THE FEBRUARY 1974
VOL. 53 NO. 2
BELL SYSTEM
TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer, R. D. Lilley, President, J. J. Scanlon, Executive Vice President and Chief Financial Officer, F. A. Hutson, Jr., Secretary. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 1038, 195 Broadway, New York, N. Y. 10007. Subscriptions $11.00 per year; single copies $1.50 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 February 1974 Number 2

Delay Distortion in Generalized Lens-Like Media

By S. E. MILLER

(Manuscript received August 16, 1973)

Explicit expressions are derived for the phase constant, the specific-group-delay constant, and the rms width of the impulse response for two-dimensional or square media having a transverse variation of index of refraction according to $n = n_1(1 - \frac{1}{2}a_u x^u - \frac{1}{2}a_r x^r)$, *in which* x *is the transverse dimension,* a_u *and* a_r *are constants with* $|a_r| \ll a_u$, *and* $(n - n_1) \ll 1$. *Use is made of an approximation which the author has previously shown yields significant results.*

The results are applied to fibers with graded-index variation, clad by an additional medium of index $n = n_1(1 - \Delta)$. *The ideal index gradient, a near-parabolic profile, gives delay distortion orders of magnitude less than for the conventional fiber with a step-change in index at the core-cladding boundary. However, it is shown that several forms of 5-percent error in the ideal gradient yield improvement of the order of 50 compared with the conventional clad fiber. The delay distortion is shown to be very sensitive to the exact index distribution in the vicinity of the ideal distribution but increasingly insensitive to perturbations in the index distribution as that distribution departs more and more from the ideal.*

I. INTRODUCTION

Optical fibers have assumed considerable importance for potential use as transmission media wherever wire pairs or coaxials are now used.

We give here some theory related to increasing the information capacity of such fibers.

Conventional fibers have a core of uniform index of refraction n, surrounded by a cladding of slightly lower index of refraction. The cladding serves to isolate the outer fiber surface from the optical field, which is confined to the core, thus permitting the fiber to be handled and bundled into cables without affecting the information transmission. More recently, fibers having continuously graded index of refraction, with maximum on the fiber's axis and lower values at increasing radial distances, were proposed and realized in practice.[1–3] Graded-index fibers have image-focusing properties[3] and provide modulated-carrier transmission with less delay distortion than conventional fibers having a step-index change at the core-cladding boundary.[4–6] Recent experimental work verified the potential of low delay distortion.[7] Current studies of graded-index fibers from Corning Glass Works show that total transmission losses under 10dB/kilometer and low delay distortion (approximately one nanosecond per kilometer) can be realized simultaneously in graded-index fibers.[8,9] This brings into focus the need for detailed knowledge about delay distortion related to the shape of the index gradient in the transverse plane. The following sections relate to that need. This study was done in parallel with that of D. Gloge and E. A. J. Marcatili.[10] The present paper presents an approximate analysis that may be extended to a wide variety of index distributions with closed-form solutions for the important wave-propagation constants.

Section II summarizes the approximate method of calculation, yielding in turn the phase constant, the specific-delay constant, and the impulse response.

Section III gives the solutions for the index distribution,

$$n = n_1(1 - \tfrac{1}{2}a_u x^u - \tfrac{1}{2}a_r x^r), \tag{1}$$

in which x is the radial coordinate, and a_u and a_r are arbitrary constants with $a_u > 0$, $a_u \gg |a_r|$, and $0.5\ a_u x^u \ll 1$. These conditions describe fibers of current interest.

Sections IV through VI discuss cases of interest, taken as simplifications of (1), which describe (*i*) an "ideal" index distribution, (*ii*) variations around the ideal, and (*iii*) other distributions which may result from convenient manufacturing processes.

This paper is concerned with the delay difference between signals launched simultaneously in the various propagating modes; the fiber-

output impulse response is derived assuming

(*i*) all propagating modes are excited equally at the fiber input,
(*ii*) the fiber structure is uniform along its length, yielding no mode conversion,
(*iii*) there are negligible losses or, equivalently, the same loss for all modes, and
(*iv*) material dispersion due to variation of the bulk index of refraction versus wavelength is ignored.

A very few modes very near cutoff are not accounted for; this is believed to yield negligible error, and the same assumption was made by Gloge and Marcatili.[10]

It is found that the delay distortion of graded-index media is a very critical function of the index distribution in the vicinity of the near-parabolic distribution which gives the lowest delay distortion. For index distributions increasingly far from the ideal, the performance is less and less sensitive to changes.

For fixed percentage error in fabrication of the index distribution, the delay distortion is *linearly* dependent on Δ, the fractional index difference between core center and the edge of the guiding region where the index becomes

$$n = n_1(1 - \Delta). \tag{2}$$

This is true even near the ideal index distribution, where previous work assuming no fabrication error indicated the delay distortion varied as Δ^2.[11]

II. OUTLINE OF THEORY

The approach taken here is based on Ref. 5, which gives an approximate method for deriving the phase constant and other relevant quantities for wave propagation in the generalized media.

We write the index n as a function of the transverse coordinate x,

$$n = n_1[1 + f(x)] \tag{3}$$

with $|f(x)| \ll 1$. For guiding media, $f(x)$ is predominantly negative; $x = 0$ at the central axis. Here we assume $f(x)$ is independent of the longitudinal coordinate. In accordance with Ref. 5 we derive a parameter a_e, which measures the transverse field width, using

$$f(a_e) = -0.1515 \left(\frac{m + 2.5}{2.5} \right)^2 \left(\frac{\lambda_0}{a_e n_1} \right)^2. \tag{4}$$

Thus the phase constant β is given by, from Ref. 5,

$$\beta = \frac{2\pi n_1}{\lambda_0}\left\{1 - \frac{1}{32}\left(\frac{\lambda_0}{n_1 a_e}\right)^2 (m+1)^2\right\}. \tag{5}$$

The normal mode field may be considered composed of plane-wave components traveling at an angle α to the longitudinal axis; from Ref. 5,

$$\alpha = \frac{\lambda_0}{n_1}\frac{(m+1)}{4a_e} \tag{6}$$

in which m is the mode number. The mth-order mode has $(m+1)$ extrema in the transverse cross section. The maximum angle that such plane-wave components can have is set by the fractional index difference Δ [defined in (2)]; any components with α larger than

$$\alpha_{\max} = \sqrt{2\Delta} \tag{7}$$

exceed the critical angle for total internal reflection and are unguided. Thus eqs. (7) and (6) taken together establish a maximum modal index $m_{\max}$, with $(m+1)_{\max}$ transverse extreme, controlled by a_e and Δ.

The specific group delay is

$$\tau = \frac{d\beta}{d\omega}, \tag{8}$$

where $\omega = 2\pi f$ is the angular frequency. The range of values τ can assume run from the value with $m = 0$ to $\tau_{\max}$ which is

$$\tau_{\max} = \tau\,|_{(m+1)_{\max}}. \tag{9}$$

It is convenient to compare the specific group delay for the guided mode τ to that for an infinite medium of index n_1 using

$$t = \left(\tau - \frac{n_1}{c}\right). \tag{10}$$

The work of Ref. 5 related specifically to two-dimensional waveguides. We extend this here to three-dimensional guides in Cartesian coordinates. We note the relation between the propagation constants has the form

$$\epsilon_r \beta_0^2 = \beta_z^2 + \beta_{x1}^2 + \beta_{x2}^2, \tag{11}$$

where ϵ_r is the dielectric constant, β_0 is the free-space phase constant, and β_z, β_{x1}, and β_{x2} are the longitudinal and transverse wave numbers respectively. In our case, β_{x1} and β_{x2} are small compared to β_z. The value of β_z depends only on the sum of the squares of the two transverse

wave numbers $\beta_{x1}^2 + \beta_{x2}^2$; further, for a square guide of width $2a_e$ we can write

$$\beta_{x1} 2a_e = (m_a + 1)\pi, \tag{12}$$

$$\beta_{x2} 2a_e = (m_b + 1)\pi. \tag{13}$$

Thus the total modal designation is

$$\rho = (m + 1) = \sqrt{(m_a + 1)^2 + (m_b + 1)^2} \tag{14}$$

for the square guide. This replaces $(m + 1)$ in the two-dimensional guide. The maximum $(m + 1)_{\max}$ can be reached with a variety of field distributions given by various values of m_a and m_b in (14).

To compute the impulse response we first note in solutions for specific group delay τ from (8) that all combinations of modes having the same m or ρ in (14) have the *same* specific group delay. Hence the fiber output at a given delay associated with ρ will be $P(\rho)$, for equal power into the fiber in each mode, where

$$P(\rho) = \frac{\pi\rho}{2} d\rho \tag{15}$$

and ρ is given by (14). The fiber impulse response as a function of time t is

$$\text{Output} = P(t) = \frac{P(\rho)}{dt} = \frac{\pi\rho}{2} \left| \frac{dt}{d\rho} \right|^{-1}. \tag{16}$$

The relation between t and ρ is obtained from (10) and (8). The range of t over which (16) is valid is found by inserting the minimum and maximum values that ρ may assume,

$$\rho_{\min} = \sqrt{2}, \tag{17}$$

$$\rho_{\max} = (m + 1)_{\max}. \tag{18}$$

For many purposes the value of $t_{\min}$ corresponding to $\rho_{\min}$ may be taken as zero since $\rho_{\max} \gg \rho_{\min}$; thus

$$t_{\min} \simeq 0, \tag{19}$$

$$t_{\max} = \left(\tau_{\max} - \frac{n_1}{c} \right). \tag{20}$$

The effect on transmission system performance is approximately the same for various shapes of the impulse response if the rms width is the same;[12] this applies in the region where the fiber impulse response degrades the system signal-power requirements by only 1 or 2 dB. We

find the desired second moment using

$$A = \int_0^{t_{\max}} P(t)dt, \tag{21}$$

$$T = \frac{1}{A}\int_0^{t_{\max}} tP(t)dt, \tag{22}$$

$$\sigma^2 = \frac{1}{A}\int_0^{t_{\max}} t^2P(t)dt - T^2. \tag{23}$$

III. A GENERAL SOLUTION

We now give the results for the index distribution according to eq. (1). The guide has width $2a$. We specify that at $x = a$, $f(x) = -\Delta$, leading to eq. (2) and

$$a_u = \frac{2(1-\delta)\Delta}{a^u}, \tag{24}$$

$$a_r = \frac{2\delta\Delta}{a^r}, \tag{25}$$

in which δ may be either positive or negative but $|\delta| \ll 1$. Using eq. (4), we find a_e,

$$\frac{1}{a_e^2} = \left\{\frac{n_1^2 a_u}{0.303b^4\lambda_0^2}\right\}^{2/(u+2)} + \frac{2a_r n_1^2}{(u+2)0.303b^4\lambda_0^2\left\{\dfrac{n_1^2 a_u}{0.303b^4\lambda_0^2}\right\}^{r/(u+2)}}. \tag{26}$$

This leads immediately to β,

$$\beta = \frac{2\pi n_1}{\lambda_0} - \frac{\pi(m+1)^2}{16(0.1515)^{2/(u+2)}\left(\dfrac{m+2.5}{2.5}\right)^{4/(u+2)}} \times \frac{[(1-\delta)\Delta]^{2/(u+2)}}{a^{2u/(u+2)}}\left(\frac{\lambda_0}{n_1}\right)^{(u-2)/(u+2)} - 2.592 \times \frac{(0.1515)^{r/(u+2)}(m+1)^2\Delta^{(u+2-r)/(u+2)}\delta}{(u+2)\left(\dfrac{m+2.5}{2.5}\right)^{(2u+4-2r)/(u+2)}(1-\delta)^{r/(u+2)}a^{2r/(u+2)}} \times \left(\frac{\lambda_0}{n_1}\right)^{(2r-u-2)/(u+2)}, \tag{27}$$

and to the number of propagating modes N,

$$N = \frac{4}{\pi}(0.7757)^{2/u}(n_1 ka)^2\Delta, \tag{28}$$

in which $k = 2\pi/\lambda_0$. The maximum modal index has two useful forms,

derived from (6) and (7),

$$\frac{(m+1)^2_{\max}}{(m+2.5)^{4/(u+2)}_{\max}} = 32(0.02424)^{2/(u+2)} \left(\frac{n_1 a}{\lambda_0}\right)^{2u/(u+2)} \Delta^{u/(u+2)}, \quad (29)$$

$$m_{\max} \simeq 5.657(0.7757)^{1/u} \left(\frac{n_1 a}{\lambda_0}\right) \Delta^{\frac{1}{2}}. \quad (30)$$

The specific group delay from (8) is

$$\tau = \frac{n_1}{c}\left\{1 + \frac{(u-2)}{(u+2)}\Delta(1-\delta)^{2/(u+2)} \frac{(m+1)^2}{(m+1)^2_{\max}} \frac{(m+2.5)^{4/(u+2)}_{\max}}{(m+2.5)^{4/(u+2)}} + Q\Delta \frac{\delta}{(1-\delta)^{r/(u+2)}} \frac{(m+1)^2}{(m+1)^{2r/u}_{\max}} \frac{(m+2.5)^{4r/n(n+2)}_{\max}}{(m+2.5)^{(2u+4-2r)/(u+2)}}\right\}, \quad (31)$$

where Q is

$$Q = 2.578(0.7757)^{r/u} \frac{(2r-u-2)}{(u+2)^2}. \quad (32)$$

We can simplify τ for $m \gg 1$ to

$$\tau = \frac{n_1}{c}\left\{1 + \frac{(u-2)}{(u+2)}\Delta(1-\delta)^{2/(u+2)} \left(\frac{m}{m_{\max}}\right)^{2u/(u+2)} + Q\Delta \frac{\delta}{(1-\delta)^{r/(u+2)}} \left(\frac{m}{m_{\max}}\right)^{2r/(u+2)}\right\}. \quad (33)$$

Using (21), (22), and (23) we find the impulse response is

$$P(t)|_{u\neq 2} = t^{2/u}\left\{\frac{\pi(u+2)^{(2u+2)/u} m^2_{\max}}{4u(u-2)^{(u+2)/u}\Delta^{(u+2)/u}(1-\delta)^{2/u}}\right\} - t^{(r+2-u)/u}\left\{\frac{\pi(u+2)^{(2+r+2u)/u} Q\delta m^2_{\max}}{4n^2(n-2)^{(2+r+u)/u}(1-\delta)^{(r+2)/n}\Delta^{(r+2)/n}}\right\} \quad (34)$$

and the rms width of the impulse response is

$$\sigma = \frac{n_1}{c}\Delta(u-2)\left(\frac{1}{3u+2} - \frac{(u+2)}{(2u+2)^2}\right)^{\frac{1}{2}} \quad (35)$$

for the case where $a_r = 0$. The minimum allowed value of t in (34) is

$$t_{\min} \simeq \frac{n_1\Delta}{c}\frac{(u-2)}{(u+2)}\frac{2}{(2.5)^{4/(u+2)}}\frac{(m+2.5)^{4/(u+2)}_{\max}}{(m+1)^2_{\max}}. \quad (36)$$

Equation (29) can be used to eliminate the last factor in (36). For the first term of (34), representing the major response due to $a_u x^u$ in (1), the maximum value of t is

$$t_{\max(u)} = \frac{n_1\Delta}{c}\frac{(u-2)}{(u+2)}. \quad (37)$$

For the second term of (34), representing the perturbing term $a_r x^r$ in (1), the maximum value of t is

$$t_{\max(r)} = \frac{n_1 \Delta}{c} \frac{\delta}{(1 - \delta)^{r/(u+2)}}. \tag{38}$$

Equation (34) is not valid for $u = 2$; for the later case the impulse response is

$$P(t)\,|_{u=2} = \frac{\pi}{r} t^{(4-r)/r} \left\{ \frac{4096(1 - \delta)^{r/4} m_{\max}^{r/2}}{32^{r/2}(41.254)^{(2-r)/2}(2r - 4)\delta\Delta} \right\}^{(2r-4)/r}. \tag{39}$$

IV. THE NEAR-PARABOLIC INDEX DISTRIBUTION

Letting $u = 2$ in the equations of the preceding section yields the near-parabolic index distribution. As shown by (35), the impulse response has zero width when $a_r = 0$ in the approximation made here. In the cylindrical fiber there is no index distribution which gives zero delay distortion among all the various modes. There is a distribution of index which minimizes the delay distortion, and we now evaluate this condition. We can use the above theory to evaluate the cylindrical waveguide by noting the two limiting conditions already known for low dispersion. Take the form

$$n = n_1(1 - \tfrac{1}{2}\alpha^2 R^2 + b_4 \alpha^4 R^4 + \cdots) \tag{40}$$

in which R is the radial transverse coordinate. In two earlier papers[6,13] it has been pointed out that the value of b_4 must be different to produce no dispersion for meridional rays versus skew rays; the difference in b_4 was found to be $\frac{1}{6}$. Kawakami and Nishizawa[6] found that b_4 must be $\frac{1}{2}$ to give no dispersion for skew rays and must be $\frac{1}{3}$ to give no dispersion for meridional rays. We can visualize minimizing the dispersion for one ray type, and thus experiencing a maximum dispersion corresponding to a change in b_4 of $\frac{1}{6}$. We do this in the approximation used here by setting $u = 2$, $r = 4$, and making

$$a_4 = -\tfrac{1}{3} a_2^2 \tag{41}$$

corresponding to no dispersion for meridional rays. From (24), (25), (31), and (41) we find for this "ideal" index distribution in the round fiber the specific group delay

$$\tau = \frac{n_1}{c} \left\{ 1 - 0.26 \frac{(m + 1)^2}{(m + 1)_{\max}^2} \Delta^2 \right\} \tag{42}$$

and

$$t_{\max} = -0.26 \frac{n_1}{c} \Delta^2. \tag{43}$$

This agrees reasonably well with the value $-n_1\Delta^2/8c$ arrived at by Gloge and Marcatili[10] using an entirely different analysis for an optimum index distribution defined differently.* The rms width of the impulse response corresponding to (41) and (42) is

$$\sigma = 0.752 \frac{n_1}{c} \Delta^2. \tag{44}$$

In contrast with this, the simple step-index fiber [represented by $a_r = 0$ and $n \to \infty$ in (1)] has an rms impulse response width of

$$\sigma = \frac{1}{\sqrt{12}} \frac{n_1}{c} \Delta. \tag{45}$$

Thus the "ideal," corresponding to (42) and (43), is smaller by a factor of 0.26Δ. Since $\Delta \simeq 0.01$ the ideal graded-index fiber has an impulse response narrower by a factor of about 400 than that for the conventional fiber.

The fourth-order term represented by (41) corresponds to

$$\delta = -\tfrac{2}{3}\Delta. \tag{46}$$

This is very little different from the simple parabola—not enough to see on Fig. 1 where curve ① represents *both* $\delta = 0$ and (46) for $\Delta \simeq 0.01$. More importantly, (46) and (41) imply that the fourth-order term decreases in size relative to the second-order term as Δ decreases. The inaccuracies of material processing are likely to prevent this as Δ becomes small. A more probable limit is a fixed value of δ in (25) as Δ changes. This results in a specific group delay

$$\tau = \frac{n_1}{c}\left\{1 + \frac{0.389\delta}{(1-\delta)} \frac{(m+1)^2}{(m+1)^2_{\max}} \Delta\right\} \tag{47}$$

and an rms width for the impulse response

$$\sigma = 0.112 \frac{n_1}{c} \Delta \frac{\delta}{(1-\delta)}. \tag{48}$$

---

* If instead of minimizing the delay distortion for either the skew rays or meridional rays we had minimized the delay distortion for an index equation (40) at the *mean* of the index values giving minimum distortion for the skew and meridional rays—i.e., at $b = 5/12$—then the maximum departure in b for any mode corresponds to a change in $b = \pm 1/12$. This corresponds to $a_4 = \pm 1/6a_2^2$, leading to

$$\tau = \frac{n_1}{c}\left\{1 \pm 0.13 \frac{(m+1)^2}{(m+1)^2_{\max}} \Delta^2\right\}$$

in place of (42). The coefficient 0.13 corresponds almost exactly to the result of Gloge and Marcatili,[10] but the present analysis indicates twice the total delay spread predicted by Gloge and Marcatili due to the $\pm$ sign.

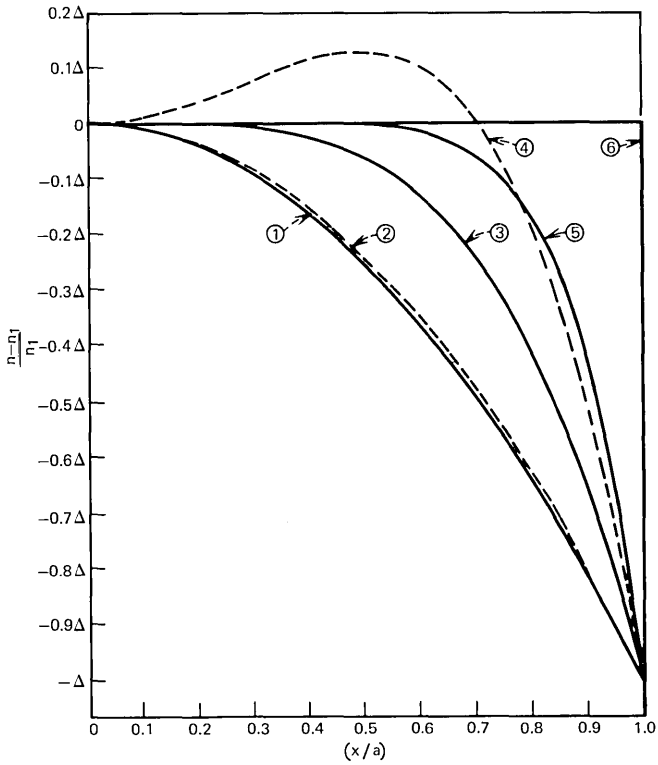

Fig. 1—Normalized index of refraction versus transverse coordinate (x/a) for the following parameters in eqs. (1) and (25): curve ①, $u = 2$, $\delta = 0$; curve ②, $u = 2$, $r = 4$, $\delta = 0.05$; curve ③, $u = 4$, $\delta = 0$; curve ④, $u = 4$, $r = 2$, $\delta = -1$; curve ⑤, $u = 8$, $\delta = 0$; curve ⑥, $u = \infty$, $\delta = 0$.

For $\delta = 0.05$, σ becomes 0.00591 $n_1\Delta/c$ which is narrower than for the step-index fiber by a factor of about 50 independent of Δ.

We note from (39) that the impulse response for the ideal index distribution perturbed by a fourth-order term ($r = 4$) is a rectangular pulse, shown as curve ② in Fig. 2. However, if the perturbation were sixth order, $r = 6$, the impulse response would vary as $t^{-\frac{1}{2}}$. Other values of r give other impulse-response shapes, which we discuss further in the next section.

Finally we note from (47) that the impulse response due to the fourth-order perturbation of the ideal distribution may either lead or lag the impulse at $\tau = n_1/c$, depending on whether δ is positive or negative [see (1) and (25)]. Similar effects due to the sign of δ are

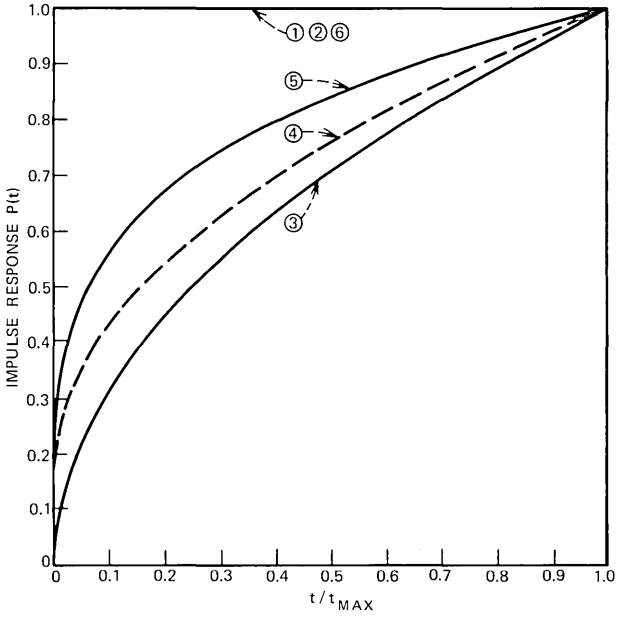

Fig. 2—Impulse response $P(t)$ versus $t/t_{\max}$ where $t = \tau - n_1/c$ and τ is given by (31) and (33). The conditions are the same as defined under Fig. 1.

found for perturbations of the nonideal index distributions and may be seen in (31).

In Section VI there is a discussion of Fig. 8 which shows the effects of perturbing the parabolic index distribution in several ways.

V. DISCUSSION OF THE INDEX DISTRIBUTION, $n = n_1(1\text{-}\frac{1}{2}a_u x^u)$

In this section we discuss the distributions obtained by setting $a_\tau = 0$ in (1) which mean $\delta = 0$ in (24) and (25).

The total spread in specific group delay for all modes $t_{\max,u}$ is given by (37), which gives a null value when $u = 2$. As already discussed, this is a simplification in the vicinity of the "ideal" distribution. However, (37) gives a valid representation as u departs significantly from the value 2. Figure 3 shows the variation in $t_{\max,u}$ versus u. The behavior in the vicinity of $u = 2$ is a form of singularity. For 5-percent error in u from the ideal, $t_{\max,u} \simeq 0.025 n_1 \Delta/c$. This compares with $0.0244 n_1 \Delta/c$ for the modal delay spread from (47) for 5-percent (at $x = a$) fourth-order perturbation of the ideal. We conclude that it is not particularly important how the ideal is perturbed.

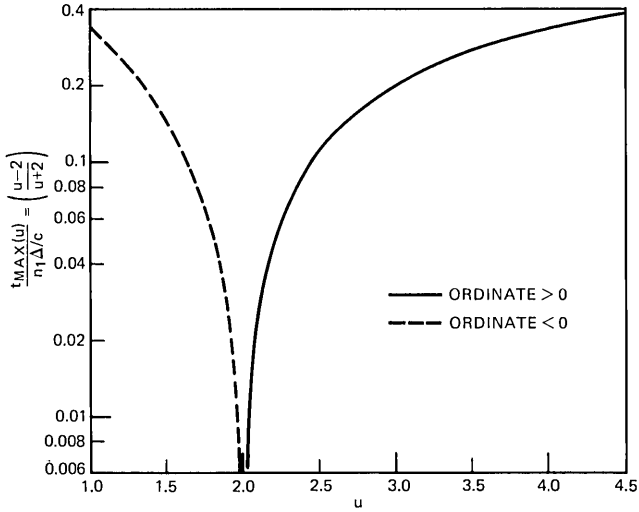

Fig. 3—Normalized modal group-delay spread versus u for $a_r = 0$ in eq. (1).

The value of $t_{\max,u}$ from (37) and plotted in Fig. 3 is not very sensitive to the value of u at values removed from $u = 2$. The reduction in rms width of the impulse response is illustrated in Fig. 4. For $u = \infty$ (the conventional step-index fiber) the value of $\sigma/(n_1\Delta/c)$ is $1/\sqrt{12}$ or 0.289. This is reduced by a factor of 2 for u near 6, and by a factor of 4 for u near 3.5. These results are identical to those of Gloge and Marcatili.[10]

The shape of the impulse response, given by (39) for $u = 2$ and by (34) for u away from 2, is plotted in Fig. 2 for several cases of interest.

The number of modes which can propagate, given by (28) for the square fiber, is plotted as a function of u^{-1} in Fig. 5. For comparison the corresponding quantity for the round fiber from Ref. 10 is also plotted. The ratio is $4/\pi$ at $u = \infty$, and near 2 for $u = 2$. The approximations made here are seen to be good, though not perfect.

VI. PERTURBATIONS OF THE GENERAL INDEX DISTRIBUTION

We discuss now some of the results for the perturbed index distributions, eq. (1). We recall the solutions have been obtained assuming $|a_r| \ll a_u$ or $|\delta| \ll 1$.

The solutions for the specific group delay, given in (31) and (33), contain the quantity Q as a factor in the perturbing term. The factor

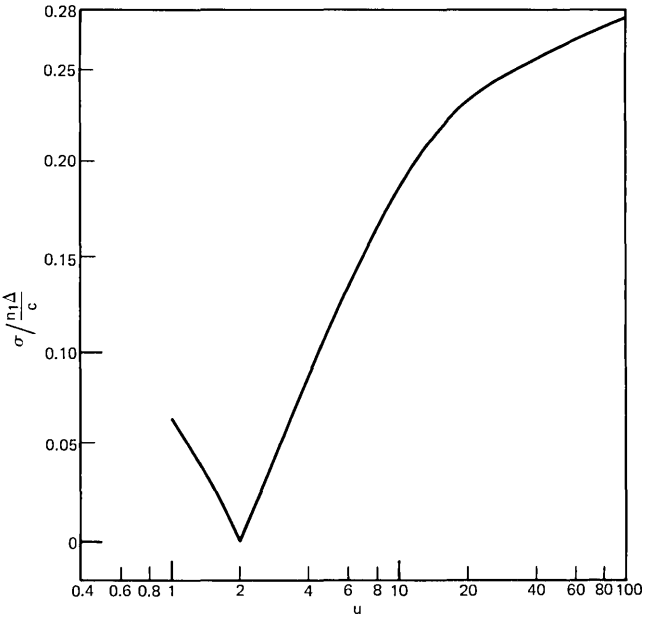

Fig. 4—Normalized rms width of the impulse response versus u for $a_r = 0$ in eq. (1).

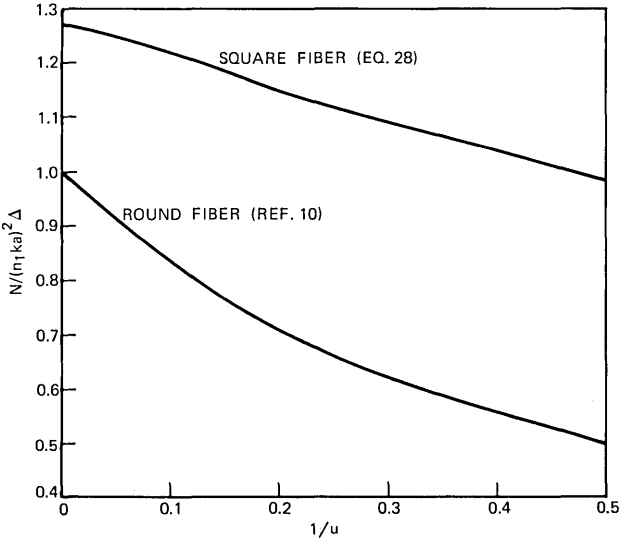

Fig. 5—Total number of propagating modes versus u^{-1} for $a_r = 0$ in eq. (1).

Q, given by (32), contains the principal effect of the exponents u and r on specific group delay. Figure 6 shows how Q varies with r for the special case $u = 2$. The region very near $r = 2$ is in question since we know the "ideal" index distribution differs slightly from $u = 2$. Elsewhere, the results should be significant and may be used in (34), (31), and (33).

More general curves for Q are given in Fig. 7. We observe that when $r > 0$ the maximum value of Q is not very dependent on u but the most sensitive region of r (giving largest Q) does depend somewhat on u. An intuitive feel for the changes in the index distribution which correspond to some of the curves in Fig. 7 can be obtained by examination of Fig. 8. Figure 8 shows the normalized index n versus transverse coordinate (x/a). The curve labeled $r = 2$ corresponds to the pure parabolic distribution. The other curves correspond to $\delta = 0.05$ with various values of r in the perturbing term and u always equal to 2.

We note in Fig. 7 that, at $u = 2$, the value of Q at $r = 10$ is much larger than at $r = \infty$. This may seem surprising, since a step-index change occurs at $(x/a) = 1.0$ when $r = \infty$. Below $(x/a) = 1$ the $r = \infty$ curve in Fig. 8 corresponds to a pure parabolic gradient between the ordinate equal zero and -0.95Δ.

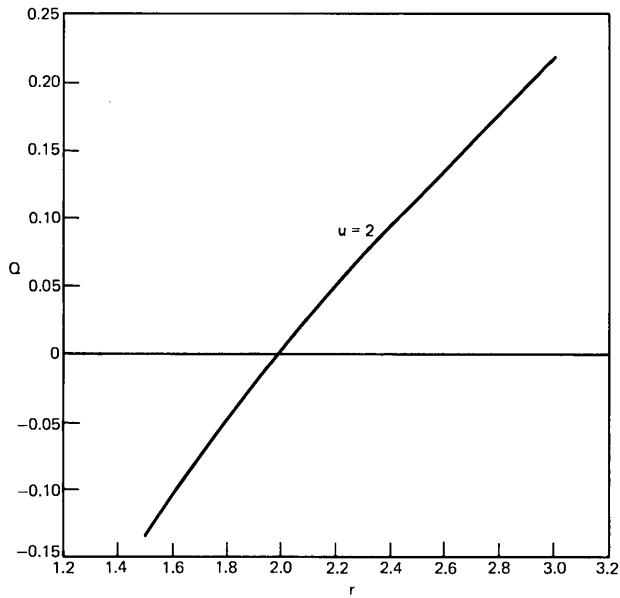

Fig. 6—Q versus r for the near-parabolic index distribution.

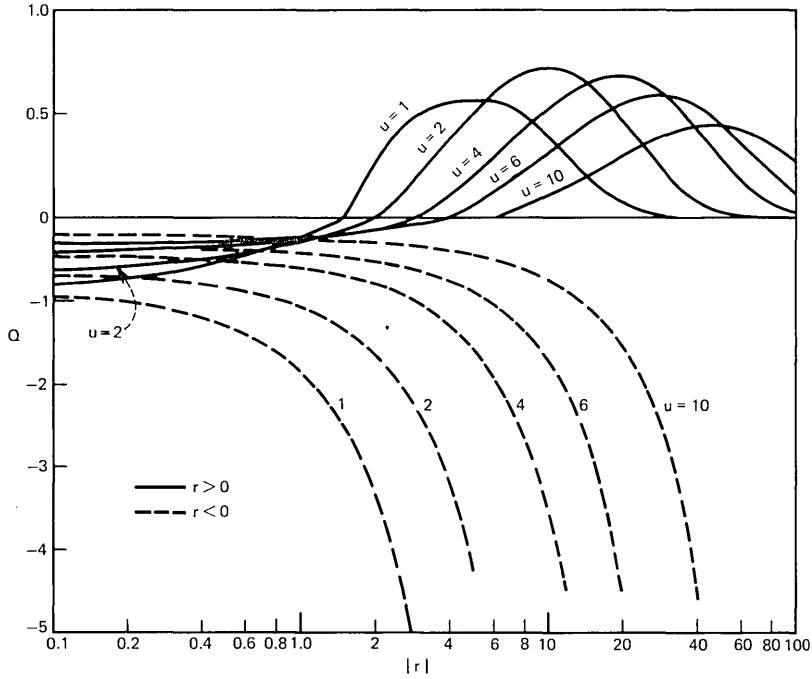

Fig. 7—Q versus $|r|$ with u as a parameter.

We also note in Figs. 8 and 7 that dips in the index distribution near (x/a) equal zero (curves for $r = -0.4$ and -0.1) yield values of $|Q|$ comparable to those for r in the range 4 to 20.

VII. CONCLUSION

The above analysis provides an approximate solution for the delay distortion to be expected in a wide variety of graded-index fibers, representable by (1) with $|a_r| \ll a_u$.

In general, gradual tapering of the index between the center of the fiber and the outer support provides reduced delay distortion. Only in the vicinity of the near-parabolic distribution is the performance highly sensitive to the exact index distribution. The "ideal" near-parabolic distribution provides a potential reduction in delay distortion of several hundred times compared to the step-index distribution of the conventional clad fiber. With an accuracy of the order of 5 percent in achieving the "ideal" distribution, the reduction in delay distortion is on the order of 50.

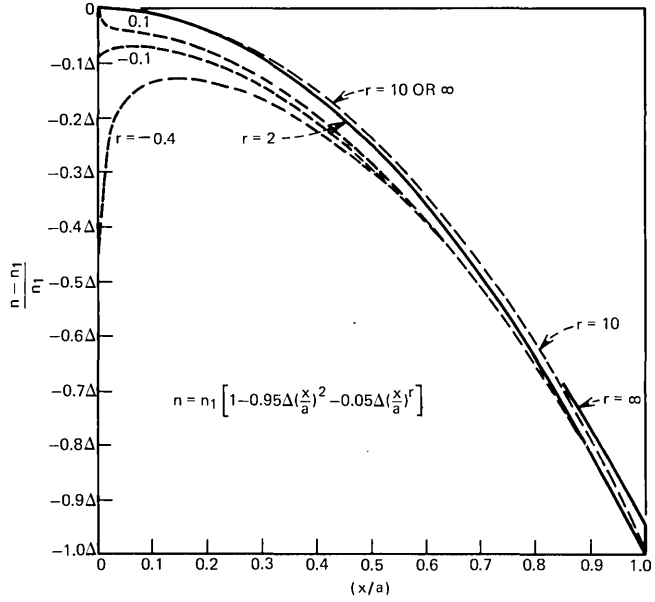

Fig. 8—Normalized index versus transverse coordinate (x/a) for several perturbations of the parabolic index distribution.

VIII. ACKNOWLEDGMENTS

It is a pleasure to acknowledge fruitful discussions with E. A. J. Marcatili and D. Gloge.

REFERENCES

1. S. E. Miller, U. S. Patent No. 3,434,774, "Waveguide for Millimeter and Optical Waves," issued March 25, 1969.
2. Nippon Selfoc Co., Ltd., "Light-Conducting Glass Fibers or Fiber Structures and Production Thereof," Japanese Patent No. 1,266,521, issued March 8, 1972.
3. T. Uchida, M. Furukawa, I. Kitano, K. Koizumi, and H. Matsumura, "A Light Focusing Fiber," IEEE J. Quan. Elec. (Digest of Technical Papers), *QE-5*, June 1969, pp. 25.
4. E. A. J. Marcatili, "Modes in a Sequence of Thick Astigmatic Lens-Like Focusers," B.S.T.J., *43*, No. 6 (November 1964), pp. 2887–2904.
5. S. E. Miller, "Light Propagation in Generalized Lens-Like Media," B.S.T.J., *44*, No. 9 (November 1965), pp. 2017–2063.
6. S. Kawakami and T. Nishizawa, "An Optical Waveguide with the Optimum Distribution of the Refractive Index with Reference to Waveform Distortion," IEEE Trans. Microwave Theory and Tech., *MTT-16*, October 1968, pp. 814–818.
7. D. Gloge, E. L. Chinnock, and K. Koizumi, "Study of Pulse Distortion in Selfoc Fibers," Elec. Letters, *8*, October 1972, pp. 526–527.
8. C. A. Burrus, E. L. Chinnock, D. Gloge, W. S. Holden, T. Li, R. D. Standley, and D. B. Keck, "Pulse Dispersion and Refractive-Index Profiles of Some

Low-Loss Multimode Optical Fibers," Proc. IEEE, *61*, October 1973, pp. 1498–1499.
9. E. L. Chinnock, L. G. Cohen, W. S. Holden, R. D. Standley, and D. B. Keck, "The Length Dependence of Pulse Spreading in the CGW-Bell-10 Optical Fiber," Proc. IEEE, *61*, October 1973, pp. 1499–1500.
10. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.
11. D. Marcuse, "The Impulse Response of an Optical Fiber With Parabolic Index Profile," B.S.T.J., *52*, No. 7 (September 1973), pp. 1169–1174.
12. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 843–886.
13. E. G. Rawson, D. R. Herriott, and J. McKenna, "Analysis of Refractive-Index Distributions in Cylindrical Graded-Index Glass Rods Used as Image Relays," Appl. Opt., *9*, March 1970, pp. 753–759.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Losses and Impulse Response in Parabolic Index Fibers With Square Cross Section

By D. MARCUSE

(Manuscript received August 13, 1973)

Mode coupling is studied in a parabolic index fiber with a lossy boundary and square cross section. Statistical deviations of the fiber axis from perfect straightness and random changes of its width are considered as causing mode coupling. The excess loss caused by these mode coupling mechanisms and the loss penalty incurred for a certain degree of narrowing of the impulse response are estimated.

I. INTRODUCTION

Multimode optical fibers whose cores have parabolic distributions of the refractive index,[1–3]

$$n = n_0 \left(1 - \frac{r^2}{a^2} \bar{\Delta}\right), \tag{1}$$

are of great practical interest for light transmission over long distances, since their delay distortion is much less serious than that of conventional clad fibers.

Since no optical fiber can ever be produced free of random imperfections, it is important to know how statistical irregularities of the fiber affect its performance. Random irregularities of the fiber axis and random changes of the effective width of the fiber cause coupling among its modes. The mode losses are functions of the mode number. Absorption losses tend to affect all modes in the same way. However, if we assume that the fiber boundary either consists of an absorptive material or is a rough surface that scatters light, we must expect that higher order modes, whose fields reach strongly into the neighborhood of the fiber boundary, suffer much higher losses than lower order modes that are confined to the vicinity of the fiber axis. Coupling of the low-order modes to the high-loss, high-order modes increases the

overall waveguide losses. One objective of our study of the effect of waveguide irregularities is thus the determination of the excess losses caused by mode coupling.

The second objective of this study of waveguide irregularities consists in determining the impulse response of the fiber. In the absence of coupling, each mode transports a fraction of the total power at its characteristic group velocity. Since the group velocities of different mode groups are not identical, pulse distortion results.[2,3] Mode coupling has the beneficial effect of improving the impulse response of the fiber. It is thus of interest to determine how much reduction of multimode pulse distortion can be achieved by random bends and random width changes of the fiber.

The effect of random bends on parabolic index fibers with circular cross section has been estimated in an earlier paper.[4] Pure diameter changes of a fiber with circular cross section leave modes with different circumferential symmetries uncoupled. Statistical irregularities are unlikely to result in pure diameter changes without distorting the circular fiber cross section. However, an analysis of more general distortions of a fiber with nominally circular cross section is difficult to perform. For this reason we discuss a fiber with parabolic index distribution (1) but with square cross section. It seems reasonable to expect that the performance of a fiber with square cross section is similar to that of a fiber with circular cross section. We expect to find the correct order of magnitude of the losses and impulse response of the round fiber by examining its close relative, the fiber with square cross section. In particular, it should be possible to assess the relative effect of random axis deformations as compared to random width changes. In a square fiber, changes of only one set of opposing walls leave groups of modes uncoupled from each other. This situation corresponds to the circular fiber with pure diameter changes that leave modes of different azimuthal symmetry uncoupled. By allowing both sets of opposing walls to change their separation randomly, we are sure that all modes are coupled to each other. This model corresponds to a nominally round fiber whose cross section is deformed in an arbitrary way that does not conserve the circular symmetry.

II. THE MODES OF THE PERFECT FIBER WITH SQUARE CROSS SECTION

The modes of an infinitely extended medium with the distribution

$$n^2 = n_0^2 \left(1 - 2\frac{r^2}{a^2}\bar{\Delta}\right) \tag{2}$$

of the square of the refractive index have the form[5]

$$E_{pq} = \frac{2\left(\sqrt{\frac{\mu_0}{\epsilon_0}}P\right)^{\frac{1}{2}} H_p\left(\sqrt{2}\,\frac{x}{w}\right) H_q\left(\sqrt{2}\,\frac{y}{w}\right) e^{-r^2/w^2}}{(n_0\pi 2^{p+q}p!\,q!)^{\frac{1}{2}}w}\, e^{-i\beta z}. \tag{3}$$

It is

$$r^2 = x^2 + y^2. \tag{4}$$

The parameter a is an arbitrary constant that, in conjunction with $\bar{\Delta}$, determines the transverse dependence of the refractive index distribution. However, in the round fiber it is convenient to associate a with the radius of the fiber boundary so that $\bar{\Delta}$ is the relative difference between the values of the refractive index on axis and at the fiber boundary.

The square of the refractive index distribution (2) does not follow precisely from (1). However, if one equation is regarded as the precise distribution of the corresponding quantity, the other holds approximately provided $\bar{\Delta}$ is small and we limit r to the range $r \leqq a$. H_p and H_q are Hermite polynomials of degree p and q, and P is the power carried by the mode. The parameter ω is defined as[5] ($k = \omega\sqrt{\epsilon_0\mu_0}$)

$$w = \left(\frac{\sqrt{2}a}{n_0 k\sqrt{\bar{\Delta}}}\right)^{\frac{1}{2}} \tag{5}$$

and determines the radius of the field distribution with $p = q = 0$. At $r = w$ the field has decayed to $1/e$ of its value on axis. E_{pq} represents the transverse component of the electric field vector. The longitudinal field components are relatively much weaker and are not being considered. The field (3) is only an approximate solution of Maxwell's equations. The propagation constant of the mode is given as[5]

$$\beta = \beta_{pq} = n_0 k\left[1 - \frac{2\sqrt{2\bar{\Delta}}}{n_0 ka}(p + q + 1)\right]^{\frac{1}{2}}. \tag{6}$$

The modes of the square-law medium are mutually orthogonal and satisfy the relation

$$\frac{\beta}{2k}\sqrt{\frac{\epsilon_0}{\mu_0}}\int_{-\infty}^{\infty} E_{pq}E^*_{p'q'}dxdy = P\delta_{pp'}\delta_{qq'}. \tag{7}$$

So far, the fiber boundary has been ignored. The mode field (3) is an (approximate) solution of the guided-wave problem if we assume that the distribution (2) extends to infinity. However, each mode decays very rapidly outside of a certain region. For a given value of p the

field oscillates as a function of x passing through p zero crossings. The shape of the function

$$H_4\left(\sqrt{2}\,\frac{x}{w}\right) e^{-x^2/w^2} \tag{8}$$

is shown in Fig. 1. At the point (see appendix)

$$x = x' = w\sqrt{p + \tfrac{1}{2}} \tag{9}$$

the oscillatory behavior of the function changes to a rapid decay. If $x' < a$ the presence of the wall does not interfere appreciably with the field distribution. However, if $x' > a$ the field distribution is severely altered by the presence of the wall. Since we are assuming that the interaction of the field with the wall causes power dissipation either by absorption or by radiation, we consider those modes whose fields reach the vicinity of the wall with high field intensity as being effectively cut off. By replacing x' in (9) with a we obtain the condition for the maximum value of p that can be allowed for low-loss modes.

$$p_c = \left(\frac{a}{w}\right)^2 - \frac{1}{2} = n_0 ka\sqrt{\frac{\Delta}{2}} - \frac{1}{2}. \tag{10}$$

Since we are assuming that the boundary of the fiber has a square cross

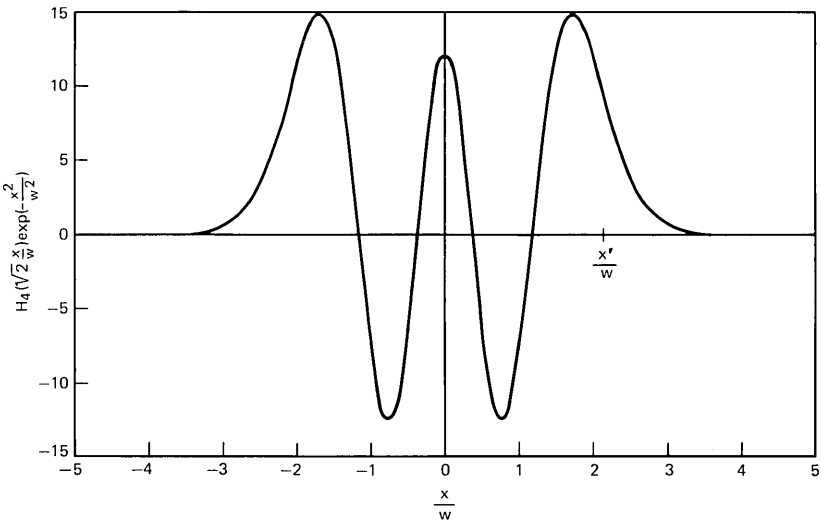

Fig. 1—Plot of the function given by eq. (8).

section, we also must impose the same "cutoff" condition on q,

$$q_c = \left(\frac{a}{w}\right)^2 - \frac{1}{2}. \tag{11}$$

We use the modes (3) of the infinite square-law medium (2) to describe the modes of the fiber with square boundary if $p < p_c$ and $q < q_c$. If either $p > p_c$ and/or $q > q_c$ we regard the modes as so lossy that they are effectively cut off. This procedure is an approximation, but it allows us to obtain estimates (whose errors are unknown) to a complicated problem.

III. COUPLING COEFFICIENTS

The coupling coefficients between two modes are defined by the general expression[6,7]

$$K_{pq,p'q'} = \frac{\omega\epsilon_0}{4iP}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy(\bar{n}^2 - n^2)E_{pq}E^*_{p'q'}. \tag{12}$$

A square-law fiber with random axis deformations can be described by the following distribution of the square of its refractive index.

$$\bar{n}^2 = n_0^2\left\{1 - \frac{2}{a^2}[(x - f)^2 + (y - g)^2]\bar{\Delta}\right\}. \tag{13}$$

We assume that f and g are both random functions of z and that f/a and g/a are small quantities. The deflection of the optical axis of the square-law medium has a far more important effect on the modes than the corresponding deviation of the fiber boundary. The deflection of the fiber boundary that results from the random bends of its axis is neglected.

Substitution of (2), (3), and (13) into (12) results in

$$K_{pq,p\pm1,q} = \frac{n_0 kw\bar{\Delta}}{ia^2}\sqrt{p + \tfrac{1}{0}}f(z) \tag{14}$$

and

$$K_{pq,p,q\pm1} = \frac{n_0 kw\bar{\Delta}}{ia^2}\sqrt{q + \tfrac{1}{0}}g(z). \tag{15}$$

All other coupling coefficients vanish. The two choices (1 or 0) that are indicated under the square-root sign in (14) and (15) belong to the corresponding upper or lower sign of the subscript on the left-hand side. Random deformations of the fiber axis couple only neighboring modes.

For random width changes of the fiber we use the distribution

$$\bar{n}^2 = n_0^2 \left\{ 1 - 2 \left[\frac{x^2}{(a+f)^2} + \frac{y^2}{(a+g)^2} \right] \bar{\Delta} \right\}. \tag{16}$$

For small values of f/a and g/a we can write this expression approximately as follows:

$$\bar{n}^2 = n_0^2 \left\{ 1 - \frac{2}{a^2} \left[x^2 \left(1 - 2\frac{f}{a} \right) + y^2 \left(1 - 2\frac{g}{a} \right) \right] \bar{\Delta} \right\}. \tag{17}$$

For random changes of the width of the guide we obtain from (2), (3), (12), and (17)

$$K_{pq,p\pm2,q} = \frac{n_0 k w^2 \bar{\Delta}}{2ia^3} \sqrt{(p \pm 1)(p + \tfrac{2}{0})} f(z) \tag{18}$$

and

$$K_{pq,p,q\pm2} = \frac{n_0 k w^2 \bar{\Delta}}{2ia^3} \sqrt{(q \pm 1)(q + \tfrac{2}{0})} g(z). \tag{19}$$

All other coupling coefficients vanish. There are nonvanishing diagonal elements in this case. However, diagonal elements couple each mode only to itself. This self-coupling is of no importance if $f(z)$ and $g(z)$ have Fourier spectra with no zero (spatial) frequency component.

IV. COUPLED POWER THEORY

Mode coupling in waveguides with random irregularities can be described by coupled power equations.[8]

$$\frac{\partial P_\nu}{\partial z} + \frac{1}{v_\nu} \frac{\partial P_\nu}{\partial t} = -\alpha_\nu P_\nu + \sum_{\mu=1}^{N} h_{\nu\mu}(P_\mu - P_\nu). \tag{20}$$

P_ν is the average power carried by the mode labeled ν, v_ν is its group velocity, and α_ν its power loss coefficient. The mode label ν is used as an abbreviation for the set of labels p, q. The power coupling coefficient $h_{\nu\mu}$ is defined as follows:[8]

$$h_{\nu\mu} = |\hat{K}_{\nu\mu}| F(\beta_\nu - \beta_\mu). \tag{21}$$

The coefficient $\hat{K}_{\nu\mu}$ is the factor of the function $f(z)$ or $g(z)$ appearing in eqs. (14), (15), (18), and (19). The spatial power spectrum of the function $f(z)$ [or $g(z)$] is defined as

$$F(\theta) = \frac{1}{L} \left\langle \left| \int_0^L f(z) e^{-i\theta z} dz \right|^2 \right\rangle. \tag{22}$$

It is assumed that $L \to \infty$ in (22). The symbol $\langle\ \rangle$ indicates an ensemble average.

Since the random processes considered here tend to couple each mode only to one of its neighbors on either side (in mode label space) the equation system (20) can be converted to a partial differential equation whose variables are not only the length coordinate z and time t but in addition the two mode labels p and q.[9,10] If the number of modes below the effective cutoff value is very large, the set of discrete modes can be regarded as a quasicontinuum. We write

$$\begin{aligned}
&\sum_{p',q'} h_{pq,p'q'}(P_{p'q'} - P_{pq}) \\
&\quad = h_{pq,p+\Delta p,q}(P_{p+\Delta p,q} - P_{pq}) + h_{pq,p-\Delta p,q}(P_{p-\Delta p,q} - P_{pq}) \\
&\qquad + h_{pq,p,q+\Delta q}(P_{p,q+\Delta q} - P_{pq}) + h_{pq,p,q-\Delta q}(P_{p,q-\Delta q} - P_{pq}) \\
&\quad \approx (\Delta p)^2 \frac{\partial}{\partial p}\left[h(p)\frac{\partial P}{\partial p}\right] + (\Delta q)^2 \frac{\partial}{\partial q}\left[h(q)\frac{\partial P}{\partial q}\right]. \qquad (23)
\end{aligned}$$

The last step follows by considering the discrete mode labels as continuous variables and replacing differences by differentials. The notation $h(p)$ and $h(q)$ serves as a reminder that, according to (14) and (15), the coupling coefficients depend only on p if q is held fixed, and (18) and (19) show that they depend only on q if p is held fixed. We thus obtain the approximate partial differential equation

$$\frac{\partial P}{\partial z} + \frac{1}{v}\frac{\partial P}{\partial t} = + (\Delta p)^2 \frac{\partial}{\partial p}\left[h(p)\frac{\partial P}{\partial p}\right] + (\Delta q)^2 \frac{\partial}{\partial q}\left[h(q)\frac{\partial P}{\partial q}\right]. \qquad (24)$$

The average mode power P is now regarded as a continuous function of z, t, p, and q. The group velocity v is a function of p and q. We have omitted the loss term. We consider the modes as lossless if the variables p and q remain below the cutoff values (10) and (11) and as having infinitely high loss if cutoff is exceeded. This fact can be incorporated into the theory as a boundary condition by requiring

$$P(p_c, q_c) = 0. \qquad (25)$$

It has been shown in Ref. 8 how the pulse propagation problem can be solved by means of a perturbation method if the solutions of (24) for the time-independent case are known. We thus consider the trial solution

$$P(z, t, p, q) = U(p)V(q)e^{-\sigma z} \qquad (26)$$

and obtain by substitution into (24)

$$(\Delta p)^2 \frac{1}{U} \frac{\partial}{\partial p} \left[h(p) \frac{\partial U}{\partial p} \right] + (\Delta q)^2 \frac{1}{V} \frac{\partial}{\partial q} \left[h(q) \frac{\partial V}{\partial q} \right] + \sigma = 0. \quad (27)$$

We separate this equation into two ordinary differential equations by introducing the separation constant κ^2:

$$\frac{d}{dp} \left[h(p) \frac{dU}{dp} \right] + \frac{\kappa^2}{(\Delta p)^2} U = 0 \quad (28)$$

and

$$\frac{d}{dq} \left[h(q) \frac{dV}{dq} \right] + \frac{\sigma - \kappa^2}{(\Delta q)^2} V = 0. \quad (29)$$

V. CALCULATION OF THE STEADY-STATE POWER LOSS

The equation system (28) and (29) together with the boundary condition (25) (and an additional one to be discussed later) defines an eigenvalue problem. The lowest order eigenvalue σ_{11} has the physical meaning of the steady-state loss of the statistical power distribution.[8] This quantity is of interest since it determines the additional losses that are caused by the statistical irregularities of the fiber.

For random deformations of the fiber axis we obtain the power coupling coefficient $h(p)$ from (14) and (21).

$$h(p) = K(\Omega)p \quad (30)$$

with

$$K(\Omega) = \left(\frac{n_0 k w \bar{\Delta}}{a^2} \right)^2 F(\Omega). \quad (31)$$

We assume that $f(z)$ and $g(z)$ have identical power spectra so that $h(q)$ follows from $h(p)$ by replacing p with q. According to (6) the difference of the propagation constants of adjacent modes can be approximated as

$$\beta_{p+1,q} - \beta_{pq} = \Omega \approx \frac{\sqrt{2\bar{\Delta}}}{a}. \quad (32)$$

This approximation is independent of the mode numbers. This means that only one spatial frequency (or actually a very narrow range of spatial frequencies) of the power spectrum $F(\Omega)$ is responsible for mode coupling. For random axis deformations we have

$$\Delta p = \Delta q = 1 \quad (33)$$

so that we must solve the differential equation

$$\frac{d}{dp}\left[p\frac{dU}{dp}\right] + \frac{\kappa^2}{K(\Omega)}U = 0. \tag{34}$$

Its solution is a Bessel function of zero order,

$$U(p) = J_0\left(2\frac{\kappa}{\sqrt{K(\Omega)}}\sqrt{p}\right). \tag{35}$$

The choice of the Bessel function instead of a Neumann function, that would also solve (34), is dictated by an additional boundary condition. Since the partial differential equation (24) can be regarded as a diffusion process, we must require that no power diffuses into the lowest order mode $p = 0$ from negative values of p. This requirement means that $\partial P/\partial p = 0$ at $p = 0$. The solution (35) satisfies this condition. The solution of (29) is similarly

$$V(q) = J_0\left(2\sqrt{\frac{\sigma - \kappa^2}{K(\Omega)}}\sqrt{q}\right). \tag{36}$$

The boundary condition (25) leads to

$$2\frac{\kappa}{\sqrt{K(\Omega)}}\sqrt{p_c} = u_\nu \tag{37}$$

and

$$2\sqrt{\frac{\sigma - \kappa^2}{K(\Omega)}}\sqrt{q_c} = u_\mu. \tag{38}$$

The roots u_ν and u_μ are defined as solutions of the equation

$$J_0(u_\nu) = 0. \tag{39}$$

Since the eigenvalues depend on the labels ν and μ, we attach these labels to σ and obtain from (37) and (38) (note, $p_c = q_c$)

$$\sigma_{\nu\mu} = \frac{K(\Omega)}{4p_c}(u_\nu^2 + u_\mu^2). \tag{40}$$

The steady-state power loss, the lowest order eigenvalue σ_{11}, follows from (5), (10) (neglecting the term $\frac{1}{2}$), (31), and $u_1 = 2.405$

$$\text{axis deformation:} \quad \sigma_{11} = 5.78\frac{\bar{\Delta}}{a^4}F(\Omega). \tag{41}$$

We have thus rederived the loss formula (42) of Ref. 4.

For random diameter changes we obtain from (18), (21), and (31), considering that the spacing (in β-space) between adjacent coupled modes is now twice as large,

$$h(p) = \frac{1}{4}\frac{w^2}{a^2}K(2\Omega)p^2. \tag{42}$$

Since the number of modes is assumed to be large, we have used the approximation $p(p-1) \approx p^2$. With $\Delta p = \Delta q = 2$ we obtain from (28) and (42)

$$\frac{d}{dp}\left[p^2\frac{dU}{dp}\right] + \frac{a^2}{w^2}\frac{\kappa^2}{K(2\Omega)}U = 0. \tag{43}$$

The solution of this differential equation is

$$U = \frac{1}{\sqrt{p}}\cos(\rho_\nu \ln p + \phi_\nu) \tag{44}$$

with

$$\rho_\nu = \sqrt{\frac{a^2}{w^2}\frac{\kappa^2}{K(2\Omega)} - \frac{1}{4}}. \tag{45}$$

The solution of (29) is correspondingly

$$V = \frac{1}{\sqrt{q}}\cos(\rho_\mu \ln q + \phi_\mu) \tag{46}$$

with

$$\rho_\mu = \sqrt{\frac{a^2}{w^2}\frac{\sigma - \kappa^2}{K(2\Omega)} - \frac{1}{4}}. \tag{47}$$

These solutions have a singular behavior at $p = 0$ or $q = 0$. However, we must keep in mind that p and q are really discrete quantities. Considering them as continuous variables is an approximate procedure that can work only for very large values of p or q where the relative difference between adjacent discrete values becomes small. Since $\ln p = 0$ for $p = 1$, we allow p and q to vary only between 1 and p_c. The requirement that no power diffuses across the lower limit of the range of the variables imposes the conditions

$$\left(\frac{dU}{dp}\right)_{p=1} = 0$$

and (48)

$$\left(\frac{dV}{dq}\right)_{q=1} = 0.$$

These conditions lead to the determination of the phase terms via

the equations

$$\tan \phi_\nu = -\frac{1}{2\rho_\nu} \tag{49}$$

and

$$\tan \phi_\mu = -\frac{1}{2\rho_\mu}. \tag{50}$$

The boundary condition (25) leads to

$$\rho_\nu = \frac{1}{\ln p_c}\left[(2\nu - 1)\frac{\pi}{2} + \arctan\frac{1}{2\rho_\nu}\right]. \tag{51}$$

ρ_ν as well as ρ_μ are solutions of this equation with integer values of ν. Since the values of ρ_ν are now known, we obtain the eigenvalue $\sigma_{\nu\mu}$ from (45) and (47)

$$\sigma_{\nu\mu} = \frac{w^2}{a^2} K(2\Omega)(\rho_\nu^2 + \rho_\mu^2 + \frac{1}{2}). \tag{52}$$

For the lowest order eigenvalue, that is, for the steady-state power loss coefficient, we obtain with the help of (5) and (31)

$$\text{width changes:} \quad \sigma_{11} = (4\rho_1^2 + 1)\frac{\bar{\Delta}}{a^4} F(2\Omega). \tag{53}$$

The solution of (51) is not a constant. It depends on the waveguide parameters through its dependence on $p_c = (a/w)^2$. A plot of ρ_1 as a function of p_c is shown in Fig. 2.

The forms of the steady-state power loss coefficients (41) and (53) are very similar. The power spectra describe the deflection of the fiber axis from its nominally straight position or the changes of the width of one of the transverse fiber dimensions. However, the excess loss caused by random changes of the width of the fiber depends on a component of the power spectrum at twice the spatial frequency compared to the excess loss for random bends of the fiber axis.

VI. CALCULATION OF THE PULSE WIDTH

The width of the impulse response of a multimode fiber that is long enough for the steady-state distribution to establish itself is given by the formula:[8]

$$\Delta t = 4\left\{L \sum_{\nu,\mu}^{N} \frac{(G_{11}, VG_{\nu\mu})^2}{\sigma_{\nu\mu} - \sigma_{11}}\right\}^{\frac{1}{2}}. \tag{54}$$

The term with $\sigma_{\nu\mu} = \sigma_{11}$ is excluded from the sum. The functions

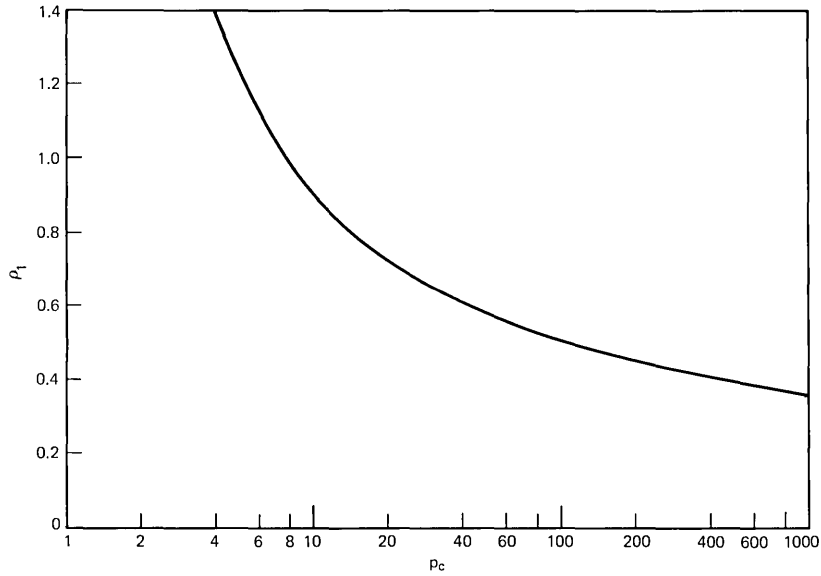

Fig. 2—Plot of the parameter ρ_1 as a function of p_c.

$G_{\nu\mu}$ are defined as

$$G_{\nu\mu} = A_{\nu\mu}U_\nu(p)V_\mu(q) \tag{55}$$

and V is by definition the difference between the inverse group velocity of the modes minus the inverse of the maximum group velocity.[4]

$$V = \frac{1}{v(p, q)} - \frac{n_0}{c} = \frac{\bar{\Delta}}{cn_0k^2a^2}(p + q)^2. \tag{56}$$

The expression in parenthesis is an abbreviated way of writing

$$(G_{11}, VG_{\nu\mu}) = \int dp \int dq G_{11}VG_{\nu\mu}. \tag{57}$$

The integrals extend over the entire range of p and q variables from either 0 or 1 to the cutoff value $p_c = q_c$. Requiring the normalization

$$\int dp \int dq G_{\nu\mu}^2 = 1, \tag{58}$$

we have for random axis deformations:

$$G_{\nu\mu} = \frac{w^2}{a^2} \frac{J_0\left(u_\nu\sqrt{\frac{p}{p_c}}\right) J_0\left(u_\mu\sqrt{\frac{q}{q_c}}\right)}{J_1(u_\nu)J_1(u_\mu)} \tag{59}$$

and for random width changes:

$$G_{\nu\mu} = \frac{A_{\nu\mu}}{\sqrt{pq}} \cos\left[\rho_\nu \ln p + \phi_\nu\right] \cos\left[\rho_\mu \ln q + \phi_\mu\right] \tag{60}$$

with

$$A_{\nu\mu} = 2\left\{\left[\ln p_c + \frac{2}{1+4\rho_\nu^2}\right]\left[\ln q_c + \frac{2}{1+4\rho_\mu^2}\right]\right\}^{-\frac{1}{2}}. \tag{61}$$

The integrals (57) have the following solutions. For axis deformations ($\mu \neq 1$):

$$\begin{aligned}(G_{11}, VG_{1\mu}) &= (G_{11}, VG_{\mu 1}) \\ &= \frac{32\bar{\Delta}p_c^2 u_1 u_\mu}{cn_0k^2a^2(u_1^2-u_\mu^2)^2}\left\{\frac{2u_1^2-1}{3u_1^2} - \frac{6(u_1^2+u_\mu^2)}{(u_1^2-u_\mu^2)^2}\right\}\end{aligned} \tag{62}$$

and for $\nu, \mu \neq 1$

$$(G_{11}, VG_{\nu\mu}) = \frac{2^7\bar{\Delta}}{cn_0k^2a^2}\frac{u_1u_\nu}{(u_1^2-u_\nu^2)^2}\frac{u_1u_\mu}{(u_1^2-u_\mu^2)^2}p_c^2. \tag{63}$$

For random width changes ($\mu \neq 1$):

$$\begin{aligned}(G_{11}, VG_{1\mu}) = (G_{11}, VG_{\mu 1}) &= -4\frac{\bar{\Delta}}{cn_0k^2a^2}\rho_1\rho_\mu p_c^2 A_{11}A_{1\mu} \\ &\times\left\{\frac{2\rho_1^2\left(1-\frac{1}{p_c}\right)}{(1+4\rho_1^2)[1+(\rho_1+\rho_\mu)^2][1+(\rho_1-\rho_\mu)^2]}\right. \\ &\times\left[(-1)^\mu + \frac{1+2(\rho_1^2+\rho_\mu^2)}{p_c\sqrt{(1+4\rho_1^2)(1+4\rho_\mu^2)}}\right] \\ &+ \frac{1}{A_{11}[4+(\rho_1+\rho_\mu)^2][4+(\rho_1-\rho_\mu)^2]} \\ &\left.\times\left[(-1)^\mu + \frac{5+2(\rho_1^2+\rho_\mu^2)}{p_c^2\sqrt{(1+4\rho_1^2)(1+4\rho_\mu^2)}}\right]\right\}\end{aligned} \tag{64}$$

and for $\nu, \mu \neq 1$

$$\begin{aligned}(G_{11}, VG_{\nu\mu}) &= 8\frac{\bar{\Delta}}{cn_0k^2a^2}p_c^2A_{11}A_{\nu\mu} \\ &\times\frac{\rho_1^2\rho_\nu\rho_\mu}{[1+(\rho_1+\rho_\nu)^2][1+(\rho_1-\rho_\nu)^2][1+(\rho_1+\rho_\mu)^2][1+(\rho_1-\rho_\mu)^2]} \\ &\times\left\{(-1)^\nu + \frac{1+2(\rho_1^2+\rho_\nu^2)}{p_c\sqrt{(1+4\rho_1^2)(1+4\rho_\nu^2)}}\right\} \\ &\times\left\{(-1)^\mu + \frac{1+2(\rho_1^2+\rho_\mu^2)}{p_c\sqrt{(1+4\rho_1^2)(1+4\rho_\mu^2)}}\right\}.\end{aligned} \tag{65}$$

Evaluation of (54) with the help of (62) and (63) yields, for random deformations of the fiber axis,

$$\Delta t = \frac{0.42\bar{\Delta} p_c^2 \sqrt{L}}{c n_0 k^2 a^2 \sqrt{K(\Omega)}} \frac{a}{w}. \tag{66}$$

Equation (66) determines the pulse width of an impulse after it has traveled a distance L (L must be large enough so that the pulse has settled down to steady state) in the presence of random deformations of the fiber axis. The pulse width for uncoupled modes is obtained from (56)

$$\Delta T = \frac{L}{v(p_c, q_c)} - \frac{L}{v(0, 0)} = \frac{4\bar{\Delta} L}{c n_0 k^2 a^2} p_c^2. \tag{67}$$

The relative improvement of the width of the impulse response caused by mode coupling is characterized by the ratio[8]

$$R = \frac{\Delta t}{\Delta T} = \frac{0.105}{\sqrt{LK(\Omega)}} \frac{a}{w}. \tag{68}$$

Mode coupling not only shortens the width of the impulse response, but it also leads to excess loss. In order to find out how much excess loss is associated with a given improvement of the width of the impulse response, we form the product of (41) with the square of (68)

$$R^2 \sigma_{11} L = 0.032. \tag{69}$$

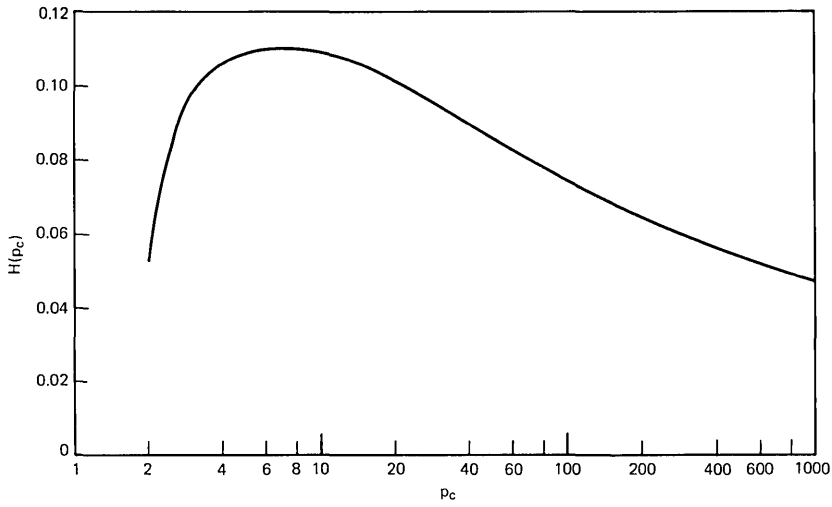

Fig. 3—Plot of the function $H(p_c)$.

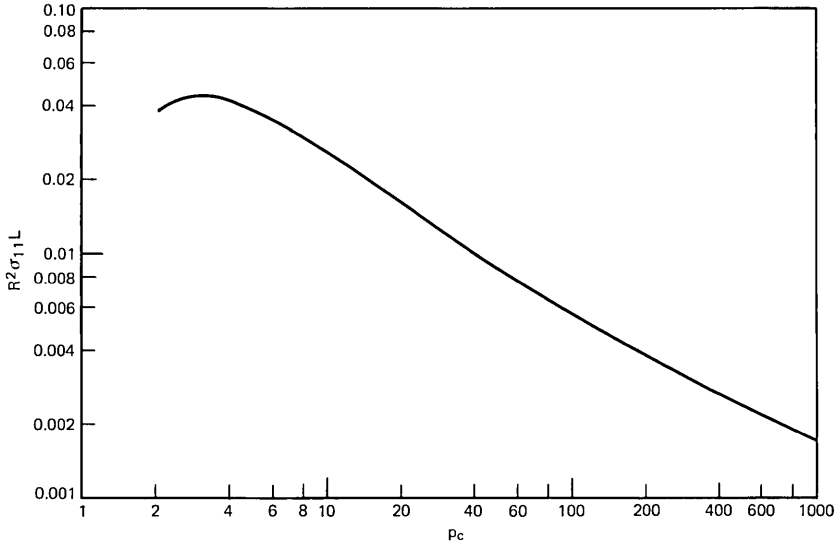

Fig. 4—Plot of the loss penalty as a function of p_c.

For random axis deformations, the product of the square of the improvement factor with the excess loss is independent of the waveguide parameters and the statistics of the random axis deformations.

For random width changes we obtain similarly

$$\Delta t = \frac{4\bar{\Delta} p_c^2}{c n_0 k^2 a^2} \sqrt{\frac{L}{K(2\Omega)}}\, H(p_c)\, \frac{a}{w}. \tag{70}$$

The function $H(p_c)$ is plotted in Fig. 3. The improvement factor is

$$R = \frac{H(p_c)}{\sqrt{LK(2\Omega)}}\, \frac{a}{w}. \tag{71}$$

Finally, we obtain the loss penalty from

$$R^2\sigma_{11}L = (2\rho_1^2 + 0.5)H^2(p_c). \tag{72}$$

This function is shown graphically in Fig. 4.

VII. DISCUSSION AND NUMERICAL RESULTS

We have derived expressions for the steady-state loss and the loss penalty of graded index fibers with square cross section for the case of random axis deformation and random width changes. The most

conspicuous difference between these two types of fiber imperfections is the fact that, whereas the Fourier components of the function $f(z)$ (describing the fiber axis) at the spatial frequency Ω are instrumental in the mode mixing process, the Fourier components at twice the spatial frequency, 2Ω, determine the mode mixing process in case of random changes of the width of the fiber. This behavior can easily be understood. Consider a Gaussian beam of arbitrary width that is injected into the fiber off axis.[11] The beam undulates periodically around the fiber axis and also changes its width periodically. The undulations around the fiber axis have a period[11]

$$\Lambda = \frac{\sqrt{2}\pi a}{\sqrt{\bar{\Delta}}} \tag{73}$$

while the width changes repeat themselves with half that period or at twice the spatial frequency.[11] Random displacements of the fiber axis couple to the deflections of the beam from its on-axis position. This deflection is driven by a Fourier component at the spatial frequency

$$\Omega = \frac{2\pi}{\Lambda} = \frac{\sqrt{2\bar{\Delta}}}{a}. \tag{74}$$

The beam width changes are correspondingly driven by changes in the gradient (width) of the parabolic index medium. It is thus clear that they respond to twice the spatial frequency.

In order to be able to associate specific rms deviations of the fiber axis or rms width changes with fiber loss we have to consider a particular statistical model. We choose (arbitrarily) a Gaussian correlation function

$$\langle f(z)f(z+u)\rangle = \bar{\sigma}^2 e^{-(u/D)^2}. \tag{75}$$

$\bar{\sigma}$ is the rms deviation of the function $f(z)$ and D its correlation length. The power spectrum of $f(z)$ is known to be[12]

$$F(\theta) = \sqrt{\pi}\bar{\sigma}^2 D e^{-(\theta D/2)^2}. \tag{76}$$

For a given value of θ this function assumes its maximum value

$$[F(\theta)]_{D=D_m} = \frac{\sqrt{2\pi}e^{-0.5}}{\theta}\bar{\sigma}^2 = \frac{1.52}{\theta}\bar{\sigma}^2 \tag{77}$$

at

$$D_m = \sqrt{2}/\theta. \tag{78}$$

Let us consider a numerical example. We use the following fiber

parameters

$$\left.\begin{aligned} a &= 4.85 \times 10^{-3} \text{ cm} \\ n_0 &= 1.56 \\ \bar{\Delta} &= 0.014 \end{aligned}\right\}. \tag{79}$$

At $\lambda = 1$ μm wavelength we have $a/w = 6.3$ or $p_c = 39.7$. The difference between the propagation constant of adjacent modes is, according to (32), $\Omega = 34.5$ cm^{-1}. With $\theta = \Omega$ we calculate the excess loss values at the peak of the power spectrum at the value of the correlation length given by (78), $D_m = 0.041$ cm. For random deviations of the waveguide axis we obtain from (41), (77), and (79) ($\bar{\sigma}$ in cm)

$$\sigma_{11} = 6.44 \times 10^6 \bar{\sigma}^2 \text{ (cm}^{-1}\text{)}. \tag{80}$$

In order to keep the excess loss below 10 dB/km $= 2.3\ 10^{-5}$ cm^{-1} we must keep the rms deviation of the fiber axis below $\bar{\sigma} = 2 \times 10^{-6}$ cm. However, this stringent tolerance requirement results from our assumption that the correlation length of the random irregularities of the fiber axis assumes its worst possible value (78). If, for example, the correlation length happens to be $D = 0.5$ cm we obtain instead of (80)

$$\sigma_{11} = 6.4 \times 10^{-25} \bar{\sigma}^2 \text{ (cm}^{-1}\text{)} \tag{81}$$

so that we can now tolerate $\bar{\sigma} = 6 \times 10^9$ cm in order to keep the excess loss below 10 dB/km. This example shows that it is impossible to predict the excess loss to be expected from a practical square-law fiber unless the statistics of its irregularities are known precisely.

For reasons of comparison we state the corresponding value for random width changes. In this case the spatial frequency that is instrumental in providing mode coupling is $2\Omega = 69$ cm^{-1}. The worst possible correlation length is now $D_m = 0.02$ cm. From (53) and Fig. 2 with $p_c = 40$ we obtain ($\bar{\sigma}$ in cm)

$$\sigma_{11} = 1.39 \times 10^6 \bar{\sigma}^2 \text{ (cm}^{-1}\text{)}. \tag{82}$$

The tolerance requirements of random width changes appear a little less stringent than those of random axis deformations. However, we have already pointed out that the excess loss value depends critically on the actual statistics of the fiber. Since the excess loss caused by random axis deviations depends on a different spatial frequency than the excess loss caused by random width changes, a loss comparison of the two effects is not possible.

Next we discuss the loss penalty that is incurred for a given improvement of the width of the impulse response of coupled mode operation compared to uncoupled mode operation. The equations (69) and (72) show that the loss penalty is independent of the statistics of the fiber irregularities. This feature makes the loss penalty a useful quantity. In case of random variations of the fiber axis, the loss penalty is even independent of the fiber parameters and is simply a dimensionless number. Let us assume that we want to achieve a ten-fold improvement of the width of the impulse response compared to the impulse response of uncoupled multimode operation. In this case we have $R = 0.1$ and obtain from (69) for random deviations of the fiber axis

$$\sigma_{11}L = 3.2 = 14 \text{ dB}. \tag{83}$$

The length L needed to incur this loss and at the same time to achieve $R = 0.1$ depends on the statistics of the irregularities. However, eq. (83) tells us that it costs 14 dB in excess loss to achieve a ten-fold relative pulse width improvement.

For random width changes, the situation is slightly different. Here the loss penalty depends somewhat on the fiber parameters. For the values used earlier we find from Fig. 4 with $p_c = 40$,

$$R^2\sigma_{11}L = 10^{-2}. \tag{84}$$

The loss penalty for $R = 0.1$ is more favorable in this case, $\sigma_{11}L = 4.34$ dB.

Fiber irregularities can be introduced intentionally in order to improve the impulse response. In the conventional fiber with a round core of constant refractive index that is surrounded by a cladding with constant index, the loss penalty for pulse distortion improvement can be reduced (in principle avoided) by tailoring the shape of the power spectrum carefully.[8] The reason that the shape of the power spectrum has an influence on the loss penalty is explained by the observation that the spacing (in β-space) between adjacent modes of the conventional fiber is dependent on the mode number, so that a band of spatial frequencies of the power spectrum takes part in the mode coupling process.

In case of the parabolic index fiber, only one spatial frequency (or actually a narrow range of spatial frequencies) is responsible for mode coupling. The shape of the power spectrum is thus immaterial, only its value at the spatial frequency Ω enters into the picture. The expressions (69) and (72) show that the loss penalty of the parabolic

index fiber is independent of the power spectrum. No loss advantage is to be gained by using especially shaped power spectra. One might think that an advantage could be gained by departing from the square-law index profile in order to change the mode spacing and sample more of the power spectrum. But as soon as the index distribution deviates slightly from the parabolic profile the uncoupled impulse response becomes much broader. The mode coupling mechanism would now have to work against a far less favorable (uncoupled) impulse response so that it seems unlikely that an advantage can be gained in this way.

Finally, we consider an example of pulse width reduction by random irregularities. We can introduce intentional deviations of the fiber axis from perfect straightness in order to cause mode coupling. Since the coupling process must be random, we could use deformation functions $f(z)$ and $g(z)$ that are sinusoidal in shape but have a random phase. This introduces a power spectrum centered around the spatial frequency of the sinusoidal process having a finite width. Instead of pursuing this idea further, we assume that we have somehow created an axis deformation whose power spectrum reaches beyond the frequency Ω of (32). For simplicity, and to have a definite case in mind, we choose

$$F(\theta) = \begin{cases} \dfrac{\pi\bar{\sigma}^2}{2\Omega} & |\theta| < 2\Omega \\ 0 & |\theta| > 2\Omega. \end{cases} \tag{85}$$

This power spectrum is flat from zero spatial frequencies to a cutoff value of $\theta = 2\Omega$ and zero for $\theta > 2\Omega$. The rms deviation $\bar{\sigma}$ of the fiber axis from a straight line appears in (85). Using the fiber parameters (79) we obtain from (68) ($\bar{\sigma}$, L in cm)

$$R = \frac{6.9 \times 10^{-5}}{\bar{\sigma}\sqrt{L}}. \tag{86}$$

A ten-fold improvement of the width of the impulse response (compared to the uncoupled case), $R = 0.1$, over a length of $L = 1$ km $= 10^5$ cm requires an rms deviation of the fiber axis of $\bar{\sigma} = 2.2 \times 10^{-6}$ cm. We already know that we pay for this improvement of the impulse response with a loss penalty of 14 dB. Very slight random deviations from perfect straightness are already very effective in providing mode coupling and improving the width of the impulse response.

For random width changes we have to allow for a wider power spectrum. Letting the power spectrum again extend twice as far as the effective spatial frequency, 2Ω in this case, forces us to divide (85) by 2. We thus find from Fig. 3 and (71)

$$R = \frac{8.37 \times 10^{-5}}{\bar{\sigma}\sqrt{L}}. \tag{87}$$

$R = 0.1$ and $L = 10^5$ cm requires $\bar{\sigma} = 2.6 \times 10^{-6}$ cm.

APPENDIX

The function

$$H_p\left(\sqrt{2}\,\frac{x}{w}\right) e^{-x^2/w^2} e^{-i\beta_p z} \tag{88}$$

describes the modes of a square-law medium defined by

$$n(x) = n_0\left(1 - \frac{x^2}{a^2}\,\bar{\Delta}\right). \tag{89}$$

The associated ray problem can be described by a paraxial Hamiltonian of the form[13]

$$H = \frac{p_x^2}{2n_0} - n(x). \tag{90}$$

The quantum mechanical treatment of this problem leads to an expression for the "energy" E of the ray that has the form[14]

$$E = -\frac{\beta_p}{k}. \tag{91}$$

We define the "turning point" of the light rays associated with the wave field (88) by the condition that p_x, which is proportional to the slope of the light ray, must vanish. That means that the ray trajectory is tangential to the optical axis as the rays turn back in their path leading them away from the axis. Using $p_x = 0$ and equating (90) and (91) we find the following condition for the turning point:

$$n(x') = \frac{\beta_p}{k}. \tag{92}$$

The propagation constant of this two-dimensional mode field is[5]

$$\beta_p = n_0 k\left[1 - 2\,\frac{\sqrt{2\bar{\Delta}}}{n_0 ka}\left(p + \frac{1}{2}\right)\right]^{\frac{1}{2}}. \tag{93}$$

Substitution of (89) and (93) into (92) leads with the help of (5) to the formula (9) for the turning point.

The physical argument advanced here serves the purpose of defining the range in which the Hermite polynomial has an oscillatory behavior. This range is given by

$$-x' \leqq x \leqq x'. \tag{94}$$

Outside of this range the Hermite polynomial grows monotonically to infinite values. However, since the Hermite polynomial enters the mode field (88) only as a product with a Gaussian function, the mode field decays rapidly without oscillation outside of the range given by (94).

REFERENCES

1. S. E. Miller, U.S. Patent No. 3,434,774, "Waveguide for Millimeter and Optical Waves," issued March 25, 1969.
2. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.
3. D. Marcuse, "The Impulse Response of an Optical Fiber With Parabolic Index Profile," B.S.T.J., *52*, No. 7 (September 1973), pp. 1169–1174.
4. D. Marcuse, "Losses and Impulse Response of a Parabolic Index Fiber with Random Bends," B.S.T.J., *52*, No. 8 (October 1973), pp. 1423–1437.
5. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold Co., 1972, p. 270.
6. A. W. Snyder, "Coupled Mode Theory for Optical Fibers," J. Opt. Soc. Am., *62*, No. 11 (November 1972), pp. 1267–1277.
7. D. Marcuse, "Coupled Mode Theory of Round Optical Fibers," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 817–842.
8. D. Marcuse, "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1199–1232.
9. D. Gloge, "Optical Power Flow in Multimode Fibers," B.S.T.J., *51*, No. 8 (October 1972), pp. 1767–1783.
10. D. Gloge, "Impulse Response of Clad Optical Multimode Fibers," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 801–816.
11. Ref. 5, pp. 272–275.
12. D. Marcuse, "Power Distribution and Radiation Losses in Multimode Dielectric Slab Waveguides," B.S.T.J., *51*, No. 2 (February 1972), pp. 429–454.
13. Ref. 5, p. 96.
14. Ref. 5, p. 104.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Transverse Coupling in Fiber Optics Part I: Coupling Between Trapped Modes

By J. A. ARNAUD

(Manuscript received August 3, 1973)

Two perturbation formulas have been proposed to evaluate the coupling between parallel optical waveguides, one involving a line integral and the other a surface integral. They are shown to be identical. The former expression is preferred because of its greater simplicity. The case of two parallel lossy dielectric slabs is discussed as an example.

I. INTRODUCTION

There has been a renewed interest during the last few years in the evaluation of the transverse* coupling between two parallel open waveguides in connection with integrated optics circuitry[2,3] and long-distance optical communication by bundles of glass fibers.

The coupling between two open waveguides can be obtained by replacing the field of one waveguide by an equivalent current and evaluating the perturbation caused by this current on the other waveguide.[4] A more direct and slightly more general (but essentially equivalent) derivation, based on Lorentz's reciprocity theorem, is given in this paper. A related result, applicable only to lossless fibers, has been used to evaluate the coupling between dielectric rods with circular cross section.[5] The perturbation formula derived in this paper involves an integral along a contour located between the two waveguides. A seemingly different perturbation formula has been recently proposed that involves a surface integral over the cross section.[6] The two formulas are shown to be in fact identical. We will not discuss in detail other coupling formulas such as the ones proposed in Refs. 7 or 3. In Ref. 7, the coupling is obtained by applying the Rayleigh-Ritz

* The word "transverse" is used here to distinguish the problem of two dielectric waveguides lying side by side, where the transfer of power takes place in transverse directions, and the axial coupling between two waveguides placed end to end, where the transfer of power takes place along the z axis (the later arrangement is discussed, for instance, in Ref. 1).

optimization technique to a variational expression. The formula obtained by this method involves surface integrals and is rather complicated. In Ref. 3, analytic expressions were obtained for the coupling between two identical rectangular fibers that agree well with numerical calculations based on the exact field equations. The approach, however, is restricted to fibers with a particular geometry.

II. GENERAL EXPRESSION OF THE COUPLING

Let the time dependence of the sources be denoted exp $(-\kappa t)$. Maxwell's equations are, in a source-free region with scalar permittivity ϵ and permeability μ_o,

$$\nabla \times \mathbf{E} = \kappa\mu_o\mathbf{H}, \tag{1a}$$

$$\nabla \times \mathbf{H} = -\kappa\epsilon\mathbf{E}. \tag{1b}$$

Any two solutions $(\mathbf{E}, \mathbf{H})$ and $(\mathbf{E}_a, \mathbf{H}_a)$ of eq. (1) satisfy the relation

$$\nabla\cdot\mathbf{J} = 0, \tag{2a}$$

where

$$\mathbf{J} = \mathbf{E}_a \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_a. \tag{2b}$$

Integrating over a volume V, Lorentz reciprocity theorem is obtained

$$\int_S (\mathbf{E}_a \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_a)\cdot\mathbf{d}S = 0, \tag{2c}$$

where S denotes the surface enclosing V, and $\mathbf{d}S$ a vector normal to S pointing outward with magnitude dS. Let the medium be uniform along z, that is, ϵ be independent of z. If

$$(\mathbf{E}, \mathbf{H}) \equiv (E_z, \mathbf{E}_t, H_z, \mathbf{H}_t) \exp(\gamma z) \tag{3}$$

denotes a solution of Maxwell's equations, then

$$(\mathbf{E}^+, \mathbf{H}^+) \equiv (-E_z, \mathbf{E}_t, H_z, -\mathbf{H}_t) \exp(-\gamma z) \tag{4}$$

is also a solution of Maxwell's equations. The arguments x, y have been omitted in the above expressions, and the subscripts t stand for "transverse." The field $(\mathbf{E}^+, \mathbf{H}^+)$ is the field adjoint to $(\mathbf{E}, \mathbf{H})$; it describes a wave propagating in an opposite direction in the same medium and at the same frequency. A more general definition of the adjoint field, applicable to nonreciprocal media, can be found in Ref. 8.

Let us now consider two open waveguides a and b uniform along the z-axis, and let S be the surface $S_a + S_a' + C_a dz$ shown in Fig. 1. The field $(\mathbf{E}_a, \mathbf{H}_a)$ in eq. (2) is taken as the field of a trapped mode on

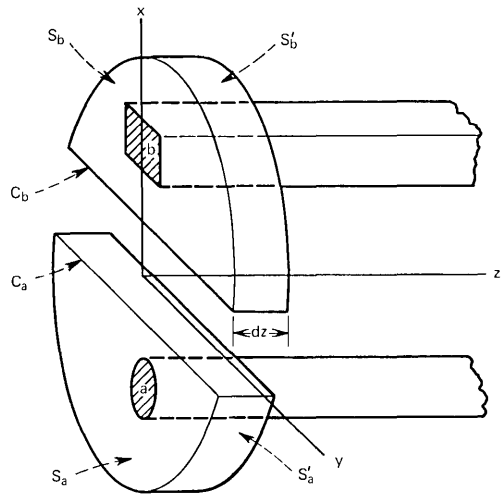

Fig. 1—Coupled dielectric waveguides.

waveguide a in the absence of waveguide b. The dependence of $(\mathbf{E}_a, \mathbf{H}_a)$ on z is denoted exp $(\gamma_a z)$. The field $(\mathbf{E}^+, \mathbf{H}^+)$ is the adjoint field of a trapped mode of the two coupled waveguides, with an exp $(-\Gamma z)$ dependence on z. Letting the spacing dz between S_a' and S_a tend to zero, eq. (2) becomes

$$(\gamma_a - \Gamma) \int_{S_a} (\mathbf{E}_a \times \mathbf{H}^+ - \mathbf{E}^+ \times \mathbf{H}_a) \cdot \mathbf{d}S_a = -\int_{C_a} (\mathbf{E}_a \times \mathbf{H}^+ - \mathbf{E}^+ \times \mathbf{H}_a) \cdot \mathbf{d}C_a, \quad (5a)$$

where $\mathbf{d}C_a$ is a vector perpendicular to the contour C_a, pointing outward. Proceeding similarly for waveguide b we obtain

$$(\gamma_b - \Gamma) \int_{S_b} (\mathbf{E}_b \times \mathbf{H}^+ - \mathbf{E}^+ \times \mathbf{H}_b) \cdot \mathbf{d}S_b = -\int_{C_b} (\mathbf{E}_b \times \mathbf{H}^+ - \mathbf{E}^+ \times \mathbf{H}_b) \cdot \mathbf{d}C_b. \quad (5b)$$

Because the coupling between the two waveguides is small, we can assume that the field $\mathbf{E}$, $\mathbf{H}$ at plane $z = 0$ is the sum of the fields of the two waveguides, that is,

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_a + \mathbf{E}_b, \\ \mathbf{H} &= \mathbf{H}_a + \mathbf{H}_b. \end{aligned} \quad (6)$$

Substituting these expressions, eqs. (6), in eqs. (5a) and (5b) we observe that the cross terms can be neglected on the left-hand sides (l.h.s.) because ($\mathbf{E}_b$, $\mathbf{H}_b$) is small when ($\mathbf{E}_a$, $\mathbf{H}_a$) is large, and vice versa. On the right-hand sides (r.h.s.), on the contrary, only the cross terms remain, as we can verify by applying Lorentz reciprocity theorem to each waveguide. Multiplying together the l.h.s. and r.h.s. of eq. (5a) and (5b) the desired equation for Γ is obtained.

$$(\Gamma - \gamma_a)(\Gamma - \gamma_b) = c_a c_b / P_a P_b, \tag{6a}$$

where

$$c_{a,b} = \int_{C_{a,b}} (\mathbf{E}_{a,b} \times \mathbf{H}^+_{b,a} - \mathbf{E}^+_{b,a} \times \mathbf{H}_{a,b}) \cdot \mathbf{d}C_{a,b}, \tag{6b}$$

$$P_{a,b} = \int_{S_{a,b}} (\mathbf{E}_{a,b} \times \mathbf{H}^+_{a,b} - \mathbf{E}^+_{a,b} \times \mathbf{H}_{a,b}) \cdot \mathbf{d}S_{a,b}. \tag{6c}$$

Because the coupling takes place only if $\gamma_a \sim \gamma_b$, the coupling c_a (resp. c_b) is independent of the choice of the contour C_a (resp. C_b) as long as it surrounds only one waveguide. By choosing the two contours as coincident in the region where the fields of the two trapped modes have a significant intensity and using eq. (4), we find that c_a is equal to c_b. It is shown in the appendix that our result, eq. (6), can be expressed in the form given in Ref. 6. The expression, eq. (6), however, is simpler to evaluate.

Let us now assume that the contours C_a, C_b coincide with the y axis and are closed at infinity where the fields vanish. The general expression, eq. (6), becomes

$$(\Gamma - \gamma_a)(\Gamma - \gamma_b) = c^2 / P_a P_b, \tag{7}$$

where

$$c = \frac{1}{2} \int_{-\infty}^{+\infty} (E_{ay} H_{bz} + E_{az} H_{by} - E_{by} H_{az} - E_{bz} H_{ay}) dy,$$

and

$$P_a = \int \int_{-\infty}^{+\infty} (\mathbf{E}_a \times \mathbf{H}_a) \cdot \mathbf{z} dx dy,$$

$$P_b = \int \int_{-\infty}^{+\infty} (\mathbf{E}_b \times \mathbf{H}_b) \cdot \mathbf{z} dx dy,$$

where $\mathbf{z}$ denotes the unit vector directed along the z axis.

Let us specialize eq. (7) to symmetrical stratified dielectric waveguides such as the slabs shown in Fig. 2. The fields are assumed to be independent of y. For TE waves the electric field has only one

component $E_y \equiv E$. We have, from Maxwell's equations, eq. (1)

$$E_x = E_z = H_y = 0, \tag{8a}$$

$$H_z = (\kappa\mu_o)^{-1} dE/dx, \tag{8b}$$

$$H_x = -(\gamma/\kappa\mu_o)E. \tag{8c}$$

Equation (7) thus reduces to the simpler form

$$(\Gamma - \gamma_a)(\Gamma - \gamma_b) = (1 - \kappa^2/\gamma^2)E_a^2 E_b^2 \Big/ \left(\int_{-\infty}^{+\infty} E_a^2 dx \int_{-\infty}^{+\infty} E_b^2 dx \right), \tag{9}$$

the fields E_a and E_b of the uncoupled waveguides being evaluated at the same point between the two waveguides.

III. COUPLING BETWEEN LOSSY DIELECTRIC SLABS

If the waveguides are homogeneous dielectric slabs of thickness $2d$ and complex permittivity ϵ we have

$$E = \exp(\mp\gamma_{xo}x) \tag{10a}$$

above or below the slabs and, for even modes,

$$E = \cosh(\gamma_x x)/\cosh(\gamma_x d) \tag{10b}$$

within the slabs (obvious changes in the origin of the x axis were made). In eqs. (10a) and (10b) we have defined

$$\gamma_{xo}^2 \equiv k_o^2 - \gamma^2, \qquad \text{Real } (\gamma_{xo}) > 0, \tag{11a}$$

$$\gamma_x^2 \equiv k_o^2 n^2 - \gamma^2, \tag{11b}$$

$$k_o^2 \equiv \kappa^2 \epsilon_o \mu_o, \tag{11c}$$

$$n^2 \equiv \epsilon/\epsilon_o.$$

The propagation constant γ is known to satisfy

$$\gamma_x \tanh(\gamma_x d) + \gamma_{xo} = 0. \tag{12}$$

(See, for instance, Ref. 6.) Substituting E from eqs. (10a) and (10b) in eq. (9) we obtain, using eq. (12),

$$\Gamma - \gamma = \pm\gamma^{-1}\gamma_x^2\gamma_{xo}^2(1 + \gamma_{xo}d)^{-1}(\gamma_x^2 - \gamma_{xo}^2)^{-1} \exp(-\gamma_{xo}D), \tag{13}$$

where D denotes the spacing between the slabs. This expression coincides with the result given in Ref. 6 when appropriate changes of notation are made.

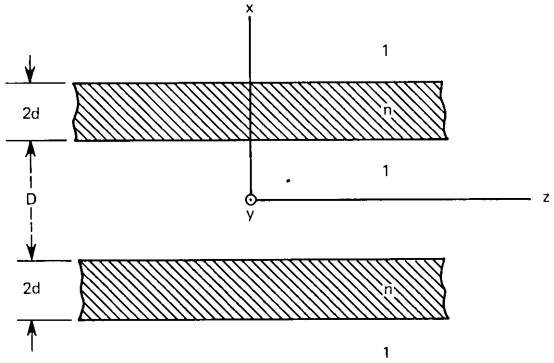

Fig. 2—Coupled dielectric slabs.

Let us now make a general comment. The coupling formula, eq. (6), rests on the existence of a divergenceless quantity, the vector **J** in eq. (2a). Coupling formulas similar to eq. (6) can be derived from other wave equations. For the case of the scalar parabolic wave equation[9] applicable to the propagation of radio waves in atmospheric ducts,* the vector **J** has components

$$J_z = 2k_o E_a^+ E \approx E_a^+ \partial E/\partial z - E \partial E_a^+/\partial z,$$
$$\mathbf{J}_t = E_a^+ \nabla_t E - E \nabla_t E_a^+,$$

where the adjoint field is

$$E_a^+(x, y, z) = E_a(x, y, -z).$$

The above expression for **J** can be obtained by analogy with the equivalent quantum mechanical problem.[9]

In conclusion, we have derived a simple coupling formula which is more general than previous similar expressions[4,5] because it is applicable to lossy fibers. In order to evaluate explicitly the coupling, one needs to know the normalized field of each waveguide, in the absence of the other, along some line located between the two waveguides. For slabs and rods with circular cross section, exact solutions are available. In general, however, we have to resort to numerical techniques or to measurements made at a convenient wavelength on a scaled version of the open waveguide. In a second part of this paper,[10] we will apply eq. (6) to mode-selecting systems.

* A similar equation is applicable to anisotropic fibers that have small transverse variation of permittivity.[8] Note that, in this approximation, a curvature of the fiber axis is equivalent to a constant gradient of refractive index.

APPENDIX

The purpose of this appendix is to show that for lossless fibers the coupling formulas given in Refs. 4 and 6 coincide.

Let $(\mathbf{E}, \mathbf{H})$ denote a field in free space

$$\begin{aligned} \nabla \times \mathbf{E} &= \kappa \mu_o \mathbf{H}, \\ \nabla \times \mathbf{H} &= -\kappa \epsilon_o \mathbf{E}, \end{aligned} \tag{14}$$

and $(\mathbf{E}_b, \mathbf{H}_b)$ a field in a dielectric with permittivity $\epsilon(\mathbf{r})$

$$\begin{aligned} \nabla \times \mathbf{E}_b &= \kappa \mu_o \mathbf{H}_b, \\ \nabla \times \mathbf{H}_b &= -\kappa \epsilon \mathbf{E}_b. \end{aligned} \tag{15}$$

It is easy to show that these fields satisfy the relation

$$\int_S (\mathbf{E} \times \mathbf{H}_b - \mathbf{E}_b \times \mathbf{H}) \cdot \mathbf{d}S = \kappa \int_V (\epsilon - \epsilon_o) \mathbf{E} \cdot \mathbf{E}_b dV \tag{16}$$

in any source-free volume V bounded by S. Let now the surface S be the surface $S_b + S_b' + C_b dz$ shown in Fig. 1, $(\mathbf{E}_b, \mathbf{H}_b)$ be the field of a trapped mode of waveguide b with an exp $(\gamma_b z)$ dependence on z, and $(\mathbf{E}, \mathbf{H})$ be the adjoint field $(\mathbf{E}_a^+, \mathbf{H}_a^+)$ of a trapped mode of waveguide a, with an exp $(-\gamma_a z)$ dependence on z. The field $(\mathbf{E}_a^+, \mathbf{H}_a^+)$ satisfies eq. (14) inside the surface S that we have just defined. If the two trapped modes are degenerate, that is, if $\gamma_a = \gamma_b$, the contributions of the two surfaces S_b and S_b' on the l.h.s. of eq. (16) cancel out. Therefore, letting dz tend to zero, eq. (16) becomes

$$\int_{C_b} (\mathbf{E}_a^+ \times \mathbf{H}_b - \mathbf{E}_b \times \mathbf{H}_a^+) \cdot \mathbf{d}C_b = \kappa \int_{S_b} (\epsilon - \epsilon_o) \mathbf{E}_a^+ \cdot \mathbf{E}_b dS_b. \tag{17}$$

A similar relation can be obtained for waveguide a. Our coupling equation, eq. (6), can therefore be written in the form given in Ref. 6, except for the fact that in eq. (6) $\mathbf{E}_a^+$ and $\mathbf{E}_b$ represent fields at the same frequency. In Ref. 6 the field $\mathbf{E}_a^+$ is defined at the opposite angular frequency $-\kappa$, that is, $\mathbf{E}_a^+$ is replaced by $\mathbf{E}_a^{+*}$, where the asterisk denotes complex conjugation. For lossless fibers this difference is unimportant because $\mathbf{E}_a^+$ can be assumed real.

REFERENCES

1. J. A. Arnaud, "Mode Coupling in First-Order Optics," J. Opt. Soc. Am., *61*, June 1971, p. 751.
2. S. E. Miller, "Integrated Optics: An Introduction," B.S.T.J., *48*, No. 7 (September 1969), p. 2059.
3. E. A. J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," B.S.T.J., *48*, No. 7 (September 1969), p. 2071.

4. J. A. Arnaud, "General Properties of Periodic Structures," Sections B and E, in *Crossed Field Microwave Devices*, Vol. 1, E. Okress ed., New York: Academic Press, 1961.
5. R. Vanclooster and P. Phariseau, "The Coupling of Two Parallel Dielectric Fibers," Physica, 7, June 1970, pp. 485 and 501.
6. D. Marcuse, "The Coupling of Degenerate Modes in Two Parallel Dielectric Waveguides," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1791–1816.
7. M. Matsuhara and N. Numagai, "Coupling Theory of Open-Type Transmission Lines and its Application to Optical Circuits," Electron. Commun. Japan, *55*, April 1972, p. 102.
8. J. A. Arnaud, "Biorthogonality Relations for Bianisotropic Media," J. Opt. Soc. Am., *63*, February 1973, p. 238.
9. V. A. Fock, "Theory of Radio-Wave Propagation in an Inhomogeneous Atmosphere for a Raised Source," Bull. Acad. Sciences de l'URSS, sér. phys. *14*, No. 1, p. 70 (1950). In Russian. Translation in V. A. Fock, *Electromagnetic Diffraction and Propagation Problems*, Oxford: Pergamon Press, 1965, Chapter 14.
10. J. A. Arnaud, "Transverse Coupling in Fiber Optics, Part II," to appear in B.S.T.J., *56*, No. 4 (April 1974).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Performance Models of an Experimental Computer Communication Network

By J. F. HAYES

(Manuscript received March 12, 1973)

This paper reports the results of a performance study of an experimental computer communication network. The network is currently being designed and built in order to test concepts and techniques that may find future application. The network consists of synchronous digital transmission lines connected in loops to a Central Switch. User traffic enters the system through multiplexers connected to the synchronous lines. The Central Switch has the two-fold function of routing and controlling traffic.

Two multiplexing techniques were examined, Demand Multiplexing (DM) and Synchronous Time Division Multiplexing (STDM). In both techniques, user messages are blocked into fixed size packets, prior to multiplexing on the line. The synchronous line can carry these packets at a maximum rate of 4000 packet slots per second. In STDM each terminal is assigned a packet slot which recurs periodically. In contrast, for DM, packets are multiplexed on the line asynchronously into unoccupied packet slots. Alternative implementations of the DM technique were studied, one where each terminal transmits and receives at a maximum rate of 4000 packets per second and another where the maximum rate is 2000 packets per second.

As part of its message-handling function, the Central Switch buffers messages in transit. This allows User Terminals to transmit and receive messages with a degree of independence from one another. However, the terminals' strategy affects the amount of storage required in the Central Switch. In order to prevent the loss of information when there is insufficient buffering, there is a mechanism to inhibit traffic from User Terminals when the Central Switch buffer is near overflow. Due to this control of traffic, there is a relationship between the amount of data that flows through the switch and the amount of buffering in the switch.

Simulation results showed that there was little difference in delay performance between the two implementations of DM. However, an analysis

comparing DM and STDM showed a great difference in performance for all but the very heaviest line loadings. This difference increases as the number of terminals sharing the T1 line increases.

Our study concentrated on two aspects of buffering in the Central Switch. We examined the relationship between throughput and the amount of storage available in the switch. The results of a simulation study showed that throughput can be quite high for all but minimal storage in the switch. Moreover, a strategy that dedicates buffers does quite well compared to common buffering. The second aspect of the study concentrated upon the User Terminal's strategy. Since each terminal acts independently, there may be strategies that make particularly high demands upon storage capacity in the Central Switch. An analysis showed that at the loadings where the system would be expected to operate, the user strategy in transmitting and receiving messages has little effect.

I. INTRODUCTION

An experimental computer communication network is currently being designed and built. The function of this network is to provide a flexible communication medium between computers, users, and peripheral devices. The network can accommodate sources with varying input-output rates and varying activity. Many of the components of the system employ techniques that are new. In order to gain insight into the operation of these components and thereby aid in design decisions, mathematical models were developed. The study of these models involved both analysis and simulation. The results are presented in the form of sets of curves.

The system under study consists of several T1 carrier lines,* configured as loops, connected to a Central Switch (see Fig. 1). The system is accessed through Terminal Interface Units (TIU) connected between User Terminals and the T1 line. In addition to forming an interface between the user and the T1 line, the TIU also does signaling. This signaling plays a role in switching calls and controlling the traffic flow.

There may be a wide variation of users accessing the system, ranging from Teletypes† to high-speed computer systems. The switch receives messages from all terminals and delivers messages to all addressed terminals so that any terminal in the system may communicate with

---

* The T1 carrier line is a digital synchronous short-haul transmission system operating at 1.544-Mb/s rate.

† Registered trademark of the Teletype Corporation.

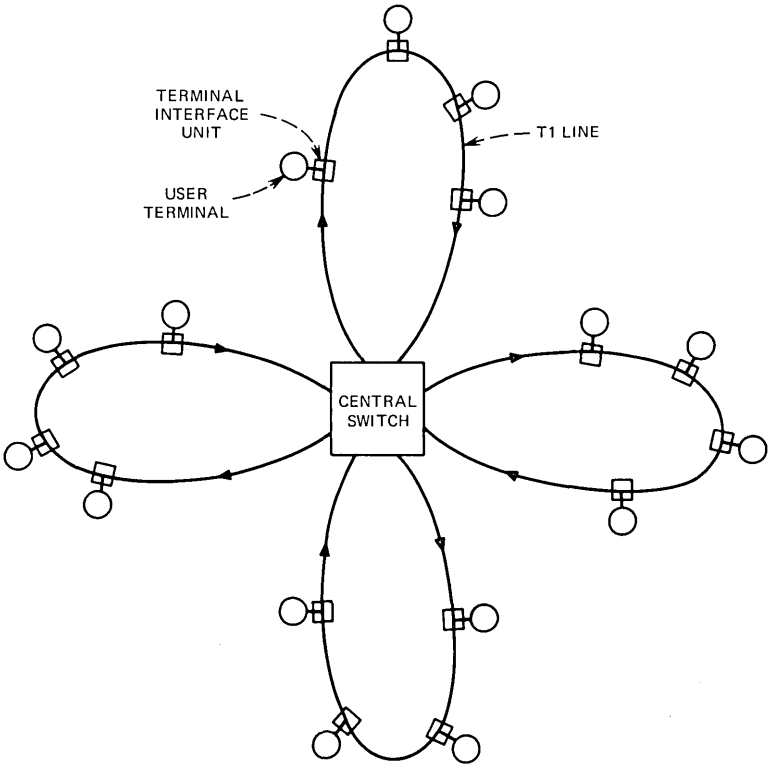

Fig. 1—Computer communication network.

any other terminal. All data pass through the switch even when two terminals are on the same T1 line.

The T1 line operates at a rate of 1.544×10^6 b/s. For purposes of synchronization and timing, the bit flow is divided into frames of 193 bits, with a flow of 8000 frames per second. The multiplexing arrangement in the system under study is such that a "network frame" consists of two adjacent T1 frames. Figure 2 indicates schematically how the 386 bits of the pair of T1 frames are allocated. The 50 bits required for framing and timing are part of the operation of the T1 carrier system. The assignment of the remaining 336 bits in the network frame is peculiar to the system under study.

User data are blocked into 256-bit packets and multiplexed on the line. Twenty-four bits of header information are attached to these information packets. (In the sequel we shall use the term "packet slot" in referring to this 280-bit block assigned to data and header.)

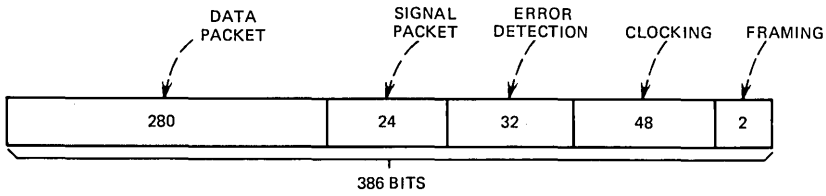

Fig. 2—Bit allocation of T1 frame pair.

Twenty-four of the 336 bits are used to carry signal packets. Signal packets convey control and routing information between the TIU and the switch. The remaining 32 bits of the frame pair are used for error detection in order to insure the integrity of the information in the signal and data packets.

From the foregoing we see that the information-carrying capability of the T1 loop is 4000 packets per second, each packet bearing 256 data bits, yielding a total information capacity of 1.024 Mb/s. There are many strategies that can be used to divide this capacity among the terminals connected to the loop. We shall evaluate the performance of two strategies, Synchronous Time Division Multiplexing and Demand Multiplexing. In Synchronous Time Division Multiplexing (STDM), each User Terminal is assigned a particular packet slot which recurs periodically. The terminal may multiplex data into its slot and receive data only in this same slot. For example, if there are ten terminals on the T1 loop and each terminal receives the same service, a particular terminal may multiplex packets on the line at a maximum rate of 400 packets per second. The time between packet multiplexing for these terminals is a constant 1/400 second.

In Demand Multiplexing (DM), packet slots are not assigned to a particular terminal. If a terminal has a packet to transmit to the switch, it inserts the packet into the first slot that is empty. Thus, unlike the STDM system, the flow of packets into the switch has no particular ordering as to originating terminal. So that the switch can sort packets according to their originating terminal, each packet has an address label in its 24-bit header. Similarly, information packets going from the switch to the terminal are not ordered and a header is required for each packet. Furthermore, each TIU must be able to recognize packets addressed to it. As we shall see, the number of bits required for addressing is relatively small.

Once a packet has been multiplexed on the loop either from the switch or from a terminal, it has priority over incoming traffic until

it reaches its destination. A terminal must wait for an empty data packet slot before it can place a waiting packet onto the line. As in the STDM, the implementation also allows a terminal to place an outgoing packet into a slot from which it is removing an incoming packet.

We consider two implementations of the DM system, which correspond to the maximum speed at which terminals can transmit or receive. In the adjacent slot seizure implementation, terminals can transmit and receive at a 4000-packet-per-second rate. We consider an alternate implementation where the terminal is constrained to operate at a 2000-packet-per-second rate. In this case, a terminal can only write into or read from alternate packet slots.

A major component of the system is the Central Switch. The function of the switch is to route and control the flow of information. All messages generated at User Terminals pass through the switch where they are passed on to their destination terminals. Now the operation of the system (see Section V) is such that, as it may not be possible to deliver a message to its destination immediately, messages are temporarily stored in the switch. Also, destination terminals have some control over the way that these stored messages are read out of the switch's buffer.

The storage capability of the switch is not unlimited; therefore, the flow of information packets into the switch must be controlled. The switch does this by informing terminal TIU of the amount of storage currently available in the switch. The terminal does not transmit information packets when there is no room in the switch, but holds them until storage is available.

As mentioned earlier, models of the system were studied in order to gain insight into performance and thereby guide design decisions. The models studied are approximations to actual system operation. We felt that the study of more exact, hence more complicated, models would have involved far more time and effort, without a corresponding increase in insight.

We study the performance of multiplexing techniques on the loop as measured by message delay. In the switch we study packet storage requirements from two points of view, throughput and user strategy. Since the switch inhibits the flow of information when the storage in the switch is used up, there is a relationship between throughput and storage capacity. We also study the effect of different user readout strategies on switch storage requirements.

II. SUMMARY

As a guide to the reader, we pause to summarize the main body of the paper before plunging into details. In Section III, analytical and simulation approaches to the loop multiplexing problem are presented. Section IV is devoted to a discussion of our results on loop multiplexing. The relationship between storage capacity in the switch and throughput is considered in Section V. The results of a simulation study of capacity and throughput are presented in Section VI. In Section VII we consider the effect of a user's strategy on storage requirements in the switch. The results of this study are presented in Section VIII.

Although the analytical and simulation techniques used in our study are not restricted to a particular message distribution, we concentrated on the case where 30 percent of the messages are 32 packets long (8192 bits) and the remaining messages are one packet in duration (256 bits). This message distribution was our best guess at the actual distribution of messages in the system and reflects the fact that most terminals will, in fact, be computers. In the sequel we use the term variable message length to designate this distribution. The case where all messages are one packet in duration was also studied to some extent. In referring to this latter distribution we use the term constant message length.

The results of our studies of loop multiplexing are presented in Section IV. Simulation results for Demand Multiplexing indicate little difference in performance between alternate and adjacent slot seizure (see Figs. 6 and 7). The simulation was carried out for the variable message length distribution which consists of a large proportion of long messages. One would expect this distribution to be especially sensitive to the minimum time required to transmit and receive these long messages. In contrast, for the constant message length distribution, these maximum speeds, 2000 packets per second (alternate slot seizure) or 4000 packets per second (adjacent slot seizure), should have much less effect since the time to transmit a single packet is the same for both.

Analytical results show, not unexpectedly, that Demand Multiplexing yields better performance than Synchronous Time Division Multiplexing (see Fig. 8). Further, as the number of terminals served by a loop increases so also does the advantage of Demand Multiplexing (see Fig. 9). However, as the loading for a particular loop configuration increases, the difference in performance between DM and STDM decreases (see Fig. 8).

Results on information storage in the switch are presented in Sections VI and VIII. Our results indicate that, for all but minimum storage allocation, storage in the switch does not markedly affect throughput (see Figs. 12 and 13). The study also showed that, for the message distributions we considered, little is gained by dynamically allocating storage in the switch, as it is needed, holding nothing in reserve. Indeed, in certain circumstances, a static storage assignment does better simply because a static assignment insures reserves in the switch.

Our study showed that, for the loadings under which the system may be expected to operate, the effect of user strategy on switch storage requirements is not pronounced (see Figs. 14 and 15).

III. LOOP MULTIPLEXING

Two techniques for multiplexing data on the line are under consideration, Synchronous Time Division Multiplexing and Demand Multiplexing. In this section we shall present models designed to evaluate the performance of each of these techniques. These models are studied using both mathematical analysis and simulation. Simulation is necessary in situations where mathematical analysis is not possible.

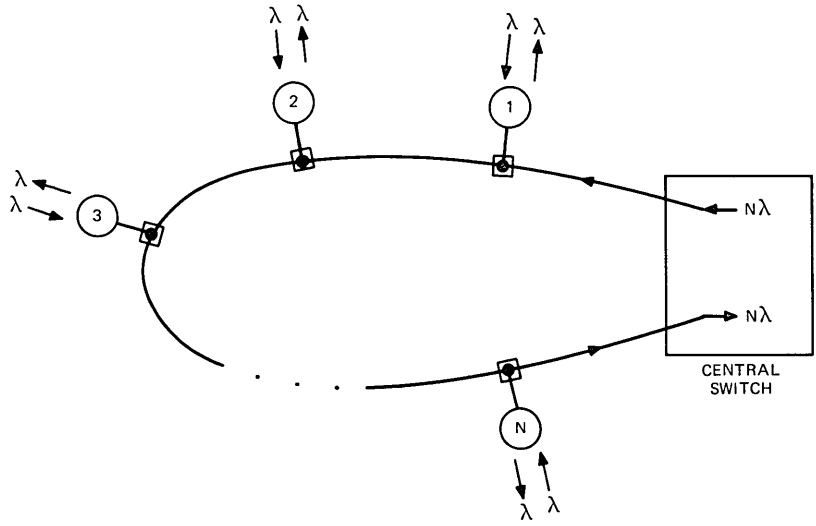

Fig. 3—Loop configuration.

A basic consideration in the design of the system is the response time to interactive users. An important component in this response time is message delay. We define message delay to be the time elapsing between the arrival of a message at a User Terminal and the departure of the last packet of the message from the terminal. Message delay is the criterion that we use to evaluate the multiplexing techniques under study.

We assumed that messages arrive at the User Terminals at a Poisson rate, and that the entire message arrives instantaneously. One would encounter this sort of behavior when a computer outputs directly from its memory to the loop, since the operation of the computer is at a much higher speed than the operation of the loop.

In both of the system configurations under study, the message delay may be broken up into two components, queuing delay and multiplexing delay. Recall that messages arrive at the station at a Poisson rate. Since it takes a nonzero amount of time to multiplex each message, there is a nonzero probability that when a message arrives at the station it must wait until previously arrived messages have been multiplexed onto the line. Once all prior messages have been multiplexed it takes additional time to multiplex the newly arrived message.

The problem of determining message delay for Synchronous Time Division Multiplexing and Demand Multiplexing with adjacent slot seizure is mathematically tractable. In order to study message delay in the DM case with alternate slot seizure, simulation is required.

A sketch of the loop model is shown in Fig. 3. N terminals are connected to the loop. The flow of data is unidirectional around the loop and is shown as counterclockwise in Fig. 3. The first terminal after the switch is labeled terminal number 1, the second terminal number 2, and so on. Messages arrive for multiplexing at a terminal at a rate of λ messages per second. We also assume that messages flow from the switch to each terminal at a rate of λ messages per second. The result is that the volume of traffic flow around the loop is symmetric. The total traffic flow from the switch to the terminals on the loop is $N\lambda$ messages per second. In the case of Demand Multiplexing this flow of return messages affects the operation of the loop. We shall ignore the interaction between the loop and the rest of the system by assuming that return message flow is Poisson. This assumption makes analysis possible and considerably simplifies simulation. We shall return to a consideration of this assumption after we present our models.

3.1 Synchronous time division multiplexing

As we have seen, the bit flow on the T1 loop is formated so that slots, into which information packets can be inserted, flow at a rate of 4000 per second. For Synchronous Time Division Multiplexing, each of these packet slots are assigned to particular terminals on a periodic basis. If there are N terminals connected to the T1 loop and each terminal is accorded equal treatment, then a packet slot is available every $T_c = NT_s$ seconds, where T_s is the duration of a packet slot. In the sequel we shall refer to T_c as the cycle time. For each terminal, we take the end of one cycle and the beginning of the next to be the end of the packet slot assigned to that terminal. We assume that a terminal may always write into its assigned slot even if it is simultaneously reading from the slot.

In order to develop an expression for the delay encountered by a message, let us assume that a message consisting of m_{L+1} packets arrives at a terminal whose buffer is empty, i.e., all previously arriving messages have been transmitted. If the message arrives w seconds before the end of a cycle, then a total of $w + (m_{L+1} - 1)T_c$ seconds elapse before the entire message is transmitted. Now if previous messages have not been transmitted, a newly arrived message suffers queuing delay as well as this multiplexing delay. For the purposes of analysis we categorize the packets of previously arrived messages into two classes: packets held over from previous cycles and packets that have arrived during the present cycle in the time interval $T_c - w$. We may write the total delay queuing and multiplexing as:

$$d_1 = qT_c + T_c \sum_{i=1}^{L} m_i + w + (m_{L+1} - 1)T_c . \tag{1}$$

In eq. (1), q is the number of packets remaining from previous cycles, L is the number of messages arriving in the interval $T_c - w$, and m_i is the number of packets in the ith of these L messages. The mean value of d_1 is shown in the appendix to be

$$\bar{d}_1 = T_c\left(\bar{m} - \frac{1}{2}\right) + \frac{\lambda T_c^2 \overline{m^2}}{2(1 - \lambda T_c \bar{m})} , \tag{2}$$

where $\bar{m}$ is the average message length in packets. Higher moments of d_1 can be found since, in eq. (1), the terms qT_c, $T_c \sum_{i=1}^{L} m_i + w$, and $(m_{L+1} - 1)T_c$ are independent random variables whose moment-generating functions can be calculated (see the appendix). Expressions

for the moment-generating function of d_1 and for the mean-square value of d_1 are given in the appendix.*

3.2 Demand multiplexing

We now consider two implementations of Demand Multiplexing, adjacent slot seizure and alternate slot seizure. With adjacent slot seizure, a typical sequence of information slots leaving the switch might look as shown in Fig. 4. For purposes of explanation in Fig. 4 we assume messages are either three packets long or one packet long. The numbers in the slots correspond to the destination of the packet. No number in a slot indicates that it is empty. When the first three slots shown in Fig. 4 pass terminal 1, it is blocked. Terminal 1 may insert an information packet into slot 4 and into successive slots up to slot 10 when it is again blocked until slot 11. Terminal 1 also removes packets from slots 7 and 9. Terminal 2 removes the packets from slots 1, 2, and 3 and may insert information packets into these slots. Terminal 2 will be blocked when slots 10, 13, 15, 16, and 17 pass. Terminal 2 will also be blocked by terminal 1, if terminal 1 multiplexes packets onto the lines. The same rules apply to all of the other terminals on the loop. A terminal is free to insert data into empty slots and slots from which it removes data. Once an information packet has been inserted into an empty slot, it has priority over any other incoming packets.

Alternate slot seizure is similar except in one important respect. Since the terminal cannot receive nor transmit on adjacent information slots, it is limited to a maximum rate of 2000 packets per second. A typical flow of information is indicated in Fig. 5. Terminal 1 is blocked in slots 1, 3, and 5 but may transmit in slots 2 and 4. If terminal 1 begins by inserting a packet in slot 8, the next slot that is available to it is slot 11. These same rules apply to all of the other terminals in the loop.

The problem of message delay for Demand Multiplexing with adjacent slot seizure has received considerable attention recently.[2–4] Expressions for message delay that are relevant to our study can be found in Ref. 5. In order to make clear the assumptions that led to these expressions we shall sketch the analysis.

Basic assumptions of our study are that each terminal on the loop receives as much traffic as it transmits (see Fig. 3) and may write into slots containing packets addressed to it. Therefore, each

* A general analysis for which STDM is a special case is given in Ref. 1.

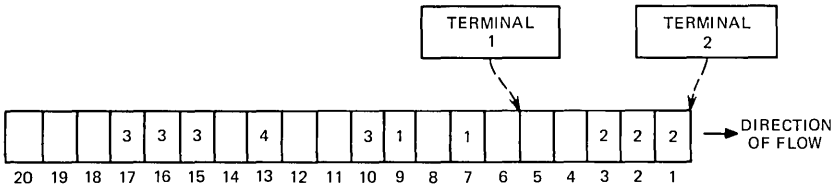

Fig. 4—Flow of data slots, adjacent slot seizure.

terminal "sees" the same traffic volume, $(N - 1)\lambda$ messages per second. The flow of traffic past each terminal consists of alternate busy and idle periods. Messages are multiplexed into line idle periods and are blocked by busy periods.

The probability distribution of message delay depends upon the distributions of the line busy and idle periods. In order to find these distributions, a basic assumption about the nature of the traffic flow out of the switch is necessary. We assume that the line busy and idle periods out of the switch are caused by the Poisson arrival of messages at the switch. This is not difficult to justify when there is light loading on the loops connected to the switch. Under light loading, messages arriving at User Terminals encounter little blocking and are conveyed immediately to the switch. Message arrival at the terminals is Poisson. For the message lengths and loadings we shall consider, the time between message arrivals is large compared to a slot time so that the discretization of message flow on the loop has little effect. As loading increases, the situation is less clear. However, messages arrive from all of the loops connected to the switch, tending to randomize message flow.

A line busy period at the output of the switch is initiated by a message arrival, which under the Poisson assumption is instantaneous. The busy period is lengthened if another message arrives while the previous message is being transmitted. The duration of a busy period is the same as the duration of the busy period of an M/G/1 queue

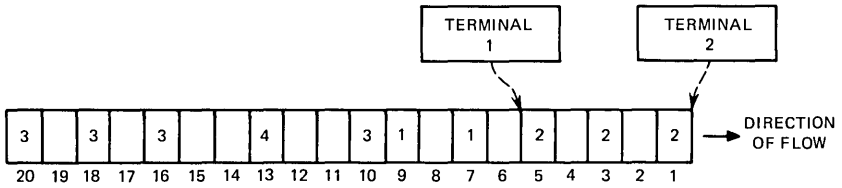

Fig. 5—Flow of data slots, alternate slot seizure.

for which results are well known.[6] Furthermore, under the Poisson assumption, the interval between message arrivals has a negative exponential distribution. Since each terminal sees $(N - 1)\lambda$ messages per second, the duration of a line idle period, as seen by a TIU, is negative exponential with mean $[(N - 1)\lambda]^{-1}$.

We assume that the statistics of line busy and idle periods are the same for all terminals on the loop. The removal of messages from the data stream may affect this assumption. Also, strictly speaking, these statistics also depend upon the multiplexing strategy of the switch. Further work involving simulation must be done to verify this assumption.

Messages are multiplexed packet-by-packet into empty slots. The multiplexing is interrupted by the advent of a busy period and is resumed when the busy period is over. Armed with the statistics of the line busy and idle periods, the message delay can be found. In the language of Queuing Theory, the model is that of a server that suffers periodic breakdowns. The results of the analysis appear as sets of curves that will be discussed presently.

From the foregoing we have an analytical approach for the calculation of delay in the case of Demand Multiplexing with adjacent slot seizure. There are inherent difficulties that preclude an analytical solution in the case of alternate slot seizure. The basic difficulty lies in calculating the durations of line busy and idle periods. For alternate slot seizure a terminal is blocked only if there are two or more interfering terminals. Thus a line idle period is terminated when a terminal begins transmitting, if there is at least one other terminal already transmitting.

Message delay in the case of alternate slot seizure was studied by means of simulation. The simulation program also could be used to study adjacent slot seizure. Although adjacent slot seizure can be analyzed, it was simulated primarily as a check.

In order to insure that the basic assumptions that underlie our model are understood, we outline the basic structure of the simulation program. The number of terminals on the loop is an input variable to the program. Input variables also determine the rate of message arrival, the length of long messages in packets, and the mixture of long and short messages. The simulation was carried out for the variable length message distribution. The basic time unit of the program corresponds to the duration of a packet slot, 1/4000 second. During each time unit a message may arrive to be multiplexed on the

line. The random message arrival is simulated by comparing the output of a pseudorandom number generator to a threshold. If the test indicates that a message has arrived, the length of the message and the terminal to which it arrived are chosen randomly. A basic assumption here is that during a packet slot time no more than one message arrives at all N terminals. For the loadings of interest, this is not a restrictive assumption.

For each of the N terminals in the loop simulation, numbers are stored indicating the current number of packets and messages in the terminal buffer. In each basic time unit, terminal buffers 1, 2, $\cdots$, N are examined in succession until a nonempty buffer is found. If adjacent slot seizure is being simulated, a packet is removed from the first nonempty buffer. In the simulation of alternate slot seizure, a packet is removed from the first nonempty buffer from which a packet was not removed in the previous time unit. After either a packet has been removed from a buffer or all buffers have been examined, the program shifts to the next basic time unit and the cycle repeats beginning with message arrival.

Our interest is in the Nth terminal as it sees traffic from $N - 1$ other terminals. The line busy and idle periods seen by this terminal are measured. The program also measures the number of messages remaining in terminal N's buffer immediately after an entire message has been removed from the buffer. This measurement was not made every time a message departs, but periodically. The length of the period between measurements was varied from run to run. The reason for this is to guard against high correlation between measurements, thereby insuring independent samples.

The measurement of buffer contents upon message departure can be related to message delay. All of these messages remaining have arrived while the departing message was in the buffer. Since we know the rate of message arrival, we can estimate the delay or the length of time the departing message resided in the buffer. Estimates of the mean and the standard deviation of delay so derived will be shown in the next section of this paper.

IV. RESULTS OF LOOP STUDY

Results of the simulation and the analysis of the previous section are shown as curves which show the mean and the standard deviation of delay as a function of loading. From these curves we draw conclusions about the relative merits of the systems under study.

4.1 Demand multiplexing

On Fig. 6 are shown plots of average delay measured in the simulation as a function of loading for 5- and 20-terminal loops. The line loading is the portion of the time that the line is occupied. For comparison the results of the theoretical calculation of average delay is also shown in Fig. 6.

The simulation also yielded estimates of the standard deviation of message delay. Results are shown on Fig. 7, where the standard deviation of delay is shown as a function of loading. As in Fig. 6, the results of analysis for adjacent slot seizure calculations are shown.

In all of the curves the message distribution is the variable length message distribution defined earlier. The resulting average message length is 10.3 packets per message. The message arrival rate at each station in the loop in terms of loading, ρ, is $\rho/(10.3N)$ messages per slot time. (Each slot time is 1/4000 second.) Thus for 0.103 loading on a 20-terminal loop, messages arrive at a rate of 0.0005 message per slot time or 2 messages per second at each terminal.

For each of the simulations care was taken to insure that statistical

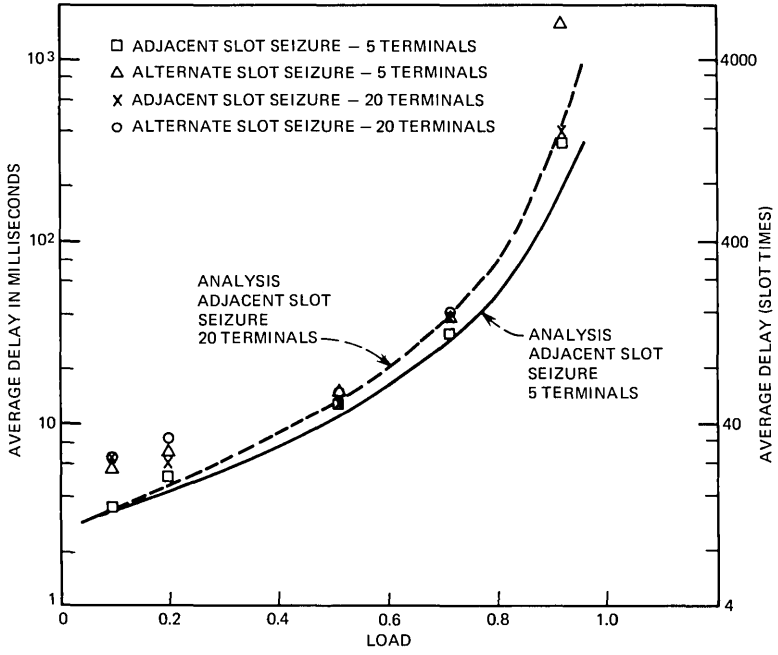

Fig. 6—Simulation results, average delay.

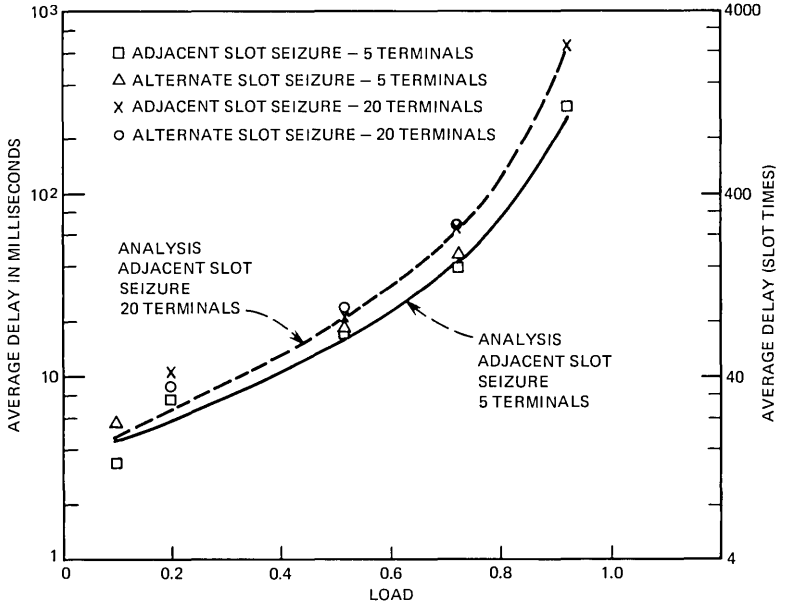

Fig. 7—Simulation results, standard deviation of delay.

equilibrium had been reached. The duration of runs and the random sequences used in the simulation were varied. The standard deviation of the estimates of the mean values of delay shown on Fig. 6 can be estimated. We assume that the measured standard deviations are the true standard deviations. The standard deviation of the mean is then the measured standard deviation divided by the square root of the number of samples. The results indicate that the standard deviation of the mean is small compared to the mean value. The standard deviation is largest relative to the mean at light loadings on the 20-terminal loop where it is approximately 5 percent of the mean.

There is a basic difficulty in making measurements at light loadings on a 20-terminal loop. Due to the relatively low departure rate, fewer independent samples can be gathered. Except for the lighter values of line loading on the 20-terminal loop there is good correspondence between the results of simulation and theory. Even at these lighter loadings the results of simulation are not so far from theory as to cast doubt on the simulation.

The simulation results show that adjacent slot seizure yields somewhat better delay performance than alternate slot seizure. For almost all values of line loading, adjacent slot seizure gives lower values of

mean delay and standard deviation of delay. Moreover, measurements made on loops with 2, 10, and 64 terminals, not shown here, yield much the same result.

The reader will notice, however, that for most values of line loading, the difference between alternate and adjacent slot seizure is not large. The difference is small enough so that ease of implementation should probably determine the choice between the two.

Because of pressures of time we did not compute message delay distributions. However, estimates of the distribution of message delay can be calculated from the simulation values of mean and standard deviation. From the Tchebychev inequality we have

$$P_r[\text{delay} \geqq \mu + k\sigma] \leqq 1/k^2,$$

where μ is the mean and σ the standard deviation of the delay. On a 20-terminal loop with alternate slot seizure this inequality shows that at 0.515 loading 90 percent of the messages suffer delays of less than 83 milliseconds. However, the Tchebychev inequality often gives a rather loose bound. Under a rather tenuous assumption about the distribution of delay, one obtains a more optimistic result. If we compare means and standard deviations of delay we see that, for the same line loadings, the means and the standard deviations are roughly the same. For an exponentially distributed random variable the mean and the standard deviation are equal. If we assume that delay is exponentially distributed we find

$$P_r[\text{delay} \geqq x] = e^{-\alpha x},$$

where $1/\alpha$ is the standard deviation. For 0.515 loading on a 20-terminal loop, 90 percent of the messages have delay less than 50.5 milliseconds.

4.2 Comparison of multiplexing techniques

A second phase of our work on loop multiplexing was devoted to a comparison of Synchronous Time Division Multiplexing and Demand Multiplexing. In order to simplify analysis we have ignored the fact that for DM information packets must contain the address of the transmitting terminal. No such addressing is required for STDM. However, such addressing information is a negligible part of an information packet. For example, for a 64-terminal loop, only 6 bits are necessary to specify the address of a terminal. We feel that the small improvement in accuracy that could be attained by considering addressing did not justify the complications introduced into the analysis.

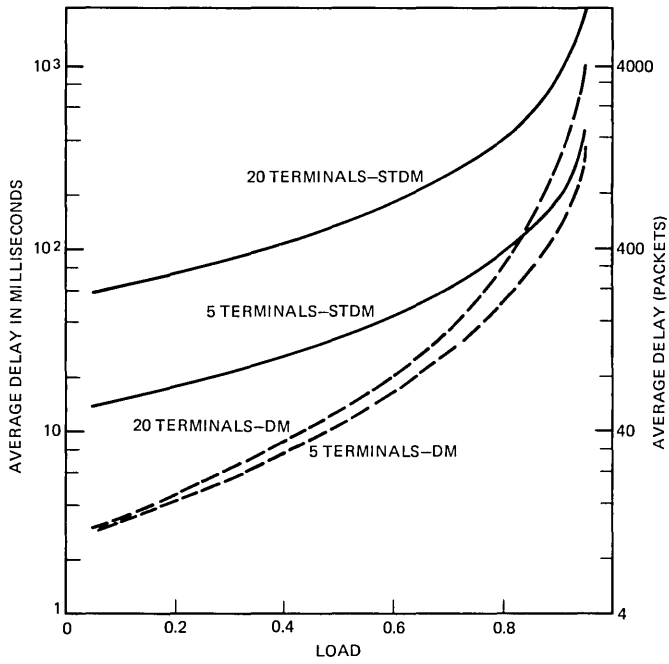

Fig. 8—Average delay versus loading in DM and STDM (30 percent of messages are 32 packets long).

Typical results are shown on Fig. 8 where delay in both packet times and milliseconds is shown as a function of line loading for the variable length message case. As the curves show, DM is clearly superior to STDM. This superiority is more pronounced at the lighter loading, where the multiplexing time is the strongest component of delay. In the absence of interfering traffic, the time required to multiplex a message in DM is an average of 10.3 slot times. In contrast, for an STDM system with N terminals, the average time required to multiplex a message is $10.3 \times N$ slot times. As the loading increases, the difference between the two systems decreases. Line traffic in the DM system interferes with message multiplexing and as the load increases so does the interference.

Similar results have been obtained for the constant length message distribution. Computations of the standard deviation of delay for both constant and variable length message distributions also show the same basic pattern.

Another view of the performance is indicated on Fig. 9 where average delay is shown as a function of the number of terminals in the loop

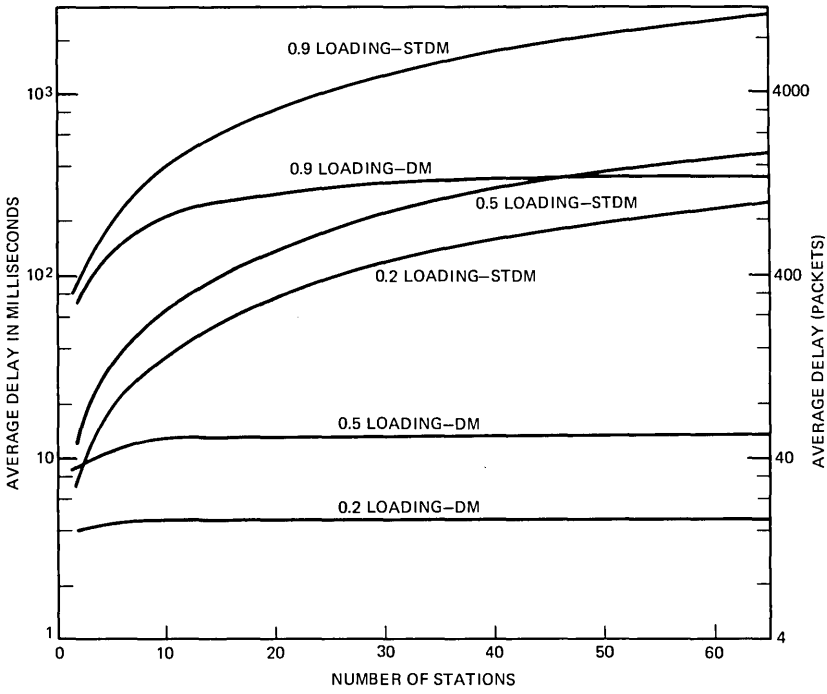

Fig. 9—Message delay versus number of stations (30 percent of messages are 32 packets long).

for fixed values of loading. In Fig. 9 the dependence of delay in the STDM system on the number of terminals in the loop is marked. There is little of this dependence in the case of DM. However, DM is more sensitive to changes in load than STDM. Notice the large jump in delay from 0.5 loading to 0.9 loading in the case of DM. Although we have not shown them, similar results obtain for the constant length message distribution.

V. SWITCH THROUGHPUT

The second phase of our work involved a study of buffering in the switch. Streams of data enter the switch from the loops connected to it. In the DM implementation entering packets are labeled as to originating terminal. In the STDM implementation each terminal has an assigned packet slot recurring periodically. System operation is such that, at any given time, each terminal in the system transmits to and receives from only one other terminal in the system. Information on which pairs of terminals are linked together is stored in the switch.

Therefore, given the origin of an information packet, the switch determines its destination by looking in a table.

A terminal can rapidly change the destination of the packets that it transmits. Stored in the Central Switch is a list of up to 64 possible correspondent terminals for each terminal. A terminal that is transmitting to terminal A, for example, may select a new destination, say terminal B. By means of signal packets (see Section I) the Central Switch is notified of this change in destination. After the change all information packets transmitted from the originating terminal are routed to terminal B. A terminal can select only from the list of its 64 correspondent terminals stored in the switch. However, this list can be altered by the originating terminal when it wishes to make connection with a new terminal or drop connection with an old. Again, signal packets are used to communicate between the terminal and the switch. The process of altering the list requires much more time then switching between terminals already on the list.

At a given instant of time, a terminal transmits to and receives from the same terminal. Further, each terminal in the system acts independently in selecting the correspondent terminal that is the destination of its packets. Thus a terminal may select a destination terminal that is, at that point in time, corresponding with a third terminal. In this event the packets that are transmitted are stored temporarily in the Central Switch. The Central Switch, again using signal packets, notifies the destination terminal that packets from a particular originating terminal are waiting to be delivered. It may happen that, for a particularly busy terminal, there may be messages from several different originating terminals stored in the switch waiting to be delivered. The receiving terminal is free to choose the order in which these messages are read out of the switch buffer.

In connection with this routing and selection procedure we use the term virtual channel as a notational shorthand. As we have seen, each User Terminal has stored in the switch a list of as many as 64 correspondent terminals. When a terminal selects the ith correspondent on this list, we say that the terminal selects the ith virtual channel. When we say that a terminal transmits and receives over virtual channel i we mean that the terminal transmits to and receives from the ith correspondent terminal on the list stored in the Central Switch.

As we have indicated, it may be necessary to store information packets in the switch before they can be delivered. As a practical necessity, the amount of storage in the switch is finite and under heavy loading conditions storage may be used up. In this situation the

switch sends signal packets to User Terminals which inhibit transmission until storage is available in the switch.

Our study of packet storage capacity in the switch focused on two aspects of the problem, throughput and user strategy. Given the random nature of the message flow in the system, there will be occasions when all of the storage assigned to a channel is used up and the transmitting terminal is inhibited. If this condition occurs often enough, there will be a significant effect on the total throughput of data. Secondly, the user through his virtual channel selection strategy can affect the amount of storage that is required in the Central Switch. As we have noted earlier, a certain amount of time is required for a User Terminal to switch from one virtual channel to another. During this switching time, the terminal cannot read packets out of the switch. If User Terminals pursue a strategy calling for frequent switches, demands on switch storage may be too large.

In order to study throughput and the effect of user strategy on buffer requirements, a simplified model was constructed. The model is shown in Fig. 10. N independent data streams carrying λ messages per second flow into N buffers. These data streams represent traffic from correspondent terminals flowing over different virtual channels to the same destination terminal. The destination terminal's changing of virtual channels is represented by the switch in Fig. 10 moving from buffer to buffer. In the model the time required to switch buffers is taken to be either zero or eight packet times (1/4000 second).

In our study of switch throughput, two kinds of buffering were considered, dedicated and common. For dedicated buffering, each of the N buffers is a fixed size. When a buffer is filled, the transmitting terminal is informed and information packets are held at the User Terminal until there is room. In the case of common buffering, a fixed amount of storage is allocated for all N buffers. Each input line uses as much input capacity as it needs. Thus one input line can use up all of the common storage. Again, when there is no more room in the buffer, data flow is inhibited. In our study of throughput, we assume that the entire contents of a buffer are removed before moving on to a new buffer. The order in which buffers are examined is fixed and empty buffers skipped. In a later section we compare this to a strategy where switching takes place after a single message has been read out of a buffer.

A good deal of previous work on buffer occupancy has been based on a Poisson arrival model for messages in the input data stream. In the Poisson model, messages arrive instantaneously with an ex-

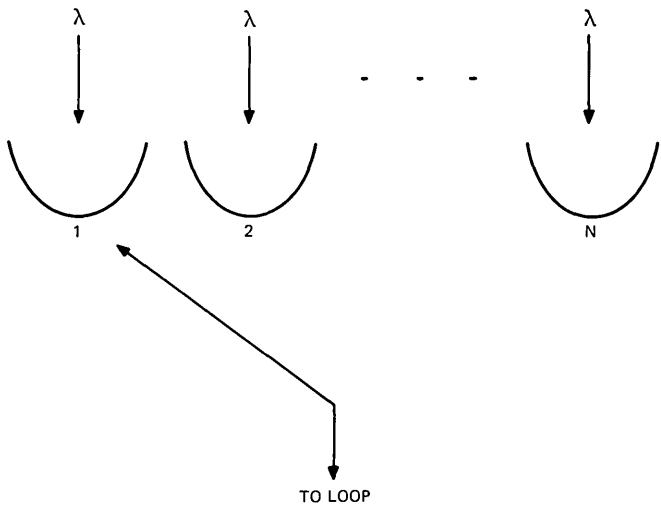

Fig. 10—Model of Central Switch.

ponentially distributed interval of time between messages. A more realistic model, for our study of throughput, is one in which messages arrive over a time interval proportional to the message length with the time between the beginning of one message and the end of the previous message being exponentially distributed. This latter model is more appropriate to buffering in the switch where the arrival and departure of messages is over T1 lines and the read-in and write-out rate of messages is the same. In Section VIII, results based on the continuous arrival model will be compared with the Poisson arrival model.

The model used in our study is indeed something of a simplification. Because terminals share the same T1 loop, messages are likely, especially in heavy loading, to be broken up when they are multiplexed. In our model we do not take this effect into account. For example, in our model a message with 32 packets would occupy 32 successive packet slots on the input line. In the actual system, there may well be gaps in these messages. Similarly, we assume that messages going to a particular destination terminal have sole access to the T1 line, when in fact the line is shared. Thus we assume that messages can be read out of buffers at will. Fortunately, these two effects tend to cancel out. In our model we read in faster and read out faster than reality. Also, we are not looking at absolute measures of performance, but are

comparing different implementations. We felt that a more complicated input output model would not improve this comparison significantly.

Even this simplified model was not amenable to analysis and a Monte Carlo simulation program was written. The basic functions of the program is to measure the throughput as a function of storage capacity and to measure the average occupancy of the buffers. Input variables to the program determine the amount of storage available, the number of input lines and buffers, the time required to switch between buffers, and the probability of message arrival.

As in the program for loop multiplexing, each cycle of the program represents a packet slot time (1/4000 second). In each cycle, the program runs through three distinct parts of the program: input, output, and measurement. In the input portion, each input line is examined in turn. If the line is free, i.e., no message currently being delivered, a random test is performed. This test corresponds to the arrival of a message at a User Terminal in the system under study. If a message has arrived, another test determines its length. If either a new message has arrived or a message is already on the line, the contents of the line's buffer is checked. A packet is inserted in the buffer only if there is room. This packet insertion in the simulation program corresponds to a User Terminal transmitting a packet to the switch. If a line's buffer is full, the input process is suspended. This corresponds to storing a message or part of a message at a User Terminal. Until all of a previously generated message is fed into a buffer, no new messages can arrive.

The program is easily changed to handle either common or dedicated buffering. For common buffering, an input variable is the total storage available. When a packet is inserted in a buffer, this number is reduced by one. For dedicated storage, the storage for each line's buffer is an input variable. As a packet is inserted in a buffer, the amount of storage available for that buffer is reduced by one.

There is a simple relationship between the probability of message arrival and load. Let p be the probability that a message is generated in a slot. It can be shown that the portion of the slots on each input line that are busy is given by the expression

$$\ell = \frac{p\bar{m}}{p\bar{m} + 1 - p},$$

where $\bar{m}$ is the mean length of a message in packets. The assumption here is that there is no limitation on the content of the buffer into which the input line feeds. N input lines feed into the buffers, conse-

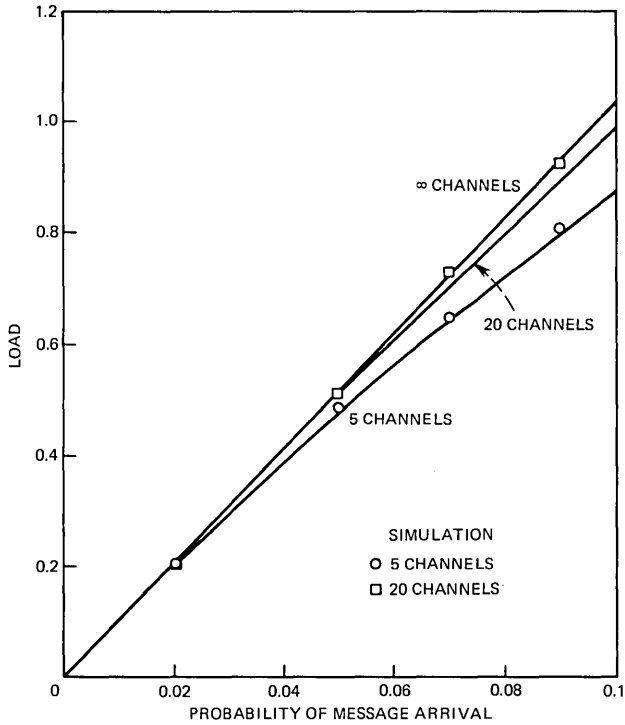

Fig. 11—Line load as a function of the probability of message arrival (30 percent of messages are 32 packets long).

quently the maximum occupancy of the line carrying messages out of the buffers is

$$L = \frac{pN\bar{m}}{p\bar{m} + 1 - p}. \tag{3}$$

The output line will attain this maximum occupancy if no time is required to switch between buffers. The relationship between loading, L, and the quantity pN, which we designate as the probability of message arrival, given in eq. (3) is plotted in Fig. 11.

In the output portion of the program, packets are removed from buffers and placed on the output line. This corresponds to a User Terminal receiving packets that have been stored in the switch. The program examines each of the N buffers in fixed order. Empty buffers are skipped and all of the contents of nonempty buffers are removed. One of the input variables to the program is the time required to switch between nonempty buffers. This corresponds to the time

required by a User Terminal to select a new virtual channel. When a packet is removed from a buffer, the amount of storage available is increased by one up to some fixed amount.

In successive simulation runs the amount of packet storage available was varied with all other parameters held constant. For very large amounts of storage, there is always room in the buffers. As the amount of storage is decreased, it is increasingly likely that packet flow from an input line is inhibited. For relatively low amounts of storage, it will often happen that there is no room in the buffer. In this case, the flow of messages will be halted frequently and the number of packets flowing into the buffers per unit time will be reduced.

In the measurement portion of the program, the main focus was on throughput. Programs were run for 20,000 cycles, and the total number of packets that were fed into buffers were measured. By varying the total amount of storage available, with all other parameters fixed, one obtains the relationship between throughput and storage. Simulation runs were made for the constant and variable length message distributions. Measurements were also made of the total number of messages in the buffers. The results of these latter measurements will be considered in a later section dealing with user strategy.

VI. RESULTS OF SWITCH THROUGHPUT STUDY

Typical results of simulation are shown on Figs. 12 and 13 for 5 and 20 input lines, respectively. In obtaining the results shown on both figures the variable length message distribution was used. The switching time is 8 packet slots. If the line rate is 4000 packets per second, the time required to switch is 2 milliseconds. The curves show normalized throughput as a function of the total packet storage with message arrival probability as a parameter. For each loading the throughput is normalized to the throughput measured at very large storage capacity.

The basic configuration of the curves is as one might expect. As the storage capacity decreases, the throughput decreases. Further, the normalized throughput decreases faster for the larger values of loading.

The results show that, even for a limited amount of storage, the throughput is high. For example, if there are two packet slots for each buffer, the throughput is over 70 percent even for high loading. Results (not shown) for the case of zero switching time show that for this same amount of storage the throughput is over 90 percent.

A surprising result shown on Figs. 12 and 13 is that dedicated storage shows better performance than common storage when there is

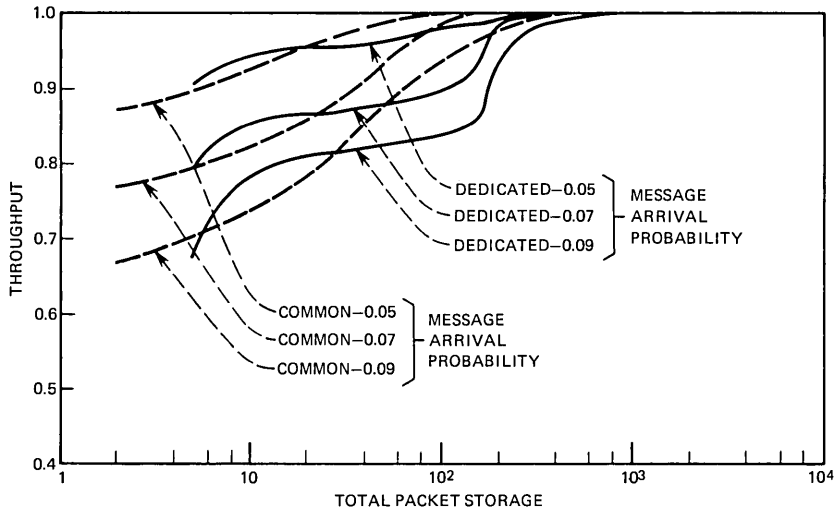

Fig. 12—Throughput versus packet storage for 5 input lines (switching time is equal to 8 slot times).

a limited amount of buffering available. A combination of factors produces this result. First of all, even though storage may be held in common, it is committed to input lines (virtual channels) in a specific

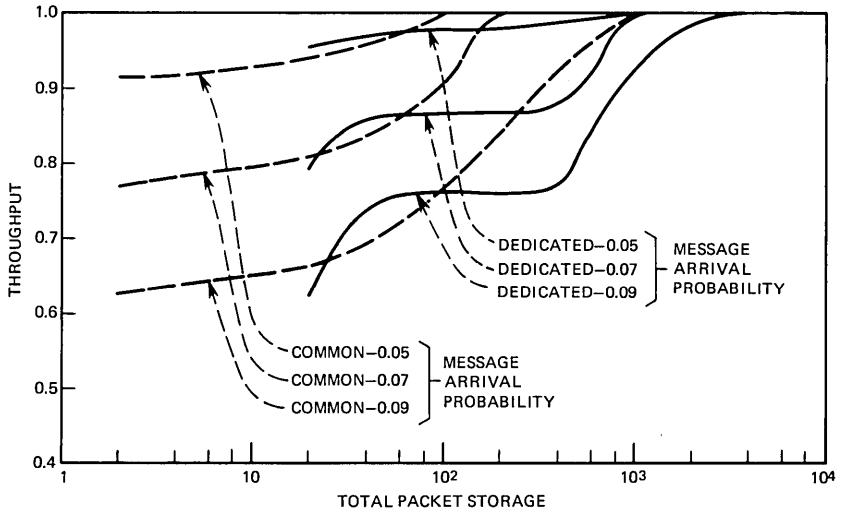

Fig. 13—Throughput versus packet storage for 20 input lines (switching time is equal to 8 slot times).

way that may be far from optimum. The preponderance of traffic is contained in messages that are 32 packets long. If the amount of storage held in common is limited, one channel may absorb all of the storage that is available in the switch. We have also simulated models where all messages are one packet long. In this case common storage is superior to dedicated. However, even in this case the difference between common and dedicated storage is not great.

In the foregoing, messages are generated regardless of whether there is room in the switch or not. We have also examined an implementation where messages are not generated until there is buffer space available. The results of a simulation study of this implementation are essentially the same as the results presented here.

Before concluding this section let us consider the reliability of the foregoing results. First of all, a good many simulation runs were made whose results are not shown here. In these runs, the number of channels, the starting points of the random sequence used in the Monte Carlo technique, and the running time of the simulation were varied. All of the results were in conformity with the results presented in this paper.

Recall that an input parameter to the program was the probability of message arrival. In Fig. 11 the theoretical relationship between loading and this quantity is shown. Also shown on Fig. 11 is the loading obtained in the simulation. As seen on Fig. 11, the results of simulation are within 5 percent of the theoretical values. This is additional indication that the simulation runs were long enough to obtain representative data sequences.

Estimates of the standard deviation of the throughput were made. This was done on each simulation run by measuring the throughput every 1000 cycles, obtaining a sequence of 20 points. (Recall that the simulation runs were 20,000 cycles long.) The mean, μ, and the variance, σ^2, of these points was calculated, giving an estimate of the mean and standard deviation of the throughput in 1000-cycle intervals. The sum of these points is the total throughput for the simulation run. If we assume that the throughputs for successive 1000-cycle intervals form a sequence of independent, identically distributed random variables, the variance of the throughput for a 20,000-cycle run is $20\sigma^2$. The coefficient of variation or the ratio of the standard deviation to the mean for a 20,000-cycle run is

$$\mathrm{CV} = \frac{20\mu}{\sqrt{20\sigma^2}}.$$

Our measurements show that in all cases the coefficient of variation is less than 0.1 and in many cases it is less than 0.05. This result means that with different starting points in the random sequences we would expect a relatively small variation around the points we have plotted on Figs. 12 and 13.

VII. BUFFER STORAGE REQUIREMENTS

At any given point in time, a User Terminal knows which virtual channels have messages waiting to be delivered. A terminal is free to select these virtual channels in any order. We assume that a terminal will not interrupt the reading out of a message in order to switch to a new channel. Since the channel selection procedure is entirely in the hands of the User Terminal, we studied the effect of different strategies on system storage requirements. Accordingly, a calculation of buffer occupancy statistics for different user strategies was performed.

As in the study of throughput we use the model shown on Fig. 10. Again, input lines correspond to virtual channels and switching between buffers corresponds to the selection of new virtual channels. In order to make the analysis tractable, we assume that messages arrive at a Poisson rate of λ messages per second over each input line. Further, we assume that eight packet slot times are required to switch buffers. In order to calculate bounds, we also consider the case where no time is required to switch.

Now if it is known which of the buffers are not empty, the worst strategy, in terms of buffer occupancy, is to always switch after reading a message out of a buffer. Thus, even if messages remain in a buffer, time is wasted in switching to a new buffer. In the sequel we shall refer to this strategy as "1-by-1." In contrast, the most efficient strategy is to cycle through the N buffers skipping empties and reading out the entire contents of nonempty buffers. This latter strategy is the one considered in the previous section on throughput. We shall refer to it as "empty before switch." An intermediate strategy involves switching at random. In this case, after a message has been read out of a buffer, one selects the next buffer at random from those having messages. It can be shown that, if there are N buffers, the probability of switching to a new buffer is $1\text{-}1/N$. Thus with probability $1/N$, two messages are read from the same buffer in succession.

We analyze the buffer occupancy of the 1-by-1 and the random strategies by using the theory of the M/G/1 queue. Messages arrive at all buffers at a Poisson rate $N\lambda$ messages per second. If the time required to switch between buffers is zero, we can take the service

time in both strategies to be either 1 slot time or 32 slot times, depending on whether a message is long or short. An upper bound on storage requirement for the 1-by-1 strategy can be found by assuming that after each message is read out one always switches to a nonempty buffer. We can analyze this situation by adding the switch time to the time required to read out each message. Thus we have read out times of 9 slot times for short messages and 40 slot times for long messages. This is an upper bound because there is a nonzero probability that all of the messages are in the same buffer and there is no reason for the station to select a new virtual channel.

For the random strategy the service time is slightly different than 1-by-1. With probability $1/N$ no switching takes place and the service time is simply the time required to multiplex a message.

Let b be a random variable denoting the number of slots required to read a message out of a buffer, including switching time. We write $b = (m + w)T_s$, where w is the time required to switch in slot times. The random variable w is independent of m. If, in the case of random switching, 8 slot times are required to switch between buffers, the probability that $w = 8$ is $1\text{-}1/N$ and the probability that $w = 0$ is $1/N$. From the analysis of the M/G/1 queue,[5] it can be shown that the mean number of messages in all N buffers is

$$\bar{n}_1 = N\lambda\bar{b} + \frac{(N\lambda)^2\overline{b^2}}{2(1 - \lambda N\bar{b})}, \tag{4}$$

where $\overline{b^i}$ is the ith moment of b and λ is the average message arrival rate in messages per second. The mean-square number of messages in the buffer is

$$\overline{n_1^2} = (N\lambda)^2\overline{b^2} + \frac{3\bar{n}_1(N\lambda)^2\overline{b^2} + (N\lambda)^3\overline{b^3}}{3(1 - \lambda N\bar{b})} + \bar{n}_1. \tag{5}$$

Our primary interest is in the number of data packets in switch buffers rather than in the number of messages. It can be shown that the mean and the mean-square number of packets in all N buffers is given respectively by

$$\bar{p}_1 = \bar{m}\bar{n}_1 \tag{6}$$

and

$$\overline{p_1^2} = \overline{n_1^2}(\bar{m})^2 + \bar{n}_1[\overline{m^2} - (\bar{m})^2]. \tag{7}$$

The foregoing considers packet storage requirements in all of the buffers in the switch. We shall focus our attention on the number of packets in individual buffers assigned to virtual channels. The mean

number of packets in each buffer is simply $\bar{p}_1$ given by eq. (6) divided by the number of buffers N. In order to calculate the variance of the number of packets in individual buffers, it is necessary to assume that the contents of a buffer are independent of the contents of any other buffer. Under this assumption the variance of the number of packets in any buffer is given by

$$V_1 = [\overline{p_1^2} - (\bar{p}_1)^2]/N. \tag{8}$$

7.1 "Empty before switch" strategy

We now consider the strategy in which the User Terminal goes cyclically from buffer to buffer emptying the entire contents of each buffer. The User Terminal will not select a virtual channel which has no messages in its buffer. The analysis of the number of packets in a buffer under this strategy is mathematically difficult. However, we can obtain an upper bound by considering a strategy in which each of the N buffers is examined in turn (even empty buffers). Since time is wasted examining buffers which are known to be empty, the upper bound follows.

The analysis of the cyclic system is contained in Ref. 6. Since the contents of each buffer is random, the time required to cycle through all buffers is a random variable. It can be shown that the mean and the mean-square values of this quantity are

$$\bar{\tau}_c = \frac{N\bar{w}T_s}{1 - N\lambda\bar{m}T_s} \tag{9}$$

$$\overline{\tau_c^2} = \frac{N(N-1)(\bar{w} + \bar{\tau}_c\lambda\bar{m})^2T_c^2 + (\overline{w^2} + 2\bar{w}\bar{\tau}_c\lambda\bar{m} + \bar{\tau}_c\lambda\overline{m^2})T_s^2}{1 - N(\lambda\bar{m}T_s)^2}. \tag{10}$$

In the sequel we shall consider the time to switch between buffers, w, as fixed; therefore, $\bar{w} = w$ and $\bar{w}^2 = w^2$. The quantities $\bar{m}T_s$ and $\overline{m^2}T_s^2$ denote the first two moments of the time required to read a message out of a buffer. The mean number of messages in the buffer is

$$\bar{n}_2 = \lambda\bar{m}T_s + \frac{\bar{\tau}_c^2}{2\bar{\tau}_c}(1 + \lambda\bar{m}T_s). \tag{11}$$

The mean number of packets is

$$\bar{p}_2 = \bar{m}\bar{n}_2. \tag{12}$$

An expression for the mean-square number of packets in each buffer can be derived from the work presented in Ref. 6. This expression is rather lengthy and provides little insight; therefore, we shall omit it.

Results of computation using this expression will be presented in the sequel.

VIII. RESULTS OF BUFFER STORAGE REQUIREMENTS STUDY

The results of computations using the formula derived in the foregoing are shown in Figs. 14 and 15 for 5 and 20 buffers, respectively. In these figures the average occupancy of each buffer is shown as a function of load, which is the product of the message arrival rate and the average time required to read out a message, $N\lambda\bar{m}$. As expected, the lowest buffer occupancy occurs in the case where no time is required to select a new virtual channel. When an 8-slot-time channel select time is required, the technique with the lower occupancy depends upon the loading. At light loading, the "empty before switch" strategy shows poorer performance because time is wasted stopping at empty buffers. It must of course be remembered that this is only an upper bound for the "empty before switch" that selects only nonempty buffers. As the loading increases, there are fewer empty buffers and the performance of the "empty before switch" strategy improves relative to the 1-by-1 strategy.

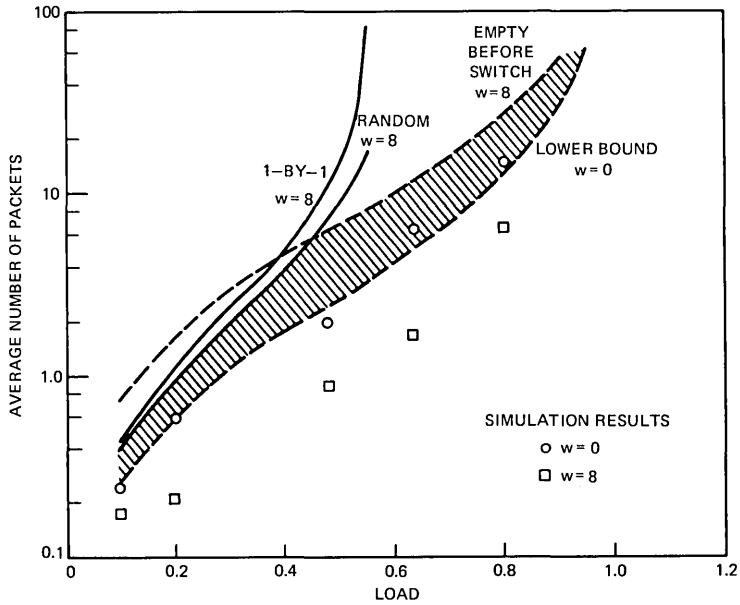

Fig. 14—Average number of packets in each buffer for 5 buffers (30 percent of messages are 32 packets long).

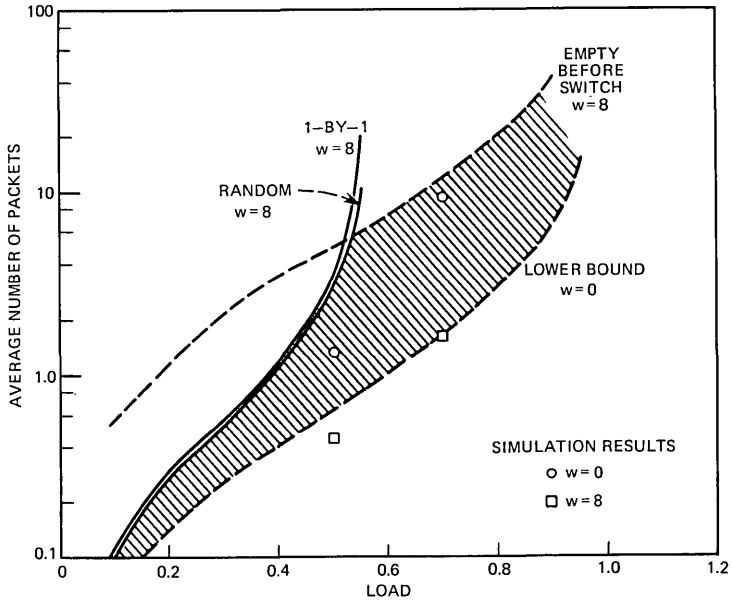

Fig. 15—Average number of packets in each buffer for 20 buffers (30 percent of messages are 32 packets long).

In the previous section we considered an "empty before switch" strategy that skipped empty buffers. As we have mentioned, the problem of calculating occupancy statistics for this technique is intractable. However, the results shown on Figs. 14 and 15 form bounds on the skipping empty technique. The shaded areas in the figures indicate the areas in which the statistics for this method lie.

If the system is operated below 0.5 loading, the difference between the different channel switching strategies is not very large. For example, for 20 channels and 0.4 loading (see Fig. 15) the average occupancy for 1-by-1 rotating strategy is 1.1 packets. For the "empty before switch" strategy skipping empties, the average occupancy is between 0.4 and 1.0 packet. As the load increases beyond 0.5 loading, the 1-by-1 strategy leads to saturation and the cyclic system is clearly superior.

Results for the standard deviations of buffer occupancy have been obtained. These results support the foregoing conclusions.

The simulation program discussed in the previous section computed means and standard deviations of buffer occupancy for an "empty before switch" strategy with skipping of empty buffers. The results are shown on Figs. 14 and 15. A comparison of analysis and simulation

indicates that for the most part the analysis gives upper bounds to the simulation. This is not unexpected since, in the simulation program, messages arrive over an interval of time, whereas for the Poisson arrival model used in the analysis, messages arrive instantaneously.

IX. ACKNOWLEDGMENTS

The system described in this paper was conceived and built by A. G. Fraser.[7] The author would like to express his appreciation to Dr. Fraser for many enlightening discussions on the system while our work was being carried out. We would also like to thank Brian W. Kernighan who, with patience and unfailing good humor, answered many questions on the computer programming involved in the work.

APPENDIX

Delay in synchronous time division multiplexing

Equation (1) of the text is the following expression for the delay in Synchronous Time Division Multiplexing:

$$d_1 = qT_c + T_c \sum_{i=1}^{L} m^3 + w + (m_{L+1} - 1)T_c. \tag{1}$$

The definitions for each of the terms on the RHS of (1) are given in the text.

The delay, d_1, is the sum of the three mutually independent random variables qT_c, $T_c \sum_{i=1}^{L} m_i + w$, and $(m_{L+1} - 1)T_c$. Thus the moment-generating function for d_1 is the product of the moment-generating function for these three variables. In this appendix we shall calculate the moment-generating functions of each of these.

Recall that in STDM dedicated packet slots are available to each terminal cyclically every T_c seconds. We take the end of one cycle and the beginning of the next to be the end of a dedicated packet slot.

Let q_j be the number of packets remaining at the end of the jth cycle and let a_j be the number of packets arriving during the jth cycle. We can write

$$\begin{aligned} q_{j+1} &= q_j - 1 + a_{j+1} && \text{for} \quad q_j + a_{j+1} > 0 \\ &= 0 && \text{for} \quad q_j + a_{j+1} = 0. \end{aligned} \tag{13}$$

Writing this in a more compact form, we have

$$q_{j+1} = q_j + a_{j+1} - U(q_j + a_{j+1}), \tag{14}$$

where $U(x)$ is such that $U(x) = 1$ for $x > 0$ and $U(x) = 0$ for $x \leqq 0$. Taking expectation on both sides of (14) and assuming equilibrium (i.e., $Eq_{j+1} = Eq_j$) we find that

$$E[u(q_j + a_{j+1})] = E[a_{j+1}].$$

But

$$E[u(q_j + a_{j+1})] = P_r[q_j + a_{j+1} > 0].$$

Therefore,

$$P_r[u(q_j + a_{j+1} = 0)] = 1 - \lambda \bar{m} T_c. \tag{15}$$

We now find the moment-generating function. From (14) we have

$$E[e^{-sq_{j+1}}] = E[e^{-sq_j - sa_{j+1} + U(q_j + a_{j+1})}]$$

$$= \sum_{k=0}^{\infty} P_r[\bar{q}_j + a_{j+1} = 0] e^{-s[k - U(k)]}$$

$$= P_r[\bar{q}_j + a_{j+1} = 0] + e^{s} \sum_{k=1}^{\infty} P_r[\bar{q}_i + a_{j+1} = k] e^{-sk}. \tag{16}$$

If we assume equilibrium has been reached, we may define

$$Q(s) = E[e^{-sq_j}]$$

for all j. From (16) after some manipulation we have

$$Q(s) = \frac{(1 - \bar{m}\lambda T_c)(e^{-s} - 1)}{e^{-s} - A(s)}, \tag{17}$$

where $A(s) \equiv E[e^{-a_j s}]$. Since messages arrive at a Poisson rate λ, it can be shown that

$$A(s) = e^{-\lambda T_c [1 - M(s)]}, \tag{18}$$

where $M(s)$ is the generating function of the messages.

By successive differentiation it can be shown that the first two moments of q are

$$\bar{q} = \frac{\overline{a^2} - \bar{a}}{2(1 - \bar{a})}, \tag{19}$$

$$\overline{q^2} = \frac{\overline{a^3} - \bar{a} + 3\bar{q}(a^2 - 1)}{2(1 - \bar{a})}, \tag{20}$$

where $\bar{a}$, $\overline{a^2}$, and $\overline{a^3}$ are the first three moments respectively of the number of packets arriving in a cycle T_c. By successive differentiation

of (18) we find that

$$\bar{a} = \lambda T_c \bar{m}, \tag{21a}$$

$$\overline{a^2} = (\lambda T_c)^2 \overline{m^2} + (\lambda T_c \bar{m})^2, \tag{21b}$$

$$\overline{a^3} = \lambda T_c m^3 + 3(\lambda T_c)^2 \bar{m} \overline{m^2} + (\lambda T_c \bar{m})^3, \tag{21c}$$

where $\bar{m}$, $\overline{m^2}$, and $\overline{m^3}$ are respectively the first three moments of the number of packets in a message.

We turn now to the second term in (1), $\bar{f} \triangleq w + T_c \sum_{i=1}^{L} m_i$. A message arrives at random during a cycle, w seconds before the end of a cycle. In the time interval $T_c - w$, L messages arrive, all of which have priority over the newly arrived message. Since message arrival is random, the quantities L and w are mutually dependent random variables. Conditioned on w, the probability that L messages arrive in the interval $T_c - w$ is

$$P_r[L \text{ messages in } T_c - w] = \frac{\lambda^L (T_c - w)^L}{L!} e^{-\lambda(T_c - w)}.$$

The random variable w is uniformly distributed in the interval $(0, T_c)$. Let $r(t)$ be the density function of the random variable $T_c m_i$. We may write

$$P_r[\tau < \bar{f} \leqq \tau + d\tau] = \frac{1}{T_c} \int_0^{T_c} dw \sum_{L=0}^{\infty} \frac{\lambda^L (T_c - w)^L}{L!} e^{-\lambda(T_c - w)} r^{(L)}(\tau - w), \tag{22}$$

where $r^{(L)}(t)$ is the L-fold convolution of $r(t)$. The Laplace-Stieltjes transform of this can be shown to be

$$F(s) = \frac{e^{-\lambda T_c [1 - R(s)]} - e^{-sT_c}}{T_c \{s - \lambda[1 - R(s)]\}}, \tag{23}$$

where $R(s)$ is the L–S transform of $r(t)$. It can be shown that $R(s) = M(T_c s)$.

The first two moments of f can be found by successive differentiation of (23):

$$\bar{f} = \frac{T_c}{2} + \frac{\lambda T_c^2 \bar{m}}{2}, \tag{24}$$

$$\overline{f^2} = \frac{T_c^5 \lambda^3 (\bar{m})^3 + 3\lambda^2 T_c^4 \bar{m} \overline{m^2} + 3\lambda T_c^2 \bar{f} + T_c^2}{3(1 - T_c \lambda \bar{m})}. \tag{25}$$

The final term to be evaluated in (1) is $(m_{L+1} - 1)T_c$. Since the

generating function of the message is $M(s)$, the generating function of this term is easily shown to be

$$G(s) = e^{sT_c}M(sT_c). \tag{26}$$

The first two moments of g are easily shown to be

$$\bar{g} = T_c(\bar{m} - 1), \tag{27}$$

$$\overline{g^2} = T_c^2(\overline{m^2} - 2\bar{m} + 1). \tag{28}$$

The mean value of delay is the sum of the terms $\bar{q}T_c$, $\bar{f}$, and $\bar{g}$. The variance of the delay can be calculated by summing the variances of qT_c, f, and g.

REFERENCES

1. A. G. Kohneim, "Service Epochs in a Loop System," presented at the 22nd Int. Symp. Computer-Communications Networks and Teletraffic, Polytech. Inst. Brooklyn, Brooklyn, N. Y., April 1972.
2. J. F. Hayes and D. N. Sherman, "Traffic Analysis of a Ring Switched Data Transmission System," B.S.T.J., *50*, No. 9 (November 1971), pp. 2947–2978.
3. R. R. Anderson, J. F. Hayes, and D. N. Sherman, "Simulated Performance of a Ring Switched Data Network," IEEE Trans. Commun., *COM-20*, No. 3 (June 1972), pp. 576–591.
4. B. Avi-Itzhak, "Heavy Traffic Characteristics of a Circular Data Network," B.S.T.J., *50*, No. 8 (October 1971), pp. 2521–2549.
5. D. R. Cox and W. L. Smith, *Queues*, London: Methuen and Co., 1961.
6. J. F. Hayes and D. N. Sherman, "A Study of Data Multiplexing Techniques and Delay Performance," B.S.T.J., *51*, No. 9 (November 1972), pp. 1983–2011.
7. A. G. Fraser, "Interconnecting Computers and Digital Equipment," internal report available upon request.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Peak-Load Traffic Administration of a Rural Multiplexer with Concentration

By S. B. GERSHWIN, R. V. LAUE, and ERIC WOLMAN

(Manuscript received June 19, 1973)

A procedure is proposed for estimating the main-station capacity of an SLM (Subscriber Loop Multiplexer) system by observing the traffic load when the system is partially filled. The procedure is intended to be usable in unattended offices, and requires only one measurement per week and very few calculations. In contrast to the usual practice of measuring load in a time-consistent busy-hour, we work with weekly peak loads, and so our method is based upon the statistical theory of extreme values. The validity and precision of the procedure have been investigated by applying it to data from a study of rural traffic and by a Monte Carlo study of its behavior. Use of this administrative procedure should give the average SLM system a capacity of about 120 rural residential customers, in contrast to the limit of 80 that would be necessary in the absence of traffic measurements.*

The technique described in this paper was developed for the SLM system and could be used, with suitable changes of numerical values, to handle any subscriber system with concentration. We also hope that, with some modification, the method will be applicable to the administration of other traffic-carrying systems.

CONTENTS

* Trademark of the Bell System.

I. INTRODUCTION

The SLM (Subscriber Loop Multiplexer) system is a digital carrier and switching system that was developed to provide economically for main-station growth and upgraded service on long rural cable routes. It is capable of serving 80 lines, all sharing 24 channels. Each of the 80 lines can be used for single- or multi-party service. (For a detailed description of the SLM system see Ref. 1.) For the purposes of this paper, which is concerned with traffic, the SLM system may be viewed as a remote line-concentrator serving 80 lines on 24 full-access channels.

The quality of service given to SLM subscribers should be kept well above levels that might lead to complaints, and to the need for hasty rearrangements that would interfere with the orderly growth of subscriber plant. Hence, service for these subscribers should not be noticeably different from that for customers served by physical pairs to the central office. This service objective will be met if blocking exceeds one-half percent in no more than a few hours per year. (It is possible to imagine a distinct service, for sparsely populated rural areas, in which a less stringent service objective would be appropriate.)

The Rural Line Study,[2,3] a study of subscriber line usage in rural areas, has confirmed that rural residential subscribers like those studied in the territory of South Central Bell can almost always be served on one SLM system with essentially no blocking in groups of 80 main stations or more. (In fact, as shown below, most rural systems should be able to serve many more than 80 main stations.) However, the load per main station does vary greatly from place to place and from customer to customer. Thus even 80 main stations will in a few cases generate enough load to cause undesirably frequent blocking in excess of one-half percent. Some means of monitoring the traffic performance of SLM systems are therefore necessary.

One such means is a register which records the total amount of a system's all-channels-busy time since the register was last read and reset to zero. But such a register, which can indicate by its readings when a system is overloaded, cannot be used to foresee an overloaded condition until too many lines have already been assigned to the system. When the register's readings exceed a specified threshold, the administrator's response must be to remove lines from the system and to serve them on other facilities.* Because of the long lead-times involved in cable planning and installation, a major goal of the work reported here was to avoid this situation.

* Administrative use of this register is described elsewhere.

Since some additional traffic-monitoring capability was necessary, it seemed best to provide a measurement which could be used to *predict* the ultimate main-station capacity of a system, and thus to guide the loading of the system and the planning of relief facilities. The natural quantity to measure is carried load; and the system's second traffic register, which was also chosen to minimize the volume of data to be taken and processed, records peak hourly carried load—that is, the highest hourly carried load that has occurred since this register was last read and reset to zero. As shown below, it is sufficient to take readings once a week. This frequency is compatible with a normal schedule of visits to unattended offices. Note that the time that this peak traffic occurred is neither recorded nor of importance to the procedure to be described, so that the usual time-consistent busy-hour is not identified.

This paper describes a simple system-loading procedure, requiring few measurements and calculations, which should assure that the fraction of calls blocked exceeds 0.5 percent in no more than a few hours during a year. Load measurements are taken on a partially filled system, and from these an estimate is made of the total number of main stations, in the same locale, that the SLM system can safely serve.

This method could be used to administer any subscriber concentrator for which peak-load traffic data can be obtained. Changes in the numbers of lines served or voice channels, in the access afforded by the switching network, or in the handling of intra-system calls would of course require modification of the numerical values used. (And additional problems can arise: For example, partial access may give rise to a need for load-balancing procedures.)

II. THE BASIC PROCEDURE

Weekly peak load measurements can be used whenever 40 or more main stations are assigned to an SLM system. We begin with N weekly peaks, $x_1, \cdots, x_N$. (N is normally equal to 4.) We simply calculate the mean,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{1}$$

and the variance,

$$v = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \tag{2}$$

of these peaks. On a chart (see Fig. 1) corresponding to the number of

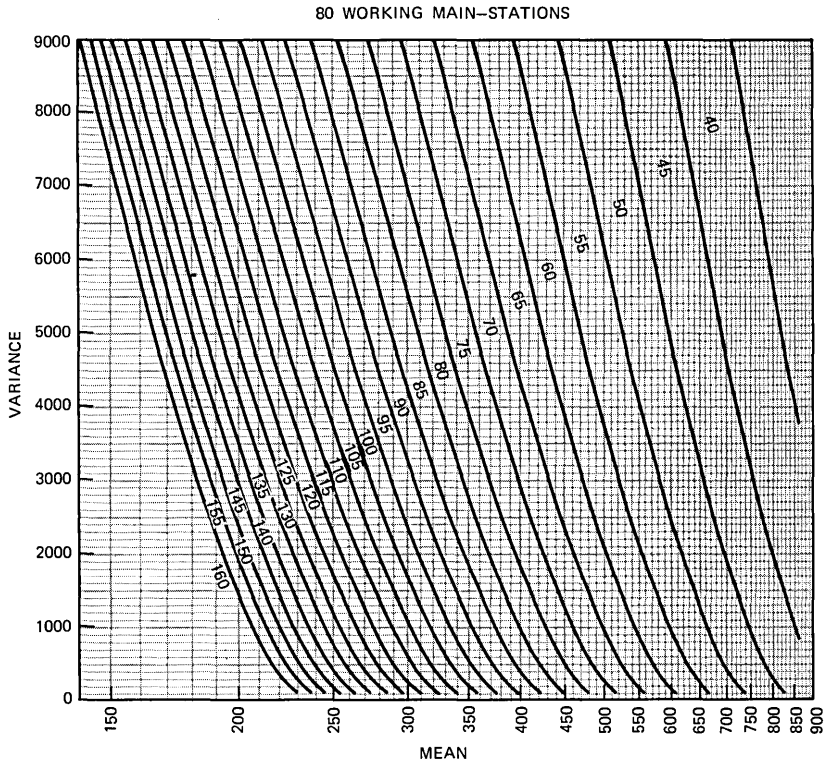

Fig. 1—Estimated main-station capacity as determined by mean and variance of weekly peak loads.

currently working main stations, we find the point defined by the coordinates $\bar{x}$ and v. This point will fall in one of the regions labeled 40, 45, $\cdots$, 160. The label of that region is the estimated main-station capacity of the SLM system.

For example, let us say we have 80 working main stations. We observe four weekly peaks and calculate a mean of 260 CCS and a variance of 1400 $(\text{CCS})^2$. Figure 1 is the chart corresponding to 80 working main stations. The point (260, 1400) falls in the region labeled 120. This is the *estimated* main-station capacity for the system in question. Repeating this estimation procedure four times (using data from 16 weeks of operation) and calculating the weighted mean of the estimates, using the respective numbers of working main stations as weights, we obtain the *predicted* capacity of the system in

main stations. Some precautions which must be observed in drawing conclusions from this process are summarized below.

III. THE DISTRIBUTION OF WEEKLY PEAK TRAFFIC LOADS

Maynard[3] showed that weekly peak traffic loads of potential SLM subscribers seemed to behave as if they came from an extreme-value distribution of the form

$$G(x) = \exp\ (-e^{-\alpha(x-u)}), \tag{3}$$

which is sometimes referred to as "Gumbel's first asymptotic distribution." It has been shown[4] that, for the so-called "exponential" class of distributions, the largest value of a random sample of size n will be asymptotically distributed (as $n \to \infty$) according to (3). The exponential class includes most well-known distributions with an infinite tail to the right, such as the normal, lognormal, and gamma.

The mean of the distribution (3) is

$$E(x) = u + \gamma/\alpha, \tag{4}$$

where $\gamma(=0.5772\cdots)$ is Euler's constant, and the variance is

$$V(x) = \frac{\pi^2}{6\alpha^2}. \tag{5}$$

Gumbel suggests that u and α be estimated by replacing $E(x)$ and $V(x)$ by their sample values, (1) and (2) respectively, and solving (5) and then (4) for α and u:[4]

$$\hat{\alpha} = \frac{\pi}{\sqrt{6v}}, \tag{6}$$

$$\hat{u} = \bar{x} - \gamma/\hat{\alpha}. \tag{7}$$

IV. THE MATHEMATICAL MODEL

With J main stations served by an SLM system, the weekly peaks will have an extreme-value distribution with parameters u_J and α_J. If we increase the number of main stations from J to K, the weekly peaks will have a new extreme-value distribution with parameters u_K and α_K. In this section we describe a method for estimating u_K and α_K when J and K are known and u_J and α_J have been estimated from measurements. That is, we want to know what the distribution of weekly peaks will look like for K main stations when we have observed this distribution with only J main stations being served.

Suppose that, during any week, there are n hours in which the weekly peak traffic load may occur. We know from experience that the weekly peak can occur during almost any waking hour.[3] However, for any given week, n will be much smaller than the number of waking hours. (In Appendix A we show that $n = 10$ seems to be an appropriate choice for our purposes.) We call these n hours (whose actual times of occurrence are not specified) the *candidate* busy-hours.

We now assume that each main station generates a load with mean μ and standard deviation σ during each of the candidate busy-hours. If customers behave independently, the load distribution in candidate busy-hours must have mean $J\mu$ and standard deviation $\sigma\sqrt{J}$. Let this distribution be F, with density $f = F'$. The weekly peak will then be the maximum value in a random sample of size n from the distribution F. If F is in the exponential class of distributions, the distribution of this maximum can be approximated by the extreme-value distribution (3). Gumbel[4] shows that the parameters u and α are given approximately by

$$F(u) = 1 - \frac{1}{n} \tag{8}$$

and

$$\alpha = nf(u). \tag{9}$$

Since the candidate-busy-hour loads are the sums of the loads from J main stations, it seems reasonable to assume that F is normal,* as suggested by the central-limit theorem. (As mentioned above, we take J and K to be at least 40.)

Let Φ be the standard unit-normal distribution function and $\phi = \Phi'$ the corresponding density. Define ν by the relation

$$\Phi(\nu) = 1 - \frac{1}{n}. \tag{10}$$

Then ν is the $1 - (1/n)$ quantile of the unit-normal distribution, for which tables and computer subroutines are available. Then from (8) and (9) it is readily seen that

$$u_J = J\mu + \nu\sigma\sqrt{J} \tag{11}$$

and

$$\alpha_J = \frac{n\phi(\nu)}{\sigma\sqrt{J}}. \tag{12}$$

* The gamma distribution was also considered. A study which led to the choice of the form of the candidate-busy-hour load distribution, and to the number n of candidate busy-hours in a week, is described in Appendix A.

If the $K - J$ subscribers to be added come from the same population as the J subscribers already being served, the candidate-busy-hour load distribution for K main stations will be normal with mean $K\mu$ and standard deviation $\sigma\sqrt{K}$. The weekly-peak-load distribution will have parameters defined by (11) and (12) with J replaced by K. From the four equations (11), (12), and the corresponding equations for K main stations, the variables μ and σ can be algebraically eliminated to yield these expressions for u_K and α_K in terms of u_J and α_J:

$$u_K = \frac{K}{J} u_J - \left[\frac{K}{J} - \sqrt{\frac{K}{J}} \right] \frac{C_n}{\alpha_J}, \tag{13}$$

$$\alpha_K = \sqrt{\frac{J}{K}} \alpha_J. \tag{14}$$

Here $C_n = n\nu\phi(\nu)$, a function of n only. Hence, for a given n, we can estimate the parameters of the weekly-peak-load distribution for K main stations by observing the weekly peaks generated by $J(<K)$ main stations.

We can choose K so that the weekly-peak-load distribution defined by u_K and α_K is such that the system satisfies the service criterion (which is described below). This entire calculation can be incorporated in a series of charts, each corresponding to a different value of J. An example is that given in Fig. 1 for $J = 80$.

V. THE SERVICE CRITERION

To introduce the service criterion, we define a *heavy-load hour* as an hour in which the offered load is such that the probability of blocking is 0.005 or greater. We first attempted to limit the frequency of such hours to no more than four times a year, but found, as shown below, that the statistical characteristics of our administrative procedure lead to a different formulation of the service criterion.

Jones proposed a model for relating the probability of blocking in a concentrator to the number of channels, the number of customer lines, the percentage of intra-system traffic, and the total source load.[5] Johnson has written a computer program which calculates load-service relations based on Jones's model.[6] Although Jones's model assumes blocked calls cleared, whereas waiting and retrials occur in a real system, we believe that these effects are compensated by the lopsided distribution of load per line,[7] so that this model is appropriate. Laue showed from the Rural Line Study that, for 80 main stations, the expected intra-system traffic (IST) should be about 18 percent,

but because of temporal and customer variation we used the more conservative value 30 percent.[2] (In subscriber concentrators which are not arranged for remote switching of intra-system calls,* such calls occupy two channels. Thus for a given offered load, IST increases the variability of the traffic and hence the blocking also.) It is known that the percentage of IST increases with the number of main stations; but the value 30 percent is used throughout because the smoothing effect of party-line interference, which we neglect, grows with the number of main stations.[2] Omission of party-line interference from the model is equivalent to treating a main station as a "traffic source." For 24 channels and an IST of 30 percent, the finite-source effect makes that load which results in a 0.005 probability of blocking a decreasing function of the number of main stations. This load varies from 526 CCS for 40 main stations to 471 CCS for 160 main stations; we call it $L(K)$, the load that causes a heavy-load hour.

From our model we have an estimate of u_K and α_K. The probability of a heavy-load hour in any week for an SLM system with K main stations assigned is then estimated as

$$\hat{P}(K) = 1 - \exp\left(-e^{-\hat{\alpha}_K[L(K)-\hat{u}_K]}\right). \tag{15}$$

Our goal is to let heavy-load hours occur about once every quarter of a year (13 weeks). We now invoke Gumbel's definition of *return period* as the mean time (in weeks) between heavy-load hours, which is

$$R(K) = [P(K)]^{-1}. \tag{16}$$

If we choose K so that the estimated return period is exactly 13 weeks, the temporal variation in the observed weekly peak loads will cause a distribution of return periods, among systems, centered around 13 weeks.

(Note that the expected number of heavy-load hours per year—in other words, the frequency of heavy-load hours—is $52/[R(K)]$ or $52 \cdot P(K)$, so that a 13-week return-period is equivalent to 4 heavy-load hours per year.)

If this distribution—the distribution of return periods that would be realized if K were chosen as just described, based on measured values of u_J and α_J, for each of many systems—were very narrow, so that most systems would turn out to have return periods not far from 13 weeks, then it would be appropriate to choose K in such a way as

* In the SLM system, as in many such systems, the cost of this capability would not be justified; and it would have the further disadvantage of preventing operator access to such calls.

to satisfy the equation

$$1 - \exp\left(-13e^{-\hat{\alpha}_K[L(K)-\hat{u}_K]}\right) = 0.5.$$

This would produce an even chance of having a heavy-load hour in any 13-week period. But as shown in the next section, this situation only occurs when u_J and α_J are estimated from more weeks' peak-load data than are conveniently obtainable in practice.

Furthermore, the scaling procedure described above makes K dependent on the value of J at which u_J and α_J were measured; that is, the charts corresponding to J main stations, of which Fig. 1 is the example for $J = 80$, differ considerably from each other. And this method of determining K is appreciably affected by using the chart for J working main stations (MS) when the actual number of MS served has varied widely during the measurement period, even with the correct mean of J. Thus the number of working main stations must be held nearly constant (actually within 10 percent) during the N weeks of peak-load measurements that yield $\bar{x}$ and v from eqs. (1) and (2). Yet we do not think it practical to impose limits on the growth of an SLM system's fill over periods exceeding four weeks in length, merely so as to obtain usable data, so long as the actual MS fill is not too high.

We resolve this difficulty by setting $N = 4$. This results in a very broad distribution of actual return-periods. We then locate this distribution so that its right-hand tail (in terms of MS capacities) is not too large: In particular, we limit the probability of a return period less than four weeks to a few percent. This is equivalent to setting the mean return-period for all systems equal to 37 weeks, and to having at least one heavy-load hour occur in any 13-week period with probability 0.3. Thus we choose K (to the nearest multiple of 5) so as to satisfy the equation

$$1 - \exp\left(-13e^{-\hat{\alpha}_K[L(K)-\hat{u}_K]}\right) = 0.3. \tag{17}$$

This is the basis for the charts such as that shown in Fig. 1.

The following section shows that this way of estimating K must be repeated four times, and the results averaged, in order to make the resulting distribution of return periods acceptably narrow. When this is done (to obtain what we call the *predicted* capacities), practically no systems should end up with return periods less than four weeks—that is, with more than 1 heavy-load hour per month. This, then, is the service criterion: *No system should have more than 1 heavy-load hour in the average month.* This criterion is extremely conservative: First, it

means that the average system has less than 2 heavy-load hours per year (return period, 37 weeks). Second, it typically implies about one originating call with dial-tone delay and one blocked incoming call *per month* for the worst systems. Systems with more frequent heavy-load hours (return periods less than four weeks) would be called overloaded and would have to have main stations removed. Detection of overloaded systems by means of all-channels-busy readings is covered elsewhere. (Two heavy-load hours can, of course, occur in one month without implying an overloaded condition.) Proper use of the administrative procedure should make the occurrence of overloads extremely rare.

VI. THE MONTE CARLO STUDY

In order to evaluate the statistical variability of the main-station capacities that would be predicted by the recommended procedure, we simulated its performance in a Monte Carlo study. For each of five samples of weekly peak loads actually observed in the Rural Line Study, we estimated the parameters of their distribution from (6) and (7) and converted these, by means of (13) and (14), to estimates of the values u_J and α_J that would have existed with $J = 40$

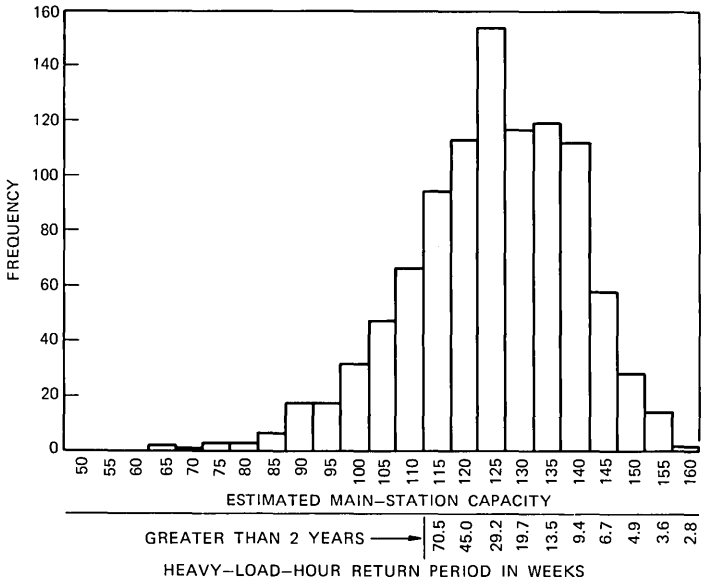

Fig. 2—Monte Carlo results: 40 working main stations; 1000 runs of 4 weeks each.

and $J = 80$ main stations working. Using these as the parameters of the extreme-value distribution (3), we repeatedly generated random samples of size four from that distribution. We calculated the mean and variance of each such sample and estimated the main-station capacity. This process was repeated 100 times for each of the five samples and for both 40 and 80 working main stations, yielding a sample distribution of the estimated main-station capacity. The spread of this distribution is attributable to the variability of the weekly peaks drawn from the distribution (3).

Figure 2 shows an example representing 1000 runs based on the 1FR* data from McComb, Mississippi, described in Ref. 3. Alongside the estimated-main-station-capacity scale is a scale which gives the mean return-period of heavy-load hours for the indicated number of main stations. Note that if the procedure were followed with only four weeks' data, some systems would be overloaded (with return periods of heavy-load hours as short as three weeks) and others, with much spare capacity, would be greatly underloaded. Table IA summarizes the Monte Carlo results for the five samples. The statistics listed are:

(*i*) The percentage of cases that would have a return period of less than the desired 13 weeks.

(*ii*) The percentage of cases that would have a return period of less than 4 weeks. These systems would be considered overloaded.

(*iii*) The percentage of cases that would be underloaded by more than 20 main stations. A system is considered underloaded if it has a return period greater than the central value of 37 weeks.

We see from Table IA that the percentage of systems that would be overloaded ranges from 2 to 14 and the percentage of systems that would be seriously underloaded ranges from 3 to 12. On the basis of these results it was decided that four weeks' data (one *measurement month*) are not enough for a final determination of main-station capacity. We therefore recommend the use of the mean of the main-station-capacity estimates of four samples of four weeks each. (This should be a weighted mean, the weights being the average numbers of working MS in the four measurement months. This weighting accounts for the greater predictive value of an estimate that is based on the traffic of a larger fraction of the stations that will ultimately be served.) A Monte Carlo study was carried out for this procedure and the results are shown in Table IB. Note that an overloaded or badly

* Single-party, flat-rate, residential service.

Table I — Percentage of systems that would have particularly high or low loads — summary of Monte Carlo results

	Return Period <13 Weeks		Return Period <4 Weeks		Underloaded by More Than 20 MS	
Working Main Stations:	40	80	40	80	40	80
A: ESTIMATED (100 Runs of 4 Weeks Each)						
Study Area						
Hanceville	28%	39%	3%	5%	7%	12%
Benton—1FR	21	37	2	4	7	7
Cleveland	28	38	2	14	9	7
Copper Hill	29	43	3	10	3	4
McComb—1FR	19	27	3	5	8	10
B: PREDICTED (1000 Runs of 4 Groups of 4 Weeks)						
Hanceville	11.2%	11.4%	0.0%	0.0%	0.3%	0.6%
Benton—1FR	8.8	19.7	0.0	0.0	0.1	0.5
Cleveland	10.3	22.6	0.0	0.3	0.1	0.1
Copper Hill	11.3	24.1	0.0	0.0	0.0	0.0
McComb—1FR	4.1	13.3	0.0	0.2	0.1	0.1

underloaded system would result very infrequently. Figure 3 shows a histogram of the sample distribution based on the McComb data. We distinguish the result of the modified procedure, using the mean of four estimated capacities, by calling it the *predicted* main-station capacity; and Fig. 3 is so labeled.

The last two pairs of columns in Table I relate to cases in which the administrative system may be said to have performed inadequately. The criterion for this is more stringent on the overload side, as necessitated by the inherent variability of the predicted capacities. This difficulty could be cured by taking many more measurements—a solution whose cost, in our judgment, would not be justified.

An alternative to the averaging procedure would be to calculate the mean and variance of 16 weekly peaks and to use the main-station-capacity charts only once. The precision of this approach was shown by a Monte Carlo study to be comparable to that resulting from averaging the four estimates from groups of four weekly peaks. But, as mentioned above, if a system is being filled during the measurement period, there is a much better chance of the number of working

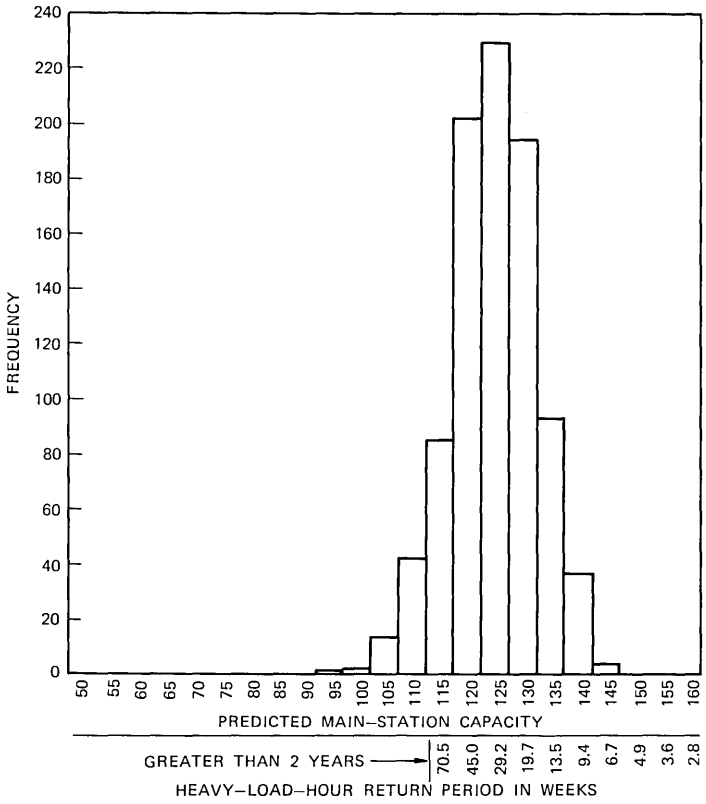

Fig. 3—Monte Carlo results: 40 working main stations; 1000 runs of 4 groups of 4 weeks.

main stations remaining nearly constant for four weeks than for 16 weeks.

The spread of the capacity estimates, as indicated by Table I, was slightly larger with 80 main stations working than it was with 40. We have found no convincing explanation for this unexpected result.

VII. DISCUSSION OF ASSUMPTIONS

Several assumptions used in the foregoing argument are known to be invalid. Nevertheless, the procedure we recommend worked well when applied to the Rural Line Study data. [Some confirmation of its efficacy comes from weekly peak loads observed in an SLM system trial in Brandon, Mississippi, which closely fit the Gumbel distribution (3).]

But in order to distinguish the realistic from the idealized features of our mathematical model, we summarize the latter in this section.

The concept of a candidate busy-hour is fictitious. However, we know of no better means of projecting the distribution of weekly peaks from the current number of working main stations to a larger number of main stations; and it seems intuitively reasonable that not all waking hours of the week should be equally likely to contain the week's highest load. Furthermore, as shown in Appendix A, the procedure does seem to be quite insensitive to the form of the assumed distribution of candidate-busy-hour loads and to the number n of candidate busy-hours. We do not believe that this aspect of our model will cause problems in practice.

It is also assumed that the loads generated during a candidate busy-hour by each of the (existing or future) working main stations are statistically independent and identically distributed. In fact, subscribers in an area do not behave independently, since they are subject to similar influences from events in their shared environment. More important is the critical assumption that the load parameters are identical for all customers. Departures from this assumption among customers already served may actually offset other errors in the model, as noted in Section V; but the possibility of adding to an SLM system a group of subscribers whose traffic characteristics differ considerably from those of the ones already being served constitutes the greatest danger in using the predictive method of administration here proposed. One way of guarding against this danger is to limit the number of main stations that may be added to a system before additional measurements are taken; and a precaution of this kind is included among the rules of administration given in Appendix B. New subscribers should be added to a system even more cautiously if they are thought to differ sociologically from those whose loads have been measured, especially if those to be added have higher incomes. Business telephones, in particular, may generate several times the loads typical of residential service.

The problem of seasonal variation has not been mentioned. No seasonal effects were observed in the Rural Line Study, but some SLM systems, especially in resort areas, will have highly seasonal loads. Our administrative procedure contains no internal safeguards against this source of error. Local knowledge will have to ensure that only weekly peak loads recorded in the busy season are used as inputs to the computations, and in some cases this may force planning to be based on fewer than four 4-week estimates of MS capacity.

In the discussion above, it has been assumed that the number of working main stations remains constant during a measurement month. The case in which inward or outward movement occurs during or between measurement months is covered by some of the rules in Appendix B.

It is not known whether any error is introduced by applying to *peak* loads the load-service relations deduced from Jones's model.

VIII. SUMMARY AND CONCLUSIONS

In this paper we describe a procedure for forecasting the main-station capacity of an SLM system when it is partially filled. The procedure is simple to use, requiring little data and few calculations. Its validity and precision have been shown to be adequate by applying it to the Rural Line Study data (as described in Appendix A) and through a Monte Carlo study of its statistical variability. An SLM system filled in the recommended way should be essentially nonblocking.

In fact, most rural systems will be able to serve many more than the nominal 80 MS. Application of our procedure to the data from the Rural Line Study led to a sample of predicted capacities whose mean slightly exceeds 120 MS, but this figure must be viewed with caution for three reasons: First, this is the mean of the limits imposed by traffic considerations alone; geographical constraints associated with transmission criteria may keep it from being attained in practice. (Lack of demand for multiparty service could also act to limit main-station fills.) Second, the Rural Line Study was conducted in eight rural areas in the territory of South Central Bell; loads would certainly be larger in some suburban applications, and as yet we have no assurance that our data are representative even of rural areas in other parts of the country. And third, the predicted capacities have a very wide spread; a small but significant fraction of systems are predicted to have a capacity of only 80 MS, confirming the early choice of that number as appropriate for rural use. Studies of suburban traffic that are now in progress should lead to an evaluation of SLM main-station capacities in the suburban environment.

In this study we have viewed the capacity of a telephone system in terms of its behavior in the presence of peak rather than average demand. (In this general sense our work has many forerunners, in the Bell System and elsewhere.) Instead of measuring the load in a time-consistent busy-hour, we record only the load carried during the busiest hour of the week. This reduces the measurements required to

only one observation per week by focusing on the hour that is most important to the quality of service, regardless of when that hour occurs. It has the corresponding disadvantage of working with a traffic statistic that is volatile (as compared to the more stable mean, for example) and therefore difficult to predict with accuracy.

Although our procedure was developed for subscriber multiplexers with concentration, we hope that modified versions of it will prove applicable to other traffic-handling systems.

IX. ACKNOWLEDGMENT

M. M. Buchner, Jr., was deeply involved in the planning which led to the idea of using a peak-load measurement for the SLM system.

APPENDIX A

Number of Candidate Busy-Hours and Their Load Distribution

In the body of this paper we assume that the candidate-busy-hour loads are normally distributed. Although the central-limit theorem supports this assumption, the gamma distribution has in some situations been found to be more descriptive of offered loads; and unlike the normal, it does not imply the existence of negative loads. This appendix summarizes a study comparing the normal and gamma distributions as bases for scaling peak loads, and also leading to the choice of $n = 10$ as the number of candidate busy-hours.

The gamma distribution function, also a member of the exponential class, takes the form

$$G(x) = \frac{1}{\Gamma(\eta)\beta^{\eta}} \int_0^x s^{\eta-1}e^{-s/\beta}ds \tag{18}$$

for $x > 0$; $G(x) = 0$ otherwise. The scale of G is determined by the parameter β and the shape by η; the mean and variance are $\beta\eta$ and $\beta^2\eta$ respectively. The distribution G is asymptotically (as $\eta \to \infty$) normal. It is considerably more difficult to manipulate algebraically than is the normal: For example, such expressions as (13) and (14), which are simple for the normal, are not available for the gamma.

With a change of variable in (18), eqs. (8) and (9) become

$$\frac{1}{\Gamma(\eta)} \int_0^{u/\beta} s^{\eta-1}e^{-s}ds = 1 - \frac{1}{n} \tag{19}$$

and

$$u\alpha = n \frac{(u/\beta)^{\eta}e^{-u/\beta}}{\Gamma(\eta)}. \tag{20}$$

The procedure requires that given n, u, and α we solve (19) and (20) for η and β. This was done by using a modified regula-falsi method of iteration. Solving for u and α when η and β are given is somewhat easier: Given η, the value of u/β can be found from a subroutine for the inverse incomplete gamma function, and from u/β the product $u\alpha$ is easily calculated from (20).

It is well-known that the sum of independent, identically distributed gamma variables is also gamma. Hence, if the candidate-busy-hour load distribution for J main stations is gamma with parameters β and η, the individual main-station mean load $\mu = \beta\eta/J$ with variance $\sigma^2 = \beta^2\eta/J$. Thus the candidate-busy-hour load distribution for K main stations is also gamma with scale parameter β and shape parameter $K\eta/J$.

We now have all that is necessary to arrive at u_K and α_K, for both the gamma and normal distributions, if we are given u_J and α_J and any value of n.

In choosing the form of the candidate-busy-hour distribution F and the number n of candidate busy-hours, we used the Rural Line Study data to evaluate the precision and bias of the estimation procedure. We chose 13 groups of lines, varying in size from 54 to 254 main stations, on which we had peak-load data in series whose lengths varied from 14 to 51 weeks. We divided each group in two in such a way that the two subgroups were approximately equal in the numbers of main stations for each class of service. We found the weekly peaks for each subgroup and, combining the two subgroups, for each whole group as well. From the subgroups' weekly peaks we predicted the distribution of the weekly peaks for each whole group and then compared these predictions with the observed whole-group weekly peaks. This was done with both normal and gamma candidate-busy-hour load distributions, for numbers of candidate busy-hours ranging from 6 to 18.

To determine the best model two measures were used: the root-mean-square deviations of the subgroup predicted values from the total-group estimated values of the mean peak load and of the 37-week-return-period load. Table II shows the results of these calculations. Since the gamma and normal assumptions perform about equally well, and since calculation of the charts (such as that shown in Fig. 1) is many times easier, faster, and cheaper with the normal, we chose it as a model for the distribution of candidate-busy-hour loads. The results of our procedure are not very sensitive to the value of n; and

Table II — Variability of predicted peak loads — square-root of the mean-squared deviation (in CCS)

Number of Candidate Busy-Hours	Mean Weekly Peak		37-Week-Return-Period Load	
	Gamma	Normal	Gamma	Normal
4	22	21	24	24
6	17	16	23	23
8	14	14	24	23
10	13	13	25	25
12	13	14	27	26
14	13	15	29	28
16	14	17	31	29
18	15	18	32	31

we took $n = 10$ because this value gave optimal performance in predicting the mean weekly peak load (the more stable statistic) and nearly optimal for the 37-week-return-period load.

Table II also gives us an evaluation of the procedure; and in particular, it tests the assumption of homogeneity within the groups of customers represented. The rms error for the predicted mean weekly peak is 13 CCS, or about four to six main stations. This means that most such predictions should be accurate to within 10 main stations. (Since these results come from 14 to 51 weeks' data, the statistical variability that would result from using only four weekly peaks is not represented here; it was investigated in the Monte Carlo study discussed in the body of this paper.)

Table III shows the predicted means and the whole-group sample means for the weekly peaks of the 13 groups, using the normal model with $n = 10$. There appears to be no bias in the prediction procedure. The last column in Table III shows the predicted main-station capacities for SLM systems if installed in the 13 study areas. These capacities range from 100 to well over 160 main stations. Excluding the five groups that are predominantly four- and eight-party subscribers, the mean main-station capacity is 124.

APPENDIX B

Rules for Traffic Administration *

In order to ensure that the conditions for validity of our mathematical model are satisfied, or nearly so, and to guard against erroneous

* The proposed procedure, together with these operational rules, is now being tested in field use. This appendix is included here only to illustrate problems that arise in reducing the peak-load approach to practice.

Table III — Comparison of predicted and observed peaks (in CCS) from the 13 Rural Line Study groups

Group Location	Number of Main Stations for Each Class of Service			No. of Weeks	Predicted Mean Weekly Peak for Total Group		Whole-Group Sample Mean Weekly Peak	Predicted Main-Station Capacity
	1FR	2FR	4 & 8FR		Subgroup I	Subgroup II		
Cullman, Ala.	49	4	33	44	230	265	246	130
Hanceville, Ala.	54	22	12	28	265	274	278	110
Jones Chapel, Ala.			254	14	413	461	426	>160
Jones Chapel, Ala.		6	114	14	324	308	308	>160
Benton, Tenn.	54			43	192	180	187	115
Benton, Tenn.		68		43	174	188	181	150
Cleveland, Tenn.	72	24		51	348	320	326	110
Copper Hill, Tenn.	71	22		40	345	323	338	100
McComb, Miss.	79			26	253	261	253	120
McComb, Miss.			205	26	461	460	477	135
Tylertown, Miss.	33	40	33	26	234	249	237	>160
Tylertown, Miss.			147	26	261	282	270	>160
Tylertown, Miss.			151	26	353	335	339	>160

predicted capacities when they are not, we give the following rules and guidelines for the use of the administrative procedure.

Data and definitions

Readings should be taken (and the WPL register reset to zero) at the same time every week. When this is not possible, each measurement week must contain no more than six weekdays and no less than four weekdays. When this condition is violated, the data should be discarded.

Measurement weeks need not be contiguous. Studies have shown that there is no serial correlation among weekly peaks. A missing week should therefore not affect the results, so long as there are four weekly peaks in each measurement month.

The number J of main stations associated with a measurement month should be the mean of the numbers of main stations being served at the times the four weekly peak loads were recorded. This mean should be rounded to the nearest multiple of ten when one of the charts such as Fig. 1 is to be chosen.

Inward or outward movement of main stations served by an SLM system will tend to increase the variance of the weekly peak loads and so to decrease the accuracy of capacity estimates. A study has shown that if the number of main stations varies more than 10 percent during a measurement month, the peak-load data should not be used to estimate a main-station capacity.

The MS capacity predicted from several measurement months is the weighted mean of the monthly estimates. The weights are the numbers of working main stations for each measurement month.

No measurements should be used when fewer than 40 main stations are being served. For this reason, the first of the charts for estimating main-station capacities is for $J = 40$ working main stations.

Restrictions on allowed fill

No system should be loaded beyond 60 main stations until the available data (used in accordance with the rules in this appendix, and consisting of at least an estimate based upon one measurement-month) indicate that it is safe to do so. (The Rural Line Study shows that even some rural systems may suffer excessive blocking with 80 main stations.)

No system may serve more than 160 main stations, with this exception: Unusual configurations in which no concentrator-blocking

is possible, such as 24 lines with 8 MS on each, are perfectly acceptable with respect to the considerations treated here.

Except for the initial fill, no more than 40 main stations may be added to an SLM system on the basis of a single predicted capacity. If a predicted capacity exceeds the present fill by more than 40 MS, the system should be allowed to grow by 40 MS and a new series of measurements taken. This rule embodies a compromise between the value of predicting main-station capacities and the danger of adding customers unlike those already served.

If the current main-station capacity estimate was calculated from fewer than four measurement months, the number of main stations may be increased to the lesser of

(*i*) the current main-station-capacity estimate minus 20, and
(*ii*) the mean of the current main-station-capacity estimate and the current number of working main stations.

These restrictions are necessary because of the statistical variability of such estimates.

REFERENCES

1. I. M. McNair, Jr., "Digital Multiplexer for Expanded Rural Service," Bell Laboratories Record, *50*, No. 3 (March 1972), pp. 80–86.
2. R. V. Laue, unpublished reports, 1970 and 1972.
3. J. M. Maynard, unpublished reports, 1971.
4. E. J. Gumbel, *Statistics of Extremes*, New York: Columbia University Press, 1958.
5. A. W. Jones, "Measurement of Non-Random Telephone Traffic with Special Reference to Line Concentrators," Third International Teletraffic Congress, Paper No. 21, Paris, 1961.
6. E. M. Johnson, unpublished report, 1970.
7. W. S. Hayward, Jr., "Traffic Engineering and Administration of Line Concentrators," Second International Teletraffic Congress, The Hague, 1958.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Time Domain Analysis and Synthesis of Notch Filters

By H. ZUCKER

(Manuscript received July 25, 1973)

The response of notch filters to sudden excitations is analyzed. Unit step and stepped trigonometric inputs are considered for the class of filters derived from low-pass networks by a frequency transformation. It is possible in some cases to approximate the transient solutions in terms of Laguerre functions and deduce general properties of notch filters from these solutions. The use of phasing sections to modify the transient response is also examined. It is shown that this method can be used to effectively reduce the overshoot in the response to a stepped trigonometric excitation.

I. INTRODUCTION

In order to accurately measure noise levels in compandored communication systems, it is necessary to set the compandor characteristics at approximately the values associated with signal transmission. This is done by applying a so-called "holding tone" which is subsequently removed by a notch filter incorporated into the measuring set. This investigation originated in connection with the design of such a notch filter for an impulse noise counter.[1] The filter has to meet both frequency and time domain requirements. The frequency response requirements could be readily met with existing filter design procedures. However, the time domain characteristics of notch filters needed investigation with regard to the suitability for the present application. The time domain requirement is that the filter when combined with a C-message weighting filter[2] and excited with a stepped trigonometric time function at the notch frequency should, in the transient state, have only a specified overshoot level. This requirement is imposed by the necessity to distinguish in the measuring set between sudden gain and phase variations and impulse noise.

To examine the transient response of notch filters, a class of such filters derived from low-pass filters by a frequency transformation is

considered. The transient response to a unit step function, and to a stepped trigonometric function with the notch frequency, is expressed in terms of the low-pass impulse response. It is shown that the low-pass and notch filter response are for both excitations related by a Hankel transform. Some general properties of the transient response are deduced by considering notch filters for which the response can be obtained approximately in terms of generalized Laguerre functions.

This investigation shows that notch filters would distort narrow time pulses of duration less than one-half the notch frequency period. The amount of distortion is related to the notch depth. The response of a notch filter to a stepped trigonometric function can be kept to a low level only after a certain time duration, which depends on the filter parameters. The transient response of a notch filter followed by a low-pass filter may still assume large values at relatively short times from the beginning of the response. A method of decreasing the transient response at such times by the use of phasing sections is also presented. The use of phasing sections is of particular importance where it is necessary to modify the transient response without affecting the frequency response.

Although this work is primarily concerned with notch filters, the methods used may also be applied to determine the transient response of high-pass and bandpass filters, when derived from low-pass filters by a frequency transformation.

II. TRANSIENT RESPONSE OF NOTCH FILTERS TO A UNIT STEP FUNCTION

A class of notch filters derived from low-pass filters by a frequency transformation[3] is considered. The transformation corresponds to replacing the inductances and capacitances in the low-pass filters with parallel and series resonant circuits, respectively. Let $T_L(s)$ be the low-pass transfer function in the complex frequency domain s. The transformation is given by

$$s = \frac{\beta z}{z^2 + \omega_o^2}, \tag{1}$$

where $\omega_o = 2\pi f_o$, f_o = notch frequency, and β is a constant. (For low-pass filters normalized such that for $s = j\omega$ the 3-dB bandwidth is at $\omega = 1$, β is the 3-dB circular bandwidth of the notch filter.) The transfer function of the notch filter in the complex frequency domain z, $T_N(z)$, is related to the low-pass transfer function $T_L(s)$ by

$$T_N(z) = T_L\left(\frac{\beta z}{z^2 + \omega_o^2}\right). \tag{2}$$

To investigate the transient response of notch filters, that response is related to the impulse response of the low-pass filter. Such a relationship is obtained by first expressing the transfer function of the low-pass filter in terms of the Laplace transform of the impulse response, $f_L(t)$,

$$T_L(s) = \int_0^\infty e^{-st} f_L(t)dt; \tag{3}$$

then from (2) the transfer function of the notch filter by

$$T_N(z) = \int_0^\infty \exp - \left[\frac{\beta z t}{z^2 + \omega_o^2} \right] f_L(t)dt. \tag{4}$$

The time response, $f_{NS}(t)$, of the notch filter to a unit step function can now be obtained by inversion of the Laplace transformation, through integration in the complex plane over the contour Γ, as follows:

$$f_{NS}(t) = \frac{1}{2\pi j} \int_{\gamma - j\infty}^{\gamma + j\infty} e^{zt} \frac{T_N(z)}{z} dz, \tag{5}$$

where γ is a positive constant. The relationship between step response of the notch filter and the impulse response of the low-pass filter is obtained by the substitution of (4) into (5) yielding

$$f_{NS} = \frac{1}{2\pi j} \int_\Gamma \frac{e^{zt}}{z} \int_0^\infty \exp - \left(\frac{\beta z u}{z^2 + \omega_o^2} \right) f_L(u)du\, dz. \tag{6}$$

In Appendix A it is shown that (6) can be expressed as follows:

$$f_{NS}(t) = \left[\int_0^\infty f_L(u)du \right] \cdot 1(t) - \int_0^t \left[\int_0^\infty f_L(u) \sqrt{\frac{\beta u}{x}} J_1(2\sqrt{\beta x u})du \right] J_o(2\omega_o \sqrt{xt - x^2})dx, \tag{7}$$

where $1(t)$ is the unit step function and $J_n(y)$ is a Bessel function of order n.

In (7) the first term is the undistorted unit step response and the second term expresses the distortion introduced by the notch filter.

The first term can be simplified by observing that from (3)

$$\int_0^\infty f_L(t)dt = T_L(0). \tag{8}$$

Without loss of generality, $T_L(s)$ can be normalized to be equal to unity at zero frequency. The expression in the brackets of the second term contains a Hankel transform,[4] of order one, hence (7) can be

written

$$f_{NS}(t) = 1(t) - 2\int_0^t \psi_1(2\sqrt{\beta x})\sqrt{\frac{\beta}{x}}\,J_o(2\omega_o\sqrt{xt - x^2})dx, \qquad (9)$$

where ψ_1 is the Hankel transform.

The integration in (9) can be performed approximately by using the method of stationary phase. This is a good approximation when the Hankel transform is a slowly varying function in comparison to the Bessel function. It is also shown in Appendix A that with the stationary phase approximation (9) can be approximated by

$$f_{NS}(t) = \left\{1.0 - \frac{\sin\omega_o t}{\omega_o}\int_0^\infty f_L(u)\sqrt{\frac{2\beta u}{t}}\,J_1(\sqrt{2\beta ut})du\right\}\cdot 1(t). \qquad (10)$$

An example considered subsequently shows that (10) is indeed a good approximation for $\beta \ll \omega_o$, i.e., for narrowband notch filters.

III. TRANSIENT RESPONSE OF NOTCH FILTERS TO A UNIT STEPPED SINE FUNCTION

The transient response to this function is obtained in a manner similar to the response to a unit step function. However, a more general response function, $f_{Nm}(t)$, is considered,

$$f_{Nm}(t) = \frac{\omega_o^{2m+1}}{2\pi j}\int_\Gamma e^{zt}\frac{T_L\left(\dfrac{\beta z}{z^2 + \omega_o^2}\right)}{(z^2 + \omega_o^2)^{m+1}}\,dz, \qquad (11)$$

providing for the possibility of a low-pass notch combination. The time response of a notch filter to a stepped sine function with the notch frequency is obtained as a special case, by setting $m = 0$. Proceeding in a manner similar to Appendix A, it can be readily shown that (11) can be expressed as follows:

$$f_{Nm}(t) = \omega_o\int_0^t\left\{\int_0^\infty f_L(u)\left[\frac{\omega_o^2(t - x)}{\beta u}\right]^{m/2}J_m(2\sqrt{\beta ux})du\right\} \cdot J_m[2\omega_o\sqrt{x(t - x)}]dx. \qquad (12)$$

Equation (12) also contains a Hankel transform but of order m. Similarly as before, (12) can be approximated by using the method of stationary phase. It can also be readily shown, from the results in Appendix A, that with this approximation (12) is given by

$$f_{Nm}(t) \approx \sin\left(\omega_o t - \frac{m\pi}{2}\right)\int_0^\infty f_L(u)\left(\frac{\omega_o^2 t}{2\beta u}\right)^{m/2}J_m(\sqrt{2\beta ut})du\cdot 1(t). \qquad (13)$$

It is shown below that, for $m = 0$, (13) is a good approximation for $\beta \ll \omega_o$.

IV. APPLICATION TO PARTICULAR TYPES OF NOTCH FILTERS

4.1 Introduction

To gain some insight into the behavior of the transient response, a low-pass transfer function is chosen for which the integrals (10) and (13) can be evaluated. As mentioned before, these integrals are Hankel transforms. A class of functions for which Hankel transforms can be evaluated in closed form are the generalized Laguerre functions.[5,6] In fact, for these functions the Hankel transforms are self-reciprocal.[5] These functions form orthogonal sets so that, in principle, any passive filter impulse response can be expanded in terms of these functions. However, the impulse response of some low-pass filters may be approximated closely by a single Laguerre function. Such a response can be used to obtain an estimate of the notch filter response.

The low-pass transfer functions considered are of the following form:

$$T_L(s) = \frac{(s/\alpha_2 + 1)^n}{(s/\alpha_1 + 1)^{n+m+1}} \tag{14}$$

with m and n integers, $n, m \geqq 0$, and $\alpha_1 > 0$.

The transfer function (14) is normalized such that $T_L(0) = 1.0$. This function can be considered as consisting of $(m + 1)$ cascaded low-pass filter sections and, for $\alpha_2 = -\alpha_1$, of n phasing sections. For physical realizability of the n cascaded sections, it is necessary and sufficient[7,8] that $|\alpha_2| \geqq \alpha_1$.

The impulse response corresponding to the transfer function (14) is[9]

$$f_L(t) = \frac{n!}{(n+m)!}\alpha_1 \left(\frac{\alpha_1}{\alpha_2}\right)^n (\alpha_1 t)^m e^{-\alpha_1 t} L_n^m[(\alpha_1 - \alpha_2)t], \tag{15}$$

where $L_n^m(x)$ are generalized Laguerre polynomials given by[10]

$$L_n^m(x) = \sum_{k=0}^{n} (-1)^k \left(\frac{n+m}{n-k}\right)\frac{x^k}{k!}. \tag{16}$$

These polynomials are oscillatory for positive arguments and monotonically increasing for negative arguments.

4.2 Unit step response

The integral (10) with $f_L(u)$ given by (15) can be reduced to a tabulated integral for the special case $n = 0$. The applicable integral

is of the form[11]

$$\int_0^\infty e^{-x^2} x^{2n+\mu+1} J_\mu(2x\sqrt{Y})dx = \frac{n!}{2} e^{-Y} Y^{1/2\mu} L_n^\mu(Y). \tag{17}$$

For this special case the low-pass transfer function is

$$T_L(s) = \frac{1}{(s+1)^{m+1}}, \tag{18}$$

where without loss of generality α_1 is set equal to one.

The transient response of the notch filter is obtained by using (10), (15), and (17) and is approximately

$$f_{NS}(t) \approx \left[1.0 - \frac{\beta}{\omega_o} e^{-(\beta/2)t} L_m^1\left(\frac{\beta}{2}t\right) \sin \omega_o t\right] \cdot 1(t). \tag{19}$$

It is of interest to determine the accuracy of (19) for different values of m. This is done in Appendix B for $m = 0$, 1, and 2 where (19) is compared with the exact solutions. It is shown that the accuracy of (19) is of order β/ω_o when expressions multiplied by both $\sin \omega t$ and $\cos \omega t$ are considered. However, expressions multiplied by $\sin \omega t$ only are of order $(\beta/\omega_o)^2$, so that near the maximum amplitudes of the second term in (19), expected near $\cos \omega_o t \approx 0$, the approximation can be said to be of order $(\beta/\omega_o)^2$. Equation (19) is therefore a very

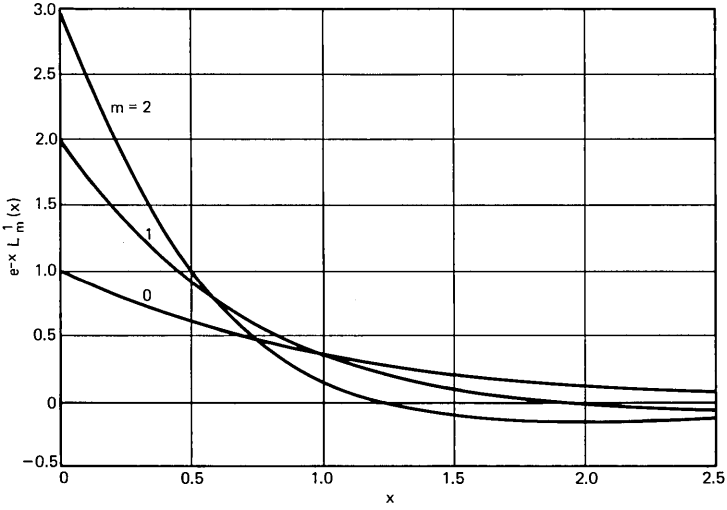

Fig. 1—Laguerre functions, $e^{-x}L_m^1(x)$.

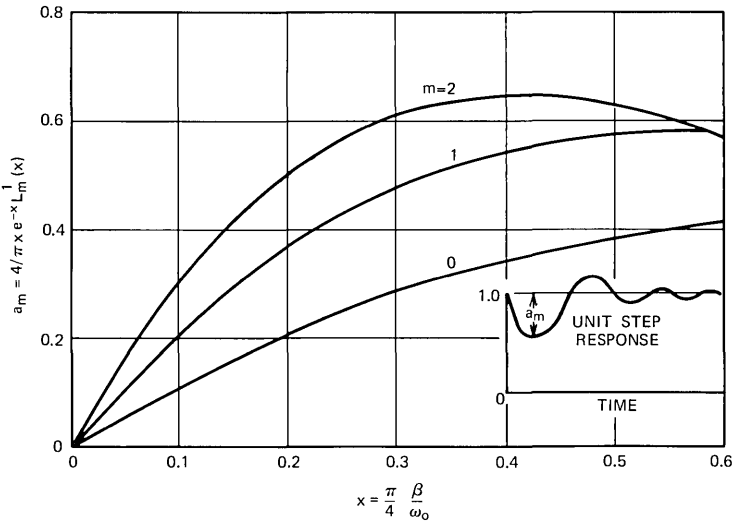

Fig. 2—Values of $a_m(x)$.

good approximation for $\beta/\omega_o \ll 1$, i.e., for notch filters in which the 3-dB bandwidth is much smaller than the notch frequency.

In Fig. 1 the function $e^{-x}L_m^1(x)$ is shown for $m = 0, 1, 2$. It follows from this figure, in conjunction with (19), that the distortion of the unit step function will be particularly pronounced in the time vicinity $T = \pi/2\omega_o$. In Fig. 2 the values of the second term in (19) are plotted as a function of $\beta T/2$ for $m = 0, 1, 2$. This graph gives an estimate of the distortion of the unit step function, and is of particular significance when considering the transient response of notch filters to pulses of short time duration. This investigation shows that pulses of duration less than one-half notch frequency period will be considerably distorted by notch filters.

To obtain a numerical estimate of the distortion, a notch filter with the following requirements is considered: (*i*) Notch frequency—1010 Hz. (*ii*) Notch depth at least −30 dB in the frequency range 995–1025 Hz. A notch filter with these requirements is derived from the low-pass transfer function (18) for $m = 0, 1, 2, 3$.

It readily follows from the transformation (1) that β can be determined from the following equations.

$$\beta = \omega_A 2\pi(f_o - f_1)\left(1 + \frac{f_o}{f_1}\right), \tag{20}$$

where f_o = notch frequency, f_1 is the lower frequency where a notch

depth A[dB] is required, and, from (18),

$$\omega_A = \sqrt{\exp - \left(\frac{A}{10(m+1)} \ln 10 \right) - 1}. \tag{21}$$

The 3-dB bandwidth B of the notch, also obtained from (1), is

$$B = \frac{\beta}{2\pi\omega_B}, \tag{22}$$

where ω_B is given by (21) with $A = -3.0$.

Table I lists the filter parameters, including the maximum value, a_m, of the second term in (19).

It is evident from Table I that the 3-dB bandwidth and the distortion term a_m decrease as the number of sections increases. The difference is particularly pronounced between $m = 0$ and $m = 1$. A further decrease in the distortion term can be obtained by decreasing β; however, this also reduces the depth of the notch.

The listed values of a_m seem to be representative of what is obtainable with notch filters of specified notch depth and 3-dB bandwidth. For example, computation of a two-section notch filter derived from a Tschebyscheff low-pass filter with 0.5-dB ripple gave a value for a_m of 0.18. The notch depth of this filter was also -30 dB and the 3-dB bandwidth 170 Hz. The lower value obtained with this filter can be attributed to the more oscillatory behavior[12] of the low-pass impulse response. Equation (10) suggests such an interpretation.

4.3 Transient response to a stepped sine function

The low-pass transfer function (14) and the corresponding impulse response (15) are again considered. With that impulse response the integral (13) is given by

$$f_{Nm}(t) = \frac{\alpha_1^{n+m+1}}{\alpha_2^n} \frac{n!}{(n+m)!} \left(\frac{\omega_o t}{2\beta} \right)^{m/2} \sin \left(\omega_o t - \frac{m\pi}{2} \right) \cdot \int_0^\infty e^{-\alpha_1 u} u^{m/2} L_n^m[(\alpha_1 - \alpha_2)u] J_m(\sqrt{2\beta u t}) du. \tag{23}$$

The integral (23) can be evaluated in closed form, again yielding Laguerre functions[6]

$$f_{Nm}(t) = \frac{n!}{(n+m)!} \left(\frac{\omega_o t}{2} \right)^m e^{-\beta t/2\alpha_1} L_n^m \left[\frac{\beta t}{2} \frac{(\alpha_2 - \alpha_1)}{\alpha_1 \alpha_2} \right] \cdot \sin \left(\omega_o t - \frac{m\pi}{2} \right). \tag{24}$$

Table I—Notch filter parameters

m	ω_A	ω_B	β[kHz]	B[Hz]	a_m
0	31.61	1	6.00	955	0.527*
1	5.53	0.64	1.05	260	0.273
2	3.00	0.51	0.57	178	0.234
3	2.15	0.43	0.41	149	0.227

* Computed from the exact expression, and occurs at a time, $t = 0.12$ ms.

It is noted that with the condition for physical realizability of the n cascaded sections in (14), $|\alpha_2| \geqq \alpha_1$, that the argument of the Laguerre functions is always positive, and hence the functions are oscillatory.

From (11), (14), and (1), (24) corresponds to the inverse of the following Laplace transform:

$$T(z) = \left[\frac{\omega_o}{z^2 + \omega_o^2}\right]\left[\frac{\omega_o^{2m}}{(z^2 + \omega_o^2 + \beta/\alpha_1 z)^m}\right] \cdot \left[\frac{z^2 + \omega_o^2}{z^2 + \omega_o^2 + \beta/\alpha_1 z}\right]\left[\left(\frac{z^2 + \omega_o^2 + \beta/\alpha_2 z}{z^2 + \omega_o^2 + \beta/\alpha_1 z}\right)^n\right]. \quad (25)$$

The terms in the brackets in (25) can be interpreted as transforms of (*i*) a stepped sine function, (*ii*) an m-section low-pass filter, (*iii*) a notch filter section, (*iv*) n phasing sections for $\alpha_2 = -\alpha_1$ or n additional notch sections for $\alpha_2 \to \infty$.

The transient response of a notch filter followed by a low-pass filter is of interest. However, (25) is restricted to a particular low-pass filter with a high-frequency cutoff in the vicinity of the notch frequency, and will not be considered further.

Setting $m = 0$, (24) simplifies to

$$f_{N0}(t) = e^{-(\beta/2\alpha_1)t} L_n^0\left[\frac{\beta t}{2\alpha_1}\left(1 - \frac{\alpha_1}{\alpha_2}\right)\right] \sin \omega_o t. \quad (26)$$

The special case $\alpha_2 \to \infty$, treated previously for the unit step response, can be obtained from (26) and is in agreement with the result obtained by performing the integration directly by using (17).

For $\alpha_2 \to \infty$ and $m = 0, 1, 2$, a comparison was made between the exact solutions and the approximate solution (26). The comparison showed the same accuracies as for the unit step response.

The time response due to a stepped cosine excitation can be obtained by differentiating (26) and dividing by ω_o. It readily follows

that, within the accuracy of (26), the same expression is obtained but with $\sin \omega_o t$ replaced by $\cos \omega_o t$.

It is noted that (26) gives the correct value for $t = 0$. This value can be obtained by using the initial value theorem of Laplace transforms.[13]

$$\lim_{s \to \infty} sF(s) = \lim_{t \to 0} f(t). \tag{27}$$

Graphs of $e^{-x}L_n^0(x)$ are shown in Fig. 3 for $n = 0, 1, 2$. These graphs give the envelope of the transient response (26) for $\alpha_2 \to \infty$. The effect of finite values of α_2 can be deduced from the graphs. For example, for α_2 negative the arguments of the Laguerre functions increase reaching maximum values of $(\beta t)/\alpha_1$ for $\alpha_1 = -\alpha_2$. Therefore, for the same βt the spacing between the zeros would decrease and the maximum values increase. For positive α_2 the opposite would be the case.

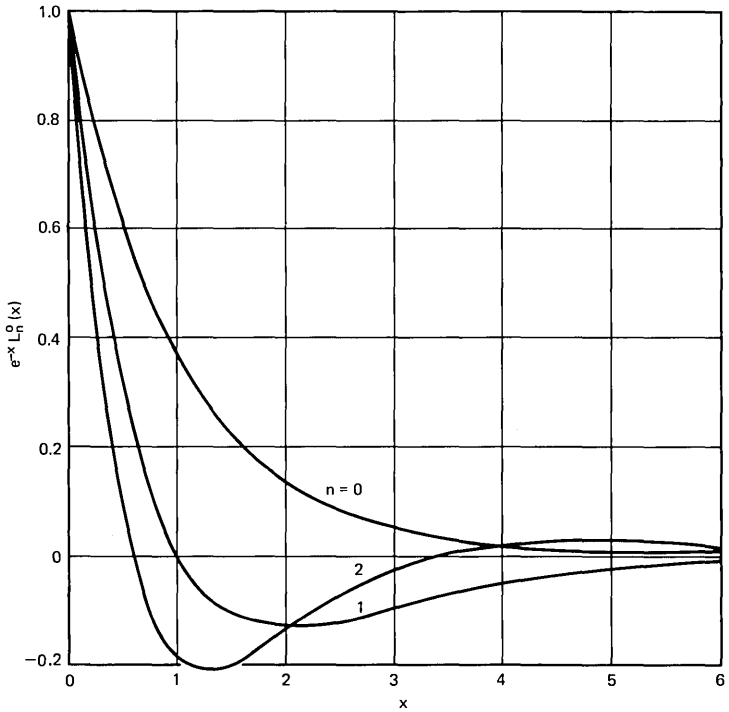

Fig. 3—Laguerre functions, $e^{-x}L_n^0(x)$.

It has been shown above that the transient response of notch filters can be obtained from the Hankel transform of the low-pass impulse response. This property can be used to deduce the qualitative characteristics of notch filters derived from conventional low-pass filters of the Bessel, Butterworth, and Tschebyscheff type. The impulse response of conventional filters, with transfer function polynomials of the same order and approximately the same bandwidth, has similarities to the impulse response considered. Therefore, the graphs shown in Fig. 3 are also representative of the transient response of notch filters derived from conventional filters.

The transient response of notch filters can be kept arbitrarily low at large times, such times being defined after the first zero of the envelope, t_o. The time t_o can be kept small by the choice of the number of sections m, and/or by choosing β/α_1 large. However, for the interval $0 \leqq t \leqq t_o$ these methods are not effective. In fact, it has been shown above from the initial value theorem (27) that, at $t = 0$, the envelope of the response is unity independent of the filter parameters. A low-pass filter combined with a notch filter would cause the transient response to be zero at $t = 0$, and behave, for small t, as t^{k-1} if k is the order with which the transfer function goes to zero as $s \to \infty$. However, with a given low-pass filter the transient response may not be reduced to a desired level at small t. An additional method of reducing the response by the use of phasing sections is discussed subsequently.

4.4 Numerical computations

To illustrate some of the properties of notch filters, numerical computations for a three-section filter ($m = 2$), with the parameters given in Table I, have been performed. Figure 4 shows the filter response to a step function and Fig. 5 the response to a stepped cosine function. The computed results are essentially in agreement with those obtained based on the approximate method. Figure 6 shows the transient response of this filter followed by a C-message weighting filter, when excited with a stepped cosine function. A comparison of Figs. 5 and 6 shows that the C-message filter reduced, as expected, the first lobe of the response, affected only slightly the second lobe and increased the subsequent lobes. Increasing β and hence the 3-dB bandwidth of the notch is not very effective in reducing the second lobe. This led to the investigation of phasing sections as a means of reducing the transient response.

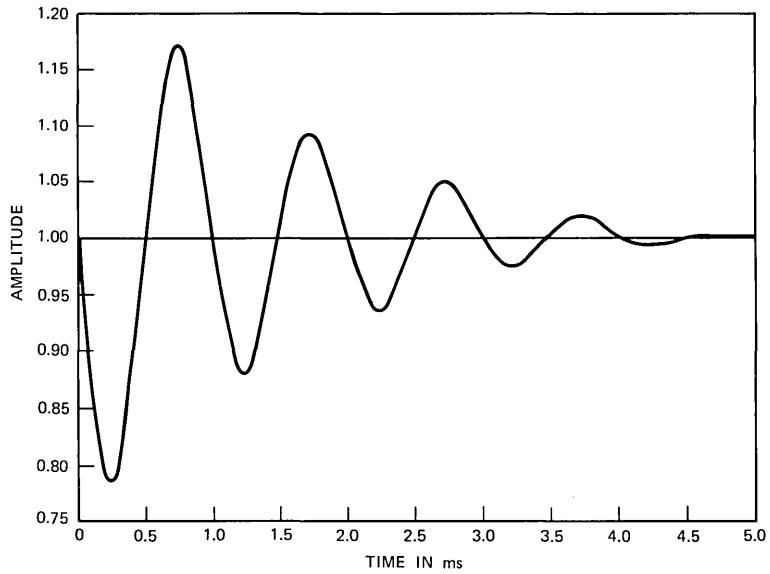

Fig. 4—Unit step response of notch filter.

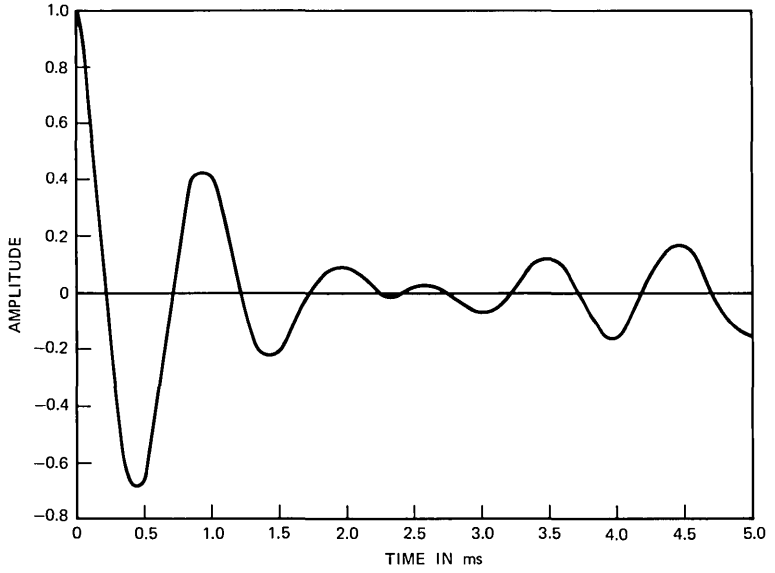

Fig. 5—Stepped cosine response of notch filter.

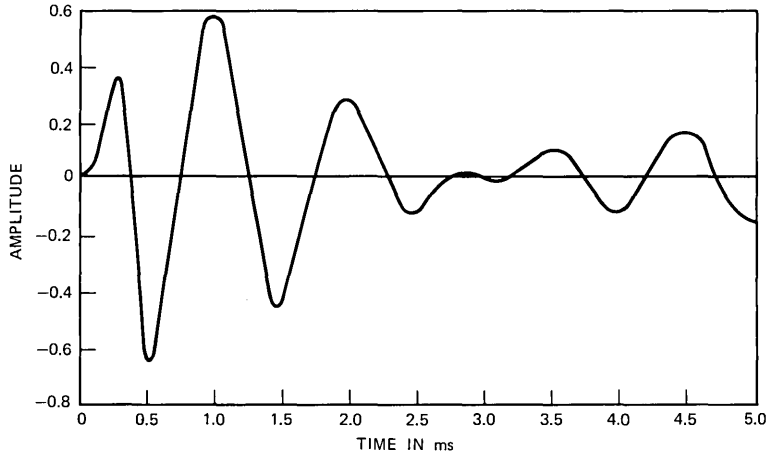

Fig. 6—Response of notch and C-message weighting filters to stepped cosine function.

V. TRANSIENT RESPONSE OF PHASING SECTIONS

The transient response of a phasing section is considered when excited by a damped sine function representing the output of the notched low-pass filter combination. Let the transfer function of the phasing section, $T_p(s)$, be given by

$$T_p(s) = \frac{s^2 - cs + d^2}{s^2 + cs + d^2} \tag{28}$$

and the Laplace transform of the damped sine function, $F_d(s)$, by

$$F_d(s) = \frac{\omega_d}{s^2 + as + b^2} \tag{29}$$

with $\omega_d = \sqrt{b^2 - (a/2)^2}$. The Laplace transform of the response, $F(s)$, is

$$F(s) = \frac{\omega_d}{s^2 + as + b^2} \frac{(s^2 - cs + d^2)}{s^2 + cs + d^2}. \tag{30}$$

Equation (30) can also be written

$$F(s) = \frac{\omega_d}{s^2 + as + b^2} - 2c\omega_d \left[\frac{A\left(s + \frac{a}{2}\right) + B}{s^2 + as + b^2} + \frac{C\left(s + \frac{c}{2}\right) + D}{s^2 + cs + d^2} \right], \tag{31}$$

where the constants A, B, C, and D are obtained by comparing (30) and (31).

It can be readily shown that the time response corresponding to (31), $f(t)$, is

$$f(t) = e^{-(a/2)t}\left[\sin \omega_d t - \frac{2cb}{r}\sin(\omega_d t + \theta_o)\right] + \frac{2cd}{r}\frac{\omega_d}{\omega_1}e^{-(c/2)t}\sin(\omega_1 t + \theta_1), \quad (32)$$

where

$$r = \sqrt{(d^2 - b^2)^2 + (c - a)[b^2 c - d^2 a]}, \quad (33)$$

$$\tan\theta_0 = \frac{(d^2 - b^2)\omega_d}{b^2(c - a) - a/2(d^2 - b^2)}, \quad (34)$$

$$\tan\theta_1 = \frac{(d^2 - b^2)\omega_1}{d^2(c - a) - c/2(d^2 - b^2)}, \quad (35)$$

and

$$\omega_1 = \sqrt{d^2 - \left(\frac{c}{2}\right)^2}. \quad (36)$$

The first term in (32) is the damped sine function and the other two terms are introduced by the phasing section. In order that the last two terms be of significance, it is necessary that these terms be comparable to the first term. This will be the case for $d = b$, for which (32) reduces to

$$f(t) = e^{-(a/2)t}\frac{c + a}{a - c}\sin \omega_d t - \frac{2c}{a - c}\frac{\omega_d}{\omega_1}e^{-(c/2)t}\sin \omega_1 t. \quad (37)$$

If, in addition, $(a/2)^2 \ll b^2$ and $(c/2)^2 \ll d^2$, (37) simplifies further and, for $(a/2 - c/2)t \ll 1$, (37) is approximately given by

$$f(t) \approx e^{-a/2t}\sin bt(1 - ct). \quad (38)$$

Equation (38) contains the damped sine function but modified by the term $(1 - ct)$. This term can be used to introduce a zero in the time vicinity where the damped sine function assumes a maximum value.

To illustrate the above, a phasing section is introduced to modify the impulse response of a C-message weighting filter. The computed impulse response of the filter is shown in Fig. 7 and has relatively large values in the time vicinity of 0.4 ms. The impulse response modified by a phasing section is shown in Fig. 8. The phasing section parameters are $c = 2 \cdot 10^3$ and $d = 10^8$. These parameters have been chosen on the basis of the above analysis. It is evident the large values of the response have been reduced, but the modified response has

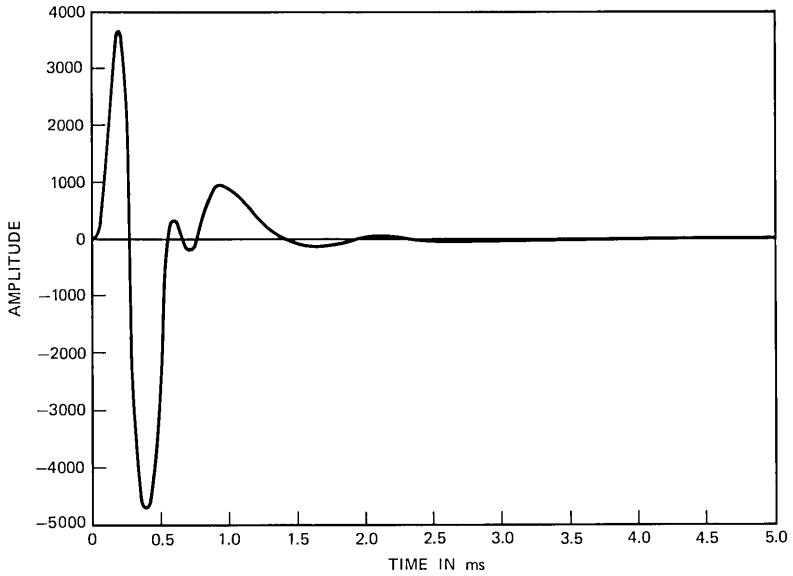

Fig. 7—Impulse response of C-message weighting filter.

appreciable values for a much longer time duration than the initial response. This behavior can be explained on the basis of Parseval's theorem,[14] since the absolute value of the Fourier transform of the response is the same with and without the phasing section.

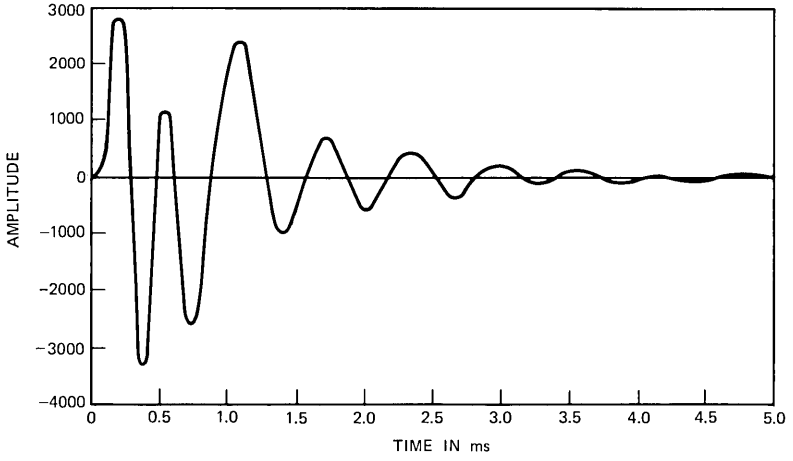

Fig. 8—Impulse response of C-message weighting filter with phasing section.

Table II—Two-section filter with phasing section (3-dB bandwidth 390 Hz, 30-dB bandwidth 80 Hz)

i	Parameters in kHz Units			
	σ_{ni}	ω_{ni}^2	σ_{di}	ω_{di}^2
1	0.09	40.27	1.35	27.91
2	0.09	40.27	1.94	51.10
3	−1.8	100.00	1.8	100.00

Phasing sections can be used to reduce the overshoot of the response of a notch filter followed by a low or bandpass filter and excited with a stepped trigonometric function at the notch frequency. Such sections are of particular importance where the transient response has to be modified without affecting the amplitude of the frequency response or where a modification is needed at times shortly after the beginning of the response. However, such sections may also introduce considerable distortions of the unit step response.

As an example, the performance of a 1010-Hz notch filter with and without a phasing section is considered. The transfer function, $T(z)$, of the filter and phasing section can be written as

$$T(z) = \prod_{i=1}^{3} \frac{z^2 + \sigma_{ni}z + \omega_{ni}^2}{z^2 + \sigma_{di}z + \omega_{di}^2}. \tag{39}$$

The parameters in (39) are listed in Table II.

This filter was derived from a Tschebyscheff low-pass filter, and an operational amplifier version was synthesized and built. Figures 9a and 9b show the computed response to a stepped cosine of the filter combined with a C-message weighting filter without and with the phasing section. Figures 9c and 9d show photos of the corresponding oscilloscope displays obtained with the actual filters. Good agreement was obtained between the computed and measured response. The effect of the phasing section on the response is evident in this figure. About a 4-dB reduction in the overshoot was obtained with the phasing section.

VI. CONCLUSIONS

The transient response of a class of notch filters which are derived from low-pass filters by a frequency transformation was investigated. General expressions for the transient response due to a unit step

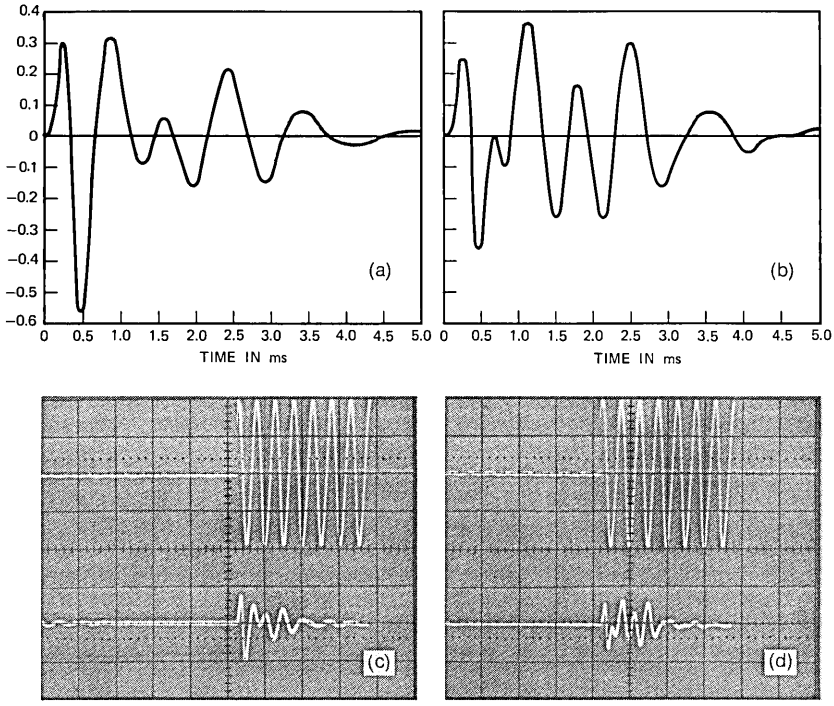

Fig. 9—Computed and measured response of notch filter with C-message weighting filter to a stepped cosine: (a) and (c) without phasing section, (b) and (d) with phasing section.

function and a stepped trigonometric function have been obtained in terms of the low-pass impulse response.

The transient response of certain types of notch filters can be formulated approximately in terms of Laguerre functions. These filters have been examined in detail and some general properties of the notch filter response have been deduced from this formulation.

Notch filters may considerably distort short time pulses (time duration less than one-half notch frequency period). The amount of distortion depends on the notch depth.

The response of notch filters to stepped trigonometric functions can be kept at low levels only after a certain time interval from the beginning of the response. The length of the time interval depends on the filter parameters.

A method of reducing the transient response at short time intervals by the use of phasing sections was presented. This method may prove

particularly useful in applications where it is necessary to modify the transient response without affecting the frequency response.

VII. ACKNOWLEDGMENTS

The author wishes to express his gratitude to E. R. Nagelberg for his critical reading of the manuscript and for his helpful comments and suggestions, Mrs. A. M. Franz for the computational assistance, Mrs. E. Y. McBride for the synthesis of the operational amplifier version of the notch filter and phasing section, and R. R. Redington who built the filter and measured its characteristics.

APPENDIX A

Transient Integrals

A.1 Unit step response

The relationship between the step response of the notch filter and the impulse response of the low-pass filter from which the notch filter is derived is given by (6) of the text,

$$f_{NS}(t) = \frac{1}{2\pi j} \int_\Gamma \frac{e^{zt}}{z} \int_0^\infty \exp\left(-\frac{\beta z}{z^2 + \omega_o^2} u\right) f_L(u)du\, dz. \quad (40)$$

After interchanging the order of the integrations and expanding the exponential function, (40) can be written

$$f_{NS}(t) = \frac{1}{2\pi j} \int_0^\infty f_L(u) \int_\Gamma \frac{e^{zt}}{z} \sum_{n=0}^{\infty} \frac{(-\beta u)^n}{n!} \left(\frac{z}{z^2 + \omega_o^2}\right)^n dz\, du. \quad (41)$$

Equation (41) contains a sum of inverse Laplace transforms of a tabulated form [15,16]

$$\frac{1}{2\pi j} \int_\Gamma \frac{e^{zt}}{z^{2v+1}} G\left(z + \frac{\omega_o^2}{z}\right) dz = \int_0^t \left(\frac{t-u}{\omega_o^2 u}\right)^v J_{2v}[2\omega_o\sqrt{ut - u^2}]g(u)du, \quad (42)$$

where $g(u)$ is the inverse Laplace transform of $G(z)$.

Equation (42) is of its own interest, since it may be used to obtain the step response of a bandpass filter derived from a low-pass filter by a frequency transformation. A direct derivation of (42) follows.

Using the definition of $G[z + (\omega_o^2)/z]$, (42) can be written

$$\frac{1}{2\pi j} \int_\Gamma \frac{e^{zt}}{z^{2v+1}} G\left(z + \frac{\omega_o^2}{z}\right) dz = \frac{1}{2\pi j} \int_\Gamma \frac{e^{zt}}{z^{2v+1}} \int_0^\infty \exp\left[-\left(z + \frac{\omega_o^2}{z}\right) u\right] g(u)du. \quad (43)$$

After interchanging the order of integration and expanding the exponential function in a power series, (43) can be written

$$\frac{1}{2\pi j}\int_\Gamma \frac{e^{zt}}{z^{2v+1}} G\left(z+\frac{\omega_o^2}{z}\right) dt = \frac{1}{2\pi j}\int_0^\infty g(u)\int_\Gamma e^{z(t-u)} \sum_{m=0}^{\infty} \frac{(-\omega_o^2 u)^m}{m!\, z^{m+2v+1}}\, dz\, du. \quad (44)$$

The inverse of each term in (44) is zero for a negative exponential argument. For a positive argument $u \leqq t$, the inverse is readily obtained; hence,

$$\frac{1}{2\pi j}\int_\Gamma \frac{e^{zt}}{z^{2v+1}} G\left(z+\frac{\omega_o^2}{z}\right) dz = \int_0^t g(u) \sum_{m=0}^{\infty} \frac{(-1)^m(\omega_o^2 u)^m}{m!} \frac{(t-u)^{m+2v}}{(m+2v)!}\, du. \quad (45)$$

The summation of the terms in (45) gives a Bessel function [17] of order $2v$, and hence (42). Using (42), (41) can be written

$$f_{NS}(t) = \int_0^\infty f_L(u)du + \int_0^\infty f_L(u)\int_0^t \sum_{n=1}^{\infty} \frac{(-\beta u)^n}{n!} \frac{x^{n-1}}{(n-1)!} \cdot J_o(2\omega_o\sqrt{xt-x^2})dx\, du. \quad (46)$$

The series in (46) can be summed yielding a Bessel function of order one; hence,

$$f_{NS}(t) = \int_0^\infty f_L(u)du - \int_0^t \int_0^\infty f_L(u)\sqrt{\frac{\beta u}{x}}\, J_1(2\sqrt{\beta x u})du \cdot J_0(2\omega_o\sqrt{tx-x^2})dx. \quad (47)$$

The integration with respect to u can be interpreted as a Hankel transform of the low-pass impulse response. For an impulse response which is not very oscillatory, and for $\beta \ll \omega_o$, the Hankel transform will be a slowly varying function in comparison to the Bessel function. Under these conditions an approximation to (47) can be obtained by using the method of stationary phase.

A.2 The stationary phase approximation

The stationary phase method[18] approximates integrals, I, of the following type:

$$I = \int_a^b g(x)e^{jk\psi(x)}dx, \quad (48)$$

where k is large and $g(x)$ is a slowly varying function. The approximation considers only contributions from the vicinity of stationary

points where $(d\psi)/dx = 0$, and is of order $(1/k)$. The approximate value of the integral (48) is

$$I \approx \sum_i g(x_i) \sqrt{\frac{2j\pi}{k\psi''(x_i)}}\, e^{jk\psi(x_i)}. \tag{49}$$

To bring (47) to a form suitable for evaluation with the stationary phase method, the Bessel function is expressed in terms of modulus and phase.[19]

$$J_o(z) = M_o(z) \cos \theta_o(z) \tag{50}$$

with

$$\theta_o(z) = z - \frac{\pi}{4} + \delta_o(z), \tag{51}$$

where $\delta_o(z)$ and $M_o(z)$ are slowly varying functions for large z with $\lim_{z\to\infty} \delta_o(z) = 0$ and $\lim_{z\to\infty} M_o(z) = \sqrt{2/(\pi z)}$.

With $z = 2\omega_0\sqrt{tx - x^2}$, the integral in (47) has a stationary point at $x = t/2$. The approximate value of (47), obtained by using (49), is

$$f_{NS}(f) \approx \int_0^\infty f_L(u)du - \sqrt{\frac{\pi t}{2\omega_0}} M_o(\omega_o t) \cos\left[\theta_o(\omega_o t) - \frac{\pi}{4}\right] \cdot \int_0^\infty f_L(u) \sqrt{\frac{2\beta u}{t}} J_1(\sqrt{2\beta ut})du. \tag{52}$$

For large values of t such that $M_o(z)$ and $\theta_o(t)$ can be approximated with their asymptotic values,

$$f_{NS}(t) \approx \int_0^\infty f_L(u)du - \frac{\sin \omega_o t}{\omega_o} \int_0^\infty f_L(u) \sqrt{\frac{2\beta u}{t}} J_1(\sqrt{2\beta ut})du. \tag{53}$$

It is of interest to note that the stationary phase method gives the correct value for the integral

$$\int_0^t J_o(2\omega_o\sqrt{ut - u^2})du = \frac{\sin \omega_o t}{\omega_0}. \tag{54}$$

This integral can be evaluated exactly by using (42) with $v = 0$ and $g(u) = 1.0$. The left-hand side of (42) is readily inverted yielding (54).

APPENDIX B

Comparison of Exact and Approximate Solutions

Consider a notch filter transfer function

$$T_N(z) = \left(\frac{z^2 + \omega_o^2}{z^2 + \beta z + \omega_o^2}\right)^{m+1}. \tag{55}$$

The Laplace transform of the time response due to a unit step function is

$$F_{NS}(z) = \frac{1}{z}\left(1 - \frac{\beta z}{z^2 + \beta z + \omega_o^2}\right)^{m+1}. \tag{56}$$

For $m = 0$, the time response is

$$f_{NS}(t) = \left(1 - \frac{\beta}{\omega} e^{-(\beta/2)t} \sin \omega t\right) \cdot 1(t), \tag{57}$$

where

$$\omega = \sqrt{\omega_o^2 - \left(\frac{\beta}{2}\right)^2}. \tag{58}$$

A comparison of (57) with (19) shows that both are of the same form but ω is replaced by ω_o. Hence, the approximation is of order $(\beta/\omega_o)^2$.

For $m = 1$, using tables of Laplace transforms,

$$f_{NS}(t) = 1 - \frac{\beta}{\omega}\left[2 + \left(\frac{\beta}{2\omega}\right)^2 - \frac{\beta}{2}t\right]e^{-(\beta/2)t} \sin \omega t + \left(\frac{\beta}{\omega}\right)^2 \beta t e^{-(\beta/2)t} \cos \omega t. \tag{59}$$

Neglecting terms of order $(\beta/\omega)^2$ against one, (59) can be written

$$f_{NS}(t) = 1 - \frac{\beta}{\omega_o} e^{-(\beta/2)t} L_1^1\left(\frac{\beta}{2}t\right) \sin \omega_o t + \left(\frac{\beta}{2\omega_o}\right)^2 \beta t e^{-(\beta/2)t} \cos \omega_o t, \tag{60}$$

where from (16)

$$L_1^1(x) = 2 - x. \tag{61}$$

The terms multiplied by $\sin \omega_o t$ are up to order $(\beta/\omega_o)^2$ the same as in (19).

For $m = 2$,

$$f_{NS}(t) = 1 - \frac{\beta}{\omega} e^{-(\beta/2)t}\left[3 + \frac{7}{8}\frac{\beta^2}{\omega^2} + \frac{3}{32}\left(\frac{\beta}{\omega}\right)^4 - \frac{\beta t}{2}\left(3 + \frac{\beta^2}{4\omega^2}\right) + \frac{1}{2}\left(\frac{\beta t}{2}\right)^2\left(1 - \frac{\beta^2}{4\omega^2}\right)\right] \sin \omega t + \frac{\beta^2}{\omega^2} e^{-(\beta/2)t} \beta t\left[\frac{7}{8} + \frac{3}{32}\frac{\beta^2}{\omega^2} - \frac{\beta t}{8}\right] \cos \omega t. \tag{62}$$

Neglecting terms of order $(\beta/\omega_o)^2$ and higher against one yields

$$f_{NS}(t) = 1 - \frac{\beta}{\omega_o} e^{-(\beta/2)t} L_2^1\left(\frac{\beta}{2}t\right) \sin \omega_o t + \frac{\beta^2}{\omega^2} e^{-(\beta/2)t} \frac{\beta t}{8} [7 - \beta t] \cos \omega_o t, \quad (63)$$

where from (16)

$$L_2^1(x) = 3 - 3x + \frac{x^2}{2}. \quad (64)$$

REFERENCES

1. D. L. Favin, "The 6A Impulse Counter," Bell Laboratories Record, March 1963, pp. 100–102.
2. Members of the Technical Staff, Bell Laboratories, *Transmission Systems for Communications*, Fourth Edition, 1970, pp. 32–34.
3. M.E. Van Valkenburg, *Introduction to Modern Network Synthesis*, New York: John Wiley, 1960, pp. 479–486.
4. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part I, New York: McGraw-Hill, 1953, p. 944.
5. W. T. Howell, "On a Class of Function Which are Self-Reciprocal in the Hankel Transform," Phil. Mag., *25*, Series 7, April 1938, pp. 622–628.
6. A. Erdelyi, et al., *Tables of Integral Transforms*, Vol. 2, New York: McGraw-Hill, 1954, p. 43.
7. E. A. Guillemin, *Synthesis of Passive Networks*, New York: John Wiley, 1957, p. 489.
8. Ref. 3, p. 343.
9. Ref. 6, Vol. 1, p. 175.
10. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integral Series and Products*, New York: Academic Press, 1965, p. 1037.
11. Ref. 10, p. 718.
12. K. W. Henderson and W. H. Kautz, "Transient Response of Conventional Filters," IRE Trans. Circuit Theory, December 1958, pp. 333–347.
13. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, New York: John Wiley, 1951, p. 267.
14. Ref. 4, p. 456.
15. Ref. 6, Vol. 1, p. 133.
16. N. W. McLachlan, *Modern Operational Calculus*, London: The MacMillan Company, 1949, p. 45.
17. Ref. 10, p. 951.
18. A. Erdelyi, *Asymptotic Expansions*, New York: Dover Publications, 1950, pp. 51–57.
19. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publications, 1965, p. 365.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

Some Comparisons Between FIR and IIR Digital Filters

By L. R. RABINER, J. F. KAISER, O. HERRMANN, and M. T. DOLAN

(Manuscript received July 27, 1973)

The purpose of this paper is to make comparisons between optimum, linear phase, finite impulse response (FIR) digital filters and infinite impulse response (IIR) digital filters which meet equivalent frequency domain specifications. The basis of comparison is, for the most part, the number of multiplications per sample required in the usual realizations of these filters—i.e., the cascade form for IIR filters, and the direct form for FIR filters. Comparisons are also made between group-delay equalized filters and linear phase FIR filters. Considerations dealing with finite word-length effects are discussed for both these filter types. A set of design charts is also presented for determining the minimum filter order required to meet given low-pass filter specifications for both digital and analog filters.

I. INTRODUCTION

Although a great deal is known about the properties of different types of digital filters, very little has been done to relate the various designs as to performance and complexity of realization. Thus the filter designer must learn the details of several design procedures before being able to make a wise decision on a suitable filter for his specific application. It is the purpose of this paper to add insight into some of the problems that have been encountered by filter designers by: (*i*) presenting new and useful design curves for digital and analog low-pass filters, and (*ii*) making several comparisons between optimum (quasi-equiripple), linear phase, FIR low-pass filters and equiripple (elliptic) IIR filters which meet equivalent frequency domain specifications. Although these results are presented for low-pass filter designs, they are easily extended to the case of bandpass, bandstop, and high-pass filters by the well-known frequency band transformations.

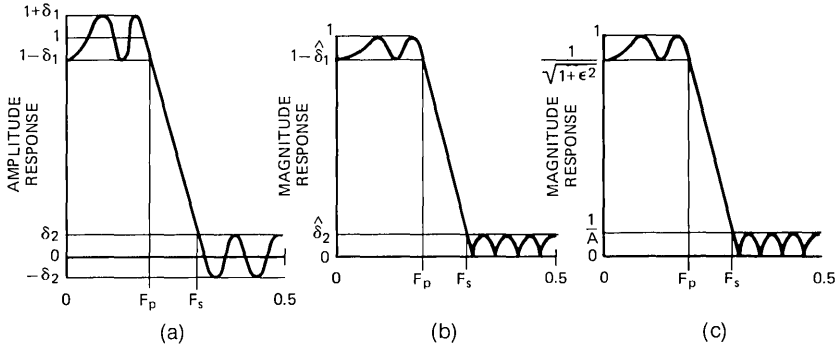

Fig. 1—Terminology used to describe low-pass filter characteristics.

The organization of the paper is as follows. After defining the terminology to be used, the design relationships between the FIR filter parameters are reviewed. The design relationships for the IIR filter parameters are developed and a novel graphical interpretation of these relations is presented in the form of useful filter design charts (applicable to both digital and analog filters). Using the design relationships, the filter orders required to achieve equivalent performance are compared for different ranges of filter parameters.

II. TERMINOLOGY

The design procedures for the two general classes of digital filters, FIR and IIR, have essentially progressed along independent paths. As a result, the terminology used in specifying the filter performance is generally not quite the same. Thus it is instructive to define the most commonly accepted definitions of the filter parameters for these two classes of filters. The relations between these parameter sets for equivalence are then established. Figure 1a shows the amplitude response of a typical optimum FIR low-pass filter and Fig. 1b shows the magnitude response of a typical elliptic low-pass filter. For the FIR case the amplitude response in the passband ($0 \leqq f \leqq F_p$)* generally oscillates between $1 + \delta_1$ and $1 - \delta_1$, where δ_1 is the passband ripple. In the stopband ($F_s \leqq f \leqq 0.5$) the amplitude response oscillates between $+\delta_2$ and $-\delta_2$ where δ_2 is the stopband ripple. For the elliptic case the magnitude response is constrained to always be less

* Throughout this paper the frequency scale has been normalized with respect to the sampling frequency. Thus the normalized sampling frequency is 1.0 and the frequency range graphed is $0 \leqq f \leqq 0.5$ or, equivalently, $0 \leqq \omega \leqq \pi$.

than 1.0. Thus, in the passband ($0 \leqq f \leqq F_p$) the magnitude response oscillates between 1 and $1 - \hat{\delta}_1$. In the stopband the magnitude response oscillates between $\hat{\delta}_2$ and 0. It is straightforward to relate δ_1, δ_2, $\hat{\delta}_1$, and $\hat{\delta}_2$ so the resulting magnitude characteristics are equivalent. If the FIR amplitude characteristic is scaled by $1/(1 + \delta_1)$ and the magnitude of the resulting amplitude response is taken, then the following relationships are obtained:

$$\hat{\delta}_1 = \frac{2\delta_1}{1 + \delta_1} \tag{1}$$

$$\hat{\delta}_2 = \frac{\delta_2}{1 + \delta_1} \tag{2}$$

$$\delta_1 = \frac{\hat{\delta}_1}{2 - \hat{\delta}_1} \tag{3}$$

$$\delta_2 = \frac{2\hat{\delta}_2}{2 - \hat{\delta}_1}. \tag{4}$$

Thus, given either (δ_1, δ_2) or ($\hat{\delta}_1$, $\hat{\delta}_2$), eqs. (1) through (4) can be used to find the equivalent specifications for the other type of filter.

Although the notation of Fig. 1b is acceptable for the magnitude response of an elliptic filter, it is not the most widely used form for these filters. Figure 1c shows the same magnitude response described in terms of passband parameter ϵ and stopband parameter A. Comparing Figs. 1b and 1c it is easy to relate ϵ and A to $\hat{\delta}_1$ and $\hat{\delta}_2$ as

$$\epsilon = \frac{\sqrt{2 - \hat{\delta}_1}\,\sqrt{\hat{\delta}_1}}{(1 - \hat{\delta}_1)} \tag{5}$$

$$A = \frac{1}{\hat{\delta}_2}. \tag{6}$$

At this point it is convenient to define the additional filter terms E, ATT, and η as

$$E = \text{(in-band) ripple} = 20 \log_{10} \sqrt{(1 + \epsilon^2)} \tag{7}$$

$$\text{ATT} = \text{stopband attenuation} = 20 \log_{10} A \tag{8}$$

$$\eta = \frac{\epsilon}{\sqrt{A^2 - 1}} = \frac{\hat{\delta}_2\sqrt{\hat{\delta}_1}\sqrt{2 - \hat{\delta}_1}}{(1 - \hat{\delta}_1)\sqrt{1 - \hat{\delta}_2^2}} = \frac{2\delta_2(\sqrt{\delta_1})}{(1 - \delta_1)\sqrt{(1 + \delta_1)^2 - \delta_2^2}}. \tag{9}$$

Thus, parameters E and ATT are a third set of parameters which describe the characteristics of the magnitude response of the elliptic

filter. The parameter η has been shown to be a basic analog filter parameter[1] which will be used in the filter design curves given in a later section.

III. FIR DESIGN RELATIONS

The five basic FIR filter parameters are F_p, F_s, δ_1, δ_2, and N, the duration of the filter impulse response in samples. For the general case of optimum, linear phase, low-pass FIR filters, there exist no simple analytical relationships between these five filter parameters, except in special cases, e.g., one passband or one stopband ripple. However, an approximate empirical relationship between the filter parameters has recently been obtained[2] which accurately satisfies known design results for a wide range of values of the filter parameters. The relationship is of the form:

$$N = 1 + \frac{D_\infty(\delta_1, \delta_2)}{\Delta F} - f(\delta_1, \delta_2)\Delta F, \tag{10}$$

where

$$\Delta F = F_s - F_p = \text{relative transition width}, \tag{11}$$

$$D_\infty(\delta_1, \delta_2) = [0.005309\ (\log_{10}\delta_1)^2 + 0.07114\log_{10}\delta_1 - 0.4761]\log_{10}\delta_2 - [0.00266\ (\log_{10}\delta_1)^2 + 0.5941\log_{10}\delta_1 + 0.4278], \tag{12}$$

and

$$f(\delta_1, \delta_2) = 0.51244\log_{10}(\delta_1/\delta_2) + 11.01. \tag{13}$$

Equation (10) can generally predict the value of N required to meet specifications on δ_1, δ_2, F_p, and F_s to within ± 2. In the cases where F_p is very close to 0, or F_s is very close to 0.5, eq. (10) tends to overestimate the required N. It should be noted that eq. (10) shows the estimate of N to be independent of specific values of F_s or F_p, but instead is dependent only on the transition width, $(F_s - F_p)$.

A simpler expression giving a less accurate estimate of N is

$$N = \frac{-10\log_{10}(\delta_1\cdot\delta_2) - 15}{14\Delta F} + 1. \tag{14}$$

This expression is a modification of the design relationship for FIR filters designed by windowing techniques (where $\delta_1 = \delta_2$). See Ref. 3, pp. 237–238.

IV. IIR DESIGN RELATIONS

One of the most general procedures for designing IIR digital filters is through the bilinear transformation of an appropriate continuous

filter. There are two equivalent techniques for obtaining the desired digital filter using the bilinear transformation and these are illustrated in Fig. 2. The technique of Fig. 2a begins with an analog low-pass filter with normalized passband cutoff frequency of 1 radian per second, and analog stopband cutoff frequency of Ω_s radians per second. This filter is bilinearly transformed[4] to give a "normalized" digital filter with passband cutoff frequency $\pi/2$ radians per second (or $f = 0.25$ on the normalized scale) and stopband cutoff frequency $\hat{\omega}_s$. The relation between the frequency variables Ω and $\hat{\omega}$ is given by

$$\Omega = \tan(\hat{\omega}/2). \tag{15}$$

Thus Ω_s and $\hat{\omega}_s$ are simply related by eq. (15) with $\Omega = \Omega_s$, and $\hat{\omega} = \hat{\omega}_s$. Finally, a digital all-pass transformation[5] is used to give the desired digital low-pass filter with passband cutoff frequency ω_p and stopband cutoff frequency ω_s. The relation between the frequency variables ω and $\hat{\omega}$ is given by

$$\tan\omega = \frac{(1-\alpha^2)\sin\hat{\omega}}{(1+\alpha^2)\cos\hat{\omega} - 2\alpha}, \tag{16}$$

where

$$\alpha = \frac{\sin(\omega_p/2) - \cos(\omega_p/2)}{\sin(\omega_p/2) + \cos(\omega_p/2)}. \tag{17}$$

The second technique (shown in Fig. 2b) begins with the identical analog low-pass filter as in the first technique and immediately performs a low-pass–to–low-pass transformation to give an analog filter with passband cutoff frequency $\hat{\Omega}_p$ and stopband cutoff frequency $\hat{\Omega}_s$. The relation between the frequency variables Ω and $\hat{\Omega}$ is

$$\Omega = \hat{\Omega}/\hat{\Omega}_p. \tag{18}$$

The resulting low-pass filter is then transformed to a digital filter using the bilinear transformation, giving the same end result as in the first technique (Fig. 2a). The relation between frequency variables ω and $\hat{\Omega}$ is

$$\hat{\Omega} = \tan(\omega/2). \tag{19}$$

In the case where the prototype normalized analog filter is an elliptic filter, it is relatively easy to derive a design formula relating the various filter parameters, both in the analog and digital cases. For either the analog or digital case the order, n, of the elliptic filter is related to the remaining filter parameters by the equation

$$n = \frac{K(k)K(\sqrt{1-k_1^2})}{K(k_1)K(\sqrt{1-k^2})}, \tag{20}$$

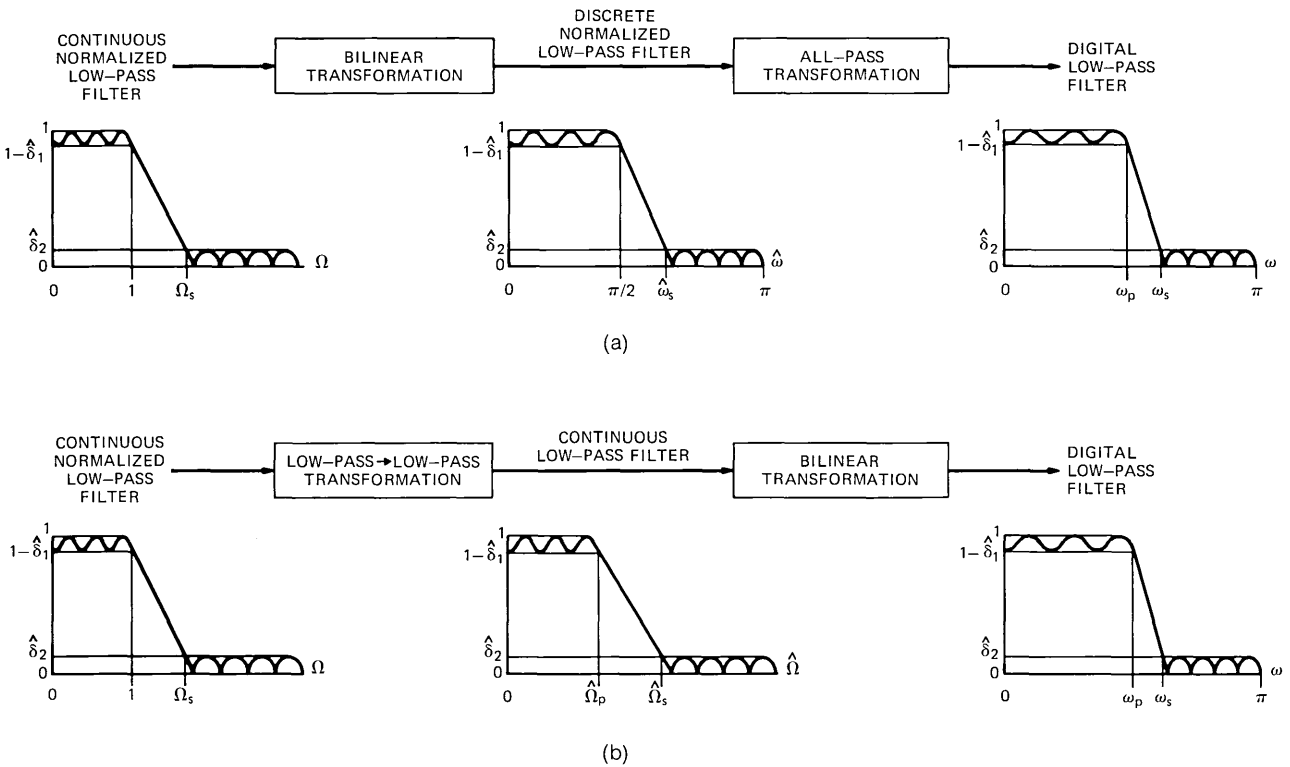

Fig. 2—Two techniques for transforming a continuous normalized low-pass filter to a digital low-pass filter.

where $K(\cdot)$ is the complete elliptic integral of the first kind and

$$k = \text{transition ratio} = \frac{1}{\Omega_s} = \frac{\tan(\omega_p/2)}{\tan(\omega_s/2)} \tag{21}$$

and

$$k_1 = \eta = \frac{\epsilon}{\sqrt{A^2 - 1}} = \frac{2\delta_2\sqrt{\delta_1}}{(1 - \delta_1)\sqrt{(1 + \delta_1)^2 - \delta_2^2}}. \tag{22}$$

Thus eq. (20) relates filter order, n, to the parameters F_p, F_s [through eq. (21)] and δ_1 and δ_2 [through eq. (22)].

For the case when the prototype filter is a Chebyshev filter (either type I—equiripple passband, monotone stopband, or type II—maximally flat passband, equiripple stopband) the design equation becomes

$$n = \frac{\cosh^{-1}(1/\eta)}{\ln \beta}, \tag{23}$$

where

$$\beta = \frac{1 + \sqrt{1 - k^2}}{k} \tag{24}$$

and η and k are defined as eqs. (21) and (22). Finally for a prototype Butterworth filter (maximally-flat magnitude, all pole) the design equation is

$$n = \frac{\ln \eta}{\ln k}. \tag{25}$$

Although eqs. (20) through (25) completely describe the design curves for both analog and digital filters, it is generally quite helpful to see the relationships between filter parameters displayed in a meaningful way. Since, in general, there are five filter parameters there is no simple way of presenting these relationships on a single plot, even in terms of well-known nomograph procedures.[6] There is, however, a simple and straightforward way of including all design relations for both digital and analog filters, for any prototype filter, using a sequence of three charts.

The first chart(s) relates the filter design parameter η to the passband and stopband ripple specifications δ_1 and δ_2 or their equivalents. The second chart(s) graphs the filter design equation relating filter order n, design parameter η, and transition ratio k. The third chart(s) relates transition ratio k to passband cutoff frequency F_p and transition bandwidth ν.

Figures 3a through 3d show four possibilities for Chart No. 1. The graphs of Figs. 3a and 3b correspond to digital filters with δ_1 as a param-

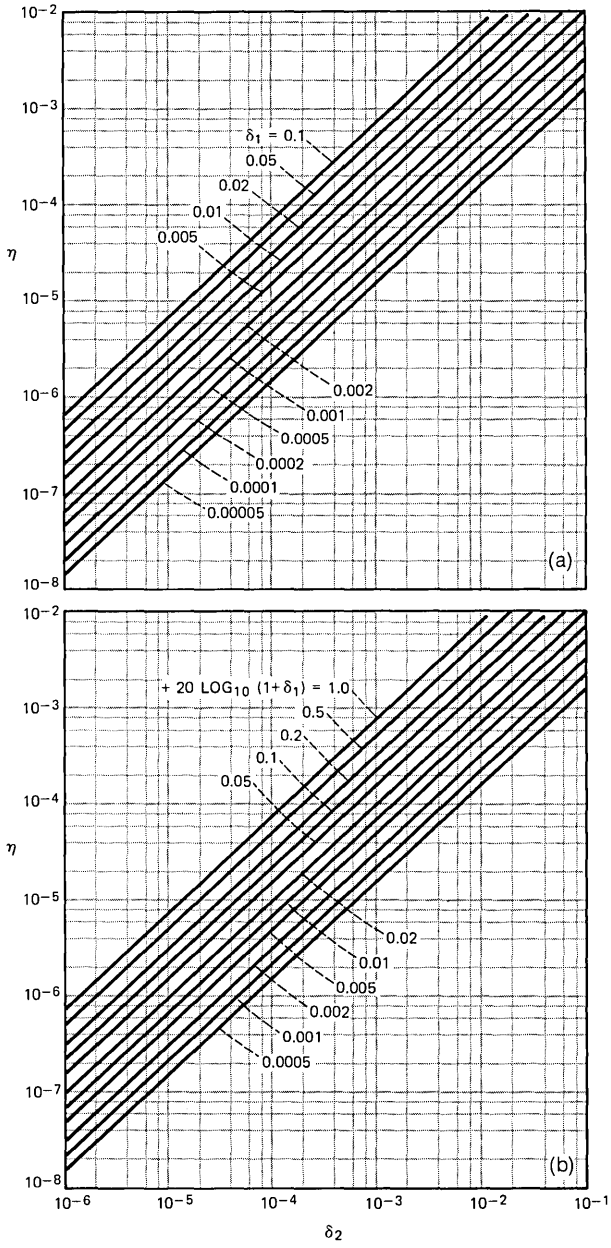

Fig. 3—Plots of η versus stopband specification, with parameter passband specification for low-pass filters.

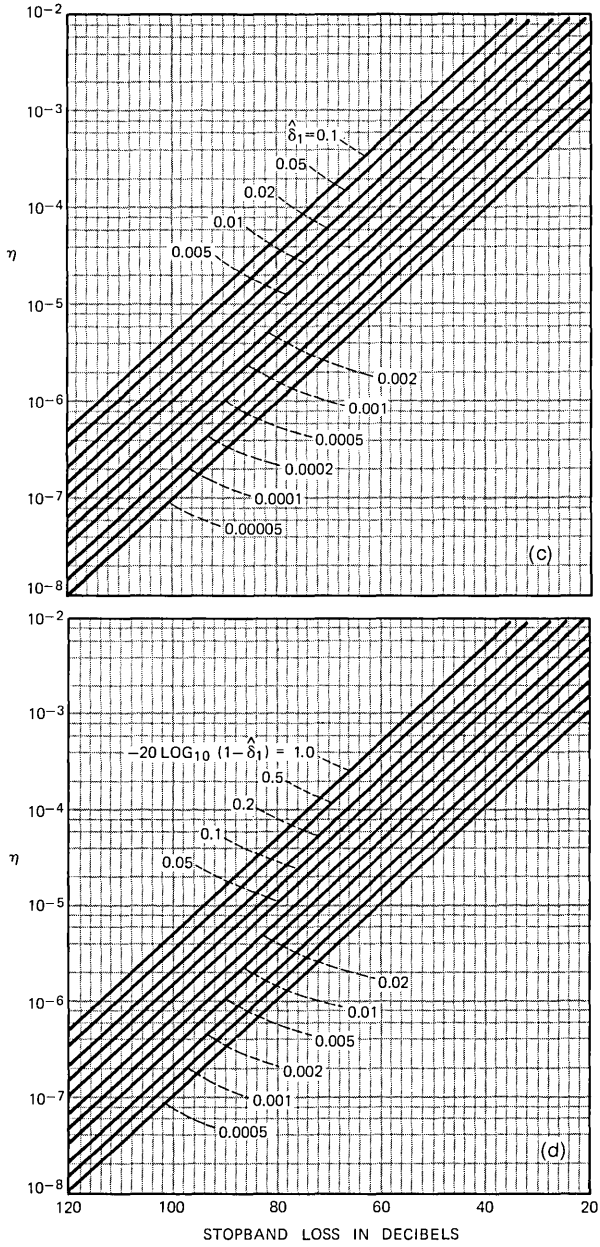

Fig. 3 (continued).

eter (Fig. 3a) or $20 \log_{10} (1 + \delta_1)$(dB) as a parameter (Fig. 3b). The graphs of Figs. 3c and 3d correspond to analog filters with absolute ripple $\hat{\delta}_1$ as a parameter (Fig. 3c) or total ripple $20 \log_{10} [1/(1 - \hat{\delta}_1)]$(dB) as a parameter (Fig. 3d).

Chart No. 2 represents the design relations particular to the prototype filters, i.e., eq. (20) for elliptic filters, eq. (23) for Chebyshev filters, and eq. (25) for Butterworth filters. For these graphs the parameter η is plotted versus transition ratio, k, with filter order, n, as the parameter. Figures 4a through 4c show the resulting graphs for elliptic filters, Chebyshev filters, and Butterworth filters, respectively. The horizontal scale on each of these graphs is a nonuniform scale which was chosen to provide a reasonably good spacing of the curves for the various values of n. The actual nonlinear scale used is represented by the equation

$$x = \frac{k + k^8}{2}, \tag{26}$$

where x is the x-axis coordinate ($0 \leqq x \leqq 1$) and k is the transition width. Thus, the scale is linear for small values of k and highly nonlinear near $k = 1.0$.

Chart No. 3 represents the relation between the transition ratio and the filter cutoff frequencies [eq. (21)]. For these graphs the passband cutoff frequency, F_p, is plotted versus transition ratio, k, for various values of normalized transition width, ν, defined as

$$\nu = F_s - F_p = \frac{\omega_s - \omega_p}{2\pi}. \tag{27}$$

Figures 5a and 5b show the resulting graphs for digital and analog filters. The scale for transition ratio is identical to the scale used for Chart No. 2.

V. USE OF CHARTS

To illustrate how to use the set of charts of Figs. 3 through 5, consider the determination of filter order n required to meet the following specifications:

$\delta_1 = 0.01$ ($\approx \pm 0.086$-dB passband ripple)
$\delta_2 = 0.0001$ (80-dB stopband loss)
passband cutoff frequency = 480 Hz
stopband edge frequency = 520 Hz
sampling frequency = 8000 Hz.

Normalizing the band-edge frequencies gives

$$F_p = \frac{480}{8000} = 0.06$$

$$F_s = \frac{520}{8000} = 0.065.$$

For the determination of filter order n for a digital filter of the elliptic type, the charts of Figs. 3a, 4a, and 5a are used (Fig. 4a specializes the design to the elliptic type). To obtain the value of η on Fig. 3a, we use the curve $\delta_1 = 0.01$ and find its intersection with the line $\delta_2 = 0.0001$ which yields a value of η approximately equal to 2×10^{-5}. To obtain the transition ratio, we use Fig. 5a by finding the intersection of the curve $\nu = F_s - F_p = 0.005$ with line $F_p = 0.06$; this yields a value of 0.923 for the transition ratio (this agrees nicely with $F_p/F_s = 0.06/0.065 = 0.923$, an alternate way of arriving at the same result). Finally, the filter order, n, can now be determined from Fig. 4a by finding the intersection of the lines $\eta = 2 \times 10^{-5}$ and transition ratio = 0.923; thus the required theoretical elliptic filter order is ≈ 11.5. In order to meet specifications on all four parameters, a 12th-order filter must be used.

However, there are several tradeoffs possible for the final filter specifications. For example, if η is held fixed at 2×10^{-5} and the transition ratio is changed to approximately 0.94 to lie on the $n = 12$ curve, then either F_s or F_p can be varied to match this new value of transition ratio. The tradeoffs here are obtained from Fig. 5a. If the transition ratio is held fixed, then for $n = 12$ we find η is $\approx 1.0 \times 10^{-5}$; from Chart No. 1 (Fig. 3a) we can observe the tradeoff as δ_1 and δ_2 are varied for this new value of η. Finally, both transition ratio and η can be varied, e.g., to 0.93 for transition ratio and 1.5×10^{-5} for η, so as to make their intersection remain on the $n = 12$ curve; now all four filter parameters can be varied to match the new values of η and transition ratio.

It is interesting to note that if a Chebyshev or Butterworth filter type is specified in place of the elliptic, the designer need only substitute Figs. 4b or 4c for Fig. 4a as Chart No. 2 and proceed as before. In both cases of the example given, the required filter order considerably exceeds the maximum limit of 20 of the curves; thus the "efficiency" of the elliptic design is clearly seen.

Clearly, this design procedure presents a tremendous amount of flexibility to the designer—more so than is generally available in most

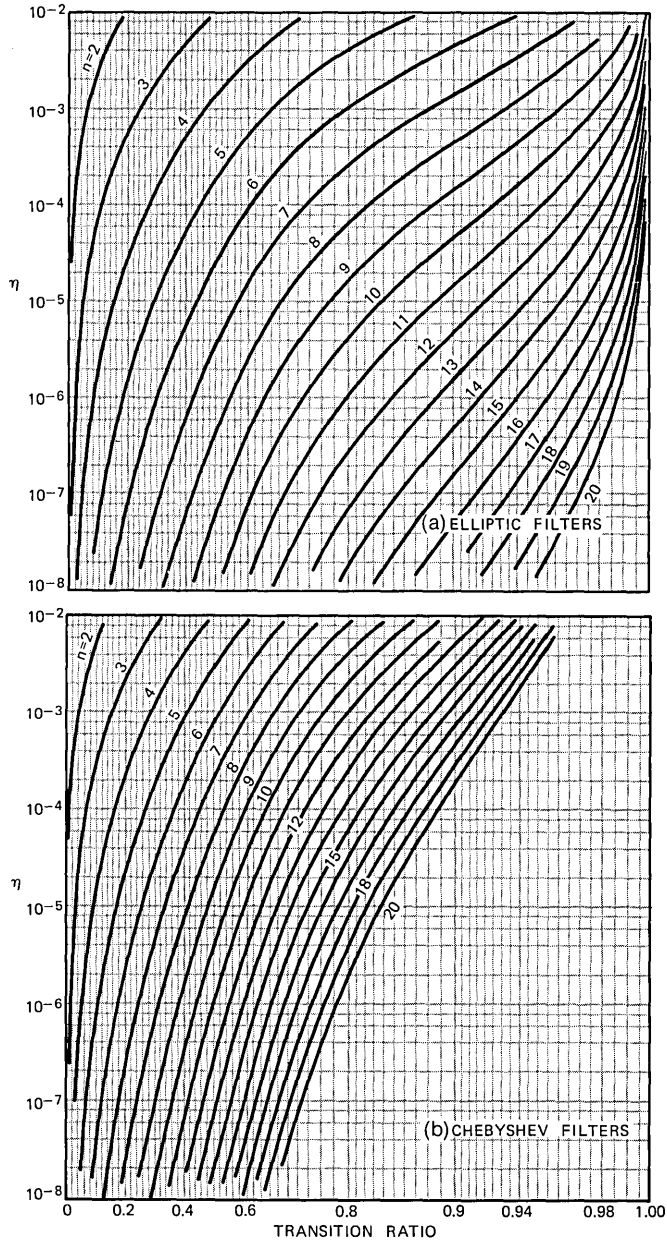

Fig. 4—Plots of η versus transition ratio as a function of filter order n for elliptic, Chebyshev, and Butterworth low-pass filters.

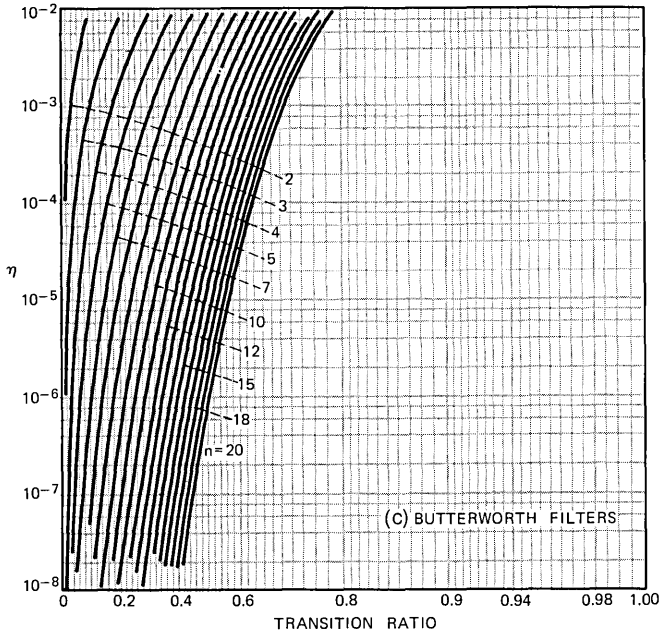

Fig. 4 (continued).

programs for filter order determination. *Furthermore, the insight into the design problem afforded by this graphical technique allows the designer to get a feeling for the way in which small changes in filter specification affect the required filter order.* Quite often the designer is willing to changes his ideas on "required" specifications, especially if it reduces the filter order necessary to meet his specifications.

VI. COMPARISONS BETWEEN OPTIMUM FIR AND ELLIPTIC DIGITAL FILTERS

Based on the design formulas of the preceding sections, it is possible to make some comparisons between optimum FIR low-pass filters and equivalent elliptic filters. The main basis of comparison will be the number of multiplications per input sample* required in the most standard realization of each filter type, i.e., the direct form for the FIR case and the cascade form for the elliptic case.[7] Direct realization of an N-point impulse response filter with linear phase requires

* The number of multiplications per input sample is a useful measure of the computational complexity of the filtering operations as it represents the number of multiply-add operations required for a software implementation of the algorithm as well as for a general hardware implementation.

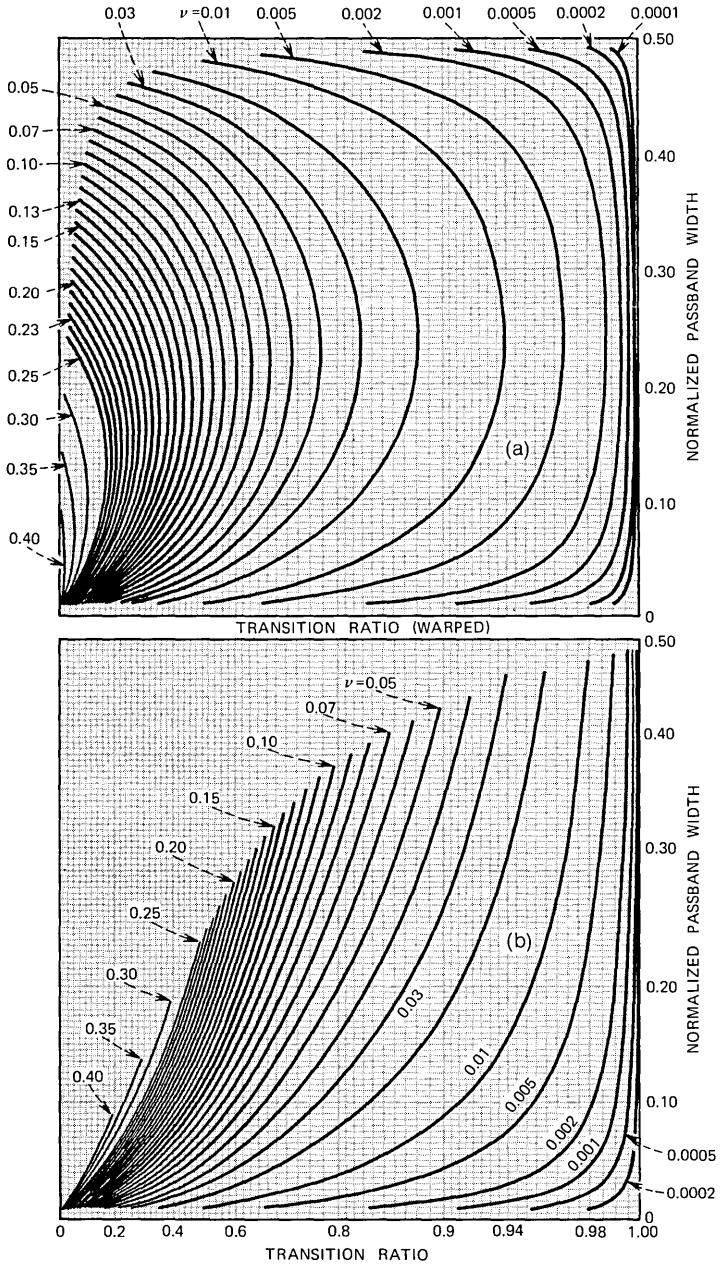

Fig. 5—Plots of passband cutoff frequency versus transition ratio as a function of transition width for discrete and continuous low-pass filters.

$[(N + 1)/2]$ multiplications per sample, whereas cascade realization of an nth-order elliptic filter (all zeros on the unit circle) requires $[(3n + 3)/2]$* multiplications per sample where $[\cdot]$ denotes "integer part of."

Thus, one basis of comparison between equivalent filter designs (i.e., both meeting the same specifications on δ_1, δ_2, F_p, and F_s) is in terms of the efficiency of the respective realizations, i.e., which structure requires fewer multiplications per sample. Equivalence between structures is attained when the condition

$$\left[\frac{3n + 3}{2}\right] = \left[\frac{(N + 1)}{2}\right] \tag{28}$$

or equivalently

$$\frac{N}{n} \approx 3 + \frac{1}{n}. \tag{29}$$

Using the appropriate filter design formulas, we have measured the quantity N/n as a function of n for a large range of values of F_p, δ_1, and δ_2. Figure 6 shows two typical sets of curves which were obtained. Figure 6a shows data for the case $F_p = 0.15$, $\delta_1 = 0.1$, $\delta_2 = 0.1$, 0.01, 0.001, 0.0001, and Fig. 6b shows data for $F_p = 0.35$, $\delta_1 = 0.00001$, and the same range of δ_2 as in Fig. 6a. Also shown in these plots is the line $N/n = 3$ for showing where the data lie with respect to the fixed portion of eq. (29). As seen in this figure, for certain values of F_p, δ_1, and δ_2, the ratio of N/n falls below the equivalence level of eq. (29), i.e., the FIR filter is more efficient than the elliptic filter. However, in general, the elliptic filter is more efficient than the optimum FIR filter, and, in the case of high-order elliptic designs, the ratio of N/n is often in the hundreds or thousands.

Based on our examination of large amounts of data, the following general observation can be made: the most favorable conditions for the FIR design are large values of δ_1, small values of δ_2, and large transition widths (i.e., small transition ratios). One also observes the following behavior:

(*i*) For values of $F_p \geqq 0.3$, the ratio N/n always exceeded $3 + 1/n$ for all values of δ_1, δ_2, and n.

(*ii*) For values of $n \geqq 7$, the ratio N/n always exceeded $3 + 1/n$ for all values of δ_1, δ_2, and F_p.

---

*This number of multiplications per sample for the IIR filter assumes that any scaling between sections is an integer power of 2 and is performed entirely by shifts of the data. If finer scaling multipliers are included between each cascade section, the realization requires $[(4n + 3)/2]$ multiplications per sample.

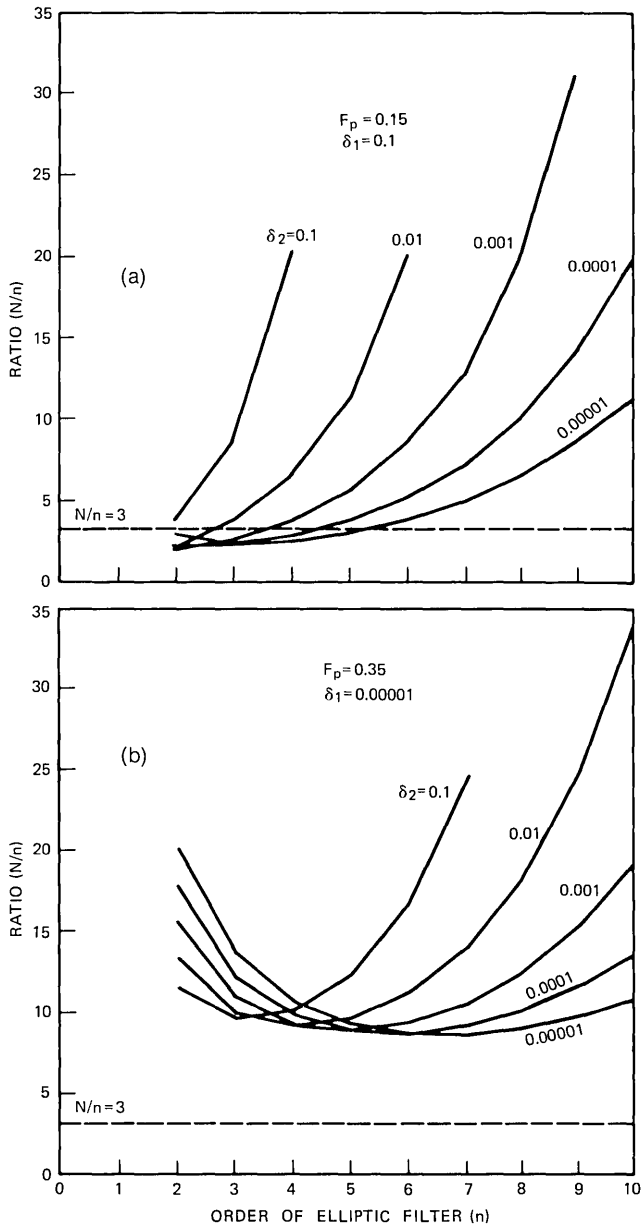

Fig. 6—Plots of the ratio N/n as a function of n for optimum FIR filters and elliptic filters meeting identical specifications on δ_1, δ_2, F_p, and F_s.

(*iii*) The smaller the value of F_p, the larger the range of δ_1, δ_2, and n for which N/n was less than $3 + 1/n$.

Since the design formula for N for the optimum FIR case is not exact but only an estimate, measurements were also made of the required theoretical value of n (elliptical filter order) to meet the specifications of optimum FIR filters which had already been designed. Typical results of these measurements are shown in Fig. 7. Figure 7a shows the theoretical order n (n need not be an integer) required to match specifications on F_p, F_s, for $\delta_1 = 0.1$, $\delta_2 = 0.1$, 0.01, 0.001, 0.0001, and 0.00001, as a function of F_p for a set of optimum FIR filters with $N = 21$. (It should be noted that as F_p varies, F_s also varies so as to achieve the desired specifications on δ_1 and δ_2.) Figure 7b shows similar measurements for $N = 41$. In Fig. 7a the theoretical point of equivalence is $n = 6.3$, whereas in Fig. 7b it is $n = 13$. From this figure it is seen that for these cases the elliptic filter is always more efficient than the equivalent FIR filter, as anticipated by the discussion in the preceding paragraphs.

In summary, elliptic filters are generally more efficient in achieving given specifications on the frequency response than optimum FIR filters. However, the FIR filters have the additional useful property that their phase is exactly linear, i.e., there is no group delay distortion. For the elliptic filter, however, there is generally a large amount of group delay distortion (concentrated primarily near the band edge). A question of both theoretical and practical importance is whether, in cases when the additional requirement of a flat delay is specified, it is more desirable to equalize the delay of an elliptic filter or to use the equivalent optimum FIR filter (with its constant group delay). In the next section we discuss various aspects of this question. It should be noted that the above alternatives are not the only possibilities for obtaining a digital filter which meets frequency domain specifications on both magnitude and group delay responses. For example, a filter can be designed, using modern optimization procedures, where the number of poles and zeros are unequal. In such cases, the comparisons between FIR and IIR filters are quite distinct from those to be discussed in the next section.

VII. COMPARISONS OF OPTIMUM FIR FILTERS AND DELAY-EQUALIZED ELLIPTIC FILTERS

Recently developed optimization procedures[8] make it possible to design an all-pass equalizer which can equalize the group delay of any

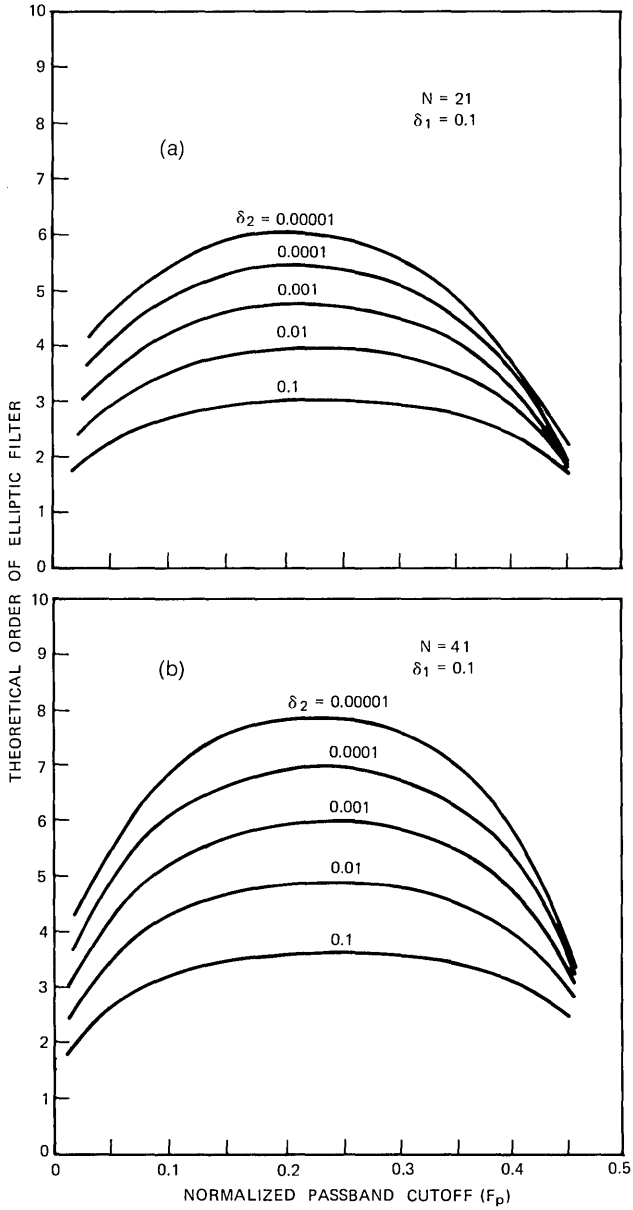

Fig. 7—Theoretical order of elliptic filters required to meet given specification on δ_1, δ_2, F_p, and F_s as a function of F_p for various values of δ_2. Optimum FIR filters with $N = 21$ meet the specifications for all filters of (a), whereas $N = 41$ is required for all filters of (b).

digital filter to any desired accuracy over a restricted band of frequencies. As an example of the use of this procedure, Fig. 8 shows plots of the group delay of a 6th-order (unequalized) elliptic filter (with parameters $\delta_1 = 0.01$, $\delta_2 = 0.0001$, $F_p = 0.24163$, $F_s = 0.34842$) and the equalized group delay using a 10th-order all-pass filter. The relative error in the equalized delay curve is 3.6 percent of the average delay in the passband. In this case the equalized elliptic filter requires 20 multiplications per sample, whereas an optimum FIR filter which achieves the same specifications requires only 11 multiplications per sample.

The difficulty with trying to equalize the group delay of a filter lies in the fact that the equalized filter must have a total delay greater than the largest delay in the unequalized filter which always occurs near the passband cutoff frequency. Thus, in the example of Fig. 8, even though the delay throughout most of the passband is between 2 and 6 samples, the delay at the edge of the band is about 15 samples. It can be shown that an all-pass equalizer of degree n_e has the

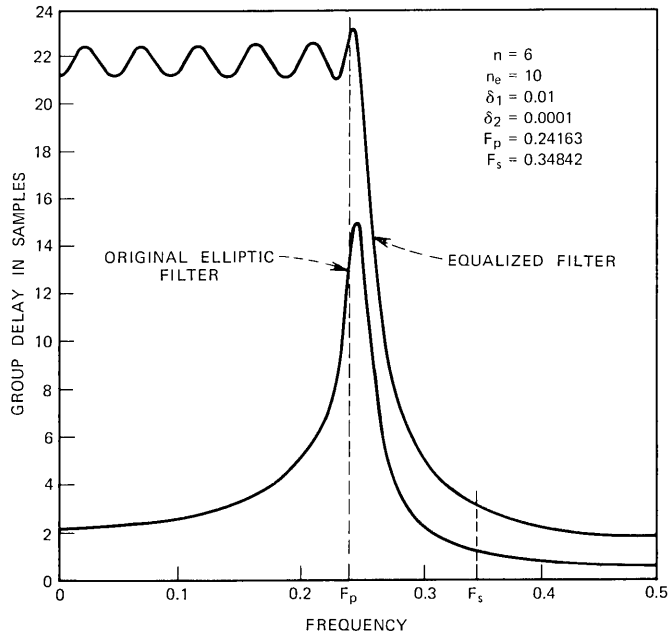

Fig. 8—The group delay of an unequalized and an equalized elliptic filter. The equalizer is of 10th degree and the elliptic filter is of 6th degree.

property

$$\frac{1}{2\pi}\int_0^{\pi} \tau_g(\omega)d\omega = 0.5n_e, \tag{30}$$

where $\tau_g(\omega)$ is the equalizer group delay and the integral is taken over half the sampling interval ($0 \leqq \omega \leqq \pi$). Since $\tau_g(\omega) \geqq 0$, i.e., group delays add, to justify eq. (30) it is sufficient to show that a first-degree all-pass equalizer has the required property. The z-transform of a first-degree all-pass equalizer is

$$H(z) = \frac{1 - z^{-1}/a}{1 - az^{-1}}, \tag{31}$$

where a is the pole position and $1/a$ is the zero position in the z-plane. The group delay is commonly defined as

$$\tau_g(\omega) = -\frac{d[\not< H(e^{j\omega})]}{d\omega}, \tag{32}$$

where $\not< H(e^{j\omega})$ is the phase of the transfer function. Using eqs. (31) and (32) we obtain

$$\tau_g(\omega) = \frac{1 - a^2}{1 + a^2 - 2a\cos\omega} \tag{33}$$

for the first-degree equalizer. Integrating eq. (33) from 0 to π and normalizing by 2π gives

$$\frac{1}{2\pi}\int_0^{\pi} \frac{1 - a^2}{1 + a^2 - 2a\cos\omega}\, d\omega = \frac{1}{\pi}\tan^{-1}\left[\frac{(1 - a^2)\tan(\omega/2)}{(1 - a)^2}\right]\Bigg|_0^{\pi} = \frac{\pi}{2\pi} = 0.5.$$

The significance of eq. (30) is that one can estimate the minimum-order equalizer required to equalize a given group delay characteristic by determining the area between the line $\tau = \tau_{\max}$ and the curve $\tau_g(\omega)$ and dividing by π, where $\tau_{\max}$ is the maximum value of $\tau_g(\omega)$ in the passband. Thus in the example of Fig. 8, the estimated order of the equalizer is approximately $(13 \times \pi/2)/\pi = 6.5$. Of course, the required order of the equalizer must be greater than the estimate given above, since this estimate assumes the delay of the equalizer exactly compensates the delay of the unequalized filter. As the degree of the equalizer is increased over the estimate, the peak error of approximation decreases monotonically.

We have used the above algorithm, along with initial estimates of equalizer order, to equalize three sets of elliptic filters. The data for

Table I — Comparisons between optimum FIR and equalized elliptic digital filters
(Set 1: $\delta_1 = 0.01$, $\delta_2 = 0.0001$)

F_p	F_s	n	N	n_e	$\bar{\tau}_g$	r	N_1*	N_2*
0.0502	0.13999	5	21	2	28.7	12.1	11	11
				4	42.7	3.4		13
0.09846	0.24111	5	21	2	14.5	11.6	11	11
				4	22.2	4.1		13
				6	29.4	0.8		15
0.14722	0.25297	6	21	4	17.6	13.1	11	14
				6	23.0	6.3		16
				8	28.5	2.6		18
0.19507	0.30647	6	21	4	13.8	16.0	11	14
				6	17.8	8.7		16
				8	22.0	4.2		18
0.24163	0.34842	6	21	6	14.5	11.1	11	16
				8	18.3	7.0		18
				10	21.8	3.6		20
0.28664	0.41668	5	21	6	11.6	8.4	11	15
				8	14.5	3.8		17
				10	17.3	1.6		19
0.33014	0.43727	5	21	6	10.7	14.7	11	15
				8	13.1	8.3		17
				10	15.7	4.5		19
0.37254	0.47479	4	21	6	8.7	19.1	11	13
				8	11.1	6.5		15
				10	13.4	3.2		17
0.41665	0.49417	3	21	8	9.6	6.3	11	14
				10	11.8	3.2		16

* N_1 is the number of multiplications per sample for the optimum FIR filter; N_2 is the number of multiplications per sample for the equalized elliptic filter.

these three sets of filters are given in Tables I through III. Included in the table are the filter specifications (δ_1, δ_2, F_p, F_s); the required elliptic order n; the required FIR filter duration N; the equalizer order n_e; the average passband delay, $\bar{\tau}_g$ (in samples), of the equalized filter; the percentage ripple, r, in the passband group delay of the equalized filter; and a comparison between the number of multiplications per sample required in both the optimum FIR filter and the equalized elliptic filters. The data in these tables indicate that to achieve equalization to within about a 3-percent error requires on the order of 30 percent more multiplications per sample for the equalized filter than for the optimum FIR design, although in most cases the unequalized elliptic filter was more efficient than the optimum FIR designs. Thus it would appear that, at least for these restricted results, if constant group delay is required in addition to the equiripple magni-

Table II — Comparisons between optimum FIR and equalized elliptic digital filters
(Set 2: $F_p = 0.25$, $\delta_1 = 0.02$, $\delta_2 = 0.001$)

F_s	n	N	n_e	$\bar{\tau}_g$	r	N_1*	N_2*
0.4893	2	11	2	3.3	1.2	6	6
			4	5.6	0.1		8
0.44816	3	13	2	4.5	9.4	7	8
			4	7.3	1.0		10
0.39146	4	19	2	5.9	25.1	10	9
			4	8.8	8.0		11
			6	11.9	2.2		13
0.34153	5	29	2	8.4	37.4	15	11
			4	10.6	21.6		13
			6	13.7	11.6		15
			8	16.9	5.6		17
			10	20.3	2.4		19
0.30639	6	45	4	13.8	34.7	23	14
			6	16.0	25.0		16
			8	18.7	16.9		18
			10	22.0	11.7		20
			12	25.5	7.9		22
			14	29.4	5.2		24
			16	32.8	3.2		26
			18	36.3	1.8		28

* N_1 is the number of multiplications per sample for the optimum FIR filter; N_2 is the number of multiplications per sample for the equalized elliptic filter.

tude characteristics, then the optimum FIR filter is always more efficient than an equalized elliptic filter. It should also be noted that the delay of the optimum FIR filter [$(N - 1)/2$ samples] was *always* less than the delay of the equalized elliptic filter.

The examples of Tables I through III considered filters where the order of the unequalized elliptic filter was six or less. It can be argued that, for higher-order elliptic designs, the relative efficiency of the elliptic filter over the optimum FIR filter is far greater than for lower-order designs; hence in these cases perhaps the equalized filter may still be more efficient than the optimum FIR design. This conjecture turns out to be untestable because high-order elliptic filters have a peak passband delay $\tau_{\max}$ which is much larger than for low-order filters, hence the order required for the equalizer becomes extremely large and thus is not even practical to consider if equalization over the entire passband is required. To illustrate this point, Fig. 9 shows the group delay of a 10th-order elliptic low-pass filter with $F_p = 0.25$. Using eq. (30) to get an estimate of n_e we arrive at a value of $n_e = 45$. Since this value of n_e is only an underbound on the actual order of the

Table III — Comparisons between optimum FIR and equalized elliptic digital filters
(Set 3: $F_p = 0.25$, $\delta_1 = 0.02$, $\delta_2 = 0.0001$)

F_s	n	N	n_e	$\bar{\tau}_g$	r	N_1*	N_2*
0.49661	2	11	2	3.3	1.2	6	6
			4	5.6	0.1		8
0.47564	3	11	2	4.5	9.1	6	8
			4	7.3	1.0		10
0.43591	4	17	2	5.8	23.3	9	9
			4	8.8	7.0		11
			6	11.8	1.7		13
0.38983	5	21	2	8.0	33.4	11	11
			4	10.3	18.0		13
			6	13.5	8.7		15
			8	16.7	3.7		17
			10	20.0	1.4		19
0.34878	6	31	4	12.8	28.9	16	14
			6	15.5	19.2		16
			8	18.2	11.8		18
			10	22.0	7.7		20
			12	25.3	4.3		22
			14	28.8	2.2		24

* N_1 is the number of multiplications per sample for the optimum FIR filter; N_2 is the number of multiplications per sample for the equalized elliptic filter.

equalizer, it is clear that it is not practical to try to obtain such a high-degree equalizer.

Another interesting question which arises when one considers the idea of equalizing an IIR filter is how does the cascade combination of an elliptic filter and an all-pass equalizer compare to the optimum IIR filter which best approximates both the desired magnitude and group delay characteristics? It is clear that the optimum IIR filter can be no worse than the cascade; the question remains as to how much

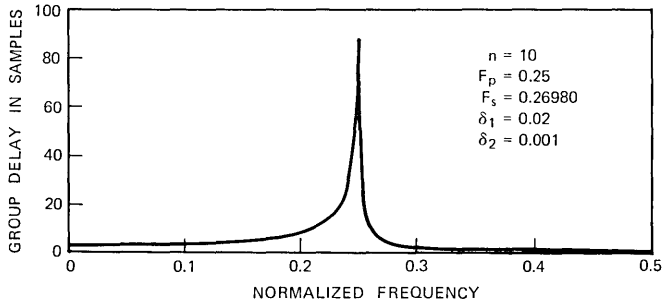

Fig. 9—The group delay of a 10th-order elliptic filter with $F_p = 0.25$.

better it can be. There is no clear-cut answer to this question. However, based on our experience with equalized elliptic filters, several observations can be made. (We shall use the z-plane pole-zero plot of a typical equalizer filter, shown in Fig. 10, to aid in understanding the nature of the equalized filter.)

(*i*) The zeros of the elliptic filter lie on the unit circle to give good stopband attenuation.
(*ii*) The zeros of the equalizer lie outside the unit circle to give positive delay.
(*iii*) The poles of the elliptic filter are constrained by the transition width requirements of the low-pass filter.
(*iv*) The poles of the equalizer lie approximately on a circle of fixed radius, and are approximately equally spaced in the passband.

If the zeros of the optimum filter are not constrained to lie on the unit circle, then each second-order section will require four multiplications per sample, rather than the three multiplications for each second-order section of the elliptic design and the two multiplications for each second-order section of the all-pass equalizer. Based on the

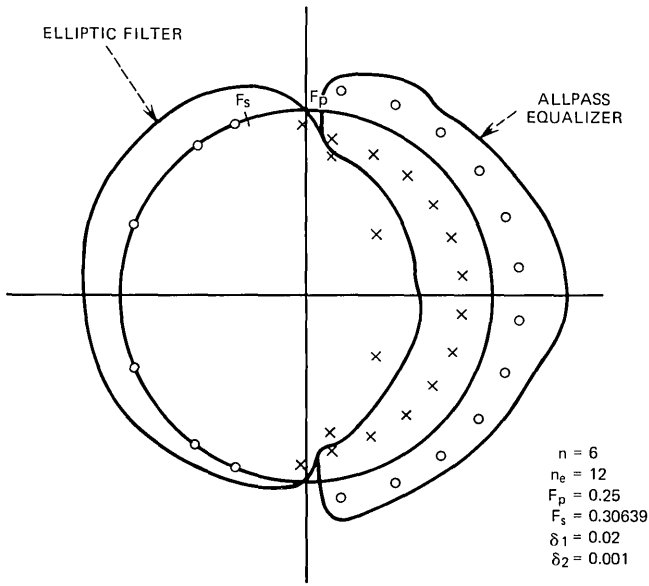

Fig. 10—The pole-zero positions of an equalized elliptic filter.

above observations, it seems unlikely that there is much to gain by using the optimum IIR filter over the equalized filter.

VIII. GENERAL DISCUSSION

In this paper we have considered only one basis for comparison between optimum FIR filters and equivalent IIR designs, that measure being the number of multiplications per sample required in the standard method of realization for each of these filter types. The justification for this measure is that in hardware (and generally in software) the number of multiplications per sample is an excellent measure of the complexity required in the implementation as well as the factor which determines the maximum throughput rate of the system.[9] However, there are many other ways for comparing these filter types when one takes into consideration the various finite word-length effects which occur in a practical design situation. In this section we review several of these design issues.

Among the various finite word-length effects are roundoff noise, both uncorrelated and correlated (e.g., limit cycles), and coefficient quantization sensitivity. For direct-form FIR realization, the peak roundoff noise can easily be made to be less than $\frac{1}{2}$ of the least significant bit by accumulating partial sums in an extended length register and then rounding the final result. For cascade IIR filters realized with fixed-point arithmetic, the roundoff noise problem is inherently related to the dynamic range problem,[7] and involves the concepts of pole-zero pairing and section ordering. Jackson[10] has shown that with reasonable pairing and ordering the uncorrelated roundoff noise variance can be minimized. However, even in the best of situations, the roundoff noise is equivalent to several bits. In terms of correlated roundoff noise, i.e., limit cycles, the direct-form FIR realization has no zero-input limit cycles (because no feedback is present), whereas the cascade IIR realization will generally exhibit zero-input limit cycles. Kaiser[11] has extensively studied these limit cycles and has developed bounds and estimates for their amplitude and frequency.

The coefficient quantization problem is one of the most difficult finite word length effects to treat analytically. Rounding of infinite precision filter coefficients to a fixed number of bits alters the overall frequency response of the filter in a complicated manner. Avenhaus[12] has shown that straight rounding of the infinite precision filter coefficients is generally inferior to optimizing the filter performance over the finite set of fixed precision filter coefficients. However, there are

no general procedures for performing this optimization, nor are there any guarantees of convergence of the existing methods. Furthermore, in many cases the advantage of optimizing finite precision coefficients over straight rounding of the infinite precision coefficients is small. Thus for the case of coefficient quantization neither direct-form realization of FIR filters nor cascade realization of IIR filters seems to offer a relative advantage here.

Thus it is difficult, if not impossible, to be quantitative in comparing FIR and IIR filters based on anything other than number of multiplications per sample. This is why we have used this measure throughout this paper.

IX. SUMMARY

In this paper some comparisons were made between equivalent FIR and IIR digital filters based on the number of multiplications per sample required to realize these filters. In the case of low-pass filters with quasi-equiripple magnitude characteristics, IIR elliptic filters could generally be realized more efficiently than equivalent linear phase FIR filters. When the additional requirement of constant group delay in the passband was added to the specifications, comparisons showed the linear phase FIR filters to be more efficient than group-delay-equalized elliptic IIR filters.

Additionally, a novel set of design charts for determining the minimum filter order required to meet given filter specifications for both digital and analog elliptic, Chebyshev, and Butterworth low-pass filters was presented. Explanation of how to use these charts to gain insight into the various filter parameter tradeoffs was also given.

REFERENCES

1. J. E. Storer, *Passive Network Synthesis*, New York: McGraw-Hill Book Co., 1957.
2. O. Herrmann, L. R. Rabiner, and D. S. K. Chan, "Practical Design Rules for Optimum Finite Impulse Response Low-Pass Digital Filters," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 769–799.
3. J. F. Kaiser, "Digital Filters," in *System Analysis by Digital Computer*, edited by F. F. Kuo and J. F. Kaiser, New York: John Wiley and Sons, 1966.
4. R. M. Golden and J. F. Kaiser, "Design of Wideband Sampled Data Filters," B.S.T.J., *43*, No. 4, Pt. 2 (July 1964), pp. 1533–1546.
5. A. G. Constantinides, "Spectral Transformations for Digital Filters," Proc. IEE, *117*, No. 8, 1970, pp. 1585–1590.
6. E. Christian and E. Eisenmann, *Filter Design Tables and Graphs*, New York: John Wiley and Sons, 1966.
7. L. B. Jackson, "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," B.S.T.J., *49*, No. 2 (February 1970), pp. 159–184.
8. A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum p-Error Criterion," IEEE Trans. Audio and Electroacoustics, *AU-20*, No. 4 (October 1972), pp. 257–263.

9. L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio and Electroacoustics, *AU-16*, No. 3 (September 1968), pp. 413–421.
10. L. B. Jackson, "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form," IEEE Trans. Audio and Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 107–122.
11. J. F. Kaiser, "An Overview on Digital Filters," Newsletter of IEEE Circuit Theory Group, *6*, No. 1 (March 1972).
12. E. Avenhaus, "On the Design of Digital Filters with Coefficients of Limited Word Length," IEEE Trans. Audio and Electroacoustics, *AU-20*, No. 3 (August 1972), pp. 206–212.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

On the Behavior of Minimax Relative Error FIR Digital Differentiators

By L. R. RABINER and R. W. SCHAFER

(Manuscript received June 28, 1973)

Optimum (in a minimax relative error sense) linear phase FIR digital differentiators can be designed in an efficient manner using a Remez optimization procedure. This paper presents data on wideband differentiators designed with even and odd values of N, the filter impulse response duration in samples. Based on these data, several interesting observations can be made, including:

(i) Differentiators with even values of N have peak relative errors which are approximately one to two orders of magnitude smaller than identical bandwidth differentiators with odd values of N, and with the same number of multiplications per sample in a direct convolution realization.

(ii) The smaller the bandwidth of the differentiator, the faster the decrease of the peak relative error with increasing N.

(iii) The larger the value of N, the faster the decrease of the peak relative error with decreasing bandwidth.

These observations lead to the conclusions that the bandwidth of a differentiator should be made as small as possible, and that even values of N should be used whenever possible. Complete tables of values of the impulse response coefficients are included for several wideband differentiators for even and odd values of N.

I. INTRODUCTION

In the past few years a great deal of work has been done in devising filter design techniques capable of obtaining optimum approximations (in the Chebyshev or minimax sense) to a prescribed frequency response characteristic. A general-purpose algorithm now exists[1] for the design of such optimum linear phase finite-duration impulse response (FIR) approximations to any desired multiband filter, differentiator, or Hilbert transformer. Since differentiators are an integral

part of many practical systems,[2–4] it is the purpose of this paper to present new data on the characteristics of optimum FIR differentiators, as an aid in making informed decisions concerning their use.

II. DISCRETE-TIME DIFFERENTIATORS

A differentiator is a system whose output is the derivative of its input. The frequency response of a differentiator is purely imaginary and is proportional to frequency. A sequence of samples of the derivative of a band-limited signal can be obtained by filtering a sequence of samples of the signal with a digital filter that approximates the ideal frequency response of a differentiator over the bandwidth of the signal. Therefore, digital filters having this type of frequency response are also called differentiators.

The frequency response of the ideal digital differentiator with a delay of τ samples is

$$\begin{aligned} H_d(e^{j\omega}) &= j\omega e^{-j\omega\tau} && 0 \leqq \omega \leqq \pi \\ &= j(\omega - 2\pi)e^{-j(\omega-2\pi)\tau} && \pi < \omega \leqq 2\pi. \end{aligned} \tag{1}$$

The impulse response corresponding to eq. (1) is obtained as the inverse Fourier transform of eq. (1) and is given by

$$h_d(n) = \frac{1}{2\pi}\left[\int_0^{\pi} j\omega e^{-j\omega\tau}e^{j\omega n}d\omega + \int_{\pi}^{2\pi} j(\omega - 2\pi)e^{-j(\omega-2\pi)\tau}e^{j\omega n}d\omega\right], \tag{2}$$

which can be written as

$$h_d(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} j\omega e^{j\omega(n-\tau)}d\omega \tag{3}$$

$$= \frac{\cos[\pi(n-\tau)]}{(n-\tau)} - \frac{\sin[\pi(n-\tau)]}{\pi(n-\tau)^2}. \tag{4}$$

For $\tau = 0$, eq. (4) gives

$$\begin{aligned} h_d(n) &= 0 && n = 0 \\ &= \frac{\cos(\pi n)}{n} && n \neq 0, \end{aligned} \tag{5}$$

whereas for $\tau = -\frac{1}{2}$ (i.e., a half-sample advance) eq. (4) gives

$$h_d(n) = \frac{-\cos(\pi n)}{\pi(n+\frac{1}{2})^2} = \frac{-4}{\pi}\frac{\cos(\pi n)}{(2n+1)^2}. \tag{6}$$

The impulse response of eq. (5), which corresponds to an ideal differentiator with zero delay, is of infinite duration and obeys the symmetry

condition

$$h_d(n) = -h_d(-n) \qquad n = 1, 2, \cdots. \tag{7}$$

The impulse response of eq. (6), which corresponds to an ideal differentiator with one-half-sample advance, is of infinite duration and obeys the symmetry condition

$$h_d(n) = -h_d(-n-1) \qquad n = 0, 1, \cdots. \tag{8}$$

In Ref. 2 it was shown that the frequency response of an ideal differentiator with zero delay had a discontinuity at half the sampling frequency [i.e., $\omega = \pi$ in eq. (1) with $\tau = 0$], whereas the frequency response of an ideal differentiator with a one-half-sample advance had no discontinuity at $\omega = \pi$. The frequency response of a half-sample-advance differentiator has a slope discontinuity at $\omega = \pi$ but, as we will see, slope discontinuities are much easier to approximate than function discontinuities. It should be noted that the output of a one-half-sample-advance differentiator is the derivative of the input signal evaluated midway between input samples. For numerical analysis applications where one desires the derivative at the sample point rather than midway between samples, the use of differentiators with zero delay is required. For most signal processing applications, either type of differentiator is generally appropriate.

We have only considered $\tau = 0$ and $\tau = -\frac{1}{2}$ as possible delays for the ideal differentiator. It can be seen from eq. (4) that these are the only values of $(-1 < \tau \leqq 0)$ such that the impulse response has desirable symmetry properties.

In order to obtain a causal approximation to the ideal differentiator which has no phase distortion (other than that corresponding to delay), it can be shown that an FIR approximation is required.[5] Therefore, consider a causal FIR filter with impulse response $h(n)$, $0 \leqq n \leqq N-1$, and frequency response

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)e^{-j\omega n}. \tag{9}$$

For this system to have exactly linear phase, the impulse response sequence must satisfy the condition

$$h(n) = -h(N-1-n) \qquad n = 0, 1, \cdots, N-1. \tag{10}$$

For N an odd integer, this means that $h(n)$ has odd symmetry about the sample at $n = (N-1)/2$. [This case corresponds to the $\tau = 0$ case above with an additional delay of $(N-1)/2$ samples.] For N

even, $h(n)$ has odd symmetry about a point halfway between the samples at $n = N/2$ and $n = (N/2) + 1$. (This case corresponds to the $\tau = -\frac{1}{2}$ case above with an additional delay of $N/2$ samples.) This implies that the frequency response of a filter satisfying eq. (10) can be expressed as

$$H(e^{j\omega}) = e^{-j\omega(N-1)/2}[jH^*(e^{j\omega})], \tag{11}$$

where $H^*(e^{j\omega})$ is a purely real function of ω. In particular, for N odd, $H^*(e^{j\omega})$ is

$$H^*(e^{j\omega}) = \sum_{n=1}^{(N-1)/2} a(n) \sin(\omega n), \tag{12a}$$

where

$$a(n) = 2h\left(\frac{N-1}{2} - n\right) \qquad n = 1, 2, \cdots, \left(\frac{N-1}{2}\right). \tag{12b}$$

Also, from eq. (10),

$$h\left(\frac{N-1}{2}\right) = 0. \tag{12c}$$

For N even, the expression for $H^*(e^{j\omega})$ is

$$H^*(e^{j\omega}) = \sum_{n=1}^{N/2} b(n) \sin[\omega(n - \tfrac{1}{2})], \tag{13a}$$

where

$$b(n) = 2h\left(\frac{N}{2} - n\right) \qquad n = 1, 2, \cdots, N/2. \tag{13b}$$

The factor $e^{-j\omega(N-1)/2}$ in eq. (11) represents a delay of $(N-1)/2$ samples, which may be accounted for when necessary. Therefore, in approximating the differentiator, the coefficients $a(n)$ and $b(n)$ must be chosen so that $jH^*(e^{j\omega})$ approximates the ideal differentiator frequency response of eq. (1) (with τ either 0 or $-\frac{1}{2}$). In many cases it is not required, and often it is not desirable, that the approximation be carried out over the entire band $0 \leqq \omega \leqq \pi$. Thus, in general, the real function $H^*(e^{j\omega})$ must approximate

$$\begin{aligned} D(e^{j\omega}) &= \omega && 0 \leqq \omega \leqq 2\pi F_p \\ &= (\omega - 2\pi) && 2\pi(1 - F_p) \leqq \omega \leqq 2\pi, \end{aligned} \tag{14}$$

where F_p is the highest frequency where a good approximation to a differentiator is desired. In the interval $2\pi F_p < \omega < 2\pi(1 - F_p)$, the frequency response of the differentiator is generally left unconstrained —although it certainly could be constrained to approximate zero over

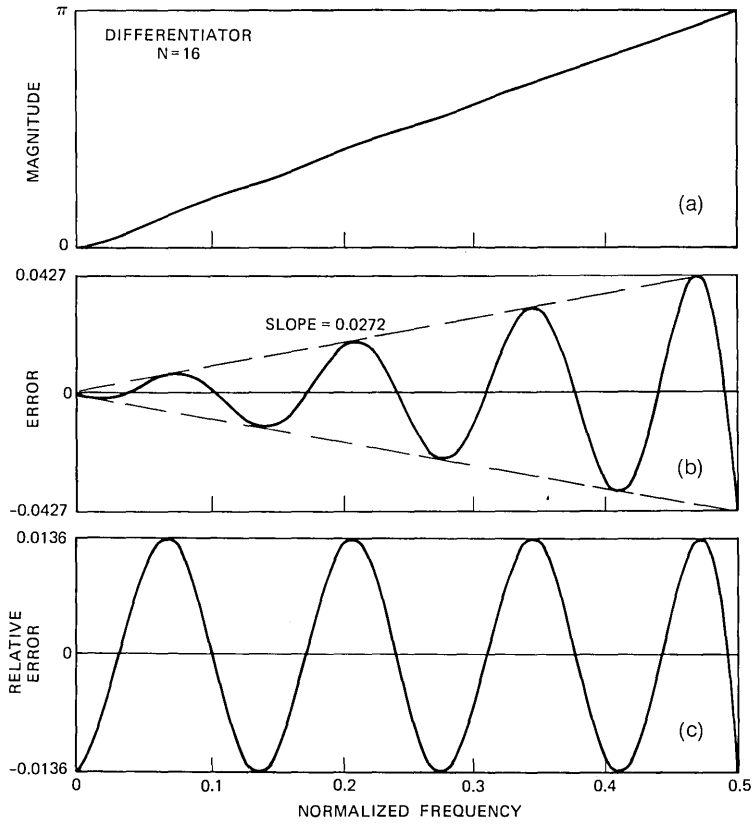

Fig. 1—The frequency response and error curves of an $N = 16$ wideband differentiator.

this interval. The frequency F_p is called the bandwidth of the differentiator. When $F_p = 0.5$, eqs. (1) and (14) are identical (to within the delay τ). This case is called a full-band differentiator.

The iterative Remez algorithm of McClellan, Parks, and Rabiner can be used to choose the values of $a(n)$ or $b(n)$ that minimize the peak relative error of approximation

$$\delta = \max_{0 \leq \omega \leq 2\pi F_p} \left[\frac{D(e^{j\omega}) - H^*(e^{j\omega})}{D(e^{j\omega})} \right]. \tag{15}$$

Our present concern is the general properties of the resulting approximations rather than the details of the approximation algorithm which are available in Ref. 1. The general properties of differentiators de-

signed by the optimization procedure are illustrated by examples given in the next section.

III. CHARACTERISTICS OF OPTIMUM DIFFERENTIATORS

The approximation to the ideal differentiator is characterized by a relative error function that is equiripple over the approximation band $0 \leqq \omega \leqq 2\pi F_p$. This is illustrated by Figs. 1 through 4 which are plots of the frequency responses of several wideband differentiators. Figure 1 shows the magnitude response, the approximation error $[D(e^{j\omega}) - H^*(e^{j\omega})]$, and the relative approximation error $[D(e^{j\omega}) - H^*(e^{j\omega})]/[D(e^{j\omega})]$ as a function of frequency for $N = 16$ and $F_p = 0.5$. The relative error curve can be seen to be equiripple

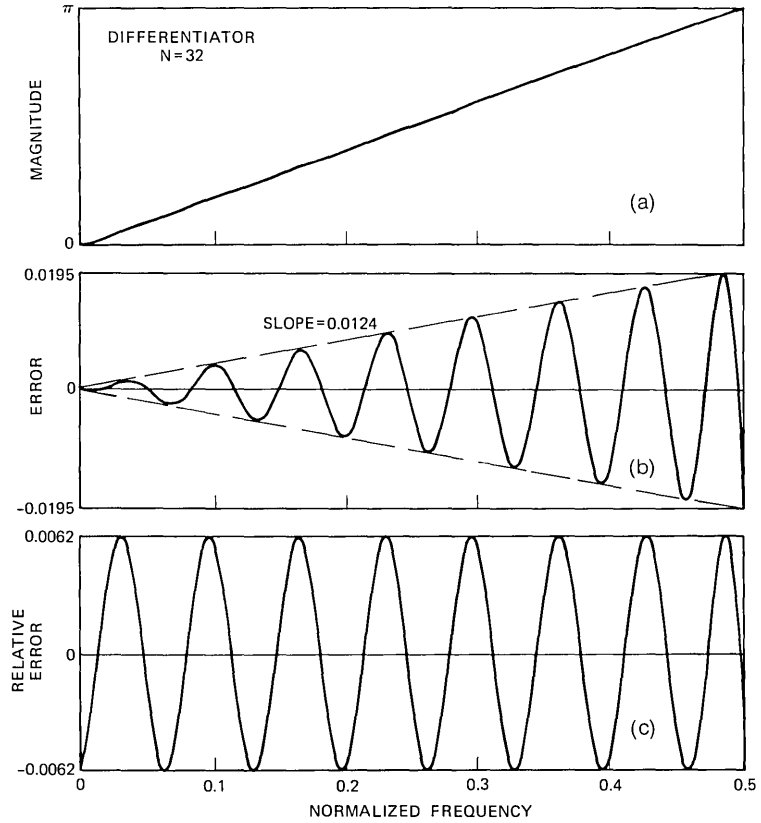

Fig. 2—The frequency response and error curves of an $N = 32$ wideband differentiator.

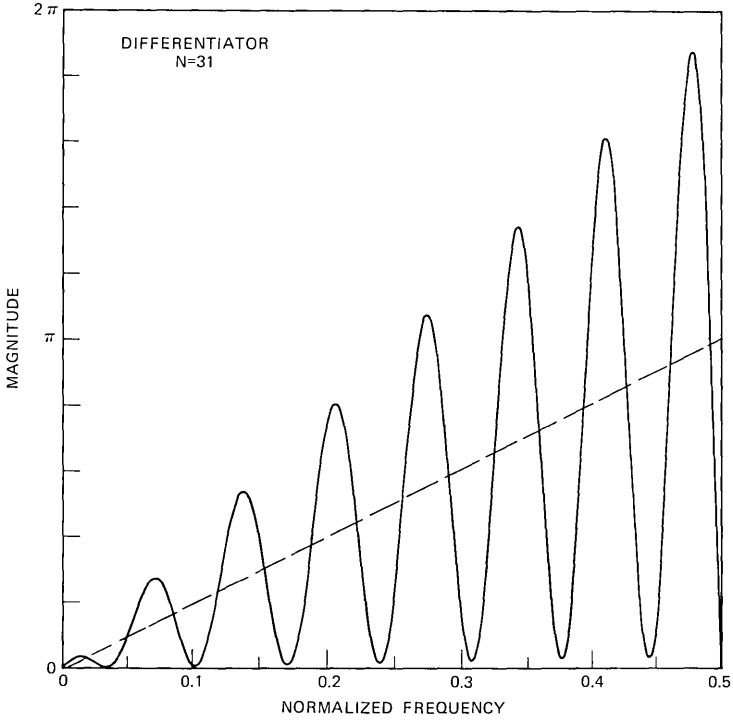

Fig. 3—The frequency response of an $N = 31$ wideband differentiator.

with a peak relative error of 0.0136. That is, the maximum error is 1.36 percent of the desired frequency response over the entire band $0 \leqq \omega \leqq \pi$. Figure 2 shows the same curves for $N = 32$ and $F_p = 0.5$. By doubling N, the peak relative error is reduced to 0.0062. Figure 3 shows the magnitude response for $N = 31$ and $F_p = 0.5$. In this case, the relative error at $f = 0.5$ is 1.0 since the desired value is π and the approximation is zero. The reason for the undesirable behavior of the approximation in this case is apparent from eqs. (12a) and (13a). When N is odd, $H^*(e^{j\omega})$ is exactly zero at $\omega = 0$ and $\omega = \pi$, independent of the choice of $a(n)$ in eq. (12a). When N is even, $H^*(e^{j\omega})$ is exactly zero only at $\omega = 0$, independent of the choice of $b(n)$ in eq. (13a). It is quite desirable to have a zero of $H^*(e^{j\omega})$ at $\omega = 0$ since this is the desired response; however, a zero of $H^*(e^{j\omega})$ at $\omega = \pi$ is at odds with the desired response at $\omega = \pi$. Thus, the inherent zero at $\omega = \pi$ for N odd is a fundamental limitation to the design of extremely wideband differentiators.[2] However, if F_p is less than 0.5, the resulting

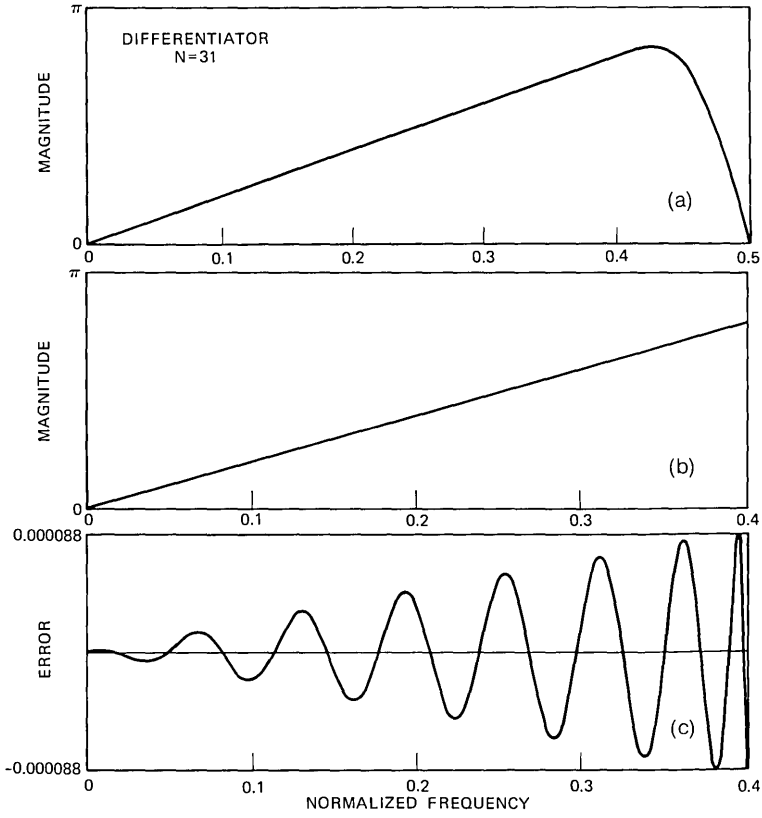

Fig. 4—The frequency response and error curve of an $N = 31$, $F_p = 0.4$ wideband differentiator.

approximation can be quite acceptable as seen in Fig. 4a for the case $N = 31$ and $F_p = 0.4$. Figure 4b shows the magnitude response plotted from $f = 0$ to $f = 0.4$, the cutoff frequency, and Fig. 4c shows the error function over this same range. The peak relative error is 0.000088 over the approximation interval.

As is clear from these examples, the basic parameters that characterize these differentiator approximations are N, F_p, and δ, the peak relative error of approximation. The examples suggest that δ can be reduced by increasing N, and by decreasing F_p. Also, there seems to be a distinct advantage in choosing even values of N provided the half-sample delay is not undesirable. To substantiate these observations, a large set of measurements of δ as a function of F_p and N were

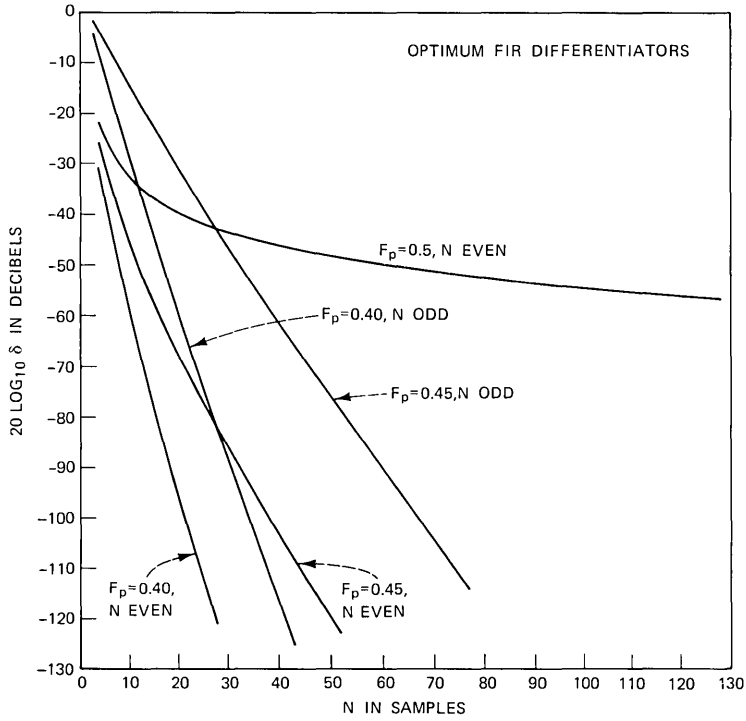

Fig. 5—The curves of 20 $\log_{10} \delta$ versus N for $F_p = 0.5, 0.45, 0.40$ for even and odd values of N.

made. The results are shown in Figs. 5 through 8* and in Tables I through VIII. Figure 5 shows the dependence of 20 $\log_{10} \delta$ upon N for F_p equal to 0.5, 0.45, and 0.40 and for even and odd values of N in the range $3 \leqq N \leqq 128$. The curve for N odd, $F_p = 0.5$, is not included since $\delta = 1.0$ independent of N. Figure 6 shows the same data as Fig. 5 with a logarithmic horizontal scale. From Figs. 5 and 6 it is seen that for the same value of F_p the values of δ for even values of N are approximately 1 to 2 orders of magnitude (20–40 dB) smaller than the values of δ for comparable odd values of N. This difference between even and odd values of N is due to the frequency response discontinuity for odd N, which is considerably more difficult to approximate than the slope discontinuity in the frequency response for even N. To substantiate this claim, the curve for $F_p = 0.5$, N even, was subtracted (on a log scale) from the $F_p = 0.45$, N even, and the

* The curves in Figs. 5 through 8 are straight-line connections of measured data points and do not represent any smoothing or fitting of the data.

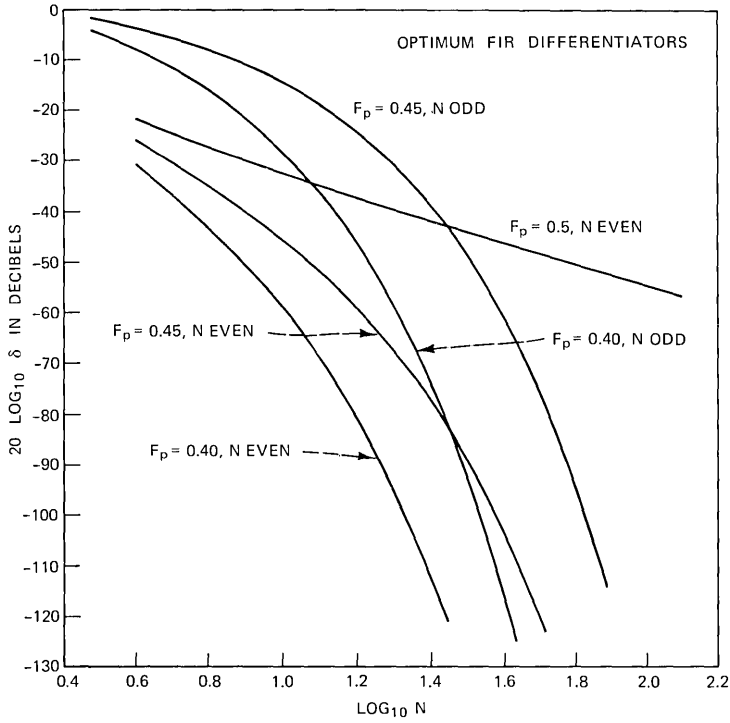

Fig. 6—The curves of 20 $\log_{10} \delta$ versus $\log_{10} N$ for $F_p = 0.5$, 0.45, and 0.40 for even and odd values of N.

$F_p = 0.40$, N even, curves and the resulting curves were replotted on the same scales as Fig. 5. It was then seen that the difference between the N even and N odd curves for fixed values of F_p was small (on the order of 2 dB) and essentially independent of N. Thus the 1 to 2 order-of-magnitude difference in the deltas is almost entirely accounted for by the frequency response discontinuity for odd values of N.

Another observation from Figs. 5 and 6 is that the smaller the bandwidth (F_p) of the differentiator, the faster the peak relative error decreases with increasing N. Thus for $F_p = 0.5$, the value of 20 $\log_{10} \delta$ decreases by only about 30 dB as N varies from 4 to 128; whereas for $F_p = 0.45$, the value of 20 $\log_{10} \delta$ decreases by about 98 dB as N varies from 4 to 52. For the cases when N is odd, the relation

$$(N - 1)(0.5 - F_p) \approx \alpha \log_{10} \delta \tag{16}$$

appears to work for δ smaller than approximately 0.01, where $\alpha \approx -0.5$. Thus, for $F_p = 0.45$, a value of $N = 41$ gives a δ of 0.000764, whereas for $F_p = 0.40$, a value of $N = 21$ gives a value of δ of 0.000878. This

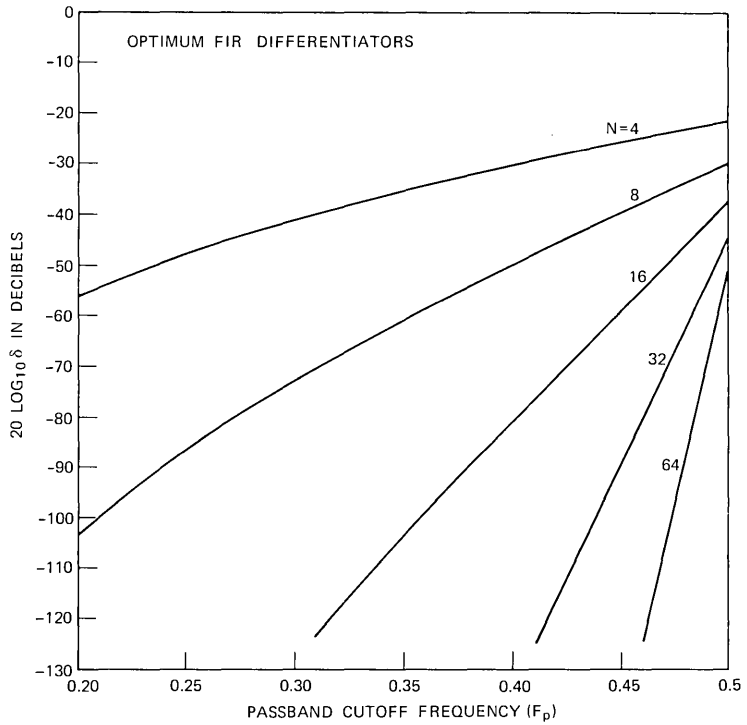

Fig. 7—The curves of 20 $\log_{10} \delta$ versus F_p for N = 4, 8, 16, 32, and 64.

inverse proportionality between transition bandwidth and filter order for fixed value of δ was originally noted by Kaiser.[6]

Figures 7 and 8 show the dependence of 20 $\log_{10} \delta$ upon F_p for even values of N (N = 4, 8, 16, 32, 64) and odd values of N (N = 5, 9, 17, 33, 65) respectively. Note that since $h((N-1)/2) = 0$ when N is odd, there are the same number of nonzero impulse response coefficients for differentiators of length 4 and 5, 8 and 9, etc. The data for even and odd values of N are presented on different figures because of the different nature of the solution in the two cases. As seen in Fig. 8, where N is odd, as F_p approaches 0.5, 20 $\log_{10} \delta$ approaches 0, independent of N. As previously discussed, this is because of the constrained zero of $H^*(e^{j\omega})$ at $\omega = \pi$ when N is odd. For even values of N, the curves are spaced apart for all values of F_p. The main observation from these figures is that the larger the value of N, the faster the peak relative error decreases with decreasing differentiator bandwidth.

The curves in Figs. 5 through 8 can be used to determine the length of impulse response required to meet a given specification of approxi-

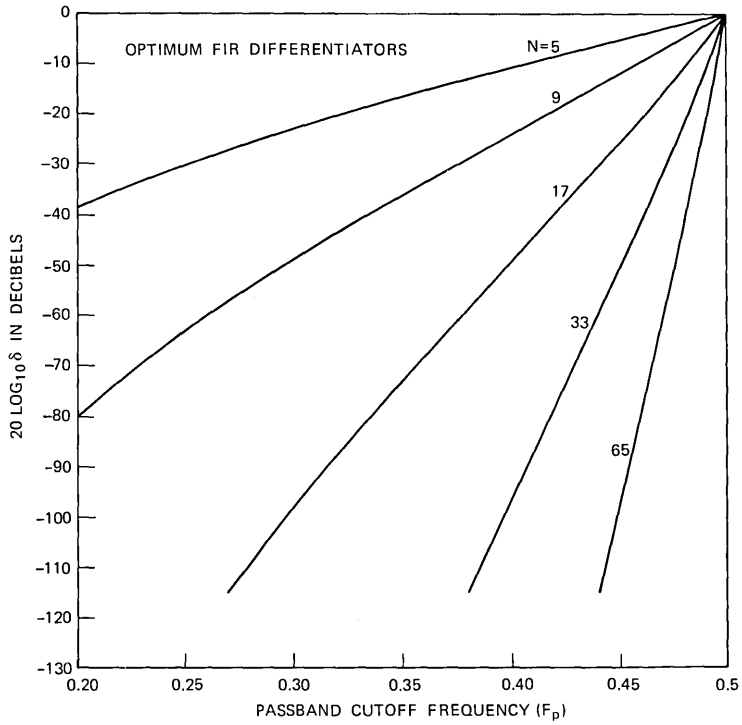

Fig. 8—The curves of 20 $\log_{10} \delta$ versus F_p for $N = 5$, 9, 17, 33, and 65.

mation error. For example, to obtain a peak relative error less than 1 percent (−40 dB) requires the following values of N (as a function of F_p):

F_p	N (odd)	N (even)
0.5	impossible	22
0.45	27	10
0.40	15	6

Similarly, to obtain a peak relative error less than 0.1 percent requires

F_p	N (odd)	N (even)
0.5	impossible	>128
0.45	41	18
0.40	21	12

These examples indicate the substantial reductions in N that result when F_p is reduced and when N is changed from odd to even.

IV. APPLICATION OF FIR DIFFERENTIATORS

The data presented above indicate that the most efficient FIR differentiators (i.e., having the smallest value of N for a given value of peak relative approximation error) are obtained when the bandwidth, F_p, is as small as possible and when N is even. These design considerations must be viewed in the light of the intended application.

In processing sequences obtained by sampling an analog signal, it is generally true that a fullband differentiator is not required, since the sampling rate is generally set somewhat higher than twice the Nyquist frequency of the analog signal. Thus, differentiator bandwidths on the order of 0.40 to 0.49 are quite reasonable for most applications.

Since even values of N result in better approximations than comparable odd values, it is generally desirable to choose N even. However, even values of N are sometimes undesirable because the delay is a nonintegral number of samples. In situations when the differentiator is part of a larger signal processing system it may be important to be able to equalize signal delays in different parts of the system. This may be more difficult to accomplish when the delay of the differentiator is not an integral number of samples. When N is large and the differentiator is to be realized as a general-purpose computer program, it may be desirable to use an FFT realization instead of a direct discrete convolution realization. In such cases, the odd symmetry of the impulse response about $(N - 1)/2$ permits the frequency response of the differentiator to be purely imaginary only when N is odd, thus reducing the storage requirements for the transform of the impulse response and the number of intermediate multiplications by a factor of two.

These two practical limitations, however, are often of little consequence in digital signal processing applications, where the overriding concern is often simply the reduction of system complexity while maintaining a prescribed performance. In such cases, it is clear from previous discussion that even values of N provide the most efficient approximations to the differentiator.

A subset of the differentiators designed in this study is given in Tables I through VIII. Included in these tables are wideband differentiators (F_p = 0.5, 0.49, 0.48, 0.45, and 0.40) for both even and odd values from $N = 3$ to $N = 50$.* The value of peak relative error,

* Only those differentiators for which $\delta < 0.1$ are included in these tables.

denoted D, is given for each case as well as the impulse response coefficients for the filter. Note that only the first half of the impulse response is given in the table; i.e., $h(n)$ for $n = 0, 1, \cdots, (N/2) - 1$. The remainder of the impulse response can be obtained using eq. (10). These data should be adequate for most design applications.

V. ACKNOWLEDGMENTS

The authors would like to acknowledge the valuable comments and criticisms provided by J. F. Kaiser and H. S. McDonald.

REFERENCES

1. J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," IEEE Trans. on Audio and Electroacoustics, *AU-21*, No. 6 (December 1973), pp. 506–526.
2. L. R. Rabiner and K. Steiglitz, "The Design of Wide-Band Recursive and Non-recursive Digital Differentiators," IEEE Trans. Audio and Elect., *AU-18*, June 1970, pp. 204–209.
3. J. L. Hall, "Simulations of Schroeder's Integrable Model for Motion of the Basilar Membrane," unpublished work.
4. J. L. Flanagan and R. M. Golden, "Phase Vocoder," B.S.T.J., *45*, No. 9 (November 1966), pp. 1493–1509.
5. A. J. Gibbs, "On the Frequency-Domain Responses of Causal Digital Filters," Ph.D. Thesis, Dept. of Electrical Engineering, University of Wisconsin, 1969.
6. J. F. Kaiser, "Digital Filters," Chapter 7 in *System Analysis by Digital Computer*, edited by F. F. Kuo and J. F. Kaiser, New York: John Wiley and Sons, 1966.

Table I—Wideband differentiators (F_p = 0.5, N even)

	$N = 4$ $D = 0.0831040$	$N = 6$ $D = 0.0470970$	$N = 8$ $D = 0.0320840$	$N = 10$ $D = 0.0240945$
0	−0.1310637	0.0508099	−0.0289328	0.0196350
1	1.3091933	−0.1630958	0.0668126	−0.0386111
2		1.2829105	−0.1471745	0.0549917
3			1.2774790	−0.1441674
4				1.2755435

	$N = 12$ $D = 0.0192165$	$N = 14$ $D = 0.0159235$	$N = 16$ $D = 0.0135730$	$N = 18$ $D = 0.0118460$
0	−0.0146791	0.0116236	−0.0095995	0.0081685
1	0.0261968	−0.0194562	0.0153294	−0.0125786
2	−0.0291034	0.0182665	−0.0126660	0.0093698
3	0.0528837	−0.0275156	0.0169834	−0.0115683
4	−0.1430612	0.0521278	−0.0269451	0.0165044
5	1.2746868	−0.1425363	0.0517493	−0.0266517
6		1.2742573	−0.1422334	0.0515315
7			1.2739692	−0.1420377
8				1.2737782

	$N = 20$ $D = 0.0104985$	$N = 22$ $D = 0.0094195$	$N = 24$ $D = 0.0085365$	$N = 26$ $D = 0.0078050$
0	−0.0071000	0.0062722	−0.0056118	0.0050796
1	0.0106302	−0.0091800	0.0080620	−0.0071883
2	−0.0072703	0.0058390	−0.0048205	0.0040624
3	0.0084292	−0.0064393	0.0051079	−0.0041422
4	−0.0111803	0.0081015	−0.0061625	0.0048465
5	0.0162637	−0.0109978	0.0079432	−0.0060168
6	−0.0264660	0.0161349	−0.0108840	0.0078424
7	0.0513867	−0.0263595	0.0160438	−0.0108017
8	−0.1419117	0.0512852	−0.0262854	0.0159807
9	1.2736673	−0.1418134	0.0512146	−0.0262332
10		1.2735768	−0.1417436	0.0511684
11			1.2735077	−0.1417041
12				1.2734707

	$N = 28$ $D = 0.0071890$	$N = 30$ $D = 0.0066600$	$N = 32$ $D = 0.0062025$	$N = 34$ $D = 0.0058025$
0	−0.0046291	0.0042600	−0.0039405	0.0036653
1	0.0064623	−0.0058789	0.0053806	−0.0049612
2	−0.0034847	0.0030260	−0.0026653	0.0023703
3	0.0034495	−0.0029082	0.0025070	−0.0021762
4	−0.0039226	0.0032396	−0.0027291	0.0023323
5	0.0047240	−0.0038139	0.0031397	−0.0026465
6	−0.0059332	0.0046549	−0.0037470	0.0030907
7	0.0077802	−0.0058820	0.0046040	−0.0037036
8	−0.0107505	0.0077387	−0.0058437	0.0045698
9	0.0159345	−0.0107147	0.0077095	−0.0058145
10	−0.0261921	0.0159024	−0.0106896	0.0076837
11	0.0511322	−0.0261619	0.0158792	−0.0106654
12	−0.1416711	0.0511027	−0.0261387	0.0158559
13	1.2734383	−0.1416415	0.0510788	−0.0261186
14		1.2734088	−0.1416173	0.0510609
15			1.2733839	−0.1416001
16				1.2733660

Table I—continued

	$N = 36$ $D = 0.0054505$	$N = 38$ $D = 0.0051375$	$N = 40$ $D = 0.0048585$	$N = 42$ $D = 0.0046080$
0	−0.0034253	0.0032145	−0.0030272	0.0028607
1	0.0046009	−0.0042886	0.0040140	−0.0037721
2	−0.0021287	0.0019267	−0.0017546	0.0016072
3	0.0019148	−0.0017005	0.0015212	−0.0013694
4	−0.0020200	0.0017690	−0.0015623	0.0013908
5	0.0022528	−0.0019478	0.0017009	−0.0014989
6	−0.0025912	0.0022079	−0.0019054	0.0016619
7	0.0030565	−0.0025573	0.0021778	−0.0018777
8	−0.0036810	0.0030238	−0.0025346	0.0021567
9	0.0045437	−0.0036587	0.0030046	−0.0025174
10	−0.0057900	0.0045295	−0.0036367	0.0029902
11	0.0076598	−0.0057714	0.0045142	−0.0036235
12	−0.0106443	0.0076422	−0.0057629	0.0044975
13	0.0158387	−0.0106280	0.0076300	−0.0057519
14	−0.0261035	0.0158233	−0.0106158	0.0076262
15	0.0510459	−0.0260878	0.0158104	−0.0106082
16	−0.1415850	0.0510323	−0.0260762	0.0158013
17	1.2733532	−0.1415734	0.0510217	−0.0260652
18		1.2733444	−0.1415624	0.0510100
19			1.2733312	−0.1415514
20				1.2733199

	$N = 44$ $D = 0.0043825$	$N = 46$ $D = 0.0041785$	$N = 48$ $D = 0.0039920$	$N = 50$ $D = 0.0038215$
0	−0.0027118	0.0025770	−0.0024552	0.0023443
1	0.0035572	−0.0033646	0.0031912	−0.0030348
2	−0.0014794	0.0013682	−0.0012698	0.0011841
3	0.0012416	−0.0011307	0.0010345	−0.0009519
4	−0.0012475	0.0011244	−0.0010198	0.0009299
5	0.0013308	−0.0011900	0.0010713	−0.0009698
6	−0.0014615	0.0012959	−0.0011583	0.0010418
7	0.0016346	−0.0014370	0.0012745	−0.0011376
8	−0.0018561	0.0016167	−0.0014213	0.0012579
9	0.0021391	−0.0018419	0.0016028	−0.0014081
10	−0.0025029	0.0021275	−0.0018293	0.0015925
11	0.0029785	−0.0024922	0.0021165	−0.0018212
12	−0.0036153	0.0029682	−0.0024834	0.0021086
13	0.0044897	−0.0036056	0.0029610	−0.0024759
14	−0.0057381	0.0044821	−0.0035984	0.0029537
15	0.0076162	−0.0057321	0.0044752	−0.0035915
16	−0.0106048	0.0076042	−0.0057262	0.0044686
17	0.0157953	−0.0105953	0.0075986	−0.0057202
18	−0.0260586	0.0157918	−0.0105840	0.0075939
19	0.0510019	−0.0260529	0.0157830	−0.0105799
20	−0.1415416	0.0509959	−0.0260495	0.0157733
21	1.2733098	−0.1415357	0.0509909	−0.0260416
22		1.2733032	−0.1415313	0.0509884
23			1.2732988	−0.1415278
24				1.2732960

Table II—Wideband differentiators ($F_p = 0.49$, N even)

	$N = 4$ $D = 0.0756235$	$N = 6$ $D = 0.0402290$	$N = 8$ $D = 0.0255765$	$N = 10$ $D = 0.0178930$
0	−0.1249465	0.0458333	−0.0244765	0.0155119
1	1.2992160	−0.1556062	0.0604056	−0.0328369
2		1.2778984	−0.1435042	0.0519839
3			1.2742206	−0.1416396
4				1.2731675

	$N = 12$ $D = 0.0132725$	$N = 14$ $D = 0.0102355$	$N = 16$ $D = 0.0081120$	$N = 18$ $D = 0.0065675$
0	−0.0108071	0.0079847	−0.0061368	0.0048544
1	0.0208881	−0.0145299	0.0107000	−0.0081958
2	−0.0265471	0.0159989	−0.0106296	0.0075254
3	0.0507873	−0.0256995	0.0153658	−0.0101326
4	−0.1411075	0.0504637	−0.0254815	0.0152063
5	1.2727987	−0.1409274	0.0503597	−0.0254177
6		1.2726648	−0.1408696	0.0503343
7			1.2726240	−0.1408602
8				1.2726199

	$N = 20$ $D = 0.0054065$	$N = 22$ $D = 0.0045055$	$N = 24$ $D = 0.0037945$	$N = 26$ $D = 0.0032230$
0	−0.0039242	0.0032252	−0.0026848	0.0022597
1	0.0064632	−0.0052075	0.0042657	−0.0035440
2	−0.0055757	0.0042694	−0.0033533	0.0026892
3	0.0071314	−0.0052518	0.0040021	−0.0031297
4	−0.0100173	0.0070397	−0.0051824	0.0039440
5	0.0151654	−0.0099887	0.0070237	−0.0051698
6	−0.0254039	0.0151620	−0.0099918	0.0070290
7	0.0503330	−0.0254130	0.0151745	−0.0100057
8	−0.1408656	0.0503509	−0.0254312	0.0151940
9	1.2726300	−0.1408860	0.0503717	−0.0254529
10		1.2726507	−0.1409083	0.0503940
11			1.2726749	−0.1409325
12				1.2726997

	$N = 28$ $D = 0.0027515$	$N = 30$ $D = 0.0023760$	$N = 32$ $D = 0.0020535$	$N = 34$ $D = 0.0017890$
0	−0.0019148	0.0016418	−0.0014118	0.0012240
1	0.0029716	−0.0025230	0.0021529	−0.0018542
2	−0.0021894	0.0018089	−0.0015120	0.0012767
3	0.0024976	−0.0020279	0.0016726	−0.0013946
4	−0.0030847	0.0024627	−0.0020003	0.0016484
5	0.0039389	−0.0030810	0.0024618	−0.0019999
6	−0.0051770	0.0039458	−0.0030894	0.0024683
7	0.0070419	−0.0051893	0.0039581	−0.0031014
8	−0.0100233	0.0070579	−0.0052056	0.0039744
9	0.0152150	−0.0100424	0.0070768	−0.0052210
10	−0.0254777	0.0152355	−0.0100622	0.0070931
11	0.0504213	−0.0254978	0.0152556	−0.0100801
12	−0.1409620	0.0504402	−0.0255185	0.0152744
13	1.2727299	−0.1409799	0.0504625	−0.0255380
14		1.2727475	−0.1410032	0.0504816
15			1.2727717	−0.1410217
16				1.2727889

Table II—continued

	$N = 36$ $D = 0.0015625$	$N = 38$ $D = 0.0013680$	$N = 40$ $D = 0.0012035$	$N = 42$ $D = 0.0010655$
0	−0.0010650	0.0009290	−0.0008146	0.0007191
1	0.0016041	−0.0013927	0.0012161	−0.0010685
2	−0.0010867	0.0009299	−0.0008017	0.0006955
3	0.0011746	−0.0009968	0.0008520	−0.0007339
4	−0.0013744	0.0011567	−0.0009814	0.0008401
5	0.0016487	−0.0013748	0.0011577	−0.0009836
6	−0.0020081	0.0016562	−0.0013820	0.0011649
7	0.0024812	−0.0020191	0.0016666	−0.0013917
8	−0.0031146	0.0024941	−0.0020314	0.0016776
9	0.0039898	−0.0031300	0.0025076	−0.0020436
10	−0.0052395	0.0040046	−0.0031441	0.0025208
11	0.0071107	−0.0052559	0.0040200	−0.0031576
12	−0.0100974	0.0071289	−0.0052706	0.0040335
13	0.0152926	−0.0101144	0.0071443	−0.0052845
14	−0.0255565	0.0153093	−0.0101310	0.0071569
15	0.0504998	−0.0255738	0.0153250	−0.0101445
16	−0.1410399	0.0505181	−0.0255892	0.0153398
17	1.2728072	−0.1410588	0.0505341	−0.0256034
18		1.2728270	−0.1410757	0.0505482
19			1.2728442	−0.1410899
20				1.2728587

	$N = 44$ $D = 0.0009435$	$N = 46$ $D = 0.0008355$	$N = 48$ $D = 0.0007420$	$N = 50$ $D = 0.0006635$
0	−0.0006352	0.0005614	−0.0004979	0.0004439
1	0.0009406	−0.0008294	0.0007336	−0.0006528
2	−0.0006060	0.0005300	−0.0004653	0.0004109
3	0.0006355	−0.0005529	0.0004835	−0.0004247
4	−0.0007235	0.0006264	−0.0005457	0.0004778
5	0.0008423	−0.0007260	0.0006293	−0.0005488
6	−0.0009909	0.0008492	−0.0007323	0.0006359
7	0.0011743	−0.0009997	0.0008573	−0.0007405
8	−0.0014027	0.0011844	−0.0010091	0.0008665
9	0.0016892	−0.0014134	0.0011947	−0.0010188
10	−0.0020555	0.0017009	−0.0014241	0.0012045
11	0.0025331	−0.0020675	0.0017119	−0.0014341
12	−0.0031702	0.0025453	−0.0020788	0.0017219
13	0.0040464	−0.0031824	0.0025566	−0.0020892
14	−0.0052977	0.0040593	−0.0031941	0.0025670
15	0.0071710	−0.0053102	0.0040709	−0.0032047
16	−0.0101577	0.0071839	−0.0053222	0.0040816
17	0.0153533	−0.0101712	0.0071961	−0.0053332
18	−0.0256184	0.0153665	−0.0101838	0.0072068
19	0.0505624	−0.0256316	0.0153794	−0.0101945
20	−0.1411034	0.0505765	−0.0256436	0.0153900
21	1.2728719	−0.1411169	0.0505881	−0.0256542
22		1.2728844	−0.1411291	0.0505982
23			1.2728967	−0.1411395
24				1.2729083

Table III—Wideband differentiators ($F_p = 0.48$, N even)

	$N = 4$ $D = 0.0685890$	$N = 6$ $D = 0.0342610$	$N = 8$ $D = 0.0203665$	$N = 10$ $D = 0.0132955$
0	−0.1194462	0.0415099	−0.0208464	0.0123838
1	1.2897494	−0.1488110	0.0549279	−0.0282322
2		1.2731593	−0.1400594	0.0492614
3			1.2710796	−0.1392143
4				1.2708226

	$N = 12$ $D = 0.0091960$	$N = 14$ $D = 0.0066075$	$N = 16$ $D = 0.0048850$	$N = 18$ $D = 0.0036880$
0	−0.0080758	0.0055801	−0.0040121	0.0029697
1	0.0169344	−0.0110911	0.0076843	−0.0055383
2	−0.0243059	0.0141089	−0.0090057	0.0061179
3	0.0488103	−0.0240382	0.0139418	−0.0088998
4	−0.1392103	0.0488703	−0.0241146	0.0140184
5	1.2709260	−0.1393617	0.0490132	−0.0242418
6		1.2711088	−0.1395291	0.0491596
7			1.2712857	−0.1396846
8				1.2714456

	$N = 20$ $D = 0.0028315$	$N = 22$ $D = 0.0022025$	$N = 24$ $D = 0.0017315$	$N = 26$ $D = 0.0013725$
0	−0.0022456	0.0017272	−0.0013462	0.0010612
1	0.0041061	−0.0031111	0.0023970	−0.0018733
2	−0.0043398	0.0031799	−0.0023882	0.0018287
3	0.0060476	−0.0042958	0.0031532	−0.0023725
4	−0.0089705	0.0061132	−0.0043558	0.0032063
5	0.0141281	−0.0090663	0.0061971	−0.0044281
6	−0.0243671	0.0142374	−0.0091603	0.0062788
7	0.0492947	−0.0244843	0.0143370	−0.0092466
8	−0.1398254	0.0494151	−0.0245867	0.0144256
9	1.2715885	−0.1399466	0.0495200	−0.0246775
10		1.2717111	−0.1400541	0.0496127
11			1.2718185	−0.1401477
12				1.2719131

	$N = 28$ $D = 0.0010945$	$N = 30$ $D = 0.0008810$	$N = 32$ $D = 0.0007135$	$N = 34$ $D = 0.0005780$
0	−0.0008419	0.0006751	−0.0005457	0.0004414
1	0.0014784	−0.0011790	0.0009488	−0.0007659
2	−0.0014238	0.0011215	−0.0008944	0.0007185
3	0.0018215	−0.0014209	0.0011234	−0.0008972
4	−0.0024200	0.0018636	−0.0014586	0.0011558
5	0.0032688	−0.0024753	0.0019117	−0.0014998
6	−0.0044975	0.0033307	−0.0025277	0.0019575
7	0.0063517	−0.0045625	0.0033857	−0.0025767
8	−0.0093227	0.0064189	−0.0046203	0.0034366
9	0.0145047	−0.0093921	0.0064786	−0.0046731
10	−0.0247589	0.0145754	−0.0094534	0.0065333
11	0.0496959	−0.0248305	0.0146379	−0.0095087
12	−0.1402316	0.0497675	−0.0248937	0.0146945
13	1.2719969	−0.1403032	0.0498316	−0.0249515
14		1.2720686	−0.1403686	0.0498901
15			1.2721348	−0.1404270
16				1.2721930

Table III—continued

	$N = 36$ $D = 0.0004745$	$N = 38$ $D = 0.0003865$	$N = 40$ $D = 0.0003195$	$N = 42$ $D = 0.0002630$
0	−0.0003616	0.0002947	−0.0002435	0.0002004
1	0.0006258	−0.0005102	0.0004213	−0.0003471
2	−0.0005831	0.0004750	−0.0003908	0.0003220
3	0.0007241	−0.0005881	0.0004825	−0.0003971
4	−0.0009268	0.0007493	−0.0006120	0.0005023
5	0.0011925	−0.0009582	0.0007785	−0.0006362
6	−0.0015400	0.0012274	−0.0009902	0.0008052
7	0.0019999	−0.0015771	0.0012610	−0.0010185
8	−0.0026210	0.0020389	−0.0016123	0.0012909
9	0.0034828	−0.0026616	0.0020753	−0.0016434
10	−0.0047202	0.0035246	−0.0026993	0.0021077
11	0.0065816	−0.0047636	0.0035632	−0.0027329
12	−0.0095586	0.0066259	−0.0048032	0.0035981
13	0.0147451	−0.0096042	0.0066668	−0.0048390
14	−0.0250027	0.0147916	−0.0096453	0.0067035
15	0.0499419	−0.0250492	0.0148333	−0.0096830
16	−0.1404791	0.0499890	−0.0250919	0.0148714
17	1.2722451	−0.1405266	0.0500314	−0.0251305
18		1.2722929	−0.1405696	0.0500707
19			1.2723365	−0.1406089
20				1.2723761

	$N = 44$ $D = 0.0002185$	$N = 46$ $D = 0.0001810$	$N = 48$ $D = 0.0001510$	$N = 50$ $D = 0.0001260$
0	−0.0001665	0.0001382	−0.0001153	0.0000961
1	0.0002884	−0.0002394	0.0002001	−0.0001671
2	−0.0002673	0.0002224	−0.0001860	0.0001561
3	0.0003292	−0.0002736	0.0002287	−0.0001920
4	−0.0004153	0.0003449	−0.0002881	0.0002416
5	0.0005243	−0.0004342	0.0003619	−0.0003032
6	−0.0006607	0.0005454	−0.0004530	0.0003786
7	0.0008310	−0.0006830	0.0005652	−0.0004709
8	−0.0010458	0.0008545	−0.0007040	0.0005843
9	0.0013195	−0.0010707	0.0008768	−0.0007241
10	−0.0016732	0.0013455	−0.0010939	0.0008979
11	0.0021385	−0.0017002	0.0013697	−0.0011156
12	−0.0027646	0.0021664	−0.0017254	0.0013924
13	0.0036307	−0.0027935	0.0021925	−0.0017489
14	−0.0048726	0.0036606	−0.0028205	0.0022170
15	0.0067378	−0.0049031	0.0036879	−0.0028453
16	−0.0097176	0.0067686	−0.0049314	0.0037137
17	0.0149065	−0.0097490	0.0067975	−0.0049574
18	−0.0251657	0.0149386	−0.0097785	0.0068242
19	0.0501065	−0.0251984	0.0149684	−0.0098059
20	−0.1406450	0.0501395	−0.0252289	0.0149961
21	1.2724119	−0.1406780	0.0501700	−0.0252568
22		1.2724452	−0.1407088	0.0501979
23			1.2724760	−0.1407371
24				1.2725043

Table IV—Wideband differentiators ($F_p = 0.45$, N even)

	$N = 4$ $D = 0.0507220$	$N = 6$ $D = 0.0209865$	$N = 8$ $D = 0.0102350$	$N = 10$ $D = 0.0054640$
0	−0.1049330	0.0313204	−0.0132691	0.0066049
1	1.2640770	−0.1318473	0.0425774	−0.0188979
2		1.2599360	−0.1307889	0.0423179
3			1.2621213	−0.1324320
4				1.2640211

	$N = 12$ $D = 0.0030870$	$N = 14$ $D = 0.0018135$	$N = 16$ $D = 0.0010970$	$N = 18$ $D = 0.0006780$
0	−0.0036000	0.0020785	−0.0012494	0.0007731
1	0.0097311	−0.0054557	0.0032324	−0.0019893
2	−0.0189385	0.0098781	−0.0056282	0.0033951
3	0.0434812	−0.0197685	0.0104785	−0.0060664
4	−0.1337806	0.0444447	−0.0204678	0.0109918
5	1.2654301	−0.1348054	0.0452000	−0.0210311
6		1.2664828	−0.1355959	0.0458000
7			1.2672905	−0.1362207
8				1.2679283

	$N = 20$ $D = 0.0004255$	$N = 22$ $D = 0.0002715$	$N = 24$ $D = 0.0001750$	$N = 26$ $D = 0.0001140$
0	−0.0004882	0.0003148	−0.0002055	0.0001357
1	0.0012579	−0.0008143	0.0005353	−0.0003569
2	−0.0021291	0.0013757	−0.0009079	0.0006095
3	0.0037159	−0.0023684	0.0015526	−0.0010402
4	−0.0064459	0.0040014	−0.0025818	0.0017137
5	0.0114147	−0.0067686	0.0042459	−0.0027693
6	−0.0214891	0.0117693	−0.0070416	0.0044582
7	0.0462857	−0.0218711	0.0120681	−0.0072772
8	−0.1367253	0.0466885	−0.0221913	0.0123242
9	1.2684422	−0.1371425	0.0470259	−0.0224649
10		1.2688667	−0.1374915	0.0473133
11			1.2692216	−0.1377887
12				1.2695236

	$N = 28$ $D = 0.0000750$	$N = 30$ $D = 0.0000495$	$N = 32$ $D = 0.0000330$	$N = 34$ $D = 0.0000220$
0	−0.0000905	0.0000609	−0.0000412	0.0000280
1	0.0002406	−0.0001637	0.0001122	−0.0000776
2	−0.0004147	0.0002853	−0.0001979	0.0001385
3	0.0007084	−0.0004888	0.0003409	−0.0002400
4	−0.0011618	0.0008008	−0.0005589	0.0003943
5	0.0018573	−0.0012723	0.0008856	−0.0006245
6	−0.0029342	0.0019861	−0.0013722	0.0009638
7	0.0046439	−0.0030806	0.0021014	−0.0014640
8	−0.0074817	0.0048076	−0.0032113	0.0022063
9	0.0125460	−0.0076617	0.0049527	−0.0033295
10	−0.0227011	0.0127401	−0.0078204	0.0050837
11	0.0475609	−0.0229076	0.0129110	−0.0079627
12	−0.1380441	0.0477767	−0.0230885	0.0130637
13	1.2697831	−0.1382662	0.0479655	−0.0232500
14		1.2700086	−0.1384607	0.0481336
15			1.2702059	−0.1386335
16				1.2703809

Table IV—continued

	$N = 36$ $D = 0.0000145$	$N = 38$ $D = 0.0000100$	$N = 40$ $D = 0.0000065$	$N = 42$ $D = 0.0000045$
0	−0.0000192	0.0000132	−0.0000091	0.0000063
1	0.0000537	−0.0000377	0.0000264	−0.0000185
2	−0.0000974	0.0000691	−0.0000490	0.0000352
3	0.0001700	−0.0001213	0.0000870	−0.0000628
4	−0.0002805	0.0002014	−0.0001451	0.0001052
5	0.0004448	−0.0003198	0.0002312	−0.0001684
6	−0.0006855	0.0004929	−0.0003569	0.0002608
7	0.0010361	−0.0007430	0.0005378	−0.0003927
8	−0.0015475	0.0011033	−0.0007961	0.0005809
9	0.0023015	−0.0016248	0.0011652	−0.0008467
10	−0.0034360	0.0023892	−0.0016958	0.0012237
11	0.0052009	−0.0035337	0.0024690	−0.0017621
12	−0.0080899	0.0053080	−0.0036223	0.0025431
13	0.0131997	−0.0082058	0.0054048	−0.0037043
14	−0.0233932	0.0133232	−0.0083101	0.0054940
15	0.0482831	−0.0235233	0.0134344	−0.0084063
16	−0.1387868	0.0484182	−0.0236399	0.0135362
17	1.2705364	−0.1389253	0.0485392	−0.0237470
18		1.2706765	−0.1390494	0.0486498
19			1.2708019	−0.1391628
20				1.2709168

	$N = 44$ $D = 0.0000030$	$N = 46$ $D = 0.0000020$	$N = 48$ $D = 0.0000015$	$N = 50$ $D = 0.0000010$
0	−0.0000044	0.0000031	−0.0000022	0.0000016
1	0.0000132	−0.0000094	0.0000066	−0.0000047
2	−0.0000251	0.0000182	−0.0000132	0.0000094
3	0.0000452	−0.0000330	0.0000239	−0.0000176
4	−0.0000767	0.0000559	−0.0000412	0.0000302
5	0.0001232	−0.0000905	0.0000666	−0.0000493
6	−0.0001910	0.0001411	−0.0001043	0.0000776
7	0.0002887	−0.0002133	0.0001587	−0.0001181
8	−0.0004266	0.0003157	−0.0002350	0.0001756
9	0.0006211	−0.0004593	0.0003421	−0.0002560
10	−0.0008938	0.0006594	−0.0004904	0.0003673
11	0.0012777	−0.0009381	0.0006959	−0.0005202
12	−0.0018231	0.0013283	−0.0009802	0.0007304
13	0.0026113	−0.0018802	0.0013760	−0.0010198
14	−0.0037787	0.0026748	−0.0019337	0.0014209
15	0.0055751	−0.0038485	0.0027338	−0.0019836
16	−0.0084933	0.0056505	−0.0039129	0.0027888
17	0.0136282	−0.0085737	0.0057202	−0.0039732
18	−0.0238434	0.0137137	−0.0086482	0.0057849
19	0.0487497	−0.0239323	0.0137919	−0.0087170
20	−0.1392649	0.0488417	−0.0240143	0.0138645
21	1.2710199	−0.1393589	0.0489262	−0.0240897
22		1.2711151	−0.1394452	0.0490041
23			1.2712024	−0.1395247
24				1.2712825

Table V—Wideband differentiators
($F_p = 0.40$, N even)

	$N = 4$ $D = 0.0296405$	$N = 6$ $D = 0.0089245$	$N = 8$ $D = 0.0031370$	$N = 10$ $D = 0.0012040$
0	−0.0861745	0.0205935	−0.0068427	0.0026524
1	1.2288833	−0.1115391	0.0300431	−0.0109331
2		1.2405798	−0.1179778	0.0335962
3			1.2484761	−0.1225086
4				1.2534103

	$N = 12$ $D = 0.0004890$	$N = 14$ $D = 0.0002065$	$N = 16$ $D = 0.0000900$	$N = 18$ $D = 0.0000400$
0	−0.0011231	0.0005036	−0.0002350	0.0001128
1	0.0045751	−0.0020735	0.0009893	−0.0004888
2	−0.0129299	0.0057108	−0.0027250	0.0013644
3	0.0362512	−0.0145085	0.0066583	−0.0032958
4	−0.1255952	0.0381971	−0.0157403	0.0074380
5	1.2567278	−0.1278179	0.0396824	−0.0167274
6		1.2590998	−0.1294942	0.0408526
7			1.2608785	−0.1308018
8				1.2622592

	$N = 20$ $D = 0.0000180$	$N = 22$ $D = 0.0000085$	$N = 24$ $D = 0.0000040$	$N = 26$ $D = 0.0000020$
0	−0.0000553	0.0000276	−0.0000141	0.0000072
1	0.0002476	−0.0001279	0.0000672	−0.0000358
2	−0.0007056	0.0003738	−0.0002014	0.0001100
3	0.0017093	−0.0009151	0.0005008	−0.0002787
4	−0.0037888	0.0020219	−0.0011121	0.0006255
5	0.0080905	−0.0042201	0.0023044	−0.0012978
6	−0.0175364	0.0086475	−0.0045990	0.0025623
7	0.0417986	−0.0182150	0.0091270	−0.0049367
8	−0.1318492	0.0425824	−0.0187905	0.0095470
9	1.2633610	−0.1327106	0.0432403	−0.0192881
10		1.2642636	−0.1334294	0.0438045
11			1.2650144	−0.1340417
12				1.2656525

	$N = 28$ $D = 0.0000010$	$N = 30$ $D = 0.0000005$
0	−0.0000038	0.0000019
1	0.0000192	−0.0000104
2	−0.0000606	0.0000336
3	0.0001568	−0.0000889
4	−0.0003569	0.0002058
5	0.0007458	−0.0004345
6	−0.0014706	0.0008608
7	0.0027963	−0.0016311
8	−0.0052373	0.0030087
9	0.0099149	−0.0055056
10	−0.0197195	0.0102394
11	0.0442895	−0.0200961
12	−0.1345657	0.0447099
13	1.2661969	−0.1350178
14		1.2666657

Table VI—Wideband differentiators
(F_p = 0.48, N odd)

	N = 31 D = 0.0904880	N = 33 D = 0.0775765	N = 35 D = 0.0669570	N = 37 D = 0.0574865
0	0.0608611	−0.0522909	0.0452050	−0.0389237
1	−0.0899683	0.0775166	−0.0671563	0.0580516
2	0.0575832	−0.0499780	0.0435425	−0.0380356
3	−0.0599504	0.0521363	−0.0454862	0.0398662
4	0.0676325	−0.0587798	0.0512472	−0.0449414
5	−0.0779065	0.0675169	−0.0587481	0.0514649
6	0.0906186	−0.0781443	0.0677554	−0.0591983
7	−0.1063470	0.0910075	−0.0784716	0.0682828
8	0.1262399	−0.1068305	0.0913845	−0.0790428
9	−0.1522623	0.1267944	−0.1072511	0.0919846
10	0.1878138	−0.1527018	0.1270991	−0.1077415
11	−0.2401009	0.1881949	−0.1529673	0.1275383
12	0.3258611	−0.2405147	0.1884984	−0.1534175
13	−0.4949906	0.3261771	−0.2407258	0.1888647
14	0.9974893	−0.4951904	0.3263414	−0.2410280
15	0.	0.9975854	−0.4953164	0.3265623
16		0.	0.9976379	−0.4954722
17			0.	0.9977334
18				0.

	N = 39 D = 0.0496845	N = 41 D = 0.0427980	N = 43 D = 0.0370790	N = 45 D = 0.0319585
0	0.0337241	−0.0291326	0.0252974	−0.0218781
1	−0.0504549	0.0437514	−0.0381107	0.0331074
2	0.0333314	−0.0292008	0.0256448	−0.0225406
3	−0.0350300	0.0308071	−0.0271405	0.0239650
4	0.0395115	−0.0347953	0.0307025	−0.0271669
5	−0.0452125	0.0398250	−0.0351626	0.0311429
6	0.0519079	−0.0456920	0.0403302	−0.0357293
7	−0.0596859	0.0524473	−0.0462408	0.0409394
8	0.0687827	−0.0602598	0.0530210	−0.0468694
9	−0.0795659	0.0694016	−0.0608706	0.0536820
10	0.0924131	−0.0801166	0.0699494	−0.0614800
11	−0.1081069	0.0929299	−0.0806174	0.0705137
12	0.1279109	−0.1086479	0.0934269	−0.0811778
13	−0.1537285	0.1283931	−0.1090871	0.0939352
14	0.1891258	−0.1541598	0.1287811	−0.1095404
15	−0.2412344	0.1894996	−0.1544809	0.1291782
16	0.3267197	−0.2415476	0.1897616	−0.1548211
17	−0.4955762	0.3269657	−0.2417578	0.1900419
18	0.9977978	−0.4957493	0.3271268	−0.2419859
19	0.	0.9978660	−0.4958426	0.3272880
20		0.	0.9979125	−0.4959485
21			0.	0.9979703
22				0.

Table VI—continued

	$N = 47$ $D = 0.0277365$	$N = 49$ $D = 0.0239425$
0	0.0190371	−0.0164918
1	−0.0289080	0.0251594
2	0.0198593	−0.0174902
3	−0.0211875	0.0187459
4	0.0240589	−0.0213377
5	−0.0276055	0.0245186
6	0.0316820	−0.0281603
7	−0.0362942	0.0322648
8	0.0415174	−0.0368924
9	−0.0474776	0.0421467
10	0.0542475	−0.0480777
11	−0.0620081	0.0548148
12	0.0710446	−0.0625739
13	−0.0816707	0.0715793
14	0.0943907	−0.0821746
15	−0.1099542	0.0948557
16	0.1295429	−0.1103751
17	−0.1551425	0.1299234
18	0.1903205	−0.1554828
19	−0.2421995	0.1906061
20	0.3274491	−0.2424311
21	−0.4960631	0.3276270
22	0.9980259	−0.4961835
23	0.	0.9980887
24		0.

Table VII—Wideband differentiators ($F_p = 0.45$, N odd)

	$N = 15$ $D = 0.0735575$	$N = 17$ $D = 0.0507440$	$N = 19$ $D = 0.0351550$	$N = 21$ $D = 0.0245205$
0	0.0621149	−0.0435352	0.0307251	−0.0218457
1	−0.1208630	0.0858820	−0.0616198	0.0445927
2	0.1395483	−0.1005756	0.0736612	−0.0545490
3	−0.1970413	0.1402768	−0.1025906	0.0762772
4	0.2916922	−0.1997274	0.1436066	−0.1061155
5	−0.4713966	0.2942249	−0.2029221	0.1470256
6	0.9854400	−0.4732577	0.2968538	−0.2059085
7	0.	0.9864199	−0.4751178	0.2992430
8		0.	0.9873796	−0.4767810
9			0.	0.9882420
10				0.

	$N = 23$ $D = 0.0171585$	$N = 25$ $D = 0.0120320$	$N = 27$ $D = 0.0084550$	$N = 29$ $D = 0.0059730$
0	0.0155952	−0.0111731	0.0080321	−0.0058000
1	−0.0324316	0.0236892	−0.0173702	0.0127850
2	0.0406871	−0.0305049	0.0229594	−0.0173278
3	−0.0573479	0.0434319	−0.0330574	0.0252509
4	0.0797751	−0.0606437	0.0464403	−0.0357494
5	−0.1095875	0.0831228	−0.0637571	0.0492781
6	0.1502102	−0.1127769	0.0861795	−0.0666162
7	−0.2086526	0.1530955	−0.1156518	0.0889542
8	0.3014107	−0.2111147	0.1556885	−0.1182549
9	−0.4782695	0.3033453	−0.2133255	0.1580256
10	0.9889982	−0.4796075	0.3050898	−0.2153078
11	0.	0.9896812	−0.4808120	0.3066433
12		0.	0.9902957	−0.4818795
13			0.	0.9908395
14				0.

	$N = 31$ $D = 0.0042255$	$N = 33$ $D = 0.0029835$	$N = 35$ $D = 0.0021205$	$N = 37$ $D = 0.0015065$
0	0.0041972	−0.0030316	0.0022041	−0.0016019
1	−0.0094342	0.0069542	−0.0051557	0.0038195
2	0.0131118	−0.0099202	0.0075348	−0.0057161
3	−0.0193447	0.0148315	−0.0114071	0.0087679
4	0.0276221	−0.0213785	0.0165958	−0.0128862
5	−0.0382866	0.0298388	−0.0233329	0.0182658
6	0.0518947	−0.0406255	0.0319384	−0.0251623
7	−0.0692328	0.0542892	−0.0428193	0.0338874
8	0.0914895	−0.0716110	0.0565198	−0.0448418
9	−0.1206189	0.0937794	−0.0738174	0.0585671
10	0.1601348	−0.1227515	0.0958981	−0.0758355
11	−0.2170932	0.1620418	−0.1247175	0.0978301
12	0.3080442	−0.2187083	0.1637913	−0.1265044
13	−0.4828449	0.3093083	−0.2201835	0.1653775
14	0.9913327	−0.4837123	0.3104607	−0.2215197
15	0.	0.9917735	−0.4845030	0.3115043
16		0.	0.9921756	−0.4852199
17			0.	0.9925410
18				0.

Table VII—continued

	$N = 39$ $D = 0.0010700$	$N = 41$ $D = 0.0007635$	$N = 43$ $D = 0.0005430$	$N = 45$ $D = 0.0003895$
0	0.0011652	−0.0008504	0.0006202	−0.0004546
1	−0.0028328	0.0021064	−0.0015651	0.0011677
2	0.0043404	−0.0033002	0.0025070	−0.0019091
3	−0.0067453	0.0051953	−0.0039977	0.0030822
4	0.0100145	−0.0077927	0.0060586	−0.0047187
5	−0.0143150	0.0112340	−0.0088112	0.0069219
6	0.0198555	−0.0156929	0.0124011	−0.0098156
7	−0.0268851	0.0213726	−0.0169985	0.0135415
8	0.0357108	−0.0285178	0.0228014	−0.0182649
9	−0.0467274	0.0374289	−0.0300468	0.0241771
10	0.0604684	−0.0484943	0.0390296	−0.0315102
11	−0.0777016	0.0622422	−0.0501326	0.0405542
12	0.0996092	−0.0794383	0.0638811	−0.0516874
13	−0.1281462	0.1012620	−0.0810377	0.0654312
14	0.1668343	−0.1296683	0.1027800	−0.0825450
15	−0.2227461	0.1681805	−0.1310632	0.1042063
16	0.3124603	−0.2238759	0.1694129	−0.1323710
17	−0.4858737	0.3133399	−0.2249101	0.1705665
18	0.9928727	−0.4864766	0.3141448	−0.2258767
19	0.	0.9931793	−0.4870273	0.3148957
20		0.	0.9934589	−0.4875406
21			0.	0.9937194
22				0.

	$N = 47$ $D = 0.0002770$	$N = 49$ $D = 0.0001985$
0	0.0003314	−0.0002432
1	−0.0008674	0.0006481
2	0.0014489	−0.0011040
3	−0.0023700	0.0018278
4	0.0036666	−0.0028563
5	−0.0054274	0.0042647
6	0.0077575	−0.0061421
7	−0.0107775	0.0085916
8	0.0146251	−0.0117291
9	−0.0194593	0.0156885
10	0.0254670	−0.0206239
11	−0.0328758	0.0267183
12	0.0419704	−0.0341940
13	−0.0531259	0.0433323
14	0.0668603	−0.0545041
15	−0.0839314	0.0682257
16	0.1055160	−0.0852521
17	−0.1335695	0.1067601
18	0.1716214	−0.1347058
19	−0.2267589	0.1726205
20	0.3155805	−0.2275936
21	−0.4880090	0.3162280
22	0.9939572	−0.4884507
23	0.	0.9941812
24		0.

Table VIII—Wideband differentiators
(F_p = 0.40, N odd)

	$N = 9$ $D = 0.0609450$	$N = 11$ $D = 0.0292625$	$N = 13$ $D = 0.0142550$	$N = 15$ $D = 0.0070280$
0	−0.0756455	0.0391813	−0.0207637	0.0111819
1	0.2086718	−0.1121159	0.0622962	−0.0352879
2	−0.4002986	0.2123949	−0.1213132	0.0715702
3	0.9466429	−0.4116991	0.2254262	−0.1332217
4	0.	0.9534363	−0.4221135	0.2364777
5		0.	0.9591518	−0.4305757
6			0.	0.9637379
7				0.

	$N = 17$ $D = 0.0034905$	$N = 19$ $D = 0.0017425$	$N = 21$ $D = 0.0008780$	$N = 23$ $D = 0.0004415$
0	−0.0060893	0.0033408	−0.0018491	0.0010251
1	0.0202278	−0.0116764	0.0067871	−0.0039493
2	−0.0429142	0.0259414	−0.0157683	0.0095916
3	0.0813764	−0.0504920	0.0315941	−0.0198219
4	−0.1433229	0.0898511	−0.0572637	0.0367532
5	0.2455792	−0.1518244	0.0972194	−0.0632651
6	−0.4374621	0.2531345	−0.1590934	0.1036232
7	0.9674450	−0.4431314	0.2595207	−0.1653232
8	0.	0.9704860	−0.4478868	0.2649402
9		0.	0.9730241	−0.4518964
10			0.	0.9751570
11				0.

	$N = 25$ $D = 0.0002235$	$N = 27$ $D = 0.0001140$	$N = 29$ $D = 0.0000580$	$N = 31$ $D = 0.0000295$
0	−0.0005721	0.0003211	−0.0001806	0.0001015
1	0.0023072	−0.0013534	0.0007945	−0.0004659
2	−0.0058456	0.0035688	−0.0021774	0.0013261
3	0.0124605	−0.0078430	0.0049320	−0.0030942
4	−0.0236760	0.0152804	−0.0098561	0.0063438
5	0.0414998	−0.0273300	0.0180098	−0.0118517
6	−0.0686617	0.0458852	−0.0307650	0.0206299
7	0.1092853	−0.0735522	0.0499183	−0.0339823
8	−0.1707635	0.1143420	−0.0779740	0.0536245
9	0.2696310	−0.1755688	0.1188552	−0.0819780
10	−0.4553459	0.2737411	−0.1798160	0.1228947
11	0.9769854	−0.4583518	0.2773477	−0.1835837
12	0.	0.9785734	−0.4609763	0.2805263
13		0.	0.9799561	−0.4632794
14			0.	0.9811659
15				0.

Table VIII—continued

	$N = 33$ $D = 0.0000150$	$N = 35$ $D = 0.0000075$	$N = 37$ $D = 0.0000040$	$N = 39$ $D = 0.0000020$
0	−0.0000572	0.0000324	−0.0000182	0.0000104
1	0.0002730	−0.0001602	0.0000946	−0.0000556
2	−0.0008061	0.0004901	−0.0002985	0.0001813
3	0.0019374	−0.0012117	0.0007581	−0.0004734
4	−0.0040743	0.0026122	−0.0016742	0.0010703
5	0.0077839	−0.0051026	0.0033417	−0.0021825
6	−0.0138180	0.0092391	−0.0061707	0.0041092
7	0.0231441	−0.0157479	0.0107053	−0.0072577
8	−0.0370061	0.0255562	−0.0176463	0.0121583
9	0.0570510	−0.0398558	0.0278845	−0.0194879
10	−0.0856320	0.0602338	−0.0425629	0.0301031
11	0.1265440	−0.0889869	0.0632195	−0.0451067
12	−0.1869606	0.1298634	−0.0921018	0.0659923
13	0.2833588	−0.1900104	0.1329205	−0.0949672
14	−0.4653230	0.2859028	−0.1928014	0.1357111
15	0.9822372	−0.4671520	0.2882201	−0.1953339
16	0.	0.9831935	−0.4688120	0.2903130
17		0.	0.9840602	−0.4703068
18			0.	0.9848390
19				0.

	$N = 41$ $D = 0.0000010$	$N = 43$ $D = 0.0000005$	$N = 45$ $D = 0.0000005$
0	−0.0000060	0.0000035	−0.0000019
1	0.0000330	−0.0000195	0.0000113
2	−0.0001106	0.0000672	−0.0000408
3	0.0002959	−0.0001844	0.0001147
4	−0.0006846	0.0004364	−0.0002774
5	0.0014247	−0.0009274	0.0006013
6	−0.0027341	0.0018140	−0.0011982
7	0.0049144	−0.0033182	0.0022305
8	−0.0083667	0.0057416	−0.0039235
9	0.0136084	−0.0094801	0.0065776
10	−0.0212934	0.0150360	−0.0105809
11	0.0322462	−0.0230414	0.0164220
12	−0.0475332	0.0342936	−0.0247140
13	0.0686105	−0.0498257	0.0362270
14	−0.0976501	0.0710606	−0.0519676
15	0.1383067	−0.1001420	0.0733295
16	−0.1976766	0.1407019	−0.1024316
17	0.2922413	−0.1998276	0.1428897
18	−0.4716800	0.2940053	−0.2017829
19	0.9855531	−0.4729332	0.2956031
20	0.	0.9862041	−0.4740651
21		0.	0.9867909
22			0.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

On the Behavior of Minimax FIR Digital Hilbert Transformers

By L. R. RABINER and R. W. SCHAFER

(Manuscript received July 9, 1973)

Optimum (in a minimax sense) linear phase FIR Hilbert transformers can be designed efficiently using a Remez optimization procedure. This paper presents useful design data on wideband Hilbert transformers with even and odd values of N, the impulse response duration (in samples) of the filter. Based on these data, the following observations can be made:

(i) In the case of equal lower and upper transition regions, Hilbert transformers with odd values of N can be realized more efficiently than those with even values of N, assuming the same peak errors of approximation for both cases. This is because every other impulse response sample is exactly zero for odd values of N.

(ii) The peak approximation error for Hilbert transformers with odd values of N is determined primarily by the minimum of the values of the lower and upper transition widths.

(iii) The peak approximation error for Hilbert transformers with even values of N is determined primarily by the lower transition width of the filter.

(iv) The smaller the bandwidth of the Hilbert transformer, the faster the decrease of peak error of approximation with decreasing bandwidth of the Hilbert transformer.

(v) The larger the value of N, the faster the decrease of peak approximation error with decreasing bandwidth of the Hilbert transformer.

These observations lead to the conclusion that the bandwidth of the Hilbert transformer should be made as small as possible, and that odd values of N should be used, whenever possible, for efficient direct form realizations. A set of tables of values of the impulse response coefficients is included for several different bandwidth Hilbert transformers and for both even and odd values of N.

I. INTRODUCTION

Hilbert transformers have been used in a variety of applications including radar systems, speech processing systems, modulation systems, and schemes for efficient sampling of a real bandpass signal.[1,2] Although the theory of discrete Hilbert transformer systems is well understood, it is only recently that design techniques have become available for obtaining optimum (in a minimax or Chebyshev sense) finite-duration impulse response (FIR) approximations to the ideal Hilbert transformer.[3,4] It is the purpose of this paper to present new design data on minimax FIR Hilbert transformers to aid in selecting the most efficient network to meet given design specifications. Herrmann[3] has already given a small table of data on FIR Hilbert transformers; however, the generality and widespread applicability of the data presented here justify this more complete elaboration of the properties of the minimax Hilbert transformer approximations.

II. DISCRETE-TIME HILBERT TRANSFORMERS

Most applications of Hilbert transformers involve the representation of a real bandpass signal in terms of a complex signal. For continuous-time signals, such complex signals are analytic functions of time and thus are called *analytic signals*. Although the concept of analyticity is meaningless for discrete-time signals, complex representation of real, discrete-time, bandpass signals can be used in a similar manner to analytic signals. Therefore, consider a real sequence $x(n)$ with Fourier transform $X(e^{j\omega})$. Corresponding to this sequence, we can construct the complex sequence

$$\tilde{x}(n) = x(n) + j\hat{x}(n) \tag{1}$$

whose Fourier transform, $\tilde{X}(e^{j\omega})$, is

$$\begin{aligned} \tilde{X}(e^{j\omega}) &= 2X(e^{j\omega}) \qquad 0 \leqq \omega < \pi \\ &= 0 \qquad\qquad\quad \pi \leqq \omega < 2\pi. \end{aligned} \tag{2}$$

From eq. (1), the Fourier transform of $\tilde{x}(n)$ is

$$\tilde{X}(e^{j\omega}) = X(e^{j\omega}) + j\hat{X}(e^{j\omega})$$

and from eq. (2) it follows that

$$X(e^{j\omega}) + j\hat{X}(e^{j\omega}) = 0 \qquad \pi \leqq \omega < 2\pi$$

and

$$\tilde{X}(e^{j\omega}) = 2X(e^{j\omega}) \qquad 0 \leqq \omega < \pi.$$

These relationships are satisfied if

$$\hat{X}(e^{j\omega}) = H_d(e^{j\omega})X(e^{j\omega}), \tag{3}$$

where

$$\begin{aligned} H_d(e^{j\omega}) &= -j \qquad 0 < \omega < \pi \\ &= +j \qquad \pi < \omega < 2\pi. \end{aligned} \tag{4}$$

Thus $\hat{x}(n)$ can be obtained by linear filtering the sequence $x(n)$ with a system having a frequency response, $H_d(e^{j\omega})$, as in eq. (4). Such a system is called an ideal Hilbert transformer and $\hat{x}(n)$ is the Hilbert transform of $x(n)$. If one allows the addition of a linear phase term in the definition of the frequency response of the "ideal" Hilbert transformer, eq. (4) becomes

$$\begin{aligned} H_d(e^{j\omega}) &= -je^{-j\omega\tau} \qquad 0 < \omega < \pi \\ &= +je^{-j(\omega-2\pi)\tau} \qquad \pi < \omega < 2\pi, \end{aligned} \tag{5}$$

where τ is the delay in samples. The impulse response of the ideal Hilbert transformer of eq. (5) can be shown to be

$$h_d(n) = \frac{2}{\pi} \frac{\sin^2\left[\frac{\pi}{2}(n-\tau)\right]}{(n-\tau)}. \tag{6}$$

For $\tau = 0$, eq. (6) gives

$$\begin{aligned} h_d(n) &= \frac{2}{\pi} \frac{\sin^2\left[\frac{\pi n}{2}\right]}{n} \qquad n \neq 0 \\ &= 0 \qquad\qquad\qquad\qquad n = 0, \end{aligned} \tag{7}$$

whereas for $\tau = -\frac{1}{2}$ (i.e., a half-sample advance) eq. (6) gives

$$h_d(n) = \frac{1}{\pi(n+\frac{1}{2})}. \tag{8}$$

The frequency response [eq. (5)] corresponding to $\tau = 0$ has discontinuities at both $\omega = 0$ and $\omega = \pi$, whereas the frequency response corresponding to $\tau = -\frac{1}{2}$ only has a discontinuity at $\omega = 0$. The impulse response of eq. (7), corresponding to $\tau = 0$, has odd symmetry, is noncausal, and is of infinite duration. In addition, all even-numbered samples of the impulse response are exactly zero—i.e., $h_d(2n) = 0$, $n = 0, \pm 1, \pm 2, \cdots$. The impulse response of eq. (8), corresponding to $\tau = -\frac{1}{2}$, is also noncausal, is of infinite duration, and obeys the symmetry condition

$$h(n) = -h(-n-1) \qquad n = 0, 1, \cdots. \tag{9}$$

This impulse response does *not* have the property that even-numbered (or odd-numbered) impulse response samples are zero.

We have only considered $\tau = 0$ and $\tau = -\frac{1}{2}$ for determining the impulse response of the ideal Hilbert transformer. All other values of $(-1 < \tau \leqq 0)$ can be shown to yield impulse responses without desirable symmetry properties and hence are not suitable candidates for approximation or implementation.

To obtain a causal approximation to the ideal Hilbert transformers with no phase distortion (except for a delay), it can be shown that an FIR approximation is required.[5] Since there are a number of subtle distinctions between FIR systems of even-length and of odd-length, the next section deals with some of the general properties of FIR Hilbert transformer approximations.

III. PROPERTIES OF FIR HILBERT TRANSFORMERS

Consider a causal FIR system with impulse response $h(n)$, $0 \leqq n \leqq N - 1$, and frequency response

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)e^{-j\omega n}. \tag{10}$$

Since the desired frequency response $H_d(e^{j\omega})$ is purely imaginary and $h(n)$ is real, a linear phase approximation to the ideal frequency response is obtained only when $h(n)$ satisfies the symmetry condition

$$h(n) = -h(N - 1 - n) \qquad n = 0, 1, \cdots, N - 1. \tag{11}$$

For N an odd integer, this means that $h(n)$ has odd symmetry about the sample at $n = (N - 1)/2$. (This case corresponds to the $\tau = 0$ case above with an additional delay of $(N - 1)/2$ samples.) For N even, $h(n)$ has odd symmetry about a point halfway between the samples at $n = N/2$ and $n = (N/2) + 1$. (This case corresponds to the $\tau = -\frac{1}{2}$ case above with an additional delay of $N/2$ samples.) This implies that the frequency response of a filter satisfying eq. (11) can be expressed as

$$H(e^{j\omega}) = e^{-j\omega(N-1)/2}[jH^*(e^{j\omega})], \tag{12}$$

where $H^*(e^{j\omega})$ is a purely real function of ω. In particular, for N odd, $H^*(e^{j\omega})$ is

$$H^*(e^{j\omega}) = \sum_{n=1}^{(N-1)/2} a(n) \sin(\omega n), \tag{13a}$$

where

$$a(n) = 2h\left(\frac{N-1}{2} - n\right) \qquad n = 1, 2, \cdots, \left(\frac{N-1}{2}\right). \tag{13b}$$

Also, from eq. (11),

$$h\left(\frac{N-1}{2}\right) = 0. \tag{13c}$$

For N even, the expression for $H^*(e^{j\omega})$ is

$$H^*(e^{j\omega}) = \sum_{n=1}^{N/2} b(n) \sin\left[\omega(n - \tfrac{1}{2})\right], \tag{14a}$$

where

$$b(n) = 2h\left(\frac{N}{2} - n\right) \qquad n = 1, 2, \cdots, N/2. \tag{14b}$$

The factor $e^{-j\omega(N-1)/2}$ in eq. (12) represents a delay of $(N-1)/2$ samples. When N is odd, this delay is an integer number of sample intervals, but when N is even, it is an integer plus one-half of a sample interval. In approximating the Hilbert transformer, the coefficients $a(n)$ and $b(n)$ must be chosen so that $jH^*(e^{j\omega})$ approximates the ideal frequency response of eq. (5). In many cases it is not required, and often it is not desirable, that the approximation be carried out over the entire band $0 \leqq \omega \leqq 2\pi$. Thus, in general, the real function $H^*(e^{j\omega})$ must approximate

$$\begin{aligned} D(e^{j\omega}) &= -1 \qquad 2\pi F_L \leqq \omega \leqq 2\pi F_H \\ &= +1 \qquad 2\pi(1 - F_H) \leqq \omega \leqq 2\pi F_L, \end{aligned} \tag{15}$$

where F_L and F_H define the lower and upper cutoff frequencies respectively. Alternatively, F_L and $0.5 - F_H$ define the lower and upper transition bandwidths respectively. It can be seen from eqs. (13a) and (14a) that $H^*(e^{j\omega})$ is constrained to be zero at $\omega = 0$, and π when N is odd, and when N is even, $H^*(e^{j\omega})$ is constrained to be zero at $\omega = 0$. Thus, fullband approximations are impossible since F_L cannot be zero, and F_H can be 0.5 only when N is even. In the transition regions the desired frequency response is left undefined; hence, in these regions, $H^*(e^{j\omega})$ is unconstrained except at the end points. In the next section it will be shown by examples that leaving these regions unconstrained can lead to unsatisfactory FIR approximations.

The impulse response of the ideal Hilbert transformer (with $\tau = 0$) has the property

$$h_d(n) = 0 \qquad n = 0, \pm 2, \pm 4, \cdots.$$

This can be shown to be a direct result of the fact that $h_d(n)$ is real

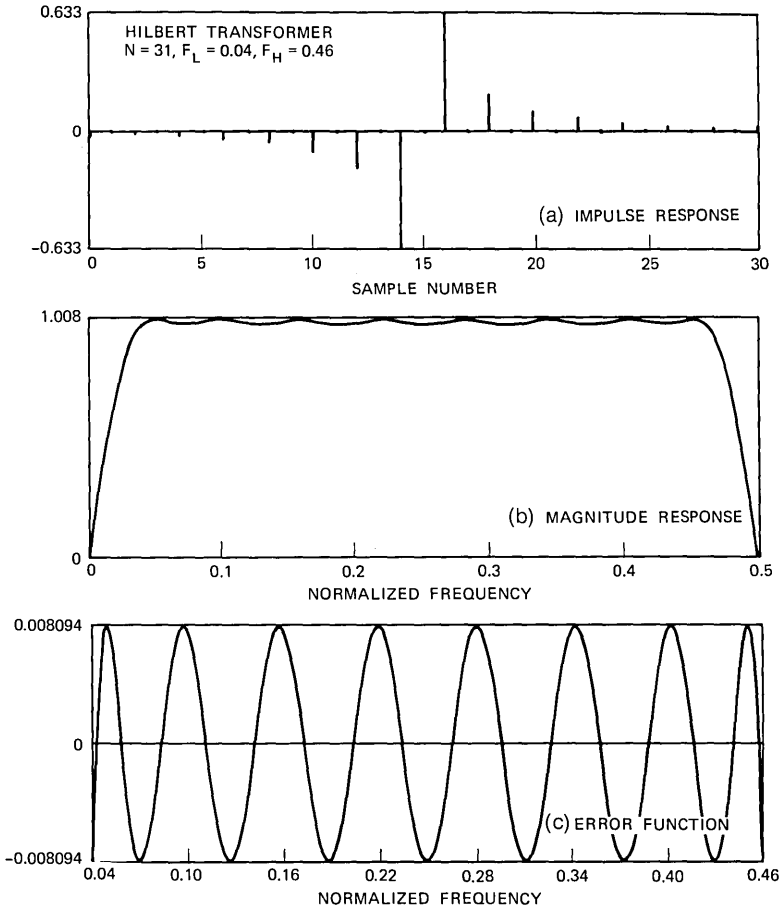

Fig. 1—The impulse response, magnitude response, and error function of an $N = 31$ Hilbert transformer with $F_L = 0.04$ and $F_H = 0.46$.

and the frequency response is an imaginary, odd, periodic function and

$$H_d(e^{j\omega}) = H_d(e^{j(\pi-\omega)}). \tag{16}$$

Similar properties can be achieved for the FIR Hilbert transformers if N is odd and $F_L = 0.5 - F_H$. To see this, assume that

$$H^*(e^{j\omega}) = H^*(e^{j(\pi-\omega)}). \tag{17}$$

Clearly this requires that $F_L = 0.5 - F_H$. Substituting eq. (13a) into eq. (17), it follows that

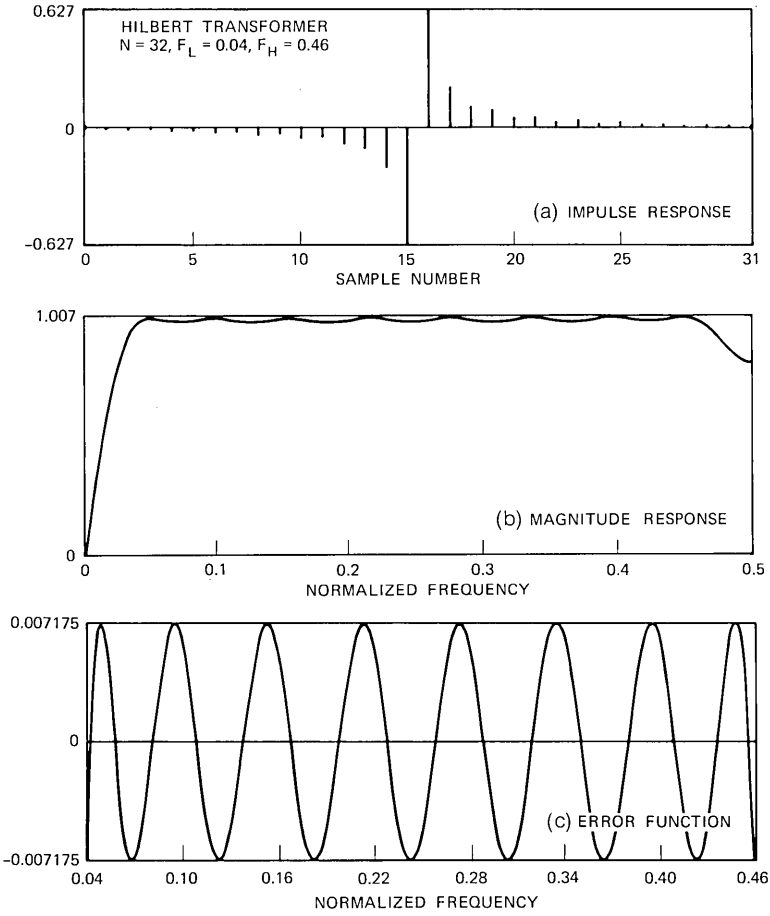

Fig. 2—The impulse response, magnitude response, and error function of an $N = 32$ Hilbert transformer with $F_L = 0.04$ and $F_H = 0.46$.

$$\sum_{n=1}^{(N-1)/2} a(n) \sin(\omega n) = \sum_{n=1}^{(N-1)/2} a(n) \sin[(\pi - \omega)n]$$

$$= \sum_{n=1}^{(N-1)/2} a(n)(-1)^{n+1} \sin(\omega n)$$

or

$$\sum_{n=1}^{(N-1)/2} a(n) \sin(\omega n)[1 - (-1)^{n+1}] = 0$$

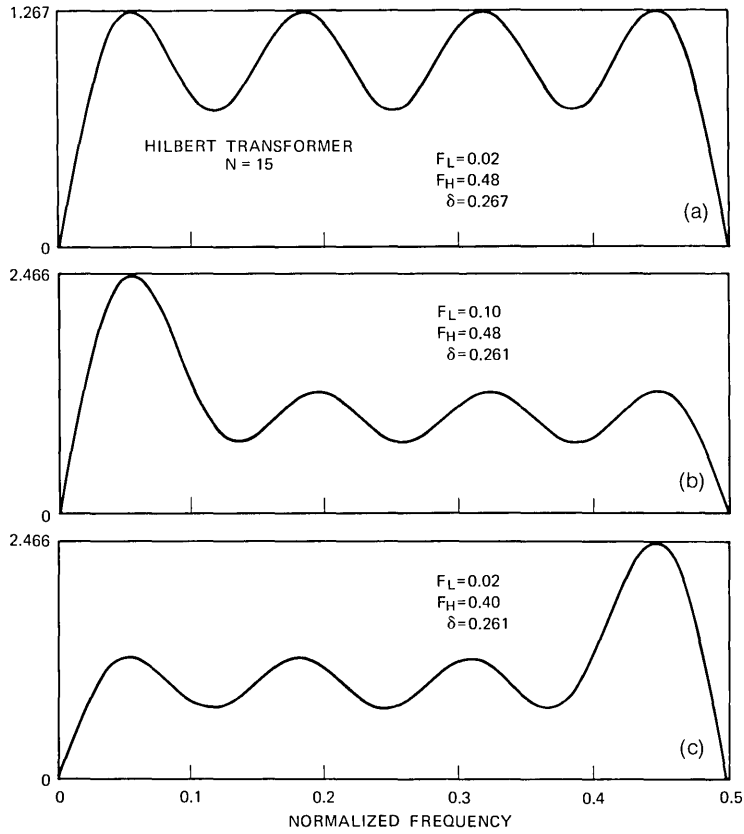

Fig. 3—The magnitude responses of three $N = 15$ Hilbert transformers with different upper and lower cutoff frequencies.

which implies

$$a(n) = 0 \qquad n \text{ even}$$

$$= \text{unconstrained} \qquad n \text{ odd.}$$

Therefore, from eqs. (11), (13b), and (13c), it can be shown that when $(N - 1)/2$ is even, $h(n) = 0$ for $n = 0, 2, 4, \cdots$ and when $(N - 1)/2$ is odd, $h(n) = 0$ for $n = 1, 3, 5, \cdots$. When N is odd and $(N - 1)/2$ is even, $h(0)$ and $h(N - 1)$ are both zero so that the actual length of the impulse response is $N - 2$ samples. Thus, for symmetrical approximations [i.e., satisfying eq. (17)], it is only necessary to consider

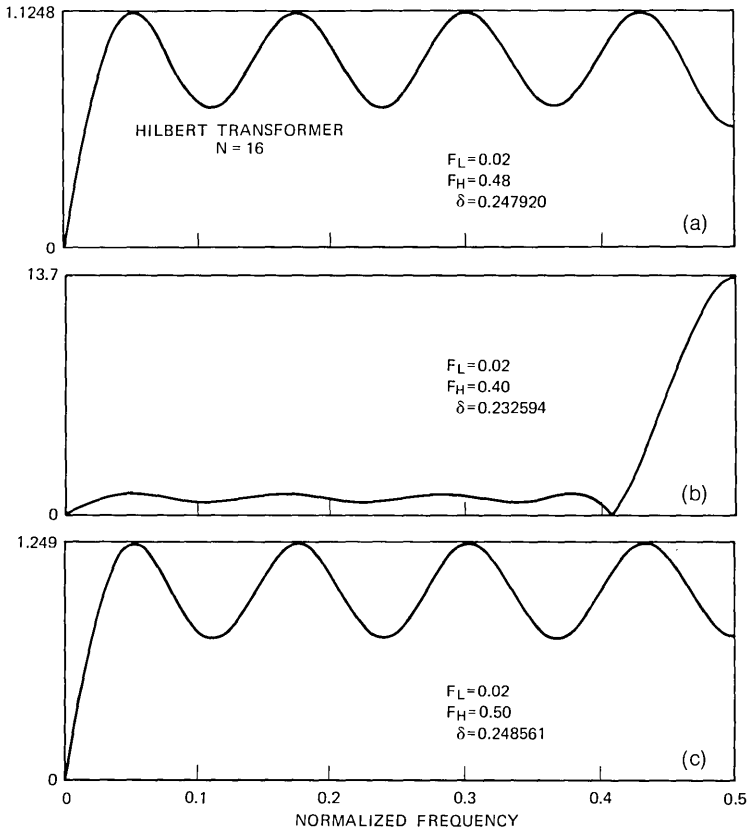

Fig. 4—The magnitude responses of three $N = 16$ Hilbert transformers with different upper and lower cutoff frequencies.

values of N such that $(N - 1)/2$ is odd, since the FIR approximations to be discussed in the next section are unique.

In a manner similar to above, it can be shown that when N is even, alternate coefficients are not zero even if $F_L = 0.5 - F_H$.

This distinction between even- and odd-length impulse responses is important in direct convolutional realizations of such systems. For this realization, the convolution summation

$$y(n) = \hat{x}(n) = \sum_{k=0}^{N-1} h(k)x(n - k)$$

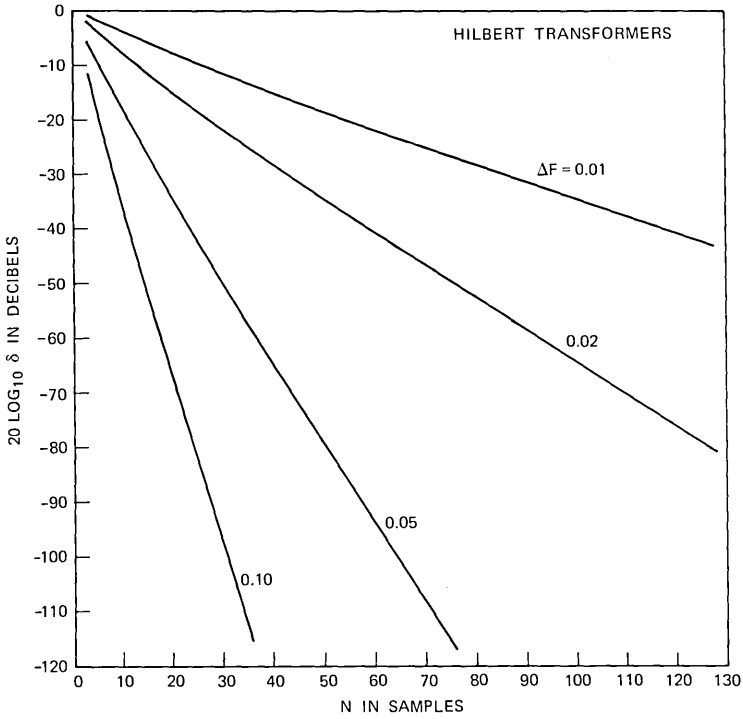

Fig. 5—The curves of 20 $\log_{10} \delta$ versus N for $\Delta F = 0.01$, 0.02, 0.05, and 0.10 for even and odd values of N.

requires $N/2$ multiplications per output sample for N even, whereas only $(N + 1)/4$ multiplications per output sample are required for N odd when alternate samples of $h(n)$ are exactly zero. Thus an odd-length filter requires about half the computation required by an even-length filter one sample shorter.

IV. PROPERTIES OF OPTIMUM FIR HILBERT TRANSFORMERS

The iterative Remez algorithm of McClellan, Parks, and Rabiner[4] can be used to choose the values of $a(n)$ or $b(n)$ that minimize the peak approximation error.

$$\delta = \max_{2\pi F_L \leq \omega \leq 2\pi F_H} [D(e^{j\omega}) - H^*(e^{j\omega})]. \tag{18}$$

The resulting approximation is the unique best approximation (in the minimax sense) to $D(e^{j\omega})$ for a given choice of N, F_L, and F_H. The

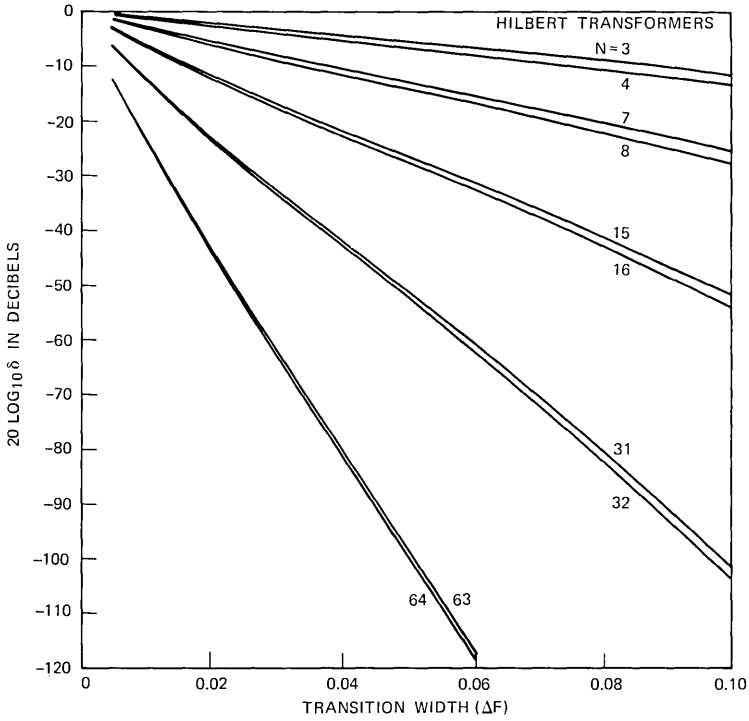

Fig. 6—The curves of 20 $\log_{10} \delta$ versus ΔF for N = 3, 4, 7, 8, 15, 16, 31, 32, 63, and 64.

details of the approximation algorithm are given in Ref. 4. Our present concern is the general properties of the minimax FIR Hilbert transformer approximations obtained using the McClellan, et al., algorithm.

The minimax approximation to the ideal Hilbert transformer is characterized by an error function that is equiripple over the approximation band $2\pi F_L \leqq \omega \leqq 2\pi F_H$. This is illustrated in Figs. 1 through 4 which show the responses of several optimum Hilbert transformer approximations. Figure 1 shows the impulse response, magnitude response, and the error function of an optimum Hilbert transformer having $N = 31$, $F_L = 0.04$, and $F_H = 0.46$. The peak approximation error is 0.008094 and the error curve is seen to be equiripple. It can be seen that $H(e^{j\omega})$ has the symmetry property of eq. (17) and therefore alternate samples of the impulse response are zero. To see that the minimax solution must satisfy eq. (17), assume that

$$H^*(e^{j\omega}) \neq H^*(e^{j(\pi-\omega)});$$

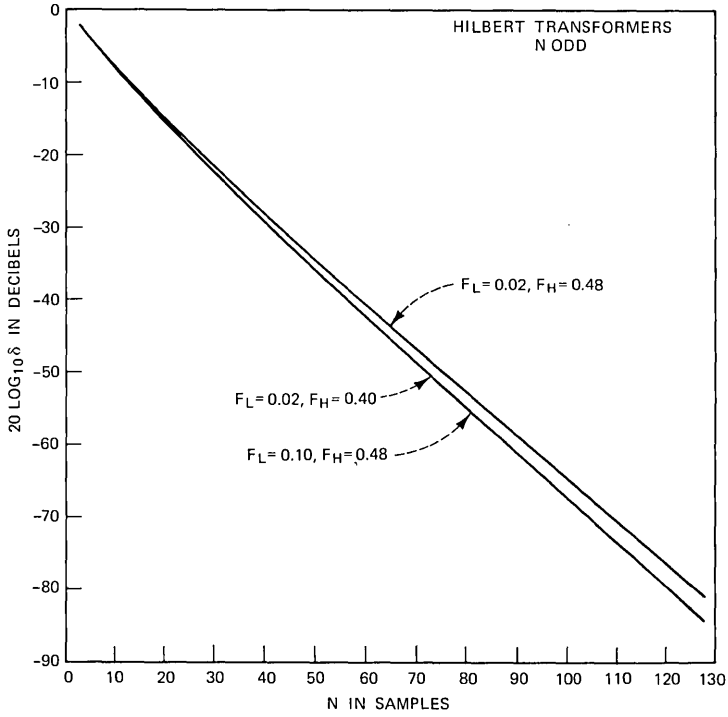

Fig. 7—The curves of 20 $\log_{10} \delta$ versus N for three sets of upper and lower cutoff frequencies with N odd.

then because $H^*(e^{j\omega}) = -H^*(e^{-j\omega})$ and because $F_L = 0.5 - F_H$, both $H^*(e^{j\omega})$ and $H^*(e^{j(\pi-\omega)})$ would satisfy the conditions for optimality, thus contradicting the uniqueness of the optimum approximation.

Figure 2 shows the same responses as in Fig. 1 for the case $N = 32$, $F_L = 0.04$, and $F_H = 0.46$. As previously noted, even though the upper and lower transition widths are equal, all the impulse response coefficients are nonzero because the frequency response cannot have the required symmetry when N is even. As seen in Fig. 2b, the magnitude response is not constrained to be zero at $f = 0.5$. Figure 2c shows the error curve to again be equiripple over the band of approximation. The peak approximation error for this case is 0.007175.

Figures 3 and 4 show the effects of making the upper and lower transition bandwidths unequal. Figure 3 shows three sets of conditions for $N = 15$; i.e., $F_L = 0.02$, $F_H = 0.48$ in Fig. 3a, $F_L = 0.10$, $F_H = 0.48$ in Fig. 3b, and $F_L = 0.02$, $F_H = 0.40$ in Fig. 3c. The peak approxi-

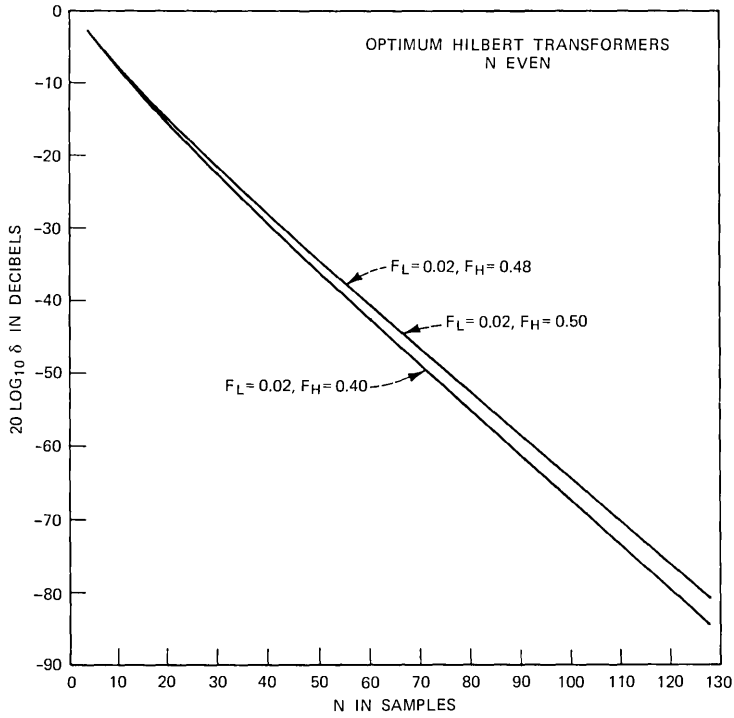

Fig. 8—The curves of 20 $\log_{10} \delta$ versus N for three sets of upper and lower cutoff frequencies with N even.

mation errors are 0.267 for 3a, 0.261 for 3b, and 0.261 for 3c. Thus, even though one of the transition widths changes by a factor of 5:1 (i.e., from 0.02 to 0.10), the change in peak error is on the order of 2 percent. Furthermore, as seen in Figs. 2b and 2c, the magnitude response of the filter peaks up significantly in the wide transition band—a generally undesirable result. Also, when the transition bandwidths are unequal, the symmetry property of the frequency response no longer holds and all the impulse response coefficients are nonzero. In conclusion, the negligible decrease in peak error obtained by using unequal transition bandwidths is more than offset by the undesirable effects in the magnitude and impulse responses. Furthermore, based on this and other similar examples, it is seen that the peak error of approximation is determined primarily by the smaller of the two transition widths when the filter impulse response duration is odd.

Figure 4 shows similar examples of unequal transition widths for

Table I—Wideband Hilbert transformers
($F_L = 0.01$, N odd)

	$N = 55$ $D = 0.0954980$	$N = 59$ $D = 0.0819670$	$N = 63$ $D = 0.0704360$	$N = 67$ $D = 0.0606110$
0	−0.0555675	−0.0477809	−0.0411397	−0.0354857
2	−0.0167258	−0.0145478	−0.0126851	−0.0110945
4	−0.0194725	−0.0169056	−0.0147109	−0.0129141
6	−0.0226148	−0.0196199	−0.0171191	−0.0149160
8	−0.0264560	−0.0228180	−0.0197933	−0.0173168
10	−0.0310395	−0.0266103	−0.0229817	−0.0200179
12	−0.0366783	−0.0311604	−0.0267924	−0.0231739
14	−0.0439134	−0.0368259	−0.0313469	−0.0269273
16	−0.0535532	−0.0440473	−0.0369689	−0.0314908
18	−0.0671782	−0.0536627	−0.0441620	−0.0371318
20	−0.0881528	−0.0672589	−0.0537578	−0.0443213
22	−0.1253143	−0.0882199	−0.0673427	−0.0538996
24	−0.2109951	−0.1253639	−0.0882996	−0.0674568
26	−0.6362148	−0.2110201	−0.1254324	−0.0883853
28		−0.6362213	−0.2110732	−0.1254904
30			−0.6362430	−0.2111059
32				−0.6362516

	$N = 71$ $D = 0.0522480$	$N = 75$ $D = 0.0450690$	$N = 79$ $D = 0.0388830$	$N = 83$ $D = 0.0336160$
0	−0.0306629	−0.0265190	−0.0229388	−0.0198861
2	−0.0097354	−0.0085562	−0.0075151	−0.0066273
4	−0.0113179	−0.0099546	−0.0087784	−0.0077367
6	−0.0131328	−0.0115623	−0.0101565	−0.0090091
8	−0.0151642	−0.0133373	−0.0117808	−0.0103953
10	−0.0175162	−0.0153870	−0.0135612	−0.0119879
12	−0.0202240	−0.0177321	−0.0155902	−0.0137883
14	−0.0233958	−0.0204320	−0.0179182	−0.0158166
16	−0.0271460	−0.0235782	−0.0206260	−0.0181339
18	−0.0316681	−0.0273088	−0.0237742	−0.0208124
20	−0.0372851	−0.0318335	−0.0274953	−0.0239473
22	−0.0444587	−0.0374340	−0.0319865	−0.0276626
24	−0.0540183	−0.0445843	−0.0375627	−0.0321489
26	−0.0675540	−0.0541292	−0.0447012	−0.0377091
28	−0.0884541	−0.0676461	−0.0542333	−0.0448244
30	−0.1255376	−0.0885235	−0.0677331	−0.0543328
32	−0.2111303	−0.1255838	−0.0885965	−0.0678187
34	−0.6362585	−0.2111601	−0.1256401	−0.0886665
36		−0.6362709	−0.2111964	−0.1256905
38			−0.6362830	−0.2112252
40				−0.6362926

$N = 16$. The lower and upper cutoff frequencies for these examples are: $F_L = 0.02$, $F_H = 0.48$ for Fig. 4a, $F_L = 0.02$, $F_H = 0.40$ for Fig. 4b, and $F_L = 0.02$, $F_H = 0.50$ for Fig. 4c. The resulting peak approximation errors are 0.247920 for 4a, 0.232594 for 4b, and 0.248561 for 4c. For the data of Fig. 4b (where the upper transition width is 5 times the lower transition width) the magnitude response becomes

Table I—continued

	$N = 87$ $D = 0.0290440$	$N = 91$ $D = 0.0251650$	$N = 95$ $D = 0.0217910$
0	−0.0172371	−0.0149759	−0.0130099
2	−0.0058530	−0.0051648	−0.0045718
4	−0.0068525	−0.0060643	−0.0053689
6	−0.0079733	−0.0070506	−0.0062800
8	−0.0092241	−0.0081894	−0.0072616
10	−0.0106323	−0.0094332	−0.0083873
12	−0.0122073	−0.0108306	−0.0096455
14	−0.0139949	−0.0124102	−0.0110350
16	−0.0160152	−0.0141947	−0.0126022
18	−0.0183265	−0.0162129	−0.0143770
20	−0.0210006	−0.0185127	−0.0163895
22	−0.0241199	−0.0211705	−0.0186922
24	−0.0278194	−0.0242786	−0.0213465
26	−0.0322920	−0.0279699	−0.0244424
28	−0.0378402	−0.0324331	−0.0281175
30	−0.0449441	−0.0379675	−0.0325665
32	−0.0544359	−0.0450542	−0.0380864
34	−0.0679017	−0.0545312	−0.0451608
36	−0.0887289	−0.0679795	−0.0546233
38	−0.1257349	−0.0887891	−0.0680547
40	−0.2112513	−0.1257759	−0.0888468
42	−0.6363005	−0.2112740	−0.1258168
44		−0.6363082	−0.2112989
46			−0.6363167

very large in the upper transition band. For this case, the peak approximation error is about 6 percent smaller than the peak error for equal transition widths. Thus the slight decrease in peak error does not compensate for the undesirable peaking in the transition region of the frequency response. On the other hand, Fig. 4c shows that setting the upper transition width to 0, i.e., letting $F_H = 0.50$ produces a negligible change in the peak error. Based on these and other examples, it has been found that for even values of N, the peak approximation error depends almost entirely on the lower transition width (because of the zero at $\omega = 0$). Thus the upper transition width should be made less than or equal to the lower transition width to minimize the peaking in the upper transition band.

V. DESIGN DATA FOR OPTIMUM HILBERT TRANSFORMERS

The basic Hilbert transformer parameters are N, F_L, F_H, and δ, the peak approximation error (or the ripple) of the filter. It is assumed that $F_H = 0.5 - F_L$, i.e., the upper and lower transition widths are equal, for the data to be presented in this section. Thus there are only three parameters: N, δ, and $\Delta F = F_L = 0.5 - F_H$. A large set of

Table II—Wideband Hilbert transformers ($F_L = 0.02$, N odd)

	$N = 31$ $D = 0.0750950$	$N = 35$ $D = 0.0554920$	$N = 39$ $D = 0.0413090$	$N = 43$ $D = 0.0308560$
0	−0.0510610	−0.0380993	−0.0286598	−0.0216528
2	−0.0315637	−0.0241883	−0.0187406	−0.0146215
4	−0.0425110	−0.0324289	−0.0250998	−0.0196329
6	−0.0577251	−0.0433438	−0.0332813	−0.0259590
8	−0.0805394	−0.0584460	−0.0441044	−0.0340917
10	−0.1196632	−0.0811443	−0.0591324	−0.0448233
12	−0.2075778	−0.1201457	−0.0816868	−0.0597557
14	−0.6350674	−0.2078690	−0.1206017	−0.0821807
16		−0.6351797	−0.2081174	−0.1209598
18			−0.6352463	−0.2083547
20				−0.6353344

	$N = 47$ $D = 0.0231120$	$N = 51$ $D = 0.0172930$	$N = 55$ $D = 0.0130540$	$N = 59$ $D = 0.0098300$
0	−0.0164094	−0.0124427	−0.0095117	−0.0072606
2	−0.0114516	−0.0090077	−0.0071168	−0.0056269
4	−0.0154645	−0.0122483	−0.0097423	−0.0077659
6	−0.0204740	−0.0162677	−0.0129935	−0.0104170
8	−0.0267802	−0.0212626	−0.0170107	−0.0136817
10	−0.0348570	−0.0275401	−0.0219890	−0.0176945
12	−0.0455267	−0.0355514	−0.0282298	−0.0226586
14	−0.0603428	−0.0461607	−0.0361989	−0.0288540
16	−0.0826818	−0.0608791	−0.0467260	−0.0367817
18	−0.1213191	−0.0831083	−0.0613745	−0.0472445
20	−0.2085553	−0.1216348	−0.0835102	−0.0618163
22	−0.6353986	−0.2087518	−0.1219297	−0.0838702
24		−0.6354638	−0.2089363	−0.1221934
26			−0.6355268	−0.2090940
28				−0.6355770

	$N = 63$ $D = 0.0074430$	$N = 67$ $D = 0.0056250$	$N = 71$ $D = 0.0042570$	$N = 75$ $D = 0.0032330$
0	−0.0055706	−0.0042726	−0.0032801	−0.0025265
2	−0.0044618	−0.0035440	−0.0028140	−0.0022373
4	−0.0062078	−0.0049725	−0.0039850	−0.0031981
6	−0.0083775	−0.0067529	−0.0054510	−0.0044073
8	−0.0110475	−0.0089470	−0.0072626	−0.0059065
10	−0.0143195	−0.0116331	−0.0094805	−0.0077446
12	−0.0183266	−0.0149084	−0.0121819	−0.0099820
14	−0.0232716	−0.0189147	−0.0154612	−0.0126922
16	−0.0294397	−0.0238353	−0.0194595	−0.0159751
18	−0.0373177	−0.0299778	−0.0243626	−0.0199615
20	−0.0477262	−0.0378112	−0.0304737	−0.0248494
22	−0.0622273	−0.0481616	−0.0382687	−0.0309326
24	−0.0841979	−0.0626016	−0.0485700	−0.0386871
26	−0.1224340	−0.0845011	−0.0629480	−0.0489418
28	−0.2092445	−0.1226550	−0.0847785	−0.0632654
30	−0.6356280	−0.2093771	−0.1228583	−0.0850334
32		−0.6356716	−0.2095022	−0.1230441
34			−0.6357140	−0.2096158
36				−0.6357524

Table II—continued

	$N = 79$ $D = 0.0024390$	$N = 83$ $D = 0.0018650$	$N = 87$ $D = 0.0014170$	$N = 91$ $D = 0.0010810$
0	−0.0019358	−0.0015010	−0.0011577	−0.0008961
2	−0.0017746	−0.0014164	−0.0011264	−0.0008980
4	−0.0025624	−0.0020630	−0.0016565	−0.0013331
6	−0.0035600	−0.0028866	−0.0023363	−0.0018946
8	−0.0048021	−0.0039174	−0.0031921	−0.0026054
10	−0.0063300	−0.0051899	−0.0042528	−0.0034906
12	−0.0081910	−0.0067435	−0.0055519	−0.0045785
14	−0.0104453	−0.0086254	−0.0071273	−0.0059016
16	−0.0131630	−0.0108928	−0.0090270	−0.0074976
18	−0.0164470	−0.0136187	−0.0113048	−0.0094123
20	−0.0204251	−0.0168997	−0.0140366	−0.0117008
22	−0.0252943	−0.0208693	−0.0173154	−0.0144355
24	−0.0313515	−0.0257215	−0.0212753	−0.0177116
26	−0.0390711	−0.0317515	−0.0261107	−0.0216618
28	−0.0492818	−0.0394368	−0.0321163	−0.0264799
30	−0.0635544	−0.0496063	−0.0397698	−0.0324618
32	−0.0852651	−0.0638299	−0.0499010	−0.0400852
34	−0.1232135	−0.0854855	−0.0640805	−0.0501796
36	−0.2097186	−0.1233750	−0.0856865	−0.0643163
38	−0.6357869	−0.2098177	−0.1235215	−0.0858747
40		−0.6358204	−0.2099064	−0.1236593
42			−0.6358499	−0.2099905
44				−0.6358782

	$N = 95$ $D = 0.0008240$
0	−0.0006935
2	−0.0007156
4	−0.0010724
6	−0.0015362
8	−0.0021265
10	−0.0028655
12	−0.0037775
14	−0.0048903
16	−0.0062348
18	−0.0078492
20	−0.0097764
22	−0.0120734
24	−0.0148104
26	−0.0180826
28	−0.0220234
30	−0.0268256
32	−0.0327847
34	−0.0403791
36	−0.0504390
38	−0.0645363
40	−0.0860506
42	−0.1237874
44	−0.2100684
46	−0.6359043

Table III—Wideband Hilbert transformers
($F_L = 0.05$, N odd)

	$N = 15$ $D = 0.0475550$	$N = 19$ $D = 0.0227350$	$N = 23$ $D = 0.0110710$	$N = 27$ $D = 0.0054480$
0	−0.0529897	−0.0272752	−0.0144218	−0.0077528
2	−0.0882059	−0.0478756	−0.0272241	−0.0158203
4	−0.1868274	−0.0931810	−0.0525858	−0.0311403
6	−0.6278288	−0.1902395	−0.0971984	−0.0564206
8		−0.6290423	−0.1929460	−0.1004278
10			−0.6299931	−0.1950992
12				−0.6307509

	$N = 31$ $D = 0.0026800$	$N = 35$ $D = 0.0013490$	$N = 39$ $D = 0.0006790$	$N = 43$ $D = 0.0003430$
0	−0.0041956	−0.0023116	−0.0012787	−0.0007098
2	−0.0092821	−0.0054978	−0.0032636	−0.0019390
4	−0.0188358	−0.0115383	−0.0071031	−0.0043804
6	−0.0344010	−0.0214763	−0.0135513	−0.0085912
8	−0.0595516	−0.0371891	−0.0237704	−0.0153599
10	−0.1030376	−0.0621930	−0.0395684	−0.0257859
12	−0.1968315	−0.1052127	−0.0644154	−0.0416246
14	−0.6313536	−0.1982643	−0.1070280	−0.0663138
16		−0.6318550	−0.1994533	−0.1085658
18			−0.6322687	−0.2004560
20				−0.6326171

	$N = 47$ $D = 0.0001730$	$N = 51$ $D = 0.0000880$	$N = 55$ $D = 0.0000450$	$N = 59$ $D = 0.0000230$
0	−0.0003957	−0.0002206	−0.0001243	−0.0000703
2	−0.0011534	−0.0006843	−0.0004080	−0.0002435
4	−0.0027042	−0.0016654	−0.0010288	−0.0006354
6	−0.0054585	−0.0034627	−0.0022014	−0.0013985
8	−0.0099769	−0.0064829	−0.0042215	−0.0027472
10	−0.0170000	−0.0112537	−0.0074746	−0.0049655
12	−0.0275785	−0.0184771	−0.0124599	−0.0084200
14	−0.0434280	−0.0291657	−0.0198467	−0.0135851
16	−0.0679619	−0.0450039	−0.0306162	−0.0211031
18	−0.1098913	−0.0693880	−0.0464284	−0.0319297
20	−0.2013159	−0.1110305	−0.0706666	−0.0477053
22	−0.6329151	−0.2020517	−0.1120458	−0.0718040
24		−0.6331695	−0.2027050	−0.1129440
26			−0.6333949	−0.2032809
28				−0.6335934

measurements of δ as a function of ΔF and N were made, and the results are shown in Figs. 5 and 6. Figure 5 shows a plot of 20 $\log_{10} \delta$ as a function of N for values of ΔF of 0.01, 0.02, 0.05, and 0.10, and for even and odd values of N in the range $3 \leqq N \leqq 128$. The curves for even and odd values of N, for fixed transition widths, are almost indistinguishable on the scales of Fig. 5, and hence a single curve is

Table III—continued

	$N = 63$ $D = 0.0000120$	$N = 67$ $D = 0.0000060$	$N = 71$ $D = 0.0000030$	$N = 75$ $D = 0.0000020$
0	−0.0000396	−0.0000224	−0.0000127	−0.0000072
2	−0.0001447	−0.0000860	−0.0000512	−0.0000305
4	−0.0003910	−0.0002402	−0.0001476	−0.0000907
6	−0.0008853	−0.0005592	−0.0003531	−0.0002226
8	−0.0017821	−0.0011533	−0.0007456	−0.0004811
10	−0.0032906	−0.0021759	−0.0014366	−0.0009467
12	−0.0056838	−0.0038306	−0.0025777	−0.0017311
14	−0.0093111	−0.0063793	−0.0043663	−0.0029833
16	−0.0146260	−0.0101567	−0.0070545	−0.0048946
18	−0.0222481	−0.0155978	−0.0109630	−0.0077079
20	−0.0331126	−0.0233034	−0.0165113	−0.0117315
22	−0.0488448	−0.0341916	−0.0242838	−0.0173711
24	−0.0728120	−0.0498758	−0.0351847	−0.0251972
26	−0.1137361	−0.0737183	−0.0508176	−0.0361023
28	−0.2037871	−0.1144451	−0.0745415	−0.0516821
30	−0.6337675	−0.2042389	−0.1150864	−0.0752934
32		−0.6339227	−0.2046464	−0.1156699
34			−0.6340625	−0.2050163
36				−0.6341892

	$N = 79$ $D = 0.0000010$
0	−0.0000041
2	−0.0000179
4	−0.0000550
6	−0.0001389
8	−0.0003074
10	−0.0006182
12	−0.0011532
14	−0.0020239
16	−0.0033761
18	−0.0053956
20	−0.0083167
22	−0.0124372
24	−0.0181511
26	−0.0260178
28	−0.0369200
30	−0.0524475
32	−0.0759556
34	−0.1161821
36	−0.2053402
38	−0.6343000

given for both.* Based on these curves, it is seen that the larger the transition bandwidth of the Hilbert transformer, the faster the decrease of peak error with increasing N. Thus for $\Delta F = 0.01$, the value of $20 \log_{10} \delta$ decreases by only about 42 dB as N varies from 3 to 128;

* Points on each curve are connected for convenience in plotting.

Table IV—Wideband Hilbert transformers
($F_L = 0.10$, N odd)

	$N = 7$ $D = 0.0515030$	$N = 11$ $D = 0.0111870$	$N = 15$ $D = 0.0025460$	$N = 19$ $D = 0.0005960$
0	−0.1270413	−0.0379718	−0.0125869	−0.0043760
2	−0.6012845	−0.1426690	−0.0517464	−0.0203793
4		−0.6102909	−0.1563345	−0.0622833
6			−0.6159002	−0.1655747
8				−0.6195926

	$N = 23$ $D = 0.0001420$	$N = 27$ $D = 0.0000340$	$N = 31$ $D = 0.0000080$	$N = 35$ $D = 0.0000020$
0	−0.0015643	−0.0005691	−0.0002098	−0.0000778
2	−0.0082383	−0.0033614	−0.0013764	−0.0005626
4	−0.0269557	−0.0119241	−0.0052972	−0.0023445
6	−0.0702312	−0.0324676	−0.0153339	−0.0072495
8	−0.1722057	−0.0764431	−0.0371398	−0.0184388
10	−0.6221851	−0.1771894	−0.0814346	−0.0411255
12		−0.6240992	−0.1810706	−0.0855177
14			−0.6255683	−0.1841662
16				−0.6267261

	$N = 39$ $D = 0.0000010$
0	−0.0000292
2	−0.0002304
4	−0.0010348
6	−0.0034149
8	−0.0091741
10	−0.0212775
12	−0.0445889
14	−0.0889475
16	−0.1867128
18	−0.6276691

whereas for $\Delta F = 0.05$, the value of $20 \log_{10} \delta$ decreases by about 112 dB as N varies from 3 to 76.

Figure 6 shows plots of $20 \log_{10} \delta$ as a function of ΔF for even and odd values of N. The actual values used were $N = 3, 4, 7, 8, 15, 16, 31, 32, 63$, and 64.† As seen in this figure, as ΔF tends to 0, $20 \log_{10} \delta$ tends to 0 dB or a peak error of 1.0, independent of N. It is also seen from Fig. 6 that the larger the value of N, the faster the decrease of peak error with increasing transition width.

From Figs. 5 and 6 it is seen that for a fixed value of δ the product of N and ΔF is approximately a constant. Thus a simple relation of

† The values $N = 3, 7, 15, 31, 63$ were chosen since they satisfy the condition that $(N - 1)/2$ be an even integer.

Table V—Wideband Hilbert transformers
($F_L = 0.02$, N even)

	$N = 28$ $D = 0.0951180$	$N = 30$ $D = 0.0814020$	$N = 32$ $D = 0.0701460$	$N = 34$ $D = 0.0602910$
0	−0.0583712	−0.0503043	−0.0435848	−0.0376976
1	−0.0113151	−0.0095838	−0.0080700	−0.0068443
2	−0.0250915	−0.0221816	−0.0196145	−0.0174179
3	−0.0167301	−0.0141454	−0.0119484	−0.0102336
4	−0.0325317	−0.0285499	−0.0251301	−0.0222922
5	−0.0247101	−0.0206725	−0.0174096	−0.0148471
6	−0.0431589	−0.0372953	−0.0325172	−0.0285854
7	−0.0372340	−0.0305387	−0.0254134	−0.0213430
8	−0.0603244	−0.0504644	−0.0431125	−0.0372679
9	−0.0602139	−0.0470364	−0.0378782	−0.0311972
10	−0.0952250	−0.0740018	−0.0601857	−0.0503464
11	−0.1181848	−0.0814343	−0.0608355	−0.0476850
12	−0.2181622	−0.1322336	−0.0948931	−0.0737466
13	−0.6290021	−0.2041128	−0.1187316	−0.0819751
14		−0.6431801	−0.2177788	−0.1319350
15			−0.6295038	−0.2045900
16				−0.6427692

	$N = 36$ $D = 0.0519110$	$N = 38$ $D = 0.0446700$	$N = 40$ $D = 0.0385670$	$N = 42$ $D = 0.0333770$
0	−0.0326595	−0.0282952	−0.0245988	−0.0214323
1	−0.0058524	−0.0050023	−0.0042702	−0.0036896
2	−0.0154507	−0.0137594	−0.0122671	−0.0109456
3	−0.0087342	−0.0075285	−0.0064963	−0.0055936
4	−0.0197750	−0.0176078	−0.0157127	−0.0140321
5	−0.0126789	−0.0109166	−0.0094489	−0.0081451
6	−0.0252849	−0.0224120	−0.0199636	−0.0178443
7	−0.0181526	−0.0155645	−0.0133761	−0.0115729
8	−0.0325485	−0.0286715	−0.0253906	−0.0225664
9	−0.0260945	−0.0220630	−0.0188546	−0.0161964
10	−0.0430437	−0.0372556	−0.0325934	−0.0287451
11	−0.0385775	−0.0318807	−0.0267466	−0.0227040
12	−0.0600130	−0.0502752	−0.0429692	−0.0372458
13	−0.0614526	−0.0483355	−0.0391895	−0.0324747
14	−0.0946252	−0.0735543	−0.0598556	−0.0501784
15	−0.1192901	−0.0825735	−0.0620113	−0.0488826
16	−0.2174115	−0.1316258	−0.0943740	−0.0733751
17	−0.6299570	−0.2051058	−0.1197944	−0.0830622
18		−0.6423602	−0.2170604	−0.1313650
19			−0.6303873	−0.2055264
20				−0.6420162

the form

$$N\Delta F \approx -0.61 \log_{10} \delta \tag{19}$$

has been suggested by Kaiser as a reasonable approximation to most cases of interest in Figs. 5 and 6. Similar inverse proportionality between filter order N and transition with ΔF was originally noted by Kaiser.[6]

Table V—continued

	$N = 44$ $D = 0.0289000$	$N = 46$ $D = 0.0249860$	$N = 48$ $D = 0.0216400$	$N = 50$ $D = 0.0187460$
0	−0.0186686	−0.0162707	−0.0141858	−0.0123663
1	−0.0031491	−0.0027127	−0.0023498	−0.0020288
2	−0.0097718	−0.0087416	−0.0078077	−0.0069733
3	−0.0048431	−0.0042104	−0.0036580	−0.0031894
4	−0.0125771	−0.0112581	−0.0100901	−0.0090601
5	−0.0070753	−0.0061883	−0.0053834	−0.0047149
6	−0.0159563	−0.0143100	−0.0128560	−0.0115401
7	−0.0100472	−0.0087497	−0.0076664	−0.0067136
8	−0.0201647	−0.0180530	−0.0161867	−0.0145598
9	−0.0139985	−0.0121846	−0.0106353	−0.0093153
10	−0.0254980	−0.0227293	−0.0203292	−0.0182394
11	−0.0194634	−0.0168032	−0.0145949	−0.0127501
12	−0.0326375	−0.0288154	−0.0256046	−0.0228635
13	−0.0273381	−0.0232943	−0.0200391	−0.0173661
14	−0.0429264	−0.0372432	−0.0326723	−0.0288867
15	−0.0397492	−0.0330275	−0.0278909	−0.0238409
16	−0.0597342	−0.0500908	−0.0428880	−0.0372447
17	−0.0625157	−0.0493928	−0.0402628	−0.0335462
18	−0.0941798	−0.0731977	−0.0596122	−0.0500194
19	−0.1202296	−0.0835088	−0.0629931	−0.0498713
20	−0.2167857	−0.1311139	−0.0939759	−0.0730461
21	−0.6307415	−0.2059084	−0.1206478	−0.0839375
22		−0.6416971	−0.2165035	−0.1308803
23			−0.6311019	−0.2062839
24				−0.6413871

As discussed earlier, it has been found that for odd values of N, the peak error is determined primarily by the smaller transition width of the filter; whereas for even values of N, the peak error is determined primarily by the lower transition width. Figures 7 and 8 present data which essentially verify these claims. Figure 7 shows curves of $20 \log_{10} \delta$ as a function of N for three sets of conditions: (*i*) $F_L = 0.02$, $F_H = 0.48$, (*ii*) $F_L = 0.10$, $F_H = 0.48$, (*iii*) $F_L = 0.02$, $F_H = 0.40$ for filters where N is odd. The curves for cases (*ii*) and (*iii*) are indistinguishable on these scales. This figure shows that the maximum differences between the peak errors for these cases is about 3.4 dB for $N = 127$, 1.5 dB for $N = 63$, and 0.7 dB for $N = 31$. Figure 8 shows similar results for even values of N. For this figure the three cases were: (*i*) $F_L = 0.02$, $F_H = 0.48$, (*ii*) $F_L = 0.02$, $F_H = 0.50$, (*iii*) $F_L = 0.02$, $F_H = 0.40$. In this case the curves for cases (*i*) and (*ii*) are indistinguishable. The maximum differences between peak errors are almost identical to the errors for comparable cases when N is odd. These figures substantiate the conclusions stated previously—that the peak ripple is determined

primarily by the smaller transition width for N odd, and by the lower transition width for N even.

VI. APPLICATION OF FIR HILBERT TRANSFORMERS

The above discussion suggests that, for direct realizations, the most efficient FIR Hilbert transformer (i.e., using the smallest number of multiplications per sample to obtain a desired value of peak approximation error) has as large a transition bandwidth as possible, and an odd number of impulse response samples. As an example, to obtain a peak error of less than 1 percent ($\delta \leqq 0.01$) requires the following values of N (as a function of ΔF):

ΔF	N (odd)	Number of Multiplications per Sample	N (even)	Number of Multiplications per Sample
0.01	119	30	118	59
0.02	59	15	60	30
0.05	27	7	24	12
0.10	11	3	12	6

whereas to obtain a peak error of less than 0.1 percent ($\delta \leqq 0.001$) requires:

ΔF	N (odd)	Number of Multiplications per Sample	N (even)	Number of Multiplications per Sample
0.01	>127	—	>128	—
0.02	95	24	94	47
0.05	39	10	38	19
0.10	19	5	18	9

The above tables indicate the substantial processing advantages of odd-length Hilbert transformers with symmetrical frequency responses.

Further discussion of the relative merits of even and odd values of N for signal processing applications is given in Ref. 7.

A subset of the Hilbert transformers designed in this study is given in Tables I through VII. These are symmetrical approximations ($F_L = 0.5 - F_H$), with $F_L = 0.01$, 0.02, 0.05, and 0.10 and both even

Table VI—Wideband Hilbert transformers ($F_L = 0.05$, N even)

	$N = 12$ $D = 0.0877240$	$N = 14$ $D = 0.0597790$	$N = 16$ $D = 0.0410810$	$N = 18$ $D = 0.0283820$
0	−0.0736839	−0.0521805	−0.0373724	−0.0269588
1	−0.0363534	−0.0241948	−0.0165007	−0.0114666
2	−0.0903566	−0.0652477	−0.0485333	−0.0367081
3	−0.0984140	−0.0611815	−0.0408488	−0.0284029
4	−0.2224184	−0.1316266	−0.0902308	−0.0658999
5	−0.6161179	−0.1891560	−0.1026126	−0.0652968
6		−0.6508407	−0.2207546	−0.1308795
7			−0.6192659	−0.1925465
8				−0.6486609

	$N = 20$ $D = 0.0196660$	$N = 22$ $D = 0.0136990$	$N = 24$ $D = 0.0095770$	$N = 26$ $D = 0.0067230$
0	−0.0195468	−0.0142530	−0.0104407	−0.0076828
1	−0.0080974	−0.0057586	−0.0041350	−0.0029765
2	−0.0280261	−0.0215342	−0.0166115	−0.0128555
3	−0.0202798	−0.0147091	−0.0107983	−0.0079913
4	−0.0496954	−0.0381941	−0.0296899	−0.0232496
5	−0.0447385	−0.0319145	−0.0233406	−0.0173478
6	−0.0902215	−0.0664464	−0.0506653	−0.0394210
7	−0.1060595	−0.0686265	−0.0478081	−0.0346831
8	−0.2194255	−0.1302985	−0.0902502	−0.0669352
9	−0.6218219	−0.1952413	−0.1087422	−0.0712032
10		−0.6469252	−0.2184181	−0.1298971
11			−0.6237772	−0.1973125
12				−0.6456253

	$N = 28$ $D = 0.0047080$	$N = 30$ $D = 0.0033280$	$N = 32$ $D = 0.0023350$	$N = 34$ $D = 0.0016500$
0	−0.0056474	−0.0041824	−0.0030839	−0.0022879
1	−0.0021606	−0.0015760	−0.0011563	−0.0008479
2	−0.0099450	−0.0077198	−0.0059708	−0.0046343
3	−0.0059617	−0.0044628	−0.0033653	−0.0025388
4	−0.0182731	−0.0144233	−0.0113703	−0.0089951
5	−0.0130602	−0.0098963	−0.0075683	−0.0058006
6	−0.0310326	−0.0246460	−0.0196285	−0.0157095
7	−0.0258302	−0.0195324	−0.0149782	−0.0115559
8	−0.0514404	−0.0404127	−0.0321135	−0.0257649
9	−0.0502551	−0.0369370	−0.0278948	−0.0213933
10	−0.0902847	−0.0673263	−0.0520567	−0.0411844
11	−0.1108639	−0.0732746	−0.0522566	−0.0388270
12	−0.2176401	−0.1295790	−0.0903025	−0.0676060
13	−0.6253137	−0.1989583	−0.1125885	−0.0750001
14		−0.6445903	−0.2170031	−0.1292842
15			−0.6265553	−0.2003315
16				−0.6437163

Table VI—continued

	$N = 36$ $D = 0.0011710$	$N = 38$ $D = 0.0008310$	$N = 40$ $D = 0.0005910$	$N = 42$ $D = 0.0004190$
0	−0.0017032	−0.0012691	−0.0009465	−0.0007052
1	−0.0006244	−0.0004591	−0.0003397	−0.0002522
2	−0.0036007	−0.0027975	−0.0021716	−0.0016832
3	−0.0019213	−0.0014545	−0.0011051	−0.0008426
4	−0.0071254	−0.0056461	−0.0044715	−0.0035365
5	−0.0044616	−0.0034385	−0.0026575	−0.0020612
6	−0.0126059	−0.0101283	−0.0081395	−0.0065367
7	−0.0089661	−0.0069853	−0.0054622	−0.0042881
8	−0.0207937	−0.0168426	−0.0136681	−0.0110982
9	−0.0165912	−0.0129692	−0.0102003	−0.0080657
10	−0.0330465	−0.0267559	−0.0217832	−0.0177909
11	−0.0295922	−0.0229337	−0.0179839	−0.0142292
12	−0.0526097	−0.0418905	−0.0338324	−0.0275664
13	−0.0538732	−0.0403563	−0.0310313	−0.0242859
14	−0.0903686	−0.0679104	−0.0530752	−0.0424512
15	−0.1139494	−0.0763652	−0.0552251	−0.0416870
16	−0.2165425	−0.1291214	−0.0904306	−0.0681302
17	−0.6275161	−0.2013870	−0.1150770	−0.0775488
18		−0.6430778	−0.2161671	−0.1289480
19			−0.6283058	−0.2023087
20				−0.6425071

	$N = 44$ $D = 0.0002970$	$N = 46$ $D = 0.0002110$	$N = 48$ $D = 0.0001500$	$N = 50$ $D = 0.0001070$
0	−0.0005261	−0.0003931	−0.0002935	−0.0002189
1	−0.0001872	−0.0001391	−0.0001037	−0.0000773
2	−0.0013044	−0.0010110	−0.0007823	−0.0006044
3	−0.0006418	−0.0004892	−0.0003738	−0.0002860
4	−0.0027965	−0.0022113	−0.0017457	−0.0013761
5	−0.0015986	−0.0012405	−0.0009647	−0.0007512
6	−0.0052500	−0.0042167	−0.0033821	−0.0027090
7	−0.0033696	−0.0026507	−0.0020894	−0.0016495
8	−0.0090193	−0.0073334	−0.0059575	−0.0048348
9	−0.0063942	−0.0050794	−0.0040451	−0.0032278
10	−0.0145676	−0.0119479	−0.0098009	−0.0080374
11	−0.0113197	−0.0090412	−0.0072487	−0.0058283
12	−0.0226020	−0.0186095	−0.0153542	−0.0126823
13	−0.0192248	−0.0153413	−0.0123211	−0.0099429
14	−0.0344657	−0.0282591	−0.0233103	−0.0193023
15	−0.0323002	−0.0254597	−0.0203150	−0.0163525
16	−0.0534295	−0.0429264	−0.0350095	−0.0288317
17	−0.0564114	−0.0428283	−0.0333997	−0.0265154
18	−0.0904417	−0.0683159	−0.0537337	−0.0433047
19	−0.1160705	−0.0785556	−0.0574287	−0.0438482
20	−0.2158094	−0.1288020	−0.0904537	−0.0684414
21	−0.6290120	−0.2030878	−0.1169158	−0.0794554
22		−0.6420267	−0.2155088	−0.1286420
23			−0.6296096	−0.2037902
24				−0.6415800

Table VII—Wideband Hilbert transformers
($F_L = 0.10$, N even)

	$N = 6$ $D = 0.0908340$	$N = 8$ $D = 0.0409350$	$N = 10$ $D = 0.0188980$	$N = 12$ $D = 0.0088710$
0	−0.1291026	−0.0685959	−0.0381031	−0.0217187
1	−0.1482639	−0.0642207	−0.0320979	−0.0171366
2	−0.6651173	−0.2171077	−0.1158537	−0.0691363
3		−0.5959958	−0.1657805	−0.0792088
4			−0.6563900	−0.2154404
5				−0.6069062

	$N = 14$ $D = 0.0042080$	$N = 16$ $D = 0.0020170$	$N = 18$ $D = 0.0009720$	$N = 20$ $D = 0.0004710$
0	−0.0125849	−0.0073860	−0.0043698	−0.0026042
1	−0.0095049	−0.0053963	−0.0031146	−0.0018192
2	−0.0429910	−0.0272417	−0.0174135	−0.0111820
3	−0.0436051	−0.0254919	−0.0153794	−0.0094478
4	−0.1183342	−0.0735061	−0.0478330	−0.0317741
5	−0.1765512	−0.0886565	−0.0513442	−0.0315749
6	−0.6513727	−0.2145760	−0.1200051	−0.0764602
7		−0.6134417	−0.1833481	−0.0950101
8			−0.6483398	−0.2140796
9				−0.6176569

	$N = 22$ $D = 0.0002300$	$N = 24$ $D = 0.0001120$	$N = 26$ $D = 0.0000550$	$N = 28$ $D = 0.0000270$
0	−0.0015625	−0.0009411	−0.0005678	−0.0003437
1	−0.0010738	−0.0006373	−0.0003815	−0.0002288
2	−0.0072008	−0.0046408	−0.0029868	−0.0019225
3	−0.0058736	−0.0036749	−0.0023155	−0.0014609
4	−0.0213129	−0.0143562	−0.0096696	−0.0065138
5	−0.0200200	−0.0129010	−0.0084074	−0.0055027
6	−0.0512512	−0.0351492	−0.0243353	−0.0169296
7	−0.0569012	−0.0362082	−0.0237777	−0.0158604
8	−0.1211777	−0.0785910	−0.0537606	−0.0377068
9	−0.1879880	−0.0995599	−0.0611011	−0.0398850
10	−0.6463321	−0.2137791	−0.1220080	−0.0801285
11		−0.6205691	−0.1913704	−0.1030239
12			−0.6448821	−0.2135097
13				−0.6227594

and odd values of N from 3 to 95.* The peak approximation error in the band $2\pi F_L \leqq \omega \leqq 2\pi F_H$, denoted D, is given for each case as well as the impulse response of the filter. Note that only the first half of the impulse response is given in the table; the last half can be obtained using eq. (11). When N is odd, only the even-indexed samples are given.

* Only those Hilbert transformers for which $\delta < 0.1$ are given in these tables.

Table VII—continued

	$N = 30$ $D = 0.0000130$	$N = 32$ $D = 0.0000070$	$N = 34$ $D = 0.0000030$	$N = 36$ $D = 0.0000020$
0	−0.0002100	−0.0001278	−0.0000778	−0.0000474
1	−0.0001373	−0.0000832	−0.0000508	−0.0000309
2	−0.0012437	−0.0007997	−0.0005133	−0.0003291
3	−0.0009210	−0.0005851	−0.0003730	−0.0002375
4	−0.0044021	−0.0029567	−0.0019805	−0.0013237
5	−0.0036033	−0.0023756	−0.0015700	−0.0010367
6	−0.0118308	−0.0082330	−0.0057153	−0.0039588
7	−0.0106521	−0.0072175	−0.0049095	−0.0033408
8	−0.0268217	−0.0191244	−0.0136464	−0.0097328
9	−0.0268094	−0.0183719	−0.0127195	−0.0088460
10	−0.0557869	−0.0398133	−0.0287628	−0.0209019
11	−0.0642873	−0.0428072	−0.0294506	−0.0206098
12	−0.1227568	−0.0814133	−0.0572528	−0.0414122
13	−0.1938141	−0.1056496	−0.0669934	−0.0453215
14	−0.6439281	−0.2134097	−0.1231647	−0.0822898
15		−0.6243396	−0.1959100	−0.1078912
16			−0.6429969	−0.2131694
17				−0.6257461

	$N = 38$ $D = 0.0000010$
0	−0.0000291
1	−0.0000187
2	−0.0002118
3	−0.0001500
4	−0.0008867
5	−0.0006795
6	−0.0027458
7	−0.0022607
8	−0.0069528
9	−0.0061396
10	−0.0152613
11	−0.0145127
12	−0.0304415
13	−0.0316100
14	−0.0585324
15	−0.0691050
16	−0.1236067
17	−0.1974694
18	−0.6423832

VII. ACKNOWLEDGMENT

The authors would like to acknowledge the valuable comments and criticisms provided by J. F. Kaiser.

REFERENCES

1. B. Gold, A. V. Oppenheim, and C. M. Rader, "Theory and Implementation of the Discrete Hilbert Transform," Proc. Symp. Comput. Process. Comm., Polytechnic Press, 1970, pp. 235–250.

2. M. R. Schroeder, J. L. Flanagan, and E. A. Lundry, "Bandwidth Compression of Speech by Analytic-Signal Rooting," Proc. IEEE, *55*, No. 3 (March 1967), pp. 396–401.
3. O. Herrmann, "Transversal Filters for Hilbert Transformation," Arch. Elek. Ubertr., *23*, No. 12 (December 1969), pp. 581–587.
4. J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," IEEE Trans. on Audio and Electroacoustics, *AU-21*, No. 6 (December 1973), pp. 506–526.
5. A. J. Gibbs, "On the Frequency-Domain Responses of Causal Digital Filters," Ph.D. Thesis, Dept. of Electrical Engineering, University of Wisconsin, 1969.
6. J. F. Kaiser, "Digital Filters," chapter 7 in *System Analysis by Digital Computer*, edited by F. F. Kuo and J. F. Kaiser, New York: John Wiley and Sons, 1966.
7. L. R. Rabiner and R. W. Schafer, "On the Behavior of Minimax Relative Error FIR Digital Differentiators," B.S.T.J., this issue, pp. 333–361.

Contributors to This Issue

Jacques A. Arnaud, Dipl. Ing., 1953, Ecole Supérieure d'Electricité, Paris, France; Docteur Ing., 1963, University of Paris; Docteur es Science, 1972, University of Paris; Assistant at E.S.E., 1953–1955; CSF., Centre de Recherche de Corbeville, Orsay, France, 1955–1966; Warnecke Elec. Tubes, Des Plaines, Illinois, 1966–1967; Bell Laboratories, 1967—. At CSF., Mr. Arnaud was engaged in research on high-power traveling-wave tubes and supervised a group working on noise generators. He is currently a supervisor studying microwave quasi-optical devices and the theory of optical wave propagation. Senior Member, IEEE; Member, Optical Society of America.

Marie T. Dolan, B.A., 1954, Montclair State Teachers College; M.S., 1966, Stevens Institute of Technology; Bell Laboratories, 1954—. Ms. Dolan has been a programmer in general mathematical research and in recent years has been concerned with computerized design of digital filters. Member, Kappa Mu Epsilon.

Stanley B. Gershwin, B.S. (Engineering Mathematics), 1966, Columbia University; M.A., 1967, and Ph.D. (Applied Mathematics), 1971, Harvard University; Bell Laboratories, 1970–1971. While at Bell Laboratories, Mr. Gershwin worked on various traffic aspects of the Subscriber Loop Multiplier. Member, IEEE, SIAM, Tau Beta Pi.

Jeremiah F. Hayes, B.E.E., 1956, Manhattan College; M.S., 1961, New York University; Ph.D., 1966, University of California, Berkeley; Faculty Member, Purdue University, 1966–1969; Bell Laboratories, 1969—. Mr. Hayes' current research interests are in the area of signal processing. Member, IEEE, Sigma Xi, Eta Kappa Nu.

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.

James F. Kaiser, E.E., 1952, University of Cincinnati, S.M., 1954, and Sc.D., 1959, Massachusetts Institute of Technology; faculty of the Massachusetts Institute of Technology, 1956–1960; Bell Laboratories, 1959—. Mr. Kaiser has been concerned with problems of data processing, digital filter design, system simulation, and computer graphics. He is coauthor of two books, *Analytical Design of Linear Feedback Controls* with G. C. Newton and L. A. Gould and *System Analysis by Digital Computer* with F. Kuo. Fellow, IEEE; member, Association for Computing Machinery, Society for Industrial and Applied Mathematics, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

Richard V. Laue, B.A. (Mathematics), 1953, Hofstra College; M.S., 1958, and Ph.D. (Applied and Mathematical Statistics), 1966, Rutgers, The State University; Bell Laboratories, 1959—. From 1959 to 1964, Mr. Laue was involved in statistical studies associated with the manufacture and selection of electronic components. Since then, he has been primarily concerned with various statistical aspects of telephone traffic systems. Member, American Statistical Association, Institute of Mathematical Statistics, Sigma Xi.

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 three books. Fellow, IEEE; member, Optical Society of America.

Stewart E. Miller, S.B. and S.M. (Electrical Engineering), 1941, Massachusetts Institute of Technology; Bell Laboratories, 1941—. Mr. Miller was concerned with microwave radar design from 1941 to 1945. At the conclusion of World War II he resumed design work on coaxial-cable carrier systems until 1949, when he joined the Radio Research Department. His work was concerned with circular electric waveguide communication, microwave ferrite devices, and other components for microwave radio systems. As Director, Guided Wave Research, he headed a group which did work leading to a current millimeter-wave system development. More recently, his interest and that of the Guided Wave Research group has shifted to the optical region and to exploration of the use of lasers and associated devices in transmission. Member, Tau Beta Pi and Eta Kappa Nu; associate member, Sigma Xi; Fellow of IEEE; and member, National Academy of Engineering. Recipient, 1972 IEEE Morris N. Liebmann Award.

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.

Ronald W. Schafer, B.S. (E.E.), 1961, and M.S. (E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Laboratories, 1968—. Mr. Schafer has been engaged in research on digital waveform processing techniques and speech communication. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America, and the IEEE G-AU Technical Committees on Digital Signal Processing and Speech Communication.

Eric Wolman, A.B., 1953, A.M., 1954, and Ph.D. (Applied Mathematics), 1957, Harvard University; Bell Laboratories, 1957—. Mr. Wolman has worked on various aspects of traffic flow in communication

systems, and now heads the Network Analysis Department. He is a member of the Evaluation Panel for the Fire Technology Division, National Bureau of Standards, and served as visiting lecturer on applied mathematics at Harvard in 1964. Member, AMS, IEEE, ORSA, Phi Beta Kappa, Sigma Xi, SIAM; Fellow, AAAS.

H. Zucker, Dipl. Ing., 1950, Technische Hochschule, Munich, Germany; M.S.E.E., 1954, Ph.D., 1959, Illinois Institute of Technology; Bell Laboratories, 1964—. Mr. Zucker was concerned with satellite communication antennas, optical resonators, and problems in the areas of electromagnetic theory and optics. More recently, he has been engaged in analytical studies related to transmission systems. Member, IEEE, Commission 6 of URSI, Eta Kappa Nu, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 2, February 1974
Printed in U.S.A.

B.S.T.J. BRIEFS

Interframe *Picturephone*® Coding Using Unconditional Vertical and Temporal Subsampling Techniques

By R. T. BOBILIN

(Manuscript received November 8, 1973)

I. INTRODUCTION

A number of articles[1–4] have described the use of horizontal redundancy removal techniques to reduce the transmission rate required for coded *Picturephone*® signals to 6.312 Mb/s. Here an extension of this work is given which uses unconditional vertical and temporal subsampling techniques to reduce the required transmission capacity to 3 Mb/s. This type of processing is unique in that it does not employ the complex conditional replenishment techniques which typically have been used to reduce the digital transmission rate to either 2 Mb/s[5] or 1.5 Mb/s.[6]

II. SYSTEM DESCRIPTION

The analog *Picturephone* signal has an interlaced frame format[7] as illustrated in Fig. 1. Two adjacent fields, each containing $125\frac{1}{2}$ alternating lines, combine to form a complete 251-line frame of video information. In the experimental system being described, two adjacent fields are "averaged" together at the coder; this averaged $125\frac{1}{2}$ line field is processed using variable-length coding techniques[3] and sent to the decoder at a 30-Hz rate. The decoder processes the received averaged field to form a two-field, 251-line frame for transmission to the station set receiver. Hence, unconditional vertical and temporal subsampling is being used to reduce the digital transmission rate from 6 Mb/s to 3 Mb/s; vertical subsampling because only $125\frac{1}{2}$ lines per frame are transmitted and temporal subsampling because averaged

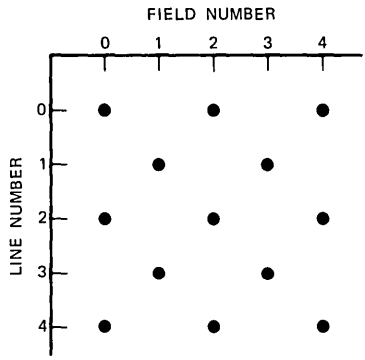

Fig. 1—Interlaced frame format.

fields are only transmitted at a 30-Hz rate instead of the original 60-Hz rate. The basis for this unconditional vertical subsampling is that the vertical spatial dimension of a *Picturephone* signal does not contain 251 lines of detail. There is no vertical aperture correction circuitry in the station set and the actual resolution is much closer to the $125\frac{1}{2}$ lines contained in one field of the interlace pattern; the vertical amplitude response of the station set transmitter varies from 8 to 27 dB down at half the field sampling rate ($125\frac{1}{2}$ video lines/2), depending on the control unit settings at the transmitter. This is not a station set design failure but a requirement to prevent objectionable vertical aliasing effects in the displayed scene. In the temporal dimension the required signal properties are harder to define because the postfiltering action of the human eye dominates system parameters. It is known that a 60-Hz field repetition rate was chosen on the basis of flicker sensitivity, not on the basis of motion rendition, but the application of general eye temporal sensitivity studies to the present *Picturephone* format is of questionable value and leads to inconclusive results. Therefore, an experimental 3-Mb/s codec was built both to verify the vertical processing predictions and to see if 30-Hz transmission is subjectively acceptable in the present *Picturephone* system.

A block diagram of the experimental system is given in Fig. 2. Prior to A/D conversion the analog video signal is deemphasized, equalized, clamped, and filtered with a crispened Gaussian filter that is 20 dB down at 1 MHz. The A/D converter uses a synchronized 1.536-MHz clock to produce an 8-bit PCM word at each sample time. The input to the variable-length coder (VLC) is either the unfiltered (no vertical or temporal filtering) output of the A/D converter or a

filtered version of the same signal. This is done so that an easy AB test can be made to determine the effects of additional vertical and temporal prefiltering. In the unfiltered case the input to the VLC during time i is field i (f_i); in the filtered case the input is the average of field i (f_i) and an estimate of field i based on the information contained in field $i-1$ ($\hat{f}_{i-1}$). If line j is in field i, then the estimate of line j from field $i-1$ is given by:

$$\hat{l}_j = \frac{l_{j-5} - 3l_{j-3} + 10l_{j-1} + 10l_{j+1} - 3l_{j+3} + l_{j+5}}{16}. \qquad (1)$$

$\hat{l}_j$ is an estimate of even fields from odd fields and vice versa. This particular estimator gives adequate filtering performance and is easy to implement. The variable-length coder and decoder are the same as described in Ref. 3 and process either every other field or every other average field. The vertical and temporal postfilter uses the received 30-Hz fields to reconstruct the proper 60-Hz even-odd field pattern for the station set receiver. Again two options were designed for the postfilter. In the vertical dimension both postfilters are the same and estimate an adjacent field using the relationship given in (1); in the

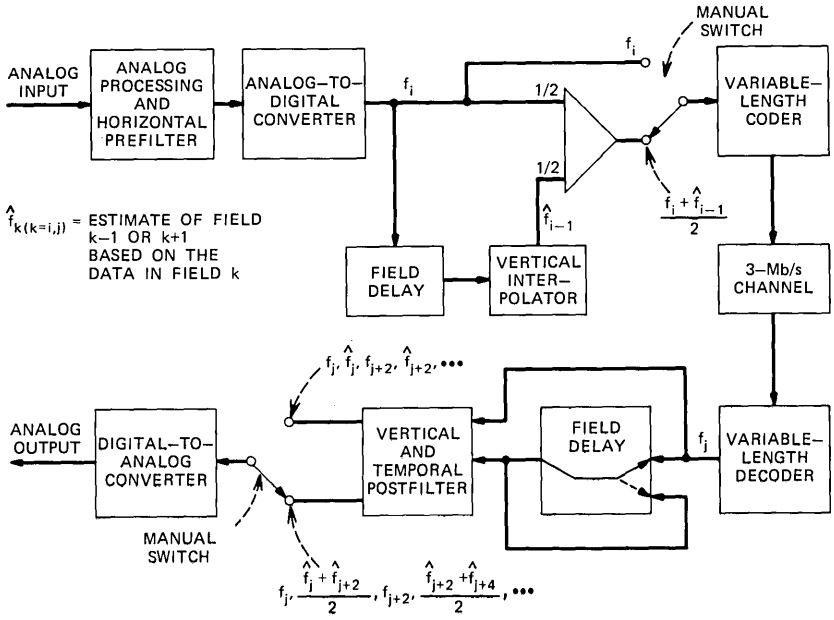

Fig. 2—Experimental system block diagram.

time domain the postfilter using the estimate $\hat{f}_j$ is a sample-and-hold postfilter, whereas the one using the estimate $(\hat{f}_j + \hat{f}_{j+2})/2$ is an interpolating postfilter. The D/A converter turns the 8-bit 1.536-MHz PCM samples into an analog format.

III. SUBJECTIVE RESULTS

When the station set transmitter is operating in its highest resolution condition, with no zoom, and when the coder prefilter is not used, then vertical aliasing is easily seen on eyeglass rims and around the lips. This aliasing can be eliminated either by using the coder vertical prefilter or by placing the station set in a partial-zoom mode. Both changes have the effect of introducing additional vertical prefiltering; the coder vertical prefilter adds 6 dB of vertical filtering at half the sampling rate or $62\frac{3}{4}$ lines per frame and the use of the zoom mode adds filtering proportional to the setting, somewhere between 8 and 27 dB at $62\frac{3}{4}$ lines per frame. The effect of the temporal prefiltering associated with the coder field filtering is insignificant.

The use of the sample-and-hold postfilter (3 dB down at 15 Hz) results in a slight jerkiness of motion (temporal aliasing) which is completely removed by using the interpolating postfilter which is 6 dB down at 15 Hz.

In order to evaluate this 3-Mb/s codec a series of subjective tests were developed. Fifteen people were given an AB test between the analog and the coded video (using field filtering at the coder and the interpolating postfilter at the decoder) and were asked if the impairment added by the codec was:

1. Not noticeable
2. Just noticeable
3. Noticeable but not objectionable
4. Objectionable
5. Extremely objectionable.

In this test each observer was given 15 seconds of analog followed by 15 seconds of the coded scene shown in Fig. 3. During each 15-second interval, Bonnie carried on a normal conversation for the first 5 seconds, moved from side to side for the next 5 seconds, and again carried on a normal conversation for the last 5 seconds. With the station set in its high-resolution mode, no zoom, the average scale rating was 2.8 indicating that for this scene the impairment added by the 3-Mb/s codec was noticeable but not objectionable; with the station set in a partial-zoom mode the average scale rating was 2.3 indicating an

Fig. 3—Bonnie.

impairment that is just noticeable. These comment scale ratings can be compared with a more complete series of tests carried out using the same horizontal processing without any vertical or temporal processing on three different scenes. This resulted in an overall average scale rating of 2.1 for the corresponding 6-Mb/s codec.[4]

IV. CONCLUSIONS

An experimental system has been built showing that unconditional vertical and temporal subsampling techniques can be used on a 6-Mb/s intraframe codec to result in a 3-Mb/s interframe codec. The impairment resulting from this 3-Mb/s codec is rated as being between just noticeable and noticeable but not objectionable. This unconditional alternate field transmission technique can also be used as a higher

activity mode in a conditional replenishment type codec as shown in Ref. 6.

REFERENCES

1. J. B. Millard and H. I. Maunsell, "The *Picturephone*® System: Digital Encoding of the Video Signal," B.S.T.J., *50*, No. 2 (February 1971), pp. 459–479.
2. R. P. Abbott, "A Differential Pulse-Code-Modulation Coder for Videotelephony Using Four Bits Per Sample," IEEE Trans. Commun. Tech., *COM-19*, No. 6, Part 1 (December 1971), pp. 907–912.
3. M. C. Chow, "Variable Length Redundancy Removal Coders for Differentially Coded Video Telephone Signals," IEEE Trans. Commun. Tech., *COM-19*, No. 6, Part 1 (December 1971), pp. 923–926.
4. R. T. Bobilin, "Prefilters, Sampling, and Transmission Rates for Intraframe Codecs for *Picturephone*® Service," B.S.T.J., *52*, No. 4 (April 1973), pp. 497–525.
5. J. C. Candy, Mrs. M. A. Franke, B. G. Haskell, and F. W. Mounts, "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1889–1917.
6. Y. C. Ching, B. Gotz, D. M. Henderson, and J. B. Millard, "Video Processing System (VPS)—A Conditional Replenishment Interframe Codec for the Digital Transmission of *Picturephone*® Signals at the T1 Rate," unpublished work.
7. W. B. Cagle, R. R. Stokes, and B. A. Wright, "The *Picturephone*® System: 2C Video Telephone Station Set," B.S.T.J., *50*, No. 2 (February 1971), pp. 271–312.

Simultaneous Measurements of Depolarization by Rain Using Linear and Circular Polarizations at 18 GHz

By R. A. SEMPLAK

(Manuscript received August 3, 1973)

I. INTRODUCTION

Limitations imposed by attenuation during heavy rain on the reliability of microwave systems are well known[1] and a recent paper[2] discussed observations of depolarization of circular polarization by rain at 18 GHz; it was concluded that depolarization by oblate raindrops poses a serious problem for the use of circular polarization. However, it is desirable that a direct comparison be made by *simultaneous* measurements of linear and circular polarizations on the same propagation path. Continuous measurements have been made during the period June 1972 through April 1973 (a total of 35 rain showers); a discussion of these follows a few remarks on the experimental system.

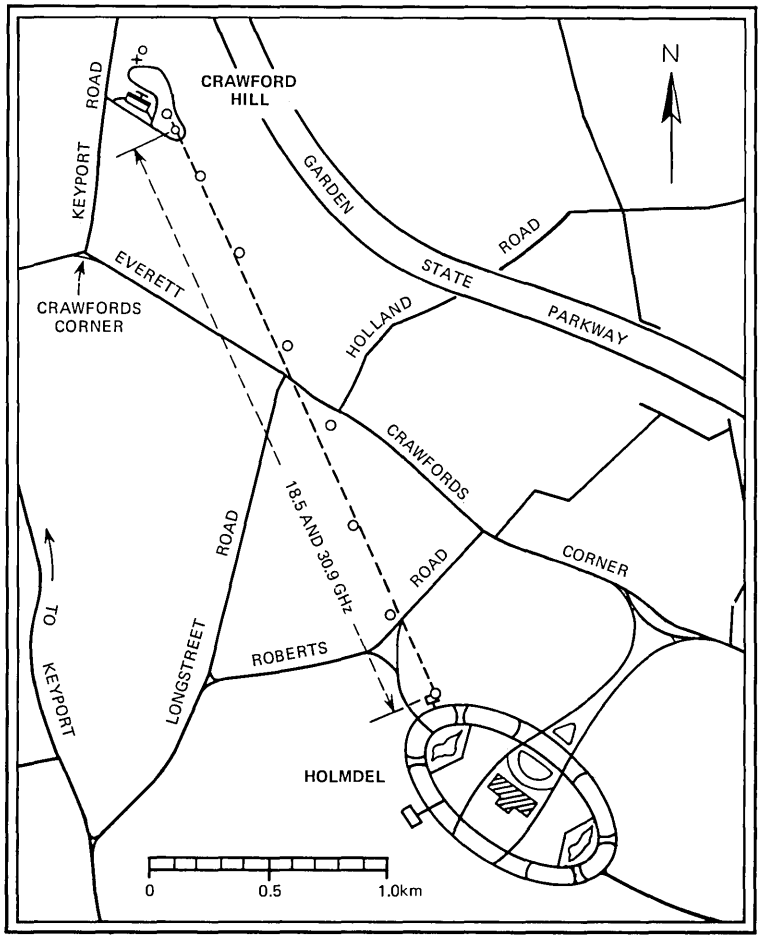

Fig. 1—The 2.6-km propagation path is indicated by the dashed line.

II. EXPERIMENTAL SYSTEM

The 2.6-km propagation path used in the experiment is shown by the dashed line in Fig. 1. Two frequency-swept solid-state transmitters (with separate antennas) are located at the lower right of the path. The circularly polarized transmitter operates at a frequency of 18.65 GHz while the linear horizontally polarized transmitter is at a frequency of 18.35 GHz.

The two receivers (with separate antennas) share a common building at Crawford Hill, Holmdel, New Jersey (upper left of propagation

path, Fig. 1). Each system has a ferrite switch which looks sequentially at the received fields; e.g., in the circularly polarized system, the desired circular polarization and then at the depolarized component. Switching rates are of the order of 17 Hz and occur much faster than the changes in attenuation produced by rain. Strip-chart recordings are made of all four components.

The clear-day discrimination for both the circular and linear systems is more than 32 dB, but none of the depolarization data below −32 dB are included.

III. DISCUSSION OF DATA

The extremes of attenuation in linear (horizontal) and circular polarizations are shown in Fig. 2 for the fades induced by the 35 rain showers that occurred during the period of June 1972 through April 1973; they have similar magnitudes but the attenuation in linear (horizontal) polarization is slightly higher than that for circular polarization.* This is as it should be, for it is known[4] that most storms consist of oblate drops whose major axes are predominantly horizontally aligned, resulting in less attenuation for vertical than for horizontal polarization, whereas in circular polarization the attenuation is something like the average of vertical and horizontal.

Comparison of the linear cross-polarization discrimination (XPD) with the simultaneously measured circular polarization discrimination (CPD) is made in Fig. 3; the curve is the median of the data. The depolarization is much stronger in circular than in linear polarization; for example, the former is −15 dB when the latter is −25 dB.

We pursue the comparison further by examining a particular fade associated with a rain shower that occurred April 1, 1973, as presented in Fig. 4. The set of two curves on the left pertains to circular polarization, the set of two curves on the right to linear polarization. The upper curve in each set is a plot of the rain-induced attenuations. As in Fig. 2 the circular polarization shows a slightly smaller attenuation than that for horizontal linear polarization. For example, the maximum attenuation for circular polarizations was 36 dB while that for linear was 38.5 dB. At the same time, from the lower curve of each set, we see that the circular polarization discrimination was only 8 dB while the cross-polarization discrimination for the linear case was 13.5 dB. Let us examine the measurements at time 19:57; for the circular case,

* The sparseness of attenuations greater than 20 dB is, of course, due to the limited number of heavy rain fades that occur during a year. Figure 3 of Ref. 3 shows that this path has rain-induced attenuations that exceed 20 dB about 20 minutes a year.

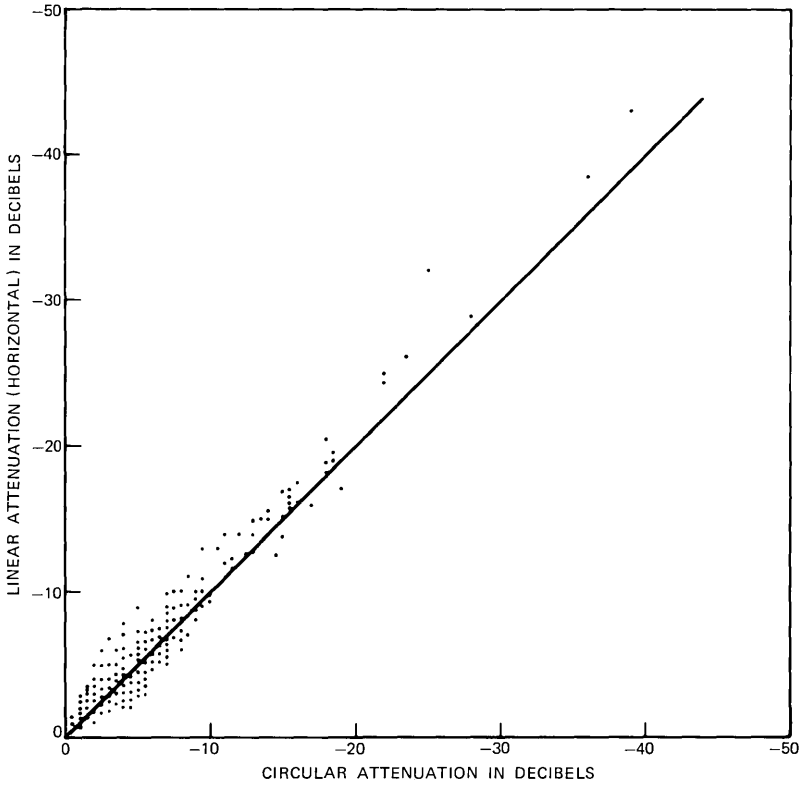

Fig. 2—Data on linear and circular rain-induced attenuation from June 1972 through April 1973.

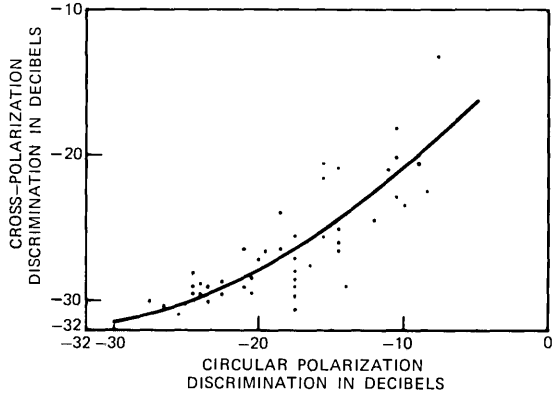

Fig. 3—Comparison of simultaneous measurements of linear polarization discrimination (XPD) and circular polarization discrimination (CPD).

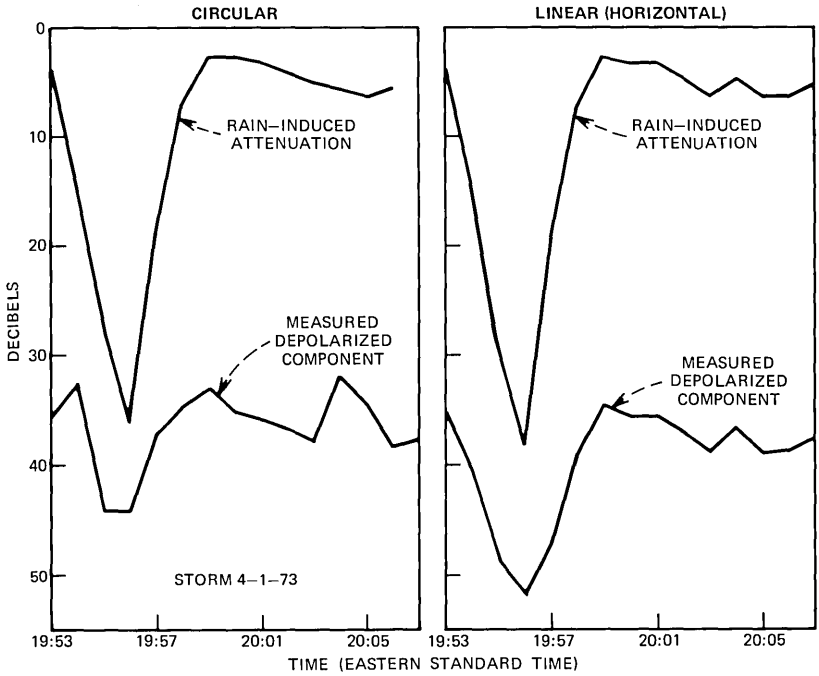

Fig. 4—Rain shower of April 1, 1973. Upper curves are rain-induced attenuation. The lower curves show the level of the depolarized component.

the desired signal was attenuated to a level of −18.5 dB and the depolarized component was only 18.5 dB below that (−37 dB). In the linear case at this same point in time, the rain-induced fade was 19 dB and the XPD was 28 dB. The linear polarization is depolarized significantly less than the circular polarization; therefore, in radio relay systems at frequencies of the order 20 GHz, linear (vertical or horizontal) polarization is believed to be preferable in systems relying on polarization discrimination.

REFERENCES

1. D. C. Hogg, "Statistics of Attenuation of Microwaves by Intense Rain," B.S.T.J., *48*, No. 9 (November 1969), pp. 2949–2962.
2. R. A. Semplak, "The Effect of Rain on Circular Polarization at 18 GHz," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 1029–1031.
3. R. A. Semplak, "Dual Frequency Measurements of Rain-Induced Microwave Attenuation on a 2.6-Kilometer Propagation Path," B.S.T.J., *50*, No. 8 (October 1971), pp. 2599–2606.
4. R. A. Semplak, "Effect of Oblate Raindrops on Attenuation at 30.9 GHz," Radio Science, *5*, No. 3 (March 1970), pp. 559–564.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Computer & Control Abstracts, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electrical & Electronic Abstracts, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Mathematical Reviews, Metals Abstracts, Science Abstracts,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System