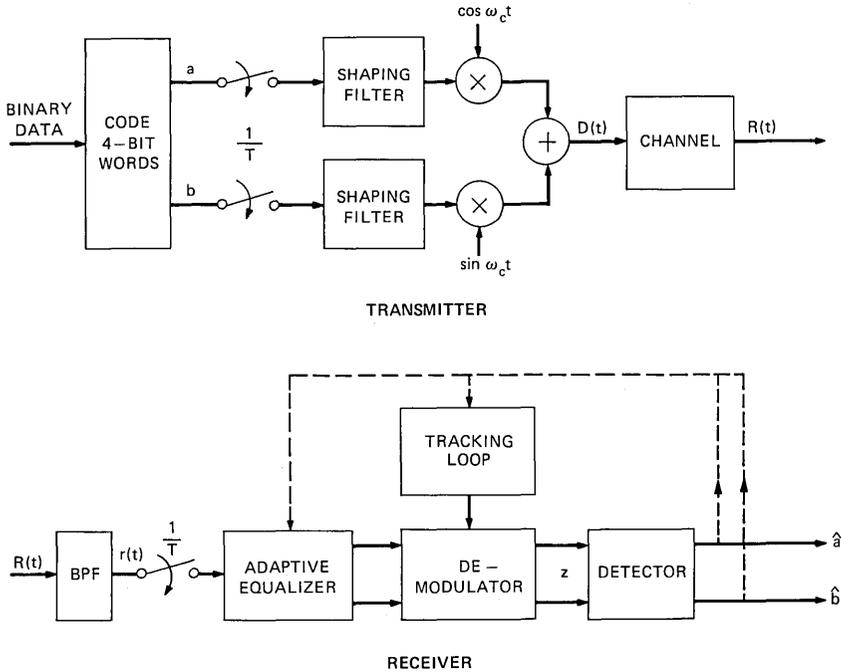
Fig. 1—An in-phase and quadrature data transmission system.

phase jitter, and $n(t)$ is additive Gaussian noise. The received signal, which is sampled at the symbol rate, is adaptively equalized and then coherently demodulated with the aid of a phase-locked loop.[5] The demodulated output, denoted by the sequence of two-tuples $z_k = \{(z_k, \check{z}_k)\}$, is then processed by the (simple) detector to give the output sequence $\{(\hat{a}_k, \hat{b}_k)\}$. The system error rate is just the probability that $(\hat{a}_k, \hat{b}_k)$ differs from the transmitted two-tuple (a_k, b_k).

2.2 Basic Model

For the purposes of this study, it will be convenient to assume that the equalizer has completely eliminated the intersymbol interference present in $x(t)$ and $y(t)$, but that the phase-locked loop has only partially compensated for the carrier phase jitter. The in-phase and quadrature demodulator outputs, at the kth sampling instant, are then given by

$$z(kT) = a_k \cos \phi_k - b_k \sin \phi_k + n_c(kT),$$
$$\check{z}(kT) = a_k \sin \phi_k + b_k \cos \phi_k + n_s(kT),$$
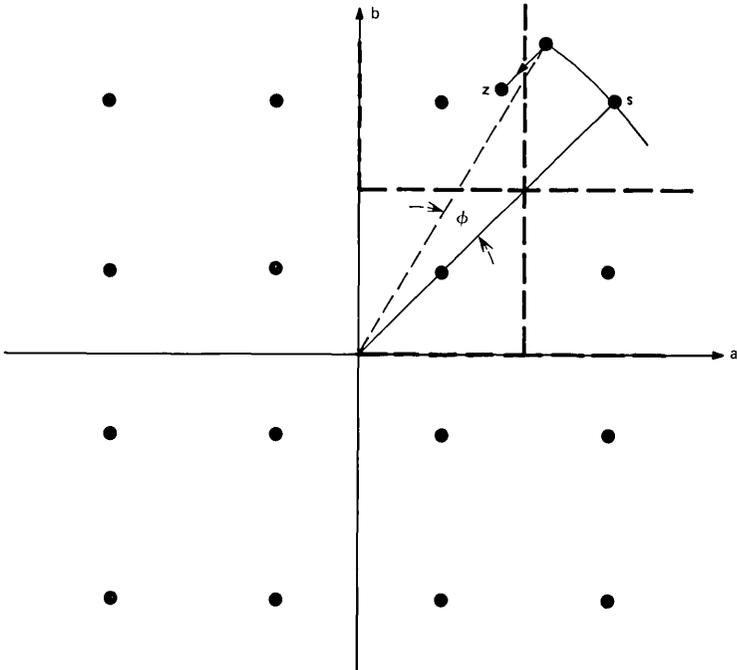
$$(3)$$

Fig. 2—Effect of Gaussian noise and phase jitter on transmitted symbol.

where ϕ_k is the (slowly varying) phase error in the tracking loop, and $n_c(kT)$ and $n_s(kT)$ are, respectively, the in-phase and quadrature Gaussian noise components.* Dropping the time index, eq. (3) can be rewritten to give the basic model

$$\mathbf{z} = R\mathbf{s} + \mathbf{n}, \tag{4}$$

where the vectors are given by†

$$\mathbf{z} = \begin{pmatrix} z \\ \overset{\vee}{z} \end{pmatrix}, \qquad \mathbf{s} = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{n} = \begin{pmatrix} n_c \\ n_s \end{pmatrix},$$

and the matrix R is the rotational (by an angle ϕ) transformation

$$R = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}. \tag{5}$$

---

* Recall that $n_c(kT)$ and $n_s(kT)$ are independent Gaussian random variables with equal variance, N_0. It should be noted that $2N_0$ is the noise power contained in the bandwidth of the received signal.

† We denote the values that a and b can assume by $a^{(i)}$ and $b^{(i)}$, respectively, and the values of the transmitted symbols by $\mathbf{s}^{(i)} = (a^{(i)}, b^{(i)})$.
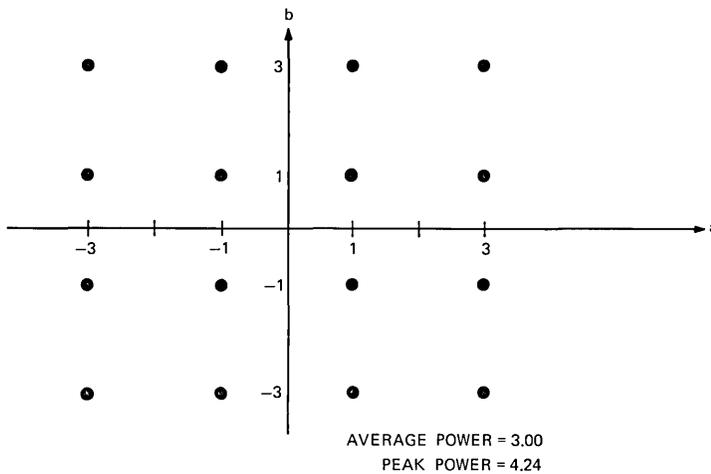
AVERAGE POWER = 3.00
PEAK POWER = 4.24

Fig. 3—Quadrature amplitude modulation.

As we show in Fig. 2, the effect of the phase jitter is to rotate the transmitted symbol, **s**, by an angle ϕ; thus the demodulator output, **z**, is dispersed in an angular manner due to phase jitter and in a circularly symmetric way due to the Gaussian noise. The receiver will make an error when these perturbations move the demodulator output across the decision boundary associated with the transmitted symbol. For a particular transmitted data sequence, the demodulated sequence will be scattered about the transmitted points in a manner which reflects the combined effects of phase jitter and Gaussian noise. For the transmitted signal constellation of Fig. 3, which is known as QAM, a typical scattered demodulated sequence is shown in Fig. 4. As one might expect, for those signal points further away from the origin, the angular displacement becomes more apparent. An estimate of this effect is given by the mean-square error between the transmitted and demodulated symbols. For small values of jitter, this error is obtained from (4) by noting that

$$\mathbf{z} - \mathbf{s}^{(j)} = \begin{pmatrix} \cos \phi - 1 & -\sin \phi \\ \sin \phi & \cos \phi - 1 \end{pmatrix} \mathbf{s}^{(j)} + \mathbf{n} \approx \phi \begin{pmatrix} -b_j \\ a_j \end{pmatrix} + \mathbf{n};$$

averaging the norm-squared of both sides gives*

$$E\|\mathbf{z} - \mathbf{s}^{(j)}\|^2 = N_0 + \sigma_\phi^2 \|\mathbf{s}^{(j)}\|^2, \tag{6}$$

---

* The notation $\|\mathbf{z}\|$ denotes the Euclidean norm of **z**; additionally the notation $\langle \mathbf{z}, \mathbf{s} \rangle$ will be used to denote the inner product of **z** and **s**.
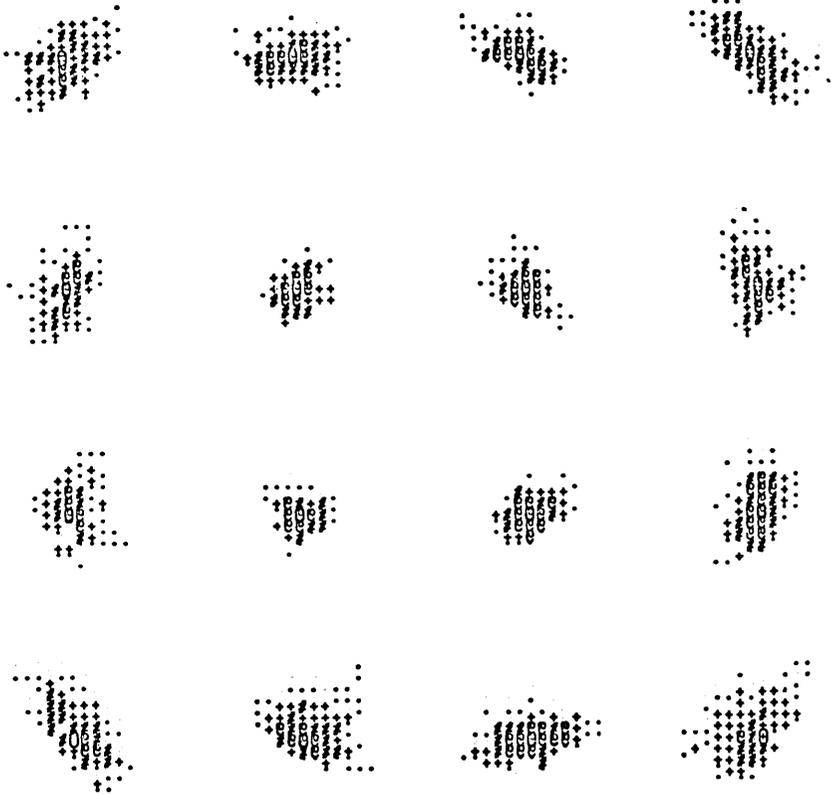
Fig. 4—Received signal points in a QAM system.

where σ_ϕ^2 is the variance of ϕ and E denotes the statistical average. Thus, because of phase jitter, signal points located further from the origin are subjected to a larger mean-square error.

2.3 *Probability Density of the Demodulator Output*

In order to evaluate the system error rate, the probability density function (pdf) of the phase error must be specified. The pdf of the phase error in a phase-locked loop that is tracking the angle of the two-dimensional data signal given by (2) is not yet known, but as explained below it can be approximated by the following (Tikhonov) density[5]:

$$p(\phi) = \frac{1}{2\pi} \frac{e^{\alpha \cos \phi}}{I_0(\alpha)}, \qquad |\phi| \leqq \pi, \tag{7}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind and α is a positive number. As is shown in Fig. 5, $\alpha = 0$ implies a completely incoherent system, while $\alpha = \infty$ corresponds to a completely coherent system. For large values of α ($\alpha > 100$), we have the useful relation

$$\sigma_\phi^2 \approx \frac{1}{\alpha} \qquad \text{(in radians)}. \tag{8}$$

Since the above density arises from a first-order phase-locked loop whose input is the sum of Gaussian noise and a sinusoid (whose phase is being tracked by the loop), it is felt that for high signal-to-noise ratio, for negligible intersymbol interference, and for slowly varying phase jitter the actual phase-error density will closely resemble the Tikhonov density. This simple model will be quite useful in studying the effect of phase jitter on the system error rate.

The pdf of the demodulator output, conditioned on the transmission of $\mathbf{s}^{(j)}$, is given by

$$p(\mathbf{z} \,|\, \mathbf{s}^{(j)}) \triangleq p_j(\mathbf{z}) = \int_{-\pi}^{\pi} p_j(\mathbf{z} \,|\, \phi) p(\phi) d\phi, \tag{9}$$

where it is noted that the output density conditioned on both $\mathbf{s}^{(j)}$ and ϕ is given by

$$p_j(\mathbf{z} \,|\, \phi) = \frac{1}{2\pi N_0} \exp\left[-\frac{1}{2N_0} \|\mathbf{z} - R\mathbf{s}^{(j)}\|^2 \right] \tag{10a}$$

$$= \frac{1}{2\pi N_0} \exp\left\{ -\frac{1}{2N_0} \big[\|\mathbf{z} - \mathbf{s}^{(j)}\|^2 + 2\langle \mathbf{z}, \mathbf{s}^{(j)}\rangle \right.$$
$$\left. - 2\langle \mathbf{z}, R\mathbf{s}^{(j)}\rangle \big] \right\}. \tag{10b}$$

Substituting the Tikhonov density into (9) gives

$$p_j(\mathbf{z}) = \frac{1}{2\pi N_0} \exp\left\{ -\frac{1}{2N_0} \big[\|\mathbf{z} - \mathbf{s}^{(j)}\|^2 + 2\langle \mathbf{z}, \mathbf{s}^{(j)}\rangle \big] \right\}.$$

$$(I_0(\alpha))^{-1} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp \frac{1}{N_0} \big[(\langle \mathbf{z}, \mathbf{s}^{(j)}\rangle + \alpha N_0) \cos \phi$$
$$+ \langle \mathbf{z}, \mathbf{s}^{(j)\perp}\rangle \sin \phi \big] d\phi, \tag{11}$$

(where if $\mathbf{s}^{(j)} = \|\mathbf{s}^{(j)}\|(\cos \alpha, \sin \alpha)$ then $\mathbf{s}^{(j)\perp} = \|\mathbf{s}^{(j)}\|(\sin \alpha, -\cos \alpha)$), and we recognize the latter integral as

$$I_0 \left(\frac{1}{N_0} \sqrt{(\langle \mathbf{z}, \mathbf{s}^{(j)}\rangle + \alpha N_0)^2 + \langle \mathbf{z}, \mathbf{s}^{(j)\perp}\rangle^2} \,\right).$$

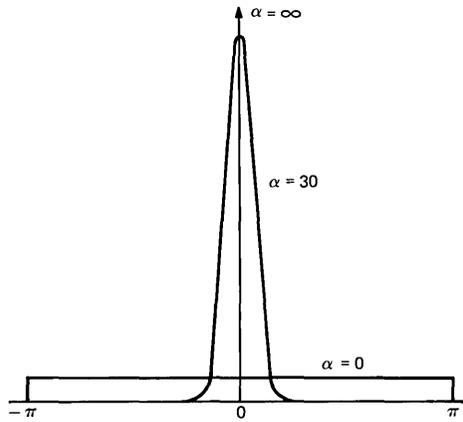Assume $\alpha = k/N_0$ (k a constant). Employing the well-known result

Fig. 5—Tikhonov phase jitter density $p\ (\phi) = \exp\ (\alpha \cos \phi)/2\pi I_o(\alpha)$.

that for large values of argument $I_0(x) \approx e^x/|\sqrt{x}|$, and assuming α is sufficiently large and N_0 is sufficiently small, we get

$$p_j(\mathbf{z}) \approx \frac{1}{2\pi N_0} \sqrt{\frac{\alpha N_0}{\sqrt{(\langle \mathbf{z}, \mathbf{s}^{(j)} \rangle + \alpha N_0)^2 + (\langle \mathbf{z}, \mathbf{s}^{(j)\perp} \rangle)^2}}} \exp \left(- \frac{1}{2N_0} \right.$$

$$\times \ \{ \|\mathbf{z}, \mathbf{s}^{(j)}\|^2 + 2\langle \mathbf{z}, \mathbf{s}^{(j)} \rangle + 2\alpha N_0$$

$$\left. - 2\sqrt{(\langle \mathbf{z}, \mathbf{s}^{(j)} \rangle + \alpha N_0)^2 + \langle \mathbf{z}, \mathbf{s}^{(j)\perp} \rangle^2} \} \right). \quad (12)$$

In the error rate computations we shall eventually make, we will have k so large compared to the practical range of $\langle \mathbf{z}, \mathbf{s}^{(j)} \rangle$ that the coefficient multiplying the exponential of $p_j(\mathbf{z})$ can be taken to be $(2\pi N_0)^{-1}$.

Thus, for our purposes,

$$p_j(\mathbf{z}) \approx \frac{1}{2\pi N_0} \exp \left[- \frac{1}{2N_0} d^2(\mathbf{z}, \mathbf{s}^{(j)}) \right], \quad (13)$$

where

$$d^2(\mathbf{z}, \mathbf{s}) \triangleq \|\mathbf{z} - \mathbf{s}\|^2 + 2\langle \mathbf{z}, \mathbf{s} \rangle + 2\alpha N_0$$

$$- 2| \sqrt{\|\mathbf{z}\|^2 \|\mathbf{s}\|^2 + 2\alpha N_0 \langle \mathbf{z}, \mathbf{s} \rangle + (\alpha N_0)^2}|. \quad (14)$$

The form of eq. (13) is very reminiscent of the density in Gaussian noise alone, and to further suggest such a similarity we refer to $d(\mathbf{z}, \mathbf{s})$ as the "distance" between $\mathbf{z}$ and $\mathbf{s}$. It is important to emphasize that this function is the key to assessing the combined effect of Gaussian noise and phase jitter on the system performance; through its use we
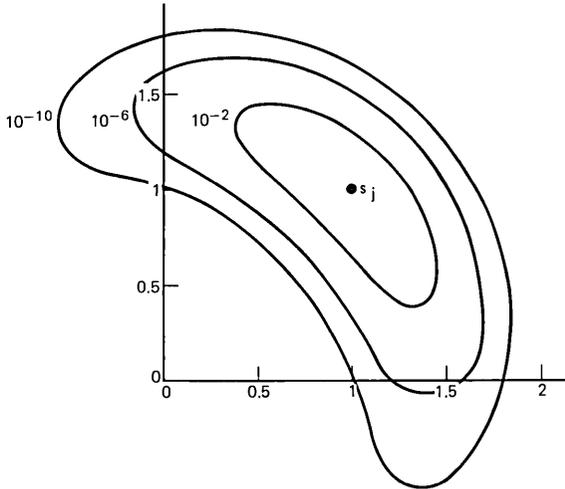
SNR = 25 dB, AND RMS PHASE JITTER = 9 DEGREES

Fig. 6—Constant distance contours: values of z for which

$$2\pi N_0 p_j(\mathbf{z}) = \exp\left\{-\frac{1}{2N_0} d^2(\mathbf{z}, \mathbf{s}_j)\right\} = 10^{-2}, 10^{-6}, 10^{-10}.$$

d^2 given by eq. (14).

are able to give a very suggestive geometric interpretation of the jitter phenomenon as well as accurate estimates of the error rate. In Fig. 6 we show some contours of constant "distance" (i.e., points which are equiprobable) about a given point. Note the angular orientation and similarity of these "banana" shaped contours to those obtained experimentally (Fig. 4).

Our procedure will be to use (13) and (14) to estimate the error rate for various constellations and detector structures; however, before we do this, we first wish to discuss some properties of the function $d(\mathbf{z}, \mathbf{s})$ which will be useful in estimating the system performance, and then to consider the conditional density $p_j(\mathbf{z})$ in the absence of a phase-locked loop.

2.4 Some Properties of the Jitter Distance $d(\mathbf{z}, \mathbf{s})$

(i) A requisite property that $d(\mathbf{z}, \mathbf{s})$ should possess is that for vanishingly small jitter the distance between points becomes Euclidean. This is easily verified by expanding (14) in terms of $1/\alpha$ and observing that

$$\lim_{1/\alpha \to 0} d^2(\mathbf{z}, \mathbf{s}) = \|\mathbf{z} - \mathbf{s}\|^2. \tag{15}$$

(*ii*) Since we do not expect the distance between an arbitrary point and the origin to be affected by phase jitter, it is reassuring to note that

$$d^2(\mathbf{z}, \mathbf{0}) = \|\mathbf{z}\|^2, \tag{16}$$

and that a similar property holds for points on the same ray, i.e.,

$$d^2(\mathbf{z}, k\mathbf{z}) = (1 - k)^2 \|\mathbf{z}\|^2, \qquad k > 0. \tag{17}$$

(*iii*) A "bonus" property is that even though (14) was derived for large α (small jitter), the expression

$$\lim_{\alpha \to 0} d^2(\mathbf{z}, \mathbf{s}) = |\ \|\mathbf{z}\| - \|\mathbf{s}\|\ |^2 \tag{18}$$

is quite reasonable in that, for completely incoherent systems, all points on the same circle are indistinguishable.

(*iv*) An interesting consequence of (14) is that points are now *closer* together than their Euclidean distance. To demonstrate this, we use the Schwartz inequality to show that the non-Euclidean part of $d^2(\mathbf{z}, \mathbf{s})$ is always negative, i.e.,

$$\langle \mathbf{z}, \mathbf{s} \rangle + \alpha N_0 \leqq |\ \sqrt{\|\mathbf{z}\|^2 \|\mathbf{s}\|^2 + 2\alpha N_0 \langle \mathbf{z}, \mathbf{s} \rangle + (\alpha N_0)^2}\ | \tag{19}$$

since squaring gives

$$\langle \mathbf{z}, \mathbf{s} \rangle^2 + 2\alpha N_0 \langle \mathbf{z}, \mathbf{s} \rangle + (\alpha N_0)^2 \leqq \|\mathbf{z}\|^2 \|\mathbf{s}\|^2 + 2\alpha N_0 \langle \mathbf{z}, \mathbf{s} \rangle + (\alpha N_0)^2$$

which upon cancelling becomes the Schwartz inequality

$$\langle \mathbf{z}, \mathbf{s} \rangle^2 \leqq \|\mathbf{z}\|^2 \|\mathbf{s}\|^2.$$

(*v*) If both the signal and noise power are scaled by the same constant for a given jitter level, α, it is clear that the conditional density is unchanged. Thus the signal-to-noise ratio and the mean-square jitter α are the natural parameters for characterizing the system.

(*vi*) A natural question to ask is whether or not $d(\mathbf{z}, \mathbf{s})$ is a convex metric* in the plane or, of more practical interest, if some sufficiently accurate approximation to $d(\mathbf{z}, \mathbf{s})$ is a convex metric in some circle centered about the origin. As we shall see later, if such an approximate representation can be obtained, some tedious error rate computations may be done quite simply. *While the requirements of symmetry and positivity can easily be shown to hold in the entire plane, J. E. Mazo has recently informed us that his results in Ref. 8 imply that $d(\mathbf{z}, \mathbf{s})$ is not a*

---

* A convex metric is a metric which possesses the midpoint property, i.e., for any two points $\mathbf{x}$ and $\mathbf{y}$ there is always a third point $\mathbf{z}$ such that $d(\mathbf{x}, \mathbf{z}) = d(\mathbf{z}, \mathbf{y}) = (1/2)d(\mathbf{x}, \mathbf{y})$.

convex metric in any circle about the origin for any value of $2N_0$. The question remains open as to whether or not $d(\mathbf{z}, \mathbf{s})$ admits an accurate convex metric approximation in a neighborhood of the origin. Appendix B reports the results of an investigation of this question.

By considering the above properties, it is apparent that a signal constellation will be relatively immune to small amounts of phase jitter either if the error rate (or minimum distance) is determined by a point at the origin and any other signal point, or if signal points on the same circle are widely separated (the more circles the greater the Gaussian noise penalty).

2.5 *Probability Density in the Absence of a Phase-Locked Loop*

In a subsequent section we will compare the performance of various signal constellations, and the following question naturally arises: Can we, by judicious signal design, eliminate the need for a phase-locked loop?[*] Preliminary to answering this question we must obtain the density of the demodulated signal in the absence of a phase-locked loop. For simplicity we model the jitter as arising from the single tone

$$\phi(t) = A \cos(\omega_j t + \psi), \tag{20}$$

where $2A$ is the peak-to-peak jitter, ω_j is the jitter frequency, and ψ is a uniformly distributed random phase. For this model, the jitter density is given by

$$p(\phi) = \begin{cases} \dfrac{1}{A\pi \sqrt{1 - \left(\dfrac{\phi}{A}\right)^2}} & |\phi| \leqq A \\ 0 & |\phi| > A \end{cases}. \tag{21}$$

To determine the density of the demodulated samples we use (9) and (10) to write

$$p_j(\mathbf{z}) = \frac{1}{2\pi N_0} \exp\left\{-\frac{1}{2N_0}\left[\|\mathbf{z} - \mathbf{s}^{(j)}\|^2 + 2\langle\mathbf{z}, \mathbf{s}^{(j)}\rangle\right]\right\}$$

$$\times \int_{-A}^{A} \exp\left\{\frac{1}{N_0}\left[(\mathbf{z}, \mathbf{s}^{(j)})\cos\phi + \langle\mathbf{z}, \mathbf{s}^{(j)}\rangle\sin\phi\right]\right\} p(\phi)d\phi$$

which for small (<12 degrees) peak-to-peak jitter becomes

$$p_j(\mathbf{z}) \approx \frac{1}{2\pi N_0} \exp\left[-\frac{1}{2N_0}\|\mathbf{z} - \mathbf{s}^{(j)}\|^2\right] M_\phi(\xi_j), \tag{22}$$

---

[*] The equalizer[4] is capable of determining the optimum demodulating static phase, so that the demodulation is noncoherent only to the degree that the untracked jitter degrades the error rate.

where $M_\phi(\cdot)$ is the moment-generating function of ϕ, and

$$\xi_j = \frac{1}{N_0} (z, s^{(j)\perp}).$$ (23)

For the jitter density given by (21) it is easy to show that

$$M_\phi(\xi_j) = I_0(A \,|\, \xi_j|);$$ (24)

thus we have the familiar form

$$p_j(z) = \frac{1}{2\pi N_0} \exp \left[-\frac{1}{2N_0} D^2(z, s^{(j)}) \right],$$ (25)

where

$$D^2(z, s) = \|z - s\|^2 - 2N_0 \, \ell n \, I_0 \left(\frac{A}{N_0} \,|\, (z, s^{(j)\perp}) \,| \right).$$ (26)

It is useful to summarize our work up to this juncture. Using the simple model (4) and the phase-error densities (7) and (21), we have derived the conditional density of the demodulated two-tuple with and without a phase-locked loop. Equations (13) and (14) and (25) and (26) are the desired expressions. In the next section we will use these densities to estimate the system error rate.

III. ESTIMATING THE ERROR RATE

In this section we give two estimates of the error rate: an asymptotic (high SNR) evaluation and an upper bound. Consider the arbitrary signal constellation shown in Fig. 7, where the decision regions are denoted by R_j. The detector, which is specified by the decision regions, will declare that $s^{(j)}$ has been transmitted if and only if the demodulated vector z falls inside R_j. Because of various practical considerations, principally ease of implementation, the mathematically optimum detector will not always be the one which is built.

3.1 Asymptotic (High SNR) Error Rate

The probability of error is given by

$$P_e = \sum_{j=1}^{M} p_j P_{ej},$$ (27)

where the p_j's are the (taken to be equal) a priori probabilities and P_{ej} is the conditional error rate. The conditional error rate is just the probability that z falls outside R_j when $s^{(j)}$ is transmitted, i.e.,

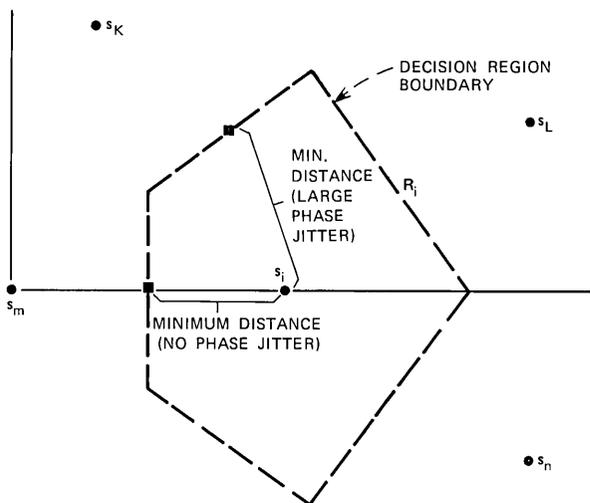$$P_{ej} = \Pr [z \notin R_j | s^{(j)} \text{ transmitted}].$$ (28)

Fig. 7—Typical decision region about signal point s_i, with minimum distances to boundary shown for Gaussian noise alone (no phase jitter) and for large jitter (angular displacements highly likely).

This quantity may be written, using the conditional density $p_j(z)$, as

$$P_{ej} = \int_{z \notin R_j} p_j(z) dz. \tag{29}$$

Let z_j^* denote a point of global minimum for the function d^2 or D^2 on the decision region boundary and let M_j be the number of times this minimum is achieved on the boundary. Let (u, v) denote an orthogonal coordinate system erected at z_j^* with the positive u axis pointed along the boundary line in a clockwise direction and v pointed outside R_j. Then, as is shown in Appendix A, for a high signal-to-noise ratio $(N_0 \to 0)$ with $\alpha N_0 = k$, the conditional error rate is given by

$$P_{ej} = \frac{M_j}{\dfrac{\partial d^2}{\partial v}} \sqrt{\frac{N_0}{2\pi \left| \dfrac{\partial^2 d^2}{\partial u^2} \right|}} \cdot \exp \left. \frac{-d^2}{2N_0} \right|_{z=z_j^*}. \tag{30a}$$

For the case of Gaussian noise alone (no phase jitter) d^2 becomes the ordinary Euclidean distance and the partials in (30a) are easily evaluated. In this case, with $\mathbf{s}^{(i)}$ the signal(s) closest to $\mathbf{s}^{(j)}$, we have

$$z_j^* = \frac{\mathbf{s}^{(i)} + \mathbf{s}^{(j)}}{2}$$

and

$$\frac{\partial d^2}{\partial v}\bigg|_{z=z_j^*} = \|\mathbf{s}^{(i)} - \mathbf{s}^{(j)}\|$$

$$\frac{\partial^2 d^2}{\partial u^2}\bigg|_{z=z_j^*} = 2.$$

Hence, for the Gaussian case,

$$P_{ej} = \frac{M_j}{2\|\mathbf{s}^{(i)} - \mathbf{s}^{(j)}\|} \sqrt{\frac{N_0}{\pi}} \exp \frac{-\|\mathbf{s}^{(i)} - \mathbf{s}^{(j)}\|^2}{8N_0}, \tag{30b}$$

a useful formula in its own right.

In terms of the distance functions $d(\cdot)$ and $D(\cdot)$, the asymptotic error rate is determined by the point on the decision boundary "closest" to the transmitted signal. Of course, in the presence of phase jitter, this point will generally differ from the closest point according to a Euclidean measurement. In Fig. 7, the indicated point on the vertical boundary segment is the closest point to $\mathbf{s}_j$ in the Euclidean sense. As phase jitter increases, those points with a radial coordinate nearly equal to that of $\mathbf{s}_j$ becomes closer to $\mathbf{s}_j$. So for large phase jitter, the point indicated on the boundary segment above $\mathbf{s}_j$ is the "closest" point. Thus the exponential decay in error rate is quite similar to the asymptotic Gaussian result since it is of the form

$$\exp \left[-d_{\min}^2(j)/2N_0\right], \tag{31}$$

where $d_{\min}(j)$ is the *minimum distance* (measured in a non-Euclidean manner) to the jth decision boundary. The minimum distance can sometimes be obtained analytically, but most often must be obtained by a computer search of the boundary.

The minimum distance to the jth decision boundary is particularly easy to determine if the function $d(\cdot,\cdot)$ is a metric which possesses the midpoint property, i.e., for each distinct pair of points $\mathbf{s}^{(i)}$ and $\mathbf{s}^{(j)}$ there exists a third point z^* such that

$$\tfrac{1}{2}d(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) = d(\mathbf{s}^{(i)}, \mathbf{z}_*) = d(\mathbf{z}_*, \mathbf{s}^{(j)}).$$

If this is the case, let $\mathbf{z}_i$ denote any point on the decision boundary between $\mathbf{s}^{(i)}$ and $\mathbf{s}^{(j)}$ and the triangle inequality gives

$$d(\mathbf{s}^{(j)}, \mathbf{z}_i) + d(\mathbf{s}^{(i)}, \mathbf{z}_i) \geqq d(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}). \tag{32a}$$

Since for maximum-likelihood detection

$$d(\mathbf{s}^{(j)}, \mathbf{z}_i) = d(\mathbf{s}^{(i)}, \mathbf{z}_i),$$

minimizing both sides of (32a) over the decision boundary gives

$$d(\mathbf{s}^{(j)}, \mathbf{z}^*) \geqq \min_{i \neq j} \tfrac{1}{2} d(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) \tag{32b}$$

where $\mathbf{z}^*$ is the point on the boundary closest to $\mathbf{s}^{(j)}$. Note that (32b) would provide an upper bound on (31). Since the midpoint $\mathbf{z}^*$ clearly lies on the boundary, $\mathbf{z}^* = \mathbf{z}_*$, and (32b) is satisfied with equality, the minimum distance is given by

$$d(\mathbf{s}^{(i)}, \mathbf{z}^*) = \min_{j \neq i} \tfrac{1}{2} d(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}). \tag{32c}$$

Since the Euclidean metric is convex, eq. (30b) could be directly obtained from (30a) by using (32c).

3.2 *An Upper Bound on the Error Rate (for Small Jitter)*

In systems which use a tracking loop, an upper bound on the system error rate may be obtained, for small jitter, by considering Fig. 8. This figure shows the transmitted point $\mathbf{s}^{(j)}$, the decision boundary R_j, and several nested regions C_j (defined by contours of constant probability), where

$$C_j = \{\mathbf{z} : d^2(\mathbf{z}, \mathbf{s}^{(j)}) \leqq c_j\}.$$

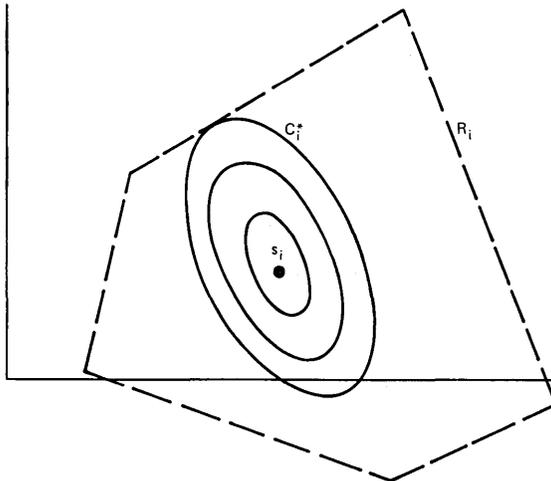Fig. 8—Nested equidistance contours [distance defined by eq. (13)] about signal s_i in arbitrary decision region. Contour c_i^*, at distance d_i^*, defines d_i^* as the shortest distance to the boundary.

If we let C_j^* denote the first contour which touches the boundary, then it is clear that

$$P_{ej} = \Pr\left[\mathbf{z} \notin R_j \mid \mathbf{s}^{(j)} \text{ transmitted}\right] \leq \Pr\left[\mathbf{z} \notin C_j^* \mid \mathbf{s}^{(j)}\right]$$

$$= \int_{\mathbf{z} \notin C_j^*} p_j(\mathbf{z}) d\mathbf{z}$$

$$= \int_{d^2\{\mathbf{z}, \mathbf{s}^{(j)}\} \geq c_j^{*2}} \exp\left[-\frac{1}{2N_0} d^2(\mathbf{z}, \mathbf{s}^{(j)})\right] d\mathbf{z}. \tag{33}$$

For small jitter, the exponent may be expanded in the first power of $1/\alpha$, to give

$$d^2(\mathbf{x}, \mathbf{s}) = \|\mathbf{x} - \mathbf{s}\|^2 - \frac{1}{\alpha N_0} \langle \mathbf{x}, \mathbf{s}^\perp \rangle^2, \tag{34}$$

for which the contours of constancy are ellipses. Transforming the ellipses into circles and changing to polar coordinates enable us to integrate (33) to get

$$P_{ej} \leq \frac{1}{\sqrt{1 - \dfrac{\|\mathbf{s}^{(j)}\|^2}{\alpha N_0}}} \exp\left[-d_{\min}^{(j)2}/2N_0\right]. \tag{35}$$

Again $d_{\min}^{(j)}$ is the minimum distance to the jth decision boundary which for convex polygonal decision boundaries can be determined analytically. For high SNR, the above bound is useful up to 1.5 degrees rms jitter. Because of the similarity of (30a) and (35), we will use only the former asymptotic results in the sequel.

IV. A COMPARISON OF VARIOUS SIGNAL CONSTELLATIONS

In the preceding section we have presented a means of evaluating the asymptotic (high SNR) error rate for a given signal constellation and detector structure. In terms of the minimum distance, measured via the appropriate noise/phase-jitter distance function to the jth decision boundary, we have

$$P_e \sim \sum_j p_j \frac{M_j}{2\|\mathbf{s}^{(i)} - \mathbf{s}^{(j)}\|} \sqrt{\frac{N_0}{\pi}} \exp\left[-\frac{1}{2N_0} d_{\min}^{(j)2}\right]. \tag{36}$$

The minimum distance will be obtained by a computer search of the decision boundary. It should be emphasized that comparisons based on the asymptotic error rate are not exact but rather indicate order-of-magnitude effects.

Our comparisons will be made by varying the following quantities:

(*i*)   signal constellations
(*ii*) signal-to-noise ratios
(*iii*) rms jitter
(*iv*) decision boundaries
(*v*)   phase-error density.

Clearly the pie may be sliced several ways, so let us first say a few words about each of the above variables.

(*i*) *Signal constellations*: For the purposes of signal evaluation we will consider the existing 16-point constellations QAM, 8–8, and (4, 90°) shown in Figs. 9a through 9c. The circular constellation (4, 90°) has signal points equally spaced (i.e., 90 degrees apart) on four circles. This large angular spacing of points on the same circle suggests that this constellation will be insensitive to small amounts of phase jitter. The ratio of outer radius to inner radius (r_2/r_1) for the 8–8 constellation is 1.59, found by Lucky[1] to minimize the error rate in Gaussian noise. 8–8 is an optimized form of AM-PM in which signal points on the outer circle do not lie on the same radial lines as those on the inner circle. It offers an order-of-magnitude improvement (over AM-PM) in error rate in the presence of Gaussian noise. We will also consider the new circular modulation formats 1–5–10, 1–6–9, 5–11, shown in Fig. 10. The optimum ratio (r_2/r_1) is very close to 2 for these constellations as we have determined by equating the three smallest nearest-neighbor distances.

(*ii*) We consider peak and average SNR's, P_{pk} and P_{avg} respectively, chosen so that P_e (no jitter) is $<10^{-5}$. These quantities are
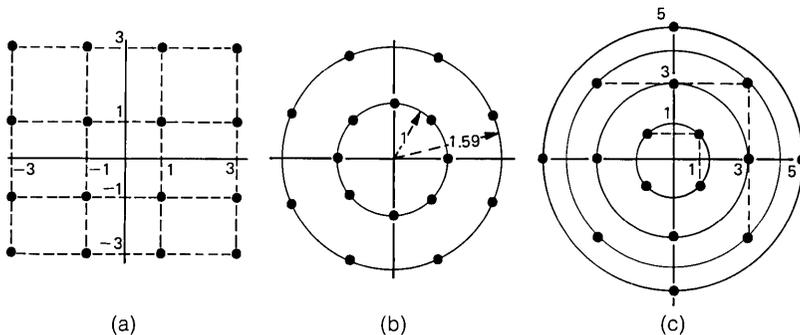

Fig. 9—Existing signal constellations: (a) quadrature amplitude modulation (QAM), $p_{peak}/p_{av} = 1.8$; (b) modified AM-PM (8–8), $p_{peak}/p_{av} = 1.43$; (c) circular constellation, (4, 90°) $p_{peak}/p_{av} = 1.85$.

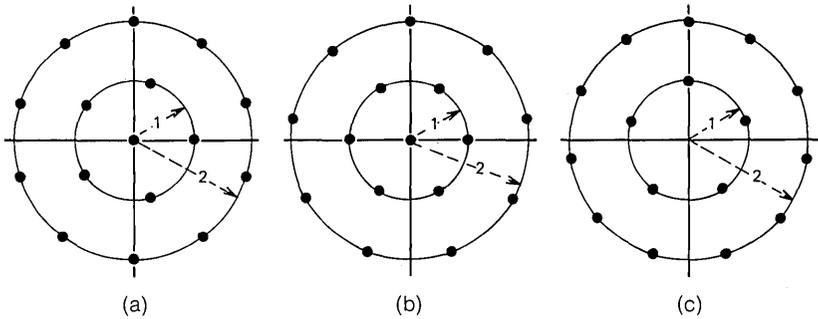Fig. 10—New signal constellations: (a) 1–5–10, $p_{peak}/p_{av} = 1.42$; (b) 1–6–9, $p_{peak}/p_{av} = 1.525$; (c) 5–11, $p_{peak}/p_{av} = 1.31$.

defined by

$$P_{pk} = \max_i \|\mathbf{s}^{(j)}\|^2/2N_0,$$

$$P_{avg} = \frac{1}{16} \sum_{j=1}^{16} \|\mathbf{s}^{(j)}\|^2/2N_0.$$

(*iii*) Our attention is focused on the practical range of residual (from a PLL) jitter of <3 degrees rms.

(*iv*) We will consider both the boundaries which are optimum in Gaussian noise (straight lines) as well as more practical boundaries for the circular formats (polar wedges).

(*v*) The Tikhonov density will be taken as representative of those systems which use a tracking loop, while the peak-to-peak density will be used to model the raw (untracked) phase jitter.

The error rate curves are grouped as follows: Figure 11 shows the error rate vs rms residual phase jitter (Tikhonov density) under average and peak power constraints for the six constellations described above. For each constraint, a Gaussian noise power (and thus an SNR) is assumed which places the curves in a useful operating range, and the receiver is presumed to use the Gaussian optimum decision region boundaries. These boundaries are piecewise-linear contours constructed from segments of perpendicular bisectors of lines joining signal points as shown in Fig. 12a for the 1–5–10 constellation. This is equivalent to deciding in favor of the signal point closest in Euclidean distance to the demodulated point. Figure 12b shows a more "practical" set of decision boundaries for the 1–5–10 constellation. Figure 11 indicates the immunity of the (4, 90°) constellation to small amounts of phase jitter at the expense of an error rate more than an order of magnitude greater
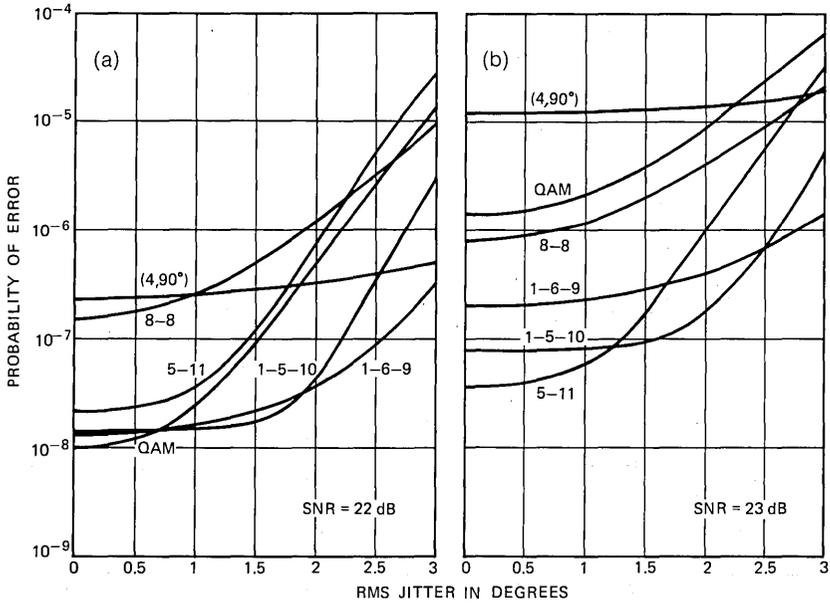
Fig. 11—Error rate vs jitter with a phase-locked loop: (a) average power constraint; (b) peak power constraint.

than those of the 1–5–10, 1–6–9, and QAM constellations. In practical operation with a tracking loop, any of these alternative constellations will almost always outperform the (4, 90°) constellation.

Figure 13 shows the error rates vs SNR, again under average and peak power constraints, in the presence of Gaussian noise alone (no phase jitter) and with 1.5 degrees rms residual phase jitter in addition
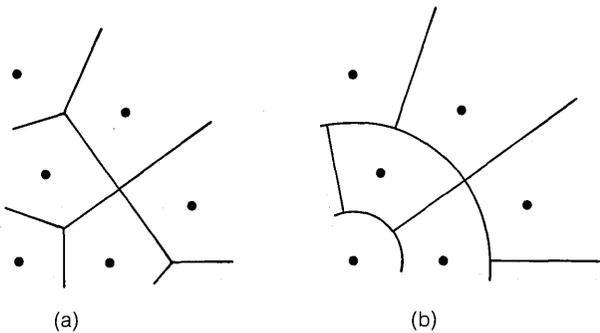

Fig. 12—Decision region boundaries for part of 1–5–10 constellation: (a) Gaussian-optimum decision region boundaries; (b) practical decision region boundaries.
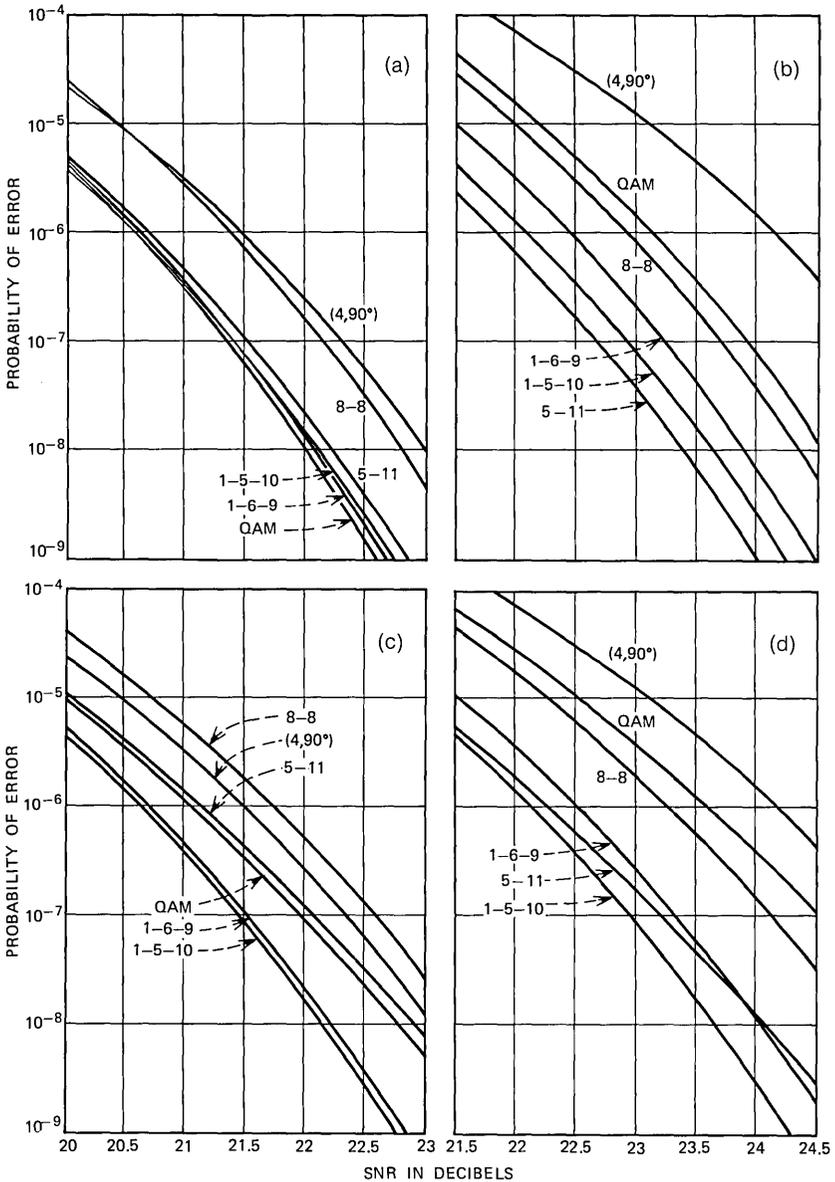
Fig. 13—Error rate vs SNR for channels with and without phase jitter; Gaussian-optimum receiver is assumed to have a phase-locked loop: (a) no jitter, average power constraint; (b) no jitter, peak power constraint; (c) rms jitter = 1.5 degrees, average power constraint; (d) rms jitter = 1.5 degrees, peak power constraint.
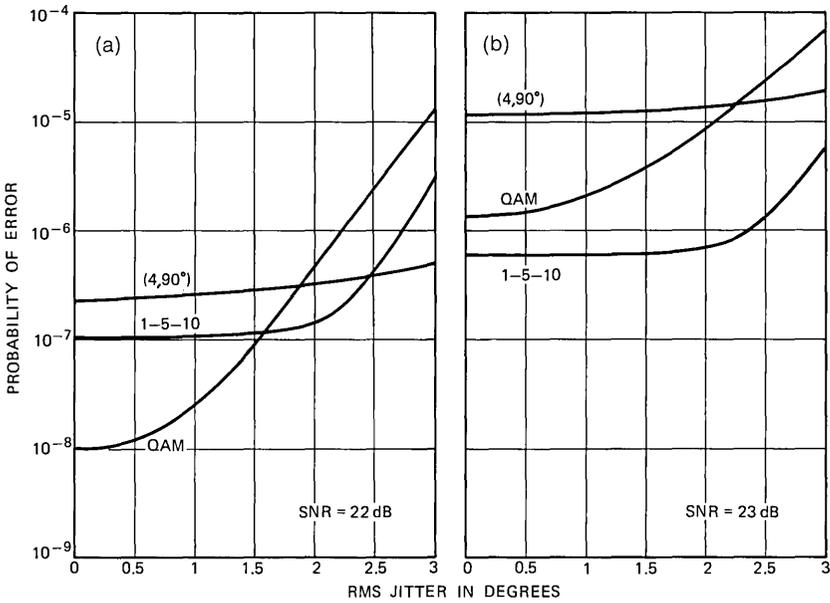
Fig. 14—Probability of error for receiver with "practical" decision region boundaries for the 1–5–10 constellation and with phase-locked loop: (a) average power constraint; (b) peak power constraint.

to the Gaussian noise. Although some of the curves shift their relative positions (at least under the average power constraint) when phase jitter is added, the good performances of 1–5–10 and 1–6–9 are maintained. QAM performs respectably and the (4, 90°) constellation comes in last.

Figure 14 is similar to a reduced Fig. 11 except that the "practical" set of decision region boundaries is presumed for the 1–5–10 constellation. As can easily be seen, the Gaussian optimum boundaries for QAM are also practical boundaries, and the jitter-immune (4, 90°) constellation is shown to best advantage by presuming Gaussian optimum boundaries. Under the average power constraint, QAM is superior to 1–5–10 below about 1.5 degrees rms jitter. Under the peak power constraint, 1–5–10 is uniformly superior to QAM. The (4, 90°) constellation does not show an advantage until the rms jitter reaches 2.5 to 3 degrees.

Figure 15 is a set of error rate vs SNR plots with practical decision region boundaries for 1–5–10. Curves are shown for Gaussian noise alone (no jitter) and 1.5 degrees rms jitter. Only the average power
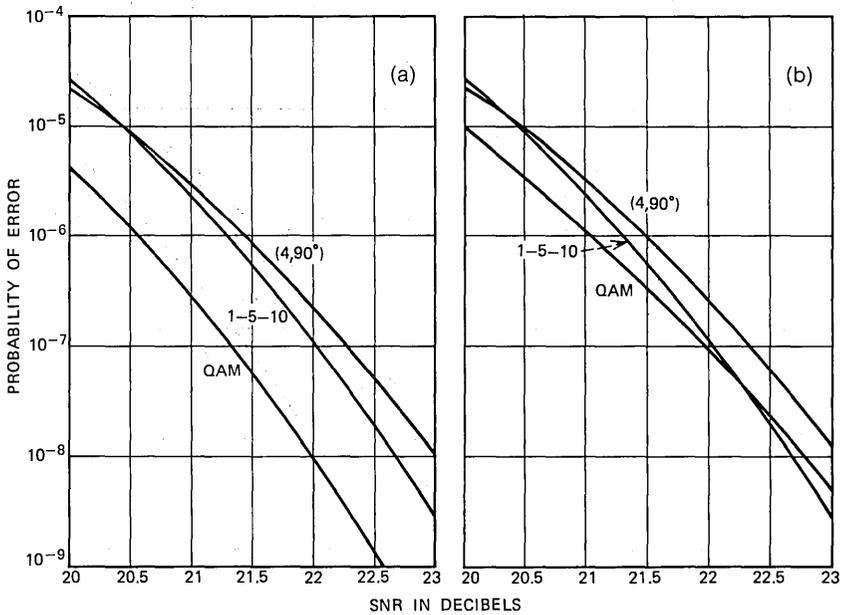
Fig. 15—Probability of error vs SNR for Gaussian-optimum receiver with "practical" decision region boundaries for the 1–5–10 constellation and with phase-locked loop: (a) no jitter, average power constraint; (b) rms jitter = 1.5 degrees, average power constraint.

constraint is presumed. As before, QAM shows an advantage in the absence of phase jitter and still does well in the presence of moderately severe residual phase jitter.

Figure 16 presents some data for receivers which do not use tracking loops. In this case the peak-to-peak density of eq. (22) describes the raw jitter. Curves are plotted vs peak jitter under average and peak power constraints. Some interesting features are the resistance of 1–5–10 up to a threshold of about 8 degrees peak-to-peak jitter and the rapid deterioration of the performance of QAM.

Figure 17 shows the performance vs SNR for the receivers without tracking loops when the peak-to-peak jitter is 12 degrees. This is the only instance for which the (4, 90°) constellation looks relatively good, but here, too, the 1–6–9 constellation performs slightly better. QAM, of course, does rather poorly.

The advantage of using a phase-locked tracking loop with QAM and 1–5–10 can be seen from the above data and a simple calculation. If the raw phase jitter is modeled by

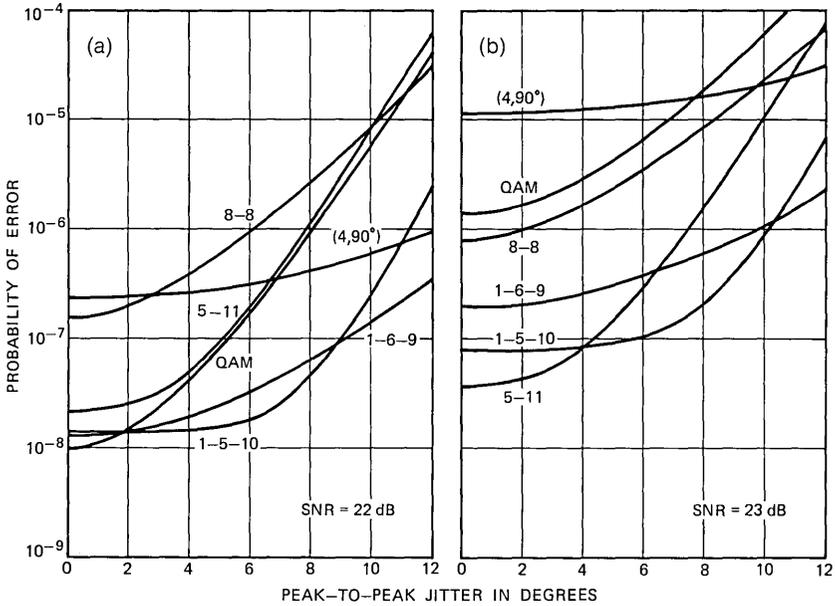$$Q(t) = A \cos [\omega_j t + \psi],$$

Fig. 16—Probability of error vs. peak-to-peak jitter for Gaussian-optimum receiver without a phase-locked loop: (a) average power constraint; (b) peak power constraint.

where ψ is uniformly distributed from $-\pi$ to π and $2A$ is the peak-to-peak jitter, then a rule of thumb[9] suggests that the residual rms jitter out of a tracking loop is of the order of $0.1 \times 2A$. For $A = 6$ degrees, this rms value is 1.2 degrees. A comparison of the curves of Figs. 11 and 16 shows that substantially lower error rates are achieved by the receiver with a tracking loop. The performance of constellation $(4, 90°)$ is relatively unaffected by the introduction of a tracking loop.

A further conclusion that can be drawn from the numerical data is that the QAM constellation, which is simple to generate and to demodulate, performs quite well in Gaussian noise alone or (with the aid of tracking loop) Gaussian noise plus phase jitter. More circular constellations, such as 1–5–10 and 1–6–9, appear to offer a moderate further advantage at the expense of greater complexity.

V. OPTIMUM SIGNAL CONSTELLATIONS UNDER A PEAK POWER CONSTRAINT

In this section we discretize the received signal space to obtain a tractable optimization problem. The discretizing is such that the M signal points are selected from a circle containing L points while the received points lie in a circle of N $(N > L)$ points (note: $M < L$). We make the following two comments concerning this approach to solving
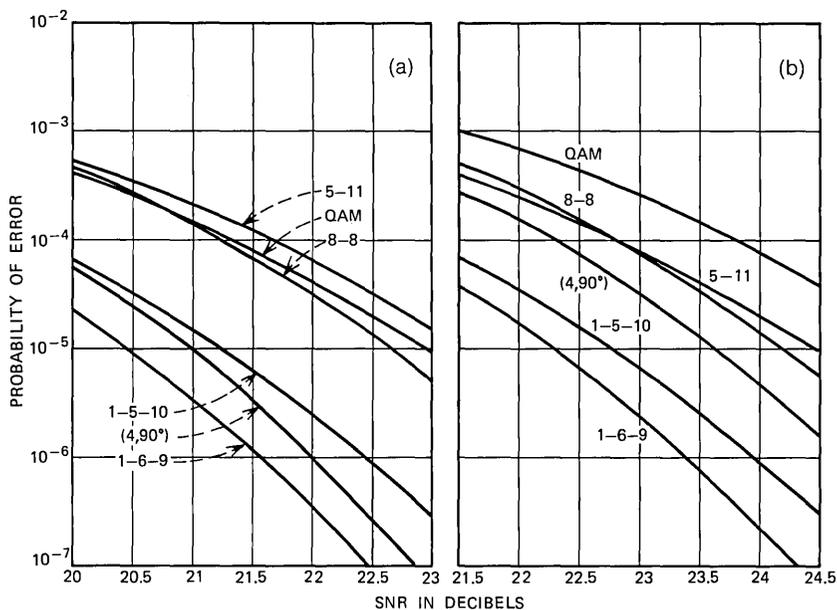
Fig. 17—Probability of error vs SNR for Gaussian-optimum receiver without a phase-locked loop: (a) peak-to-peak jitter = 12 degrees, average power constraint; (b) peak-to-peak jitter = 12 degrees, peak power constraint.

the problem:

(*i*) the level of discretization must be fine enough to provide a good approximation to the continuous problem, and

(*ii*) the outer radius must be chosen so that for all practical purposes the probability that a received point lies outside the outer circle is negligible. The peak power constraint simply means that no signal points can be selected outside the L circle.

5.1 *Discrete Maximum-Likelihood Formulation*

If we denote the received point, z, by "*i*" and the transmitted signal, $s^{(j)}$, by "*j*," then the perturbation of the transmitted signal due to Gaussian noise and phase jitter may be summarized by a transition matrix whose elements are defined by

$$p(i \mid j) = \text{Pr [receiving "}i"\mid \text{transmitting "}j"]$$
$$i = 1, 2, \cdots N \qquad (37)$$
$$j = 1, 2, \cdots M.$$

The transition probabilities may be computed by integrating the condi-

tional densities $p_j(z)$ over an appropriate region. It is convenient to work with the maximum-likelihood receiver (i.e., the optimum decision boundaries are used) which receives "i" and declares that "ℓ_i" was transmitted, where

$$p(i\,|\,\ell_i) > p(i\,|\,j), \; j \neq \ell_i \qquad j, \, \ell_i = 1, \, 2, \, \cdots M. \tag{38}$$

Note that we have (temporarily) fixed the M points in the signal constellation. It is easy to see that the probability of being correct is given by

$$\Pr\,[\text{correct}] = \sum_{i=1}^{N} \Pr\,[\text{correct}\,|\,\text{receive "}i\text{"}]\,\Pr\,[\text{receive "}i\text{"}], \tag{39}$$

but by Bayes rule

$$\Pr\,[\text{correct}\,|\,\text{receive "}i\text{"}]$$

$$= \Pr\,[\text{send "}\ell_i\text{"}\,|\,\text{receive "}i\text{"}]$$

$$= \frac{\Pr\,[\text{receive "}i\text{"}\,|\,\text{send "}\ell_i\text{"}]\,\Pr\,[\text{send "}i\text{"}]}{\Pr\,[\text{receive "}i\text{"}]}. \tag{40}$$

Substituting (40) in (39) and recalling that the transmitted signals are equiprobable gives

$$\Pr\,[\text{correct}] = \frac{1}{M} \sum_{i=1}^{N} \Pr\,[\text{receive "}i\text{"}\,|\,\text{send "}\ell_i\text{"}]$$

$$= \frac{1}{M} \sum_{i=1}^{N} p(i\,|\,\ell_i), \tag{41}$$

and the optimum constellation is the M signals (or columns) that maximize

$$\sum_{i=1}^{N} p(i\,|\,\ell_i). \tag{42}$$

Note that since $p(i\,|\,\ell_i)$ is the maximum entry in the ith row of the transition matrix, the problem is one of selecting M out of L columns such that the sum of the row maxima is maximized. The error rate may be determined from (41).

5.2 *Optimum Constellations*

A heuristic program for solving the combinational optimization problem posed by (42) has been developed by Kernighan and Lin.[6] Their process is based upon iterative improvement of either known initial constellations or random initial starts. For each start, a "locally

optimum" solution is found in the sense that no change of position of a single signal can improve the criterion. The heuristic process is very fast and, for the resolution we require, 20 random starts can be pursued to completion in 25 seconds. For particular values of rms phase jitter and noise power we find, among the best of 20 local optima, reaffirmation of known solutions and in some instances new competitive constellations. For example, for a peak signal energy of SNR = 22 dB, Figs. 18, 19, and 20 give the best among the 20 local optima for an rms of 0, 1.5, and 3 degrees, respectively. As expected, the 0-degree solution has a 5–11 character and the 1.5-degree solution has a 1–5–10 character. On the other hand, the 3-degree solution is somewhat of a surprise; it has a 1–6–9 character. The (4, 90°) constellation, which is

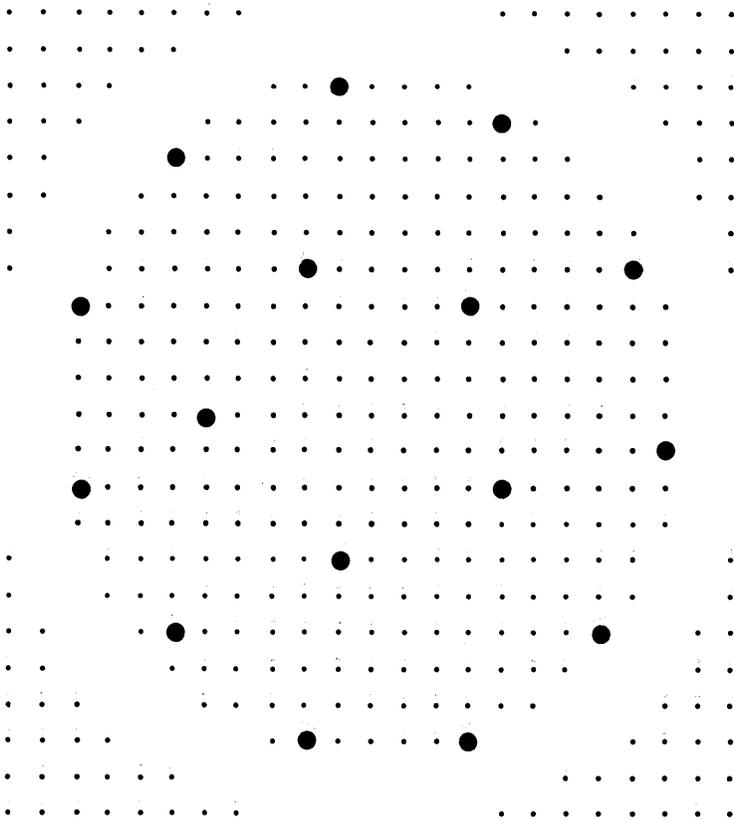Fig. 18—Optimum signal constellation: Peak SNR = 27 dB, no jitter (courtesy of B. W. Kernighan and S. Lin).
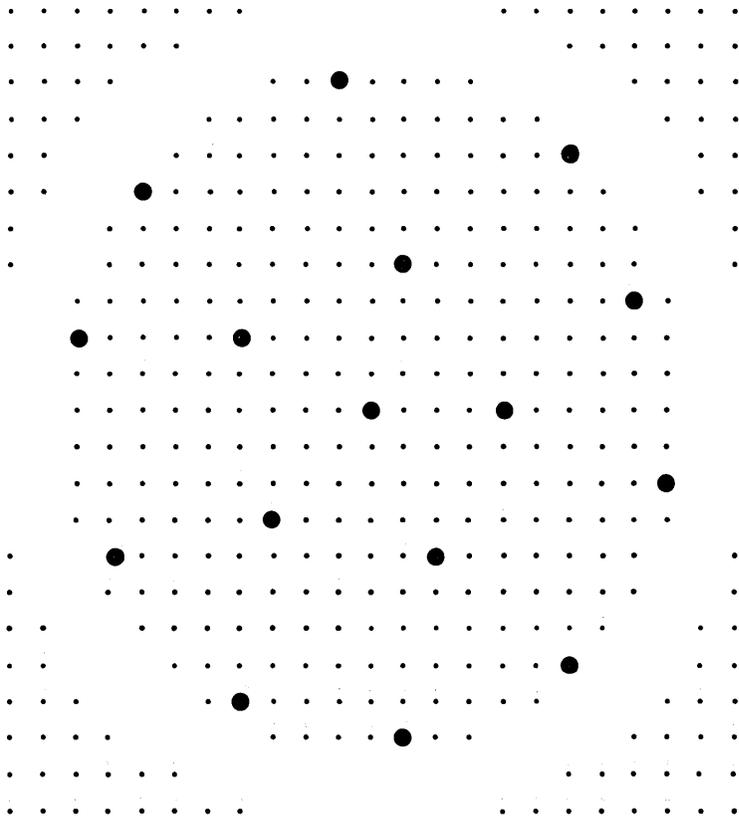
Fig. 19—Optimum signal constellation: Peak SNR = 27 dB, 1.5 degrees rms jitter (courtesy of B. W. Kernighan and S. Lin).

best at 3 degrees among heretofore considered designs, has an error rate only a few percent worse than that of the 1–6–9 constellation. A byproduct of the development of the above procedure is the demonstration of the fact that numerical quadrature routines offer a competitive alternative to asymptotic techniques and bounding methods for the estimation of system error rates.

VI. CONCLUSIONS

Comparisons have been made of several well-known two-dimensional signal formats in the presence of Gaussian noise and phase jitter, at high signal-to-noise ratios and under both peak and average power constraints. It has been demonstrated that, under an average
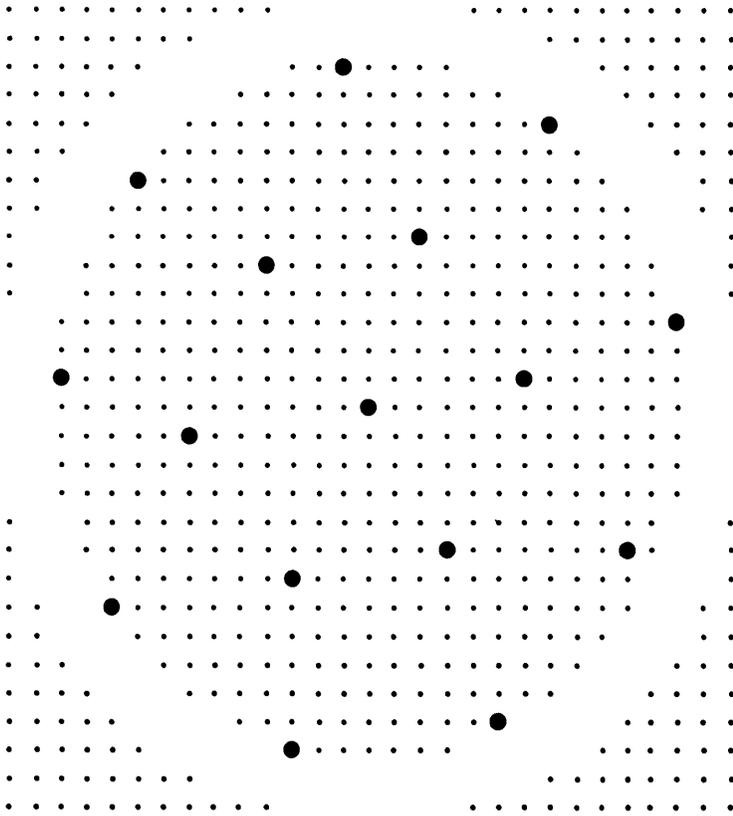
Fig. 20—Optimum signal constellation: Peak SNR = 27 dB, 3 degrees rms jitter (courtesy of B. W. Kernighan and S. Lin).

power constraint for systems which have a high-quality phase-locked loop (rms residual jitter <1 degree), QAM had the lowest error rate of all candidate constellations. If the residual jitter is <1.5 degrees rms, the 1–5–10 constellation becomes extremely attractive since it is immune to phase jitter in this range and provides the same asymptotic (no-jitter) error rate as QAM. For small amounts of jitter, 1–5–10 and QAM have a 2-dB SNR advantage over the (4, 90°) constellation which is "immune" to phase jitter. Both these constellations offer a 0.5-to-1-dB advantage in SNR over conventional AM/PM signaling techniques. Thus, under an average power constraint, both QAM and 1–5–10 merit consideration.

For a peak power constraint, in addition to making comparisons similar to the above, we have been able to attain the optimum signal

constellations for various levels of jitter. Our comparisons indicate that, for jitter <1 degree, QAM suffers a 1.5-dB SNR penalty with respect to the 5–11 constellation, while 1–5–10 suffers a 0.1-dB penalty. At 1.5 degrees rms jitter, 1–5–10 again is superior to both QAM and 5–11 (by 4 and 1 dB respectively).

Based upon the peak and average power constraints, the new modulation format 1–5–10 appears to make extremely efficient use of available signal power and for a slight increase in modulation/demodulation complexity offers considerable immunity to moderate (<1.5 degrees rms) residual phase jitter. QAM systems which employ high-quality phase-locked loops will also be operating very efficiently, provided that the residual phase error is <0.8 degree rms.

VII. ACKNOWLEDGMENTS

APPENDIX A

Method of Laplace

Let $g(x)$ and $h(x)$ be continuous real functions on $[a, b]$ where $h(x)$ is also twice continuously differentiable. Then, if $h(x)$ attains a single maxima at c $(a < c < b)$, we have that

$$\int_a^b g(x)e^{(1/k)h(c)}dx \sim g(c)e^{(1/k)h(a)}\sqrt{\frac{-2\pi k}{h''(c)}} \qquad (k \to 0)$$

(read $\sim$ as asymptotic to). The proof is not difficult and the key steps can be found in Papoulis[10] or in Jones.[11] This analysis technique for estimating an integral for large values of the parameter k is called the method of Laplace.

An immediate application of this method used in the body of this paper is

$$I_0(\alpha) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\alpha\cos\theta}d\theta \sim \frac{e^\alpha}{\sqrt{2\pi\alpha}} \qquad (\alpha \to \infty).$$

For estimating system error rates, a certain two-dimensional version of Laplace's method is needed. Particularly, we shall investigate the following two-dimensional integral:

$$I \triangleq \iint_{v \geq 0} \exp\left\{\frac{1}{k}h(\mathbf{z})\right\}d\mathbf{z} \qquad [\mathbf{z} = (u, v)]$$

for small k. The function $h(z)$ we shall be concerned with is assumed to have the following properties:

(i) $h(z)$ is twice continuously differentiable such that $h_v(0) < 0$, $h_{uv}(0) \neq 0$, and 0 is the unique point of maximum for $h(z)$ in $v \geqq 0$.[*] Thus we also assume h_{uu} of $(0) < 0$.

Let $G_r = \{v \geqq 0, u^2 + v^2 \leqq r\}$

(ii) For some $r > 0$,

$$\lim_{k \to 0} \int\!\!\int_{G_r} \exp\left\{\frac{1}{k} h(z)\right\} \Big/ I \to 1.$$

Rewriting I as

$$e^{(1/k)h(0)} \int\!\!\int_{\{v \geqq 0\}} \exp\left\{\frac{1}{k}\left[h(z) - h(0)\right]\right\},$$

it is easy to conclude that for each $\epsilon > 0$

$$I \sim e^{(1/k)h(0)} \int_{-\epsilon}^{\epsilon}\int_{0}^{\epsilon} \exp\left\{\frac{1}{k}\left[h(z) - h(0)\right]\right\}.$$

We shall proceed to integrate with the exponent in the integrand replaced by its local representation

$$e^{(1/k)h(0)} \int_{-\epsilon}^{\epsilon}\int_{0}^{\epsilon} \exp\left\{\frac{1}{k}\left[h_v(0)v + h_{uu}(0)u^2/2 + h_{uv}(0)uv\right]\right\} dvdu.$$

The absence of $h_u(0)$ and $h_{vv}(0)$ follow directly from (i). Integrating (dv) we get

$$I \sim e^{(1/k)h(0)} \int_{-\epsilon}^{\epsilon} k\left\{\frac{\exp\left\{\frac{1}{k}\left[h_v(0)\epsilon + h_{uu}(0)u\epsilon\right]\right\} - 1}{h_v(0) + h_{uv}(0)u}\right\}$$

$$\times e^{(1/k(h_{uu}(0)(u^2/2))}du.$$

We appeal to the first paragraph to integrate each term involved in this subtraction. Take ϵ small enough to avoid the singularity at $u = -h_v(0)/h_{uv}(0)$. The first term is asymptotic to

$$k^{3/2}\sqrt{\frac{-2\pi}{h_{uu}(0)}}\left\{\exp\frac{1}{k}\left[h(0) + h_v(0)\epsilon - \frac{h_{\mu\nu}^2}{h_{uu}}\epsilon^2\right]\right\}$$

---

[*] In this appendix, subscripts are used to denote partial derivatives, e.g.,

$$h_v(0) = \left.\frac{\partial h}{\partial v}\right|_{(u,\,v)=(0,\,0)}.$$

while the second is

$$-\frac{k^{3/2}}{h_v(\mathbf{0})} \sqrt{\frac{-2\pi}{h_{uu}(\mathbf{0})}} \, e^{h(\mathbf{0})/k}.$$

Since for ϵ small enough $h_v(\mathbf{0})\epsilon - [h_{\mu v}^2(\mathbf{0})/h_{uu}(\mathbf{0})]\epsilon^2 < 0$, we conclude

$$I \sim \frac{-k^{3/2}}{h_v(\mathbf{0})} \sqrt{\frac{-2\pi}{h_{uu}(\mathbf{0})}} \, e^{h(\mathbf{0})/k}.$$

In error rate computations one is often integrating over the exterior of a convex polygon. In the body of this paper we encounter the case where $h(z)$ has a finite number of global maxima on the boundary, at most one on each side, and none at the vertices. Let $\mathbf{z}_n^*$ be the nth local maximum. Map the exterior half-space containing $\mathbf{z}_n^*$ into the upper-half plane via a rotation, composed with a translation taking $\mathbf{z}_n^* \to \mathbf{0}$. Then the method of the last paragraph can be applied. The process is repeated for each maxima and the results sum to the required asymptotic estimate of the exterior integral. The fact that the exterior half-spaces containing distinct points of maxima may overlap is of no consequence. Furthermore, in our applications, $h(\mathbf{z})$ is symmetric with respect to each $\mathbf{z}_k^*$ and the process need only be completed once and the answer multiplied by the multiplicity of the maxima.

APPENDIX B

The Nature of $d(x, y)$

B.1 Introduction

Let V be a vector space endowed with a scalar product $\langle \mathbf{x}, \mathbf{y} \rangle$ and a norm derived therefrom. It is not known whether $d_\gamma(\mathbf{x}, \mathbf{y})\colon V \times V \to |R|$ defined by

$$d_\gamma(x, y) = (\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\gamma^{-1} - 2|\sqrt{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + 2\gamma^{-1}\langle \mathbf{x}, \mathbf{y}\rangle + \gamma^{-2}}|)^{1/2}$$

is a metric for any values of γ $(0 < \gamma < \infty)$. Nor is it known whether $d_\gamma(\mathbf{x}, \mathbf{y})$ is a metric in any sphere centered about the origin. Similarly, the status of the midpoint property is also unknown to us. However, from Ref. 8, we know that these two properties cannot hold simultaneously for any γ in any sphere about the origin.

Concerning the metric question, the difficulty (as usual) is the triangle inequality. By definition $d_\gamma(\mathbf{x}, \mathbf{y})$ is symmetric in its arguments. Positivity follows easily from the Schwartz inequality,

$$d_\gamma(\mathbf{x}, \mathbf{y}) \geq \{\|\mathbf{x} - \mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle + 2\gamma^{-1}$$
$$- 2|\sqrt{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + 2\gamma^{-1}\|\mathbf{x}\|\|\mathbf{y}\| + \gamma^{-2}}|\} = |\,\|\mathbf{x}\| - \|\mathbf{y}\|\,| \geqq 0.$$

Notice $d_\gamma(\mathbf{x}, \mathbf{y}) > 0$ unless $x = y$. If $d_\gamma(x, y)$ is a metric on $V \times V$ for a particular value of γ then it is a metric for all γ $(0 < \gamma < \infty)$; this is easily obtained by employing the mapping $\mathbf{z} \to |\sqrt{\gamma}|\mathbf{z}$.

B.2 One-Dimensional Case

In the special case $V = R$ we can show $d_1(x, y)$ is a metric. In one dimension we have (the "1" subscript will now be suppressed)

$$d(x, y) = \begin{cases} |x - y| & xy + 1 \geqq 0 \\ |\sqrt{(x + y)^2 + 2^2}| & xy + 1 \leqq 0 \end{cases}.$$

To show the triangle inequality, first notice that, if three points a, b, and c are on the same side of zero, the distances are all Euclidean. So for the remaining cases to be investigated we assume one point has a different sign than the other two. Notice $d(x, y) = d(-x, -y)$ so we lose no generality by assuming $c \geqq b \geqq 0 \geqq a$. Two subcases remain: (i) only $d(c, a)$ is non-Euclidean; (ii) $d(c, a)$ and $d(b, a)$ are non-Euclidean. For (i) the distances involved are $(c - b)$, $(b - a)$, and $|[(a + c)^2 + 2^2]^{1/2}|$. Now $|[(a + c)^2 + 2^2]^{1/2}| \leqq c - a$ since, by squaring, this is equivalent to $1 + ac \leqq 0$. To show

$$(c - b) \leqq (b - a) + |[(a + c)^2 + 2^2]^{1/2}|$$

and

$$(b - a) \leqq (c - b) + |[(a + c)^2 + 2^2]^{1/2}|,$$

it is enough to show

$$[(c + a) - 2b]^2 \leqq [(a + c)^2 + 2^2]$$

since squaring both sides when the right-hand side is positive can only weaken the inequality. The last inequality can be simplified to $b^2 - bc \leqq 1 + ab$ for which the left-hand side is negative and the right-hand side is not. On the other hand, (ii) is immediate since $c - b$, $\sqrt{(a + c)^2 + 4}$, and $\sqrt{(b + a)^2 + 4}$ can be identified as sides of a triangle with apex $(-a, 2)$ and base points c and b. Notice d does not have the midpoint property since $d(10, -10) = 2$, yet the only points y satisfying $d(10, y) = 1$ are 9 and 11, but both $d(9, -10)$ $d(11, -10)$ exceed 2.

This last observation shows that the open spheres in this metric space are not all connected. In two dimensions, the boundaries of certain open spheres are disconnected; specifically, it can be shown that for certain values of $q > 0$

$$\{y \,|\, d(\mathbf{x}, \mathbf{y}) = q\}$$

is a disconnected set if $\|\mathbf{y}\|^2 \gamma > 1$.

B.3 *Approximating $d_\gamma(x, y)$ on the Unit Circle*

As mentioned in the text, an important open question is whether $d_\gamma(x, y)$ can be accurately approximated by a convex metric. There are, of course, many ways in which one can approximate $d_\gamma(x, y)$. In this section we dispose of two approximations which suggest themselves immediately.

Let us view $d_\gamma(\mathbf{x}, \mathbf{y})$ on the unit circle. Notice, as $\gamma \to 0$,

$$| \sqrt{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + 2\gamma^{-1}\langle\mathbf{x}, \mathbf{y}\rangle + \gamma^{-2}} |$$
$$\approx \gamma^{-1} \left(1 + \gamma\langle\mathbf{x}, \mathbf{y}\rangle + \gamma^2 \left(\frac{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 - \langle\mathbf{x}, \mathbf{y}\rangle^2}{2} \right) \right).$$

Hence, for small phase jitter,

$$d_\gamma^2(x, y) \approx \|\mathbf{x} - \mathbf{y}\|^2 - \gamma\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \sin^2\theta,$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$, or what is the same,

$$d_\gamma(\mathbf{x}, \mathbf{y}) \approx \|\mathbf{x} - \mathbf{y}\| \left(1 - 4\gamma \frac{A^2(\mathbf{x}, \mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|^2} \right)^{1/2},$$

where $A(\mathbf{x}, \mathbf{y})$ is the area of the triangle $0xy$. Let h_{xy} denote the altitude of $0xy$ perpendicular to xy; then

$$d_\gamma(\mathbf{x}, \mathbf{y}) \approx \|\mathbf{x} - \mathbf{y}\|(1 - \gamma h_{xy}^2)^{1/2}.$$

Since $h \leq 1$,

$$\|\mathbf{x} - \mathbf{y}\|(1 - \gamma h_{xy}^2)^{1/2} \geq 0$$

with equality if and only if $\mathbf{x} = \mathbf{y}$ so long as $\gamma < 1$; also, the left-hand side is symmetric. Notice

$$\Delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|(1 - h_{xy}^2)^{1/2}$$

has the midpoint property. Indeed, if z lies on the line segment joining $\mathbf{x}$ and $\mathbf{y}$,

$$d(\mathbf{x}, z) + d(z, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})$$

since $h_{xz} = h_{xy} = h_{yz}$.

So we are strongly motivated now to see if $\Delta(\mathbf{x}, \mathbf{y})$ is a metric on the unit disc. Note that if F is any finite subset of the unit disc, then for γ sufficiently small, $d(\mathbf{x}, \mathbf{y})$ is a metric on F. This follows from the fact that for noncolinear triples, the triangle inequality holds properly for the metric $\|\mathbf{x} - \mathbf{y}\|$. Perhaps surprisingly, the triangle inequality for $\Delta_\gamma(\mathbf{x}, \mathbf{y})$ does *not* hold for all triplets in the unit disc. This was discovered by linearizing $\Delta(\mathbf{x}, \mathbf{y})$ to get $\delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|(1 - \gamma/2h_{xy}^2)$ which is likewise a valid approximation for $d(\mathbf{x}, \mathbf{y})$ for small γ.

For $\delta(\mathbf{x}, \mathbf{y})$, the midpoint property and all requirements for a metric hold except for the triangle inequality. The triangle inequality question is easier for us to investigate for $\delta(\mathbf{x}, \mathbf{y})$ than for $\Delta(\mathbf{x}, \mathbf{y})$ as it develops into a geometrical extremal problem which we can handle. In solving the extremal problem, a class of vector triplets in the unit disc emerge for which no value of $\gamma > 0$ exists such that the triangle inequality holds uniformly for the class. Although these triplets arise in the analysis of $\delta(\mathbf{x}, \mathbf{y})$, they serve just as ruinously for $\Delta(\mathbf{x}, \mathbf{y})$.

In order to discuss this geometric extremal problem, we require some notation. Suppose we have a triangle (nontrivial—positive area) in the unit disc with side lengths a, b, and c and opposing vertices A, B, and C, respectively. Let h_a, h_b, and h_c denote the distance from the origin to the line containing the side of length a, b, and c, respectively.

Write the "triangle inequality" expression for $\delta(\mathbf{x}, \mathbf{y})$:

$$\|\mathbf{x} - \mathbf{y}\| \left(1 - \frac{r}{2} h_{xy}^2\right) \overset{?}{\leqq} \|\mathbf{x} - \mathbf{z}\| \left(1 - \frac{r}{2} h_{xz}^2\right)$$
$$+ \|\mathbf{z} - \mathbf{y}\| \left(1 - \frac{r}{2} h_{zy}^2\right).$$

Changing to the notation just introduced, we have directly that the triangle inequality holds for $0 \leqq \gamma \leqq \boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = \inf \Gamma(a, b, c)$.

$$\Gamma(a, b, c) \overset{\Delta}{=} \left\{ \frac{a + b - c}{ah_a^2 + bh_b^2 - ch_c^2} \right\}$$

and the infimum is over all triangles for which the bracketed expression is positive. This geometric extermal problem is disposed of by substituting a triangle of the form depicted in Fig. 21. It follows easily $\lim_{\epsilon \to 0} \Gamma = 0$. Thus $\boldsymbol{\gamma} = 0$ and the only $d(\mathbf{x}, \mathbf{y})$ metric is the obvious one. A straightforward substitution of the three vectors depicted above into the "triangle inequality" for $\Delta(\mathbf{x}, \mathbf{y})$ yields again that for no fixed $\gamma > 0$ can the triangle inequality hold for this family of triangles.

Strikingly, for the class of triplets in Fig. 21, it is easily demonstrated, via substitution, that the triangle inequality for $d_\gamma(x, y)$ holds for an open interval including $\gamma = 0$!

B.4 *A Metric on the Boundary of the Disc*

We end on a positive note: $\delta(x, y)$ is a metric on the boundary of the unit disc for a nontrivial interval $[0, \boldsymbol{\gamma}]$.* It is enough to show inf

---

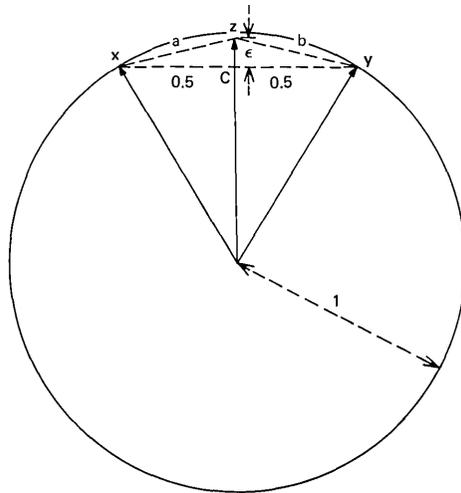* We anticipate applications in phase-modulation systems.

Fig. 21—Infimizing family of vectors.

$\Gamma(a, b, c) > 0$ where the infimum is over those triangles circumscribed by the unit circle for which Γ is positive. Now $h_a^2 = 1 - a^2/4$ and similarly for h_b and h_c so

$$\Gamma(a, b, c) = \left(1 - \frac{1}{4} \frac{a^3 + b^3 - c^3}{a + b - c}\right)^{-1}.$$

Dividing gives

$$(a, b, c) = \left(1 - \frac{1}{4}\left[a^2 + b^2 + c^2 - ab + ac + bc - \frac{3abc}{a + b - c}\right]\right)^{-1}.$$

So $\gamma > 0$ if and only if

$$\sup \frac{abc}{a + b - c} < \infty.$$

At this point, we must digress and recall some plane geometry from Ref. 12. First $abc/4$ is the area of the triangle and $2^{-1}(a + b + c)$ is called a semiperimeter. Given any triangle ABC, extend the two lines emanating from the apex A. Construct bisectors to the two exterior angles complementary to the angles B and C respectively. The bisectors meet in a point equidistant from the three lines containing a, b, and c. The circle tangent to these three lines is called an excircle. See Fig. 22. The center of the excircle (where the bisectors meet) is called the excenter and of course the radius is called the exradius. Each triangle has three excircles. Finally from Ref. 12 we need:
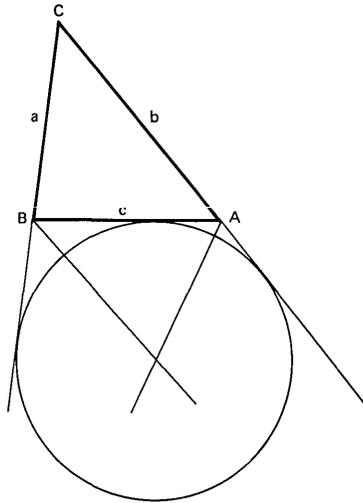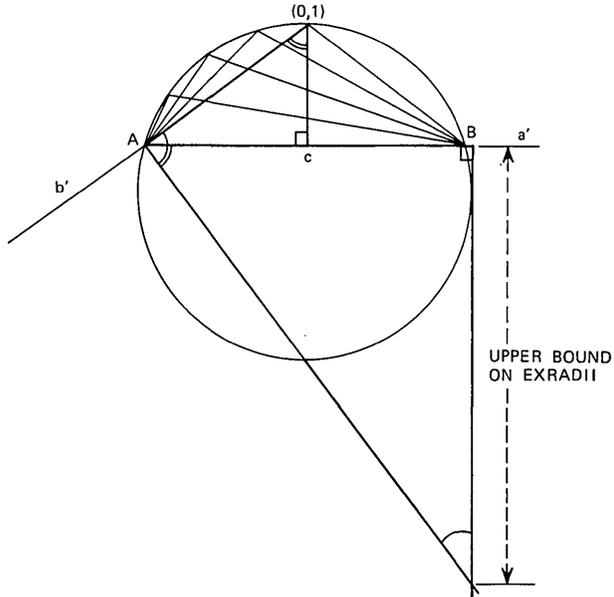
Fig. 22—Excircle tangent to c.

Fig. 23—Upper exradius bound for excircles tangent to c. Since $0 \leqq c \leqq 2$, the exradii are uniformly bounded.

Theorem: An exradius of a triangle is equal to the ratio of the area to the difference between the semiperimeter and the side to which the excircle is tangent internally.

So if the set of exradii of triangles circumscribed by the unit circle is uniformly bounded away from infinity, then $\gamma > 0$ and $\{\delta_\gamma\}_{0 < \gamma < \gamma}$ are metrics.

To complete the proof, consider any triangle with vertices on the unit disc boundary. With an isometric transformation of the circle into itself, we can situate c horizontally with apex C above c in the left half-plane. The more obtuse the exterior angles at A and B, the larger the excircle tangent to c. So take a horizontal line a' through B and a line b' going through A and $(0, 1)$ and replace a and b with a' and b'. Clearly, bisectors constructed using a' and b' will intersect at a point further away from c than the excenter of any excircle tangent to c. By similar triangles, the primed bisectors intersect at a distance $2(1 \pm \sqrt{1 - c^2/4})$ from c where the sign is plus if c lies in the lower half-plane and negative otherwise (Fig. 23).

REFERENCES

1. Lucky, R. W., and Hancock, J. C., "On the Optimum Performance of N-ary Systems Having Two Degrees of Freedom," IEEE Trans. Commun. Syst. (March 1962).
2. Salz, J., Sheehan, J. R., and Paris, D. J., "Data Transmission by Combined AM and PM," B.S.T.J., *50*, No. 7 (September 1971), pp. 2399–2419.
3. Thomas, C. M., "Amplitude Phase-Keying with M-ary Alphabets: A Technique for Bandwidth Reduction," Int. Telemetering Conf. Proc., *VIII* (October 1972).
4. Gitlin, R. D., Ho, E. Y., and Mazo, J. E., "Passband Equalization of Differentially Phase-Modulated Data Signals," B.S.T.J., *52*, No. 2 (February 1973), pp. 129–238.
5. Viterbi, A. J., *Principles of Coherent Communication,* New York: McGraw-Hill, 1966, pp. 86–96 and 118–119.
6. Kernighan, B. W., and Lin, S., "Heuristic Solution of a Signal Design Optimization Problem," Proc. Seventh Annual Princeton Conf. on Information Sciences and Systems, March 1973.
7. Foschini, G. J., Gitlin, R. D., and Weinstein, S. B., "Optimization of Two-Dimensional Signal Constellations in the Presence of Gaussian Noise," to be published in IEEE Transactions on Communications.
8. Mazo, J. E., "A Note on Metrics and Metric Convexity," unpublished work.
9. Falconer, D. D., unpublished work.
10. Papoulis, A., *The Fourier Integral and Its Application,* New York: McGraw-Hill, 1962.
11. Jones, D. S. "Asymptotic Behavior of Integrals," SIAM Review, *14*, No. 2 (April 1972) pp. 286–317.
12. Davis, D. R., *Modern College Geometry,* Reading, Mass.: Addison-Wesley, 1957.

A Theory of Traffic-Measurement Errors for Loss Systems With Renewal Input

By S. R. NEAL and A. KUCZURA

A theory of traffic-measurement errors for loss systems with renewal input is developed. The results provide an accurate approximation for the variance of any differentiable function of one or more of the following basic traffic measurements taken during a given time interval:

   (i) The total number of attempts (peg count)
   (ii) The number of unsuccessful attempts (overflow count)
   (iii) The usage based on discrete samples (TUR measurement) or on continuous scan.

   The approximation is given in terms of the individual variances and covariance functions of the three measurements. Asymptotic approximations for these moments are obtained using the concept of a generalized renewal process, and are shown to be sufficiently accurate for telephone traffic-engineering purposes.

   As an application of the theory, we examine the variances of the standard estimates of the load and peakedness (variance-to-mean ratio) of an input traffic stream for a time interval of one hour. Other possible applications to Bell System trunking problems are discussed.

I. INTRODUCTION

In the Bell System, there are a number of traffic measurements which can be made on any given trunk group. For a standard time interval $(0, t]$ of one hour, the three most important measurements are:

   (i) $A(t)$, the number of attempts (peg count);
   (ii) $O(t)$, the number of unsuccessful attempts (overflow count); and
   (iii) $L_d(t)$, an estimate of usage based on 36 discrete samples (TUR measurement).

When all three measurements are available, several statistics can be formed to estimate traffic parameters of interest. For instance, the

ratio $O(t)/A(t)$ is an estimate of call congestion. Two other important parameters are the peakedness (variance-to-mean ratio) and the load of the input traffic. An estimate of the load is given by the function

$$\hat{\alpha} = \frac{L_d(t)/36}{1 - \dfrac{O(t)}{A(t)}},$$

while an estimate of peakedness is a complicated function of $A(t)$, $O(t)$, and $L_d(t)$ which is usually obtained by iteration using the Equivalent Random method.[1]

Since the trunk-engineering procedures are based on such estimates, it is important to know their statistical accuracy. For instance, it would be useful to know the error inherent in a prediction of the required size of a trunk group (to obtain a specified grade of service) based on the estimates of offered load and peakedness of the input traffic. Such a result could be used to determine the number of single-hour measurements necessary to ensure a desired accuracy in the prediction, to determine the optimum number of measurements from a cost-effectiveness point of view, or to evaluate the consequences for trunk provisioning of using a given number of measurements.

Many results concerning the accuracy of the individual traffic measurements (*i*) through (*iii*) have been obtained previously, but most of these have assumed the arrival process to be Poisson. For example, assuming Poisson arrivals, the variance of the usage measurement $L_d(t)$ was obtained by Beneš,[2] and the variance of the measured call-congestion $O(t)/A(t)$ was given by Descloux.[3] More recently, the variance of $O(t)/A(t)$ was obtained for arbitrary renewal input by Kuczura and Neal.[4] The variances of some nonstandard traffic counts were considered by Descloux,[5] and numerical results were obtained for the case of Poisson input.

Using the concept of a multidimensional renewal process, we develop a general theory of errors which provides an estimate of the variance of any differentiable function of the measurements (*i*) through (*iii*). Consequently, our results can be used to answer many questions similar to those mentioned above. The variances of the estimates of offered load and peakedness of the input traffic will be derived in Section IV as examples which illustrate our general theory.

Section II contains the derivation of an approximation for the variance of a function of the three traffic measurements. The approximation is given in terms of the individual variances of (*i*), (*ii*), and (*iii*) and the covariance functions between them.

Section III contains the mathematical model used to derive the variances and covariances. Section V contains a summary and an outline of other possible applications.

II. STANDARD ERRORS OF FUNCTIONS OF RANDOM VARIABLES

For completeness, we present those results from the theory of standard errors[6] which are required below. Let ξ_1, ξ_2, and ξ_3 be random variables and g a real-valued function. Assume that ξ_i has a mean θ_i, $g(\xi_1, \xi_2, \xi_3)$ has finite mean and variance, and g is differentiable at the point $(\theta_1, \theta_2, \theta_3)$. Using a Taylor series expansion we have, to first order,

$$g(\xi_1, \xi_2, \xi_3) - g(\theta_1, \theta_2, \theta_3) \approx \sum_{i=1}^{3} (\xi_i - \theta_i) \frac{\partial g}{\partial \theta_i}. \tag{1}$$

We are assuming that the observation period will be sufficiently large so that the contribution of the higher-order terms can be neglected. This assumption will be justified in our model.

Taking the expectation of (1), we see that the mean value of g is approximately $g(\theta_1, \theta_2, \theta_3)$ since $E[\xi_i - \theta_i] = 0$. We also have

$$\text{Var}\,[g(\xi_1, \xi_2, \xi_3)]$$

$$\approx E\left[\,\sum_{i=1}^{3} (\xi_i - \theta_i) \frac{\partial g}{\partial \theta_i}\right]^2$$

$$= \sum_{i=1}^{3} \left(\frac{\partial g}{\partial \theta_i}\right)^2 \text{Var}\,[\xi_i] + 2 \frac{\partial g}{\partial \theta_1} \frac{\partial g}{\partial \theta_2} \text{Cov}\,[\xi_1, \xi_2]$$

$$+ 2 \frac{\partial g}{\partial \theta_1} \frac{\partial g}{\partial \theta_3} \text{Cov}\,[\xi_1, \xi_3] + 2 \frac{\partial g}{\partial \theta_2} \frac{\partial g}{\partial \theta_3} \text{Cov}\,[\xi_2, \xi_3]. \tag{2}$$

After setting $\xi_1 = A(t)/t$, $\xi_2 = O(t)/t$, and $\xi_3 = L_d(t)/36$ the above relation becomes the starting point for our theory of traffic-measurement errors. It approximates the variance of any differentiable function of the measurements in terms of their variances and covariances.

In the next section we derive expressions for the required moments. These, together with the first partial derivatives of g, approximately determine the variance of g. If the function g is too complicated to be differentiated analytically, differencing may be used to approximate the partial derivatives. An example of this procedure is given in Section IV where we discuss the variance of an estimate of the peakedness of a stream of offered traffic.

III. MATHEMATICAL MODEL

Consider a system of c servers serving customers whose arrival epochs constitute a renewal process. We assume that the interarrival times are independent and identically distributed according to the distribution F having mean $1/\lambda$, and that the service times are independent and identically distributed according to a negative-exponential distribution with unit mean. If all servers are occupied when a customer arrives, he leaves and has no further effect on the system. If an idle server is available when a customer arrives, service begins immediately.

Let $(0, t]$ denote a time interval of length t which commences at a stationary point for the arrival process.* (Such a point is often said to be chosen at random on the time axis.) Let $A(t)$ be the number of arrivals and $O(t)$ the number of blocked attempts in $(0, t]$. Finally, let

$$L(t) = \int_0^t C(u)du \, , \tag{3}$$

where $C(u)$ is the number of busy servers at time u, be the total usage in $(0, t]$. Note that $L(t)/t$ is a continuous-scan estimate of the carried load.

As was pointed out in Section II, the individual first two moments of $A(t)$, $O(t)$, and $L_d(t)$ and the corresponding three covariance functions are sufficient to obtain an estimate of the variance of any function of these measurements. By numerical experimentation, we found that the covariance functions $\text{Cov}[A(t)/t, \ L_d(t)/36]$ and $\text{Cov}[O(t)/t, \ L_d(t)/36]$ are, for our purposes, well approximated by $\text{Cov}[A(t)/t, \ L(t)/t]$ and $\text{Cov}[O(t)/t, \ L(t)/t]$, respectively. However, the variance of $L(t)$ can be significantly smaller than the variance of $L_d(t)$ so that, in general, we must use $\text{Var}[L_d(t)]$ in our applications.

In the next two sections, we derive the individual and joint moments of $A(t)$, $O(t)$, and $L(t)$. In Section 3.3, we obtain the variance of $L_d(t)$.

3.1 A Multidimensional Renewal Process

Assume that the system described above is in statistical equilibrium†, let t_n, $n = 0, 1, 2, \cdots$, be the instant of time at which the nth overflow

---

* That is, the time until the first arrival after $t = 0$ has the remaining life-time distribution $H(t) = \lambda \int_0^t [1 - F(x)]dx$.

† That is, the system has been in operation sufficiently long prior to $t = 0$ so that system-state probability distribution at $t = 0$ is the limiting (or stationary) distribution $P\{C(0) = k\} = p_k = \lim_{u \to \infty} P\{C(u) = k\}$. It follows that for any $t \geq 0$, $P\{C(t) = k\} = p_k$, i.e., the process $\{C(t), t \geq 0\}$ is stationary.

occurs, $t_0 < 0 < t_1 < t_2 < \cdots$, and define $X_n = t_n - t_{n-1}$. Now let K_n, $n = 1, 2, \cdots$, be the number of arrivals in $(t_{n-1}, t_n]$ and

$$I_n = \int_{t_{n-1}}^{t_n} C(u)du, \quad n = 1, 2, \cdots,$$

be the total usage in $(t_{n-1}, t_n]$.

Since holding times are exponential and arrival epochs constitute a renewal process, the sequence of times t_n, $n = 0, 1, 2, \cdots$, are regeneration or renewal points in our model. Hence X_n, K_n, and I_n, $n = 1, 2, \cdots$, are sequences of independent and identically distributed random variables.

If we now define the row vector

$$\mathsf{X}_n = (1, K_n, I_n), \quad n = 1, 2, \cdots,$$

then it follows that $\{\mathsf{X}_n, X_n\}$, $n = 1, 2, \cdots$, is a multidimensional renewal process.[7] Moreover, setting

$$\eta(t) = \Sigma \mathsf{X}_n,$$

where the sum is taken over all n such that $0 < t_n \leqq t$, we see that for large t,

$$\eta(t) \approx (O(t), A(t), L(t)).$$

Since this formulation corresponds to the concept of a generalized renewal process as communicated by J. M. Hammersley in the discussion of W. L. Smith's paper,[7] his results apply directly to our model. In particular, we shall use his equations (25) and (26) to compute the moments of $\eta(t)$.

Let $\mu_n(c) = E(X_1^n)$ be the nth moment of the interoverflow times from a group of c servers and

$$\nu_n = \int_0^\infty \xi^n dF(\xi) \cdot$$

For brevity, we denote the arrival intensity ν_1^{-1} by λ. From Ref. 4, we have the first moments of $A(t)$ and $O(t)$ already computed:

$$E[A(t)] = \lambda t,$$
$$E[O(t)] = \frac{t}{\mu_1(c)} \cdot \tag{4}$$

From eq. (25) of Ref. 7, we have

$$E[L(t)] = \frac{t\omega_1(c)}{\mu_1(c)}, \tag{5}$$

where

$$\omega_n(c) = E[I_1^n].$$

Again, from Ref. 4 and Ref. 7, the variances and covariances of the three measurements for large t, omitting terms which behave as $o(t)$, are given by

$$\text{Var}\,[A(t)] \sim \frac{t}{\nu_1^3}[\nu_2 - \nu_1^2],$$

$$\text{Var}\,[O(t)] \sim \frac{t}{\mu_1^3(c)}[\mu_2(c) - \mu_1^2(c)],$$

$$\text{Var}\,[L(t)] \sim \frac{t}{\mu_1^3(c)}\{\mu_1^2(c)\omega_2(c) + \mu_2(c)\omega_1^2(c)$$
$$- 2\mu_1(c)\omega_1(c)E[X_1 I_1]\}, \quad (6)$$

$$\text{Cov}\,[A(t), O(t)] \sim \frac{t}{\mu_1^2(c)}\{\lambda\mu_2(c) - E[K_1 X_1]\},$$

$$\text{Cov}\,[O(t), L(t)] \sim \frac{t}{\mu_1^3(c)}\{\mu_2(c)\omega_1(c) - \mu_1(c)E[X_1 I_1]\},$$

$$\text{Cov}[A(t), L(t)] \sim \frac{t}{\mu_1^2(c)}\{\mu_1(c)E[K_1 I_1] + \lambda\mu_2(c)\omega_1(c)$$
$$- \lambda\mu_1(c)E[X_1 I_1] - \omega_1(c)E[X_1 K_1]\}.$$

We now need to compute the various moments and joint moments of X_1, K_1, and I_1 appearing on the right-hand side of (4), (5), and (6) in order to evaluate the approximate expressions for the moments and joint moments of $A(t)$, $O(t)$, and $L(t)$. Note that the mean and variance of $A(t)$ are known since $\lambda = \nu_1^{-1}$ and ν_2 are computed directly from F.

3.2 The Joint Distribution of K_1, X_1, and I_1

The development here parallels that of Section 2.2 in Ref. 4. Let $h_c(w, r, n)$ be the joint density function defined by

$$h_c(w, r, n) = \frac{\partial^2}{\partial w \partial r} P\{X_1 \le w, I_1 \le r, K_1 = n\}.$$

By considering the two mutually exclusive events {the cth trunk remains busy throughout $(0, w)$} and {its complement}, and using a renewal-type argument, we arrive at the following integral equation:

$$h_c(w, r, n) = e^{-w}h_{c-1}(w, r - w, n) + \sum_{k=1}^{n-1}\int_0^w\int_0^u\int_0^{r-v} e^{-v}h_{c-1}(u, s, k)$$
$$\times h_c(w - u, r - s - v, n - k)ds\,dv\,du, \quad (7)$$

in which the time variables u and v run concurrently from an overflow epoch.

The following boundary conditions hold:

$$h_c(w, r, n) = 0, \quad \text{for} \quad r > cw \quad \text{or} \quad r < 0,$$
$$h_c(w, cw, n) = e^{-cw}f(w)\delta_{1,n},$$

where $f = F'$ and $\delta_{1,n} = 1$ for $n = 1$ and is zero otherwise.[*]

If we define

$$\gamma_c(x, y, z) = \sum_{n=1}^{\infty} \int_0^{\infty} \int_0^{\infty} e^{-xw-vr}h_c(w, r, n)z^n dr dw, \qquad (8)$$

then it follows from (7) that

$$\gamma_c(x, y, z) = \frac{(y + 1)\gamma_{c-1}(x + y + 1, y, z)}{y + 1 - \gamma_{c-1}(x, y, z) + \gamma_{c-1}(x + y + 1, y, z)}. \qquad (9)$$

Motivated by the work of Riordan[8] and the success of the approach taken in Ref. 4, we set

$$\gamma_c(x, y, z) = \frac{(y + 1)D_c(x, y, z)}{D_{c+1}(x, y, z)}, \qquad (10)$$

where $D_0(x, y, z) = 1$, and, as can be seen by setting $c = 0$ in (8),

$$D_1(x, y, z) = \frac{y + 1}{z\phi(x)}, \qquad (11)$$

where

$$\phi(x) = \int_0^{\infty} e^{-xt}dF(t).$$

Furthermore, for $m \geqq 1$,

$$D_{m+1}(x, y, z)$$
$$= D_m(x, y, z) + \left[\frac{y + 1}{z\phi(x)} - 1 \right] D_m(x + y + 1, y, z). \qquad (12)$$

If we now define

$$\lambda_j = \lambda_j(x, y, z) = 1 - \frac{y + 1}{z\phi(x + jy + j)}, \qquad (13)$$

then using (11) and (12) and mathematical induction one can show that

$$D_m(x, y, z) = 1 + \sum_{j=1}^{m} (-1)^j \binom{m}{j} \lambda_0 \lambda_1 \cdots \lambda_{j-1}. \qquad (14)$$

---

[*] In our model we assume that the interarrival-time probability distribution function is differentiable. However, with more formalism, the same results can be obtained for the more general case; e.g., the one-point distribution function for constant interarrival times.

Now, since

$$E[X_1^i I_1^j K_1^k] = (-1)^{i+j} \frac{\partial^{i+j+k}}{\partial x^i \partial y^j \partial z^k} \gamma_c(x, y, z) \Big|_{\substack{x=y=0 \\ z=1}}$$

for $k = 0, 1$ and $i, j \geqq 0$, we can compute the required moments directly by means of differentiation. Omitting all of the details of the operations indicated, we obtain

$$\begin{aligned} \mu_1(c) &= \nu_1 D_c \\ \omega_1(c) &= D_c - 1. \end{aligned} \tag{15}$$

where

$$D_c = D_c(1, 0, 1) = 1 + \sum_{j=1}^{c} (-1)^j \binom{c}{j} \Lambda_0 \Lambda_1 \cdots \Lambda_{j-1},$$

with

$$\Lambda_k = 1 - \frac{1}{\phi(k+1)}.$$

Note that D_c is the reciprocal of the generalized Erlang-B blocking probability B_c. Moreover, with the aid of the results obtained in Ref. 4, we have

$$\omega_2(c) = 2[\omega_1(c) + 1] \sum_{j=1}^{c} [\omega_1(j) + 1] - 2[D_c^{(100)} + D_c^{(010)}],$$

$$\mu_2(c) = \frac{\nu_2}{\nu_1} \mu_1(c) + 2\mu_1(c) \sum_{k=1}^{c} \mu_1(k) - 2\nu_1 D_c^{(100)}, \tag{16}$$

where

$$D_c^{(ijk)} = \frac{\partial^{i+j+k}}{\partial x^i \partial y^i \partial z^k} D_c(x, y, z) \Big|_{\substack{y=0 \\ x=z=1}}. \tag{17}$$

The required derivatives are given by

$$D_c^{(100)} = \sum_{j=1}^{c} (-1)^j \binom{c}{j} \Lambda_0 \Lambda_1 \cdots \Lambda_{j-1} \left[\frac{\Omega_0'}{\Lambda_0} + \frac{\Omega_1'}{\Lambda_1} + \cdots + \frac{\Omega_{j-1}'}{\Lambda_{j-1}} \right],$$

$$D_c^{(010)} = \sum_{j=1}^{c} (-1)^j \binom{c}{j} \Lambda_0 \Lambda_1 \cdots \Lambda_{j-1} \left[\frac{\Omega_0''}{\Lambda_0} + \frac{\Omega_1''}{\Lambda_1} + \cdots + \frac{\Omega_{j-1}''}{\Lambda_{j-1}} \right],$$

where Ω_k' is the derivative of $\lambda_k(x, 0, 1)$ evaluated at $x = 1$, i.e.,

$$\Omega_k' = \frac{\phi'(k+1)}{\phi^2(k+1)},$$

and Ω_k'' is the derivative of $\lambda_k(1, y, 1)$ evaluated at $y = 0$, i.e.,

$$\Omega_k'' = \frac{k\phi'(k+1)}{\phi^2(k+1)} - \frac{1}{\phi(k+1)}.$$

Similarly, we have the joint moments

$$E[X_1I_1] = 2\nu_1[\omega_1(c) + 1] \sum_{j=1}^{c} [\omega_1(j) + 1]$$
$$- (1 + \nu_1)D_c^{(100)} - \nu_1 D_c^{(010)},$$

$$E[K_1X_1] = \mu_1(c) + \frac{2}{\nu_1}\mu_1(c) + \sum_{k=1}^{c} \mu_1(k) + \nu_1 D_c^{(001)} - D_c^{(100)}, \qquad (18)$$

$$E[K_1I_1] = 2[\omega_1(c) + 1] \sum_{j=1}^{c} [\omega_1(c) + 1]$$
$$- D_c^{(100)} - D_c^{(010)} + D_c^{(001)},$$

where

$$D_c^{(010)} = \sum_{j=1}^{c} (-1)^j \binom{c}{j} \Lambda_0\Lambda_1\cdots\Lambda_{j-1} \left[\frac{1}{\Lambda_0} + \frac{1}{\Lambda_1} + \cdots + \frac{1}{\Lambda_{j-1}} - j\right].$$

3.3 *Variance of Discrete-Scan Estimate of Usage*

Our present mathematical model assumes that the measurement of usage, $L(t)$, is made by means of continuous scanning, as can be seen from the definition of $L(t)$ in eq. (3). In practice, however, usage is estimated by discrete scanning. The number of busy trunks is sampled at constant intervals of time, say τ, and the integral in (3) is replaced by the finite sum

$$L_d(t) = \sum_{k=1}^{n(t)} C(k\tau), \qquad (19)$$

where

$$n(t) = \max (k : k\tau \leqq t).$$

This procedure introduces a sampling error in the evaluation of the integral. As we shall see later, the difference between the variances of $L(t)$ and $L_d(t)$ can sometimes be large enough that the discrete-scan variance of usage must be used to estimate accurately the variance of g in eq. (2). In this section we indicate how the variance of $L_d(t)$ is computed.

Let $n = n(t)$ be the number of discrete samples in $(0, t]$. In trunking applications, t is usually taken to be one hour (about 20 mean holding times) and, since τ is normally set at 100 seconds, $n = 36$. Since the process $\{C(t), t \geqq 0\}$ is stationary, from (19) we have

$$\text{Var}\,[L_d(t)] = nR(0) + 2 \sum_{j=1}^{n} (n - j)R(j\tau), \qquad (20)$$

where R is the covariance function defined by

$$R(t) = E[C(0)C(t)] - E[C(0)]E[C(t)]. \qquad (21)$$

Since $P\{C(0) = k\} = P\{C(t) = k\} = p_k,$

$$R(t) = \sum_{k=0}^{c} kp_k \sum_{j=0}^{c} jP_{kj}(t) - m_1^2, \tag{22}$$

where

$$P_{kj}(t) = P\{C(t) = j \mid C(0) = k\}, \tag{23}$$

and

$$m_1 = \sum_{k=1}^{\infty} kp_k. \tag{24}$$

The problem of determining the transition function in (23) has been treated by Takács in Chapter 4 of Ref. 9. However, he uses a renewal point for $t = 0$, i.e., his origin is chosen at a point immediately after an arrival has occurred. His result, though not directly applicable, can be modified in a straightforward manner to take account of our different location of the origin. We state here the analogue of his Theorem 3 for the case of a stationary origin and give a proof in the appendix. We use $1/\mu$ to denote the mean service time throughout the statement and proof of the theorem.

Theorem: Let $t = 0$ be a stationary point for the arrival process described above. Then the Laplace transform of (23) is given by

$$\pi_{kj}(s) = \int_0^{\infty} e^{-st} P_{kj}(t) dt = \sum_{i=j}^{c} (-1)^{i-j} \binom{i}{j} \beta_{ki}(s), \tag{25}$$

where

$$\beta_{ki}(s) = \frac{1 - \phi(s + i\mu)}{\phi(s + i\mu)} \frac{\psi_{ki}(s)}{(s + i\mu)} + \binom{k}{i} \frac{1}{(s + i\mu)} \left[1 - \frac{\bar{\phi}(s + i\mu)}{\phi(s + i\mu)} \right],$$

$$\psi_{ki}(s) = \left[C_i(s) \Big/ \sum_{j=0}^{c} \binom{c}{j} \frac{1}{C_j(s)} \right]$$

$$\times \left\{ \left[\sum_{j=i}^{c} \binom{c}{j} \frac{1}{C_j(s)} \right] \left[\sum_{j=0}^{i} \binom{k}{j} \frac{1}{C_{j-1}(s)} \frac{\bar{\phi}(s + j\mu)}{\phi(s + j\mu)} \right] \right.$$

$$\left. - \left[\sum_{j=0}^{i-1} \binom{c}{j} \frac{1}{C_j(s)} \right] \left[\sum_{j=i+1}^{c} \binom{k}{j} \frac{1}{C_{j-1}(s)} \frac{\bar{\phi}(s + j\mu)}{\phi(s + j\mu)} \right] \right\},$$

$$C_j(s) = \prod_{i=0}^{j} \frac{\phi(s + i\mu)}{1 - \phi(s + i\mu)}, \qquad j = 0, 1, 2, \cdots,$$

$$C_{-1}(s) = 1,$$

$$\phi(s) = \int_0^{\infty} e^{-st} dF(t),$$

and $\bar{\phi}(s)$ is the Laplace-Stieltjes transform of the distribution (24), that is,

$$\bar{\phi}(s) = \frac{\lambda}{s}[1 - \phi(s)].$$

Taking the Laplace transform of (22), substituting for $\pi_{kj}(s)$, and simplifying, we obtain the following expression for $\rho(s)$, the Laplace transform of the covariance function $R(t)$:

$$\rho(s) = \frac{m_2}{s+1}\left\{1 - \frac{\lambda}{s+1}\left[\frac{1-\phi(s+1)}{\phi(s+1)}\right]\right\} - \frac{m_1^2}{s}$$

$$+ \left[1\Big/(s+1)\sum_{j=0}^{c}\binom{c}{j}\frac{1}{C_j(s)}\right]\left\{\frac{\lambda m_1}{s}\sum_{j=1}^{c}\binom{c}{j}\frac{1}{C_j(s)}\right.$$

$$+ \frac{\lambda m_2}{(s+1)}\frac{C_0(s)}{C_1(s)}\sum_{j=1}^{c}\binom{c}{j}\frac{1}{C_j(s)}$$

$$- \lambda^2 \sum_{j=2}^{c}\frac{1}{(s+j)C_j(s)}\left[\frac{1-\phi(j)}{\phi(j)}B_j\right.$$

$$\left.\left. + \frac{1-\phi(j+1)}{\phi(j+1)}B_{j+1}\right]\right\}, \quad (26)$$

where

$$B_j = C_j\sum_{i=j}^{c}\binom{c}{i}\frac{1}{C_i}\Big/\sum_{i=0}^{c}\binom{c}{i}\frac{1}{C_i},$$

$$C_j = \prod_{i=1}^{j}\frac{\phi(i\mu)}{1-\phi(i\mu)}, \qquad j = 1, 2, \cdots,$$

$$C_0 = 1,$$

and m_1 and m_2, the first and second moments of the distribution $\{p_k\}$, are given by[9]

$$m_1 = \sum_{k=1}^{c}kp_k = \frac{\lambda}{\mu}(1 - B_c),$$

$$m_2 = \sum_{k=1}^{c}k^2p_k = m_1 + \lambda[B_1 - cB_c].$$

Note that m_1 is the carried load and B_c is the generalized Erlang-B blocking

$$B_c = 1\Big/\sum_{j=0}^{c}\binom{c}{j}\frac{1}{C_j}.$$

Equation (26) has been inverted analytically for the case of Poisson input.[2] When $\phi(s)$ does not have the simple expression of this special

case, analytical inversion appears to be complicated. However, for the purpose of computing the variance of $L_d(t)$ for our trunking application, it is unnecessary to obtain an explicit inverse of $\rho(s)$. We found that the numerical inversion scheme described by Eisenberg[10] is computationally efficient and gives satisfactory results.

To illustrate the difference between continuous-scan and discrete-scan measurements for the case when the input traffic is of the overflow type, we computed the estimates of the variances of $L(t)/t$, the continuous-scan estimate of the carried load, and $L_d(t)/36$, the discrete-scan estimate, for various trunk-group sizes and $t = 20$ mean holding times–i.e., about one hour. For these results, the interarrival-time distribution of the arriving traffic was obtained by using the Interrupted Poisson process with a three-moment match.[11]

The case for ten trunks is typical and is presented in Fig. 1 where $\sigma_{L_d} = \sqrt{\text{Var}\,[L_d(t)/36]}$ and $\sigma_L = \sqrt{\text{Var}\,[L(t)/t]}$ vs α are graphed for $z = 1, 2,$ and 4.

Since the variance of $L_d(t)/36$ must be at least as large as the variance of $L(t)/t$, our results show that the asymptotic estimate of Var $[L(t)/t]$ has a small positive bias, especially for low loads. Our simulation results verify this observation and also indicate that the asymptotic approximation becomes more accurate as the input load increases. Notice that the variance of $L_d(t)/36$ is about equal to the variance of $L(t)/t$ at low loads. As the load increases, the relative error introduced by discrete scanning can increase substantially. Finally, we found that for fixed load and peakedness, the relative difference between Var $[L_d(t)/36]$ and Var $[L(t)/t]$ decreases as the trunk-group size increases (an effect not shown in the figure).

IV. TWO APPLICATIONS

We give two applications of our results, in which we obtain the accuracy of the estimates of two traffic parameters. In the first example, the parameter is the offered load as given in Section I. For the second example, we discuss an estimate of the peakedness of the offered traffic.

4.1 *Accuracy of an Estimate of Offered Load*

Suppose we have observations $A(t)$, $O(t)$, and $L_d(t)$ recorded. Then for the measurement period $(0, t]$, $\hat{\alpha}$, an estimate of the offered load (in erlangs), is given by

$$\hat{\alpha} = g[A(t), O(t), L_d(t)] = \frac{L_d(t)/36}{1 - \dfrac{O(t)}{A(t)}}. \tag{27}$$
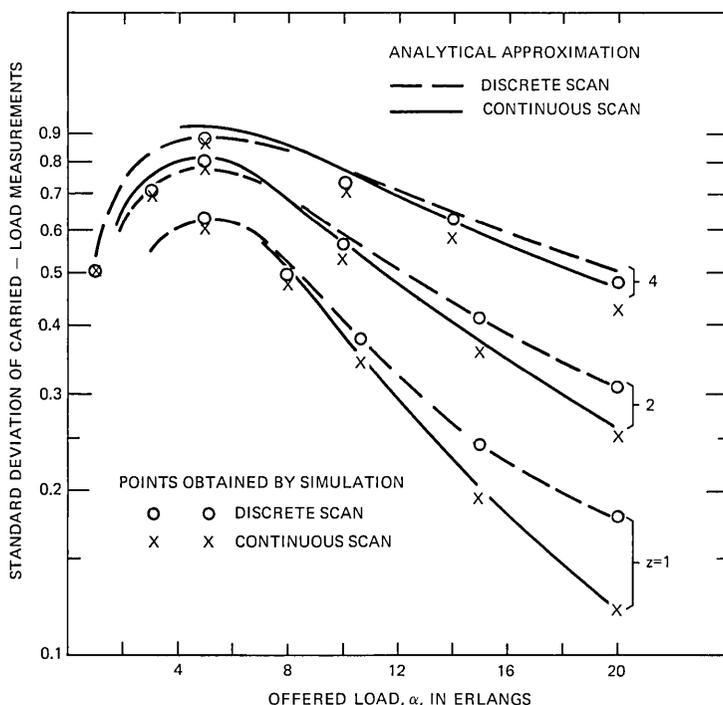
Fig. 1—Standard deviation of carried load measurements vs offered load using discrete-scan and continuous-scan measurements on a 10-trunk group for $t = 20$ mean holding times.

Obtaining the required derivatives of g as indicated in eq. (2), substituting into (2), and simplifying, we have the following expression for the variance of the offered-load estimate in (27):

$$\text{Var}\,[\hat{\alpha}] = \frac{1}{t^2(1 - B_c)^2} \left\{ \text{Var}\,[O(t)] + \text{Var}\,[L_d(t)] \left(\frac{t}{36}\right)^2 \right.$$

$$+ B_c^2 \,\text{Var}\,[A(t)] + 2\,\text{Cov}\,[L(t), O(t)]$$

$$\left. - 2B_c\,\text{Cov}\,[L(t), A(t)] - 2B_c\,\text{Cov}\,[O(t), A(t)] \right\}. \quad (28)$$

Now using eqs. (6), (15), (16), and (18) to substitute for the various quantities on the right-hand side of (28), we can compute $\text{Var}\,[\hat{\alpha}]$.

To test the approximation (28), we computed the variance of $\hat{\alpha}$, as outlined above, for trunk-group sizes of $c = 10$ and $c = 40$ trunks, for input traffic streams of the overflow type having different combinations of load and peakedness values. We also used a computer simulation to
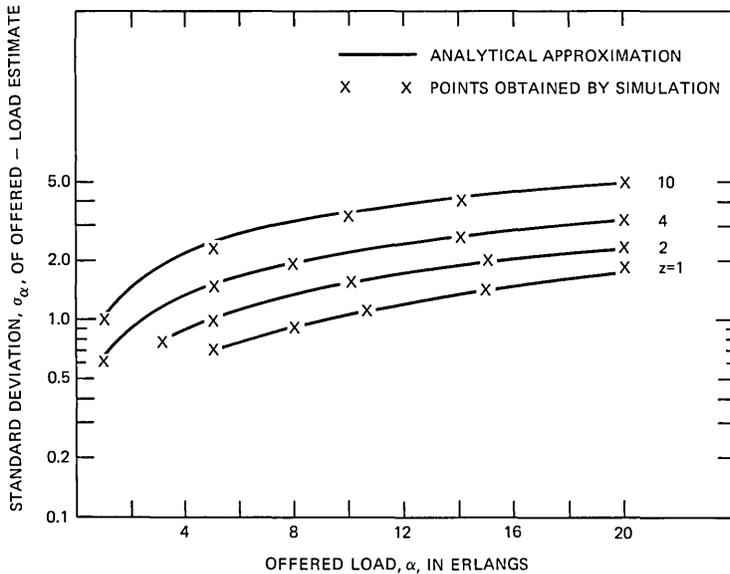
Fig. 2—Standard deviation of offered-load estimate vs offered load for $c = 10$ trunks. The measurement interval is 20 mean holding times.

estimate Var $[\hat{\alpha}]$ at several points. The numerical results are displayed in Fig. 2 for $c = 10$ trunks and Fig. 3 for $c = 40$ trunks where $\sigma_\alpha = \sqrt{\text{Var}[\hat{\alpha}]}$ vs α is displayed for $z = 1, 2, 4,$ and 10 (again for $t = 20$ mean holding times).
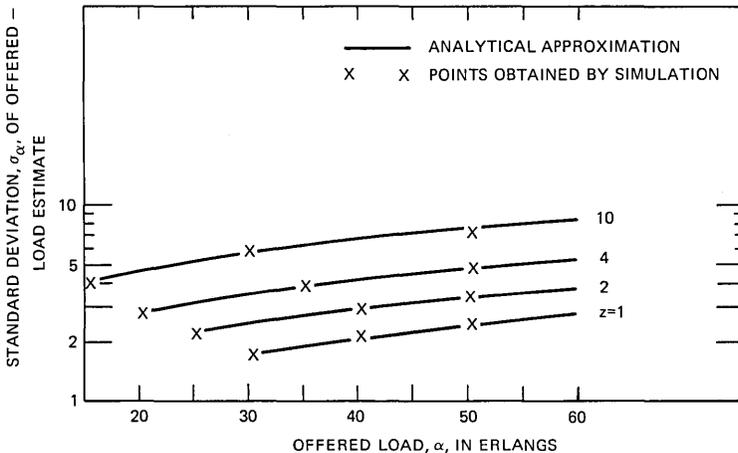

Fig. 3—Standard deviation of offered-load estimate vs offered load for $c = 40$ trunks. The measurement period is 20 mean holding times.

The simulation results indicate that for $c = 10$ and $c = 40$ the asymptotic approximation for Var $[\hat{\alpha}]$ is quite accurate for all ranges of load and peakedness of interest in trunking applications.

We obtained almost identical results for Var $[\hat{\alpha}]$ regardless of whether we used Var $[L_d(t)/36]$ or Var $[L(t)/t]$. Apparently, for low loads (low blocking probability) the accuracy of $\hat{\alpha}$ is dominated by the accuracy of the usage measurements while at high loads (blocking near 1) the accuracy of the call-congestion estimate is the dominant factor. Since the relative difference between Var $[L_d(t)/36]$ and Var $[L(t)/t]$ is small for low blocking probabilities, we see that the accuracy of $\hat{\alpha}$ is not significantly affected by the TUR sampling error.

4.2 Accuracy of an Estimate of Traffic Peakedness

When all three of the measurements $A(t)$, $O(t)$, and $L_d(t)$ are available for a final trunk group of c trunks in an alternate-route network, the peakedness z of the input traffic is estimated in the following manner: First an estimate of offered load $\hat{\alpha}$ is determined as described in the preceding section. Then an estimate $\hat{z}$ of z is obtained by iterative methods (using the Equivalent Random method[1]), such that a single overflow stream having load $\hat{\alpha}$ and peakedness $\hat{z}$ would experience the call congestion $O(t)/A(t)$ or, equivalently, the resulting carried load would be $L_d(t)/36$.

Thus, there is a well-defined procedure for determining a unique value for $\hat{z}$ corresponding to $A(t) \geqq O(t) > 0$ and $L_d(t) > 0$, i.e., we have the estimate

$$\hat{z} = g \left[\frac{A(t)}{t}, \frac{O(t)}{t}, \frac{L_d(t)}{36} \right]$$

in the required form. However, there is no explicit analytical expression for g which can be used to obtain the derivatives needed in (2) to obtain the variance of $\hat{z}$.

In such cases, it is natural to estimate the partial derivatives by first differences. For example,

$$\left. \frac{\partial g}{\partial x_1} \right|_{\theta_1, \theta_2, \theta_3} \approx \frac{g[\theta_1(1 + \Delta), \theta_2, \theta_3] - g[\theta_1, \theta_2, \theta_3]}{\theta_1 \Delta}, \tag{29}$$

where Δ is a small positive number. Numerical experimentation indicated that $\Delta = 0.001$ gives sufficient accuracy for the present application. Using the first-difference approximations as illustrated in (29) for the derivatives in (2) we have an estimate for the variance of $\hat{z}$.

We computed the resulting approximation for $c = 10$ and $c = 40$ trunks for a range of offered loads α, several values of peakedness z,
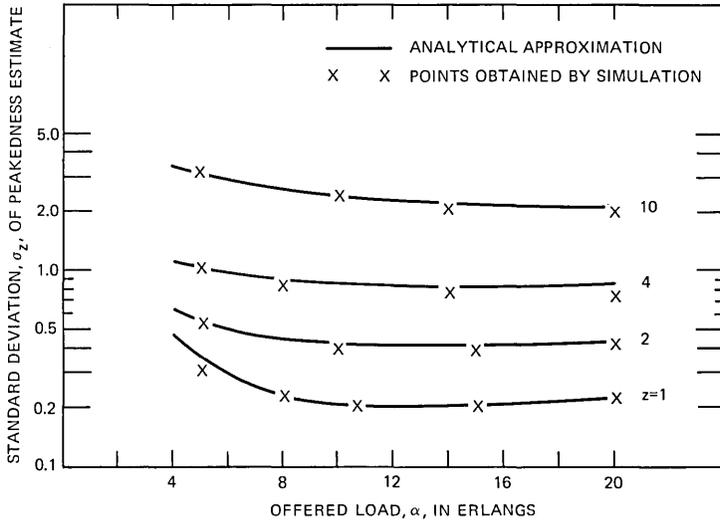
FIG. 4—Standard deviation of peakedness estimate vs offered load for $c = 10$ trunks. The measurement interval is 20 mean holding times.

and $t = 20$ mean holding times. We also compared our approximation with results obtained by simulation. The results are displayed in Figs. 4 (for $c = 10$) and 5 (for $c = 40$) where $\sigma_z = \sqrt{\mathrm{Var}\,[\hat{z}]}$ vs α is given for $z = 1, 2, 4,$ and 10. Note that $\sigma_z/z \approx 0.2$ for large α, independent of c.
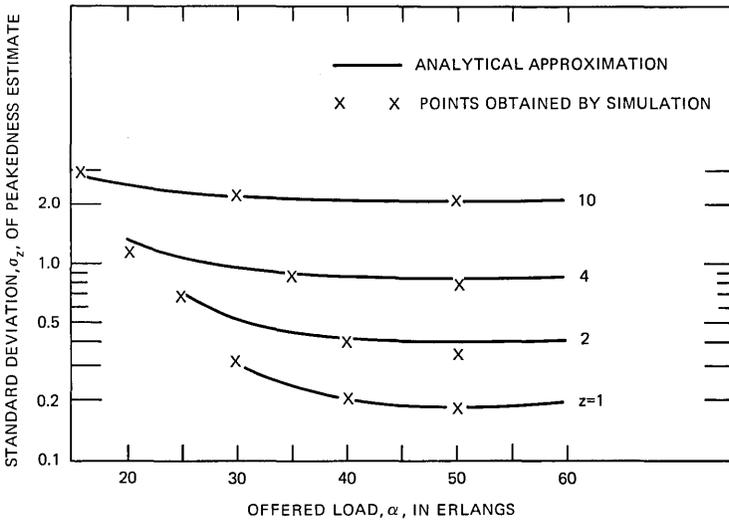

Fig. 5—Standard deviation of peakedness estimate vs offered load for $c = 40$ trunks. The measurement period is 20 mean holding times.
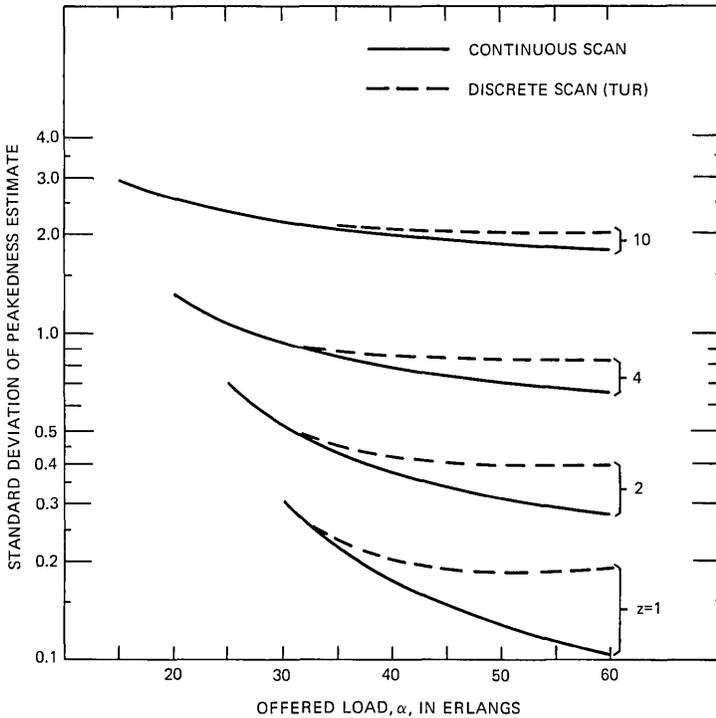
Fig. 6—Comparison of the standard deviation of peakedness estimates using discrete-scan and continuous-scan measurements of carried load on a 40-trunk group.

In general, the simulation results are in good agreement with the approximation. The curves are plotted either for $\alpha \geqq z$, or else for call-congestion exceeding 0.01, the range of interest for trunking applications. When α is smaller than z or the call-congestion is much smaller than 0.01 (not shown in the figures), the value obtained from the approximation for Var $[\hat{z}]$ occasionally tends to be larger than that obtained by simulation. Hence, the approximation may not be adequate for such applications.

In the preceding section, we noted that essentially the same results were obtained for Var $[\hat{\alpha}]$ regardless of whether we used Var $[L(t)/t]$ or Var $[L_d(t)/36]$ in the computations. In contrast, the variance of $\hat{z}$ is very sensitive to the variance of the usage measurements. That is, Var $[L_d(t)/36]$ is required to obtain an accurate approximation for Var $[\hat{z}]$. The errors which can result from using Var $[L(t)/t]$ instead of Var $[L_d(t)/36]$ are illustrated in Fig. 6 for the case of $c = 40$ trunks.

V. SUMMARY AND OTHER APPLICATIONS

5.1 Summary

We derived an approximation for the variance of any differentiable function of the three basic traffic measurements–namely, peg count, overflow count, and usage (TUR). The approximation is expressed in terms of the first partial derivatives or first differences of the function, and the individual variances and covariances of the measurements. Except for the variance of the TUR measurements, asymptotic approximations for the required moments were obtained by an application of Hammersley's generalized renewal theory.

The variance of the TUR was given as a sum involving the covariance function for the number of busy servers at equally spaced scan intervals. The Laplace transform of the covariance function was derived and inverted numerically using an inversion technique described by Eisenberg.

The results were then applied to obtain approximations for the variances of estimates of offered load and peakedness (variance-to-mean ratio) of a stream of traffic of the overflow type submitted to a loss system. The approximations were in good agreement with results obtained by simulation.

5.2 Other Applications

5.2.1 Offered Load Estimates Based on Usage Measurements

At present, $A(t)$ and $O(t)$ are not always measured on primary high-usage trunk groups. Estimates of the single-hour offered load and call-congestion for such groups are obtained from the TUR measurement $L_d(t)$ with an iterative procedure based on the Erlang-B theory. Using the techniques presented above, one could compare the accuracy of these estimates with that which would be obtained by using all three measurements. It should then be possible to evaluate the difference in statistical accuracy that results from using (or not using) the additional measurements.

5.2.2 The Optimum Number of Single-Hour Measurements

Normally, 20 single-hour measurements are used to obtain estimates $\hat{s}$ of the correct number of trunks s required to obtain a specified grade of service. For example, on final trunk groups, the 20 single-hour estimates of call-congestion $O(t)/A(t)$ and usage $L_d(t)$ are first averaged and then used to estimate an average load and peakedness of the input traffic. These average values are used to obtain $\hat{s}$.

It has been proposed that the number of measurements be reduced in order to lower the data handling costs. However, reducing the number of measurements would increase the variance of $\hat{s}$, i.e., decrease the accuracy of the trunk estimates. It appears that an optimum number of measurements could be determined by minimizing a function of the form

$$\mathfrak{C}(N) = \mathfrak{K} \operatorname{Var}[\hat{s}(N)] + C_N, \qquad (30)$$

where C_N is the cost of taking and processing N measurements, $\hat{s}(N)$ is the estimate of s based on N measurements, and $\mathfrak{K}$ is a cost associated with inaccurate trunk estimates. The precise form of the function might require modification. However, the basic idea is to trade off an increase in the accuracy of the provisioning process due to more accurate trunk estimates (as N increases) against a corresponding increase in cost.

It appears that one can obtain an approximation for $\operatorname{Var}[\hat{s}(N)]$ using an extension of the ideas presented in Sections II and III in order to account for the effects of day-to-day variation in the offered loads. However, a realistic model to justify (30) or to obtain $\mathfrak{K}$ and C_N will require further study.

APPENDIX

We prove here the theorem stated in Section 3.3. We shall need the following lemma.

Lemma: For the model described in the text, let $Y(t)$ be the number of busy servers at time t, $t = 0$ be a stationary point, and $Y(0) = i$. Now let Y_n be the number of busy servers found by the nth arrival and t_n be the time of the nth arrival. For $n = 1, 2, \cdots,$ and $r = 0, 1, \cdots, c$, define

$$A_{ir}^{(n)}(s) = E\left\{ e^{-st_n} \binom{Y_n}{r} \,\middle|\, Y(0) = i \right\}, \qquad (31)$$

and

$$\psi_{ir}(s) = \sum_{n=1}^{\infty} A_{ir}^{(n)}. \qquad (32)$$

Then we have

$$
\psi_{ir}(s) = \left[C_r(s) \,\middle/\, \sum_{j=0}^{c} \binom{c}{j} \frac{1}{C_j(s)} \right]
$$
$$
\cdot \left\{ \left[\sum_{j=r}^{c} \binom{c}{j} \frac{1}{C_j(s)} \right] \left[\sum_{j=0}^{r} \binom{i}{j} \frac{1}{C_{j-1}(s)} \frac{\bar{\phi}(s+j\mu)}{\phi(s+j\mu)} \right] \right.
$$
$$
\left. - \left[\sum_{j=0}^{r-1} \binom{c}{j} \frac{1}{C_j(s)} \right] \left[\sum_{j=r+1}^{c} \binom{i}{j} \frac{1}{C_{j-1}(s)} \frac{\bar{\phi}(s+j\mu)}{\phi(s+j\mu)} \right] \right\}, \qquad (33)
$$

where

$$C_j(s) = \prod_{i=0}^{j} \frac{\phi(s + i\mu)}{1 - \phi(s + i\mu)}, \qquad j = 0, 1, 2, \cdots,$$

$$C_{-1}(s) \equiv 1,$$

$$\phi(s) = \int_0^{\infty} e^{-st} dF(t),$$

and

$$\bar{\phi}(s) = \int_0^{\infty} e^{-st} d\bar{F}(t) = \lambda \int_0^{\infty} e^{-st} [1 - F(t)] dt$$

$$= \frac{\lambda}{s} [1 - \phi(s)].$$

The proof below is essentially the same as the proof of Lemma 1 of Ref. 9, modified to account for the stationary origin. For $n = 1, 2, \cdots$, we have

$$E\left\{ e^{-st_{n+1}} \binom{Y_{n+1}}{r} \,\middle|\, Y_n = j, t_{n+1} - t_n = x \right\}$$

$$= \begin{cases} e^{-(s+r\mu)x} \binom{j+1}{r} E\{e^{-st_n} | Y_n = j\}, & \text{for } j < c, \\ e^{-(s+r\mu)x} \binom{c}{r} E\{e^{-st_n} | Y_n = c\} & \text{for } j = c, \end{cases}$$

because under the given conditions Y_n has a binomial distribution with parameters $j + 1$ (for $j = 0, 1, \cdots, c - 1$) or c (for $j = c$) and $e^{-\mu x}$. If we remove the condition $t_{n+1} - t_n = x$, that is, multiply by $dP\{t_{n+1} - t_n \leq x\}$ and integrate over all x, we obtain

$$E\left\{ e^{-st_{n+1}} \binom{Y_{n+1}}{r} \,\middle|\, Y_n = j \right\} = \phi(s + r\mu) \binom{j+1}{r} E\{e^{-st_n} | Y_n = j\},$$

for $j = 0, 1, \cdots, c - 1$ and

$$E\left\{ e^{-st_{n+1}} \binom{Y_{n+1}}{r} \,\middle|\, Y_n = c \right\} = \phi(s + r\mu) \binom{c}{r} E\{e^{-st_n} | Y_n = c\}.$$

If we multiply the corresponding equations by $P\{Y_n = j\}$ and add them for $j = 0, 1, \cdots, c$, then we get

$$A_{ir}^{(n+1)}(s) = \phi(s + r\mu) \left[A_{ir}^{(n)}(s) + A_{i,r-1}^{(n)}(s) - \binom{c}{r-1} A_{ic}^{(n)}(s) \right], \quad (34)$$

for $r = 1, 2, \cdots, c$ and

$$A_{i0}^{(n)}(s) = \bar{\phi}(s)[\phi(s)]^{n-1}.$$

Since $Y(0) = i$ and $t = 0$ is a stationary point, we have

$$A_{ir}^{(1)}(s) = \binom{i}{r} \bar{\phi}(s + r\mu).$$

Forming the sum (32) we get

$$\psi_{ir}(s) - \binom{i}{r} \bar{\phi}(s + r\mu)$$

$$= \phi(s + r\mu) \left[\psi_{ir}(s) + \psi_{i,r-1}(s) - \binom{c}{r-1} \psi_{ic}(s) \right],$$

or

$$\psi_{ir}(s) = \frac{\phi(s + r\mu)}{[1 - \phi(s + r\mu)]} \left[\binom{i}{r} \frac{\bar{\phi}(s + r\mu)}{\phi(s + r\mu)} \right.$$

$$\left. + \psi_{i,r-1}(s) - \binom{c}{r-1} \psi_{ic}(s) \right]. \quad (35)$$

Dividing both sides of this equation by $C_r(s)$ we get

$$\frac{\psi_{ir}(s)}{C_r(s)} = \frac{\psi_{i,r-1}(s)}{C_{r-1}(s)} + \frac{\binom{i}{r} \frac{\bar{\phi}(s + r\mu)}{\phi(s + r\mu)} - \binom{c}{r-1} \psi_{ic}(s)}{C_{r-1}(s)}.$$

Adding these equations over $r, r - 1, r - 2, \cdots, 1$ we obtain

$$\frac{\psi_{ir}(s)}{C_r(s)} = \sum_{j=0}^{r} \binom{i}{j} \frac{\bar{\phi}(s + j\mu)}{\phi(s + j\mu)} \frac{1}{C_{j-1}(s)} - \psi_{ic}(s) \sum_{j=1}^{r} \binom{c}{j-1} \frac{1}{C_{j-1}(s)},$$

$$i = 1, 2, \cdots, c. \quad (36)$$

Setting $r = c$ in (36) we get

$$\psi_{ic}(s) = \sum_{j=0}^{c} \binom{i}{j} \frac{\bar{\phi}(s + j\mu)}{\phi(s + j\mu)} \cdot \frac{1}{C_{j-1}(s)} \bigg/ \sum_{j=0}^{c} \binom{c}{j} \frac{1}{C_j(s)}. \quad (37)$$

Substituting (37) into (36) we obtain $\psi_{ir}(s)$ for $r = 1, 2, \cdots, c$. If $r = 0$, then

$$\psi_{i0}(s) = \sum_{n=1}^{\infty} \bar{\phi}(s)[\phi(s)]^{n-1}$$

$$= \frac{\bar{\phi}(s)}{1 - \phi(s)}.$$

This completes the proof of the lemma.

We now prove the theorem stated in Section 3.3. Again, the proof is a slight modification of the proof of Theorem 3 in Chapter 4 of Ref. 6.

Let us define the binomial moments

$$B_{ir}(t) = \sum_{k=r}^{c} \binom{k}{r} P_{ik}(t), \qquad i, r = 0, 1, \cdots, c.$$

From the definition it follows that

$$P_{ik}(t) = \sum_{r=k}^{c} (-1)^{r-k} \binom{r}{k} B_{ir}(t), \qquad (38)$$

so that setting

$$\beta_{ir}(s) = \int_{0}^{\infty} e^{-st} B_{ir}(t) dt$$

and forming the Laplace transform of (38), we get eq. (25) of the theorem. It remains to determine $\beta_{ir}(s)$.

Let $Y(t)$ be the number of busy servers at time t and let $t = 0$ be a stationary point. The times between those successive arrivals which find j servers busy are independent and identically distributed random variables. Hence, the sequence of epochs immediately preceding those arrivals which find j servers busy constitutes a renewal process. If $Y(0) = i$, we will denote the renewal function of such an imbedded renewal process by $M_{ij}(t)$.

It may be helpful to recall that $M_{ij}(t)$ is the expected number of those calls which arrive in the time interval $(0, t]$ and find exactly j servers busy, given that initially there are i servers busy. Hence, we can write

$$M_{ij}(t) = \sum_{n=1}^{\infty} P\{t_n \leqq t, Y_n = j \mid Y(0) = i\}, \qquad (39)$$

where t_n and Y_n have the same meaning as in the statement of the lemma.

If no calls arrive in $(0, t]$, then $Y(t)$ has the binomial distribution with parameters i and $e^{-\mu t}$, and $B_{ir}(t)$ is given by

$$\binom{i}{r} e^{-r\mu t} [1 - \bar{F}(t)],$$

where

$$\bar{F}(t) = \lambda \int_{0}^{t} [1 - F(x)] dx.$$

If one or more calls arrive in $(0, t]$, let the last call's arrival epoch be u and at that instant let the number of busy servers be j. Now $Y(t)$ has

the binomial distribution with parameters $j + 1$ (if $j = 0, 1, \cdots,$ $c - 1$) or c (if $j = c$) and $e^{-\mu(t-u)}$. Thus, together we have

$$B_{ir}(t) = \binom{i}{r} e^{-r\mu t}[1 - \bar{F}(t)]$$

$$+ \sum_{j=r-1}^{c-1} \binom{j+1}{r} \int_0^t e^{-r\mu(t-u)}[1 - F(t - u)]dM_{ij}(u)$$

$$+ \binom{c}{r} \int_0^t e^{-r\mu(t-u)}[1 - F(t - u)]dM_{ic}(u). \quad (40)$$

If we introduce the Laplace-Stieltjes transform

$$\mu_{ij}(s) = \int_0^\infty e^{-st}dM_{ij}(t),$$

then from (40) we have

$$\beta_{ir}(s) = \frac{1 - \phi(s + r\mu)}{(s + r\mu)}\left[\binom{i}{r}\frac{1 - \bar{\phi}(s + r\mu)}{1 - \phi(s + r\mu)}\right.$$

$$\left. + \sum_{j=r-1}^{c-1} \binom{j+1}{r} \mu_{ij}(s) + \binom{c}{r} \mu_{ic}(s)\right]. \quad (41)$$

From (39) we have

$$\mu_{ij}(s) = \sum_{n=1}^\infty P\{Y_n = j\}E\{e^{-st_n}| Y_n = j, Y(0) = i\}$$

and hence by (31) and (32) we get

$$\sum_{j=r}^c \binom{j}{r} \mu_{ij}(s) = \sum_{n=1}^\infty E\left\{e^{-st_n} \binom{Y_n}{r}\middle| Y(0) = i\right\}$$

$$= \psi_{ir}(s).$$

Thus, $\beta_{ir}(s)$ can be written in the following form

$$\beta_{ir}(s) = \frac{1 - \phi(s + r\mu)}{s + r\mu}\left[\binom{i}{r}\frac{1 - \bar{\phi}(s + r\mu)}{1 - \phi(s + r\mu)} + \psi_{ir}(s)\right.$$

$$\left. + \psi_{i,r-1}(s) - \binom{c}{r-1}\psi_{ic}(s)\right].$$

If we take relation (35) into consideration, then this formula can be simplified to

$$\beta_{ir}(s) = \frac{1 - \phi(s + r\mu)}{\phi(s + r\mu)}\cdot\frac{\psi_{ir}(s)}{(s + r\mu)} + \binom{i}{r}\frac{1}{(s + r\mu)}\left[1 - \frac{\bar{\phi}(s + r\mu)}{\phi(s + r\mu)}\right].$$

This completes the proof of the theorem.

REFERENCES

1. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
2. Beneš, V. E., "The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement," B.S.T.J., *40*, No. 1 (January 1961), pp. 117–148.
3. Descloux, A., "On the Accuracy of Loss Estimates," B.S.T.J., *44*, No. 6 (July-August 1965), pp. 1139–1164.
4. Kuczura, A., and Neal, S. R., "The Accuracy of Call-Congestion Measurements for Loss Systems with Renewal Input," B.S.T.J., *51*, No. 10 (December 1972), pp. 2197–2208.
5. Descloux, A., "On Markovian Servers with Recurrent Input," Proc. Sixth Int. Teletraffic Cong., Munich, West Germany, 1970, pp. 331/1–331/6.
6. Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*, Vol. 1, New York: Hafner, 1958, chapter 10.
7. Hammersley, J. M., Discussion of paper by W. L. Smith, "Renewal Theory and Its Ramifications," J. Roy. Stat. Soc., Series B, *20*, No. 2, 1958, pp. 289–295.
8. Riordan, J., *Stochastic Service Systems*, New York: John Wiley and Sons, 1962, pp. 36–40.
9. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962, chapter 4.
10. Eisenberg, M., "A Program for the Analysis of a Class of Electronic Switching Systems," unpublished work, April 15, 1971.
11. Kuczura, A., "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *52*, No. 3 (March 1973), pp. 437–448.

Switching Networks of Planar Shifting Arrays

By R. S. KRUPP and L. A. TOMKO

An array of shift registers that may operate in two orthogonal directions can be called a planar shifting array. This article shows how two basic building blocks, fashioned from planar shifting arrays, may be interconnected to form a time-division switching network of arbitrary size. The characteristics of magnetic bubble and charge-coupled devices are compatible with the concept of planar arrays, and it is in these emerging technologies that switching networks of planar shifting arrays may become practical.

I. INTRODUCTION

Switching machines in the Bell System have grown in both number and capacity to meet the growing traffic demand. Early machines consisted of a small amount of distributed logic embodied in electromechanical devices but, as technology has permitted, the machines have evolved into largely solid state systems with central processor control. The environment in which switching machines must operate has also changed from a relatively small collection of analog voice grade circuits to an overwhelming number of circuits of various bandwidths with an increasing proportion of digital facilities. It is the purpose of this paper to look at a possible future realization of one portion of a switching machine that might have advantages in meeting future requirements in a largely digital environment.

We may consider that a switching machine consists of three major subdivisions. One is a switching network, which makes cross connections for each call. A controller, used to direct the operation of the network, is another. Finally, some interface is needed between the network, the controller, and external circuits. The subject of this paper is a switching network, one that may reduce the complexity of the tasks of the other two subdivisions of a switching machine as well as have advantages of its own in conjunction with some emerging technologies.

Most switching networks in the past have been "space-division" networks; that is, a spatially distinct path is assigned to each of the many simultaneous calls that might pass through the network. The present paper, however, refers to a "time-division" network, in which interleaved samples of several simultaneous calls may share parts of the same spatial path. A potential savings of equipment is implied by this time-sharing process, and its merits have already resulted in plans for a large-scale digital electronic switching machine with some time sharing in the switching network (the No. 4 ESS, now under development in Bell Laboratories).[1]

The possibility of constructing a somewhat different time-division digital switching network from two basic building blocks is explored in this paper. The two blocks, or subsystems, may be realized in the form of planar shifting arrays which are basically shift registers that can perform shifting operations in two orthogonal directions. These capabilities seem to be consistent with those of the emerging technologies of magnetic bubble and charge-coupled devices,[2,3] which do shift data on a plane. Although these types of devices may be particularly well suited for use as planar shifting arrays, the concepts presented below are not restricted to implementation by any particular type of device.

The first building block, a time-slot interchanger, is the only actual switching element in the system. The other type of block, a mass serial-to-parallel converter, performs a time-space mapping and thereby acts as the interconnection links between successive stages of time-slot interchangers. Networks of arbitrary size and blocking probability can be fashioned from these two building blocks.

In the sections that follow, the basic interconnections of time-slot interchangers and mass serial-to-parallel converters necessary to perform multistage switching functions are explained, and some logical details of the two building blocks are presented. Although the network is only a part of a total switching machine, the concepts presented here may simplify the interface with digital external circuits and may help reduce the burden on the central control processor.

II. NETWORK ARCHITECTURE

In this section, multistage switching network structures are described that use pure time-division techniques only. Such a network may be diagrammed using two types of functional blocks, as in Figure 1.

Each block labelled TSI denotes one time-slot interchanger—a familiar subsystem in the time-division switching art. The attached notation $N \times M$ indicates that the TSI rearranges words from an
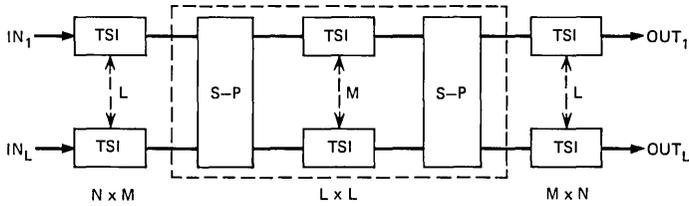
Fig. 1—Three-stage switching network.

N-slot input frame into an M-slot output frame. The TSI is the only actual switching element in the network. It corresponds to an $N \times M$ array of crosspoints in an analogous space-division network.

Each block labelled S–P denotes a mass serial-to-parallel converter. The S–P does not perform switching functions, but corresponds to the links between crosspoint stages in a multistage space-division network. Its basic function is that of interchanging the space and time coordinates of each word it handles. As inputs to one S–P, the figure shows L time-division multiplexed lines, each carrying M time slots per frame. The output is M lines with L slots per frame. An input word in slot m of line ℓ will always be routed by the S–P to slot ℓ of output line m, where $1 \leq m \leq M$ and $1 \leq \ell \leq L$. Thus, the mth words of all L input frames are combined to form a single output frame on line m, and conversely the input frame on line ℓ gets distributed into the ℓth slots of all the M output frames.

A three-stage switching network is diagrammed in Figure 1 using the two functional blocks. To illustrate its operation, suppose that a word in slot n of input line ℓ must be sent to time slot ν on output line λ, where $1 \leq n, \nu \leq N$ and $1 \leq \ell, \lambda \leq L$. First an intermediate time slot m is assigned for some $1 \leq m \leq M$, which permits the middle stage to complete a connection. Thereafter:

(i) The ℓth input TSI switches word n to slot m.
(ii) The first S–P places this word in slot ℓ on line m.
(iii) The mth intermediate TSI switches word ℓ to slot λ.
(iv) The second S–P places the word in slot m on line λ.
(v) The λth output TSI switches word m to slot ν, and the task is finished.

The portion of the figure surrounded by a dashed line performs the same function as the time-shared space-division switch (TSSDS) stages of the No. 4 ESS.[1] That is, all words that enter the first S–P in time slot m will exit the second S–P in the same slot. In their passage through
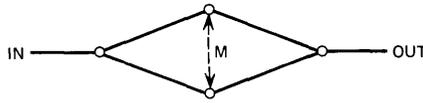
Fig. 2—Spiderweb graph for Figure 1.

the S–P's and the mth intermediate TSI, however, these words may be spatially rearranged to different output lines. A direct realization of the same TSSDS function using magnetic bubbles has been suggested by P. I. Bonyhard.[4]

Figure 2 is the probability linear graph[5] (also called spiderweb or Lee graph) of all paths for one call in Figure 1. It is the standard Lee graph for a three-stage network and yields the usual blocking formulas. In particular, if $M \geq N - 1$ and we assume independent occupancy p of all input and output time slots, then a modification of C. Y. Lee's argument[5] yields the mismatch blocking probability first given by M. Karnaugh[6] in 1954:

$$P_B = (p^2)^{M-N+1}[1 - (1 - p)^2]^{2N-M-2}(N - 1)!^2/$$
$$M!(2N - M - 2)!. \quad (1)$$

On the other hand, if $M \leq N - 1$ and we assume all intermediate time slots m have independent occupancy p, a more appropriate estimate is:

$$P_B = [1 - (1 - p)^2]^M. \quad (2)$$

When $M \geq 2N - 1$ in Figure 1, P_B vanishes and the network is non-blocking.[7] Figure 3 illustrates such a case for $M = 2N$. A part of the time-space interchange is accomplished, not by the S–P, but by splitting each input and output TSI into two separate TSI's. Now each TSI in the network handles input and output frames of equal size. This might prove convenient for certain applications or implementations. Note that Figure 3 may also be interpreted as a pair of three-stage networks whose inputs and outputs are tied in parallel. A third, inactive, network could be placed in parallel as well, and held in reserve against failure of one of the first two.

There is an unavoidable signal delay of at least one frame for each stage of the TSI's. This could contribute to the cost of echo suppression on long toll circuits; however, the duration of the frame itself may be shortened within the switch to mitigate this effect.

The three-stage network of Figure 1 has a capacity of $C = LN$ terminations. Such networks may be nested to achieve a higher switching capacity. A five-stage example having capacity $C = JLN$ is dia-
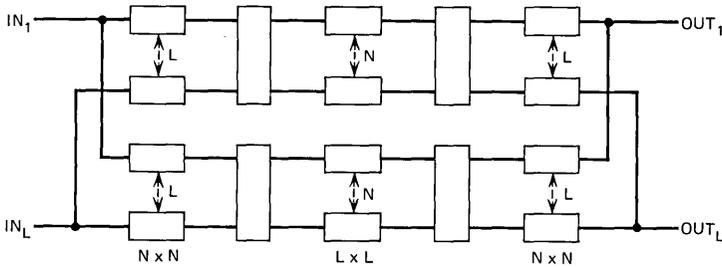
Fig. 3—Nonblocking three-stage network.

grammed in Figure 4. Shown are M three-stage TSSDS sections within dashed lines, as well as one large overall TSSDS which rearranges JL words spatially in each of M time slots. Blocking probabilities for the five-stage nested network are obtained by arguments similar to those for eqs. (1) and (2), using the spiderweb graph in Figure 4.

A different kind of five-stage network organization is shown in Figure 5. No section of this network performs the TSSDS function. In fact, a word entering its first S–P in time slot m may exit its last S–P in any time slot μ for $1 \leqq m, \mu \leqq M$, depending upon the path it follows through the network. The number M^2 of possible pairs (m, μ) yields the same number of possible paths for routing a message, as the spiderweb graph in Figure 5 shows. At full occupancy, this graph reduces to
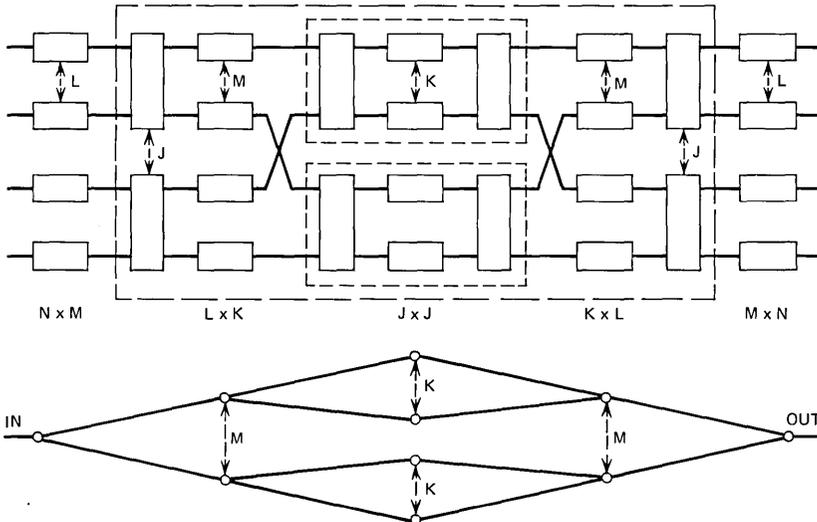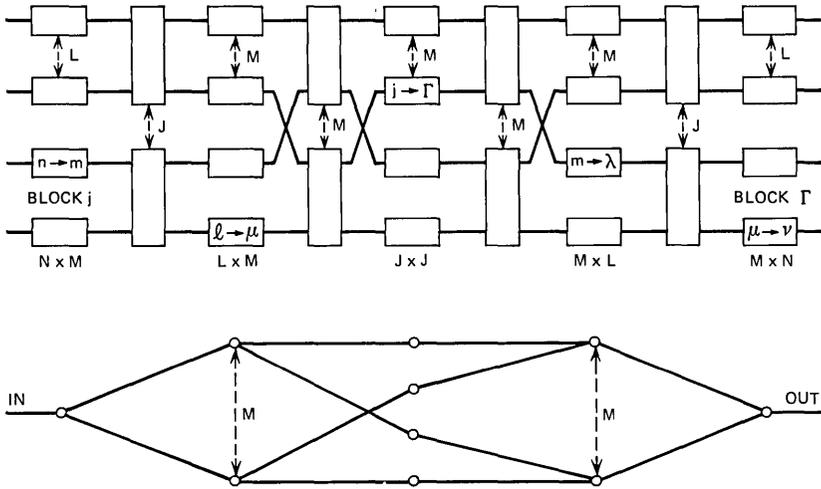

Fig. 4—Nested five-stage network.

Fig. 5—Cross-connected five-stage network.

Figure 2, but with $M' = (M - N + 1)^2$ center-stage switches, so that (2) applies with link occupancies $p = N(L - 1)/M^2$ to yield the following blocking probability expression:

$$P_B = [1 - (1 - p)^2]^{(M-N+1)^2}. \tag{3}$$

Thus, parameters $L = N = 48$ and $M = 60$ yield $P_B < 10^{-10}$ at any occupancy level; essentially, this is a nonblocking network. The same parameters in a nested arrangement produce a much higher degree of blocking, $P_B > 0.027$ at full occupancy, based on (1) for the case $p = 1$.

To illustrate operation of the network in Figure 5, suppose that a word in time slot n on line ℓ of input block j must be transmitted to slot ν on line λ of output block Γ for $1 \leqq n,\nu \leqq N$ and $1 \leqq \ell,\lambda \leqq L$ with $1 \leqq j,\Gamma \leqq J$. First, intermediate time slots m and μ must be chosen for some $1 \leqq m,\mu \leqq M$ which permits completion of the connection. Then TSI's in the five stages perform the following switching operations consecutively to complete the task, as indicated on Figure 5:

   (i) Switch word n into slot m (Stage 1).
  (ii) Switch word ℓ into slot μ (Stage 2).
 (iii) Switch word j into slot Γ (Stage 3).
 (iv) Switch word m into slot λ (Stage 4).
  (v) Switch word μ into slot ν (Stage 5).

The networks of Figures 4 and 5 each have a center section consisting of the third-stage TSI's and the two rows of S–P's to which they

attach. A total of J independent input and output blocks of LN termi-
nations each connect to the center. Each pair of blocks consists of an
input S–P, an output S–P, and their attached TSI's in the first, second,
fourth, and fifth stages. This structure immediately suggests an ap-
propriate strategy for growing the network capacity in modules of LN
terminations.

III. NETWORK PATH SEARCH

Section II mentioned the need to choose one or two intermediate
time slots m before setting up a message path through the network. This
amounts to specifying which of the many possible paths in the spider-
web graph is currently free and will be used. The decision might be
made by a central control processor after consulting its memory records
of the current status of the network. In this section, an alternate pro-
cedure is described by which some simple operations within the network
itself can identify suitable paths for routing a new call. A possible ad-
vantage is reduced demands upon processor time and memory for call
processing.

To provide necessary information about network status, a "busy-
bit" is used. This is a bit in each word which travels through the net-
work with that word and serves to indicate whether or not the word is
part of a call currently in progress. The busy-bit may occur in every
word of a message or less frequently, perhaps every tenth or hundredth
frame. For concreteness, assume the busy-bit is "one" for a call in
progress and "zero" otherwise.

To determine occupancy of a time slot on some line in the network,
we merely consult the busy-bit in that slot. For example, in routing a
word from input line ℓ of Figure 1 to output line λ, an intermediate time
slot m was chosen that was vacant both at the output of the ℓth first-
stage TSI and at the input of the λth third-stage TSI. Such m may be
found by comparing the two corresponding streams of busy-bits. This
is illustrated by Figure 6a for the case $\ell = \lambda = 1$. Output from the first
input TSI and input to the first output TSI are sent to a NOR gate.
When simultaneous zeroes are found in the busy-bit positions of words
m, an output pulse is produced by the gate identifying an available
path through the network by its timing.

Other means of sampling and matching busy-bits are possible. For
instance, the stream of busy-bits from the ℓth input TSI will exit the
first S–P simultaneously in word time ℓ, while the busy-bits entering
the second S–P at word time λ are those for the λth output TSI.
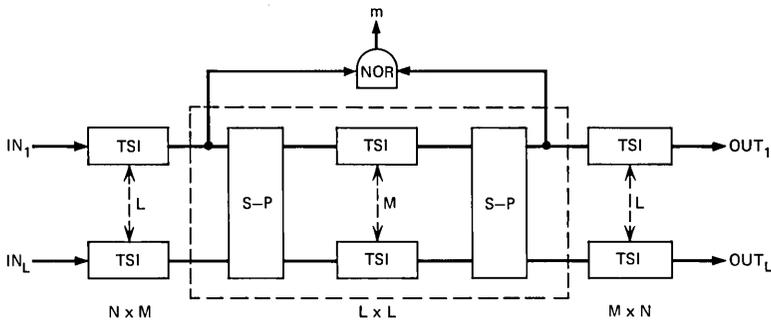Hence, timed sampling pulses to the two S–P's can select the two de-

Fig. 6a—Path search in three-stage network.

sired strings of busy-bits rather than use spatial selection of an input and output TSI. Such an option might serve to reduce the amount of control and signal wiring associated with the network.

Once an intermediate time slot m is chosen, a central processor could issue the necessary orders to the first-, second-, and third-stage TSI's to establish the path. But again some additional logical apparatus in the network could perform the same task. The significance of this observation is that the dependence of the network upon external intervention may be reduced. Indeed, for each new message, the principal external control required would be a specification of which input and output terminations must be connected. Such a structure might lend itself, for example, to applications in which all processor functions are performed by a remote computer that sends its instructions to the network over a data link.

The path search procedure in the nested five-stage network of Figure 4 would consist of two nested three-stage path searches. That is, the busy-bit streams of the input and output TSI's are compared, as above, to determine which one m of the M intermediate three-stage networks will carry the message. Then the mth three-stage network is searched for some intermediate time slot $1 \leq k \leq K$ which will complete the connection.

A different search procedure is required for the cross-connected network of Figure 5, since all M^2 paths in the spiderweb graph must be tested. This involves sorting the M busy-bit streams leaving the second-stage TSI's of the input block j by using them as inputs to an S–P with M output lines. The S–P outputs are then compared by a set of OR gates to the M busy-bit streams entering the fourth-stage TSI's of the output block Γ. A "zero" at word time m from gate μ identifies a possible path in Figure 6b. Note that, if time slot m from the first-stage
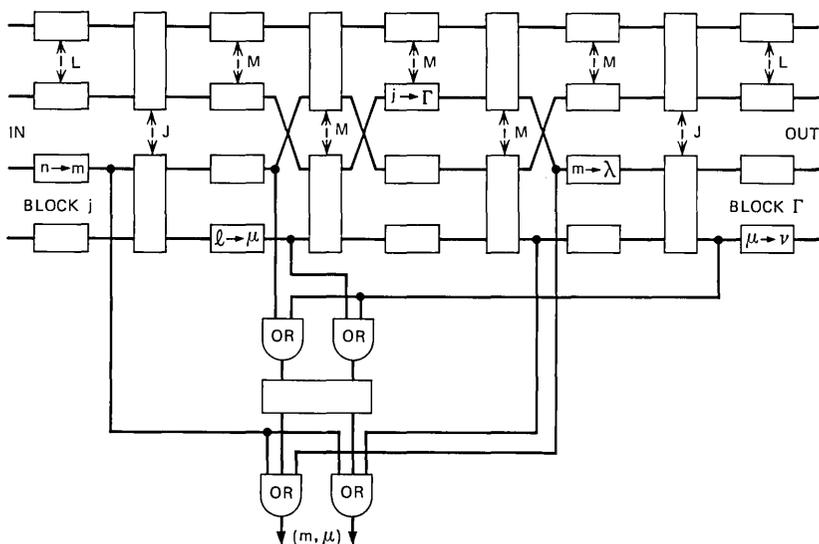
Fig. 6b—Path search in five-stage network.

TSI or time slot μ into the fifth-stage TSI is busy, then path $(m,\ \mu)$ cannot be used. To eliminate such possibilities, all the OR gates should be pulsed during each busy first-stage slot m, and all gates μ corresponding to busy fifth-stage slots should be pulsed for the entire frame, as is accomplished in Figure 6b by feeding parallel "ones" to the S–P. This last path search requires only one pass instead of two nested steps.

IV. PLANAR SHIFTING ARRAYS

To understand the role of planar array devices in the realization of the functions described above, it is helpful to study the building blocks at the logical level, with a notation suggestive of the two-dimensional nature of the devices. The device technologies mentioned in Section I differ considerably in their physical principles. Nevertheless, certain common features of their operation may be abstracted to aid in discussing their use for switching applications. To this end, a "planar shifting array" (PSA) notation will be introduced, exemplified by Figure 7. Circles O represent fixed word-storage locations in one-dimensional or two-dimensional shift registers. Diamonds ◇ stand for shifting apparatus, which can move the contents of each circle to the next in line, under control of clock pulses A and B. Gates $\boxed{n}$ allow transfer of individual words between the adjacent circles under control of switch-
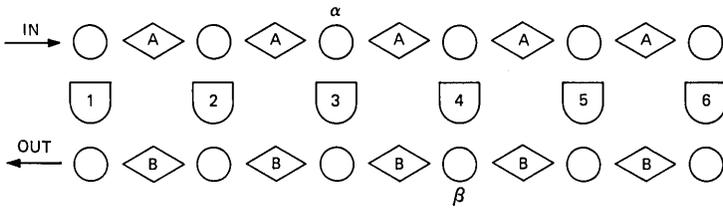
Fig. 7—Switch elements for time-slot interchanger.

ing signals n. The distinction between $\diamondsuit$ and $\bigsqcup$ is one of function rather than physical structure. In this section, a word, rather than a bit, is the basic quantum of information. It is not necessary to specify whether this word is in a serial or parallel digital representation or some other form.

A basic scheme for performing switching functions in a time-slot interchanger appears in Figure 7. Gates 1 to 6 allow transfer of input words from a frame stored in the upper shift register into output slots in a frame stored in the lower shift register. The input and output frames move, relative to one another, in the two adjacent shift registers, and the switching strategy is: For each input word α, wait until its destined output slot β lies directly underneath and then transfer through the appropriate gate.

It is important to note a particular difficulty. If both the input and output frames are shifted simultaneously, then word α in the diagram will pass slot β halfway between gates 3 and 4, so that it is not possible to transfer. This we will call the "half-word" problem. At least three kinds of solution can be suggested:

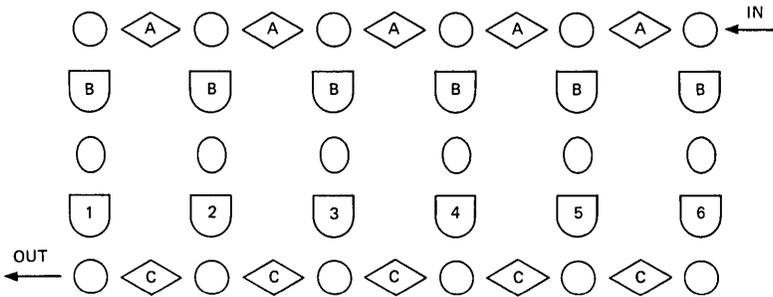(i) Alternate clock pulses A and B, so that only one frame shifts at a time.

Fig. 8—Switch elements for time-slot interchanger with buffer.

(ii) Omit one set of clock pulses, say A, so that one frame is held in stationary buffers.

(iii) Let each circle contain only a half-word, so that two consecutive locations hold an entire word. Then it is never necessary to transfer between first and second half-word locations.

A fourth solution, based on the technique of bubble expansion, has been found by P. I. Bonyhard.[4]

The first solution above may be illustrated by Figure 7. The basic switching cycle would consist of the following four steps:

(i) Diamonds ◇A◇ move the input words right one slot.

(ii) Gates ⊡n may transfer input words to output slots.

(iii) Diamonds ◇B◇ move the output words left one slot.

(iv) Gates ⊡n may transfer input words to output slots.

For an M-word frame, switching is completed in M cycles. Each shift register operates M times, and the M gates each have as many as $2M$ chances to operate in every frame.

The second solution to the half-word problem is illustrated by Figure 8. The diamonds ◇A◇ read in an entire input frame, which is stored in buffers ◯ by simultaneous operation of all gates ⊡B at the end of the input frame. Then the following two-step switching cycle:

(i) diamonds ◇C◇ move the output words left one slot,

(ii) gates ⊡n may transfer words from buffers to output,

is repeated $2M$ times, and the M gates each have M chances to operate in an M-word frame. This scheme has the additional advantage that the input and output clocks need not be synchronous. Each solution to the half-word problem has advantages with respect to particular hardware.

Figure 9 uses the PSA notation to illustrate a simple scheme for implementing the S–P function. The S–P shown has $L = 4$ input lines, each carrying six-word frames, and $M = 6$ output lines, each with four time slots per frame. The diamonds ◇A◇ operate to load an entire frame into the S–P from each input line. Then the diamonds ◇B◇ operate to unload an entire frame from the S–P onto each output line. Loading and unloading may proceed at twice the external basic word rate so that the S–P empties in time to receive the next frame, or else an alternate S–P may handle the next frame while the first one is unloading. Other schemes permit the S–P to load and unload simultaneously. While the S–P is drawn as a distinct unit in preceding figures,
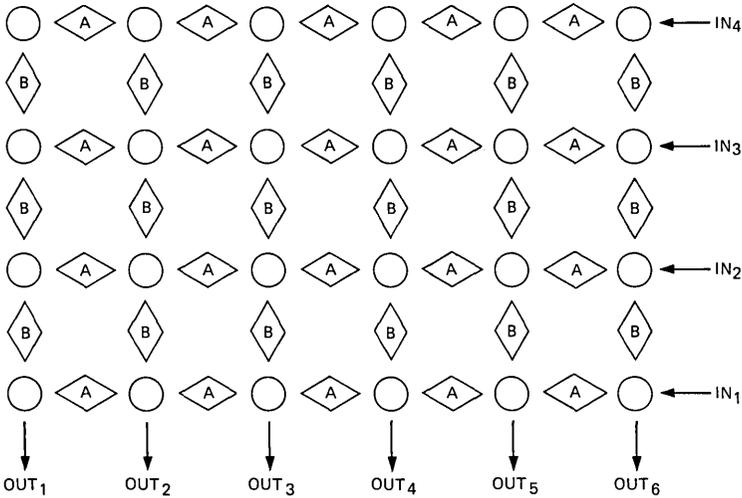
Fig. 9—Mass serial-to-parallel converter.

only its function need be distinct. The S–P as a device may be broken up and its parts integrated onto the same chips as the various TSI's which it serves. Schemes to obviate the S–P completely using specific device properties have been proposed by W. F. Chow and P. I. Bonyhard.[4]

V. MEMORY STRUCTURE

In the switching schemes described above, each gate $\boxed{n}$ should operate at the same word times in each consecutive frame; it is natural to place it under the control of a recirculating local memory. This could consist of a shift register fabricated from the same materials as the switching elements and preferably integrated onto the same chips. Such an arrangement is diagrammed in Figure 10, using the PSA notation. The switching elements appear at the right. The rest of the diagram is the local memory array to control these elements. Each shift register in the control memory moves its contents one word to the right under the action of clock pulse C whenever a pulse A or B occurs. A word in shift register n is read when it reaches the box $\boxed{n}$, which is considered to generate a pulse n that operates gates $\widehat{n}$. For example, the gating operations might be realized through repulsion between control bubbles and message bubbles. Gates $\boxed{n}$ at the left start new words down the shift registers to accomplish recirculation of the local memory.

The size of the local memory array depends on the number of words Q in each register and the number R of such registers, which is the same as the number of switching gates. Although the exact size depends on the particular solution chosen for the half-word problem presented above, we can conclude in each case that the local memory in a TSI grows quadratically in frame size (as QR), while the number of switching elements grows only linearly (as R). For moderate-size frames (say, 20 words or more), each TSI becomes a large recirculating memory plane with a small quantity of logic elements at the edges. A word in the control memory corresponds most closely to a crosspoint in a space-division switch, since the number of crosspoints also grows as the square of the number of circuits, and each crosspoint stores just one bit of information.

VI. CONTROL INSERTION AND ERASURE

The switching operations of a TSI are fixed from frame to frame on a short time scale (say, thousands of frames); thus, the contents of its local memory array are fixed. On some longer time scale, though, it is necessary to be able to change portions of local memory; for instance, in response to external signals emanating from a central control processor.

Specifically, one requires the capability to address single word positions in the local memory in order to close or open a "crosspoint" by writing or erasing a bit. In the case of erasure, it would be sufficient to simultaneously erase all memory locations affecting a given input word, or alternatively a given output slot. This is due to the following two properties of the TSI:

(A) Each input word is switched into one output slot at most.
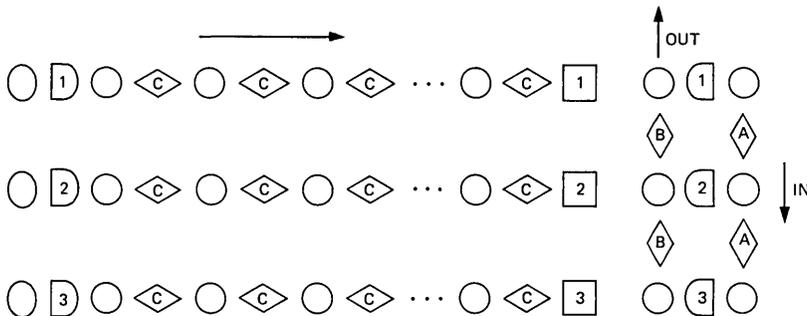(B) Each output slot receives at most one input word.

Fig. 10—Recirculating memory array for TSI gates.

Hence, one can provide for erasure of a control word without knowing exactly where it is in the local memory by erasing a whole class of words.

The fact that local memory recirculates will greatly simplify the problem of addressing specific locations, since this allows access to any location from the edge of the array. Indeed, one efficient and convenient scheme would be to build an analog of the switching elements and operate it "backwards" to insert control bits into the local memory. Insertion could be directed by two external control pulses whose timing would specify an input word and the output slot for which it is intended. Such an approach is illustrated in Figure 11, in which switching elements are deleted to concentrate attention on the memory plane. Control insertion elements appear at the right-hand side. A bit is inserted in order to control gate n at word time m as follows:

(i) At the start of a frame, ⬚D gates a "one" into the vertical shift register, where it propagates downward.

(ii) After n of the A clock pulses, gates I operate to place the "one" in buffer ◯ at the right of the nth recirculating shift register in local memory.

(iii) In the next frame, gates J operate with the mth clock pulse C, to place the "one" in word m of memory line n. This is indicated formally by reading the "one" at ⬚n , but the actual details of bit injection will depend on the type of hardware.

It is clear now that any word location in the local memory may be addressed employing a pair of time-coded pulses I and J. The scheme shown would be particularly appropriate to the case of buffered input, since n then becomes the number of the input word to be switched.
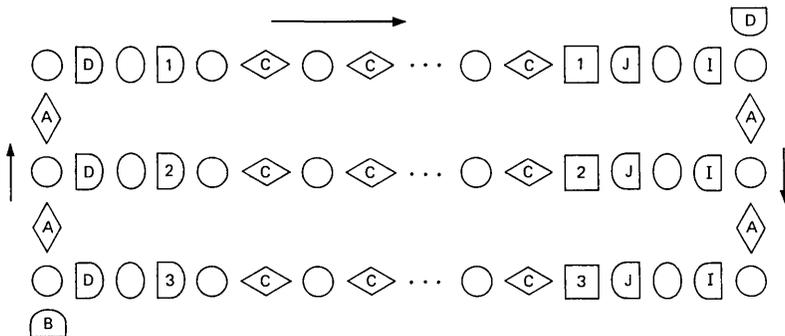

Fig. 11—Addressing and erasure schemes for the TSI local memory.

Indeed, the nth shift register contains only those locations that can affect input word n. By property (A) above, at most one location on this memory line has nonzero contents. It is sufficient now to erase the entire shift register, by interrupting recirculation, for example, before a new control word is inserted.

The left side of Figure 11 shows a simple scheme for erasure in the case of buffered input, which also guarantees that at most one nonzero word is resident in each memory line. Once each frame, the buffers $\bigcirc$ at the left are loaded with "ones." Gates $\boxed{n}$ then start these "ones" down the various shift registers at the appropriate times to accomplish recirculation. Clearly, at most one nonzero word can circulate. To erase memory line n, the "one" is simply deleted in its recirculation buffer for a single frame.

A single time-coded pulse will suffice to specify erasure of any word in the local memory above, if the buffers are loaded serially. Just such an arrangement appears at the left side. Gate $\boxed{B}$ enters "ones" in the vertical shift register at each input word time, and these are loaded into the buffers by gates $\boxed{D}$ at the start of each frame. By deleting the B pulse at the nth word time, a "zero" is sent to the buffer for the nth memory line, erasing it and opening any "crosspoint" which might affect the nth input word. The deletion of pulse B could coincide with pulse I, which routes a new control word to memory line n. Then the pair of external signals I and J would accomplish erasure together with address insertion. Examples at device level of this procedure have been given by P. I. Bonyhard and W. F. Chow.[4]

VII. BUSY–BITS FOR PATH TAKE–DOWN

The preceding section considered two schemes for erasure in the local memory:

(i) A direct instruction from the central control processor.
(ii) Automatic erasure when a new control word is inserted.

A third scheme is to provide for automatic erasure when the message that is being switched terminates. This might be accomplished through use of the busy-bits introduced previously. Recirculation of a word in local memory would be made contingent on the presence of a "one" in the busy-bit position of the particular input word which that control word switches. When the message terminates, the path along which it was routed through the switch is taken down simply by sending through one word with a "zero" busy-bit. This function is analogous to that of

the sleeve-lead in electromechanical switches and might be considered an "electronic sleeve-lead" application of busy-bits.

This strategy of making recirculation in the local memory contingent on busy-bits can be easily implemented in Figure 11. It is merely necessary to read the busy-bit stream for the input frame into the left-hand shift register and load it into the recirculation buffers once each frame. Now each busy-bit will prime the recirculation of the memory register associated with the word of that busy-bit.

VIII. CONCLUSION

The possibility of constructing a time-division switching network using two building blocks of planar arrays was discussed in this paper. The compatibility of planar arrays with the emerging technologies of magnetic bubble, charge-coupled, and bucket-brigade devices[2,3] may lead to application of these concepts in the construction of relatively inexpensive large-capacity switching machines.

The need for external connections to a network composed of these building blocks can be minimized by including path search and path maintenance functions with the blocks. As a result, a relatively small amount of information must be exchanged between the network and the supervisory processor, and some of the processing burden on the controller is shared by the network itself.

Some examples of compatible network architecture have been given, although no specific design is proposed. The process of switching is accomplished by time-slot interchangers within the network rather than a space-division crosspoint array, but the task performed by some of the described subnetworks, when viewed from their external ports, is the same as a time-shared space-division switch.

ACKNOWLEDGMENTS

REFERENCES

1. Vaughan, H. E., "An Introduction to No. 4 ESS," paper presented at the International Switching Symposium, Massachusetts Institute of Technology, Cambridge, Mass., June 6, 1972.

2. Bobeck, A. H., Fischer, R. F., Perneski, A. J., Remeika, J. P., and Van Uitert, L. G., "Application of Orthoferrites to Domain Wall Devices," IEEE Trans. Magnetics, *Mag-5*, No. 3 (September 1969), pp. 544–553.
3. 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-4*, No. 3 (June 1969).
4. Chow, W. F., and Bonyhard, P. I., unpublished work (1971) and private conversations.
5. Lee, C. Y., "Analysis of Switching Networks," B.S.T.J., *34*, No. 6 (November 1955), pp. 1287–1315.
6. Karnaugh, M., unpublished work, August 1954.
7. Clos, Charles, "A Study of Non-Blocking Switching Networks," B.S.T.J., *32*, No. 2 (March 1953), pp. 406–424.

Charge Distribution in Buried-Channel Charge-Coupled Devices

By W. H. KENT

This paper studies charge distribution in buried-channel charge-coupled devices. Detailed development of a one-dimensional electrostatic model is presented and a numerical solution of the resulting nonlinear potential equations is described. Graphical results show the charge-filling mechanism and the relationship between the oxide-semiconductor interface potential and total free positive charge.

I. INTRODUCTION

This paper describes a numerical determination of the distribution of charge in a one-dimensional model of a buried-channel charge-coupled device (CCD).[1] Several calculations have recently been made of the static potential in CCD's in the *absence* of stored charge.[2-4] In addition, the motion of the stored charge under dynamic conditions has been studied by means of essentially one-dimensional models which do not involve a true knowledge of the distribution of stored charge or the charge-carrying capacity of the CCD.[2,5] However, so far it has not been possible to calculate even the static stored-charge distribution in a two-dimensional model of a buried-channel CCD, much less to follow the motion of this charge under dynamic conditions. In this paper a start is made on the problem by calculating the static distribution of stored charge in a one-dimensional model of a buried-channel CCD. The resulting information on the charge distribution is of interest in itself. However, an additional important objective has been to find numerical techniques which can be extended to the two-dimensional problem.

The paper is divided into three parts. Section I treats the physics of the model and Section III gives numerical results. Section II deals briefly with the numerical techniques used and it may be omitted by the reader without loss of continuity.
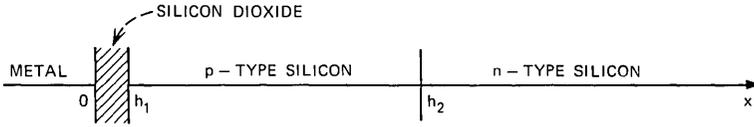
Fig. 1—Buried-channel device structure.

The buried-channel device has a layered structure which will be modeled in one dimension as follows (Fig. 1):

$0 \leqq x \leqq h_1$:   silicon dioxide (SiO$_2$) with relative dielectric constant $\epsilon_1/\epsilon_o = 4$. ($\epsilon_o$ is capacitivity of free space.)

$h_1 \leqq x \leqq h_2$:   p-type silicon with acceptor number density $N_A(x)$ and relative dielectric constant $\epsilon_2/\epsilon_o = 12$.

$h_2 \leqq x \leqq \infty$:   n-type silicon uniformly doped with constant donor number density N_D and dielectric constant $\epsilon_3 = \epsilon_2$.

The point $x = 0$ is a perfectly conducting boundary held at a potential V_o. The potentials $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$, in the SiO$_2$, p-type, and n-type regions, respectively, satisfy the one-dimensional Poisson equation,

$$\frac{d^2\phi_i(x)}{dx^2} = -\rho_i(x)/\epsilon_i, \qquad i = 1, 2, 3, \tag{1}$$

where the $\rho_i(x)$, the lineal charge densities, are nonlinear functions of the potentials $\phi_i(x)$.

The functional forms of the $\rho_i(x)$ are determined by the assumptions:

   (i) The SiO$_2$ is a perfect insulator.

   (ii) The generation and recombination rates for holes and electrons are zero.

   (iii) Hole and electron currents are zero at the time of observation.

   (iv) The flat-band voltage is zero.

   (v) The injected free holes and electrons are separately in equilibrium.

Conditions (i) through (v) define a static device for which the $\rho_i(x)$ are:

$$\left.\begin{array}{l} \rho_1(x) = 0, \\ \rho_2(x) = q[p(x) - N_A(x)], \\ \rho_3(x) = q[N_D + p(x) - n(x)], \end{array}\right\} \tag{2}$$

where $n(x)$ and $p(x)$ are the number densities of free electrons and holes, respectively. The most general expression for $n(x)$ is

$$n(x) = \int_{E_c(x)}^{\infty} g(E)F_e(E)dE, \tag{3}$$

where $g(E)$ is the density of states, $F_e(E)$ is the Fermi distribution for electrons, and $E_c(x)$ is the conduction band edge. Substitution for $g(E)$ and $F_e(E)$ yields

$$n(x) = N_c^o \int_{E_c(x)}^{\infty} \frac{(E - E_c(x))^{\frac{1}{2}} dE}{1 + \exp ((E - E_f)/kT)}. \tag{4}$$

This specific choice for the density of states corresponds to the simplest possible band structure. E_f is the equilibrium Fermi level for electrons and N_c^o is the constant

$$N_c^o = 4\pi(2m_c/h^2)^{\frac{3}{2}}, \tag{5}$$

where m_c is the effective mass of an electron in silicon and h is Planck's constant.[6]

A similar expression for the distribution of holes is

$$p(x) = \int_{-\infty}^{E_v(x)} g(E) F_h(E) dE, \tag{6}$$

where $F_h(E) = 1 - F_e(E)$; substitution for $g(E)$ and $F_h(e)$ yields

$$p(x) = N_v^o \int_{-\infty}^{E_v(x)} \frac{(E_v(x) - E)^{\frac{1}{2}} dE}{1 + \exp ((E_{fh} - E)/kT)}. \tag{7}$$

$E_v(x)$ is the valence band edge and E_{fh} is defined as the pseudo-Fermi level[7] for the holes. N_v^o is the constant given by

$$N_v^o = 4\pi(2m_v/h^2)^{\frac{3}{2}}, \tag{8}$$

where m_v is the effective hole mass in silicon.[8] $E_c(x)$ and $E_v(x)$ are functions of $\phi_i(x)$ given by:

$$E_c(x) = E_c^o - q\phi_i(x), \tag{9}$$

$$E_v(x) = E_v^o - q\phi_i(x). \tag{10}$$

E_v^o and E_c^o are the valence and conduction band edges at $x = \infty$, and q is the magnitude of an electronic charge. $p(x)$ and $n(x)$ are functions of the yet undetermined constants E_f and E_{fh}. Later they will appear only in the forms

$$\eta = (E_c^o - E_f)/kT,$$
$$\eta' = (E_c^o - E_{fh})/kT.$$

Since only difference terms appear, no energy reference need be established for the model. η will be determined from the electron charge neutrality condition at $x = \infty$. η' is fixed by the total amount of free positive charge $Q_+ = \int_{h_1}^{\infty} p(x) dx$ in the device. In reality, determina-

tion of η' is difficult, so the alternative scheme of choosing η' and calculating the resulting Q_+ is employed. Substituting eqs. (2), (4), and (7) into eq. (1) results in the system of differential equations which characterize the device described above. In the appendix these equations are reduced in a straightforward manner by using the Boltzman approximation; justification for its use is given. The resulting equations which will be solved are (' denotes d/dy):

$$\psi_1''(y) = 0, \qquad\qquad\qquad\qquad 0 \leqq y \leqq h_1/\lambda, \quad (11)$$

$$\psi_2''(y) = \sigma - (m_v/m_c)^{\frac{3}{2}} e^{\rho_o - \psi_2(y)}, \qquad h_1/\lambda \leqq y \leqq h_2/\lambda, \quad (12)$$

$$\psi_3''(y) = e^{\psi_3(y)} - 1 - (m_v/m_c)^{\frac{3}{2}} e^{\rho_o - \psi_3(y)}, \qquad h_2/\lambda \leqq y \leqq \infty, \quad (13)$$

where

$$\psi = q\phi/kT,$$
$$y = x/\lambda,$$
$$\lambda = \sqrt{kT\epsilon_2/q^2 N_D},$$
$$N_v = N_v^o(kT)^{\frac{3}{2}},$$
$$N_c = N_c^o(kT)^{\frac{3}{2}},$$

ϵ_g = the bandgap energy in kT's,

$\eta = (E_c^o - E_f)/kT$ is a constant dependent on doping levels,

$\eta' = (E_c^o - E_{fh})/kT$ is a parameter depending on the pseudo-Fermi level,

$$\rho_o = \eta + \eta' - \epsilon_g,$$
$$\sigma = N_A/N_D.$$

m_v and m_c are the effective masses of holes and electrons in silicon; they are $1.08m_o$ and $0.59\ m_o$, respectively.[9] Equations (11), (12), and (13) satisfy boundary conditions

$$\psi_1(0) = V_o \qquad\qquad\qquad (14)$$

and, consistent with eqs. (9) and (10),

$$\psi_3(\infty) = 0. \qquad\qquad\qquad (15)$$

At $y = h_1/\lambda$ and $y = h_2/\lambda$ the continuity conditions are

$$\psi_1(h_1/\lambda) = \psi_2(h_1/\lambda)$$
$$\epsilon_1\psi_1'(h_1/\lambda) = \epsilon_2\psi_2'(h_1/\lambda) \qquad\qquad (16)$$

$$\psi_2(h_2/\lambda) = \psi_3(h_2/\lambda)$$
$$\psi_2'(h_2/\lambda) = \psi_3'(h_2/\lambda). \qquad\qquad (17)$$

II. NUMERICAL SOLUTION

The system of differential equations (11), (12), and (13) together with boundary equations (14), (15), (16), and (17) are solved numerically by the method of finite differences.

For computational convenience the length of the device will be truncated from the whole half-line to the segment $[0, L]$; the distance L is chosen such that $|\psi_3(\infty) - \psi_3(L)|$ and $|\psi_3'(\infty) - \psi_3'(L)|$ are suitably small. $[0, L]$ is partitioned by a mesh of N points and the solution at each point y_i is denoted by ψ^i. The mesh lengths (distance between successive points) are δ_o, δ_p, and δ_n in the oxide, p-region, and n-region, respectively, and are chosen such that $y_{N_1} = h_1/\lambda$ and $y_{N_2} = h_2/\lambda$. The boundaries are $\psi^o = \psi_1(0) = V_o$ and $\psi^N = \psi_3(L) \approx \psi_3(\infty) = 0$. The subscripts on $\psi(y)$ can now be dropped since the superscripts identify the solution point. The second derivative is approximated by the second difference

$$\frac{d^2\psi(y)}{dy^2} = (\psi^{i+1} - 2\psi^i + \psi^{i-1})/\delta^2$$

in each region; δ is one of δ_o, δ_n, or δ_p as appropriate. Using this approximation to discretize eqs. (11), (12), and (13) results in the matrix equation

$$[A][\psi^i] = [\rho(y_i, \psi^i)] \tag{18}$$

which has rows generated by eq. (19); $\rho(y_i, \psi^i)$ is defined in the three regions by the right-hand side of eq. (19).

$$\psi^{i+1} - 2\psi^i + \psi^{i-1} = \begin{cases} 0, & i < N_1, \\ (\sigma - (m_v/m_c)^{\frac{3}{2}}e^{\rho_o - \psi^i})\delta_p, & N_1 < i < N_2, \\ (e^{\psi^i} - 1 - (m_v/m_c)^{\frac{3}{2}}e^{\rho_o - \psi^i})\delta_n, & i > N_2. \end{cases} \tag{19}$$

Rows corresponding to the solutions at the boundary points y_{N_1} and y_{N_2} are obtained from eqs. (16) and (17) where the first derivative is approximated by

$$\mp\psi'(y) \approx (+11\psi(y) - 18\psi(y \pm \delta) \\ + 9\psi(y \pm 2\delta) - 2\psi(y \pm 3\delta))/6\delta, \tag{20}$$

and δ is the appropriate mesh size for each region (the positive or forward derivative uses points to the left of y). Equation (20) results from simultaneous solution of the Taylor expansions for $\psi(y \pm \delta)$, $\psi(y \pm 2\delta)$, and $\psi(y \pm 3\delta)$ with terms $0(\delta^4)$ and higher dropped; the result is accurate to $0(\delta^3)$ which is consistent with the second difference approximation to $d^2\psi/dy^2$. In the oxide layer, a first-difference approximation

$$\psi'(x) = (\psi^i - \psi^{i-1})/\delta_o \tag{21}$$

is adequate since the solution in this region is linear. The boundary equations are:

for $i = N_1$,

$$-(6\delta_p\epsilon_1/\delta_o\epsilon_2)\psi^{N_1-1} + (11 + 6\delta_p\epsilon_1/\delta_o\epsilon_2)\psi^{N_1}$$
$$- 18\psi^{N_1+1} + 9\psi^{N_1+2} - 2\psi^{N_1+3} = 0; \quad (22)$$

for $i = N_2$,

$$(-1/\delta_p)(2\psi^{N_2-3} - 9\psi^{N_2-2} + 18\psi^{N_2-1}) + 11\psi^{N_2}(1/\delta_n + 1/\delta_p)$$
$$- (1/\delta_n)(18\psi^{N_2+1} - 9\psi^{N_2+2} + 2\psi^{N_2+3}) = 0. \quad (23)$$

The matrix A in eq. (18) is tridiagonal except for the N_1th and N_2th rows.

For each choice of the parameter ρ_o, the $\{\psi^i\}$ of eq. (18) are solved for iteratively using a combination of successive under- and over-relaxation[10] (SOR) on the equation

$$[\psi^i] = [A]^{-1}[\rho(y_i, \psi^i)]. \quad (24)$$

The procedure described here differs slightly from the usual SOR in that the transformed equation (24) is solved instead of eq. (18). For the ith row of eq. (24), the $j + 1$th estimate of ψ^i is given as

$$\psi^i_{j+1} = \psi^i_j + \omega(\check{\psi}^i - \psi^i_j), \quad (25)$$

where ω is a relaxation parameter with values $0 < \omega \leq 2$. $\check{\psi}^i$ is a Newton's method solution of eq. (26), a transcendental equation in ψ^i resulting from the ith row of eq. (24) with the remaining $N - 2$ variables ψ^k held constant. The coefficients a_{ik}^{-1} in eq. (26) are elements of A^{-1}.

$$\psi^i - \sum_{k=1}^{i-1} a_{ik}^{-1}\rho(y_k, \psi^k_{j+1}) - a_{ii}^{-1}\rho(y_i, \psi^i) - \sum_{k=i+1}^{N-1} a_{ik}^{-1}\rho(y_k, \psi^k_j) = 0. \quad (26)$$

Criteria for convergence of the process, as well as the choice of the value of ω, is based on the residual r_j defined at the jth iteration as

$$r_j = \sum_{i=1}^{N-1} |\psi^i_j - \psi^i_{j-1}|. \quad (27)$$

It can be shown that if (18) were a linear system of equations and if ψ^i were the true solution at y_i then

$$\sup_{1 \leq i \leq N} |\psi^i - \psi^i_{j+1}| \Big/ \sup_{1 \leq i \leq N} |\psi^i| \leq C(\omega)r_{j+1}, \quad (28)$$

where $C(\omega)$ is a constant dependent on the choice of ω.[11] For the optimum value of ω, $C(\omega)$ is $0(N)$ while for values of ω only slightly different

from the optimum $C(\omega)$ can be $0(N^2)$. In the computations presented in the next section, $N \leqq 100$ so the iterative process was stopped when the residual terms were less than 10^{-8}; this allowed margin for the fact that (18) is nonlinear. A discussion of the choice of ω is necessarily even more heuristic. It was found that for a mesh of $N = 50$ points and large negative values of ρ_o corresponding to small positive charge densities (i.e., $e^{\rho_o - \psi^i} \ll 1$ for all i), the iterative scheme was convergent for any choice of ω and any initial estimate of the $\{\psi^i\}$. For less-negative values of ρ_o and the corresponding larger values of $e^{\rho_o - \psi^i}$, the scheme was convergent for over-relaxation ($\omega > 1$) only if $r_j < 10$ for all j (approximate figure); but using under-relaxation the process was well behaved with r_j's as great as 10^4. It should be pointed out that due to the exponential nature of the right-hand side of eq. (19) even a very good initial estimate typically resulted in $r_1 > 10^3$ when over-relaxation was applied. Although the SOR process described above may well be stable[12] regardless of r_j values, the magnitude of the exponential terms in (19) limits machine computations. The combination of under- and over-relaxation detailed below eliminates this problem.

Good initial estimates for the $\{\psi^i\}$ were obtained by choosing the ρ_o values with equal spacing $\Delta\rho$; the first ρ_o value being the small positive charge-density case described above. With $\Delta\rho < 10$ a linear estimate of the $\{\psi^i\}$ for successive values of ρ_o was adequate. The iterations were under-relaxed ($\omega = 0.5$) until the residual term was less than 10; successive iterations were over-relaxed so as to increase the rate of convergence. The choice of ω for the SOR steps was not critical; values in the range $1 < \omega \leqq 1.5$ had approximately the same rate of convergence. Values above 1.5 did not converge for all values of ρ_o.

Total positive charge in the device for a given value of ρ_o is calculated by

$$Q_+ = \int_{y_{N_1}}^{y_{N_2}} e^{\rho_o - \psi_2(y)} dy + \int_{y_{N_2}}^{L} e^{\rho_o - \psi_3(y)} dy. \tag{29}$$

Using a piecewise linear approximation for $\psi(y)$ on $[y_k, y_{k+1}]$,

$$\psi(y) \cong \psi^k + ((y - y_k)/\delta)(\psi^{k+1} - \psi^k), \tag{30}$$

and summing the integrals over each such interval in $[y_{N_1}, L]$ gives the approximation

$$Q_+ \cong \delta_p \sum_{i=N_1}^{N_2-1} e^{\rho_o}(e^{-\psi^i} - e^{-\psi^{i+1}})/(\psi^{i+1} - \psi^i)$$

$$+ \delta_n \sum_{i=N_2}^{N-1} e^{\rho_o}(e^{-\psi^i} - e^{-\psi^{i+1}})/(\psi^{i+1} - \psi^i), \tag{31}$$

which can be modified to include the case $\psi^{i+1} = \psi^i$. Dimensional charge in coulombs can be calculated from Q_+ by $Q = Q_+ N_D \lambda$.

The operational scheme may be summarized as follows:

(i) Choose a value of ρ_o corresponding to small total charge Q_+ ($\rho_o \rightarrow -\infty$ in eq. (24) $\Rightarrow Q_+ \rightarrow 0$),

(ii) Solve for $\psi(y)$, thus determining $p(y)$,

(iii) Integrate $p(y)$ to find Q_+,

(iv) Increment ρ_o by $\Delta\rho$ and repeat (ii), (iii), and (iv).

The technique of starting with $Q_+ \approx 0$ and slowly adding positive charge to the device is important since it is this scheme of operation, in conjunction with the successive under- and over-relaxation, that avoids the exponential overflow limitations in machine computation.

III. COMPUTATIONAL RESULTS

In this section, numerical results for two specific device configurations will be given. The first device has a constant doping profile (N_A = a constant) and dimensions

$$h_1/\lambda = 0.48,$$
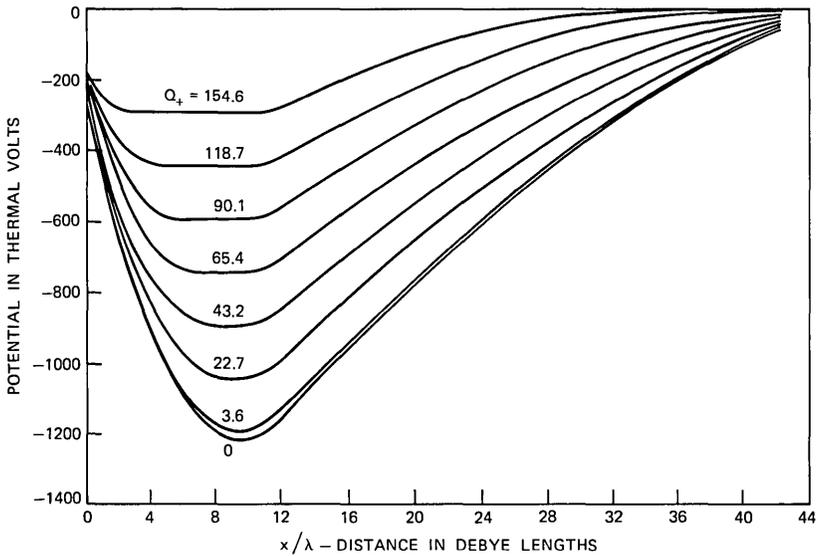$$h_2/\lambda = 12.48,$$
$$L = 42.48.$$

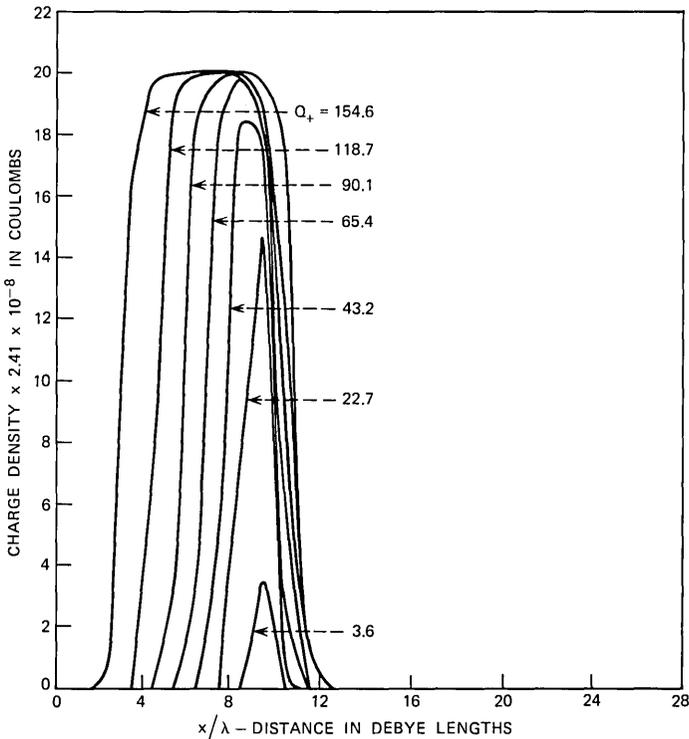Fig. 2—Potential versus distance for a buried-channel device with uniformly doped p-region and $V_o = -4$ volts.

Fig. 3—Charge density versus distance for a buried-channel device with uniformly doped p-region.

λ in this case is approximately 0.415 μm. Doping levels are

$$N_A = 2 \times 10^{15} \text{ cm}^{-3},$$
$$N_D = 1 \times 10^{14} \text{ cm}^{-3},$$

so

$$\sigma = N_A/N_D = 20.$$

N_c is calculated by[13]

$$N_c = 4.831 \times 10^{15}(m_c/m_o)^{\frac{3}{2}}T^{\frac{3}{2}},$$

with $T = 300°$K and m_c as before.

Using eq. (46), η is found to be

$$\eta = 12.09.$$

The nondegeneracy condition for this device may now be stated using eq. (45)

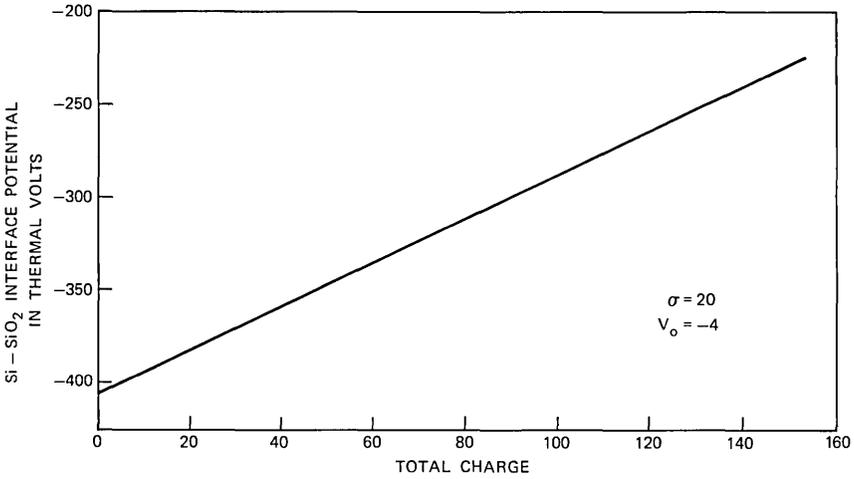$$\epsilon_g - \eta' + \psi(y) > 3.5,$$

Fig. 4—Potential at the Si-SiO$_2$ interface versus total positive charge in a buried-channel device with a uniform p-region and $V_o = -4$ volts.

or

$$\eta' - \epsilon_g - \psi(y) < -3.5.$$

Adding η to both sides of the inequality gives

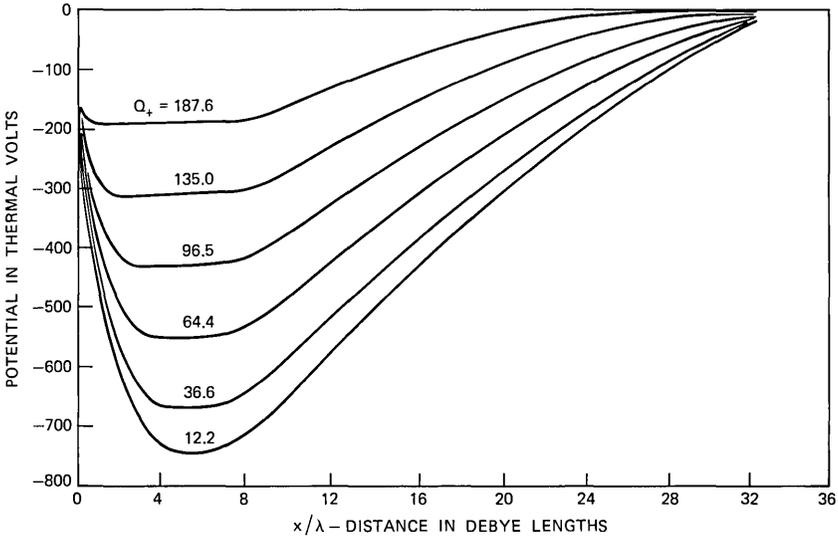$$\rho_o - \psi(y) < \eta - 3.5.$$

Fig. 5—Potential versus distance for a buried-channel device with a Gaussian p-region doping profile and $V_o = -4$ volts.
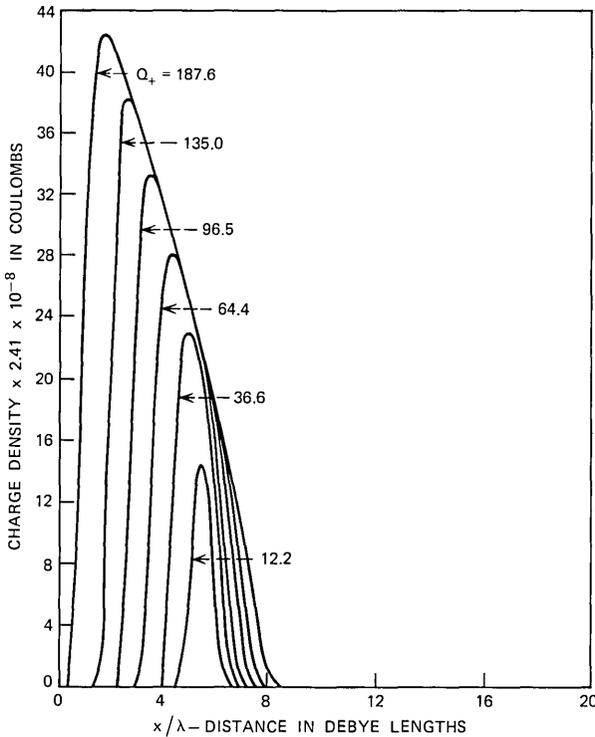
Fig. 6—Charge density versus distance for a buried-channel device with a Gaussian doping profile.

Then for each choice of the parameter ρ_o, the solution set $\{\psi^i\}$ must satisfy:

$$\rho_o - \psi^i < 8.59. \tag{32}$$

From eq. (12) it is clear that if one assumes the Boltzmann approximation to hold, then as long as $Q_+ < \sigma(h_2 - h_1)/\lambda$ the equilibrium condition causes the lineal charge density to have constant sign so

$$\rho_o - \psi^i < \log \sigma(m_c/m_v)^{\frac{3}{2}} \tag{33}$$

for all i; for this device, eqs. (32) and (33) are always consistent if $N_A < 1.3 \times 10^{18} \ cm^{-3}$.

Computation was performed using a 90-point mesh and took approximately 4.8 minutes of processor time on a Honeywell 6000-series machine.

Figures 2, 3, and 4 summarize the computational results. Figure 2 shows potential solutions $\{\psi^i\}$ versus the points y_i for a family of ρ_o
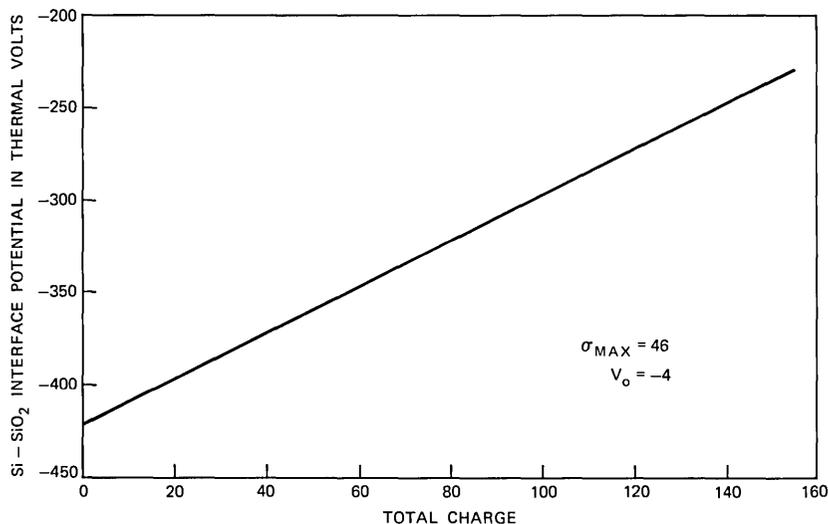
Fig. 7—Potential at the Si-SiO₂ interface versus total positive charge in a buried-channel device with a Gaussian p-region doping profile and $V_o = -4$ volts.

values with $V_o = -4$ volts. The range of total charge values is indicated. Figure 3 is a plot of the linear charge density with Q_+ values indicated. Figure 4 is Q_+ plotted against ψ^{N_1} (the oxide-interface potential).

A realistic modification to the device studied thus far is to allow σ to have y variation. The second set of results presented are for the same device as described above but with the p-region having a doping profile

$$\sigma(y) = (D_s + 1)e^{-[(y-h_1)/(h_2-h_1)]^2 \ln (D_s+1)} - 2. \tag{34}$$

The values for h_1, h_2, and V_o are as before, $L = 32.48$, and $D_s = 46$. [This value of D_s corresponds to an average doping level in the p-layer of $\bar{N}_A = 1.9 \times 10^{15}/cm^3$, and eq. (34) describes a doping profile as if the p-layer were formed by drive-in diffusion.[14]] The solutions $\{\psi^i\}$ must still satisfy eq. (32). Figures 5, 6, and 7 are a summary of results.

Comparison of the two cases shows that for the doping profile in eq. (34) the potential minimum is greater and the "channel" is shifted toward the oxide. In both cases the added positive charge is contained entirely in the p-region and it fills the region starting from the side remote from the oxide interface (see Figs. 2 and 5).

IV. ACKNOWLEDGMENTS

Special thanks goes to J. McKenna for his many contributions during the course of this work and for his comments regarding the preparation of this paper.

APPENDIX

Substituting eqs. (2), (4), and (7) into eq. (1) results in ($'$ denotes d/dx):

$$\phi_1''(x) = 0, \qquad 0 \leqq x \leqq h_1, \tag{35}$$

$$\phi_2''(x) = \frac{q}{\epsilon_2} \left\{ N_A - N_v^o \int_{-\infty}^{E_v(x)} \frac{(E_v(x) - E)^{\frac{1}{2}} dE}{1 + \exp\left[(E_{fh} - E)/kT\right]} \right\},$$
$$h_1 \leqq x \leqq h_2, \tag{36}$$

$$\phi_3''(x) = \frac{q}{\epsilon_2} \left\{ N_c^o \int_{E_c(x)}^{\infty} \frac{(E - E_c(x))^{\frac{1}{2}} dE}{1 + \exp\left[(E - E_f)/kT\right]} - N_D \right.$$
$$\left. - N_v^o \int_{-\infty}^{E_v(x)} \frac{(E_v(x) - E)^{\frac{1}{2}} dE}{1 + \exp\left[(E_{fh} - E)/kT\right]} \right\},$$
$$h_2 \leqq x \leqq \infty. \tag{37}$$

Transform the integral involving $E_c(x)$ by making the substitutions:

$$\epsilon = (E - E_c(x))/kT,$$
$$E_c(x) = E_c^o - q\varphi(x),$$
$$\eta = (E_c^o - E_f)/kT.$$

For the integrals involving $E_v(x)$ make the substitutions:

$$\epsilon = (E_v(x) - E)/kT,$$
$$E_v(x) = E_v^o - q\varphi(x),$$
$$\eta' = (E_c^o - E_{fh})/kT,$$
$$\epsilon_g = (E_c^o - E_v^o)/kT.$$

For both cases let

$$\psi(x) = q\varphi(x)/kT$$

and make the change of variables

$$y = x/\lambda,$$

where

$$\lambda = \sqrt{kT\epsilon_2/q^2 N_D}.$$

These substitutions and some straightforward manipulation reduce eqs. (35), (36), and (37) to

$$\psi_2''(y) = 0, \qquad 0 \leqq y \leqq h_1/\lambda, \tag{38}$$

$$\psi_2''(y) = \frac{N_A}{N_D} - \frac{N_v^o}{N_D} (kT)^{\frac{3}{2}} \int_0^{\infty} \frac{\epsilon^{\frac{1}{2}} d\epsilon}{1 + \exp\left(\epsilon + \epsilon_g - \eta' + \psi_2(y)\right)},$$
$$h_1/\lambda \leqq y \leqq h_2/\lambda, \tag{39}$$

$$\psi_3''(y) = \frac{N_c^o}{N_D} (kT)^{\frac{3}{2}} \int_0^\infty \frac{\epsilon^{\frac{1}{2}} d\epsilon}{1 + \exp\left[\epsilon + \eta - \psi_3(y)\right]} - 1$$

$$- \frac{N_v^o}{N_D} (kT)^{\frac{3}{2}} \int_0^\infty \frac{\epsilon^{\frac{1}{2}} d\epsilon}{1 + \exp\left(\epsilon + \epsilon_g - \eta' + \psi_3(y)\right)},$$

$$h_2/\lambda \leqq y \leqq \infty. \quad (40)$$

These equations are further simplified by making the classical Boltzmann approximation for the electrons:

$$\int_0^\infty \frac{\epsilon^{\frac{1}{2}} d\epsilon}{1 + \exp\left(\epsilon + \eta - \psi(y)\right)} \cong \int_0^\infty \frac{\epsilon^{\frac{1}{2}} d\epsilon}{\exp\left(\epsilon + \eta - \psi(y)\right)} \quad (41)$$

which holds since $\eta - \psi(y)$ is always positive by several kT (see Fig. 8). The last result is integrable and simplifies the approximation to

$$\Gamma(\tfrac{3}{2}) \exp\left(\psi(y) - \eta\right), \quad (42)$$

where $\Gamma(\cdot)$ is the usual gamma function. This approximation has been shown to be less than one percent in error if[15]

$$\eta - \psi(y) > 3.5. \quad (43)$$

Similarly, for the holes,

$$\int_0^\infty \frac{\epsilon^{\frac{1}{2}} d\epsilon}{1 + \exp\left(\epsilon + \epsilon_g - \eta' + \psi(y)\right)} \cong \Gamma(\tfrac{3}{2}) \exp\left(\eta' - \epsilon_g - \psi(y)\right) \quad (44)$$

if

$$\epsilon_g - \eta' + \psi(y) > 3.5. \quad (45)$$

The validity of this last approximation is not clear since $\psi(y)$ is dependent on η' and $\psi(y)$ becomes large and negative. Note that *requiring* the inequalities (43) and (45) to hold is equivalent to requiring the device to be nondegenerate. Computational results show this approximation to be consistent.

η is expressed in terms of the constants N_c, N_v, and N_D by requiring electron charge neutrality at $x = \infty$. It should be stated that the correct condition to use at this point is complete charge neutrality. This requirement, however, is computationally indistinguishable from the algebraically simpler condition used here. Using eqs. (40) and (15), the neutrality condition is

$$(N_c/N_D)e^{-\eta}\Gamma(\tfrac{3}{2}) - 1 = 0,$$

so

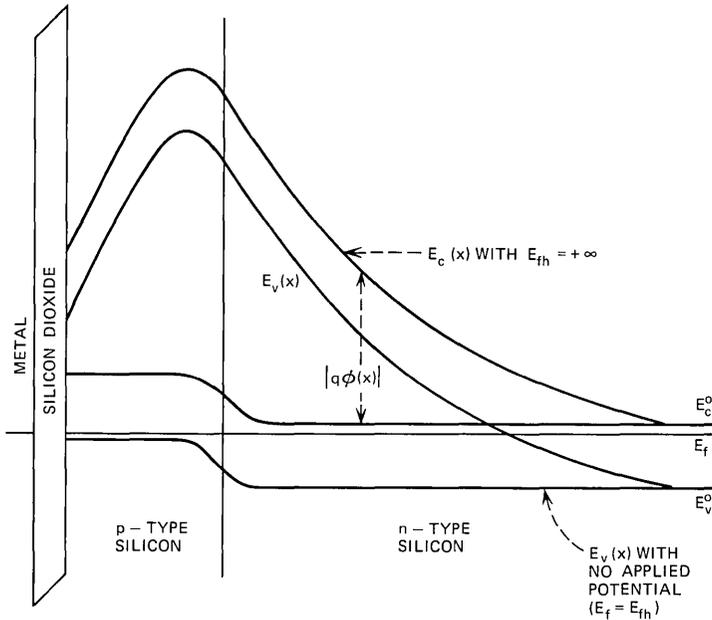$$N_c/N_D = e^\eta/\Gamma(\tfrac{3}{2}). \quad (46)$$

Fig. 8—Band diagram for buried-channel device.

Equation (46) defines the Fermi level for the electrons. Using eqs. (5) and (8),

$$N_v/N_D = (N_v/N_c) \cdot (N_c/N_D) = (m_v/m_c)^{\frac{3}{2}} e^\eta / \Gamma(\tfrac{3}{2}), \qquad (47)$$

where m_v and m_c are the effective masses of holes and electrons in silicon; they are 1.08 m_o and 0.59 m_o, respectively.[10] Using eqs. (42), (44), (46), and (47) in eqs. (38), (39), and (40) results in equations (11), (12), and (13) which are to be solved.

REFERENCES

1. Walden, R. H., Krambeck, R. H., Strain, R. S., McKenna, J., Schryer, N. L., and Smith, G. E., "The Buried Channel Charge Coupled Device," B.S.T.J., *51*, No. 7 (September 1972), pp. 1635–1640.
2. Amelio, G. F., "Computer Modeling of Charge-Coupled Device Characteristics," B.S.T.J., *51*, No. 3 (March 1972), pp. 705–730.
3. McKenna, J., and Schryer, N. L., "The Potential in a Charge Coupled Device With No Mobile Minority Carriers and Zero Plate Separation," B.S.T.J., *52*, No. 5 (May-June 1973), pp. 669–696.
4. McKenna, J., and Schryer, N. L., "The Potential in a Charge-Coupled Device With No Mobile Minority Carriers," unpublished work.
5. Strain, R. J., and Schryer, N. L., "A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer," B.S.T.J., *50*, No. 6 (July-August), pp. 1721–1740.
6. Blakemore, J. S., *Semiconductor Statistics*, New York: Pergamon Press, 1962, pp. 75–77.

7. Grove, A. S., *Physics and Technology of Semiconductor Devices*, New York: John Wiley and Sons, 1967, p. 162.
8. Blakemore, J. S., op. cit., p. 81.
9. Blakemore, J. S., op. cit., pp. 58–63.
10. Varga, R. S., *Matrix Iterative Analysis*, Englewood Cliffs, N. J.: Prentice-Hall, 1962, p. 59.
11. McKenna, J., unpublished work.
12. Richtmeyer, R. D., and Morton, K. W., *Difference Methods for Initial Value Problems*, 2nd Ed., New York: Interscience, 1970, p. 9.
13. Blakemore, J. S., op. cit., p. 82.
14. Grove, A. S., op. cit., p. 43.
15. Blakemore, J. S., op. cit., p. 358.

Contributors to This Issue

WIN AUNG, B.S. (M.E.), 1963, University of Rangoon, Rangoon, Burma; M.S. (M.E.), 1966, and Ph.D. (M.E.), 1969, University of Minnesota; Bell Laboratories, 1969—. Mr. Aung has been engaged in fluid mechanics studies of reentry vehicles and development of general thermal design techniques for new electronic systems. More recently, he has worked on manufacturing process evaluation and reliability investigation of new interconnection devices for electronic systems. Member, American Society of Mechanical Engineers, Sigma Xi.

DAVID S. K. CHAN, S.B. (Electrical Engineering) and S.M. (Electrical Engineering), 1972, Massachusetts Institute of Technology; Bell Laboratories, 1970—. Mr. Chan has worked on digital switching and transmission systems, digital circuitry, and digital filtering. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

GERARD J. FOSCHINI, B.S.E.E., 1961, Newark College of Engineering; M.E.E., 1963, New York University; Ph.D. (Mathematics), 1967, Stevens Institute of Technology; Bell Laboratories, 1961—. Mr. Foschini initially worked on real-time program design. Since 1965 he has mainly been engaged in analytical work concerning the transmission of signals. Currently he is working in the area of data communication theory. Member, Sigma Xi, Mathematical Association of America, American Men of Science, New York Academy of Sciences.

RICHARD D. GITLIN, B.E.E., 1964, City College of New York; M.S., 1965, and D.Eng.Sc., 1969, Columbia University; Bell Laboratories, 1969—. Mr. Gitlin is presently concerned with problems in data transmission. Member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

D. GLOGE, Dipl. Ing., 1961, Dr. Ing., 1964, Technical University of Braunschweig, Germany; Bell Laboratories, 1965—. Mr. Gloge's work has included the design and field testing of various optical transmission media and the application of ultra-fast measuring techniques to optical component studies. He is presently engaged in transmission research related to optical fiber communication systems.

OTTO HERRMANN, Dipl.-Ing. (Electrical Engineering), 1956, and Dr.-Ing. (Electrical Engineering), 1965, University of Aachen, Germany; venia legendi, 1971, University of Erlangen, Nuremberg, Germany. Mr. Herrmann has worked on problems concerning approximation theory as applied to analog and digital filter design. From 1959 to 1971 he was a Teaching and Research Assistant at the University of Aachen, University of Karlsruhe, and University of Erlangen. He was at Bell Laboratories during the summer of 1972 on leave from the Technical Faculty at the University of Erlangen. Presently, he teaches courses in communications, analog computation, and digital signal processing at the University of Erlangen. Member, Nachrichtentechnische Gesellschaft.

W. H. KENT, B.S.E.E., 1968, Michigan Technological University; M.S.E.E., 1969, University of Illinois; Bell Laboratories, 1968—. Since joining Bell Laboratories, Mr. Kent has been involved in studies concerning electromagnetic sensing of and transmission in coaxial and twisted-pair cable; he is currently in the Loop Transmission Engineering Center. Member, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi.

ROY STEPHEN KRUPP, S.B. (Mathematics, Physics), 1960, Massachusetts Institute of Technology; M.I.T. Aerophysics Laboratory, 1960–65; S.M., 1967 and Ph.D., 1970 (Aeronautics and Astronautics), Massachusetts Institute of Technology; Bell Laboratories, 1970—. A member of the Toll Switching Systems Studies Department, Mr. Krupp has worked at modeling the toll network and on studies of time-division switching networks. His general interests include combinatorics, fluid mechanics, and various branches of applied mathematics.

ANATOL KUCZURA, B.S., (Engineering Physics), 1961, University of Illinois; M.S. (Mathematics), 1963, University of Michigan; M.S.E.E., 1966, New York University; Ph.D. (Mathematics), 1971, Polytechnic Institute of Brooklyn; Bell Laboratories, 1963–1973. From 1963 to 1966, Mr. Kuczura worked in military systems engineering. Since 1966, he has been engaged in research on the application of probability theory and stochastic processes to the analysis of telephone traffic and queuing. Mr. Kuczura is now Director, Systems Analysis, at the North Electric Company's Paul H. Henson Research Center. Member, ORSA, SIAM, American Mathematical Society, Mathematical Association of America, AAAS, Chi Gamma Iota, Pi Mu Epsilon.

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 Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell 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 Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of two books. Fellow, IEEE; member, Optical Society of America.

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 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 Mathematical Society.

S. D. PERSONICK, B.E.E., 1967, City College of New York; S.M., 1968, E.E., 1969, and ScD.., 1969, Massachusetts Institute of Technology; Bell Laboratories, 1967—. Mr. Personick is engaged in studies of optical communication systems.

LAWRENCE R. RABINER, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; Chairman of the IEEE G-AU Technical Committee on Digital Signal Processing; vice-president of the G-AU AdCom, associate editor of the G-AU Transactions; member of the technical committees on speech communication of both the IEEE and Acoustical Society.

WOLFGANG O. SCHLOSSER, Dr. Ing., 1964, Technische Hochschule, Darmstadt, Germany; Research Associate, Technische Hochschule,

Braunschweig, Germany, 1963–1966; Bell Laboratories, 1966—. Mr. Schlosser's work has included the design of microwave IMPATT oscillators and the design of millimeter-wave phase switches and PIN diodes. He is now working on optical communication subsystems. Member, IEEE.

L. A. TOMKO, B.S., 1966, Oklahoma State University; M.S., 1967 and Ph.D., 1971, University of Illinois; Bell Laboratories, 1970—. Mr. Tomko has been engaged in toll switching system exploratory studies. He is presently involved with common channel interoffice signaling implementation. Member IEEE.

STEPHEN B. WEINSTEIN, B.S.E.E., 1960, Massachusetts Institute of Technology; M.S.E.E., 1962, University of Michigan; Ph.D., 1966, University of California at Berkeley; research engineer at Philips Research Laboratories, Eindhoven, Netherlands, 1967–1968; Bell Laboratories, 1968—. Mr. Weinstein's technical interests include data communication, statistical estimation theory, and information retrieval. Member, IEEE, Sigma Xi.

B. S. T. J. BRIEFS

The Effect of Rain on Circular Polarization at 18 GHz

By R. A. SEMPLAK

(Manuscript received February 14, 1973)

Limitations imposed by attenuation during heavy rain on the reliability of microwave systems are well known,[1] and some calculations of the depolarization of linearly[2] and circularly[3] polarized waves have been made. Recently, measurements of rain-induced rotation[4] of linear polarization at 30 GHz indicated that depolarization by large oblate raindrops[5] may limit efficient utilization of a microwave channel where orthogonal polarizations are employed. However, an advantage in using circular polarization is that less-stringent mechanical stability may be required of some antennas; also, in satellite systems, circular polarization is not affected by Faraday rotation. As the effect of rain on transmission of circular polarization had not been measured, an experiment was initiated. Data have been collected for the period June 1, 1972, through January 24, 1973.

A frequency-swept Gunn oscillator operating at a frequency of 18.5 GHz is used as a source in circular polarization on a 2.6-km path oriented in a southeasterly direction from Crawford Hill, Holmdel, New Jersey, the site of the receiver. The receiver has a ferrite switch which looks sequentially at the received fields, i.e., the desired circular polarization and then the depolarized component, the opposite sense, are observed.* Paper strip chart recordings are made of both the desired and depolarized components.

The clear-day polarization discrimination of the system is better than 32 dB. In view of the ambiguities[6] associated with measuring cross polarization below −32 dB, none of the data below −32 dB are included.

The total data obtained from the circularly polarized system are shown in Fig. 1 where the depolarized component (the field measured in the undesired sense) is plotted as a function of the rain-induced attenuation. One can see that rain has a strong depolarizing effect. For example, the very deep rain-induced attenuations of 39–40 dB have

---

* The switching rate is 17 Hz; this is much faster than the changes in attenuation produced by rain.
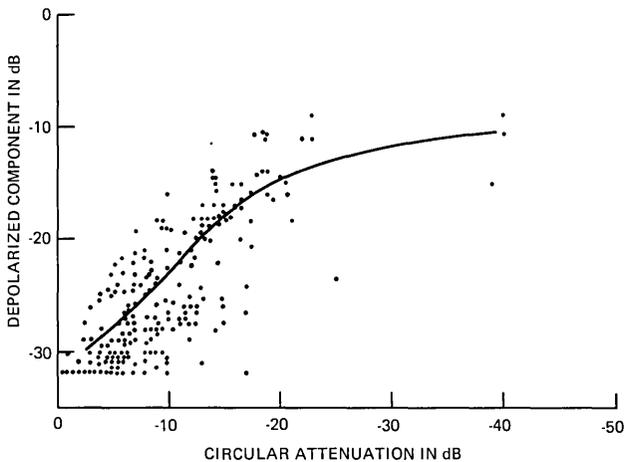
Fig. 1—Data on circular polarization from June 1, 1972, through January 24, 1973. Depolarized component plotted as a function of rain-induced attenuation.

depolarized components that are only 9–15 dB below that level. The curve shown in Fig. 1 is an estimate of the median of the data. From the figure, we observe, for example, that a rain-induced fade of 20 dB has associated with it a depolarized component with a median value of the order of −15 dB. Likewise, for a modest rain-induced fade of 10 dB, the median value of the depolarized component is about 23 dB below that level.

From measured data at 18 GHz it is concluded that there are serious polarization discrimination problems for circular polarization during periods of rain. However, circular depolarization should not be as serious for frequencies of 60 GHz and higher, since at these frequencies the small raindrops have the strongest effect on transmission, and these small drops tend to be spherical rather than oblate. Measurements not discussed here show by comparison that the attenuation of circularly polarized waves by rain lies between that for horizontally and vertically polarized waves.[7]

REFERENCES

1. Hogg, D. C., "Statistics on Attenuation of Microwaves by Intense Rain," B.S.T.J., 48, No. 9 (November 1969), pp. 2949–2962.
2. Saunders, M. J., "Cross Polarization at 18 and 30 GHz Due to Rain," IEEE Trans. Ant. and Prop., AP-19, No. 2 (March 1971), pp. 273–277.
3. Hogg, D. C., "Depolarization of Microwaves in Transmission Through Rain," AGARD Conference Preprint CPP-107, p. 6–1; Telecommunications Aspects on Frequencies Between 10 & 100 GHz, September 1972.

4. Semplak, R. A., "Rain-Induced Polarization Rotation," to be published in Radio Science.
5. Morrison, J. A., Cross, M. J., and Chu, T. S., "Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 4 (April 1973), pp. 599–604.
6. Chu, T. S., private communication.
7. Semplak, R. A., "Effect of Oblate Raindrops on Attenuation at 30.9 GHz," Radio Sci., *5*, No. 3 (March 1970), pp. 559–564.

Attenuation Through the Clear Atmosphere at 30, 19, and 13 GHz for Low Elevation Angles

By PAUL S. HENRY

I. INTRODUCTION

Synchronous satellite service for Alaska, and possibly other places, requires that ground station antennas point at low elevation angles. For example, from Point Barrow (71°N, 155°W), such a satellite would never be more than 11 degrees above the horizon, and for satellite longitudes 45 degrees east or west of Point Barrow, the elevation is only 5 degrees. At such low angles, the attenuation of a nominally clear atmosphere is significant in the 18- and 30-GHz bands of proposed domestic satellite systems. There are also satellite bands near 13 GHz, where the attenuation is expected to be somewhat lower. Predictions of this attenuation have been made,[1] but as a check direct measurements have been obtained with the Crawford Hill Sun Tracker as reported below.

II. APPARATUS AND PROCEDURE

The experimental setup, described in detail elsewhere,[2] is briefly this: the antenna temperatures of the sun and a nearby patch of sky are compared by means of a radiometer. The temperature difference, ΔT, is related to the excess attenuation above the solar noon value, A, by the formula

$$A \text{ (dB)} = 10 \log [\Delta T / \Delta T_0], \tag{1}$$

where ΔT_0 is the antenna temperature difference at solar noon on a clear day. As the antenna follows the setting sun, attenuation as a function of elevation angle is measured directly.

The measured attenuations are increases above the noontime value. Thus the absolute attenuation is the measured value plus the atmospheric attenuation at noon; the latter has been estimated to be 0.3, 0.3, and 0.1 dB at 30, 19, and 13 GHz.[1] For the remainder of this paper, quoted attenuations are referred to noontime as the zero of attenuation.

A small correction must be applied to the Sun Tracker raw data to account for different antenna temperatures in the "sun" and "sky" positions due to differences in atmospheric radiation. This correction is readily determined by measuring ΔT for low elevation angles of the antenna beam, while the sun is high in the sky and thus out of the beam. The magnitude of the correction is less than 3 percent of the measured attenuation in dB for elevations of 5 degrees or greater. Below 5 degrees, the correction rises to a maximum of 8 percent at 3 degrees elevation.

III. DATA

The measured attenuations, corrected as described above, are shown as a function of elevation angle in Fig. 1. The data, collected during four sunsets in August 1972, are shown as bars spanning the full range of the values observed. During the measurements the absolute surface humidity was about 12.5 gm/m³, which is typical of summertime New Jersey. The curves through the bars represent the average behavior of the data. At 5-degree and 10-degree elevations are indicated the
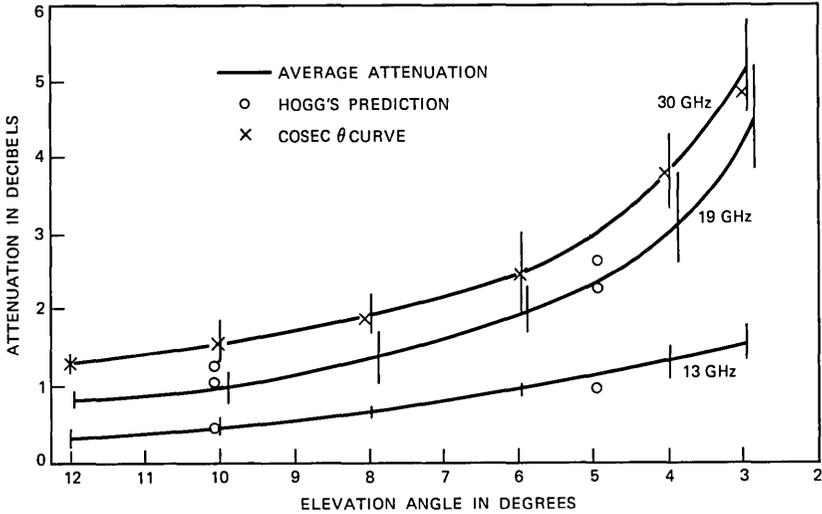

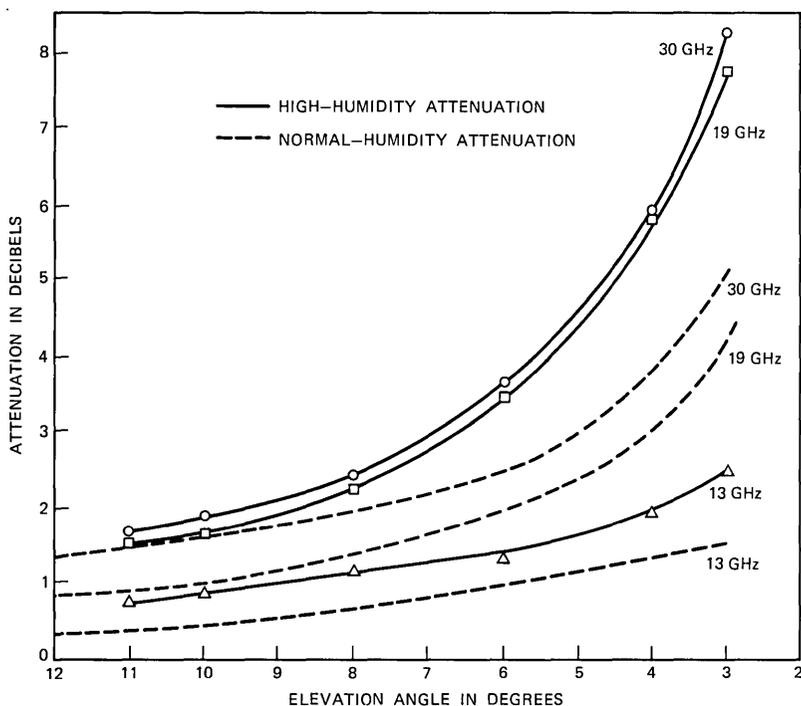FIG. 1—Attenuation vs elevation angle, normal summer weather (average humidity 12.5 gm/m³).

FIG. 2—Attenuation vs elevation angle, humid weather (August 7, 1972, humidity 23 gm/m³).

attenuation predictions of Hogg, normalized to zero dB at noon. Also shown for the 30-GHz attenuation are the points corresponding to a cosec θ law, normalized to agree with observation at 10 degrees and 12 degrees.

Figure 2 is a plot of measured attenuation versus elevation for very humid air. These measurements were made on August 7, 1972, shortly after a rainstorm when the atmosphere had cleared. The surface humidity was 23 gm/m³ at the time of observation. For comparison, the curves from Fig. 1 also are shown. The increased attenuation is clearly visible. Even larger attenuations are possible. Hogg and Semplak[3] have calculated that very humid weather can result in attenuations more than double those shown in Fig. 2.

IV. EXTRAPOLATION TO OTHER ATMOSPHERIC CONDITIONS

The atmospheric attenuation is due primarily to oxygen and water vapor. In summertime New Jersey, the fraction of the total attenuation attributable to water vapor is roughly 50, 75, and 50 percent at

30, 19, and 13 GHz, respectively.[1] Thus one can convert attenuations measured at one humidity to values corresponding to a different water vapor content. For example, we can predict the attenuations that should have been observed under the conditions of Fig. 2 (23 gm/m³) by doubling (very nearly) the contributions due to water vapor in Fig. 1 (12.5 gm/m³). This simple extrapolation rule predicts the observed attenuation to an accuracy of 10 to 15 percent.

Compared with New Jersey, the Point Barrow atmosphere contains essentially no water. In winter the humidity is 0.5 gm/m³ and in summer 5 gm/m³.[4] Therefore, a first approximation for conversion of the attenuations of Fig. 1 to Alaskan conditions consists in neglecting the water vapor entirely and simply reducing the observed attenuations by 50, 75, and 50 percent at 30, 19, and 13 GHz.

V. ERRORS

There are two main sources of systematic error. The first is occultation of the setting sun by objects near the horizon. At the Crawford Hill location the Sun Tracker has a clear view down to about 3 degrees elevation. The second error is an increase in antenna beamwidth in the vertical plane due to the gradient of atmospheric refraction within the antenna beam.[5] The forward gain of the antenna is thus reduced, resulting in an apparent increase in attenuation. The magnitude of this effect depends on the size of the source being observed. A "point" source, such as a synchronous satellite, would show an attenuation about 0.1 dB above the values reported here.

Strictly speaking, the points calculated by Hogg shown in Fig. 1 should not be compared directly with the data. Although the attenuation at solar noon has been subtracted from them (they are normalized to 0 dB at noon), they still do not correspond to the conditions of this experiment. Hogg assumed a humidity of 10 gm/m³—a value 20 percent below the prevailing humidity during the observations. A rough correction to Hogg's values would involve scaling his attenuations (in dB) up by 10, 15, and 10 percent at 30, 19, and 13 GHz.

Fluctuations in the data are attributable to three sources: (i) error in reading attenuations on the chart recorder, (ii) error in reading the time on the chart, and (iii) changes in humidity over the course of the observations. The first-mentioned error, called δA below, is estimated to be about 0.2 dB rms, one-fifth of the smallest chart division. The timing error is significant because it leads to an uncertainty in the elevation angle, $\delta\theta$ below, at which a particular measurement was made.

A reasonable estimate for this error is 1 minute rms, which means $\delta\theta$ is 0.25 degree. Finally, the diurnal changes in humidity, α, were about 20 percent. All these errors add in quadrature to give a resultant overall error estimate

$$E^2 = (\delta A)^2 + S^2(\delta\theta)^2 + (f\alpha A)^2, \qquad (2)$$

where S is the slope of the attenuation versus elevation angle curve (see Fig. 1), f is the fraction of the attenuation due to water vapor, and A is the measured attenuation. The errors represented by the three terms of eq. (2) are of comparable magnitude. Thus the noise in the data must be ascribed to both instrumental and "external" sources.

The rms fluctuations predicted by eq. (2) can be compared with the observed scatter in the measurements at various frequencies and angles. Typically the two differ by only 0.2 to 0.3 dB, indicating that the stochastic processes operating in this experiment are reasonably well understood.

VI. CONCLUSIONS

The 13-, 19-, and 30-GHz attenuation of a clear atmosphere at low elevation angles has been measured by the Crawford Hill Sun Tracker. The results are in good agreement with predictions. Extrapolation of the measurements to Alaskan conditions yields attenuations substantially below those measured in New Jersey. A significant improvement in the measurements using the Crawford Hill Sun Tracker can be made only under conditions of reduced and/or more stable atmospheric water vapor content.

VII. ACKNOWLEDGMENTS

REFERENCES

1. Hogg, D. C., "Effective Antenna Temperatures Due to Oxygen and Water Vapor in the Atmosphere," J. Appl. Phys., *30* (1959), p. 1417.
2. Wilson, R. W., "Sun Tracker Measurements of Attenuation by Rain at 16 and 30 GHz," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1383–1404.
3. Hogg, D. C., and Semplak, R. A., "The Effect of Rain and Water Vapor on Sky Noise at Centimeter Wavelengths," B.S.T.J., *40*, No. 5 (September 1961), pp. 1331–1348.
4. Bean, B. R., and Dutton, E. J., *Radio Meteorology*, National Bureau of Standards Monograph 92, 1966, pp. 277–290.
5. Howell, T. F., and Shakeshaft, J. R., "Attenuation of Radio Waves by the Troposphere over the Frequency Range 0.4–10 GHz," J. Atmosph. Terr. Phys., *29*, (1967), pp. 1559–1571.