THE BELL SYSTEM TECHNICAL JOURNAL

OCTOBER 1982
VOL. 61, NO. 8

The Tap-Leakage Algorithm: An Algorithm for the Stable Operation of a Digitally Implemented, Fractionally Spaced, Adaptive Equalizer 1817
R. D. Gitlin, H. C. Meadors, and S. B. Weinstein

Using Magnetic Bubble Memories to Provide Recorded Announcements 1841
J. E. Rowley and J. Bernardini

Application of Graph Theory to the Solution of a Nonlinear Optimal Assignment Problem 1863
M. Malek-Zavarei

Stochastic Theory of a Data-Handling System With Multiple Sources 1871
D. Anick, D. Mitra, and M. M. Sondhi

Spatial Subsampling in Motion-Compensated Television Coders 1895
J. D. Robbins and A. N. Netravali

The Detection of Long Error Bursts During Transmission of Video Signals 1911
K. Janac and N. J. A. Sloane

Considerations for Single-Mode Fiber Systems 1919
K. Ogawa

Error Probability of Partial-Response Continuous-Phase Modulation With Coherent MSK-Type Receiver, Diversity, and Slow Rayleigh Fading in Gaussian Noise 1933
C.-E. Sundberg

Comparisons on Blocking Probabilities for Regular Series Parallel Channel Graphs 1965
D. Z. Du and F. K. Hwang

On the Physical Limits of Digital Optical Switching and Logic Elements 1975
P. W. Smith

(Contents continued on back cover)

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

D. E. PROCKNOW, *President, Western Electric Company*
I. M. ROSS, *President, Bell Telephone Laboratories, Incorporated*
W. M. ELLINGHAUS, *President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

A. A. PENZIAS, *Chairman*

M. M. BUCHNER, JR.
A. G. CHYNOWETH
R. P. CLAGETT
T. H. CROWLEY
B. P. DONOHUE, III
I. DORROS
R. A. KELLEY
R. W. LUCKY
R. L. MARTIN
J. S. NOWAK
L. SCHENKER
G. SPIRO
J. W. TIMKO

EDITORIAL STAFF

B. G. KING, *Editor*
PIERCE WHEELER, *Managing Editor*
LOUISE S. GOLLER, *Assistant Editor*
H. M. PURVIANCE, *Art Editor*
B. G. GRUBER, *Circulation*

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May–June and July–August combined issues, by the American Telephone and Telegraph Company, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary. Editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, 101 J. F. Kennedy Parkway, Short Hills, N.J. 07078. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

ISSN0005-8580

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 61 October 1982 Number 8

The Tap-Leakage Algorithm: An Algorithm for the Stable Operation of a Digitally Implemented, Fractionally Spaced Adaptive Equalizer

By R. D. GITLIN, H. C. MEADORS, JR., and S. B. WEINSTEIN

(Manuscript received January 19, 1982)

A fractionally spaced equalizer is a nonrecursive adaptive filter whose tap weights are spaced a fraction of a symbol interval apart. Such an equalizer can significantly enhance modem performance in the presence of severe linear distortion, when compared with a conventional synchronous equalizer whose taps are spaced a symbol interval apart. However, a digitally implemented, fractionally spaced equalizer generally will exhibit long-term instability when the conventional tap-adjustment algorithm is used. This occurs because, in contrast to the synchronous equalizer, a fractionally spaced equalizer generally will have many sets of tap values, which result in nearly equal values of mean-squared error (mse). Some of these tap settings—which invariably will be attained because of biases in the digital tap-updating circuitry—are large enough to cause register overflows and consequent performance deterioration. In this paper we report how a simple modification in the tap-adjustment algorithm provides a solution to the above problem. The modified tap-adjustment algorithm prevents the buildup of large coefficient values by systematically "leaking" or decreasing the magnitudes of all the equalizer tap weights. For an experimental modem operating at 9.6 kb/s, it has been demonstrated that the tap-leakage adjustment algorithm prevents the accumulation of large equalizer tap values,

while permitting the full performance gain of a fractionally spaced equalizer to be realized.

I. INTRODUCTION

Fractionally spaced equalizers (FSEs), which are nonrecursive, tapped, delay-line adaptive filters, are currently receiving much attention[1-7] because of the significant performance advantage they provide when compared with a conventional synchronous equalizer. Since an FSE has the capability to adaptively realize the optimum linear receiver, it can greatly improve the performance of a modem in the presence of severe linear distortion. More specifically, significant performance improvements have been observed owing to the ability of an FSE to compensate effectively for delay distortion at the limits of private-line, voice-grade channel conditioning.[8] However, in laboratory experiments with a digitally implemented FSE it was noticed that after an extended period of operation some of the equalizer tap weights would invariably become large, while the mean-squared error (mse) remained at a satisfactory level. The taps generally would become so large that one or more registers, which compute partial sums of the equalizer output, would overflow, and the modem performance was then substantially degraded. This phenomenon is a consequence of the fact that an FSE, in contrast to a conventional synchronous equalizer, generally has many sets of tap values that correspond to roughly the same mse. Included in the set of tap values that correspond to the minimum mse are some tap coefficients of relatively large magnitude. These large tap values can be attained because of the cumulative effect of noise or any bias in the digital circuitry that performs the equalizer updating. Even though the value of the mse is satisfactory, some of these tap values will be large enough to cause occasional overflow of the partial sums computed to form the equalizer output. The purpose of this paper is twofold: to elucidate why a fractionally spaced equalizer has so many apparently "good" sets of tap values, and to indicate how equalizer operation can be stabilized by simply modifying the conventional estimated-gradient tap-adjustment algorithm.

The "almost unique" nature of the fractionally spaced equalizer coefficients is discussed in Section II. In Section III we describe the reasons for the occurrence of large tap values, and in Section IV a modified adjustment algorithm, dubbed the tap-leakage algorithm, is proposed to remedy the observed equalizer instability. The results of laboratory experiments are reported in Section V.

II. DOES A FRACTIONALLY SPACED EQUALIZER HAVE A UNIQUE OPTIMUM SETTING?

2.1 Fractionally spaced equalizers

To answer the question posed by the title of this section, we refer to the simplified baseband data transmission system shown in Fig. 1a.

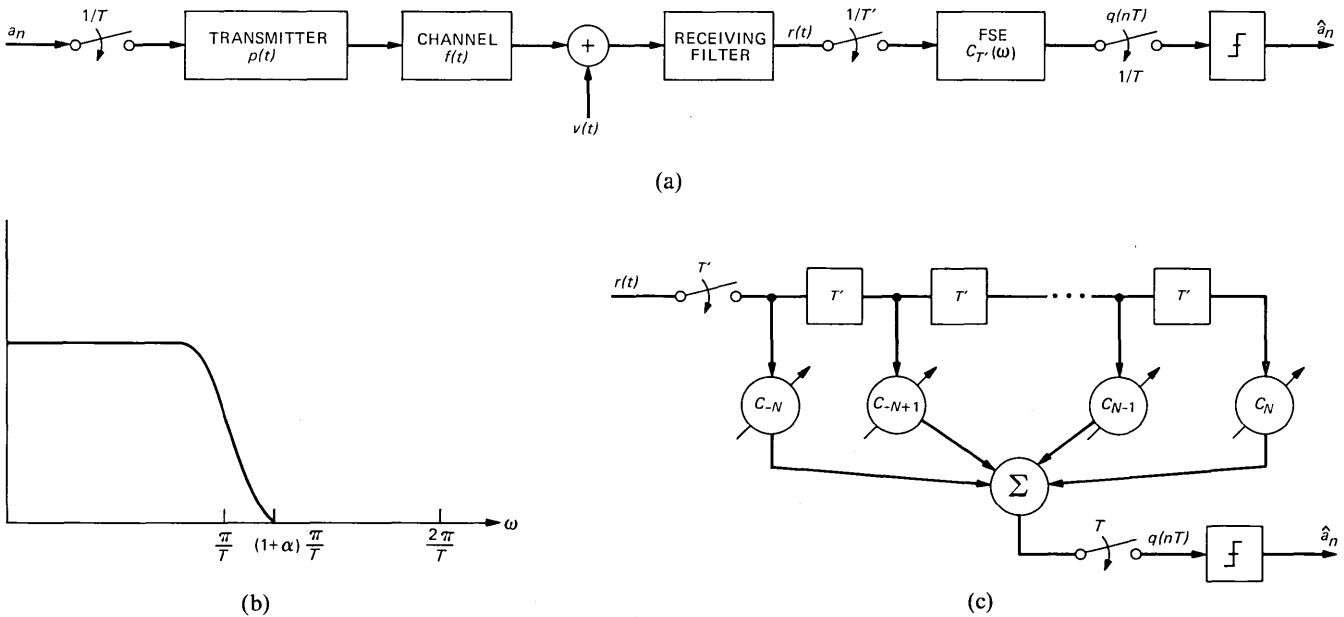


Fig. 1—Simplified (baseband) pulse amplitude modulation (PAM) data transmission system incorporating an FSE. (a) Transmission system. (b) Spectrum of the transmitted pulse. (c) A fractionally spaced equalizer.

For the purposes of exploring the phenomenon of large-tap buildup, this baseband model will suffice. Referring to the figure: $\{a_n\}$ are the discrete-valued multilevel data symbols, $1/T$ is the symbol rate, $p(t)$ is the band-limited transmitter pulse (whose spectrum is shown in Fig. 1b), $f(t)$ is the channel impulse response, and $v(t)$ is the additive background noise. Note that the receiving filter output, $r(t)$, is sampled at the rate $1/T'$, and the samples are then passed through the tapped delay-line equalizer (shown in Fig. 1c) having $(2N + 1)$ delay elements spaced $T'(<T)$ seconds apart and weighting coefficients $\{c_n\}$. The FSE output

$$q(nT) = \sum_{m=-N}^{N} c_m r(nT - mT'), \qquad n = 1, 2, \cdots \tag{1}$$

is computed at the symbol rate and quantized (sliced) to provide the data decision, $\hat{a}_n$. The transmitted pulse spectrum, shown in Fig. 1b, generally will be band-limited to $(1 + \alpha)\pi/T$ radians/second where the rolloff factor, α, varies between 0 and 1. An FSE with tap spacing T' seconds will have a transfer function, C_T, (ω), with period $2\pi/T'$, and, as shown in Fig. 2, if $T' < T/(1 + \alpha)$, the transfer function of the FSE will span the entire spectral range of the transmitted signal. This enables the equalizer to exert complete control over the amplitude and delay distortion present in the region $0 < |\omega| < (1 + \alpha)\pi/T$. Consequently, the FSE can compensate for these distortions directly, rather than filtering the aliased (folded) spectrum, as is done by the conventional equalizer.[9] As an example of this feature, consider the ability of the FSE to compensate for delay distortion by synthesizing the phase characteristic conjugate to that of the received pulse. This operation will leave the noise power at the equalizer output unchanged from the received noise power. The conventional synchronous equalizer, because it has a transfer function that has period $2\pi/T$, can only equalize the folded spectrum, i.e., the compensation characteristics on either

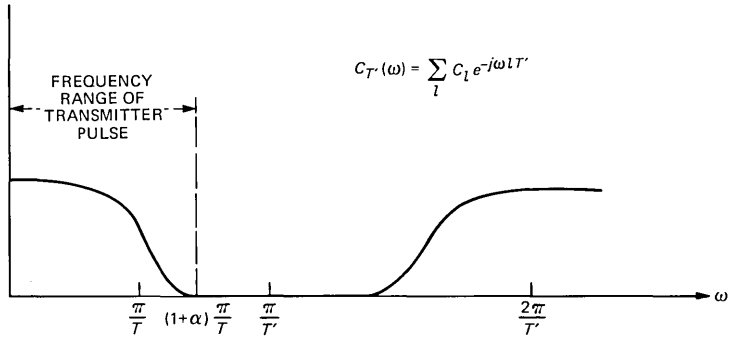


Fig. 2—Transfer function of a fractionally spaced equalizer.

side of π/T are restricted to be the conjugate of one another. If the channel characteristics are such that the folded spectrum has a relative null at a particular frequency, then the noise at the output of the synchronous equalizer can be significantly enhanced and the performance degraded proportionately.

From Fig. 2 it should also be evident that when the noise becomes vanishingly small, there is legitimate concern as to what function(s) the equalizer will synthesize in the region, $(1 + \alpha)\pi/T < |\omega| < 2\pi/T$, where there is no signal energy. It is known[7,8] that for an infinitely long equalizer the transfer characteristic that minimizes the mse is

$$C_T, (\omega) = \frac{X^*(\omega)}{\sum_l \left| X\left[\omega + l\frac{2\pi}{T}\right]\right|^2 + N_0}, \qquad |\omega| \leq \frac{2\pi}{T}, \tag{2}$$

where the asterisk denotes the complex conjugate, $X(\omega)$ is the transfer function of the received pulse,† $x(t)$, presented to the equalizer input, and N_0 is the noise spectral density. As long as $N_0 \neq 0$, the equalizer function is zero whenever the received signal has no power; however, as $N_0 \to 0$ the derivation of (2) is no longer valid, and, moreover, (2) approaches 0/0 in the region $(1 + \alpha)\pi/T < |\omega| < \pi/T$. Clearly, as the noise vanishes, an infinitely long FSE can synthesize the required channel characteristic in the region $0 < |\omega| < (1 + \alpha)\pi/T$, and an arbitrary—and nonunique—characteristic in the remaining frequency band. An interesting question, then, is what happens to the optimum tap setting for a finite length FSE as the noise becomes vanishingly small.

2.2 *Uniqueness of solution for finite length FSE as the noise vanishes*

The equalized mse is defined as

$$E = \langle [q(nT) - a_n]^2 \rangle = \langle e_n^2 \rangle, \tag{3}$$

where the brackets denote the ensemble average with respect to the data symbols and the noise, and e_n is the equalizer output error, $q(nT) - a_n$, at $t = nT$. The mse is readily evaluated as the quadratic form

$$E = \mathbf{c}'A\mathbf{c} - 2\mathbf{c}'\mathbf{x} + \langle a_n^2 \rangle, \tag{4}$$

where the prime denotes the transposed vector, $\mathbf{c}'$ is the tap vector $(c_{-N}, \cdots, c_0, \cdots, c_N)$, $\mathbf{x}'$ is the truncated impulse-response vector $|x(NT'), \cdots, x(-NT')|$, and A is the channel-correlation matrix. More specifically, the channel vector is given by

$$\mathbf{x} = \langle a_n \mathbf{r}_n \rangle, \tag{5}$$

† Thus $x(t)$ is the convolution of the transmitter, channel, and receiver filters.

where the received vector is given by $\mathbf{r}'_n = |r(nT + NT'), \cdots, r(nT), r(nT - NT')|$, and the A matrix is defined by

$$A = \langle \mathbf{r}_n \mathbf{r}'_n \rangle. \tag{6}$$

The klth element of the channel correlation matrix is given by

$$A_{kl} = \sum_{m=-\infty}^{\infty} x(mT - kT')x(mT - lT') + \sigma^2 \delta_{k-l}, \tag{7}$$

where σ^2 is the noise variance and δ_k is the Kronecker delta. Note that A is not a Toeplitz matrix, as it would be for a synchronous equalizer ($T' = T$). If A is nonsingular then the optimum setting and the corresponding minimum mse are obtained from (4) by differentiation, and are given by

$$\mathbf{c}_{\text{opt}} = A^{-1}\mathbf{x} \tag{8a}$$

$$E_{\text{opt}} = 1 - \mathbf{x}'A^{-1}\mathbf{x}, \tag{8b}$$

where $\langle a_n^2 \rangle$ is taken to be unity.

As seen from (7), the matrix A is the sum of two matrices, and as will be evident from the discussion that follows, the channel-dependent component of A is always positive semidefinite. Since the other component of the channel-correlation matrix, $\sigma^2 I$, is positive definite, then A will also be positive definite, and we can conclude that when there is noise present, the optimum tap setting is unique.

We now consider the situation as the noise becomes vanishingly small; clearly, the optimum tap setting will be unique if, and only if, A is nonsingular. A sufficient condition for A to be nonsingular is the nonvanishing of the quatratic form $\mathbf{u}'A\mathbf{u}$, for any *nonzero* test vector $\mathbf{u}$ with components $[u_i]$. Let us consider in detail this quadratic form, which we write from (7) as:

$$\begin{aligned}
\mathbf{u}'A\mathbf{u} &= \sum_{m,n=-N}^{N} u_m A_{mn} u_n \\
&= \sum_{m,n=-N}^{N} u_m u_n \sum_{l=-\infty}^{\infty} x(lT - nT')x(lT - mT') \\
&= \sum_{l=-\infty}^{\infty} \left[\sum_{m=-N}^{N} u_m x(lT - mT') \right]^2 \geq 0.
\end{aligned} \tag{9}$$

The above inequality establishes the positive semidefinite nature of the matrix A, and we see from (9) that $\mathbf{u}'A\mathbf{u}$ can vanish only if*

* The authors gratefully acknowledge discussions with J. E. Mazo that led to this development.

$$\sum_{m=-N}^{N} u_m x(lT - mT') = 0 \qquad l = 0, \pm 1, \pm 2, \cdots . \tag{10}$$

If we define the periodic Fourier transform

$$U'_T(\omega) = \sum_{m=-N}^{N} u_m e^{j\omega mT'}, \qquad |\omega| \le \frac{\pi}{T'}, \tag{11}$$

then we can proceed further by noting that

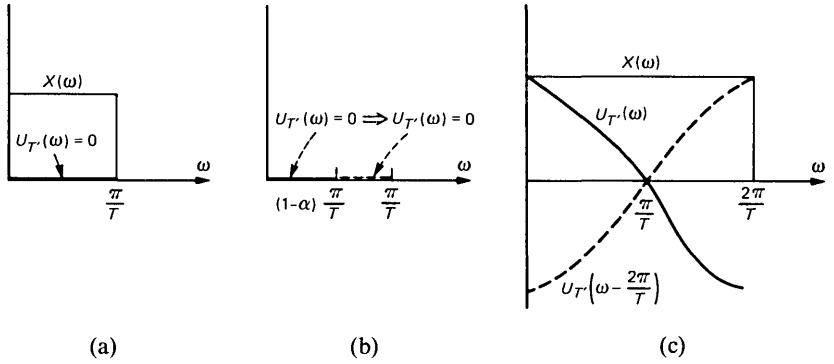
$$\begin{aligned}
\sum_{m=-N}^{N} u_m x(lT - mT') &= \sum_{m=-N}^{N} u_m \int_{-\infty}^{\infty} X(\omega) e^{j\omega(lT - mT')} \frac{d\omega}{2\pi} \\
&= \int_{-\infty}^{\infty} \left[\sum_{m=-N}^{N} u_m e^{-j\omega mT'} \right] X(\omega) e^{-j\omega lT} \frac{d\omega}{2\pi} \\
&= \int_{-\infty}^{\infty} U_{T'}(\omega) X(\omega) e^{-j\omega lT} \frac{d\omega}{2\pi} \\
&= \sum_{k} \int_{(2k-1)\frac{\pi}{T}}^{(2k+1)\frac{\pi}{T}} U_{T'}(\omega) X(\omega) e^{-j\omega lT} \frac{d\omega}{2\pi} \\
&= \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \left[\sum_{k} U_{T'} \left(\omega + \frac{k2\pi}{T} \right) \right. \\
&\qquad \left. \cdot X \left(\omega + \frac{k2\pi}{T} \right) \right] e^{-j\omega lT} \frac{d\omega}{2\pi} .
\end{aligned} \tag{12}$$

The right-hand side of (12) is recognized as the sample, at $t = lT$, of a function whose Fourier transform, $Z_{\text{eq}}(\omega)$, is contained in the brackets. If (12) is to be zero for every value of l, then it must be that the Fourier transform inside the integral vanishes completely, i.e.,

$$Z_{\text{eq}}(\omega) \equiv \sum_{k} U_{T'} \left(\omega + \frac{k2\pi}{T} \right) X \left(\omega + \frac{k2\pi}{T} \right) = 0, \qquad |\omega| \le \frac{\pi}{T}. \tag{13}$$

In Fig. 3a we show the situation when there is no excess bandwidth, and since the sum, (13), reduces to one term, the only way for $Z_{\text{eq}}(\omega) \equiv 0$ is for either $X(\omega) \equiv 0$ or $U_{T'}(\omega) \equiv 0$. Since this implies that $U_{T'}(\omega) \equiv 0$, it would violate the nonzero requirement on $\mathbf{u}$. Thus, we can conclude for this case that A is positive definite. A similar sketch for the less than 100-percent excess bandwidth case is shown in Fig. 3b, where it is noted that only the $k = 0, \pm 1$ terms contribute to the sum, (13). However, in the nonrolloff region, $|\omega| \le (1 - \alpha)\pi/T$, only the $k = 0$ term influences the sum. For channels that do not vanish over the entire nonrolloff region, it is clear that for $Z_{\text{eq}}(\omega)$ to vanish it

CONDITIONS FOR WHICH $Z_{eq}(\omega) = \sum_k U_{T'}\left(\omega + \frac{k2\pi}{T}\right) X\left(\omega + \frac{k2\pi}{T}\right) = 0, (\omega) \leq \frac{\pi}{T}$



(a) (b) (c)

Fig. 3—Sketches associated with eq. (13). (a) $\alpha = 0$. (b) $0 < \alpha < 1$. (c) $\alpha = 1$.

is required that $U_T(\omega)$ vanish at least over the entire nonrolloff region. Since $U_{T'}(\omega)$ is a finite-term Fourier series, it cannot vanish over an interval without vanishing everywhere, which in turn would again make $\mathbf{u} = 0$. Note that if the channel vanished over a portion of the nonrolloff region, then since $Z_{eq}(\omega)$ is a finite-term Fourier series, its energy could not be totally concentrated in the region where there was no channel energy. Thus, the solution still would be unique. It is worth noting that in the extreme case of 100-percent excess bandwidth, $Z_{eq}(\omega)$ can vanish. For example, in Fig. 3c we sketch the situation for a constant $X(\omega)$, and with $U_{T'}(\omega) = \cos \omega T/2$ it is apparent that $Z_{eq}(\omega) \equiv 0$. Thus, for a finite-length FSE with an excess bandwidth of less than 100 percent, we can conclude that even as the noise becomes vanishingly small, the A matrix is nonsingular and there is a unique optimum tap setting.

We digress for a moment to point out that for a finite-length synchronous equalizer where $T' = T$, (13) indicates that since $U_T(\omega + k2\pi/T) = U_T(\omega)$, we can conclude that if the folded channel spectrum does not vanish completely, then there is always a unique tap setting.

III. THE TAP-WANDERING PHENOMENON

3.1 Motivation, background, and infinite-precision considerations

We have shown that for a finite-length fractionally spaced equalizer and the practical range of interest—where the excess bandwidth is on the order of 10 to 50 percent—even with vanishingly small noise there will always be a unique best tap setting. One issue of interest is the

"closeness" (or ill-conditioning) of the A matrix to a singular matrix. This is important for two reasons. First, the distribution of the eigenvalues of A, which is a measure of the ill-conditioning of the matrix, influences the rate of convergence of the equalizer taps to their optimum setting.[10] Second, and more importantly, we observe, from (4) that the contours of equal mse are elliptical, and the eccentricity of these contours is directly related to the eigenvalue distribution. In the appendix it is shown that for an infinitely long equalizer with $T' = T/2$, half the eigenvalues are zero; in Fig. 4 we illustrate some constant mse contours for a finite-length $T/2$ equalizer, whose optimum tap setting is denoted by $\mathbf{c}_{\text{opt}}$. Recall[10] that even with an infinite-precision analog implementation, the use of a finite step-size in the conventional estimated-gradient tap-adjustment algorithm results in a steady-state mse that exceeds* E_{opt}. This is depicted in Fig. 4, where the boldface contour is the mse that can be attained with the chosen step-size. Owing to the random component in the algorithm's correction term, the taps will wander along the constant mse contour, and there will be a certain probability that the taps will become so large that one or more registers will saturate.† Thus, even in an analog implementation, random tap wandering can, in principle, lead to degraded performance.

3.2 A model for tap drifting in digital equalizer

It has been observed in laboratory experiments with a digitally implemented FSE, that under control of the conventional estimated-gradient tap adjustment algorithm,[10]

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha[e_n\mathbf{r}_n], \qquad n = 1, 2, 3, \cdots \tag{14}$$

the equalizer taps inevitably drift close to the shaded (large tap) region of Fig. 4. In (14), $\mathbf{c}_n$ is the tap vector at $t = nT$, α is a positive value called the step-size, which influences both the convergence rate of the equalizer and the steady-state mse, and the brackets around $e_n\mathbf{r}_n$ indicate that this increment is quantized to a specified number of bits. The term $[e_n\mathbf{r}_n]$ will have a deterministic component proportional to the desired gradient, and a random component owing to both the manner in which the digital quantization is performed and the influence of the noise and data-dependent terms. Generally, $[e_n\mathbf{r}_n]$ will also possess a deterministic component owing to bias inevitably present in a digital implementation. A typical mechanism for such a bias is the two's complement type of quantizing characteristic shown in Fig. 5. To quantify our discussion, we denote the bias by a time-invariant

* The mse E_{opt} is achieved when the taps are at their optimum values, $\mathbf{c}_{\text{opt}}$.

† For a synchronous equalizer the ellipses will not be very eccentric, and the tap wandering will not do any damage.

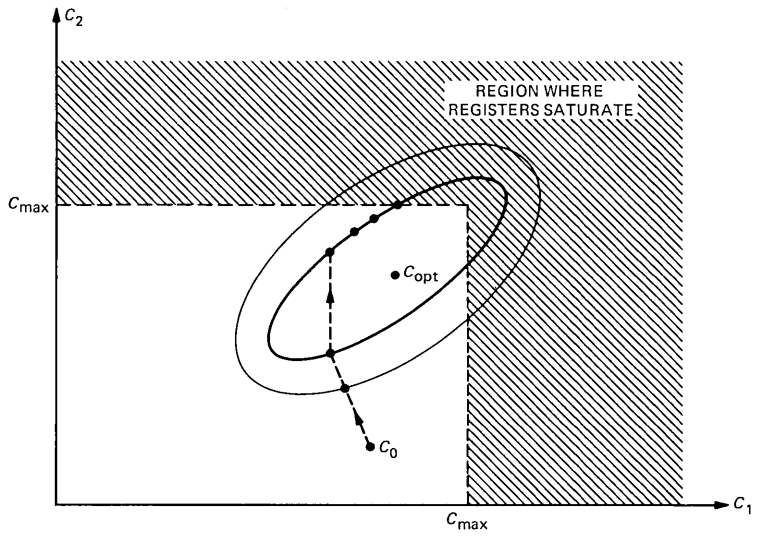


Fig. 4—Contours of equal mse and tap convergence.

vector, **b**, and model the adjustment algorithm as

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha(e_n\mathbf{r}_n + \mathbf{b}), \tag{15}$$

where the bias vector has equal components, and b has a magnitude of less than half a quantization interval. Equation (15) ignores, except for the bias, the effect of limited precision on the algorithm.* Since it has been observed in the laboratory that tap wandering results in a systematic buildup of some tap values, the model expressed by (15) should be useful in relating the magnitude of the bias to the other system parameters. It is bias component that can drive, in a *deterministic* manner,† the tap vector towards the tap region corresponding to large tap values. Since the equalizer output is formed as a series of partial sums, of the form $\sum_m c_m r_{n-m}$, it is clear that large-tap values can lead either to an overflow of a partial sum or saturation of a tap. As the taps grow, occasional register overflows begin to occur, resulting in noise-like "hits" on the equalizer output. The occurrence of such a "hit" is a function of the specific pattern of data samples contained in the equalizer. Continued growth of the taps increases the frequency of the "hits" as more data patterns can produce these events. The error rate can become very high, relative to what constitutes acceptable performance, but the frequency of occurrence of "hits" is still low

* Reference 10 discusses the effect of limited precision on the mse or an FSE in the absence of a bias term. Note that the bias as "seen" by the taps is αb.

† As opposed to the random wandering associated with the self-noise of the algorithm.

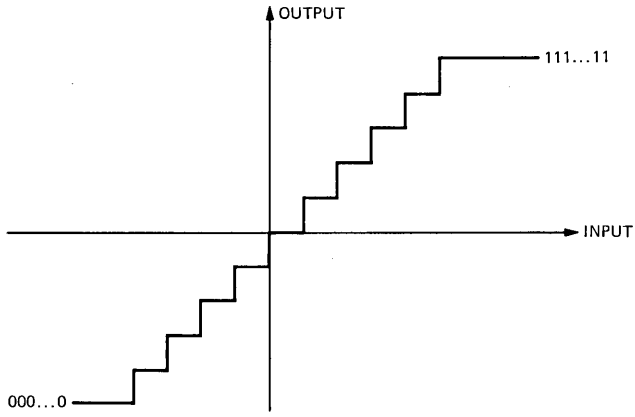


Fig. 5—Quantizing characteristic that produces a bias in the tap-adjustment algorithm.

enough that the mse is almost unaffected. Growth continues until the tap coefficients themselves begin to saturate. At this point severe degradation occurs as the degrees of freedom of the system are reduced. In fact, with an experimental digitally implemented FSE, with a symbol rate of 2400 symbols/second, overflows typically begin to occur within several minutes of operation. The overflows produce noise-like hits and a degraded equalizer output.

3.3 The mean tap error and the mean-squared error

To assess the effects of the bias quantitatively we first define the tap-error vector,

$$\boldsymbol{\epsilon}_n \equiv \mathbf{c}_n - \mathbf{c}_{\text{opt}}, \tag{16}$$

and then use the model described by (15) to write the tap-error evolution as

$$\boldsymbol{\epsilon}_{n+1} = \boldsymbol{\epsilon}_n - \alpha(e_n \mathbf{r}_n + \mathbf{b}). \tag{17}$$

The mse at the nth iteration,* E_n, is given by

$$E_n = E_{\text{opt}} + \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle. \tag{18}$$

Our intent is to study the excess mse,

$$q_n \equiv \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle, \tag{19}$$

and the mean tap error vector, $\langle \boldsymbol{\epsilon}_n \rangle$, in the presence of the bias, $\mathbf{b}$.

* In arriving at (18), we assume that iterations are infrequent enough so that successive vectors $\{\mathbf{r}_n\}$ are independent. In practice, adjustments are generally made at the symbol rate, and the algorithm is observed to behave as if the $\{\mathbf{r}_n\}$ were independent. This phenomenon is discussed in Ref. 11.

From (17) we have

$$\langle \boldsymbol{\epsilon}_{n+1} \rangle = \langle \boldsymbol{\epsilon}_n \rangle - \alpha \langle e_n \mathbf{r}_n \rangle - \alpha \mathbf{b} = (I - \alpha A) \langle \boldsymbol{\epsilon}_n \rangle - \alpha \mathbf{b}, \tag{20}$$

and thus the steady-state mean tap error satisfies

$$\langle \boldsymbol{\epsilon} \rangle = A^{-1} \mathbf{b}. \tag{21}$$

If λ_i and $\mathbf{p}_i$, respectively, denote the ith eigenvalue and eigenvector of A, then

$$\langle \boldsymbol{\epsilon} \rangle = \sum_{-N}^{N} \frac{\mathbf{p}_i' \mathbf{b}}{\lambda_i} \mathbf{p}_i . \tag{22}$$

Clearly, if there is a small eigenvalue whose eigenvector is *not* orthogonal to $\mathbf{b}$, then the steady-state tap error can be quite large. It is interesting to note that for a small number of taps, one would expect that the eigenvalues not be small, i.e., the equalizer would not possess enough degrees of freedom to realize a somewhat arbitrary transfer function beyond the rolloff region. Consider the one-tap equalizer (or automatic gain control) where $A = \langle r^2(nT) \rangle$, and consequently

$$\langle \boldsymbol{\epsilon} \rangle = \frac{\mathbf{b}}{\langle r^2(nT) \rangle}. \tag{23}$$

Thus, for a one-tap equalizer, the tap error is directly proportional to the magnitude of the bias, and the buildup of a large tap value is prohibited. However, in the limit as the number of taps becomes infinite, it is shown in the appendix that with $T' = T/2$, half the eigenvalues are zero, while the other half tend to uniformly sample the aliased (with respect to the symbol rate) squared magnitude of the channel transfer function. Moreover, the eigenvectors corresponding to the zero eigenvalues have most of their energy concentrated near $1/T$ Hz (and are thus close to being orthogonal to $\mathbf{b}$), while the ith eigenvector corresponding to the nonzero eigenvalues approaches a sinusoid of radian frequency $\omega_i = i/N\ \pi/T$. For practical, finite-length equalizers, these limiting conditions will only be approximated, and there will be small eigenvalues whose corresponding eigenvector is not orthogonal to $\mathbf{b}$. Consequently, $\langle \boldsymbol{\epsilon} \rangle$ can become as large as the largest ratio $[\mathbf{p}_i' \mathbf{b}]/\lambda_i$, and the steady-state tap error would then be biased away from the optimum value.

We now discuss the effect of the bias term on the equalized mse,

$$E_n = E_{\text{opt}} + \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle , \tag{24}$$

where the tap error evolves according to (17). In particular we will examine the size of the residual mse,

$$q_n \equiv \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle . \tag{25}$$

An exact analysis (or tight bounds) of the behavior of q_n is an extremely difficult problem; however, by assuming that the sequence $\{\mathbf{r}_n\}$ is independent,[11] and that when the taps are at or near their optimum settings, the squared output error is relatively insensitive to the transmitted data pattern,* it is possible to establish simple, but useful, relationships between the relevant system parameters. From (17) we have that

$$q_{n+1} = \langle \boldsymbol{\epsilon}'_{n+1} A \boldsymbol{\epsilon}_{n+1} \rangle = \langle [\boldsymbol{\epsilon}'_n - \alpha(e_n \mathbf{r}'_n + \mathbf{b}')] A [\boldsymbol{\epsilon}_n - \alpha(e_n \mathbf{r}_n + \mathbf{b})] \rangle, \quad (26)$$

and by using the above assumptions we have (see Ref. 10)

$$q_{N+1} \cong [1 - 2\alpha\bar{\lambda} + \alpha^2 \lambda_M (2N+1)\langle r_n^2 \rangle] q_n + \alpha^2 \lambda_M (2N+1)(\langle r_n^2 \rangle E_{\text{opt}} + b^2), \quad (27)$$

where λ_M is the maximum eigenvalue of A, and where

$$\bar{\lambda} = \frac{1}{2N+1} \sum_{-N}^{N} \lambda_i \quad (28)$$

is the average eigenvalue. Thus, the steady-state fluctuation about the minimum mse is

$$q_\infty \cong \frac{\alpha \lambda_M (2N+1)[\langle r^2(nT) \rangle E_{\text{opt}} + b^2]}{2\bar{\lambda} - \alpha \lambda_M (2N+1)\langle r^2(nT) \rangle}. \quad (29)$$

To assess the effect of the bias on q_∞, we note that the bias "seen" by a tap component, αb, will be approximately $2^{-B} C_{\max}$, where B is the number of bits used to represent the tap weights and $C_{\max}$ is the maximum tap value. If the equalized signal is assumed to have unity power, then $C_{\max} \approx 1/[(2N+1)\langle r_n^2 \rangle]^{1/2}$ and α is typically[10] on the order of $1/|(2N+1)\cdot\langle r_n^2 \rangle|$. Thus, $b^2/|r^2(nT)|$ will be on the order of $2^{-2B}(2N+1)$, and when the equalized output signal power is unity, E_{opt} is roughly the inverse of the output signal-to-noise ratio. With typical parameters like $(2N+1) = 60$, $B = 12$, and $E_{\text{opt}} = 0.001$, it is clear that the effect of the bias on q_∞ is negligible. Thus, owing to the quantizing bias, there can be, on the average, a buildup of one or more large-tap weights, while the mse is relatively unaffected. In other words, under the influence of a bias the taps would still remain on the boldface mse contour of Fig. 4, but would spend most of the time near the shaded region, and the system would be subject to random overflows, or hits. This phenomenon has been repeatedly observed experimentally, and in the next section we will describe a very simple means of controlling the tap wandering.

* When the taps are at their optimum values, the error is known to be uncorrelated with the received samples.

IV. THE TAP-LEAKAGE EQUALIZER ADJUSTMENT ALGORITHM

As we have discussed in the previous section, some or all of the tap weights in an FSE can reach unacceptably large values when the conventional tap adjustment algorithm, (14), is used. A simple means of controlling large-tap buildup is by minimizing either of the augmented cost functions

$$J_1 = E + \mu \sum_{i=-N}^{N} c_i^2, \tag{30a}$$

$$J_2 = E + \mu \sum_{i=-N}^{N} |c_i|, \tag{30b}$$

where, as in (3), E is the mse and μ is a suitably chosen (small) constant. The cost function J_1 ascribes a quadratic penalty to the magnitude of the tap vector, while J_2 provides a magnitude penalty, i.e., the cost function is penalized whenever the tap vector builds up excessively. Since the taps are to be adjusted adaptively, we cannot interpret μ as a Lagrange multiplier. The use of a Lagrange multiplier would be appropriate if we were actually able to minimize J in a deterministic manner by using the true gradient. However, since the gradient of E with respect to $\mathbf{c}$, $A\mathbf{c} - \mathbf{x}$, is not available, we must implement a stochastic algorithm analogous to (14). Thus, μ must be chosen beforehand by using some prior knowledge of the system parameters.

4.1 Increased steady-state mse: true-gradient algorithm

As a preliminary calculation, let us first consider the degradation in the minimum attainable steady-state mse caused by choosing $\mathbf{c}$ to minimize J_1 instead of E. Note that, for the moment, we are neglecting the bias and only assessing the increased mse caused by minimizing the augmented cost function, J_1, via the true-gradient algorithm. From (4) we observe, for binary transmission, that the taps will attempt to minimize the modified criterion:

$$J_1 = \mathbf{c}'(A + \mu I)\mathbf{c} - 2\mathbf{c}'\mathbf{x} + 1. \tag{31}$$

$$= \mathbf{c}'B\mathbf{c} - 2\mathbf{c}'\mathbf{x} + 1, \tag{32}$$

where $B = A + \mu I$. The matrix B has the same eigenvectors as A, while the eigenvalues of B are $\lambda + \mu$. Note that the contours of equal values of J_1 are still ellipses but the maximum-to-minimum eigenvalue ratio governing the tap wandering is now $(\lambda_{max} + \mu)/(\lambda_{min} + \mu)$, where λ_{max} and λ_{min} are the maximum and minimum eigenvalues of A, respectively. Thus, by choosing μ properly, the eccentricity can be controlled, and the equalizer tap vector is now determined as if the noise power were increased from σ^2 to $\sigma^2 + \mu$. Of course, the use of a

tap vector selected on the basis of the pseudo-noise power, $\sigma^2 + \mu$, will increase the steady-state mse.* Note that the steady-state tap vector will now satisfy

$$B\mathbf{c} = \mathbf{x}$$

or

$$\mathbf{c}(\mu) = B^{-1}\mathbf{x} = (A + \mu I)^{-1}\mathbf{x}, \tag{33}$$

where $\mathbf{c}(\mu)$ denotes the steady-state tap vector corresponding to the chosen value of μ. Thus, the minimum attainable mse is given by

$$E(\mu) = \mathbf{c}'(\mu)A\mathbf{c}(\mu) - 2\mathbf{c}'(\mu)\mathbf{x} + 1, \tag{34}$$

and the increased mse, $E(\mu) - E_{\text{opt}}$, is

$$E(\mu) - E_{\text{opt}} = [\mathbf{c}(\mu) - \mathbf{c}_{\text{opt}}]'A[\mathbf{c}(\mu) - \mathbf{c}_{\text{opt}}]. \tag{35}$$

To make a more detailed evaluation of the increase in mse, we let

$$\epsilon(\mu) = \mathbf{c}(\mu) - \mathbf{c}_{\text{opt}} \tag{36}$$

denote the tap-error vector. Recalling the diagonalization

$$A = \sum_{i=-N}^{N} \lambda_i \mathbf{p}_i \mathbf{p}_i', \tag{37}$$

and the fact that inverse matrices have the same eigenvectors and inverse eigenvalues, we find that

$$\begin{aligned}\epsilon(\mu) = \mathbf{c}(\mu) - \mathbf{c}_{\text{opt}} &= [(A + \mu I)^{-1} - A^{-1}]\mathbf{x} \\ &= \sum_i \left(\frac{1}{\lambda_i + \mu} - \frac{1}{\lambda_i}\right) \mathbf{p}_i \mathbf{p}_i' \mathbf{x} \\ &= -\sum_i \frac{\mu}{\lambda_i(\lambda_i + \mu)} \mathbf{p}_i' \mathbf{x} \cdot \mathbf{p}_i. \end{aligned} \tag{38}$$

Substituting (38) into (35), and using the orthogonality property of distinct eigenvectors, we find that

$$E(\mu) - E_{\text{opt}} = \mu^2 \sum_{i=-N}^{N} (\mathbf{p}_i'\mathbf{x})^2 \frac{1}{\lambda_i(\lambda_i + \mu)^2}. \tag{39}$$

Thus, to a first approximation, the increased mse grows only as the square of the leakage parameter, μ, while the eigenvalue distribution—and the range in which the taps can wander—can be favorably altered, in a significant manner, by using even a very small value of μ.

---

* The fluctuation about the minimum mse, caused by the finite step size, will also be examined.

4.2 The adaptive tap-leakage algorithm

In a manner analogous to the commonly used estimated gradient algorithm, the adaptive tap-leakage algorithms are constructed, from (30a), by minimizing the augmented instantaneous squared error, $e_n^2 + \mu \mathbf{c}_n' \mathbf{c}_n$, and, from (30b), by minimizing $e_n^2 + \mu \sum_{i=-N}^{N} |c_n^{(i)}|$. The first algorithm modifies the gradient by a term proportional to the tap vector itself, giving the algorithm

$$\begin{aligned} \mathbf{c}_{n+1} &= \mathbf{c}_n - \alpha[e_n \mathbf{r}_n + \mu \mathbf{c}_n] \\ &= (1 - \alpha\mu)\mathbf{c}_n - \alpha e_n \mathbf{r}_n, \end{aligned} \tag{40a}$$

while the second algorithm is of the form

$$\begin{aligned} \mathbf{c}_{n+1} &= \mathbf{c}_n - \alpha(e_n \mathbf{r}_n + \mu \operatorname{sgn} \mathbf{c}_n) \\ &= \mathbf{c}_n - \alpha\mu \operatorname{sgn} \mathbf{c}_n - \alpha e_n \mathbf{r}_n, \end{aligned} \tag{40b}$$

where the sgn operation is applied individually to each component of the tap vector. Note that from an implementation point of view, the algorithm can be modified with almost no hardware change other than applying a systematic decrement to the magnitude of each tap. The second algorithm (40b) has the practical advantage that adjustments will continue to be made no matter how small any tap weight becomes, while the first algorithm has the "advantage" of analytical tractability.

Consider now the mean tap error when the leakage algorithm (40a) is used. In the presence of digital bias, the algorithm is modeled as

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha[e_n \mathbf{r}_n + \mathbf{b} + \mu \mathbf{c}_n]. \tag{41}$$

Subtracting $\mathbf{c}_{\text{opt}}$ from both sides of (41), and solving for the steady-state average tap error we find

$$\begin{aligned} \langle \boldsymbol{\epsilon} \rangle &= (A + \mu I)^{-1}\mathbf{b} + (A + \mu I)^{-1}\mu A^{-1}\mathbf{x} \\ &= \sum_{i=-N}^{N} \frac{\mathbf{p}_i'\mathbf{b}}{\lambda_i + \mu} \mathbf{p}_i + \mu \sum_{i=-N}^{N} \frac{\mathbf{p}_i'\mathbf{x}}{\lambda_i(\lambda_i + \mu)} \mathbf{p}_i. \end{aligned} \tag{42}$$

The first term, on the right-hand side of (42), is similar to (22), but note that the eigenvalues have been modified to eliminate the catastrophic effects that can accompany vanishingly small eigenvalues. The second term, which is proportional to the leakage parameter, is similar to (39), and represents the increased tap error caused by the minimization of J_1 and not E. The leakage parameter, μ, must be chosen sufficiently large so that the first term is properly controlled in magnitude, but not so large that the magnitude of the second term becomes appreciable. In general, the choice of μ is best done empirically, but if μ is chosen to be in the range where $\lambda_{\text{minimum}} + \mu \approx \mu$, then equating the magnitude of the two terms in (42) gives

$$\mu \approx \frac{\sum_{i=-N}^{N} (\mathbf{p}_i'\mathbf{b})^2}{\sum_{i=-N}^{N} \frac{(\mathbf{p}_i'\mathbf{x})^2}{\lambda_i^2}}. \tag{43}$$

For any given channel, (43) can be evaluated, but as far as the average tap error is concerned, it is sufficient to choose μ in such a way that the magnitude of $\langle \epsilon \rangle$ is within the nonshaded region of Fig. 4. From (42), one reasonable choice is to make μ roughly equal to the smallest value of λ_i for which $\mathbf{p}_i'\mathbf{x}$ is significantly larger than λ_i.

To assess the effect of the tap-leakage algorithm, (40a), on the steady-state mse we recall eq. (18)

$$E_n = E_{\text{opt}} + q_n,$$

where E_{opt} is the minimum mse when both $\mathbf{b}$ and μ are zero. From (41) we can compute an approximation to the steady-state value, q_∞, in a manner similar to the computation of (28). For the adaptive tap-leakage algorithm we find that the fluctuation about the minimum mse is

$$q_\infty \approx \frac{\alpha\lambda_M[(2N+1)\cdot\langle r^2(nT)\rangle E_{\text{opt}} + \mathbf{b}'\mathbf{b}] + 2\alpha\mu\mathbf{b}'\mathbf{x} + \mu^2\mathbf{x}'A^{-1}\mathbf{x}}{2\bar{\lambda} + \mu(1-\alpha\bar{\lambda}) - \alpha[\lambda_M(2N+1)\langle r^2(nT)\rangle + \mu^2]}. \tag{44}$$

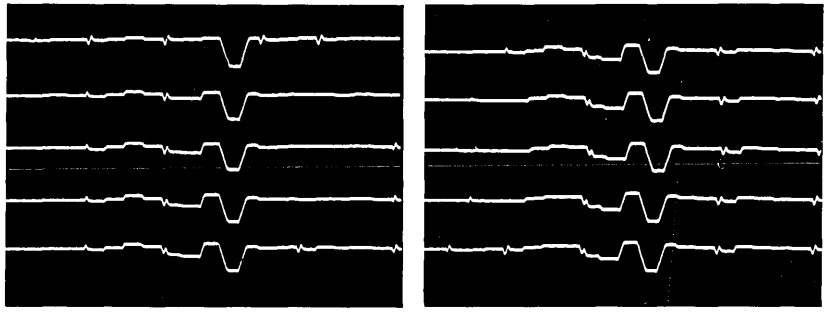
Obviously, (44) is very channel dependent, but when μ is chosen on the order of a small eigenvalue, then the fluctuation about the minimum mse is a rather insensitive function of the leakage parameter, while the magnitude of the mean tap error, (42), can be effectively controlled by the proper choice of μ.

In the next section we will discuss the results of laboratory experiments that use the tap-leakage algorithm to control the potentially unstable operation of a fractionally spaced equalizer.

V. LABORATORY EXPERIMENTS

For an experimental 9.6-kb/s data transmission system using 16-point quadrature amplitude modulation, the tendency of the tap coefficients of a digitally implemented equalizer with $T/2$ sample spacing to drift is demonstrated by the waveforms of Figs. 6a and 6b. These waveforms are analog representations of the values of one component of a set of complex tap coefficients associated with data samples taken $T(=1/2400)$ second apart. The equalizer is physically constructed by using two conventional T-spaced interleaved structures, requiring four tap coefficient component distributions to totally describe the state of the equalizer.

The equalizer goes through a conventional start-up procedure, except that timing recovery and carrier phase adjustments are suppressed



(a) (b)

Fig. 6—Tap buildup as a function of time. (a) Three-minute intervals. (b) Five-minute intervals.

so that the distribution of tap coefficients will not change through interaction with either of these operations. The transmitted signal is fed back to the receiver, after passing through appropriate attenuators, thus avoiding any time varying channel characteristics, and assuring a low noise environment.

The first trace of Fig. 6a illustrates the distribution of components among the particular collection of coefficients immediately after start-up. A single large negative component is noted, with all other components relatively insignificant. The coefficients are updated in the usual fashion, via (14), without the addition of a tap-leakage adjustment. Subsequent traces in Fig. 6a are taken at 3-minute intervals. The traces of Fig. 6b are a continuation of Fig. 6a with the separation in time extended to five minutes. A clear pattern of buildup in the amplitude of the taps, particularly those immediately preceding the original dominant tap, is demonstrated. A similar compensatory buildup occurs among those tap coefficient components not displayed, such that the mse is essentially unchanged over the duration of the test. This deterministic growth of tap amplitudes will eventually lead to saturation of shift register accumulators used in forming the various components of the equalizer passband outputs. The observed output signal constellation will display frequent apparent noise-like hits of large amplitude, and an unacceptable output error rate results.

The tap-coefficient components in the laboratory configuration are stored in 24-bit shift registers. The 12 most significant bits are used in the multiplication to form tap-product outputs. The remaining bits were to average out the effects of tap updating, which is normally done according to the rule

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha(e_n \mathbf{r}_n).$$

The components of the updating quantity $-\alpha(e_n\mathbf{r}_n)$ are stored in 12-bit

words which are added to the 20 most significant bits of the coefficients during the normal steady-state mode of operation. To counteract bias in the arithmetic, the updating is changed to the tap-leakage algorithm

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \beta \operatorname{sgn} \mathbf{c}_n - \alpha(e_n \mathbf{r}_n), \tag{45}$$

where $\beta = \alpha\mu$ of (40b). In the experimental setup, a count of 1 is added or subtracted to the 23rd most significant bit of each component of each tap coefficient once each symbol interval. In steady-state operation, the 12-bit updating signal will typically show activity in a minimum of the five least significant bits. In general, therefore, the leakage term is quite small compared with the conventional updating term.

The effect of introducing this leakage is shown in the waveforms of Fig. 7. The top trace shows the coefficient distribution 40 minutes after initial start-up, at the time the leakage is enabled. Subsequent traces were taken at 30-second intervals. Within two minutes the coefficient components had been virtually restored to the state that existed immediately after start-up.

The experimental arrangement allows the leakage to be scaled over a wide range. Viewing the 24-bit coefficient as an integer, the leakage increment can be made 2^r, $r = 0, 1, \cdots, 7$ ($r = 1$ is the case displayed). Since the coefficients are chosen to span an analog range of ± 4, this corresponds to β as defined in (45) ranging from 2^{-21} to 2^{-14}, with $\beta = 2^{-20}$ displayed. The value of α used in the experiment is $\alpha = 2^{-11}$. It is observed that $\beta = 2^{-21}$, the lowest possible level of continuous leakage, is adequate to suppress tap drift, indicating the extreme low level of the system bias to which the FSE updating algorithm appears susceptible.

The larger the value of β that is chosen, the more the leakage will degrade the equalizer performance, although the degradation is negligible for β less than 2^{-17}. In normal data-set operation, it is observed

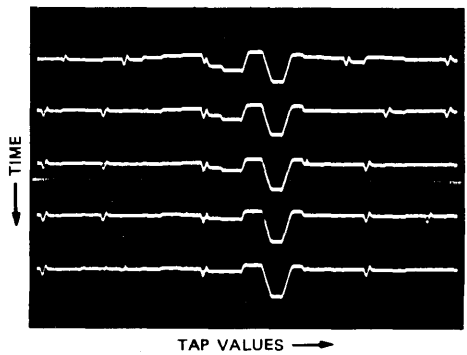


Fig. 7—The effect of the tap-leakage algorithm (30-second intervals between traces).

that a substantially rapid shift in the sampling epoch, which may occur during the timing recovery operation, will greatly accelerate the buildup of tap coefficients. The choice of $\beta = 2^{-18}$ will allow rapid response to this situation without perceivable performance degradation.

VI. CONCLUSIONS

Effective control of tap drifting for a fractionally spaced equalizer, at a 9.6-kb/s data rate, has been demonstrated by employing the easily implemented tap-leakage algorithm. The tap-leakage algorithm, or some variation of it, might be appropriate for any digitally implemented adaptive system that has too many degrees of freedom and that exhibits coefficient wandering.

APPENDIX

Asymptotic Distribution of the Eigenvalues and Eigenvectors for Synchronous and Fractionally Spaced Equalizers

In this appendix we describe the eigenvalues and eigenvectors of infinitely long synchronous and fractionally spaced equalizers.

A.1 Synchronous equalizer

First we recall the eigenvalues and eigenvectors of an infinitely long synchronous equalizer. From (7) we have the eigenvalue equation

$$\sum_{l=-N}^{N} A_{k-l} p_l = \lambda p_k, \qquad N \le k \le N, \tag{46}$$

where λ is an eigenvalue and $\mathbf{p}' = (p_{-N}, \cdots, p_0, \cdots, p_N)$ is the associated eigenvector. As $N \to \infty$, taking the Fourier Transform of both sides of (46) yields

$$A(\omega)P(\omega) = \lambda P(\omega), \qquad |\omega| \le \frac{\pi}{T}, \tag{47}$$

where

$$\begin{aligned} A(\omega) &= \left| \sum_k X\left(\omega + \frac{k2\pi}{T}\right) \right|^2 + \sigma^2 \\ &= |X_{\text{eq}}(\omega)|^2 + \sigma^2, \qquad |\omega| < \frac{\pi}{T}. \end{aligned} \tag{48}*$$

The only way for (47) to be satisfied [with $P(\omega) \not\equiv 0$] is for $P(\omega)$ to be concentrated at a single frequency [unless $A(\omega)$ has the same value at

* Recall that the Nyquist-equivalent spectrum $X_{\text{eq}}(\omega)$ is defined as $X_{\text{eq}}(\omega) = \sum_k X(\omega + k2\pi/T)$.

more than one frequency]. If we let

$$\omega_i = \frac{i}{N}\frac{\pi}{T},$$

then the solution to (47) is

$$\lambda_i = A(\omega_i)$$
$$-N \le i \le N$$
$$P_i(\omega) = \delta(\omega - \omega_i). \tag{49}$$

Thus, for a synchronous equalizer the asymptotic ($N \to \infty$) eigenvalues uniformly sample the folded-channel plus noise spectrum, and the eigenvectors are the corresponding sinusoids.

A.2 Fractionally Spaced Equalizer

Here the channel-correlation matrix, while symmetric, is not Toeplitz; thus Fourier Transform techniques do not yield the eigenvalues and eigenvectors in the above short order. For convenience we consider the noiseless situation, and the eigenvalue equation becomes

$$\sum_{l=-N}^{n} A(kT', lT')p(lT') = \lambda p(kT') \qquad -N < k < N, \tag{50}$$

where the channel-correlation matrix has elements

$$A(kT', lT') = \sum_m x(mT - kT')x(mT - lT'). \tag{51}$$

With $T' = T/2$ and $N \to \infty$ we write (50) for even and odd values of k

$$\sum_{l\,\text{even}} A\left(k\frac{T}{2}, l\frac{T}{2}\right) p\left(l\frac{T}{2}\right) + \sum_{l\,\text{odd}} A\left(k\frac{T}{2}, l\frac{T}{2}\right) p\left(l\frac{T}{2}\right) = \lambda p\left(k\frac{T}{2}\right), \; k \text{ even} \tag{52}$$

$$\sum_{l\,\text{even}} A\left(k\frac{T}{2}, l\frac{T}{2}\right) p\left(l\frac{T}{2}\right) + \sum_{l\,\text{odd}} A\left(k\frac{T}{2}, l\frac{T}{2}\right) p\left(l\frac{T}{2}\right) = \lambda p\left(k\frac{T}{2}\right), \; k \text{ odd}. \tag{53}$$

Now (52) and (53) can be written respectively as

$$\sum_l A(kT, lT)p(lT) + \sum_l A\,(kT, lT) + \frac{T}{2}\,p\left(lT + \frac{T}{2}\right) = \lambda p(kT), \qquad -\infty < k < \infty \tag{54}$$

and

$$\sum_l A\left(kT+\frac{T}{2}, lT\right)p(lT)+\sum_l A\left(kT+\frac{T}{2}, lT+\frac{T}{2}\right)p\left(l\frac{T}{2}\right) = \lambda p\left(kT+\frac{T}{2}\right), \quad (55)$$

where both equations hold for all integer values of k, and, more importantly, the various component matrices are now all Toeplitz.* If we let

$$\tilde{X}_{\text{eq}}(\omega) \triangleq X(\omega) - X\left(\omega-\frac{2\pi}{T}\right) - X\left(\omega+\frac{2\pi}{T}\right), \qquad |\omega| \le \frac{\pi}{T} \quad (56)$$

and

$$\tilde{P}(\omega) \triangleq P(\omega) - P\left(\omega\frac{2\pi}{T}\right) - P\left(\omega+\frac{2\pi}{T}\right), \qquad |\omega| \le \frac{\pi}{T}, \quad (57)$$

then taking the synchronous Fourier Transform (i.e., with respect to the T seconds sampling interval) of (54) and (55) gives

$$|X_{\text{eq}}(\omega)|^2 P(\omega) + X_{\text{eq}}(\omega)\tilde{X}^*_{\text{eq}}(\omega)\tilde{P}(\omega) = \lambda P(\omega) \qquad 0 \le |\omega| \le \frac{\pi}{T} \quad (58)$$

and

$$\tilde{X}_{\text{eq}}(\omega)X^*_{\text{eq}}(\omega)P(\omega) + |\tilde{X}_{\text{eq}}(\omega)|^2\tilde{P}(\omega) = \lambda\tilde{P}(\omega). \quad (59)$$

Note that $p(kT + T/2)$ has the Fourier Transform

$$e^{-j\omega\frac{T}{2}}\tilde{P}(\omega),$$

while the Transform of $p(kT)$ is of course $P(\omega)$. Arguing as we did for the synchronous equalizer, we see that the ith eigenvectors $P_i(\omega)$ and $\tilde{P}_i(\omega)$ must again be delta functions at $\omega_i = i/N\ \pi/T$. Upon setting the determinant of the pair (58) and (59) to zero, we see that the eigenvalues satisfy

$$\lambda_i^2 - \lambda_i||X_{\text{eq}}(\omega_i)|^2 + |\tilde{X}_{\text{eq}}(\omega_i)|^2| = 0, \quad (60)$$

and thus for each value of ω_i there are two eigenvalues

$$\lambda_i = 0$$

$$\lambda_i = |X_{\text{eq}}(\omega_i)|^2 + |X_{\text{eq}}(\omega_i)|^2 = \sum_k \left|X\left(\omega_i+\frac{k2\pi}{T}\right)\right|^2. \quad (61)$$

In contrast to the synchronous equalizer, half of the eigenvalues are

* For example, $A(kT, lT + T/2) = \sum_m x(mT - kT)x(mT - lT + T/2) = \sum_n x(nT)x|nT + (k - l)T + T/2|$.

exactly zero, while the other half are samples of the aliased magnitude-squared channel transfer function. Not surprisingly, the eigenvalues are independent of the receiver sampling phase. Once the eigenvalues are determined we can solve for the eigenvectors. The ith eigenvector associated with the *zero* eigenvalue is

$$p_i\left(n\frac{T}{2}\right) = \begin{cases} \tilde{X}^*_{\text{eq}}(\omega_i)e^{j\omega_i nT}, & n \text{ even} \\ -X^*_{\text{eq}}(\omega_i)e^{j\omega_i\left(n+\frac{1}{2}\right)T}, & n \text{ odd,} \end{cases} \tag{62}$$

while the eigenvector associated with the *nonzero* eigenvalue is

$$p_i\left(n\frac{T}{2}\right) = \begin{cases} X_{\text{eq}}(\omega_i)e^{j\omega_i nT}, & n \text{ even} \\ \tilde{X}_{\text{eq}}(\omega_i)e^{j\omega_i\left(n+\frac{1}{2}\right)T}, & n \text{ odd.} \end{cases} \tag{63}$$

At this point we remark that when ω_i is not in the rolloff region then $X_{\text{eq}}(\omega_i) = \tilde{X}_{\text{eq}}(\omega_i)$, and (63) describes a sinusoid of frequency ω_i, since the even and odd portions of $p_i(n\ T/2)$ mesh together in a continuous manner [i.e.,

$$p_i(n\ T/2) = X_{\text{eq}}(\omega_i)e^{j\omega_i n\,T/2}].$$

However, (62) describes a function that changes sign and oscillates almost a full cycle in T seconds. Consequently, $p_i(n\ T/2)$, as given by (62), will have most of its spectral energy concentrated near $1/T$ Hz. When ω_i is in the rolloff region, the frequency content of (62) and (63) will differ somewhat from the above extreme cases but the general results will still be as above.

REFERENCES

1. A. Gersho, private communication, 1969.
2. D. M. Brady, "An Adaptive Coherent Diversity Receiver for Data Transmission Through Dispersive Media," Conference Record ICC 1970, pp. 21-35 to 21-40.
3. D. M. Brady, "Adaptive Signal Processor for Diversity Radio Receivers," U. S. Patent No. 3,633,107, filed June 4, 1970, issued January 4, 1972.
4. L. Guidoux, "Equaliseur Autoadaptif à Double Echantillonnage," L'Onde Electrique, *55* (January 1975), pp. 9-13.
5. O. Macchi and L. Guidoux, "A New Equalizer: the Double Sampling Equalizer," Ann. Telecommun. *30* (1975), pp. 331-8.
6. G. Ungerboeck, "Fractional Tap-Spacing Equalizers and Consequences for Clock Recovery for Data Modems," IEEE Trans. on Commun., *COM-24*, No. 8 (August 1976), pp. 856-64.
7. S. U. H. Qureshi and G. D. Forney, Jr., "Performance and Properties of a $T/2$ Equalizer," Conference Record NTC 1977 (December 1977).
8. R. D. Gitlin and S. B. Weinstein, "Fractionally-Spaced Equalization: An Improved Digital Transversal Equalizer," B.S.T.J., *60,* No. 2 (February 1981), pp. 275-96.
9. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication,* New York: McGraw-Hill, 1968.
10. R. D. Gitlin and S. B. Weinstein, "On the Required Tap Weight Precision for Digitally-Implemented Adaptive Equalizers," B.S.T.J., *58,* No. 2 (February 1979), pp. 301-21.
11. J. E. Mazo, "On The Independence Theory of Equalizer Convergence," B.S.T.J., *58,* No. 5 (May-June 1979), pp. 963-93.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Using Magnetic Bubble Memories to Provide Recorded Announcements

By J. E. ROWLEY and J. BERNARDINI

(Manuscript received September 20, 1981)

Recorded announcement systems have traditionally been electromechanical in nature, and have required routine maintenance and adjustments. An electronic system, the 13A Announcement System, was designed to replace many of these older systems and eliminate the need for routine maintenance. It is the Bell System's first application of magnetic bubble memories, and can provide from one to eight announcements of up to 24 seconds in length each. Entering the fifth year of production, the 13A is being installed in almost every type of central office of the telecommunications network.

I. INTRODUCTION

In the Bell System, audio recorders and announcement systems are used for a variety of applications to provide customers and operators with information. One group of applications are those associated with calls that cannot be completed. Announcements in this group tell the customer why the call was not completed. Examples of problems explained by these announcements are: non-working numbers, incorrectly dialed calls, damaged equipment and heavy traffic conditions. In general the announcements associated with these conditions are non-revenue producing. These announcements are fairly brief, many lasting only twelve seconds. Another major group of announcements are those which are customer requested. These announcements are longer, often about a minute in length. Examples of these are: the weather, stock market quotations, the news, department store sales, jokes, sports phone, etc. In general this group of announcements are revenue producing.

Traditionally, the machines used to provide these announcements have been categorized as being either light-duty or heavy-duty ma-

chines. The light-duty machines were generally designed for customer premises or for central offices where their use was not continuous. The heavy-duty service machines were characterized by a much higher calling rate, requiring that the machines run continuously 24 hours a day.

The majority of the recording machines presently being used by the Bell System are approximately 25 years old in design. These recorders are all basically electromechanical in technology, employing magnetic drums for analog storage. In addition these designs require magnetic heads, mechanical drives, and a variety of moving parts. The problem associated with these designs has been the need for frequent routine maintenance in the form of lubrication, mechanical adjustments, and the replacement of worn parts. There is also a high cost associated with the special mechanical skills required for this type of maintenance. Other drawbacks of these electromechanical systems include their large physical size and their large power requirements.

In contrast a solid-state design employing digital storage has the advantage of no moving parts and, hence, no routine maintenance. It also has the advantage of high reliability, extremely long life and small size. This technology also leads itself to modular construction with standard components and circuit boards.

A study conducted in 1975 showed that for the shorter announcements a memory system employing magnetic bubble technology would be economically attractive. The Bell System had extensive development under way in this technology. It was clear that the magnetic bubble memory architecture was ideal for use in a solid-state announcement system and that the nonvolatility which the bubble memory offered was very desirable for this application. The large storage capacity of the bubble memory was another important consideration. All of these factors combined to make magnetic bubble memories the ideal choice for a storage medium.

In the attempt to define what type of announcement system would be appropriate for the present and near future market, a survey of operating companies was conducted in 1975. This survey showed that 90 percent of the shorter announcement market could be served by an eight-channel system. It also indicated that 85 percent of the market would be served by a system capable of providing recordings up to 24 seconds in length. To minimize digital storage requirements various digital coding techniques were considered. Based on human factors tests, adaptive delta modulation (ADM) at 24 kb/s was selected as the best compromise of bit rate and intelligibility.

With message length, bit rate, and multiple channel capability defined, the basic design requirements of a solid-state announcement system, coded the 13A, were established.

II. GENERAL DESCRIPTION OF THE 13A ANNOUNCEMENT SYSTEM

The design of the 13A system evolved as a multichannel system that can record and play back up to eight messages. Each message can be, and usually is, different. With appropriate distribution circuits, each message can serve up to 500 customers simultaneously. For many applications, the 13A is used as a direct replacement for up to eight 7A electromechanical machines. Figure 1 shows the 13A as compared to eight commonly used 7A machines in standard central office frames. Figure 2 shows the 13A fully equipped with eight message modules.

To minimize the cost of the system, circuitry that is common to all messages resides on three circuit packs at the left side of the system. A minimum 13A system requires one additional message-module circuit pack, which contains the memory and other circuitry necessary for that particular message. The remaining seven positions at the right side of the 13A can be equipped or not, as desired.

There are two different message module circuit packs. Both types offer a variable-length message feature. One records announcements from 3 to 12 seconds in length, and the other can record messages up to 24 seconds long. The message length can be adjusted in 3-second increments by means of a thumb-wheel switch on the message module pack. Each pack can be set to a different message length, if desired, contingent upon its associated switching system.

The 13A requires a maximum of 77 watts from the -48 volt central office battery for eight channels. This is an improvement over previous announcement machines, such as the 7A, which requires 30 watts from 110 volts ac for one channel.

Each message channel of the 13A has a START input which controls the start of a message. After a closure to ground on this input, the message starts at the beginning with an average waiting time of 2.5 seconds. As long as the start input is grounded, the message will continue to repeat with a 3-second silence interval between message repetitions.

In addition to the message audio, several other outputs are available on a per-channel basis. To provide interchangeability between the 7A and the 13A, 7A output signals and their nomenclature were adopted. These outputs are contact closures that provide timing information that can be used by trunk circuits to ensure that the customer hears the message from the beginning, rather than from any point within the message.

Each message-module channel also provides a voice-alarm output. A contact closure on this pair of leads signals an absence of audio output from the message channel. A loss of power to the 13A system will also result in a voice alarm. A system input called VATEST can be used to test the operation of the voice-alarm circuit.

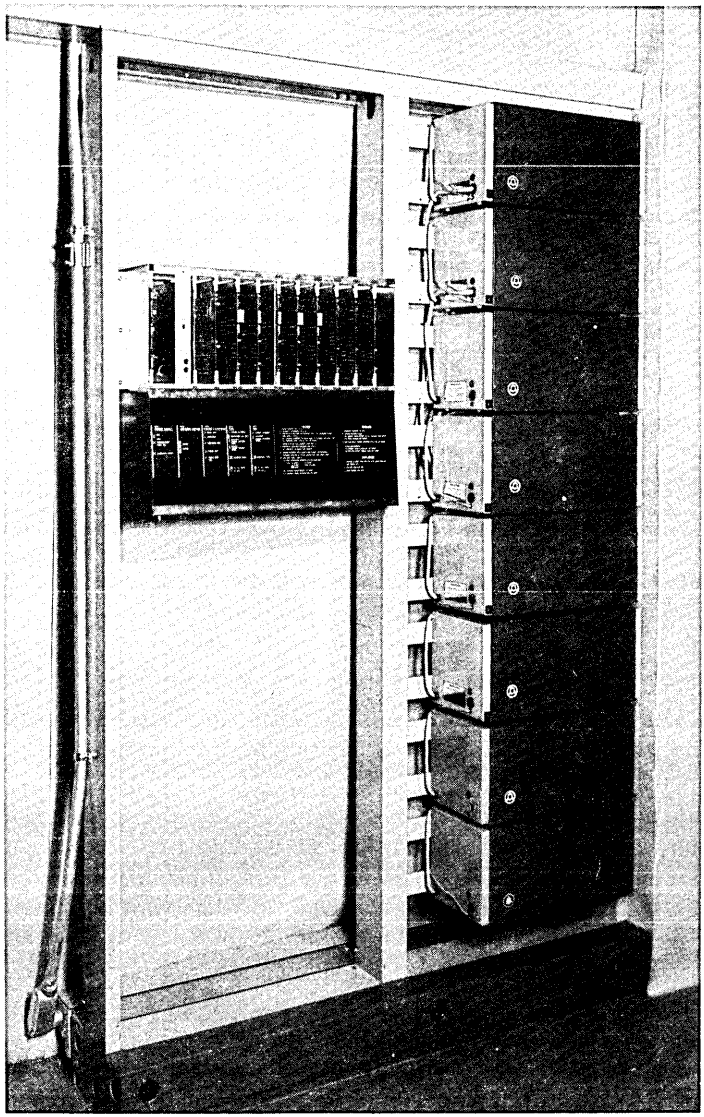


Fig. 1—Frame-mounted 13A Announcement System (left) compared to eight 7A units.

There are two methods for recording a message on the 13A. One method is to use a G3CR-type handset. The preferred method is to use a prerecorded tape and dub the message from a cassette recorder (typically) onto the 13A. The latter method produces the highest-quality announcement, since the tape can be prepared in a quiet environment by a professional speaker. The 13A provides a tape input

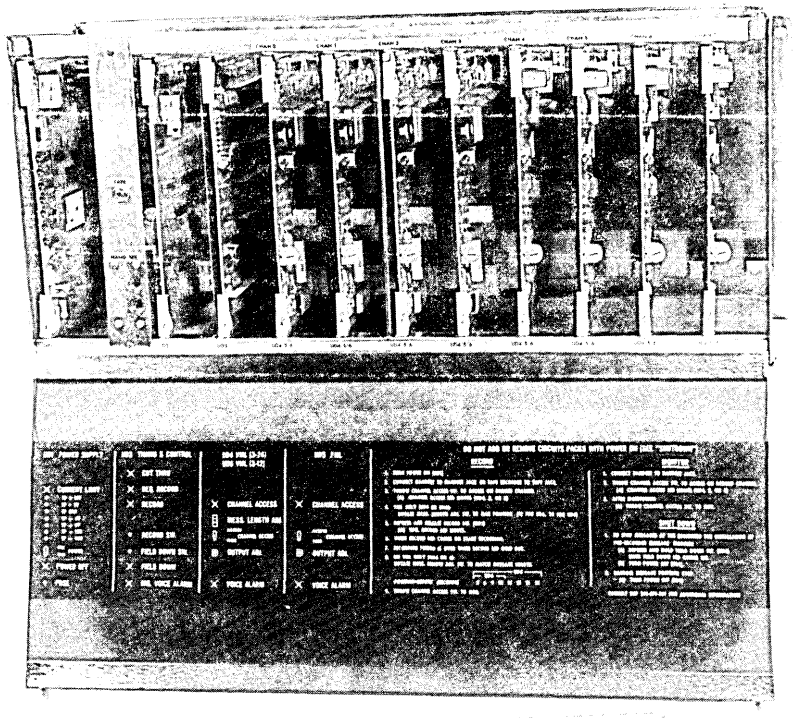


Fig. 2—The 13A Announcement System with eight message modules.

jack for this purpose. With the introduction of special information tones on standard intercept, vacant code, no circuit, and reorder announcements in 1981, dubbing from prerecorded tapes will become the standard method for recording these messages.

A simplified block diagram of the 13A system is shown in Fig. 3.[1] In recording, the input speech is filtered, converted into a digital representation using adaptive delta modulation (ADM), and written into the 29A magnetic bubble memory. Reproducing the message involves recovering the data from the memory, conversion back to an analog signal, and amplification for driving the output, which interfaces with trunk circuits. A large portion of the 13A circuitry is used to generate signals that are peculiar to the 29A memory. Therefore, some knowledge of the bubble memory is helpful in understanding the 13A system.

III. THE 29A MAGNETIC BUBBLE MEMORY

The 29A package, shown in Fig. 4, is a 32-pin dual in-line package 2.49 inches long and 1.2 inches wide. A permalloy outer case protects

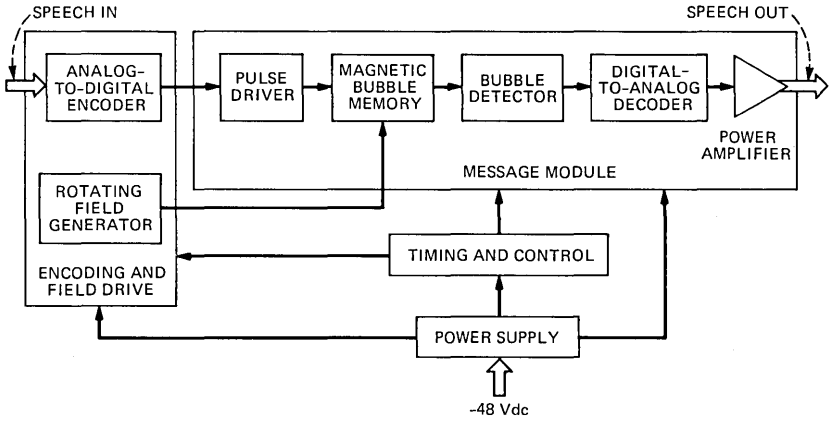


Fig. 3—Block diagram of the 13A Announcement System.

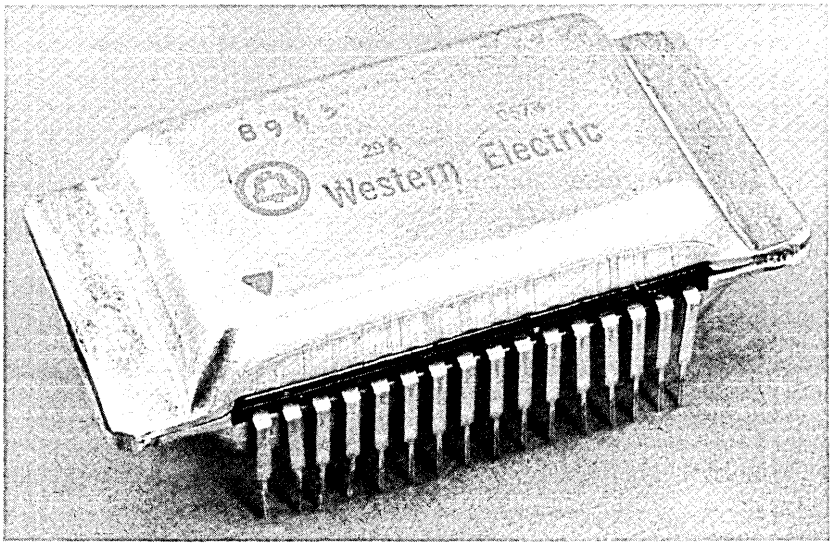
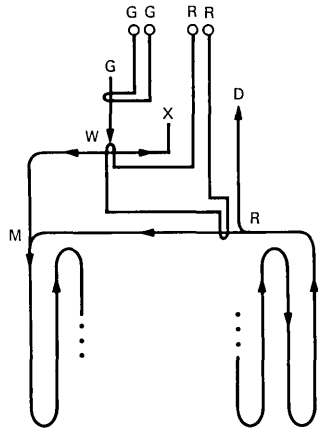


Fig. 4—The 29A bubble memory.

the bubble chips, rotating field coils, erase coil, and bias magnet from stray external fields.[2]

The package contains four 68,121-bit chips for a total of 272,484 bits. At the 24-kb/s encoding rate, storage capacity is approximately 3 and 12 seconds for chip and package, respectively. Each chip, approximately 5 mm by 6 mm, is organized as a single serial shift register. Ones and zeros are represented by the presence or absence of bubbles, respectively. Generation of new information and detection of the



D ACTIVE DETECTOR
G DATA INPUT GENERATOR
W WRITE TRANSFER
R READ REPLICATE OR TRANSFER
M PASSIVE MERGE POINT
X GUARD RAIL ANNIHILATOR

OPERATION	IN-PLANE FIELD ROTATIONS
DATA LOOP	68,121
R TO D	60
G TO W	8
R TO M EQUALS W TO M	81

Fig. 5—The 29A chip organization.

existing memory contents are both done outside the 68,121-bit shift register loop.

A line diagram of the chip organization is shown in Fig. 5. A bubble is generated at point G by an appropriate current pulse in the G conductor. Eight steps later, the bubble is at the write transfer point, W. If no current pulse is present in the R conductor, the bubble will travel along the path which ends in an X and be annihilated. If the proper current pulse is present in the R conductor, the bubble will take the other path, which will put the bubble into the data loop at point M. This is a passive merge point, and no signals are required to merge the generated bubble into the 68,121-bit loop.

Once in the main loop, the bubbles require no conductor signals to keep them circulating in the loop. However, no bubbles will go to the detector unless a proper current pulse is applied to the R conductor. Two different types of pulses can accomplish this. A transfer pulse, applied to the R conductor, will cause all bubbles at point R to travel to the detector. A series of 68,121 of these could be used to clear the memory. Or, in combination with the generator lead, a smaller number of transfer pulses could be used to rewrite a portion of the 68,121 locations. Another pulse, a replicate pulse, will cause a bubble at point R to split into two. One bubble stays in the main loop, the other travels to the detector. This pulse type is used for nondestructive readout of the memory's contents. The transfer pulse is not used in the 13A, since the 13A does not write a portion of a chip. In the 13A, clearing of the 29A memory is done by using the z coil.

In the z axis, which is perpendicular to the plane of the chip, there is a bubble-stabilizing bias field, nominally 160 Oe, which is supplied by a pair of permanent magnets inside the package. This field can be

aided or opposed by a current in the z coil. The z coil is used in manufacture to test the bias margins of the four chips. A much larger z coil current, aiding the bias field, will cause collapse of the bubble domains and can be used to bulk erase the entire memory.

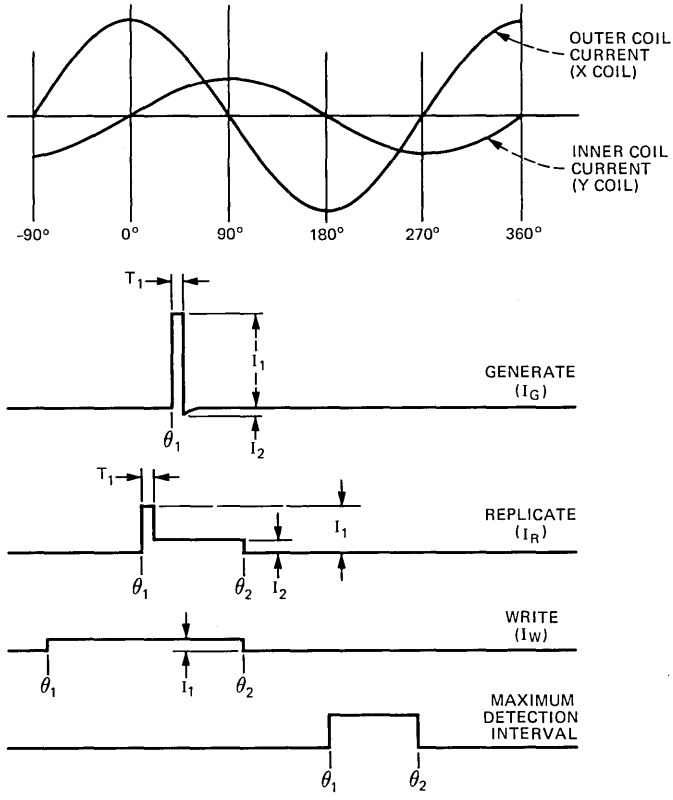
In a semiconductor shift register, a clock signal advances the memory. In the 29A, an in-plane, constant-magnitude, rotating magnetic field does this. One 360-degree revolution of the rotating field advances the bubbles by one position. The rotating field is formed by applying two sinusoidal currents, 90 degrees out of phase, to the inner and outer field coils. In the 13A, these coils are driven by voltage sources, so the inductance of the coils must be tightly controlled in order to obtain control of the current amplitudes. At the 48-kHz rotating field frequency, the required inner coil drive voltage is 25.15 volts peak to peak, and the current is 0.50 ampere peak to peak. The outer coil requires 25.15 volts peak to peak at 1.34 amperes peak to peak. The tolerance on both voltages is ±5 percent. Because the Q of the two coils is different, the desired 90-degree phase shift in currents is achieved by having the outer coil voltage lead the inner coil voltage by 95 ± 3 degrees.

The timing of the other functions of generate, replicate, and strobe (for data detection) is related to the phase angle of the rotating field. The timing requirements, as well as the current amplitudes, are shown in Fig. 6. It should be noted that the function drives must be current sources, since the function resistances have a tolerance of ±20 percent and the current tolerance, for example, of high replicate current is 110 mA ± 13.6 percent.

In the 29A package, the generators of all four chips are connected in series, and one pair of generator leads is brought out. Similarly, the detectors and dummy detectors are connected in series. The replicate leads are brought out individually, and they function as chip-select leads, determining which chip is being written into or read from.

The bubble arrives at the detector 60 rotating field cycles after being replicated at point R. The bubble field is detected by a magnetoresistive detector. Since other fields (especially the rotating field) are picked up, a similar dummy detector, under which no bubbles pass, is included to cancel out unwanted signals. Both active and dummy detectors are driven by constant current sources of 4.5 mA ± 0.2 mA. The match between the two current sources must be better than 1 percent. One side of the active and dummy detectors is connected together inside the 29A, and external circuitry must place this node at ac ground. The remaining two leads, one active and one dummy, must be amplified differentially. Since the detectors are insensitive to field polarity, the unwanted common mode signal is mostly at twice the rotating field rate and is about 80 mV peak to peak.

The desired bubble signature occurs during the strobe interval and is seen differentially across the active and dummy leads. To obtain a digital result, the signal is processed as shown in Fig. 7. A "1" (bubble) signature is typically 5 mV in magnitude, and a "0" response is somewhat smaller, as shown in Fig. 8. A technique known as dc restoration is used after this waveform is obtained. This references the signal level to the level at the start of the strobe interval. That is, within the strobe interval:



SIGNAL	I_1 mA	I_2 mA	T_1 μs	θ_1 DEGREES	θ_2 DEGREES
GENERATE	280 ± 35	<10	0.4 ± 0.1	40 ± 15	--
REPLICATE	110 ± 15	30 ± 6	0.7 ± 0.1	10 ± 10	110 ± 15
WRITE	30 ± 6	--	--	-75 ± 10	110 ± 15
MAXIMUM DETECTION INTERVAL	--	--	--	185°	276°

Fig. 6—The 29A timing requirements.

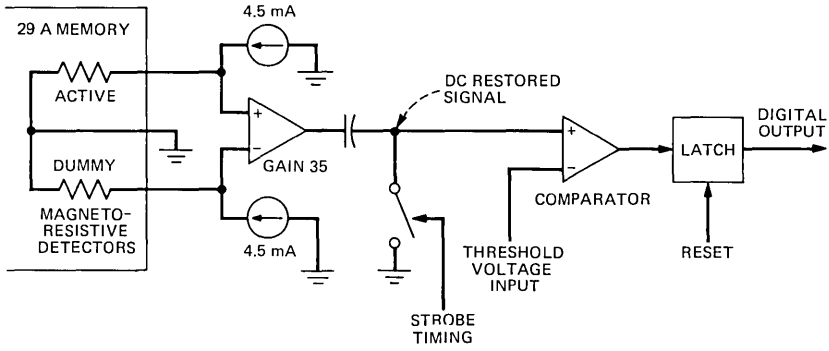


Fig. 7—Memory output signal processing.

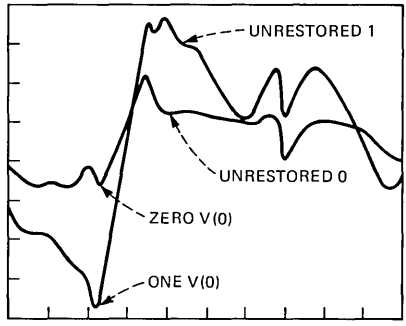


Fig. 8—Amplified bubble signal before dc restoration.

$$V(t)_{\substack{\text{AFTER}\\\text{RESTORE}}} = V(t)_{\substack{\text{BEFORE}\\\text{RESTORE}}} - V(0)$$

where V(0) is the signal amplitude at the start of the strobe interval. Since V(0) is higher for 0's than for 1's, this process increases the separation between 0's and 1's, as shown in Fig. 9.

The specifications of 29A outputs are based on dc restored waveforms. Output amplitudes vary widely from module to module. In the 13A, these variations are dealt with by varying the amplitude of a decision threshold until it is one-third of the distance between the 0 and 1 amplitudes.

One advantage of bubble memories is that data can be retained during loss of power, provided the rotating field signals are turned off properly. The rotating field must turn on and off within the ±45-degree segment as shown in Fig. 10. Failure to do this will result in loss of memory contents.

The 29A memories undergo extensive testing during manufacture.[3] This includes test of worst case amplitude and timing margins at room temperature, as well as a ±3 Oe bias margin test. In addition, each

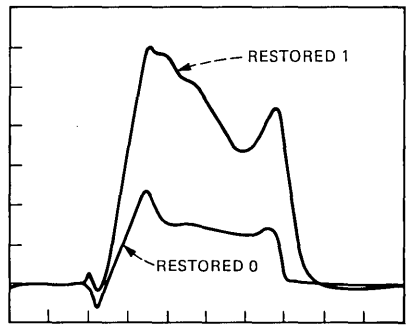


Fig. 9—Amplified bubble signal after dc restoration.

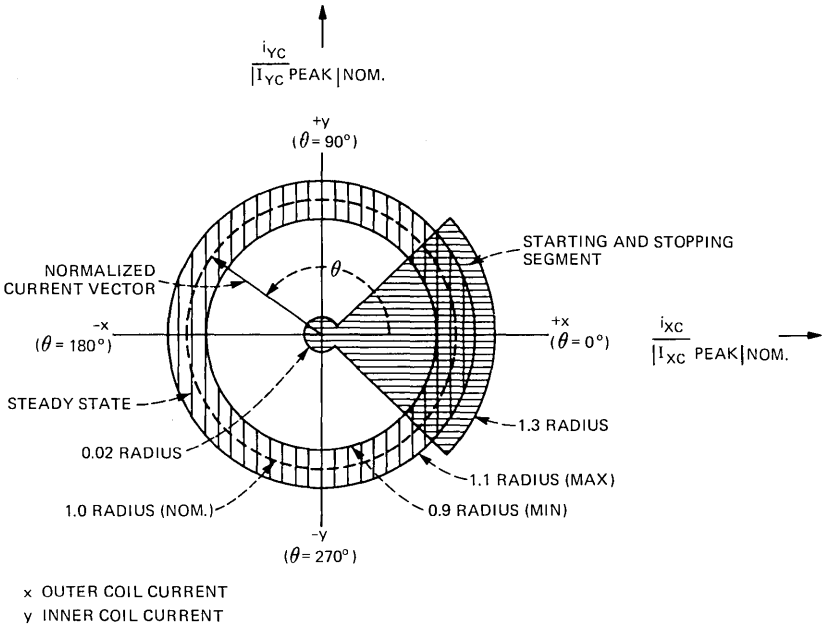


Fig. 10—Normalized coil-current polar plot showing the 29A start/stop requirements.

device must pass a less severe test at both 13°C and 60°C case temperature.

IV. THE AUDIO ENCODING-DECODING CHAIN

The 13A, viewed as a digitized audio system, is shown in Fig. 11. The signal flow can be traced through the audio-processing stage, the audio-encoding section, and the bubble memory. After the memory, the signal goes to a decoding circuit and then to an output amplifier.

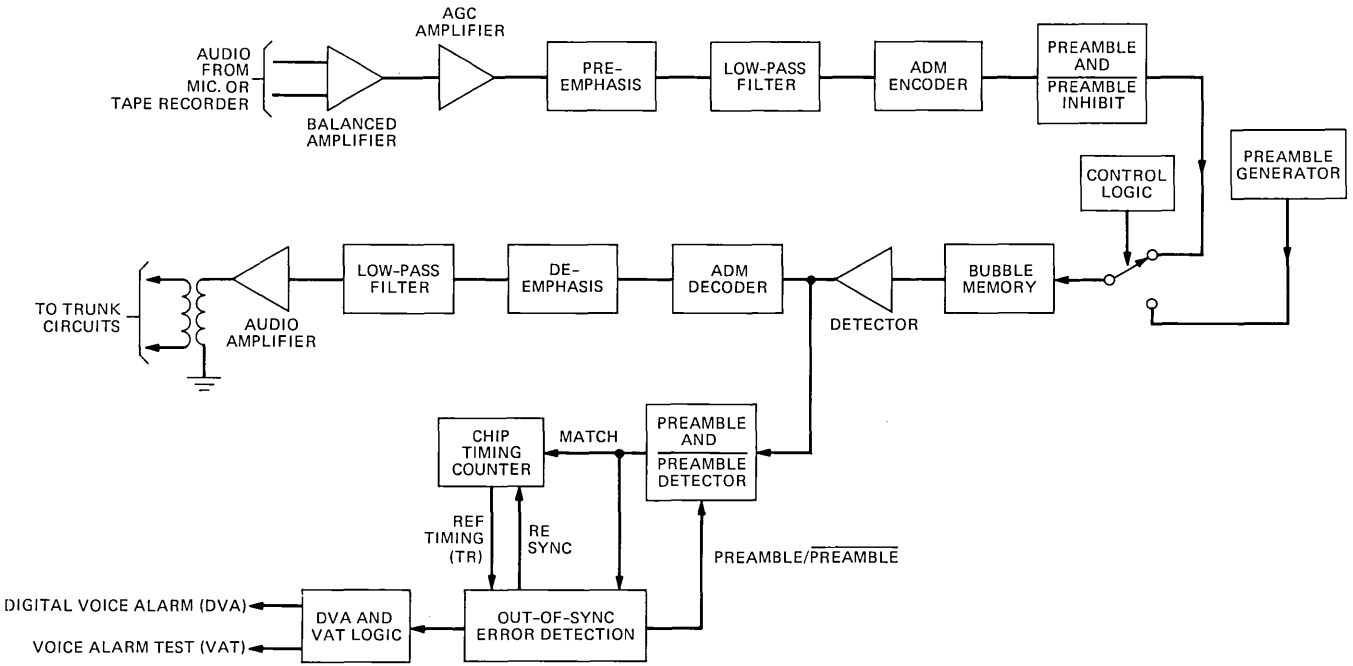


Fig. 11—Audio-chain and error-detection circuitry.

The first stage is a balanced amplifier which provides an input for a T1-type transmitter or a tape recorder source. The input of this stage is designed to bias a T1-type transmitter for up to 850 ohms of loop resistance and provide ac coupling for a tape recorder input.

The AGC amplifier provides 40 dB of input dynamic range to allow for various speaker levels, record levels, and loop resistance losses. With the input dynamic range compressed to 4 dB, an equalizer provides a signal increase at a slope of 6 dB/octave above 1 kHz and a signal decrease at 6 dB/octave below 1 kHz. This preemphasis is coupled with a corresponding deemphasis at the decoder output so that a flat frequency response is maintained. The overall preemphasis/deemphasis was subjectively determined to improve the response of the encoder/decoder chain by eliminating quantization noise at the decoder output. The last analog stage before encoding is a third-order, 2.8-kHz, low-pass filter defining the 13A bandwidth.

The filtered audio signal is digitized using an adaptive delta modulator (ADM) encoder operated at 24 kb/s. This is an ADM chip-set originally developed for the *SLC**-40 system and operated at approximately 38 kb/s.[4] For the 13A, the reduced clock rate is made possible because of the input AGC and the reduced importance of speaker identification. The output of the ADM encoder is processed by a preamble pattern inhibiter. The purpose of this circuitry is to inhibit bit patterns of 15 or more 1's and patterns of 15 or more 0's. A pattern of 15 or more 1's is modified by changing the eighth 1 to a 0. A pattern of 15 or more 0's is modified by changing the eighth 0 to a 1. This procedure ensures a unique bit pattern for use as a data preamble. With the ADM encoder used in the 13A, patterns of greater than 14 1's or 14 0's have a low frequency of occurrence. This procedure, therefore, introduces a negligible distortion into the ADM encoding. The preamble is generated by inserting a bit sequence of 15 1's surrounded by 0's as the first data loaded into the bubble memory during a message recording. The bubble memory output is converted to logic levels by the detector amplifier. The detector drives both an ADM decoder and circuitry associated with preamble detection, synchronization and error detection (see Section VIII). The ADM decoder decodes the digitized speech into audio. The inverse equalizer normalizes the speech to a flat spectrum and the low-pass filter eliminates quantizing noise residing outside of the audio band.

The final stage in this chain is a low-output-impedance amplifier capable of driving up to 500 trunk circuits at a maximum level of -9 volume units. Referring to Fig. 11 it should be noted that all circuitry

---

* *SLC* is a trademark of Western Electric.

after the preamble generator is duplicated for each message channel. All circuitry before this point is shared by all message channels for both record and playback operations.

V. TIMING AND FIELD DRIVE

A major consideration in the use of bubble memories is the overhead represented by the necessary circuitry for control and timing-signal generation. In the 13A, this circuitry takes the better part of two printed-circuit boards. However, this circuitry is shared by all message modules in the system.

Figure 12 shows a functional diagram of the circuitry for the field drive, field-cycle timing, and chip-cycle timing signals. The field drive used in the 13A is sinusoidal. The basic system timing is derived from the crystal oscillator which drives the control and timing counter. The control counter in turn supplies a clock to the quadrature sinewave generator, chip-timing counter and the timing Read Only Memory (ROM). The start and stop timing for the quadrature sinewave generator and for the application of the generator outputs to the power amplifiers is provided by the timing ROM. The timing ROM also provides basic timing for the field-cycle timing signals previously described. The chip-timing counter provides signals necessary for timing during the chip cycle.

On the message-module boards, where the bubble memories reside, are capacitors in parallel with both the inner and outer coils. These form parallel resonant circuits at the 48-kHz rotating field frequency. They reduce the power required for steady-state operation by a factor of 10, but they also change the start-up and shutdown loads that the power amplifiers drive.

To maintain nonvolatility with bubble memories it is important that the power up and power down cycles occur within the ±45-degree quadrant of the field drive cycles. When a system is powered up, the timing ROM starts the quadrature sinewave generator, allowing enough time for the generator to stabilize; then the timing ROM provides the start signal for the closure of the analog switches. The analog switch associated with the outer-coil drive is closed first. After a delay corresponding to 95-degrees of the field drive cycle, the inner coil drive analog switch is closed. At this point the inner and outer coil field drive signals must meet the previously outlined amplitude and phase requirements.

For nonvolatility upon power down, the inner and outer coil drive analog switches are shut off simultaneously at the point in the field drive cycle where the outer coil current is a maximum and the inner coil current is a minimum. This results in a gradual decay of the outer coil current and a slight oscillatory decay for the inner coil current.

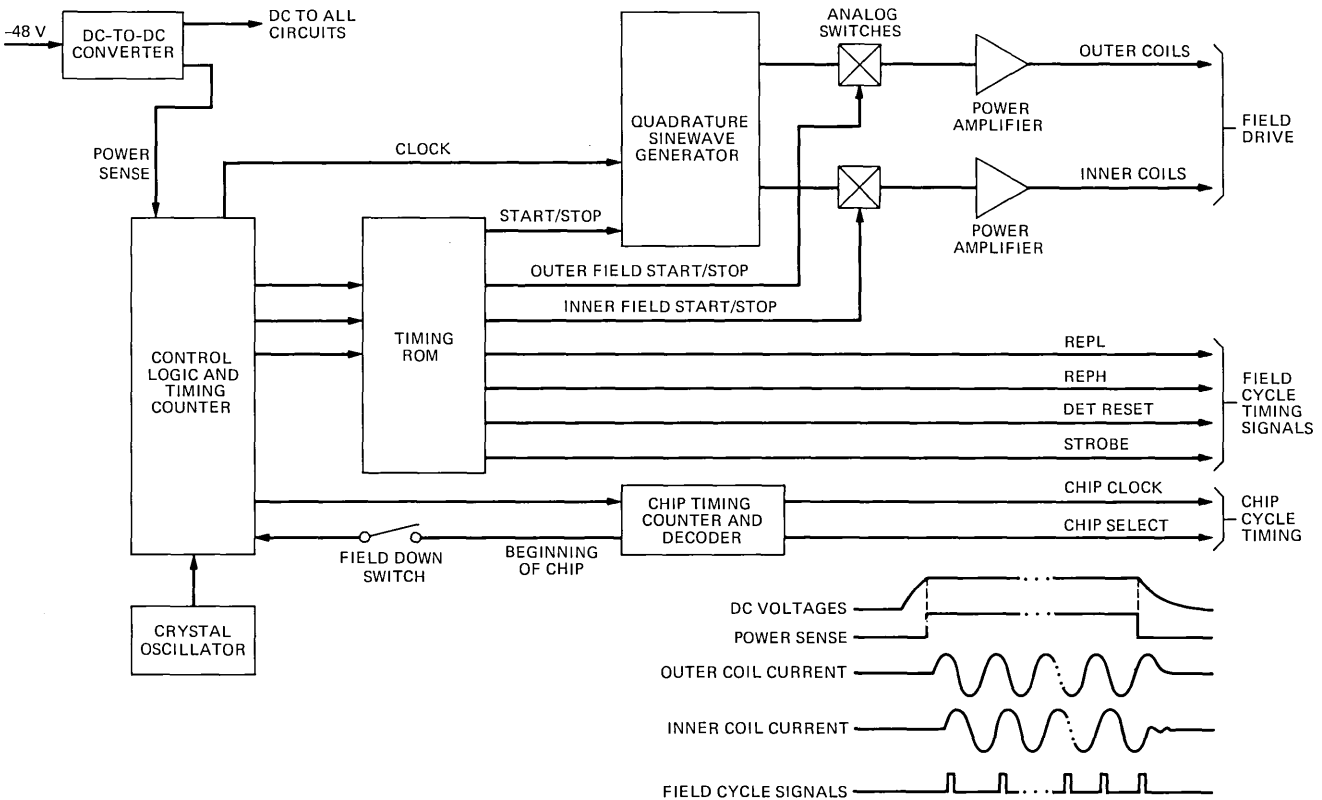


Fig. 12—Field and chip timing circuitry.

The resultant field vector lies within the ±45-degree sector necessary to maintain nonvolatility.

The final links in the chain for the inner and outer coil field drives are the associated power amplifiers. To maintain nonvolatile bubble memories during steady-state, transient, and variable-loading conditions, these amplifiers must be carefully designed. During steady-state operation, the amplifier must supply a sinusoidal signal whose amplitude must stay within ±5 percent and whose phase must not vary by more than ±1 degree for a load variation of 1 to 16 field coils. When the field drive signals are first turned on, the field drive loads appear capacitive. For this case, the power amplifiers must be capable of supplying up to 16 amperes at the sinusoidal peak voltage (12.57 volts). On turn-off, the power amplifiers must exhibit the transient response and low output impedances necessary to dampen the field coil energy. The power amplifiers also include circuitry to inhibit extraneous signals from reaching the field coils when the power supply voltages are below 13 volts. This is necessary to deal with transients occurring during low-voltage conditions when the 13A system is powered up or down.

VI. POWER CONSIDERATIONS

To ensure a nonvolatile system, the 13A must sense loss of power and turn the rotating fields and timing signals off before power supply voltages decay to the point that the control logic will no longer function. On power up, the system must hold the fields and timing signals off until power supply voltages are within normal limits. Figure 12 shows the functional circuitry and signals related to the 13A power start-up and shutdown operations.

When the dc voltages from the power supply are within 5 percent of their nominal value, the power sense signal is sent to the control logic. The power sense signal initializes all the logic and starts the sinewave generator via the timing ROM. Allowing sufficient time for the sinewave generator to stabilize, the field drive and field cycle timing signals are switched on.

If the -48 volt power is lost, or the power switched off, the power sense signal is switched low when the dc voltages have decayed by 5 percent. This signal, delayed by the control logic, switches off the field drive signals and associated field cycle timing signals at the proper time in the rotating field cycle. The maximum delay is one cycle of the rotating field, or 21 μs.

Although the previously described power-down sequence maintains 13A nonvolatility, it causes the messages to be stopped at a random point in the chip cycle. Message boards powered down in this manner cannot be interchanged between different 13A's without introducing

preambles at different relative positions. The ambiguous preambles would not permit proper synchronization of the 13A system. This, of course, would not be a problem for a single-channel 13A. To permit interchangeability of message modules for multichannel 13A's, the 13A can also be powered down by first employing the Field Down switch. Operating this switch (before switching the power off) turns the field drive signals off near the beginning of the chip cycle. This ensures that all message modules fielded down in this manner will have preambles in the same relative position. This feature cannot be ensured for emergency power down, or loss-of-power cases, because the control timing would have to delay the field down switching for up to one chip cycle (2.84 seconds) to reach the proper point in the chip cycle.

VII. RECORD AND PLAYBACK CONTROL

The record and playback operation of the 13A is controlled locally via circuit-card edge switches and light-emitting diodes (LEDs). The audio source can be either a telephone handset or a tape recording. Message modules can also be prerecorded at one site and installed in systems at other sites, provided the proper precautions are taken using the Field Down switch.

The record operation starts with the record-request signal generated from a card-edge push-button switch. This sets the record-request flip flop, which in turn lights a record-request LED on circuit pack UD3's card-edge. When the chip counter arrives at the appropriate point of the chip cycle, the record process is started by pulsing the z coil of all accessed message modules to clear the data recorded previously. Next, the preamble pattern is generated. The record-request LED is turned off and the record LED turned on as a signal to input the audio. At the end of each chip, the generated data is switched to the next chip until the last-chip signal occurs. The last-chip signal turns off the record LED, which signals the end of the record interval.

The playback operation can be controlled either locally or remotely. With the selected channel started remotely or accessed locally, the playback operation begins. This involves replication and detection of bubble signals, decoding the ADM data, and amplification for driving trunk circuits. At the end of each chip cycle, the replication process is switched to the next chip in the bubble module. If the playback involves two bubble memories, switching between them is not immediate. Owing to the delay between the replicator and the detector inside the 29A, switching of the detector from the first to the second module must follow switching of replication by 30 data cycles.

The setting of the thumb-wheel message-length switch determines which chip is the last chip in the message. At the end of the last-chip signal, the message ends and, if the start or access signals are present,

a 2.84-second silence interval begins, after which the message repeats. During the silence interval, several relay signals are given to signal the end of the message and the start of the next message repetition.

VIII. PREAMBLE SYNCHRONIZATION AND ERROR DETECTION

In the 13A system, the stored-message-synchronization and error-detection techniques were combined into common logic. The design requirements which prompted this approach were based on a need to provide synchronization of stored messages to the control counter and inexpensive means of logic fault detection. It was considered important that the fault-detection scheme be capable of detecting stuck-at-one(SA1) and stuck-at-zero(SA0) conditions in the data paths, and at the same time tolerate small or transient errors in the data. It was also considered desirable that the system be able to detect and recover from conditions when the control counter was out of synchronization with the messages.

Referring to Fig. 11, the data stream of the ADM encoded speech is first processed to inhibit preamble patterns and their complement from the data stream. A preamble of 15 1's is then first generated and stored in the bubble memory. When reading the data from the bubble memory, the data is first applied to the preamble detector noninverted. The output of the preamble detector, the match signal, is used to time synchronize the chip timing counter. Normally the first preamble match signal that occurs after powering up is employed to provide this synchronization. Following this operation, all 13A messages are chip-cycle synchronized to the control logic. All further occurrences of the match signal are compared with the chip-timing counter signal, TR. With a normal 13A system, match signals occur timed to TR. In this normal mode, every 60 seconds (enough time for two cycles of the longest 13A message) the data from the bubble store is inverted and applied to the preamble detector logic. During this cycle, the preamble detector in effect searches for 15 0's. Since this pattern was inhibited from the memory, any match signal generated in this cycle is interpreted as an error condition.

To make the system tolerant to transient errors, a threshold of two or more false preambles during a 60-second interval was established. If two or more false preamble errors are detected, the internal digital voice alarm (DVA) is set. To cover the possibility that the control counter has slipped out of sync with the stored messages, a sync retry cycle is initiated at this point. It starts by permitting the next preamble match signal to resync the chip timing counter. If another two or more false preambles occur, the external voice alarm test (VAT) signal is set. This same alarm is set for any case of a false preamble match signal during the data complement cycle. When an alarm is set during the

noncomplement data cycle, repeated attempts to resync the control counter to the preamble match signals are made. If a resync condition is achieved the alarm is reset and the system is returned to a normal state.

IX. PHYSICAL DESIGN AND CIRCUIT PACK DESCRIPTION

The five types of circuit packs used in the 13A are 8 by 11 inches. Except for circuit packs UD1 and UD2, all have three 963B-20 connectors for a total of 60 I/O pins per pack. UD1 and UD2 need only two connectors each. All packs are class 1 (25-mil paths and spaces) double-sided printed-circuit boards. All controls and indicators for operation of the 13A are mounted on the front edge of the circuit packs.

A major objective in the physical design of the 13A system was to eliminate hand wiring. This is accomplished by the backplane, shown in Fig. 13. This double-sided printed-wiring board interconnects all the circuit packs and the packs to the 123 I/O pins. Screw terminals are provided for -48V power and ground return. All other I/O terminals are 0.045-inch-square wire-wrap pins. Only four wires are needed to connect the handset and tape jacks located on the mounting strip at the front of the system.

Overall dimensions of the 13A system are 9.88 inches high by 12 inches deep by 23 inches wide. Compatibility with all the types of offices and frames in which the 13A is used is achieved by ordering different brackets to mount the system in the various frames. On the front of the unit there is a smoked plastic door, hinged at the bottom. All system indicators can be seen through the closed door at either the top or bottom. The strip across the center of the door is opaque. When

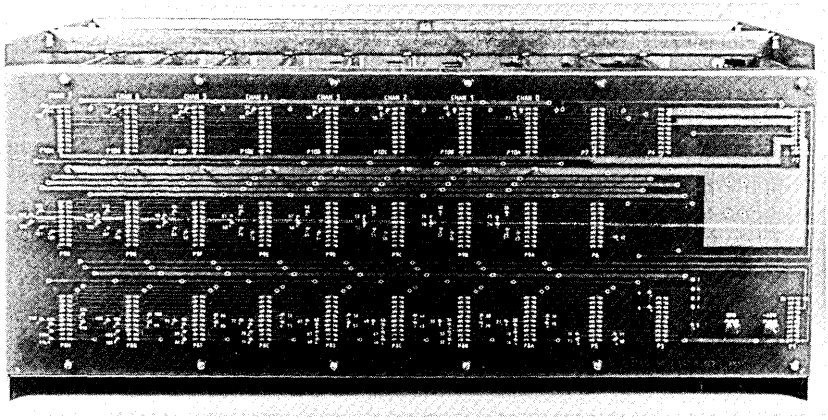


Fig. 13—The 13A Announcement System backplane.

the door is opened, the instructions on the inside of the door are visible. These instructions, shown in Fig. 2, contain a summary of the procedures for monitoring an existing message, recording a new message, and shutdown of power. All system indicators (LEDs) and control switches are also identified here.

The first circuit pack in the 13A is the power supply, which is coded UD1. It is a pulse-width-modulated dc-to-dc converter. It converts the -48 volt supply to +15.75V at 1.6A, -15.75V at 0.5A, and +5.2V at 3.2A. Also, two low-current supplies, -4V at 100 mA and -8V at 100 mA, are derived from the -15.75V. When supplying these currents, which are a maximum for a fully equipped 13A, UD1 draws 1.6A from the -48V supply. The voltage outputs have a tolerance of ±5 percent, except for the +5V supply. It has a tolerance of ±3.5 percent. Circuitry is also provided for electronic shutdown if the voltages exceed their tolerance or if currents exceed specified values. Each message board adds a resistor in parallel with the current-sense resistor on UD1, changing the current-sense limit in proportion to the number of message boards in the system. Also, the +5.2V and +15.75V outputs are monitored, and both must be higher than the minimum values to result in a high level on the PWR SENSE output.

The second common circuit pack is UD2, the encoder driver. It provides the circuitry for the input speech processing and encoding. This circuit pack also generates, amplifies, and provides analog switching of the field drive signals.

The other common circuit pack is UD3. This circuit pack contains the basic digital control and timing for the 13A. It controls the rotating field generation, provides bubble memory timing signals and chip sequencing, and placement of the recording of new messages. UD3 also provides the logic for message synchronization and error detection.

UD4 is a variable-length message module that stores and plays back up to 24 seconds of speech. It has a variable message length (VML) that can be adjusted from 3 to 24 seconds in 3-second increments. UD4 stores the digitally encoded speech in two 29A bubble memories. It converts the stored information back into analog speech, and it amplifies and buffers the output to drive up to 500 trunk circuits. Signaling closures are also provided by this circuit pack.

UD6 is a variable-length message module that stores and plays back up to 12 seconds of speech. It is similar to UD4. In fact, the two codes use the same printed-circuit board. A UD4 has the second 29A and associated components, a UD6 does not.

X. MANUFACTURING AND APPLICATION STATUS

The 13A is manufactured at the Western Electric Columbus Works. Manufacturing information was released on October 1, 1977. The first

13A was shipped to New York Telephone Company in February 1978. By the end of 1981, approximately 4188 13As had been shipped, and production should continue at the rate of about 110 systems per month. This requires 1670 29A memories per month. Each circuit pack must pass a circuit-pack test. Testing of UD3, UD4, and UD6 packs is automated, using computer-controlled test sets. The only adjustments during circuit-pack test are made on the message-module packs to match the detection threshold with the particular 29A device. In addition to circuit-pack tests, each 13A is given a system test before shipment.

The first 60 systems manufactured went to a special application in New York City. This is the Automatic Dial Coin Zone (ADCZ) System, which uses announcements to automate handling of two-message-unit coin phone calls in the New York City area. Each 13A is equipped with six UD6 message-module packs, for a total of 360 29A's in the 60 systems. All systems were placed in service on July 1, 1978.

At present 13A systems are also in service in Step-by-Step, No. 1 and No. 5 Crossbar, No. 1/1A ESS, No. 2/2B ESS, and No. 3 ESS offices. In No. 3 ESS applications, the 13A is installed and shipped with the office from the manufacturing facility. For No. 1/1A and No. 2/2B ESS the 13A's are installed and shipped as part of an announcement frame. In all other applications the 13A's are shipped directly to central offices for installation.

XI. CONCLUSION

The 13A is the first application of magnetic bubble memories in the Bell System. The 13A performance and modular design has permitted it to satisfy a very large percentage of the Bell System announcement market. It is being used in all types of central offices, and will be used in No. 5 ESS.

The 13A compares very favorably to the widely used electromechanical 7A system. The 13A is lower in cost for systems employing three or more channels, requires 77 watts for an eight-channel system versus 30 watts for a single 7A, and eliminates the need for costly periodic maintenance.

Magnetic bubble memory technology has proven to be an excellent match to the needs of the 13A Announcement System. By using magnetic bubble memories as the storage medium, the 13A offers nonvolatile storage of recorded announcements. The high storage density of magnetic bubble memories has enabled the 13A to occupy as little as one-eighth the space previously taken by eight 7As. The high density and low cost of magnetic bubble memories relative to other nonvolatile and writable storage devices enable the 13A to be very cost competitive with electromechanical systems.

In its fifth year of production and at its current manufacturing rate of 110 per month, the 13A has proven to be a successful product. This is true both in terms of the use of magnetic bubble memories and the 13A modular design.

REFERENCES

1. R. D. Trupp, "Improving Recorded-Announcement Service with Magnetic Bubbles," Bell Laboratories Record, *55,* No. 9 (Oct. 1977), pp. 249–52.
2. A. H. Bobeck and I. Danylchuk, "Characterization and Test Results for a 272K Bubble Memory Package," IEEE Trans. on Magnetics, *MAG-13,* No. 5 (Sept. 1977), pp. 1370–72.
3. R. Kowalchuk and J. Neuhausel, "Magnetic Bubble Memory, Part I: Bubble Technology and Memory Device Manufacture, Part II: Production Testing of Bubble Memory Message Modules," WE Engineer, *23,* No. 2 (April 1979), pp. 2–18.
4. R. J. Canniff, "Signal Processing in SLC-40, 40 Channel Rural Subscriber Carrier," ICC 1975 Conf. Rec., *3,* pp. 40.7–40.11.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Application of Graph Theory to the Solution of a Nonlinear Optimal Assignment Problem

By M. MALEK-ZAVAREI

(Manuscript received November 5, 1981)

This paper poses an assignment problem with a nonlinear objective function. It is formulated as an integer programming problem and a graph model is used to determine its exact solution. Application of the problem in long-range homing in telephone networks is also discussed.

I. INTRODUCTION

Consider a process in which at each stage an *originating center* must be connected to exactly one of a set of *terminating centers*. The originating center has a *load* and each terminating center has a *capacity*. The originating center can be connected to a terminating center only if the terminating center has enough capacity for its load. The originating center's load and the terminating centers' capacities vary with time (i.e., from one stage to another), but they are assumed to be known at all times. The problem is to determine the optimal connection configuration at each stage of time. The costs involved are the *transmission cost* and the *rearrangement cost*. Both costs are nonlinear functions of the originating center's load. The transmission cost is the cost of connection of the originating center to a terminating center. The rearrangement cost is incurred if, in transition from one stage to the next, the connection of the originating center to a terminating center has to be changed because of insufficient capacity.

Such an assignment problem may be encountered in many applications. One important application arises in the design of hierarchical telephone networks.[1-3] The process of connecting a switching center to a center in the next level of hierarchy in the backbone route of such networks is called *homing*. The problem of determining the optimal homing configuration of a switching center over several stages during a study period can be formulated as the above assignment problem.[4]

In such a case the originating center could be an end office and the terminating centers could be toll centers (i.e., switching centers in the next level of hierarchy). The load of the originating center would then correspond to the traffic volume of the end office and the capacities of terminating centers would correspond to the switching capacities of the toll centers. The transmission cost would be the cost of trunks homing the end office on a toll center and the rearrangement cost would correspond to the cost of changing the homing configuration.

In this paper a graph model will be developed for the above nonlinear assignment problem. The solution to the problem will be converted to determining a shortest path on this graph. The algorithm developed for the solution is easily programmable on the digital computer.

II. FORMULATION OF THE PROBLEM

Consider N stages of time indicated by $t = 1, 2, \cdots N$. Let $TC(t)$ represent the set of available terminating centers at stage t. The originating center OC must be connected to a terminating center TC_k in this set at stage t. Let

$$x_k(t) = \begin{cases} 1 & \text{if } OC \text{ homes on } TC_k \in TC(t) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Indicate the load of the originating center at stage t by $S(t)$ and the capacity of terminating center TC_k at stage t by $C_k(t)$. The costs and the constraints of the problem can then be formulated as follows.

2.1 Costs

The transmission cost per unit distance is assumed to be a nonlinear function f of the originating center's load, as shown in Fig. 1. It includes a fixed cost that is independent of the originating center's load. (In the case of the homing problem in telephone networks, this fixed cost would represent the cost of preparing for establishing a transmission facility, e.g., laying a cable.) The fixed cost will be incurred only for a new connection. For increasing the capacity of an existing connection, only the incremental cost will be incurred.

Thus, the total transmission and rearrangement costs can be formulated as

$$\sum_{t=1}^{N} \sum_{TC_k \in TC(t)} \{f[S(t)] - f[S(t-1)]x_k(t-1)\}x_k(t)d_k \tag{2}$$

where $S(0) = 0$ and d_k is the distance between OC and TC_k. Note that $S(t)$ is known a priori for $t = 1, 2, \cdots, N$.

Equation (2) includes the rearrangement cost, i.e., the cost incurred if in transition from one stage to the next the terminating center assignment changes. However, if such a change occurs, the old con-

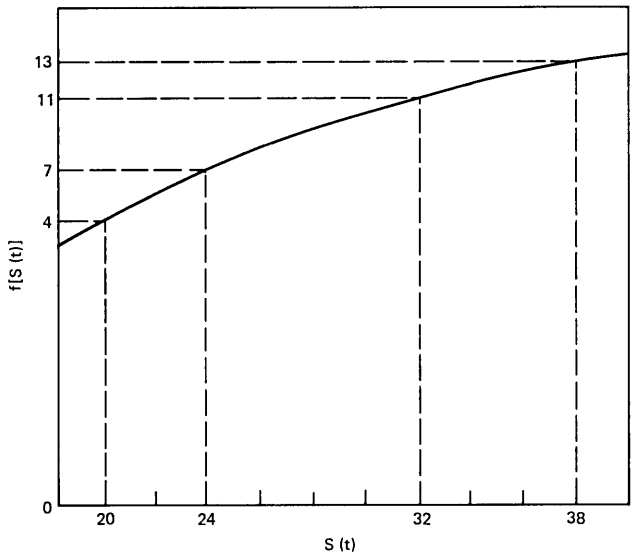


Fig. 1—Transmission cost per unit distance as a function of the originating center's load.

nection will represent a saving in the cost that must be considered. This saving is, in general, a nonlinear function of the rearranged load. We call this function $g[S(t)]$, which is obtained empirically for each application. Thus, eq. (2) must be modified as follows to represent the total transmission and rearrangement costs

$$\sum_{t=1}^{N} \sum_{TC_k \in TC(t)} \{f[S(t)] - f[S(t-1)]x_k(t-1)\}x_k(t)d_k - \sum_{t=1}^{N-1} \sum_{TC_k \in TC(t)} \frac{1}{2} g[S(t)]|x_k(t+1) - x_k(t)|. \quad (3)$$

The reason for including the coefficient ½ in the second part of eq. (3) is as follows. If in transition from one stage to the next the terminating center assignment changes from TC_k to $TC_{k'}$, the absolute value term will contribute 2 instead of 1 (1 for k and 1 for k').

2.2 Constraints

Since the originating center must be connected to exactly one terminating center at each stage, we must have

$$\sum_{TC_k \in TC(t)} x_k(t) = 1 \quad \text{for } t = 1, 2, \cdots, N. \quad (4)$$

Also, the load of the originating center cannot exceed the capacity of the terminating center to which it may be connected; thus

$$S(t)x_k(t) \leq C_k(t) \quad \text{for } all\ TC_k \in TC(t),\ t = 1, 2, \cdots, N. \tag{5}$$

The problem can then be formulated as follows. Minimize the objective function (3) subject to constraints (4), (5) and

$$x_k(t) = 0 \text{ or } 1 \quad \text{for } all\ TC_k \in TC(t),\ t = 1, 2, \cdots, N. \tag{6}$$

III. A GRAPH MODEL FOR THE PROBLEM

The above problem is a nonlinear integer programming problem.[5] For most practical applications this problem will have a considerable number of variables and constraints. For example, for 10 stages and 8 terminating centers there will be $8 \times 10 = 80$ variables and $10 + 8 \times 10 = 90$ constraints in the problem. The nonlinearity of the cost function and the large number of variables and constraints render the available standard integer programming techniques (such as the branch and bound method[5]) impractical for the solution of this problem. In this section a graph model will be developed for the problem, which aids in obtaining an exact solution for it.

Define a matrix A whose rows and columns correspond to stages and terminating centers, respectively. The (t, k) element of matrix A is defined as

$$a(t, k) = \begin{cases} 1 \text{ if } OC \text{ can be connected to } TC_k \text{ at stage } t \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

Element $a(t, k)$ of matrix A can easily be obtained by determining whether $TC_k \in TC(t)$ has enough capacity for the originating center at stage t. Hence, constraint (5) will be used in the construction of matrix A. To incorporate constraint (4) and to perform the required minimization, define a directed graph[6] G whose nodes are elements 1 of matrix A. The arcs of G connect node pairs in consecutive rows of A in the direction that t increases. Then each path in G that connects a node in the first row of matrix A to a node in its last row forms a feasible solution to the problem where each node (t, k) of graph G included in such a path corresponds to $x_k(t) = 1$. Note that since exactly one node in each row of matrix A is included in such a path, constraint (4) will be satisfied.

Costs will now be assigned to the nodes and the arcs of graph G in such a way that the solution of the problem converts to solving a shortest-path problem on G. The cost associated with a vertical arc in graph G (i.e., an arc connecting a node pair in the same column of matrix A) is zero. A cross arc of graph G (i.e., an arc connecting a node pair in different columns of matrix A) from row t to row $t + 1$ has cost $-\frac{1}{2}g[S(t)]$. Each node of graph G corresponding to terminating center TC_k and stage t is assigned the cost $\{f[S(t)] - f[S(t-1)]x_k(t-1)\}d_k$. If the arc connecting the predecessor node of (t, k) to it is a cross arc,

then the cost of node (t, k) is $f[S(t)]d_k$. If the arc connecting the predecessor node of (t, k) to it is a vertical arc, then the cost of node (t, k) is $\{f[S(t)] - f[S(t-1)]\}d_k$. With these assignments, the costs of the nodes and the arcs in graph G correspond, respectively, to the first and the second parts in expression (3). Hence, to solve the problem, a minimum cost path in graph G from nodes in the first row to nodes in the last row of matrix A must be found.

IV. AN ALGORITHM FOR THE SOLUTION

Since graph G includes no cycles, a labeling method similar to that used in the "shortest-path algorithm"[6] may be employed to determine paths of minimum cost. Label each node in the first row of A by its corresponding cost. The labels of nodes in the other rows of matrix A will be determined according to the following rule:

$$\text{Label of node } (t, k) = \underset{\text{all predecessor nodes}}{\text{Min}} \left\{ \begin{array}{l} \text{Node label of its} \\ \text{predecessor node} \\ + \text{ cost of the arc} \\ \text{connecting them} \\ + \text{ proper node cost} \end{array} \right\} \tag{8}$$

where

$$\text{proper node cost of node } (t, k) = \left\{ \begin{array}{l} f[S(t)]d_k \text{ if the arc} \\ \text{arriving at node } (t, k) \text{ is} \\ \text{a cross arc,} \\ \{f[S(t)] - f[S(t-1)]\}d_k \text{ if} \\ \text{the arc arriving at node} \\ (t, k) \text{ is a vertical arc.} \end{array} \right. \tag{9}$$

With the above labeling method, the label of each node will be the minimum cost of path(s) in graph G connecting nodes in the first row of matrix A to it. An optimal solution to the problem is obtained by finding the node(s) with minimum label in the last row of matrix A and determining its predecessor(s) from which the label was produced and continuing to the first row of matrix A. Each node thus obtained will then identify the proper terminating center assignment at the corresponding stage.

The above method is essentially a forward dynamic programming technique.[7] It can also handle the case where a terminating center has been initially assigned to the originating center. In such a case, simply adjoin a row on top of the first row of matrix A such that all its elements are zero except the one corresponding to the initial terminating center assigned to the originating center, which is 1. In that case all paths in graph G will originate from this new node (and its

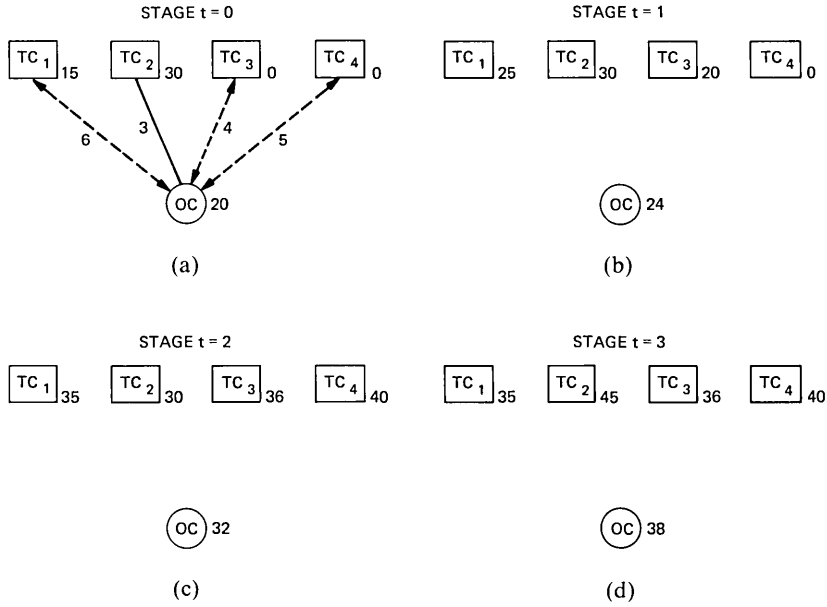


Fig. 2—Load, capacities, and distances in the example for stages $t = 0$ through $t = 3$.

label can be assumed to be zero). Note that for cost calculation in such a case, $S(0)$ will have its actual value, not zero.

V. EXAMPLE

We will illustrate the above method by applying it to a problem involving three stages, $t = 1, 2, 3$, and four terminating centers TC_k, $k = 1, 2, 3, 4$. Assume that initially (i.e., at $t = 0$) terminating center TC_2 has been assigned to the originating center. This is illustrated in Fig. 2a where the load of the originating center and the capacities of the terminating centers (in proper units) are indicated adjacent to them. Also, the distance between the originating center and each terminating center is indicated in Fig. 2a (i.e., $d_1 = 6$, $d_2 = 3$, $d_3 = 4$, $d_4 = 5$ in proper units). Figures 2b, 2c and 2d show the originating center's load and the terminating centers' capacities in stages 1, 2 and 3, respectively. Note that

$$S(0) = 20, S(1) = 24, S(2) = 32, S(3) = 38.$$

This configuration results in the matrix A shown in Fig. 3a. The corresponding graph G in which nodes are distinguished by circles and arcs are shown by directed links between the nodes is shown in Fig. 3b. Assume, for convenience, that the cost characteristic shown in Fig. 1 does not change with time and that

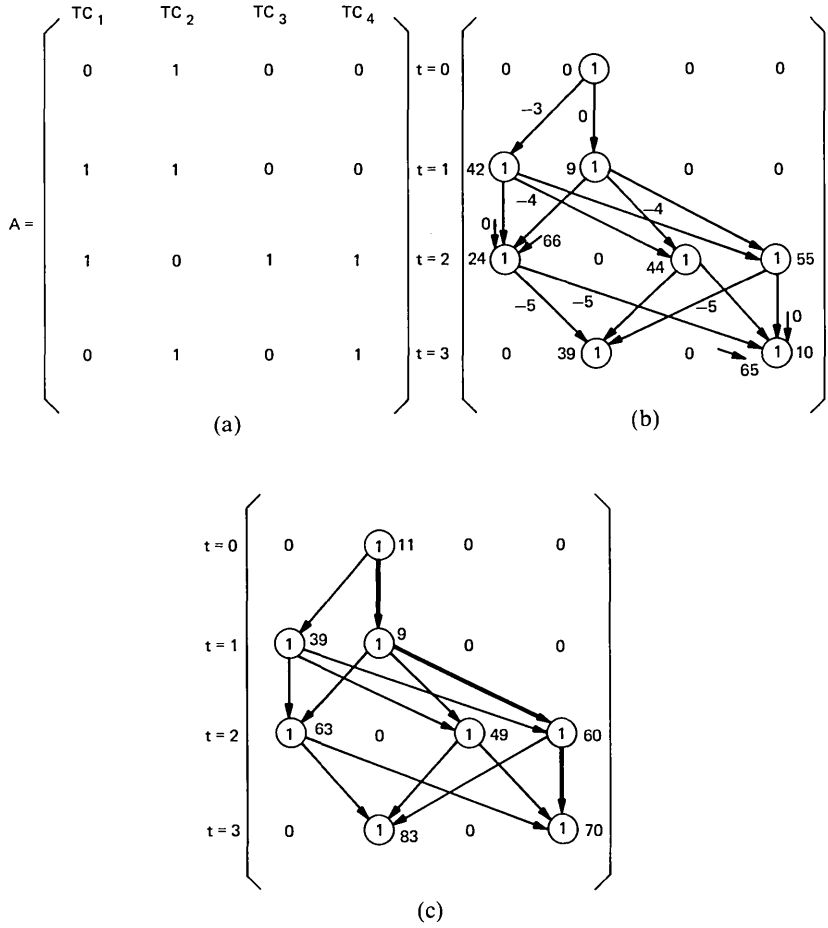


Fig. 3—Construction of the optimal solution in the example. (a) The matrix. (b) The graph with costs of arcs and nodes. (c) Node labels and the optimal path (in heavy lines).

$$f(20) = 4,\ f(24) = 7,\ f(32) = 11,\ f(38) = 13,$$

as indicated in Fig. 1. Also assume that

$$g(20) = 6,\ g(24) = 8,\ g(32) = 10.$$

From the above information the proper node costs can be calculated. For example, for node (1,1) (corresponding to stage $t = 1$ and terminating center TC_1) the proper node cost is

$$f[S(1)]d_1 = (7)(6) = 42,$$

and for node (1,2) (corresponding to stage $t = 1$ and terminating center TC_2) it is

$$\{f[S(1)] - f[S(0)]\}d_2 = (7 - 4)(3) = 9.$$

The costs of arcs and the proper node costs in graph G are written adjacent to them in Fig. 3b. Note that all the vertical arcs have zero cost and all the cross arcs from any stage to the next have the same (negative) cost. The node labels obtained by the above procedure are shown in Fig. 3c. The path distinguished by heavy lines in Fig. 3c is the minimum-cost path corresponding to the optimal solution. Thus, the optimal assignments are as follows:

stage	0	1	2	3
terminating center	TC_2	TC_2	TC_4	TC_4

REFERENCES

1. M. Malek-Zavarei, "Optimal Scheduling of Switching Centers for Telephone Systems," Proc. World Telecommunications Forum, Geneva, Switzerland, Oct. 6–8, 1975.
2. J. J. Pilliod, "Fundamental Plans for Toll Telephone Plants," B.S.T.J., *31* (September 1952), pp. 832–50.
3. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, London: Oliver & Boyd, 1960.
4. M. Malek-Zavarei, "Computer-Aided Design of Telephone Systems," Computers & Electrical Engineering, *4*, No. 1 (January 1977), pp. 37–51.
5. R. S. Garfield and G. L. Nemhauser, *Integer Programming*, New York: Wiley, 1972.
6. L. R. Ford, Jr. and D. R. Fulkerson, *Flows in Networks*, Princeton, N.J.: Princeton University Press, 1962.
7. S. E. Dreyfus and A. M. Law, *The Art and Theory of Dynamic Programming*, New York: Academic Press, 1977.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Stochastic Theory of a Data-Handling System with Multiple Sources

By D. ANICK,* D. MITRA, and M. M. SONDHI

(Manuscript received December 1, 1981)

In this paper we consider a physical model in which a buffer receives messages from a finite number of statistically independent and identical information sources that asynchronously alternate between exponentially distributed periods in the 'on' and 'off' states. While on, a source transmits at a uniform rate. The buffer depletes through an output channel with a given maximum rate of transmission. This model is useful for a data-handling switch in a computer network. The equilibrium buffer distribution is described by a set of differential equations, which are analyzed herein. The mathematical results render trivial the computation of the distribution and its moments and thus also the waiting time moments. The main result explicitly gives all the system's eigenvalues. While the insertion of boundary conditions requires the solution of a matrix equation, even this step is eliminated since the matrix inverse is given in closed form. Finally, the simple expression given here for the asymptotic behavior of buffer content is insightful, for purposes of design, and numerically useful. Numerical results for a broad range of system parameters are presented graphically.

I. INTRODUCTION

1.1 Physical model

A data-handling switch receives messages from many, say N, information sources, which independently and asynchronously alternate between the 'on' and 'off' state. The on periods as well as the off periods are exponentially distributed for each source. These two distributions, while not necessarily identical, are common to all sources;

* Presently at the Department of Mathematics, University of California, Berkeley.

also, the sources are mutually independent. Without loss of generality, the unit of time is selected to be the average on period; with this unit of time, the average off period is denoted by $1/\lambda$. Again, without loss of generality, the unit of information is chosen to be the amount generated by a source in an average on period. In these units an on source transmits at the uniform rate of 1 unit of information per unit of time. Thus, when r sources are on simultaneously, the instantaneous receiving rate at the switch is r. The switch stores or buffers the incoming information that is in excess of the maximum transmission rate, c, of an output channel. (Thus, c is also the ratio of the output channel capacity to an on source's transmission rate.)

As long as the buffer is not empty, the instantaneous rate of change of the buffer content is $r - c$. Once the buffer is empty, it remains so as long as $r \leq c$. We assume that the buffer is infinite and that the following stability condition is satisfied:

$$\frac{N\lambda}{c(1+\lambda)} < 1. \tag{1}$$

The left-hand side is the traffic intensity, ρ.

Discussions with A. G. Fraser[1] suggest that the above is a useful model for a switch in a computer network. In such an application, the output channel rate may be in the range from 5 kb/s to 56 kb/s. For one specific source type, the slow terminals, the message rate may be taken to be 300 b/s, which gives 16 ⅔ and 186 ⅔ as the extreme values of c. A representative value of λ for this source type is 0.4, as computed from the fact that $\lambda/(1+\lambda)$ is the long term on time fraction. The stability condition is satisfied with N as large as 3.5c. Other sources, such as screen terminals and computers, will have quite different statistics. In the interests of generality, we have not placed any further restrictions on the system parameters.

We first derive in a straightforward manner the set of differential equations that governs the equilibrium buffer distribution. We obtain a set of mathematical results that renders trivial the computation of the distribution and its moments and thus also the waiting time moments. The main result explicitly gives all the system's eigenvalues. This is achieved by using the generating function method. Insertion of the boundary conditions in the differential equations requires the solution of a matrix equation, which in many cases of practical interest is dimensionally quite large. However, even this step is eliminated since the matrix inverse is given in closed form. Finally, simple expressions for the moments of the distribution and the asymptotic behavior of buffer content are obtained.

The physical model described above is related to the model in our primary reference, a powerful paper by L. Kosten.[2] Kosten's model is

a limiting case of our model with $N \to \infty$, $\lambda \to 0$ in a manner so as to give a finite traffic intensity, ρ. Also, pooling the instants of commencement of the on periods of all the sources yields, by assumption, a Poisson process. The above physical model and its variants have also been proposed in other papers.[3-6]

1.2 Motivations and discussion of results

Two broad questions supply most of our motivation. In this connection we acknowledge the benefit of several discussions with our colleague A. G. Fraser. The first question concerns the right buffer size to use for a predetermined number of sources and grade of service. The other question, which is of operational significance, concerns the selection of the maximum number of sources to be allowed in the system, the reasoning being that the incremental source disproportionately affects the grade of service for all the sources.

The study of these questions requires that the number of sources in the system, N, be finite. We also examined the conventional belief that the traffic intensity is a reliable indicator of overflow probabilities.

We would like to draw the reader's attention to an important aspect of our problem, namely, numerical stability. Underlying this problem is the fact proven below that the set of linear differential equations governing the behavior of the equilibrium probabilities [see eq. (8)] has 'unstable' eigenvalues. (This is cause for calling the system of equations 'inherently unstable.') Thus, if the boundary conditions are such that any of these modes are excited, then the solution grows at an exponential rate. In the mathematical model this does not happen. However, the situation during computation is quite different. The inevitable errors, no matter how small, incurred during numerical integration are liable to excite the unstable modes and lead to solutions that blow up.

The above observations apply as well when the Laplace transforms of the equilibrium probabilities are available and are to be numerically inverted. The boundedness of solutions is, in principle, obtained by the exact cancellation of unstable factors in the numerator and denominator of the transform. Of course, a straightforward numerical inversion cannot be expected to preserve this feature.

Therefore, it appears inevitable that any method that counters the inherent instability must depend on the a priori segregation of the stable modes from the unstable modes. This in turn depends on the availability of complete information on the eigenvalues and eigenvectors—information that is generally costly to obtain. We compose the solution to eq. (8) in the form

$$\mathbf{F}(x) = \sum_{i:\mathrm{Re}z_i \leq 0} \mathbf{A}_i e^{z_i x}, \qquad x \geq 0, \tag{2}$$

where the z_i's appearing above are a subset of the eigenvalues, and the coefficient vectors $\{\mathbf{A}_i\}$ depend on the boundary conditions and eigenvectors. While the problem of numerical stability does not arise when the solution is computed in the above form, its effectiveness depends on the efficiency of the computation of the eigenvalues and coefficient vectors. An example of the efficiency achieved is that we are able to obtain all the eigenvalues by solving only a set of quadratic equations.

We should mention that Kosten[2] is fully cognizant of the problem of numerical stability. Kosten's solution method consists of obtaining the initial conditions and then numerically integrating the differential equation while continually filtering away the component of the numerical solution that exists in the span of the eigenvectors associated with the unstable eigenvalues.

A noteworthy feature of our primary solution method is that it manages to avoid requiring the numerical solution of matrix equations. This is achieved by avoiding the direct procedure (see Section III) which requires the solution of a dense set of $\lfloor c \rfloor + 1$ linear algebraic equations.[2]* Instead, we require the solution of a typically much larger set of $N - \lfloor c \rfloor$ linear algebraic equations which, however, we obtain in closed form.

Numerical results are discussed in Section VI. We observe substantial departures from Kosten's results in cases where N is small. More generally, we observe for identical traffic intensities, rather different probabilities of overflow for different values of N. We graphically demonstrate the quite acceptable quality of an approximation to the overflow probabilities provided by a relatively simple asymptotic formula. The formula states that, for x large, the probability of buffer content exceeding x behaves as Ae^{-rx}, A some constant and

$$r \triangleq \frac{(1-\rho)(1+\lambda)}{1-c/N}. \tag{3}$$

The positive parameter r is thus, like traffic intensity, a predictor of overflow probabilities. Small values of r may be associated with high probabilities of overflow and low grades of service.

1.3 Mathematical model

If at time t the number of on sources equals i, two elementary events can take place during the next interval Δt, i.e., a new source can start or a source can turn off. Since the on and off periods are exponentially distributed, the respective probabilities are $(N-i)\lambda\Delta t$ and $i\Delta t$. Compound events have probabilities $0(\Delta t^2)$. The probability of no change is $1 - \{(N-i)\lambda + i\}\Delta t + 0(\Delta t^2)$.

---

* We let $\lfloor c \rfloor$ denote the integer part of c. A tacit assumption is that $c < N$, since otherwise the buffer is always empty.

Let $P_i(t, x)$, $0 \le i \le N$, $t \ge 0$, $x \ge 0$, be the probability that at time t, i sources are on and the buffer content does not exceed x. Now,

$$P_i(t + \Delta t, x) = \{N - (i - 1)\}\lambda \Delta t P_{i-1}(t, x) + (i + 1)\Delta t P_{i+1}(t, x) + [1 - \{(N - i)\lambda + i\}\Delta t]P_i\{t, x - (i - c)\Delta t\} + 0(\Delta t^2). \quad (4)$$

Passing to the limit $\Delta t \to 0$:

$$\frac{\partial P_i}{\partial t} + (i - c)\frac{\partial P_i}{\partial x} = (N - i + 1)\lambda P_{i-1} - \{(N - i)\lambda + i\}P_i + (i + 1)P_{i+1}. \quad (5)$$

We are interested only in time-independent, equilibrium probabilities,

$$F_i(x) \triangleq \text{equilibrium probability that } i \text{ sources are on and buffer content does not exceed } x. \quad (6)$$

Therefore, we set $\partial P_i/\partial t = 0$ and obtain, for $i \in [0, N]$,

$$(i - c)\frac{dF_i}{dx} = (N - i + 1)\lambda F_{i-1} - \{(N - i)\lambda + i\}F_i + (i + 1)F_{i+1}, \quad (7)$$

where the understanding is that $F_i = 0$ if i is not in the stated interval. In matrix notation,

$$\mathbf{D}\frac{d}{dx}\mathbf{F}(x) = \mathbf{M}\mathbf{F}(x), \qquad x \ge 0, \quad (8)$$

where $\mathbf{D} = \text{diag}\{-c, 1 - c, 2 - c, \cdots, N - c\}$ and

$$\mathbf{M} = \begin{bmatrix} -N\lambda & 1 & & & & \\ N\lambda & -\{(N-1)\lambda + 1\} & 2 & & & \\ & (N-1)\lambda & -\{(N-2)\lambda + 2\} & 3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 2\lambda & -\{\lambda + (N-1)\} & N \\ & & & & \lambda & -N \end{bmatrix}.$$

Note that the initial conditions to the differential equations are as yet unspecified. Considerations relating to their determination may be found in Section III.

Let*

$$G(x) \triangleq Pr(\text{buffer content} > x) = 1 - \mathbf{1}'\mathbf{F}(x), \qquad x \ge 0. \quad (9)$$

We refer to $G(x)$ as the 'probability of overflow beyond x', or, loosely, as just the 'probability of overflow.' Also observe that

* We let $\mathbf{1}$ denote the vector with unity for all its components and prime denote transposition.

$$F_i(\infty) = \frac{1}{(1+\lambda)^N}\binom{N}{i}\lambda^i, \qquad 0 \le i \le N, \tag{10}$$

since $F_i(\infty)$ is the probability that i out of N sources are on simultaneously. Obviously,

$$\sum_{i=0}^{N} F_i(\infty) = 1.$$

In the analysis to follow, we assume that c is not an integer. (When c is an integer, one of the differential equations in (7) degenerates to an algebraic equation that may be used to eliminate one of the unknown components of $\mathbf{F}$.)

In Ref. 2 the elements of the matrix $\mathbf{M}$ below the diagonal are identical, i.e., independent of row number, and the diagonal element is accordingly adjusted to give column sum 0, as in (8).

The work of Arthurs and Shepp considers various models related to the one considered here; the emphasis of their analysis is on obtaining Laplace transforms of the probabilities.[6] Cohen[7] obtains a broad range of results for the case $c = 1$.

II. EIGENVALUES AND EIGENVECTORS

2.1 Computing the eigenvalues

Let z be some eigenvalue of $\mathbf{D}^{-1}\mathbf{M}$ and let ϕ be the associated right eigenvector. That is,

$$z\mathbf{D}\phi = \mathbf{M}\phi. \tag{11}$$

Equation (11) is also

$$z(i-c)\phi_i = \lambda(N+1-i)\phi_{i-1} - \{(N-i)\lambda + i\}\phi_i + (i+1)\phi_{i+1}, \qquad 0 \le i \le N. \tag{12}$$

Let $\Phi(x)$ denote the generating function of ϕ, i.e.,

$$\Phi(x) \triangleq \sum_{i=0}^{N} \phi_i x^i. \tag{13}$$

By multiplying (12) by x^i and summing over i we expect to obtain an equation in $\Phi(x)$ and $\Phi'(x)$ [for example, $\sum ix^i\phi_i = x\Phi'(x)$]. In fact,

$$\frac{\Phi'(x)}{\Phi(x)} = \frac{zc - N\lambda + N\lambda x}{\lambda x^2 + (z+1-\lambda)x - 1}. \tag{14}$$

In preparing to solve the differential equation, we define r_1 and r_2 to be the distinct real roots, $r_1 > 0 > r_2$, of the quadratic in the denominator of the right-hand side, i.e.,

$$r_1 = \{-(z+1-\lambda) + \sqrt{(z+1-\lambda)^2 + 4\lambda}\}/2\lambda \tag{15a}$$

$$r_2 = \{-(z+1-\lambda) - \sqrt{(z+1-\lambda)^2 + 4\lambda}\}/2\lambda. \tag{15b}$$

Equation (14) may now be written as

$$\frac{\Phi'(x)}{\Phi(x)} = \frac{c_1}{x - r_1} + \frac{c_2}{x - r_2}, \tag{16}$$

where the residues are computed to be

$$c_2 = N - c_1 \tag{17a}$$

$$c_1 = \frac{zc - N\lambda + N\lambda r_1}{\lambda(r_1 - r_2)}. \tag{17b}$$

The solution to (16) is

$$\Phi(x) = (x - r_1)^{c_1}(x - r_2)^{N - c_1}, \tag{18}$$

where, as in the rest of the paper, we have assumed $\phi_N = 1$.

There is an observation on (18) to be made that is central to the present derivation. Observe that by its definition in (13), $\Phi(x)$ is a polynomial in x of degree N. Since r_1 and r_2 are distinct, this is possible if and only if c_1, defined in (17b), is an integer in $[0, N]$. Denoting this integer by k we get

$$\Phi(x) = (x - r_1)^k(x - r_2)^{N-k}, \quad K = 0, 1, \cdots, N. \tag{19}$$

If in (17b) we write k for c_1, use (15) to substitute expressions for r_1 and $r_1 - r_2$, rearrange, and square, then we obtain the following family of quadratics in the unknown eigenvalue z,

$$A(k)z^2 + B(k)z + C(k) = 0, \quad k = 0, 1, \cdots, N$$

where,

$$A(k) \triangleq (N/2 - k)^2 - (N/2 - c)^2$$

$$B(k) \triangleq 2(1 - \lambda)(N/2 - k)^2 - N(1 + \lambda)(N/2 - c) \tag{20}$$

$$C(k) \triangleq -(1 + \lambda)^2\{(N/2)^2 - (N/2 - k)^2\}.$$

We denote by $z_1^{(k)}$ and $z_2^{(k)}$ the two roots associated with the kth quadratic.

To recapitulate, we have shown that all the roots of the above family of $N + 1$ quadratics are eigenvalues, as defined in (11). The reader will find in the following section an enumeration of the properties of the roots and eigenvalues.

We observe that the above argument takes the place of the argument "$\Phi(x)$ should be an entire function of x" employed by Kosten.[2]

2.2 Properties of the roots of the quadratics

The theorem below, which is presented in conjunction with Fig. 1, is a collection of various properties of the roots.

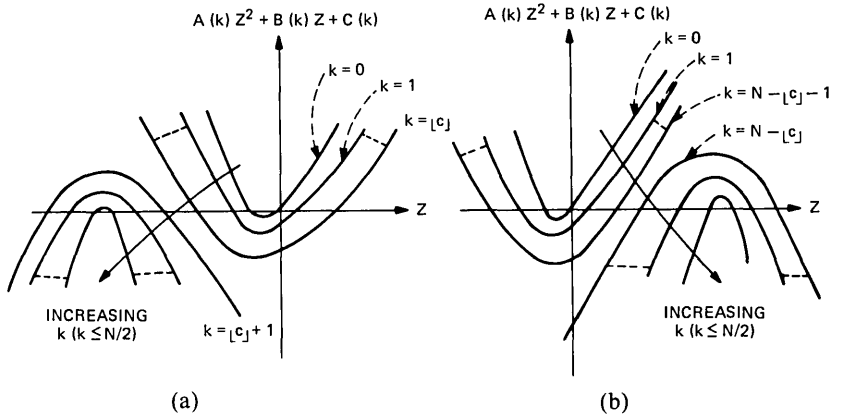


Fig. 1—Sketches of graphs of quadratics, N odd. If N is even, the quadratic for $k = N/2$ has a repeated real root. (a) $c < N/2$. (b) $N/2 < c$.

Theorem

(i) The quadratics for k and k' are identical when $N/2 - k = k' - N/2$.

(ii) For each $k < N/2$ the corresponding quadratic has two real and simple roots. When N is even and $k = N/2$, the corresponding quadratic has a real repeated root.

(iii) $k' < k \leq N/2 \Rightarrow A(k')z^2 + B(k')z + C(k') > A(k)z^2 + B(k)z + C(k), \quad \forall z.$

(iv) The roots of the quadratic corresponding to any k are distinct from those of the quadratic corresponding to any k', provided $k' < k \leq N/2$.

(v) Ignoring the multiplicities, there are $N - \lfloor c \rfloor$ negative roots, 1 root at 0 and $\lfloor c \rfloor$ positive roots. (If the inequality in the stability condition is reversed, then there is 1 less negative root and 1 more positive root.)

(vi) The set of eigenvalues coincide exactly with the set of roots of the quadratics.

(vii) The largest negative eigenvalue is $-(1 + \lambda - N\lambda/c)/(1 - c/N)$. □

The proof of the theorem is given in the appendix.

We will employ the following convention for the eigenvalues:

$$z_{N-\lfloor c \rfloor - 1} < \cdots < z_1 < z_0 < z_N = 0 < z_{N-1} < \cdots < z_{N-\lfloor c \rfloor}. \quad (21)$$

With this convention, z_k and z_{N-k} are roots of the kth quadratic, i.e., $\{z_k, z_{N-k}\} = \{z_1^{(k)}, z_2^{(k)}\}$.

Since we shall later require the stable or negative eigenvalues, let us be explicit about their computation. Let N be first odd and employ the notation $z_1^{(k)} < z_2^{(k)}$. The stable subset is given in braces.

If $c < N/2$:

$$\left\{\begin{aligned} &z_1^{(k)} = [-B(k) - \sqrt{B^2(k) - 4A(k)C(k)}]/2A(k), \\ &\qquad\qquad 0 \le k \le \lfloor c \rfloor. \\ &z_{1,2}^{(k)} = [-B(k) \mp \sqrt{B^2(k) - 4A(k)C(k)}]/2A(k), \\ &\qquad\qquad \lfloor c \rfloor + 1 \le k < N/2 \end{aligned}\right\} \quad (22a)$$

and if $N/2 < c$:

$$\{z_1^{(k)} = [-B(k) - \sqrt{B^2(k) - 4A(k)C(k)}]/2A(k), \qquad 0 \le k \le N - \lfloor c \rfloor - 1.\}. \quad (22b)$$

When N is even, the only change is that the set in (22a) is augmented by $-N(1 + \lambda)/(N - 2c)$, which is one of the repeated roots of the quadratic associated with $k = N/2$.

2.3 Eigenvectors

We mention the procedure for obtaining the eigenvectors from the eigenvalues. Given an eigenvalue z, we compute in order, using (15) and (17), the quantities r_1, r_2, and k. These are used in (19) to yield the coefficients of the given polynomial and thus the eigenvector coefficients. Therefore, for the ith component of the eigenvector*

$$\phi_i = (-1)^{N-i} \sum_{j=0}^{k} \binom{k}{j}\binom{N-k}{i-j} r_1^{k-j} r_2^{N-k-i+j}, \qquad 0 \le i \le N. \quad (23)$$

Of particular subsequent interest are the eigenvectors $\boldsymbol{\phi}_N$ and $\boldsymbol{\phi}_0$, corresponding respectively to the eigenvalues $z_N = 0$ and $z_0 = (1 + \lambda - N\lambda/c)/(1 - c/N)$. The vector $\boldsymbol{\phi}_N$ may either be obtained directly or by noting that it is the appropriately normalized vector of equilibrium probabilities $\mathbf{F}(\infty)$ given in (10). Therefore,

$$\boldsymbol{\phi}_N = \frac{1}{\lambda^N}\left[1, N\lambda, \cdots, \binom{N}{i}\lambda^i, \cdots, \lambda^N\right]' \quad \text{and} \quad \mathbf{1}'\boldsymbol{\phi}_N = \left(\frac{1+\lambda}{\lambda}\right)^N. \quad (24)$$

The vector $\boldsymbol{\phi}_0$ is obtained by following the procedure mentioned in the previous paragraph. We find that

$$r_1 = 1 - \frac{N}{c}, \qquad r_2 = \frac{1}{\lambda} \cdot \frac{1}{N/c - 1}, \qquad k = N,$$

$$\Phi(x) = \left\{x + \left(\frac{N}{c} - 1\right)\right\}^N. \quad (25)$$

* We let $\boldsymbol{\phi}_i$ *and* ϕ_i, respectively, denote the ith eigenvector and the ith component of the generic eigenvector.

Hence,

$$(\phi_0)_i = \binom{N}{i}\left(\frac{N}{c} - 1\right)^{N-i}, \qquad 0 \le i \le N \tag{26}$$

and

$$\mathbf{1}'\phi_0 = \left(\frac{N}{c}\right)^N.$$

This example serves to illustrate a noteworthy point: Recall that for a specific k we may obtain an eigenvalue z using (20); for the same z we may obtain k' from (15) and (17), as outlined in the first paragraph of this section. It may be that $k \neq k'$; however, it is always true that $|N/2 - k| = |N/2 - k'|$. This should not be surprising in view of statement (i) of the theorem, and it is due to the operations, including squaring, that allow us to go from (15) and (17) to (20).

2.4 Left eigenvectors

Sometimes, as in Section 3.3 below, not only the right, but also the left, eigenvectors ψ,

$$z\psi\mathbf{D} = \psi\mathbf{M} \tag{27}$$

are required to be known. It is reasonable to expect that the procedure in Section 2.1 can be repeated to obtain the generating function of the left eigenvectors, but we have not found this approach to be tractable. However, the procedure outlined below may be used.

There is a diagonal matrix τ that symmetrizes $\mathbf{M}$, i.e.,

$$\tau^{-1}\mathbf{M}\tau = (\tau^{-1}\mathbf{M}\tau)'. \tag{28}$$

In fact,

$$\tau_i = \left\{\lambda^i \binom{N}{i}\right\}^{1/2}, \qquad 0 \le i \le N, \tag{29}$$

where τ_i is the ith diagonal element of τ.

Define $\tilde{\psi}$ and $\tilde{\phi}$ to be the left and right eigenvectors obtained when $\mathbf{D}$ is replaced by $\tau^{-1}\mathbf{D}\tau$ and $\mathbf{M}$ by $\tau^{-1}\mathbf{M}\tau$. It is easy to see from the defining relations that

$$\tilde{\phi} = \tau^{-1}\phi \quad \text{and} \quad \tilde{\psi} = \tau\psi. \tag{30}$$

Now notice the important fact that since $\tau^{-1}\mathbf{M}\tau$ is symmetric, it has identical left and right eigenvectors. Hence,

$$\tau^2\psi = \phi$$

or, component-wise,

$$\boxed{\lambda^i \binom{N}{i} \psi_i = \phi_i, \qquad 0 \le i \le N.} \tag{31}$$

Thus, the left eigenvector may be obtained from the right eigenvector.

III. THE SOLUTION

The solution to the differential equations in (8) with $\mathbf{F}(0) = \mathbf{f}$, can be written as

$$\mathbf{F}(x) = \sum_{i=0}^{N} e^{z_i x} \frac{\psi_i' \mathbf{D}\mathbf{f}}{\phi_i' \mathbf{D}\psi_i} \phi_i. \tag{32}$$

It is clear, however, that as only bounded solutions are allowed,

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{i=0}^{N-\lfloor c \rfloor - 1} e^{z_i x} \frac{\psi_i' \mathbf{D}\mathbf{f}}{\phi_i' \mathbf{D}\psi_i} \phi_i, \tag{33}$$

where, according to our convention in (21), the z_i's appearing in the above expression are all negative. The term $\mathbf{F}(\infty)$ in (33) is identical to the $i = N$ term in (32). Recall that $\mathbf{F}(\infty)$ is already known, as shown in eq. (10). With appropriate identification, (33) also may be written as

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{i=0}^{N-\lfloor c \rfloor - 1} e^{z_i x} a_i \phi_i. \tag{34}$$

Our primary solution method developed below in Sections 3.1 and 3.2 depends on the explicit solution of the coefficients a_i in the form appearing in (34). In Section 3.3 we give a second, contrasting, method in which the initial condition vector $\mathbf{f}$ is numerically solved and substituted in (33).

3.1 A key property of the solution at the boundary x = 0

If the number of sources on at any time exceeds c, then the buffer content increases and the buffer cannot stay empty. It follows that

$$F_i(0) = 0, \qquad \lfloor c \rfloor + 1 \le i \le N. \tag{35}$$

By supplementing (35) with the tri-diagonal structure of the matrix $\mathbf{D}^{-1}\mathbf{M}$ we may make further deductions regarding the behavior of $\mathbf{F}(x)$ when x is small. Observe that an application of $\mathbf{D}^{-1}\mathbf{M}$ on $\mathbf{F}(0)$ will diminish by 1 the number of trailing elements that are zero, and that each additional application will have the same effect until $(\mathbf{D}^{-1}\mathbf{M})^{N-\lfloor c \rfloor - 1}\mathbf{F}(0)$ has only its last component equal to zero and $(\mathbf{D}^{-1}\mathbf{M})^{N-\lfloor c \rfloor}\mathbf{F}(0)$ has none. Thus,

$$\{(\mathbf{D}^{-1}\mathbf{M})^j \mathbf{F}(0)\}_i = 0, \qquad \lfloor c \rfloor + 1 + j \le i. \tag{36}$$

Now recollect that on account of the governing differential equations for $\mathbf{F}(\cdot)$ in (8),

$$\mathbf{F}^{(j)}(0) = (\mathbf{D}^{-1}\mathbf{M})^j \mathbf{F}(0). \tag{37}$$

Thus, from (36) we find that

$$F_i^{(j)}(0) = 0, \qquad \lfloor c \rfloor + 1 + j \le i, \tag{38}$$

and, in particular, the following relation, which we shall find most useful:

$$F_N^{(j)}(0) = 0, \qquad j = 0, 1, \cdots, N - \lfloor c \rfloor - 1. \tag{39}$$

Thus, not only is the event "all sources are on and buffer is empty" of probability zero, as already is known from (35), but the growth of the probability is also slow when the buffer content is small.

3.2 Procedure for obtaining the solution

We proceed to obtain the coefficients $\{a_i\}$ in the solution expression (34). Recall that by convention $\{\phi_i\}_N = 1$, so that from (34) we find that

$$F_N(x) = \left(\frac{\lambda}{1+\lambda}\right)^N + \sum_{i=0}^{N-\lfloor c \rfloor - 1} a_i e^{z_i x}, \qquad x \geq 0. \tag{40}$$

The above, taken with (39), implies the following set of equations:

$$\sum_{j=0}^{N-\lfloor c \rfloor - 1} (z_j)^i a_j = -\left(\frac{\lambda}{1+\lambda}\right)^N \delta_{oi}, \qquad 0 \leq i \leq N - \lfloor c \rfloor - 1. \tag{41}$$

Equation (41) in matrix form is

$$\mathbf{V}\mathbf{a} = -\left(\frac{\lambda}{1+\lambda}\right)^N \mathbf{e}, \tag{42}$$

where $V_{ij} = (z_j)^i$, $\mathbf{a} = (a_0, a_1, \cdots, a_{N-\lfloor c \rfloor - 1})'$ and $\mathbf{e} = (1, 0, \cdots, 0)'$.

The key observation is that $\mathbf{V}$ is a Vandermonde matrix. Well known results on such matrices allow us to solve (42) explicitly. Note that $\mathbf{V}$ is nonsingular because the eigenvalues $\{z_i\}$ are distinct,[8] as previously established in the theorem (see Section 2.2). Therefore,

$$|\mathbf{V}| = \prod_{0 \leq i < j \leq N - \lfloor c \rfloor - 1} (z_i - z_j). \tag{43}$$

This formula, applied to the minors, which also are related to Vandermonde matrices, gives

$$\boxed{a_j = -\left(\frac{\lambda}{1+\lambda}\right)^N \prod_{\substack{i=0 \\ i \neq j}}^{N-\lfloor c \rfloor - 1} \frac{z_i}{z_i - z_j}, \qquad 0 \leq j \leq N - \lfloor c \rfloor - 1.} \tag{44}$$

To summarize, the above procedure for obtaining the equilibrium probabilities $\mathbf{F}(x)$ and the probability of overflow $G(x)$ is based on using the expressions

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{i=0}^{N-\lfloor c \rfloor - 1} e^{z_i x} a_i \boldsymbol{\phi}_i \tag{45}$$

and

$$G(x) = -\sum_{i=0}^{N-\lfloor c \rfloor - 1} e^{z_i x} a_i (\mathbf{1}' \boldsymbol{\phi}_i). \tag{46}$$

Only the stable eigenvalues appear in the above forms and they are explicitly given in (22); $\{\phi_i\}$ is obtained from the generating function in (18) and the coefficients $\{a_i\}$ appear in (44). Note that $\mathbf{1}'\boldsymbol{\phi} = \Phi(1)$, so that $\{\phi_i\}$ need not be computed explicitly to compute $G(x)$. $\mathbf{F}(\infty)$ is given in (10).

3.3 An alternative procedure for obtaining the solution

Under certain conditions the following procedure may be considered a viable alternative to the one described above. Here we compute $\mathbf{F}(0) = \mathbf{f}$ and use it in the solution series (33).

Recall from (35) that only the leading $\lfloor c \rfloor + 1$ elements of $\mathbf{f}$ are nonzero and need to be computed. Precisely the same number of equations are forthcoming by requiring of the initial conditions that the $\lfloor c \rfloor$ unstable modes are not excited (compare with Kosten's "illegal eigenvalues"[2]) and that $\mathbf{F}(\infty)$ is normalized, i.e.,

$$\left.\begin{aligned} \boldsymbol{\psi}_i'\mathbf{D}\mathbf{f} &= 0 \qquad i = N-1, N-2, \cdots, N-\lfloor c \rfloor \\ \boldsymbol{\psi}_0'\mathbf{D}\mathbf{f} &= -c + \frac{N\lambda}{1+\lambda} \end{aligned}\right\}. \tag{47}$$

In matrix form,

$$\left.\begin{aligned} \mathbf{f}' &= \{\mathbf{f}_1', \mathbf{0}'\} \\ \text{and } \mathbf{A}\mathbf{f}_1 &= \left(-c + \frac{N\lambda}{1+\lambda}\right)\mathbf{e} \end{aligned}\right\}, \tag{48}$$

where the coefficients in (47) multiplying the unknown $\mathbf{f}$ have been arranged to form the matrix $\mathbf{A}$ in (48). We note parenthetically that $\mathbf{A}$ is not sparse. The $(\lfloor c \rfloor + 1)$-dimensional matrix equation in (48) has to be solved.

Recall that our primary procedure in Section 3.2 and the one shown above require solutions of matrix equations. However, in the former case we were able to explicitly obtain the solution even though the dimension there, $N - \lfloor c \rfloor$, is typically much greater. In the absence of an explicit inverse for A, we expect the above procedure to be useful only for small c.

IV. ASYMPTOTICS

4.1 Probability of overflow

Here we examine the behavior of $G(x)$, the probability of overflow beyond x, for large values of x. The asymptotic formulas obtained are useful for the following reasons: As can be seen from the numerical results in Section VI, they often describe the system behavior rather well in all but the regions of lesser importance, where x is small; also,

the analytic formulas are simple, even though they contain the essential information.

Since the form of the solution in (34) is a sum of exponential terms, the departure of $\mathbf{F}(x)$ from $\mathbf{F}(\infty)$ will be dominated by the exponential with the largest exponent. Hence,*

$$\mathbf{F}(x) - \mathbf{F}(\infty) \sim a_0 \phi_0 e^{-rx} \tag{49}$$

and

$$G(x) \sim -a_0(\mathbf{1}'\phi_0)e^{-rx}, \tag{50}$$

where $-r(=z_0)$ is the largest negative eigenvalue of $\mathbf{D}^{-1}\mathbf{M}$. Using statement (vii) of the theorem, we find that

$$r = \frac{(1 + \lambda - N\lambda/c)}{1 - c/N} = \frac{(1 + \lambda)(1 - \rho)}{1 - c/N} \tag{51}$$

where in the latter, more suggestive form, ρ is the traffic intensity. The coefficient a_0 and the eigenvector ϕ_0 are given in (45) and (26). Collecting terms, we find that

$$\boxed{G(x) \sim \rho^N \left\{ \prod_{i=1}^{N-\lfloor c \rfloor - 1} \frac{z_i}{z_i + r} \right\} e^{-rx}.} \tag{52}$$

In the event that the alternative procedure given in Section 3.3 is used, then the following asymptotic formula is more relevant.

$$G(x) \sim Ae$$

where

$$A = \left(\frac{N - c}{N - 2c + c/\rho} \right)^N \left[\frac{\rho}{c(1 - \rho)} - \frac{1}{N - c} \right] \sum_{i=0}^{\lfloor c \rfloor} \frac{(c - i) f_i}{\lambda^i (N/c - 1)^i}. \tag{53}$$

4.2 Limiting equations for an infinite number of sources

Here we bridge some of the results presented in this paper and Kosten's results. We will show that some of Kosten's important expressions are obtained by passing to the appropriate limit, namely,

$$N \to \infty, \qquad \lambda \to 0 \quad \text{and} \quad N\lambda = \bar{\lambda}. \tag{54}$$

(Our notation is close to Kosten's with one notable exception: $\bar{\lambda}$ is Kosten's λ.) The results obtained in the limit may be interpreted to be applicable when the number of sources grows large and the fraction of time that each source is on decreases in such a manner that the traffic intensity approaches the fixed constant $\bar{\lambda}/c$.

In this section we follow Kosten and normalize the generating

* By $a(x) \sim b(x)$ we mean that $a(x)/b(x) \to 1$ as $x \to \infty$.

function to yield $\phi_0 = 1$ (rather than, as in the rest of the paper, $\phi_N = 1$), so that

$$\Phi(x) = \left(1 - \frac{x}{r_1}\right)^k \left(1 - \frac{x}{r_2}\right)^{N-k}, \tag{55}$$

where r_1 and r_2 are as in (15). It may be shown that, in the limit (54)

$$r_1 \to 1/(z+1) \tag{56}$$

$$r_2 \to -N(z+1)/\bar{\lambda}; \tag{57}$$

and, furthermore, on using (55),

$$\Phi(x) \to e^{x\bar{\lambda}/(z+1)}\{1 - x(1+z)\}^k, \tag{58}$$

which is Kosten's expression for the generating function.

To establish another correspondence, observe that on substituting (17b), the expression for r_1 and r_2 in (15), we obtain

$$k = \frac{zc - N\lambda + N\lambda[-(z+1-\lambda) + \sqrt{(z+1-\lambda)^2 + 4\lambda}]/2\lambda}{\sqrt{(z+1-\lambda)^2 + 4\lambda}}. \tag{59}$$

The limit of the right-hand side may be obtained and found to give Kosten's key equation

$$k = z(cz + c - \bar{\lambda})/(z+1)^2. \tag{60}$$

The correspondence in the eigenvector components is particularly illuminating, showing that

$$\phi_i = \lambda^i \sum_{j=0}^{k} \binom{k}{j}\binom{N-k}{i-j} (-\lambda)^{-j} (r_1)^{i-2j},$$

and, from (23), that

$$\phi_i \to \bar{\lambda}^i \sum_{j=0}^{k} \binom{k}{j} \frac{1}{(i-j)!} (-\bar{\lambda})^{-j} \left(\frac{1}{1+z}\right)^{i-2j}. \tag{61}$$

A final correspondence concerns the important rate parameter r appearing in the asymptotic formulas in Section 4.1. We find that

$$r = \frac{(1+\lambda)(1-\rho)}{1 - c/N} \to 1 - \bar{\rho}, \tag{62}$$

where $\bar{\rho}$ is simply the limiting traffic intensity $\bar{\lambda}/c$.

We should also mention that certain key results of this paper, such as those pertaining to our primary solution method in Sections 3.1 and 3.2, have no parallel in Ref. 2.

V. MOMENTS

We give below expressions for the moments of the equilibrium buffer

content, which allows for their easy computation once the elements of the exponential series solution in (34) are available.

We observe that the nth moment

$$E(x^n) = \int_0^\infty x^n d\{\mathbf{1}'\mathbf{F}(x)\} = n \int_0^\infty x^{n-1} G(x) dx. \tag{63}$$

Since

$$G(x) = - \sum_{i=0}^{N-\lfloor c \rfloor - 1} e^{z_i x} a_i (\mathbf{1}'\boldsymbol{\phi}_i),$$

as we saw in eq. (46), we obtain

$$E(x^n) = \frac{n!}{(-1)^{n+1}} \sum_{i=0}^{N-\lfloor c \rfloor - 1} \frac{a_i(\mathbf{1}'\boldsymbol{\phi}_i)}{z_i^n}. \tag{64}$$

If the alternative procedure given in Section 3.3 is used to obtain the solution, then (64) may again be used with the identification

$$a_i = \frac{\boldsymbol{\psi}_i' \mathbf{D}\mathbf{f}}{\boldsymbol{\psi}_i' \mathbf{D}\boldsymbol{\phi}_i}. \tag{65}$$

VI. NUMERICAL RESULTS

In Figs. 2 through 5 we have held the source statistic λ to a constant

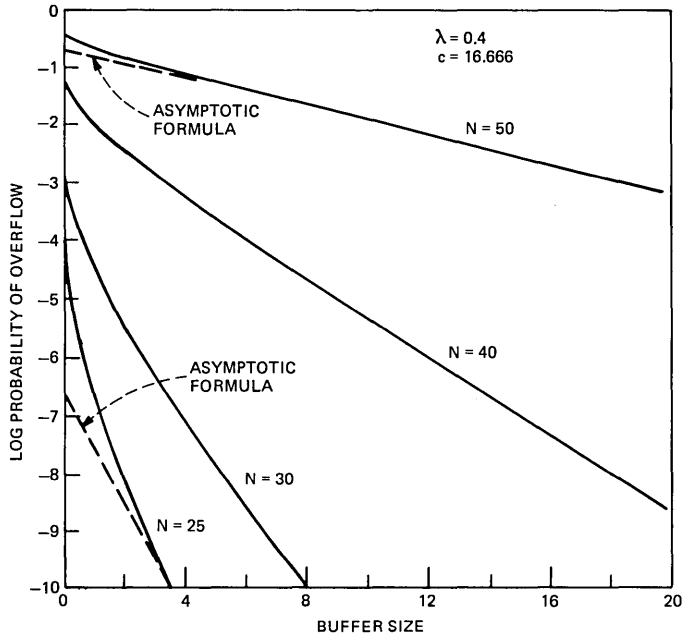


Fig. 2—Probability of overflow vs buffer size with λ and c constant. For N = 25, 30, 40, and 50, the traffic intensity ρ is 0.43, 0.51, 0.69, and 0.86, respectively.

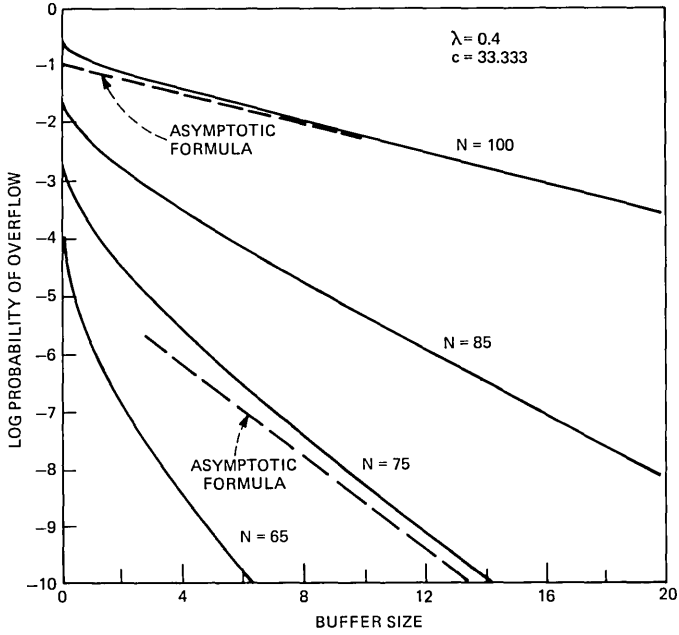


Fig. 3—Probability of overflow vs buffer size with λ and c constant. For N = 65, 75, 85, and 100, the traffic intensity ρ is 0.56, 0.65, 0.73, and 0.83, respectively.

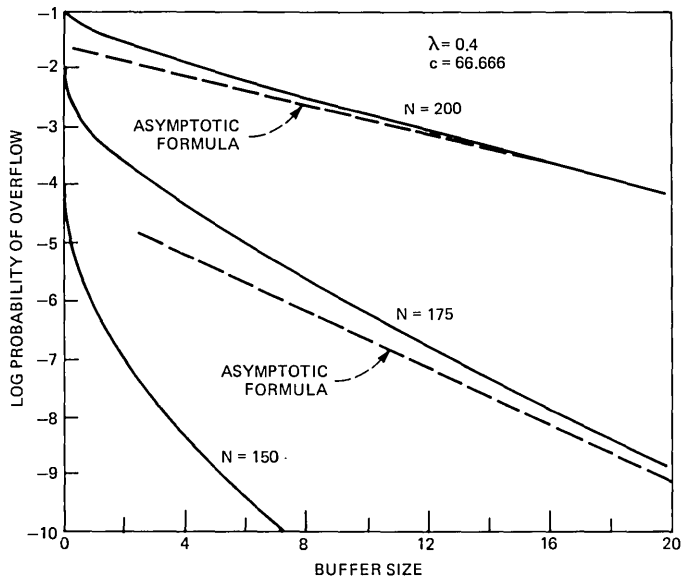


Fig. 4—Probability of overflow vs buffer size with λ and c constant. For N = 150, 175, and 200, the traffic intensity ρ is 0.64, 0.75, and 0.86, respectively.

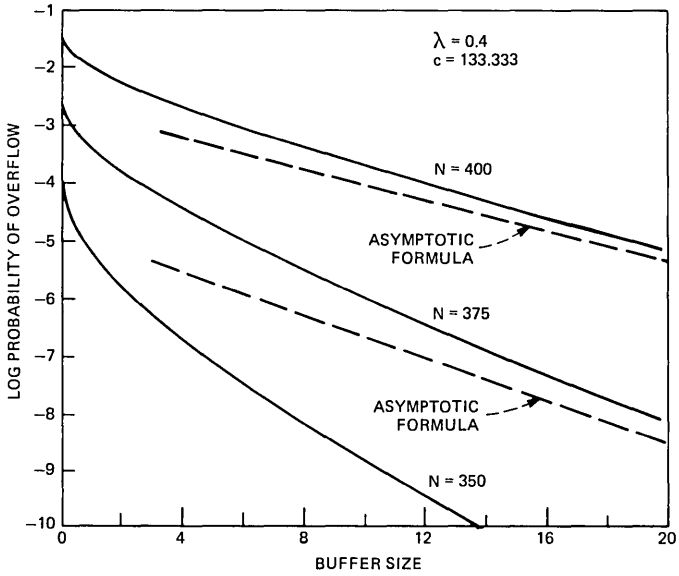


Fig. 5—Probability of overflow vs buffer size with λ and c constant. For $N = 350$, 375, and 400, the traffic intensity $\rho = 0.75$, 0.80, and 0.86, respectively.

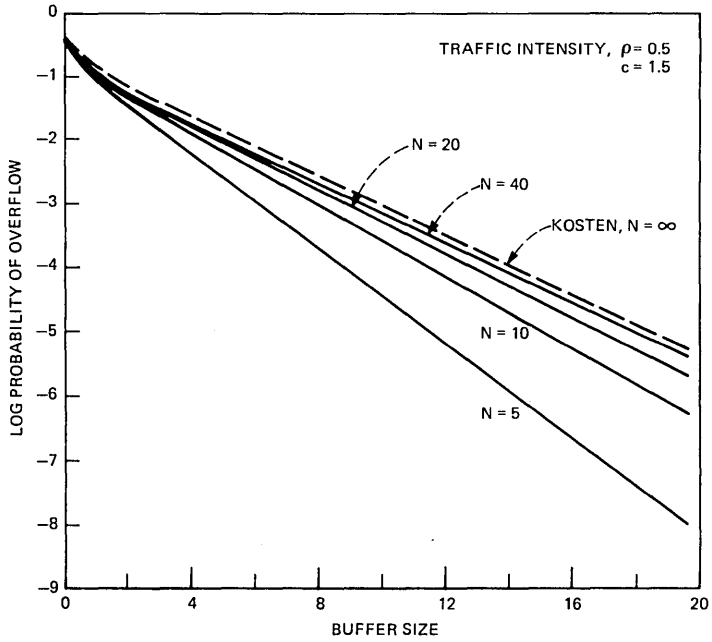


Fig. 6—Probability of overflow vs buffer size with traffic intensity ρ and c constant. For $N = 5$, 10, 20, and 40, the parameter λ is 0.18, 0.08, 0.04, and 0.02, respectively. The curve for $N = \infty$ is from Ref. 2.

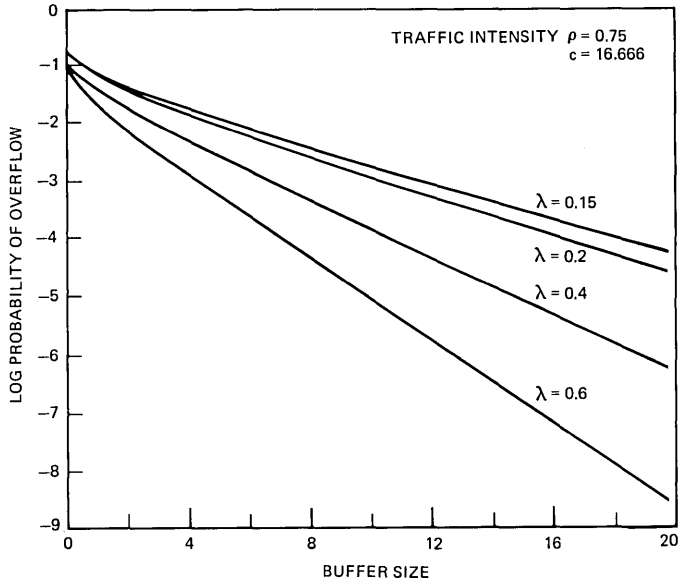


Fig. 7—Probability of overflow vs buffer size with traffic intensity ρ and c constant. For $\lambda = 0.6$, 0.4, 0.2, and 0.15. The number of sources $N = 33$, 44, 75, and 96, respectively.

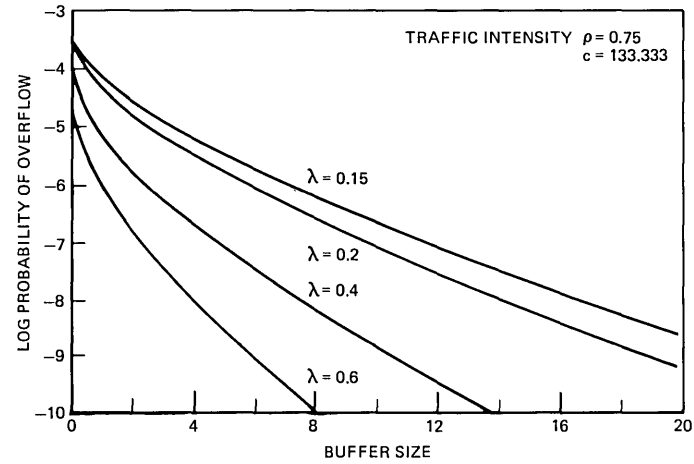


Fig. 8—Probability of overflow vs buffer size with traffic intensity ρ and c constant. For $\lambda = 0.6$, 0.4, 0.2, and 0.15, the number of sources $N = 267$, 350, 600, and 767, respectively.

value, 0.4. Each figure has a distinctive value of c, the ratio of output to input transmission rates. The four values of c are chosen for the cases where an on source transmits at 300 b/s and the output channel rates are 5 kb/s, 10 kb/s, 20 kb/s, and 40 kb/s. These figures show

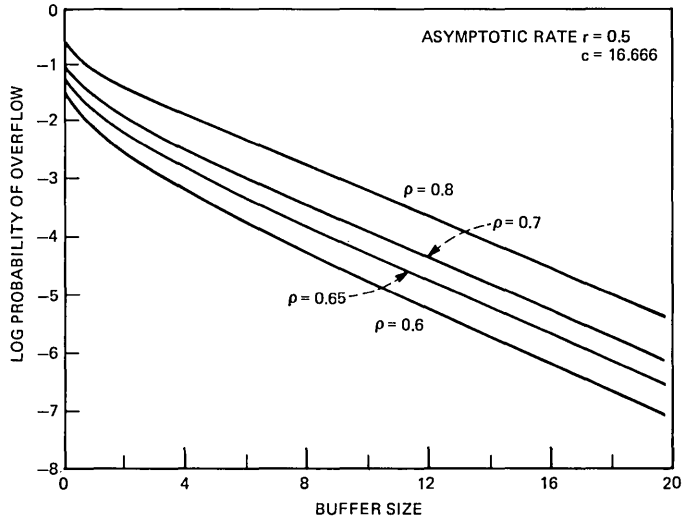


Fig. 9—Probability of overflow vs buffer size for constant asymptotic rate r [see eq. (51)] and c. For $\rho = 0.6$, 0.65, 0.7, and 0.8, the number of sources $N = 127$, 85, 63, and 41, respectively.

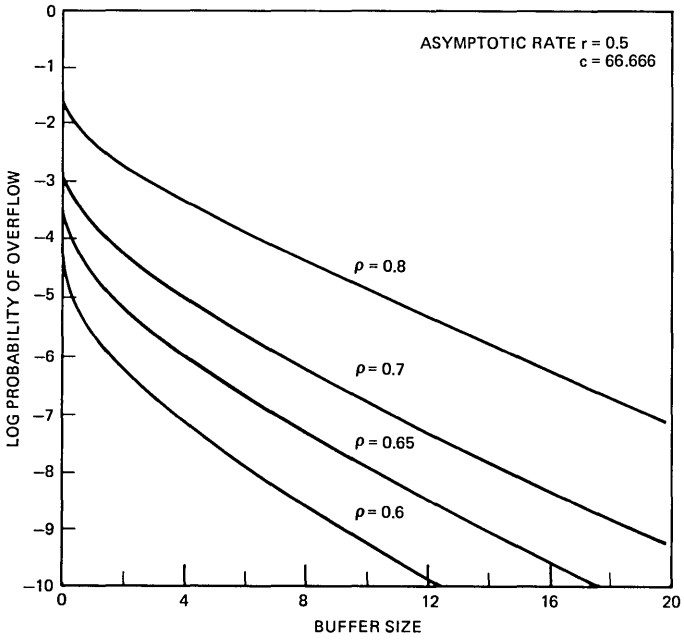


Fig. 10—Probability of overflow vs buffer size for constant asymptotic rate r and c. For $\rho = 0.6$, 0.65, 0.7, and 0.8, the number of sources $N = 507$, 338, 253, and 164, respectively.

rather clearly the effect of incremental sources. For example, we see from Fig. 2 that for a fixed grade of service as given by the probability of overflow = 10^{-5}, a 33-percent increase in the number of sources from 30 to 40 requires about a 300-percent increase in buffer size from 2 to 8. Recall from the discussion in Section 1.1 that the unit of information, and thus of the buffer as well, is the amount generated by one source in the average on period.

Observe in Figs. 2 through 5 the generally acceptable quality of the approximation to the probability of overflow provided by the asymptotic formula in eq. (52).

In each of Figs. 6 through 8 we have a constant traffic intensity, ρ, and a constant ratio of transmission rates, c. Note in particular that any two curves will have different source statistics λ and different N. The figures in this series illustrate the difference between the model considered in this paper and Kosten's limiting model. The figures also demonstrate rather emphatically the limitations of using only the traffic intensity as a predictor of overflow behavior. For example, in Fig. 7 we see that for constant traffic intensity and buffer size = 20, probability of overflow varies from about 10^{-4} to about 10^{-9}, depending on λ.

In Figs. 9 and 10 we examine the proposition that the rate parameter

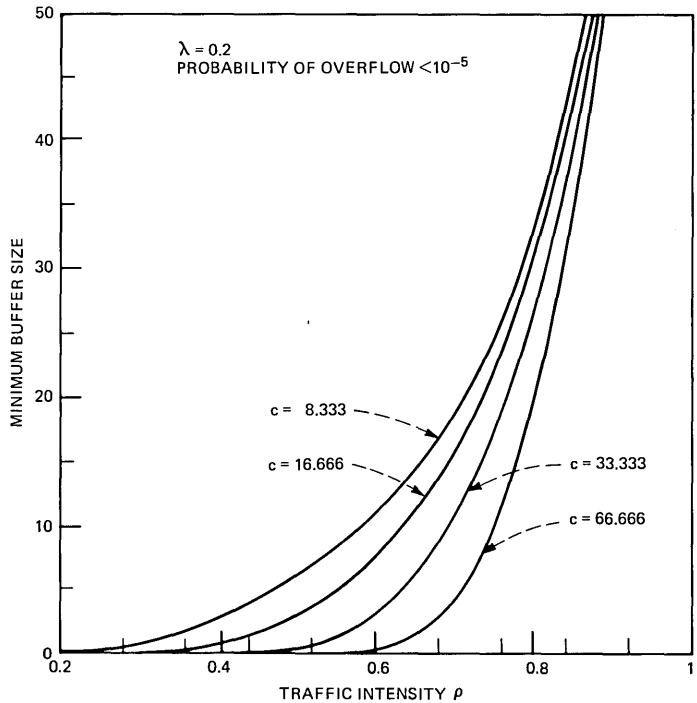


Fig. 11—Minimum buffer size required to satisfy traffic intensity ρ with constrained probability of overflow. The constant is λ.

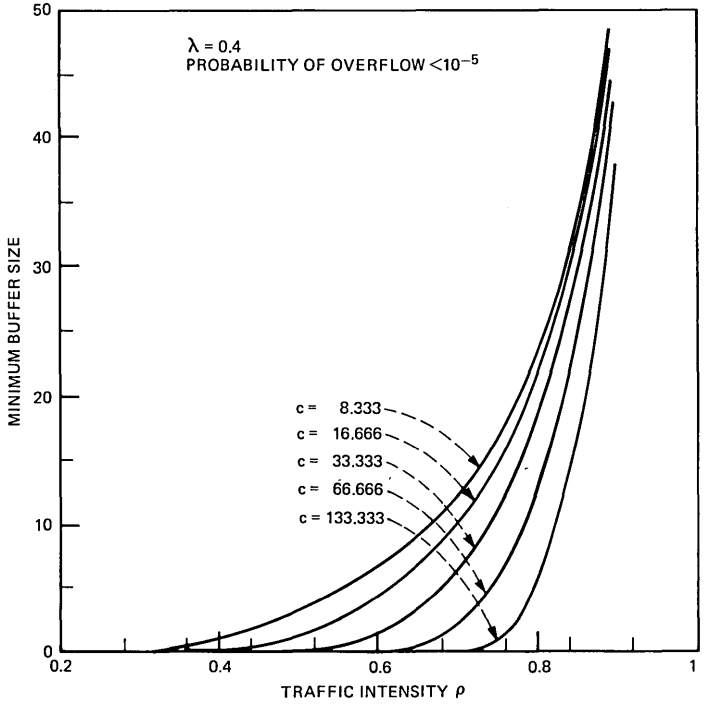


Fig. 12—Minimum buffer size required to satisfy traffic intensity ρ with constrained probability of overflow. The constant is λ.

r, which gives the slopes of the curves obtained from the asymptotic formula, is a useful single index of overflow behavior. Equations (3) and (51) give r. The strength and limitations of the index are contrasted in Figs. 9 and 10.

Figures 11 and 12 are motivated by the design problem in which it is required to estimate the buffer size needed to meet various traffic conditions with a specified grade of service and fixed source statistics. The slackening requirements in the buffer size with increasing c, which are observed in both figures, denote the economies of scale that stem from using higher capacity output channels. In comparing Figs. 11 and 12 we observe that for the same traffic intensity, buffer requirements are less stringent when λ is greater, i.e., when the source type is on for higher fractions of time.

REFERENCES

1. A. G. Fraser, private communication.
2. L. Kosten, "Stochastic Theory of a Multi-Entry Buffer (1)", Delft Progress Report, Vol. *1*, 1974, pp. 10–18.
3. H. Rudin, "Buffered Packet-Switching: A Queue with Clustered Arrivals", Int. Switching Symp. Rec., M.I.T., June 1972.

4. O. Hashida and M. Fujika, "Queueing Models for Buffer Memory in Store-and-Forward Systems", Int. Teletraffic Congress, Stockholm, June 1973, pp. 323/1-323/7.
5. W. W. Chu and L. C. Liang, "Buffer Behavior for Mixed Input Traffic and Single Constant Output Rate", IEEE Trans. on Commun., Vol. *COM-20*, No. 2 (April 1972), p. 230-35.
6. E. Arthurs and L. Shepp, unpublished work, 1979.
7. J. W. Cohen, "Superimposed Renewal Processes and Storage with Gradual Input," Stochastic Processes and Their Applications, Vol. 2 (January 1974), pp. 31-58.
8. R. Bellman, *Introduction to Matrix Analysis*, 2d ed., New York: McGraw-Hill, 1970, p. 193.

APPENDIX

Proof of Theorem

(*i*) Observe that $A(k)$, $B(k)$, and $C(k)$, as given in eq. (20), depend on k only through $(N/2 - k)^2$.

(*ii*) Roots are real if $B^2(k) - 4A(k)C(k) \geq 0$, and distinct as well if the expression is positive. View the expression as a function of $(N/2 - k)^2$. Observe that it is a quadratic in $(N/2 - k)^2$ and, furthermore, that it is concave since the leading coefficient of $(N/2 - k)^4$ is $4(1 - \lambda)^2 - 4(1 + \lambda)^2 < 0$. Thus, the minimum of $B^2(k) - 4A(k)C(k)$ for $(N/2 - k)^2$ in $[0, (N/2)^2]$ is at one of the corner points where the respective values are 0 and positive.

(*iii*) We claim that

$$\{A(k') - A(k)\}z^2 + \{B(k') - B(k)\}z + \{C(k') - C(k)\} > 0, \qquad \forall z. \tag{66}$$

To prove eq. (66) we need to observe that

$$A(k') - A(k) > 0 \tag{67}$$

and

$$\{B(k') - B(k)\}^2 - 4\{A(k') - A(k)\}\{C(k') - C(k)\} < 0. \tag{68}$$

(*iv*) It follows immediately from (*iii*) that the graphs of the quadratics are nonintersecting. See Fig. 1.

(*v*) We first consider $c < N/2$ and later its opposite. Consider in turn $k = 0$, $0 < k \leq c$, and $c < k \leq N/2$.

When $k = 0$, it turns out that $C(k) = 0$, so that the quadratic reduces to

$$A(0)z\{z + B(0)/A(0)\}. \tag{69}$$

Hence

$$z_1^{(0)} = -\frac{B(0)}{A(0)} = \frac{-(1 + \lambda - N\lambda/c)}{1 - c/N} \tag{70}$$

and

$$z_2^{(0)} = 0. \tag{71}$$

Observe that the stability condition implies that $z_1^{(0)} < 0$; reversal of the inequality in the stability condition gives $z_1^{(0)} > 0$.

Now consider $0 < k < c$. Observe that $A(k) > 0$ and $C(k) < 0$, so that $z_1^{(k)} < 0 < z_2^{(k)}$.

Finally, consider $c < k < N/2$. It is easy to see that $A(k) < 0$ and $C(k) < 0$. We show below that $B(k) < 0$, from which it will follow that $z_1^{(k)} < 0$ and $z_2^{(k)} < 0$. If $\lambda \geq 1$ then the defining expression for $B(k)$ in eq. (20) shows that $B(k) < 0$. If $\lambda < 1$ then the following equivalent expression shows that $B(k) < 0$:

$$B(k) = 2(1-\lambda)\left[\left(\frac{N}{2}-k\right)^2 - \left(\frac{N}{2}-c\right)^2\right] - 2c\left(\frac{N}{2}-c\right)\left(1-\lambda+\frac{N\lambda}{c}\right). \quad (72)$$

The above completes the considerations related to $c < N/2$.

For the opposite situation where $N/2 < c$, the case of $k = 0$ is unchanged. For $0 < k < N - c$, it is easy to see that $A(k) > 0$ and $C(k) < 0$, so that $z_1^{(k)} < 0 < z_2^{(k)}$.

Now consider $N - c < k < N/2$. We show below that $B(k) > 0$ (note the contrast with the situation for $c < N/2$), which when taken with $A(k) < 0$ and $C(k) < 0$, which are easy to see, gives $0 < z_1^{(k)} < z_2^{(k)}$. If $\lambda \leq 1$ then the defining expression for $B(k)$ in eq. (20) shows that $B(k) > 0$. Assume now that $1 < \lambda$. Then,

$$\begin{aligned} B(k) &= N(1+\lambda)\left(c-\frac{N}{2}\right) + 2(1-\lambda)\left(\frac{N}{2}-k\right)^2 \\ &> N(1+\lambda)\left(c-\frac{N}{2}\right) + 2(1-\lambda)\left(c-\frac{N}{2}\right)^2 \\ &= 2c\left(c-\frac{N}{2}\right)\left(1-\lambda+\frac{N\lambda}{c}\right) > 0. \end{aligned} \quad (73)$$

(*vi*) From our derivation of the quadratics it is clear that all roots are eigenvalues of $\mathbf{D}^{-1}\mathbf{M}$. As we have isolated exactly $(N+1)$ distinct values for the roots, there cannot be an eigenvalue that is not one of the roots.

(*vii*) It follows from (*iii*) and (*v*) that the largest negative eigenvalue is a root of the quadratic corresponding to $k = 0$. The negative root in this case is $-(1+\lambda-N\lambda/c)/(1-c/N)$, as shown in eq. (70).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Spatial Subsampling in Motion-Compensated Television Coders

By J. D. ROBBINS and A. N. NETRAVALI

(Manuscript received December 9, 1981)

Motion-compensated television coders generate data at a nonuniform rate, which is smoothed by a buffer for transmission over a channel of constant bit rate. Spatial subsampling is one of the methods used to prevent overflow of the buffer, which would otherwise occur for scenes with complex motion from frame to frame. In this paper we evaluate the effects of spatial subsampling on the performance of motion-compensated coders. In particular, we find that, although the quality of motion estimation does degrade in the presence of subsampling, the degradation is not substantial. Use of 2:1 horizontal subsampling, for example, results in bit rates that are 50 percent lower compared with no subsampling for motion-compensated coders. This percentage is approximately the same for the conditional-replenishment coders. Spatial subsampling generally results in blurring of the picture. We describe a technique for adaptive interpolation that results in blurring of only the unpredictable areas of the picture. The subjective quality in the presence of subsampling is thus improved considerably. In conclusion, our techniques for subsampling in a motion-compensated coder reduce the bit rate approximately by the same factor as subsampling in a conditional-replenishment coder, but result in a much better picture quality.

I. INTRODUCTION

Television signals contain a significant amount of frame-to-frame redundancy. Interframe coders attempt to exploit this redundancy by

(*i*) Segmenting each television frame into two parts, one part that is predictable from the previous data, and one part that is unpredictable.

(*ii*) Transmitting two types of information: (a) addresses specifying the location of the picture elements in the unpredictable area, and (b)

information (usually quantized prediction error) by which the intensities of the unpredictable area can be updated.

(*iii*) Matching the coder bit rate to the channel rate. Since the motion in a real television scene occurs randomly and in bursts, the amount of information about the unpredictable area will change as a function of time. It is transmitted over a constant bit-rate channel by, (a) storing it in a buffer prior to transmission to smooth out the transmitted information rate, and (b) using the buffer fullness to regulate the encoded bit rate by varying the amplitude, spatial, and temporal resolution of the television signal. Intensities of the picture elements (pels) in the unpredictable areas are transmitted by predictive coding. In conditional-replenishment coding,[1-4] quantized values of frame difference, element difference, and line difference (or a combination thereof) are transmitted. In motion-compensated coders,[5-9] estimates of interframe translation of objects are obtained, and more efficient predictive coding is performed by taking differences of elements from the previous frame that are appropriately translated. The translation is equal to the displacement of the object. In our previous papers[5-9] we described several methods of displacement estimation and locally adaptive prediction to reduce the bit rate of interframe coders. Displacement estimation methods were recursive, which minimized the motion-compensated prediction error by a steepest-descent algorithm.

As mentioned before, most interframe coders require a buffer to smooth the output of the coder for transmission over a channel. One method of reducing the size of the buffer or preventing buffer overflow is to control the resolution by spatial subsampling. Typically, if the buffer starts to fill rapidly, resolution is decreased; resolution is increased if the buffer is nearly empty. It is not known how to optimally control the resolution of the unpredictable area for a given channel-bit rate and buffer size. However, many excellent resolution-channel algorithms have been designed on a trial and error basis. In this paper, we are concerned with spatial subsampling in motion-compensated coders, for the purpose of

(*i*) Investigating to what extent spatial subsampling adversely affects the recursive-displacement estimation algorithm

(*ii*) Modifying the displacement estimator to increase its efficiency in the presence of subsampling

(*iii*) Evaluating a new algorithm for adaptive interpolation that blurs only the unpredictable area (as compared with the "moving area") in a conditional-replenishment coder

(*iv*) Presenting simulation results on synthetic scenes with known displacement, and real scenes containing complex motion.

Our simulations are restricted to horizontal subsampling by factors

of 2:1 and 4:1. Simulations indicate that spatial subsampling does degrade our displacement estimation. However, the degradation is not very serious. It is known (and confirmed by our simulations) that 2:1 and 4:1 subsampling reduces the bit rates of conditional-replenishment coders approximately by a factor of two and four, respectively. We found that in the case of motion-compensated coders, 2:1 and 4:1 subsampling reduces the bit rates of a motion-compensated coder by similar factors. In most conditional-replenishment coders, subsampling blurs the "moving areas" (i.e., pels for which amplitude of the frame difference is above a certain threshold); our adaptive interpolation algorithm blurs only the unpredictable areas. Thus, improvement in prediction reduces the blurred areas, thereby improving the picture quality.

II. ALGORITHM

In this section, we describe the modifications to our displacement estimation algorithm. It is worthwhile, however, to look at the basic algorithm described in our earlier works.[5] Let $I(\mathbf{x}_k, t)$ denote the intensity of a scene at the kth sample point $\mathbf{x}_k$ in the scanning order of a frame, and let $I(\mathbf{x}_k, t - \tau)$ denote the intensity at the same spatial location in the previous frame. If the scene consists of an object that is undergoing pure translation under uniform illumination, then, disregarding the background,

$$I(\mathbf{x}_k, t) = I(\mathbf{x}_k - D, t - \tau), \tag{1}$$

where $\mathbf{D}$ is the displacement (two-component vector) of the object in one frame interval, τ. The pel-recursive algorithm obtains an estimate of $\mathbf{D}$ (i.e., $\hat{\mathbf{D}}$) by recursively minimizing the square of the displaced-frame difference at the current pel location. The displaced-frame difference $DFD(\cdot, \cdot)$ is defined by

$$DFD(\mathbf{x}_k, \hat{\mathbf{D}}) = I(\mathbf{x}_k, t) - I(\mathbf{x}_k - \hat{\mathbf{D}}, t - \tau)\,. \tag{2}$$

The minimization is performed by a steepest-descent algorithm of the form

$$\hat{\mathbf{D}}_{k+1} = \hat{\mathbf{D}}_k - \tfrac{1}{2}\epsilon\nabla_{\mathbf{D}}[DFD(\mathbf{x}_{k+1}, \hat{\mathbf{D}}_k)]^2\,, \tag{3}$$

where $\nabla_{\mathbf{D}}[\cdot]$ is the two-dimensional gradient with respect to $\mathbf{D}$. Equation (3) can be expanded to

$$\hat{\mathbf{D}}_{k+1} = \hat{\mathbf{D}}_k - \epsilon DFD(\mathbf{x}_{k+1}, \hat{\mathbf{D}}_k)\nabla I(\mathbf{x}_{k+1} - \hat{\mathbf{D}}_k, t - \tau)\,, \tag{4}$$

where $\nabla = \nabla_x$ is the two-dimensional spatial-gradient operator with respect to horizontal and vertical coordinates of vector $\mathbf{x}$. We use a finite-difference approximation for the gradient, which is formed by

element difference, $EDIF_k$, and line difference, $LDIF_k$, using the pel closest to the point $\mathbf{x}_{k+1} - \mathbf{D}_k$ in the previous frame. The above displacement estimator requires multiplication at each iteration, which is undesirable for hardware implementation and is therefore simplified to:

$$\hat{\mathbf{D}}_{k+1} = \hat{\mathbf{D}}_k - \epsilon \operatorname{sgn}|DFD(\mathbf{x}_{k+1}, \hat{\mathbf{D}}_k)| \cdot \operatorname{sgn}|\nabla I(\mathbf{x}_{k+1} - \hat{\mathbf{D}}_k, t - \tau)|, \quad (5)$$

where

$$\operatorname{sgn}(z) = \begin{cases} -1, \text{ if } Z < -T \\ 0, \text{ if } |Z| \leq T \\ +1, \text{ otherwise.} \end{cases} \quad (6)$$

The above recursion to update $\hat{\mathbf{D}}_k$ is carried out only in the moving areas of the current frame, i.e., for those pels where

$$\sum_{j=-p}^{+p} |I(\mathbf{x}_{k+j}, t) - I(\mathbf{x}_{k+j}, t - \tau)| \geq \text{Threshold}.$$

The motion-compensated coder predicts intensity, $I(x_k, t)$, using either the previous frame intensity, $I(\mathbf{x}_k, t - \tau)$, or displaced-previous-frame intensity, $I(\mathbf{x}_k - \hat{\mathbf{D}}_{k-1}, t - \tau)$, based on which predictor results in less error for certain already transmitted neighbors. We note that if a point $\mathbf{x} - \mathbf{D}$ does not lie on the grid formed by the pels, then interpolation is required to evaluate $I(\mathbf{x}_k - \hat{\mathbf{D}}_{k-1}, t - \tau)$. The displacement at either the previous pel or the previous line element is used to form the displaced-previous-frame prediction of the present pel. This allows the receiver to compute displaced-previous-frame predictions without explicit transmission of the displacement. If the magnitude of the prediction error exceeds a predetermined threshold, the coder transmits a quantized version of the prediction error and the necessary addressing information to the receiver.

Above we described the basic motion-compensation algorithm. We now describe its modifications relevant to the spatially subsampled television signal. Although the algorithm we describe below can be applied to signals subsampled in a variety of ways, we restrict ourselves to horizontal (along a scan line) subsampling by a factor of two and four to one. Modifications and details of each component of the motion-compensated system are given below.

2.1 Displacement estimator

Figure 1 shows the pel arrangement used for the updating process of the displacement estimator. Since the subsampled and subsequently interpolated pels contain more noise, they are given less importance in the updating process. This is done in two ways. The frame-difference signal at subsampled pels is given less weight in determining where

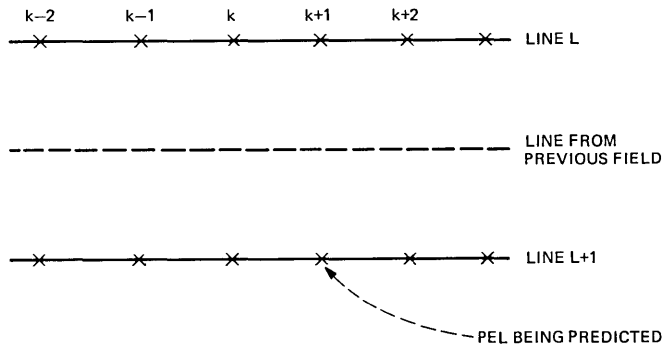


Fig. 1—Pel configuration for a displacement-updating process. Displacement is recursively updated only at those pels where the weighted sum of frame difference in a window around pel k is above a threshold.

the displacement is updated and a smaller updating constant ('epsilon') is used at subsampled pels. Thus, the estimator works as follows:

$$\hat{\mathbf{D}}_k = \hat{\mathbf{D}}_{k-1} - \epsilon SGN|\, DFD(\mathbf{x}_k, \hat{\mathbf{D}}_{k-1})\,| \cdot SGN \left| \frac{EDIF_k}{LDIF_k} \right|, \tag{7}$$

where

$$\epsilon = \begin{cases} \epsilon_1, & \text{if pel } x_k \text{ is subsampled} \\ \epsilon_2, & \text{otherwise .} \end{cases} \tag{8}$$

We have found that the estimator performance is improved by choosing $\epsilon_1 < \epsilon_2$ (both positive numbers). The recursion is carried out only at those pels where a weighted sum of the magnitude of the frame difference in a window around pel k exceeds a given threshold. Thus, $\hat{\mathbf{D}}_{k-1}$ is updated only if

$$\sum_{j=-p}^{p} w_j |\, FDIF(\mathbf{x}_{k+j})\,| \geq THRESH_1\,; \tag{9}$$

otherwise $\hat{\mathbf{D}}_k = \hat{\mathbf{D}}_{k-1}$. The weights $\{w_j\}$ are nonnegative and are lower for subsampled pels compared with the nonsubsampled pels. The threshold, $THRESH_1$, is preselected and optimized by simulations, and the displacement $\hat{\mathbf{D}}_k$ was used for the prediction of the element at location k in the next line (see Fig. 1).

2.2 Predictor and predictor selection

As in our earlier works, we have used only predictors based on intensities in the previous frame. The previous-frame and displaced-previous-frame predictions are calculated for each pel, whether subsampled or not. Of course, other predictions (e.g., intrafield predictors such as previous element or line) can be used to augment our prediction strategy. Having computed both the previous-frame and displaced-

frame predictors, we use the following rule to switch adaptively between them on a pel-by-pel basis. Referring to Fig. 1, we use the frame-difference predictor if

$$\sum_{j=-m}^{+m} w_j |FDIF(\mathbf{x}_{k+j})| \leq \sum_{j=-m}^{+m} w_j |DFD(\mathbf{x}_{k+j}, \hat{\mathbf{D}}_k)|; \tag{10}$$

otherwise we use the displaced-frame predictor. The above inequality is evaluated for a window of size $(2m + 1)$ pels centered around pel k. Again, less weight is given to pels that are subsampled, i.e., w_j is lower for subsampled pels. In the calculation of $FDIF(\cdot)$ and $DFD(\cdot, \cdot)$, sometimes the use of interpolated pels in the previous frame may be required.

2.3 Subsampling and interpolation

We considered several patterns for subsampling. Some patterns were such that they did not change from line to line, field to field, or frame to frame, whereas some were intentionally staggered. Some of these are described in Section 2.4 Having selected a subsampling pattern, we then interpolated the missing pels using an adaptive technique. If the magnitude of the prediction error for the nearest nonsubsampled pels to the right and left was below a threshold, then the intensity of the subsampled pel was replaced by its prediction. If, on the other hand, either the closest right or left nonsubsampled neighbor had prediction-error magnitude above the threshold, then a simple linear (horizontal) interpolation was used to reconstruct the intensity of the subsampled pel. The threshold used is the same as the one that determines whether the quantization error is transmitted. This type of adaptive interpolation improves with the quality of prediction. The conditional-replenishment coders subsample the "moving area" pels, which are determined by the frame difference signal. This has the effect of blurring the entire moving area even if more sophisticated predictors are used. Our strategy, on the other hand, replaces all the "predictable" pels (as determined by the closest neighbors) by their prediction rather than by spatial interpolation. Thus, only the unpredictable areas are blurred by spatial interpolation.

2.4 Transmitted information

As previously mentioned, we transmitted to the receiver the quantized prediction error of every nonsubsampled pel where magnitude of prediction error was above a threshold, called the replenishment threshold. This classifies pels into predictable and unpredictable pels. A 35-level symmetric quantizer was used with representative levels at 0, 3, 6, 11, 16, 21, 28, 35, 44, 53, 64, 77, 92, 109, 128, 149, 178, and 197

(on an 8-bit scale of 0 to 255). We obtained decision levels by averaging the adjacent representative levels. A code set for representing the quantizer levels was not designed, but entropies were computed. In addition, horizontal run lengths of predictable and unpredictable pels were transmitted. Here again, entropies of the predictable- and unpredictable-pel run lengths were calculated. This assumes that in practice separate code sets will be used for run lengths of predictable and unpredictable pels.

III. SIMULATIONS AND RESULTS

Computer simulations were performed on two types of scenes. The first was a synthetic scene that was computer-generated. It was a damped radial cosine in intensity with a radius of 60 pels, which translated from frame to frame by a given amount. The pattern is described mathematically by the intensity function

$$I(R) = 100 \cdot \exp(-0.01R)\cos(2\pi R/P),\ 0 \leq 60\,,$$

where R is the radial distance from the center [taken to be (100,100)] and

$$P = (1 - R/60)10 + 10\,.$$

This function is displayed on a 256- by 256-element raster in two interlaced fields of 128 lines each. The pattern is shown as Fig. 6 in Ref. 7. The other scene, called Judy, is a head and shoulders view of a person engaged in active conversation. This consisted of 50 frames obtained by taking a Nyquist-rate sampling of a video signal having a 1-MHz bandwidth. Each sample was quantized uniformly to 8 bits. Four frames of this sequence are shown in Fig. 4 of Ref. 5.

3.1 Synthetic Scene

The simulations on the synthetic scene were restricted to 2:1 subsampling with a subsampling pattern that did not change from line to line and field to field (referred to as a nonstaggered pattern). The purpose of this simulation was to evaluate the degradation in the performance of the displacement estimator. Therefore, only the displacement was calculated, without using it for coding. Figures 2 and 3 show the relative displacement error as a function of the iteration number. The iteration number in this case refers to only those instances where the displacement was actually updated. Owing to several factors [e.g., the setting of the threshold in eq. (9)], a given iteration number in a subsampled case may not be at the same location in the nonsampled case. Figure 2 shows the case when the pattern moves at 4 pels per frame, whereas Fig. 3 shows the results for displacement of 5 pels per frame. The parameters of the displacement estimator (of

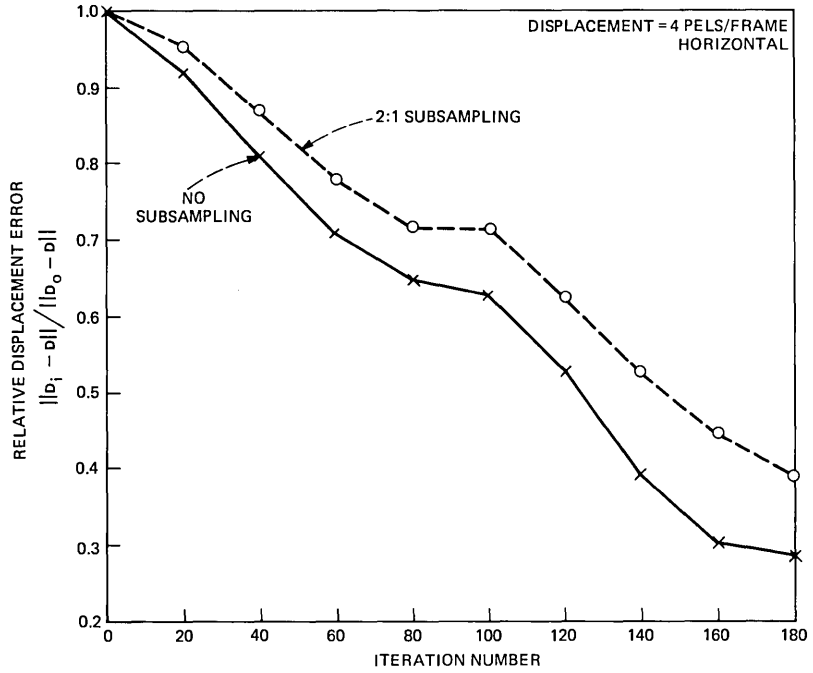


Fig. 2—Relative error in displacement for a synthetic moving pattern at 4 pels per frame as a function of iteration number.

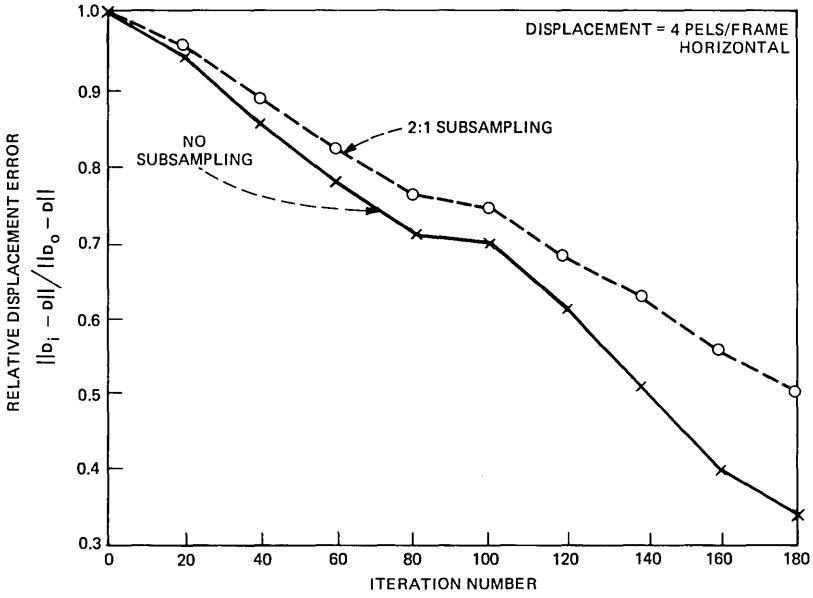


Fig. 3—Relative error in displacement for a synthetic moving pattern at 5 pels per frame as a function of iteration number.

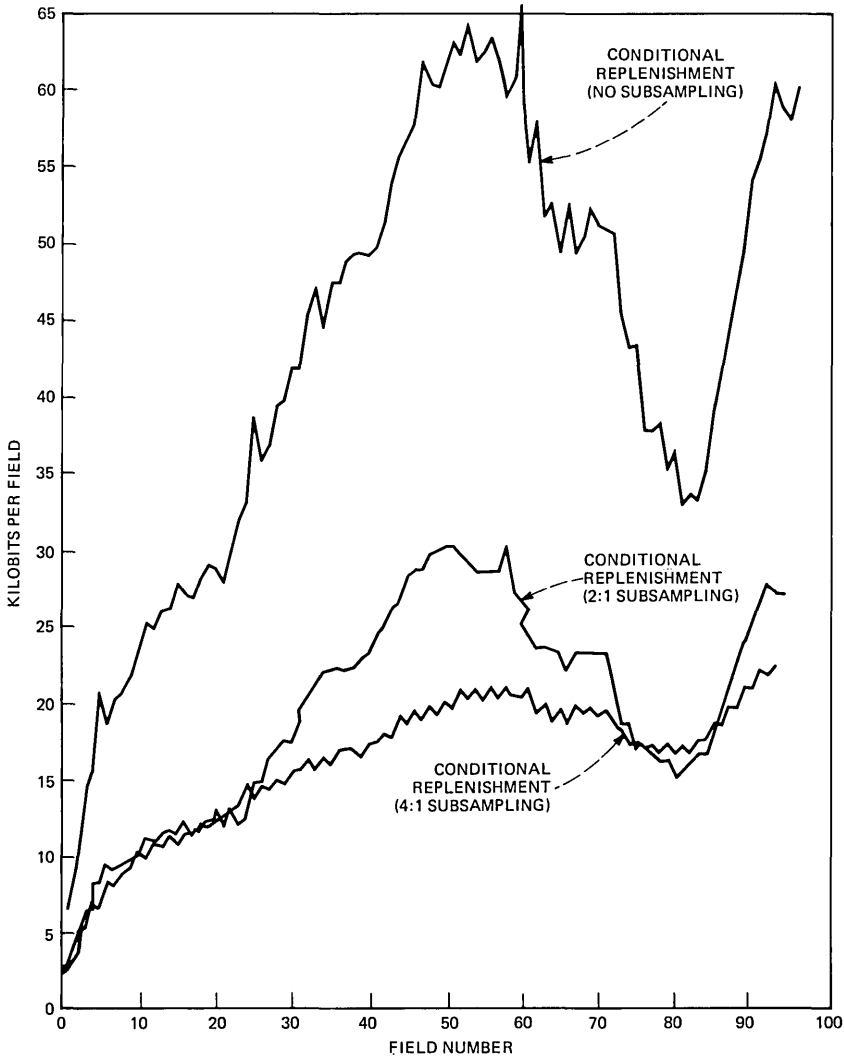


Fig. 4—Bits required per field for conditional-replenishment scheme using a fine (511-level) quantizer.

Section II) were selected by trial and error. It is seen from both figures that the convergence of the displacement estimator is more rapid when there is no subsampling. The degradation appears to be somewhat less for the 5-pel/frame case as compared with the 4-pel/frame case. The effect of this degradation on the motion-compensated coder is evaluated in Section 3.2.

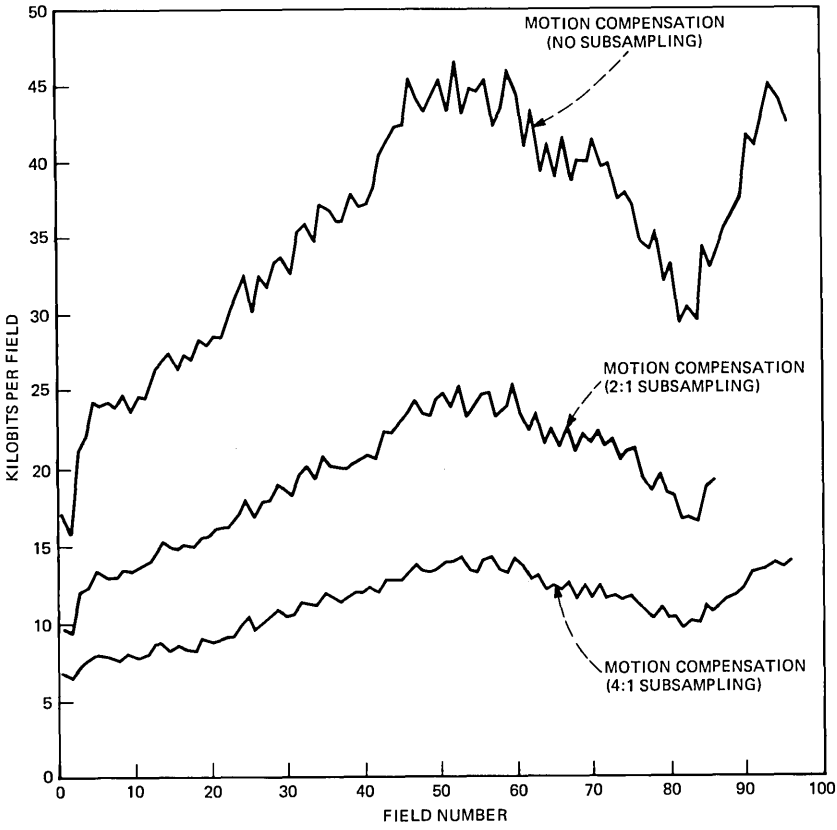


Fig. 5—Bits required per field for motion-compensation scheme using a fine (511-level) quantizer.

3.2 Real Scene

Before the performance of the subsampled motion-compensated coders is given, it is important to note the parameters that were chosen for simulations. These choices were based on trial and error. While this may not have resulted in optimum choices, our choices may not be too far from the optimum.

The epsilon of eq. (8) was different for subsampled pels compared with nonsubsampled pels; ϵ_1 was 1/64, and ϵ_2 was 1/32. The parameters in the update condition of eq. (9) were taken to be: $p = 1$, and $THRESH_1 = 4$ (on a scale of 0 to 255, 8-bits).The weight given to subsampled pels was 1 and to nonsubsampled pels was 2. Thus, through ϵ and these weights, subsampled pels were given less "importance" in the displacement estimation process. The predictor selection [eq. (10)] was done by using a window of 3 (i.e., $m = 1$), and the subsampled pels were given weight $w_j = 1$, whereas nonsubsampled pels were given weight

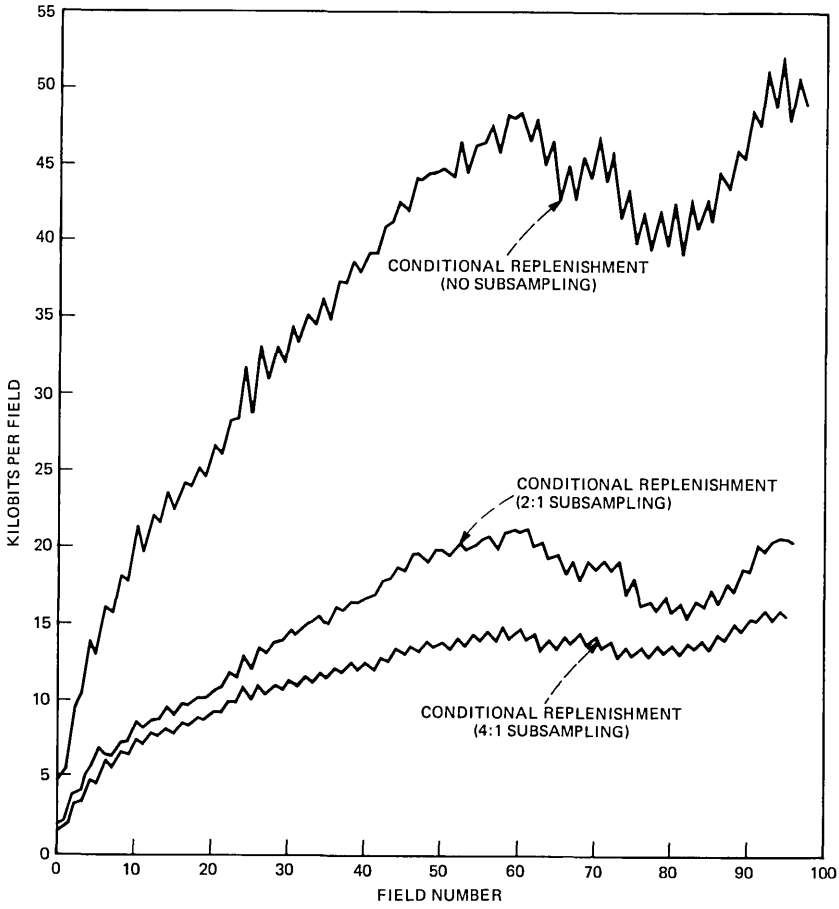


Fig. 6—Bits required per field for conditional-replenishment scheme using a coarse 35-level quantizer.

$w_j = 4$. The displacement estimator was not initialized at the beginning of each scanning line; thus, the estimate from the last pel of the previous line was used as the initial estimate for the first pel of the next line.

The performance was measured by the number of coded bits that were required for each field (approximated by appropriate entropies). When the coarse quantizer of Section 2.4 was used, then the total squared-interpolation error was also calculated for each field. Figures 4 and 5 show the coded bits for both conditional replenishment and motion compensation, when no coarse quantization was performed (i.e., quantizer with 511 levels was used). It is obvious that 2:1 subsampling reduces the bits/field by approximately two for both conditional

replenishment and motion compensation. The decrease in bit rates is high for those fields with a large amount of motion. The 4:1 subsampling reduces the bit rates of motion-compensation schemes much more significantly compared with conditional replenishment. However, this may be a peculiarity of the particular scene we used for simulation. We used a few more scenes and found that 4:1 subsampling, in general, reduced the rate by a factor of two compared with 2:1 subsampling for both conditional replenishment and motion compensation.

Using the 35-level quantizer mentioned earlier, the bit rates are plotted in Figs. 6 and 7. Figure 6 shows conditional replenishment and Fig. 7 shows motion compensation. These figures indicate that motion compensation results in approximately 60-percent reduction for no subsampling, approximately 80-percent reduction for 2:1 subsampling, and about 90-percent reduction for 4:1 subsampling, compared with

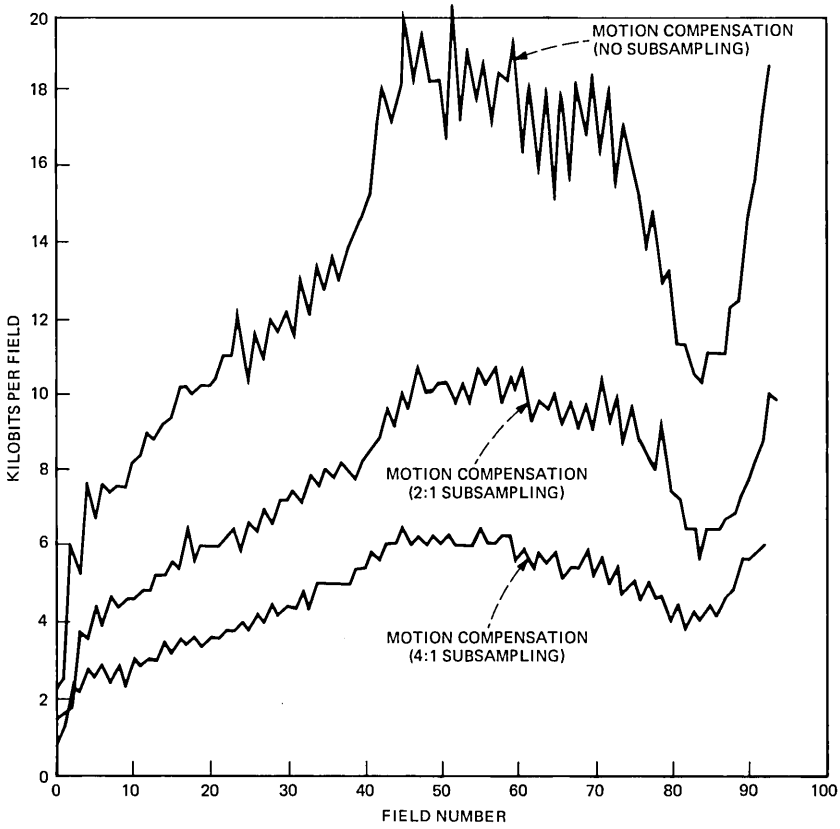


Fig. 7—Bits required per field for motion-compensation scheme using a coarse 35-level quantizer.

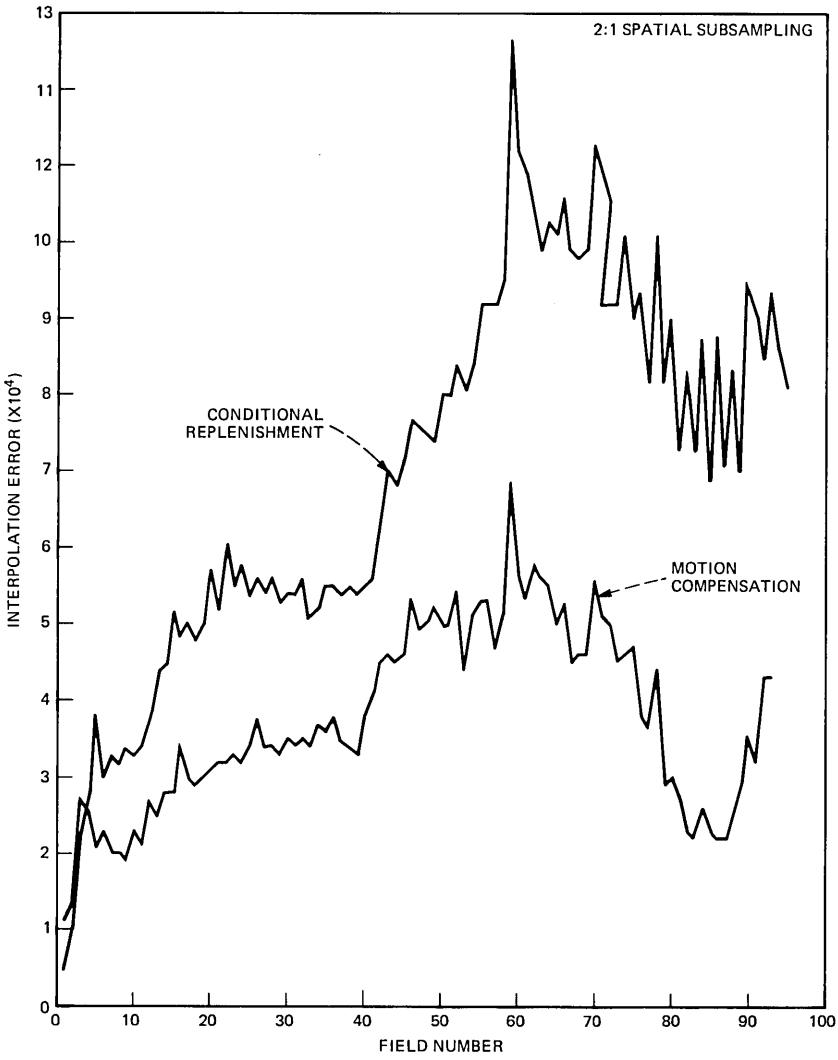


Fig. 8—Plot of squared-interpolation error summed over entire field versus field number of 2:1 spatial subsampling. Adaptive interpolation is used in the case of motion compensation. Fixed-linear one-dimensional spatial interpolation is used for conditional replenishment.

conditional replenishment. Obviously these conclusions are scene-dependent. However, for scenes containing significant translational motion, such conclusions may remain valid. It is always difficult to evaluate quality of short segments of scenes. We made informal observations to compare pictures resulting from conditional replenishment and motion compensation. Subsampling results in visible blur-

ring. However, it was found that, owing to our adaptive interpolation scheme, motion compensation blurred a much smaller area than did conditional replenishment. Also, the blurred areas in motion compensation appear to be much more fragmented and somewhat randomly distributed, which also decreases their visibility. Figures 8 and 9 show the plots of squared-interpolation error per field versus the field number for both 2:1 and 4:1 subsampling. Curves for both frame-difference conditional replenishment and motion compensation are shown. In the case of 2:1 subsampling, the interpolation error decreases by almost a factor of two using motion compensation. This decrease is even greater for 4:1 subsampling.

In our simulations we also tried staggering the subsampling pattern

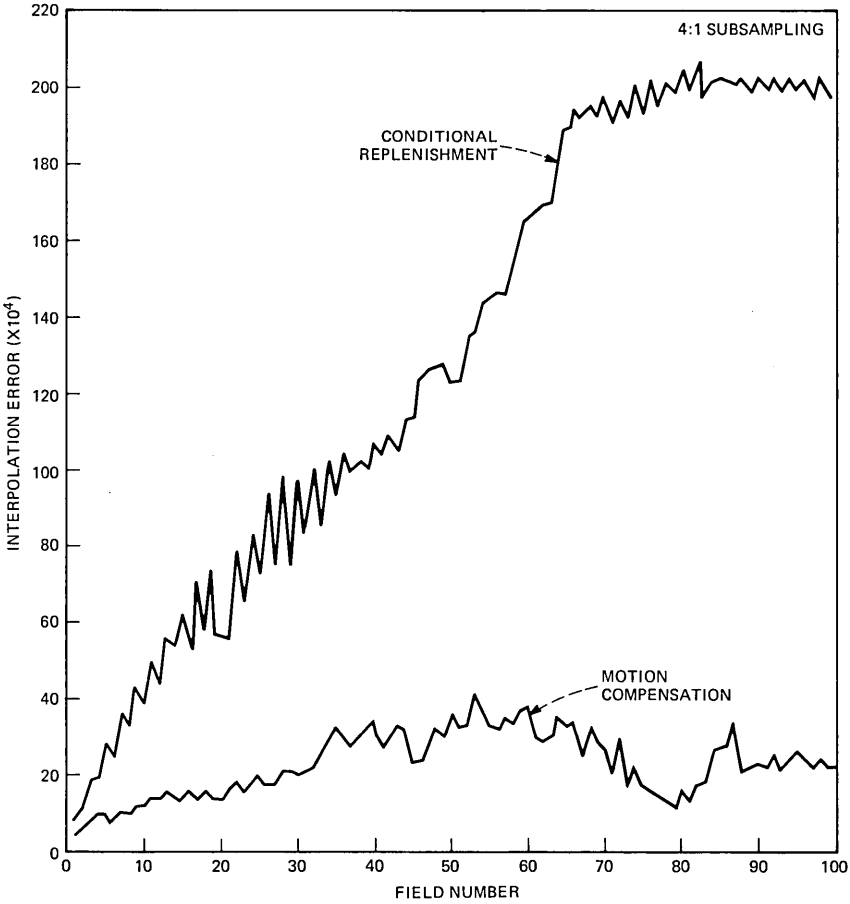


Fig. 9—Plot of squared interpolation error summed over an entire field versus field number for a 4:1 spatial subsampling.

from line to line, field to field, and frame to frame. This makes the quantization and the interpolation noise appear random and less patterned. However, field-to-field or frame-to-frame staggering results in annoying flicker, as expected. We thus found line-to-line staggering to be most useful. Although staggering improved the quality of pictures for both conditional replenishment and motion compensation, the improvement was somewhat higher for conditional replenishment. The required bits per field did not change significantly because of staggering. Thus, we might conclude that line-to-line staggered-subsampling patterns improve the quality of pictures without any significant increase in the bit rates.

IV. SUMMARY AND CONCLUSIONS

We have presented in this paper schemes for motion compensation in the presence of spatial subsampling, which is required in interframe coders to prevent buffer overflow.We also described an adaptive interpolation scheme that blurred only the "unpredictable" area during subsampling. Computer simulations were performed in synthetic scenes to evaluate degradation of the displacement estimator in the presence of subsampling. It was found that, although the quality of displacement estimation was degraded in the presence of spatial subsampling, the effect on the bit rates was not significant. Compared with conditional replenishment, motion compensation reduced the bit rates by a factor of two or more even during subsampling. Thus, a 2:1 subsampled motion-compensated coder results in about one quarter of the bit rate of conditional replenishment and about one half of the bit rate of the 2:1 subsampled conditional-replenishment coder. Our adaptive interpolation scheme blurs only the "unpredictable area" rather than the "moving area" that is blurred in subsampled conditional-replenishment coders. Since "unpredictable area" is a subset of moving area and is fragmented randomly, the blurring caused by subsampling in the case of a motion-compensated coder is much less visible. This is also borne out by the total-interpolation error, which decreases by more than a factor of two using adaptive interpolation.

REFERENCES

1. F. W. Mounts, "A Video Encoding System Using Conditional Picture-Element Replenishment," B.S.T.J., *48,* No. 7 (September 1969), pp. 2545–54.
2. B. G. Haskell, F. W. Mounts, and J. C. Candy, "Interframe Television Coding of Videotelephone Pictures," Proc. IEEE, *60,* No. 7 (July 1972), pp. 792–800.
3. T. Ishiguro, K. Iinuma, Y. Iijima, T. Koga, S. Azaini, and T. Mune, "Composite Interframe Coding of NTSC Color Television Signals," 1976 Nat. Telecommun. Conf. Rec., Vol. *1,* Dallas, Texas, November 1976, pp. 6.4-1 to 6.4-5.
4. B. G. Haskell, "Frame Replenishment Coding of Television," in *Image Transmission Techniques,* W. K. Pratt, ed., New York: Academic Press, 1978.
5. A. N. Netravali and J. D. Robbins, "Motion-Compensated Television Coding: Part 1," B.S.T.J., *58,* No. 3 (March 1979), pp. 631–70.

6. J. D. Robbins and A. N. Netravali, "Interframe Coding Using Movement Compensation," Int. Commun. Conf. (June 1979), pp. 23.4/1-5.
7. J. A. Stuller and A. N. Netravali, "Transform Domain Motion Estimation," B.S.T.J., *58,* No. 7 (September 1979), pp. 1673–703.
8. A. N. Netravali and J. A. Stuller, "Motion-Compensated Transform Coding," B.S.T.J., *58,* No. 7 (September 1979), pp. 1703–18.
9. A. N. Netravali and J. D. Robbins, "Motion-Compensated Coding: Some New Results," B.S.T.J., *59,* No. 4 (November 1980), pp. 1735–45.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

The Detection of Long Error Bursts During Transmission of Video Signals

By K. JANAC and N. J. A. SLOANE

(Manuscript received January 4, 1982)

In this paper we describe a simple threshold code for monitoring digital communication channels which, with high probability, will detect the presence of an error burst of length at least 70, say, but will ignore shorter bursts. This code has applications to digital systems, such as video-conferencing systems, which have an existing error-correction procedure that can handle small numbers of random errors and bursts of length less than some fixed number, but which are adversely affected by longer bursts.

I. INTRODUCTION

Sophisticated techniques are available for removing redundancy in TV pictures.[1,2] These reduce the cost of transmitting video signals over telephone lines, as, for example, in video-conferencing systems.[3,4] The compressed video signal is represented in digital form and is protected by coding[5] against the most frequently occurring channel errors.[1] Certain other, less common, error patterns, however, are also detrimental to the system, causing the picture to disappear momentarily, a phenomenon that is annoying to the viewer. In this paper we describe a coding scheme that detects a class of error patterns that contributes to this degradation of the system. The scheme can also be used for other digital transmission systems where similar coding procedures are employed.

II. MODEL OF A DIGITAL VIDEO-COMMUNICATION SYSTEM

Figure 1 shows a simplified model of a digital video-communication system, as used, for example, in video-conferencing services.[1-3] This model emphasizes the error-correcting mechanism. Video data (a) from the camera is first passed through a picture coder or data compressor (b) that removes some of the redundancy from the picture.

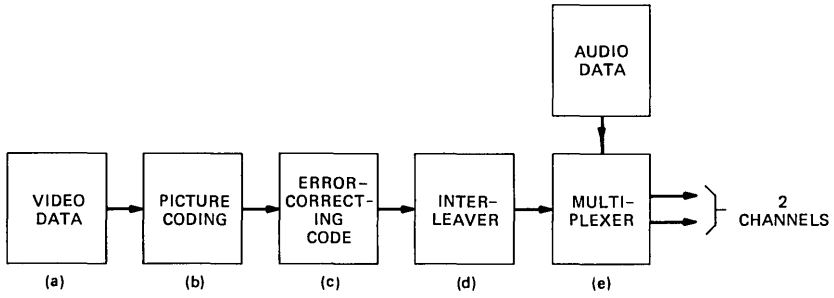


Fig. 1—Error-correcting mechanism in a typical video communication system.

The resulting bit stream is encoded (c) by an error-correcting code, a typical code being the [255, 239, 5] double-error-correcting Bose-Chaudhuri-Hocquenghem (BCH) code.[6] The output from the encoder is interleaved (d), say, to a depth of 35. (This is explained in greater detail below.) Finally, a multiplexer (e) combines the video data from the interleaver with audio and control information, and distributes the result among two high-speed (typically 1.544 Mb/s) channels for transmission. At the receiving end the inverse operations are performed. The audio and control data have a much lower rate than the video data, and can usually be ignored in the analysis of errors. (Other systems use one or even four transmission channels, but it is not difficult to show that our proposed coding scheme applies equally well to these situations.)

To understand how errors on one of the channels affect the picture, we must consider how the interleaver and multiplexer together distribute the bits among the channels. This is explained in Fig. 2. The interleaver consists of an array (b_{ij}) of 35 x 255 bits (Fig. 2a). Code words from the BCH encoder enter the array by rows and are read out by columns. Figure 2b shows the same bits distributed among the two channels. We see that 35 bits now separate any pair of adjacent bits in a code word.

Thus an error burst* of length ≤ 34 on a channel will affect at most one bit from a code word, a burst of length 35 to 69 at most two bits, a burst of length 70 to 104 at most three bits, and so on. If not more than two bits in a code word are in error, this will be corrected by the BCH code. Thus error bursts of length ≤ 69 on any channel are not significant. On the other hand, bursts of length ≥ 70 may cause three or more errors in a code word, and this cannot be corrected by the BCH

* Definition of burst: Suppose $x_1, \cdots, x_n$ is transmitted and $y_1, \cdots, y_n$ is received. The difference $e_1, \cdots, e_n = y_1 - x_1, y_2 - x_2, \cdots, y_n - x_n$ is the error pattern. If the vector $e_1, \cdots, e_n$ is 0 except for a string of length b, which begins and ends with a 1, we say that a burst of length b has occurred. For example, 00110101000 contains a burst of length 6.

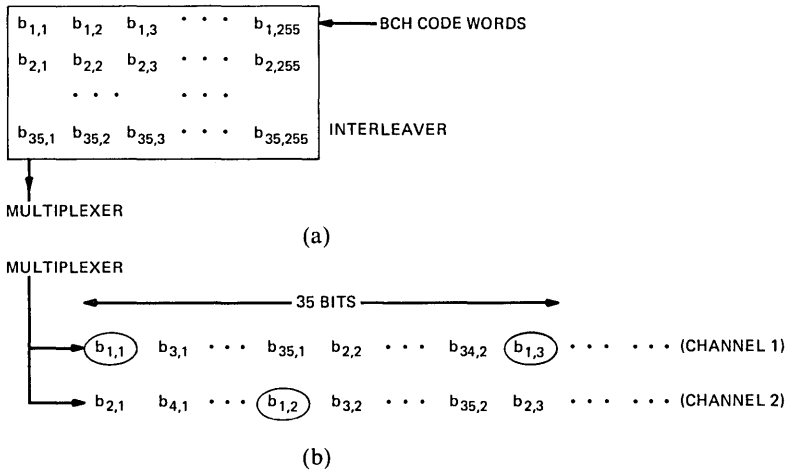


Fig. 2—The interleaver and multiplexer distributing bits among channels. (a) 35 BCH code words are loaded by rows into the array in the interleaver. (b) The same bits after they have been read out of the array by columns and multiplexed between the two channels. Bits from the first row of the array are marked with circles.

code. However, in most cases, the presence of such an error pattern will be "detected," either by the BCH code or by the parity checks associated with the picture coder. When this happens, the receiver normally substitutes the appropriate portion of the previous frame for the missing segment. But if too many errors of this type are detected in one frame, the whole frame must be retransmitted, instead of, for example, just sending the difference from the previous frame. This leaves the screen momentarily blank, which is disturbing to the viewer and is what we are trying to avoid. The penalty for a burst of length greatly exceeding 70 is much higher, for more rows of the array will be affected.

The problem then is to detect whether either channel has suffered from an error burst of length greater than 70. When such a channel has been identified, the appropriate corrective action can be taken (possibly including switching the transmission to a backup channel if the long bursts occur frequently).

The problem is somewhat unusual, since most existing burst detection schemes (compare with Ref. 7) are designed to detect the presence of any burst of length up to some number b_1, while here we are only concerned with detecting bursts of length greater than b_1.

III. BURST DETECTING CODE

The code we propose is a simple threshold scheme, to be used separately on each channel. There are many possible variations of this

code, but we shall just describe the simplest version. The data is usually transmitted down each channel in blocks of some fixed length, say 4200. We wish to detect the presence of an error burst of length ≥ 70, occurring either inside a block or overlapping two adjacent blocks.

3.1 Encoding

Our encoding procedure takes the last 70 bits of the previous block and the 4200 bits of the current block, a total of $4270 = 70 \times 61$ bits, which we denote by

$$u_1, u_2, \cdots, u_{4270},$$

and forms the following check sums:

$$v_1 = u_1 + u_{71} + u_{141} + \cdots + u_{4201},$$

$$v_2 = u_2 + u_{72} + u_{142} + \cdots + u_{4202},$$

$$\cdots$$

$$v_{70} = u_{70} + u_{140} + u_{210} + \cdots + u_{4270},$$

the addition being carried out modulo 2. In other words, the 4270 bits are divided into 61 subblocks of length 70 and the modulo 2 sum $\mathbf{v} = (v_1, \cdots, v_{70})$ of these subblocks is determined, without carries. This check vector $\mathbf{v}$ is then transmitted along with the current block. (Of course, parity checks have been used since the beginning of coding theory, but as far as we know, their use in this particular configuration is new.)

3.2 Decoding

When the bit stream on this channel reaches the receiver, the check sums are recomputed from the same 4270 bits, producing a vector $\mathbf{w} = (w_1, \cdots, w_{70})$, say. The decoder determines the number N of places where $\mathbf{w}$ differs from the received version of $\mathbf{v}$. Of course, if there are no errors, then $\mathbf{w} = \mathbf{v}$ and $N = 0$. The decoder compares N with a fixed threshold θ. If $N \leq \theta$, the decoder decides that the channel is acceptable (and there is only a low probability that a burst of length ≥ 70 has occurred). If $N > \theta$, the decoder decides that there is a significant probability that a burst of length ≥ 70 has occurred, and requests that the appropriate corrective action be taken. The basic configuration of the scheme is illustrated on one channel by the block diagram shown in Fig. 3. A similar arrangement is realized in all four channels (two in each direction). The implementation of this scheme clearly presents no difficulties.

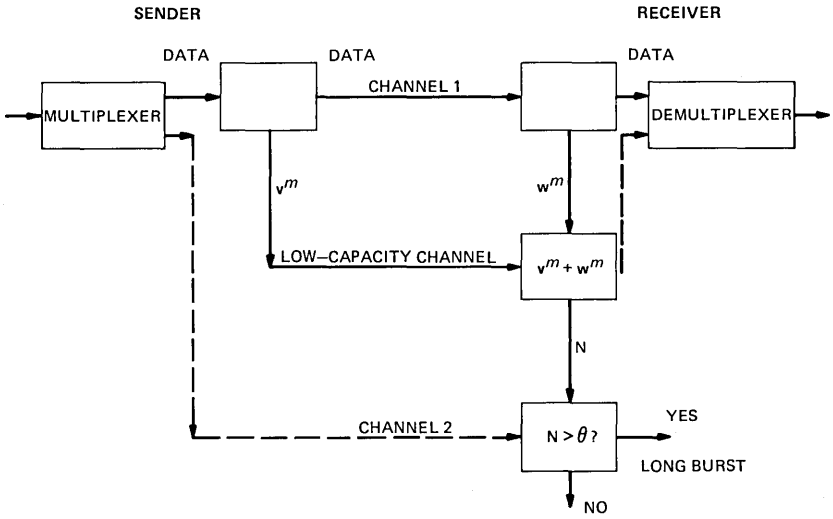


Fig. 3—Simplified block diagram of the proposed scheme for one channel showing the mth data block $\mathbf{w}^m$.

3.3 Choice of Threshold

The following argument shows that an initial threshold of $\theta = 22$ is a good, conservative choice. (However, θ can be left as a variable parameter in the system.) Let us assume that a random burst of errors of length b will begin at any point in the bit stream, and will complement each of the following b bits with probability 1/2. On the average, $b/2$ bits will be wrong. If b is less than 70, each of the errors will affect just one check bit, and so there will be about $b/2$ places where $v_i \neq w_i$, implying $N \approx b/2$. On the other hand, if $b \geq 70$, all the check bits will be involved, and we can expect $N \approx 35$. However, these are just the average values of N, and the actual values will be different. The variance σ^2 of N is roughly $b/4$, which is 17.5 when $b = 70$. Therefore, an initial threshold of $\theta = 35 - 3\sigma \approx 22$ is recommended. The optimal value can then be determined by experiment. A more precise determination of the threshold cannot be made at this time owing to the uncertainty about the density of 1's inside a burst and of the distribution of burst lengths.

The problem is essentially one of hypothesis testing, of distinguishing between the "null hypothesis" that the channel is acceptable and the hypothesis that a burst of length ≥ 70 has occurred. Let us assume that the burst lengths have an exponential distribution, i.e., that a burst of length $b = 0, 1, 2, \cdots$ occurs with probability

$$c_1 e^{-\lambda b},$$

where λ is a parameter and $c_1 = 1 - e^{-\lambda}$. Then two important quantities associated with our code are the "false alarm" probability, which is

$$\alpha = \text{Prob}(N > \theta \mid \text{channel is good}) = \sum_{b=\theta+1}^{69} c_1 e^{-\lambda b} \sum_{N=\theta+1}^{b} \binom{b}{N} \frac{1}{2^b}, \tag{1}$$

and the probability of an undetected burst, which is

$$\beta = \text{Prob}(N \leq \theta \mid \text{channel is bad}) = \sum_{b=70}^{\infty} c_1 e^{-\lambda b} \sum_{N=0}^{\theta} \binom{70}{N} \frac{1}{2^{70}}. \tag{2}$$

If λ is known, these two probabilities can be calculated easily from eqs. (1) and (2) and the best value of θ chosen accordingly. Presumably, β should be made much smaller than α, but this decision must be made by the system designer.

Some possible variations on this scheme follow: (i) **v** could be sent over a separate, slower channel. (ii) The number of check bits, i.e., the length of **v**, could be made greater than 70, thus supplying more information about the state of the channel. [Equations (1) and (2) then need to be slightly modified.] (iii) The decoder could use two thresholds, and declare the channel good if $N < \theta_1$, bad if $N > \theta_2$, and questionable if $\theta_1 \leq N \leq \theta_2$.

IV. SUMMARY

The existing error-detecting procedures for digital communication channels do not distinguish long error bursts from other types of errors, even though in the case of compressed video signals these long bursts contribute to fast retransmissions of the whole frame, leaving the screen momentarily blank, which is annoying to the viewer. We have proposed a simple threshold code that is compatible with the present video-compression and communication-channels systems, and will detect such long bursts with high probability. Formulas are given for the probability of an undetected burst and of a false alarm. The code can be used for monitoring the burstiness of digital communication channels for other services.

REFERENCES

1. B. G. Haskell, P. L. Gordon, R. L. Schmidt, and J. V. Scattaglia, "Interframe Coding of 525-Line Monochrome Television at 1.5M bits/sec," IEEE Trans. Commun., *Com-25* (November 11, 1977), pp. 1339–48.
2. "New Technique Could Allow Video Transmission Over Phone Lines," Bell Lab. Rec., *58*, No. 9 (October 1980), pp. 302–3.
3. E. F. Brown, J. O. Limb, and B. Prasada, "A Continuous Presence Video Conferencing System," Nat. Telecomm. Conf. Rec., *3*, Birmingham, Alabama, December 4–6, 1978, pp. 34.1.1–34.1.4.

4. N. Mokhoff, "The Global Video Conference," IEEE Spectrum, *17,* No. 9 (September 1980), pp. 45–7.
5. N. Iinuma et al., "NETEC-6: Interframe Encoder for Color Television Signals," NEC Res. and Develop., *44* (January 1977), pp. 92–5.
6. F. J. MacWilliams and N. J. A. Sloane, "The Theory of Error-Correcting Codes," 3rd printing, Amsterdam: North-Holland Publishing Company, 1981.
7. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2d ed., Cambridge, MA: MIT Press, 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Considerations for Single-Mode Fiber Systems

By K. OGAWA

(Manuscript received January 12, 1982)

The intrinsic low-dispersion and low-attenuation properties of single-mode fibers between 1.3 and 1.6 μm make them attractive for use in high-capacity, long-haul digital systems. In this paper we discuss some fundamental performance limitations—such as attenuation, bandwidth, mode-partition noise, burst-type error, and receiver sensitivities—for systems with bit rates above 274 Mb/s. Also, we discuss the maximum capacity achievable by either using a single channel at the minimum-dispersion wavelength, or multiple wavelength-multiplexed channels with equal, but necessarily lower, bit rates. We conclude that the characteristics of present laser diodes limit repeater spacing to lengths far less than the potential capacity expected from single-mode fibers. For total capacity of bit rates less than 1 Gb/s, wavelength multiplexing is found to offer longer repeater spacings than single-wavelength systems.

I. INTRODUCTION

The intrinsic low-dispersion property of single-mode fibers makes them attractive for high-capacity, long-haul lightwave systems,[1,2] especially in the wavelength region between 1.1 and 1.7 μm, where low attenuation (less than 1 dB) has been demonstrated.[3] Minimum dispersion can be achieved over a range of wavelengths above 1.3 μm by controlling Ge-doping density and core size.[4-6] Low dispersion over a wide range of wavelengths can be achieved by introducing a certain amount of waveguide dispersion.[7,8] As in any communication system, there are three basic origins of system-performance limitations and degradations: (*i*) sources, (*ii*) media, and (*iii*) receivers. Some limitations, such as receiver sensitivity and source-output power, are determined by a single component. However, many other performance limitations and degradations are caused by the interaction of two components.

Laser diodes as sources have several characteristics that result in

Table I—Degradation factors of laser diodes

	Characteristics of Laser	Related Fiber Characteristics	Degradation Effects to System
Center wavelength	Manufacturing variation ±3 nm		Bandwidth (with fiber dispersion)
	Temperature variation 0.5 nm/°C		Bandwidth (with fiber dispersion)
	Mode skipping (jumps) 6 ~ 10 Δλ		Burst error
Spectrum (half-rms width)	Half rms width 2 ~ 4 nm	Chromatic dispersion	Bandwidth of fiber
	Spectrum broadening		
	Mode partition ($0 \leqq k \leqq 1$)		Mode-partition noise
Intrinsic noise	Quantum noise Transverse mode fluctuation	N.A. and Δ (coupling to fiber)	Noise Power fluctuation
Reflection	Spectrum noise Self-pulsing	Fiber end Pigtail length	Noise
Bandwidth	1 ~ 2 GHz		Bandwidth of transmitter

severe limitations on the performance of a single-mode fiber system. As indicated in Table I, these characteristics, such as spectral distribution,[6] longitudinal mode-partition effects,[9-11] and abrupt jumps[10,12] of the longitudinal or transverse lasing modes,[10] cause a degree of system degradation that depends on the characteristics of the fiber (chromatic dispersion). There are also fundamental performance limitations and degradation factors for single-mode fiber systems caused by fiber attenuation and bandwidth,[13,14] mode-partition noise,[9-11,15] and limited receiver sensitivity.[16-19]

Because of these limitations and their mutual interactions, it is a problem when designing a lightwave system with a given bit rate per fiber to decide whether to use a single high-speed digital signal or several lower-speed channels multiplexed on different wavelength lightwaves. This paper will discuss this problem and develop some general guidelines for solving it.

II. PERFORMANCE LIMITATIONS AND DEGRADATIONS

Table I shows detrimental aspects of laser diodes and Table II shows degradation factors caused by fibers. In this paper, we will focus on five fundamental factors that limit system performance: (*i*) fiber attenuation, (*ii*) fiber bandwidth, (*iii*) mode-partition noise, (*iv*) burst-type errors caused by mode jumps, and (*v*) receiver sensitivity.

Table II—Characteristics of fiber

Loss		
Intrinsic loss	● Infrared absorption ● Ultraviolet absorption ● Rayleigh scattering	● Wavelength dependency
OH-absorption		● 1.24 μm ● 1.38 μm
Imperfection of waveguide	● Boundary of core and cladding ● Bubbles ● Core size variation	Refractive index Difference dependency
Microbending	● Coating ● Stress	● Core size and refractive index difference ● Wavelength dependency
Splicing loss		
Dispersion		
Material dispersion	● Doping material	● Wavelength dependency
Waveguide dispersion	● Core size and refractive index difference	● Wavelength dependency

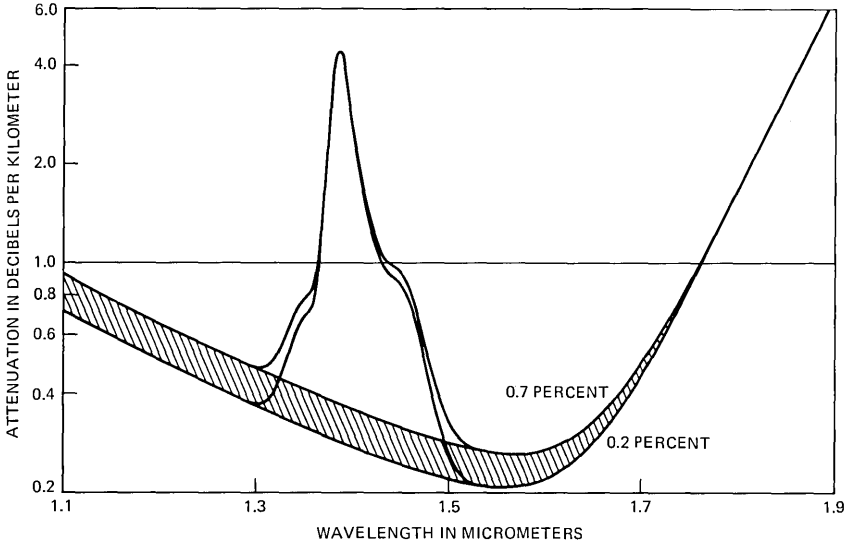


Fig. 1—Fiber attenuation spectrum of a Ge-doped core single-mode fiber.

2.1 Fiber attenuation

Figure 1 shows the theoretical attenuation of single-mode fiber (Ge-doped core) caused by Rayleigh scattering, infrared absorption, and ultraviolet absorption. The crosshatched area in Fig. 1 shows the theoretical limitation (refractive index difference $\Delta = 0.2$ to 0.7 percent) without the water absorption peaks at 1.38 and 1.24 μm. The water

absorption peaks at 1.24 and 1.38 μm are expected to be eliminated eventually by improvement of the dehydration process. There are additional attenuation factors such as (*i*) imperfection losses of the waveguide,[20] (*ii*) microbending losses,[21] and (*iii*) splicing losses.[21,22] Considering these additional losses, probably 0.5 dB/km total fiber loss at 1.3 μm is the theoretical limit, and 1 dB/km can be assumed to be obtainable for practical fiber systems.

2.2 Bandwidth, mode-partition noise, and burst error by mode jumps

Bandwidth, mode-partition noise, and burst error caused by mode jumps are all generic functions of laser and fiber characteristics. If laser diodes are stabilized someday either by injection locking,[23] a new laser-cavity configuration,[24] or other methods (such as the use of an external modulator),[25] then these will not limit system performance so severely. However, at present these three factors are important considerations for single-mode systems.

2.2.1 Bandwidth limitations

The bandwidth of a system using a well-stabilized single-longitudinal-mode laser would be limited by the degenerate-polarized mode dispersion of the fiber,[26] not by the laser spectrum. Since most laser diodes show spectral broadening under direct-current modulation, the bandwidth is determined by the source spectrum and the chromatic dispersion of the fiber. Figure 2 shows the bandwidth of single-mode fibers. Two curves are shown for source rms half-widths of 1 nm and 2 nm. The wavelength of minimum dispersion is assumed to be 1.3 μm. The overall transmission bandwidth is determined by chromatic dispersion of the fiber and by the laser spectrum, as follows:[10,14]

$$f_{6\text{dB}} = \frac{187.3}{Z \cdot \sigma \cdot \left| \frac{d\tau}{d\lambda} \right|} \text{[GHz]} \quad \text{off the wavelength of minimum dispersion}$$

$$f_{6\text{dB}} = \frac{616.4}{Z \cdot \sigma^2 \cdot \left| \frac{d^2\tau}{d\lambda^2} \right|} \text{[GHz]} \quad \text{at the wavelength of minimum dispersion,} \qquad (1)$$

where $f_{6\text{dB}}$[GHz] is the bandwidth of fiber (3-dB optical bandwidth), Z is its length, $\left| \frac{d\tau}{d\lambda} \right|$ [ps/nm·km] is the differential chromatic dispersion $\left| \frac{d^2}{d\lambda^2} \right|$ [ps/nm^2·km] is the third-order chromatic dispersion, and σ is the half-rms width of laser spectrum.

If we permit a 3-dB power penalty at an error rate of 10^{-9}, then the signal bit rate, B(b/s), and fiber bandwidth, $f_{6\text{dB}}$, are related by[19]

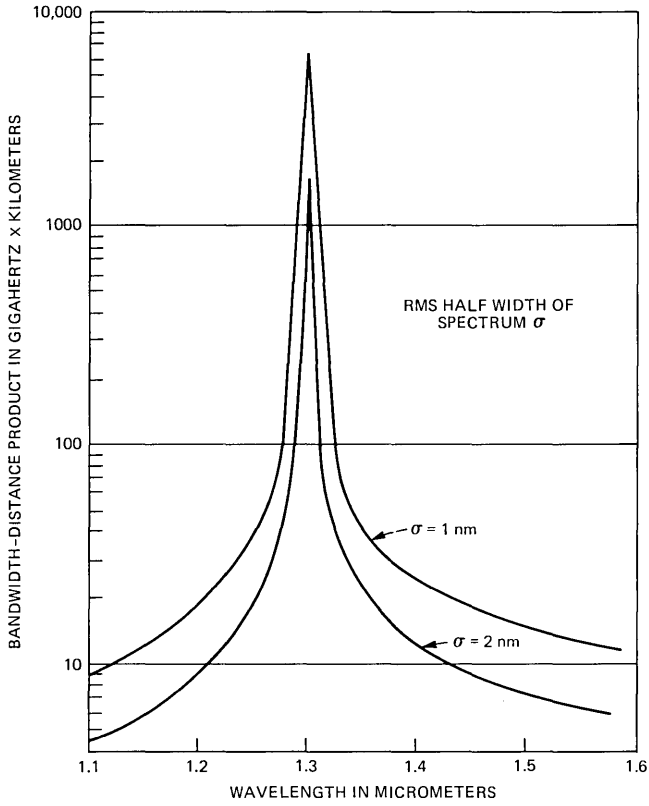


Fig. 2—Bandwidth of a single-mode fiber vs. wavelength.

$$f_{6\text{dB}} = 0.55B. \tag{2}$$

Combining equations (1) and (2), we obtain the performance limitation caused by bandwidth as a bit rate-distance product, $B \cdot Z$, as shown below.

$$B \cdot Z \leqq \frac{340.5}{\sigma \cdot \left| \frac{d\tau}{d\lambda} \right|} \text{[Gb/s·km]} \quad \text{off the wavelength of minimum dispersion}$$

$$B \cdot Z \leqq \frac{11207}{\sigma^2 \cdot \left| \frac{d^2\tau}{d\lambda^2} \right|} \text{[Gb/s·km]} \quad \text{at the wavelength of minimum dispersion.} \tag{3}$$

2.2.2 *Mode-partition noise*

As discussed in Ref. 10, mode-partition noise can become a serious performance limitation. The bit-rate-distance product caused by this

partition noise is expressed for the condition that the asymptotic error rate be 10^{-17} (or a power penalty or 1.5 dB at 10^{-9} error rate)

$$B \cdot Z \leqq \frac{130}{\sigma \cdot \left| \frac{d\tau}{d\lambda} \right| \sqrt{k}} \text{[Gb/s·km]} \quad \text{off the wavelength of minimum dispersion}$$

$$B \cdot Z \leqq \frac{1173}{\sigma^2 \cdot \left| \frac{d^2\tau}{d\lambda^2} \right| \sqrt{k}} \text{[Gb/s·km]} \quad \text{at the wavelength of minimum dispersion,} \qquad (4)$$

where k is the mode-partition noise-suppression factor ($0 \leqq k \leqq 1$). Measured results indicate that k lies between 0.4 and 0.7 and a theoretical analysis of k shows that it cannot be zero and is about 0.1 even with dc operation. A comparison of eqs. (3) and (4) shows that the effect of partition noise is greater than that of bandwidth when k is greater than 0.1. Therefore, mode-partition noise is the dominant effect on the system. Figure 3 shows the bit-rate-distance product caused by the mode-partition noise in the worst case, which is $k = 1$. Two curves are shown for source rms half-widths of 1 nm and 2 nm. The wavelength of minimum dispersion is assumed to be 1.3 μm.

2.2.3 Burst error caused by mode jumps

Because of temperature variations or drive-current variations, some laser diodes show sudden jumps of the center wavelength. This causes a burst of errors until the retiming circuit locks into the shifted phase. This burst-type error will happen randomly at any repeater. To keep the system error rate within system requirements, there are two design approaches:

(*i*) Error-rate objective—worst short-term error rate must be under 10^{-9} when this burst error is a frequent random phenomenon.

(*ii*) Error-free-seconds objective—the error-free seconds must be at least 95 percent when burst errors occur occasionally.

If we assume that mode jumps at each repeater are rather frequent random phenomena, unlike fading or lightning, then a 6400-km terminal-to-terminal error-rate objective should be applied to this burst-type error caused by mode jumps. From Ref. 8, the bit-rate-distance product in the worst case is then given by the relation

$$B \cdot Z \leqq \frac{225}{\left| \frac{d\tau}{d\lambda} \right| \cdot \Delta\lambda} \text{[Gb/s·km]},$$

where $\Delta\lambda$[nm] is the center-wavelength shift for a typical jump.

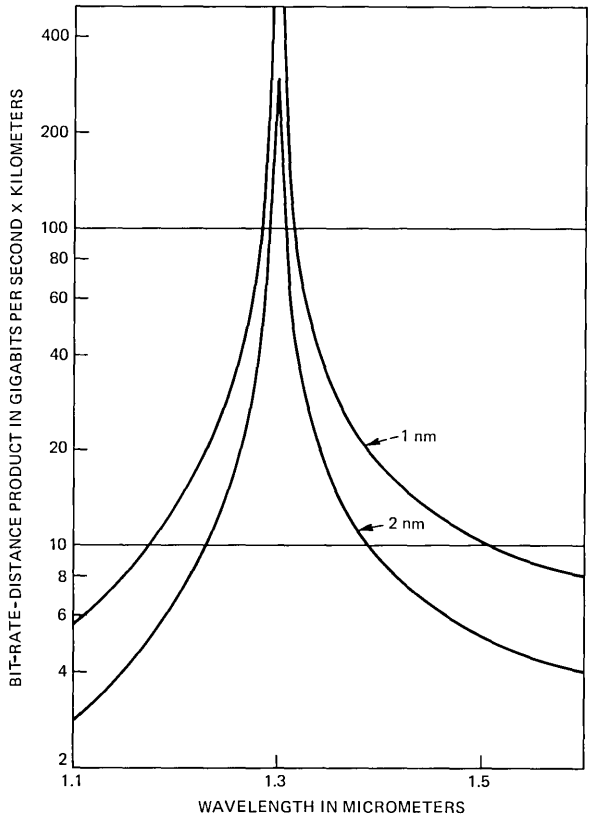


Fig. 3—Bit-rate-distance product caused by mode-partition noise vs. wavelength.

2.3 Receiver sensitivity

There are two basic design approaches used to obtain highly sensitive receivers: (*i*) InGaAs-pin photodiode[27,28] with an ultra-low noise amplifier, and (*ii*) Ge[29] or InGaAs avalanche photodiode[30] with a low-noise amplifier. For the first approach, there are several device technologies that provide ultra-low noise amplifiers above 274 Mb/s: (*i*) Si-microwave bipolar transistors,[31] (*ii*) GaAs-MESFETs,[32] and (*iii*) Si short-channel MOSFETs.[18,33] Although Si bipolar transistors will deliver better noise characteristics than FETs above 300 Mb/s, theoretically FETs are still useful for high bit-rate amplifiers because the cut-off frequency (~4 GHz) of practical bipolar transistors is lower than that of FETs (10 ~ 25 GHz). An avalanche photodiode APD receiver delivers much higher sensitivities, theoretically, even with its intrinsic excess noise factor. However, the sensitivity achieved using a practical

GeAPD is comparable to InGaAs pin with a ultra-low noise amplifier. In general, the practical value of sensitivity can be expressed as

$$\bar{P} = -34 + 10 \log B[\text{dBm}],$$

where B is [Gb/s].

III. REPEATER SPAN

The achievable repeater spacing at different bit rates above 274 Mb/s can now be found assuming 0 dBm of average output power from laser diodes into the fiber. Figures 4 and 5 show two different cases: Fig. 4 shows that the half-rms width of laser spectrum is 2 nm and mode jumping $\Delta\lambda$ is less than 3 nm. Solid lines indicate performance limitations due to fiber attenuation (0.5 dB/km and 1.0 dB/km). The dot-dash lines are the fiber bandwidth of 1.3 μm and 1.275 μm. However, the mode-partition noise limitations indicated by the dashed

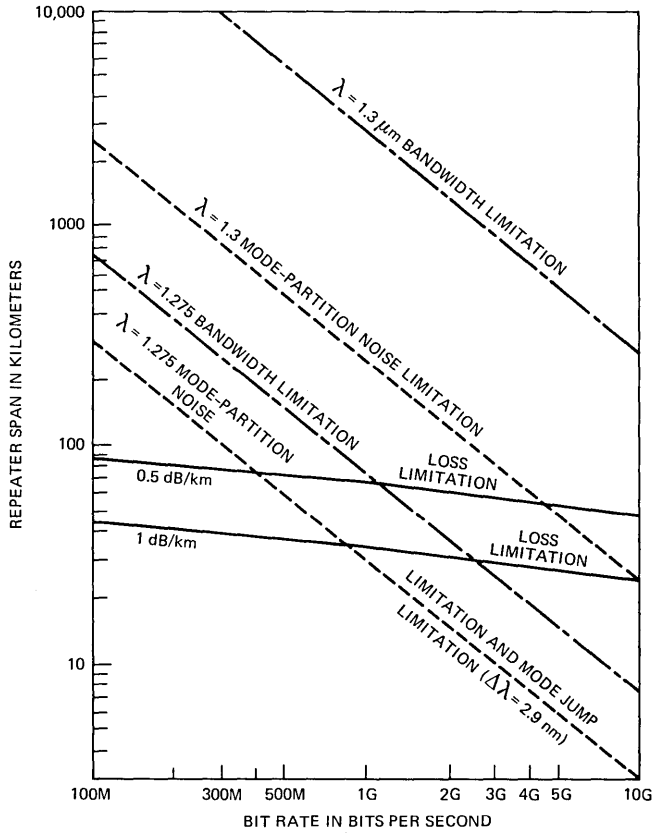


Fig. 4—Theoretical limits of repeater span due to the mode-partition noise limitation.

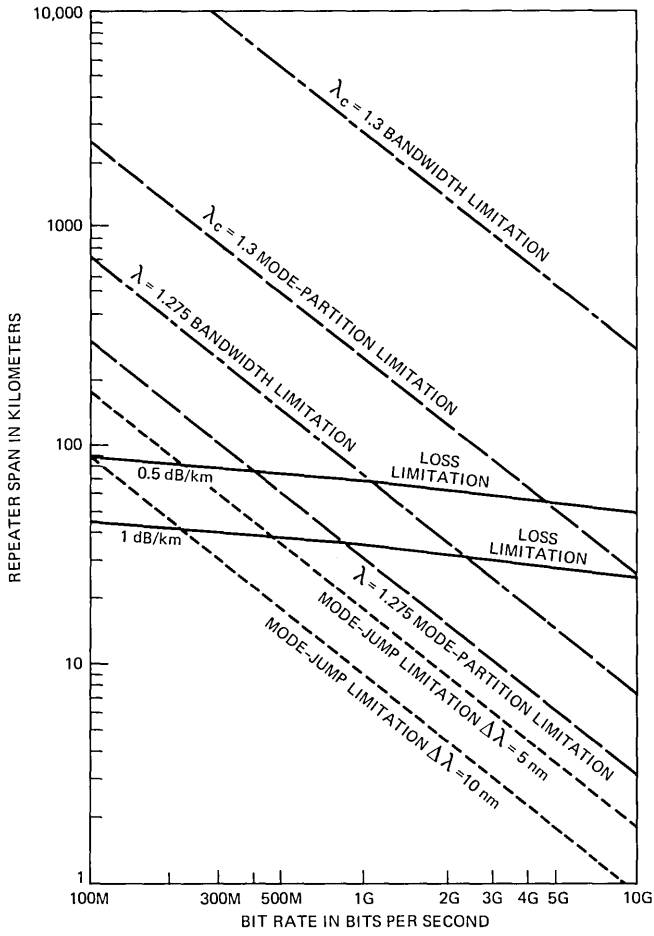


Fig. 5—Theoretical limits of repeater span due to the burst-error limitation.

lines are much more severe than the fiber bandwidth limitations. Also, when the center wavelength shift caused by a mode jump is 2.9 nm, the burst-error limitation is the same as the mode-partition noise limitation for 1.275 μm. Figure 5 shows that the half-rms width of laser spectrum is 2 nm and mode jumping $\Delta\lambda$ is 5 to 10 nm. When the center wavelength shift due to a mode jump becomes large (5 nm to 10 nm), the system performance is limited by the burst errors caused by mode jumps. Two dotted lines indicate the performance limitation caused by mode jumps. For both figures, two solid lines indicate the fiber attenuation limitations; one is 0.5 dB/km and the other is 1 dB/km. The dashed lines indicate the performance limitations caused by the mode-partition noise when the mode-partition factor, k, is 1; the dotted

line is the performance limitation of the burst-errors caused by mode jumps. The chain line indicates the bandwidth limitations caused by laser spectral width. As we see in Figs. 4 and 5, the fiber bandwidth limitation is no longer dominant in comparison with mode-partition noise or mode-jump penalties. The difference between Figs. 4 and 5 is due to the effect of mode jumps. This shows that when we use laser diodes with narrow spectra, the mode jumps become the dominant effect.

IV. MAXIMUM TRANSMISSION CAPACITY

Laser diodes have about 1- to 2-GHz modulation bandwidth. The following criteria are used to maximize transmission capacity by either increasing signal speed or using wavelength multiplexing.

(*i*) Total data rates are to be $n \times 274$ Mb/s ($n = 1, 2, 4$).

(*ii*) For wavelength multiplexing, the separation of wavelengths is 25 nm. Since InGaAsP/InP laser diodes have a wavelength temperature dependency of 0.5 nm/°C,[2] we assume stabilization of the temperature is within ±10°C. We also assume that the initial variation of laser wavelength is within ±2.5 nm, and that the separation band required to suppress the interference from a neighboring channel is 10 nm. Therefore, a total value of 25 nm becomes the separation of neighboring wavelengths.

(*iii*) The transmission distance and bit rate for each wavelength-multiplexed channel is the same.

(*iv*) The laser spectrum half-width is 2 nm and mode jumps are less than 3 nm.

(*v*) We ignore the insertion losses of the wavelength multiplexing and demultiplexing devices.

(*vi*) We also ignore the mode-partition noise caused by the optical filter of wavelength multiplexer, which must be carefully designed to minimize that noise.

Figure 6 shows the total capacity and repeater spacing. There are two cases: one for 0.5 dB/km, and the other for 1-dB/km fiber loss. The solid line is for a basic transmission rate of 274 Mb/s. The small numbers on each dot are the orders of multiplexing. The dashed lines are the cases of 548 Mb/s. The chain lines are the cases of 1096 Mb/s. Each number ($1 \cdots 5$) indicates the number of wavelength multiplexed channels. There are two dominant performance limitations: (*i*) loss limitation, and (*ii*) mode-partition noise limitation. In loss limitation, the wavelength multiplexing technique is advantageous to increase its capacity. In the mode-partition noise limitation, increasing the bit-rate of single channel is a more powerful way to increase capacity and repeater spacing.

Once fiber attenuation and repeater spacing are given, what bit rate

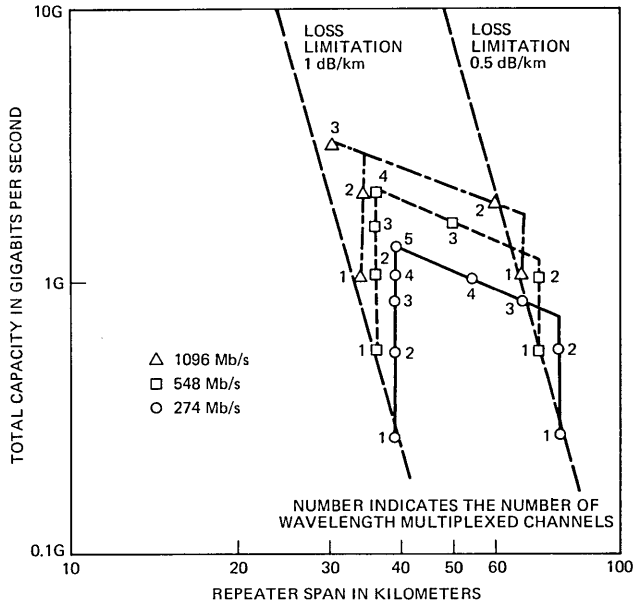


Fig. 6—Maximum capacity and repeater span.

is the best choice? Let us choose a fiber attenuation of 0.5 dB/km and the repeater spacing of about 60 km. A 274 Mb/s-multiplexed system can achieve its maximum capacity of 3 × 274 Mb/s with three-wavelength multiplexing. The 548 Mb/s system and the 1096 Mb/s system can achieve maximum capacities of 1096 Mb/s and 2192 Mb/s, respectively; further wavelength multiplexing reduces the repeater spacing to a much lower span than 60 km. Thus, the maximum capacity can be achieved by 1096 Mb/s systems using two-wavelength multiplexing. The choice of bit rate, 274 Mb/s or 1096 Mb/s, will be determined by availabilities of components and transmission capacity demands.

V. CONCLUSION

We have discussed performance limitations for single-mode fiber systems and introduced formulas for several performance limitations. Also, in Fig. 6, we have shown that wavelength multiplexing increases transmission capacity. We conclude that

(*i*) For more than 1 Gb/s capacity, system performance is limited by fiber dispersion, especially mode-partition noise. In this case a single high-bit-rate channel is the better choice to achieve the maximum capacity and longest repeater spacing.

(*ii*) For less than 1 Gb/s capacity, wavelength multiplexing be-

comes the most effective technique to increase total capacity, to achieve longer repeater spacing, and to lower the bit rate for each channel. The lower channel bit rate eases component requirements.

(*iii*) Mode-partition noise and burst errors caused by mode jumps severely limit performance.

Results of the bit-rate-distance product we discussed here are far less than the theoretical 400 to 3000 Gb/s·km that can be expected from single-mode fibers. To achieve such a high-capacity single-mode fiber system, further research effort is required, especially in the areas of: (*i*) fiber and cabling improvements, (*ii*) fabrication, stabilization, and mode behavior of laser diodes, (*iii*) high-sensitivity receiver design using low-noise components, (*iv*) a high-gain detector such as InGaAs-APD or InGaAs-photo-FETs, and (*v*) high-speed integrable components such as GaAs-MESFETs, and Si-MOSFETs.

REFERENCES

1. J. Yamada and T. Kimura, "Single-Mode Optical Fiber Transmission Experiments at 1.3 μm Wavelength," Rev. of the Elect. Commun. Lab., *27,* Nos. 7 to 8 (July–August 1979), pp. 611–29.
2. T. Kimura, "Single-Mode Systems and Components for Longer Wavelengths," IEEE Trans. on Circuit and System, *CAS-26,* No. 12 (December 1979), pp. 987–1010.
3. S. Tomaru, M. Kasu, M. Kawachi, and T. Edahiro, "VAD Single Mode Fiber with 0.2 dB/km Loss," Elec. Lett., *17,* No. 2 (January 1981), pp. 92–3.
4. D. E. Payne and W. A. Gambling, "Zero Dispersion in Optical Fibers," Elec. Lett., *11,* No. 8 (April 17, 1975), pp. 176–8.
5. S. Kobayashi, S. Shibata, N. Shibata, and T. Izawa, "Refractive-Index Dispersion of Doped Fused Silica," 1977 IOOC, Technical Digest, B8-3 (July 1977), pp. 309–12.
6. L. G. Cohen, C. Lin, and W. G. French, "Tailoring Zero Chromatic Dispersion into the 1.5 ~ 1.6 μm Low-Loss Spectral Region of Single Mode Fibers," Elec. Lett., *15,* No. 12 (June 1979), pp. 334–5.
7. K. Okamoto, T. Edahiro, A. Kawana, and T. Miya, "Dispersion Minimization in Single-Mode Fibers Over A Wide Spectral Range," Elec. Lett., *15,* No. 22 (October 1979), pp. 729–31.
8. L. G. Cohen and W. L. Mammel, "Tailoring the Shapes of Dispersion Spectra to Control Bandwidths in Single-Mode Fibers," Seventh European Conf. on Optical Commun., September 8 to 11, 1981, pp. 3.3-1 to 3.3-4.
9. Y. Okano, K. Nakagawa, and T. Itoh, "Laser Mode Partition Noise Evaluation for Optical Fiber Transmission," IEEE Trans. Commun. *COM-28,* No. 2 (February 1980), pp. 238–43.
10. K. Ogawa, "Analysis of Mode Partitioning Noise for Laser Diode System," IEEE J. Quantum Elec., *17,* No. 5 (May 1982), pp. 849–55.
11. K. Ogawa and R. W. Vodhanel, "Analysis and Measurement of Mode Partition Noise," Topical Meeting on Optical Fiber Commun. (OFC82) *THDD-4,* Phoenix, Arizona, April 13–15, 1982, pp. 58–9.
12. T. Ito, S. Machida, K. Nawata, and T. Ikegami, "Intensity Fluctuation in Each Longitudinal Mode of a Multimode AlGaAs Laser," IEEE J. Quantum Elec., *QE-13,* No. 8 (August 1977), p. 574.
13. K. Nakagawa and T. Ito, "Detailed Evaluation of an Attainable Repeater Spacing for Fiber Transmission at 1.3 μm and 1.55 μm Wavelengths," Elec. Lett., *15,* No. 24 (November 1979), pp. 776–7.
14. D. Gloge, K. Ogawa, and L. G. Cohen, "Baseband Characteristics of Long-Wavelength LED System," Elec. Lett., *16,* No. 10 (May 1980), pp. 366–7.
15. K. Nawata, S. Machida, and T. Ito, "An 800 Mb/s Optical Transmission Experiment Using a Single-Mode Fiber," IEEE J. Quantum Elec., *QE-14,* No. 2 (February 1978), pp. 98–103.
16. D. R. Smith, R. C. Hooper, R. P. Webb, and M. F. Sanders, "PIN Photodiode

Hybride Optical Receivers," Proc. of the Optical Commun. Conf., Amsterdam, The Netherlands, September 17–19, 1979, pp. 13.4/1–4.
17. K. Ogawa and E. L. Chinnock, "GaAs-FET Transimpedance Front-End Design for a Wide Band Optical Receiver," Elect. Lett., *15,* No. 20 (September 1979), pp. 650–3.
18. K. Ogawa, B. Owen, and H. J. Bell, "A Long-Wavelength Optical Receiver Using a Short Channel Si-MOSFET," Conf. on Lasers and Elect. Fiber-Optics (CLEO), Washington, D.C., *WM-3,* June 10–12, 1981, pp. 56–7.
19. D. Gloge, A. Albanese, C. A. Burrus, E. L. Chinnock, J. A. Copland, A. G. Dentai, T. P. Lee, T. Li, and K. Ogawa, "High-Speed Digital Lightwave Communication Using LEDs and PIN Photodiodes at 1.3 μm," B.S.T.J., *59,* No. 8 (October 1980), pp. 1365–82.
20. D. Marcuse, "Loss Analysis of Single-Mode Fiber Splice," B.S.T.J., *56,* No. 5 (May–June 1977), pp. 703–18.
21. J. Sagai, "Microbend Loss of Single-Mode Fiber," Technical Meeting of IECE on Optical Quantum Elec., *OQE 79-49* (July 1979), pp. 45–60.
22. H. Tsuchiya, I. Hatakeyama, and N. Shimizu, "Splicing Loss and Connector," Technical Meeting of IECE on Optical Quantum Elec., *OQE 78-44* (June 1978), pp. 101–8.
23. S. Kobayashi and T. Kimura, "Coherence of Injection Phase-Locked AlGaAs Semiconductor Laser," Elec. Lett., *16,* No. 17 (January 16, 1980), pp. 608–70.
24. Y. Sakakibara, K. Furuya, K. Utaka, and Y. Suematsu, "Single-Mode Oscillation Under High-Speed Direct Modulation in GaAsInP/InP Integrated Twin-Guide Lasers with Distributed Bragg Reflectors," Elec. Lett., *16,* No. 12 (June 1980), pp. 456–7.
25. R. C. Alferness, private communication.
26. N. Imoto, N. Yoshizawa, J. Sakaki, and H. Tsuchiya, "Birefringence in Single-Mode Optical Fiber Due to Elliptical Case Reformation and Stress Anisotropy," IEEE J. Quantum Elec., *QE-16,* No. 11 (November 1980), pp. 1269–71.
27. R. F. Leheny, R. E. Nahory, M. A. Pollack, E. D. Beebe, and J. C. DeWinter, "Characterization of $In_{0.53}Ga_{0.47}As$ Photodiodes Exhibiting Low Dark Current and Low Junction Capacitance," IEEE J. Quantum Elec., *QE-17,* No. 2 (February 1981), pp. 227–31.
28. T. P. Lee, C. A. Burrus, Jr., and A. G. Dentai, "InGaAs/InP p-i-n Photodiodes for Lightwave Communications at the 0.95 ~ 1.65 μm Wavelength," IEEE J. Quantum Elec., *QE-17,* No. 2 (February 1981), pp. 232–8.
29. T. Mikawa, S. Kagawa, T. Kaneda, T. Sakurai, H. Ando, and O. Mikami, "A Low-Noise $n + np$ Germanium Avalanche Photodiode," IEEE J. Quantum Elec., *QE-17,* No. 2 (February 1981), pp. 210–6.
30. N. Susa, H. Nakagome, H. Ando, and H. Kanbe, "Characteristics in InGaAs/InP Avalanche Photodiodes with Separated Absorption and Multiplication Regions," IEEE J. Quantum Elec., *QE-17,* No. 2 (February 1981), pp. 243–9.
31. R. Paski, private communication.
32. K. Ogawa, "Noise Caused by GaAs MESFETs in Optical Receivers," B.S.T.J., *60,* No. 6 (July–August 1981), pp. 923–8.
33. P. I. Suciu, E. N. Fuls, and H. J. Boll, "High-Speed NMOS Circuits Made with X-Ray Lithography and Reactive Sputter Etching," IEEE Elec. Device Lett., *EDL-1,* No. 1 (January 1980), pp. 10–1.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Error Probability of Partial-Response Continuous-Phase Modulation with Coherent MSK-Type Receiver, Diversity, and Slow Rayleigh Fading in Gaussian Noise

By C.-E. SUNDBERG*

(Manuscript received January 21, 1982)

This paper considers a class of constant-amplitude modulation schemes with good spectral main lobe and tail behavior. Detection is assumed to be coherent and the receiver is of offset-quadrature type, i.e., minimum shift keying (MSK) type, consisting of a linear filter in each quadrature arm followed by simple processing. Analytical error-probability formulas are derived for various modulation schemes and receiver filters for ideal diversity with maximal-ratio and selection combining. Independent slow Rayleigh fading in Gaussian noise is assumed and several numerical examples are given. Asymptotic behavior of the error probability for large signal-to-noise ratios is derived, and the relationship between the degree of smoothing in the partial-response continuous-phase modulation and the asymptotic error probability is shown for fading channels with and without diversity.

I. INTRODUCTION

The transmission of information over radio channels with multiple changing-propagation paths is subject to fading, i.e., random time variations of the receiver signal strength. For digital transmission over a fading channel, the time variations cause a varying error probability for all types of digital-modulation methods.

The application of bandwidth-efficient constant-amplitude modulation schemes to digital land-mobile radio has been considered in several papers.[1-5] Other investigations have shown that transmission

* This work was done while Mr. Sundberg was a consultant at Bell Laboratories.

over such channels is subject to fading.[6,7] In this paper we will consider slow Rayleigh fading where the density function for the signal-to-noise ratio (s/n) is

$$f(\gamma) = \frac{1}{\Gamma} e^{-\gamma/\Gamma}, \tag{1}$$

where Γ is the average signal-to-noise ratio. We define slow fading as the time-varying s/n γ, which is approximately constant over several transmitted bits (symbols). It will be seen below that the type of detectors considered for the modulation schemes in this paper operate over at most that number of symbols.

By combining a number of channels with independent Rayleigh fading, the density function for the resulting signal-to-noise ratio (1) can be improved. This is called diversity. Thus, with decreased probability of very low signal-to-noise ratios, the average error probability is improved by means of diversity.

This paper presents analytical, easy to use, bit-error probability formulas for smoothed continuous-phase constant-amplitude modulation with a simple coherent receiver and diversity. From these calculations we conclude that the increase in error probability owing to the degree of smoothing is smaller for a fading channel than for a nonfading channel. In Section IV we show an example of the difference between quadriphase shift keying (QPSK) and the schemes $3RC$ and $4RC$, which depict smoothed constant-amplitude modulations with narrow spectral main lobe and low spectral tails. The nonfading channel is approached with large numbers of diversity branches. The fading channel is approached with few branches of diversity. The gradual change from the two extreme cases is readily apparent.

1.1 Probability of error

Coherent transmission is considered. For quadriphase shift keying, 4-PSK (QPSK), there is a special case of the more general error-probability formula considered below. For QPSK and Binary Phase Shift Keying, 2-PSK (BPSK) the error probability for the Gaussian channel with ideal transmission and optimal detection is

$$P(\gamma) = Q(\sqrt{2\gamma}), \tag{2}$$

where $\gamma = E_b/N_o$ is the s/n (per information bit), E_b is the energy per information bit, and N_o is the (one-sided) spectral density for the additive white Gaussian noise. The function $Q(\cdot)$ is the error function associated with the normal distribution, i.e.,

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt. \tag{3}$$

For the family of modulation schemes considered in this paper, the error probability $P(\gamma)$ for the Gaussian channel is a linear sum of Q-function terms. We obtain the average bit-error probability for the fading case by averaging $P(\gamma)$ over the density function in eq. (1), i.e.,

$$P = \int_o^\infty f(\gamma)P(\gamma)d\gamma. \tag{4}$$

The averaging is performed with different density functions for the diversity cases below with different combiner strategies. The problem dealt with in this paper is finding the average bit-error probability (4) for the family of modulation schemes and diversity cases described briefly below.

The solution to the error-probability averaging in (4) is of interest in cellular systems with frequency reuse where cochannel interference is the main interference source and where diversity is used to improve the signal-to-noise density function.[3,6,7] When we consider so-called time-division retransmission schemes, it is realistic to consider space diversity with more than two diversity branches.[8-10]

1.2 Modulation schemes

Before we address the fading and diversity problems, we will give a short description of the modulation schemes considered in this paper. A family of binary constant-amplitude digital modulation schemes is defined by the transmitted signal

$$S(t, \underline{\alpha}) = \sqrt{\frac{2E}{T}} \cos[2\pi f_o t + \phi(t, \underline{\alpha})], \tag{5}$$

where $E = E_b$ is the symbol (bit-) energy, T is the symbol time, and f_o is the carrier frequency. The information-carrying phase is

$$\phi(t, \underline{\alpha}) = 2\pi h \sum_{i=-\infty}^{\infty} a_i q(t - iT), \tag{6}$$

where $\underline{\alpha} = \cdots \alpha_{n-2}, \alpha_{n-1}, \alpha_n, \alpha_{n+1}, \alpha_{n+2} \ldots$ is a sequence of independent binary symbols $\alpha_i \in \{-1, +1\}$ (we will consider only binary schemes here) and h is the modulation index. The phase response is defined by

$$q(t) = \int_{-\infty}^{t} g(\tau)d\tau, \tag{7}$$

where $g(t)$ is a time-limited pulse defining instantaneous frequency.[11-13] The above family of schemes is considered in this paper only for modulation index $h = 1/2$. For this case, such modulation schemes as minimum shift keying (MSK)—or fast frequency shift keying (FFSK),[1]

tamed-frequency modulation (TFM),[2] and Gaussian MSK (GMSK)[3,4]—are contained in the family (5). Different modulation schemes are obtained by changing the pulse shape $g(t)$. Thus, for MSK

$$g(t) = \begin{cases} 0 & t < 0, \quad t > T \\ \dfrac{1}{2T} & 0 \le t \le T, \end{cases} \tag{8}$$

and for TFM[2,12]

$$g(t) = \tfrac{1}{8}[g_o(t - T) + 2g_o(t) + g_o(t + T)], \tag{9}$$

where

$$g_o(t) = \frac{1}{T}\left[\frac{\sin\left(\frac{\pi t}{T}\right)}{\left(\frac{\pi t}{T}\right)} - \frac{\pi^2}{24} \cdot \frac{2\sin\left(\frac{\pi t}{T}\right) - \frac{\pi t}{T}\cos\left(\frac{\pi t}{T}\right) - \left(\frac{\pi t}{T}\right)^2 \sin\left(\frac{\pi t}{T}\right)}{\left(\frac{\pi t}{T}\right)^3}\right]. \tag{10}$$

The TFM pulse is infinite in time. Below we consider time-truncated versions of the pulse (9).

Raised-cosine pulses of various lengths LT[11-12] are also considered. For this case

$$g(t) = \begin{cases} \dfrac{1}{2LT}\left[1 - \cos\left(\dfrac{2\pi t}{T}\right)\right] & 0 \le t \le LT \\ 0 \quad \text{otherwise.} \end{cases} \tag{11}$$

The notation LRC denotes a raised-cosine pulse of length LT or a modulation scheme (5), (6) based on the pulse LRC, shown in eq. (11). Additional details on schemes based on LRC can be found in Refs. 11, 12, and 13.

Duobinary MSK, i.e.,

$$g(t) = \begin{cases} \dfrac{1}{4T} & 0 \le t \le 2T \\ 0 \quad \text{otherwise} \end{cases} \tag{12}$$

is also considered. This pulse is also denoted as $2REC$.[12]

Schemes based on pulses $g(t)$ of length larger than the symbol time T are so-called partial-response schemes. Controlled intersymbol in-

terference is introduced. This improves the power spectra[12,14] and the detection efficiency under certain conditions.[12,15]

Figures 1 and 2 show the motivation for considering partial-response continuous-phase modulation schemes with overlapping pulses $g(t)$ of increasing length L. Figure 1 shows how the power spectral density (normalized to the total power) becomes more narrow for larger L values. The sidelobes are also low due to the large number of continuous derivatives in $g(t)$.[11,12,14] Figure 2 compares QPSK, MSK, and $3RC$. Note that the raised-cosine schemes have constant amplitude. By changing $g(t)$, further improvements of the spectral tails can be achieved.[2,12]

1.3 Detectors

We show in our references that the $h = 1/2$ schemes can be detected with good efficiency in an MSK detector[1] or an MSK-type detector (modified offset quadrature receiver),[13] which consists of a "matched" filter in each quadrature arm followed by very simple digital processing. In its most simple form, this processing consists of differential decoding in each quadrature arm[1,2] with differential encoding employed at the transmitter. For MSK, the receiver in Fig. 3a is optimum. MSK consists of linear modulation in each quadrature arm. A matched receiver filter whose impulse response is a half-cycle sinusoid of length $2T$ is the filter function $a_1(t)$ and $a_2(t)$ in this case.

An MSK receiver with modified filters was proposed in Ref. 2 for partial response schemes in the case of TFM. For the partial response case with overlapping pulses $g(t)$, the receiver shown in Fig. 3a is suboptimum. However, for $h = 1/2$ and for reasonably small L ($L \leq 4$), this receiver gives good results for low and intermediate signal-to-noise ratios.[2,13] The error probability shown in the graphs below is for binary decisions based on the filter output compared to a threshold which is 0.

Figures 3b and 3b show two examples of the receiver filter $a_1(t)$.[13] The receiver filter $a_2(t)$ is identical to $a_1(t)$. The first example is the so-called SPAM (selected pulse amplitude modulation) filter,[13] where we have used the same strategy in filter selection as was shown in Ref. 2. Note that the filter in Fig. 3b is truncated to $N_T = 4$ symbols. Figure 3c shows the so-called MIN (minimum energy signal) filter. It is shown in some detail in Ref. 13 that the preferable filter depends on the s/n . Filter SPAM is good for low s/n's, while filter MIN is good for high s/n's.[13] For $3RC$ the difference in performance is small, though.

Figure 4 shows the "detection eye," i.e., the input $\cos\{\phi(t, \underline{\alpha})\}$ to the filter $a_1(t)$ for the $3RC$ scheme. (For further details see Ref. 13). The eye in the other quadrature arm is the same, but offset one symbol interval T.

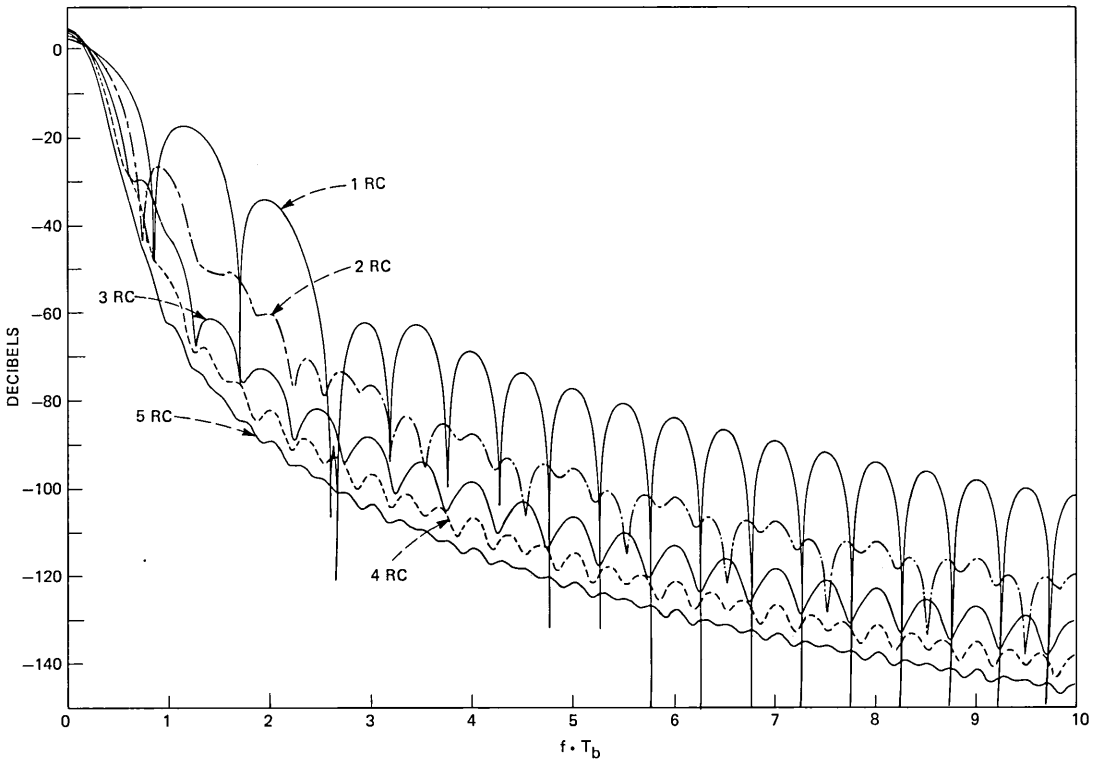


Fig. 1—Power spectra for binary RC schemes when $h = 1/2$ and $T_b = T$.

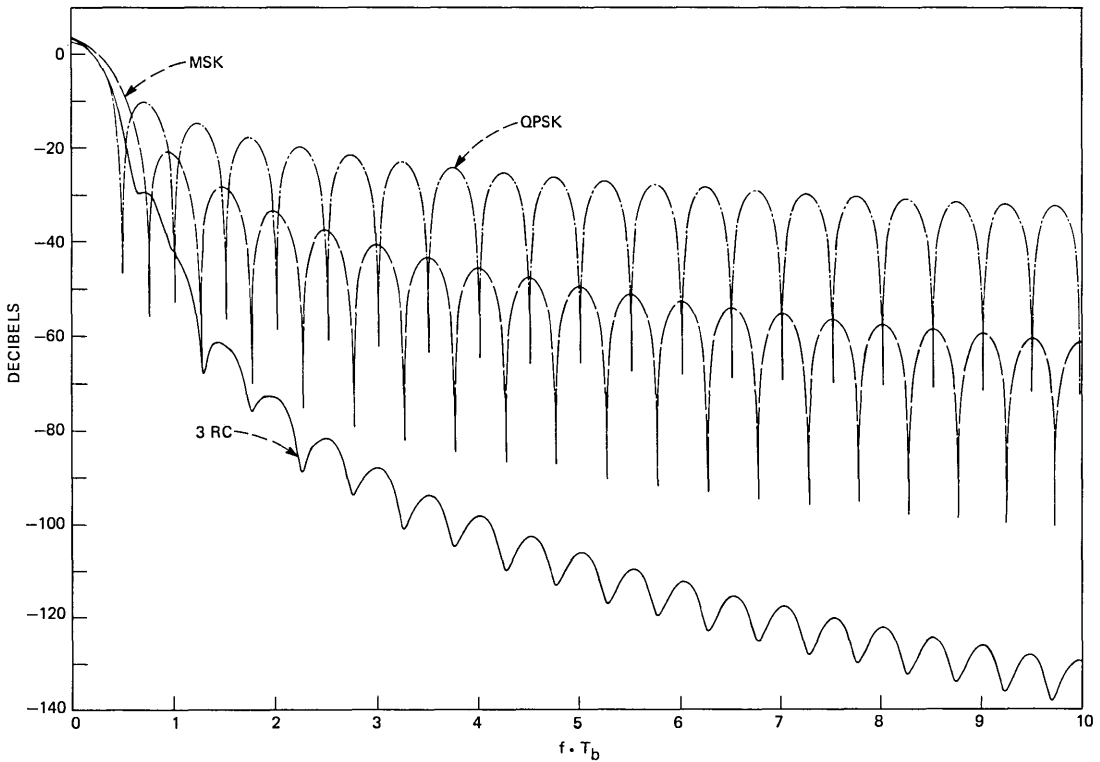
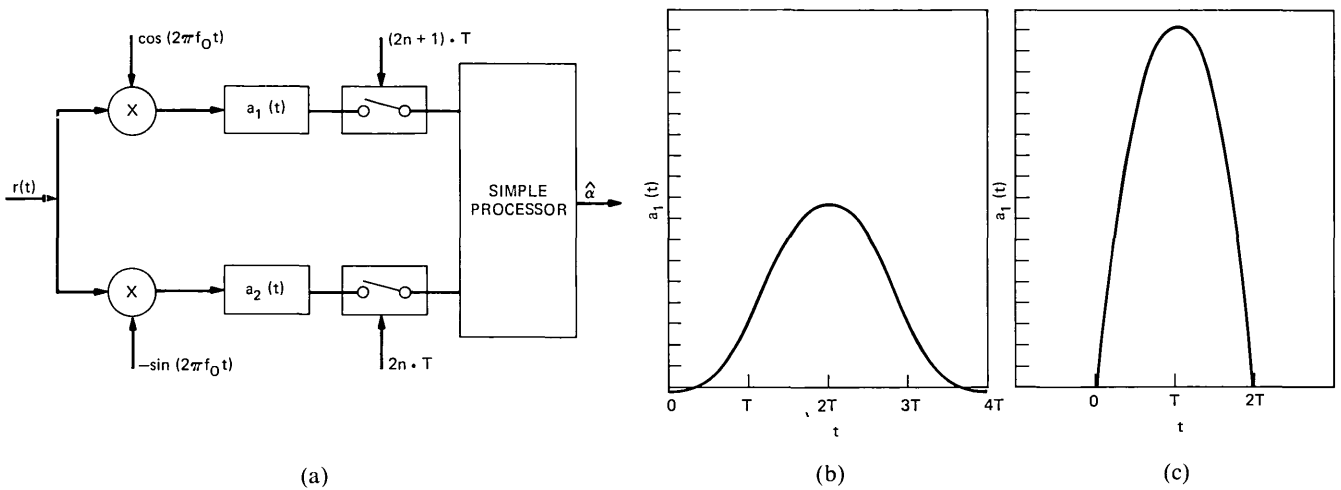


Fig. 2—Power spectra for binary $3RC$ with $h = 1/2$, QPSK, and MSK.



(a) (b) (c)

Fig. 3a—Receiver structure for the modified offset quadrature receivers for partial-response CPM.

Fig. 3b—The SPAM filter for a binary, $h = 1/2$, $3RC$ receiver.

Fig. 3c—The MIN filter for a binary, $h = 1/2$, $3RC$ receiver.

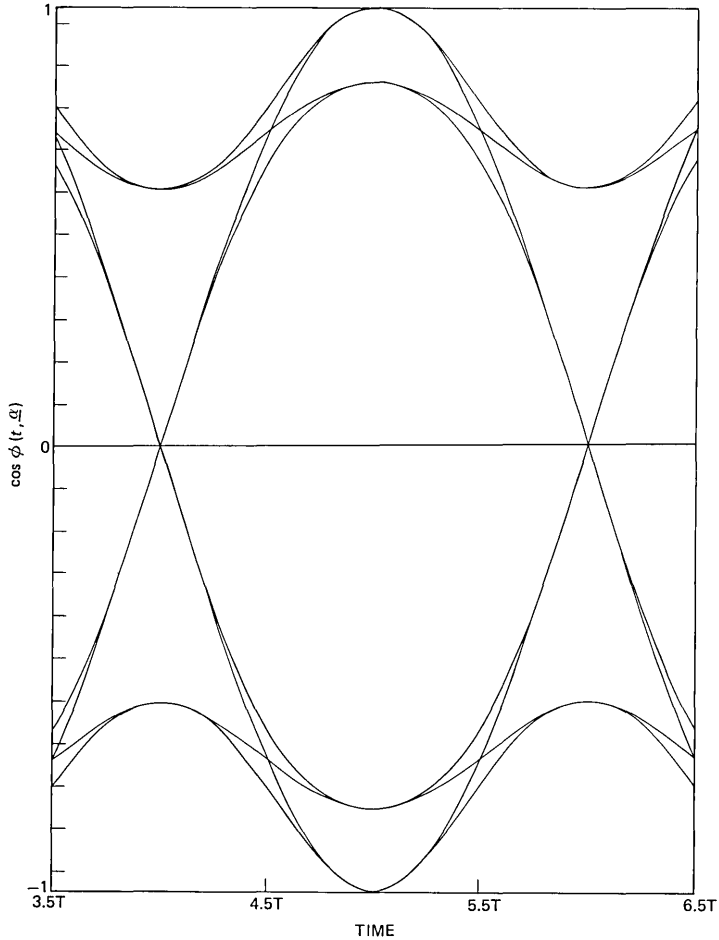


Fig. 4—Cos $\phi(t, \underline{\alpha})$ eye pattern of a binary $3RC$, $h = 1/2$, scheme. Note how open the eye is at the detection point ($t = 5T$ in this time scale).

1.4 Error probability in Gaussian noise

The average error probability P for a given modulation scheme—given pulse shape $g(t)$—and for a given receiver filter $a(t) = a_1(t) = a_2(t)$ can be evaluated exactly. For any given data sequence, the decision after the filter is a comparison between a Gaussian variable and a threshold which is zero in Fig. 3a. (For details see Refs. 13, 15, 16, 17, and 18.) It is shown in Ref. 13 that various filters are best at various levels of P (various values of s/n). Note that by error probability P we consistently refer to the average error probability associated with one binary decision after the receiver filter $a(t)$. Because of the phase ambiguity of π, the bit-error probability of the modulation

scheme, P_{bit}, is affected by the resolution of this ambiguity.[1,2] With differential decoding we have[19]

$$P_{bit} = 2P(1 - P). \tag{13}$$

This will be discussed somewhat in Section V. Until then the calculations will be done using P (before the differential decoding).

In Ref. 13 we see that the average error probability P for the partial-response scheme with an MSK-type receiver is

$$P = \sum_{i=1}^{m} C_i Q\left(\sqrt{d_i^2 \frac{E_b}{N_o}}\right). \tag{14}$$

Here $E_b = E$, N_o is the spectral density of the one-sided Gaussian noise, $Q(\cdot)$ is the error function associated with the normal distribution, d_i^2 is the squared Euclidean distance associated with a signal corresponding to data sequence number i received in the fixed filter $a(t)$ (see Refs. 13 and 15 for details), and C_i is the probability of that specific signal. There are at most $m = 2^{N_T+L+1}$ different signals in (14), where N_T is the filter length in bit intervals (some of the distance values are the same for several data sequences due to symmetry. (See Ref. 13 for details.)

Assuming independent data symbols with $p(-1) = p(+1) = 1/2$, then

$$C_i = 1/m. \tag{15}$$

For QPSK and MSK we have $d^2 = 2$; thus,

$$P = Q\left(\sqrt{\frac{2E_b}{N_o}}\right) \tag{16}$$

with the optimum receiver filter.

For the general partial-response case, the sum (14) consists of several terms with several squared distance values d_i^2, where one of them is the minimum squared Euclidean distance. Methods for the calculation of d_i^2, $i = 1 \cdots m$, are given in Ref. 13. The parameters d_i^2 are independent of signal-to-noise ratio. They depend only on the data sequence, the pulse shape $g(t)$, and the receiver filter $a(t)$.

Figures 5 and 6 show the calculated average error probability P, using eq. (14), for a number of modulation schemes where the receiver filter is the SPAM filter.[13] The technique described in Ref. 2 was used for selecting the filter. Several different pulse shapes are considered (see Ref. 13 for details). For comparison, the error probability for QPSK is also shown. It is evident that the longer and smoother pulse $g(t)$, the better the power spectrum and the larger the penalty in E_b/N_o compared to QPSK. However, the penalty—using, for example, $3RC$—is only 0.5 dB at $P = 10^{-3}$. The penalty for TFM at $P = 10^{-3}$ is about 1.0 dB. For lower P values the penalty is larger.

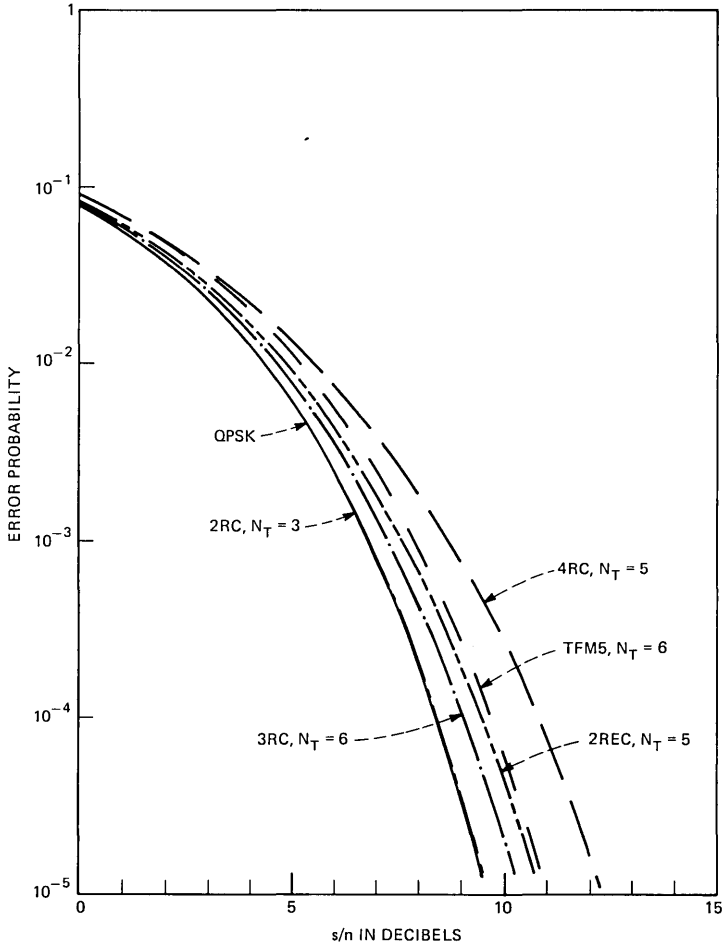


Fig. 5—Error probability in detecting the phase node sequence (output of the receiver filter) for SPAM receivers for various binary, $h = 1/2$, schemes.

The rest of the paper is organized as follows: In Section II we derive formulas for error probability for the partial-response continuous-phase modulation schemes with M branch diversity and maximal-ratio combining. Section III contains the corresponding results for selection combining, and Section IV presents some numerical results. Section V contains a discussion and conclusions.

II. MAXIMAL-RATIO-COMBINING

First we consider the problem of calculating the average error probability

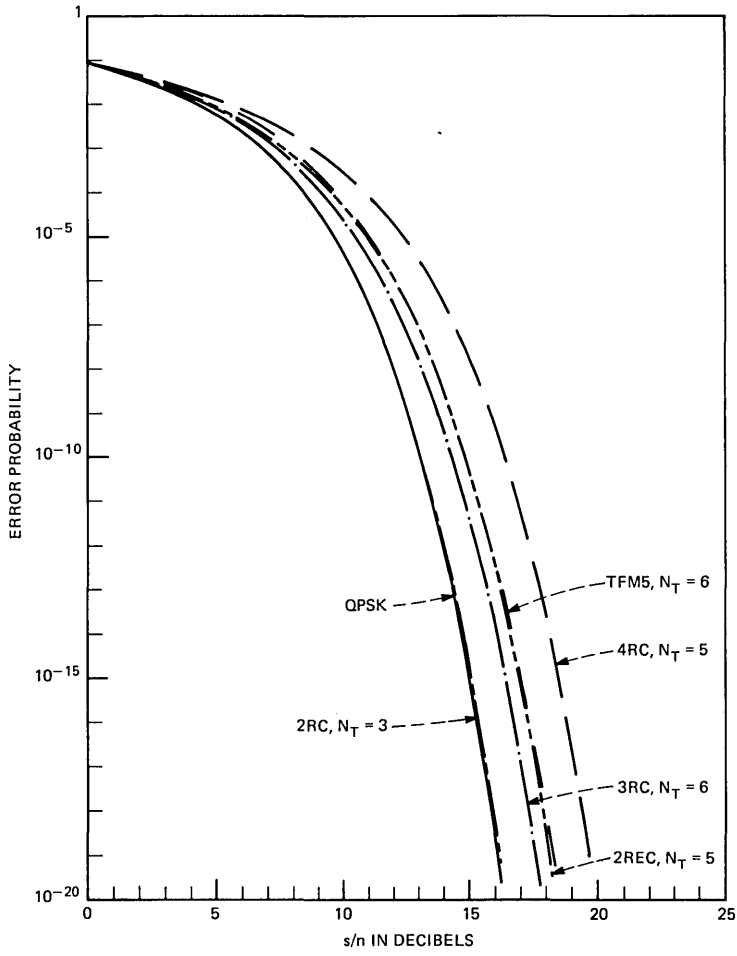


Fig. 6—Error probability in detecting the phase node sequence (output of the receiver filter) for SPAM receivers for various binary, $h = 1/2$, schemes.

$$P = \int_o^\infty f(\gamma)P(\gamma)d\gamma, \tag{17}$$

where $P(\gamma)$ is the average error probability of the modulation scheme (with coherent detection) for Gaussian noise at the s/n γ given by (14),

$$P(\gamma) = \sum_{i=1}^{m} C_i Q(\sqrt{d_i^2 \gamma}), \tag{18}$$

and where $f(\gamma)$ is the density function for γ. Note that the receiver

filter $a(t)$ operates over N_T symbols, typically smaller than 6. Slow fading is assumed and the averaging is performed over one "block" of length N_T, much the same way as is shown in Ref. 20.

The instantaneous s/n in diversity branch k is assumed to be γ_k, with equal average value for all branches. For Rayleigh fading, the γ_k have the probability density function of eq. (1).

Assume ideal maximal-ratio combining.[6,7] This combiner must know each path magnitude and phase to perform perfect combining and must have the property that the output signal-to-noise ratio is the sum of the instantaneous branch s/n's, i.e.,

$$\gamma = \sum_{k=1}^{M} \gamma_k. \tag{19}$$

The random variable γ has the density function[6,7]

$$f(\gamma) = \frac{1}{\Gamma}\left(\frac{\gamma}{\Gamma}\right)^{M-1} \frac{1}{(M-1)!} e^{-\gamma/\Gamma} \tag{20}$$

for M-branch maximal-ratio combining, assuming independent Rayleigh fading in each branch. The average receiver-output s/n is

$$E\{\gamma\} = M\Gamma. \tag{21}$$

References 21 and 22 show that the average bit-error probability for BPSK (QPSK) with coherent detection, maximal-ratio combining, independent Rayleigh fading, and Gaussian noise is

$$P = \frac{1}{2}\left\{1 - \sqrt{\frac{\Gamma}{1+\Gamma}}\left[1 + \frac{1}{1!2}(1+\Gamma)^{-1} + \frac{1\cdot 3}{2!2^2}(1+\Gamma)^{-2} + \cdots + \frac{1\cdot 3\cdot 5\cdot \cdots \cdot (2M-3)}{(M-1)!2^{M-1}}(1+\Gamma)^{-(M-1)}\right]\right\}, \tag{22}$$

where Γ is the average s/n per branch and where M is the number of branches, $M \geq 2$. Corresponding formulas are also described in Refs. 23 through 25. For $M = 1$, see Ref. 7.

The average error probability for a modulation scheme with bit-error probability $P(\gamma) = Q(\sqrt{2\alpha\gamma})$ for the additive white Gaussian channel is given by eq. (22) with Γ replaced by $\alpha\Gamma$ (see Refs. 21 and 22).

For the case of coherent MSK-type reception of partial-response continuous-phase modulated signals (see Section I), the error probability after the detection filter but before differential detection is a weighted sum of Q-functions, as shown in eq. (14). Thus, the bit-error probability for the fading and diversity case is given by

$$P = \sum_{i=1}^{m} \frac{C_i}{2} \left\{ 1 - \sqrt{\frac{\frac{d_i^2\Gamma}{2}}{1 + \frac{d_i^2\Gamma}{2}}} \right.$$
$$\cdot \left[1 + \frac{1}{1!2} \left(1 + \frac{d_i^2\Gamma}{2} \right)^{-1} + \frac{1 \cdot 3}{2!2^2} \left(1 + \frac{d_i^2\Gamma}{2} \right)^{-2} \right.$$
$$\left. \left. + \cdots + \frac{1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2M-3)}{(M-1)!2^{(M-1)}} \left(1 + \frac{d_i^2\Gamma}{2} \right)^{-(M-1)} \right] \right\}, \quad (23)$$

where C_i, d_i^2, $i = 1 \cdots m$ depend on the modulation scheme and the receiver filter used (see Section I). Numerical calculations of (23) for various modulation schemes and various M are presented in Section IV below.

For large s/n's Γ, the formula (22) reduces to[6,7,21]

$$P \approx \frac{(2M-1)!}{M!(M-1)!} \left(\frac{1}{4\Gamma} \right)^M. \quad (24)$$

It is easy to adapt this formula to the partial-response continuous-phase modulation case. For large s/n's eq. (23) can be written as

$$P \approx \frac{(2M-1)!}{M!(M-1)!} \left(\frac{1}{4\Gamma} \right)^M \cdot \sum_{i=1}^{m} \frac{C_i}{\left(\frac{d_i^2}{2} \right)^M}. \quad (25)$$

To evaluate the asymptotic s/n difference between different schemes, it is convenient to write eq. (25) as

$$P \approx \frac{(2M-1)!}{M!(M-1)!} \left\{ \frac{1}{4\Gamma \left[\frac{1}{\sum_{i=1}^{m} \frac{C_i}{\left(\frac{d_i^2}{2} \right)^M}} \right]^{\frac{1}{M}}} \right\}^M \quad (26)$$

The special case of BPSK (2PSK) is given by $m = 1$, $C_i = 1, d_i^2 = 2$.

Thus, the asymptotic E_b/N_o difference between two schemes, e.g., between QPSK (BPSK) and some partial-response CPM schemes is given by the factor

$$\frac{1}{\left[\sum_{i=1}^{m} \frac{C_i}{\left(\frac{d_i^2}{2} \right)^M} \right]^{\frac{1}{M}}} \quad (27)$$

From the relationships above and (17) we can draw the following conclusions:

For $M = 1$, the difference in E_b/N_o between a partial-response scheme and QPSK is smaller than the difference given by $d^2_{\min}$ alone. All the d_i^2 affect the difference with equal weight $C_i = 1/m$. For increasing M values it can easily be concluded from the average in eq. (27) that the smallest d_i^2, i.e., $d^2_{\min}$, will play an increasing role and dominate the difference for large M, just as for the Gaussian channel.

These relative performance differences for various M values are illustrated by the numerical examples in Section IV.

III. SELECTION COMBINING

In this section we will derive formulas corresponding to eqs. (22), (23), and (24) for the case of diversity with ideal selection combining.

The ideal selection-diversity combiner is defined here as a device that selects that diversity branch with the largest s/n for bit decisions. The same branch is used for all symbols over one time interval for the receiver filter, under the assumption of slow fading.

Let γ_k be the instantaneous signal-to-noise ratio in branch k with average value Γ equal for all branches. For Rayleigh fading, the γ_k have the probability density function of eq. (1). As shown in Refs. 6 and 7, the probability density function for the ideal selection-combiner output s/n γ is

$$f(\gamma) = \frac{M}{\Gamma} e^{-\gamma/\Gamma}(1 - e^{-\gamma/\Gamma})^{M-1}, \tag{28}$$

where Γ is the average s/n per branch and M is the number of branches. The average receiver-output s/n in this case is

$$E\{\gamma\} = \Gamma \cdot \sum_{k=1}^{M} \frac{1}{k}, \tag{29}$$

which of course increases more slowly with increasing M than the corresponding average for maximal ratio combining (21).[6,7]

First we derive an analytical solution to (17) for selection combining and coherent BPSK. It is then easy to apply this solution to the error probability for partial-response continuous-phase modulation (14). The probability density function (28) can be written as

$$f(\gamma) = \frac{M}{\Gamma} \sum_{j=0}^{M-1} (-1)^j \binom{M-1}{j} e^{-\frac{\gamma}{\Gamma}(1+j)}. \tag{30}$$

The average bit-error probability for BPSK, coherent-detection and M-branch diversity with selection combining is

$$\begin{aligned} P &= \int_o^\infty f(\gamma) Q(\sqrt{2\gamma}) d\gamma \\ &= \frac{M}{\Gamma} \sum_{j=o}^{M-1} (-1)^j \binom{M-1}{j} \int_o^\infty Q(\sqrt{2\gamma}) e^{-\frac{\gamma}{\Gamma}(1+j)} d\gamma. \end{aligned} \tag{31}$$

We evaluate (31) for the more general family of modulation schemes where the bit-error probability for the Gaussian channel is $P(\gamma) = Q(\sqrt{2\alpha\gamma})$; here α is a constant ≤ 1.

The integral above can be evaluated analytically as

$$\int_o^\infty Q(\sqrt{2\alpha\gamma}) e^{-\gamma/\Gamma(1+j)} d\gamma$$

$$= \int_o^\infty \left\{ \int_{\sqrt{2\alpha\gamma}}^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \right\} e^{-\frac{\gamma}{\Gamma}(1+j)} d\gamma$$

$$= \frac{1}{2} \cdot \frac{\Gamma}{1+j} \left(1 - \frac{1}{\sqrt{1 + \frac{1+j}{\alpha\Gamma}}} \right). \tag{32}$$

Thus, with (31) we have P:

$$P = \frac{M}{2} \sum_{j=0}^{M-1} \frac{(-1)^j}{(1+j)} \cdot \binom{M-1}{j} \left(1 - \frac{1}{\sqrt{1 + \frac{1+j}{\alpha\Gamma}}} \right), \tag{33a}$$

which can be written more compactly as

$$P = \frac{1}{2} \sum_{j=0}^{M} \frac{(-1)^j \binom{M}{j}}{\sqrt{1 + \frac{j}{\alpha\Gamma}}}. \tag{33b}$$

For the general case with an MSK-type receiver for a partial-response continuous-phase modulation scheme with $h = 1/2$, as shown in eq. (14), we arrive at the analytical error-probability formula

$$P = \sum_{i=1}^{m} C_i \cdot \frac{1}{2} \sum_{j=0}^{M} \frac{(-1)^j \binom{M}{j}}{\sqrt{1 + \frac{2j}{d_i^2 \Gamma}}}, \tag{34}$$

where C_i, d_i^2, $i = 1 \cdots m$ are defined by the modulation scheme and by the receiver. Numerical examples for specific transmitted-signal formats, receiver filters, and number of diversity branches M are given for (34) in Section IV below.

Next, we evaluate the dominating term in (34) for large signal-to-noise ratios Γ. For simplicity we do this for BPSK ($\alpha = 1$). This can then easily be extended to the general case (34), as we will see. For large Γ's, eq. (33) with $\alpha = 1$ can be written as

$$P = \frac{1}{2}\sum_{i=0}^{M}(-1)^i\binom{M}{i} \cdot\left\{1+\sum_{k=1}^{\infty}(-1)^k\frac{1\cdot 3\cdot\ \cdots\ \cdot(2k-1)}{2^k k!}\left(\frac{i}{\Gamma}\right)^k\right\}. \tag{35}$$

Since

$$\sum_{i=0}^{M}(-1)^i\binom{M}{i} = 0 \tag{36}$$

we have

$$P = \frac{1}{2}\sum_{k=1}^{\infty}(-1)^k\frac{1\cdot 3\cdot\ \cdots\ \cdot(2k-1)}{2^k\cdot k!}\left(\frac{1}{\Gamma}\right)^k\cdot\sum_{i=0}^{M}(-1)^i\binom{M}{i}i^k. \tag{37}$$

The quantity

$$S = \sum_{i=0}^{M}(-1)^i\binom{M}{i}i^k \tag{38}$$

is related to the Stirling numbers of the second kind (see Ref. 26, Chapter 24 and Ref. 27, p. 33). S is 0 for $k < M$ and

$$S = (-1)^M\cdot M! \tag{39}$$

for $k = M$. S can also be evaluated for $k > M$, as shown in Refs. 26 and 27. Thus, for large Γ's, the error probability behaves like

$$P \cong \frac{1\cdot 3\cdot\ \cdots\ \cdot(2M-1)}{2^{M+1}}\frac{1}{\Gamma^M}. \tag{40}$$

The power of $1/\Gamma$ is the same as for M-branch diversity with maximal-ratio combining. The multiplying coefficient is larger, however.

For high-average-branch s/n's Γ, the asymptotic error-probability behavior for coherent BPSK with M-branch diversity is shown in Table I below.

Table 1—Asymptotic behavior of average probability of error for large average per-branch signal-to-noise Γ for selection and maximal-ratio combining using *BPSK* (QPSK).

M	Selection Combining	Maximal-Ratio Combining
1	$\frac{1}{4\Gamma}$	$\frac{1}{4\Gamma}$
2	$\frac{3}{8}\frac{1}{\Gamma^2}$	$\frac{3}{16}\frac{1}{\Gamma^2}$
3	$\frac{15}{16}\frac{1}{\Gamma^3}$	$\frac{5}{32}\frac{1}{\Gamma^3}$
4	$\frac{105}{32}\frac{1}{\Gamma^4}$	$\frac{35}{256}\frac{1}{\Gamma^4}$

For partial response $h = 1/2$ continuous phase-modulation with an MSK-type receiver, the asymptotic behavior for selection combining is

$$P = \frac{1 \cdot 3 \cdot \cdots \cdot (2M-1)}{2^{(M+1)}} \left\{ \frac{1}{\Gamma \left[\frac{1}{\sum_{i=1}^{m} \frac{C_i}{\left(\frac{d_i^2}{2}\right)^M}} \right]^{\frac{1}{M}}} \right\}^M \tag{41}$$

Comparing formula (41) with the corresponding one for maximal-ratio combining (24), it can immediately be seen that the relative asymptotic performances of various modulation schemes are the same for the selection combiner as for the maximal-ratio combiner. This is also illustrated by the curves in Section IV.

IV. NUMERICAL RESULTS

Formula (23) for maximal-ratio combining and (34) for selection combining are used to calculate the error probability P in the figures below. Again note that the error probability is that found after the filter in the quadrature arm in Fig. 1.

Figure 7 shows the error probability P versus the average per-branch (and per-bit) signal-to-noise ratio Γ for the modulation scheme $3RC$, $h = \frac{1}{2}$, with a receiver filter of MIN-type.[13] Formula (23) was used and QPSK results are shown for comparison. Note that the difference between QPSK and $3RC$ grows with M. Also compare the difference in Figs. 5 and 6 for the nonfading-additive white Gaussian channel. As was predicted by the formula in Section II above, the asymptotic difference between QPSK and $3RC$ is smaller for $M = 1$ than for $M = 16$ and for the additive white Gaussian channel. This is due to the fact that, for low M, low signal-to-noise ratios dominate the density function $f(\gamma)$, as shown in eq. (1). For this region of γ, the difference between the error probability for QPSK and $3RC$ is small for the additive white Gaussian channel (see Figs. 5 and 6).

Figure 8 shows the same modulation and diversity schemes as Fig. 7 with the exception that for $3RC$ a SPAM-receiver filter is used (see the introduction and Ref. 13). Notice that the differences between the corresponding curves in Figs. 7 and 8 are small. The MIN filter is better than the SPAM filter for high s/n's over the additive white Gaussian channel (see Figs. 5 and 6 and Ref. 13). The SPAM filter is better than the MIN filter for low s/n's. This explains the small but noticeable differences between the $3RC$ curves in Figs. 7 and 8.

Figures 9 and 10 show the $3RC$-MIN and $3RC$-SPAM cases for selection combining with QPSK results shown for comparison. Formula (34) was

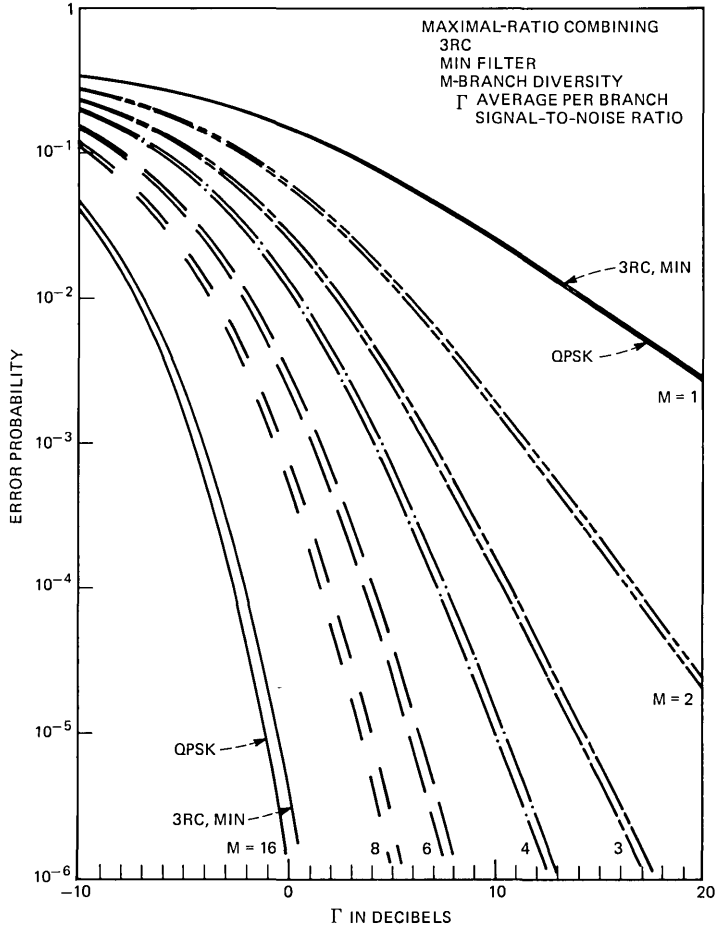


Fig. 7—Error probability P vs average-per-branch s/n Γ with $3RC$, $h = 1/2$, receiver with MIN-filter. (The curves are shown without differential encoding/decoding.)

used for the computations. Note that the same relative behavior is present in Figs. 9 and 10 as was shown in Figs. 7 and 8. Also note the larger per-branch s/n required for selection combining compared to maximal-ratio combining for equal M.[6,7]

The exact analytical formulas (23) and (34) are easy and straightforward to evaluate. For example, for $M = 16$ and for the $3RC$-SPAM case with a filter of length 4 binary symbols, $16x2^8$ terms are used in the sums. In general, $Mx2^{N_T+L+1}$ terms are summed where N_T is the filter impulse response length in bits.[13]

Figures 11 to 14 show the error probability P versus the average-per-branch s/n Γ for an increasing number of diversity branches with

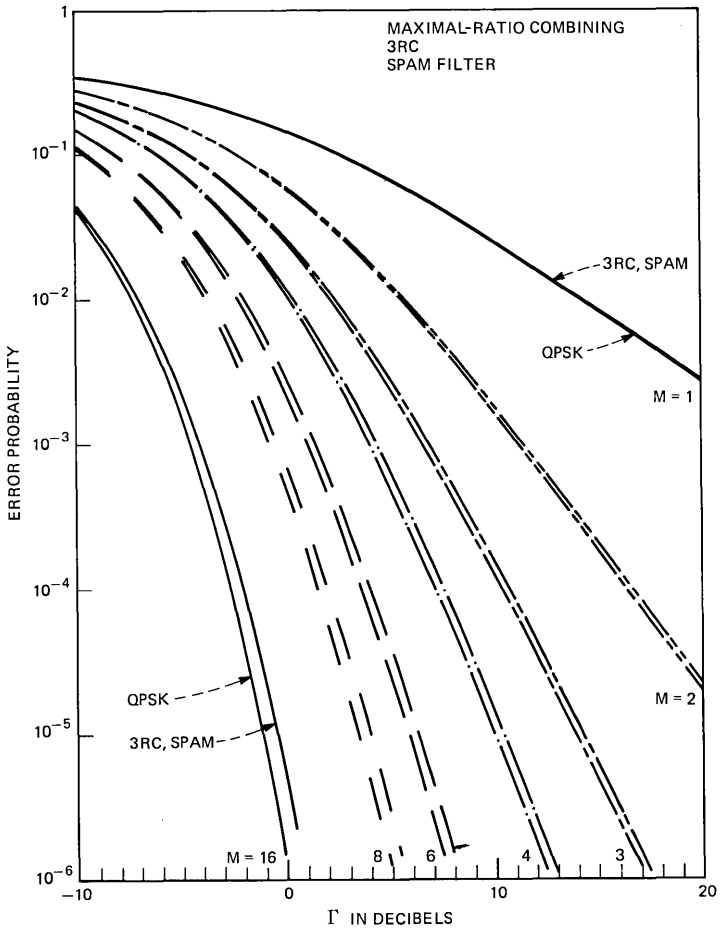


Fig. 8—Error probability P vs average-per-branch s/n Γ with $3RC$ and SPAM-filter M-branch diversity with maximal-ratio combining.

maximal-ratio combining. Again, QPSK (BPSK) results are shown for comparison. The same group of modulation schemes and receiver filters are used. Note the widening spread for increasing M, as expected. The corresponding case for selection combining is shown in Figs. 15 through 17 (and in Fig. 11 for the $M = 1$ case).

Figure 18 shows another way of presenting the relative error-probability results. This figure shows the required receiver output s/n $(M \cdot \Gamma)$ to achieve bit-error probability 10^{-3} as a function of the number of branches of diversity M. Maximal-ratio combining is assumed. The modulation schemes are coherent $2RC$, $3RC$, and $4RC$ with SPAM filters,[13] with QPSK shown for reference. For increasing M, the s/n for

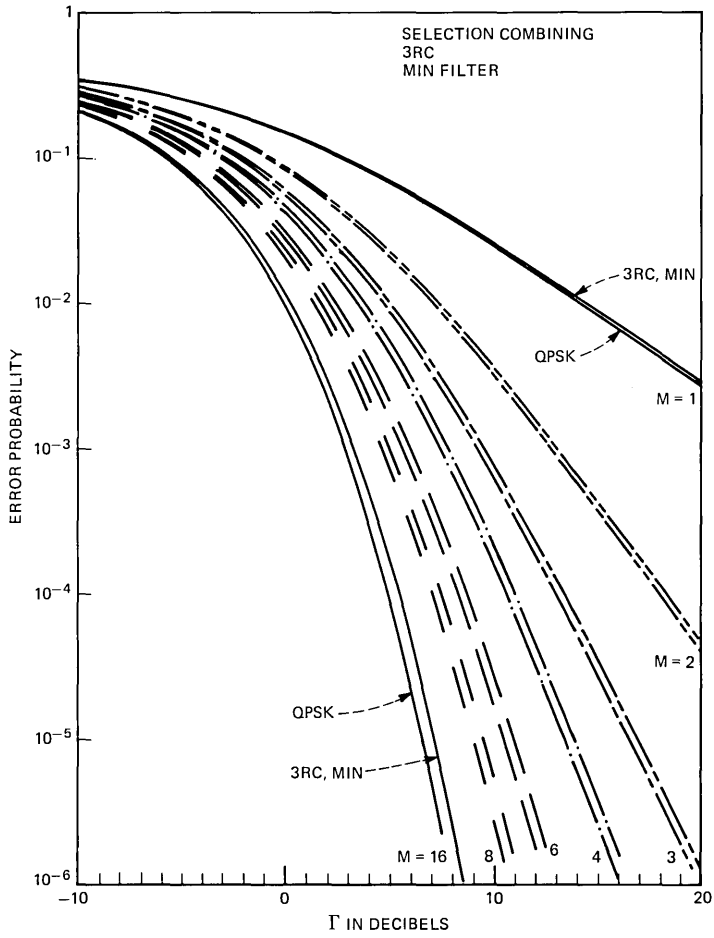


Fig. 9—$3RC$ with MIN filter and M-branch diversity with selection combining.

the additive Gaussian channel is approached. Note the relative position of the curves for $2RC$, $3RC$, and $4RC$. For further numerical results, see Ref. 28.

V. DISCUSSION AND CONCLUSIONS

The curves in Figs. 7 through 18 are calculated with the assumption that the receiver can resolve phase ambiguity of 180 degrees in the detection process.[1,2] This is normally done by employing differential encoding and decoding[1–3,17] and, likewise, for QPSK. For this case, the bit-error probability for the Gaussian channel is[19]

$$P_{bit} = 2P(1 - P), \tag{42}$$

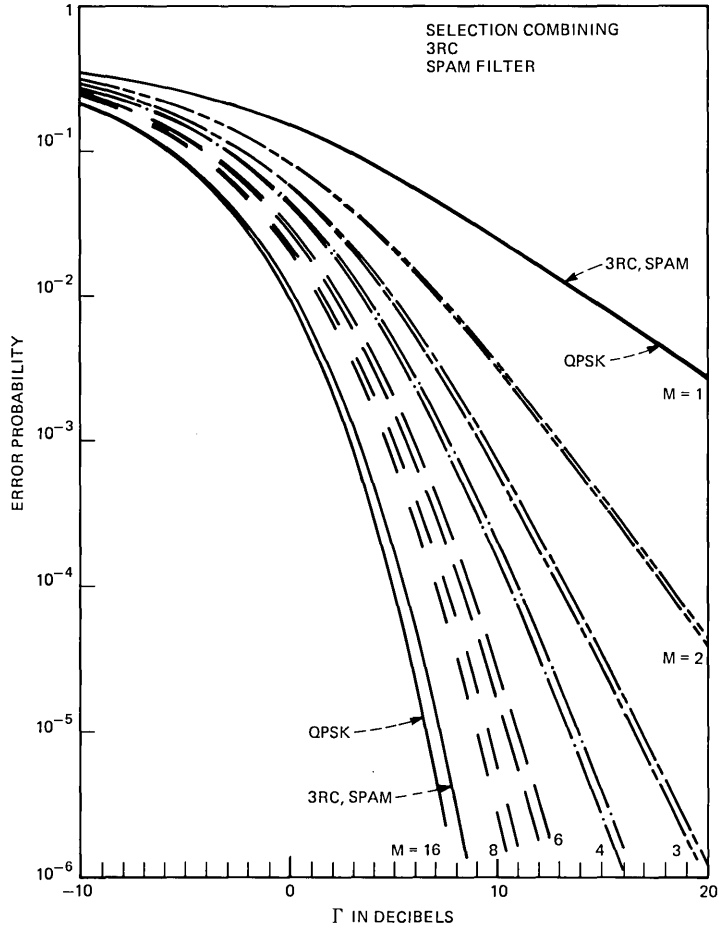


Fig. 10—$3RC$ with SPAM-filter and M-branch diversity with selection combining.

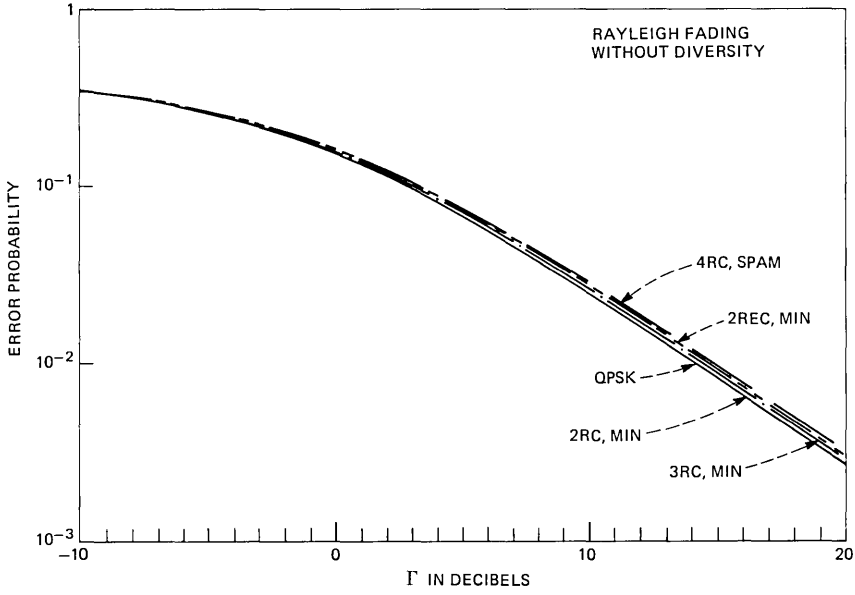


Fig. 11—Error probability P vs s/n Γ for various selected $h = 1/2$, binary CPM schemes with selected receiver filters. On each curve the modulation format and the receiver filter are given.

where P is the error probability in the symbol decision after the linear filter in each quadrature arm. Thus, for the fading and diversity case, the same relative curves are obtained as upper bounds

$$P_{bit} \leq 2P \tag{43}$$

on the bit-error probability with differential encoding/decoding. In absolute numbers, all error probabilities P must be multiplied by 2. In principle, averaging of the type in (17) can be performed numerically for (42) but, especially for large s/n's and large numbers of diversity branches, the bound based on eq. (43) is very tight.

The calculations of the bit-error probability for TFM with fading are given in Ref. 28. This lies, as expected, between that of $3RC$ and $4RC$.

The performance of GMSK in fading for various values of the parameter B_bT—affecting the length of the pulse $g(t)$—can be estimated by comparisons to the raised-cosine pulses.[3,4] Compare the spectra and eye patterns in Refs. 3 and 4 with those for raised-cosine pulses in Refs. 11 through 13. Approximate fading and diversity ($M = 2$) error-probability behavior is given in Ref. 3 for some cases of GMSK. A detailed comparison of raised-cosine schemes and GMSK is given in Ref. 28.

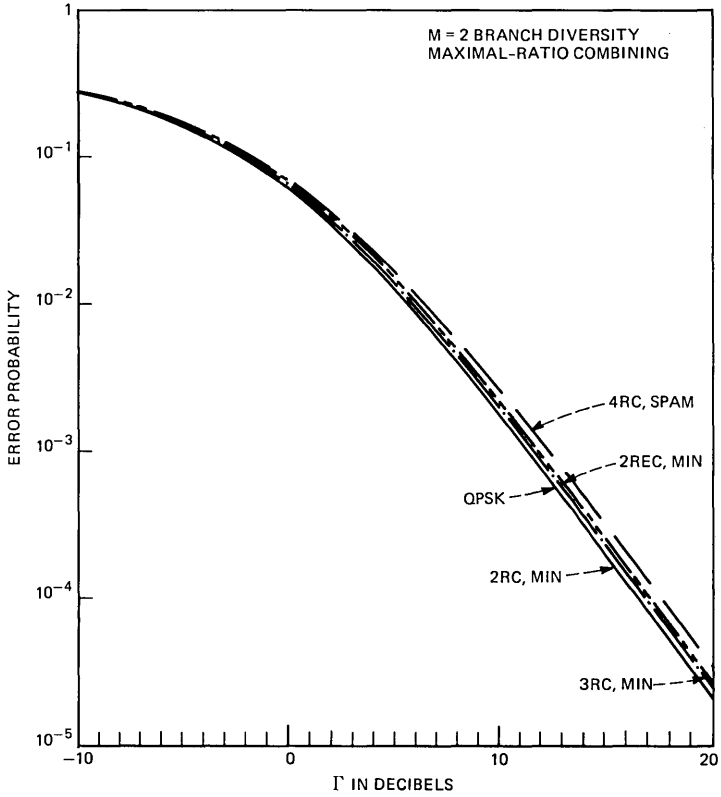


Fig. 12—Error probability P vs average-per-branch s/n Γ for the group of modulation schemes in Fig. 11. $M = 2$ branch diversity with maximal-ratio combining is used.

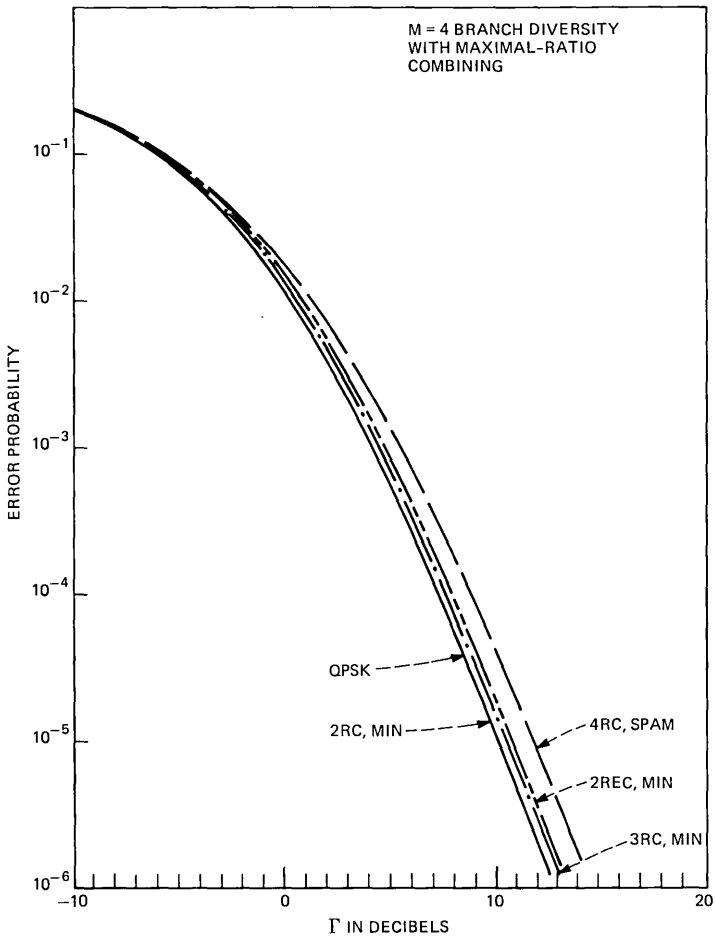


Fig. 13—Error probability P vs average-per-branch s/n Γ for $M = 4$ branch diversity with maximal-ratio combining.

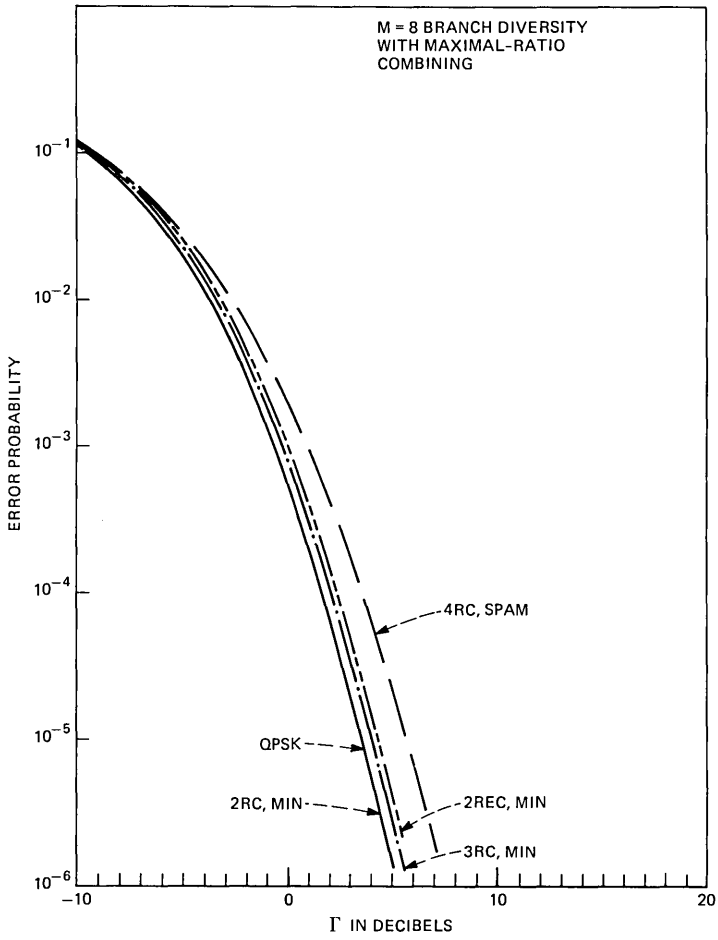


Fig. 14—Error probability P vs average-per-branch s/n Γ for $M = 8$ branch diversity with maximal-ratio combining.

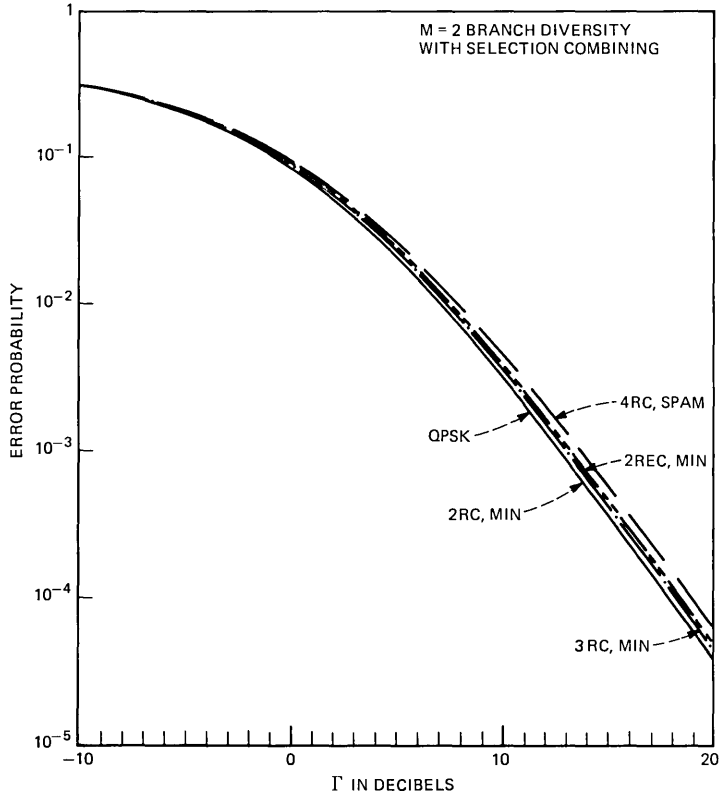


Fig. 15—Error probability P vs average-per-branch s/n Γ for $M = 2$ branch diversity with selection combining.

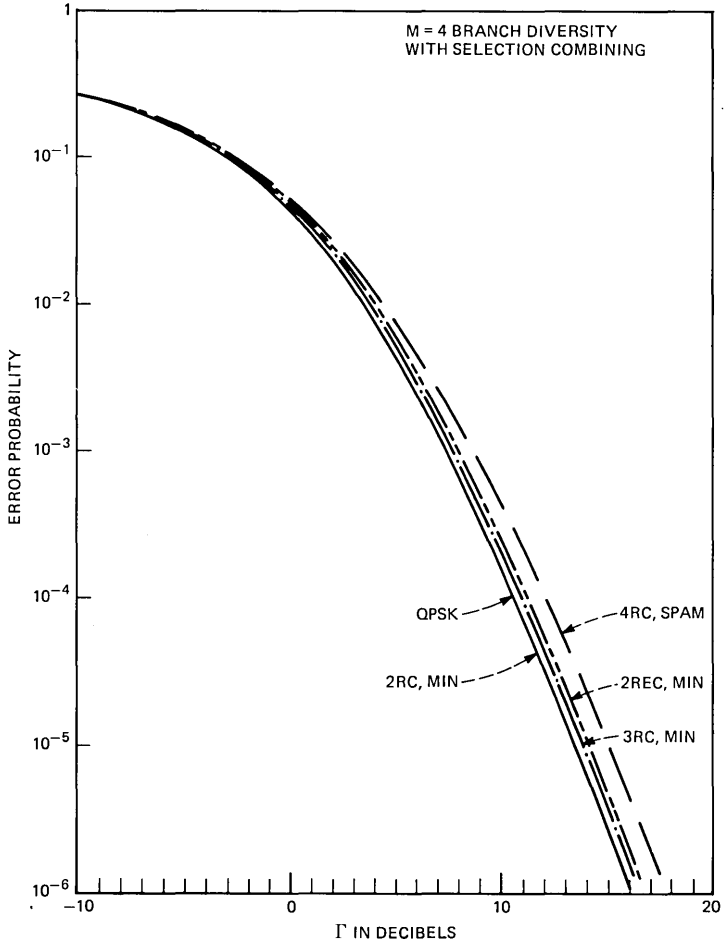


Fig. 16—Error probability P vs average-per-branch s/n Γ for $M = 4$ branch diversity with selection combining.

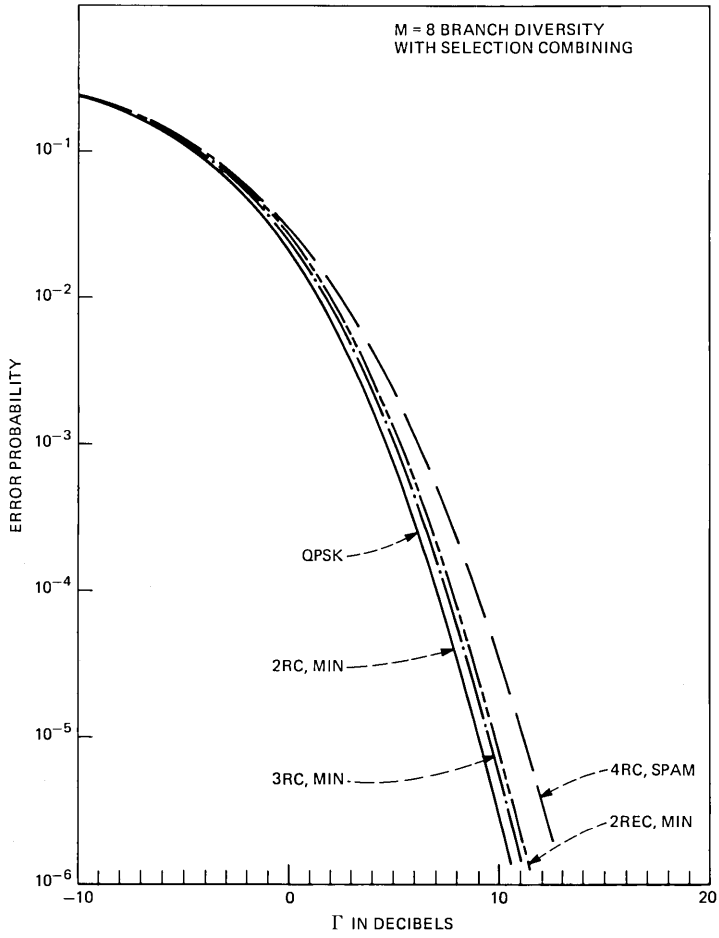


Fig. 17—Error probability P vs average-per-branch s/n Γ for $M = 8$ branch diversity with selection combining.

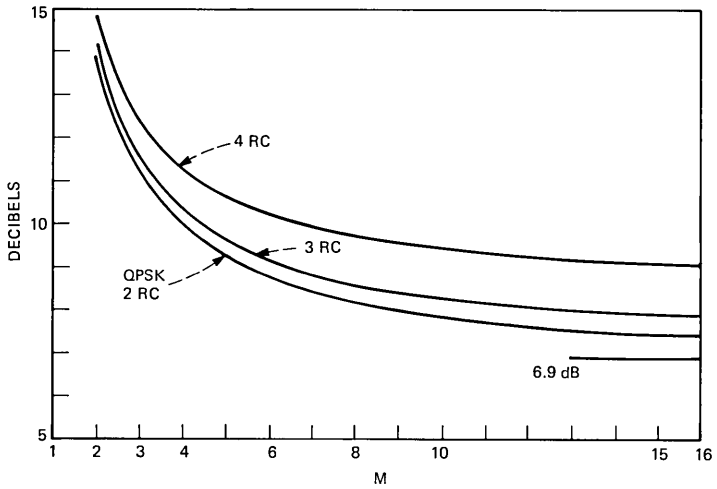


Fig. 18—Required receiver output s/n to achieve 10^{-3} error probability as a function of the number of branches for coherent detection of QPSK and for $2RC$, $3RC$, and $4RC$ with SPAM filters.

This paper contains formulas for the calculations of the bit-error probability for a class of modulation schemes with attractive spectral behavior. The class contains such modulation schemes as MSK, FFSK, TFM, and GMSK, as defined above. A simple suboptimum coherent detector can be used with performance approaching the optimum detector. The channel is assumed to be a slow Rayleigh fading channel with Gaussian noise, and diversity is employed to combat multipath fading. We assume the receiver uses either ideal maximal-ratio combining or selection combining. Analytical formulas are derived for both cases, and simple asymptotic expressions for large signal-to-noise ratios are also derived and discussed. It is noted from the numerical calculations and also from the asymptotic formulas that the difference in E_b/N_o between various modulation schemes in the considered class at a given error probability decreases as the number of branches of diversity decreases.

VI. ACKNOWLEDGMENT

We owe our thanks to Arne Svensson who computed the error-probability curves in this paper.

REFERENCES

1. R. deBuda, "Coherent Demodulation of Frequency-Shift Keying with Low Deviation Ratio," IEEE Trans. Commun., *COM-20* (June 1972), pp. 429–35.
2. F. deJager and C. B. Dekker, "Tamed Frequency Modulation, A Novel Method to

Achieve Spectrum Economy in Digital Transmission," IEEE Trans. Commun., *COM-26*, No. 5 (May 1978), pp. 534–42.
3. K. Murota, K. Kinoshita, and K. Hirade, "Spectrum Efficiency of GMSK Land Mobile Radio," ICC *81*, Denver, June 1981, Conf. Rec., pp. 23.8.1–23.8.5.
4. K. Murota and K. Hirade, "GMSK Modulation for Digital Mobile Radio Telephony," IEEE Trans. Commun., *COM-29*, No. 7 (July 1981), pp. 1044–50.
5. D. Muilwijk, "Tamed Frequency Modulation—A Bandwidth-Saving Digital Modulation Method Suited for Land Mobile Radio," Philips Telecommun. Rev., *37*, No. 1 (March 1979), pp. 35–49.
6. W. C. Jakes, Jr., *Microwave Mobile Communications*, New York: Wiley, 1974.
7. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.
8. Y. S. Yeh and D. O. Reudink, "Efficient Spectrum Utilization for Mobile Radio Systems Using Space Diversity," Proceedings—IEE Conf. on Radio Spectrum Conversion Techniques, London, July 7–9, 1980.
9. P. S. Henry and B. S. Glance, "A New Approach to High Capacity Digital Mobile Radio," B.S.T.J., *60*, No. 8 (October 1981), pp. 1891–1904.
10. C-E. Sundberg, unpublished work.
11. T. Aulin and C-E. Sundberg, "Continuous Phase Modulation-Part I: Full Response Signaling," IEEE Trans. Commun., *COM-29*, No. 3 (March 1981), pp. 196–209.
12. T. Aulin, N. Rydbeck, and C-E. Sundberg, "Continuous Phase Modulation-Part II: Partial Response Signaling," IEEE Trans. Commun., *COM-29*, No. 3 (March 1981), pp. 210–25.
13. T. Aulin, C-E. Sundberg, and A. Svensson, "MSK-Type Receivers for Partial Response Continuous Phase Modulation," Int. Conf. Commun., Philadelphia, Pennsylvania, June 1982, Conf. Rec., pp. 6E3.1–6E3.6.
14. T. Aulin and C-E. Sundberg, "Numerical Calculation of Spectra for Digital FM Signals," Nat. Telecommun. Conf., NTC *81*, New Orleans, Louisiana, December 1981, Conf. Proc., pp. D.8.3.1–D.8.3.7.
15. T. Aulin, C-E. Sundberg, and A. Svensson, "Viterbi Detectors with Reduced Complexity for Partial Response Continuous Phase Modulation," Nat. Telecommun. Conf., NTC *81*, New Orleans, Louisiana, December 1981, Conf. Proc., pp. A.7.6.1–A.7.6.7.
16. W. Hirt and S. Pasupathy, "Suboptimal Reception of Binary CPSK Signals," Proc. IEE, Part F, *128*, No. 3 (June 1981), pp. 125–34.
17. P. Galko and S. Pasupathy, "On a Class of Generalized MSK," Proc., Int. Conf. Commun., ICC *81*, Denver, June 1981, pp. 2.4.1–2.4.5.
18. P. Galko and S. Pasupathy, "Generalized MSK," Int. Elec. Electron. Conf. and Expo., Toronto, Canada, October 5–7, 1981, Conf. Proc.
19. W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, New York: Prentice Hall, 1973.
20. C-E. Sundberg, "Block Error Probability for Noncoherent FSK with Diversity for Very Slow Rayleigh Fading in Gaussian Noise," IEEE Trans. Commun., *COM-29*, No. 1 (Janaury 1981), pp. 57–60.
21. D. O. Reudink, private communication.
22. B. Glance, private communication.
23. P. Bello and B. D. Nelin, "Predetection Diversity Combining with Selectively Fading Channels," IRE Trans. Commun. Syst., *CS-20* (March 1962), pp. 32–42.
24. K. Brayer, ed., *Data Communications via Fading Channels*, New York: IEEE Press, 1975.
25. O. Yue, "Frequency-Hopping, Multiple-Access, Phase-Shift-Keying System Performance in a Rayleigh Fading Environment," B.S.T.J., *59*, No. 6 (July–August 1980), pp. 861–79.
26. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publications, 1965.
27. J. Riordan, *An Introduction to Combinatorial Analysis*, New York: Wiley, 1958.
28. C-E. Sundberg and A. Svensson, "Calculation of the Exact Error Probability for Partial Response Continuous Phase Modulation with Coherent MSK-type Receivers, Diversity with Maximal Ratio Combining and Selection Combining and Slow Rayleigh Fading in Gaussian Noise," Telecommun. Theory, University of Lund, Technical Report TR-158, September 1981.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Comparisons on Blocking Probabilities for Regular Series Parallel Channel Graphs

By D. Z. Du and F. K. HWANG

(Manuscript received January 19, 1982)

We give a sufficient condition for one regular series parallel channel graph to be superior to another with the same number of stages. The main mathematical tools used for doing this are the recently developed results on majorization over a partial order.

I. INTRODUCTION

An s-stage channel graph is a graph whose vertices can be partitioned into s subsets (stages) $V_1, V_2, \cdots, V_s$, with V_1 and V_s each containing a single vertex (called the source and the sink, respectively), and whose edges can be partitioned into $s-1$ subsets $E_1, E_2, \cdots, E_{s-1}$ such that

(*i*) Edges in E_i connect vertices in V_i to vertices in V_{i+1},

(*ii*) Each vertex in V_i, $1 < i < s$, is connected to at least one vertex in each of V_{i-1} and V_{i+1}.

A channel graph is regular if for each i, the numbers of edges in E_{i-1} and E_i coincident to a vertex in V_i are independent of which vertex is chosen.

A series combination of an s-stage channel graph G and a t-stage channel graph H is a union of G and H into an $(s+t-1)$-stage channel graph, with the sink of G identified with the source of H. A parallel combination of two s-stage channel graphs is a union of these two graphs into another s-stage channel graph with the source and the sink of one graph being identified with the source and the sink, respectively, of the other graph. A channel graph is series parallel if it is either an edge or is constructable from two smaller series parallel channel graphs by either a series or a parallel combination. A series parallel canopy is a special case of a series parallel channel graph in which parallel combinations are allowed only when at least one of the two component subgraphs consists solely of a single edge.

Each edge in a channel graph can be in one of two states, occupied

or idle. In this paper, we follow Lee's assumption[1] that the states of the edges are independent and that each edge in E_i has probability p_i, called the occupancy for E_i, of being occupied. The blocking probability of a channel graph is the probability that every *channel*—by which we mean a path from source to sink consisting of one edge from each E_i—contains at least one occupied edge. An s-stage channel graph is said to be superior to another s-stage channel graph if the blocking probability of the former never exceeds that of the latter, independent of the occupancies for the E_i (common to both graphs).

Chung and Hwang[2] showed that a regular series parallel channel graph (hereafter referred to as rspcg) without multiple edges can be uniquely represented by its degree vector. They also proved that in the case of two s-stage regular series parallel canopies, a necessary and sufficient condition for one graph to be superior to the other is that the degree vector of the former "majorizes" that of the latter. They conjectured that the same condition might also hold for rspcg's. However, counterexamples to the sufficiency of the condition for rspcg's were given in Refs. 3 and 4. In this paper, we give a sufficient condition for one s-stage rspcg to be superior to another, with multiple edges between two vertices allowed, by using the recently developed results of majorization over a partial order.[5,6]

II. MAJORIZATION OVER A PARTIAL ORDER

A set of numbers $A = \{a_1 \geq a_2 \geq \cdots \geq a_n\}$ is said to be *weakly submajorized*[7] by another set of numbers $B = \{b_1 \geq b_2 \geq \cdots \geq b_n\}$ if

$$\sum_{i=1}^{k} a_i \leq \sum_{i=1}^{k} b_i \qquad \text{for each} \quad k = 1, \cdots, n.$$

If, in addition,

$$\sum_{i=1}^{k} a_i = \sum_{i=1}^{k} b_i,$$

then A is simply said to be majorized by B.

The above concept of set majorization has been extended to majorization over a partial order.[5,6] Let $P = \{S, \rightarrow\}$ denote a partial order on S, where S is a set of n elements and $s_i, s_j \in S$, $s_i \rightarrow s_j$ indicates that s_i is greater than s_j in P. Let $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_n\}$, where A and B can be thought of as two sets of weights for the elements in S. Then A is said to be weakly submajorized by B on P if for every filter S_j of S,

$$\sum_{s_i \in S_j} a_i \leq \sum_{s_i \in S_j} b_i,$$

where S_j is a filter if $s_i \in S_j$ and $s_k \to s_i \Rightarrow s_k \in S_j$. If equality holds for $S_j = S$, then A is simply said to be majorized by B on P.

Lemma 1. Suppose A is weakly submajorized by B on P. Then there exists $C = \{c_1, c_2, \cdots, c_n\}$, where $c_i \geq 0$ for all i, such that $A + C$ is majorized by B on P.

Proof: The proof is by induction on n. For $n = 1$, Lemma 1 is true by setting $c_1 = b_1 - a_1$. For general n, without loss of generality, assume that s_n is a minimal element in P. Set $c_n = b_n - a_n$. Then $A' = \{a_1, a_2, \cdots, a_{n-1}, a_n + c_n\}$ is still weakly submajorized by B on P, since for any filter S_j containing s_n,

$$\sum_{s_i \in S_j} b_i - \sum_{s_i \in S_j} a_i' = \sum_{s_i \in S_j - \{s_n\}} b_i - \sum_{s_i \in S_j - \{s_n\}} a_i \geq 0.$$

Next consider the partial order P on $S - \{s_n\}$. By our inductive assumption, there exists nonnegative $c_1, c_2, \cdots, c_{n-1}$ such that $(a_1 + c_1, a_2 + c_2, \cdots, a_{n-1} + c_{n-1})$ is majorized by $(b_1, b_2, \cdots, b_{n-1})$. Lemma 1 follows immediately.

We quote a result from Ref. 5:

Theorem 1: Let $f(x_1, x_2, \cdots, x_n)$ be a function defined over the domain D. Let $P = (X, \to)$ denote a partial order, where $X = \{x_1, x_2, \cdots, x_n\}$. Then

$$f(a_1, a_2, \cdots, a_n) \leq f(b_1, b_2, \cdots, b_n)$$

for all A majorized by B on P if and only if f is such that for every i and j,

$$x_i \to x_j \Rightarrow \frac{\partial f}{\partial x_i} \geq \frac{\partial f}{\partial x_j} \qquad \text{over all} \quad X \in D.$$

We now generalize Theorem 1 into Theorem 2.

Theorem 2: Let $f(x_1, x_2, \cdots, x_n)$ be a function defined over the domain D such that f is monotone nonincreasing in each of its arguments. Let $P = (X, \to)$ denote a partial order, where $X = \{x_1, x_2, \cdots, x_n\}$. Then

$$f(a_1, a_2, \cdots, a_n) \geq f(b_1, b_2, \cdots, b_n)$$

for all A weakly submajorized by B on P if and only if f is such that for every i and j,

$$x_i \to x_j \Rightarrow \frac{\partial f}{\partial x_i} \leq \frac{\partial f}{\partial x_j} \qquad \text{over all} \quad X \in D.$$

Proof:

(*i*) Assume $f(a_1, a_2, \cdots, a_n) \geq f(b_1, b_2, \cdots, b_n)$ for all A weakly submajorized by B on P. Then, in particular,

$$f(a_1, a_2, \cdots, a_n) \geq f(b_1, b_2, \cdots, b_n),$$

or equivalently,

$$-f(a_1, a_2, \cdots, a_n) \leq -f(b_1, b_2, \cdots, b_n)$$

for all A majorized by B. From Theorem 1, a necessary condition for this to happen is that for every i and j,

$$x_i \to x_j \Rightarrow \frac{\partial(-f)}{\partial x_i} \geq \frac{\partial(-f)}{\partial x_j},$$

or equivalently

$$\frac{\partial f}{\partial x_i} \leq \frac{\partial f}{\partial x_j} \qquad \text{over all} \quad X \in D.$$

(*ii*) Assume that for every i and j

$$x_i \to x_j \Rightarrow \frac{\partial f}{\partial x_i} \leq \frac{\partial f}{\partial x_j} \qquad \text{over all} \quad X \in D.$$

Let A be weakly submajorized by B on P and let $\sum_{i=1}^{n} b_i - \sum_{i=1}^{n} a_i = c \geq 0$. From Lemma 1, there exists nonnegative C such that $A + C$ is majorized by B. From Theorem 1,

$$-f(a_1 + c_1, a_2 + c_2, \cdots, a_n + c_n) \leq -f(b_1, b_2, \cdots, b_n).$$

Since f is monotone nonincreasing in each x_i, it follows that $f(a_1, a_2, \cdots, a_n) \geq f(a_1 + c_1, a_2 + c_2, \cdots, a_n + c_n) \geq f(b_1, b_2, \cdots, b_n)$. ∎

III. THE MAIN RESULTS

Theorem 2 will be used for comparing two rspcg's. To do this, however, we first have to define a partial order such that an rspcg can be represented as a set of weights for the elements of the partial order. This can be done by using the Takagi graph characterization of an rspcg.

An (i, j, r) multiplex, $1 \leq i \leq j \leq s$, of an s-stage channel graph G is an s-stage channel graph formed from the union of r copies of G, with the copies being merged into a single copy from stage 1 to stage i and from stage j to stage s. A channel graph is called a Takagi graph[8,9] if it can be obtained as a multiplex of a smaller Takagi graph, where the smallest Takagi graph of s stages is an s-stage path. An (i, j, r) multiplex can also be represented by the equation $m_{ij} = r$, where m_{ij} is called the *multiplex index* and r is the value of the index. Therefore, a Takagi graph can be represented by a set $\{m_{ij} = k\}$ called a multiplex set. Figure 1 illustrates how the Takagi graph $\{m_{13} = 3, m_{24} = 2\}$ is constructed. It is clear that adding or deleting a multiplex index with value one has no effect on the Takagi graph. Up to this

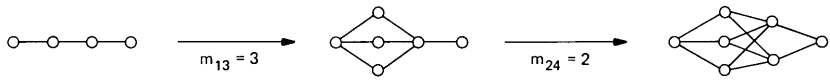


Fig. 1—A Takagi graph.

equivalence, it has been proved (see Ref. 4) that there exists a one-to-one mapping between multiplex sets and Takagi graphs, regardless of the ordering of the multiplex indices in the set. Furthermore, it is straightforward to verify that for an rspcg the product of all the values in its multiplex set equals the total number of distinct channels.

Let m_{ij} and m_{pq} denote two multiplex indices. Then m_{ij} is said to *cross* m_{pq} if $i < p < j < q$, and to *contain* m_{pq} if $i \leq p < q \leq j$. The following has been proved in Ref. 10:

Theorem 3: A channel graph is an rspcg if and only if it is a Takagi graph without crossing multiplex indices.

We define a partial order P_{1s} on the set of multiplex indices $\{m_{ij}: 1 \leq i < j \leq s\}$ by: $m_{ij} \rightarrow m_{pq}$ if m_{ij} contains m_{pq}. Then the multiplex set of any s-stage rspcg can be considered as a set of weights for the elements of P_{1s} (if m_{ij} is not in the multiplex set, we define $m_{ij} = 1$). For a given multiplex set M, we define M_{ij} to be the subset of M consisting of all multiplex indices contained by m_{ij}. We also let P_{ij} denote the partial order P restricted on M_{ij}. For fixed occupancies p_1, p_2, $\cdots$, p_{s-1}, let $B(M)$ denote the blocking probability for the Takagi graph with multiplex set M. Then, from Theorem 3, we have

$$B(M_{ij}) = \left\{1 - \prod_{l \in L_{ij}} (1 - p_l) \prod_{m_{pq} \in N_{ij}} [1 - B(M_{pq})]\right\}^{m_{ij}},$$

where $L_{ij} = \{l: m_{l,l+1} = 1, m_{ij} \rightarrow m_{l,l+1}$, but there does not exist $m_{uv} > 1$ such that $m_{ij} \rightarrow m_{uv} \rightarrow m_{l,l+1}\}$ and where $N_{ij} = \{m_{pq}: m_{pq} > 1, m_{ij} \rightarrow m_{pq}$, but there does not exist $m_{uv} > 1$ such that $m_{ij} \rightarrow m_{uv} \rightarrow m_{pq}\}$. We quote a result from Ref. 2:

Lemma 2: For given constants $c_1, c_2, \cdots, c_n$, all lying between zero and one, define

$$f(x_n) = (1 - c_n)^{x_n};$$

$$f(x_k, x_{k+1}, \cdots, x_n) = \{1 - c_k[1 - f(x_{k+1}, x_{k+2}, \cdots, x_n)]\}^{x_k}$$

$$\text{for} \quad k = 1, 2, \cdots, n - 1.$$

Suppose the vector $(\ln a_1, \ln a_2, \cdots, \ln a_n)$ is weakly submajorized by the vector $(\ln b_1, \ln b_2, \cdots, \ln b_n)$ where $\ln a_i$ and $\ln b_i$ are nonnegative for all i. Then

$$f(a_1, a_2, \cdots, a_n) \geq f(b_1, b_2, \cdots, b_n). \qquad \blacksquare$$

In particular, for any $w > 1$ and $i < j$, we have $f(a_1, a_2, \cdots, a_i, \cdots, a_j w, \cdots, a_n) \geq f(a_1, a_2, \cdots, a_i w, \cdots, a_j, \cdots, a_n)$. Therefore, we also have:

Corollary:

$$\frac{\partial f(x_1, x_2, \cdots, x_n)}{\partial \ln x_i} \leq \frac{\partial f(x_1, x_2, \cdots, x_n)}{\partial \ln x_j} \qquad \text{for} \quad i \leq j.$$

We are now ready to prove Theorem 4.

Theorem 4: An s-stage rspcg with the multiplex set $\{m_{ij} = a_{ij}\}$ is superior to another s-stage rspcg with the multiplex set $\{m_{ij} = b_{ij}\}$ if $\{ln\ b_{ij}\}$ is weakly submajorized by $\{ln\ a_{ij}\}$ on P_{1s}.

Proof: A straightforward induction proof shows that $B(M_{1s})$ is monotone nonincreasing in each $m_{ij} \in M_{1s}$. Therefore, if we can prove that for every $m_{uv} \rightarrow m_{xy}$,

$$\frac{\partial B(M_{1s})}{\partial \ln m_{uv}} \leq \frac{\partial B(M_{1s})}{\partial \ln m_{xy}},$$

then Theorem 4 will follow immediately from Theorem 2.

Consider a path Z from the top of P_{1s} to the bottom of P_{1s} which contains m_{uv} and m_{xy}. Let r_i, $i = 1, 2, \cdots, n$, denote the value of the ith multiplex index on this path. Suppose we hold every other m_{ij} constant except those on Z. Then $B(M_{1s})$ can be expressed as a function of $r_1, r_2, \cdots, r_n$ alone since all other m_{ij} are now constants. To be more specific, we have

$$B(r_n) = (1 - c_n)^{r_n}$$

and

$$B(r_k, r_{k+1}, \cdots, r_n) = \{1 - c_k[1 - B(r_{k+1}, r_{k+2}, \cdots, r_n)]\}^{r_k} \qquad \text{for} \quad k = 1, 2, \cdots, n-1.$$

From the Corollary of Lemma 2, we conclude that $i < j$ implies

$$\frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln r_i} \leq \frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln r_j}.$$

In particular, we have

$$\frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln m_{uv}} \leq \frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln m_{xy}}.$$

The proof is now complete by noting

$$\frac{\partial B(M_{1s})}{\partial \ln m_{uv}} = \frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln m_{uv}}$$

and

$$\frac{\partial B(M_{1s})}{\partial \ln m_{xy}} = \frac{\partial B(r_1, r_2, \cdots, r_n)}{\partial \ln m_{xy}}.$$

Define $\bar{M}_{1s} = \{m_{ij} \in M_{1s}: L_{ij} \neq \phi\}$ and define the partial order $\bar{P}_{1s}$ accordingly.

Theorem 5: An s-stage rspcg with the multiplex set $\{m_{ij} = a_{ij}\}$ *is superior to another s-stage rspcg with the multiplex set* $\{m_{ij} = b_{ij}\}$ *only if* $\{ln\ b_{ij}\}$ *is weakly submajorized by* $\{ln\ a_{ij}\}$ *on* $\bar{P}_{1s}$ *(associated with the* $\{a_{ij}\}$ *set).*

Proof: Consider two s-stage rspcg's A and B with multiplex numbers $\{m_{ij} = a_{ij}\}$ and $\{m_{ij} = b_{ij}\}$, respectively. Suppose there exists a filter $M \subset \bar{M}_{1s}$ such that

$$\sum_{m_{ij} \in M} \ln a_{ij} < \sum_{m_{ij} \in M} \ln b_{ij}.$$

Consider a set of occupancies $p_1, p_2, \cdots, p_{s-1}$ such that $p_k = 0$ if $m_{k,k+1}$ is contained by any m_{ij} not in M. Clearly, if all edges from stage i to stage j are idle, then we can set a_{ij} and b_{ij} to 1 without affecting the blocking probabilities of A and B. But now, owing to the assumption

$$\sum_{m_{ij} \in M} \ln a_{ij} < \sum_{m_{ij} \in M} \ln b_{ij},$$

the product of all a_{ij} in A is less than the product of all b_{ij} in B, or equivalently, there are fewer paths in A than in B. But it is well known that when the occupancies of all edges approach one,[4] then the blocking

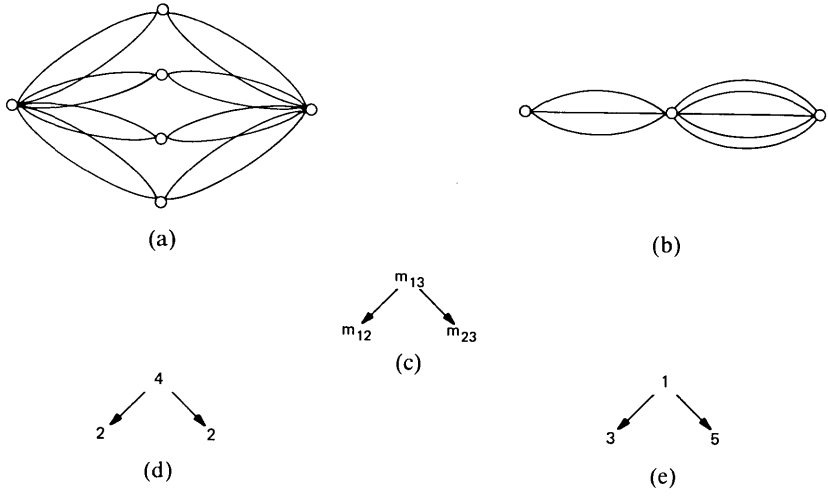


Fig. 2—Graphs for Example 1.

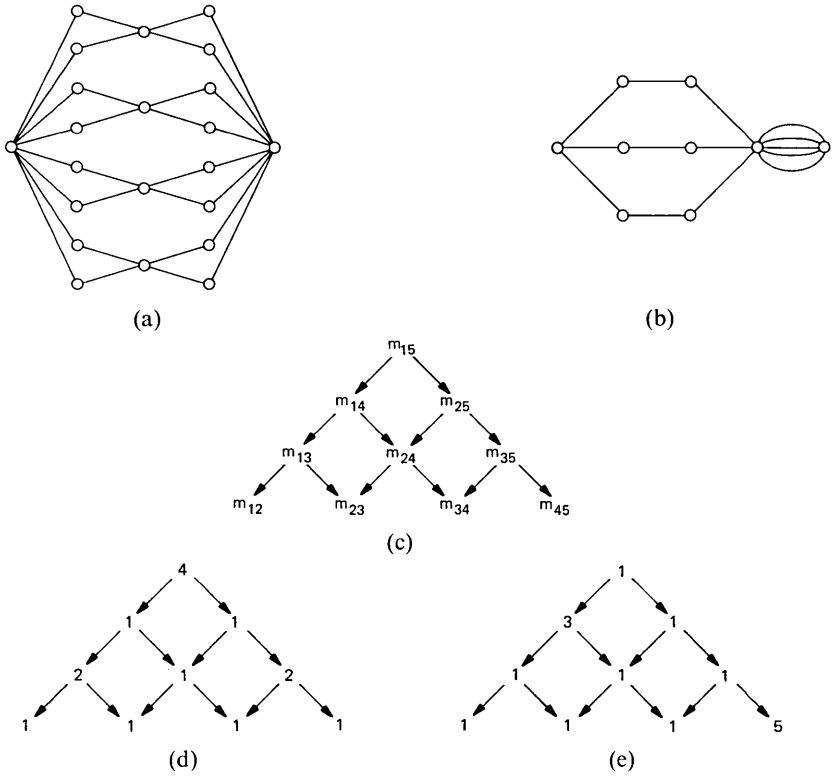


Fig. 3—Graphs for Example 2.

probability of a channel graph with fewer channels exceeds the blocking probability of a channel graph with more channels. Therefore, A cannot be superior to B. ∎

IV. EXAMPLES

Example 1. Figure 2(a) shows the Takagi graph $A = \{m_{12} = m_{23} = 2, m_{13} = 4\}$. Figure 2(b) shows the Takagi graph $B = \{m_{12} = 3, m_{23} = 5\}$. Figure 2(c) shows the partial order P_{13}. Figure 2(d) shows the weights of A on P_{13}. Figure 2(e) shows the weights of B on P_{13}.

It is easily seen that M_{13} has only four filters, $\{m_{13}\}$, $\{m_{12}, m_{13}\}$, $\{m_{23}, m_{13}\}$ and $\{m_{12}, m_{23}, m_{13}\}$, and product of the weights of A is greater than that of B in every case. From Theorem 4, the first graph is superior to the second graph.

Example 2. Figure 3(a) shows the Takagi graph $A = \{m_{15} = 4, m_{13} = m_{35} = 2\}$. Figure 3(b) shows the Takagi graph $B = \{m_{14} = 3, m_{45} = 5\}$. Figure 3(c) shows the partial order P_{15}. Figure 3(d) shows the weights of A on P_{15}. Figure 3(e) shows the weights of B on P_{15}.

Consider the filter $M = \{m_{14}, m_{45}, m_{35}, m_{15}, m_{25}\}$. The product of the weights of A on M is 8 while the product of the weights of B on M is 15. Hence, A is not superior to B. Note that A can still be preferable to B (or B preferable to A) in many other senses. But one does not dominate the other as far as the strong property of superiority is concerned.

V. CONCLUSION

Channel graphs, of which regular series parallel channel graphs form an important subclass, have been extensively used in modeling and analyzing blocking probabilities of switching networks. A popular concept in comparing the blocking characteristics of two channel graphs is to see whether one is superior to the other under arbitrary traffic loads. We give a sufficient condition for superiority in comparing regular series parallel channel graphs.

REFERENCES

1. C. Y. Lee, "Analysis of Switching Networks," B.S.T.J., *34,* No. 6 (November 1955), pp. 1287–1315.
2. F. R. K. Chung and F. K. Hwang, "On Blocking Probabilities for a Class of Linear Graphs," B.S.T.J., *57,* No. 8 (October 1978), pp. 2915–25.
3. H. W. Berkowitz, "A Counterexample to a Conjecture on the Blocking Probabilities of Linear Graphs," B.S.T.J., *58,* No. 5 (May–June 1979), pp. 1107–08.
4. F. K. Hwang, "Superior Channel Graphs," Proc. 9th International Teletraffic Congress, Terremolino, Spain 1979, paper no. 543.
5. F. K. Hwang, "Majorization on a Partially Ordered Set," Proceedings of Amer. Math. Soc., *76,* No. 2 (September 1979), pp. 199–203.
6. F. K. Hwang, "Generalized Schur Functions," Bull. Inst. Math., Academia Sinica, *8,* No. 4 (December 1980), pp. 513–16.
7. A. W. Marshall and I. Olkin, Inequalities, *Theory of Majorization and Its Applications,* New York: Academic Press, 1979.
8. K. Takagi, "Design of Multistage Link Systems with Optimal Channel Graphs," Rev. Elec. Commun. Lab., *17,* No. 10 (October 1969), pp. 1205–26.
9. K. Takagi, "Optimal Channel Graph of Link System and Switching Network Design," Rev. Elec. Commun. Lab., *20,* Nos. 11–12 (November–December 1972), pp. 962–85.
10. X. M. Chang, D. Z. Du and F. K. Hwang, "Characterizations for Series Parallel Channel Graphs," B.S.T.J., *60,* No. 6 (July–August 1981), pp. 887–92.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

On the Physical Limits of Digital Optical Switching and Logic Elements

By P. W. SMITH

(Manuscript received February 5, 1982)

In this paper we identify and discuss some fundamental physical mechanisms that will provide limits on the speed, power dissipation, and size of optical switching elements. Illustrative examples are drawn primarily from the field of bistable optical devices. We compare the limits for optical switching elements with those for other switching technologies, and present a discussion of some potential applications of optical switching devices. Although thermal effects will preclude their wide application in general-purpose computers, the potential speed and bandwidth capability of optical devices, and their capability for parallel processing of information, should lead to a number of significant applications for specific operations in communication and computing fields.

I. INTRODUCTION

A number of recent developments have increased the interest in digital optical signal-processing devices and techniques. Laser technology has now advanced to the point that lasers are being used in consumer electronics. Optical fiber communication systems are being widely installed. Integrated-optics spectrum analyzers have been developed.

In the research stage it has been shown that optical fibers can be used to transmit information at rates approaching 1 THz.[1,2] This rate is much beyond the capabilities of any presently known electronic light detector. Thus, to utilize this information-handling capacity, some form of optical signal processing will have to be performed before the light signals are converted to electronic ones.

Low-power integrated-optics light switches[3] and low-energy integrated-optical bistable devices[4] capable of performing optical logic have been demonstrated. It is tempting to propose that such digital

optical switching elements be used to construct high-speed computers as well as repeaters and terminal equipment for optical communications systems. To examine these possibilities we need to understand: (*i*) What are realistic possibilities for speed, power dissipation, and size for optical switching elements? and (*ii*) What are the fundamental limits imposed by the physics of the nonlinear interactions, and by the available optical materials?

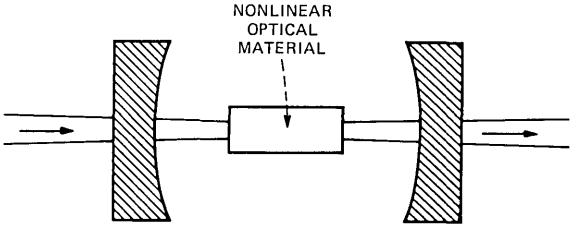
Previous studies of these optical device limits have been made by several authors. The pioneering work of Keyes[5] examined several nonlinear optical processes and concluded that thermal considerations imposed severe limits on the use of optical logic elements. A similar assessment was made by Landauer.[6] A more optimistic conclusion was reached by Fork,[7] who suggested that by using suitable interactions and resonant structures, competitive optical elements could be realized. Recently Kogelnik[8] has examined the role of integrated-optics devices and has concluded that they may well be most useful in performing functions that cannot be provided by other technologies.

In this paper we will attempt to provide a perspective on the ultimate limits of optical switching elements, and the areas in which one might expect optical signal processing to offer a significant advantage over other technologies.

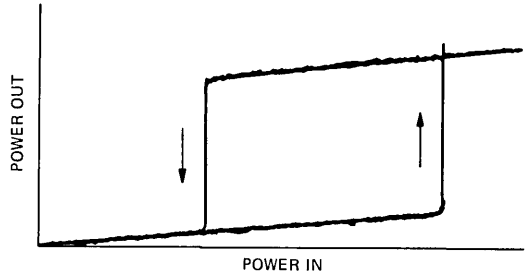
II. BISTABLE OPTICAL DEVICES

In this paper we will draw examples from the field of bistable optical devices. This is to some extent due to a bias of the author, as he has worked on these devices for several years. However, in many respects, bistable optical devices are the most basic binary optical systems, and they have a demonstrated capability for low-energy latching operation. In addition, these devices are extremely versatile and can function as optical limiters, differential amplifiers, and optical logic elements. Their transmission can also be controlled by another optical beam creating an "optical triode."

A generic bistable optical device is shown in Fig. 1. It consists of a Fabry-Perot resonator containing a nonlinear optical material. This nonlinearity can be either a saturable absorption (an absorption that decreases with increasing light intensity), or a nonlinear refractive index (a refractive index that increases or decreases with increasing light intensity). The bistability arises from the simultaneous requirements that the intensity of light inside the resonator, and thus the transmitted light intensity, depends on the resonator tuning (or the loss in the resonator), and the resonator tuning (or the loss in the resonator) depends on the intensity of light in the resonator. The resonator transmission exhibits optical hysteresis, as shown in Fig. 1b.



(a)



(b)

Fig. 1—(a) A generic bistable optical device. (b) Typical output power characteristic for a bistable optical device.

Optical bistability has been demonstrated using both intrinsic devices, which use materials with an intrinsic optical nonlinearity, and hybrid devices, which use a detector and an electrooptic modulator to create an artificial nonlinear medium.

Many types of bistable optical devices have been developed and studied. Small, integrated-optics hybrid devices have been operated with less than a picojoule of optical energy.[9] Two-dimensional arrays of bistable elements using a liquid-crystal light valve have been demonstrated for image-processing applications.[10] Nonresonant devices have been developed that use no Fabry-Perot resonator and thus have a broad frequency response and can switch very rapidly with a suitably fast-responding nonlinear material.[11,12]

III. SPEED AND POWER LIMITATIONS

In this section we will discuss the fundamental physical mechanisms that limit the performance of these devices. Although we will specifically consider bistable optical devices, the results will be generally applicable to any passive digital optical switching elements.

The switching speed of a bistable optical device is limited by the buildup time of the resonator, and by the response time of the nonlinear medium. In principle, the resonator response time can be made

negligible by reducing the length of the resonator, or by using a nonresonant configuration. The ultimate limit is set by the response time of the nonlinearity. Materials exhibiting strong electronic nonlinearities are known with response times of $<10^{-14}$ seconds. To operate a device with such a short response time, however, requires high light powers.

The switching power and switching speed of a bistable optical device are not independent. For example, if the response time of a given device is dominated by the resonator buildup time, the response time can be reduced by a factor of two by halving the length of the resonator. However, as only half the length of nonlinear material is now available, twice the switching power must be used to reach the switching threshold.

R. W. Keyes[5] has discussed several physical processes that limit the switching power and speed of optical devices. For any (nonreversible) switching operation, it can be shown that a minimum energy of the order of kT must be dissipated (k is Boltzman's constant and T is the absolute temperature). Quantum mechanical considerations lead to the assertion that a switching operation must dissipate at least h/τ of energy (h is Planck's constant and τ is the switching time). These limits are shown in Fig. 2, which is a plot of the power required for a switching operation as a function of switching time, i.e., the time for which this power must be applied. The frequency label on the horizontal axis is appropriate if switching is being done repetitively so that a switching time limit implies a limit on the data rate.

Keyes has discussed the limitations imposed by the heat dissipated in a switching element. For continuous operation, this heat sets an upper limit on the achievable switching rate (a higher rate would result in an unacceptable temperature rise in the device). The region affected by such thermal considerations is also shown in Fig. 2, assuming a value of heat transfer coefficient (100 W/cm^2) that is appropriate for liquid-cooled elements and a maximum acceptable temperature rise of 20°C. In many cases a lower temperature rise may be required because of the rapid refractive index change with temperature exhibited by most optical materials. It is important to note that a switching device can operate in this "thermal transfer" region provided that it is operated at less than the maximum repetition rate, or that not all of the switching power is dissipated in the device.

Keyes considered the case of an optical switching device that operates by absorbing light that saturates an atomic transition and changes the optical properties of the material. To achieve appreciable saturation, the condition

$$\sigma I > hc/\lambda\tau_D \tag{1}$$

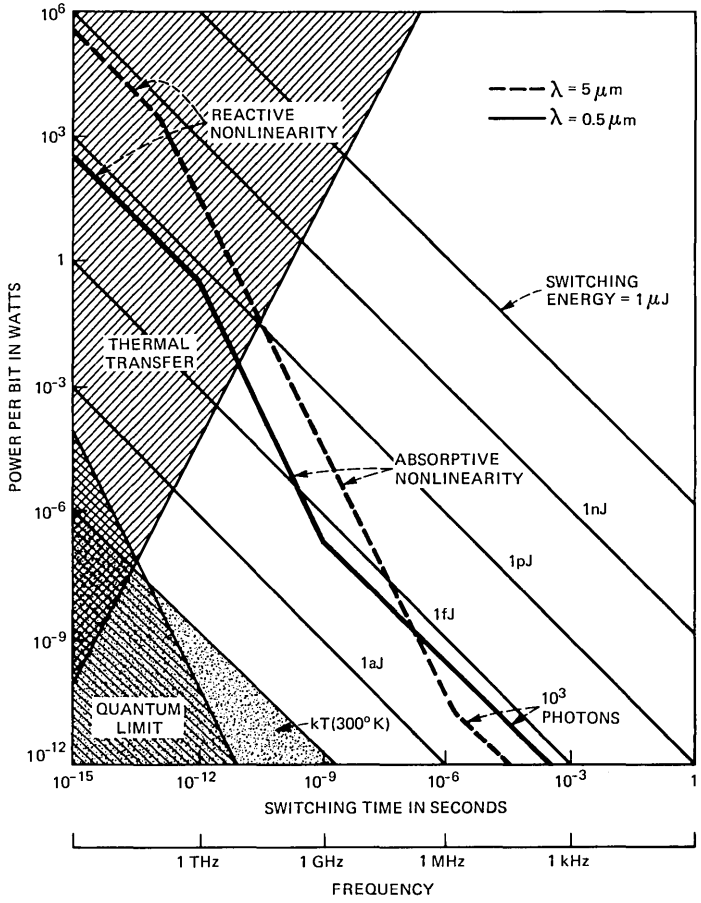


Fig. 2—Limitations on optical switching devices. The frequency scale (at bottom) applies for repetitive switching. The heavy lines indicate limits on optical switching devices for the case of $\lambda = 0.5$ μm (solid lines) and $\lambda = 5$ μm (dashed lines).

must be satisfied. Here, σ is the peak absorption cross section of the transition, I is the light intensity, h is Planck's constant, c is the velocity of light, λ is the wavelength of the light, and τ_D is the decay time of the transition. From the expression for the Einstein stimulated emission coefficient, we can write

$$\sigma = \frac{4\pi^2|\mu|^2 T_2}{3\epsilon_0 h\lambda}, \tag{2}$$

where $|\mu|$ is the dipole moment of the transition, T_2 is the inverse line width of the transition, and ϵ_0 is the vacuum permittivity.

The maximum intensity for a given input power is obtained in a

waveguide geometry for which the input power, P, is given by

$$P \sim I\lambda^2 \tag{3}$$

(see Section IV). Combining eqs. 1, 2 and 3 we obtain:

$$P > \frac{3\epsilon_0 h^2 \lambda^2 c}{4\pi^2 |\mu|^2 T_2 \tau_D}. \tag{4}$$

To evaluate eq. 4 we will assume a large dipole moment for an atom given by

$$|\mu| = ea_0, \tag{5}$$

where e is the electronic charge and a_0 is the Bohr radius. We will further take the response time, τ, of the switching element to be

$$\tau = \tau_D \sim T_2. \tag{6}$$

(T_2 cannot be less than τ_D, and a larger T_2 implies a larger switching energy.) With these assumptions we find

$$P > 1.2 \times 10^{-24} \times \frac{\lambda^2}{\tau^2} W, \tag{7}$$

where λ is in μm and τ is in seconds. This limit is plotted on Fig. 2 for $\lambda = 0.5$ μm and $\lambda = 5$ μm and labeled "Absorptive Nonlinearity." Two points should be noted. First, Keyes shows that a similar limit should also apply for the case of a second-order nonlinear effect [$\chi^{(2)}$ process] or for the case of a nonlinearity based on self-induced transparency. Second, it has recently been shown that it is possible to find systems in which one can use excitonic resonances[13] to obtain effective dipole moments appreciably greater than ea_0. An example using such a system is described in Section VI.

A different limit is found for the case of a third-order nonlinearity [$\chi^{(3)}$ process]. Let us consider a material exhibiting an optical Kerr effect, i.e., the refractive index has a term proportional to the light intensity. To obtain an appreciable effect, we require a change in phase shift through the medium

$$\Delta\phi > \pi. \tag{8}$$

If the refractive index change is $n_2 I$, where n_2 is the optical Kerr coefficient, then eq. 8 becomes

$$\frac{2\pi n_2 I \ell}{\lambda} > \pi, \tag{9}$$

where ℓ is the length of the element. The delay time is governed by the length

$$\tau = n_0 \ell / c, \tag{10}$$

where n_0 is the (linear) refractive index of the material. Combining eqs. 9, 10, and 3, we obtain

$$P > \frac{n_0 \lambda^3}{2 n_2 c \tau}. \tag{11}$$

Note that the dependence of P on τ is different from that in eq. 7.

How can we estimate n_2? We can argue that the nonlinear refractive-index term should be of the order of unity for light fields of the order of the atomic fields. Thus, we could write

$$n_2 I = n_0 \tag{12}$$

for light fields of the order of e/a_0^2. Now

$$I = \frac{\epsilon_0 n_0 c E^2}{2}, \tag{13}$$

where E is the electric field strength of the light. We combine eqs. 12 and 13 to obtain

$$n_2 = \frac{32\pi^2 \epsilon_0 a_0^4}{ce^2}. \tag{14}$$

From eq. 14 we obtain $n_2 = 2.9 \times 10^{-17}[\mathrm{W/cm^2}]^{-1}$. Much larger electronic nonlinearities are known, however. The polydiacetylene PTS has the largest-known value;[12] it is $n_2 \sim 6 \times 10^{-12}[\mathrm{W/cm^2}]^{-1}$. The reason for this large value is that the electrons are relatively unconfined along the chain axis of the PTS molecules. Thus, effective distances much larger than a_0 are encountered, and as $n_2 \propto a^4$, nonlinearities much larger than our estimate are found. The limits represented by eq. 11 are shown in Fig. 2 for $\lambda = 0.5$ μm and $\lambda = 5$ μm and labeled "Reactive Nonlinearity." They are evaluated using the value of n_2 for PTS.

A third limit shown in Fig. 2 is derived from statistical considerations. A number of photons large compared with unity is necessary to define a switching state. We have somewhat arbitrarily taken 10^3 photons as this statistical limit, i.e.,

$$P\tau = 10^3 hc/\lambda. \tag{15}$$

This limit is also plotted in Fig. 2 for $\lambda = 0.5$ μm and $\lambda = 5$ μm.

A word of caution is in order here. These limits that we have identified are "fuzzy" in that it may not be possible to do as well as these limits, or it may be possible to do somewhat better than these limits would indicate. In general, many kT of energy will be required for stable switching devices. On the other hand, the use of a high-finesse optical resonator will lower the required switching energy. These limits are intended to be used as a guide and an indication of the underlying physical mechanisms.

An optical resonator decreases the required switching power at the expense of a reduction in the bandwidth. A bistable optical device utilizing a lossless material with a refractive nonlinearity has a switching power that varies as

$$P \propto 1/F^2, \tag{16}$$

where F is the finesse of the resonator. In the region where the switching time is limited by the resonator,

$$\tau \propto F \tag{17}$$

so that the switching energy, defined as the product of the switching power and switching time, is given by

$$P\tau \propto 1/F. \tag{18}$$

If the switching time is limited by the response time of nonlinear material, τ will not depend on F, and

$$P\tau \propto 1/F^2. \tag{19}$$

The limits shown in Fig. 2 were computed assuming no resonator, i.e., for $F \sim 1$. We see that with high-finesse resonators, switching energies appreciably below the limits shown in Fig. 2 should be possible. The 10^3 photon limit will still apply, however.

IV. SIZE LIMITATIONS

To obtain the largest light intensity for a given input power, it is usually desirable to focus the input light. The light can be focussed to a cross-sectional area of $\sim\lambda^2$, but will diffract rapidly if not confined by some waveguiding structure. For this reason the lowest-power switching devices are likely to be those in which the light is guided in an optical dielectric waveguide with cross-sectional dimensions of $\sim\lambda$. (Kogelnik[8] has pointed out that with a smaller dielectric waveguide, the guided mode will extend beyond the waveguide walls so that the minimum light-beam cross section will always be of the order of λ. Light can be confined to smaller cross sections using metallic waveguides, but in this case large absorption losses will result.)

The minimum length of the waveguide (i.e., the device) will depend on the strength of the optical nonlinearity and on the finesse of the optical resonator. In many cases the linear absorption loss of the nonlinear medium will determine the resonator dimensions. Miller[14] has shown that for a high-finesse waveguide resonator containing a material with a nonlinear refractive index and a linear absorption loss, the lowest switching power will occur for a length of resonator such that

$$1 - R = A, \tag{20}$$

where R is the reflectivity of each of the resonator mirrors, and A is the absorption loss per pass through the nonlinear medium. If we write $A = \alpha\ell$, where ℓ is the length of the medium and α is the absorption coefficient, this condition requires

$$\alpha\ell = \pi/2F \tag{21}$$

or

$$\ell = \pi/2\alpha F. \tag{22}$$

Thus, the length of a device that is optimized for minimum switching power will be determined by the absorption coefficient and the finesse of the resonator. If the nonlinearity is caused by the absorption of light, then compact, fast, and efficient elements will require a large value of α. Under optimized conditions the switching power, P, varies as

$$P \propto 1/F. \tag{23}$$

The optimum length of the device varies with R so that the resonator response time, τ, is independent of F. Thus, in this case we find the switching energy

$$P\tau \propto 1/F. \tag{24}$$

In Table I, we show recent results from the literature on bistable optical devices. We have taken the switching time to be the "recovery time" of the device, i.e., the time constant for the return to equilibrium in the absence of a driving signal. It is important to note that in many cases a device can be switched on (or off) much more rapidly by the application of a short, intense driving pulse. The data in Table I shows a wide range of switching powers and speeds. As might be expected, the fastest devices require the highest switching powers. The lowest

Table I—Experiments: recent results

	Switching Power (watts)	Switching Time (seconds)	Switching Energy (joules)
Bistable Fabry-Perot resonators			
CS_2	3×10^{5}	5×10^{-10}	1.5×10^{-4}
Na vapor	10^{-2}	10^{-5}	10^{-7}
GaAs	2×10^{-1}	4×10^{-8}	8×10^{-9}
InSb	10^{-2}	$<5 \times 10^{-7}$	$<5 \times 10^{-9}$
Hybrid bistable Fabry-Perot resonator			
$LiNbO_3$	10^{-5}	5×10^{-8}	5×10^{-13}
Bistable liquid-crystal matrix			
Hughes liquid crystal light valve	5×10^{-7}	4×10^{-2}	2×10^{-8}
Nonlinear interface			
Glass—CS_2	2×10^{5}	2×10^{-12}	4×10^{-7}

reported switching energy is for a hybrid, integrated-optical bistable device. This low energy is possible because of the very large effective nonlinearity created by the electrooptic modulator with electrical feedback. Although for some applications long switching times or large switching powers are acceptable, in general one wishes to minimize each of these parameters. How much could we improve present bistable devices by shrinking device dimensions to provide shorter transit times and higher light intensities for a given input power?

We have extrapolated current experimental results to λ^2/n^2 cross-section waveguide devices with a length adjusted for minimum switching energy. We have assumed a finesse of 30 for the Fabry-Perot resonator and have considered the response time to be the undriven recovery time of the device. The results are shown in Fig. 3. It is interesting to note that with known devices and materials it should be possible to approach rather closely the fundamental limits for optical

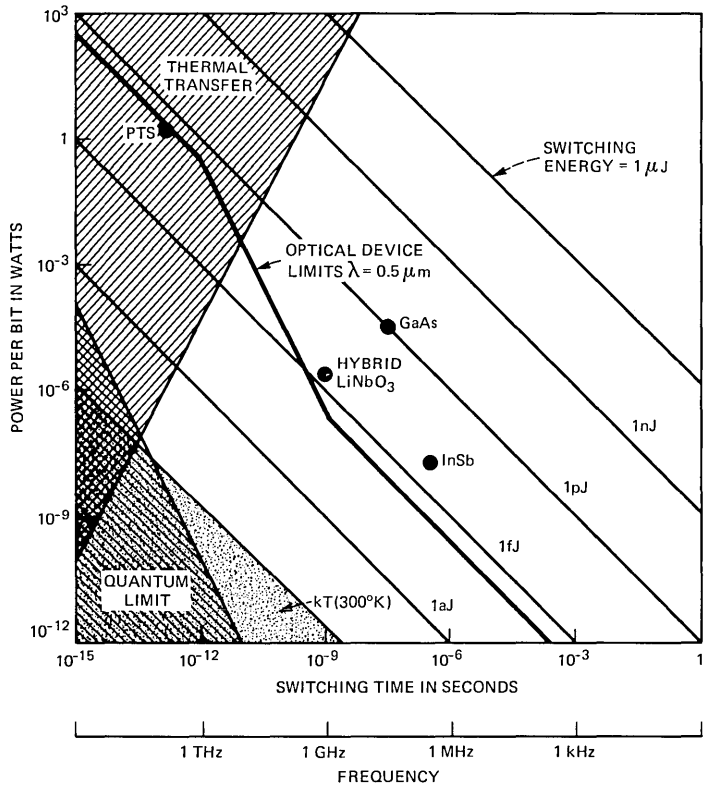


Fig. 3—The points represent performance limits extrapolated from current experimental values for waveguide Fabry-Perot resonators containing the polydiacetylene PTS, the semiconductor GaAs, the semiconductor InSb, and a hybrid device using the electrooptic crystal $LiNbO_3$.

devices; however, as we see in Table I, present laboratory devices are not yet developed to the point where they are close to these limits.

V. COMPARISON WITH OTHER SWITCHING TECHNOLOGIES

How do these projected results and limits that we have derived compare with those for other switching technologies? In Fig. 4 we show how the limits for bistable optical devices compare with the best reported values for two well-established switching technologies—semiconductor electronic devices, and Josephson devices. It is also interesting to see how these devices compare with a biological switching device—a neuron.

It is clear that in the 10^{-6} through 10^{-11} second region, one cannot hope to switch with substantially less power than that required for semiconductor electronic devices, and appreciably lower switching powers are possible with Josephson technology. In the 10^{-12} through 10^{-14} second region, however, optical devices appear to have no com-

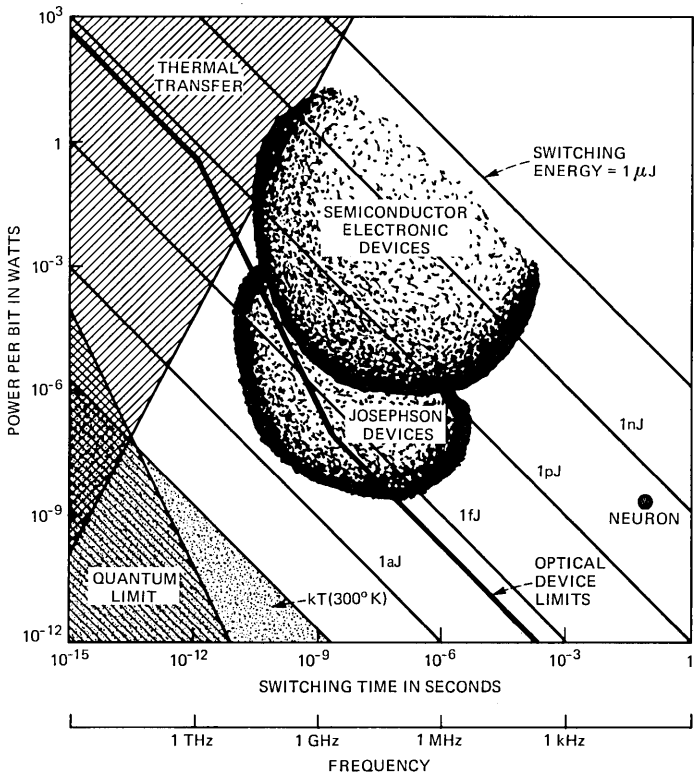


Fig. 4—A comparison with the operating range of two other switching technologies: semiconductor electronic devices, and Josephson devices. The operating point for a biological "switch"—a neuron—is also shown for comparison.

petition. This unique capability for sub-picosecond switching is one of the most exciting aspects of optical switching technology.

The switching power required in this short-time region puts the operating point well within the "thermal transfer" region discussed earlier. For this reason, it does not appear feasible to design a high-speed, general-purpose digital optical computer. However, for many applications these thermal limits may not present severe problems. Two points are worth noting. First, devices using a reactive nonlinearity do not depend on the absorption of the incident light. Thus, most of the switching power is transmitted by the device and the power dissipated is much less than the power required for switching. Second, for some applications, fast switching operations are required at relatively low duty cycles. In both cases the temperature rise in the switching elements will be much lower than the maximum value used in computing the "thermal transfer" region boundary.

There are many other factors that relate to the choice of a switching technology that cannot be shown on a power-time plot. In many cases it is desirable to perform some signal-processing operation on a light signal, either because the incoming signal is in the form of light or because freedom from electromagnetic interference is desired. Optical switching devices typically operate at room temperature. In many cases, they have extremely large bandwidths and can be adapted for many special functions such as rapid parallel processing of information. For these reasons, there will be cases where optical switching systems will be used, even in an area of Fig. 4 in which other technologies show a switching-energy advantage.

VI. ILLUSTRATIVE EXAMPLES

We have shown that because of thermal problems associated with the high packing densities required for rapid operation, optical switching elements are unlikely to be used as building blocks for a general-purpose computer. For certain specific applications, such as the integrated-optical spectrum analyzer recently developed for microwave signal processing, and optical computers for picture processing and pattern recognition,[15-17] special-purpose optical computers have already demonstrated their usefulness. We should also point out that fast optical switching devices are opening up a new time region for scientific studies, and picosecond spectroscopy is rapidly becoming an important field of research.[18]

In this section we will consider a few specific applications of optical switching elements, and see where an extrapolation of current technology might lead. In many cases nonlinear materials with a suitable combination of properties are not currently available. The materials

we have chosen for these examples illustrate the wide range of properties that can be obtained.

6.1 Low-energy optical switch

The lowest "switch-off" energy currently demonstrated is 0.5 pJ for a hybrid bistable device (see Table I). Extrapolation of this figure with a $LiNbO_3$ device with minimum waveguide dimensions and assuming current detector technology, we find 1 fJ operation should be possible. A similar limit is found by extrapolating current figures for InSb devices at 5 μm wavelength. These figures might be reduced still further by using optical resonant structures related to those currently being employed for surface-enhanced Raman studies. (See, for example, Ref. 19.)

6.2 High-speed optical switch

The highest-speed operation will be obtained with a device utilizing a nonlinear material with an electronic nonlinearity. Such nonlinearities are believed to have response times in the range of 10^{-14} seconds. For minimum device response time, a nonresonant configuration should be used. A suitable configuration might be the self-focussing, bistable optical switch described in Ref. 20. By focusing the input to a spot size on the order of the wavelength of the light at the nonlinear material, adequate discrimination between "on" and "off" states could be obtained with a device length of 20 wavelengths. For a device using as nonlinear material the polydiacetylene PTS[12] and light of 1 μm wavelength, the response time would be 0.1 ps and the peak pulse switching power would be 100 W. This power is low enough that it might be reached with a mode-locked semiconductor laser diode.

6.3 4 x 4 optical switching network

If picosecond speed is not required, an optical distribution network could be formed with integrated optics technology on a $LiNbO_3$ substrate, as demonstrated by Schmidt and Buhl.[21] This example is somewhat different from the others we have given, in that electrical signals are used to control the distribution of the optical signals. Kogelnik[8] has addressed the question of the limits for the stepped $\Delta\beta$ couplers that comprise the switching units. He shows that the limiting electrical energy, E_{SW}, needed for one switch of the optical path is given by

$$E_{SW} \cdot \tau = 100 \text{ pJ} \times \text{ps}, \tag{25}$$

where τ is the transit time through the device. A directional coupler switch with a switching time of 110 ps and a transit time of 3 ps has

already been demonstrated.[22] Switching times of 30 ps should be possible with 1 μm electrode gaps.[22] $LiNbO_3$ waveguides suffer severe problems at optical power levels higher than ~100 μW in the visible range. Much higher power levels are possible, however, if near infrared light is used.

6.4 Image amplifier

An optical image amplifier could be made using an array of bistable elements, as shown in Fig. 5a. Each element in the array is a self-focussing bistable element similar to that described in Ref. 20. A typical output characteristic for each element is shown in Fig. 5b. It can be seen that a weak input signal (I_S) produces a strong modulated signal at the output when the input light level corresponds to the

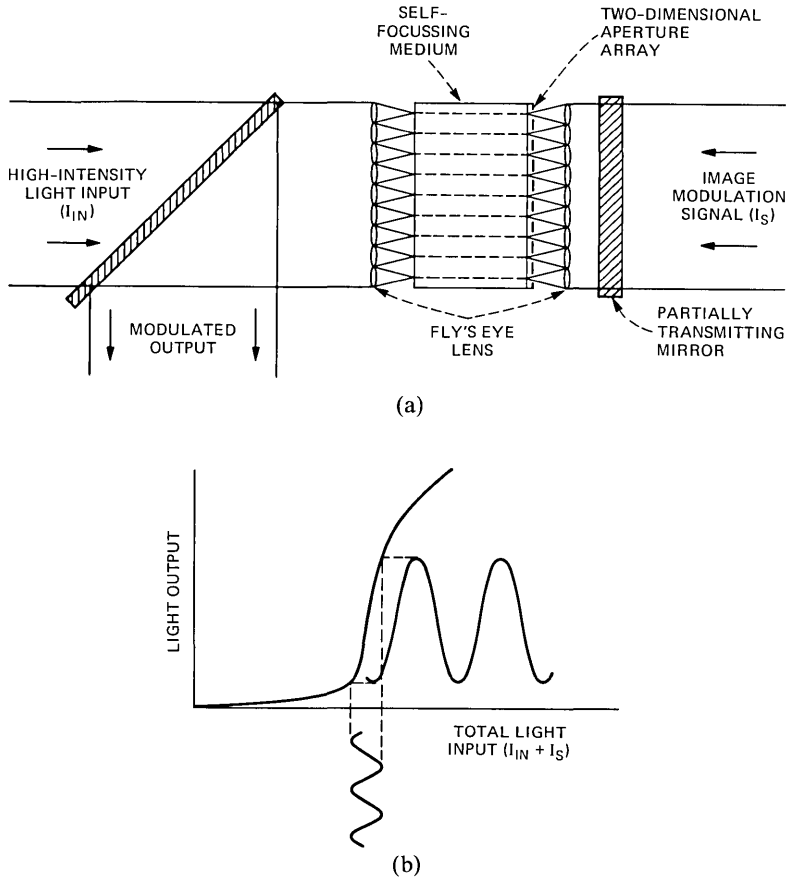


Fig. 5—An optical image amplifier. (a) Schematic diagram. (b) Operating characteristic of each element.

"knee" in the characteristic curve. Thus, each spatial element behaves as an "optical triode."

One possible nonlinear medium for this application would be a liquid suspension of sub-micron dielectric particles. As shown in Ref. 23, this medium exhibits a large nonlinear coefficient, although it has a slow response time on the order of a second. Such times may be acceptable for certain image-processing operations. If each element of the input image is focussed by the composite lens to a spot size of $\sim\lambda^2$, with an aqueous suspension of quartz particles one would require an input power of 10 mW/resolution element at $\lambda = 0.5\ \mu$m and 1 mW/resolution element at $\lambda = 5\ \mu$m. Recent calculations[24] indicate that input powers of $\sim 100\ \mu$W/resolution element and a response time of $\sim 1\ \mu$sec might be possible using suitably doped GaAs as the nonlinear medium. The verification of these ideas must await further experiments.

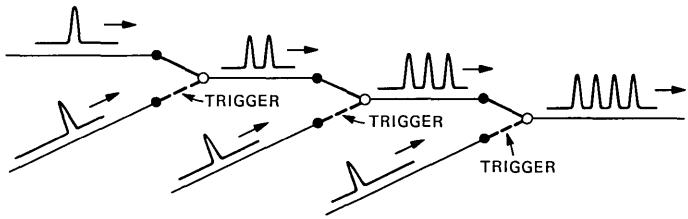
6.5 Optical time-division multiplexer and demultiplexer

As our final example, let us consider a high-speed optical time-division multiplexer (or demultiplexer) that might be used to multiplex picosecond optical pulses into a high-capacity optical fiber, and demultiplex the signals at the receiver to obtain low enough bit rates to allow handling by optical detectors and subsequent electronic systems.

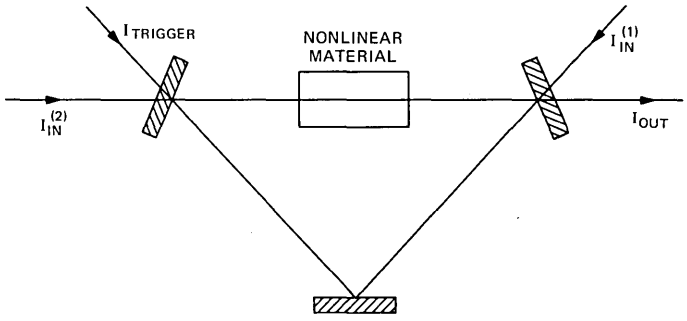
A multiplexer can be made from a number of triggerable switching elements, as shown in Fig. 6a. A trigger pulse with the proper time synchronization is required to multiplex pulses as shown. Each element could be made from a properly designed bistable optical device, as shown in Fig. 6b. This bistable device consists of a suitable nonlinear optical material in a ring resonator. The ring geometry allows separation of the inputs and outputs. However, some polarization selectivity may have to be employed to avoid interference effects between the two input beams $I_{IN}^{(1)}$ and $I_{IN}^{(2)}$. Let us assume here that pulses in these two beams are never present simultaneously in the element. The output intensity depends on the total input intensity $I_{IN}^{(1)} + I_{IN}^{(2)} + I_{TRIG}$, as shown in Fig. 6c. If the input pulses are of intensity slightly less than the critical intensity corresponding to the "knee" of the curves, the output in the absence of a trigger input will consist solely of $I_{IN}^{(1)}$. However, during the time that a small trigger signal I_{TRIG} is present, the output will consist solely of $I_{IN}^{(2)}$. Because of the sharp "knee" in the characteristic curves, only a small I_{TRIG} is required to accomplish this switching.

In a similar way, one can perform demultiplexing by using an optical "tap," as shown in Fig. 7a. A similar bistable ring resonator serves as a triggerable "tap," as shown in Fig. 7b; the output characteristics are shown in Fig. 7c.

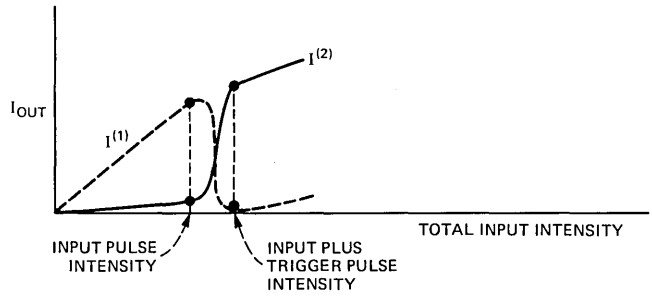
An appropriate nonlinear material for use in these devices might be



(a)



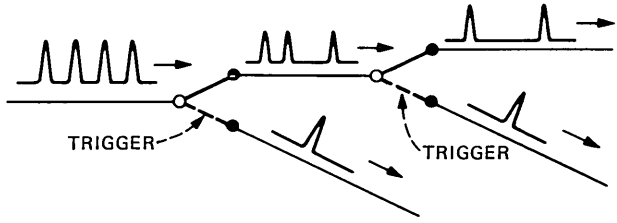
(b)



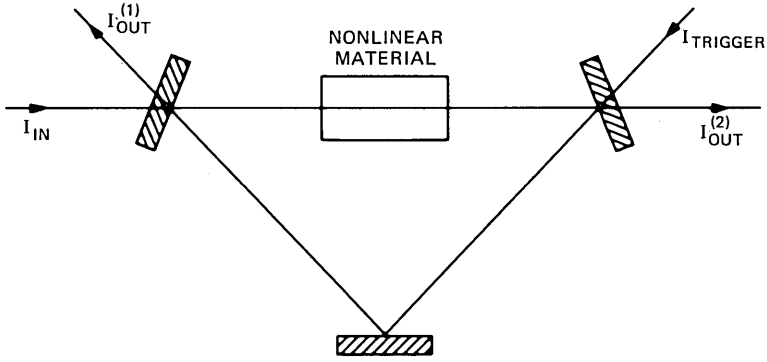
(c)

Fig. 6—Optical time-division multiplexer. (a) Overall schematic diagram. (b) Ring triggerable bistable element. (c) Output characteristic of ring bistable element.

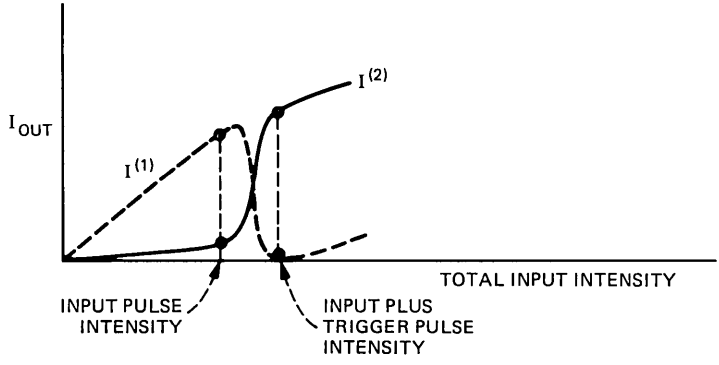
the semiconductor CdS. It has recently been shown[13] that large nonlinear effects are found near the biexitonic resonance line. By using light at $\lambda \sim 4836$Å, we will obtain a sufficiently large nonlinearity to allow the device to operate at a pulse power level of about 1 mW with a time response of ~1 ps. This material requires cooling to liquid helium temperatures, however, and the wavelength range over which this operation can be obtained is small (~5Å). A material such as PTS will avoid these difficulties, but will increase the peak power required



(a)



(b)



(c)

Fig. 7—Optical time-division demultiplexer. (a) Overall schematic diagram. (b) Ring triggerable bistable element. (c) Output characteristic of ring bistable element.

to the 1 W level. As this power is only required for 1 ps, however, the operating energy would still be a very reasonable 1 pJ.

VII. CONCLUSIONS

Having identified the physical limits for optical switching devices and discussed some specific examples, let us try to draw some general conclusions with regard to their future applications. The strong points of optical switching devices are:

(*i*) Speed: With an electronic nonlinearity or free-carrier generation in semiconductors, sub-picosecond switching times are possible.

(*ii*) Bandwidth: With a nonresonant bistable optical device or a nonlinear interface, a large fraction of the visible light bandwidth can be used.

(*iii*) Ability to treat directly signals already in the form of light.

(*iv*) Capability for parallel processing: With a liquid crystal bistable array, image processing has already been demonstrated.[15]

The weak points of optical switching devices are:

(*i*) High power is required for fast switching. This will tend to create thermal problems unless highly transparent materials are used.

(*ii*) Materials do not yet exist that have the ideal combination of properties for these devices.

(*iii*) Theoretical and practical problems involved in waveguide and microresonator formation in λ^3 volumes have yet to be overcome.

(*iv*) The minimum size of an optical switching element cannot be reduced below a volume of about λ^3 unless lossy metallic structures are used.

The field of digital optical switching is a dynamic one in which rapid progress is being made. New materials and devices are being proposed and studied. Let us close by proposing some areas where future work should be directed.

There is clearly a need for good optical quality, high-nonlinearity materials with low absorption coefficients. The development of such materials will have a great impact on future applications. Techniques must be developed for fabricating optical waveguides and optical resonant structures with dimensions on the order of optical wavelengths. For many applications, the problem of regeneration of signals associated with 'fan out' is important.[7] Some types of bistable lasers may well find use here. Finally, the general problem of the optics—electronics interface (i.e., making optics compatible with electronics)—is one that will require a good deal of attention if the most effective use is to be made of the tremendous potential of digital optical switching.

REFERENCES

1. L. F. Mollenauer, R. H. Stolen, and J. P. Gordon, "Experimental Observation of Picosecond Pulse Narrowing and Solitons in Optical Fibers," Phys. Rev. Lett., *45* (September 1980), pp. 1095–7.
2. A. Hasegawa and Y. Kodama, unpublished work.
3. R. C. Alferness, "Guided Wave Devices for Optical Communication," IEEE J. Quantum Elec., *QE17* (June 1981), pp. 946–58.
4. P. W. Smith and W. J. Tomlinson, "Bistable Optical Devices Promise Subpicosecond Switching," IEEE SPECTRUM, *18* (June 1981), pp. 26–33.
5. R. W. Keyes, "Power Dissipation in Information Processing," Science *168* (May 1970), pp. 796–801, and R. W. Keyes, "Physical Limits in Digital Electronics," Proc. IEEE, *63* (May 1975), pp. 740–67.
6. R. Landauer, "Optical Logic and Optically Accessed Digital Storage," *Optical Information Processing*, Nesterikhin, Stroke, and Kock, Eds., New York: Plenum, 1976, pp 219–54.
7. R. L. Fork, unpublished work.
8. H. Kogelnik, "Limits in Integrated Optics," Proc. IEEE, *69* (February 1981), pp. 232–8.
9. P. W. Smith, I. P. Kaminow, P. J. Maloney, and L. W. Stulz, "Self-Contained Integrated Bistable Optical Devices," Appl. Phys. Lett., *34* (January 1979), pp. 62–4.
10. U. H. Gerlach, U. K. Sengupta, and S. A. Collins, Jr., "Single-Spatial Light Modulator Bistable Optical Matrix Device," Opt. Eng., *19* (July/August 1980), pp. 452–5.
11. E. Garmire, J. H. Marburger, and S. D. Allen, "Incoherent Mirrorless Bistable Optical Devices," Appl. Phys. Lett., *32* (March 1978), pp. 302–22; P. W. Smith, W. J. Tomlinson, P. J. Maloney, and J.-P. Hermann, "Experimental Studies of a Nonlinear Interface," IEEE J. Quantum Elec., *QE-17* (March 1978), pp. 340–8; and J. E. Bjorkholm, P. W. Smith, W. J. Tomlinson, and A. E. Kaplan, "Optical Bistability Based on Self-Focusing," Opt. Lett., *6* (July 1981), pp. 345–7.
12. J.-P. Hermann and P. W. Smith, "Nonlinear Fabry-Perot Containing the Polydiacetylene PTS," Proc. of the XI Int. Quantum Elec. Conf., Paper T6, (June 1980), pp. 656–7.
13. A. Maruani and D. S. Chemla, "Active Nonlinear Spectroscopy of Biexcitons in Semiconductors: Propagation Effects and Fano Interferences," Phys. Rev. B, *23* (January 1981), pp. 841–60.
14. D. A. B. Miller, "Refractive Fabry-Perot Bistability with Linear Absorption: Theory of Operation and Cavity Optimization," IEEE J. Quantum Elec., *QE-17* (March 1981), pp. 306–11.
15. M. T. Fatehi, S. A. Collins Jr., and K. C. Wasmundt, "The Optical Computer Goes Digital," Optical Spectra, *15*, No. 1 (January 1981), pp. 39–44.
16. N. G. Basov et al., "Methods of Realization of an Optical Processor with Variable Operators," Sov. J. Quantum Elec, *8* (March 1978), pp. 307–12.
17. A. Huang, Y. Tsunoda, J. W. Goodman, and S. Ishihara, "Optical Computation Using Residue Arithmetic," Appl. Opt., *18* (January 1979), pp. 149–62.
18. *Picosecond Phenomena II*, R. M. Hochstrasser, W. Kaiser, and C. V. Shank, Eds., Berlin: Springer Verlag, 1980.
19. R. L. Fork and J. P. Gordon, unpublished work.
20. J. E. Bjorkholm, P. W. Smith, W. J. Tomlinson, and A. E. Kaplan, "Optical Bistability Based on Self-Focusing," Optics Lett., *6* (July 1981), pp. 345–7.
21. R. V. Schmidt and L. L. Buhl, "Experimental 4 x 4 Optical Switching Network," Elec. Lett., *12* (October 1976), pp. 575–7.
22. R. C. Alferness, N. P. Economou, and L. L. Buhl, "Fast Compact Optical Waveguide Switch Modulator," Appl. Phys. Lett., *38* (February 1981), pp. 214–7.
23. P. W. Smith, A. Ashkin, and W. J. Tomlinson, "Four-Wave Mixing in an Artificial Kerr Medium," Optics Lett., *6* (June 1981), pp. 284–6.
24. D. A. B. Miller, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

An Inversion Technique for the Laplace Transform

By D. L. JAGERMAN

(Manuscript received February 8, 1982)

In this paper we use an approximation sequence defined by the Widder Laplace transform inversion formula to provide a practical method for inverting the Laplace transform. The approximation sequence converges uniformly and retains essential structural characteristics of the original function, e.g., nonnegativity, monotonicity, and convexity. Thus, we approximate a distribution function by distribution functions and use enhancement techniques to increase the speed of convergence and to capture the quality of exponential decay. Also, we present a practical computational method illustrated by examples.

I. INTRODUCTION

The purpose of this paper is to summarize the techniques presented in the paper "An Inversion Technique for the Laplace Transform"[1] and to make available a useful reference of properties of the 'approximation sequence,' and a new numerical method developed since the publication of Ref. 1. The inversion, or approximation sequence, retains the essential structural characteristics of the original function, e.g., nonnegativity, monotonicity, and convexity. Thus, we approximate a distribution function by distribution functions. For application to queueing theory, this may be considered quite important. The basic inversion sequence, together with error estimates, is discussed in Section II; also, two enhancement procedures are given—namely, the construction of a sequence that is more rapidly convergent than the approximation sequence and which was not given in Ref. 1, and a method of accurate approximation to functions that decay exponentially. Section III discusses the new numerical method, and Section IV presents two examples of numerical inversion along with controls. Except for the new material whose derivations are given here, all proofs can be found in Ref. 1.

II. INVERSION SEQUENCE

Consider a function $f(t)$ for which the Laplace transform, $\tilde{f}(s)$, defined by

$$\tilde{f}(s) = \int_0^\infty e^{-st} f(t)\,dt \tag{1}$$

exists for $s > c$ and $f(t) = 0(e^{ct})(t \to \infty)$; then a sequence of functionals, $(L_n)_0^\infty$, is defined by

$$L_n f(t) = f_n(t) = \frac{(-1)^n}{n!} s^{n+1} \frac{d^n \tilde{f}(s)}{ds^n}\bigg|_{s=\frac{n+1}{t}}. \tag{2}$$

One has

$$\lim_{n\to\infty} f_n(t) = f(t) \tag{3}$$

uniformly in every finite closed interval throughout which $f(t)$ is continuous. Equations (2) and (3) constitute a variant of Widder's theorem.[2] The sequence $[f_n(t)]_0^\infty$ is called the "approximation sequence" of $f(t)$.

The following lists some important properties of $f_n(t)$ valid for $t \geq 0$, $n \geq 0$; a dot indicates differentiation. Assumptions on $f(t)$ are to hold in $(0, \infty)$.

(*i*) $a \leq f(t) \leq b \Leftrightarrow a \leq f_n(t) \leq b$,

$$f_n(0+) = f(0+),$$
$$\dot{f}_n(0+) = \dot{f}(0+),$$
$$f_n(\infty) = f(\infty),$$
$$\dot{f}_n(\infty) = \dot{f}(\infty),$$

(*ii*) $f(t)$ monotone $\Rightarrow f_n(t)$ monotone in the same sense. $f(t)$ completely monotone, absolutely monotone, convex, log-convex implies the same property, respectively, for $f_n(t)$.

(*iii*) $f(t) \geq 0 \Rightarrow \dfrac{f_n(t)}{n+1}$ monotone decreasing in n, $f(t)$ convex $\Leftrightarrow$ $f(t) \leq f_n(t), \ddot{f}(t) \geq 0 \Rightarrow f_n(t)$ monotone decreasing in n.

(*iv*) If $f(t)*g(t) = \displaystyle\int_0^\infty f\left(\frac{t}{u}\right) g(u)\,\frac{du}{u}$ (Mellin convolution), then $f_n(t)$ is the approximation sequence of $f(t) \Rightarrow f_n(t)*g(t)$ is the approximation sequence of $f(t)*g(t)$.

The pointwise error, $\epsilon_n(t; f)$, is defined by

$$\epsilon_n(t; f) = f_n(t) - f(t). \tag{4}$$

One has

$$\epsilon_n(t; f) \sim \sum_{l=1}^{\infty} \frac{1}{l!} G_l(-n-1, -n-1) t^l f^{(l)}(t), \qquad n \to \infty, \tag{5}$$

in which $G_l(-n-1, -n-1)$ are Poisson-Charlier polynomials.[3] Set $\beta_l = G_l(-n-1, -n-1)$; then the following recursion is satisfied:

$$\beta_{l+1} = \frac{l}{n+1}[\beta_l + \beta_{l-1}], \qquad \beta_0 = 1, \qquad \beta_1 = 0. \tag{6}$$

The terms of eq. (5) through order n^{-2} are

$$\epsilon_n(t; f) \sim \frac{1}{2}\frac{1}{n+1} t^2 f^{(2)}(t) + \frac{1}{(n+1)^2}\left[\frac{1}{3} t^3 f^{(3)}(t) + \frac{1}{8}\frac{n+3}{n+1} t^4 f^{(4)}(t)\right], \qquad n \to \infty. \tag{7}$$

Some bounds for $\epsilon_n(t; f)$ are

$$|\epsilon_n(t; f)| \leq \frac{t^2}{2n+2} \sup_{x>0} |\ddot{f}(x)|, \tag{8}$$

for pointwise error, and

$$|\epsilon_n(t; f)| \leq \alpha_n \sup_{t>0} |t\dot{f}(t)| \tag{9}$$

uniform over all t. The α_n satisfies

$$\alpha_n \sim \sqrt{\frac{2}{\pi n}}, \qquad n \to \infty. \tag{10}$$

It appears, therefore, that the convergence of $f_n(t)$ to $f(t)$ is not rapid. In fact, from eq. (7) it is $0(1/n)$ and is not improved by assuming a higher degree of smoothness for $f(t)$. However, one may decrease the error, $\epsilon_n(t; f)$, for a given n in at least two ways by: (i) modifying the sequence $f_n(t)$ to obtain a higher convergence rate, and (ii) modifying $f_0(t)$ to improve its approximation to $f(t)$. The first way may destroy the desirable properties of the approximation sequence as previously listed because the transformation from $f(t)$ to the members of the new sequence may no longer be positive; the second way can be implemented to preserve all the desirable properties with respect to a function different from the original $f(t)$.

The following procedure improves the convergence rate but does not preserve positivity. Define the functionals, σ_n, by

$$\sigma_n f = \sigma_n(t) = \left(2 + \frac{1}{n}\right) f_{2n}(t) - \left(1 + \frac{1}{n}\right) f_n(t); \tag{11}$$

then, from eq. (7)

$$\sigma_n(t) - f(t) \sim -\frac{1}{3}\frac{1}{(n+1)(2n+1)} t^3 f^{(3)}(t) - \frac{1}{8}\frac{2n^2+9n+5}{(n+1)^2(2n+1)^2} t^4 f^{(4)}(t), \qquad n \to \infty, \quad (12)$$

and hence

$$\sigma_n(t) - f(t) = 0(1/n^2). \quad (13)$$

Often, $f(t)$ decreases exponentially fast for $t \to \infty$; the inversion method may not capture this effect well for large t if the decrease is very rapid. This problem is largely overcome by the following technique, which constitutes an example of the second method of improving the approximation.

If $\tilde{f}(s)$ converges for $s \geq -\beta(\beta > 0)$, and $0 \leq \alpha \leq \beta$, then $\tilde{f}(s-\alpha)$ exists and is the transform of $g(t) = e^{\alpha t} f(t)$. If $g_n(t)$ is the approximation sequence for $e^{\alpha t} f(t)$ and

$$f_{n,\alpha}(t) = e^{-\alpha t} g_n(t), \quad (14)$$

then

$$\lim_{n\to\infty} f_{n,\alpha}(t) = f(t). \quad (15)$$

The error $\epsilon_{n,\alpha}(t; f) = f_{n,\alpha}(t) - f(t)$ can be much smaller than $\epsilon_n(t; f)$, corresponding to $\alpha = 0$, so that a real improvement in accuracy can result; of course,

$$\epsilon_{n,\alpha}(t; f) = e^{-\alpha t} \epsilon_n(t; g). \quad (16)$$

The approximation $f_{n,\alpha}(t)$ imitates the rapid exponential decrease of $f(t)$ with increasing accuracy as α approaches β. Since $g_n(t)$ is an approximation sequence, it possesses all of the properties given earlier, many of which carry over to $f(t)$.

III. NUMERICAL EVALUATION

A type of generating function may be constructed for $[f_n(t)]_0^\infty$. Let

$$G(z, t) = \frac{1}{t}\tilde{f}\left(\frac{1-z}{t}\right) \quad (17)$$

and define $[a_n(t)]_0^\infty$ by

$$G(z, t) = \sum_{n=0}^{\infty} a_n(t) z^n; \quad (18)$$

then

$$f_n(t) = a_n\left(\frac{t}{n+1}\right). \quad (19)$$

This permits the use of numerical procedures for obtaining coefficients in a power series. The following is such a method. The notation

$$e_q(x) = e^{i\frac{2\pi x}{q}} \tag{20}$$

will be used. Let $0 < r < 1$, q prime and $q > n$, and define the functional S by

$$Sf = \frac{n+1}{tqr^n} \sum_{j=1}^{q} e_q(-nj)\tilde{f}\left\{\frac{n+1}{t}[1 - re_q(j)]\right\}, \tag{21}$$

then, in terms of $f_n(t)$, one has

$$Sf = f_n(t) + \sum_{l=1}^{\infty} f_{n+lq}\left(\frac{n+lq+1}{n+1}t\right) r^{lq}. \tag{22}$$

For the purpose of computing $f_n(t)$, Sf in the form of eq. (21) will be used.

We will now assume that $f(t)$ is bounded so one may take $|f(t)| \leq A$. Clearly, from eq. (22) and property 1,

$$|Sf - f_n(t)| \leq A\frac{r^q}{1-r^q} \tag{23}$$

and

$$\lim_{r\to 0+} Sf = f_n(t). \tag{24}$$

Thus, Sf is an accurate approximation to $f_n(t)$ when r is small; however, if the round-off error of each term of eq. (21) is bounded by ϵ, and the total round-off error by η, then

$$\eta \leq \epsilon\frac{n+1}{t}r^{-n}. \tag{25}$$

It follows that a choice of r may be made to ensure the round-off error, η, is not too large, in accordance with eq. (25). The parameter q may now be chosen to render the truncation error given by eq. (23) comparably small.

IV. EXAMPLES

A Fortran program for the evaluation of Sf, as shown in Eq. (21), was written by Bharat Doshi. Double precision arithmetic was used with a round-off error of 10^{-15}. As a rough estimate, it was assumed that the computations required to form each term of the summation in eq. (21) resulted in a round-off error of 10^{-11}; thus the choice $\epsilon = 10^{-11}$ was made. The choices

$$r = 0.83, \quad q = 127; \quad n = 50,$$
$$r = 0.91, \quad q = 251; \quad n = 100 \tag{26}$$

lead to about five units error in the sixth decimal; the values of q given result in negligible truncation errors. These errors in computing f_{50} and f_{100} were considered acceptably small.

The function

$$f(t) = 1/2e^{-t} + 1/2e^{-3t} \tag{27}$$

is chosen for the first example. It represents a complementary distribution function whose values may be accurately computed as a control, and which shows the enhancement of accuracy obtainable by use of eqs. (11) and (14). Tables I and II illustrate these results. Since

$$\tilde{f}(s) = \frac{1}{2}\frac{1}{s+1} + \frac{1}{2}\frac{1}{s+3}, \tag{28}$$

the choice $\alpha = 1$ is the appropriate value to use.

It may be observed, from Table I, that the error is halved in going from f_{50} to f_{100} as expected from the $0(1/n)$ behavior of eq. (7), while the decrease of error from f_{100} to σ_{50} is better than 1/50 as seen from the $0(1/n^2)$ rate of eq. (13). Table II shows that the effect of α is greater for larger t since the exponential term that is being tracked becomes dominant.

The second example concerns the complementary busy-period distribution for an $M/M/1$.[4] The arrival rate is ρ and the service rate is $\mu = 1$. For this case

$$f(t) = 1 - \rho^{-1/2} \int_0^t e^{-(1+\rho)x} I_1(2x\sqrt{\rho}) \frac{dx}{x}, \tag{29}$$

and

$$\tilde{f}(s) = \frac{1}{2\rho s}[\rho - 1 - s + \sqrt{(\rho + 1 + s)^2 - 4\rho}], \tag{30}$$

Table I—$\alpha = 0$

t	f	f_{50}	f_{100}	σ_{50}
2	0.06891	0.07203	0.07048	0.06890
4	0.00916	0.01064	0.00990	0.00915
8	0.00017	0.00030	0.00023	0.00016

Table II—$\alpha = 1$

t	f	$f_{50,1}$	$f_{100,1}$	$\sigma_{50,1}$
2	0.06891	0.06911	0.06901	0.06891
4	0.00916	0.00916	0.00916	0.00916
8	0.00017	0.00017	0.00017	0.00017

Table III—$\rho = 0.9$, $\alpha = 0$

t	f	f_{50}	f_{100}	σ_{50}
3	0.2919	0.2942	0.2931	0.2919
6	0.1950	0.1967	0.1959	0.1950
9	0.1514	0.1528	0.1521	0.1514

Table IV—$\rho = 0.5$, $\alpha = 0$

t	f	f_{50}	f_{100}	σ_{50}
3	0.1803	0.1834	0.1819	0.1803
6	0.0764	0.0785	0.0775	0.0764
9	0.0399	0.0415	0.0407	0.0399

Table V—$\rho = 0.1$, $\alpha = 0$

t	f	f_{50}	f_{100}	σ_{50}
3	0.0732	0.0774	0.0753	0.0732
6	0.0088	0.0103	0.0096	0.0088
9	0.0013	0.0018	0.0016	0.0013

Table VI—$\rho = 0.1$, $\alpha = 0.4675$

t	f	$f_{50,\alpha}$	$f_{100,\alpha}$	$\sigma_{50,\alpha}$
3	0.0732	0.0742	0.0737	0.0732
6	0.0088	0.0090	0.0089	0.0088
9	0.0013	0.0014	0.0014	0.0013

in which $I_1(2x\sqrt{\rho})$ is a modified Bessel function. The evaluation of the control was accomplished by using a quadrature procedure. The values $\rho = 0.9, 0.5$, and 0.1 were used. The best choice of α is, from Section II, $(1 - \sqrt{\rho})^2$; however, when α is small, one could simply set $\alpha = 0$, since there is little improvement in using the best α. For the case $\rho = 0.1$, one has $\alpha = 0.4675$; accordingly, a comparison was made with $\alpha = 0$. Tables III through VI sample the results obtained.

V. CONCLUSION

It may be observed that the best α for $\rho = 0.1$ noticeably improved the approximation. The performance of the σ_n functionals in all cases created a marked improvement in accuracy; however, one should remember that the desirable properties of the L_n functionals are not possessed by the σ_n. Nonetheless, there is a property of the σ_n, important in numerical work, which follows from eq. (11). Namely, if δ is the round-off error in the computation of $f_n(t)$ and $f_{2n}(t)$, then $(3 + 2/n)\delta$, $n \geq 1$ bounds the induced round-off error in σ_n (i.e., round-off errors are not significantly magnified).

An interesting use of the inversion technique follows. A function $f(t)$ is given in the form

$$f(t) = \int_0^t g(t, x)dx, \tag{31}$$

with $g(t, x)$ known; it is required to evaluate $f(t)$ over a wide range of values of t. Quadrature methods are accurate, for t small, or can be designed, for t large, but do not apply equally well over all values of t, including the important transition region from small to large. If the transform $\tilde{f}(s)$ can be obtained, then the inversion method described here can be used to obtain sufficiently accurate values of $f(t)$ over the entire range of t, since eq. (9) shows that the errors, $\epsilon_n(t; f)$, are uniformly bounded.

REFERENCES

1. D. L. Jagerman, "An Inversion Technique for the Laplace Transform with Application to Approximation," B.S.T.J., *57,* No. 3 (March 1978), pp. 669–710.
2. I. I. Hirschman and D. V. Widder, *The Convolution Transform*, Princeton, NJ: Princeton University Press, 1955.
3. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53,* No. 3 (March 1974), pp. 525–51.
4. R. B. Cooper, *Introduction to Queueing Theory*, New York: Macmillan Co., 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Approximate Mean Waiting Times in Transient *GI/G/1* Queues

By D. L. JAGERMAN

(Manuscript received February 8, 1982)

In this paper we give an approximation method for obtaining the probability the server is busy and the mean waiting time as seen by the nth arriving customer for the GI/G/1 queueing system. Transient behavior is the key issue of the method. The approximation consists of a pair of recursion formulae whose state variables are the probability of delay and the mean waiting time. Any initial state may be prescribed for the 0th arriving customer. Programming is very easy, and the computation is rapid. The procedure is useful for rush-hour analyses and for studying the recovery of a system from temporary overload.

I. INTRODUCTION

This paper presents an approximation method for obtaining the probability the server is busy and the mean waiting time as seen by the nth arriving customer for the $GI/G/1$ queueing system. Thus, transient behavior is the key issue of the method. The approximation consists of a simultaneous pair of recursion formulae whose state variables are the probability of delay and the mean waiting time. It is assumed that the 0th arriving customer finds the queue in some prescribed state from which the successive states are computed. Naturally, for the nth arrival when n is large, the computations provide approximations to the corresponding eqilibrium quantities when equilibrium exists. The procedure, however, is not limited to queues possessing an equilibrium state. For methods specially adapted to approximating the equilibrium quantities, we refer to a paper by A. A. Fredericks,[1] in which the approximation $A_{0,1}$ of that paper essentially corresponds to the equilibrium results obtained here; we also refer to Fredericks for an application to computer systems.[2]

Transient analysis is of particular importance in studying the recovery of a system from temporary overload. This can occur after a short

downtime during which the buffers become full and the machine must now work off the accumulated load. Another situation calling for transient analysis occurs in the case of minicomputers receiving hourly data from an electronic switching system (ESS). The hourly rush of traffic requires a transient analysis to determine the buildup and falloff of delays. This corresponds to the general problem of rush-hour situations. Knowledge of the transient behavior of a queue also indicates the number of customers needed in the arrival stream until approximate equilibrium conditions are recovered from some temporary overload (i.e., the relaxation time may be estimated).

In Section II we obtain the general recursion formulae and the corresponding formulae defining the equilibrium results. In Section III we specialize the recursions to the $GI/M/1$ queue and present the exact solutions obtained using the Takaćs[3] method as a control. We show the relationship between the exact probability of delay and mean waiting time, and also conclude that the approximate equilibrium quantities are, in fact, exact. We discuss a number of numerical examples and compare them with exact results. In Section IV we reduce the general recursions to the $M/G/1$ case. A relation is again obtained between the probability of delay and the mean waiting time. The interesting property is proved that the approximate probability of delay and mean waiting time as seen by the nth arrival satisfies the same relation as the exact quantities; thus, the approximation method is shown to preserve certain properties of the exact solution. We discuss a number of numerical examples and compare them with exact results. In Section V we present several numerical examples of $GI/G/1$ queues. Exact solutions, however, are derived only for equilibrium values.

In all the numerical examples given, the queues start empty; however, this is done only to facilitate obtaining the controls. Any initial state may, of course, be used in the recursion formulae. A synopsis of the important recursions and definitions of symbols is given in Section VI.

II. RECURSION FORMULAE

Let $A(x)$ be the distribution of time between arrivals with mean arrival rate λ, and $B(x)$ the distribution of service time with mean service rate μ; the random variable ξ, which is service time minus interarrival time, has the distribution

$$K(x) = \int_{0_-}^{\infty} B(x + y)dA(y). \tag{1}$$

If W_n is the waiting time of the nth arrival, then we have the recursion

$$W_n = [W_{n-1} + \xi_n]^+, \tag{2}$$

in which

$$[x]^+ = x, \qquad x \geq 0$$
$$= 0, \qquad x \leq 0$$

and the ξ_n are iid with the common distribution $K(x)$. Let $W_n(x)$ be the distribution function of W_n; then $\hat{W}_n(s)$ is defined by

$$\hat{W}_n(s) = \int_{0_-}^{\infty} e^{-sx} dW_n(x), \tag{3}$$

and hence

$$\hat{W}_n(s) = Ee^{-sW_n}. \tag{4}$$

A recursion relating $\hat{W}_n(s)$ to $\hat{W}_{n-1}(s)$ will now be developed. From eqs. (2) and (4) we have

$$\hat{W}_n(s) = Ee^{-s(W_{n-1}+\xi_n)^+}, \tag{5}$$

and hence

$$\hat{W}_n(s) = E \int_{-\infty}^{\infty} e^{-s(W_{n-1}+x)^+} dK(x), \tag{6}$$

in which the expectation is over the distribution of W_{n-1}. Let

$$\hat{K}(s) = \int_{-\infty}^{\infty} e^{-sx} dK(x); \tag{7}$$

then the functions

$$\hat{K}_+(s) = \int_{0_-}^{\infty} e^{-sx} dK(x) \tag{8}$$

and

$$\hat{K}_-(s) = \int_{-\infty}^{0_-} e^{-sx} dK(x) \tag{9}$$

will be used in the development. From eq. (6), we find that

$$\hat{W}_n(s) = \hat{W}_{n-1}(s)\hat{K}_+(s) + \int_{-\infty}^{0_-} Ee^{-s[W_{n-1}+x]^+} dK(x). \tag{10}$$

The recursion of eq. (10) is exact but, to obtain a simple, explicit recursion, an approximation will be used at this point for the distribution of W_{n-1}. The more accurate the approximation chosen, the

more accurate the final recursion will be, but also, presumably, the more difficult to use; accordingly, a simple exponential approximation will be used.

Let

$$\alpha_n = EW_n, \tag{11}$$

and

$$J_n = 1 - W_n(0+); \tag{12}$$

then we use the approximation

$$\begin{aligned} W_n(x) &\simeq 1 - J_n e^{-\frac{J_n}{\alpha_n}x}, \qquad x \geq 0, \\ &= 0, \quad x < 0. \end{aligned} \tag{13}$$

From eq. (13) we compute

$$Ee^{-s[W_{n-1}+x]^+} = 1 - \frac{J_{n-1}s}{\frac{J_{n-1}}{\alpha_{n-1}} + s} e^{\frac{J_{n-1}}{\alpha_{n-1}}x}, \qquad x < 0. \tag{14}$$

Thus, eqs. (10) and (14) yield the recursion

$$\hat{W}_n(s) = \hat{W}_{n-1}(s)\hat{K}_+(s) + \hat{K}_-(0) - \frac{J_{n-1}s}{\frac{J_{n-1}}{\alpha_{n-1}} + s}\hat{K}_-\left(-\frac{J_{n-1}}{\alpha_{n-1}}\right). \tag{15}$$

A pair of recursion relations will now be obtained for J_n, α_n. Since

$$\lim_{s\to\infty} \hat{W}_n(s) = 1 - J_n, \tag{16}$$

$$-\hat{W}_n'(0) = \alpha_n, \tag{17}$$

and

$$\lim_{s\to\infty} \hat{K}_+(s) = 0, \tag{18}$$

we obtain from eq. (15)

$$J_n = \hat{K}_+(0) + J_{n-1}\hat{K}_-\left(-\frac{J_{n-1}}{\alpha_{n-1}}\right), \tag{19}$$

and

$$\alpha_n = \left[\hat{K}_+(0) + \hat{K}_-\left(-\frac{J_{n-1}}{\alpha_{n-1}}\right)\right]\alpha_{n-1} - \hat{K}_+'(0). \tag{20}$$

Equations (19) and (20) constitute the approximate recursions sought.

If the queueing system possesses equilibrium values, then approximations designated by J, α, $\hat{W}(s)$ are obtained from

$$J = \hat{K}_+(0) + J\hat{K}_-\left(-\frac{J}{\alpha}\right), \tag{21}$$

$$\alpha = \left[\hat{K}_+(0) + \hat{K}_-\left(-\frac{J}{\alpha}\right)\right]\alpha - \hat{K}'_+(0), \tag{22}$$

and

$$\hat{W}(s) = \frac{1}{1-\hat{K}_+(s)}\left[\hat{K}_-(0) - \frac{Js}{\frac{J}{\alpha}+s}\hat{K}_-\left(-\frac{J}{\alpha}\right)\right]. \tag{23}$$

The recursions of eqs. (15), (19), and (20), and all subsequent recursions derived from them, apply to arbitrary initial conditions and to stable or unstable queues.

III. *GI/M/1* QUEUEING SYSTEM

To ascertain the quality of the approximations obtained by eqs. (19) and (20), a control is needed. This will be provided by the explicit solution formulated by Takács[3] for the *GI/G/1* system. Accordingly, let

$$G(z, s) = \sum_{n=0}^{\infty} \hat{W}_n(s) z^n, \tag{24}$$

$$\hat{A}(s) = \int_0^{\infty} e^{-sx} dA(x), \tag{25}$$

and

$$\hat{B}(s) = \int_0^{\infty} e^{-sx} dB(x); \tag{26}$$

then

$$\hat{K}(s) = \hat{A}(-s)\hat{B}(s), \tag{27}$$

and the factorization

$$1 - z\hat{K}(s) = \Gamma_+(z, s)\Gamma_-(z, s) \tag{28}$$

defines the functions $\Gamma_+(z, s)$, $\Gamma_-(z, s)$. The function $\Gamma_+(z, s)$ is to be analytic for $R_e s > 0$ and not to vanish for $R_e s \geq 0$, while $\Gamma_-(z, s)$ is to be analytic for $R_e s < 0$ and not to vanish for $R_e s \leq 0$. Further, define the projection operator T by

$$TEe^{-s\eta} = Ee^{-s\eta+}, \tag{29}$$

in which η is an arbitrarily given random variable. Then the exact solution for the generating function, $G(x, s)$, is

$$G(z, s) = \frac{1}{\Gamma_+(z, s)} T\left[\frac{\hat{W}_0(s)}{\Gamma_-(z, s)}\right]. \tag{30}$$

The only exact solutions to be studied in this paper will correspond to the queueing system starting empty; hence, $\hat{W}_0(s) \equiv 1$. For this case the projection of eq. (30) may be evaluated; we find that

$$G(z, s) = \frac{1}{\Gamma_+(z, s)\Gamma_-(z, 0)}. \tag{31}$$

To apply eq. (31) to the $GI/M/1$, we consider

$$\hat{K}(s) = \hat{A}(-s)\frac{\mu}{\mu + s} \tag{32}$$

and

$$1 - z\hat{K}(s) = \frac{\mu + s - \mu z\hat{A}(-s)}{\mu + s}. \tag{33}$$

Let $\delta(z) > 0$ be defined by

$$\mu - \delta - \mu z\hat{A}(\delta) = 0; \tag{34}$$

then we find that

$$1 - z\hat{K}(s) = \frac{\delta + s}{\mu + s}\frac{\mu + s - \mu z\hat{A}(-s)}{\delta + s}. \tag{35}$$

Thus,

$$\Gamma_+(z, s) = \frac{\delta + s}{\mu + s}, \tag{36}$$

$$\Gamma_-(z, s) = \frac{\mu + s - \mu z\hat{A}(-s)}{\delta + s}, \tag{37}$$

and

$$\Gamma_-(z, 0) = \frac{\mu(1 - z)}{\delta}. \tag{38}$$

Hence, from eq. (31), we obtain

$$G(z, s) = \frac{\delta}{\mu(1 - z)}\frac{\mu + s}{\delta + s}. \tag{39}$$

Let j_n, a_n designate the exact values of $1 - W_n(0+)$, and EW_n, respectively, and let

$$j(z) = \sum_{n=0}^{\infty} j_n z^n, \tag{40}$$

$$a(z) = \sum_{n=0}^{\infty} a_n z^n; \tag{41}$$

then, from

$$j(z) = \frac{1}{1-z} - G(z, \infty), \tag{42}$$

$$a(z) = -\frac{\partial}{\partial s} G(z, s)|_0 \tag{43}$$

and eq. (39), we get

$$j(z) = \frac{1 - \delta/\mu}{1-z}, \tag{44}$$

and

$$a(z) = \frac{\delta^{-1} - \mu^{-1}}{1-z}. \tag{45}$$

Let $L_n = j_n - j_{n-1}$, $A_n = a_n - a_{n-1}$ and $L(z), A(z)$ be the corresponding generating functions; then elimination of δ in eqs. (44) and (45) yields

$$A(z) = \frac{1}{\mu} \frac{L(z)}{1 - L(z)} \tag{46}$$

and

$$a(z) = \frac{1}{\mu} \frac{j(z)}{1 - L(z)}. \tag{47}$$

In particular eq. (47) implies the recursion relating j_n, a_n, as follows:

$$a_n = \frac{1}{\mu} j_n + \sum_{k=1}^{n-1} A_{n-k}(j_k - j_{k-1}). \tag{48}$$

From eq. (32), we find that

$$\hat{K}_+(s) = \frac{\mu}{\mu + s} \hat{A}(\mu), \tag{49}$$

$$\hat{K}_-(s) = \frac{\mu}{\mu + s} [\hat{A}(-s) - \hat{A}(\mu)], \tag{50}$$

and

$$\hat{K}_+(0) = \hat{A}(\mu), \qquad \hat{K}'_+(0) = -\frac{1}{\mu} \hat{A}(\mu). \tag{51}$$

Hence, eqs. (19) and (20) specialized to the $GI/M/1$ are

$$J_n = \hat{A}(\mu) + \frac{J_{n-1}}{1 - \frac{1}{\mu} \frac{J_{n-1}}{\alpha_{n-1}}} \left[\hat{A}\left(\frac{J_{n-1}}{\alpha_{n-1}}\right) - \hat{A}(\mu) \right] \tag{52}$$

and

$$\alpha_n = \left\{ \hat{A}(\mu) + \frac{1}{1 - \frac{1}{\mu} \frac{J_{n-1}}{\alpha_{n-1}}} \left[\hat{A}\left(\frac{J_{n-1}}{\alpha_{n-1}}\right) - \hat{A}(\mu) \right] \right\} \alpha_{n-1} + \frac{1}{\mu} \hat{A}(\mu). \tag{53}$$

We will now show that the limiting forms of J_n, α_n, that is J, α are exact; thus $J = j$, $\alpha = a$. Equations (52) and (53) show that

$$J = \hat{A}(\mu) + \frac{J}{1 - \frac{1}{\mu}\frac{J}{\alpha}}\left[\hat{A}\left(\frac{J}{\alpha}\right) - \hat{A}(\mu)\right], \tag{54}$$

and

$$\alpha = \left\{\hat{A}(\mu) + \frac{1}{1 - \frac{1}{\mu}\frac{J}{\alpha}}\left[\hat{A}\left(\frac{J}{\alpha}\right) - \hat{A}(\mu)\right]\right\}\alpha + \frac{1}{\mu}\hat{A}(\mu). \tag{55}$$

Hence,

$$\frac{\hat{A}\left(\frac{J}{\alpha}\right) - \hat{A}(\mu)}{1 - \frac{1}{\mu}\frac{J}{\alpha}} = 1 - \frac{1}{J}\hat{A}(\mu) \tag{56}$$

and, from eqs. (55) and (56),

$$\alpha = \frac{1}{\mu}\frac{J}{1 - J} \tag{57}$$

and

$$J = \hat{A}[\mu(1 - J)]. \tag{58}$$

Since eqs. (57) and (58) are the exact relations defining j, a, the statement is proved.

An $M/M/1$ will be used to start the numerical examples. For the $M/M/1$, one has ($\rho = \lambda/\mu$)

$$J_n = \frac{1}{1 + \rho}\left(\rho + \frac{J_{n-1}}{1 + \frac{1}{\lambda}g_{n-1}}\right) \tag{59}$$

and

$$\alpha_n = \frac{1}{1 + \rho}\left[\left(\rho + \frac{1}{1 + \frac{1}{\lambda}g_{n-1}}\right)\alpha_{n-1} + \frac{\rho}{\mu}\right],$$

in which the designation $g_n = J_n/\alpha_n$ will henceforth be used.

Since $J_0 = 0$, $\alpha_0 = 0$, the computations are started with

$$J_1 = \hat{K}_+(0) = \frac{\rho}{1 + \rho}, \tag{60}$$

and

Table I—$M/M/1$, $\rho = 0.2$.

n	J	α	j	a
1	0.1667	0.1667	0.1667	0.1667
2	0.1898	0.2176	0.1898	0.2176
3	0.1962	0.2368	0.1962	0.2364
4	0.1985	0.2445	0.1985	0.2440
5	0.1994	0.2477	0.1993	0.2473

Table II—$M/M/1$, $\rho = 0.8$.

n	J	α	j	a
1	0.4444	0.4444	0.4444	0.4444
2	0.5542	0.7517	0.5542	0.7517
3	0.6047	0.9959	0.6084	0.9912
4	0.6354	1.2016	0.6418	1.1890
5	0.6570	1.3804	0.6650	1.3578

Table III—$M/M/1$, $\rho = 2.0$.

n	J	α	j	a
1	0.6667	0.6667	0.6667	0.6667
2	0.8148	1.2593	0.8148	1.2593
3	0.8719	1.8233	0.8807	1.8189
4	0.9012	2.3727	0.9172	2.3603
5	0.9191	2.9131	0.9362	2.8922

$$\alpha_1 = -K'_+(0) = \frac{1}{\mu}\frac{\rho}{1+\rho}, \tag{61}$$

which are exact. The value $\mu = 1$ will be used in all examples. Tables I through III below present approximate and exact values for $\rho = 0.2$, 0.8, and 2.0.

For the next example, the renewal stream is

$$\hat{A}(s) = \frac{1}{(1+s)(1+2s)}. \tag{62}$$

The recursion relations are

$$J_n = \frac{1}{6}\left[1 + J_{n-1}\frac{5+2g_{n-1}}{(1+g_{n-1})(1+2g_{n-1})}\right], \tag{63}$$

and

$$\alpha_n = \frac{1}{6}\left[1 + \frac{5+2g_{n-1}}{(1+g_{n-1})(1+2g_{n-1})}\right]\alpha_{n-1} + \frac{1}{6}. \tag{64}$$

Some numerical results are given in Table IV.

As a further example the case $D/M/1$ is considered. We have

$$\hat{A}(s) = e^{-sT}. \tag{65}$$

The recursion relations are

Table IV—$GI/M/1$, $\rho = 1/3$.

n	J	α	j	a
1	0.1667	0.1667	0.1667	0.1667
2	0.1991	0.2269	0.1991	0.2269
3	0.2100	0.2538	0.2101	0.2534
4	0.2147	0.2670	0.2148	0.2662
∞	0.2192	0.2808	0.2192	0.2808

Table V—$D/M/1$, $T = 2$, $\rho = 0.5$.

n	J	α	j	a
1	0.1353	0.1353	0.1353	0.1353
2	0.1720	0.1903	0.1720	0.1903
3	0.1867	0.2179	0.1868	0.2175
4	0.1938	0.2331	0.1940	0.2324
5	0.1977	0.2419	0.1978	0.2410
∞	0.2032	0.2550	0.2032	0.2550

$$J_n = e^{-T} + J_{n-1}\frac{e^{-Tg_{n-1}} - e^{-T}}{1 - g_{n-1}}, \tag{66}$$

and

$$\alpha_n = \left(e^{-T} + \frac{e^{-Tg_{n-1}} - e^{-T}}{1 - g_{n-1}}\right)\alpha_{n-1} + e^{-T}. \tag{67}$$

We find that the exact result for j_n is

$$j_n = \sum_{k=1}^{n} \frac{k^{k-1}}{k!} T^{k-1} e^{-kT}. \tag{68}$$

The equilibrium values for $T = 2$ are $J = j = 0.2032$, $\alpha = a = 0.2550$. Table V presents the numerical results.

V. $M/G/1$ QUEUEING SYSTEM

For the $M/G/1$ queue, we have

$$\hat{K}(s) = \frac{\lambda}{\lambda - s}\hat{B}(s); \tag{69}$$

hence

$$1 - z\hat{K}(s) = \frac{\lambda - s - \lambda z\hat{B}(s)}{\lambda - s}. \tag{70}$$

Define $\delta(z) > 0$ by

$$\lambda - \delta - \lambda z\hat{B}(\delta) = 0; \tag{71}$$

then

$$1 - z\hat{K}(s) = \frac{\delta - s}{\lambda - s}\frac{\lambda - s - \lambda z\hat{B}(s)}{\delta - s}, \tag{72}$$

and hence

$$\Gamma_+(z, s) = \frac{\lambda - s - \lambda z\hat{B}(s)}{\delta - s}, \tag{73}$$

$$\Gamma_-(z, s) = \frac{\delta - s}{\lambda - s}, \tag{74}$$

and

$$\Gamma_-(z, 0) = \frac{\delta}{\lambda}. \tag{75}$$

Accordingly, the generating function, $G(z, s)$, is

$$G(z, s) = \frac{\lambda}{\delta}\frac{\delta - s}{\lambda - s - \lambda z\hat{B}(s)}; \tag{76}$$

thus, the generating functions $j(z)$, $a(z)$ are

$$j(z) = \frac{1}{1 - z} - \frac{\lambda}{\delta}, \tag{77}$$

and

$$a(z) = \frac{1}{1 - z}\left(\frac{1}{\delta} - \frac{1}{\lambda}\frac{1 - z\rho}{1 - z}\right). \tag{78}$$

Elimination of δ in eqs. (77) and (78) yields

$$a(z) = \frac{1}{\lambda}\frac{1}{1 - z}\left[\frac{z\rho}{1 - z} - j(z)\right], \tag{79}$$

which implies that the relation between j_n, a_n is

$$a_n = a_{n-1} + \frac{1}{\mu} - \frac{1}{\lambda}j_n, \tag{80}$$

$$a_n = \frac{n}{\mu} - \frac{1}{\lambda}\sum_{k=1}^{n} j_k. \tag{81}$$

This may be compared with eq. (48) for the $GI/M/1$.

From eq. (69), we have

$$\hat{K}_-(s) = \frac{\lambda}{\lambda - s}\hat{B}(\lambda), \tag{82}$$

$$\hat{K}_+(s) = \frac{\lambda}{\lambda - s}[\hat{B}(s) - \hat{B}(\lambda)], \tag{83}$$

$$\hat{K}_+(0) = 1 - \hat{B}(\lambda), \tag{84}$$

and

$$\hat{K}'_+(0) = \frac{1}{\lambda}[1 - \hat{B}(\lambda)] - \frac{1}{\mu}. \tag{85}$$

Thus, the recursion formulae eqs. (19) and (20) become, for $M/G/1$,

$$J_n = 1 - \hat{B}(\lambda) + J_{n-1}\frac{\hat{B}(\lambda)}{1 + \frac{1}{\lambda} g_{n-1}}, \tag{86}$$

$$\alpha_n = \left[1 - \hat{B}(\lambda) + \frac{\hat{B}(\lambda)}{1 + \frac{1}{\lambda} g_{n-1}}\right] \alpha_{n-1} + \frac{1}{\mu} - \frac{1}{\lambda}[1 - \hat{B}(\lambda)]. \tag{87}$$

Insofar as approximations imitate characteristics of an original, we may better apply the approximations. We will now show that the approximations J_n, α_n, satisfy eq. (80). From eq. (86) we find that

$$\frac{\hat{B}(\lambda)}{1 + \frac{1}{\lambda} g_{n-1}} = \frac{J_n - 1 + \hat{B}(\lambda)}{J_{n-1}}, \tag{88}$$

$$\alpha_{n-1} = \frac{1}{\lambda} \frac{J_{n-1}[J_n - 1 + \hat{B}(\lambda)]}{J_{n-1}\hat{B}(\lambda) - J_n + 1 - \hat{B}(\lambda)}, \tag{89}$$

and, from eq. (87),

$$\alpha_n - \alpha_{n-1} = -\frac{J_{n-1}\hat{B}(\lambda) - J_n + 1 - \hat{B}(\lambda)}{J_{n-1}} \cdot \alpha_{n-1} + \frac{1}{\mu} - \frac{1}{\lambda}[1 - \hat{B}(\lambda)]. \tag{90}$$

Thus, substitution of α_{n-1} from eq. (89) into the dexter of eq. (90) yields

$$\alpha_n = \alpha_{n-1} + \frac{1}{\mu} - \frac{1}{\lambda} J_n, \tag{91}$$

and

$$\alpha_n = \alpha_0 + \frac{n}{\mu} - \frac{1}{\lambda} \sum_{k=0}^{n} J_k. \tag{92}$$

From eqs. (81) and (92), we have

$$a_n - \alpha_n = -\alpha_0 + J_0 - \frac{1}{\lambda} \sum_{k=1}^{n} (j_k - J_k); \tag{93}$$

hence, for a stable queue,

$$j_k - J_k \to 0, \qquad k \to \infty. \tag{94}$$

Since $j_k \to \rho$, then also $J_k \to \rho$, $k \to \infty$. Unlike $GI/M/1$, however, $\alpha \neq a$. The equality that occurs in $GI/M/1$ was to be expected since the waiting-time distribution was approximated by an exponential in a portion of the integration producing eq. (15), and the exact equilibrium

distribution is, in fact, exponential. This does not occur in $M/G/1$, however, since the waiting time is not exponential. The value for a is simply

$$a = \frac{\lambda}{2}\frac{B''(0)}{1-\rho}, \tag{95}$$

while solution of the equilibrium form of eqs. (86) and (87) shows that

$$\alpha = \frac{1}{\mu}\frac{\rho - 1 + \hat{B}(\lambda)}{(1-\rho)[1-\hat{B}(\lambda)]}; \tag{96}$$

thus, $\alpha \simeq a$ for small λ.

For the first numerical example, consider

$$\hat{A}(s) = \frac{\rho}{\rho + s}, \qquad \hat{B}(s) = \frac{1}{\left(1 + \frac{1}{2}s\right)^2}; \tag{97}$$

then

$$J_n = \frac{\rho}{(2+\rho)^2}\left(4 + \rho + J_{n-1}\frac{4}{\rho + g_{n-1}}\right), \tag{98}$$

$$\alpha_n = \frac{\rho}{(2+\rho)^2}\left(4 + \rho + \frac{4}{\rho + g_{n-1}}\right)\alpha_{n-1} + 1 - \frac{1}{\rho}\left(1 - \frac{4}{(2+\rho)^2}\right). \tag{99}$$

We have

$$J_1 = \frac{\rho(4+\rho)}{(2+\rho)^2}, \qquad \alpha_1 = 1 - \frac{1}{\rho}\left[1 - \frac{4}{(2+\rho)^2}\right] \tag{100}$$

for the queue starting empty. Table VI presents some numerical values for $\rho = 0.5$.

As another example let us consider the following $M/D/1$:

$$\hat{A}(s) = \frac{1}{1+2s}, \qquad \hat{B}(s) = e^{-s}. \tag{101}$$

The recursion formulae are

$$J_n = 1 - e^{-1/2} + J_{n-1}\frac{e^{-1/2}}{1 + 2g_{n-1}}, \tag{102}$$

and

$$\alpha_n = \left(1 - e^{-1/2} + \frac{e^{-1/2}}{1 + 2g_{n-1}}\right)\alpha_{n-1} + 2e^{-1/2} - 1. \tag{103}$$

The values are summarized in Table VII below.

V. *GI/G/1* QUEUEING SYSTEM

For the next group of examples, the control will provide only

equilibrium values. The following uses an interrupted Poisson arrival stream[4] defined by

$$\hat{A}(s) = \frac{3.004s + 0.913216}{s^2 + 4.308s + 0.913216}. \tag{104}$$

The service time distribution is given by

$$\hat{B}(s) = \frac{3 + 5s}{(1 + 2s)(3 + 2s)}. \tag{105}$$

Thus,

$$\hat{K}(s) = \frac{(0.913216 - 3.004s)(3 + 5s)}{(0.223586 - s)(4.084414 - s)(1 + 2s)(3 + 2s)}, \tag{106}$$

$$\hat{K}_-(s) = \frac{0.0516472}{0.223586 - s} + \frac{0.672764}{4.084414 - s}, \tag{107}$$

$$\hat{K}_+(s) = \frac{0.182021}{1 + 2s} + \frac{1.266801}{3 + 2s}, \tag{108}$$

and

$$\hat{K}_+(0) = 0.604288, \qquad \hat{K}'_+(0) = -0.645553. \tag{109}$$

The recursion formulae are

$$J_n = 0.604288 + J_{n-1}\left(\frac{0.0516472}{0.223586 + g_{n-1}} + \frac{0.672764}{4.084414 + g_{n-1}}\right), \tag{110}$$

and

$$\alpha_n = \left(0.604288 + \frac{0.0516472}{0.223586 + g_{n-1}} + \frac{0.672764}{4.084414 + g_{n-1}}\right) \cdot \alpha_{n-1} + 0.645553. \tag{111}$$

Table VI—$M/G/1$, $\rho = 0.5$.

n	J	α	j	a
1	0.3600	0.2800	0.3600	0.2800
2	0.4245	0.4310	0.4266	0.4269
3	0.4515	0.5280	0.4547	0.5174
4	0.4666	0.5948	0.4698	0.5777
5	0.4762	0.6423	0.4789	0.6198
∞	0.5000	0.7778	0.5000	0.7500

Table VII—$M/D/1$, $\rho = 0.5$.

n	J	α	j	a
1	0.3935	0.2131	0.3935	0.2130
2	0.4443	0.3244	0.4482	0.3166
3	0.4655	0.3933	0.4701	0.3764
4	0.4773	0.4387	0.4812	0.4140
5	0.4846	0.4694	0.4876	0.4388
∞	0.5000	0.5415	0.5000	0.5000

Table VIII—$GI/G/1$, interrupted Poisson, $\rho = 0.7$.

n	J	α
1	0.6043	0.6456
2	0.7123	1.1509
3	0.7498	1.5762
4	0.7703	1.9470
5	0.7842	2.2769
6	0.7947	2.5743
7	0.8031	2.8452
∞	0.8824	8.0160

Table IX—Approximate and exact waiting-time distributions.

x	$w(x) \simeq$	$w(x) =$
0	0.1176	0.1183
1	0.2190	0.2212
2	0.3014	0.3026
3	0.3728	0.3717
4	0.4363	0.4323
5	0.4933	0.4863
6	0.5448	0.5350
7	0.5911	0.5789
8	0.6330	0.6187
9	0.6706	0.6547
10	0.7045	0.6872

Table VIII presents the numerical results. The exact equilibrium values are $j = 0.8817$, $a = 8.547$.

We will use this example to illustrate eq. (23) for the approximate distribution function. Using the approximate equilibrium values J, α, as in Table VIII, we find that

$$w(x) \simeq 1 - 0.894602e^{-0.11008x} + 0.0439472e^{-0.306135x} - 0.0317493e^{-0.969454x}. \tag{112}$$

The exact waiting-time distribution is

$$w(x) = 1 - 0.8420579e^{-0.0990363x} - 0.03963069e^{-0.8959713x}. \tag{113}$$

The numerical comparison is given in Table IX.

As another $GI/G/1$ example, we consider

$$\hat{A}(s) = \frac{1}{(1+s)(1+2s)}, \tag{114}$$

and

$$\hat{B}(s) = \tfrac{1}{2}e^{-.7s} + \tfrac{1}{2}e^{-1.3s}. \tag{115}$$

Here the service consists of a step function with two values. The recursive equations are

$$J_n = 0.157825 + J_{n-1}\left(\frac{1.226734}{1 + 2g_{n-1}} - \frac{0.384559}{1 + g_{n-1}}\right), \tag{116}$$

and

$$\alpha_n = \left(0.157825 + \frac{1.226734}{1 + 2g_{n-1}} - \frac{0.384559}{1 + g_{n-1}}\right)\alpha_{n-1} + 0.068909. \tag{117}$$

We show the numerical values in Table X. The exact equilibrium values are $j = 0.1798$, $\alpha = 0.09266$.

Table X—*GI/G/1*, two-step service time distribution, $\rho = 1/3$.

n	J	α
1	0.1578	0.06891
2	0.1741	0.08688
3	0.1782	0.09278
4	0.1795	0.09485
5	0.1800	0.09558
∞	0.1802	0.0960

For the last example, a $D/G/1$ will be considered. Let

$$\hat{A}(s) = e^{-sT}, \qquad \hat{B}(s) = \frac{1}{\left(1 + \frac{1}{2}s\right)^2}; \tag{118}$$

then

$$\hat{K}(s) = \frac{e^{sT}}{\left(1 + \frac{1}{2}s\right)^2}, \qquad \hat{K}_+(s) = \frac{e^{-2T}}{\left(1 + \frac{1}{2}s\right)^2} + \frac{2Te^{-2T}}{1 + \frac{1}{2}s}, \tag{119}$$

and

$$\hat{K}_-(s) = \hat{K}(s) - \hat{K}_+(s). \tag{120}$$

The recursion equations are

$$J_n = e^{-2T}(1 + 2T) + J_{n-1}\hat{K}_-(-g_{n-1}), \tag{121}$$

and

$$\alpha_n = [e^{-2T}(1 + 2T) + \hat{K}_-(-g_{n-1})]\alpha_{n-1} + e^{-2T}(1 + T). \tag{122}$$

Tables XI through XIII show values for $T = 2$, $10/9$, and 0.5, respectively.

To obtain exact equilibrium values for the stable queues, define the roots $\delta_1(z)$, $\delta_2(z)$ by

$$1 - \tfrac{1}{2}\delta_1(z) = \sqrt{z}e^{-1/2T\delta_1(z)}, \tag{123}$$

Table XI—$D/G/1$, $t = 2$, $\rho = 0.5$.

n	J	α
1	0.0916	0.05495
2	0.1085	0.07016
3	0.1136	0.07561
4	0.1154	0.07775
5	0.1162	0.07863
∞	0.1167	0.07927

Table XII—$D/G/1$, $T = 10/9$, $\rho = 0.9$.

n	J	α
1	0.3492	0.2288
2	0.4614	0.3822
3	0.5185	0.5025
4	0.5545	0.6032
5	0.5801	0.6906
∞	0.7590	2.0631

Table XIII—$D/G/1$, $T = 0.5$, $\rho = 2$.

n	J	α
1	0.7358	0.5518
2	0.8875	1.0716
3	0.9362	1.5822
4	0.9566	2.0893
5	0.9672	2.5945
10	0.9850	5.1101
20	0.9928	10.1252

$$\tfrac{1}{2}\delta_2(z) - 1 = \sqrt{z}e^{-1/2T\delta_2(z)}; \tag{124}$$

then

$$j(z) = \frac{1}{1-z}\left[1 - \frac{1}{4}\delta_1(z)\delta_2(z)\right], \tag{125}$$

and

$$a(z) = \frac{1}{1-z}\left[\delta_1(z)^{-1} + \delta_2(z)^{-1} - 1\right]. \tag{126}$$

Thus, the equilibrium values are

$$j = 1 - \tfrac{1}{4}\delta_1(1)\delta_2(1), \qquad a = \delta_1(1)^{-1} + \delta_2(1)^{-1} - 1; \tag{127}$$

and, hence, for $T = 2$, we have $j = 0.1164$, $a = 0.07842$, and for $T = 10/9$, $j = 0.7587$, $a = 1.9895$.

VI. SYNOPSIS

We list the recursions and definitions of symbols here for ready reference.

6.1 Definitions

$A(x)$: interarrival time distribution.
$B(x)$: service time distribution.
$W_n(x)$: waiting-time distribution of nth arrival.

$$K(x) = \int_{0-}^{\infty} B(x + y)dA(y).$$

λ: mean arrival rate.
μ: mean service rate.

$$\rho = \lambda/\mu.$$

$$\hat{K}(s) = \int_{-\infty}^{\infty} e^{-sx}dK(x).$$

$$\hat{K}_+(s) = \int_{0_-}^{\infty} e^{-sx}dK(x).$$

$$\hat{K}_-(s) = \int_{-\infty}^{0_-} e^{-sx}dK(x).$$

$$\hat{A}(s) = \int_{0_-}^{\infty} e^{-sx}dA(x).$$

$$\hat{B}(s) = \int_{0_-}^{\infty} e^{-sx}dB(x).$$

$$\hat{W}_n(s) = \int_{0_-}^{\infty} e^{-sx}dW_n(x).$$

j_n: probability nth arrival sees server is busy.
J_n: approximate evaluation of j_n.
a_n: mean waiting time of nth arrival.
α_n: approximate evaluation of a_n.

$$g_n = J_n/\alpha_n.$$

All symbols without the subscript n designate equilibrium values.

6.2 Recursion formulae

6.2.1 General—Transient

$$J_n = \hat{K}_+(0) + J_{n-1}\hat{K}(-g_{n-1}),$$

$$\alpha_n = [\hat{K}_+(0) + \hat{K}_-(-g_{n-1})]\alpha_{n-1} - \hat{K}'_+(0),$$

and

$$\hat{W}_n(s) = \hat{W}_{n-1}(s)\hat{K}_+(s) + \hat{K}_-(0) - \frac{s}{s+g_{n-1}} J_{n-1}\hat{K}_-(-g_{n-1}).$$

6.2.2 General—Equilibrium

$$J = \hat{K}_+(0) + J\hat{K}_-(-g),$$

$$\alpha = [\hat{K}_+(0) + \hat{K}_-(-g)]\alpha - \hat{K}'_+(0),$$

and

$$\hat{w}(s) = \frac{1}{1-\hat{K}_+(s)}\left[\hat{K}_-(0) - \frac{s}{s+g} J\hat{K}_-(-g)\right].$$

6.2.3 GI/M/1

$$J_n = \hat{A}(\mu) + \frac{J_{n-1}}{1-\frac{1}{\mu}g_{n-1}}[\hat{A}(g_{n-1}) - \hat{A}(\mu)],$$

$$\alpha_n = \left\{\hat{A}(\mu) + \frac{1}{1-\frac{1}{\mu}g_{n-1}}[\hat{A}(g_{n-1}) - \hat{A}(\mu)]\right\}\alpha_{n-1} + \frac{1}{\mu}\hat{A}(\mu),$$

$$J = \hat{A}[\mu(1-J)],$$

and

$$\alpha = \frac{1}{\mu}\frac{J}{1-J}.$$

6.2.4 M/G/1

$$J_n = 1 - \hat{B}(\lambda) + J_{n-1}\frac{\hat{B}(\lambda)}{1+\frac{1}{\lambda}g_{n-1}},$$

$$\alpha_n = \left[1 - \hat{B}(\lambda) + \frac{\hat{B}(\lambda)}{1+\frac{1}{\lambda}g_{n-1}}\right]\alpha_{n-1} + \frac{1}{\mu} - \frac{1}{\lambda}[1-\hat{B}(\lambda)],$$

$$J = j = \rho,$$

$$\alpha = \frac{1}{\mu}\frac{\rho - 1 + \hat{B}(\lambda)}{(1-\rho)[1-\hat{B}(\lambda)]},$$

and

$$a = \frac{\lambda}{2} \frac{B''(0)}{1-\rho}.$$

REFERENCES

1. A. A. Fredericks, "A Class of Approximations for the Waiting Time Distribution in a $GI/G/1$ Queueing System," B.S.T.J., *61,* No. 3 (March 1982), pp. 295–325.
2. A. A. Fredericks, "Analysis of a Class of Schedules for Computer Systems with Real Time Applications," *Performance of Computer Systems*, M. Arrato, A. Butrimenko, and E. Gelinbe, eds., Amsterdam: North-Holland Publishing Co., 1979, pp. 201–16.
3. Lajos Takaćs, "On a Linear Transformation in the Theory of Probability," Acta Scientiarum Mathematicum, Szeged (June 1972), pp. 15–24.
4. A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *52,* No. 3 (March 1973), pp. 437–48.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

An Analysis of the Carrier-Sense Multiple-Access Protocol

By D. P. HEYMAN

(Manuscript received February 16, 1982)

In this paper we analyze the throughput and delay characteristics of the Carrier-Sense Multiple-Access protocol with a queueing model. The effects of finite buffer size, bursty arrivals, and collision detection with exponential rescheduling are examined.

The Carrier-Sense Multiple-Access protocol could be used on a bus in a packet switch. It works by sending a signal to all sources when the bus is occupied. A source postpones transmission to the bus when this signal is heard. Since the signal takes a positive amount of time to reach the sources, two or more sources occasionally will attempt to use the bus at about the same time. When this occurs, all the packets are destroyed and rescheduled.

We conclude that: (i) it is important to choose the mean rescheduling time correctly, and (ii) performance degrades significantly when compound Poisson arrivals (peaked traffic) replace Poisson arrivals (smooth traffic).

I. INTRODUCTION

The Carrier-Sense Multiple-Access (CSMA) protocol has been proposed for resolving conflicts when several sources attempt to use a single channel. In this paper we investigate the throughput and delay characteristics of CSMA.

1.1 Background

When several sources attempt to use a single channel, a protocol (in this context, protocol is synonymous with queue discipline) is required to allocate the channel among the sources. When the channel is occupied, a signal is sent indicating that state of affairs. When a source wants to use the channel, it first listens for this signal. If the signal is not heard, the source starts transmitting. If the signal is heard, the

source postpones transmitting and tries again at another time. The advantage of this system is that a device to control the sources is not required. The disadvantage is that occasionally messages will be destroyed because two sources will transmit at the same time. This is a consequence of the fact that signals travel at finite speeds, so there is a delay between the epoch when one source seizes an idle channel and the epochs when the other sources can first hear the busy-channel signal. Thus, soon after a source seizes the channel, another source may sense that the channel is free even though the channel is busy. When this occurs, both messages are destroyed because their bits have been merged. At the end (earlier if collision detection is used) of each transmission, the source determines whether the transmission was successful. If it was, the source goes about its business; if it was not, each message is rescheduled as if the channel were sensed as busy.

The CSMA protocol is used in the Ethernet* local-area distribution system and has been considered for part of a digital switch. Since CSMA is envisioned as a protocol for packet networks, we will refer to the arrivals as packets and the channel as a bus.

1.2 Relation to other work

The first study of CSMA is Kleinrock and Tobagi.[1] In addition to the version of CSMA studied here, which they call nonpersistent CSMA, they investigated various forms of persistent CSMA and slotted CSMA that are not alluded to in this paper. Their paper implicitly assumes that the rescheduling delay is infinite. This means that packets which find the bus occupied or which are destroyed in a collision are abandoned. For the most part, we assume, as their paper does, that packets arrive according to a Poisson process. In a subsequent paper by Kleinrock and Tobagi,[2] the rescheduling times are modelled as geometrically distributed random variables. They used a finite number of sources and assumed the arrivals to be quasi-random (i.e., finite-source Poisson). Also, they introduced an unnatural discretization of the time scale, which causes some small distortions in the results. A continuous version of this model is described and solved in Halfin.[3] Our analysis shares many features with the analysis found in Ref. 2, particularly the exploitation of the regeneration epochs.

Tobagi and Hunt[4] add collision detection to the model of Kleinrock and Tobagi.[2] Collision detection is a feature that informs a source that its packet has been destroyed soon after the collision occurs. Our model of collision detection is based on the model presented in Ref. 4.

Rappaport[5] and Rappaport and Bose[6] describe an elaborate model of a protocol similar to CSMA.

---

* Ethernet is a trademark of Xerox Corporation.

1.3 Summary of results

The models considered in this paper can be easily solved numerically. The following conclusions were reached by considering many numerical examples.

(*i*) It is important to choose the mean rescheduling time correctly; performance is not sensitive to changes in this number near its best value.

(*ii*) If the mean rescheduling time is based on a designed load but the realized load is different, the realized performance is almost as good as if the mean rescheduling time were based on the realized load.

(*iii*) When the traffic intensity is no larger than 0.7, the throughput is insignificantly lower than the traffic intensity.

(*iv*) Performance is significantly degraded when compound Poisson arrivals (peaked traffic) replace Poisson arrivals (smooth traffic).

(*v*) When the traffic intensity is no larger than 0.7, collision detection lowers the average waiting time significantly and reduces the sensitivity of the throughput to the mean rescheduling time. For higher traffic intensities, collision detection significantly increases the throughput and decreases the average delay.

1.4 Outline of this paper

Our model is described in detail in Section II, and the method of solution is outlined. The details of the solution are presented in Section III. Numerical examples and empirical conclusions from the model are given in Section IV. In Section V, we consider the effects of bursty traffic by introducing compound Poisson arrivals. We return to Poisson arrivals in Section VI and examine the benefits of collision detection. Appendix A records the transition probabilities omitted from the text, and Appendix B lists the most important symbols used in the text.

II. MODEL AND OVERVIEW OF THE SOLUTION METHOD

Our setting is a queue of the $M/D/1/K$ type. Let λ be the rate of the Poisson arrival process. There is a single server with a constant service time per packet. At most K packets may be in the system (queue plus server) at any time. Packets that arrive when the system is full are lost forever and have no effect on the system. Our use of constant service times reflects the assumption that all packets have the same length.

The parameter K has two potential interpretations. One is to suppose there is a buffer that can hold $K - 1$ packets, so the maximum number of packets in the queue is $K - 1$. The other is to suppose there are K ports and a packet that cannot seize a port is lost. We will show that proper limiting probabilities will not exist without this bound on the number of packets in the system.

When a packet is on the bus (i.e., when it enters service), a signal is sent that warns other packets that the bus is occupied. (Equivalently, a signal indicating the bus is free is turned off when a packet is on the bus.) Let h denote the one-way propagation delay for this signal. It is common to all potential users. The propagation time between a source and the bus depends on the distance between them. If we calculate h for the source that is farthest from the bus, the assumption that h, the one-way propagation delay, is common to all users will underestimate the performance measures. This propagation delay has two effects. The first effect is that for the first h time units after a packet occupies a previously idle bus, other packets that arrive and want to use the bus will do so because the signal that the bus is occupied has not reached them. When this happens, we say that a "collision" has occurred and the first h time units of a service time are called the "vulnerable" period. The second effect is that the signal that the bus is free is not received until the bus has been free for h time units.

When a collision occurs, the bits in both packets are scrambled, and so both packets must be retransmitted. We assume that the collision is detected at the end of the unsuccessful transmission. Then each source involved in the collision draws a number from an exponential distribution with mean $1/\alpha$, which is the length of time until the source next attempts to use the bus. Furthermore, all sources that attempt to transmit a packet and hear the signal that the bus is occupied make their next attempt in this fashion. We call α the "retry rate."

Our objective is to express the throughput (equivalently, the asymptotic departure rate or the asymptotic proportion of time that the bus is successfully transmitting packets) and average waiting time of a packet (i.e., the departure time minus the arrival time) to the system parameters λ, K, h, and α.

2.1 Outline of solution

Choose time units so that the processing time of a packet is unity. Consider a packet, $\mathscr{P}$, that gets on an empty bus at time t. This packet will relinquish the server at $t + 1$ and the signal that $\mathscr{P}$ is on the bus will cease being sensed at some random time τ with $t + 1 + h \leq \tau \leq t + 1 + 2h$. If no packets arrive during the vulnerable period, then $\tau = t + 1 + 2h$.

Suppose we assume that $\mathscr{P}$ holds the bus for a (nonrandom) length of time ν, with $1 \leq \nu \leq 1 + 2h$, at the end of which $\mathscr{P}$ relinquishes the server and the signal that $\mathscr{P}$ is on the bus ceases. Retries that occur during the vulnerable period destroy the messages involved in the collision. These retries are not considered to have seized the bus.

When $\nu = 1$, these assumptions will provide an upper bound for the throughput and a lower bound for the average number of customers

present (in a buffer or on the bus) in the steady state. When $\nu = 1 + 2h$, a lower bound for the throughput and an upper bound for the average number of customers present is obtained. Little's theorem (the queueing formula $L = \lambda W$) implies that setting $\nu = 1$ provides an upper bound for the average waiting time and that setting $\nu = 1 + 2h$ provides a lower bound. Therefore, we will solve a model with constant service times ν.

Since h is typically much smaller than one (we use $h = 0.01$ in our examples), we expect that the bounds will be close together. That is the case in our numerical examples. Therefore, either of the bounds is a good approximation to the true measure of performance.

Let $X(t)$ be the number of packets present at time t. We want to obtain

$$p_i = \lim_{t\to\infty} P\{X(t) = i\} \tag{1}$$

for each $i = 1, 2, \cdots, K$. These "limiting probabilities" are easily shown to be independent of the distribution of the initial state $X(0)$. To obtain $\{p_i\}$, we embed a discrete Markov chain at customer "ejection" epochs, which are defined as follows. Let C_m be the epoch where the mth packet to get on an empty bus leaves the bus. This may be the end of a successful transmission where the packet leaves the system or it may be the end of a destroyed transmission where the packet rejoins the queue. To encompass both types of events, we call C_m the mth ejection epoch.

Let $Y_m = X(C_m^+) \triangleq \lim_{s\downarrow 0} X(C_m + s)$; it represents the number of customers present just after C_m. Since the arrivals are Poisson and retrials are governed by an exponential distribution, it is easy to see that $\{Y_m; m = 0, 1, \cdots\}$ is a Markov chain. It is also easy to see that this chain is irreducible and aperiodic (details are omitted) so that the limits

$$\pi_i \triangleq \lim_{m\to\infty} P\{Y_m = i\}, \qquad i = 0, 1, \cdots, K,$$

exist and are independent of Y_0, are positive, and sum to one.

Three steps are used to obtain $\{p_i\}$ from $\{\pi_i\}$. The first step is to relate $\{\pi_i\}$ to $\{p_i'\}$ where

$$p_i' \triangleq \lim_{n\to\infty} P\{Y_n = i \text{ and transmission is successful}\},$$

$$i = 0, 1, \cdots, K-1. \tag{2}$$

The relationship is obtained by calculating the state-dependent probability of having a collision. The second step is to use a "rate up equals rate down" argument to show

$$\lambda p_i = \zeta p_i', \qquad i = 0, 1, \cdots, K-1, \tag{3}$$

where ζ is the ejection rate. The third step is to obtain p_0 by a renewal-reward argument. Then eq. (3) is used to calculate ζ and the remaining p_i's.

III. SOLVING THE MODEL

In this section we give details of the solution of the model described in Section II.

3.1 Collision probabilities

Let c_n be the probability that n packets arrive in an interval of length, v; then

$$c_n = e^{-\lambda v}\frac{(\lambda v)^n}{n!}, \qquad n = 0, 1, \cdots .$$

A well-known property of Poisson arrival processes (e.g., Corollary 5-13 in Heyman and Sobel[7]) is that the arrival epochs are iid and uniform over $[0, v]$ when it is given that $n > 0$ arrivals occurred during $[0, v]$. Thus, for a service interval that starts with i packets in the queue, we have

$$\begin{aligned}\eta_n &\triangleq \mathrm{P}\{\text{no arrivals prior to } h | n \text{ arrivals in a service interval that starts with } i \text{ in queue}\}\\ &= \left(\frac{v-h}{v}\right)^n, \qquad n = 0, 1, \cdots \text{ and any } i. \end{aligned} \tag{4}$$

Use the memoryless property of the exponential distribution to obtain

$$\begin{aligned}\delta_i &\triangleq \mathrm{P}\{\text{no retries prior to } h | n \text{ arrivals in a service that starts with } i \text{ in queue}\}\\ &= e^{-i\alpha h}, \qquad i = 0, 1, \cdots \text{ and any } n. \end{aligned} \tag{5}$$

Since a collision is avoided if, and only if, there are no arrivals and no retries during the vulnerable period, we have

$$\begin{aligned}\bar{d}_n(i) &\triangleq P\{\text{no collision and } n \text{ arrivals in a service interval} | \text{start with } i \text{ in queue}\}\\ &= \eta_n c_n \delta_i, \; i, n \geq 0, \end{aligned} \tag{6}$$

and

$$\begin{aligned}d_n(i) &\triangleq P\{\text{collision and } n \text{ arrivals in a service interval} | \text{start with } i \text{ in queue}\}\\ &= (1 - \eta_n\delta_i)c_n, \; i, n \geq 0. \end{aligned} \tag{7}$$

Equations (6) and (7) are used to calculate the transition probabilities of the embedded Markov chain $\{Y_m; m = 0, 1, \cdots\}$.

3.2 Transition probabilities

Recall from Section 2.1 that Y_m is the number of packets in the system just after the mth ejection epoch. All of these packets must be in the queue because of the propagation delay in broadcasting that the bus is available. Let

$$p_{ij} \triangleq P\{Y_{m+1} = j \mid Y_m = i\}. \tag{8}$$

In this section we present formulas for computing p_{ij}, $0 \le i, j \le K$. It will be convenient to call both exogenous packet arrivals and retries "arrivals"; the former are "outside" arrivals. Then, for $2 \le i \le K-1$ and $0 \le i+n \le K-2$, we may write

$$\begin{aligned} p_{i,i+n} &= P\{\text{next arrival is from outside} \mid i \text{ in queue}\} \\ &\quad \times [P\{n \text{ outside arrivals during service and no collision} \mid \text{service starts with } i \text{ in queue}\} \\ &\quad + P\{n-1 \text{ outside arrivals during service and a collision} \mid \text{service starts with } i \text{ in queue}\}] \\ &\quad + P\{\text{next arrival is a retry} \mid i \text{ in queue}\} \\ &\quad \times [P\{n+1 \text{ outside arrivals during service and no collision} \mid \text{service starts with } i-1 \text{ in queue}\} \\ &\quad + P\{n \text{ outside arrivals during service and a collision} \mid \text{service starts with } i-1 \text{ in queue}\}]. \end{aligned}$$

Use the memoryless property of the exponential distribution to obtain

$$P\{\text{next arrival is from outside} \mid i \text{ in queue}\} = \frac{\lambda}{\lambda + i\alpha}.$$

Set $\beta = \lambda/\alpha$ for notational simplicity; then

$$p_{i,i+n} = \frac{\beta}{i+\beta}[\bar{d}_n(i) + d_{n-1}(i)] + \frac{i}{i+\beta}[\bar{d}_{n+1}(i-1) + d_n(i-1)]. \tag{9}$$

To obtain the remaining entries of (p_{ij}) we need to account for boundary conditions. The details and the formulas are given in Appendix A.

It is easy to see that (p_{ij}) is irreducible and aperiodic. Therefore, there is a unique stationary distribution that is also the limiting distribution. Furthermore, since the continuous-time process $\{X(t); t \ge 0\}$ starts from scratch (i.e., it regenerates) each time the Y-process reaches zero, the X-process is regenerative and the mean time between regeneration points is finite. This implies the limits in eq. (1) exist. A similar argument establishes that the limits in eq. (2) exist.

We can use eq. (9) to explain why the finite capacity, K, is essential for our analysis. Suppose $K = \infty$; then eq. (9) is valid for $i \ge 2$ and every $n \ge 0$. Equation (9) is also valid for $n = -1$ if we set $d_j(i) = 0$ whenever $j < 0$. Define

$$g_i = \sum_{1}^{\infty} n p_{i,i+n} - p_{i,i-1};$$

it is the expected size of a jump out of state i in the Y-process. Intuitively, if $\lim_{i\to\infty} g_i > 0$, the Y-process is tending to infinity and $\lim_{m\to\infty} P\{Y_m \leq j\} = 0$ for all j. From eq. (9) we can calculate (details are omitted)

$$g_i = \lambda\nu + \frac{\beta(1 - e^{-i\alpha h}e^{-\lambda h})}{i + \beta} - \frac{i[e^{-\alpha(i-1)}(e^{-\lambda\nu} - e^{-\lambda h}) + e^{-\alpha i}e^{-\lambda\nu}]}{i + \beta};$$

and so, if $h > 0$,

$$\lim_{i\to\infty} g_i = \lambda\nu. \tag{10}$$

This phenomenon can be described physically. When j is large, $e^{-j\alpha h}$ is small so the probability that an ejection is a completion is very small. This means that once state j is reached (as it must be because the process is irreducible), the number of customers present grows monotonically with very high probability.

The corollary in Kaplan[8] states that an irreducible and aperiodic Markov chain for which eq. (10) holds is not ergodic if $p_{ij} = 0$ whenever $j < i - k$ for some k that is independent of i. At most, one departure can occur at a time, so $p_{ij} = 0$ whenever $j < i - 1$. Therefore, the Y-process is not ergodic when $K = \infty$.

3.3 Solving the balance equations

Instead of solving $\pi = \pi P$ to obtain $\{\pi_i\}$, we will solve balance equations between two sets of states. Let

$$p_i(\geq j) \triangleq P\{Y_{m+1} \geq j \mid Y_m = i\} = \sum_{k=j}^{K} p_{ik}.$$

In the steady state, the rate of transitions between states $\{0, 1, \cdots, i\}$ and states $\{i + 1, i + 2, \cdots, K\}$ must be the same in both directions, so*

$$\pi_{i+1}p_{i+1,i} = \sum_{k=0}^{i} \pi_k p_k(\geq i + 1), \qquad i = 0, 1, \cdots, K - 1. \tag{11}$$

We can solve eq. (11) by replacing π_i by x_i and setting $x_0 = 1$, and using eq. (11) to obtain x_{i+1} from $x_0, \cdots, x_i$. By setting

$$\pi_i = x_i \Big/ \sum_{j=0}^{K} x_j$$

we obtain a nonnegative solution of eq. (11) that sums to one.

---

* The basic idea was developed by Robert Morris and Eric Wolman. The details for Markov chains are given in Theorem 7-13 of Heyman and Sobel.[7]

3.4 Departure point probabilities

Recall that C_m is the mth ejection epoch. Let us write $C_m = D$ when C_m is a departure epoch. Now

$$P\{C_{m+1} = D,\ Y_{m+1} = j\}$$
$$= \sum_{i=0}^{K} P\{C_{m+1} = D,\ Y_{m+1} = j \mid Y_m = i\} P\{Y_m = i\}$$

for every j and m. We will see below that $P\{C_{m+1} = D,\ Y_{m+1} = j \mid Y_m = i\}$ does not depend on m, so denote it by q_{ij}. Then

$$p'_j = \lim_{m\to\infty} P\{C_{m+1} = D,\ Y_{m+1} = j\} = \sum_{i=0}^{K} q_{ij}\pi_i,$$
$$j = 0, 1, \cdots, K-1. \tag{12}$$

Observe that, for $1 \le j \le K-2$ and $1 \le i \le j-1$, $P\{C_{m+1} = D,\ Y_{m+1} = j \mid Y_m = i\}$

$= P$ {next arrival is from outside$|i$ in queue}
$\times P$ {$j - i$ outside arrivals during service and no collision|service starts with i in queue}
$+ P$ {next arrival is a retry$|i$ in queue}
$\times P$ {$j - 1 + 1$ outside arrivals during service and no collision|service starts with $i - 1$ in queue}

$$= \frac{\beta}{\beta + i}\bar{d}_{j-i}(i) + \frac{i}{\beta + i}\bar{d}_{j-i+1}(i-1).$$

The remaining entries are obtained by accounting for boundary behavior. The results are:

$$q_{00} = c_0\eta_0 = c_0, \tag{13a}$$

$$q_{10} = \frac{c_0}{1+\beta}, \tag{13b}$$

and

$$q_{i0} = 0 \quad \text{for} \quad i \ge 2. \tag{13c}$$

For $1 \le j \le K-2$,

$$q_{0j} = \bar{d}_j(0), \tag{14a}$$

$$q_{ij} = \frac{\beta}{\beta + i}\bar{d}_{j-i}(i) + \frac{i}{\beta + i}\bar{d}_{j-i+1}(i-1), \qquad 1 \le i \le j, \tag{14b}$$

$$q_{j+1,j} = \frac{j+1}{\beta + j + 1}\bar{d}_0(j), \tag{14c}$$

and

$$q_{ij} = 0, \qquad i \geq j + 2. \tag{14d}$$

When $j = K - 1$ we obtain

$$q_{0,K-1} = \sum_{K-1}^{\infty} \bar{d}_n(0), \tag{15a}$$

$$q_{i,K-1} = \frac{\beta}{\beta + 1} \sum_{K-1-i}^{\infty} \bar{d}_n(i) + \frac{i}{\beta + i} \sum_{K-i}^{\infty} \bar{d}_n(i-1),$$

$$1 \leq i \leq K - 1 \tag{15b}$$

$$q_{K,K-1} = \delta_{K-1}.$$

In eq. (46) we show that the infinite sums in eq. (15) have representations as finite sums containing no more than K terms. Since $C_{m+1} = D$ and $Y_{m+1} = K$ cannot occur simultaneously, $q_{iK} = 0$ for every i.

Notice that (the subscript ∞ denotes a limit)

$$n_c \triangleq \sum_{0}^{K-1} p'_j = P\{C_\infty = D,\ Y_\infty \leq K - 1\}$$

$$= P\{\text{an ejection epoch is a departure epoch}\}$$

$$= P\{\text{no collision}\}, \tag{16}$$

where the last two probabilities are steady-state quantities and can be interpreted as long-run proportions.

3.5 Converting $\{p'_i\}$ to $\{p_i\}$

Let $E(t)$ be the number of ejections by time t. Since $\{E(t);\ t \geq 0\}$ regenerates when an ejection leaves the system empty (i.e., when $Y_m = 0$ for some m),

$$\zeta \triangleq \lim_{t \to \infty} E(t)/t$$

exists. It is the ejection rate. It is also the rate at which packets gain access to an idle bus, so ζ is the arrival rate of new and rescheduled packets, which is denoted by G in Kleinrock and Tobagi.[1]

The rate at which the number of packets present jumps from i to $i + 1$ is λp_i. The rate at which it jumps from $i + 1$ to i is $\zeta p'_i$. In the steady state these rates must be equal so we have

Lemma 1: $\lambda p_i = \zeta p'_i, \qquad i = 0, 1, \cdots, K - 1.$ (17)

Lemma 1 can be proved rigorously. The proof uses standard methods and is omitted.

There are $K + 1$ unknowns in eq. (17) and K equations. We will find

p_0 by an independent argument; the remaining p_i's and ζ are obtained from eq. (17).

Let T_0 be the amount of time $X(\cdot)$ is zero in an arbitrary cycle of the X-process, and let M be the expected length of a regeneration cycle. A basic property of regenerative processes (see, e.g., Theorem 6–7 in Heyman and Sobel[7]) is that

$$p_0 = E(T_0)/M. \tag{18}$$

Since $X(t) = 0$ if, and only if, t is in an idle period,

$$E(T_0) = 1/\lambda. \tag{19}$$

When $Y = i$, let ψ_i be the average time between en ejection epoch and the next ejection epoch. We have

$$\psi_i = \begin{cases} \nu + \dfrac{1}{i\alpha + \lambda} & \text{if} \quad i < K \\ \nu + \dfrac{1}{K\alpha} & \text{if} \quad i = K, \end{cases} \tag{20}$$

because ψ_i is ν plus the expected time to the next arrival after the ejection epoch. Let m_i be the mean number of visits of the Y-process to state i during an arbitrary regeneration cycle of the X-process. Thus, m_i is the mean number of visits of the Y-process to state i between visits of the Y-process to state zero, because the X-process regenerates whenever the Y-process enters state zero. It is easy to show (see Exercise 7-78 in Heyman and Sobel[7]) that

$$m_i = \pi_i/\pi_0, \qquad i = 0, 1, \cdots, K. \tag{21}$$

Using eqs. (20) and (21) we obtain

$$M = \sum_{i=0}^{K} \psi_i m_i = \frac{1}{\pi_0}\left[\sum_{0}^{K-1} \pi_i \left(\nu + \frac{1}{i\alpha + \lambda}\right) + \pi_K \left(\nu + \frac{1}{K\alpha}\right)\right]$$

$$= \frac{1}{\pi_0}\frac{1}{\lambda}\left(\rho + \sum_{0}^{K-1} \frac{\beta\pi_i}{\beta + i} + \frac{\beta\pi_K}{K}\right), \tag{22}$$

where $\rho = \lambda\nu$.

Substituting eqs. (19) and (22) into (18) yields

$$p_0 = \pi_0 \Big/ \left(\rho + \sum_{0}^{K-1} \frac{\beta\pi_i}{\beta + i} + \frac{\beta\pi_K}{K}\right). \tag{23}$$

Obviously [and formally from eqs. (12) and (13)], we find that

$$p_0' = \pi_0. \tag{24}$$

Combining eq. (17) with $i = 0$ and eqs. (23) and (24) yields

$$\zeta = \lambda p_0 / p_0' = \lambda \Big/ \left(\rho + \sum_0^{K-1} \frac{\beta \pi_i}{\beta + i} + \frac{\beta \pi_K}{K} \right). \tag{25}$$

Equation (25) shows that ζ can be computed when the balance equations are solved.

3.6 The throughput, occupancy, and average waiting time

The throughput, θ, is the asymptotic departure rate. Since the departure process regenerates whenever the ejection process regenerates, θ is well defined. The asymptotic departure rate equals the asymptotic rate at which packets are accepted, because all accepted packets eventually depart and the number of packets present is no larger than K; therefore

$$\theta = \lambda \sum_0^{K-1} p_i. \tag{26}$$

Combining eq. (26) with eqs. (16) and (17) yields

$$\theta = \zeta n_c. \tag{27}$$

This equation states that the departure rate equals the ejection rate multiplied by the asymptotic proportion of ejections that do not suffer a collision. Equation (27) illuminates the essential trade-offs involved in using CSMA. One expects that ζ increases and n_c decreases as λ and α increase. This means that for each λ, there is a value of α that maximizes θ. Let $\theta^*(\lambda)$ be the largest value of θ that can be achieved when λ is specified. We also expect that $\theta^*(\lambda)$ will first increase with λ and then start decreasing.

Let ϕ be the long-run proportion of time that the bus is occupied. The regenerative arguments used above can be used to prove that ϕ is well defined and that

$$\phi = \nu \sum_0^K m_i / M. \tag{28}$$

Using eqs. (22) through (25) in conjunction with eq. (28) yields

$$\phi = \nu \zeta. \tag{29}$$

This equation can be obtained from Little's theorem by regarding the bus as "the system." Then ζ is the arrival rate (since only packets that depart could have arrived) and ν is the expected time a packet is on the bus. Little's theorem asserts that $\zeta\nu$ is the average number of packets on the bus, which is the proportion of time that the bus is occupied. Combining eqs. (27) and (29) yields

$$\phi = \nu\theta / n_c. \tag{30}$$

Suppose that we attempt to increase θ by increasing the arrival rate, and K and α are adjusted to keep n_c constant (i.e., a fixed proportion of collisions). Equation (27) shows that increases in θ must be accompanied by decreases in the occupancy of the bus, and that doubling the throughput would halve the occupancy.

From eqs. (17) and (25), p_i for $1 \leq i \leq K-1$ is obtained in the obvious way and p_K is obtained from $p_K = 1 - \sum_0^{K-1} p_i$. Then

$$L = \sum_1^K i p_i$$

is the average number of packets present in the steady state. Since θ is the arrival rate of packets that enter the system, Little's theorem yields

$$W = L/\theta, \tag{31}$$

where W is the average length of time that a packet is in the system.

3.7 Relation to the model of Kleinrock and Tobagi

In this section we explain how the "basic equation for the throughput" [eq. (3) in Ref. 1] can be obtained from the model in this paper. This will clarify the similarities and differences between the two models.

In Ref. 1 Assumption 1 states that the average retry interval is large compared with the packet transmission time. In our notation, this assumption is that $1/\alpha$ is large compared with ν. Since there is no parameter corresponding to α in eq. (3) of Ref. 1, it appears that they have set $\alpha = 0$. Let us do that. This means that every transmission that is destroyed by a collision stays in the queue forever. Consequently, Assumption 2 in Ref. 1, which states that the interarrival times of the point process consisting of packet arrival epochs and retry epochs is a Poisson process, is valid because there are no retry epochs and the arrival epochs form a Poisson process.

Another consequence of $\alpha = 0$ is that without a finite buffer, $\lim_{t\to\infty} P\{X(t) = i\} = 0$ for all i because each packet transmission has a positive probability of being destroyed by a collision. Therefore, we make these two assumptions.

Assumption 3. There is a finite buffer that can hold $K - 1 \geq 0$ packets.

Assumption 4. If a packet transmission is destroyed when the buffer is full, that packet is flushed from the system.

The next assumption is required to replicate the collision process in Ref. 1.

Assumption 5. Packets that arrive when the buffer is full and a packet is being transmitted will destroy that transmission.

With these assumptions and $\alpha = 0$, the arguments and formulas in Sections 3.2 through 3.5 can be used to obtain the results given below, but we will bypass that roundabout route and give a direct argument (which is essentially the argument in Ref. 1).

In the steady state, whenever a packet seizes the bus, there are $K - 1$ packets in the buffer. The mean time between entries to state $K - 1$ is, via eq. (20), $\nu + 1/\lambda$, and $1/\lambda$ is the mean length of stay in state $K - 1$, so

$$p_{K-1} = \frac{1/\lambda}{\nu + 1/\lambda} = \frac{1}{1 + \lambda\nu}. \tag{32}$$

The probability that a transmission is not destroyed by a collision is $e^{-\lambda h}$, so

$$\theta = \lambda p_{K-1} e^{-\lambda h}. \tag{33}$$

Equation (7) of Ref. 1 and $\alpha = 0$ show that

$$\nu = 1 + 2h - (1 - e^{-\lambda h})/\lambda. \tag{34}$$

Substituting eqs. (32) and (34) into eq. (33) yields

$$\theta = \frac{\lambda e^{-\lambda h}}{\lambda(1 + 2h) + e^{-\lambda h}},$$

which is eq. (3) in Ref. 1.

We can conclude that eq. (3) in Ref. 1 is valid when $\alpha = 0$ and the derivation in Ref. 1 requires assumptions 3, 4, and 5. Our model does not use assumptions 4 and 5 so it should produce a greater throughput when all other factors are the same; this is demonstrated numerically in Section IV. Since $K = 1$ is allowed, this derivation explains why eq. (3) in Ref. 1 becomes the single-server Erlang loss formula when there are no collisions, i.e., why $h \downarrow 0$ yields eq. (9) in Ref. 1.

IV. NUMERICAL RESULTS

A Fortran program to solve the equations in Section III was implemented in double-precision arithmetic on a PDP-1170. Because of memory limitations and the design of the program, $K \leq 30$ was required. The running time on a CRT display terminal with a 9600-baud line is less than a twinkling of an eye.

In all the numerical examples we use $h = 0.01$. This is the value used by other authors. It is a reasonable value to use for a system such as Ethernet with loops of about 300 meters carrying 128-byte packets at 10 megabits/sec. In general,

$$h\nu = \frac{lr}{sc},$$

where

l = loop length in meters,

r = transmission rate in bits/second,

s = packet size in bits, and

c = speed of light in m/second.

4.1 The upper and lower bounds are close

In Table I we present various cases that demonstrate that the lower and upper bounds for θ and W are close. The subscript "avg" stands for average and the entries are obtained by setting $\nu = 1 + h$. In all these cases, $K = 20$.

Table I suggests the approximations

$$W_{\text{avg}} = \tfrac{1}{2}(W_{lb} + W_{ub}) \quad \text{and} \quad \theta_{\text{avg}} = \tfrac{1}{2}(\theta_{lb} + \theta_{ub}).$$

We make no claim for the general validity of these approximations. Based on the encouraging results in Table I, and on several dozen other examples not reported here, we will henceforth use the case $\nu = 1 + h$ as an approximation; there will be no explicit mention that our numerical results are approximate.

4.2 The effects of changing α

The first two pairs of columns in Table I show that for fixed λ, different values of α yield different values of the performance measures. In this section we explore the consequences of changing α.

When K and λ are given, the maximum achievable throughput is attained when there are no lost service times from collisions and instantaneous retries, i.e., by an $M/D/1/K$ queue. Let the throughput of the $M/D/1/K$ queue be denoted by $\theta_{\max}$. When $K = 20$ and $\lambda = 0.7$, $\theta_{\max} = 0.700$ to three decimal places. In Table II we show that good choices of α will achieve $\theta = 0.699$, but the throughput for poor choices of α have a much lower value. In particular, very small values of α do

Table I—Evidence that the bounds for θ and W are close

λ	0.7	0.7	0.5	0.9	1.0	3.0
α	0.01	3.0	0.5	1.0	1.0	2.0
θ_{ub}	0.459	0.673	0.818	0.803	0.798	0.560
θ_{avg}	0.457	0.667	0.812	0.796	0.790	0.555
θ_{lb}	0.455	0.660	0.806	0.788	0.782	0.549
W_{ub}	42.1	11.4	19.6	19.9	22.7	35.8
W_{avg}	41.9	10.2	19.1	19.3	22.3	35.5
W_{lb}	41.7	9.1	18.7	18.7	21.9	35.1
avg n_c	0.991	0.828	0.923	0.861	0.824	0.570
avg ϕ	0.466	0.814	0.888	0.933	0.947	0.984

not do well. This confirms the observation given in Section 3.7 that the model in Ref. 1, in which $\alpha = 0$, underestimates the throughput.

The explanation for the qualitative properties shown in Table II is as follows. When α is very small, a packet that upon arrival finds another packet in service will spend a long time waiting to retry. This makes collisions rare but tends to keep the buffer positions full, which makes the throughput and bus utilization low. As α is increased, the retries are more frequent, and although there are more collisions, this is more than offset by the shorter times spent waiting to retry. When α is made sufficiently large, the retries occur so rapidly that too many packets are destroyed and performance degrades.

A significant feature of the data in Table II is that θ does not vary much for $0.5 \leq \alpha \leq 1.6$. The variation within the interval $0.5 \leq \alpha \leq 1.4$ is in the fourth decimal place. The value of α that achieves the lowest value of W also achieves a high value of θ, but the value of α that achieves the highest value of θ has a value of W about 9 percent larger than the best value of W we have found.

Let θ^* denote the largest value of θ that is found for fixed values of λ, h, and K. Numerical results not reported here have shown that $\theta^*/\theta_{\max}$ increases as λ decreases. This is not surprising because lower input rates result in fewer collisions. Since $\theta^* = 0.6993$ and $\theta_{\max} = 0.700$, $\theta^*/\theta_{\max}$ is very close to one. This suggests that CSMA will provide good throughput performance when $\lambda \leq 0.7$, and that collision detection schemes will not increase throughput very much when $\lambda \leq 0.7$.

4.3 The effects of changing λ

Take $K = 20$ and $h = 0.01$. Choose α so that the throughput is maximized. The effects of changing λ are shown in Table III.

Notice that θ increases with λ for $\lambda \leq 2$ and then θ decreases very slightly when $\lambda = 3$. This indicates that the phenomenon of lower throughput with higher arrival rate will not occur in the normal operating range of $\lambda \leq 1$. Table III also shows that α should not increase as λ increases. This property also appeared in every example we tried.

Table II—Performance measures vs. α when K = 20 and $\lambda = 0.7$

α	0.001	0.01	0.1	0.5	0.8	1.0
θ	0.362	0.457	0.660	0.6989	0.6993*	0.6992
W	53.9	41.9	22.8	8.34	6.51	6.51
n_c	0.993	0.991	0.979	0.968	0.963	0.963
ϕ	0.368	0.468	0.681	0.729	0.734	0.734
α	1.4	1.6	2.0	3.0	4.0	5.0
θ	0.6986	0.6980	0.696	0.667	0.556	0.423
W	5.53	5.52	5.87	10.2	24.1	42.1
n_c	0.949	0.943	0.927	0.828	0.612	0.437
ϕ	0.743	0.747	0.758	0.814	0.917	0.977

The average waiting times shown in Table III are about the same as the corresponding quantities for 1000 sources shown in Fig. 3 of Ref. 2.

Suppose we design a system for a nominal arrival rate and from time to time the actual arrival rate differs from the nominal arrival rate. In Table II we see that performance suffers if a poor value of α is chosen. Now we show that moderate changes in λ will not cause much degradation in the throughput rate or the average waiting time.

We assume that $K = 10$ and $h = 0.01$ are held fixed. The nominal arrival rate is 0.7 and $\alpha = 1.6$ yields the largest throughput rate. We also assume that when λ changes, the new value is maintained long enough to ensure that steady-state performance measures are adequate. If we achieve good performance with arrival rates λ_1 and λ_2, it is reasonable to suppose that we will achieve reasonable performance when λ is changing from λ_1 to λ_2.

We will use the notations θ^* and W^* for the best values of θ and W we have found by varying α. The entries marked θ and W are for $\alpha = 1.6$.

Table IV shows that, for the values chosen, deviations from the nominal load will not cause serious performance degradation if α is kept fixed.

4.4 The effects of changing K

As K is increased, more packets can be stored, so the throughput should increase. This cannot be carried too far because as $K \to \infty$, $\lim_{n\to\infty} Y_n = \infty$ (as shown in Section 3.2), which causes the probability of a collision to approach 1 and $\theta \to 0$. We have not found a value of K to demonstrate this numerically.

Table III—Performance measures vs. λ when K = 20 and h = 0.01

λ	0.7	0.9	1.0	2.0	3.0
α	0.8	0.6	0.5	0.5	0.4
θ	0.699	0.813	0.817	0.818	0.817
θ_{max}	0.700	0.898	0.975	1.00	1.00
W	6.51	18.9	21.4	23.9	24.1
n_c	0.963	0.911	0.914	0.913	0.905
ϕ	0.734	0.901	0.904	0.905	0.912

Table IV—θ and W vs. λ when $\alpha = 1.6$, K = 10, and h = 0.01

λ	0.5	0.6	0.7	0.8	0.9	1.0
θ	0.500	0.599	0.692	0.764	0.801	0.812
θ^*	0.500	0.599	0.692	0.764	0.801	0.815
W	2.30	3.07	4.37	6.22	8.06	9.39
W*	2.18	2.65	4.18	6.09	8.05	9.38

In Table V below we use the following notations:

θ^* = largest value of θ found

α_θ^* = value of α that achieves θ^*

W^* = smallest value of W found

α_W^* = value of α that achieves W^*

$W(\alpha_\theta^*)$ = W when $\alpha = \alpha_\theta^*$

$\theta(\alpha_W^*)$ = θ when $\alpha = \alpha_W^*$

$\theta_{\max}$ = throughput of $M/D/1/K$ queue.

In Table V, θ^* and W^* increase with K; α_W^* and α_θ^* decrease with K; $\alpha_W^* \geq \alpha_\theta^*$, and the difference is small and gets smaller as K increases; $W(\alpha_\theta^*)$ is not much bigger than W^*; and $\theta(\alpha_W^*)$ is not much smaller than θ^*.

4.5 Conclusions

From the data presented in this section, and from dozens of unreported sets of calculations, we conclude that:

(*i*) The bounds on θ and W are close, and choosing $\nu = 1 + h$ yields a good approximation.

(*ii*) It is important to choose a good value of α, and performance is not sensitive to changes in α near the best value.

(*iii*) The value of α that minimizes W is close to the value of α that maximizes θ.

(*iv*) The best value of α decreases as λ and K increase.

(*v*) For $\lambda \leq 0.7$, θ^* is essentially the same as $\theta_{\max}$ when $K \geq 5$.

(*vi*) Although α should vary with λ, θ and W do not significantly differ from their best values if α is held fixed while λ varies over a moderately wide interval.

V. BATCH ARRIVALS

In this section we replace the assumption that packets arrive according to a Poisson process with the assumption that packets arrive

Table V—Performance measures vs. K when $\lambda = 0.9$ and $h = 0.01$

K	5	10	15	20	30
$\theta_{\max}$	0.842	0.885	0.895	0.898	0.900
θ	0.771	0.801	0.810	0.813	0.814
$\theta(\alpha_W^*)$	0.771	0.798	0.808	0.811	0.814
α_θ^*	3.0	1.4	0.8	0.6	0.4
α_W^*	4.0	1.8	1.0	0.7	0.4
W^*	3.61	8.05	13.2	18.8	30.5
$W(\alpha_\theta^*)$	3.66	8.12	13.4	18.9	30.5

according to a compound Poisson process. In Fuchs and Jackson,[9] statistical analysis of call arrival times are given. Two of their conclusions are as follows: The exponential distribution is a reasonably good approximation of the time between bursts, and the size of a burst (measured in various ways) has a geometric distribution. The purpose of this investigation is to find out how sensitive the performance measures are to the assumption of Poisson arrivals. We will see that in this model, the throughput can be significantly lower with bursty arrivals than with Poisson arrivals with same rate.

Specifically, we assume that bursts arrive according to a Poisson process with rate λ_b, and the bursts $B_1, B_2, \cdots$ are iid with

$$P\{B_1 = i\} = (1 - \xi)\xi^i, \qquad i = 1, 2, \cdots. \tag{35}$$

It would be nice to interpret this process as one where messages arrive according to a Poisson process with rate λ_b and the jth message consists of a random number of packets with a geometric distribution. Unfortunately, we cannot do that in the context of the model described in Section II. This arrival process does not conform to the finite buffer interpretation of the model because only the current lead packet in each message would compete for the bus; we, on the other hand, assume that all packets in the buffer compete for the bus. This arrival process does not conform to the interpretation that there are K ports because that would require that if the first packet (i.e., the message) finds a free port, then all the packets in the message would enter the system. We assume that only those packets that find a port gain access to the system.

5.1 Details of the arrival process

Let $A(t)$ be the number of arrivals during $(0, T)$. Since $A(t)$ is a compound Poisson process, we know that there are constants M and V, such that

$$E[A(t)] = Mt$$

and

$$\mathrm{Var}[A(t)] = Vt$$

with $V \geq M$. We call M the arrival rate. From eq. (35) we obtain

$$E(B_1) = \frac{1}{1 - \xi} \quad \text{and} \quad \mathrm{Var}(B_1) = \frac{\xi}{(1 - \xi)^2}.$$

Standard calculations yield

$$E[A(t)] = \lambda_b t E(B_1) = \frac{\lambda_b t}{1 - \xi}$$

and

$$\begin{aligned}\mathrm{Var}[A(t)] &= \lambda_b t\,\mathrm{Var}(B_1) + \lambda_b t[E(B_1)]^2 \\ &= \frac{\lambda_b t(\xi+1)}{(1-\xi)^2}.\end{aligned}$$

In an obvious way we obtain

$$\lambda_b = \frac{2M^2}{M+V} \quad \text{and} \quad \xi = \frac{V-M}{M+V}.$$

Letting $z = V/M \geq 1$ yields

$$\lambda_b = \frac{2M}{1+z} \quad \text{and} \quad \xi = \frac{z-1}{z+1}. \tag{36}$$

Equation (36) relates the parameters we might obtain from measurements, M and z, to the parameters of the model, λ_b and ξ. Let

$$b(n,j) = P\{B_1 + B_2 + \cdots B_j = n\};$$

it is the probability that n packets are contained in j bursts. The sum of iid geometric random variables has a negative binomial distribution,

$$b(n,j) = \begin{cases} 0 & \text{if } n < j \\ \binom{n-1}{n-j}(1-\xi)^j \xi^{n-j} & \text{if } n \geq j \end{cases}$$

for all nonnegative integers j and n. We can compute $b(n, j)$ by recursion from

$$\begin{aligned} &b(0,0) = 1, \qquad b(1,1) = 1-\xi, \\ &b(j,j) = (1-\xi)b(j-1,j-1), \qquad j \geq 2, \end{aligned}$$

and

$$b(j+k,j) = \frac{j+k-1}{k}\xi b(j+k-1,j), \qquad k \geq 1 \quad \text{and} \quad j \geq 0.$$

5.2 Changes in the Poisson model

In this section we describe the changes in the equations in Section III that are caused by compound Poisson arrivals.

Let $c_b(j)$ be the probability that j batches arrive in an interval of length v; therefore,

$$c_b(j) = e^{-\lambda_b v}\frac{(\lambda_b v)^j}{j!}. \tag{37}$$

Let

$\bar{d}_n(i,j) = P\{$no collision and n arrivals in a service interval$|j$ batches arrive and start with i in queue$\}$;

then

$$\bar{d}_n(i) = \sum_{j=0}^{n} \bar{d}_n(i,j)c_b(j). \tag{38}$$

Expand on the argument used to obtain eq. (6) to conclude that

$$\bar{d}_n(i,j) = \delta_i \eta_j b(n,j). \tag{39}$$

Similarly,

$$d_n(i) = \sum_{j=0}^{n} (1 - \delta_i \eta_j) b(n,j) c_b(j). \tag{40}$$

The transition probabilities described in Section 3.2 and Appendix A are given in terms of the d and $\bar{d}$ functions, so they do not change form. The infinite sums, used in Section A.1 of Appendix A, have to be recomputed; this is done in Section A.2 of Appendix A.

The steady-state balance equations depend only on the transition probabilities, so everything in Section 3.3 is still valid. No changes are required for the equations in Section 3.4.

Since the arrivals do not form a Poisson process, time-average probabilities need not equal arrival epoch probabilities and we cannot use the arguments in Section 3.5 to obtain the steady-state probability that i packets are present. This means we have not obtained L and W for this model. We can obtain the throughput by using the arguments in Section 3.5.

Let ζ be the ejection rate and $p_0 = \lim_{t\to\infty} P\{X(t) = 0\}$. In the steady state, $\lambda_b p_0$ is the rate at which transitions leave state 0 and $\zeta\pi_0$ is the rate at which transitions enter state 0. Since these rates are equal, we have

$$\zeta = \lambda_b p_0 / \pi_0. \tag{41}$$

Equations (18) through (25) are valid for compound Poisson arrivals when λ is replaced by λ_b. The throughput is obtained from eq. (27).

5.3 Numerical examples

Let θ_z be the throughput when z is the variance to mean ratio. In Tables VI and VII we use the value of α that maximizes θ, and $h = 0.01$.

In Tables VI and VII we see that batch arrivals can significantly degrade throughput, and the degradation increases with z and decreases (except, possibly, for very large λ) with λ.

In Section IV we saw that θ is strongly influenced by α. That observation suggests that the throughput values in Tables I and II might improve if α varied with z. Our numerical experience suggests that, in the vicinity of the best α, θ_z is not sensitive to changes in α.

Table VI—Throughput vs. M for K = 5 and h = 0.01

M	0.1	0.5	0.7	0.9	1.0	2.0	5.0
α	20.0	4.0	4.0	3.0	3.0	2.0	2.0
θ_1	0.100	0.497	0.667	0.771	0.798	0.828	0.819
θ_2	0.069	0.382	0.535	0.663	0.709	0.819	0.813
θ_5	0.035	0.205	0.304	0.403	0.451	0.741	0.804
θ_2/θ_1	0.689	0.769	0.868	0.860	0.888	0.989	0.993
θ_5/θ_1	0.349	0.412	0.456	0.523	0.565	0.895	0.982

Table VII—Throughput vs. M for K = 10 and h = 0.01

M	0.1	0.5	0.7	0.9	1.0	2.0	5.0
α	10.0	1.5	1.6	1.4	1.2	0.9	0.6
θ_1	0.100	0.500	0.692	0.801	0.815	0.821	0.816
θ_2	0.069	0.400	0.590	0.737	0.775	0.816	0.805
θ_5	0.035	0.235	0.371	0.512	0.576	0.798	0.788
θ_2/θ_1	0.690	0.800	0.853	0.920	0.951	0.994	0.987
θ_5/θ_1	0.350	0.470	0.536	0.639	0.707	0.972	0.966

The largest improvement that we could achieve by changing α was 0.003.

VI. COLLISION DETECTION

Suppose that a time units after the vulnerable period ends, the bus is examined for a collision. When a collision is detected, the packet transmission is aborted and that packet is returned to the buffer. Collision detection reduces the time spent transmitting garbage, so it will increase throughput and reduce delays.

6.1 Changes in the model without collision detection

Two changes in the equations in Section III are required to describe collision detection. The first change is in eq. (7). Let

$$\eta_n^a = P\{\text{no outside arrivals prior to } h | n \text{ outside arrivals prior to } a + h\};$$

then

$$\eta_n^a = \left(\frac{a}{a+h}\right)^n, \qquad n = 0, 1, \cdots. \tag{42}$$

Let c_n^a be the probability than n outside arrivals occur in an interval of length $a + h$, i.e.,

$$c_n^a = e^{-\lambda(a+h)} \frac{(a+h)^n}{n!}, \qquad n = 0, 1, \cdots;$$

then

$$d_n^a(i) \triangleq P\{\text{collision and } n \text{ outside arrivals by } a + h | \text{service starts with } i \text{ in queue}\}$$

$$= (1 - \eta_n^a \delta_i) c_n^a, \qquad i, n \geq 0, \tag{43}$$

where δ_i is given by eq. (5).

The equations in Sections 3.2, 3.3, and 3.4 are given in terms of the functions d and $\bar{d}$, so no changes are required. In Section 3.5 we need to change eq. (20) because the expected length of time that a packet is on the bus is not ν. Let

$$p_c(i) = P\{\text{a collision occurs} \mid \text{service starts with } i \text{ in the queue}\}.$$

There are no collisions if, and only if, no arrivals (inside or outside) occur within a time interval of length h. Thus,

$$1 - p_c(i) = \delta_i c_0^a$$

or

$$p_c(i) = 1 - \delta_i c_0^a, \qquad i = 0, 1, \cdots, K - 2. \tag{44a}$$

When $i = K - 1$, no outside arrivals are permitted, so

$$p_c(K - 1) = 1 - \delta_{K-1}. \tag{44b}$$

Let T_i be the transmission time of a packet, given that i packets were present at the last ejection epoch, and let $\nu_i = E(T_i)$; then

$$T_i = \begin{cases} 1 + h \text{ if no collision} \\ a + h \text{ if a collision} \end{cases},$$

and so

$$\begin{aligned} \nu_i &= \frac{\lambda}{i\alpha + \lambda} \{(1 + h)[1 - p_c(i)] + (a + h)p_c(i)\} \\ &\quad + \frac{i\alpha}{i\alpha + \lambda} \{(1 + h)[1 - p_c(i - 1)] + (a + h)p_c(i - 1)\} \\ &= 1 + h - (1 - a) \frac{\beta p_c(i) + i p_c(i - 1)}{i + \beta}, \quad i = 0, 1, \cdots, K - 1. \end{aligned} \tag{45}$$

We replace the ν in eq. (20) by ν_i, which induces some obvious changes in eqs. (22) through (25).

6.2 Numerical examples

In Tables II and III we saw that when $\lambda = 0.7$, the throughput in the vicinity of the best α (0.8) is very close to 0.7; but when $\alpha \leq 0.1$ or $\alpha \geq 3.0$, the throughput can be much smaller than 0.7. With collision detection, with $a = 0.02$, throughputs of about 0.7 can be achieved with α as large as 5.0. Collision detection reduces the average waiting time by about one-third. This example suggests that when a throughput close to the maximum possible can be achieved without collision

Table VIII—Performance measures vs. λ when $K = 20$, $h = 0.01$, and $a = 0.02$

λ	0.9	1.0	2.0	3.0
α	4.5	3.0	2.5	2.5
θ	0.891	0.935	0.943	0.942
W	8.53	14.9	20.6	20.9
n_c	0.718	0.673	0.628	0.619
ϕ	0.910	0.958	0.969	0.970

Table IX—Performance measures without collision detection

λ	0.9	1.0	2.0	3.0
α	4.5	3.0	2.5	2.5
θ	0.434	0.569	0.606	0.602
W	44.2	33.3	32.4	32.9
n_c	0.444	0.585	0.625	0.620
ϕ	0.989	0.983	0.980	0.981

detection, collision detection will lower the average waiting time and make the throughput less sensitive to α.

Now we investigate the effects of collision detection when the throughput is significantly smaller than the arrival rate.

By comparing Tables III and VIII we can see the effects of collision detection. Throughput increases significantly for each λ and W decreases. An indirect effect of collision detection is that larger values of α are best. This reduces the time spent waiting for a retry, makes the probability of no collision smaller, and yields a larger occupancy for the bus. If the values of α shown in Table VIII were used without collision detection, performance would degrade significantly, as shown in Table IX. From the last two rows of Table IX we deduce that the bus wastes a lot of time transmitting packets that have been destroyed.

6.3 Acknowledgment

I would like to thank S. Halfin for his valuable observations.

REFERENCES

1. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part I—Carrier Sense Multiple Access Modes and Their Throughput—Delay Characteristics," IEEE Trans. on Commun., *23,* No. 12 (December 1975), pp. 1400–16.
2. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part IV—Stability Considerations and Dynamic Multiple Control in Carrier Sense Multiple Access," IEEE Trans. on Commun., *25,* No. 10 (October 1977), pp. 1103–19.
3. S. Halfin, unpublished work.
4. F. A. Tobagi and V. B. Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Proc. of the LACN Symp., (May 1979), pp. 217–44.
5. S. S. Rappaport, "Demand Assigned Multiple Access Systems Using Collision Type Request Channels: Traffic Capacity Comparisons," IEEE Trans. on Commun., *27,* No. 9 (September 1979), pp. 1325–31.

6. S. S. Rappaport and S. Bose, "Demand-Assigned Multiple-Access Systems Using Collision-Type Request Channels: Stability and Delay Considerations," IEEE Proc., *128,* Pt. E, No. 1 (January 1981), pp. 37–43.
7. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research,* Vol. I, New York: McGraw-Hill, 1982.
8. M. Kaplan, "A Sufficient Condition for the Nonergodicity of a Markov Chain," IEEE Trans. on Information Theory, *25,* No. 4 (July 1979), pp. 470–1.
9. E. Fuchs and P. E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models," Commun. ACM, *13,* No. 12 (December 1970), pp. 752–7.

APPENDIX A

Transition Probabilities

In Appendix A we record the transition probabilities omitted from the text.

A.1 Transition probabilities for Section 3.2

Let

$$S_{\bar{d}}(j, i) = \sum_{n=j}^{\infty} \bar{d}_n(i)$$

and

$$S_d(j, i) = \sum_{n=j}^{\infty} d_n(i).$$

The former is the probability that more than $j - 1$ packets arrive in a service interval and there are no collisions when i packets are in the buffer at the start of the service interval. The latter is the corresponding probability when there is a collision. For computational purposes it is important to represent these infinite sums as finite sums. Observe that

$$S_{\bar{d}}(j, i) = S_{\bar{d}}(0, i) - \sum_{n=0}^{j-1} \bar{d}_n(i), \qquad j \geq 1,$$

and

$$S_d(j, i) = S_d(0, i) - \sum_{n=0}^{j-1} d_n(i), \qquad j \geq 1;$$

then

$$S_{\bar{d}}(0, i) = \sum_{n=0}^{\infty} \bar{d}_n(i) = \delta_i \sum_{0}^{\infty} \left(\frac{\nu - h}{\nu}\right)^n e^{-\lambda\nu} \frac{(\lambda\nu)^n}{n!} = \delta_i e^{-\lambda h}. \tag{46}$$

Since $S_{\bar{d}}(0, i) + S_d(0, i) = 1$ is easily established,

$$S_d(0, i) = 1 - e^{-\lambda h}; \tag{47}$$

this equation also can be obtained by calculating the sum explicitly.

To modify eq. (9) when $i = 0$ we have to delete terms with negative arguments. When $i + n = K - 1$ we need to recognize that state $K - 1$ is reached if $n \geq K - 1 - i$ packets arrive and there is no collision. When $i + n = K$, we need to recognize that state K is reached if, and only if, at least $K - 1 - i$ packets arrive and there is a collision.

The following transition probabilities are obtained:

$$p_{0,0} = d_0(0)$$

$$p_{0,j} = \bar{d}_j(0) + d_{j-1}(0) \qquad 1 \leq j \leq K-2$$

$$p_{0,K-1} = \sum_{K-1}^{\infty} \bar{d}_n(0) + d_{K-2}(0) = S_{\bar{d}}(K-1, 0) + d_{K-2}(0)$$

$$p_{0,K} = \sum_{K-1}^{\infty} d_n(0) = S_d(K-1, 0)$$

$$p_{1,0} = \frac{1}{1+\beta}\bar{d}_0(0)$$

$$p_{1,1} = \frac{\beta}{1+\beta}\bar{d}_0(1) + \frac{1}{1+\beta}[\bar{d}_1(0) + d_0(0)]$$

$$p_{1,j} = \frac{\beta}{1+\beta}[\bar{d}_{j-1}(1) + d_{j-2}(1)] + \frac{1}{1+\beta}[\bar{d}_j(0) + d_{j-1}(0)], \qquad 2 \leq j \leq K-2$$

$$p_{1,K-1} = \frac{\beta}{1+\beta}[S_{\bar{d}}(K-2, 1) + d_{K-3}(1)] + \frac{1}{1+\beta}[S_{\bar{d}}(K-1, 0) + d_{K-2}(0)]$$

$$p_{1,K} = \frac{\beta}{1+\beta}S_d(K-2, 1) + \frac{1}{1+\beta}S_d(K-1, 0).$$

For $2 \leq i \leq K-2$,

$$p_{i,i-1} = \frac{i}{i+\beta}\bar{d}_0(i-1)$$

$$p_{i,i} = \frac{\beta}{i+\beta}\bar{d}_0(i) + \frac{i}{i+\beta}[\bar{d}_1(i-1) + d_0(i-1)]$$

$$p_{i,j} = \frac{\beta}{i+\beta}[\bar{d}_{j-i}(i) + d_{j-i-1}(i)] + \frac{i}{i+\beta}[\bar{d}_{j-i+1}(i-1) + d_{j-i}(i-1)], \qquad i+1 \leq j \leq K-2,$$

$$p_{i,K-1} = \frac{\beta}{i+\beta}[S_{\bar{d}}(K-1-i, i) + d_{K-2-i}(i)] + \frac{i}{i+\beta}[S_{\bar{d}}(K-i, i-1) + d_{K-1-i}(i-1)]$$

$$p_{i,K} = \frac{\beta}{i+\beta} S_d(K-i-1, i) + \frac{i}{i+\beta} S_d(K-i, i-1)$$

$$p_{K-1,K-2} = \frac{K-1}{K-1+\beta} \bar{d}_0(K-2)$$

$$p_{K-1,K-1} = \frac{\beta}{K-1+\beta} S_{\bar{d}}(0, K-1) + \frac{K-1}{K-1+\beta}[S_{\bar{d}}(1, K-2) + d_0(K-2)]$$

$$p_{K-1,K} = \frac{\beta}{K-1+\beta} S_d(0, K-1) + \frac{K-1}{K-1+\beta} S_d(1, K-2)$$

$$p_{K,K-1} = \delta_{K-1}$$

$$p_{K,K} = 1 - \delta_{K-1}.$$

A.2 The effect of batch arrivals

Equations (46) and (47) have to be modified for the compound arrival process. From eqs. (38) and (39) we obtain

$$\begin{aligned} S_{\bar{d}}(0, i) &= \sum_{n=0}^{\infty} \bar{d}_n(i) = \sum_{n=0}^{\infty} \sum_{j=0}^{n} \bar{d}_n(i, j) c_b(j) \\ &= \sum_{j=0}^{\infty} c_b(j) \sum_{n=j}^{\infty} \bar{d}_n(i, j) \\ &= \sum_{j=0}^{\infty} c_b(j) \sum_{n=j}^{\infty} \delta_i \eta_j b(n, j) \\ &= \delta_i \sum_{j=0}^{\infty} c_b(j) \sum_{n=j}^{\infty} \left(\frac{\nu - h}{\nu}\right)^j b(n, j). \end{aligned} \tag{48}$$

To evaluate the double sum, let

$$a_n = \sum_{j=0}^{n} b(n, j) c_b(j);$$

it is the probability that n customers arrive during a service interval. Let

$$\hat{A}(z) = \sum_{n=0}^{\infty} z^n a_n = \sum_{n=0}^{\infty} z^n \sum_{j=0}^{n} b(n, j) c_b(j). \tag{49}$$

Comparing eqs. (48) and (49) we see that the double sum in eq. (48) is $\hat{A}(\cdot)$ evaluated at $z = (\nu - h)/\nu$.

Standard generating-function arguments yield

$$\hat{A}(z) = \hat{C}[\hat{B}(z)], \tag{50}$$

where

$$\hat{B}(z) = \sum_{k=0}^{\infty} z^k P\{B_1 = k\} = \frac{z(1-\xi)}{1-\xi z}$$

and

$$\hat{C}(z) = \sum_{j=0}^{\infty} z^j c_b(j) = e^{-\lambda_b \nu(1-z)}.$$

Substitution into eq. (50) yields

$$\hat{A}(z) = \exp\left(\lambda_b \nu \frac{z-1}{1-z\xi}\right). \tag{51}$$

Evaluating eq. (51) at $z = (\nu - h)/\nu$ and substituting the result into eq. (48) yields

$$S_{\bar{d}}(0, i) = \delta_i \exp\left[\frac{-\lambda_b \nu h}{\nu - (\nu - h)\xi}\right].$$

As before,

$$S_d(0, i) = 1 - S_{\bar{d}}(0, i)$$

is obtained easily.

APPENDIX B

List of Symbols

The following is a list of symbols and their definitions as used in this paper.

- α = retry rate
- h = one-way propagation delay
- K = system capacity in packets
- λ = Poisson arrival rate
- M = compound Poisson arrival rate
- n_c = $P\{\text{no collision}\}$
- ν = service time (constant)
- ϕ = proportion of time bus is occupied
- θ = throughput
- $\theta_{\max}$ = throughput of an $M/D/1/K$ queue
- W = average wait (in queue plus in service) of a packet
- z = variance to mean ratio of compound Poisson process

ζ = ejection rate of packets from the bus.

An asterisk on a symbol means the best value found. The subscripts *lb*, *ub*, and avg stand for lower bound, upper bound, and average, respectively.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Implementing and Testing New Versions of a Good, 48-Bit, Pseudo-Random Number Generator

By C. S. ROBERTS

(Manuscript received October 27, 1981)

*In this paper we describe the design, testing, and use of drand48—a good, pseudo-random number generator based upon the linear congruential algorithm and 48-bit integer arithmetic. The drand48 subroutine is callable from C-language programs and is available in the subroutine library of the UNIX * operating system. Versions coded in assembly language now exist for both the PDP-11 and VAX-11 computers; a version coded in a "portable" dialect of C language has been produced by Rosler for the Western Electric 3B20 and other machines. Given the same initialization value, all these versions produce the identical sequence of pseudo-random numbers. Versions of drand48 in the assembly language of other computers or for other programming languages clearly could be implemented, and some output results have been tabulated to aid in testing and debugging such newly coded subroutines. Timing results for drand48 on the PDP-11/45, the PDP-11/70, the VAX-11/750, and the VAX-11/780 are also presented and compared.*

I. INTRODUCTION

The work described in this paper arose when one day the author found himself in need of a good, pseudo-random number generator that would execute and produce identical results on two different computers (in this case, the 16-bit PDP-11 and the 32-bit VAX-11, both manufactured by Digital Equipment Corporation). Good, pseudo-random number generators often require multiple-precision arithmetic; hence, to achieve speed they are usually implemented in assembly language and are dependent upon the word length of the computer. If

* *UNIX* is a trademark of Bell Laboratories.

this is not enough of a barrier to portability, add the fact that the recoding of a pseudo-random number generator for a different computer is an error-prone endeavor. The author has himself found bugs in several pseudo-random number subroutines that were supposedly "correctly" coded. (The author admits to being facetious, but do such bugs make the output from a generator more or less random?)

This paper does not present a magical method to eliminate these difficulties. However, it does present enough data and intermediate results so that a person may code on any computer a good generator based upon 48-bit integer arithmetic and then begin to test the new version for bugs. In addition, we describe proper usage of routines drand48, lrand48, mrand48, erand48, nrand48, and jrand48—C-language callable functions to generate pseudo-random numbers—currently available at Bell Laboratories in the subroutine library of the *UNIX** operating system for the Western Electric 3B20, the PDP-11, the VAX-11, and other computers.

The pseudo-random number generator considered in this paper is based upon the well known linear congruential algorithm, e.g., see Section 3.2.1 of Knuth.[1] The next number in the pseudo-random sequence is generated according to the formula

$$X_{n+1} = (aX_n + c)_{\text{mod } m} \quad n \geq 0.$$

We choose the value of m to be 2^{48}; hence, 48-bit integer arithmetic is required. The values of the multiplier a and the addend c are chosen as follows:

$$a = 5\text{DEECE66D}_{16} = 273673163155_8$$

$$c = \text{B}_{16} = 13_8.$$

While other equally good choices for m are clearly possible, we chose 2^{48} for the following reasons, based upon matters of taste and judgment. The word length of many popular computers is a multiple of 16 bits; hence, it seemed wise and convenient to choose $\log_2 m$ to be an integer multiple of 16. A period of 2^{48} for the generated pseudo-random sequence seemed long enough for most purposes, while 2^{32} seemed dangerously short. Assuming it takes 10^{-4} second to generate a pseudo-random number (see the timing results in Section IV), a period of 2^{48} corresponds to 893 years, while 2^{32} would only require 119 hours for the complete cycle to be generated. The argument for 48 bits becomes even more compelling if we assume that a future processor or customized hardware might be a factor of 10 or 100 faster.

The value of a, chosen above, was one of the multiplier values judged by Coveyou and MacPherson[2] to be satisfactory according to the "spectral test," one of the most demanding of the canonical tests

for overall pseudo-random number-generator quality. The above values of a, c, and m ensure that the generated pseudo-random sequence will have the maximum possible period, i.e., 2^{48} (see Theorem A, Section 3.2.1.2 of Knuth[1]). The testing done by Coveyou and MacPherson[2] was demanding enough and the properties of pseudo-random sequences based upon the linear congruential algorithm are well enough understood[1] that further statistical testing of the generator was deemed unnecessary.

The linear congruential algorithm with these choices for the parameters yields a generator that exceeds the requirements of most users and is much better than many generators in common use today. Obviously, this is not the "optimum" pseudo-random number generator; it will not meet all the requirements of any potential user, and the author makes no such claims. However, in more than four years of use in numerous programs at Bell Laboratories, it has served well, even in situations where other generators have failed.

II. PSEUDO-RANDOM GENERATOR RESULTS

The pseudo-random sequence generator described and specified in Section I was coded in assembly language three times—once on the PDP-11 using the 16-bit integer arithmetic instructions "mul" and "add,"[3] once on the PDP-11 using the 64-bit floating-point arithmetic instructions "mulf" and "addf,"[3] and once on the VAX-11 using the 32-bit integer extended multiply and add instruction "emul."[4] After a bit of debugging, all three independent versions finally produced the identical sequence of 48-bit pseudo-random integers. For purposes of comparison with future newly coded subroutines, a portion of that sequence is presented below (in hexadecimal) as Table I.

Since it is common practice to treat the output from a pseudo-random number generator as a pure binary fraction, thus yielding a uniform distribution over the interval 0 to 1, the numbers in Table I were so treated and then multiplied by 4096 to yield the decimal

Table I—A portion of the pseudo-random sequence X_i of 48-bit integers

1234ABCD330E	657EB7255101	D72A0C966378	5A743C062A23	72534ABF62F2
5195D97A8D15	E2ECF94AEFFC	03FD3CD49657	9586EFCA2D16	28CC61DEF669
623B341D40C0	B0E5A9A111CB	0F1160B4F57A	E65CDA1020FD	29DE25BD59C4
28B8E8F5507F	8876EDD9601E	9AA93190E0D1	952BC3577F08	451CD3C24673
63F661075102	4B1C4CBD49E5	BE0C7218348C	4C6C2C9427A7	135676A8EC26
67ACF11EB039	DB7D1EF03E50	F124D606681B	A9AF4526958A	D8B2A2FFA7CD
00B48E98A054	765E7C77BBCF	8858368AF12E	C9B2484004A1	43FF29D69E98
FB95A6FE16C3	4E897866E312	99D1A468DAB5	9BD4C9FFBD1C	3662639AACF7

Table II—Pseudo-random integers Y_i corresponding to the X_i in Table I

291	1623	3442	1447	1829	1305	3630	63	2392	652
1571	2830	241	3685	669	651	2183	2474	2386	1105
1599	1201	3040	1222	309	1658	3511	3858	2714	3467
11	1893	2181	3227	1087	4025	1256	2461	2493	870
3628	1247	622	1383	1587	2636	3086	2472	2177	1881
2672	1340	3876	1507	3866	30	2115	1117	99	2424
839	3595	243	1068	1240	3651	2040	2908	1173	3542
2767	1877	3930	3173	1542	936	1452	1230	2743	2944

integers in Table II. More precisely, each integer Y_i in Table II was computed from the corresponding value of X_i in Table I by the formula

$$Y_i = \lfloor (X_i)(2^{-48})(4096) \rfloor = \lfloor 2^{-36} X_i \rfloor.$$

For some purposes, Table II may be more convenient or appropriate than Table I.

Tables I and II should be valuable to a person attempting to code and test a new version of the generator subroutine. After initializing with $X_0 = 1234ABCD330E_{16}$, the new version, if correct, should reproduce Table I (or Table II). The fact that three independent codings of the 48-bit linear congruential algorithm gave the same results gives the author substantial confidence that Tables I and II are indeed accurate. In addition, Lawrence Rosler of Bell Laboratories has independently verified Tables I and II using both C- and assembly-language code he produced for the Honeywell H6000 series of 36-bit computers.[5]

The author has found that an empirical but effective heuristic for quickly evaluating the general quality of any pseudo-random number generator is to display the output bits on a bilevel display device, and this has been done in Fig. 1 for the generator defined in Section I of this paper. (A bilevel display device consists of an $n \times m$ rectangular array of dots, each of which can be made either white or black.) Figure 1 was generated using only the leftmost 32 bits of each 48-bit X_i; 2^{13} values of X_i were generated, thus giving a total of 2^{18} bits. These bits were displayed on a 512 × 512 CRT display device with a 1 being displayed as white and a 0 being displayed as black. Good quality pseudo-random number generators produce displays similar in appearance to Fig. 1—random salt and pepper effect with no visible patterning. For comparison, Fig. 2 shows the display produced by an inferior generator.* Note the prominent nonrandom patterns visible in

* The author apologizes for the generally poor quality of Figs. 1 and 2 and assures the reader that the result is much more striking and apparent when the original CRT or photographic plates are viewed. Glossy prints such as these simply do not reproduce well.

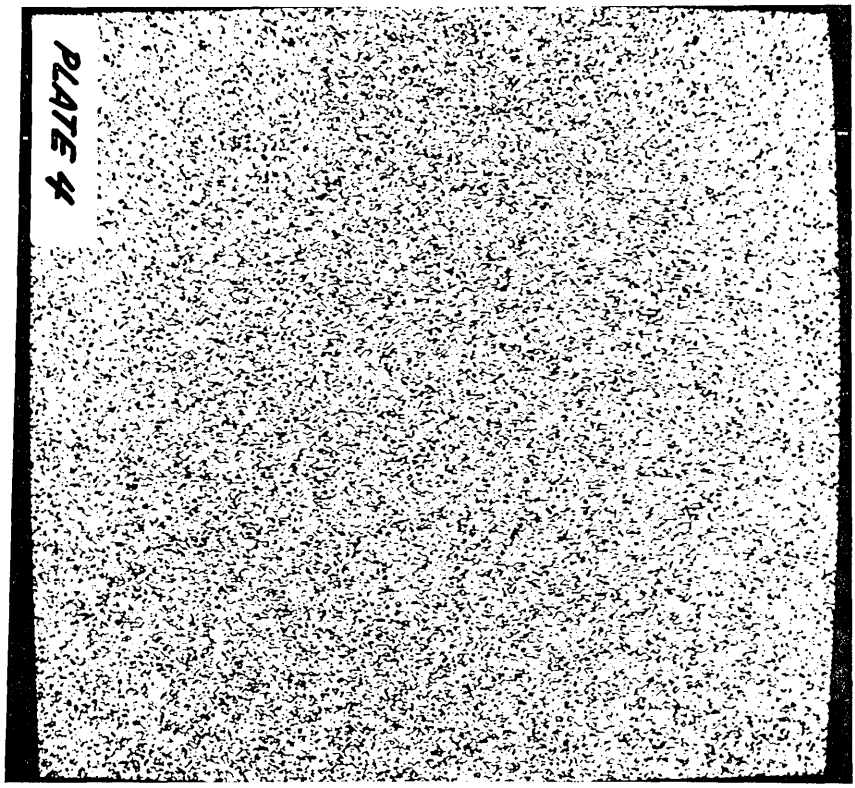


Fig. 1—Bilevel dot display generated by drand48.

Fig. 2, which was produced by a 1977 version of the library routine "rand," described in Section III of the *UNIX Programmer's Manual.* (This 1977 version used the linear congruential algorithm with $m = 2^{15}$; the version of "rand" currently in the UNIX library (1982) is better since it uses $m = 2^{31}$.)

III. SUBROUTINES FOR USE WITH THE C PROGRAMMING LANGUAGE

As an example of how the calling sequences to the pseudo-random number generator might be designed, we now describe some routines that were coded for use with the C programming language.[6] So far the author has himself implemented versions of these routines only for the PDP-11 and VAX-11 computers. However, a version for other computers that support the C language has been implemented by Rosler[7] and is now in use on the Western Electric 3B20, the Honeywell H6000, the IBM-370, and the Motorola MC68000 computers.

The pseudo-random number generator was implemented in assembly language as one subroutine having nine entry points. Six of the

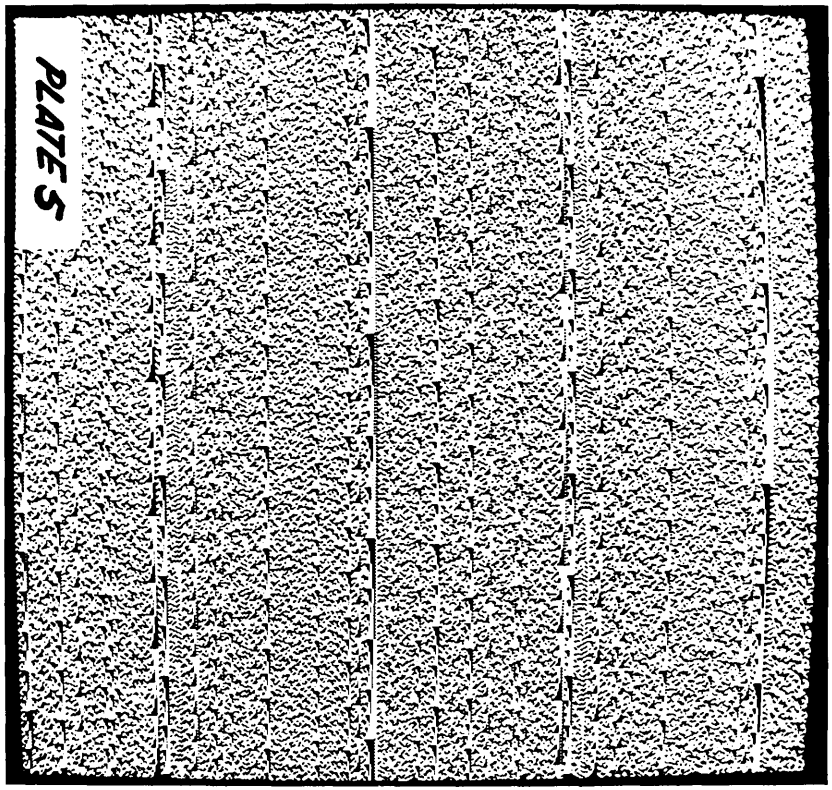


Fig. 2—Bilevel dot display generated by an inferior pseudo-random number generator.

entry points generate the next pseudo-random X_i in the sequence and then convert the leftmost bits of it into the particular type of data item desired—floating-point fraction, positive integer, or signed integer. The entry points erand48, nrand48, and jrand48 can be called without first having to invoke a special initialization entry point. Calls to the entry points drand48, lrand48, and mrand48 should be preceded by at least one call to one of the initialization entry points—either srand48, seed48, or lcong48.

3.1 Function drand48

Example C usage:

```
double fnext, drand48( );
fnext = drand48( );
```

The next pseudo-random number is returned as a non-negative binary fraction in double precision floating-point format; i.e., the value returned is $2^{-48}X_i$. The values returned are uniformly distributed over the interval [0, 1). Thus, in the above C-code example, the value range

for fnext is $0 \leq$ fnext < 1. Either srand48, seed48, or lcong48 should be invoked before calling drand48.

3.2 Function lrand48

Example C usage:

```
long int lnext, lrand48( );
lnext = lrand48( );
```

The next pseudo-random number is returned as a non-negative integer in long integer format. The long integer is formed by taking the leftmost 31 bits of X_i, i.e., the value returned is $\lfloor 2^{-17} X_i \rfloor$. In the above C-code example, the value range for lnext is $0 \leq$ lnext $< 2^{31}$. Either srand48, seed48, or lcong48 should be invoked before calling lrand48.

3.3 Function mrand48

Example C usage:

```
long int mnext, mrand48( );
mnext = mrand48( );
```

The next pseudo-random number is returned as a signed integer in long integer format. The long integer is formed by taking the leftmost 32 bits of X_i to be a signed integer in two's-complement format. Hence, the leftmost bit of X_i determines the sign of the output value. In the above C-code example, the value range for mnext is $-2^{31} \leq$ mnext $< 2^{31}$. Either srand48, seed48, or lcong48 should be invoked before calling mrand48.

3.4 Functions erand48, nrand48, and jrand48

These functions are identical to drand48, lrand48, and mrand48, respectively, in the characteristics of the data value returned. The difference is that erand48, nrand48, and jrand48 allow, and require, the calling program to provide the storage for the current 48-bit X_i value, while drand48, lrand48, and mrand48 provide this storage internally for themselves. For those programs that require only a single stream of pseudo-random numbers, drand48, lrand48, and mrand48 are a little more convenient and simpler to use. However, erand48, nrand48, and jrand48 allow multiple "independent" streams of pseudo-random numbers to be generated, i.e., subsequent numbers in each stream will not depend upon how many times the routines are called by other parts of a program to generate numbers for other streams. This property can be a big asset for certain statistical computations and for program debugging.

3.4.1 Function erand48

Example C usage:

```
double fnext, erand48( );
```

```
short int xsubi[3];
fnext = erand48(xsubi);
```

3.4.2 Function nrand48

Example C usage:

```
long int lnext, nrand48( );
short int xsubi[3];
lnext = nrand48(xsubi);
```

3.4.3 Function jrand48

Example C usage:

```
long int mnext, jrand48( );
short int xsubi[3];
mnext = jrand48(xsubi);
```

3.5 Function srand48

Example C usage:

```
long int seedval;
seedval = 0x1234ABCD;
srand48(seedval);
```

This is an initialization entry point that sets the value of X_0; the multiplier a and the addend c are set to the values specified in Section I. The leftmost 32 bits of X_0 are taken from the argument passed to srand48 when it is called (seedval in the above C-code example). The rightmost 16 bits of X_0 are arbitrarily set to $330E_{16}$. Hence, the above C-code example sets the value of X_0 to $1234ABCD330E_{16}$.

3.6 Function seed48

Example C usage:

```
short int seed16v[3], *shp, *seed48( );
seed16v[0] = 0x330E;
seed16v[1] = 0xABCD;
seed16v[2] = 0x1234;
shp = seed48(seed16v); /* pointer to previous Xi stored in shp */
        /* or alternatively, */
seed48(seed16v); /* pointer to previous Xi just ignored */
```

This is an initialization entry point that sets the value of X_0 to the 48 bits specified by the argument passed to seed48; the multiplier a and the addend c are set to the values specified in Section I. In addition, the previous value of X_i is automatically stored in a 48-bit internal buffer, used only by seed48, and the value returned is a pointer to this buffer. The argument is an array of three 16-bit integers. The above C-code example sets the value of X_0 to $1234ABCD330E_{16}$.

The pointer to the previous value of x_i is useful if a restart from that point is desired at a later time. The 48-bit X_i value must be copied out

of the internal buffer before seed48 is called again or it will be destroyed. The following code sequence, for example, restarts a program with a saved value of X_i.

```
short int newx[3], oldx[3], *shp, *seed48( ), i;
shp = seed48(newx); /* initialize with whatever is in newx */
for (i = 0; i < 3; i++) oldx[i] = shp[i]; /* save previous Xi in
oldx */
    ....
    ....
    ....
seed48(oldx); /* reinitialize with oldx */
```

3.7 Function lcong48

Example C usage:

```
short int param[7];
param[0] = 0x330E; param[1] = 0xABCD; parem [2] = 0x1234;
param[3] = 0xE66D; param[4] = 0xDEEC; param[5] = 0x5;
param[6] = 0xB;
lcong48(param);
```

This is an initialization entry point that sets the values of X_0, a, and c; hence, different 48-bit linear congruential generators may be created by specifying different values for the multiplier a and the addend c. The argument passed to lcong48 is an array of seven 16-bit integers. The first three specify a 48-bit value of X_0; the next three specify a 48-bit value of the multiplier a, and the last one specifies a 16-bit value of the addend c. Hence, the above C-code examples set $X_0 = 1234ABCD330E_{16}$, $a = 5DEECE66S_{16}$, and $c = B_{16}$.

IV. TIMING RESULTS

Table III presents the time required to generate, using function drand48, 10^6 pseudo-random numbers on five different computer hardware configurations. More precisely, Table III gives the time required to execute the following short C-language program:

```
main( ) {
register int i, j, h;
```

Table III—Time required to generate 10^6 pseudo-random numbers

Computer	Time (sec)
PDP-11/45	440
PDP-11/45 with Fabritek cache	340
PDP-11/70	162
VAX-11/750	200
VAX-11/780	96

```
double nfd, drand48( );
int lli, llj;
short int nn[500];
long int seedval;
seedval = 0x1234ABCD; srand48(seedval);
for(i = 0; i < 500; i++) nn[i] = 0;
lli = 1000; llj = 1000;
nfd = 500;
for(i = 0; i < lli; i++) for(j = 0; j < llj; j++)
  {h = nfd*drand48( );
    nn(h)++;
  }
}
```

Execution of drand48 accounts for 80 percent, approximately, of the times listed in Table III. For the timing tests on the PDP-11/45 and PDP-11/70, our version of drand48 that employs the floating-point arithmetic instructions was used, since it is substantially faster than the version that employs the integer arithmetic instructions. The first three entries in Table III represent separate executions of the same binary machine code; hence, the time differences reflect brute-force speed differences of the processor and/or memory hardware. The same is true for the last two entries in Table III. The comparison between the PDP-11 and the VAX-11 is, however, more subtle since a complete recoding of drand48 is involved here. Essentially, when comparing the VAX-11 with the PDP-11 in Table III, we are comparing the time required to accomplish the identical "function" on both computers using a near optimally coded version of drand48 on each machine. The faster time for the VAX-11/780 must be attributed to at least two factors: a more powerful and capable instruction set than the PDP-11, and faster basic hardware.

V. DISCUSSION

We have described the design of drand48, a good, pseudo-random number-generator subroutine, and presented enough output data so that this subroutine can be recoded for a new processor and quickly tested for bugs. While the reproduction by a new drand48 implementation of either Table I or II surely is not a logically complete test of correctness, it is, at least, a first-order indication that will catch many of the common types of programming errors. The drand48 subroutine is not a panacea, an "ultimate" generator for all purposes. There are numerous other algorithms for generating pseudo-random sequences, and each one has its advantages, disadvantages, advocates, and detractors. The most sophisticated requirements will always have to be met by custom design of the generator and extensive statistical testing.

However, it is not at these most sophisticated users that drand48 is aimed.

We live in an age when new computers are coming into existence very rapidly. The *UNIX* operating system and the C programming language are already running on the Western Electric 3B20 processor, and the 3B05 is not far behind. Other new processors have been or are being introduced by other manufacturers. As existing programs migrate to these new machines, it would be desirable for the associated pseudo-random number generator to continue to produce the same output sequences. A subroutine written in a portable, high-level programming language is one possible solution, and drand48 has proven itself amenable to such an approach. However, for applications in which speed is of paramount importance, subroutines written in assembly language are useful. (The implementation of drand48 in the "portable" dialect of C ran significantly slower[7] than the assembly language versions described in this paper.) The drand48 subroutine has worked well on the PDP-11 and VAX-11 computers; the author hopes that the information provided in this paper will make it possible for new implementations of drand48 and its variants to be accurately coded for at least some of the new computers that will ultimately become popular in the future.

VI. ACKNOWLEDGMENT

The author wishes to thank Lawrence Rosler for providing him with information on the portable C version of drand48 and for his helpful comments on preliminary drafts of this paper.

REFERENCES

1. D. E. Knuth, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, Reading, MA: Addison-Wesley, 1969.
2. R. R. Coveyou and R. D. MacPherson, "Fourier Analysis of Uniform Random Number Generators," J. Assoc. Comp. Mach., *14*, No. 1 (January 1967), pp. 100–19.
3. *PDP-11/70 Processor Handbook*, Maynard, MA: Digital Equipment Corporation, 1976.
4. *VAX-11/780 Architecture Handbook*, Vol. 1, Maynard, MA: Digital Equipment Corporation, 1977.
5. L. Rosler and N-P. Nelson, unpublished work.
6. B. W. Kernighan and D. M. Ritchie, *The C Programming Language*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
7. L. Rosler, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Some Extremal Markov Chains

By J. E. MAZO

(Manuscript received March 10, 1982)

Using a Markov chain model for the motion of a particle through a V-node network, we consider the quantities n_{ij}, which are the average number of steps taken by the particle in traveling from an originating node, i, to a destination node, j. A figure-of-merit, N, for the entire network is introduced by averaging n_{ij} over i and j. We investigate which networks minimize or maximize N, either when no restriction is placed on the Markov chain, or when we restrict it so that it corresponds to random routing. By the latter we mean that at each node the particle "selects at random" lines from an undirected network graph. We show that for random routing, the complete graph has $N = (V - 1)$ and is the minimizing graph. The maximizing graph is unknown, but we establish that the worst behavior of N increases at least with the cube of the number of vertices, but no worse than the 3.5 power. Properties of the class of graphs known as barbells are useful here. The minimizing unrestricted chain corresponds to placing the nodes on a circle and proceeding unidirectionally from one node to the next. Here, $N = V/2$.

I. INTRODUCTION AND RESULTS

Our work is set against the background of the Markov chain model for the movement of a "particle" or "message" through a network of V nodes. Thus, suppose a particle originates at node i and is destined for node $j \neq i$. The particle wanders through the network toward its destination via a Markov chain, going to node n from node m with probability p_{mn}. The quantity p_{mn} is the (m, n) element of the transition matrix P of the Markov chain whose states correspond to the nodes of the network. Denote the average number of steps required by the particle to reach its destination by n_{ij}, and assume that any node is accessible from any other (the Markov chain is irreducible). Then we introduce the figure-of-merit, N, for any such chain by averaging n_{ij} over the $(V - 1)$ possible destinations and V origins of particles:

$$N \equiv \frac{1}{V(V-1)} \sum_{i=1}^{V} \sum_{\substack{j=1 \\ j \neq i}}^{V} n_{ij}. \tag{1}$$

Our problem is: Which chains minimize or maximize N, either for a random-walk Markov chain or for an unrestricted chain?

An unrestricted chain means that no restrictions are placed on the transition matrix P other than irreducibility and the requirement that $p_{ii} = 0$ for all i. The random-walk chain is a special case of interest and is defined as follows. Draw any connected undirected graph on the V nodes. If m and n are not joined by an edge of this graph set $p_{mn} = 0$, while if they are so joined set $p_{mn} = 1/l_m$, where l_m is the number of edges of the graph leaving node m. Thus, at any node, the particle chooses from the available lines "at random." This random walk on the graph has previously been used by Kleinrock in Ref. 1, where it is referred to as random routing.

Our results are:

(*i*) For random routing, the average number of steps N is minimized when the graph of the network is the complete graph. For this case $N = (V - 1)$.

(*ii*) For the unrestricted chain, N is minimized when the nodes are placed consecutively on a circle and we proceed deterministically from 1 to 2 to 3, etc. Here $N = V/2$.

(*iii*) By choosing $p_{12} = p_{21} = 1 - \epsilon$, $\epsilon \to 0$, N can be arbitrarily large for the unrestricted chain (when $V > 2$).

(*iv*) We have not been able to determine the "worst" graph for random routing, but we can show for large V that $O(V^3) \leq N_{\text{worst}} \leq O(V^{3.5})$. The barbell graphs of Mitra-Weiss[2] and Landau-Odlyzko[3] are good candidates for bad graphs.

II. THE MINIMAL WALK

In this section, we demonstrate that the complete graph is the only graph which minimizes N for the random-walk problem.

The fact that the symmetry of the complete graph requires $n_{ij} = N$ for all $i \neq j$ allows a simple demonstration of the fact $N = (V - 1)$ for this case. If the particle originates at node i, we go directly to our destination j with probability $1/(V - 1)$, requiring only one step, or we go to another node with probability $(V - 2)/(V - 1)$ and then require an average of $(1 + N)$ steps to reach j from i. Thus,

$$N = \frac{1}{V-1} + \frac{V-2}{V-1}[1 + N]. \tag{2}$$

Solving (2) yields $N = (V - 1)$.

To show $N > (V - 1)$ for other graphs is more involved. We first

give properties of those transition matrices that correspond to random routing, ending with (12) which gives the stationary probabilities for those chains. We then derive (24), which is an expression for the average of the first passage times with which we are concerned. Finally, we obtain our result by giving a lower bound to (24).

Standard results on Markov chains or positive matrices may be used without reference when needed. For the former, the reader may consult Ref. 4, while Ref. 6 is a useful source for matrices.

In this paper, we denote transposition, complex conjugation, and hermitian conjugate by the symbols T, $*$, and $\dagger$, respectively.

For random routing, the transition matrix P is given by

$$P = DA, \tag{3}$$

where A is the symmetric adjacency matrix of the graph and is defined by

$$a_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } i \text{ and } j \text{ are joined by an edge} \\ 0 & \text{otherwise,} \end{cases}$$

and D is a diagonal matrix with diagonal elements

$$d_{ii} = 1/l_i, \tag{4}$$

l_i being the ith row sum of A and also, therefore, the number of edges of the graph incident on node i. The matrix P is stochastic, that is, it has nonnegative elements and the rows sum to one. Further, the assumed connectivity of the network graph implies that P is irreducible. The facts that P is stochastic and irreducible imply that the largest positive eigenvalue of P is unity and has multiplicity one. All other eigenvalues of P have modulus less than, or equal to, one. Set $\lambda_1 = 1$ and let $\lambda_2, \cdots, \lambda_V$ be the remaining eigenvalues of P.

We now investigate the eigenvalue and eigenvector structure of P. The matrices $P = DA$ and $Q = D^{1/2}AD^{1/2}$ differ by a similarity transformation ($P = D^{1/2}QD^{-1/2}$) and so have the same eigenvalues. Since Q is real symmetric, the λ_i are real. Further Q has a complete set of orthonormal eigenvectors $\phi^{(i)}$

$$\begin{aligned} Q\phi^{(i)} &= \lambda_i\phi^{(i)} \\ \phi^{(i)\dagger}\phi^{(j)} &= \delta_{ij}. \end{aligned} \tag{5}$$

If we denote the eigenvectors of P and P^T, which correspond to λ_i by $\mathbf{U}^{(i)}$ and $\mathbf{W}^{(i)}$, respectively, then we clearly have

$$\begin{aligned} \mathbf{U}^{(i)} &= \alpha_i D^{1/2}\phi^{(i)} \\ \mathbf{W}^{(i)} &= \alpha_i^{-1} D^{-1/2}\phi^{(i)*} \\ \mathbf{W}^{(i)T}\mathbf{U}^{(j)} &= \delta_{ij}. \end{aligned} \tag{6}$$

The α_i in (6) are any convenient constants. In addition, we have the spectral representation

$$P = \sum_{i=1}^{V} \lambda_i \mathbf{U}^{(i)} \mathbf{W}^{(i)T}. \tag{7}$$

Since the rows of P sum to unity, the eigenvector $\mathbf{U}^{(1)}$ may be chosen to be

$$\mathbf{U}^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{8}$$

The eigenvector $\mathbf{W}^{(1)}$,

$$P^T \mathbf{W}^{(1)} = \mathbf{W}^{(1)}, \tag{9}$$

is then the stationary probability vector for the Markov chain

$$\mathbf{W}^{(1)} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_V \end{pmatrix}, \tag{10}$$

obeying the normalization of (6).

Equation (9) is easy to solve when P corresponds to random routing. Define a vector $\mathbf{Y}$ by $(\mathbf{Y})_i = l_i$, so $D\mathbf{Y} = \mathbf{U}^{(1)}$. Then, using (3),

$$P^T \mathbf{Y} = AD\mathbf{Y} = A\mathbf{U}^{(1)} = \mathbf{Y}. \tag{11}$$

Thus, the stationary probability vector for the chain has components

$$p_i = \frac{l_i}{2L} > 0, \tag{12}$$

where L equals the number of edges in the graph. Using (12) in (6), we find

$$(\phi^{(1)})_i = \sqrt{p_i}. \tag{13}$$

We next derive the expression (24) for N that will be our point of departure. Let $f_{ij}(n)$, $n = 1, 2, \cdots$ be the first passage probability from node i to node j at time n. Then, if P^n is the nth power of the transition matrix P, we have (Feller, in Ref. 4, p. 352)

$$\begin{aligned} f_{ij}(1) &= P_{ij} \\ f_{ij}(n) &= (P^n)_{ij} - \sum_{k=1}^{n-1} (P^k)_{jj} f_{ij}(n-k), \qquad n \geq 2. \end{aligned} \tag{14}$$

In terms of the generating functions, defined for $|s| < 1$ by

$$F_{ij}(s) = \sum_{n=1}^{\infty} f_{ij}(n) s^n \tag{15}$$

$$P_{ij}(s) = \sum_{n=1}^{\infty} (P^n)_{ij} s^n, \tag{16}$$

(14) is equivalent to

$$F_{ij}(s) = \frac{P_{ij}(s)}{1 + P_{jj}(s)}, \qquad |s| < 1. \tag{17}$$

Since all eigenvalues λ_i of P satisfy $|\lambda_i| \leq 1$, (16) shows that (17) may be rewritten for $i \neq j$ as

$$F_{ij}(s) = \frac{[I - sP]^{-1}_{ij}}{[I - sP]^{-1}_{jj}}, \qquad \begin{matrix} i \neq j \\ |s| < 1. \end{matrix} \tag{18}$$

Denoting the ith component of $\mathbf{U}^{(\mu)}$ by $U_i^{(\mu)}$, and similarly for $\mathbf{W}^{(\mu)}$, set

$$U_i^{(\mu)} W_j^{(\mu)} = b_{ij}^{(\mu)} \tag{19}$$

and note from (8) and (10) that

$$b_{ij}^{(1)} = p_j. \tag{20}$$

Then, using the spectral representation (7), (18) may be written

$$F_{ij}(s) = \frac{p_j + (1 - s) \sum_{\mu=2}^{V} \frac{1}{1 - s\lambda_\mu} b_{ij}^{(\mu)}}{p_j + (1 - s) \sum_{\mu=2}^{V} \frac{1}{1 - s\lambda_\mu} b_{jj}^{(\mu)}}. \tag{21}$$

In this form, $F_{ij}(s)$ may be analytically continued to a neighborhood of $s = 1$. Then, from (15) and a standard Abelian theorem we have

$$n_{ij} = \sum_{n=1}^{\infty} n f_{ij}(n) = \frac{dF_{ij}(s = 1)}{ds}. \tag{22}$$

Using (21) to (22), we obtain

$$n_{ij} = \frac{1}{p_j} \sum_{\mu=2}^{V} \frac{1}{1 - \lambda_\mu} (b_{jj}^{(\mu)} - b_{ij}^{(\mu)}). \tag{23}$$

Finally, using the definition (1) and relations (6) and (19), we see that the average number of steps, N, required to reach a destination with random routing for V nodes may be written

$$N = \frac{1}{V - 1} \sum_{\mu=2}^{V} \frac{1}{1 - \lambda_\mu} \left[\sum_{j=1}^{V} \frac{1}{p_j} |\phi_j^{(\mu)}|^2 - \frac{1}{V} \left| \sum_j \frac{\phi_j^{(\mu)}}{\sqrt{p_j}} \right|^2 \right]. \tag{24}$$

Again, p_j are the stationary probabilities of the Markov chain governed by transition matrix $P = DA$. The λ_μ, $\mu \geq 2$, are the nonunity eigenvalues of P and $Q = D^{1/2}AD^{1/2}$, the latter matrix having orthonormal eigenvectors $\phi^{(\mu)}$. Also $\phi^{(1)}$, the eigenvector of Q associated with the eigenvalue $\lambda_1 = 1$, is given by (13).

We now lower bound the right member of (24). A few known facts about the quantities λ_μ, $\phi^{(\mu)}$, and p_j will be exploited, but their great interdependence (they are all determined by P) will be ignored.

Set

$$a_\mu = \sum_{j=1}^{V} \frac{1}{p_j} |\phi_j^{(\mu)}|^2 - \frac{1}{V} \left| \sum_j \frac{\phi_j^{(\mu)}}{\sqrt{p_j}} \right|^2 \tag{25}$$

and rewrite (24) as

$$N = \frac{1}{V-1} \sum_{\mu=2}^{V} \frac{a_\mu}{1-\lambda_\mu}. \tag{26}$$

We assert $a_\mu > 0$, since, by Cauchy's inequality,

$$\frac{1}{V} \left| \sum_j \frac{\phi_j^{(\mu)}}{\sqrt{p_j}} \cdot 1 \right|^2 \leq \frac{1}{V} \left(\sum_j \frac{|\phi_j^{(\mu)}|^2}{p_j} \right) \left(\sum_j 1 \right) = \sum_j \frac{|\phi_j^{(\mu)}|^2}{p_j}. \tag{27}$$

However, equality can hold only if

$$\frac{\phi_j^{(\mu)}}{\sqrt{p_j}} \propto 1,$$

that is, if

$$\phi_j^{(\mu)} \propto \sqrt{p_j} = \phi_j^{(1)}.$$

Since $\phi^{(\mu)}$ is orthogonal to $\phi^{(1)}$ for $\mu \geq 2$, this cannot happen.†

With the positivity of the a_μ established, minimize (26) over the λ_μ holding the $a_\mu > 0$ fixed. During this minimization, we only impose two constraints on the $\lambda^{(\mu)}$. First, $\lambda_\mu < 1$, and second $\sum_{\mu=2}^{V} \lambda_\mu = -1$. The first has been amply discussed already while the second follows from the fact that the diagonal elements of the transition matrix P are zero and so

$$0 = \operatorname{tr} P = \sum_{\mu=1}^{V} \lambda_\mu. \tag{28}$$

Since $\lambda_1 = 1$, the second constraint follows. Introducing the second constraint via a Lagrange multiplier (call it $-\beta$), we find that the unique stationary point of

† Incidentally, if we write $n_{ij} = \sum_\mu \tau_{ij}^{(\mu)}/(1-\lambda_\mu)$ it is not necessary that $\tau_{ij}^{(\mu)} \geq 0$. However, $\tau_{ij}^{(\mu)} + \tau_{ji}^{(\mu)} \geq 0$.

$$\sum_{\mu=2}^{V} \frac{a_\mu}{1-\lambda_\mu} - \beta\Sigma\lambda_\mu, \tag{29}$$

satisfying the two constraints, occurs when

$$\lambda_\mu = 1 - \sqrt{a_\mu}\left(\frac{V}{\sum_{\nu=2}^{V}\sqrt{a_\nu}}\right). \tag{30}$$

Therefore, applying (30) to (26) gives

$$N \geq \frac{1}{V(V-1)}\left(\sum_{\mu=2}^{V}\sqrt{a_\mu}\right)^2. \tag{31}$$

Next, use the fact that the arithmetic mean exceeds the geometric mean:

$$\begin{aligned} N &\geq \frac{V-1}{V}\left(\frac{1}{V-1}\sum_{\mu=2}^{V}\sqrt{a_\mu}\right)^2 \\ &\geq \frac{V-1}{V}\left[\left(\prod_{2}^{V}\sqrt{a_\mu}\right)^{\frac{1}{V-1}}\right]^2 = \frac{V-1}{V}\left(\prod_{\mu=2}^{V} a_\mu\right)^{\frac{1}{V-1}}. \end{aligned} \tag{32}$$

Motivated by the definition of the a_μ [see (25)], introduce the real symmetric matrix M having elements

$$M_{ij} = \frac{1}{p_i}\delta_{ij} - \frac{1}{V}\frac{1}{\sqrt{p_i p_j}}. \tag{33}$$

Then,

$$a_\mu = \phi^{(\mu)\dagger} M \phi^{(\mu)}. \tag{34}$$

Note that

$$M\phi^{(1)} = \sum_{j=1}^{V}\left(\frac{1}{p_i}\delta_{ij} - \frac{1}{V}\frac{1}{\sqrt{p_i p_j}}\right)\sqrt{p_j} = 0, \tag{35}$$

while we have already shown

$$\phi^\dagger M \phi > 0 \quad \text{if} \quad (\phi, \phi^{(1)}) = 0. \tag{36}$$

Thus, M is positive semi-definite, having precisely one zero eigenvalue. Fix $p_i > 0$, $i = 1, \cdots, V$, and proceed to lower bound (32) by writing

$$\prod_{\mu=2}^{V} a_\mu \geq \min \prod_{\nu=2}^{V} \psi^{(\nu)\dagger} M \psi^{(\nu)}, \tag{37}$$

the minimum in (37) being taken over all orthonormal sets of vectors $\psi^{(\nu)}$ which satisfy

$$(\psi^{(\mu)}, \psi^{(\nu)}) = \delta_{\mu\nu}$$

$$(\psi^{(\mu)}, \phi^{(1)}) = 0, \qquad \mu, \nu = 2, \cdots, V. \tag{38}$$

An inequality of Ky Fan[5] implies that under the conditions of (38) we have

$$\min \prod_{\nu=2}^{V} \psi^{(\nu)\dagger} M \psi^{(\nu)} = \mu_1 \mu_2, \cdots, \mu_{V-1}, \tag{39}$$

where μ_i, $i = 1, \cdots, V-1$ are the nonzero eigenvalues of the matrix M, which was defined in (33). Thus,

$$N \geq \frac{V-1}{V} [\mu_1 \mu_2 \cdots \mu_{V-1}]^{\frac{1}{V-1}}. \tag{40}$$

If g_i are the components of an eigenvector of M associated with eigenvalue μ, the equation

$$\sum_{j=1}^{V} M_{ij} g_j = \mu g_i \tag{41}$$

yields, when (33) is substituted for M_{ij},

$$g_i = k \frac{1}{\sqrt{p_i}\left(\frac{1}{p_i} - \mu\right)}, \tag{42}$$

k being a normalization constant. Since for $\mu \neq 0$ we must have

$$0 = \sum_{i=1}^{V} g_i \phi_i^{(1)} = \sum_{i=1}^{V} g_i \sqrt{p_i}, \tag{43}$$

we obtain, using (42),

$$\sum_{i=1}^{V} \frac{1}{\left(\frac{1}{p_i} - \mu\right)} = 0. \tag{44}$$

Equation (44) determines the $(V-1)$ positive eigenvalues μ_i. Clearing fractions in (44) yields

$$\frac{1}{D} \sum_{i=1}^{V} \prod_{\substack{j=1 \\ j \neq i}}^{V} \left(\frac{1}{p_j} - \mu\right) = 0 \tag{45}$$

with

$$D = \prod_{i=1}^{V} \left(\frac{1}{p_i} - \mu\right). \tag{46}$$

The denominator D in (45) may be discarded and then the product of the nonzero eigenvalues of M is simply read off the remaining poly-

nomial. Since the coefficient of μ^{V-1} in the numerator polynomial in (45) is $V(-1)^{V-1}$,

$$\mu_1 \cdots \mu_{V-1} = \frac{1}{V} \sum_{i=1}^{V} \prod_{\substack{j=1 \\ j \neq i}}^{V} \frac{1}{p_j}. \tag{47}$$

The product of positive numbers having a fixed sum is maximized when each number is the same, so we have

$$\prod_{\substack{j=1 \\ j \neq i}}^{V} p_j \leq \left[\frac{(1 - p_i)}{V - 1} \right]^{V-1}$$

or

$$\mu_1 \cdots \mu_{V-1} \geq \frac{1}{V} \sum_{i=1}^{V} \left(\frac{V - 1}{1 - p_i} \right)^{V-1}. \tag{48}$$

Finally, the minimum of (48) subject to $\sum_{i=1}^{V} p_i = 1$ occurs when $p_i = \frac{1}{V}$, or,

$$\mu_1 \cdots \mu_{V-1} \geq \frac{1}{V} \sum_{i=1}^{V} \left(\frac{V - 1}{1 - \frac{1}{V}} \right)^{V-1} = V^{V-1}. \tag{49}$$

Combining (49) with (40) produces our desired result:

$$N \geq \frac{V - 1}{V} (V^{V-1})^{\frac{1}{V-1}} = (V - 1). \tag{50}$$

It is easy to work backwards through the argument to see that the complete graph is the only one that can achieve $N = (V - 1)$. Clearly, equality in the last step of the argument can only be attained if $p_i = \frac{1}{V}$. Substituting this into (24) and using the orthonormality of the $\phi^{(\mu)}$ yields

$$N = \frac{V - 1}{V} \sum_{\mu=2}^{V} \frac{1}{1 - \lambda_\mu}, \tag{51}$$

which only equals the minimum when $\lambda_\mu = -\frac{1}{V - 1}$, $\mu = 2, \cdots, V$. Using this, and the fact that $\lambda_1 = 1$ in (7) yields

$$\begin{aligned} P &= \mathbf{U}^{(1)}\mathbf{W}^{(1)T} - \frac{1}{V - 1} \sum_{i=2}^{V} \mathbf{U}^{(i)}\mathbf{W}^{(i)T} \\ &= -\frac{1}{V - 1} \sum_{i=1}^{V} \mathbf{U}^{(i)}\mathbf{W}^{(i)T} + \frac{V}{V - 1} \mathbf{U}^{(1)}\mathbf{W}^{(1)T}. \end{aligned} \tag{52}$$

The first dyad sum in the right member of (52) is the identity matrix, while the last dyad in (52) is the matrix that has all elements equal to $1/V$. Thus, P is precisely the transition matrix associated with the complete graph.

III. THE MINIMAL CHAIN

Now we turn to finding the minimum value of N for the unrestricted chain. In the introduction we stated that for this we should place the nodes on a circle and go unidirectionally from one node to the next. Clearly, in this case, $\sum_j n_{ij}$ does not depend on i, and, in fact,

$$N = \frac{1}{V-1}\sum_{j=2}^{V} n_{1j} = \frac{1}{V-1}[1 + 2 + 3 + \cdots + (V-1)] = \frac{V}{2}. \quad (53)$$

We must now show that no other setup can do as well. For this end, expressions such as (23) are not useful since the λ_μ, $\mathbf{U}^{(\mu)}$ and $\mathbf{W}^{(\mu)}$ may be complex and it would be difficult to pick out even real combinations, let alone positive ones that might be lower bounded. Rather, we retreat to an obvious generalization of the argument we used to derive (2). This generalization reads

$$n_{ij} = p_{ij} + \sum_{\substack{k \\ k \neq j}} p_{ik}(1 + n_{kj}), \qquad \begin{matrix} i, j = 1, \cdots, V \\ j \neq i \end{matrix} \quad (54)$$

or, since P is stochastic,

$$n_{ij} - \sum_{\substack{k \\ k \neq j}} p_{ik} n_{kj} = 1, \qquad \begin{matrix} i, j = 1, \cdots, V. \\ j \neq i \end{matrix} \quad (55)$$

Equation (55) is our new point of departure.

Let $\mathbf{N}^{(j)}$, $j = 1, \cdots, V$, be the $(V-1)$ dimensional vector whose components are n_{ij}, with j fixed and $i \neq j$. Also let $\tilde{P}(j), j = 1, \cdots, V$, be the $(V-1) \times (V-1)$ matrix obtained by crossing out the jth row and the jth column of P. The $\tilde{P}(j)$ still have positive elements, but the irreducibility of P implies that not all rows of $\tilde{P}(j)$ can sum to one, and so the largest eigenvalue of $\tilde{P}(j)$ is strictly less than one. The equations represented by (55) may now be written

$$[I - \tilde{P}(j)]\mathbf{N}^{(j)} = \mathbf{u}, \qquad j = 1, \cdots, V, \quad (56)$$

where $\mathbf{u}$ is the $(V-1)$ dimensional vector having all components unity. From (56),

$$\mathbf{N}^{(j)} = [I - \tilde{P}(j)]^{-1}\mathbf{u} = [I + \tilde{P}(j) + \tilde{P}^2(j) + \cdots]\mathbf{u}. \quad (57)$$

If $\tilde{P}^k_{rs}(j)$ denotes the (r, s) element of the kth power of $\tilde{P}(j)$, then (1) and (57) yield

$$N = \frac{1}{V(V-1)} \sum_{j=1}^{V} \sum_{r,s=1}^{V-1} \sum_{k=0}^{\infty} \tilde{P}_{rs}^{k}(j), \tag{58}$$

where by $\tilde{P}^0(j)$ we mean the identity matrix for $(V-1)$ dimensions. Equation (58) succumbs to the application of the following:

Lemma: *Let P be any stochastic $V \times V$ matrix and let $\tilde{P}(j)$, $j = 1, \cdots, V$ be the $(V-1) \times (V-1)$ matrix obtained by crossing out the jth row and jth column of P. Then*

$$\sum_{j=1}^{V} \sum_{r,s=1}^{V-1} \tilde{P}_{rs}^{k}(j) \geq \begin{cases} 0 & \text{for} \quad k \geq 0 \\ V(V-k-1) & \text{for} \quad 0 \leq k \leq V-2. \end{cases} \tag{59}$$

Proof: The positivity of the sum for all k is trivial since $\tilde{P}(j)$ has nonnegative elements. If $k = 0$, $P^0(j)$ is the identity for all j, so the sum of all its elements if $(V-1)$, and we obtain $V(V-1)$ when we sum over j. Next consider $k = 1$ and make use of the stochastic nature of P, that is $\sum_s P_{rs} = 1$.

$$\sum_j \sum_{r,s} \tilde{P}_{rs}(j) = \sum_j \sum_{\substack{r \neq j \\ s \neq j}} P_{rs} = \sum_j \sum_{r \neq j} (1 - P_{rj})$$

$$= V(V-1) - \sum_j \sum_r P_{rj} + \sum_r P_{rr} \geq V(V-2).$$

Now proceed by induction assuming that the lemma is true for k and show it true for $(k+1)$

$$\sum_{j=1}^{V} \sum_{r,s=1}^{V-1} \tilde{P}_{rs}^{k+1}(j) = \sum_j \sum_{\substack{n_1 \neq j \\ \vdots \\ n_{k+2} \neq j}} P_{n_1 n_2} P_{n_2 n_3} \cdots P_{n_{k+1} n_{k+2}}$$

$$= \sum_j \sum_{\substack{n_1 \neq j \\ \vdots \\ n_{k+1} \neq j}} P_{n_1 n_2} P_{n_2 n_3} \cdots P_{n_k n_{k+1}}$$

$$- \sum_j \sum_{\substack{n_1 \neq j \\ \vdots \\ n_{k+1} \neq j}} P_{n_1 n_2} \cdots P_{n_{k+1} j}$$

$$\geq V(V-k-1) - \sum_j \sum_{n_1, n_2 \cdots} P_{n_1 n_2} \cdots P_{n_{k+1}, j}. \tag{60}$$

The inequality follows from the induction assumption (true for k) and is further strengthened by extending the range of summation on the negative term. In the negative term, the sums over $j, n_{k+1}, \cdots, n_2$ yield unity, and, finally, the sum over n_1 gives V. The conclusions of the lemma follow.

Now apply the lemma to (59).

$$N \geq \frac{1}{V(V-1)} \sum_{k=0}^{V-2} \left[\sum_j \sum_{r,s} \tilde{P}_{rs}^k(j) \right]$$

$$\geq \frac{1}{V(V-1)} \sum_{k=0}^{V-2} V(V-k-1) = \frac{V}{2}, \tag{61}$$

as desired.

The condition for equality in (61) obviously requires that

$$\sum_j \sum_{\substack{n_1 \neq j \\ \vdots \\ n_k \neq j}} P_{n_1 n_2} \cdots P_{n_k j} = \sum_j \sum_{n_1, \cdots, n_k} P_{n_1 n_2} \cdots P_{n_k j}, \qquad k = 1, \cdots, V-1. \tag{62}$$

Since all terms on the right member of (62) are nonegative, any term present there but not appearing on the left must be zero. In particular, we must have for any j

$$P_{jj} = 0 \qquad \text{if} \quad k = 1 \tag{63}$$

$$\sum_{n_2 \cdots n_k} P_{j n_2} \cdots P_{n_r j} = 0 \qquad k = 2, \cdots, V-1. \tag{64}$$

These equations state that it is impossible to return to any initial state j in less than V steps. This, plus irreducibility, implies the unidirectional movement or a circle.

The reason why the lemma is exact in this case is that each $\tilde{P}(j)$ is, then—except for a reordering of nodes—a canonical Jordan block with all zeros in the matrix except for $(V-2)$ ones on the appropriate off-diagonal.

IV. THE WORST WALK

We have been unable to describe the worst setup for random routing. For $V = 3$, 4, and 5, numerical work shows that arranging the vertices on a straight line gives the worst cases. In fact, for V nodes on a straight line it is possible to show that

$$N = \frac{V^2 - 1}{3}, \tag{65}$$

which is significantly worse than $(V-1)$, the best attainable with random routing. However, for large V one can do worse than (65). Our result for this situation is

$$O(V^3) \leq N_{\text{worst}} \leq O(V^{3.5}). \tag{66}$$

In (66) N_{worst} denotes the largest value of N obtained from any of the connected network graphs on the V vertices.

For the lower bound, assume $V = 6m - 1$ and consider the bar-

Fig. 1—An 11-node barbell.

bell graphs described by Landau-Odlyzko.[3] This class of graphs has $(2m - 1)$ nodes connected in a straight line with complete graphs of $2m$ nodes attached to each end of the line by a single edge of the graph. If $m = 1$, this only describes a straight line, but for $m = 2$ or greater, the barbell nature is evident. The case for $V = 11$ is shown in Fig. 1.

In Ref. 3 the authors show that for such graphs λ_2, the second largest eigenvalue of $P = DA$, satisfies

$$\lambda_2 \geq 1 - \frac{\gamma}{V^3} \tag{67}$$

with $\gamma = 54 + O\left(\frac{1}{V}\right)$. Then from (26), (67), and (34),

$$N \geq \frac{1}{V-1}\frac{a_2}{1-\lambda_2} \geq O(V^2)\phi^{(2)\dagger}M\phi^{(2)}$$

$$\geq O(V^2)\mu_{V-1}, \tag{68}$$

μ_{V-1} being the smallest nonzero eigenvalue of M. Equation (44) shows

$$\mu_{V-1} > \frac{1}{\max p_i}. \tag{69}$$

Since the nodes of the complete graphs in the barbell each have $O(V)$ incident edges and there are $O(V^2)$ edges in the barbell, (12) implies that $\max p_i = O\left(\frac{1}{V}\right)$. Equations (68) and (69) then give

$$N \geq O(V^2)O(V) = O(V^3). \tag{70}$$

Although we will not give the details here, it is possible to show that, in fact, $N = O(V^3)$ for barbell graphs.

Our upper bound will be based on (57), but first we need the result that the largest eigenvalue of $\tilde{P}(j)$ satisifes

$$\lambda_1(j) \equiv \lambda_1[\tilde{P}(j)] \leq 1 - O\left(\frac{1}{V^3}\right). \tag{71}$$

Just as the matrix $\tilde{P}(j)$ was defined in Section III, now introduce matrices $\tilde{D}(j)$ and $\tilde{A}(j)$ by eliminating the jth row and column of D and A, respectively. Then,

$$\tilde{P}(j) = \tilde{D}(j)\tilde{A}(j). \tag{72}$$

Also set

$$\tilde{Q}(j) = \tilde{D}^{1/2}(j)\tilde{A}(j)\tilde{D}^{1/2}(j) \tag{73}$$

so that $\tilde{Q}(j)$ is symmetric and has the same eigenvalues as $\tilde{P}(j)$. In fact, we have

$$\lambda_1[\tilde{P}(j)] = \lambda_1[\tilde{Q}(j)] = \max_{\sum y_i^2=1} \mathbf{y}^\dagger \tilde{Q}(j)\mathbf{y}$$

$$= \max_{\sum l_i z_i^2=1} \mathbf{z}^\dagger A(j)\mathbf{z} = \max_{\substack{x_j=0 \\ \sum l_i x_i^2=1}} \mathbf{x}^\dagger A\mathbf{x}. \tag{74}$$

Note that in (74) we have returned to the $V \times V$ adjacency matrix A by introducing the constraint $x_j = 0$. If S is the set of those (ordered) pairs of indices which corresponds to vertices connected by an edge of the graph, we have[†]

$$\mathbf{x}^\dagger A\mathbf{x} = \sum_{(m,n)\in S} x_m x_n = \frac{1}{2}\sum_S [x_m^2 + x_n^2 - (x_m - x_n)^2]$$

$$= \sum_m l_m x_m^2 - \frac{1}{2}\sum_S (x_m - x_n)^2. \tag{75}$$

Using (75) in (74) we easily obtain

$$\lambda_1(j) = 1 - \frac{1}{2} \min_{\substack{x_j=0 \\ \sum l_i x_i^2=1}} \sum_S (x_m - x_n)^2. \tag{76}$$

Let x_k be the component of x having the largest square. Then

$$1 = \sum l_i x_i^2 \leq (V-1) l_{\max} x_k^2, \tag{77}$$

or

$$x_k^2 \geq \frac{1}{l_{\max}(V-1)}, \tag{78}$$

where $l_{\max}$ (or $l_{\min}$) is the maximum (or minimum) of the l_i, $i = 1, \cdots, V$.

Now, from the connectivity of the graph, vertex j is attached to another vertex t. So, using (76),

$$\lambda_1(j) \leq 1 - x_t^2, \tag{79}$$

since $x_j = 0$. If $t = k$, (78) and (79) yield

$$\lambda_1(j) \leq 1 - \frac{1}{(V-1)l_{\max}} \qquad (t = k). \tag{80}$$

If $t \neq k$, there exists a chain of r distinct edges joing vertices t and k.

[†] The remainder of this demonstration is entirely inspired by Ref. 3.

Clearly, $r \leq d$, where d is the diameter of the graph. Then, using the basic trick of Ref. 3,

$$x_k - x_t = (x_k - x_{k_1}) + (x_{k_1} - x_{k_2}) + \cdots + (x_{k_{r-1}} - x_t) \tag{81}$$

and so, by Cauchy's inequality,

$$(x_k - x_t)^2 \leq \frac{r}{2} \sum_S (x_m - x_n)^2 \leq \frac{d}{2} \sum_S (x_m - x_n)^2. \tag{82}$$

Equations (76) and (82) thus yield

$$\lambda_1(j) \leq 1 - \frac{(x_k - x_t)^2}{d}. \tag{83}$$

Combine (79) and (83) by averaging and then minimize over the numerical value of x_t to obtain, with (78), for $t \neq k$,

$$\lambda_1(j) \leq 1 - \frac{1}{2}\left[x_t^2 + \frac{(x_k - x_t)^2}{d}\right] \leq 1 - \frac{1}{2(1+d)(V-1)l_{\max}}. \tag{84}$$

Both cases (80) and (84) are included in

$$\lambda_1(j) \leq 1 - \frac{1}{2(1+d)(V-1)l_{\max}} \leq 1 - \frac{1}{2V^3}, \tag{85}$$

since $d \leq (V-1)$ and $l_{\max} \leq (V-1)$.

To complete the upper bound on N, start with (58) and write (suppressing the j-dependence of the matrices in the notation)

$$\begin{aligned} N &= \frac{1}{V(V-1)} \sum_{j=1}^{V} \mathbf{u}^\dagger [I - \tilde{P}(j)]^{-1} \mathbf{u} \\ &\leq \frac{\|\mathbf{u}\|^2}{V(V-1)} \sum_j \|[I - \tilde{P}(j)]^{-1}\| \\ &= \frac{1}{V} \sum_j \|\tilde{D}^{1/2}[I - \tilde{Q}]^{-1}\tilde{D}^{1/2}\| \\ &\leq \frac{1}{V} \sum_j \|\tilde{D}^{1/2}\| \, \|\tilde{D}^{-1/2}\| \, \|[I - \tilde{Q}]^{-1}\| \\ &\leq \sqrt{\frac{l_{\max}}{l_{\min}}} \frac{1}{V} \sum_{j=1}^{V} \frac{1}{1 - \lambda_1(j)}. \end{aligned} \tag{86}$$

From $l_{\min} \geq 1$, $l_{\max} \leq (V-1)$, (85) and (86) we thus have

$$N \leq 2V^3\sqrt{V-1}, \tag{87}$$

as desired.

Is it possible that the complete graph is optimal for random routing because adding any edge to a graph decreases N? No. Many edges must be added to the straight-line graph to form the barbell, but the former has $N = O(V^2)$ while for the latter, $N = O(V^3)$.

REFERENCES

1. L. Kleinrock, *Communication Nets*, New York: McGraw-Hill, 1964, Chap. 6.
2. D. Mitra and A. Weiss, "Analysis of Delay-Differential Equations Arising in Communication Network Synchronization," in Proc. IEEE Int. Symp. Circuits and Systems, Houston, Texas, 1980, pp. 839–43.
3. H. J. Landau and A. M. Odlyzko, "Bounds for Eigenvalues of Certain Stochastic Matrices," Linear Algebra and Applications, *38* (June 1981), pp. 5–15.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, New York: John Wiley, 1957, 2nd ed.
5. E. F. Beckenbach and R. Bellman, *Inequalities*, New York: Springer-Verlag, 1965, p. 76, Theorem 24.
6. Peter Lancaster, *Theory of Matrices*, New York: Academic Press, 1969.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Quality Evaluation Plan Using Adaptive Kalman Filtering

By M. S. PHADKE

(Manuscript received March 5, 1982)

An important function of the Bell Laboratories Quality Assurance Center and the Western Electric Quality Assurance Directorate is to audit the quality of the products manufactured and the services provided by the Western Electric Company to determine if the intended quality standards are met. Until the sixth period of 1980, the t-rate system was used to make inference on the product quality. Starting the seventh period of 1980, the Quality Measurement Plan (QMP) has been implemented. The QMP is based on an empirical Bayes model of the audit-sampling process using the current and the preceding five periods of data. Because it ignores the time order of the data, it is slow in responding to drifts in the process mean. The Quality Evaluation Plan (QEP) has been designed to take into account the time order of the data and to be more sensitive to drifts in the process mean. In this paper we present the Quality Evaluation Plan, which uses the entire time series of data on a given product to determine if that product meets the quality standard. The time series is modeled by a stochastic process, which allows for the possibility that the process mean may drift or fluctuate around a fixed value. An adaptive Kalman filtering theory is developed for filtering out the sampling variance and obtaining the best estimate of the true defect index and its confidence interval. Thus, in QEP the best estimate of the true defect index is obtained by a combination of adaptive exponential smoothing and shrinkage to the mean. The QEP computations are recursive, and the total computing efforts of QEP and QMP are roughly equal. The paper contains several examples to illustrate the QEP.

I. INTRODUCTION

An important function of the Bell Laboratories Quality Assurance Center and the Western Electric Quality Assurance Directorate is to

audit the quality of the products manufactured, and the services provided by the Western Electric Company to determine if the intended quality standards are met. This is achieved by dividing the products and services into some 3000 homogeneous classes. A small sample is taken from each class during each period (there are eight rating periods in a year). Based on this data, an inference is made in each period regarding the compliance of each class to the quality standard.

Until the sixth period of 1980, the *t*-rate system, evolved from the work of Dodge and others,[1] was used to rate the product quality. Starting with the seventh period of 1980, the Quality Measurement Plan (QMP) was implemented. The QMP, developed by A. B. Hoadley,[2] is based on an empirical Bayes model of the audit-sampling process. It uses the current and the preceding five periods of data. It represents a considerable improvement in the statistical power for detecting substandard quality as compared with the old rules based on the *t*-rate. However, QMP ignores the time order of the observations, so it is less sensitive to drifts in the process mean. The Quality Evaluation Plan (QEP) has been designed to take into account the time order of the data and to be more sensitive to drifts in the process mean.

The object of this paper is to present the Quality Evaluation Plan, which uses the entire time series of data on a given class to determine if that class meets the quality standard. The time series is modeled by a stochastic process, which allows for the possibility of the process mean to (*i*) drift or (*ii*) fluctuate around a fixed value. An adaptive Kalman filtering theory is developed for filtering out the sampling variance and obtaining the best estimate of the true defect index and its confidence interval. Some of the salient features of QEP are: (*i*) the best estimate of the defect index is obtained by an adaptive exponential smoothing process, making QEP more responsive to shifts and drifts in the process mean; (*ii*) the QMP model is a special case of the general model proposed here; and (*iii*) the computational method is recursive.

This paper is divided into nine sections. We describe the model in Section II. Section III gives the Kalman filter solution of the model. Adaptive estimates of the model parameters are developed in Section IV, and in Section V we modify the Kalman filter solution of Section III to reflect the fact that the model parameters are estimated. The solution thus obtained is the adaptive Kalman filter. The construction of the box chart for displaying the results, and the rules for the exception report are spelled out in Section VI. The selection of the starting values for the estimation of the model parameters and the adaptive Kalman filter solution is discussed in Section VII. The algorithm has been tried on a number of rating classes and also on simulated data. We present representative examples in Section VIII.

Some numerical comparison between QEP and QMP is also presented in that section. Finally, in Section IX, we discuss the features of the QEP and other potential applications of the adaptive Kalman filtering methodology developed in this paper. A summary of the QEP formulae is given in Appendix C.

Parts of the derivation of QEP are heuristic. The heuristic has a sound theoretical foundation under two assumptions: (*i*) the audit sample size for a rating class does not vary in time by orders of magnitude, and (*ii*) the maximum likelihood estimates of the time series parameters fall within their feasible region. These assumptions are satisfied in about 95 percent of the audit examples. QEP appears to work for the other 5 percent as well, but this has not been fully tested.

II. DESCRIPTION OF THE MODEL

Let θ_t denote the true defect index in period t for the particular rating class under study. Thus,

$$\theta_t = \frac{\left(\begin{array}{c}\textit{Total number of defects present in} \\ \textit{the production of period } t\end{array}\right)}{\left(\begin{array}{c}\textit{Total number of defects allowed in} \\ \textit{that production under the quality standard}\end{array}\right)}.$$

In deriving the present QMP, Bruce Hoadley[2] assumed that over a time window of six periods the successive values of θ_t are independently and identically distributed around a fixed mean, called the long-term process mean. Consequently, the time order of the past observations is ignored in estimating the current defect index. Hence, QMP responds to a drift in the defect index only through having the moving window, which means a slow response. In our model we will overcome this deficiency by explicitly allowing for drift and serial correlation.

The mathematical analysis of serially correlated data is greatly simplified when the random variables involved are normally distributed. The audit problem can be put in this framework by the square root transformation described in the following paragraph.

For the chosen sample, let e_t be the expected number of defects under standard quality, x_t be the observed number of defects, and $I_t = x_t/e_t$ be the observed defect index; then x_t has a Poisson distribution with mean $e_t\theta_t$. It is well known that the distribution of $\sqrt{x_t}$ can be approximated by the Gaussian density with mean $\sqrt{e_t\theta_t}$ and variance ¼. Let $Y_t = \sqrt{I_t}$. The distribution of Y_t is approximately normal with mean $\sqrt{\theta_t}$ and variance $0.25/e_t$. We shall denote $\sqrt{\theta_t}$ by ζ_t and refer to it as the transformed true defect index, or simply as the true defect index.

When the observations are defined in terms of demerits or defectives we will take x_t and e_t equal to the observed and the expected equivalent defects, respectively, as defined by Hoadley.[3] In this case the distribution of x_t is approximately Poisson with mean $\theta_t e_t$, so we can still use the square root transform defined in the previous paragraph.

Autoregressive-integrated-moving average models with appropriate trend terms may be used to characterize a wide range of serial correlations and trends. However, since the available data on each product is limited, it is essential that we keep the structure simple, involving only a few parameters. Thus, we propose the following model for the variation of ζ_t:

$$\zeta_t = m_t + \nu_{1t}, \tag{1}$$

where m_t is the trend term (including the mean) and ν_{1t} is the deviation from the trend. The successive values of ν_{1t} will be assumed to be independently distributed with zero mean and variance σ_{1t}^2.

Since the exact nature of the drift is not known to begin with, we shall assume that m_t is a random walk. We found in control engineering literature[4] that the random walk model serves well in tracking a variety of trends in unknown parameters, and therefore we chose to use it in the present problem. Thus,

$$m_t = m_{t-1} + \nu_{2t}, \tag{2}$$

where ν_{2t} is a sequence of independently distributed random variables with mean zero and variance σ_{2t}^2. Further, the sequence ν_{2t} will be assumed to be independent of the sequence ν_{1t}.

Equations (1) and (2) thus characterize the variation of the defect index—the component m_t describes the low-frequency (smooth) changes, while ν_{1t} describes the high-frequency changes in ζ_t. If we take $\sigma_{2t}^2 = 0$ and $\sigma_{1t}^2 = \sigma_1^2 =$ constant, then these equations imply that the ζ_t's and hence θ_t's are independently and identically distributed. Thus, the QMP model is a special case of the general model of this paper.

The transformed observed defect index, Y_t, is the transformed true defect index plus the sampling error, η_t. Thus,

$$Y_t = \zeta_t + \eta_t. \tag{3}$$

As discussed earlier, the expected value of Y_t is ζ_t and the variance of Y_t is $0.25/e_t$, so η_t has zero mean and its variance is equal to $0.25/e_t$. We assume that the successive random variables η_t are independent. Also, since the origins of η_t, ν_{1t} and ν_{2t} are unrelated, we assume that these three series are mutually uncrosscorrelated. Further, the distributions of ν_{1t}, ν_{2t}, and η_t are assumed to be normal. The justification for this assumption comes from the fact that Y_t's are approximately normally distributed.

The problem at hand is to make an inference on θ_n given data up to and including the nth time period. In particular, we wish to determine the posterior probability of the event that θ_n exceeds one.

III. KALMAN FILTER SOLUTION

In the Kalman filter terminology, ζ_n and m_n are the unobserved state variables about which we wish to make inference using the observations $Y_1, \cdots, Y_n$. Let us, for now, assume that the model parameters σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ are known for $t = 1, \cdots n$ and that the means and the variances of ζ_0 and m_0 are known. Then the Kalman filter provides recursive formulae for estimating the posterior means and variances of ζ_n and m_n. The derivation of the general Kalman filter may be found in a number of books (e.g., see Jazwinsky[5] or Gelb[4]). A simple derivation for the special case of the audit model is given in Appendix A. The desired recursive formulae are given below.

Conditional on the data up to time t, the distribution of m_t is normal with mean $\hat{m}_t$ and variance q_t,

$$\text{i.e.,} \quad m_t | t \sim N(\hat{m}_t, q_t),$$

$$\text{where} \quad \hat{m}_t = \omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t \tag{4}$$

$$q_t = (1 - \omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2) \tag{5}$$

$$\omega_{2t} = \frac{\sigma_{1t}^2 + \sigma_{\eta t}^2}{\sigma_{1t}^2 + \sigma_{\eta t}^2 + \sigma_{2t}^2 + q_{t-1}}. \tag{6}$$

Likewise, conditional on the data up to time t, the distribution of ζ_t is normal with mean $\hat{\zeta}_t$ and variance p_t,

$$\text{i.e.,} \quad \zeta_t \sim N(\hat{\zeta}_t, p_t),$$

$$\text{where} \quad \hat{\zeta}_t = \omega_{2t}\omega_{1t}\hat{m}_{t-1} + (1 - \omega_{2t}\omega_{1t})Y_t \tag{7}$$

$$p_t = (1 - \omega_{2t}\omega_{1t})\sigma_{\eta t}^2 \tag{8}$$

$$\omega_{2t}\omega_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}}. \tag{9}$$

To use these recursive equations the starting values m_0 and q_0 must be specified. The choice of these values is discussed in Section VII. For now, we note that as $t \to \infty$, the effect of the starting values reduces to zero.

Notice that eq. (4) is an adaptive, exponential smoothing equation. The smoothing constant, ω_{2t}, is a function of time and is determined by the relative values of the different variances as given by eq. (6). Observe that $V(Y_t | m_t) = \sigma_{1t}^2 + \sigma_{\eta t}^2$ so that $\sigma_{1t}^2 + \sigma_{\eta t}^2$ measures the uncertainty in using Y_t for estimating m_t; also, $\sigma_{2t}^2 = V(m_t | m_{t-1})$ and $q_{t-1} = V(m_{t-1} | t - 1)$. Thus, $\sigma_{2t}^2 + q_{t-1}$ is a measure of uncertainty in

using $\hat{m}_{t-1}$ for estimating m_t. It is clear from eqs. (4) and (6) that the weights given to Y_t and $\hat{m}_{t-1}$ are inversely proportional to their respective uncertainties in estimating m_t.

Equations (7) and (9) are the analogous equations for computing the posterior mean of ζ_t. Note that $V(Y_t|\zeta_t) = \sigma_{\eta t}^2$ is the uncertainty in using Y_t to estimate ζ_t. Further, $V(\zeta_t|m_{t-1}) = \sigma_{1t}^2 + \sigma_{2t}^2$ and $V(m_{t-1}|t-1) = q_{t-1}$ so that $\sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}$ is the uncertainty in using $\hat{m}_{t-1}$ to estimate ζ_t. The weights on Y_t and $\hat{m}_{t-1}$ are thus seen to be inversely proportional to the respective uncertainties.

From eq. (8) we note that the posterior variance of ζ_t conditional on data up to time t is smaller than $\sigma_{\eta t}^2$ by the factor $(1 - \omega_{2t}\omega_{1t})$. Thus, the factor $(1 - \omega_{2t}\omega_{1t})$ represents the advantage of filtering in estimating ζ_t. Similarly, from eq. (5) we see that the factor $(1 - \omega_{2t})$ is the benefit of filtering in estimating m_t.

To compare the QMP model given in Ref. 2 with the QEP model we shall rewrite eqs. (7) and (9) as follows:

$$\hat{\zeta}_t = \omega_{1t}\hat{m}_t + (1 - \omega_{1t})Y_t$$

$$\omega_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \sigma_{1t}^2}.$$

Analogous to the QMP, eq. (7) expresses $\hat{\zeta}_t$ as a weighted sum of $\hat{m}_t$, the estimated current mean level, and Y_t, the current observation. The weight ω_{1t} is analogous to the shrinkage constant given in Ref. 2, and we will also call it a shrinkage constant.

The discussion of this section was based on the assumption that σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ are known quantities. However, in the audit problem σ_{1t}^2 and σ_{2t}^2 are not known and must be estimated from the observed data. In the following section we will derive the estimates of σ_{1t}^2 and σ_{2t}^2, and in Section V we will modify eqs. (4) through (9) to accommodate the fact that σ_{1t}^2 and σ_{2t}^2 are estimated.

IV. ESTIMATION OF THE MODEL PARAMETERS

Consider the case where $\sigma_{1t}^2 = \sigma_1^2 =$ constant, $\sigma_{2t}^2 = \sigma_2^2 =$ constant, and $\sigma_{\eta t}^2 = \sigma_\eta^2 =$ constant.

Let us define $Z_t = Y_t - Y_{t-1}$. Under the assumed model $E(Z_t) = 0$ and the autocovariances of Z_t are given by

$$E(Z_t^2) = 2\sigma_1^2 + \sigma_2^2 + 2\sigma_\eta^2, \tag{10}$$

$$E(Z_t Z_{t-1}) = -\sigma_1^2 - \sigma_\eta^2, \tag{11}$$

and

$$E(Z_t Z_{t-l}) = 0, \qquad l \geq 2. \tag{12}$$

Thus, Z_t is a first-order moving average [MA(1)] process of zero mean, i.e., Z_t can be represented as

$$Z_t = a_t + \beta a_{t-1}, \tag{13}$$

where a_t is a white noise series of variance σ^2. The parameters β and σ^2 are related to σ_1^2, σ_2^2, and σ_η^2 through the autocovariance function; i.e.,

$$E(Z_t^2) = (1 + \beta^2)\sigma^2 = 2\sigma_1^2 + \sigma_2^2 + 2\sigma_\eta^2, \tag{14}$$

and

$$E(Z_t Z_{t-1}) = \beta\sigma^2 = -\sigma_1^2 - \sigma_\eta^2. \tag{15}$$

Solving eqs. (14) and (15) we have

$$\sigma_1^2 = -\beta\sigma^2 - \sigma_\eta^2 \tag{16}$$

and

$$\sigma_2^2 = (1 + \beta)^2\sigma^2. \tag{17}$$

The nonnegativity of σ_1^2 and the invertibility of the model given by eq. (13) impose the following restriction on β: $-1 < \beta \leq -\sigma_\eta^2/\sigma^2$. Thus, the feasible region for β and σ^2 is the one enclosed by the lines: $\beta = -1$, $\beta\sigma^2 = -\sigma_\eta^2$, and $\sigma^2 = \infty$. The region is shown in Fig. 1.

The parameters β and σ^2 can be estimated using a suitable time-series method. Once β and σ^2 are known, eqs. (16) and (17) may be used to estimate σ_1^2 and σ_2^2.

Contrary to the assumption made at the beginning of this section, σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ may in fact vary from period to period. Through eqs. (14) and (15) this implies that β and σ^2 may vary with time. In other words, Z_t is like a $MA(1)$ process with changing parameters. Therefore,

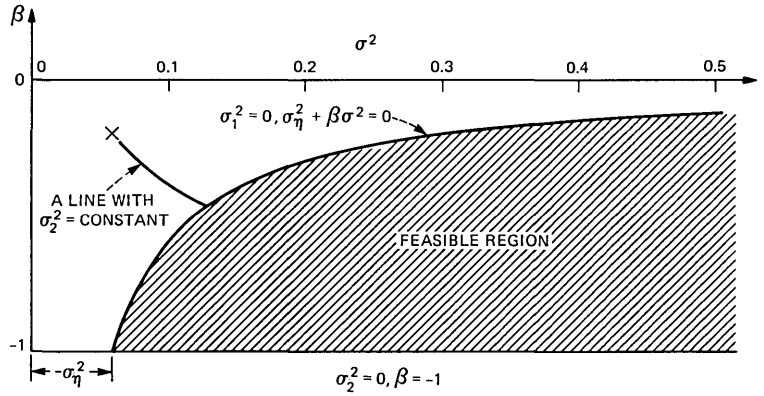


Fig. 1—Feasible region for β and σ^2.

we need an adaptive, recursive estimation method for estimating β and σ^2, rather than the usual time series estimation methods. The recursive method of Phadke[6] will therefore be used here. The method discounts the past data exponentially and thus can respond to changes in the model parameters. The necessary recursion formulae are given below. Appendix B gives the derivation of these formulae.

$$\hat{\beta}_t = \beta_0 - R_t^{-1} v_t \tag{18}$$

$$\hat{\sigma}_t^2 = S_t(\hat{\beta}_t)/A_t, \tag{19}$$

where

$$v_t = \lambda v_{t-1} + 2a_t \frac{da_t}{d\beta} \tag{20}$$

$$R_t = \lambda R_{t-1} + 2\left(\frac{da_t}{d\beta}\right)^2 \tag{21}$$

$$a_t = Z_t - \beta_0 a_{t-1} \tag{22}$$

$$\frac{da_t}{d\beta} = -a_{t-1} - \beta_0 \frac{da_{t-1}}{d\beta} \tag{23}$$

$$S_t(\beta_0) = \lambda S_{t-1}(\beta_0) + a_t^2 \tag{24}$$

$$S_t(\hat{\beta}_t) = S_t(\beta_0) + (\hat{\beta}_t - \beta_0) v_t + \tfrac{1}{2}(\hat{\beta}_t - \beta_0)^2 R_t \tag{25}$$

$$A_t = \lambda A_{t-1} + 1. \tag{26}$$

The choice of the starting values for these recursions will be studied in Section VII. The parameter λ, $0 < \lambda \leq 1$, determines how fast the old data is discounted in estimating the model parameters. $\lambda = 1$ implies that the entire past data is used. The smaller the λ the faster the past data is discounted.

The estimates $\hat{\sigma}_t^2$ and $\hat{\beta}_t$ are uncorrelated and have the following approximate variances:

$$V(\hat{\beta}_t) \simeq 2\hat{\sigma}_t^2/R_t \tag{27}$$

$$V(\hat{\sigma}_t^2) \simeq 2\hat{\sigma}_t^4/A_t. \tag{28}$$

The estimated values of σ^2 and β may be substituted in eqs. (16) and (17) to compute σ_{1t}^2 and σ_{2t}^2. When $\sigma_{\eta t}^2$ varies with time, we should use the exponentially smoothed value, $\overline{\sigma_{\eta,t}^2}$, as defined below:

$$\overline{\sigma_{\eta,t}^2} = \lambda \overline{\sigma_{\eta,t-1}^2} + (1 - \lambda)\sigma_{\eta t}^2, \tag{29}$$

in place of σ_η^2 in eq. (16). Thus,

$$\hat{\sigma}_{1t}^2 = -\hat{\beta}_t \hat{\sigma}_t^2 - \overline{\sigma_{\eta,t}^2},$$

and

$$\hat{\sigma}_{2t}^2 = (1 + \hat{\beta}_t)^2 \hat{\sigma}_t^2.$$

In the rare case of extreme variations (order of magnitude variations) in $\sigma_{\eta t}^2$, eq. (29) has not been fully tested, so we urge caution for such cases.

It is possible that eqs. (18) and (19) would yield infeasible values of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$. In that case we propose the following truncation rules, which take us close to the maximum point on the feasible boundary.

Step 1: Truncate $\hat{\beta}_t$ to the region $[-1, 0]$. Denote the truncated value by β_t^*, where

$$\beta_t^* = \begin{cases} \beta_0 - \nu_t/R_t & \text{if } -1 \le \beta_0 - \nu_t/R_t \le 0 \\ 0 & \text{if } \beta_0 - \nu_t/R_t > 0 \\ -1 & \text{if } \beta_0 - \nu_t/R_t < -1 \end{cases}.$$

Step 2: Compute σ_t^{*2} and $\hat{\sigma}_{2t}^2$:

$$\sigma_t^{*2} = S_t(\beta_t^*)/A_t = \{S_t(\beta_0) + (\beta_t^* - \beta_0)\nu_t + \tfrac{1}{2}(\beta_t^* - \beta_0)^2 R_t\}/A_t$$

$$\hat{\sigma}_{2t}^2 = (1 + \beta_t^*)^2 \sigma_t^{*2}.$$

Step 3: If β_t^* and σ_t^{*2} belong to the feasible region, i.e., if $(-\beta_t^* \sigma_t^{*2} - \overline{\sigma_{\eta t}^2}) \ge 0$ then $\hat{\beta}_t = \beta_t^*$, $\hat{\sigma}_t^2 = \sigma_t^{*2}$, and $\hat{\sigma}_{1t}^2 = -\hat{\beta}_t \hat{\sigma}_t^2 - \overline{\sigma_{\eta,t}^2}$.

Step 4: If β_t^* and σ_t^{*2} do not belong to the feasible region, i.e., if $(-\beta_t^* \sigma_t^{*2} - \overline{\sigma_{\eta,t}^2}) < 0$, then set $\hat{\sigma}_{1t}^2 = 0$. Compute $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ by solving the following two equations:

$$\hat{\sigma}_{1t}^2 = -\hat{\beta}_t \hat{\sigma}_t^2 - \overline{\sigma_{\eta t}^2} = 0,$$

and

$$\hat{\sigma}_{2t}^2 = (1 + \hat{\beta}_t)^2 \hat{\sigma}_t^2.$$

The resulting $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ are given by

$$\hat{\beta}_t = \frac{-(2 + \hat{\sigma}_{2t}^2/\overline{\sigma_{\eta,t}^2}) + \sqrt{(2 + \hat{\sigma}_{2t}^2/\overline{\sigma_{\eta,t}^2})^2 - 4}}{2}$$

and

$$\hat{\sigma}_t^2 = -\overline{\sigma_{\eta,t}^2}/\hat{\beta}_t.$$

Note that when these truncations are applied there will be a larger degree of approximation involved in using eqs. (27) and (28) for computing the variances of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$. However, these variances enter only into the secondary terms of the adaptive Kalman filter to be derived in the next section. Consequently, we may ignore the effect of truncation.

V. ADAPTIVE KALMAN FILTER

In deriving the Kalman filter solution of Section III, it was assumed that σ_{1t}^2 and σ_{2t}^2 are known quantities. Since in the audit problem these parameters are estimated using the observed data, we need to make due modifications to the Kalman filter solution. We will refer to the resulting formulae as the adaptive Kalman filter.

Consider the distribution of m_t conditional on data up to time t:

$$\hat{m}_t = E(m_t|t) = EE(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) = E(\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t);$$

hence,

$$\hat{m}_t \simeq \hat{\omega}_{2t}\hat{m}_{t-1} + (1 - \hat{\omega}_{2t})Y_t \tag{30}$$

where

$$\hat{\omega}_{2t} = \frac{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2}{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1}}. \tag{31}$$

The distribution of ω_{2t} conditional on data up to time t is very complicated. So the expected value of ω_{2t} cannot be simply computed. Therefore, in eq. (30) we have approximated $E(\omega_{2t}|t)$ by $\hat{\omega}_{2t}$, the maximum posterior density point. This approximation would be very good when $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lie inside the feasible region. However, if $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lie on the boundary of the feasible region, $\hat{\omega}_{2t}$ will be a less accurate approximation of $E(\omega_{2t}|t)$. The extent of the inaccuracy will depend on the values of R_t and A_t. For large R_t and A_t the likelihood of β_t and σ_t^2 will drop very rapidly as one goes away from the feasible boundary nearest to the maximum likelihood point. Thus, the inaccuracy would be smaller for larger values of R_t and A_t.

Now consider the variance of m_t given data up to time t:

$$\begin{aligned} q_t = V(m_t|t) &= EV(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) + VE(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) \\ &= E[(1 - \omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2)] + V[\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t]; \end{aligned}$$

hence,

$$q_t \simeq (1 - \hat{\omega}_{2t})(\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2) + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t}). \tag{32}$$

The effect of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lying on the feasible boundary will be to introduce an inaccuracy in the term $(1 - \hat{\omega}_{2t})(\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2)$ of eq. (32), as discussed above. Knowing the variance of β_t and σ_t^2, the variance of ω_{2t} can be derived via the Taylor series approximation as follows. We have

$$\begin{aligned} \omega_{2t} &= \frac{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2}{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1}} \\ &= \frac{-\beta_t\sigma_t^2 - \overline{\sigma_{\eta t}^2} + \sigma_{\eta t}^2}{(1 + \beta_t + \beta_t^2)\sigma_t^2 + q_{t-1} + \sigma_{\eta t}^2 - \overline{\sigma_{\eta t}^2}} \end{aligned}$$

$$\simeq \hat{\omega}_{2t} + \left(\frac{\partial \omega_{2t}}{\partial \beta_t}\right)(\beta_t - \hat{\beta}_t) + \left(\frac{\partial \omega_{2t}}{\partial \sigma_t^2}\right)(\sigma_t^2 - \hat{\sigma}_t^2).$$

Hence, to the first-order approximation the variance of ω_{2t} is

$$\begin{aligned} V(\omega_{2t}) &\simeq \left(\frac{\partial \omega_{2t}}{\partial \beta}\right)^2 V(\hat{\beta}_t) + \left(\frac{\partial \omega_{2t}}{\partial \sigma^2}\right)^2 V(\hat{\sigma}_t^2) \\ &= \{2\hat{\sigma}_t^6[1 + \hat{\omega}_{2t}(1 + 2\hat{\beta}_t)]^2/R_t \\ &\quad + 2\hat{\sigma}_t^4[\hat{\beta}_t + (1 + \hat{\beta}_t + \hat{\beta}_t^2)\hat{\omega}_{2t}]^2/A_t\} \\ &\quad \div (\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1})^2. \end{aligned} \tag{33}$$

Note that the smoothing constant ω_{2t} is restricted to the interval [0, 1]. The most noninformative distribution on this interval is the uniform distribution whose variance is 1/12. The $\hat{\omega}_{2t}$ computed by eq. (31) clearly adheres to the interval [0, 1]. However, because of the approximations involved, the computed $V(\omega_{2t})$ may come out larger than 1/12. In that case we propose to truncate it to 1/12.

When $\hat{\beta}_t$ and $\hat{\sigma}_{1t}^2$ lie on the boundary of their feasible region, the use of the Taylor series approximation would yield inaccurate estimates of $V(\omega_{2t})$. Since the contribution of this variance is secondary in computing $V(m_t|t)$, we may ignore the effect of truncation.

We can proceed in an analogous way to compute $E(\zeta_t|t) = \hat{\zeta}_t$ and $V(\zeta_t|t) = p_t$ to yield:

$$\hat{\zeta}_t = \hat{\omega}_{2t}\hat{\omega}_{1t}\hat{m}_{t-1} + (1 - \hat{\omega}_{2t}\hat{\omega}_{1t})Y_t, \tag{34}$$

where

$$\hat{\omega}_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \hat{\sigma}_{1t}^2} \tag{35}$$

and

$$p_t = (1 - \hat{\omega}_{2t}\hat{\omega}_{1t})\sigma_{\eta t}^2 + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t}\omega_{1t}), \tag{36}$$

where

$$\begin{aligned} V(\omega_{2t}\omega_{1t}) &\simeq [2\hat{\sigma}_t^6(1 + 2\hat{\beta}_t)^2\hat{\omega}_{1t}^2\hat{\omega}_{2t}^2/R_t \\ &\quad + 2\hat{\sigma}_t^4(1 + \hat{\beta}_t + \hat{\beta}_t^2)\hat{\omega}_{1t}^2\hat{\omega}_{2t}^2/A_t] \\ &\quad \div (\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1})^2. \end{aligned} \tag{37}$$

For the reasons discussed in the case of $V(\omega_{2t})$, if eq. (37) yields a value of $V(\omega_{2t}\omega_{1t}) > 1/12$, then we will truncate it to 1/12.

The effects of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lying on the feasible boundary are similar to those explained in connection with $\hat{m}_t$ and q_t.

VI. BOX CHART AND THE EXCEPTION REPORT

In Ref. 2 Bruce Hoadley has proposed a format for displaying the

conditional distribution of θ_t given data up to time t. He has also proposed exception rules in terms of this distribution. We shall use the same reporting format and the exception rules.

The conditional distribution of θ_t will be summarized by a box chart that shows the 99, 95, 5, and 1 percentiles of the distribution, the best estimate of θ_t, denoted by $\hat{\theta}_t$, the mean level, $\hat{M}_t$, and the current defect index I_t. By applying the inverse square root transformation, we have

$$\hat{M}_t = \hat{m}_t^2 \quad \text{and} \quad \hat{\theta}_t = \hat{\zeta}_t^2.$$

The quantiles of θ_t are once again obtained by squaring the quantiles of ζ_t. Since ζ_t is restricted to be positive, and we have approximated its density by the normal distribution, we may have to truncate some of the extreme quantiles to zero. If we take this fact into account, the desired quantiles of θ_t are:

$$Q_1 = 99\% \text{ quantile} = [\max(\hat{\zeta}_t - 2{\cdot}326\sqrt{p_t}, 0)]^2$$

$$Q_2 = 95\% \text{ quantile} = [\max(\hat{\zeta}_t - 1{\cdot}645\sqrt{p_t}, 0)]^2$$

$$Q_3 = 5\% \text{ quantile} = (\hat{\zeta}_t + 1{\cdot}645\sqrt{p_t})^2$$

$$Q_4 = 1\% \text{ quantile} = (\hat{\zeta}_t + 2{\cdot}326\sqrt{p_t})^2.$$

A sample box chart is shown in Fig. 2.

The exception rules are:

(*i*) Below Normal: A rating class will be declared below normal if the posterior probability of θ_t being larger than one exceeds 0.99.

(*ii*) Alert: An alert will be declared for a rating class if the posterior probability of θ_t being greater than one exceeds 0.95 but it is less than or equal to 0.99.

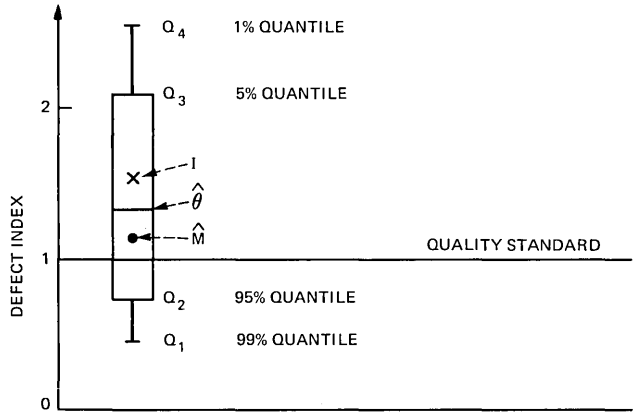


Fig. 2—Sample box chart of the conditional distribution of θ_t.

So in terms of the quantiles derived above, we will declare below normal if $Q_1 > 1$ and alert if $Q_1 \leq 1$ but $Q_2 > 1$.

VII. CHOICE OF STARTING VALUES

The Kalman filter solution described in Section III and the estimation of model parameters described in Section IV are both recursive procedures that must be appropriately initialized. Note that since each of these procedures discounts the past data, the effect of initialization diminishes to zero as more data is accumulated on any rating class. So any biases introduced by the initialization process are transient and temporary. The best way to choose the initial values is by analyzing the historical data on all rating classes. Pending such an analysis, we shall tentatively choose the initial parameter values, as follows.

We will take $\hat{m}_0 = 1.0$ and $\hat{q}_0 = 0.134$. This amounts to choosing a very diffused prior distribution on the mean level. On the square-root defect-index scale the lower and the upper one percentiles of this distribution are 0.149 and 1.851, respectively; while on the defect-index scale the lower and the upper one percentiles are 0.022 and 3.428, respectively. The mean and the median of this distribution are equal to one on either scale. Consistent with this we will choose $Y_0 = 1.0$.

The parameter $\overline{\sigma^2_{\eta,0}}$ should be taken equal to the variance of the transformed defect index associated with the planned equivalent expectancy for a period's sample for the particular rating class. In the present analysis we will take $e_0 = e_1$ and $\overline{\sigma^2_{\eta,0}} = \sigma^2_{\eta,0} = 0.25/e_0$.

The parameter λ determines how many periods of data are effectively used in estimating the time series parameters β_t and σ^2_t and, hence, the parameters σ^2_{1t} and σ^2_{2t}. We will take $\lambda = 0.95$, which implies that effectively $1/(1 - \lambda) = 20$ periods of data are used in estimating the model parameters.

We also need to specify the values of β_0, a_0, $da_0/d\beta$, $S_0(\beta_0)$, e_0, R_0, and A_0. All these variables enter into the recursive maximum likelihood estimation of β and σ^2. We shall take $\beta_0 = -0.6$, which is an approximate midpoint of the feasible range of β. The quantities a_0 and $da_0/d\beta$ will be taken equal to their respective expected values, namely, zero in each case; and A_0 will be set equal to its steady-state value, namely, $1/(1 - \lambda)$. We will take $\nu_0 = 0$, $S_0(\beta_0) = 0.625/[e_0(1 - \lambda)]$, and $R_0 = 20.0/e_0$.

The above starting values imply that at $t = 0$ the mean and the variance of β are respectively -0.6 and 0.063. The variance of the uniform distribution on the $(-1, 0)$ interval is $1/12 = 0.083$. Since the feasible interval for β is smaller than $(-1, 0)$, the variance of 0.063 represents a fairly diffused initialization.

Also, the above starting values imply that the mean and the variance of σ^2 at $t = 0$ are 2.5 $\sigma^2_{\eta,0}$ and $0.625(\sigma^2_{\eta,0})^2$. Therefore, by the gamma

density assumption, the 95-percent confidence interval on σ^2 is (1.2 $\sigma_{\eta 0}^2$, 4.28 $\sigma_{\eta 0}^2$), which is a very wide interval.

The values of σ_1^2 and σ_2^2 implied by the above starting values are 0.5 $\sigma_{\eta 0}^2$ and 0.4 $\sigma_{\eta 0}^2$, respectively.

VIII. ILLUSTRATIVE EXAMPLES

To illustrate the properties of the quality evaluation plan we shall now present six examples. The first three examples are the simulated examples, while the latter three use real audit data.

Example 1: Figure 3a shows the response of QEP to a sudden shift in the quality level. For the first ten periods the observed defect index fluctuates randomly around the fixed level 3.0. From the eleventh to the twentieth period the observed defect index is fixed at 1.0. In each period the expectancy at standard is 5.0. Notice that starting with the eleventh period the estimated mean level rapidly approaches the new mean level. Also, starting with the eleventh observation the product gets off the exception report. Figure 3b shows the corresponding results for QMP. It is clear that in terms of both $\hat{M}_t$ and $\hat{\theta}_t$ the response of QEP is quicker than the response of QMP.

Example 2: Figure 4a displays the response of QEP to a linear trend in the quality level. As in the case of Example 1, for the first ten periods the observed defect index randomly fluctuates around the fixed level 3.0. From the eleventh to the twentieth period the observed defect index has a linear downtrend, as plotted. In all twenty periods the expectancy at standard is 5.0. Notice that both $\hat{M}_t$ and $\hat{\theta}_t$ follow the trend with a small lag. Also note that the QEP algorithm recognizes that the process has a drift rather than random fluctuation. Consequently, $\hat{M}_t$ and $\hat{\theta}_t$ are very close while following the drift. Figure 4b shows the results of QMP for the same data. Here again, $\hat{M}_t$ and $\hat{\theta}_t$ follow the trend, but the lag is much larger. This is manifested in the fact that QEP gets the product off below normal in the seventeenth period while with QMP that happens a period later.

Example 3: This example illustrates that QEP and QMP have similar behaviors when the defect index fluctuates about a fixed value for a long period of time. Figure 5a shows the results of QEP when the defect index fluctuates around the fixed level of 2.0, while Fig. 5b shows the results of QMP with the same data. Note that both the methods declare below normals and alerts in the same periods.

Example 4: Figure 6a gives the data for repaired remreed grids, rating class OC038TT, for periods 7801 through 7904. The periods are numbered 1 through 12 in the figure. The QEP results are also shown in the figure. Similar results with QMP are plotted in Fig. 6b. In response to the drift in the quality we see that QEP attaches a heavier weight to the current data. Consequently, with QEP the mean level, $\hat{M}_t$

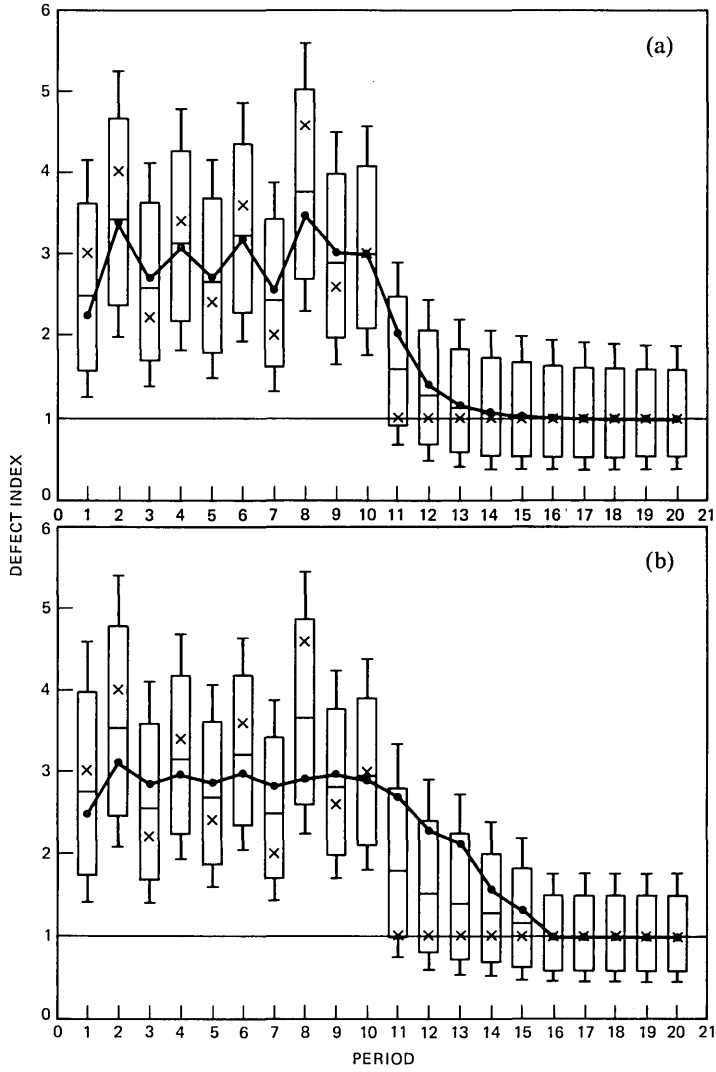


Fig. 3—Response to a sudden shift in the quality level for: (a) QEP, and (b) QMP.

follows the drift more closely than with QMP. In the period 5, recognizing the drift, QEP takes the product off the exception report while QMP still calls it an alert. Also period 7 is an alert according to QMP while, according to QEP, it is off the exception report. These differences between QEP and QMP are clearly seen to be the result of the fact that QEP exponentially discounts the past data, while QMP considers every observation in the six-period window to be equally important.

Example 5: Figures 7a and 7b give the results of QEP and QMP,

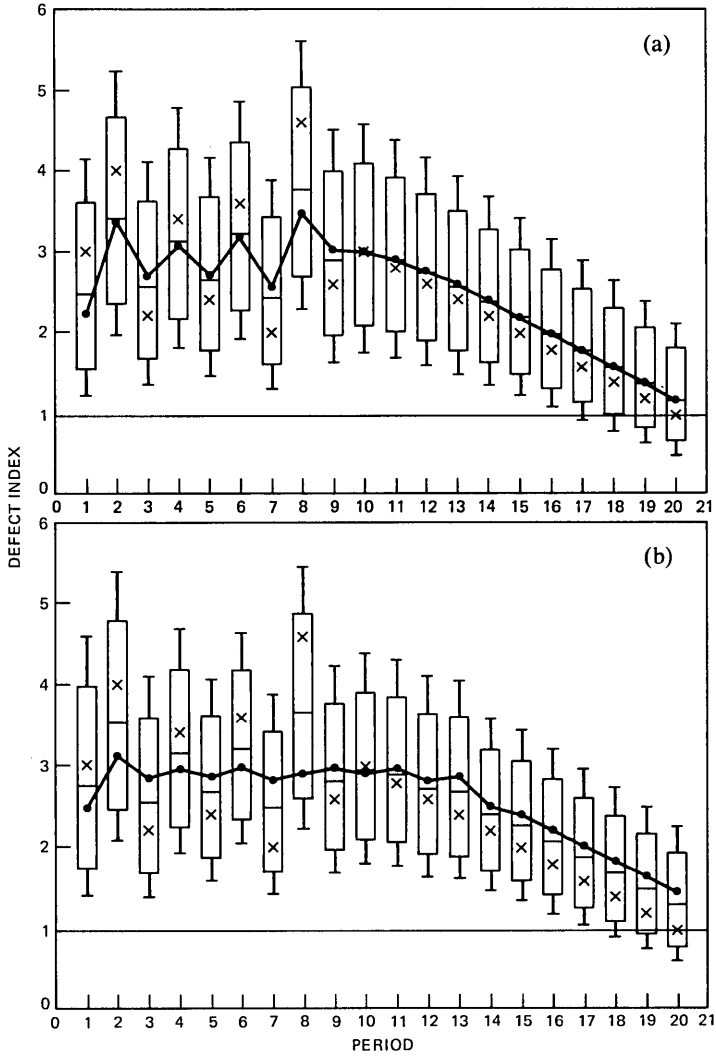


Fig. 4—Response to a linear trend in the quality level for: (a) QEP, and (b) QMP.

respectively, for the rating class OH060CM, consisting of modular telephone chords. The periods covered by the chart are 7701-7808. As we saw in Example 4, QEP follows the drift more closely. In terms of the exception report, there are several differences. In periods 8, 15, and 16 QMP declares below normal, while QEP calls it only an alert. In period 10 QMP calls it an alert while QEP does not declare any exception. These differences are once again a result of the fact that QEP recognizes the drift and hence heavily discounts the past data.

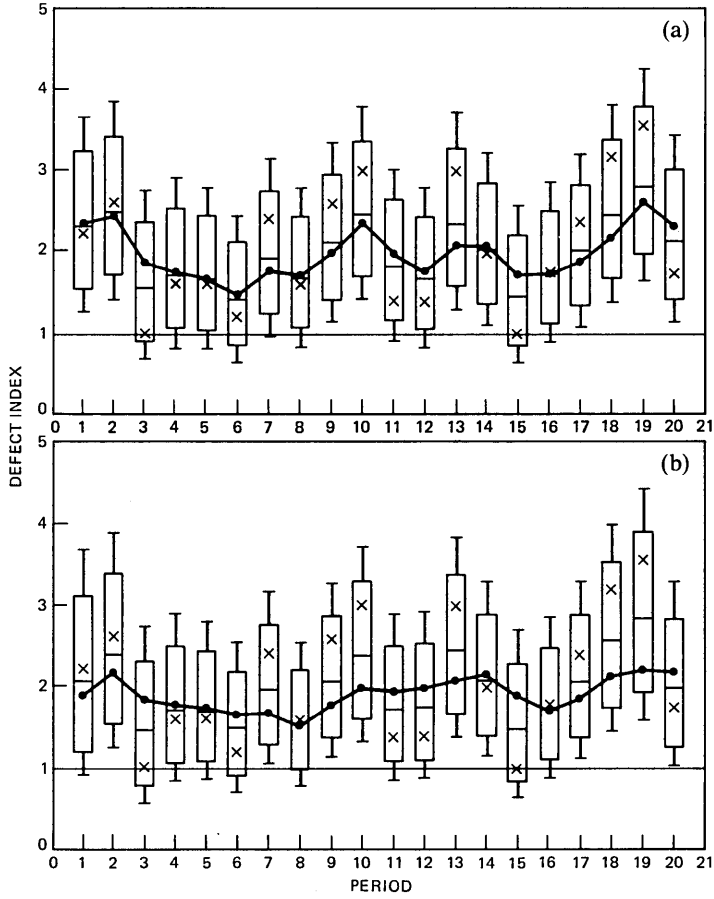


Fig. 5—Response to a random fluctuation in the quality level for: (a) QEP, and (b) QMP.

Example 6: The last example to be considered is the rating class MV104MJ. The results of both the methods are shown in Fig. 8. In this example the quality fluctuates more or less randomly about a fixed mean and, as expected, the two methods give comparable results.

The average values of the weights ω_{1t} and ω_{2t}, and the equivalent expectancies e_t for the three audit examples are tabulated in Table I. Notice that average value of ω_{2t} for OC038TT and OH060CM is 0.48 compared with 0.55 for MV104MJ. This is a direct consequence of the fact that MV104MJ does not exhibit a drift while the others do. The high-frequency fluctuation about the mean function M_t is depicted by ω_{1t}. Relative to the sampling variance $(0.25/e_t)$ OC038TT exhibits a smaller fluctuation than OH060CM. This concurs with the average values of ω_{1t} for these rating classes.

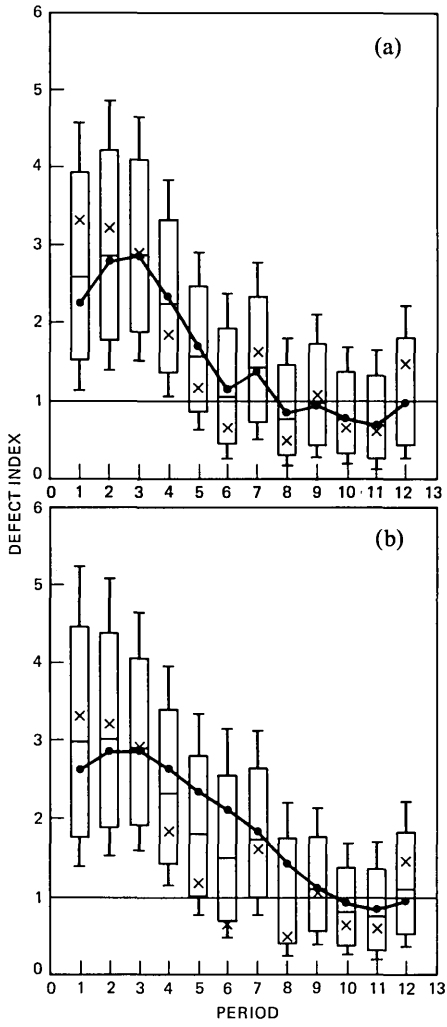


Fig. 6—Results for the rating class OC038TT, periods 7801 through 7904 for (a) QEP, and (b) QMP.

Through the preceding examples it is quite apparent that QEP and QMP could give somewhat different results. Now the key question is: Which method yields a more precise estimate of the unobserved "true defect index"? The only way to answer this question decisively is to take a 100-percent sample of a number of rating classes to find out the true defect indices and compare them with the QEP and QMP results. This is obviously an impossible task. A feasible way to answer the question is by using the models to predict one step ahead and compare the mean-squared prediction errors. Note that $\hat{M}_{t-1}$ is a predictor of I_t.

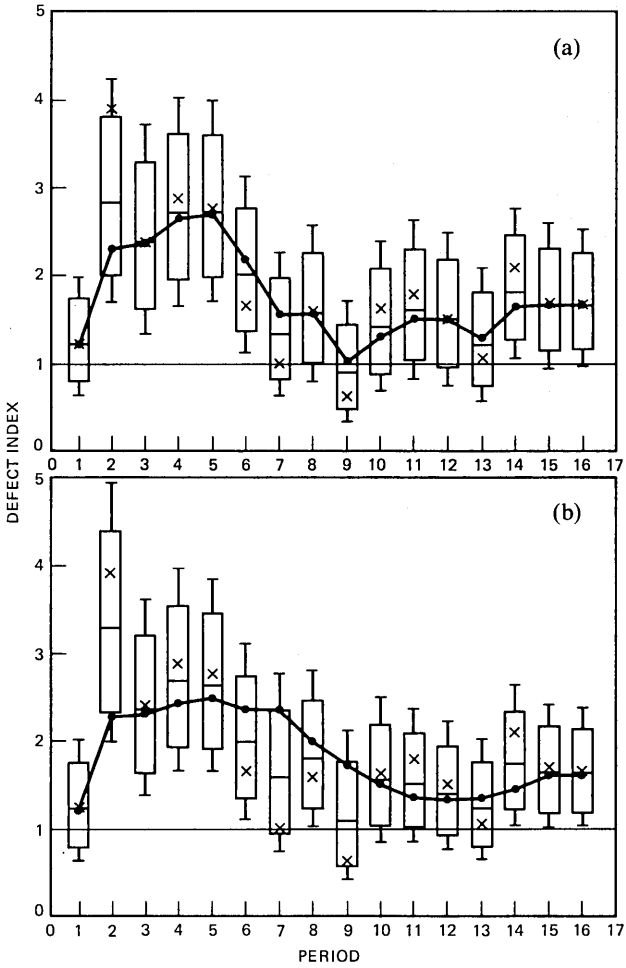


Fig. 7—Results for the rating class OH060CM, periods 7701 through 7808 for: (a) QEP, and (b) QMP.

The mean-squared errors for the three audit data examples, viz. Examples 4, 5, and 6, are given in Table II. For the rating class OC038TT we notice that the mean-squared prediction error (m.s.p.e.) of QMP is 33 percent larger than that of QEP; for OH060CM the m.s.p.e. of QMP is 11 percent larger, and for MV104MJ the m.s.p.e. of QMP is only 3 percent larger. Thus, whenever there is a drift in the quality we may expect QEP to perform better than QMP, whereas if the quality fluctuates randomly around a fixed mean, both QMP and QEP would give similar results.

Effect of truncation: In addition to the numerical examples cited

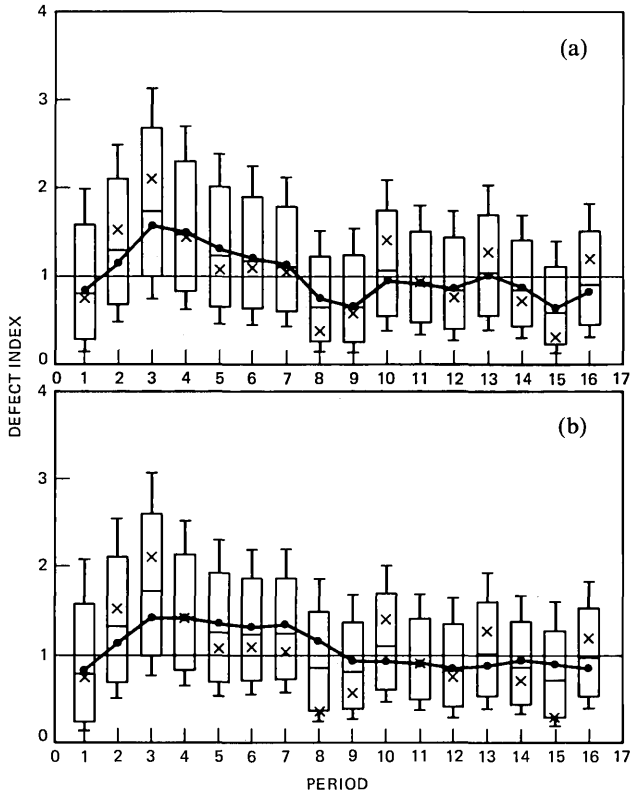


Fig. 8—Results for the rating class MV104MJ, periods 7701 through 7808 for: (a) QEP, and (b) QMP.

above, a limited numerical study was made with thirteen representative rating classes. Each rating class had about fourteen periods of data. This represents a total of 182 test periods. Among these examples, truncation occurred in only 7 percent of the periods. Except for one case, all the truncations caused $\hat{\sigma}_{1t}^2 = 0$. These cases of truncation could be recognized broadly as situations where the variance of the observed defect indices was much smaller than that for the Poisson distribution. In each case of truncation the confidence intervals computed by QEP looked reasonable and comparable to those obtained by QMP, so we can tentatively conclude that the effect of truncation is negligible. Of course, an extensive trial of QEP may suggest some modifications to the truncation rules.

One such modification may be to view the likelihood function of β_t and σ_t^2 as the posterior-probability density function. Then the Bayes estimates of β_t and σ_t^2 may be used in place of the maximum likelihood estimates used in this paper. Because of the complexity of the feasible

Table I—Computed QEP weights for the examples

Rating Class	Average Value of ω_{1t}	ω_{2t}	e_t
OC038TT	0.83	0.48	3.7
OH060CM	0.67	0.48	7.9
MV104MJ	0.75	0.55	4.6

Table II—Comparison of the mean-squared prediction error

Rating Class	Mean Squared Prediction Error QEP	QMP
OC038TT	0.69	0.92
OH060CM	0.94	1.04
MV104MJ	0.29	0.30

region, computing Bayes estimates would involve extensive numerical effort, which may be unnecessary.

IX. DISCUSSIONS

In summary, the QEP model consists of two parts—the system model and the observation model. The system model states that the transformed true-defect index is equal to the process mean that follows the random walk model plus process fluctuation, which is statistically independent from period to period. The random walk model takes care of the process drift. The observation model states that the transformed observed defect index is equal to the transformed true-defect index plus sampling error with a known variance. The different parameters of the QEP model are estimated from the observed data by the recursive, exponentially discounted, maximum likelihood method. The successive transformed true defect indices and the process mean levels are then estimated by the adaptive Kalman filtering algorithm.

From the derivation of the plan and the illustrative examples the following advantages of QEP are apparent:

(*i*) The QEP model takes into account the time order of the observations, while in QMP the time order of the observations is ignored.

(*ii*) The best estimate of the process mean level is obtained by an adaptive exponential smoothing procedure. This makes QEP more responsive to the shifts and drifts in the process level. This is evidenced by the lower mean-squared prediction error for the examples discussed in Section VIII.

(*iii*) The QMP model is a special case of the QEP model. However, the two algorithms are quite different.

(*iv*) The computations are recursive. The entire past data are summarized by ten numbers.

(*v*) The computational efforts of QEP and QMP algorithms are comparable.

In the light of the advantages listed above it is proposed that QEP be considered as a serious alternative to QMP for official rating. In preparation for using QEP it is suggested that it be tried on all rating classes for a number of rating periods, and the resulting exception reports carefully compared with those from the QMP and the *t*-rate system. Such a study would aid us in fine tuning the starting conditions, quantifying the effect of truncation, and perhaps in making some other minor modifications for improving the performance of the QEP.

For small expectancies, the square root transform of the Poisson distribution has a significantly different variance than 0.25, assumed in Section II. Since the audit samples can at times be very small it would be necessary to use the correct variances. A study of this aspect will be done in a later memorandum.

The adaptive Kalman filtering methodology derived in this paper, with appropriate extensions and modifications, can be put to many other applications. In the field of quality control, Phadke[7] had developed a sequential empirical Bayes acceptance sampling plan. The adaptive Kaman filtering method developed in this paper would be particularly suited for updating the empirical prior distribution. Another potential application is in combining the traditional $\bar{X}$ and R control charts into a single box chart. Here the adaptive Kalman filter would permit one to take into account serial correlation in the data as well as process drifts and shifts, and changes in the process variance. Yet another application is in adaptive time series forecasting.

X. ACKNOWLEDGMENTS

I thank A. B. Hoadley for several discussions and suggestions at various stages of this project. I also thank Lakshman Sihna for useful discussions on the Kalman filtering theory. The box plots were generated using S. G. Crawford's software. Dolat Salsman and R. A. Cayford helped in developing the QEP computer program.

REFERENCES

1. H. F. Dodge, "A Method of Rating Manufactured Products," B.S.T.J. (April 1928), pp. 350–68.
2. A. B. Hoadley, "The Quality Measurement Plan," B.S.T.J., *60,* No. 2 (February 1981), pp. 215–273.
3. A. B. Hoadley, unpublished work.
4. Arthur Gelb, ed., *Applied Optimal Estimation*, Cambridge, MA: M.I.T. Press, 1974.

5. A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.
6. M. S. Phadke, unpublished work.
7. M. S. Phadke, unpublished work.

APPENDIX A

Derivation of the Kalman filter solution

Let the conditional distribution of m_{t-1} given data up to time $t-1$ be normal with mean $\hat{m}_{t-1}$ and variance q_{t-1}, i.e.,

$$m_{t-1}|t-1 \sim N(\hat{m}_{t-1}, q_{t-1}). \tag{38}$$

Eq. (2) expresses m_t as a sum of two independent normal random variables m_{t-1} and v_{2t}. Since the mean and variance of v_{2t} are respectively 0 and σ_{2t}^2, it follows that

$$m_t|t-1 \sim N(\hat{m}_{t-1}, \sigma_{2t}^2 + q_{t-1}). \tag{39}$$

Substituting eq. (1) in eq. (3) we have

$$Y_t = m_t + v_{1t} + \eta_t, \tag{40}$$

which implies that

$$Y_t|m_t \sim N(m_t, \sigma_{1t}^2 + \sigma_{\eta t}^2). \tag{41}$$

In the Bayesian framework we may view eq. (39) as a prior distribution on m_t, and Y_t as an observation of m_t with the distribution specified by eq. (41). Applying the Bayes theorem the distribution of m_t conditional on data up to time t is seen to be

$$\begin{aligned} f(m_t|t) &\propto \exp\left\{-\frac{(m_t-\hat{m}_{t-1})^2}{2(\sigma_{2t}^2+q_{t-1})} - \frac{(Y_t-m_t)^2}{2(\sigma_{1t}^2+\sigma_{\eta t}^2)}\right\} \\ &\propto \exp\left\{-\frac{(m_t-\hat{m}_t)^2}{2q_t}\right\}, \end{aligned} \tag{42}$$

where

$$\hat{m}_t = \omega_{2t}\hat{m}_{t-1} + (1-\omega_{2t})Y_t, \tag{4}$$

$$\omega_{2t} = (\sigma_{1t}^2 + \sigma_{\eta t}^2)/(\sigma_{1t}^2 + \sigma_{\eta t}^2 + \sigma_{2t}^2 + q_{t-1}), \tag{5}$$

and

$$q_t = (1-\omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2). \tag{6}$$

From eq. (42) it can be inferred that the distribution of m_t conditional on data up to time t is normal with mean $\hat{m}_t$ and variance q_t.

Equations (7) through (9), used for computing the conditional distribution of ζ_t, can be derived analogously as follows. First by substituting eq. (2) in (1) we have

$$\zeta_t = m_{t-1} + \nu_{1t} + \nu_{2t}; \tag{43}$$

hence,

$$\zeta_t | t-1 \sim N(\hat{m}_{t-1}, \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}). \tag{44}$$

From eq. (3) we have

$$Y_t | \zeta_t \sim N(\zeta_t, \sigma_{\eta t}^2). \tag{45}$$

Treating eq. (44) as a prior distribution for ζ_t and applying the Bayes theorem, we readily obtain the distribution of ζ_t conditional on data up to time t as

$$f(\zeta_t | t) \propto \exp\left\{-\frac{(\zeta_t - \hat{m}_{t-1})^2}{2(\sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1})} - \frac{(Y_t - \zeta_t)^2}{2\sigma_{\eta t}^2}\right\}$$

$$\propto \exp\left\{-\frac{(\zeta_t - \hat{\zeta}_t)^2}{2p_t}\right\}, \tag{46}$$

where

$$\hat{\zeta}_t = \omega_{1t}\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{1t}\omega_{2t})Y_t, \tag{7}$$

$$\omega_{1t}\omega_{2t} = \sigma_{\eta t}^2/(\sigma_{\eta t}^2 + \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}), \tag{8}$$

and

$$p_t = (1 - \omega_{1t}\omega_{2t})\sigma_{\eta t}^2. \tag{9}$$

Thus, the conditional distribution of ζ_t on data up to time t is normal with mean $\hat{\zeta}_t$ and variance p_t. Equations (7) through (9) form the desired recursive equations for computing $\hat{\zeta}_t$ and p_t.

APPENDIX B

Estimation of β and σ^2

Given the observed transformed defect indices $Y_0, Y_1, Y_2, \cdots, Y_n$ one can compute $Z_t = Y_t - Y_{t-1}$ for $t = 1, \cdots, n$. The Z_t series follows $MA(1)$ model given by eq. (13). The exponentially discounted probability density function of $a_1, \cdots, a_t$ is given by

$$p(a_1, \cdots, a_t | \sigma^2) = \prod_{j=1}^{t} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-a_j^2/2\sigma^2)\right]^{\lambda_j}, \tag{47}$$

where $\lambda_j = \lambda^{t-j}$. Thus, the exponentially discounted probability density function of $Z_1, \cdots, Z_t$ conditional on the knowledge of a_0, β, and σ^2 is given by

$$p(Z_1, \cdots, Z_t | a_0, \beta, \sigma^2) = \prod_{j=1}^{t} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-a_j^2/2\sigma^2)\right]^{\lambda_j}, \tag{48}$$

where a_j is related to Z_j and β via the recursion relation in eq. (13), i.e.,

$$a_j = Z_j - \beta a_{j-1}. \tag{49}$$

Thus, the conditional, exponentially discounted, log-likelihood function is

$$L_t(\beta, \sigma^2) = -\tfrac{1}{2}(A_t \ln \sigma^2 + S_t/\sigma^2), \tag{50}$$

where

$$A_t = \sum_{j=1}^{t} \lambda^{t-j}, \tag{51}$$

and

$$S_t = \sum_{j=1}^{t} \lambda^{t-j} a_j^2. \tag{52}$$

By differentiating eq. (50) with respect to β and σ^2 and equating the derivatives to zero, it can be shown that L_t is maximum at $(\hat{\beta}_t, \hat{\sigma}_t^2)$, where $\hat{\beta}_t$ is the minimum point of $S_t(\beta)$ and $\hat{\sigma}_t^2 = S_t(\hat{\beta}_t)/A_t$.

In the neighborhood of a point β_0, we can approximate a_j by the linear function:

$$a_j(\beta) = a_j(\beta_0) + (\beta - \beta_0) \left. \frac{da_j(\beta)}{d\beta} \right|_{\beta_0}. \tag{53}$$

Substituting this approximation in eq. (52) we have

$$S_t(\beta) \simeq S_t(\beta_0) + (\beta - \beta_0)\nu_t + \tfrac{1}{2}(\beta - \beta_0)^2 R_t, \tag{54}$$

where ν_t, R_t and $S_t(\beta_0)$ obey the recursion relations shown in eqs. (20) through (26). It is easy to verify that $\hat{\beta}_t$, given by eq. (18) minimizes the approximate $S_t(\beta)$ of eq. (54), and eq. (19) gives $\hat{\sigma}_t^2$.

The matrix of second partial derivatives of L_t is

$$-\begin{bmatrix} \dfrac{R_t}{2\sigma^2} & 0 \\ 0 & \dfrac{A_t}{2\sigma^4} \end{bmatrix},$$

so by the Fisher-information theory, the estimates $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ are uncorrelated and their approximate variances are as given by eqs. (27) and (28).

The above recursive procedure also has a Bayesian interpretation, as given by Phadke.[6]

APPENDIX C

Summary of the Formulae

C.1 Initial conditions

$$\hat{m}_0 = 1.0$$

$$q_0 = 0.134$$

$$Y_0 = 1.0$$

$$\beta_0 = -0.6$$

$$\lambda = 0.95$$

$$a_0 = \frac{da_0}{d\beta} = 0$$

$$e_0 = e_1$$

$$S_0(\beta_0) = \frac{0.625}{e_0(1-\lambda)}$$

$$\nu_0 = 0$$

$$R_0 = 20.0/e_0$$

$$A_0 = 1/(1-\lambda)$$

$$\overline{\sigma_{\eta,0}^2} = 0.25/e_0$$

C.2 Recursive formulae

$$I_t = x_t/e_t$$

$$Y_t = \sqrt{I_t}$$

$$Z_t = Y_t - Y_{t-1}$$

$$\sigma_{\eta,t}^2 = 0.25/e_t$$

$$a_t = Z_t - \beta_0 a_{t-1}$$

$$\frac{da_t}{d\beta} = -a_{t-1} - \beta_0 \frac{da_{t-1}}{d\beta}$$

$$S_t(\beta_0) = \lambda S_{t-1}(\beta_0) + a_t^2$$

$$\nu_t = \lambda \nu_{t-1} + 2a_t \frac{da_t}{d\beta}$$

$$R_t = \lambda R_{t-1} + 2\left(\frac{da_t}{d\beta}\right)^2$$

$$A_t = \lambda A_{t-1} + 1$$

$$\beta_t^* = \beta_0 - R_t^{-1}\nu_t, \qquad -1 \le \beta_t^* \le 0$$

$$S_t(\beta_t^*) = S_t(\beta_0) + (\beta_t^* - \beta_0)\nu_t + \tfrac{1}{2}(\beta_t^* - \beta_0)^2 R_t$$

$$\sigma_t^{*2} = S_t(\beta_t^*)/A_t$$

$$\overline{\sigma^2_{\eta,t}} = \lambda\, \overline{\sigma^2_{\eta,t-1}} + (1-\lambda)\sigma^2_{\eta,t}$$

$$\hat{\sigma}^2_{1t} = -\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta t}}, \qquad \hat{\sigma}^2_{1t} \geq 0$$

$$\hat{\sigma}^2_{2t} = (1+\beta^*_t)^2 \sigma^{*2}_t$$

If $-\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta t}} \geq 0$, then

$$\hat{\beta}_t = \beta^*_t \;\&\; \hat{\sigma}^2_t = \sigma^{*2}_t.$$

If $-\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta n}} < 0$, then

$$\hat{\beta}_t = \frac{-(2+\hat{\sigma}^2_{2t}/\overline{\sigma^2_{\eta t}}) + \sqrt{(2+\hat{\sigma}^2_{2t}/\overline{\sigma^2_{\eta t}})^2 - 4}}{2} \text{ and}$$

$$\hat{\sigma}^2_t = -\overline{\sigma^2_{\eta t}}/\hat{\beta}_t.$$

$$\hat{\omega}_{1t} = \frac{\sigma^2_{\eta t}}{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}}$$

$$\hat{\omega}_{2t} = \frac{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}}{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t} + \hat{\sigma}^2_{2t} + q_{t-1}}$$

$$\hat{m}_t = \hat{\omega}_{2t}\hat{m}_{t-1} + (1-\hat{\omega}_{2t})\, Y_t$$

$$\hat{\xi}_t = \hat{\omega}_{1t}\hat{m}_t + (1-\hat{\omega}_{1t})\, Y_t$$

$$V(\omega_{2t}) = \frac{\{2\hat{\sigma}^6_t[1+\hat{\omega}_{2t}(1+2\hat{\beta}_t)]^2/R_t\} + \{2\hat{\sigma}^4_t[\hat{\beta}_t + (1+\hat{\beta}_t+\hat{\beta}^2_t)\hat{\omega}_{2t}]^2/A_t\}}{(\hat{\sigma}^2_{1t} + \hat{\sigma}^2_{2t} + \sigma^2_{\eta t} + q_{t-1})^2}; \; V(\omega_{2t}) \leq \frac{1}{12}$$

$$V(\omega_{1t}\omega_{2t}) = \frac{[2\hat{\sigma}^6_t(1+2\hat{\beta}_t)^2\hat{\omega}^2_{1t}\hat{\omega}^2_{2t}/R_t] + [2\hat{\sigma}^4_t(1+\hat{\beta}_t+\hat{\beta}^2_t)\hat{\omega}^2_{1t}\hat{\omega}^2_{2t}/A_t]}{(\hat{\sigma}^2_{1t} + \hat{\sigma}^2_{2t} + \sigma^2_{\eta t} + q_{t-1})^2}; \; V(\omega_{1t}\omega_{2t}) \leq \frac{1}{12}$$

$$q_t = (1-\hat{\omega}_{2t})(\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}) + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t})$$

$$p_t = (1-\hat{\omega}_{1t}\hat{\omega}_{2t})\sigma^2_{\eta t} + (Y_t - \hat{m}_{t-1})^2 V(\omega_{1t}\omega_{2t})$$

C.3 Points for the box chart

Current defect index : I_t

Best estimate of the defect index: $\hat{\theta}_t = \hat{\xi}^2_t$

The mean level : $\hat{M}_t = \hat{m}^2_t$

99% quantile : $Q_1 = [\max(\hat{\xi}_t - 2.326\sqrt{p_t}, 0)]^2$

95% quantile : $Q_2 = [\max(\hat{\xi}_t - 1.645\sqrt{p_t}, 0)]^2$

5% quantile : $Q_3 = (\hat{\xi}_t + 1.645\sqrt{p_t})^2$

1% quantile : $Q_4 = (\hat{\xi}_t + 2.326\sqrt{p_t})^2$

CONTRIBUTORS TO THIS ISSUE

David Anick, B.S., 1976 (Mathematics), M.I.T.; Ph.D., 1980 (Mathematics), M.I.T., 1980—. Mr. Anick is presently at the University of California, Berkeley. He has worked in algebraic topology and is actively conducting research in noncommutative graded algebras. Member, AMS.

Jerry E. Bernardini, RCA Institutes, 1963; B.S.E.E., Fairleigh Dickinson University, 1969; M.S.E.E., Stanford University, 1970; Bell Laboratories, 1963—. Mr. Bernardini was initially involved with circuit design for the Ocean System Departments and then with design for the Safeguard System Project. From 1975 to 1979 he was involved with the design of magnetic storage systems. He is presently engaged in data communication product evaluation with the Business Services Department of AT&T. Member, IEEE.

Ding Z. Du, 1967, Middle School, Qiqihaer, China. Mr. Du pursued independent studies of mathematics and became a graduate student at the Institute of Applied Mathematics, Academia Sinica, Beijing, China, in 1978. He is currently interested in mathematical programming and combinatorics. Mr. Du worked with F. K. Hwang of Bell Laboratories in 1980, while Mr. Hwang taught at the Institute of Applied Mathematics, Beijing, China.

Richard D. Gitlin, B.E.E., 1964, City College of New York; M.S., 1965, and D. Eng. Sc., 1969, Columbia University; Bell Laboratories, 1969—. Mr. Gitlin is Head of the Advanced Data Systems Department in the Data Communications Laboratory. He is a member of the Communication Theory Committee of the IEEE Communications Society, Editor for Communication Theory of the IEEE Transactions on Communications, and is a member of the Editorial Advisory Board of the Proceedings of the IEEE. Senior Member, IEEE; Member, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

Daniel P. Heyman, B. Mgt. E., (Management Engineering), 1960, Rensselaer Polytechnic Institute; M.I.E. (Industrial Engineering), 1962, Syracuse University; Ph.D. (Operations Research), 1966, University of California, Berkeley; U.S. Air Force, 1960–1963; Bell Laboratories, 1966—. Mr. Heyman has worked on a variety of operations research problems at Bell Laboratories, including studies of queues, inventories, and financial markets. He was a visiting faculty member at Yale University for the 1976–1977 academic year. Member, Sigma Xi, Alpha Pi, Mu, Operations Research Society of America.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., 1964, City University of New York; M.E.S., 1966, Ph.D., 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang has been engaged in research in discrete mathematics, computing algorithms, statistical theory, and switching networks. Member, SIAM.

D. L. Jagerman, B.E.E. (Mathematics), 1949, Cooper Union; M.S., and Ph.D. (Mathematics), 1954 and 1962, respectively, New York University; Bell Laboratories, 1963—. Mr. Jagerman has been engaged in mathematical research on quadrature, interpolation, and approximation theory, especially related to the theory of widths and metrical entropy, with application to the storage and transmission of information. For the past several years, he has worked on the theory of difference equations and queueing, especially with reference to traffic theory and computers.

Karel Janac, M.S. (Electrical Engineering), 1950, Ph.D. (Technical Science), 1956, Technical University in Prague. For 13 years, up to 1968, he was associated with the Czechoslovak Academy of Sciences working on problems related to the experimental solution of probabilistic problems, random process generation, computers, network analysis, and synthesis and solution of nonlinear problems. He also taught at the Charles University in Prague. From 1968 through 1977 he was associated with Electronic Associates, Inc., where he was involved in scientific computations, simulation, system analysis, and development of algorithms for parallel computers. At Bell Laboratories, Mr. Janac has worked on stiff communication problems and circuit analysis programs. He is presently engaged in studies of digital communication networks and related performance analysis. He is the coauthor of two books, *Solution of Non-Linear System* and *Equipment for Random Process Generation.*

Manu Malek-Zavarei, B.S.E.E., 1965, University of Tehran; M.S.E.E., 1968, and Ph.D., 1970, University of California, Berkeley.

Mr. Malek-Zavarei first joined the Traffic Studies Center at Bell Laboratories in 1970. He left in 1971 to join Shiraz University (formerly Pahlavi University) in Iran. He was a Professor and the Chairman of the Electrical Engineering Department when he left Shiraz University in June 1980. He spent the academic year 1976–77 with Hughes Aircraft Company in Canoga Park, CA on sabbatical leave from Shiraz University. During the academic year 1980–1981, he was a Visiting Professor with the Department of Electrical and Computer Engineering, University of New Mexico. He rejoined Bell Laboratories in June 1981 where he is presently working on the development of generic design techniques for special-service circuits. Mr. Malek-Zavarei is the

author of a book, *Telephone Switching Systems* (in Persian, Shiraz University Press, 1979), and author or co-author of over 30 technical papers in the areas of communication networks, control systems, and optimal control, particularly of time-delay systems. He is a Senior Member of IEEE and a member of Sigma Xi and Eta Kappa Nu Societies.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematics and Statistics Research Center. Member, IEEE.

Howard C. Meadors, Jr., S.B., 1960, S.M., 1962, E.E., 1964, Massachusetts Institute of Technology; Ph.D., 1976, Polytechnic Institute of New York; Bell Laboratories, 1966—. Mr. Meadors has been primarily concerned with the design of automatic, adaptive equalizers for high-speed data communications. He is a member of the IEEE, Eta Kappa Nu, Sigma Xi.

Debasis Mitra, B.Sc., 1965, and Ph.D., 1967 (Electrical Engineering), London University; United Kingdom Atomic Energy Authority Research Fellow, 1966–1967; Bell Laboratories, 1968—. Mr. Mitra has worked on the stability analysis of nonlinear systems, semiconductor networks, growth models for new communication systems, speech waveform coding, nonlinear phenomenon in digital signal processing, adaptive filtering, and network synchronization. Most recently, he has been involved in the analytic and computational aspects of stochastic networks and computer communications. He is a Supervisor in the Mathematics of Physics and Networks Department. Senior member, IEEE, SIAM.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, Ph.D. (Electrical Engineering), 1970, Rice University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At Bell Laboratories, he has worked on various aspects of signal processing. He is presently Head of the Visual Communication Research Department and a Visiting Professor in the Department of Electrical Engineering

at Rutgers University. Member, Tau Beta Pi, Sigma Xi; Senior Member, IEEE.

Kinichiro Ogawa, B.S.E.E., 1966, M.S.E.E., 1968, University of Tokyo, D.Sc. (Electrical Engineering), Washington University; Nippon Telegraph & Telephone Public Corporation, 1968–1976; Bell Laboratories, 1976—. At N.T.T., Mr. Ogawa worked on long-haul analog coaxial cable systems and on video transmission. He also worked on the development of pulse-code modulated/frequency division multiplex digital systems. At Bell Laboratories, he has been involved in research on optical data links and digital lightwave systems. Member, IEEE, Optical Society of America, IECE of Japan.

Madhav Phadke, B. Tech (Mechanical Engineering), 1969, Indian Institute of Technology, Bombay; M.S. (Statistics), 1972, and Ph.D. (Mechanical Engineering), 1973, University of Wisconsin at Madison. Prior to joining Bell Laboratories, Mr. Phadke worked as a research associate in the Statistics Department of the University of Wisconsin and as a Visiting Scientist at the IBM Watson Research Center. He has published articles in professional journals covering topics in system identification, process control, time series analysis, mechanical design, quality assurance, and air pollution data analysis. His current interests are in the applications of Japanese methods of quality control and productivity improvement in design, development, and manufacturing of products. Member, Phi Kappa Phi, Sigma Xi, American Statistical Association, the Americal Society for Quality Control.

John D. Robbins, B.S.E.E. (Highest Honors), 1976, M.S., 1981, Rutgers University; Bell Laboratories, 1976—. Mr. Robbins is currently a Member of the Technical Staff in the Visual Communications Research Department and a doctoral student at Rutgers University. His research interests include picture processing, parallel and distributed computation, and graphics. Member, Tau Beta Pi, Eta Kappa Nu.

Charles S. Roberts, B.S. (Chemistry), 1959, Carnegie-Mellon University, Pittsburgh, PA; Ph.D. (Physics), 1963, Massachusetts Institute of Technology, Cambridge, MA; Bell Laboratories, 1963—. From 1963 to 1968 Mr. Roberts was a member of the Radiation Physics Research Department at Bell Laboratories, where he conducted research on the Van Allen Belts. He made contributions to the theory of energetic electron loss and lifetime in the Van Allen belts and was a coexperimenter on five NASA-sponsored earth satellite experiments designed to measure characteristics of the belts. In 1968, he became Supervisor, Operating Systems Group in the Murray Hill Computation Center; in

1969 and 1970 he served as Head of the Murray Hill Computation Center. In October, 1970, he returned to the Research Area as Head, Information Processing Research Department, and in September, 1973, he assumed his current position as Head of the Interactive Computer Systems Research Department at Bell Laboratories, Holmdel, NJ. He is the author of twenty technical papers on Van Allen belts physics and information processing and he holds two patents on information processing inventions. His current research interests include computer communications, operating systems, computer architecture, information retrieval, and data base management. Member IEEE, the Association for Computing Machinery, the American Physical Society, and Sigma Xi.

James E. Rowley, B.S.E.E., 1970, Iowa State University; M.S.E.E., 1971, Purdue University; Bell Laboratories, 1970—. Mr. Rowley is engaged in development of voice response systems. Member, Eta Kappa Nu, Tau Beta Pi, IEEE.

N. J. A. Sloane was born in Beaumaris, Wales, on October 10, 1939. He received B.E.E. and B.A. degrees from the University of Melbourne in 1959 and 1960, respectively, and M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1964 and 1967, respectively.

From 1956 to 1961 he worked for the Postmaster General's Department of the Commonwealth of Australia. From 1963 to 1965 he was a Research Assistant with the Cognitive Systems Research Program at Cornell University, and was an instructor in electrical engineering at Cornell University from 1966 to 1967. In 1967 he became an assistant professor of electrical engineering at Cornell, and remained at that post until 1969. Since 1969 he has been a Member of the Technical Staff at Bell Laboratories, engaged in research in coding theory, communication theory, and combinatorial mathematics. He is the author of four books: *A Handbook of Integer Sequences* (New York: Academic, 1973); *A Short Course on Error-Correcting Codes* (New York: Springer-Verlag, 1975); (with F. J. MacWilliams) *The Theory of Error-Correcting Codes* (Amsterdam: North Holland Publishing Company, 1977); and (with M. Harwit) *Hadamard Transform Optics* (New York: Academic, 1979).

Mr. Sloane is a Fellow of the IEEE and a member of the American Mathematical Society and the Mathematical Association of America. He was awarded the Chauvanet Prize in 1979 by the Mathematical Association of America. He was the editor of IEEE Transactions on Information Theory from 1978 to 1980.

Peter W. Smith, BSc. (Mathematics and Physics), 1958, MSc.,

1961, Ph.D., 1964 (Physics), McGill University, Montreal, P. Q., Canada; Canadian Marconi Company, 1958–1959; Bell Laboratories, 1964—. Mr. Smith has conducted research on mode selection and mode-locking of lasers. He pioneered the development of waveguide gas lasers and dye vapor laser systems. Recently he developed and demonstrated hybrid bistable optical devices, and is currently involved in studies of ultra-rapid optical switching elements. In 1970 he spent nine months at the University of California, Berkeley, as Visiting MacKay Lecturer in the Department of Electrical Engineering; in 1978–1979 he was a visiting research scientist at the Laboratoire d'Optique Quantique, Ecole Polytechnique, Palaiseau, France.

Fellow IEEE, fellow, Optical Society of America. Member American Physical Society, associate editor of Optics Letters. He served as associate editor of the IEEE Journal of Quantum Electronics from 1976 to 1979, and was guest editor for a special issue of the IEEE Journal of Quantum Electronics on Optical Bistability (March 1981). Treasurer, IEEE Quantum Electronics and Applications Society, Chairman, Committee on Planning and Finances.

Man Mohan Sondhi, B.Sc. (Physics), Honours degree, 1950, Delhi University, Delhi, India, D.I.I.Sc. (Communications Engineering), 1953, Indian Institute of Science, Bangalore, India; M.S., 1955; Ph.D. (Electrical Engineering), 1957, University of Wisconsin, Madison, Wisconsin; Bell Laboratories, 1962—. Before joining Bell Laboratories, Mr. Sondhi worked for a year at the Central Electronics Engineering Research Institute, Pilani, India and taught for a year at the University of Toronto. At Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modelling of auditory and visual processes, and acoustical inverse problems. From 1971 to 1972 Mr. Sondhi was a guest scientist at the Royal Institute of Technology, Stockholm, Sweden.

Carl-Erik W. Sundberg, M.S.E.E., 1966, and Dr. Techn., 1975, Lund Institute of Technology, University of Lund, Sweden; Bell Laboratories, 1981—. Mr. Sundberg is an Associate Professor in the Department of Telecommunication Theory, University of Lund, and a consultant in his field. He is Director of the consulting company SUNCOM, Lund. During 1976 he was with the European Space Research and Technology Centre (ESTEC), Noordwijk, The Netherlands, as an ESA Research Fellow. He has been a Consulting Scientist at LM Ericsson and SAAB-SCANIA, Sweden, and at Bell Laboratories. His research interests include source coding, channel coding (especially decoding techniques), digital modulation methods, fault-tolerant systems, digital mobile radio systems, spread spectrum sys-

tems, and digital satellite communication systems. He has published a large number of papers in these areas during the last few years. Senior Member, IEEE; member, SER, Sveriges Elektroingenjörers Riksförening.

Steven B. Weinstein, B.S.E.E., 1960, Massachusetts Institute of Technology; M.S.E.E., 1962, University of Michigan; Ph.D. (E.E.), 1966, University of California at Berkeley; Philips Research Laboratories, Eindhoven, Netherlands, 1967–1968; Bell Laboratories, 1968–1980. Mr. Weinstein's technical interests include data communications, data communications security, and microcomputer systems. Senior member. IEEE.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Aho A. V., Wyner A. D., Yannakakis M., Ullman J. D., **Bounds on the Size and Transmission Rate of Communication Protocols.** Comput Math 8(3):205–214, 1982.

Anderson T. W., Hsiao C., **Formulation and Estimation of Dynamic-Models Using Panel Data.** J Econom 18(1):47–82, 1982.

Bassett G., Koenker R., **An Empirical Quantile Function for Linear-Models with IID Errors.** J Am Stat A 77(378):407–415, 1982.

Beutler F. J., Melamed B., **Multivariate Poisson Flows on Markov Step Processes.** J Appl Prob 19(2):289–300, 1982.

Boyles S. M., Exoo G., **A Counterexample to a Conjecture on Paths of Bounded Length.** J Graph Th 6(2):205–209, 1982.

Conway J. H., Sloane N. J. A., **23 Constructions for the Leech Lattice.** P Roy Soc A 381(1781):275–283, 1982.

Doverspike R. D., **Some Perturbation Results for the Linear Complementary-Problem.** Math Progr 23(2):181–192, 1982.

Du D. Z., Hwang F. K., **Balanced Howell Rotations of the Twin Prime Power Type.** T Am Math S 271(2):415–421, 1982.

Du D. Z., Hwang F. K., **Symmetric Skew Balanced Starters and Complete Balanced Howell Rotations.** T Am Math S 271(2):409–413, 1982.

Elken T. R., **Combining a Path Method and Parametric Linear-Programming for the Computation of Competitive Equilibria.** Math Progr 23(2):148–169, 1982.

Halfin S., **Linear Estimators for a Class of Stationary Queuing-Processes.** Operat Res 30(3):515–529, 1982.

Hawkins D. T., **Online Information-Retrieval Bibliography—5th Update.** Online Rev 6(2):147–208, 1982.

Hohenberg P. C., Langer J. S., **Non-Equilibrium Phenomena—Outlines and Bibliographies of a Workshop.** J Stat Phys 28(1):193–226, 1982.

Kantor W. M., **Spreads, Translation-Planes and Kerdock Sets.** SIAM J Alg 3(2):151–165, 1982.

Morrison J. A., **The Solution of an Integral-Equation Related to Optimal Linear-Estimation Problem.** SIAM J A Ma 42(3):588–607, 1982.

Odlyzko A. M., **Periodic Oscillations of Coefficients of Power-Series That Satisfy Functional-Equations.** Adv Math 44(2):180–205, 1982.

Sethi R., **Pebble Games for Studying Storage Sharing.** Theor Comp 19(1):69–84, 1982.

Slepian D., **Solution of Certain Integral-Equations With Difference Kernels.** SIAM J Num 19(3):614–622, 1982.

ECONOMICS

Bawa V. S., **Stochastic-Dominance—a Research Bibliography.** Manag Sci 28(6):698–712, 1982.

Jordan J. S., Radner R., **Rational-Expectations in Microeconomic Models—Overview.** J Econ Theo 26(2):201–223, 1982.

Jovanovic B., **Inflation and Welfare in the Steady-State.** J Polit Ec 90(3):561–577, 1982.

Jovanovic B., **Selection and the Evolution of Industry.** Econometric 50(3):649–670, 1982.

Rosenthal R. W., **A Dynamic-Model of Duopoly With Customer Loyalties.** J Econ Theo 27(1):69–76, 1982.

ENGINEERING

Alferness R. C., **Low-Loss Fiber-Coupled Wave-Guide Directional Coupler Modulator.** Electr Lett 18(12):490–491, 1982.

Boyd G. D., Cheng J., Thurston R. N., **A Multiplexible Bistable Nematic Liquid-Crystal Display Using Thermal Erasure.** Appl Phys L 40(11):936–938, 1982.

Brews J. R., **Dopant Density From Maximum-Minimum Capacitance Ratio of Implanted MOS Structures.** Sol St Elec 25(5):375–379, 1982.

Campbell J. C., Dentai A. G., Copeland J. A., Holden W. S., **Optical AND Gate.** IEEE J Q El 18(6):992–995, 1982.

Cho A. Y., Kollberg E., Zirath H., Snell W. W., Schneider M. V., **Single-Crystal Metal-Semiconductor Microjunctions Prepared by Molecular-Beam Epitaxy.** Electr Lett 18(10):424–425, 1982.

Cleveland W. S., Diaconis P., McGill R., **Variables on Scatterplots Look More Highly Correlated When the Scales are Increased.** Science 216(4550):1138–1141, 1982.

Cox D. C., Arnold H. W., Hoffman H. H., **Measurement Bounds on Rain-Scatter Coupling Between Space-Earth Radio Paths.** IEEE Antenn 30(3):493–497, 1982.

Cox D. C., Arnold H. W., **Results From the 19-GHz and 28-GHz COMSTAR Satellite Propagation Experiments at Crawford Hill.** P IEEE 70(5):458–488, 1982.

Dragone C., **A First-Order Treatment of Aberrations in Cassegrainian and Gregorian Antennas.** IEEE Antenn 30(3):331–339, 1982.

Dudderar T. D., Simpkins P. G., **The Development of Scattered-Light Speckle Metrology.** Opt Eng 21(3):396–399, 1982.

Dutta N. K., Nelson R. J., **Temperature-Dependence of the Lasing Characteristics of the 1.3 InGaAsP-InP and GaAs-Alo.36Ga0.64As DH Lasers.** IEEE J Q El 18(5):871–878, 1982.

Etzel M. H., Jenkins W. K., **The Design of Specialized Residue Classes for Efficient Recursive Digital Filter and Realization.** IEEE Acoust 30(3):370–380, 1982.

Grabbe P., Hu E. L., Howard R. E., **Trilevel Lift-Off Process for Refractory-Metals.** J Vac Sci T 21(1):33–35, 1982.

Hall T. M., Wagner A., Thompson L. F., **Ion-Beam Exposure Characteristics of Resists—Experimental Results.** J Appl Phys 53(6):3997–4010, 1982.

Knapp J. A., Lapeyre G. J., Smith N. V., Traum M. M., **Modification of a Cylindrical Mirror Analyzer for Angle-Resolved Electron-Spectroscopy.** Rev Sci Ins 53(6):781–784, 1982.

Lazay P. D., Pearson A. D., **Developments in Single-Mode Fiber Design, Materials, and Performance at Bell Laboratories.** IEEE J Q El 18(4):504–510, 1982.

Lien R. M., **Build Small-Part Molds That Double for Prototyping and Production-Runs.** Plast Eng 38(5):61–63, 1982.

Logan R. A., Vanderziel J. P., Mikulyak R. M., **A Closely Spaced (50 μm) Array of 16 Individually Addressable Buried Heterostructure GaAs-Lasers.** Appl Phys L 41(1):9–11, 1982.

Nagal S. R., Mac Chesney J. B., Walker K. L., **An Overview of the Modified Chemical Vapor-Deposition (MCVD) Process and Performance.** IEEE J Q El 18(4):459–476, 1982.

Nassau K., Wemple S. H., **Material Dispersion Slope in Optical-Fibre Waveguides.** Electr Lett 18(11):450–451, 1982.

Ogawa K., **Analysis of Mode Partition Noise in Laser Transmission-Systems.** IEEE J Q El 18(5):849–855, 1982.

Olson J. R., Nichols R. H., **Correlation-Measurements of Surface-Reflected Underwater Acoustic-Signals at Several Sea States.** J Acoust So 71(6):1453–1457, 1982.

Pearsall T. P., **ZAP—Introducing the Zero-Bias Avalanche Photodiode.** Electr Lett 18(12):512–514, 1982.

Rabiner L. R., Rosenberg A. E., Wilpon J. G., Zampini T. M., **A Bootstrapping Training Technique for Obtaining Demi-Syllable Reference Patterns.** J Acoust So 71(6):1588–1595, 1982.

Ross I. M., **R-and-D in the United-States—Its Strengths and Challenges.** Science 217(4555):130–131, 1982.

Rowe H. E., **Bounds on the Number of Signals With Restricted Cross-Correlation.** IEEE Commun 30(5):966–977, 1982.

Sessler G. M., West J. E., Von Seggern H., **Electron-Beam Method for Detecting Charge-Distributions in Thin Polyethyleneterephthalate Films.** J Appl Phys 53(6):4320–4327, 1982.

Sinha A. K., Cooper J. A., Levinstein H. J., **Speed Limitations Due to Interconnect Time Constants in VLSI Integrated-Circuits.** Elec Dev L 3(4):90–92, 1982.

Smith R. W., Tomlinson W. J., **Optically Bistable Devices.** Electtrotecn 69(3):199–206, 1982.

Sze S. M., **The Development of Semiconductor-Devices in the 1970s and 1980s—an Overview.** Electtrotecn 69(4):297–304, 1982.

Stone J., Cohen L. G., **Tunable InGaAsP lasers for Spectral Measurements of High Bandwidth Fibers.** IEEE J Q El 18(4):511–513, 1982.

Thornber K. K., **Current Equations for Velocity Overshoot.** Elec Dev L 3(3):69–71, 1982.

Tsang W. T., Logan R. A., **As Strip Buried-Heterostructure Lasers Prepared by Hybrid Crystal-Growth.** Electr Lett 18(10):397–398, 1982.

Unger B. A., **How Not to 'Zap' a Circuit.** Bell Lab Re 60(5):123–127, 1982

Williams G. F., Capasso F., Tsang W. T., **The Graded Bandgap Multilayer Avalanche Photodiode—A New Low-Noise Detector.** Elec Dev L 3(3):71–73, 1982.

Winters J. H., **Spread Spectrum in a 4-Phase Communication-System Employing Adaptive Atennas.** IEEE Commun 30(5):929–936, 1982.

PHYSICAL SCIENCES

Abrahams S. C., Zyontz L. E., Bernstein J. L., **Cobalt Cyanourate—Crystal-Structure of a Component From Cobalt-Hardened Gold Electroplating Baths.** J Chem Phys 76(11):5458–5462, 1982.

Aeppli G., et al., **Spin Correlations Near the Ferromagnetic Spin-Glass Crossover Point in Amorphous Fe-Mn Alloys.** Phys Rev B 25(7):4882–4885, 1982.

Aharonshalom E., Wudl F., Bertz S. H., Walsh W. M., Rupp L. W., Chaikin P. M., Burns M. J., Andres K., Schwenk H., **Tetrakis (Deuteriomethyl) Tetraselenafulvalene.** Molec Cryst 86(1–4):1775–1781, 1982.

Alferness R. C., **Tunable Electrooptic Wave-Guide Te↔Tm Converter Wavelength Filter.** Appl Phys L 40(10):861–862, 1982.

Applebury M. L., **Picosecond Spectroscopy of Visual Pigments.** Meth Enzym 81(PH):354–368, 1982.

Ashkin A., Dziedzic J. M., Smith P. W., **Continuous-Wave Self-Focusing and Self-Trapping of Light in Artificial Kerr Media.** Optics Lett 7(6):276–278, 1982.

Aspnes D. E., **Bounds to Average Internal Fields in Two-Component Composites.** Phys Rev L 48(23):1629–1632, 1982.

Aspnes D. E., Kelso S. M., Olson C. G., Lynch D. W., **Direct Determination of Sizes of Excitations From Optical Measurements on Ion-Implanted GaAs.** Phys Rev L 48(26):1863–1866, 1982.

Aspnes D. E., Kelso S. M., **Properties and Performance of Grazing-Incidence Mirror Systems.** Nucl Instru 195(1–2):175–181, 1982.

Ballman A. A., Brown H., Blitzer L. D., **Growth of Antimony Doped InP Single-Crystal.** J Cryst Gr 57(3):516–518, 1982.

Bally J., Lane A. P., **Observations of 2 μm Molecular-Hydrogen Emission From NGC-2071, Cepheus-A, and GL-961.** Astrophys J 257(2):612–619, 1982.

Banavar J. R., Grest G. S., Jasnow D., **Anti-Ferromagnetic Potts and Ashkin-Teller Models in 3 Dimensions.** Phys Rev B 25(7):4639–4650, 1982.

Beck S. M., Brus L. E., **The Resonance Raman-Spectra of Aqueous Phenoxy and Phenoxy-D5 Radicals.** J Chem Phys 76(10):4700–4704, 1982.

Beckman O., et al., **Low Field Simultaneous AC and DC Magnetization Measurements of Amorphous $(Fe_{0.20}\ Ni_{0.80})_{75}P_{16}B_6Al_3$ and $(Fe_{0.68}Mn_{0.32})_{75}P_{16}B_6Al_3$.** Phys Scr 25(6):676–678, 1982.

Beckman O., Figueroa E., Gramm K., Lundgren L., Rad K. V., Chen H. S., **Spin-Wave and Scaling Law Analysis of Amorphous $(Fe_XNi_{1-X})_{75}P_{16}B_6Al_3$ by Magnetization Measurements.** Phys Scr 25(6):726–730, 1982.

Beneking H., Cho A. Y., Dekkers J. J. M., Morkoc H., **Buried-Channel GaAs-MESFETs on MBE Material—Scattering Parameters and Intermodulation Signal Distortion.** IEEE Device 29(5):811–813, 1982.

Beni G., Hackwood S., Jackel J. L., **Continuous Electro-Wetting Effect.** Appl Phys L 40(10):912–914, 1982.

Bertz S. H., Dabbagh G., Cotte P., **New Copper Chemistry. 3. New Preparations of Ethyl 3,3-Diethoxypropionate and (Ethoxycarbonyl)Malondialdehyde-Cu(I)-Catalyzed Acetal Formation from Conjugated Triple Bond.** J Org Chem 47(11):2216–2217, 1982.

Blitz L., Fich M., Stark A. A., **Catalog of CO Radial-Velocities Toward Galactic H-II Regions.** Astroph J S 49(2):183–206, 1982.

Bosch M. A., Kang K. S., Hackwood S., Beni G., Shay J. L., **Optical Writing on Blue, Sputtered Iridium Oxide-Films.** Appl Phys L 41(1):103–105, 1982.

Boutique J. P., Riga J., Verbist J. J., Delhalle J., Fripiat J. G., Andre J. M., Haddon R. C., Kaplan M. L., **Electronic Structure of Naphtho < 1.8-CD, 4,5-C'D'*BIS < 1,2,6*Thiadiazines and <1,2,6*Selenadiazines—Abinitio Calculations and Photo-Electron Spectra.** J Am Chem S 104(10):2691–2697, 1982.

Brinkman W. F., Cladis P. E., **Defects in Liquid-Crystals.** Phys Today 35(5):48–54, 1982.

Broughton J. Q., Gilmer G. H., Weeks J. D., **Molecular-Dynamics Study of Melting in 2 Dimensions—Inverse-12th-Power Interaction.** Phys Rev B 25(7):4651–4669, 1982.

Brown T. R., Kincaid B. M., Ugurbil K., **NMR Chemical-Shift Imaging in 3 Dimensions.** P Nas Biol 79(11):3523–3526, 1982.

Carmeli B., Nitzan A., **First Passage Times and the Kinetics of Unimolecular Dissociation.** J Chem Phys 76(11):5321–5333, 1982.

Cava R. J., Santoro A., Murphy D. W., Zahurak S., Roth R. S., **The Structures of Lithium-Inserted Metal-Oxides—$Liero_3$ and Li_2Reo_3.** J Sol St Ch 42(3):251–262, 1982.

Celler G. K., Trimble L. E., Ng K. K., Leamy H. J., Baumgart H., **Seeded Oscillatory Growth of Si Over SiO_2 by CW Laser Irradiation.** Appl Phys L 40(12):1043–1045, 1982.

Chang T. Y., Leheny R. F., Nahory R. E., Silberg E., Ballman A. A., Caridi E. A., Harrold C. J., **Junction Field-Effect Transistors Using $In_{0.53}Ga_{0.47}As$ Material Grown by Molecular-Beam Epitaxy.** Elec Dev L 3(3):56–58, 1982.

Cheng J., Thurston R. N., Boyd G. D., Meyer R. B., **A Nematic Liquid-Crystal Storage Display Based on Bistable Boundary-Layer Configurations.** Appl Phys L 40(12):1007–1009, 1982.

Cheng K. Y., Cho A. Y., **Silicon Doping and Impurity Profiles in $Ga_{0.47}In_{0.53}As$ and $Al_{0.48}In_{0.52}As$ Grown by Molecular-Beam Epitaxy.** J Appl Phys 53(6):4411–4415, 1982.

Chung F. R. K., Erdos P., Graham R. L., **Minimal Decompositions of Hypergraphs Into Mutually Isomorphic Sub-Hypergraphs.** J Comb Th A 32(2):241–251, 1982.

Cohen J. D., Lang D. V., **Calculation of the Dynamic-Response of Schottky Barriers with a Continuous Distribution of Gap States.** Phys Rev B 25(8):5321–5350, 1982.

Crisci R. J., Draper C. W., Preece C. M., **Cavitation Erosion of Laser-Surface-Melted Self-Quenched Fe-Al Bronze.** Appl Optics 21(10):1730–1731, 1982.

Dahlberg S. C., Reinganum C. B., **Photochromic Switching in Semiconducting Films of Zinc Dithizonate.** J Chem Phys 76(11):5515–5518, 1982.

Degani J., Leheny R. F., Nahory R. E., Shah J., **High-Field Transport Characteristics of Minority Electron In P-$In_{0.53}Ga_{0.47}As$.** Thin Sol Fi 89(1):19–20, 1982.

Dicenzo S. B., Wertheim G. K., Buchanan D. N., **Epitaxy of CuI on Cu(111).** Appl Phys L 40(10):888–890, 1982.

Drummond T. J., Morkoc H., Cheng K. Y., Cho A. Y., **Current Transport in Modulation-Doped $Ga_{0.47}In_{0.53}As$ $Al_{0.48}In_{0.52}As$ Heterojunctions at Moderate Electric Fields.** J Appl Phys 53(5):3654–3657, 1982.

Dutta N. K., **Temperature-Dependence of Threshold Current of GaAs Quantum Well Lasers.** Electr Lett 18(11):451–453, 1982.

Efstathiou G., et al., **The Stability and Masses of Disk Galaxies.** M Not R Ast 199(3):1069–1088, 1982.

Finn P. L., Cladis P. E., **Cholesteric Blue Phases in Mixtures and in an Electric Field.** Molec Cryst 84(1–4):159–192, 1982.

Forrest S. R., Kim Q. K., Smith R. G., **Optical-Response Time of In0.53Ga0.47As/InP Avalanche Photo-Diodes.** Appl Phys L 41(1):95–98, 1982.

Forrest S. R., Kaplan M. L., Schmidt P. H., Feldmann W. L., Yanowski E., **Organic-on-Inorganic Semiconductor Contact Barrier Devices.** Appl Phys L 41(1):90–93, 1982.

Fowkes F. M., Butler B. L., Schissel P., Butler G. B., Delollis N. J., Hartman J. S., Hoffman R. W., Inal O. T., Miller W. G., Tompkins H. G., **Basic Research Needs and Opportunities at the Solid Solid Interface—Adhesion, Abrasion and Polymer-Coatings.** Mater Sci E 53(1):125–136, 1982.

Frankenthal R. P., **The Anodic Corrosion of Gold in Concentrated Chloride Solutions.** J Elchem So 129(6):1192–1196, 1982.

Frerking M. A., Langer W. D., **Detection of Pedestal Features in Dark Clouds—Evidence for Formation of Low Mass Stars.** Astrophys J 256(2):523–529, 1982.

Gedanken A., Robin M. B., Yafet Y., **The Methyl-Iodide Multi-Photon Ionization Spectrum With Intermediate Resonance in the A-Band Region.** J Chem Phys 76(10):4798–4808, 1982.

Goodwin C. A., Brossman J. W., **MOS Gate Oxide Defects Related to Treatment of Silicon-Nitride Coated Wafers Prior to Local Oxidation.** J Elchem So 129(5):1066–1070, 1982.

Graedel T. E., Langer W. D., Frerking M. A., **The Kinetic Chemistry of Dense Inter-Stellar Clouds.** Astroph J S 48(3):321–368, 1982.

Greenblatt M., Murphy D. W., Di Salvo B. J., Eibschutz M., Zahurak S. M., Waszczak J. V., **Preparation and Properties of Fe-Substituted V_6O_{13}.** J Sol St Ch 42(2):212–216, 1982.

Haddon R. C., **Molecular-Orbital Studies of the Hetero(A,B)Polyenes and Hetero(A,B)Annulenes <(A,B) = (C,C), (C,N), (B,N), (B,O)*.** Pur A Chem 54(5):1129–1142, 1982.

Hasegawa A., Kodama Y., **Amplification and Reshaping of Optical Solitons A Glass-Fiber.** Optics Lett 7(6):285–287, 1982.

Hauser J. J., **Electrical and Structural-Properties of Ag-Te, Ag-Se, Ag-S, and Ag-I Diffusion Couples.** J Appl Phys 53(5):3634–3638, 1982.

Ikezi H., Giannetta R. W., Platzman P. M., **Non-Linear Equilibria of the Electron-Charged Surface of Liquid-Helium.** Phys Rev B 25(7):4488–4494, 1982.

Jackson S. A., Platzman P. M., **Temperature-Dependent Effective Mass of a Self-Trapped Electron on the Surface of a Liquid-Helium Film.** Phys Rev B 25(7):4886–4889, 1982.

Jin S., Brasen D., Mahajan S., **Coercivity Mechanisms in Fe-Cr-CO Magnet Alloys.** J Appl Phys 53(6):4300–4303, 1982.

Johari G. P., Goodby J. W., Johnson G. E., **Molecular Relaxations in a Glass of Cholesteric Liquid-Crystal.** Nature 297(5864):315–317, 1982.

Johnson M. K., Raye C. L., Foley M. A., Kim J. K., **Pictures and Images—Spatial and Temporal Information Compared.** B Psychon S 19(1):23–26, 1982.

Kamgar A., **Subthreshold Behavior of Silicon MOSFETs at 4.2-K.** Sol St Elec 25(7):537–539, 1982.

Khanarian G., Cais R. E., Kometani J. M., Tonelli A. E., **Kerr Effect and Dielectric Study of the Co-Polymer Poly(Styrene-CO-P-Halogenated-Styrene).** Macromolec 15(3):866–869, 1982.

Kohl P. A., **The High-Speed Electrodeposition of Sn Pb Alloys.** J Elchem So 129(6):1196–1201, 1982.

Kvick A., Liminga R., Abrahams S. C., **Neutron-Diffraction Structural Study of Pyroelectric $Ba(No_2)2.H_2O$ at 298-K, 102-K, and 20-K.** J Chem Phys 76(11):5508–5514, 1982.

Lake G. R., Norman C., **Triaxiality and the Galactic-Center.** AIP Conf PR1982(83):189–193, 1982.

Lang D. V., Cohen J. D., Harbison J. P., **Measurement of the Density of Gap States in Hydrogenated Amorphous Silicon by Space-Charge Spectroscopy.** Phys Rev B 25(8):5285–5320, 1982.

Larson R. G., Davis H. T., **Conducting Backbone in Percolating Bethe Lattices.** J Phys C 15(11):2327–2331, 1982.

Leamy H. J., **Charge Collection Scanning Electron-Microscopy.** J Appl Phys 53(6):R51–R80, 1982.

Leon J. S., Pless V., Sloane N. J. A., **Self-Dual Codes Over GF(5).** J Comb Th A 32(2):178–194, 1982.

Leventhal M., et al., **Time Variable Positron-Annihilation Radiation From the Galactic-Center Direction.** AIP Conf PR1982(83):132–138, 1982.

Levinson M., Benton J. L., Temkin H., Kimerling L. C., **Defect States in Electron Bombarded N-InP.** Appl Phys L 40(11):990–992, 1982.

Lewerenz H. J., Aspnes D. E., Miller B., Malm D. L., Heller A., **Semiconductor Interface Characterization in Photoelectrochemical Solar-Cells—The P-InP (111)A Face.** J Am Chem S 104(12):3325–3329, 1982.

Liebman S. A., Ahlstrom D. H., Starnes W. H., Schilling F. C., **Short-Chain Branching in Polyethylene and Polyvinyl-Chloride Using Pyrolysis Hydrogenation Gas-Chromatography and C-13 Nuclear Magnetic-Resonance Analysis.** J Macr S Ch A17(6):935–950, 1982.

Marcus R. B., Sheng T. T., Lin P., **Polysilicon SiO_2 Interface Microtexture and Dielectric-Breakdown.** J Elchem So 129(6):1282–1289, 1982.

Marcus R. B., Sheng T. T., **The Oxidation of Shaped Silicon Surfaces.** J Elchem So 129(6):1278–1282, 1982.

Mc Crory J. C., Rosamilia J. M., **Modification of the Electrochemical Behavior of Copper by Azole Compounds.** J Elec Chem 136(1):105–118, 1982.

Meixner A. E., Chen C. H., Geerk J., Schmidt P. H., **High-Energy Electron-Energy-Loss Spectroscopy of Niobium and Niobium-Oxygen Solid-Solutions.** Phys Rev B 25(8):5032–5036, 1982.

Murayama K., Bosch M. A., **Hot Photo-Luminescence in Amorphous As_2S_3.** Phys Rev B 25(10):6542–6544, 1982.

Nakahara S., **Direct Observations of Inclusions in Electrodeposited Films by Transmission Electron-Microscopy.** J Elchem So 129(5):C201–C212, 1982.

Nakahara S., Mc Coy R. J., **Microstructural Determination of Fast Diffusing Species in Thin-Film Diffusion Couples.** Thin Sol Fi 88(3):285–290, 1982.

Nelson D. F., **Exact Solution for Space-Charge Broadened Packets in Semiconductors.** Phys Rev B 25(8):5267–5275, 1982.

O'Connor P., Pearsall T. P., Cheng K. Y., Cho A. Y., Hwang J. C. M., Alavi K., **In0.53Ga0.47As FETs With Insulator-Assisted Schottky Gates.** Elec Dev L 3(3):64–66, 1982.

Odagaki T., Lax M., **Hopping Conduction in One-Dimensional Random Chains.** Molec Cryst 85(1–4):129–136, 1982.

Patterson G. D., **Depolarized Rayleigh Spectroscopy of Small Alkanes With Picosecond Relaxation-Times.** J Chem Phys 76(9):4356–4360, 1982.

Peisach J., Powers L., Blumberg W. E., Chance B., **Stellacyanin—Studies of the Metal-Binding Site Using X-Ray Absorption-Spectroscopy.** Biophys J 38(3):277–285, 1982.

Pierce R. D., Caruso R., Mogab C. J., **Thermal Annealing of Implantation Strain in Bubbles Garnet-Films—Stability of Ion-Implanted Propagation Devices.** J Appl Phys 53(6):4480–4484, 1982.

Pindak R., Moncton D., **Two-Dimensional Systems.** Phys Today 35(5):57–62, 1982.

Platzman P. M., **Surface High-Energy Electron-Diffraction.** Phys Rev B 25(8):5046–5049, 1982.

Poate J. M., Brown W. L., **Laser Annealing of Silicon.** Phys Today 35(6):24–30, 1982.

Pryde G. A., Kelleher P. G., Hellman M. Y., Wentz R. P., **The Hydrolytic Stability of Some Commercially Available Polycarbonates.** Polym Eng S 22(6):370–375, 1982.

Rabinovich E. M., Johnson D. W., Mac Chesney J. B., Vogel E. M., **Preparation of Transparent High-Silica Glass Articles From Colloidal Gels.** J Non-Cryst 47(3):435–439, 1982.

Raghavachari K., **A Theoretical-Study of the Reaction Surface for The H_2O-Li_2O System.** J Chem Phys 76(11):5421–5426, 1982.

Reinhart F. K., Logan R. A., Sinclair W. R., **Electrooptic Polarization Modulation in Multielectrode $Al_xGa_{1-x}As$ Rib Wave-Guides.** IEEE J Q El 18(4):763–766, 1982.

Reynaud F., Lasserre A., Bonner W. A., Mahajan S., **Radiation-Damage in InP Single-Crystals Irradiated Between 200°C and 320°C in a High-Voltage Electron-Microscope (2 MV).** Scrip Metal 16(5):567–570, 1982.

Richton R. E., Farrow L. A., **BCl_3 Absorption of CO_2 Laser Lines.** J Chem Phys 76(11):5256–5259, 1982.

Rosenbaum T. F., Rupp L. W., Thomas G. A., Walsh W. M., Chen H. S., Banavar J. R., Littlewood P. B., **Electron-Spin Resonance Indication of the Ferromagnet-Spin Glass-Transition in Amorphous FeMn Alloys.** Sol St Comm 42(10):725–727, 1982.

Salathe R. P., Gilgen H. H., Binkert T., Reinhart F. K., Logan R. A., **Efficient Electro-Luminescence From Laser-Irradiated (Al,Ga)As-Heterostructure Diodes.** J Appl Phys 53(5):3769–3771, 1982.

Satija I. I., Hohenberg P. C., **Irrelevant Operators and Momentum-Shell Recursion-Relations in D=2+XI-Dimensions.** J Stat Phys 28(1):83–97, 1982.

Schwartz G. P., Sunder W. A., Griffiths, J. E., **The In-P-O Phase-Diagram—Construction and Applications.** J Elchem So 129(6):1361–1367, 1982.

Shah J., Pinczuk A., Alexander F. B., Bagley B. G., **Excitation Wavelength Dependence of Luminescence Spectra of A-Si-H.** Sol St Comm 42(10):717–720, 1982.

Smith N. V., Kevan S. D., **General Instrumentation Considerations in Electron and Ion Spectroscopies Using Synchrotron Radiation.** Nucl Instru 195(1–2):309–321, 1982.

Smolinsky G., Truesdale E. A., Wang D. N. K., Maydan D., **Reactive Ion Etching of Silicon-Oxides with Ammonia and Trifluoromethane—The Role of Nitrogen in the Discharge.** J Elchem So 129(5):1036–1040, 1982.

Stamatoff J., Feuer B., Guggenheim H. J., Tellez G., Yamane T., **Amplitude of Rippling in the P-Beta Phase of Dipalmitoylphosphatidylcholine Bilayers.** Biophys J 38(3):217–226, 1982.

Stevens J. R., Patterson G. D., Carroll P. J., Alms G. R., **The Central Lorentzian in the Depolarized Rayleigh Spectra of Liquid CCl_4 and $GeCL_4$.** J Chem Phys 76(11):5203–5207, 1982.

Stone J., Chraplyvy A. R., Burrus C. A., **Gas-In-Glass—A New Raman-Gain Medium—Molecular-Hydrogen in Solid-Silica Optical Fibers.** Optics Lett 7(6):297–299, 1982.

Sze S. M., **Citation Classic—Physics of Semiconductor-Devices.** CC/Eng Tech 1982(27):28, 1982.

Taborek P., **Critical Cone in Phonon-Induced Desorption of Helium.** Phys Rev L 48(25):1737–1741, 1982.

Tamargo M. C., Reynolds C. L., **Influence of Cooling Rate and Melt Configuration on Rake Lines in the Active Layer of $Al_XGa_{1-X}As$ Lasers.** J Cryst Gr 57(2):349–352, 1982.

Tewksbury S. K., **Observation of Discrete Current Levels in Metal-Oxide-Semiconductor Field-Effect Transistors Switched Into Weak Inversion Between 10 and 25 K.** Appl Phys L 41(1):62–64, 1982.

Tewksbury S. K., **Transient-Response of N-Channel Metal-Oxide-Semiconductor Field-Effect Transistors During Turnon at 10–25°K.** J Appl Phys 53(5):3865–3872, 1982.

Thurston R. N., Cheng J., Boyd G. D., **Optical-Properties of a New Bistable Twisted Nematic Liquid-Crystal Boundary-Layer Display.** J Appl Phys 53(6):4463–4479, 1982.

Tomlinson W. J., Gordon J. P., Smith P. W., Kaplan A. E., **Reflection of a Gaussian-Beam at a Non-Linear Interface.** Appl Optics 21(11):2041–2051, 1982.

Tonelli A. E., **Conformational Characteristics of Poly(N-Vinyl Pyrrolidone).** Polymer 23(5):676–680, 1982.

Tonelli A. E., Schilling F. C., Cais R. E., **F-19 NMR Chemical-Shifts and the Microstructure of Fluoro Polymers.** Macromolec 15(3):849–853, 1982.

Tsang W. T., Logan R. A., Van Der Ziel J. P., **A New Lateral Selective-Area Growth**

by Liquid-Phase Epitaxy—The Formation of a Lateral Double-Barrier Buried-Heterostructure Laser. Appl Phys L 40(11):942-944, 1982.

Tsui D. C., Stormer H. L., Gossard A. C., **Two-Dimensional Magnetotransport in the Extreme Quantum Limit.** Phys Rev L 48(22):1559-1562, 1982.

Tu C. W., Chang R. P. H., Schlier A. R., **Surface Etching Kinetics of Hydrogen Plasma on InP.** Appl Phys L 41(1):80-82, 1982.

Tully J. C., Muhlhausen C. W., Ruby L. R., **Stochastic Trajectory Studies of Chemical Processes at Surfaces.** Ber Bun Ges 86(5):433-437, 1982.

Tyson J. A., Baum W. A., Kreidl T., **Deep CCD Images of 3C 273.** Astrophys J 257(1):L1-L5, 1982.

Vasile M. J., Stevie F. A., **Reaction of Atomic Fluorine With Silicon—The Gas-Phase Products.** J Appl Phys 53(5):3799-3805, 1982.

Wang T. T., **Properties of Piezoelectric Poly (Vinylidene Fluoride) Films Irradiated by Gamma-Rays.** Ferroelectr 41(1-4):347-357, 1982.

Weber T. A., Annan N. D., **Molecular-Dynamics of Small Alkanes In an External Force-Field.** Molec Phys 46(1):193-209, 1982.

Whalen M. S., Stone J., **Index of Refraction of N-Type InP at 0.633 μm and 1.15 μm Wavelengths as a Function of Carrier Concentration.** J Appl Phys 53(6):4340-4343, 1982.

Wilson B. A., Kerwin T. P., **Time-Resolved Photo-Luminescence in A-Si-H-Sub-band-Gap Excitation.** Phys Rev B 25(8):5276-5284, 1982.

Wolff R. S., Carlson E. R., **Masing and Non-Masing Silicon Monoxide Emission From Evolved Stars.** Astrophys J 257(1):161-170, 1982.

Worlock J. M., **Electrons in Novel Two-Dimensional Structures (Editorial).** Nature 297(5865):360-361, 1982.

Wudl F., et al., **Ditetramethyltetraselenafulvalenium Fluorosulfonate—The Effect of a Dipolar Anion on the Solid-State Physical-Properties of the (TMTSF)2X Phase.** J Chem Phys 76(11):5497-5501, 1982.

Wudl F., **Is There a Relationship Between Aromaticity and Conductivity.** Pur a Chem 54(5):1051-1058, 1982.

SOCIAL AND LIFE SCIENCE

Arkin W., et al., **The Consequences of a Limited Nuclear War in East and West-Germany.** Ambio 11(2-3):163-173, 1982.

Fishburn P. C., Gehrlein W. V., **Majority Efficiencies for Simple Voting Procedures—Summary and Interpretation.** Theor Decis 14(2):141-153, 1982.

Graham R., Buhler J., **The Art of Juggling.** Recherche 13(135):856-857, 1982.

Lamola A. A., **Quantum Yield and Equilibrium Position of the Configurational Photo-Isomerization of Bilirubin Bound to Human-Serum Albumin.** 35(5):649-654, 1982.

Samuel A. G., **Phonetic Prototypes.** Perc Psych 31(4):307-314, 1982.

Weiss B., **The Decline of Late Bronze-Age Civilization as a Possible Response to Climatic-Change.** Clim Change 4(2):173-198, 1982.

CONTENTS, NOVEMBER 1982

Part 1

Feeder Planning Methods for Digital Loop Carrier
B. Bulcha, L. E. Kodrich, D. B. Luber, W. J. Mitchell, M. A. Schwartz, and F. N. Woomer

Evaluation of Private Networks
R. L. Kaufman and A. J. M. Kester

A Description of the Bell Laboratories Scanned Acoustic Microscope
P. Sulewski, D. J. Bishop, and R. C. Dynes

A Statistical Model of Multipath Fading on a Space-Diversity Radio Channel
W. D. Rummler

Matrix Analysis of Mildly Nonlinear, Multiple-Input, Multiple-Output Systems With Memory
A. A. M. Saleh

Computing the Distribution of a Random Variable Via Gaussian Quadrature Rules
M. H. Meyers

A 9.6-kb/s DSP Speech Coder
R. E. Crochiere, R. V. Cox, J. D. Johnston, and L. A. Seltzer

An Improved Model for Isolated Word Recognition
J. M. Tribolet, L. R. Rabiner, and J. G. Wilpon

A Theoretical Model of Transient Heat Transfer in a Fire-stopped Cable Bundle
P. B. Grimado

Planning and Conducting Field-Tracking Studies
S. J. Amster, G. G. Brush, and B. Saperstein

Part 2

COMPUTING SCIENCE AND SYSTEMS

Database Work at Bell Laboratories
A. V. Aho

Structure of a *UNIX* Database File System
M. J. Rochkind

Making *UNIX* Operating Systems Safe for Databases
P. J. Weinberger

A Real-Time Database Management System for No. 5 Ess
D. K. Barclay, E. R. Byrne, and F. K. Ng

Database Administration System—Architecture and Design Issues
C. C. Wang and C. P. Huang

The Implementation of a Distributed Database Management System to Support Logical Subnetworks
D. Cohen

An Architectural Overview of a Prototype Videotex System
N. H. Goguen

Human Factors in Data Access
T. K. Landauer, S. T. Dumais, G. W. Furnas, and L. M. Gomez

Design and Implementation of a Production Database Management System
T. C. Chiang and G. R. Rose

A Directed-Hypergraph Database: A Model for the Local Loop Plant
A. J. Goldstein

Issues in the Design of a Distributed Record Management System
J. P. Linderman

The Intelligent Store: A Content-Addressable Page Manager
W. D. Roome

Part 3

THE D4 DIGITAL CHANNEL BANK FAMILY

J. J. Lang, Guest Editor

Overview
J. Chernak and J. J. Lang

The Channel Bank
C. R. Crue, W. B. Gaunt, Jr., J. H. Green, J. E. Landry, and D. A. Spires

The Maintenance Bank
R. E. Benjamin and H. H. Mahn

The *SLC*-96 System
Y-S Cho, J. W. Olson, and D. H. Williamson

Dataport—Digital Access Through D4
T. J. Aprille, S. Narayanan, P. G. St. Amand, and F. E. Weber

Dataport—Channel Units for Digital Data System Subrates
T. J. Aprille, D. V. Gupta, and P. G. St. Amand

Dataport—Channel Units for Digital Data System 56-kb/s Rate
B. J. Dunbar, D. V. Gupta, M. P. Horvath, R. E. Sheehey, and S. P. Vermax

Digital Terminal Physical Design
W. G. Albert, A. G. Favale, J. R. Hall, and D. H. Klockow

Custom-Integrated Circuits for Digital Terminals
R. M. Goldstein, J. D. Leggett, G. L. Mowery, and K. F. Sodomsky

Thin-Film Dual Active Filter for Pulse Code Modulation Systems
R. L. Adams, J. S. Fisher, O. G. Petersen, and I. G. Post

Errata

J. C. Baumhauer, Jr. and A. M. Brzezinski, "The EL2 Electret Transmitter: Analytical Modeling, Optimization, and Design," B.S.T.J., *58,* No. 7 (September 1979), pp. 1557–1578.

On page 1568, line 17, change

$$J^{\circ}(\bar{t}) = 10.6 \times 10^{-6}\text{m}^2/\text{N}$$

to

$$J^{\circ}(\bar{t}) = 15.5 \times 10^{-10}\text{m}^2/\text{N}$$

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, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* 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.

An Inversion Technique for the Laplace Transform 1995
D. L. Jagerman

Approximate Mean Waiting Times in Transient GI/G/1 Queues 2003
D. L. Jagerman

An Analysis of the Carrier-Sense Multiple-Access Protocol 2023
D. P. Heyman

Implementing and Testing New Versions of a Good 48-Bit Pseudo-Random Number Generator 2053
C. S. Roberts

Some Extremal Markov Chains 2065
J. E. Mazo

Quality Evaluation Plan Using Adaptive Kalman Filtering 2081
M. S. Phadke

CONTRIBUTORS TO THIS ISSUE 2109

PAPERS BY BELL LABORATORIES AUTHORS 2117

CONTENTS, NOVEMBER ISSUE 2125

ERRATA 2127