


THE FEBRUARY 1983
VOL. 62, NO. 2, PART 1



BELL SYSTEM
TECHNICAL JOURNAL

Digital Communications Over Fading Radio Channels G. J. Foschini and J. Salz	429
Chromatic Dispersion Measurements in Single-Mode Fibers Using Picosecond InGaAsP Injection Lasers in the 1.2- to 1.5-μm Spectral Region C. Lin, A. R. Tynes, A. Tomita, P. L. Liu, and D. L. Philen	457
High-Frequency Impedance of Proton-Bombarded Injection Lasers B. W. Hakki, W. R. Holbrook, and C. A. Gaw	463
An Analysis of the Derivative Weight-Gain Signal From Measured Crystal Shape: Implications for Diameter Control of GaAs A. S. Jordan, R. Caruso, and A. R. Von Neida	477
Growth, Complexity, and Performance of Telephone Connecting Networks V. E. Beneš	499
Upper and Lower Bounds on Mean Throughput Rate and Mean Delay in Memory-Constrained Queueing Networks E. Arthurs and B. W. Stuck	541
CONTRIBUTORS TO THIS ISSUE	583
PAPERS BY BELL LABORATORIES AUTHORS	587
CONTENTS, MARCH ISSUE	593

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

D. E. PROCKNOW, *President,*
I. M. ROSS, *President,*
W. M. ELLINGHAUS, *President,*

Western Electric Company
Bell Telephone Laboratories, Incorporated
American Telephone and Telegraph Company

EDITORIAL COMMITTEE

A. A. PENZIAS, *Chairman*

M. M. BUCHNER, JR.	R. A. KELLEY
A. G. CHYNOWETH	R. W. LUCKY
R. P. CLAGETT	R. L. MARTIN
T. H. CROWLEY	J. S. NOWAK
B. P. DONOHUE, III	L. SCHENKER
I. DORROS	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 (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N.Y., N.Y. 10007, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year \$35; 2 years \$63; 3 years \$84; foreign—1 year \$45; 2 years \$73; 3 years \$94. Subscriptions to Part 2 only are \$10 (\$11 Foreign). Single copies of the Journal are available at \$5 (\$6 foreign). Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to The Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

Single copies of material from this issue of The Bell System Technical Journal may be reproduced for personal, noncommercial use. Permission to make multiple copies must be obtained from the editor.

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.

Printed in U.S.A. Second-class postage paid at Short Hills, N. J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

© 1983 American Telephone and Telegraph Company.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62

February 1983

Number 2, Part 1

Copyright © 1983 American Telephone and Telegraph Company. Printed in U.S.A.

Digital Communications Over Fading Radio Channels

By G. J. FOSCHINI and J. SALZ

(Manuscript received August 19, 1982)

A major contribution to system outage in a terrestrial digital radio channel is deep fading of the frequency transfer characteristic, which in addition to causing a precipitous drop in received signal-to-noise ratio (s/n) also causes signal dispersion that can result in severe intersymbol interference. Because the temporal variation of the channel is slow compared to the signaling rate, the information theoretic channel capacity and the "Efficiency Index" in bits/cycle—a figure-of-merit we use for the communication techniques considered—can be viewed as random processes. Starting from an established mathematical model characterizing fading channels (derived from extensive measurements), we estimate the probability distribution of channel capacity and the distributions of efficiency indices for different communications techniques. The repertoire of communication methods considered involves quadrature amplitude modulation with adaptive linear and decision feedback equalization, and maximum likelihood sequence estimation. For specific outage objectives the maximum number of bits per cycle achievable by each technique is estimated. The sensitivity of the distributions to bit-error-rate objective and unfaded s/n is assessed. For certain desired operating points the efficacy of adaptive equalization is demonstrated. There are some operating points where adaptive equalization alone is not adequate and therefore space diversity should be considered. An estimate of the effect of frequency diversity is also included.

I. INTRODUCTION AND SUMMARY

Fading of terrestrial digital radio channels owing to multipath reception is a prime cause of system outage. For a specific hop a mathematical model of these fades has been developed by W. D. Rummler^{1,2} from extensive measurements of the channel frequency power transfer characteristic over time. The radio channel has a time-varying frequency characteristic, with additive Gaussian noise; however, the temporal variations are sufficiently slow in comparison to the data symbol rate that the characteristics can be represented as a random ensemble of static frequency power transfer functions. The presence of additive noise implies that each member of the ensemble is limited to a maximum rate of transmission of data, depending on the communication method. For each specific communication technique, the stochastic nature of the channel makes it meaningful to consider the probability distribution of data rates that can be supported at a certain bit-error-rate objective.

The purpose of this paper is to explore the relative performance of various communication techniques employing quadrature amplitude modulation (QAM), distinguished by the type of equalization method used. These techniques include variants of linear equalization, decision feedback equalization, and maximum likelihood sequence estimation (MLSE). For these methods a unified set of Chernoff bounds on the probability of error is obtained. Given a communication method, a channel impulse response, an error-rate objective, a received unfaded channel s/n , a channel bandwidth, and a signaling rate we use the Chernoff bounds to estimate the maximum number of bits per cycle of bandwidth (not necessarily integer-valued) for which the constraints are met. By computing the maximum number of bits per cycle supported by each member of a large representative population of channels, we obtain the cumulative probability distribution function. One can use the cumulative distribution curve to determine the probability of outage at a prescribed bit rate.

The information theory bound on the number of bits per cycle attainable is also derived. In a sequence of plots we compare the different schemes with each other and with the information theory limit.

If $F(r)$ is the probability distribution function of data rates associated with a communication method and we set an outage objective, ξ , then the value r_ξ for which $F(r_\xi) = \xi$ represents the maximum data rate at which it is possible to transmit and meet the outage objective. We present and discuss these distributions in the context of desired long-haul and short-haul outage objectives and rates associated with the digital hierarchy constraints. The efficacy of adaptive equalization is established. The advantage of decision feedback and MLSE over

optimum linear equalization is not very substantial. There are some desired operating points for which space diversity should be considered.

For a fair comparison of different communication techniques, the transmitter filter shape must be optimized for a fixed transmitter power. We found the performance to be insensitive to whether or not the transmitter filter is optimized and we provide a theoretical guideline to indicate when this optimization becomes significant.

Our results indicate that optimized equalizer structures yield data rates only a few bits/cycle below channel capacity. It therefore appears that higher dimensional constellations³ spanning two to four symbol intervals could go a long way toward obtaining that which can be expected practically. Although we did not analyze higher dimensional signal design or optimize the constellation in QAM, it is reasonable to expect that these techniques can offer at most an equivalent few dB increase in s/n. Another method of achieving coding gain "of the order of 3-4 dB" is described in Ref. 4. Moreover, the real limitations on the selection of signal points in a practical system will most likely arise from the nonlinear operation of radio frequency (RF) power amplifiers rather than from s/n limitations.

We argue the merits of adaptive transversal equalization and provide numerical support for our claims. This is not to say that fixed or even adjustable bump and/or slope equalizers in the frequency domain could not provide adequate performance in some cases. However, fluctuating (and sometimes nonminimum) phase distortion associated with fading and other linear filters admits robust and stable compensation via adaptive transversal filters. These structures with adjustable taps can automatically equalize any phase characteristic without noise enhancement and therefore are natural candidates in these applications, especially at a high number of levels where even small amounts of phase distortion can degrade system performance.

Our analysis was carried out with ideal models and an infinite number of taps. The actual number of taps needed in any application would be determined from experiments and/or more detailed analysis.

II. THE EQUALIZED QAM SYSTEM—IDEALIZED MODEL

The use of equalizers to mitigate the effects of intersymbol interference and noise in voiceband data transmission is by now standard practice. We are thus naturally led to consider the application of these techniques in digital data transmission over the radio channel where slowly varying, frequency-selective fading is the predominant impairment. Here we review and derive the applicable mathematical theory that will be used in the sequel to evaluate the system performance indices.

To focus on basics and avoid extensive numerical analysis, we consider idealized equalizer models represented as transversal filters with an infinite number of taps. Tap adjustment algorithms are well established and our formulas are derived under the assumption that the taps have converged to their optimum values.

Our analyses are based on the digital communications model depicted in Fig. 1. To appreciate the applicability and generality of this baseband model to digital radio communications, we observe that, for any bandpass linear channel, the output waveform, when the input is any linearly modulated data wave, can be represented as

$$s(t) = \text{Re} \left\{ \sum_n \tilde{a}_n \tilde{h}(t - nT + t_0) \exp[i(2\pi f_0 t + \theta)] \right\},$$

where $\text{Re}\{\cdot\}$ stands for the "real part." The data symbols $\{\tilde{a}_n\}$ transmitted at T -second intervals, are statistically independent and, in two-dimensional modulation systems such as QAM, they assume complex values. The overall equivalent baseband impulse response, $\tilde{h}(t)$, is also complex-valued. The real part represents the in-phase response, while the imaginary part is the quadrature component. The frequency, f_0 , is the carrier frequency, θ is the carrier phase, and, t_0 is the timing phase. Ideal demodulation with a known carrier frequency f_0 and carrier phase θ implies a translation of the received bandpass signal to baseband. The real part of the resulting complex signal represents the in-phase modulation, while the imaginary part is the quadrature modulation. This then is the rationale, in addition to economics of notation, for using the complex baseband model depicted in Fig. 1.

We restrict our treatment to ideal Nyquist systems with no excess bandwidth. This permits less cumbersome calculations without loss of physical insights. We also derive our formulas by assuming flat transmitting filters and prove later that in-band optimum shaping yields imperceptible additional benefits. Also neglected is adjacent channel interference, as ideal bandlimiting eliminates this problem.

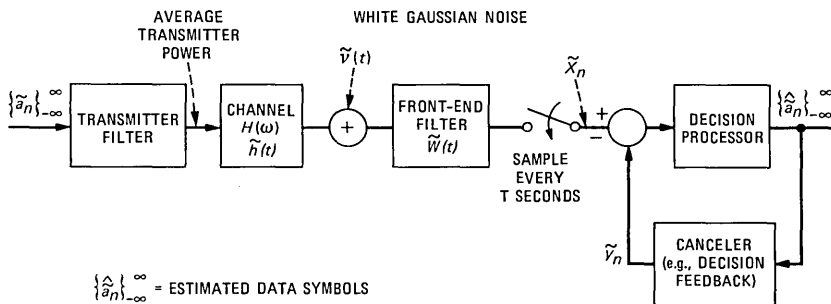


Fig. 1—Complex baseband model for QAM data transmission.

We now return to Fig. 1 and discuss the various functions and notations indicated. Without loss of generality we assume that the complex data symbols, $\{\tilde{a}_n = a_n + ib_n\}_{-\infty}^{\infty}$, take on values on a set of positive and negative odd integers with equal probability. Accordingly,

$$E\tilde{a}_n = 0 \quad \text{and} \quad E|\tilde{a}_n|^2 = 2 \frac{L^2 - 1}{3},$$

where $E(\cdot)$ denotes mathematical expectation and L (even) is the maximum number of data levels assumed by a_n and b_n . Thus, in QAM L^2 data points are available for conveying information and the source therefore generates

$$R = \frac{\log_2 L^2}{T} \text{ bits/sec.} \quad (1)$$

For a given channel bandwidth, \mathcal{W} , the efficiency index is defined as

$$I = R/\mathcal{W} = \frac{2 \log_2 L}{\mathcal{W}T} \text{ bits/cycle.} \quad (2)$$

As we shall see, the relationship among P_e -probability of error, s/n , \mathcal{W} , T and $H(\omega)$ -channel frequency characteristics is rather complicated. The determination of the relationship for different communication techniques is our chief task in the sequel.

From a mathematical point of view, the fading radio channel is characterized by a slowly varying linear distorting filter whose base-band equivalent complex impulse response is the Fourier transform of the transfer function $H(\omega)$, shifted to zero frequency:

$$\tilde{h}(t) = h_1(t) + ih_2(t) = \int_{-2\pi\mathcal{W}}^{2\pi\mathcal{W}} H(\omega) e^{i\omega t} \frac{d\omega}{2\pi}.$$

At the receiver the added complex noise process, $\tilde{v}(t) = v_1(t) + iv_2(t)$, is assumed to be white Gaussian with $v_1(t)$ independent of $v_2(t)$ and each possessing a double-sided spectral density, $N_0/2$. So,

$$\begin{aligned} E|\tilde{v}(t)|^2 &= Ev_1^2(t) + Ev_2^2(t) \\ &= N_0\delta(0), \end{aligned}$$

where $\delta(\tau)$ is the Dirac delta function. The average transmitted signal power, P_0 , for a flat transmitting filter can easily be calculated. However, for our purposes a more relevant quantity is the received, unfaded signal power

$$P = K^2 P_0 = 2 \frac{L^2 - 1}{3} \frac{K^2}{T^2},$$

where K is a constant that includes the effects of amplifiers, antennas,

and the unfaded channel loss. Also, the added average noise power in the Nyquist band, $\mathcal{W} = 1/2T$, is

$$P_v = \frac{N_0}{T}.$$

Thus the unfaded received s/n , a most important system parameter, is

$$s/n = \rho = 2 \frac{L^2 - 1}{3} \frac{K^2}{N_0} \frac{1}{T}. \quad (3)$$

The receiver structures under consideration consist of a perfect demodulator followed by a front-end filter possessing the complex impulse response $\tilde{W}(t)$, a sampler, a decision device, and a canceler. The design of an optimum receiver entails the selection of $\tilde{W}(t)$ and the canceler for a particular channel characteristic. Since the channel characteristics are usually unknown to the receiver, these components must be determined adaptively.

To understand the function of the canceler, consider the signal sample at the output of filter $\tilde{W}(t)$,

$$\tilde{x}_n = \sum_{k=-\infty}^{\infty} \tilde{r}_k \tilde{a}_{n-k} + \tilde{z}_n, \quad -\infty \leq n \leq \infty, \quad (4)$$

where $\tilde{r}_k = \tilde{r}(kT + t_0)$ is the overall complex-sampled system impulse response and \tilde{z}_n is the Gaussian noise output sample. Ideally the canceler strives to synthesize the value

$$\tilde{y}_n = \sum_{k \in S} \tilde{r}_k \tilde{a}_{n-k} \quad (5)$$

and to subtract it from (4) where the set of integers S is defined as $\{k \in S: k = -N_1 \dots -1, 1 \dots N_2\}$. The canceler's ability to synthesize these values presumes that some past ($k = 1, \dots, N_2$) and/or future ($k = -1, \dots, -N_1$) transmitted data symbols are perfectly detected and, moreover, that the set of complex numbers, \tilde{r}_k , are adaptively estimated.

The front-end filter, $\tilde{W}(t)$, is usually determined adaptively by minimizing the mean-squared error (MSE) between the sample, $\tilde{x}_n - \tilde{y}_n$, and the expected data symbol \tilde{a}_n :

$$MSE[N_1, N_2, \tilde{W}(t)] = E|\tilde{x}_n - \tilde{y}_n - \tilde{a}_n|^2, \quad (6)$$

and the optimum filter, $\tilde{W}(t)$, is chosen to achieve

$$(MSE)_0 = \min_{\tilde{W}(t)} MSE[N_1, N_2, \tilde{W}(t)] = MSE[N_1, N_2, \tilde{W}_0(t)]. \quad (7)$$

Since (6) is a quadratic functional of $\tilde{W}(t)$, a unique minimum can always be found. It is standard to represent the linear filter $\tilde{W}(t)$ by a

transversal structure and in practice the search for the minimum is accomplished by varying the taps of this filter until a minimum of the time average of the squared error is found. Clearly, to realize such a minimization procedure, estimates of the transmitted data symbols must be used.

III. SYSTEM PERFORMANCE—GENERAL

To get at the efficiency index of a system, the error rate as a function of data rate for any choice of the canceler set, $\{S\}$, and front-end filter, $\tilde{W}(t)$, must be explicitly expressed. Unfortunately, exact relationships are not mathematically tractable for the simplest of systems and so we must employ upperbounds. Fortunately, for the systems under consideration, it is possible to obtain exponentially tight inequalities.

With this approach in mind, note that after perfect cancellation, the decision variable, from (4) and (5) becomes

$$\begin{aligned} s_n &= \tilde{x}_n - \tilde{y}_n \\ &= \tilde{r}_0 \tilde{a}_n + \sum_{k \notin J} \tilde{r}_k \tilde{a}_{n-k} + \tilde{z}_n, \end{aligned} \quad (8)$$

where now the set J is $S \cup 0$, $\{k \in J: k = -N_1 \dots 0 \dots N_1\}$. Decisions in QAM are made on the real part of s_n and, separately, on the imaginary part of s_n . Simple calculations give

$$\text{Re}(s_n) = \mu_0 a_n - v_0 b_n + \sum_{k \notin J} (\mu_k a_{n-k} - v_k b_{n-k}) + z_{n1},$$

and

$$\text{Im}(s_n) = \mu_0 b_n + v_0 a_n + \sum_{k \notin J} (\mu_k b_{n-k} + v_k a_{n-k}) + z_{n2}, \quad (9)$$

where

$$\tilde{r}_k = \mu_k + i v_k,$$

and

$$\tilde{z}_k = z_{k1} + i z_{k2} = \int_{-2\pi W}^{2\pi W} \tilde{v}(t) \tilde{W}(t) dt.$$

For an L -level system, slicing levels are placed at $0 \pm 2\mu_0 \dots \pm \mu_0(L-2)$ and compared with the received samples $\text{Re}(s_n)$ and $\text{Im}(s_n)$. An error occurs whenever the noise plus intersymbol interference (in-phase and quadrature) exceed in magnitude the distance from the transmitted level to the nearest decision threshold, μ_0 . However, the outside two levels can be in error in one direction only.

Now denote the event of an error committed in the "real" rail by E_r ,

and in the “imaginary” rail by E_i . Then the probability of system error, P_e , is the probability of either (or both) E_r or E_i occurring,

$$P_e = P(E_r \cup E_i) \leq P(E_r) + P(E_i), \quad (10)$$

where

$$P(E_r) = \left(1 - \frac{1}{L}\right) \Pr \left[\left| z_{n1} - \sum_{k \notin J} (\mu_k a_{n-k} - v_k b_{n-k}) + v_0 b_n \right| \geq \mu_0 \right]$$

and

$$P(E_i) = \left(1 - \frac{1}{L}\right) \Pr \left[\left| z_{n2} - \sum_{k \notin J} (\mu_k G_{n-k} + v_k a_{n-k}) - v_0 a_0 \right| \geq \mu_0 \right]. \quad (11)$$

Because of symmetry, $P(E_r) = P(E_i) = P(E)$, and so we only need to upperbound $P(E)$.

We adopt a bounding procedure introduced by B. Saltzberg⁵ to analyze the error rate in an unequalized baseband system. We have extended Saltzberg’s approach to our systems and it can be shown that

$P(E; A, B, \delta)$

$$\leq 2 \exp \left\{ \frac{\left[\mu_0 - (L-1) \left(\sum_{k \in A} |\mu_k| + |v_k| + \delta v_0 \right) \right]^2}{2 \left\{ \sigma_{z_1}^2 + \frac{L^2-1}{3} \left[\sum_{k \in B} \mu_k^2 + v_k^2 + (1-\delta)v_0^2 \right] \right\}} \right\}. \quad (12)$$

The set of integers A and B form a partition on the set of integers not included in J . That is,

$$A \cup B = \Omega = \{k: k \notin J\}$$

and

$$A \cap B = \phi.$$

The variable $\delta = 1$ or 0 , and

$$\sigma_{z_1}^2 = \frac{N_0}{2} \int_{-\infty}^{\infty} |\tilde{W}(t)|^2 dt.$$

The sharpest upperbound is obtained by minimizing (12) with respect to the sets A , B , and δ . Algorithms for carrying out this minimization can be devised readily.

IV. SYSTEM PERFORMANCE

4.1 Discussion

Equation (12) is a rather general upperbound on the error rate for any passband linear data transmission system and it will now be specialized to include the effects of the different choices of equalizers. Before proceeding with the detailed numerical analysis, we need to make a connection between the mean-squared error (MSE), which is minimized by equalizers, and the system probability of error, which, ideally, should be the quantity minimized.

A straightforward but tedious approach for getting at the error rate might be to first determine the filter, $\tilde{W}(t)$, which minimizes the MSE for any particular equalization scheme, calculate the overall resulting impulse response, and then use eq. (12) to upperbound the error rate. This approach can be circumvented by exploiting the explicit relationship between the minimum MSE and the value of the overall impulse response at $t = t_0$ when the optimum filter, $\tilde{W}_0(t)$, is used.

The optimum structure of the minimum mean-squared error receiver can be shown to consist⁶ of a matched filter in cascade with a transversal filter combined with a linear intersymbol interference canceler. The implication of this structure is that the resulting overall system transfer function is a real function of frequency. Or, the complex-sampled impulse response, $\{\mu_k + iv_k\}_{-\infty}^{\infty}$, must be a real number at $k = 0$, which results in $v_0 = 0$. This follows from the Fourier Transform representation of $\tilde{r}(t)$, from which we see that at $t = 0$ the integrand is real and nonnegative. Indeed the overall phase characteristic has been removed by the matched filter (without enhancing the noise*). Using the fact that $v_0 = 0$ and careful numerical analysis of the available channel characteristics, our calculations showed that for all practical purposes the bound (12) becomes

$$P(E, S) \leq 2 \exp \left\{ \frac{-\mu_0^2}{2 \left[\sigma_{z_1}^2 + \sigma^2(L) \sum_{k \notin S} (\mu_k^2 + v_k^2) \right]} \right\}, \quad (13)$$

where we set $\sigma^2(L) = \frac{L^2 - 1}{3}$.

As will become apparent, the argument of the exponential function in (13) can be directly related to the minimum mean-squared error.

* A fractionally spaced ($T/2$) transversal filter can automatically synthesize any matched filter and thus eliminate phase distortion and also compensate for timing phase (see Ref. 7.)

Towards this end we recall a well-known⁶ result that states that the best achievable MSE has the simple representation,

$$(MSE)_0 = 2\sigma^2(L)(1 - \mu_0), \quad (14)$$

where μ_0 is the sample at $t = t_0$ at the output of the optimum filter. Also, when the optimum filter, $\tilde{W}_0(t)$, is used, a straightforward calculation of the resulting MSE gives

$$(MSE)_0 = 2\sigma^2(L)(1 - \mu_0)^2 + 2\sigma^2(L) \sum_{k \notin S} (\mu_k^2 + v_k^2) + 2\sigma_{z_1}^2. \quad (15)$$

Relationships (14) and (15) make it possible to write (13) as

$$P(E; S) \leq 2 \exp \left\{ -\frac{1}{(MSE)_0} \left[1 - \frac{(MSE)_0}{2\sigma^2(L)} \right]^2 \right\} \\ \sim 2 \exp -\frac{1}{(MSE)_0} \quad \text{for } N_0 \rightarrow 0, \quad (16)$$

relating error rate and minimum MSE. This is an extremely useful inequality since $(MSE)_0$ as a function of channel characteristics is often explicitly known for different equalizer structures.

It is also interesting to note (this has been pointed out before⁸) that the filter, $\tilde{W}(t)$, that minimizes MSE also minimizes the upperbound on P_e . This is true because the same quadratic functionals in $\tilde{W}(t)$ are involved in the optimization of both expressions.

We are now in a position to specialize our formulas to the various equalizer structures under investigation.

The six examples that follow do not require the knowledge of channel phase characteristics to compute performance. Implicit in each of these schemes is the complete removal of phase distortion, which can be accomplished without noise enhancement. Only a magnitude characterization of the channel transfer response was available at the time the work reported here was done. While departure from flatness of the magnitude fundamentally affects performance, theoretically, departure of phase from linear has no effect on attainable performance. Therefore, the lack of phase characterization of the channels was not an obstacle to our study. However, a complex characterization of the channel would be useful in determining the minimum number of required taps in the designs of the equalizers.

4.2 Pure phase equalization

In this particular equalizer, $N_1 = N_2 = 0$ (where N_1 and N_2 are the lengths of the precursive and postcursive cancelers, respectively). We choose $\tilde{W}(t)$ so that

$$\begin{aligned}
 W(\omega) &= e^{-i\phi(\omega)}, |\omega| \leq \frac{\pi}{T} \\
 &= 0, |\omega| > \frac{\pi}{T},
 \end{aligned}$$

where $W(\omega)$ is the Fourier transform of $\tilde{W}(t)$ and $\phi(\omega)$ is the channel phase characteristic. For this choice of filter, only the magnitude of the channel transfer function enters into the computation of the bound, as shown in eq. (13).

Using the well-known Poisson sum formula along with some algebra, it is possible to write (13) more explicitly, i.e.,

$$P_e \leq 2 \exp \left\{ -\frac{\rho}{2\sigma^2(L)} \frac{\langle H \rangle^2}{1 + \rho \langle (H - \langle H \rangle)^2 \rangle} \right\}, \quad (17)$$

where we used the shorthand notation

$$\langle \cdot \rangle = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} [\cdot] d\omega$$

and H is used in place of $|H(\omega)|$.

4.3 Linear equalization

Here again $N_1 = N_2$ (no canceler) and $\tilde{W}(t)$ is chosen to minimize (6). The expression for the optimum MSE in this case has been shown to be⁹

$$(MSE)_0 = 2\sigma^2(L) \left\langle \frac{1}{1 + \rho H^2} \right\rangle. \quad (18)$$

This formula is directly used in (16) to calculate the upperbound on error rate:

$$P_e < 2 \exp \left[-\frac{1}{2\sigma^2(L)} \left(\left\langle \frac{1}{1 + \rho H^2} \right\rangle \right)^{-1} \right].$$

4.4 Inverse equalization

In this case $N_1 = N_2 = 0$ (no cancellation) we choose $\tilde{W}(t)$ to be the inverse of the channel frequency characteristics,

$$\begin{aligned}
 W(\omega) &= H^{-1}(\omega), |\omega| \leq \frac{\pi}{T} \\
 &= 0, |\omega| > \frac{\pi}{T}.
 \end{aligned}$$

Here the channel is clearly perfectly equalized so that intersymbol interference is completely eliminated; the penalty is increased output

noise power. For this simple scheme, it is possible to express the error rate exactly but for reasons of uniformity we use the upperbound

$$P_e \leq 2 \exp - \left\{ \frac{\rho}{2\sigma^2(L)} \frac{1}{\langle \frac{1}{H^2} \rangle} \right\}. \quad (19)$$

4.5 Decision feedback

For this equalization system $N_1 = 0$ and $N_2 = \infty$ and again we choose $\tilde{W}(t)$ to minimize (6). In this type of equalizer, the causal or postcursor intersymbol interference is entirely eliminated and an expression for the optimum MSE is known,^{10,11} as shown below.

$$(MSE)_0 = \sigma^2(L) \exp\{-\langle \ln[1 + \rho H^2] \rangle\}. \quad (20)$$

This is used in (16) to express an upperbound on error rate.

4.6 The ideal equalizer

In this utopian scheme the precursor and postcursor cancelers become infinite, $N_1 = N_2 = \infty$, so that all the intersymbol interference is eliminated. In this ideal situation we obtain the very best possible result, namely, the matched filter bound, which is a lower bound on P_e . This scheme assumes that it is possible to detect each pulse $\tilde{a}_k \tilde{r}(t - nT)$ optimally by a matched filter without incurring interference from all other pulses. The filter, $\tilde{W}(t)$, in this case is chosen to be matched to the channel characteristic, i.e.,

$$\begin{aligned} W(\omega) &= H^*(\omega), \quad |\omega| \leq \frac{\pi}{T} \\ &= 0, \quad |\omega| > \frac{\pi}{T}, \end{aligned}$$

where * denotes the complex conjugate.

For this idealization the upperbound on error rate is simply,

$$P_e \leq 2 \exp \left\{ - \frac{\rho}{2\sigma^2(L)} \langle H^2 \rangle \right\}. \quad (21)$$

No other detection scheme can do better. In the next section we use these formulas to calculate the various efficiency indices.

Before concluding this section, we remark that there is one more easy case and one extremely difficult case that might be considered as candidates for making comparisons. Suppose that no filtering, other than out-of-band elimination of noise, were performed at the receiver. What performance can one expect? While we cannot answer this question exactly because channel phase characteristics are unavailable

at this writing, we would expect performance to be worse than removing the phase characteristic entirely—a situation we will examine.

The second approach, which is a very difficult one to analyze, involves the use of a finite-state Viterbi decoder. Nevertheless, we will report a bound on the performance of this processor. Specifically, the performance of an infinite canceler (the matched filter) is superior to the maximum likelihood (Viterbi) decoder. As shown later, for the channels considered, the matched filter bound is close in performance to decision feedback. Consequently, we shall see that the performance of maximum likelihood sequence estimation is tightly bracketed because it is superior to decision feedback.

4.7 Information theory bound on communications efficiency index

In this section we discuss a formula for the maximum number of bits per cycle that can be attained for a given $H(\omega)$. If $H(\omega)$ were constant in frequency the formula for the efficiency index in bits per cycle would be simply

$$I = \log_2(1 + \rho |H|^2).$$

It is reasonable to expect that if $H(\omega)$ is frequency-dependent, the maximum efficiency index would be

$$I = \frac{1}{\Omega} \int \log_2(1 + \rho |H(\omega)|^2) d\omega, \quad (22)$$

where the integral is over a frequency band of size $\Omega = 2\pi\mathcal{W}$. Indeed this is the case. To outline a derivation, we note first that A. Kolmogorov has generalized Shannon's notion of capacity to provide a very fundamental definition that gives a useful starting point for developing capacity formulas in nonstandard situations such as the one at hand.¹² M. S. Pinsker¹³ was able to derive from the Kolmogorov approach a formula for the amount of information in a stationary Gaussian process about another stationary Gaussian process related to it. Specifically, if $S_x(\omega)$ and $S_y(\omega)$ are the power spectral densities of the processes and $S_{xy}(\omega)$ is the cross-spectral density, the formula for the amount of information is

$$- \int \ln \left(1 - \frac{|S_{xy}(\omega)|^2}{S_x(\omega)S_y(\omega)} \right) d\omega.$$

If we require the transmitter output to be Gaussian, then since the additive noise is Gaussian, Pinsker's formula can be applied to the case where x is the transmitted process and y is the received process to obtain (22). Requiring the transmitter output to be Gaussian is really not a limitation since, when the additive noise is Gaussian, one can prove that capacity is attainable with a Gaussian transmitter output

using the methods discussed in Refs. 14 through 16. Thus, (22) gives the efficiency index formula.

References 16 and 17 also provide approaches to establishing (22).

4.8 Information theory limit on index when transmitter is optimized

So far we have treated ρ , eq. (3), as a constant. In this section we set the stage for exploring the advisability of optimizing the output power spectral density to maximize the efficiency index. Since we are now allowing in-band shaping of the transmitter filter frequency characteristic, we will consider ρ as a function of ω and write $\bar{\rho}$ to denote the previously considered case where ρ is constant over the band.

Although the analysis in this section is focused on the information theory limit on the efficiency index, the decision feedback index involves the identical functional form and so our analysis will be applicable to decision feedback as well.

We will compare the previously discussed index

$$I(\text{flat}) = \frac{1}{\Omega} \int \log_2(1 + \bar{\rho} |H(\omega)|^2) d\omega$$

with

$$I(\text{opt}) = \frac{1}{\Omega} \int \log_2(1 + \rho_0(\omega) |H(\omega)|^2) d\omega,$$

where $\rho_0(\omega)$ is the function maximizing I under a constraint on $\int \rho(\omega) d\omega$, the received signal-to-noise ratio in the absence of fading. This constraint is equivalent to a constraint on the transmitter output power, since in the absence of fading the channel has a flat loss characteristic.

This optimization problem is known¹⁷ and yields easily to the calculus of variations. The solution is called "water pouring." The name stems from the graphical interpretation that if $\bar{\rho}\Omega$ is the constraint on $\int \rho(\omega) d\omega$, the optimum $\rho(\omega)$, which we denote by $\rho_0(\omega)$, is obtained by forming a vessel with base $|H(\omega)|^{-2}$ and vertical sides at the band edges. One pours "water," that is, area, of amount $\bar{\rho}\Omega$ into the vessel and $\rho_0(\omega)$ is given by the depth of the water at ω . It is clear that this construction obeys the constraint. Generally, if $\bar{\rho}$ is sufficiently small, the poured water will not touch both of the vertical sides of the vessel. In such situations, it would be advantageous to limit the transmitted power to a frequency band less than Nyquist. In our case, however, the unfaded signal-to-noise ratio is so great that, for the simulated channels, the water level always meets both vertical sides. In other words, the transmitted power always occupies the full Nyquist band. Thus, $\rho_0(\omega) = A - |H(\omega)|^{-2}$, where A is chosen so that $\int \rho_0(\omega) d\omega = \bar{\rho}\Omega$, that is, $A\Omega - \int |H(\omega)|^{-2} d\omega = \bar{\rho}\Omega$ or

$$\rho_0(\omega) = \bar{\rho} + \frac{1}{\Omega} \int |H(v)|^{-2} dv - |H(\omega)|^{-2}.$$

Therefore,

$$\frac{I(\text{flat})}{I(\text{opt})} = \frac{\int \log_2(1 + \bar{\rho}|H(\omega)|^2) d\omega}{\int \log_2 \left[\frac{|H(\omega)|^2}{\Omega} \int |H(v)|^{-2} dv + \bar{\rho}|H(\omega)|^2 \right] d\omega}.$$

We expand the logarithm for $\bar{\rho}$ large to find an asymptotic representation. We get, after a cumbersome derivation,

$$\begin{aligned} \frac{I(\text{flat})}{I(\text{opt})} = 1 - \frac{1}{2\bar{\rho}^2 \log_2 \bar{\rho}} \left(1 + \frac{\int \log_2 |H|^2 d\omega}{\Omega \log_2 \bar{\rho}} \right)^{-1} \\ \cdot \left[\frac{\int \frac{1}{|H|^4} d\omega}{\Omega} - \left(\frac{\int \frac{1}{|H|^2} d\omega}{\Omega} \right)^2 \right] + O \left(\frac{1}{\bar{\rho}^3 \log_2 \bar{\rho}} \right). \quad (23) \end{aligned}$$

The last multiplier in the perturbation expression represents the variance associated with a random sampling of a specific $|H(\omega)|^{-2}$. This multiplier is zero if $|H(\omega)|^{-2}$ is a constant. We note, for later use, that for $\bar{\rho} = 10^{6.3}$, (i.e., a 63-dB s/n in the absence of fading) we have

$$\left(\frac{1}{2\bar{\rho}^2 \log_2 \bar{\rho}} \right) < 10^{-14}.$$

4.9 Communication efficiency index

The relationships in eqs. (17) through (21) are unifying expressions for the error rate for the five equalization cases. From Section 1.0 we have $I = 2\log_2 L$ and $\sigma^2(L) = [(L^2 - 1)/3]$. These two equations in conjunction with the P_e formulas enable us to determine I as a function of the P_e objective, ρ , the channel, and the equalization scheme. Specifically,

$$I = \log_2 \left[\frac{G(H, \rho)}{\ln(P_e/2)} + 1 \right],$$

where $G(H, \rho)$ is a function that depends on the communication method. With the channel response considered to be a random function, I is a random variable, and we can determine its probability distribution function for each communication scheme. The quantities ρ and P_e are parameters of the distribution.

V. MODEL FOR THE FADING CHANNEL

Now we describe the mathematical model for frequency-selective fading owing to multipath reception. This mathematical model for the random functions $|H(\omega)|^2$ is due to W. D. Rummler^{1,2} and is based on measurements of a 26.4-mile hop between Palmetto and Atlanta, Georgia. The measurements of frequency-selective fades were made on a 25.3-MHz channel situated in the 6-GHz band during the heavy fading month of June (1977).

Rummler's model uses a two-ray representation of the signal, which was quite adequate for fitting the experimental records. (The model is not necessarily intended to depict the underlying physical mechanism for a fade. The true mechanism could involve a much more complex ray combination.) Also, it is not possible to deduce the phase characteristic associated with any particular amplitude characteristic. It has been experimentally determined that this kind of a channel cannot always be viewed as minimum phase.¹⁸

In the model, the $|H(\omega)|^2$ functions are 68-degree sections of scaled, displaced cosine waves. Specifically, conditional on a fade occurring

$$|H(\omega)|^2 = a^2 |1 + b^2 - 2b \cos(\omega\tau + \theta)|,$$

where:

(i) $b = 1/10^{B/20} > 0$ with B an exponential random variable with mean 3.8.

(ii) The parameter, a , is a log normal random variable with dependence on the parameter, b . Specifically, $a = 1/10^{A/20}$, where A is normal with a mean of $24.6(B^4 + 500)/(B^4 + 800)$ dB and a standard deviation of 5 dB.

(iii) The phase, θ , is independent of a and b and has a constant density on each section $|\theta| > \pi/2$ and $|\theta| < \pi/2$ with $P\{|\theta| < \pi/2\} = 5 \cdot P\{|\theta| > \pi/2\}$.

(iv) The scale factor τ is a constant = 6.31 nanoseconds.

In the model, the channel is in the faded state for 8060 seconds in a normal heavy fading month. Thus the channel can be viewed as being in one of two states where:

$$P \{\text{unfaded state}\} = 0.99689$$

$$P \{\text{faded state (Rummler model operative)}\} = 0.00311.$$

In what follows we employ this model to estimate the outage distributions for various communication methods. The model should be regarded as a working assumption valuable in gaining initial insight into the potential of the communication techniques we consider. However, we emphasize that more measurements may be required to refine Rummler's model to accommodate different geographical situations and wider bandwidths than 25 MHz, and to sharpen the accu-

racy of the representation of the more severe fades that are of major concern in what follows.

VI. OUTAGE OBJECTIVES AND SOME PROSPECTIVE INDEX VALUES

From the proposed performance objectives for the digital transmission network,¹⁹ we have that the round trip system availability objective is 99.98 percent. So the probability of outage is 0.0002 round trip or 0.0001 one way. The 0.0001 breaks down as 0.00005 for fading, 0.000025 for equipment failure, and 0.000025 for maintenance and plant errors. Thus, for a 4000-mile system composed of 156 hops, each with a nominal length of 25.6 miles, we get a per hop outage probability of 3.2×10^{-7} for fading. This corresponds to about 10 seconds of outage per year. If we assume the year is composed of three heavy fading months and nine months with no fading, we obtain that the probability of outage in a heavy fading month is 1.28×10^{-6} . For a 250-mile short-haul system, the outage objective is 16 times less stringent on a hop, namely, 2.05×10^{-5} .

For the purpose of discussion we shall later consider the possibility of accommodating two DS-3 digital signals in a 20-, 30-, and 40-MHz channel. Each DS-3 signal corresponds to 672 64 kb/s voice circuits, so that two DS-3 signals correspond to about 90 Mb/s. Thus, for 20-, 30-, and 40-MHz channels we require 4.5, 3, and 2.25 bits per cycle, respectively. We will use 10^{-4} as the probability of bit-error threshold for registering outage. The sensitivity to this threshold will also be analyzed.

VII. COMPUTER PROGRAM

A comprehensive FORTRAN program was written to compute and display outage distributions. The program is composed of three main segments.

The first segment simulates the power transfer characteristic for the channel in the faded state. It uses a PORT routine to generate random numbers uniformly distributed on $[0, 1]$ and functions of these are evaluated to produce the random variables with the three densities underlying Rummler's model. The variables A and B are appropriately correlated. The random channel characteristics are then computed and a file containing them is produced. The file contains 25,000 characteristics.

The second segment calculates the efficiency index for each channel and then computes the probability distribution function of the indexes. Various options and parameters can be chosen in exercising this stage. These include:

(i) Method of communication (i.e., type of linear equalization, decision feedback equalization, MLSE, and the information theoretic optimum processing)

(ii) Transmitter spectrum, i.e., optimization of the transmitter spectral density for a given average power constraint or flat power spectral density

(iii) Probability of bit-error objective

(iv) Unfaded signal-to-noise ratio at the receiver input

(v) Channel bandwidth.

Option (ii) is only available for the decision feedback and the information theoretic optimum communication schemes. Extending the option to the other schemes seemed inadvisable because of the closeness of the results, as will be seen later.

The number 25,000 was found through computational experience to stabilize the density tail in the range of interest and yet not be wasteful of computer resources. Since the number 25,000 is very close to the number of experimental records of fade characteristics, we could have worked from original experimental data. We elected to work with the Rummler model since it is weighted to track the worse fades, which are our interest here, and since the model is widely accepted.

The final segment of the computer program provides labeled plots of the outage distribution functions. It uses the graphic package DISSPLA.

VIII. PRINCIPAL RESULTS

8.1 Preparatory remarks

For the purpose of presenting our principal results we will need the following notation for the outage distribution functions:

F_{PH} : phase distortion removed

F_{LIN} : optimum linear equalization

F_{DF} : postcursor intersymbol interference (ISI) removed

F_{MF} : all ISI removed (matched filter bound)

F_{IT} : information theory limit (Shannon).

The efficiency index distributions were computed for 30-MHz channels. Strictly speaking, the notion of an index in bits per cycle is imprecise in that $F(I)$ (the probability distribution of bits per cycle) would change if calculated at 20 MHz or at 40 MHz. However, by calculation we established that, for the purposes of the discussion that follows, treating $F(I)$ as invariant over the 20- to 40-MHz range of bandwidths is an adequate approximation.

In the actual development of systems of the kind we have idealized, much more detailed performance analysis is required than that reported here. One important aspect we have not considered is the effect of excess bandwidth associated with practical filter designs with rolloff factors other than zero. To get a preliminary idea of the effect of rolloff of an amount α on the communication efficiency indexes, we would simply scale the distributions abscissas by an amount $1/(1 + \alpha)$.

Equivalently, we would inflate the desired number of bits per cycle by $1 + \alpha$ before going to the curves. Thus, in considering the accommodation of two DS-3 signals, with $\alpha = 1/3$, we would inflate the 4.5-, 3-, and 2.25-bit per cycle values corresponding to 20-, 30-, and 40-MHz channels by 1.333 to obtain 6, 4, and 3 bits per cycle. We would then consult the derived curves at these values to obtain the outage probabilities. For the purpose of discussion in the section that follows, we use these three inflated bits per cycle values along with the long-haul and short-haul objectives of 1.3×10^{-6} and 2.1×10^{-5} , respectively, given in Section VI. Subsequently, an alternative means of accounting for α will be given.

8.2 The graphs

The most striking features of the outage distribution functions $F(I; P_e, \rho)$ are exhibited in Fig. 2. The beneficial effects of adaptive equalization are apparent. The three equalization schemes yield roughly similar results; however, as one looks more closely at the extreme outage tail, F_{LIN} , F_{DF} , and F_{MF} begin to depart from each other. F_{IT} is displaced over two bits per cycle to the right of F_{MF} , while

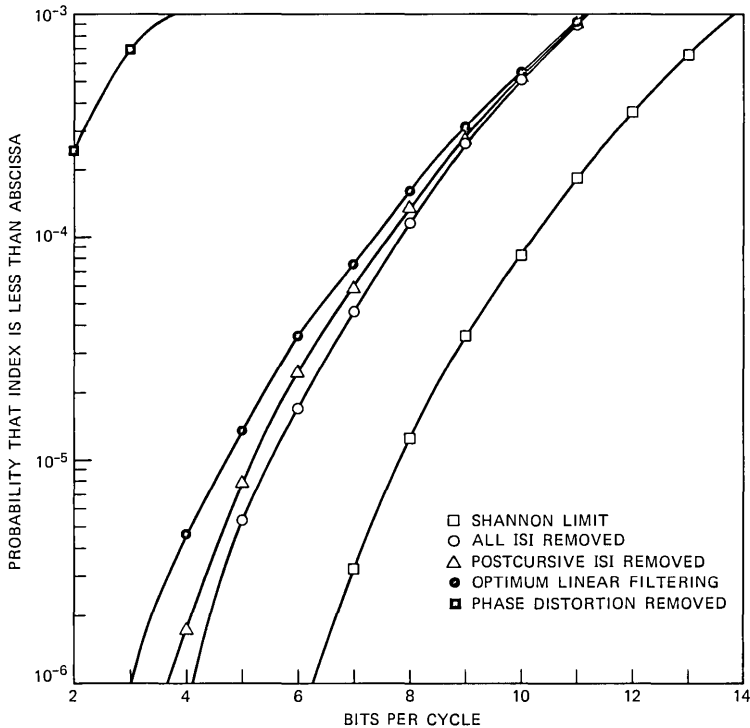


Fig. 2—Comparison of index distribution tails at 63-dB s/n, p.e. $< 10^{-4}$ (p.e. does not apply to the Shannon limit curve).

F_{PH} is very substantially to the left of F_{LIN} . For the 40-MHz band (3 bits/cycle), both outage objectives are met with linear equalization. For the 30-MHz band (4 bits/cycle) and the long-haul objective, linear and decision feedback equalization are not adequate and some coding or use of maximum likelihood sequence estimation are possible solutions. However, it may be practical to overcome the shortfall by some other means, such as improving the amplifier noise figure. For the 20-MHz channel (6 bits/cycle) the long-haul objective is not met. Also, this efficiency is so close to the information theory limit that any attempt to achieve it by coding may be ill-advised because of complexity. On the other hand, with some moderate coding the short-haul objective for 20-MHz channels should be attainable. For the other two bands, short-haul objectives are roundly met.

The plot for the equalizer that inverts the channel is not shown, as it is not perceptibly different from that for the optimum linear equalizer. This is expected since the optimum linear filter is essentially inverting the channel at the high signal-to-noise ratios we are considering.

Figures 3, 4, and 5 show the sensitivity of $F(I, P_e, \rho)$ to P_e . The

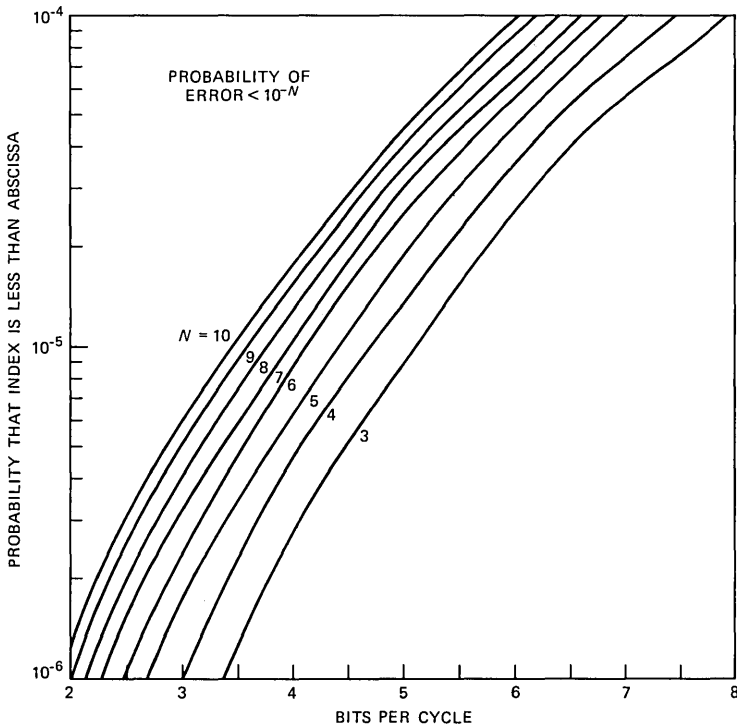


Fig. 3—Index distribution tail sensitivity to probability of error for linear equalization at 63-dB s/n.

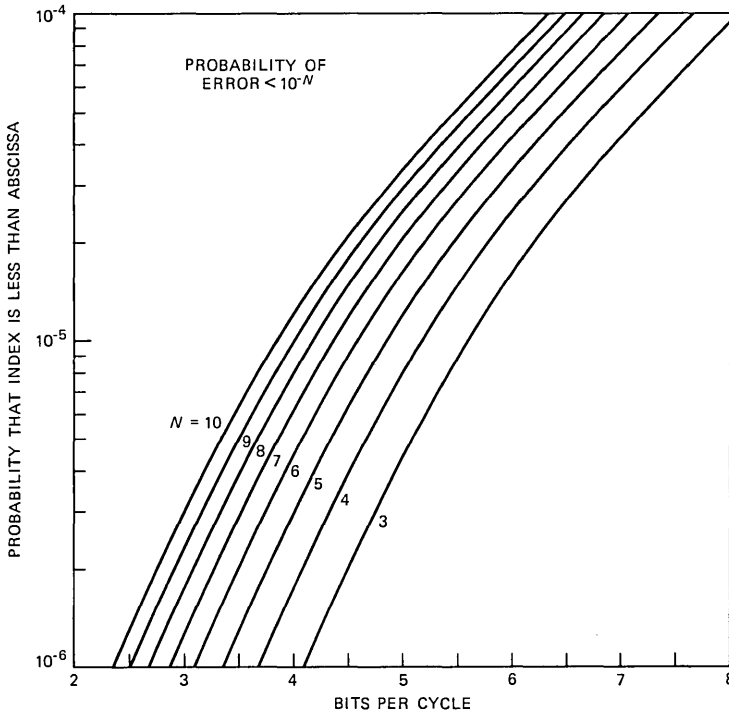


Fig. 4—Index distribution tail sensitivity to probability of error for decision feedback at 63-dB s/n.

sensitivity is especially small at the low error rates needed for data (as opposed to voice) transmission. An asymptotic analysis shows that, for large ρ , the curves translate to the left in accordance with a $\log_2 N$ shift, where 10^{-N} is the P_e objective. This insensitivity is an illustration of the well-known result²⁰ that once a pulse code modulation (PCM) operating point is achieved it takes a very small improvement to make the error rate an order of magnitude smaller. In fact, if at some operating data rate the probability of error turns out to be $10^{-5} - 10^{-6}$, one should be able to design an error-correcting code of small redundancy and moderate complexity that could improve the error rate by several orders of magnitude.

Figures 6 through 9 illustrate the sensitivity to signal-to-noise ratio. The translation in all cases is roughly 1/3-bit/cycle/dB. Note the curves for the Shannon limit have an ordinate range of 4 to 10 bits per cycle, while the others range from 2 to 8 bits per cycle. No sensitivity for F_{PH} is given since, unlike all the other distributions, there is negligible improvement as ρ increases. This is because, as ρ increases, the effect of intersymbol interference (ISI) remains and nothing is being done to mitigate it. In the other four cases, ISI tends to be

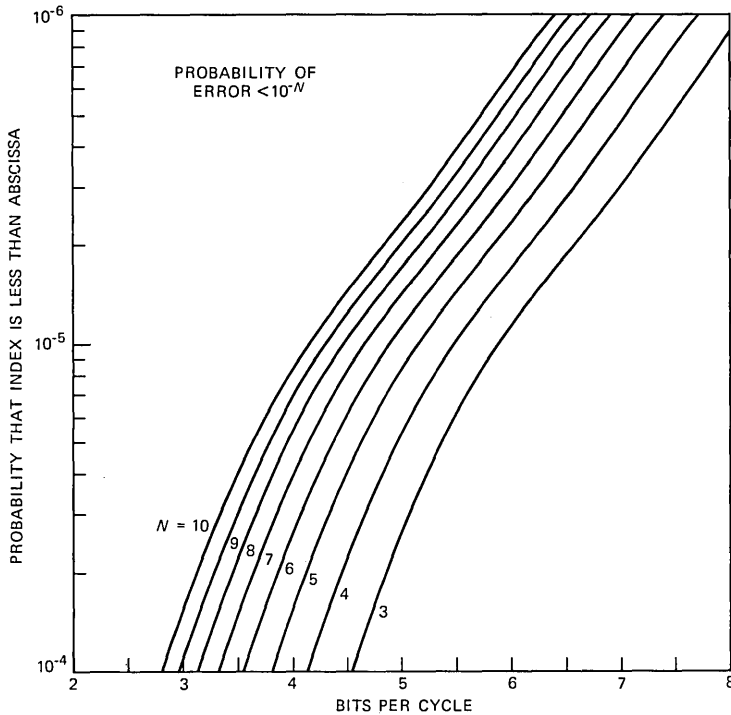


Fig. 5—Index distribution tail sensitivity to probability of error for the case of all ISI removed.

eliminated as ρ increases (the channels can be perfectly equalized without an inordinate amount of noise enhancement).

In Section 8.1 we mentioned the $(1 + \alpha)^{-1}$ scaling as a method of accounting for rolloff. This assumed $\mathcal{W} = 1/T$ so that $(1 + \alpha)\mathcal{W}$ is the actual bandwidth. Suppose instead that the real bandwidth is fixed at \mathcal{W} but the data rate is slowed by an amount $(1 + \alpha)$, leaving the average transmitter power and N_0 constant. Then the true s/n is increased by $10 \log_{10}(1 + \alpha)$. From this alternative perspective the suggested $(1 + \alpha)^{-1}$ scaling would be supplemented by a shift to the right of the probability distribution function tail by approximately $(10/3)\log_{10}(1 + \alpha)$ bits per cycle. Whether in estimating the effect of rolloff one takes the perspective that the symbol rate or the bandwidth is fixed is a matter of convenience.

Next, we consider optimization of the transmitter power spectral density. There is a practical question as to whether such an optimization could be achieved since the fade characteristic, which is first determined at the receiver, would need to be relayed back to the transmitter in time to be useful. However, the question is academic since we demonstrated that, even if an optimized transmitter could be

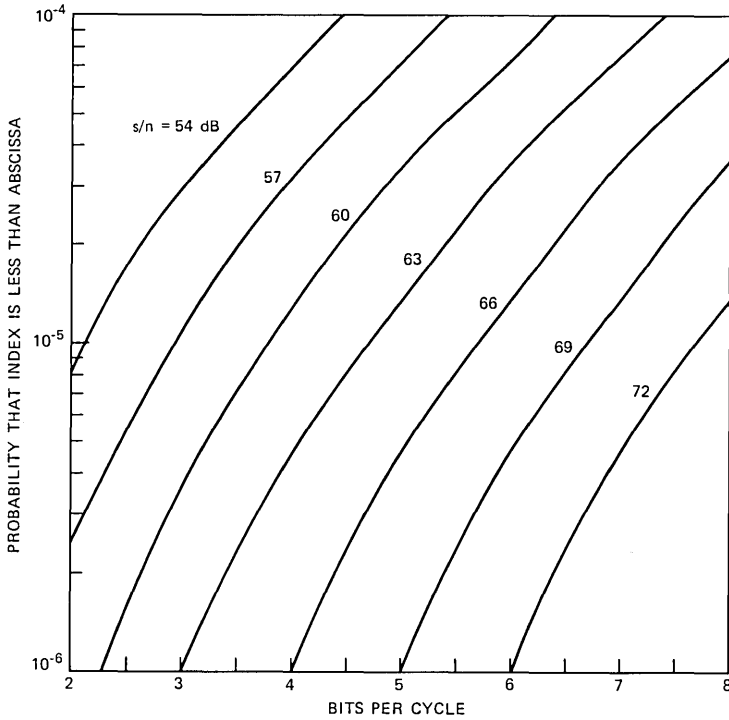


Fig. 6—Index distribution tail sensitivity to s/n for linear equalization, p.e. $< 10^{-4}$.

adapted in real time, the performance benefit would be negligible.

To understand why it is not worthwhile to optimize the transmitted power spectral density, we first consider the information theory limit that was analyzed in Section 4.7. Our detailed numerical work has shown that a plot of the tail of the index distribution for the information theory limit under the assumption of an optimized transmitter would be imperceptibly different from the F_{DF} tail plotted in Fig. 2. This closeness of the two distributions follows from the fact that, for the severest fades in our data base of 25,000 channels, $|H|^{-2}$ is of the order of 10^{-6} and the terms involving $|H|^{-4}$ ($\sim 10^{-12}$) in the perturbation expression, eq. (23), are not enough to overcome the 10^{-14} multiplier.

Since the decision feedback index has the same form as (22), we can also conclude that the distribution tail corresponding to decision feedback would not be significantly altered if the optimum transmitter were used.

The decision feedback index and information theory limit on the index give imperceptible benefits when the power spectral density is optimized; therefore, it seems extremely unlikely that there is any

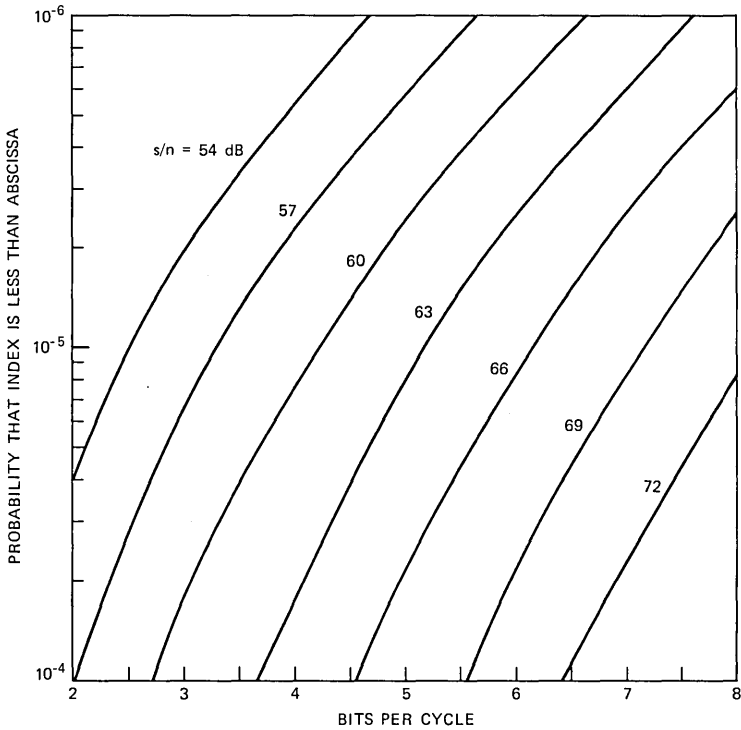


Fig. 7—Index distribution tail sensitivity to s/n for decision feedback, p.e. $< 10^{-4}$.

worthwhile benefit associated with optimizing the transmitter in the other cases.

IX. INITIAL ESTIMATE OF THE EFFECT OF FREQUENCY DIVERSITY

In implementations, digital radio systems are often protected with frequency diversity. In such systems impairments such as fading and equipment outages prompt the switching of communication traffic to a protection channel situated at a different frequency. The notation $m \times n$ means that m protection channels back up n working channels. So long as a protection channel is not occupied by an impaired channel, or is not itself impaired, it is available for temporary use in any of the n working systems. Some illustrations are 2×10 and 1×11 at 4 GHz, while at 6 GHz 2×6 and 1×7 are examples.

For FM systems the factor expressing the improvement in outage associated with frequency diversity is given by the expression $100/DG \cdot f_0 H^2$ in eq. (24) [corresponding to (34) and (35) in Ref. 21]. The parameter f_0 represents the carrier frequency in gigahertz and D denotes the path distance measured in miles. The parameter G depends

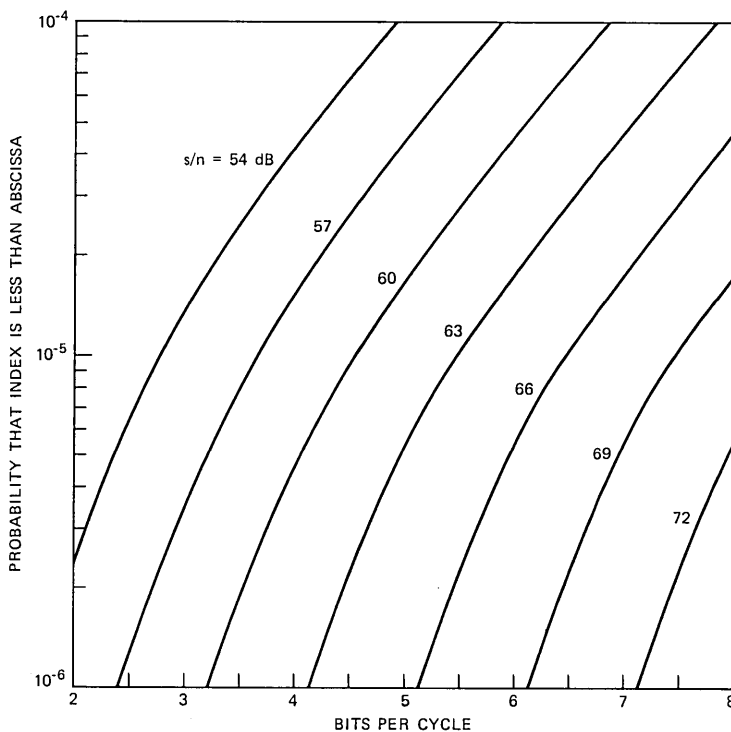


Fig. 8—Index distribution tail sensitivity to s/n for the case of all ISI removed, p.e. $< 10^{-4}$.

on the details of the frequency protection. G incorporates combinatorial effects corresponding to using m channels to back up n as well as empirical expressions involving the individual frequencies of channels involved. The term H is commonly expressed in decibels as $-20 \log H$ and is called fade margin. The fade margin is the smallest loss relative to the unfaded received signal at which the system fails. The notation H for the voltage level agrees with the previous use of H in this paper so long as the channel has a flat characteristic.

As pointed out in Ref. 22 the notation of a flat fade margin is considered meaningless in digital radio systems since the frequency-selective aspect of the fade characteristic appears necessary to describe performance of a channel. As of this writing we are not aware of any method in the extant literature for extending our results to include the effect of frequency diversity. However, for the special case of optimum linear equalization there is a way to introduce an equivalent flat fade margin so as to enable the use of (24) in making a (preliminary) estimate of the diversity effect. The estimation method was discovered in the course of generalizing the computer program to compute index

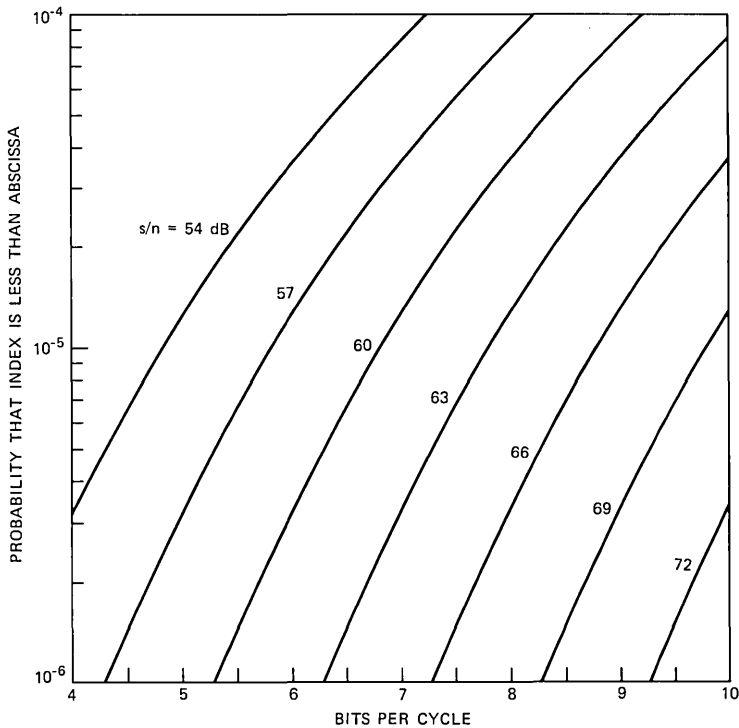


Fig. 9—Distribution tail sensitivity to s/n for the Shannon limit.

distributions for arbitrary bandwidths. The generalization was needed to investigate—and, as it turned out, to substantiate—the bandwidth insensitivity of the distribution in the 20- to 40-MHz range. The generalized program was also exercised for bandwidths an order of magnitude smaller and it was observed that the F_{LIN} distribution changed imperceptibly.* So the F_{LIN} tail in the range of interest here can be correctly obtained by using the univariate samples $|H(0)|^2$. Treating (18) as an equality and solving for $|H(0)|^2$ and then substituting in (24) gives an estimate of the improvement owing to frequency diversity.

For an illustration refer to Fig. 2 for which $s/n = 63$ dB and $Pe < 10^{-4}$. We see that with linear equalization, at the long-haul outage objective of 1.3×10^{-6} , 3.2 bits per cycle can be supported and the corresponding number for the short-haul objective is 5.3 bits per cycle. Using (24) and the G values from Ref. 21 we show in Table I the

* This is not true of the individual values of I and no claim is made for invariance of F_{DF} , F_{MF} , or F_{IT} .

Table I—Estimates of improved bits/cycle indices using frequency diversity

	Channel at 4 GHz		Channel at 6 GHz	
Short Haul	(1 × 11)	7.1	(1 × 7)	6.5
	(2 × 10)	7.8	(2 × 6)	7.1
Long Haul	(1 × 11)	5.3	(1 × 7)	4.6
	(2 × 10)	5.8	(2 × 6)	5.3

following estimates of improved indices when frequency diversity is employed.

X. ACKNOWLEDGMENT

Discussions with N. Amitay, A. Gieger, L. J. Greenstein, V. K. Prabhu, B. G. King, G. Vannucci, A. Vigants, C. B. Woodworth, W. D. Rummmler and Y. S. Yeh were valuable in the course of the work reported here.

XI. POSTSCRIPT

Application of multilevel QAM in the radio channels might be inhibited by the amplitude (AM-AM) and (AM-PM) nonlinearities present in RF power amplifiers. A method for solving this problem without sacrificing amplifier power efficiency will be described in a forthcoming paper.²³

REFERENCES

1. W. D. Rummmler, "A New Selective Fading Model: Application to Propagation Data," *B.S.T.J.*, 58, No. 5 (May-June 1979), pp. 1032-71.
2. W. D. Rummmler, "More on the Multipath Fading Channel Model," *IEEE Trans. Commun.*, COM-29, No. 3 (March 1981), pp. 346-52.
3. A. Gersho and V. B. Lawrence, unpublished work.
4. G. Ungerbock, "Channel Coding with Multilevel/Phase Signals," *IEEE Trans. Inform. Theory*, IT-28, No. 1 (January 1982), pp. 55-67.
5. B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," *IEEE Trans. Inform. Theory*, IT-14, No. 4 (July 1968), pp. 563-8.
6. M. S. Mueller and J. Salz, "A Unified Theory of Data Aided Equalization," *B.S.T.J.*, 60, No. 9 (November 1981), pp. 2023-38.
7. 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.
8. R. D. Gitlin and J. E. Mazo, "Comparison of Some Cost Functions for Automatic Equalization," *IEEE Trans. Commun.*, COM-21, No. 3 (March 1973), pp. 233-7.
9. T. Berger and D. W. Tufts, "Optimal Pulse Amplitude Modulation, Part I, Transmitter-Receiver Design and Bounds from Information Theory," *IEEE Trans. Inform. Theory*, IT-13, No. 2 (April 1967), pp. 196-208.
10. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," *B.S.T.J.*, 52, No. 8 (October 1973), pp. 1341-73.
11. D. D. Falconer and G. J. Foschini, "Theory of Minimum Mean-Square-Error QAM Systems Employing Decision Feedback Equalization," *B.S.T.J.*, 52, No. 10 (December 1973), pp. 1821-48.
12. S. P. Lloyd, "On a Measure of Stochastic Dependence," *Theory of Probability and Its Application*, No. 7, 1962, pp. 312-22.

13. M. S. Pinsker, "A Quantity of Information of a Gaussian Random Stationary Process, Contained in a Second Process Connected with it in a Stationary Manner," *Doklady Akad. Nank S.S.S.R., New Series*, 99, 1954, pp. 213-16.
14. R. K. Mueller and G. J. Foschini, "The Capacity of Linear Channels with Additive Gaussian Noise," *B.S.T.J.* 49, No. 1 (January 1970), pp. 81-94.
15. J. L. Holsinger, "Digital Communication over Fixed Time-Continuous Channels with Memory-with Special Application to Telephone Channels," Lincoln Laboratory, Technical Report 366, Lexington, MA, October 1964.
16. J. B. Thomas, "Statistical Communication Theory," New York: John Wiley and Sons, 1969, Chapter 8.
17. R. M. Fano, "Transmission of Information," New York: M.I.T. Press and John Wiley and Sons, Inc., 1961.
18. B. G. King, unpublished work.
19. M. A. Rezny and J. S. Wu, unpublished work.
20. B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The Philosophy of PCM," *Proc. IEEE* (November 1948), pp. 1324-31.
21. A. Vigants and M. V. Pursley, "Transmission Unavailability of Frequency—Protected Microwave FM Radio Systems Caused by Multipath Fading," *B.S.T.J.*, 58, No. 8 (October 1979), pp. 1779-96.
22. A. Gieger and W. T. Barnett, "Effects of Multipath Propagation on Digital Radio," *IEEE Trans. Commun., COM-29*, No. 9 (September 1981), pp. 1345-52.
23. A. A. M. Saleh and J. Salz, "Adaptive Linearization of Power Amplifier Nonlinearity in Digital Radio Systems," *B.S.T.J.*, 62, No. 4 (April 1983).

Chromatic Dispersion Measurements in Single-Mode Fibers Using Picosecond InGaAsP Injection Lasers in the 1.2- to 1.5- μm Spectral Region

By C. LIN, A. R. TYNES, A. TOMITA, P. L. LIU and
D. L. PHILEN

(Manuscript received September 23, 1982)

We describe the use of picosecond InGaAsP injection lasers for measuring chromatic dispersion in single-mode fibers in the 1.2- to 1.5- μm spectral region. Injection lasers at various wavelengths and a single-mode fiber-to-fiber switch are used in the pulse delay and pulse-broadening measurements. The simplicity and the compactness make the setup useful for field measurements and quality control.

I. INTRODUCTION

Modal and chromatic dispersion measurements in multimode and single-mode fibers provide important information about the bandwidth limitations in optical fiber transmission. A near-infrared, fiber Raman laser with subnanosecond pulses in the 1- to 1.7- μm region can be used for dispersion measurement in both multimode and single-mode fibers.¹⁻³ While this infrared-fiber, Raman-laser-based measurement system has been very useful and widely adopted, it has its limitations in terms of time resolution (by the mode-locked laser pulsewidth, ~ 140 ps) and is not suitable for field measurements because of its large size.

In this paper we describe a simpler dispersion measurement system based on picosecond InGaAsP injection lasers and ultrafast InGaAs p-i-n detectors. This is similar to the measurement system we used for measuring high-bandwidth multimode fibers,⁴ except now in order to study the chromatic dispersion in single-mode fibers by pulse delay measurements, we need to use injection lasers at different wavelengths in the 1.3- μm spectral region. Besides being more compact, this system also has a better time resolution than the fiber Raman laser setup.

II. EXPERIMENTAL APPROACH

The experimental approach is straightforward. Figure 1 shows the setup in which several InGaAsP injection lasers with wavelengths in the 1.2- to 1.5- μm spectral region are used for pulse delay and pulse-broadening measurements in single-mode fibers. A picosecond electrical pulser is used to drive one or two injection lasers at a time to obtain optical pulses of 30 to 80 ps in duration. The laser pulses coming out of single-mode fiber pigtailed are selected by a low-loss, single-mode, fiber-to-fiber switch⁵ and sent through the test fiber, the output of which is detected with a pigtailed InGaAs fast-pin photodiode.⁶ The optical pulsewidth and wavelength-dependent pulse delay information can be stored in the digital oscilloscope for further processing.

The InGaAsP injection lasers used are supplied by Lasertron Co. and have their center wavelengths near 1.21, 1.26, 1.315, 1.335, and 1.525 μm when pulsed to give short optical pulses. The technique of generating short optical pulses is that of gain-switching,^{7,8} which manifests itself as controlled relaxation oscillation⁹ and requires proper adjustment of the pumping level to give a single, short, optical pulse. Either the short electrical pulses from a comb generator⁷ and step-recovery-diode circuit¹⁰ or a high radio frequency (RF) sinusoidal drive¹¹ can be used as the pump, except it has been shown that for low repetition frequencies (a few hundred MHz or less), short electrical pulse pumping results in shorter optical pulses.⁹ The laser spectral width is typically ~ 7 nm owing to multi-longitudinal-mode oscillation in the short-pulse generation.^{7,8}

Figures 2a and b illustrate a typical measurement result obtained with a 6-km-long single-mode test fiber. The laser pulses at 1.335 and 1.315 μm are selected with the single-mode fiber switch and are sent

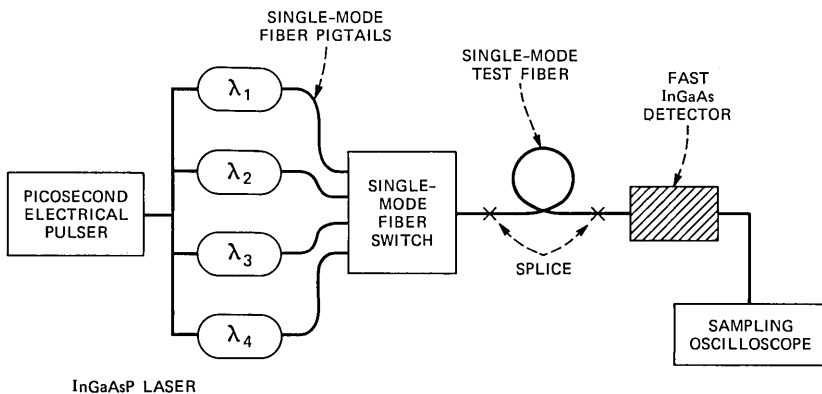


Fig. 1—The schematic of the experimental setup. A 4×1 switch is shown. Five injection lasers are actually used in our preliminary measurements. The fifth laser has a connectorized fiber pigtail for direct connection to the test fiber.

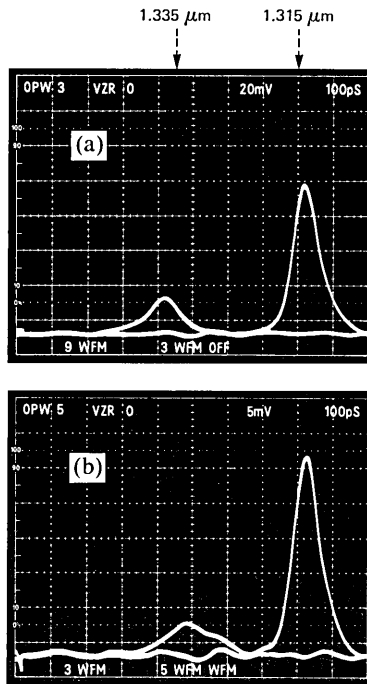


Fig. 2—The pulse pair at 1.335 μm and 1.315 μm (a) before and (b) after the 6-km-long single-mode fiber.

through the test fiber. The pulses before and after the fiber as recorded and stored are shown in Figs. 2a and b, respectively. These are both doubly exposed oscilloscope pictures showing the relative pulse delay change and pulsewidth change owing to dispersion in the test fiber. Note that the pulse at 1.315 μm being near the minimum dispersion wavelength, λ_0 , experiences no broadening, while the pulse at 1.335 μm shows considerable broadening owing to chromatic dispersion in the fiber. The display is adjusted to show the relative delay difference between the two optical pulses before and after the test fiber. The change in the delay difference ($\Delta\tau$) is ~ 60 ps, with the 1.335- μm pulse experiencing more delay because the 1.315- μm pulse lies much closer to the minimum dispersion wavelength, λ_0 . Figures 3a and b show similar results for the pulse pair at 1.21 and 1.315 μm . The 1.21- μm pulse has experienced more delay and broadening than does the 1.315- μm pulse. In time, the 1.21- μm pulse falls behind the 1.315- μm pulse after the 6-km fiber, even though it is ahead of the latter by more than 2 ns. The relative delay change $\Delta\tau$ is 3.1 ns.

Similar wavelength-dependent pulse-delay change and pulse-broadening results are obtained for the pulses at 1.26 and 1.525 μm with respect to 1.315 μm . The experimental results $\Delta\tau(\lambda)$ (in ns/km) are

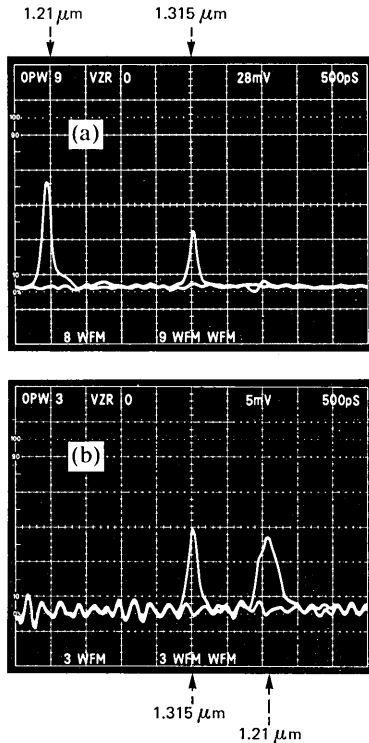


Fig. 3—The pulse pair at 1.21 μm and 1.315 μm (a) before and (b) after the 6-km-long single-mode fiber.

plotted as dots in Fig. 4a. The solid line is a third-order, Chebyshev polynomial fit ($F(\lambda) = a + b\lambda + c\lambda^2 + d\lambda^3$) on an HP-85 computer. The derivative of this fitted polynomial is a second-order polynomial whose root (zero-crossing point) gives λ_0 , the minimum chromatic dispersion wavelength. Figure 4b plots the obtained chromatic dispersion $M(\lambda)$ in ps/nm-km. The obtained λ_0 is $\sim 1.314 \mu\text{m} \pm 0.004 \mu\text{m}$, in reasonable agreement with the 1.309-μm value measured with the fiber Raman laser setup.

III. DISCUSSION AND SUMMARY

Compared with the fiber Raman laser setup, the combination of the picosecond injection lasers and the single-mode fiber switch provides a simple, compact, measurement setup for single-mode fiber chromatic dispersion measurements. In addition, the setup has a better time resolution owing to the short pulse duration and the jitter-free characteristics. While it has a lower dynamic range than the fiber Raman laser, with its 10- to 15-dB dynamic range, typical single-mode

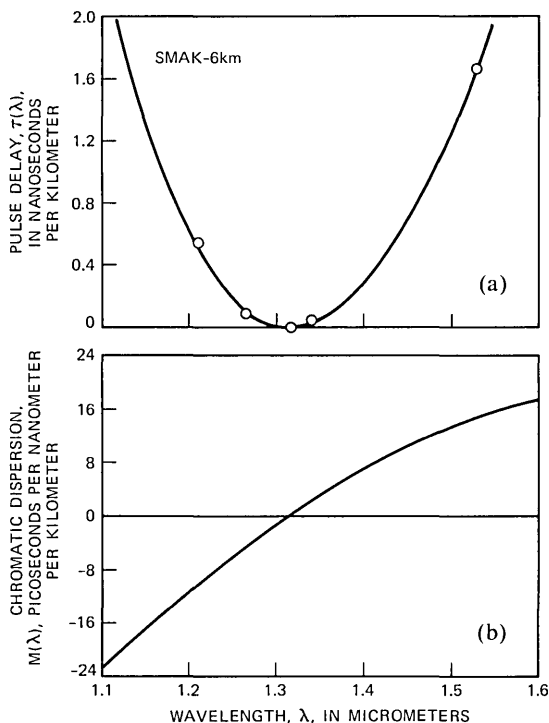


Fig. 4—(a) The time delay vs. wavelength data (dots) with respect to the reference wavelength 1.315 μm . The solid curve is the fitted third-order polynomial. (b) Chromatic dispersion obtained by differentiating the fitted polynomial of (a).

fiber lengths of 5 to 10 km can be measured over the 1.2- to 1.5- μm spectral range, and more than 20 km can be measured over the low-loss region close to 1.3 and 1.55 μm . In practice, the discrete wavelength coverage limits the wavelength resolution. For studying fibers with new or unconventional dispersion characteristics over a wide spectral range, more injection lasers are needed (e.g., 10) for more wavelength coverage. In single-mode fibers for which the design phase has passed and a specific design concerning chromatic dispersion is established, this setup is useful for quality control and field testing by measuring the pulse delay and pulse broadening at a number of selected wavelengths.

IV. ACKNOWLEDGMENTS

We would like to thank R. Knerr and S. Kaufman for supplying the single-mode fiber switch, W. C. Young and L. Curtis for the single-mode fiber connectors, and P. F. Glodis for providing the single-mode fibers.

REFERENCES

1. L. G. Cohen, P. Kaiser, and C. Lin, "Experimental techniques for evaluation of fiber transmission loss and dispersion," *Proc. IEEE*, *68* (October 1980), pp. 1203-9.
2. W. F. Love, "Bandwidth spectral characterization of $G_xO_2-P_2O_5-SiO_2$ multimode optical waveguides," Digest of 6th European Conference on Optical Communication, York, U.K., 1980, pp. 113-15.
3. M. Horiguchi, Y. Ohmori, and H. Takata, "Profile dispersion characteristics in high bandwidth optical fibers," *Appl. Opt.*, *19* (September 1980), pp. 3159-69.
4. C. Lin, P. L. Liu, T. P. Lee, C. A. Burrus, F. T. Stone, and A. J. Ritger, "Measuring high bandwidth fibers in the $1.3 \mu m$ region with picosecond InGaAsP injection lasers and ultra fast InGaAs detectors," *Electron. Lett.*, *17* (June 1981), pp. 438-40.
5. W. C. Young and L. Curtis, "Cascaded multipole switches for single-mode and multimode optical fibers," *Electron. Lett.*, *17* (August 1981), pp. 571-3.
6. T. P. Lee, C. A. Burrus, K. Ogawa, and A. G. Dentai, "Very-high-speed back-illuminated InGaAs/InP pin punch-through photodiodes," *Electron. Lett.*, *17* (June 1981), pp. 431-2.
7. C. Lin, P. L. Liu, T. C. Damen, D. J. Eilenberger, and R. L. Hartman, "Simple picosecond pulse generation scheme for injection lasers," *Electron. Lett.*, *16* (July 1980), pp. 600-2.
8. P. L. Liu, C. Lin, I. P. Kaminow, and J. J. Hsieh, "Picosecond pulse generation from InGaAsP lasers at 1.25 and $1.3 \mu m$ by electrical pulse pumping," *IEEE J. Quantum Electron.*, *QE-17* (May 1981), pp. 671-4.
9. C. Lin, unpublished work.
10. P. T. Torphammer and S. T. Eng, "Picosecond pulse generation in semiconductor lasers using resonance oscillation," *Electron. Lett.*, *16* (July 1980), pp. 587-9.
11. H. Ito, H. Yokoyama, S. Murata, and H. Inaba, "Generation of picosecond optical pulses with highly rf modulated AlGaAs DH laser," *IEEE J. Quantum Electron.*, *QE-17* (May 1981), pp. 663-70.

High-Frequency Impedance of Proton-Bombarded Injection Lasers

By B. W. HAKKI, W. R. HOLBROOK, and C. A. GAW

(Manuscript received August 11, 1982)

Experimental and theoretical results are given of an investigation of the capacitance behavior with frequency of GaAs injection lasers. It is shown that, for shallow-bombarded stripe geometry lasers, the zero-bias capacitance decreases rapidly beyond a certain frequency. This is interpreted in terms of confinement of the low-level radio frequency current under the stripe at high frequencies. Comparison of the experimental results with the analytically derived expressions provides a measure of the material resistivity adjoining the active region. This inferred material resistivity is shown to be in good agreement with results obtained by more direct measurements. Finally, the general conclusions are also applicable to other optoelectronic devices operating at high frequency, such as light-emitting diodes.

I. INTRODUCTION

Injection lasers are inherently suited for high-frequency (>50 MHz) applications.^{1,2} In the Bell System, the commonly used GaAs double-heterostructure (DH) stripe geometry laser is obtained by either shallow- or deep-proton bombardment.^{3,4} For deep bombardment, the active stripe is well defined electrically by the high-resistivity material that confines it. In the case of shallow bombardment, as well as oxide stripe geometry lasers, current can flow laterally, beyond the stripe, through the conductive cladding layer.⁵⁻⁸

Previously, the impedances of several laser geometries were measured at high frequencies, and the results interpreted in terms of phenomenological equivalent circuits.⁹⁻¹³ In this paper, we will restrict the analysis and measurements to zero-biased shallow-bombarded laser diodes, for simplicity. We will show that, for shallow-bombarded stripe geometry lasers, the low-level (radio frequency) RF current at

high frequencies is confined under the stripe. The interpretation of measurements in terms of the fundamentally derived analytical expressions provides a useful measure of relevant material properties. Finally, although the main emphasis is on stripe geometry lasers, the analysis is applicable to other optoelectronic devices. Thus, the approach can be applied to predict the small-signal, unbiased impedance of light-emitting diodes (LEDs) and photodetectors operating at high frequencies.

II. EXPERIMENT

The injection laser is a standard GaAs DH grown by either liquid phase epitaxy (LPE) or molecular beam epitaxy (MBE) technology.^{14,15} It consists of 2- μm -thick N-Ga_{1-x}Al_xAs cladding, 0.1- to 0.2- μm Ga_{1-y}Al_yAs active, 1.5- μm P-Ga_{1-x}Al_xAs cladding, and, finally, 0.5- μm p-GaAs contact layer. Shallow-proton bombardment is used to define stripe widths of either 5- or 10- μm dimension. For these measurements, it is essential to obtain an accurate measure of the stripe width, $2S$, and the distance of separation, d_A , between the active region and the semi-insulating damaged region. This is done by etching the mirror facets in a dilute A/B solution.¹⁶ An example of such an etched mirror obtained by a scanning electron microscope (SEM) is shown in Fig. 1a. Figure 1b is a schematic of the various layers. The distance, d_A , is between the unetched region and the active layer. It is assumed that the etch stops abruptly where the conductivity changes from p-type to semi-insulating. (In general, this is a fair assumption, although it may not be accurate to better than 0.1 μm .)

Every diode is subjected to dc current-voltage and ac impedance tests. The dc measurement determines rectification properties, and current-voltage dependence over the range between 10^{-12} to 10^{-2} amperes. The ac impedance test used RF signal levels, at or below 50 mV, performed on unbiased diodes. At those RF signal levels, the diode conduction current is less than 10^{-11} amperes. The laser impedance is measured at zero bias over a frequency range extending from 10^4 to 5×10^7 Hz. Over this frequency range the capacitive current is typically larger than 10^{-6} amperes. Hence, in the present impedance measurements, the junction capacitive current is well over 10^5 times greater than the diode conduction current. Therefore, for our purposes, the junction can be assumed to be a capacitance with negligible rectified current flow.

Usually, two capacitance bridges are used at different frequency ranges, with an overlap in frequency to ensure accuracy. These impedance bridges measure the equivalent parallel capacitance and conductance of the diode. As an example, Fig. 2 shows the results of two such measurements and two experimental values when results were

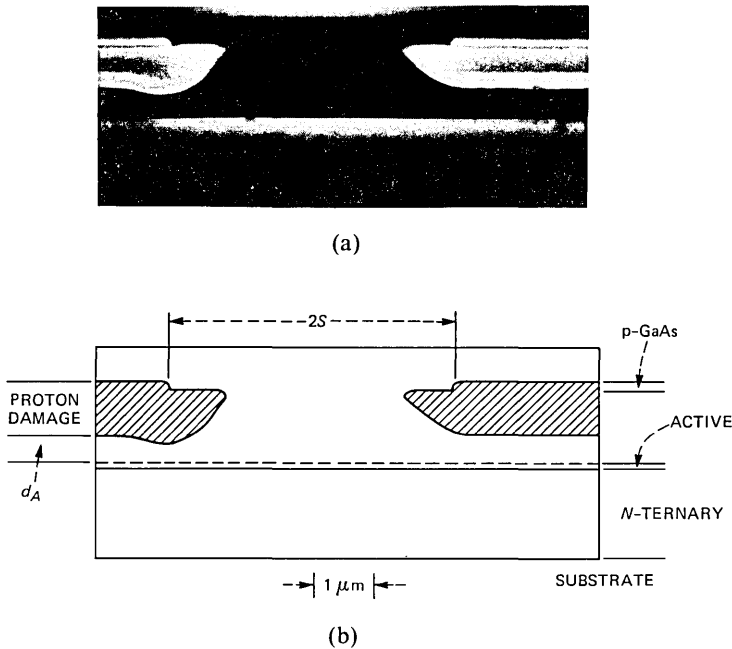


Fig. 1—Mirror of shallow-bombarded stripe geometry DH GaAs layer after etching in dilute A/B etch. (a) Photomicrograph of the etched mirror as obtained by an SEM. In this case, the p-active layer has also been etched. (b) Layer schematic of the same mirror.

obtained on two different instruments. In Fig. 2a, the zero bias capacitance of diode A14 (grown by LPE) is constant as a function of frequency up to 1 MHz. Above 1 MHz, the equivalent parallel circuit capacitance decreases rapidly with an increase in frequency. The same general behavior is observed on another diode from a different slice. The results of diode B22 (grown by MBE) are shown in Fig. 2b. The zero bias capacitance decreases abruptly for frequencies in excess of 10^5 Hz. We also find that, for both diodes, the equivalent parallel circuit conductances G increase with frequency f according to the relation $G \propto f^n$, where n is between 1.25 and 1.5. However, accurate measurements of conductance are more difficult to obtain than those of capacitance.

The results of Fig. 2 show the abrupt drop in capacitance at frequencies that may differ by as much as an order of magnitude. Furthermore, some devices exhibit a more gradual reduction in the capacitance-frequency behavior.¹⁷ Figure 3 shows an example of such a soft roll-off in the measured capacitance. The solid curve is the modified analysis, which accounts for asymmetry and metallization capacitance. The dashed theoretical curve is for the ideal symmetric

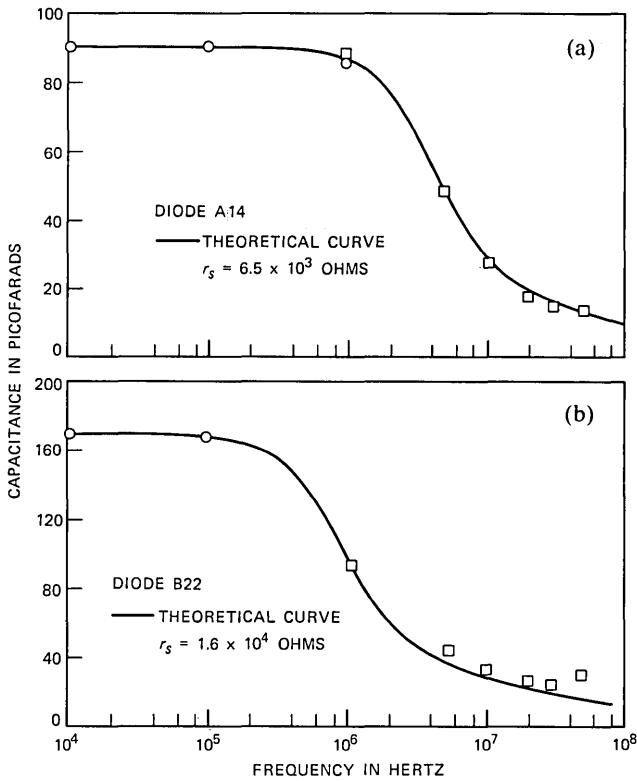


Fig. 2—Capacitance dependence on frequency for (a) diode A14 and (b) diode B22. The solid curves are obtained from analytical expressions. In (a) and (b), d_A is 0.6 and 0.21 μm , respectively.

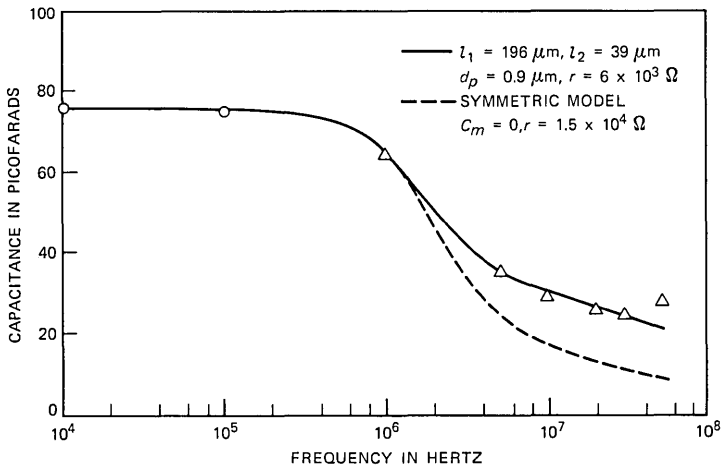


Fig. 3—Capacitance dependence on frequency for a severely eccentric stripe geometry device.

model. The distance between the stripe and the two edges, l_1 and l_2 , respectively, as well as the thickness of the damaged region, d_p , are obtained from SEM measurements. Here, the high-frequency capacitance does not decrease as rapidly as that for the cases shown in Fig. 2.

III. ANALYSIS

3.1 The ideal symmetric case

Consider first the simple and ideal case of a stripe centered within the chip. In addition, assume that the proton-damaged region is sufficiently thick (greater than 2.5 micrometers), so that the only relevant capacitance is the junction capacitance. The RF current flows through the diode as shown schematically in Fig. 4a. There is a direct stripe capacitance, C_s , defined by the width $2S$ and length L of the stripe. The RF current flows through the adjoining cladding layer, which, together with the junction, forms a transmission line. This transmission line is shown schematically in Fig. 4b. It consists of a distributed resistance, r , and a distributed capacitance, c . The analysis of such a transmission line is straightforward. The cladding sheet resistance, r_s , is equal to ρ/d_A , where ρ is the average material resistivity. The capacitance per unit area is c_a given by C_0/A , where C_0 is the low-frequency total diode capacitance and A is the total diode area. In the equivalent distributed parameter circuit shown in Fig. 4b, the resistance per unit length r is equal to r_s/L . Similarly, the capacitance per unit length c is equal to $c_a L$. At any point x , the current and voltage are given, respectively, by

$$i(x, t) = I(x)e^{j\omega t} \quad (1)$$

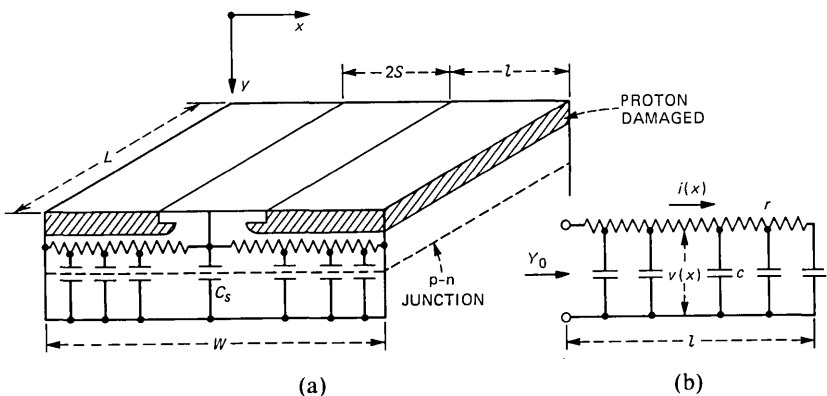


Fig. 4—Schematic of shallow-proton-bombarded, ideally symmetric, stripe geometry laser and equivalent circuit. (a) The various layers and the electrical elements. (b) The distributed parameter transmission line underneath the proton-bombarded region.

and

$$v(x, t) = V(x)e^{j\omega t}. \quad (2)$$

The spatial parts of the current and voltage are related by the equations

$$\frac{dV}{dx} = -rI \quad (3)$$

and

$$\frac{dI}{dx} = -j\omega cV. \quad (4)$$

These are almost standard transmission line equations that have to be solved with the boundary condition $I(\ell) = 0$, where $\ell = (W/2)\text{-}S$. This open-circuit boundary condition is appropriate when surface leakage is negligible. The input admittance, Y_0 , is given by

$$Y_0 = G_0 + jB_0, \quad (5)$$

where

$$G_0 = \frac{\phi \sinh \phi - \sin \phi}{2r\ell \cosh \phi + \cos \phi}, \quad (6a)$$

$$B_0 = \frac{\phi \sinh \phi + \sin \phi}{2r\ell \cosh \phi + \cos \phi} \quad (6b)$$

and

$$\phi = \ell\sqrt{2\omega rc}. \quad (6c)$$

From the susceptance given in (6b), an equivalent distributed parallel capacitance, C_d , is obtained in the form

$$\frac{C_d}{C'_0} = \frac{1 \sinh \phi + \sin \phi}{\phi \cosh \phi + \cos \phi}, \quad (7)$$

where $C'_0 = c_a\ell L$ is the low-frequency capacitance of the junction whose area is ℓL . A plot of C_d/C'_0 is shown in Fig. 5 as a function of ϕ . It is seen that the equivalent distributed parallel capacitor value remains equal to the low-frequency value as long as $\phi < 1$. However, for $\phi > 1$, the capacitance decreases at the rate of $1/\phi$. This decrease at high frequency is due to the fact that as the frequency increases, a relatively larger fraction of the reactive current is shunted by the distributed capacitance at small distances. At sufficiently high frequencies most of the current is bypassed at small distance by the capacitance.

The choice of a parallel circuit, unfortunately, does not provide a straightforward connection between conductance and the material

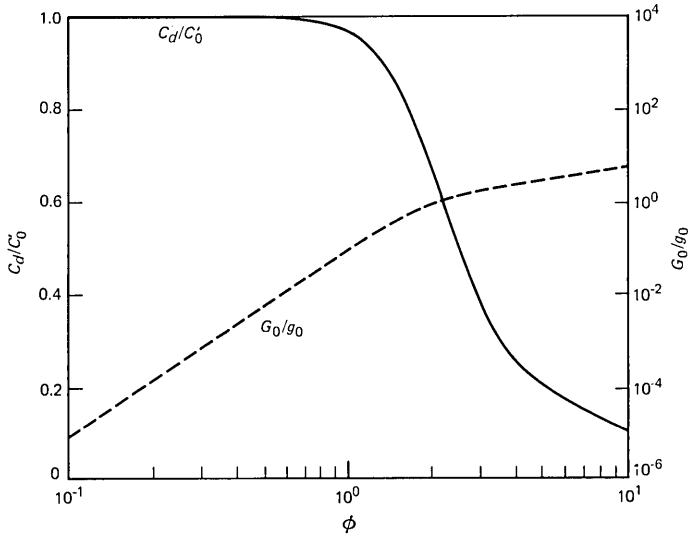


Fig. 5—Equivalent parallel circuit capacitance and conductance for the distributed transmission line of Fig. 4b. The variation is in terms of the normalized parameter, ϕ , defined in the text.

resistivity. Nevertheless, let $g_0 = Ld_A/\rho\ell$, which is the dc conductance between $x = 0$ and ℓ . Then, G_0/g_0 is plotted as a function of ϕ in Fig. 5. It is seen that the normalized conductance increases as ϕ^4 . Since ϕ varies as $f^{1/2}$, it follows that the conductance increases as f^2 , which is somewhat faster than the experimentally observed variation.

The above results can be applied to the ideally symmetric stripe geometry configuration of Fig. 4a. The total equivalent parallel capacitor of the diode C_p , comprising the direct stripe and distributed capacitances, respectively, is given by

$$C_p = C_0 \left[\frac{A_s}{A} + \left(1 - \frac{A_s}{A} \right) \frac{\sinh \phi + \sin \phi}{\phi(\cosh \phi + \cos \phi)} \right], \quad (8)$$

where $A_s = 2SL$ is the area of the stripe, C_0 is the total capacitance of the diode at low frequency, and the rest of the parameters are as defined previously.

Figure 2a shows a plot of eq. (8), with a sheet resistance chosen at a value of 6.5×10^3 ohms. The stripe width is $5 \mu\text{m}$, the diode width is $250 \mu\text{m}$, and the length is $380 \mu\text{m}$. The stripe width is obtained by etching the mirror, as in Fig. 1. The only adjustable parameter in the curve of Fig. 2a is the sheet resistance. A similar theoretical curve is plotted in Fig. 2b for diode B22. Here, the stripe width is $10 \mu\text{m}$, and the sheet resistance is 1.6×10^4 ohms. It is seen that the capacitance in Fig. 2b starts to decrease at a lower frequency because of the higher

sheet resistance. The high sheet resistance in diode B22 is due to the smaller distance of separation between the damaged region and the active layer.

3.2 Asymmetric stripe with metallization capacitance

Two deviations from the earlier symmetric model are observed in some devices. First, the stripe can be eccentric. Second, the proton-damaged layer can be as thin as 0.5 micrometers. These two factors cause the high-frequency behavior to deviate from the ideal model. This has been noted in independent measurements by C. W. Thompson et al.¹⁷

The metallization can contribute a significant additional capacitance if the metallization is separated from the conducting semiconductor by a thin insulating layer. Such a condition can exist in oxide stripe geometry lasers as well as in some proton-bombarded lasers. However, this capacitance has unique characteristics, since it is not operative at low frequencies, where the electrodes are virtually connected. But at high frequencies, charge storage between the metal and the semiconductor can take place. A schematic of the various layers is shown in Fig. 6a. The proton-damaged layer thickness is d_p . This semi-insulating layer can, at high frequencies, contribute a metallization capacitance C_m given approximately for GaAs by

$$C_m \approx 10/d_p \quad pF, \quad (9)$$

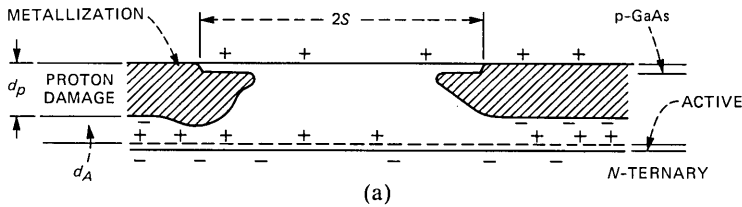


Fig. 6—Schematic of stripe geometry with thin damaged layer. (a) Charge buildup. (b) Equivalent transmission line circuit. The metallization capacitance results from charge storage across the proton-bombarded region.

where d_p is measured in micrometers, and the chip is $250 \times 380 \mu\text{m}$, which is typical for our diodes. For oxide-defined stripe geometry lasers, the metallization capacitance is associated with the oxide layer itself.

The metallization capacitance is not effective at low frequencies. This is obvious from Fig. 6a, where conduction through the stripe region bypasses the charge across the damaged region. The equivalent transmission line circuit is shown schematically in Fig. 6b.¹⁸ The circuit parameters are as defined previously, and solution of the transmission line equations follows the usual procedures. The total current branches into two components, I_0 and I'_0 , respectively. The two admittances are given by

$$G'_0 = \frac{\omega c c' \ell}{(c + c')} \left[\frac{\sinh \phi - \sin \phi}{\phi (\cosh \phi + \cos \phi)} \right] \quad (10a)$$

$$G_0 = \omega c \ell \left[\frac{\sinh \phi - \sin \phi}{\phi (\cosh \phi + \cos \phi)} \right] \quad (10b)$$

$$B'_0 = \frac{\omega c c' \ell}{(c + c')} \left[1 - \frac{\sinh \phi + \sin \phi}{\phi (\cosh \phi + \cos \phi)} \right] \quad (10c)$$

$$B_0 = \omega c \ell \frac{(\sinh \phi + \sin \phi)}{\phi (\cosh \phi + \cos \phi)}, \quad (10d)$$

where

$$\begin{aligned} \phi &= \ell \sqrt{2\omega r(c + c')}, \\ c' &= C_m/W, \end{aligned} \quad (10e)$$

and W is the total width of the chip.

Since the two admittances are in parallel, the equivalent parallel circuit capacitance is

$$C = \frac{C_0}{C_0 + C_m} \left[C_0 \frac{\sinh \phi + \sin \phi}{\phi (\cosh \phi + \cos \phi)} + C_m \right], \quad (11)$$

where C_0 is the low-frequency junction capacitance, and C_m is the metallization capacitance, as given in (9). It is clear from (11) that at low frequency, the ϕ function goes to one, and the capacitance becomes equal to the low-frequency junction capacitance. On the other hand, at high frequency, the ϕ function goes to zero, and the capacitance becomes equal to the metallization and the junction capacitances in series.

When the stripe is located off-center, the capacitance can be calculated in a straightforward manner. The total chip capacitance becomes

$$C_T = C_0 \left\{ \frac{A_s}{A} + \left(\frac{1 - A_s/A}{C_0 + C_m} \right) \left[\frac{C_0}{(\ell_1 + \ell_2)} (\ell_1 F_1 + \ell_2 F_2) + C_m \right] \right\}, \quad (12)$$

where

$$F_i = \frac{\sinh \phi_i + \sin \phi_i}{\phi_i(\cosh \phi_i + \cos \phi_i)}, \quad (13)$$

$$\phi_i = \ell_i \sqrt{2\omega r(c + c')}, \quad (14)$$

$$\ell_1 + \ell_2 + 2S = W, \quad (15)$$

ℓ_1 and ℓ_2 are the lengths of the two distributed line sections, respectively, $2S$ is the stripe width, A_s is the total area of the stripe, and A is the area of the diode.

An example of a case of extreme asymmetry is shown in Fig. 3. Measurement of eccentricity indicates that the stripe is displaced by $\approx 70 \mu\text{m}$ from its center position. The measured thickness of the proton-damaged region is $0.9 \mu\text{m}$, which results in a metallization capacitance of $11 pF$. When the expression for the asymmetric case, as given in (12), is applied to the results of Fig. 3, the agreement between theory and experiment is adequate for a sheet resistance of $6 \times 10^3 \Omega$. On the other hand, the symmetric model, shown as a dashed curve in Fig. 3, cannot fit the data over the extended frequency range. In addition, the symmetric model overestimates the sheet resistance.

3.3 Material resistivity

Having obtained a value of sheet resistance and layer thickness, d_A , the material resistivity can be derived. However, care should be exercised in accounting for proton damage spread into the conductive layer.

If $y = 0$ is the point at the edge of the damaged region where the local carrier concentration is (mathematically) zero, then for $y > 0$, the carrier concentration does not recover abruptly as a step function. Instead, the carrier concentration actually recovers as an error function of distance.¹⁹ For our purposes, this recovery can be represented approximately by the relation

$$p(y) = p_0(1 - e^{-y/y_0}), \quad (16)$$

where p_0 is the background-free hole concentration, and y_0 determines the rate at which carrier removal decreases with distance.

For current flow parallel to the junction plane, the average resistivity ρ over a thickness d_A is

$$\rho = \rho_0 / \left\{ 1 - \frac{y_0}{d_A} \left[1 - \exp(-d_A/y_0) \right] \right\}, \quad (17)$$

where ρ_0 is the resistivity of the undamaged material. From published data of proton-bombarded GaAs, it would seem that the value of y_0 is $0.13 \mp .03 \mu\text{m}$.¹⁹⁻²¹

The above equations can now be applied to the results of Fig. 2. For diode A14 we obtain $\rho = 0.39 \Omega\text{-cm}$; and $\rho_0 = 0.3 \Omega\text{-cm}$. For diode B22, we obtain $\rho = 0.34 \Omega\text{-cm}$, and $\rho_0 = 0.17 \Omega\text{-cm}$. Several diodes from these slices were measured. The results were averaged and given in Table I. The average resistivity in the cladding, adjacent to the active region, in slice *A* is $0.28 \mp 0.09 \Omega\text{-cm}$. Direct measurement of resistivity in this material by J. L. Zilko gives a value of $0.21 \Omega\text{-cm}$.²² Hence, to within statistical diode-to-diode variation, the agreement is adequate. The resistivity of the cladding in slice *B*, given in Table I, is $0.18 \mp 0.02 \Omega\text{-cm}$. These resistivities can also be used to estimate the free-hole concentration. For the typical 40-percent aluminum in the ternary, the hole mobility should be about $100 \text{ cm}^2/V\text{s}$.^{23,24} The deduced-hole concentrations for slices *A* and *B* are listed in Table I. To check these deduced values with direct-hole concentration measurements, a third slice *C* was evaluated. The hole concentration, in slice *C* in the cladding, as deduced from the capacitance-frequency measurement, was $(2.1 \mp 0.6) \times 10^{17} \text{ cm}^{-3}$. On the other hand, hole-concentration measurements on this slice, by the automatic feedback method,²¹ gave values of $(1.9 \mp 0.4) \times 10^{17} \text{ cm}^{-3}$.²⁵ Hence, again the capacitance-frequency method gave values consistent with those obtained by other methods.

Stripe geometry lasers with deep-proton bombardment were also evaluated. In these devices the proton damage extends beyond the active region. Here, as opposed to the results for shallow bombardment, the capacitance remains small and nearly invariant with frequency. This again lends support to the model of RF current confinement in shallow-bombarded stripe geometry lasers.

IV. CONCLUSION

Experiments conducted on shallow proton-bombarded stripe geometry lasers show that the capacitance decreases abruptly above a certain frequency. These results are interpreted in terms of a distributed parameter transmission line model. It is shown that the frequency beyond which the capacitance decreases depends on: the sheet resistance of the material above the active region, capacitance per unit area,

Table I—Material evaluation obtained from capacitance frequency measurements

Slice	Slice Average				
	r_s ohms	ρ ohm-cm	d_A μm	ρ_0 ohm-cm	$p_0 \text{ cm}^{-3}$
A	$(6 \mp 2) \times 10^3$	0.37 ∓ 0.12	0.6	0.28 ∓ 0.09	$(2.5 \mp 0.9) \times 10^{17}$
B	$(1.5 \mp 0.1) \times 10^4$	0.36 ∓ 0.03	0.22 ∓ 0.02	0.18 ∓ 0.02	$(3.4 \mp 0.4) \times 10^{17}$

and the dimension of the diode. These analytical expressions, when applied to experimental results, allow one to obtain an estimate of the material resistivity adjoining the active region. The deduced values of resistivity and doping concentration are in good agreement with the values obtained by other methods.

Although the analytical results are applied to laser structures, they are also applicable to other optoelectronic devices. For instance, the same general conclusions can also be drawn about the small signal impedance of LEDs operating at high frequency. For configurations other than the simple stripe geometry, the spatial distribution of RF current obviously changes. However, the qualitative results are the same.

V. ACKNOWLEDGMENT

The authors are indebted to A. E. Bakanowski, H. E. Elder, E. I. Gordon, and C. W. Thompson, Jr., for useful discussions, and to A. Y. Cho and M. C. Tamargo for providing the laser material. The authors are also grateful to J. L. Zilko for providing independent resistivity measurements, and to C. G. Bergey and E. W. Bonato for detailed measurements.

REFERENCES

1. S. E. Miller, E. A. J. Marcatili, and Tingye Li, "Research Toward Optical-Fiber Transmission Systems—Part I," *Proc. IEEE*, *61* (December 1973), pp. 1703-26.
2. S. E. Miller, Tingye Li, and E. A. J. Marcatili, "Research Toward Optical-Fiber Transmission Systems—Part II," *Proc. IEEE*, *61* (December 1973), pp. 1726-51.
3. J. C. Dymont, L. A. D'Asaro, J. C. North, B. I. Miller, and J. E. Ripper, "Proton-Bombardment Formation of Stripe Geometry Heterostructure Lasers for 300°K CW Operation," *Proc. IEEE*, *60* (June 1972), pp. 726-28.
4. B. W. Hakki, "Carrier and Gain Spatial Profiles in GaAs Stripe Geometry Lasers," *J. Appl. Phys.*, *44* (November 1973), pp. 5021-28.
5. B. W. Hakki, "GaAs Double Heterostructure Lasing Behavior Along the Junction Plane," *J. Appl. Phys.*, *46* (January 1975), pp. 292-302.
6. W. B. Joyce, "Current Crowding and Carrier Confinement in Double Heterostructure Lasers," *J. Appl. Phys.*, *51* (May 1980), pp. 2394-2401.
7. H. Yonezu, I. Sakuma, K. Kobayashi, T. Kamejima, M. Unno, and Y. Nannichi, "A GaAs/Al_xGa_{1-x}As Double Heterostructure Planar Laser," *Jpn. J. Appl. Phys.*, *12* (October 1973), pp. 1585-92.
8. W. P. Dumke, "Current Thresholds in Stripe-Contact Injection Lasers," *Solid State Electron.*, *16* (November 1973), pp. 1279-98.
9. F. K. Reinhart, "Reverse-Biased Gallium Phosphide Diodes as High-Frequency Light Modulators," *J. Appl. Phys.*, *39* (June 1968), pp. 3426-34.
10. K. Aiki, M. Nakamura, T. Kuroda, K. Umeda, R. Ito, N. Chinone, and M. Maeda, "Transverse Mode Stabilized (AlGa)As Injection Lasers with Channeled-Substrate-Planar Structure," *IEEE J. Quantum Electron.*, *14* (February 1978), pp. 89-94.
11. K. Furuya, Y. Suematsu, and T. Hong, "Reduction of Resonance-Like Peak in Direct Modulation Due to Carrier Diffusion in Injection Laser," *Appl. Optics*, *17* (June 1978), pp. 1949-52.
12. H. Kuwahara, Y. Daido, and H. Furuta, "Measurements of Impedance of DH Semiconductor Lasers," *Proc. IEEE*, *65* (September 1977), pp. 1412-13.
13. J. M. Dumant, Y. Guillausseau, and M. Monerie, "Small Signal Modulation of DH Laser Diodes: Effect of the Junction Capacitance," *Optics Commun.*, *33* (February 1980), pp. 188-92.

14. L. R. Dawson, "Near-Equilibrium LPE Growth of GaAs-Ga_{1-x}Al_xAs Double Heterostructures," *J. Cryst. Growth*, **27** (December 1974), pp. 86-96.
15. A. Y. Cho and H. C. Casey, Jr., "GaAs-Al_xGa_{1-x}As Double-Heterostructure Lasers Prepared by Molecular Beam Epitaxy," *Appl. Phys. Letters*, **25** (September 1974), pp. 288-90.
16. J. C. Campbell, S. M. Abott, and A. G. Dentai, "A Comparison of 'Normal' Lasers and Lasers Exhibiting Light Jumps," *J. Appl. Phys.*, **51** (August 1980), pp. 4010-13.
17. C. W. Thompson, Jr., H. E. Elder, and A. E. Bakanowski, private communication.
18. A. E. Bakanowski, private communication.
19. H. Matsumura, and K. Stephens, "Electrical Measurement of the Lateral Spread of the Proton Isolation Layer in GaAs," *J. Appl. Phys.*, **48** (July 1977), pp. 2779-83.
20. B. R. Pruniaux, J. C. North, and G. L. Miller, "Compensation of n-Type GaAs by Proton Bombardment," *Proc. 2nd Int. Conf. on Ion Implantation*, edited by I. Ruge and J. Graul, New York: Springer-Verlag, 1971, pp. 212-27.
21. G. L. Miller, "A Feedback Method for Investigating Carrier Distributions in Semiconductors," *IEEE Trans. Electron. Dev.*, *ED-19* (October 1972), pp. 1103-8.
22. J. L. Zilko, and P. J. Anthony, "Conductivity Profiling of GaAs/GaAlAs Multilayer Structures," *J. Electrochem. Soc.*, **128** (April 1981), pp. 871-74.
23. S. Zukotynski, S. Sumski, M. B. Panish, and H. C. Casey, Jr., "Electrical Properties of Ge-Doped p-type Al_xGa_{1-x}As," *J. Appl. Phys.*, **50** (September 1979), pp. 5795-99.
24. A. J. SpringThorpe, F. D. King, and A. Becke, "Te and Ge-Doping Studies in Ga_{1-x}Al_xAs," *J. Electron. Materials*, **4** (February 1975), pp. 101-18.
25. B. W. Hakki, "p-GaAs/P-Ga_{1-x}Al_xAs Isotype Heterojunctions in Doubleheterostructure Laser Material," *J. Appl. Phys.*, **52** (October 1981), pp. 6054-58.

An Analysis of the Derivative Weight-Gain Signal From Measured Crystal Shape: Implications for Diameter Control of GaAs

By A. S. JORDAN, R. CARUSO, and A. R. VON NEIDA

(Manuscript received May 11, 1982)

A commercially viable GaAs device technology for field-effect transistors, integrated circuits, and lasers is critically dependent on the availability of high-quality, single-crystal boules with controlled diameter. We have modeled a diameter control scheme based on the monitoring of crystal weight for liquid-encapsulated Czochralski (LEC) growth of GaAs. The presence of the B_2O_3 liquid encapsulant and significant capillary forces make the direct interpretation of the weight-gain signal and its time derivative (DWGS) more complicated in comparison with pulled materials such as oxide crystals. We have formulated a realistic model for the LEC growth of axisymmetric crystals and have derived the differential equation relating the time evolution of the DWGS to radius and length. We show that the magnitude of the DWGS at the crystal's "shoulder" is inversely related to the radius of curvature. Furthermore, the meniscus by itself gives rise to a precursor or early warning in the signal, which means that the maximum in DWGS precedes the maximum in shape by a few hundred seconds. The existence of a secondary maximum or aftershock in the signal that is the sole consequence of liquid encapsulation is also demonstrated. Excellent agreement has been obtained between DWGS and the signal predicted from the measured shape of a grown crystal. Thus, prospects for automatic diameter control are encouraging.

I. INTRODUCTION

GaAs is one of the key semiconductor materials that serves as a substrate for light-emitting diodes (LEDs), lasers, and field-effect transistors (FETs). Paralleling the development pattern of growth techniques for other single crystals, the issues of primary interest have

evolved from questions of quality (elimination of twinning and defects, doping uniformity) to that of economy of size. However, scaling up the dimensions of GaAs crystals grown by the liquid-encapsulated Czochralski (LEC) technique necessitates introducing sophisticated schemes for diameter control. By achieving satisfactory control, it will also be possible to run for extended periods of time with minimal supervision, and to attain higher crystal yields, as well as a reduction in defect generation brought on by shape change.

Since diameter control in Czochralski growth has been the subject of an excellent recent review by Hurle,¹ here a brief outline of previous efforts with respect to LEC growth will suffice. Unlike Si² and a wide variety of oxide crystals³ (e.g., GGG, LiTaO₃) for which successful diameter measurement and control systems have been developed, the realization of a viable system for the LEC growth of III-V compounds has been much more difficult to achieve. This is largely due to effects associated with the growth chamber under high pressure, the presence of the encapsulating layer of B₂O₃(ℓ), and the phenomenon of anomalous density ($d_{\text{liquid}} > d_{\text{solid}}$) in some semiconductor materials.

A number of approaches to the control problem have been taken. Pruett and Lien,⁴ and van Dijk et al.⁵ have employed an X-ray beam passing through the high-pressure growth apparatus for GaP, designed to make the melt a high absorber of radiation relative to all other materials in the radiative path of the system. The subsequent use of an image intensifier gave an accurate television picture of the growth interface. Small changes in diameter could then be electronically extracted from the video signal and utilized for its control.⁵

Alternatively, considerable success has been achieved meeting the demand for large, closely controlled diameter GaP single crystals using a floating die technique. Cole et al.⁶ employed a Si₃N₄ ring floating in the GaP melt beneath the B₂O₃(ℓ). This modified form of Stepanov ring^{6,7} creates a long-term stable growth regime, permitting diameter tolerances of ± 1 mm for a 50-mm diameter boule. The technique has also been applied to GaAs with different degrees of success, apparently depending on the crystallographic growth direction.⁸

In a third technique, advocated by Bardsley et al.,⁹ a weight signal or a derivative weight-gain signal obtained by detecting the apparent weight change of either the crucible or crystal during Czochralski growth is used to control the power output and consequently the diameter. Bardsley et al. have considered both the theory⁹ and its implementation by an analogue servo-system.¹⁰ Although these authors have proposed reasonable initial postulates and expressions to calculate the weight gain, the detailed analysis is limited only to small perturbations in diameter and the formalism precludes growth by the LEC technique.

For some time now we have routinely measured the apparent weight of LEC-pulled GaAs crystals with a high sensitivity load cell placed in the high-pressure chamber in series with the crystal pull rod. The output of the cell is recorded as a dc level. Then, the derivative of the signal is taken electronically with a device developed especially for that purpose, and is also recorded. This activity has served as a useful qualitative guide in the manual control of diameter. The major objective of this paper is to develop the fundamental theory that governs the relationship between the shape of an axisymmetric crystal with arbitrary variations in its diameter (including the shoulder) grown by the LEC technique and the instantaneously detected derivative weight-gain signal. Besides gaining new physical insight, we expect that by means of these investigations we can focus on the important physical factors and analytical techniques essential to the eventual control of crystal diameter.

As a first step, a tractable model is formulated for LEC growth with a meniscus. The treatment leads to an explicit expression for the derivative weight-gain signal exclusively in terms of the crystal cross section and its first and second derivatives, in addition to geometrical and material parameters. Next, the numerical techniques required to perform computer simulations are outlined. Then, the key features of the signal are investigated by using the probability density function of the lognormal distribution for the crystal contour. Among the effects considered in detail are the shape of the shoulder and the presence or absence of the $B_2O_3(\mathcal{L})$ encapsulant and meniscus. Furthermore, the derivative weight-gain signal is predicted for four contour lines—generated by consecutive 90° axial rotations—of a specially prepared GaAs boule and compared with the signal measured during growth. Finally, in light of these modeling calculations, the prospective techniques for diameter control are discussed.

II. THEORY

In this section, by a consideration of the relevant features of the LEC growth process, a realistic model is developed for the derivative weight-gain signal in a form suitable for computer analysis. To make the calculations tractable, but without a serious loss of generality, the following set of assumptions and constraints has been introduced:

(i) The crystal and crucible are axisymmetric with circular cross sections. Hence, the crystal's radius, r , can always be prescribed as a function of the axial location, L , which is monotonically related to the time elapsed since seeding.

(ii) The crucible containing the melt and the encapsulant is a right circular cylinder of radius, R . Less restrictively, departures from

a right cylinder are permitted below the liquid level prevailing at the completion of growth.

(iii) The effects of crucible and crystal rotation and the concomitant viscous drag are neglected.

(iv) As the initial melt level falls by distance ℓ (Fig. 1), the entire amount of the liquid is transferred to the solid.

(v) Seepage from the $B_2O_3(\ell)$ encapsulating column of mass, m_B , and volume, V_B , to the melt/crucible interface (walls) is minute.

(vi) The solid-liquid interface is planar. This allows the separation of the associated heat transfer problem from the calculations.

(vii) Capillary forces between the melt, $B_2O_3(\ell)$, and solid result in a meniscus of height, h , with a subtended angle between the vertical

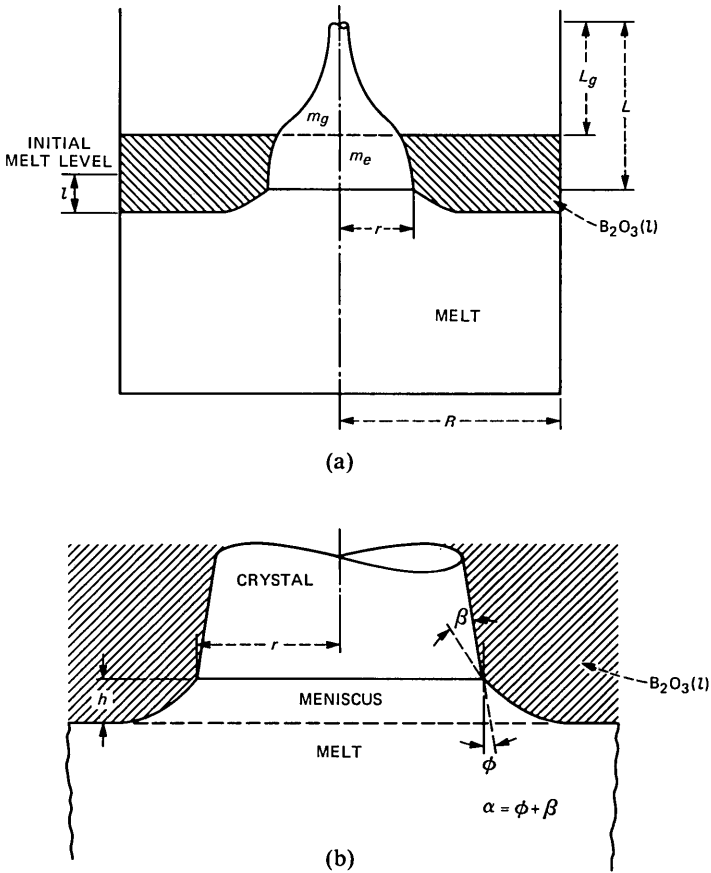


Fig. 1—(a) Schematic geometry of LEC pulling. (b) Enlarged view of the solid-liquid interface region.

and the tangent to the "skirt" of the meniscus (joining angle), α (Fig. 1).

(viii) As the crystal is pulled at constant rate p , contact with the melt is uninterrupted.

(ix) Physical constants required in the calculations are average values representative near the melting point of the crystal.

The preceding conditions in combination with the geometrical description of LEC pulling in Fig. 1 and Archimedes' principle permit writing the weight signal, W_s , corresponding to detected crystal mass, m_s , as the sum

$$W_s = m_s g = \begin{array}{cccccc} W_1 & -W_2 & +W_3 & +W_4 & -W_5 & \\ \text{true} & \text{buoyancy} & \text{vertical} & \text{weight of} & \text{buoyancy} & \\ \text{weight} & \text{correction} & \text{projection} & \text{cylindrical} & \text{correction to} & \\ & \text{to } W_1 & \text{of capillary} & \text{meniscus} & W_3 + W_4 & \\ & & \text{force} & & & \end{array} \quad (1)$$

where

$$W_1 = (m_e + m_g)g = mg \quad (2a)$$

$$W_2 = V_e \rho_B g \quad (2b)$$

$$W_3 = 2\pi r \sigma \cos \alpha \quad (2c)$$

$$W_4 = r^2 \pi h \rho_\ell g \quad (2d)$$

$$W_5 = V_{\text{men}} \rho_B g \quad (2e)$$

The symbols m , V , and ρ without subscripts designate the mass, volume, and density of the growing crystal, respectively. The same letters with subscripts B , ℓ , e , and g refer to the B_2O_3 encapsulant, melt, and that portion of the crystal that resides either in the encapsulant or the gas phase, respectively. The symbols σ and V_{men} represent the surface tension and meniscus volume, respectively.

Since $V_e = (m - m_g)/\rho$, with the abbreviation $k_1 = \rho_B/\rho$, $W_1 - W_2$ in eqs. (2a) and (2b) can be transformed into

$$W_1 - W_2 = m(1 - k_1)g + k_1 m_g g. \quad (3)$$

Then, differentiating eqs. (3), (2c), (2d), and 2e with respect to time, t , and using eq. (1), we obtain the derivative weight-gain signal (DWGS) in the form

$$\begin{aligned} \frac{dW_s/g}{dt} = & (1 - k_1) \frac{dm}{dt} + k_1 \frac{dm_g}{dt} + \frac{2\pi}{g} \sigma \cos \alpha \frac{dr}{dt} \\ & - \frac{2\pi\sigma}{g} r \sin \alpha \frac{d\alpha}{dt} + \rho_\ell \left(2r\pi h \frac{dr}{dt} + r^2 \pi \frac{dh}{dt} \right) - \rho_B \frac{dV_{\text{men}}}{dt}. \quad (4) \end{aligned}$$

It can be readily shown that dm/dt is a function of the other

variables. Clearly, if the crystal is axisymmetric, one has

$$\frac{dm}{dt} = \rho\pi r^2 \frac{dL}{dt}. \quad (5)$$

Moreover, the continuity of growth provides the relation

$$L = \ell + pt - h, \quad (6)$$

or in derivative form

$$\frac{dL}{dt} = \frac{d\ell}{dt} + p - \frac{dh}{dt}. \quad (7)$$

In addition, the conservation of matter during crystallization as the liquid level falls leads to

$$m = \rho_\ell(\pi R^2 \ell - V_{\text{men}}) \quad (8)$$

or

$$\frac{dm}{dt} = \rho_\ell \left(\pi R^2 \frac{d\ell}{dt} - \frac{dV_{\text{men}}}{dt} \right). \quad (9)$$

A combination of eqs. (5), (7), and (9) permits the elimination of $d\ell/dt$; hence eq. (5) can be rewritten as

$$\frac{dm}{dt} = \rho\pi r^2 \frac{\left(\frac{k dV_{\text{men}}}{\pi dt} + pr^2 - r^2 \frac{dh}{dt} \right)}{\bar{R}^2 + p - \frac{dh}{dt}}, \quad (10)$$

where $k = \rho_\ell/\rho$ and $\bar{R}^2 = kR^2 - r^2$.

Substituting eq. (10) into eq. (4) for the DWGS yields after several algebraic operations

$$\begin{aligned} \frac{dW_s/g}{dt} = & k_1 \frac{dm_g}{dt} + \frac{p(1-k_1)\pi\rho_\ell r^2 R^2}{\bar{R}^2} - \frac{2\pi\sigma}{g} r \sin \alpha \frac{d\alpha}{dt} \\ & + \frac{dV_{\text{men}}}{dt} \left[(1-k_1)\rho_\ell \frac{r^2}{\bar{R}^2} - \rho_B \right] + r^2 \pi \frac{dh}{dt} \rho_\ell \left[1 - (1-k_1) \frac{R^2}{\bar{R}^2} \right] \\ & + 2\pi \frac{dr}{dt} \left[rh\rho_\ell + \frac{\sigma}{g} \cos \alpha \right]. \quad (11) \end{aligned}$$

In essence, the general form of the DWGS is given by eq. (11). Since all the variables directly depend on L or indirectly through r on L , it is convenient to replace the derivatives d/dt by $d/dL \cdot dL/dt$. Accordingly, we have

$$\begin{aligned} \frac{dW_s/g}{dt} = & k_1 \frac{dm_g}{dL} \frac{dL}{dt} + \frac{p(1-k_1)\pi\rho_\ell r^2 R^2}{\bar{R}^2} - \frac{2\pi\sigma}{g} r \sin \alpha \frac{d\alpha}{dL} \frac{dL}{dt} \\ & + \frac{dV_{\text{men}}}{dL} \frac{dL}{dt} \left[(1-k_1)\rho_\ell \frac{r^2}{\bar{R}^2} - \rho_B \right] + r^2\pi \frac{dh}{dL} \frac{dL}{dt} \rho_\ell \left[1 - (1-k_1) \frac{R^2}{\bar{R}^2} \right] \\ & + 2\pi \frac{dr}{dL} \frac{dL}{dt} \left(rh\rho_\ell + \frac{\sigma}{g} \cos \alpha \right). \quad (12) \end{aligned}$$

Equation (12) completely describes the DWGS once the few remaining unknown functions are evaluated. These are the axial derivative of the crystal mass outside the $\text{B}_2\text{O}_3(\ell)$ (dm_g/dL), the macroscopic growth rate (dL/dt), and the meniscus (h , dh/dL , V_{men} , and dV_{men}/dL) in terms of $r(L)$ and the required physical and geometrical constants. The crystal shape $r(L)$ we consider here as an input signal and the choice of different functional forms will be postponed to the next section.

2.1 Evaluation of dm_g/dL

From Fig. 1 we conclude that the total cylindrical volume occupied by the segment of the growing crystal in V_e , the encapsulant, and the meniscus are conserved. Thus we can write the integral relation

$$\int_0^L r^2\pi dL - \int_0^{L_g} r^2\pi dL + V_{\text{men}} + V_B = \pi R^2(L - L_g + h), \quad (13)$$

where the two integrals on the left-hand side are equal to

$$\int_{L_g}^L r^2(L)\pi dL.$$

Thus we can express m_g/ρ as

$$m_g/\rho = \int_0^{L_g} r^2\pi dL = \int_0^L r^2\pi dL + V_{\text{men}} + V_B - \pi R^2(L - L_g + h).$$

Differentiating the above equation with respect to L gives

$$\frac{dm_g/\rho}{dL} = r^2\pi + \frac{dV_{\text{men}}}{dL} - \pi R^2 \left(1 + \frac{dh}{dL} - \frac{dL_g}{dL} \right). \quad (14)$$

To obtain dL_g/dL we define a function $\phi(L, L_g) = 0$ by subtracting the right-hand side from both sides of eq. (13). Then, employing the well-known result for the differentiation of implicit functions we have

$$\frac{dL_g}{dL} = - \frac{\partial\phi/\partial L}{\partial\phi/\partial L_g}, \quad (15)$$

where

$$\frac{\partial\phi}{\partial L} = r^2(L)\pi + \frac{dV_{\text{men}}}{dL} - \pi R^2 \left(1 + \frac{dh}{dL} \right)$$

and

$$\frac{\partial\phi}{\partial L_g} = -r^2(L_g)\pi + \pi R^2.$$

The actual value of L_g can only be determined by a numerical procedure. From eq. (13) we get

$$\frac{\int_0^L r^2\pi dL + \frac{m_B}{\rho_B} + V_{\text{men}}}{\pi R^2} - (L + h) = \frac{\int_0^{L_g} r^2 dL}{R^2} - L_g, \quad (16)$$

where the left-hand side includes only known quantities.

While the entire crystal is submerged in B_2O_3 , $m_g = 0$ and $dm_g/dL = 0$. In that case the counterpart of the volume balance in eq. (13) is of the form

$$\int_0^L r^2\pi dL + V_B + V_{\text{men}} = \pi R^2(L + h + \Delta), \quad (17)$$

where Δ is the distance of the tip of the crystal from the $B_2O_3(\ell)$ - gas interface. Rearranging eq. (17) yields

$$\frac{\int_0^L r^2\pi dL + \frac{m_B}{\rho_B} + V_{\text{men}}}{\pi R^2} - (L + h) = \Delta. \quad (18)$$

Comparing eqs. (16) and (18) we observe that the left-hand sides are identical. As the crystal protrudes through the encapsulant, Δ changes from positive [eq. (18)] to negative [eq. (16)]. This provides an important clue in the computation of L_g . The numerical solution of eq. (16) for L_g starts when $\Delta = 0$.

2.2 Determination of the macroscopic growth rate, dL/dt

The macroscopic growth rate given in eq. (7) can be rewritten as

$$\frac{dL}{dt} = \frac{d\ell}{dL} \frac{dL}{dt} - \frac{dh}{dL} \frac{dL}{dt} + p. \quad (19)$$

Hence, by a rearrangement of eq. (19) dL/dt becomes

$$\frac{dL}{dt} = \frac{p}{1 - d\ell/dL + dh/dL}. \quad (20)$$

To eliminate the melt coordinate ℓ we use the relation for the conservation of melt [eq. (8)], which gives

$$\ell = \frac{m + \rho_\ell V_{\text{men}}}{\rho_\ell \pi R^2} = \frac{\pi \int_0^L \rho r^2 dL + \rho_\ell V_{\text{men}}}{\rho_\ell \pi R^2}. \quad (21)$$

Differentiating eq. (21) with respect to L yields

$$\frac{d\ell}{dL} = \frac{r^2}{kR^2} + \frac{dV_{\text{men}}/dL}{\pi R^2}. \quad (22)$$

Finally, introducing eq. (22) into eq. (20) gives the macroscopic growth rate in the form

$$\frac{dL}{dt} = \frac{p}{1 - \left(\frac{r^2}{kR^2} + \frac{dV_{\text{men}}/dL}{\pi R^2} \right) + \frac{dh}{dL}}. \quad (23)$$

2.3 Transformation of the crystal length to time coordinate

In general, r is known as a function of L . On the other hand the DWGS is recorded as a function of time. Therefore, a transformation from L to t is necessary in the analysis. From the continuity condition [eq. (6)] t can be expressed as

$$t = \frac{L + h(L) - \ell(L)}{p}. \quad (24)$$

Substituting eq. (21) into eq. (24) yields

$$t = \left(L + h - \frac{\int_0^L r^2 dL}{kR^2} - \frac{V_{\text{men}}}{\pi R^2} \right) / p. \quad (25)$$

2.4 Meniscus height and related properties

Mika and Uelhoff¹¹ have determined the meniscus shape and interface height, h , occurring during Czochralski growth by a numerical solution of the Euler-Laplace differential equation. Although, in principle, numerical results for the meniscus could be generated concurrently with the computation of the DWGS, here we prefer to rely on faster closed form solutions of the capillary equation. Fortunately, as shown by Mika and Uelhoff,¹¹ the analytical approximation of Tsivinskii¹² describes the exact values of h with excellent precision over a wide range of joining angles, α , and radii. We have verified that for $\alpha \approx -10^\circ$ and $r \approx 0.5$ cm the error in using Tsivinskii's approximation is less than ~ 2 percent. Subsequent computer simulations have demonstrated that in practical situations, except when there is a very rapid

drop in radius, the stated limit on α is obeyed. The fact that for a short time after seeding r is less than 0.5 cm is of minor importance owing to the initial insensitivity of the DWGS to h .

Therefore, Tsivinskii's equations¹² for the interface height and shape will provide the basis of our DWGS analysis. In our notation Tsivinskii's equation for h is of the form

$$h = A[(1 - \sin \alpha + u^2)^{1/2} - u], \quad (26)$$

where

$$u = \frac{A \cos \alpha}{4r(L)}$$

and A is a constant. According to Egorov, Tsivinskii and Zatulovskii's modification for LEC growth of Tsivinskii's original work A is given by¹³

$$A = \sqrt{\frac{2\sigma}{(\rho_L - \rho_B)g}}. \quad (27)$$

The joining angle α is the sum of the growing angle, ϕ , and contact (wetting) angle, β (Fig. 1), i.e.,

$$\alpha = \phi + \beta. \quad (28)$$

In general, β is a constant, while the growing angle is a function of the derivative dr/dL according to⁹

$$\tan \phi = \frac{dr}{dL}, \quad \phi = \tan^{-1} \frac{dr}{dL}. \quad (29)$$

Thus, h [eq. (26)] can be expressed in terms of r and dr/dL at the solid-liquid interface.

Consequently, the required first derivative of h , dh/dL , includes the second derivative d^2r/dL^2 . This can be readily shown by the substitution of the combined eqs. (28) and (29) into eq. (26). Then, using standard trigonometric relations for angle-sums and inverse functions and differentiating the expression thus obtained with respect to L we find

$$\frac{dh}{dL} = \frac{A}{\left[1 + \left(\frac{dr}{dL}\right)^2\right]^{1/2}} \left[\frac{uv - \lambda}{(1 - \sin \alpha + u^2)^{1/2} - v} \right], \quad (30)$$

where

$$\lambda = \frac{1}{2} \frac{d^2r/dL^2}{1 + \left(\frac{dr}{dL}\right)^2} \left(\cos \beta - \frac{dr}{dL} \sin \beta \right)$$

$$v = -\frac{A}{4r^2} \left[\frac{rd^2r/dL^2}{1 + \left(\frac{dr}{dL}\right)^2} \left(\frac{dr}{dL} \cos \beta + \sin \beta \right) + \frac{dr}{dL} \left(\cos \beta - \frac{dr}{dL} \sin \beta \right) \right].$$

Another derivative of interest in eq. (12) is $d\alpha/dL$. From eqs. (28) and (29) we have

$$\frac{d\alpha}{dL} = \frac{d(\tan^{-1}dr/dL + \beta)}{dL} = \frac{1}{1 + (dr/dL)^2} \frac{d^2r}{dL^2}. \quad (31)$$

A property closely related to the interface height is the volume of the skirt-shaped meniscus, V_{men} (Fig. 1), and its axial derivative dV_{men}/dL . The quantity V_{men} can only be evaluated by a numerical technique. Egorov et al.¹³ provide the x, y coordinates of the meniscus in the form of the integral

$$x = \int_h^y \frac{Q}{\sqrt{1-Q^2}} dy + r(L), \quad (32)$$

where

$$Q = \sin \alpha - \left(\frac{1}{A^2} + \frac{\cos \alpha}{2rh} \right) (y^2 - h^2).$$

Owing to axial symmetry summing up narrow segments of area $x^2\pi$ provides the volume of the meniscus as

$$V_{\text{men}} = \pi\delta \sum_{y_i=0}^{y_i=h-\delta} x^2(y_i), \quad (33)$$

where δ is the thickness of a segment.

Apart from performing numerical differentiation, no simple method exists for the evaluation of dV_{men}/dL . If the shape of the meniscus is that of a right cylinder, then V_{men} and dV_{men}/dL are given by

$$V_{\text{cyl}} = r^2\pi h$$

and

$$\frac{dV_{\text{cyl}}}{dL} = 2r\pi h \frac{dr}{dL} + r^2\pi \frac{dh}{dL}. \quad (34)$$

Simulations indicate that the numerical derivative dV_{men}/dL is closely approximated by

$$\frac{dV_{\text{men}}}{dL} \approx \frac{V_{\text{men}}}{V_{\text{cyl}}} \frac{dV_{\text{cyl}}}{dL}. \quad (35)$$

III. RESULTS AND DISCUSSION

3.1 Computing methods

We have succeeded in reducing the theoretical determination of the DWGS to a differential equation in r , dr/dL and d^2r/dL^2 . Therefore, if the shape of the crystal is given, the DWGS is calculable; conversely, the DWGS completely defines the radius of the crystal at any moment of interest. Here, we address ourselves to the first part of the problem, though the eventual objective is diameter control emphasizing the DWGS.

The program to evaluate the DWGS was written in HP Basic. A key feature in the efficient organization of the computations is the use of subroutines for dm_g/dL [eq. (14)], dL_g/dL [eq. (15)], dL/dt [eq. (23)], h [eq. (26)], dh/dL [eq. (30)], $d\alpha/dL$ [eq. (31)], V_{men} [eq. (33)], dV_{men}/dL [eq. (35)], as well as for $r(L)$ and its first and second derivatives.

Numerical integration was necessary to obtain V_{men} and the crystal volumes between O and L and O and L_g . It was found advantageous to evaluate the crystal volumes in the main routine because during the simulated growth of a crystal incrementing L in steps of 0.1 cm, the accumulating volume elements could be retained in memory.

As mentioned earlier, the quantity L_g is only required after the seed emerges from the $B_2O_3(\ell)$ encapsulant. Mathematically speaking, this occurs when the left-hand side of eq. (16) or (18) becomes negative. To determine L_g , at each increment of L the left-hand side of eq. (16) is evaluated. Then, a linear search is performed, substituting trial values of L_g into the right-hand side of eq. (16) in the range $L'_g - 0.2$ to $L'_g + 0.4$ cm in steps of 0.01 cm, where L'_g is the previous solution to eq. (16) at $L - 0.1$ cm. In case the 0.01-cm mesh around L'_g was too coarse to yield a solution, an option with 0.001-cm increments was available.

There are several mathematical tools available to describe the shape of the grown crystal, $r(L)$. From the actual cross section a large number of r , L coordinates can be generated in tabular form. Then, one obvious choice is to employ an interpretation scheme based on point-by-point tabular values of $r(L)$ in addition to numerical first and second derivatives. A more sophisticated alternative is to determine the Fourier coefficients of the crystal shape. However, we have learned by experience that the most efficient and satisfactory scheme to convert the tabular data into functional form is by using cubic spline regression.

A cubic spline function is a piecewise polynomial of degree three that joins adjacent polynomials in "knots." At the knots the functional values as well as the first and second derivatives are continuous. Ordinarily, a spline passes through all the points within the interval bounded by two knots. In contrast, in spline regression a cubic least-square fit for the same points is obtained. We have adapted, for the

present application, Wold's spline regression technique in terms of B -splines.¹⁴

Fortunately, the key features of the DWGS can be illuminated without invoking advanced data-fitting techniques. A judiciously chosen elementary transcendental function representing $r(L)$, which has n continuous derivatives, will be found suitable to illustrate the effect of various factors governing the course of the DWGS.

3.2 Major influences on the derivative weight-gain signal

3.2.1 Crystal shape

Typical LEC crystals, including GaAs, exhibit a pronounced "shoulder" as the radius rapidly expands from the seed. Boule-to-boule variation in the rate of approach to and radius of curvature of the shoulder is not uncommon. A mathematical description of the shoulder, in general, and a prototype crystal, in particular, is possible by means of the probability density function of the lognormal distribution.¹⁵ Accordingly, we can write for a seed diameter of 0.6 cm

$$r(L) = C \left[\frac{1}{sL\sqrt{2\pi}} \exp - \frac{(\ln^2 L/L_m)}{2s^2} \right] + 0.3. \quad (36)$$

The median length, L_m , and the constant s are chosen such that the maximum r is 2.8 cm at $L = 2.2$ cm. From these conditions we have

$$L_m = 2.2 \exp(s^2)$$

and a lengthy expression for C that is also only dependent on the standard deviation, s . Hence, the function $r(L)$ can be constructed if s is given.

In Figs. 2, 3, and 4 we present the crystal shape $r(L)$ for $s = 0.4, 0.8,$ and 1.2 , respectively. It can be seen that with decreasing s the radius of curvature at the maximum decreases (i.e., $|d^2r/dL^2|$ increases). Furthermore, it can be shown that the rate of approach to the shoulder (dr/dL) rapidly rises as s diminishes.

Substituting eq. (36) and its first and second derivatives into the DWGS [eq. (12)] and the subsidiary equations yields the signal and the crystal cross section as a function of time, both of which are shown in Figs. 2, 3, and 4. The required parametric values are listed in Table I. The results demonstrate that the DWGS is extremely sensitive to dr/dL and $|d^2r/dL^2|$. In fact, the steeper the rise in the shape, the sharper the peak in the DWGS and the larger its absolute magnitude.

Two other characteristic features of LEC pulling that are of crucial importance can also be observed in Figs. 2, 3, and 4. First, the maximum in DWGS precedes in time the maximum in shape by a few hundred seconds. This time lag we shall designate as the "precursor." The second property is the additional maximum in DWGS at a time when

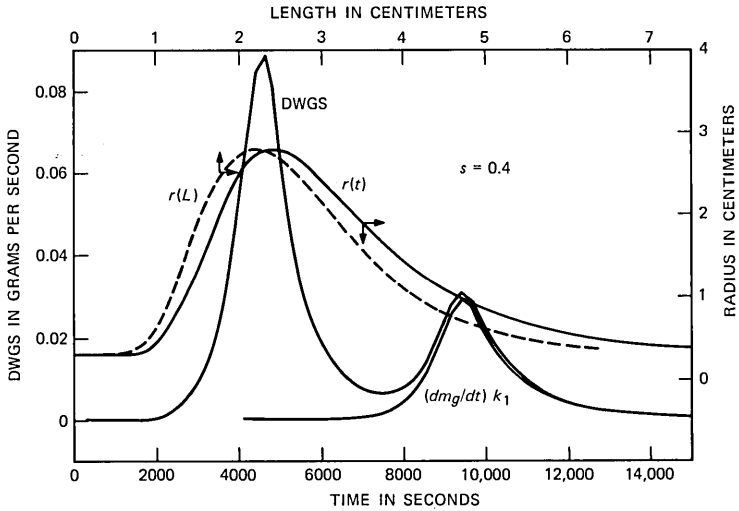


Fig. 2—The DWGS and crystal cross section as a function of time for LEC pulling. The shape versus length curve (---) is based on eq. (36), setting $s = 0.4$. The time derivative of the unencapsulated weight [dm_g/dt , eqs. (14) and (19)] is also shown.

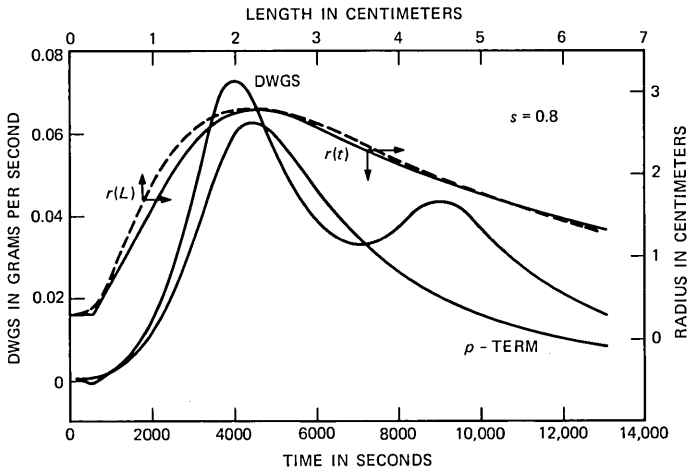


Fig. 3—The DWGS and crystal cross section as a function of time for LEC pulling. The shape versus length curve (---) is based on eq. (36), setting $s = 0.8$. The pull rate term in eq. (12)—the factor usually employed in conventional diameter control—is also shown.

the crystal's radius has already declined. We shall refer to this phenomenon as the "aftershock."

We have investigated the source of the precursor and aftershock. Simulations show that neither r nor its derivatives bears any responsibility. Examining the variation of the interface height with time (Fig.

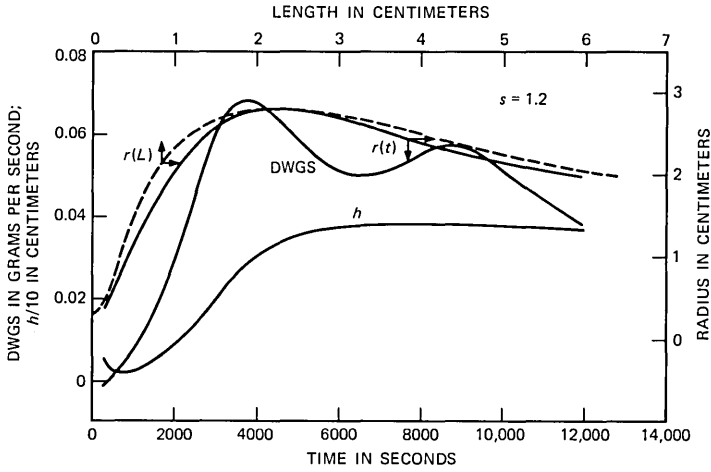


Fig. 4—The DWGS and crystal cross section as a function of time for LEC pulling. The shape versus length (---) curve is based on eq. (36), setting $s = 1.2$. The meniscus height [eq. (26)] is also shown.

Table I—Parameters for the calculation of the DWGS

Density of crystalline GaAs, d [g/cm^3]	5.17
Density of molten GaAs, d_l [g/cm^3]	5.71
Density of $\text{B}_2\text{O}_3(l)$, d_B [g/cm^3]	1.55
Mass of $\text{B}_2\text{O}_3(l)$, m_B [g]	129.5
Pull rate, ρ [cm/s]	3.6×10^{-4}
Crucible radius, R [cm]	3.9
Wetting angle, β [deg]	15
Capillary constant, A [cm]	0.4
Surface tension, σ [dyne/cm]	333

4) we note that starting with a small but finite value at the seed, h saturates beyond the shoulder ($h \approx 0.38$ cm).^{*} Hence, no clue can be extracted from the form of h . Concentrating on the term including the pull rate in eq. (12), one concludes that this conventional description of the DWGS³ holds quite well in the early phase of growth but departs from reality near the shoulder and beyond (Fig. 3). The maximum in the p term and the shoulder perfectly coincide and no secondary maximum appears.

If, however, one examines the derivative dm_g/dt [eqs. (14) and (23),

^{*} The fact that at $t = 0$, $r \neq 0$, thus $h > 0$ is finite contradicts eq. (24). Therefore, to correct eq. (25) the residual time

$$t_0 = \left[h(L=0) - \frac{V_{\text{men}}(L=0)}{\pi R^2} \right] / p$$

is always subtracted from p .

Fig. (2)], it is found that this quantity rises from zero to a maximum value precisely at the time of the aftershock and that the magnitude of the DWGS in this regime is essentially dm_g/dt . In view of the fact that dm_g/dt is specifically associated with $B_2O_3(\ell)$ encapsulation, a reasonable hypothesis is that the aftershock is a consequence of the LEC technique.

3.2.2 Meniscus shape and liquid encapsulation

To isolate the factors leading to the precursor and aftershock we have evaluated the DWGS corresponding to the lognormal $r(L)$ [eq. (36), $s = 0.8$] in the absence of liquid encapsulation. In this case the major equations still hold provided ρ_B , V_B , m_B , m_g , and dm_g/dt are taken as zero. In Fig. 5 we present the DWGS with and without interposing a meniscus between the melt and the growing crystal. It is immediately obvious that in both instances the aftershock disappears from the DWGS. Moreover, if the meniscus is also removed, the precursor is missing, i.e., the times to reach the peak in DWGS and cross section exactly coincide. In fact by simultaneously eliminating $B_2O_3(\ell)$ and the meniscus from the growth system, only the contribution of the conventional pull rate term³ is retained in eq. (12).

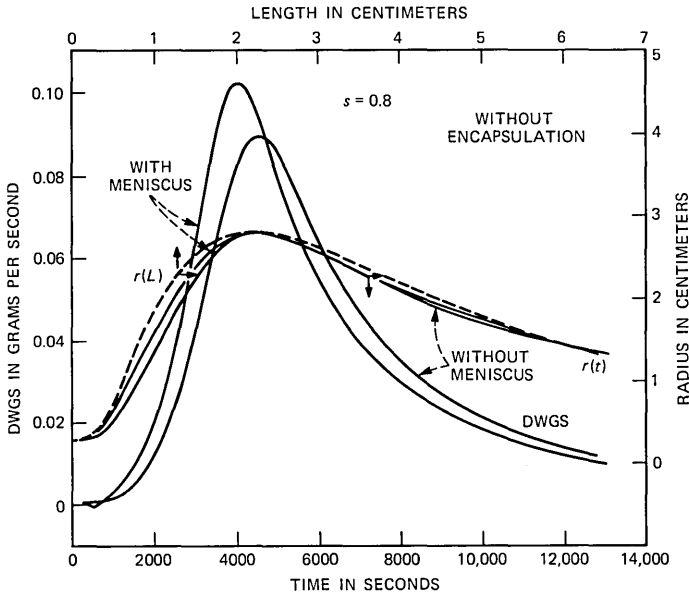


Fig. 5—The DWGS and crystal cross section as a function of time in normal Czochralski pulling without encapsulation. Curves with and without a meniscus present are given. The shape versus length curve (---) is based on eq. (36), setting $s = 0.8$. When the meniscus is absent, the pull rate term in eq. (12) exactly coincides with the curve shown.

Further confirmation of these effects is offered in Fig. 6. Here the calculation of the DWGS is repeated for LEC pulling without a meniscus. As we expected, though the precursor is missing the aftershock reappears. Apparently, up to and over the shoulder the DWGS is governed by the p term. Near the secondary maximum the derivative dm_g/dt becomes the dominant factor.

A more physical interpretation of the aftershock is possible by plotting the time-dependent radius, $r(L_g)$, of the crystal as it just protrudes through the $B_2O_3(\ell)$. In Fig. 6 we show both $r(L)$ and $r(L_g)$ as a function of time. Besides an expected displacement of the radius along the time axis one notes that the maximum in $r(L_g)$ occurs exactly at the time of the secondary maximum in DWGS. From these considerations a surprisingly simple explanation emerges. As more of the shoulder region becomes uncovered from the encapsulant, the DWGS registers the rapidly increasing true weight gain as opposed to the previously measured apparent (buoyancy-reduced) weight gain of the crystal. Hence, at the time of the maximum aftershock both the freshly grown layers at L , submerged in $B_2O_3(\ell)$, and the sizable exposed portions at L_g are detected.

The discovery of the precursor and the aftershock has a significant bearing on the diameter control of GaAs. The precursor is an early warning signal that predicts a maximum in radius a few hundred seconds before the actual event. In other words the DWGs is already

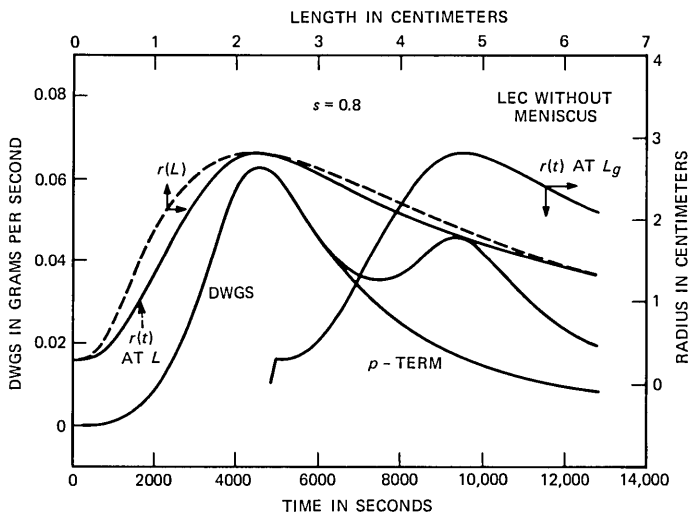


Fig. 6—The DWGS and crystal cross section as a function of time in LEC pulling, free of a meniscus. The radius is shown as a function of time at the solid-liquid interface, L , and at the top of the B_2O_3 layer, L_g . The pull rate term is also shown. The shape versus length curve (---) is based on eq. (36), setting $s = 0.8$.

dropping while the diameter is still increasing. Thus the controller has sufficient time to react to an unwanted change.

When the aftershock is observed there is a good chance that the controller misinterprets it as an unexpected gain in diameter and, if undesirable, will respond accordingly. Then, of course, a visible reduction in diameter will occur. By the time this is noticed, corrective action taken cannot restore the loss in diameter control. At best, beyond the shoulder a "sinusoidal" cross section with small amplitude results.

In view of these observations one cannot base the diameter control algorithm for GaAs on the customary p term³ in eq. (12). To be sure, from the seed up to near the shoulder perhaps it suffices. However, for the bulk of the boule, control by means of such a traditional method is clearly inadequate.

3.3 Comparison with experiment

We have extracted the salient features of the DWGS arising during LEC pulling by the use of the lognormal probability function for the crystal's cross section. To compare the experimentally determined DWGS with the theoretical one the idealized crystal shape must be replaced by that of a GaAs boule grown by the LEC method under closely controlled conditions. Some of the growth parameters of interest for this specially prepared crystal are listed in Table I. The weight and length of the crystal are 523.3 grams and 6.6 cm, respectively.

The crystal coordinates (r versus L) can be derived from a two-dimensional projection employing a digitizer. Owing to the eccentricity of the actual $\langle 100 \rangle$ boule, a single view is insufficient. We have obtained planar projections with sharp contours by a photographic technique. The crystal was placed on the top of a light box, backlit, and photographed using a high-contrast film. Then, the procedure was repeated following a 90-degree rotation around the axis. In this manner, the two photos provided a total of four cross sections. Digitizing the shape in 0.5-mm intervals 133 pairs of (r , L) coordinates were collected for each of the four contours. At the shoulder the maximum error between the crystal and its projection is less than 0.5 percent. The computed mean weight is 528.6 g with a standard deviation of ± 9.3 g, which should be compared with the measured weight of 523.3 g. Clearly, an excellent tabular description of the crystal's cross section has been accomplished. The data points for one of the four views are shown in Fig. 7.

The alternatives to represent the tabular data have been outlined earlier. In principle, because of close spacing, there would be no obstacle to a simple interpolation of the $r(L)$ values. However, as we illustrated in Fig. 7, using the numerical first and second derivatives may be problematic on account of noise. Indeed, when the DWGS was

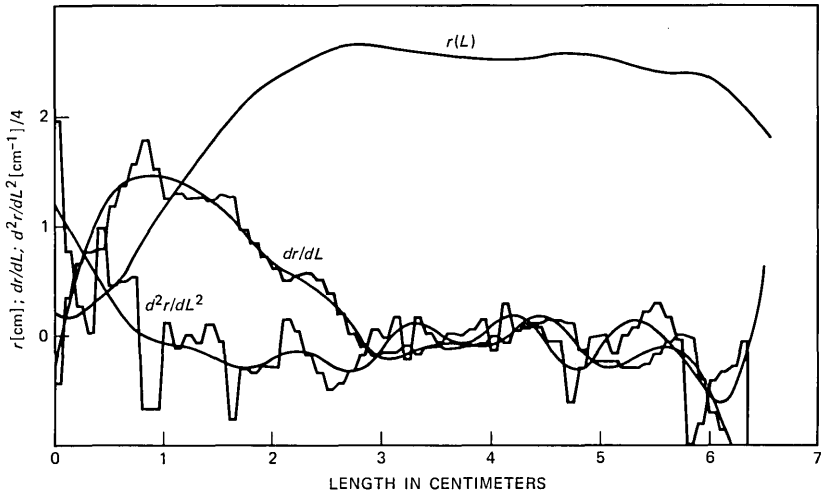


Fig. 7—One of the four experimental contours of a GaAs crystal and its cubic spline-regression representation (—). The cubic spline regression of the numerical first derivative dr/dL is also shown. In addition, the numerical second derivative d^2r/dL^2 is given, together with the quadratic expression derived from the cubic spline of dr/dL .

computed from the numerical derivatives severe discontinuities appeared in the results. This suggests that the steep “jumps” in d^2r/dL^2 seriously affect at some locations the shape and magnitude of the DWGS.

A different approach is the finite Fourier transform of the $r(L)$ points. However, owing to slow convergence a large number of terms is required. Eight terms were found to be sufficient to describe the seed and shoulder regions. Beyond the shoulder 18 terms were necessary to fit the data, albeit with concomitant high-frequency noise. Since the number of required terms was unpredictable and the computation inefficient, Fourier representation of the $r(L)$ data was abandoned. Nonetheless, the DWGS based on Fourier analysis provided a reasonable description of the measurements.

Cubic spline regression¹⁴ has proven to be the most promising method to cast the tabular data on crystal cross sections into functional form. At the middle of the axis, the crystal was separated into two slightly overlapping domains and within each 12 equally spaced knots were positioned. The outstanding spline fit to one of the contours is given in Fig. 7. There are two avenues open to obtain the first derivative dr/dL . One is a simple differentiation of the shape's spline, which results in a continuous quadratic representation. Unfortunately, in this case the second derivative becomes a continuous network of connected straight lines (the third derivative is discontinuous) leading

to a jagged DWGS. A more fruitful approach is to fit the numerical first derivative itself by a cubic spline. Then, the noisy second derivative is described by a quadratic expression. In Fig. 7 we show the cubic spline fit to the numerical first derivative data as well as to the quadratic function passing through the numerical d^2r/dL^2 . We note that a reasonably good description of the derivatives has been obtained by the spline regression technique. Therefore, all four contours were treated in a like manner.

Having thus established the shape and shape derivatives for an actual GaAs crystal, one can readily evaluate the corresponding DWGS by means of the analytical tools described earlier. Among the input parameters listed in Table I, all the densities were obtained from a recent critical evaluation of Jordan.¹⁶ The selected contact angle (15 degrees) is consistent with theoretical and experimental findings on Ge and Si.^{17,18,19} The capillary constant A and the surface tension, σ , are connected via eq. (27). Though the recommended value for A by Egorov et al. is 0.48 cm,¹³ we have achieved a slightly better fit to the experimental DWGS using $A = 0.4$ cm, which is equivalent to $\sigma = 333$ dynes/cm. The surface tension of GaAs thus obtained is consistent with that for other III-V compounds (InP²⁰ and GaSb²¹)Si and Ge.²²

The gross appearance of the DWGS is not overly sensitive to a change in A . Nonetheless, a detailed examination reveals that the peak corresponding to the shoulder becomes higher and broader as A increases from 0.28 to 0.48 cm. With respect to the rest of the DWGS one can conclude that the peak-to-valley dynamic range diminishes as A drops. Among the effects of the other parameters, we have observed a rise in DWGS with an increase in p and a reduction of the time scale with a decrease in crucible radius.

In Fig. 8 we have reproduced the experimental derivative weight-gain signal as a function of time, as obtained on a carefully calibrated strip-chart recorder during the LEC growth of GaAs. The predicted DWGSs for the four contour lines as well as the time-dependent cross sections are also shown. Clearly the theoretical curves envelop the measured values, providing an excellent overall description.

The computed DWGSs exhibit both the early warning precursor and aftershock. In the inflection before the first peak we can discern the time when the seed first protrudes through the $B_2O_3(\ell)$. Since an actual crystal may possess additional cross-sectional bulges beside the shoulder, the secondary maximum in DWGS reflects both the memory of that earlier maximum in radius and the current expansion. To a small extent the DWGS is influenced by the input shape derivative function. For example, in the trough above the shoulder, a spline of the first derivative with many more knots would lead to a smoother transition to the next peak. Likewise, near the last peak, an improved

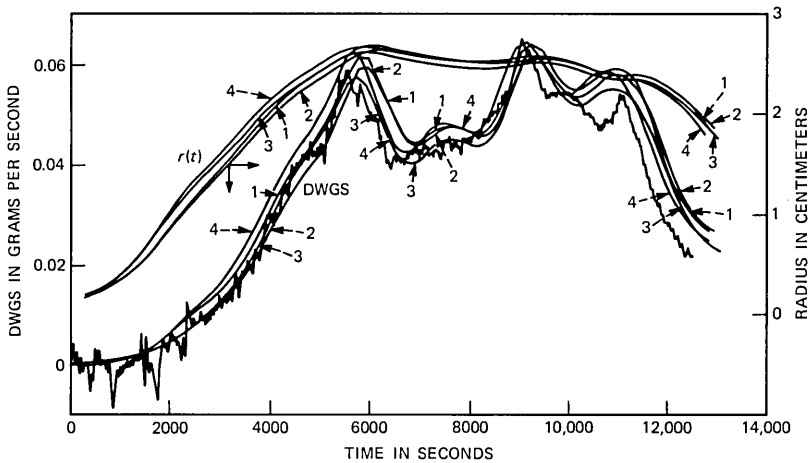


Fig. 8—DWGS (—) and crystal cross section as a function of time for four experimental contour lines of LEC-pulled GaAs. The measured DWGS was traced from the original strip-chart recording.

agreement with the experimental data is possible at the cost of a finer derivative spline. Near the bottom of the crystal the accord with the measured DWGS is more limited, perhaps on account of the increasingly negative turn in the joining angle α (between -20 and -40 degrees). Under those circumstances Tsivinskii's approximation for the meniscus properties becomes inaccurate.¹¹

3.4 Prospects for diameter control

The excellent agreement between predicted and experimental DWGS strongly suggests that diameter control of GaAs by the derivative measurement is feasible. There are two approaches to diameter control implied by the modeling calculations. One is a real-time instantaneous determination of the radius during growth from the DWGS by solving the differential equation (12). Then, depending on whether r is increasing or decreasing from a preset limit, the RF input power is changed by a suitable amount.

The other method is the construction of an "ideal" crystal with appropriate tolerances on a drawing board. Taking into account the tolerances for this crystal, a band of DWGS can be evaluated as has been done for Fig. 8. Subsequently, the control of diameter is reduced to the problem of adjusting the power input in such a fashion that a template established by the band of DWGSs is followed.

IV. ACKNOWLEDGMENTS

We appreciate the continuing interest of J. W. Nielsen in this work and his critical reading of the manuscript. We thank L. J. Oster for the

expertly grown GaAs crystal and for his careful determination of the derivative weight-gain signal.

REFERENCES

1. D. T. J. Hurlle, "Control of Diameter in Czochralski and Related Crystal Growth Techniques," *J. Crystal Growth*, **42**, No. 1 (December 1977), pp. 473-82.
2. R. E. Lorenzini, F. S. Nuff, and D. J. Blair, "An Overview of Si Crystal Growing Processes," *Solid State Techn.*, **17**, No. 2 (February 1974), pp. 33-6.
3. A. J. Valentino and C. D. Brandle, "Diameter Control of Czochralski Grown Crystals," *J. Crystal Growth*, **26**, No. 1 (November 1974), pp. 1-5.
4. H. D. Pruettt and S. Y. Lien, "X-ray Imaging Technique for Observing LEC Crystal Growth," *J. Electrochem. Soc.*, **121**, No. 6 (June 1974), pp. 822-6.
5. H. J. A. van Dijk, C. M. G. Jochem, G. J. Scholl, and P. van der Werf, "Diameter Control of LEC Grown GaP Crystals," *J. Crystal Growth*, **21**, No. 3 (February 1974), pp. 310-12.
6. M. Cole, R. M. Ware, and M. A. Whitaker, paper presented at ECCG1, Zurich (1976).
7. P. Schmaker, A. Kramer, and W. Staehlin, paper presented at ECCG1, Zurich (1976).
8. R. M. Ware, private communication.
9. W. Bardsley, D. T. J. Hurlle, and G. C. Joyce, "The Weighing Method of Automatic Czochralski Crystal Growth, Part I," *J. Crystal Growth*, **40**, No. 1 (September 1977), pp. 13-20.
10. W. Bardsley, D. T. J. Hurlle, G. C. Joyce, and G. C. Wilson, "The Weighing Method of Automatic Czochralski Crystal Growth, Part II," *J. Crystal Growth*, **40**, No. 1 (September 1977), pp. 21-8.
11. K. Mika and W. Uelhoff, "Shape and Stability of Menisci in Czochralski Growth and Comparison with Analytical Approximations," *J. Crystal Growth*, **30**, No. 1 (August 1975), pp. 9-20.
12. S. V. Tsvinskii, *Inzh. Fiz Zh.*, **5** (1962), p. 59.
13. L. P. Egorov, S. N. Tsvinskii, and L. M. Zatulovskii, "The Shape of the Melt Column in Growing a Crystal from Under a Flux Layer," *Bull. Acad. Sci. USSR, Phys. Ser.*, **40**, No. 7 (1976), pp. 172-4.
14. S. Wold, "Spline Functions in Data Analysis," *Technometrics*, **16**, No. 1 (February 1974), pp. 1-11.
15. Y. L. Maksoudian, *Probability and Statistics with Applications*, Scranton, PA: International Text Book Co., 1969, pp. 123-4.
16. A. S. Jordan, "An Evaluation of the Thermal and Elastic Properties Affecting GaAs Crystal Growth," *J. Crystal Growth*, **49**, No. 4 (August 1980), pp. 631-42.
17. P. I. Antonov, "Capillary Phenomena in Crystal Growth" in *Growth of Crystals*, Vol. 6A, ed., N. N. Sheftal, New York: Consultants Bureau, 1968.
18. T. Surek and B. Chalmers, "The Direction of Growth of a Crystal in Contact with Its Melt," *J. Crystal Growth*, **29**, No. 1 (May 1975), pp. 1-11.
19. T. Surek, "The Meniscus Angle in Ge Crystal Growth from the Melt," *Scripta Met.*, **10**, No. 5 (May 1976), pp. 425-31.
20. A. S. Popov and L. Demberel, "Surface Tension of Molten Stoichiometric InP," *Kristall Und Technik*, **12**, No. 11 (1977), pp. 1167-70.
21. M. Ya. Dashevskii, G. V. Kukuladze, V. B. Lazarev, and M. S. Mirgalovskii, "Surface Effects in GaSb Melts," *Inorg. Materials*, **3**, No. 9 (September 1967), pp. 1360-66.
22. J. C. Phillips and J. A. VanVechten, "Macroscopic Model of Formation of Vacancies in Semiconductors," *Phys. Rev. Lett.*, **30**, No. 6 (February 1973), pp. 220-3.

Growth, Complexity, and Performance of Telephone Connecting Networks

By V. E. BENEŠ

(Manuscript received December 10, 1981)

In an effort to free telephone traffic theory from some of its dependence on independence assumptions, and to reap some benefit from its traditional state equations, a systematic search is made to find relationships between load, loss, size, structure, and other network parameters that are simple, universal, and informative. Three principal topics are covered:

(i) A load-loss-size formula, linking some half-dozen network parameters by a rational function, and used repeatedly to give

(ii) Lower bounds on the number X of crosspoints in networks

(iii) Asymptotic results about blocking, growth, and complexity of selected network structures in passing from finite to "infinite" sources at constant load.

The major results in (ii) imply that for all practical networks on N terminals, the crosspoint count X must grow like $N \log N$, i.e., incurring loss by restricting access or concentrating cannot avoid the $N \log N$ growth rate known to be exacted by nonblocking networks. The chief result under (iii) is that as a constant load is spread over N terminals, then the number X of crosspoints needed to keep loss less than $\epsilon > 0$ need grow only linearly with N , at a rate dependent on ϵ , while the usage (erlangs carried per terminal) goes to zero.

I. INTRODUCTION

The relationships between traffic carried and traffic lost, between load and loss, have always been at the center of interest in telephone traffic theory. Since the time of Erlang,¹ over fifty years ago, the principal problems of traffic theory have been analytical: to predict mathematically, from the structure and mode of use of a switching or connecting network, and from the assumed stochastic behavior of the customers, how much traffic the network will carry on the average,

and how much it will *lose* as a result of blocking, overload, suboptimal routing, or incomplete searches for paths. As telephone networks have become larger, two more design parameters of interest have emerged and now command attention: the *size* of the network as measured by the number of crosspoints, and its *complexity* as measured, for example, by the number of stages of switching it has.

The probabilistic principle that it is very unlikely for more than a moderate number of customers to want to talk simultaneously has been the theoretical basis of traffic theory since its start. We can view it as an unrefined analog of the principle in information theory that separates a relatively small class of events that exhaust most of the probability from a remaining large class of very unlikely events. This principle has led quite naturally to the use of concentrators, and of networks in which blocking, mismatch, and overflow all can and do occur as it were by design. It is a function of traffic theory to articulate this principle in mathematical models for operating telephone networks, and to use such models to examine its implications for the growth and complexity of networks, as well as their loads and losses.

Even for the simplest stochastic models, progress with these tasks and problems has been very slow because of the combinatorial complexity of the network, the very large number of network states, and the lack of approximate methods. Thus, it is particularly important to find relationships between load, loss, size, and other network parameters that are simple, universal, and useful, even for very large networks. They should be simple in, for example, *not* requiring solution of very high-order systems of equations, universal in being relatively independent of network structure, and useful in providing inequalities, estimates of performance, and information about the growth of cost and complexity with network size.

In this paper we try systematically to sketch out some of these relationships and associated ideas. The results are of necessity spotty, and no claim is made of completeness or originality, only of rigor. Three principal topics are taken up here: (i) a load-loss formula, linking some half-dozen network and performance parameters by a rational function; (ii) lower bounds on the number of crosspoints in a network; (iii) asymptotic results about blocking, growth, and complexity of selected network structures in the limit of passage from finite sources to Poisson arrivals, with total offered traffic held constant.

II. SUMMARY

The organization of the sequel is as follows: By way of some background, we start with discussions of blocking, loss, concentration, etc., and of their relation to the basic principles of telephone traffic theory and engineering. After various preliminary sections on model-

ling, we call attention to a (known) generalized Erlang formula that connects some of the important parameters of an operating network. We note its technical consequences, and use them repeatedly in the rest of the paper.

Next we take up the problem of the growth of the number of crosspoints with the number N of terminals. The question we try to answer is this: When can the $N \log N$ order of growth necessary for nonblocking networks be reduced by allowing a fixed, small probability of blocking, using, for example, concentrators, or other forms of incomplete access? The answer is that it cannot, unless we consider a familiar, special kind of low traffic limit in which line usage vanishes. It is shown, quite generally, that networks arranged in stages *must* grow like $N \log N$ if certain (very reasonable and mild) traffic, access, and symmetry conditions are met. This result, similar to known results for nonblocking networks, implies that neither judicious concentration nor a nonzero loss can lower the order of growth from the $N \log N$ exacted by the nonblocking case.

The final sections describe various large networks with a simple structure vis-à-vis blocking; their loss probability can be calculated exactly in spite of the astronomical number of states. These networks are based on such structures as trunk groups, frames, and remote concentrators, all familiar to the traffic engineers. We are interested in studying these exact solutions as we let the network grow while keeping the total traffic constant; this kind of growth amounts to adding more and more customers, each of whom contributes less and less traffic, and results in a passage from finite to infinite sources (a Poisson process of arrivals) at constant offered load. The blocking formulas can be studied in this limit, and they lead to close connections with the classical Erlang function, $E(c, a)$. As an application, we can give methods of synthesizing very large networks with prescribed blocking probabilities. In particular, as a constant offered load is spread over more and more customers, the number of crosspoints sufficient to achieve less than ϵ in blocking need grow only linearly with the number of customers.

There is a bibliography of background reading and related work following the references.

III. STATEMENT OF RESULTS

3.1 *Generalized Erlang formula*

Using some standard “dynamic” assumptions² to describe random traffic, we show that half-a-dozen parameters, all characteristic of network size and performance, are related by a simple, rational function. These parameters are:

N = number of terminals on a side of a two-sided network
 = number of inlets = number of outlets
 m = (mean) carried load = equilibrium average number of calls in progress
 λ = calling rate per pair of idle customers
 bl = probability of blocking, from the "wire-chief's" point of view
 σ^2 = variance of number of calls in progress,

and the formula states that,² very simply,

$$1 - bl = \frac{1}{\lambda} \frac{m}{(N - m)^2 + \sigma^2}. \quad (1)$$

For some purposes the parameters

$p = m/N$ = line usage = erlangs carried per inlet (outlet)
 $a = \lambda N^2$ = total offered load (when everyone is idle)

are more significant or convenient, and using them the formula is recast as

$$a(1 - bl) = \frac{m}{(1 - p)^2 + \sigma^2 N^{-2}}.$$

Indeed, there are many ways of twisting and inverting the basic formula, each one illuminating some special aspect; several such will appear later. It can be shown that if $N \rightarrow \infty$ while $a = \lambda N^2$ remains constant, then p and $\sigma^2 N^{-2}$ go to zero, and we have Erlang's original result, the "tautology"

$$a(1 - bl) = m$$

or

$$\text{blocking} = \frac{\text{load lost}}{\text{load offered}}.$$

The formula (1), really a generalization of Erlang's loss formula, is useful for the following applications:

- (i) Order of magnitude estimates
 - (ii) Asymptotic analyses for large networks, with or without a passage from finite to infinite sources
 - (iii) Bounds on the number of switches in a network
 - (iv) Growth and complexity bounds for networks with given load and loss constraints
 - (v) Synthesis of networks having prescribed parameters.
- All these applications are illustrated in the text that follows.

3.2 Concentration

The principle that high occupancy states are very unlikely has

suggested the use of concentration, and it is pertinent to assess the effect and value of concentration in large networks. Some results of this kind are in the next three subsections, and they warrant these conclusions:

(i) Practical networks must grow like constant $N \log N$, whether there is concentration or not. Concentration affects primarily the value of the constant, and of course the blocking and the carried load.

(ii) The extent of possible concentration is limited by the loss and the carried load. As might be expected, higher line usage and lower blocking imply less concentration.

3.3 Growth without concentration

In the prototypical networks without concentration (e.g., those made of stages of square switches), the number of crosspoints for N customers must grow like $N \log N$ no matter what load is carried. Several arguments are given for similar lower bounds, some purely combinatorial, others involving traffic concepts and parameters. In particular, if blocking is to be kept less than ϵ , and total traffic $a = \lambda N^2$ is kept constant while N increases, the requisite networks must grow like $N \log N$ if they are made of stages of square switches: especially, for s stages

$$X = \text{number of crosspoints} \geq N \log N + sN + \log(1 - \epsilon) - a.$$

3.4 Growth with full access, allowing concentration

Some simple and mild combinatorial properties, possessed by all practical networks, mandate an $N \log N$ order of growth in the number X of crosspoints, even when concentration is permitted. A network provides "full access" if every inlet can reach every outlet by some path. A network is said to be "arranged in stages" (or "made" of stages) if its terminals are partitioned into sets T_1, T_2, \dots, T_{s+1} , such that T_1 consists of the inlets, T_2 to T_s are sets of internal nodes or junctors, and T_{s+1} are the outlets, with crosspoints placed only between T_i and T_{i+1} , $i = 1, \dots, s$. (See Fig. 1). Here s is the number of stages, and every call traverses each T_i exactly once, in the specified order or its reverse. Finally, a network arranged in stages is called "symmetric" if it looks the same from each terminal in any given T_i ; we content ourselves with this informal definition here; a precise one can be given in terms of group theory.³

We prove this fundamental telephonic fact: A symmetric network that provides full access and is arranged in stages must have at least

$$en \log N \quad e = 2.71828 \dots$$

crosspoints, where N is the number of inlets (outlets, too), and n is the "neck size," defined as the size of the smallest T_i :

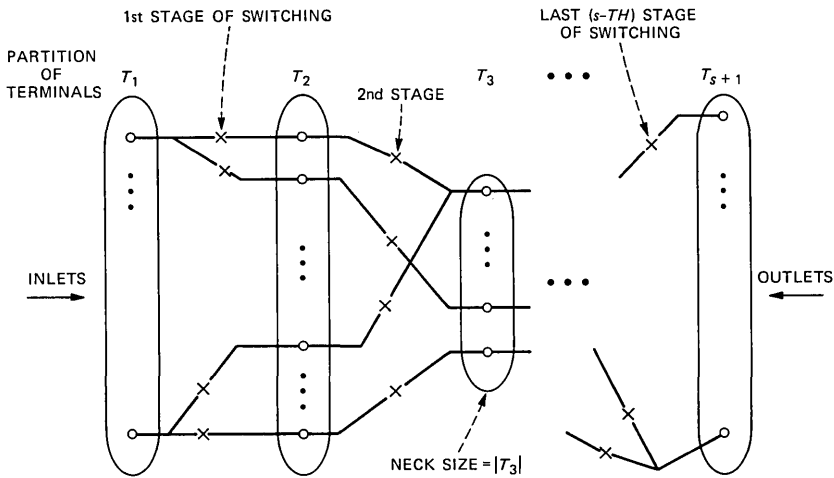


Fig. 1—Network arranged in stages.

$$n = \min_{1 \leq i \leq s+1} |T_i|, |A| = \text{cardinality of } A.$$

The ratio n/N is a global measure of concentration or expansion. If, in particular, all the T_i are equinumerous, as happens if the network is made of stages of square switches, then the neck size is N , and there must be at least $eN \log N$ crosspoints, regardless of the traffic characteristics.

3.5 Growth without full access, allowing concentration

The condition of full access is so reasonable that few engineers would consider a network that lacks it. Nevertheless, probabilistic arguments yield $N \log N$ lower bounds for the crosspoint count X even when this condition is dropped. The point is that if the network is to carry a reasonable load, the “neck size” n cannot be too small; especially, it follows from the Erlang formula (1) that n must exceed $(1 - \sqrt{1 - p}) N$ for line usage $p = m/N$. Similar lower bounds involving also the required loss can be derived. These lower bounds put a limit on how much one can concentrate (measuring global concentration by neck size), and they lead to $N \log N$ lower bounds for X even when there is not full access. For example, any network arranged in stages has

$$X \geq \frac{1}{2} e(1 - \sqrt{1 - p}) N \log N + o(N)$$

if each customer’s line carries p erlangs. Thus any sequence of networks that grow in the strong sense that p is bounded away from zero must grow like $N \log N$, if they are arranged in stages.

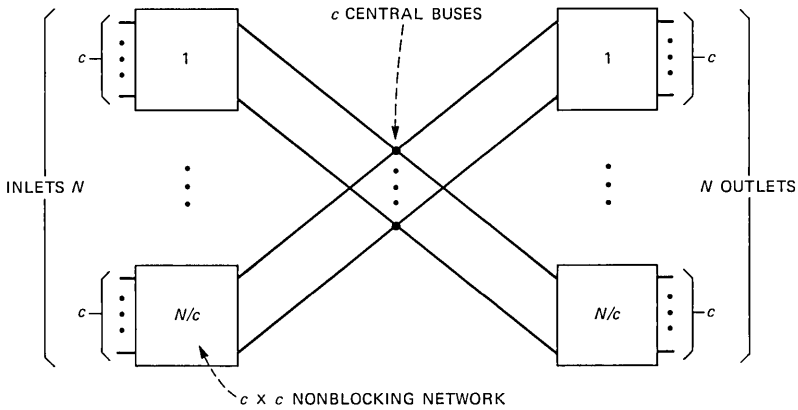


Fig. 2—Central bus network.

3.6 Asymptotic results

Three network structures lead, in the model to be used, to statistical equilibrium equations that have the “product form” solution, and so all their interesting parameters can be calculated exactly from a partition function, and their behavior in the limit $N \rightarrow \infty$, $a = \lambda N^2 = \text{constant}$, studied. They are (see Figs 2 through 4.)

(i) The central bus concept—two large N -to- c nonblocking concentrators back to back, with c central buses

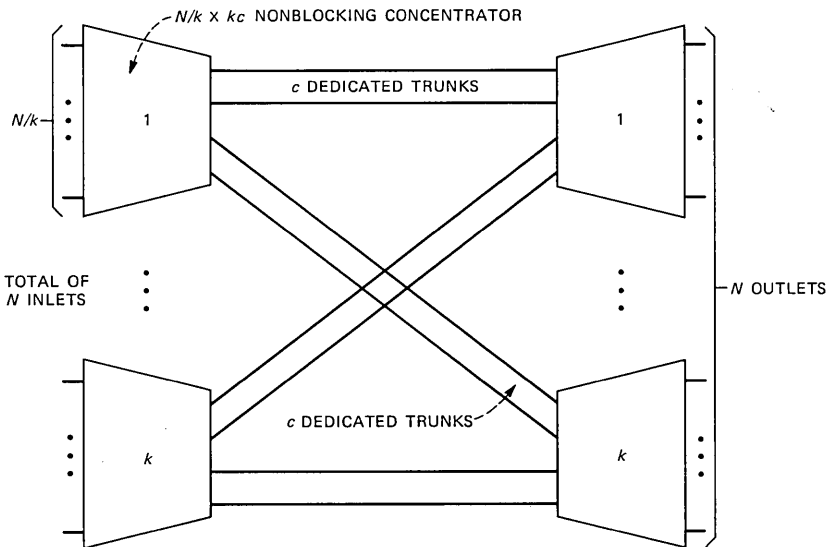


Fig. 3—Frame network.

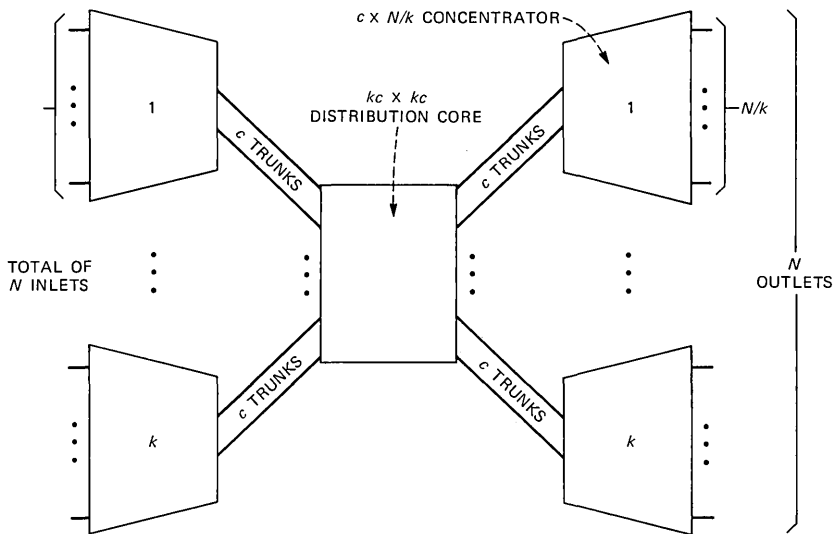


Fig. 4—Remote concentrator network.

(ii) The frame concept—large concentrators connected pairwise by dedicated groups of c trunks

(iii) The remote concentrator concept—large concentrators, each connected to a central distribution core by its own group of c trunks. In the “weak” limit as $N \rightarrow \infty$, $a = \lambda N^2 = \text{constant}$, the first two structures are (not surprisingly) closely connected to Erlang’s formula for loss: their probabilities of loss are bounded above by, and approach, the value $E(c, a)$. The third is more subtle and elusive, since loss can occur in either one of two relevant trunk groups, but also more interesting. We can find $E(\cdot, \cdot)$ -type bounds on the loss, but the question is whether asymptotically

$$\text{loss} \leq 1 - (1 - b)^2 + o(1),$$

where b is the chance that all c trunks on one (any) concentrator are busy, in the limit, remains open. This quadratic upper bound is what the loss would be in the limit if the two relevant trunk groups were independent, each with “blocking” b . Very few of the many “blocking polynomial” approximate formulas used in practice have been vindicated by so much as an inequality proved in a dynamical model.

It follows from our analyses that for each $\epsilon > 0$, and each of the three kinds of network structure considered, there exist arbitrarily large networks of that kind, with loss less than ϵ , and total offered traffic $a = \lambda N^2 = \text{constant}$, whose crosspoint count X grows at most linearly with N . We view this as a limit result in the “weak” direction, since a constant amount of traffic $a = \lambda N^2$ is being divided up among

N lines, $N \rightarrow \infty$. The carried loads converge, and so the usage $p = m/N$, the erlangs carried per line, goes to zero. This linear growth is not inconsistent with the $N \log N$ orders mentioned earlier; the latter follow from combinatorial properties, or are the result of a different "limiting direction" in which p is bounded below away from zero. The above weak limit result is inappropriate for cases in which, as the network grows, each new customer is to supply a fixed number p_0 of added erlangs carried; this latter situation enjoins $N \log N$ growth. All the results about linear growth are given rigorous proofs, for the Markov process models adopted, by a passage from finite to infinite sources. In particular, no independence or other ad hoc assumptions are made to simplify the blocking estimate.

IV. COMPROMISES AND TRADE-OFFS

No matter what technology is used to build it, whether Strowger switches, crossbars, solid-state crosspoints, or time-division, the design of a telephone connecting network is inevitably a compromise between the competing criteria of cost and performance. Restricting attention to the traffic and operational aspects, it is nearly a truism that an overengineered network rich in switches will give unblemished service at an unacceptable cost, while one meagerly endowed with switches can only provide poor service at bargain cost. The trick the switching engineer must perform is to come up with designs that avoid these naive extremes.

The engineer's task is also affected by more subtle considerations, such as the following: the same pool of switching gear can be organized in an efficient way that is combinatorially optimal for connecting many pairs of customers in different patterns, i.e., realizing many assignments readily. Unfortunately, these efficient ways involve many stages and so are usually very complex and difficult to control, because putting up a call or removing one requires a lot of information and many decisions. Or the same gear can be hooked up in an inefficient, simple network of a few stages, which is easy to operate, but since it lacks combinatorial power it will have noticeably higher loss. Examples can be found in systems in which equipment is dedicated to handle certain geographically defined kinds of traffic (see Fig. 3).

Obviously, then, there are many trade-offs available to the designer between complexity, cost, control, and the various performance parameters, such as load and loss. Part of what traffic theory must provide, as a proper theoretical underpinning to network design, is an account of the options that are available (or unavailable) in the way of equipment, load, control and structural complexity, growth, and incurred loss. Such an account would describe the achievable regions in parameter space, some of the outer limits of possible designs, and the

achievable rates of growth of various indices of performance and complexity.

In the past such information for the large networks, which are of chief interest, has been sparse and difficult to obtain by means other than simulation and admittedly fuzzy theory. Thus, it is pertinent to have an account that retains its accuracy and force as the size of the network increases without bound. To give a specific example of this kind of information, we can put the following questions: For x , p , and ϵ specified positive numbers, are there arbitrarily large networks that have blocking probability at most ϵ , line usage at least p , and a number of crosspoints per terminal at most x ? (To define blocking, let us assume for definiteness that call arrivals are random from finite sources by pairs, and holding times exponential—a “usual” model, that no one can quarrel with.) If there are such networks, how complex are they? That is, how fast does their complexity grow as measured by number of stages, amount of data needed to select a route, etc.?

V. THE VALUE OF BLOCKING

It is widely believed among telephone switching engineers that a positive probability of blocking is worth a great deal of switching and control equipment. Probably for this good reason alone, the famous nonblocking networks first invented by Charles Clos⁴ have never been utilized on any but a small scale. Put another way, the canard says that to eliminate the last little bit of blocking will take an inordinate increase in equipment. But precisely what does the word “inordinate” mean here? Can we replace it by a mathematical function $E(b)$, which increases as the blocking b decreases, and whose interpretation is somehow that to achieve blocking b you need at least $E(b)$ in equipment? And how is $E(b)$ related to the structure of the network, to its size measured by number N of terminals on a side?

The remark that started the preceding paragraph is really only the tip of the iceberg: it is not exaggerating to claim that the mathematical basis of traffic engineering is the observation that very likely only a moderate number of telephone customers will want to talk to each other at the same time. The engineer provides enough lines, junctors, switches, trunks, etc., to take care of this overwhelmingly probable case, plus maybe a little extra for hedging his bets. The traffic theorist provides probabilistic models that make precise the meanings of “average,” “likely,” and “probable” in this setting. Of course the loads and losses the engineer seeks or achieves are subject to correction from customers’ complaints, public commissions, and supervisors. But no matter what the numbers may be, the principle of not having to provide for the very unlikely events is universally accepted by all of those concerned. This principle amounts to public acceptance of pos-

itive blocking probability for public telephony; it stops designs from passing the inevitable knees in the cost curves, beyond which costs grow very fast for each increment in desired performance.

Naturally, then, we are interested in quantitative and mathematical expressions of the above principle. One can sensibly ask how much switching equipment can actually be saved by allowing a blocking probability of $\epsilon > 0$. Where in the system can you (and should you) save it? In particular, how is allowing a blocking of ϵ related to the rate of growth of the switching gear incurring that blocking as the number N of terminals gets large? It is known⁵ that if $\epsilon = 0$ the needed gear in crosspoints is bounded above and below by

$$\text{const. } N \log N.$$

These results are purely combinatorial, and involve neither probability models nor traffic parameters. But is $N \log N$ still the right order of growth when $\epsilon > 0$, and the problem is posed in the context of our "usual" model, with a traffic parameter entering, such as the calling rate λ per idle inlet-outlet pair? The answer depends on just how λ varies relative to N as $N \rightarrow \infty$. If it is constant or decreases so slowly that the usage p is bounded away from zero, then $N \log N$ growth is necessary; if λN^2 remains constant, then for any $\epsilon > 0$, X need grow only linearly with N .

VI. THE KINDS AND ORIGINS OF LOSS

When we think about the probability of blocking in a network, it is useful to ask where and how it originates. At high levels of offered load, most of the loss incurred might be due primarily to frequent outright overloads of critical "bottlenecks," even without the occurrence of any combinatorial niceties such as mismatch. At lower levels of offered load, the fraction of loss owing to overloads is very small because the high occupancy states have very small probability, and most of the loss is due to mismatch in states of moderate occupancy and high total probability. It is not always possible to draw this distinction exactly in practice, nor is it necessary. It is important for theoretical purposes, though, because it suggests exactly analyzable large network structures (and models for them), in which the blocking is either all overload, or overload with certain simple kinds of mismatch. Such models can be used to study the trade-offs among the traffic and growth parameters; several are described in the latter half of this work.

VII. THE VIRTUE OF CONCENTRATION

Let it be agreed that for many terminals N and small calling rate λ per idle terminal-pair, the chance that many customers will want to

talk simultaneously is very small; if the principle of not providing for the very unlikely event is to be taken seriously, how should switching equipment be arranged? A standard and traditional method is to *concentrate* traffic.

The most efficient networks known today concentrate traffic from very many lightly loaded terminals into a heavily loaded central distribution core, in which link occupancy may run as high as 0.8 to 0.9, and then expand it back out in an inverse manner. The number of terminals entering the core is typically much smaller than the number of inlets, a feature that leads to a natural bottleneck or what we later call a "neck size." The blocking incurred can be thought of as being of two kinds, or arising from two sources: concentrator blocking, the inability of free inlets or outlets to get a line to the distribution core, and internal blocking in the core itself. Each of these sources of loss may in turn be due mostly to overload or to mismatch.

Thus, to reap the economic and operational values consequent on allowing blocking, the possible or maximum numbers of calls in progress at various places in the network are intentionally limited so as to save switching gear, the argument being as before that while having more calls in progress at these places is possible, it is so (or sufficiently) unlikely that there is no point in providing for it. We ask, how effectively can such limits curb the growth without impairing service?

Now the known nonblocking networks achieve their perfect operation by systematic *expansion*, the provision of more paths than can actually be used at one time; this is the antithesis of the concentrator-cum-distribution core idea usually used in practice. These nonblocking networks exhibit $N \log N$ growth, as they must. Since concentration is the opposite of expansion, it should lead to a saving in crosspoints. How big a saving is it? Especially, when can concentration reduce the order of growth to something slower than $N \log N$?

VIII. ASYMPTOTICS FOR LARGE NETWORKS

In seeking to answer some of the preceding questions, we shall examine the behavior of network parameters as the number N of terminals on a side becomes arbitrarily large. To fix ideas we begin with some thought-experiments that will lead to more specific questions.

Accepting for a moment the conventional wisdom that all efficient large networks use concentrating switches, we imagine a sequence of such networks with more and more terminals, and ask: What can happen to the efficiency of these networks as they grow? Is it possible to keep the loss below a specified amount without having the crosspoint count or the number of stages grow very fast? Are there large networks which, though they may not be in the running as red-hot field designs

for a particular technology, nevertheless are of interest because their load, loss, and cost can be easily and accurately calculated or estimated? If there are such networks, what combinatorial features account for the ease of calculation? Where are “most” of the crosspoints, in the concentration stages, or in the distribution core?

IX. LIMIT DIRECTIONS

Needless to say, some care must be taken in carrying out the analyses needed to answer these questions. For most purposes, it is enough to make more exact the way in which the traffic and performance parameters are to vary as N grows; usually some group or function of them is constrained to stay in a given set. Such constraints define “directions” in which limits or other asymptotics are being sought, and they provide useful ways of looking at the performance of very large networks. Two such directions, leading to very different growth rates, will be of interest here.

9.1 The “weak” limit

For example, we can let the offered load per idle pair λ get small and N get big, so that $a = \lambda N^2$ is a fixed constant. This amounts to letting the process of attempted calls become Poisson with rate a ; the corresponding limit process is sometimes called a “passage from finite to infinite sources,” and in traffic theory is often associated with familiar notions such as the Poisson approximation to the binomial distribution. We shall show by examples that some interesting limits of this kind exist and can be evaluated, leading to functions and concepts well-known in traffic theory, such as the Erlang loss $E(c, a)$. Such calculations lead to information about the growth rate of cost and complexity for large networks that have specified load and loss.

9.2 The strong constraint

A very different condition, to be used later, is that the usage $p = m/N$ of our sequence of growing networks be bounded away from zero. It says roughly that each new customer, as N grows, adds a fixed amount of carried load to the network, at least. This condition, incompatible with the “weak” limit, leads to $N \log N$ growth in crosspoints, and is not, as far as we know, associated with any limits in distribution, the way the weak limit is. That is why we call it a condition and not a limit. It is physically natural for networks to grow in this manner, and so this is an important condition to consider.

X. NETWORK STATES

We shall use a model for the structural and combinatorial aspects of

a connecting network. This model arises by considering the network structure to be given by a graph G whose vertices are the terminals of the network, and whose edges represent crosspoints between terminals by pairs, with some of the terminals designated as inlets or outlets. Calls in the network are described by paths on G from an inlet to an outlet. Thus, a connecting network ν is a quadruple $\nu = (G, I, \Omega, S)$, where G is a graph depicting networking structure, I is the set of vertices of G which are inlets, Ω is the set of outlets, and S is the set of permitted or physically meaningful states. It is possible that $I = \Omega$ (one-sided network), that $I \cap \Omega = \phi$ (two-sided network), or that some intermediate condition obtain, depending on the "community of interest" aspects of the network ν . Variables w, x, y , and z at the end of the alphabet denote states, while u and v denote a typical inlet and a typical outlet, respectively.

A possible state x can be thought of as a set of disjoint chains on G , each joining I to Ω . Not every such set of chains need represent a state in S : wastefully circuitous chains may be excluded from S . The set S is partially ordered by inclusion \leq , where $x \leq y$ means that state x can be obtained from state y by removing zero or more calls. It is reasonable that if y is a state and x results from y by removal of some chains, then x should be a state too, i.e., S should be closed under "hangups." It can be seen from this requirement that the set S of permitted states has the structure of a semilattice, that is, a partially ordered system whose order relation is definable in terms of a binary operation \cap that is idempotent, commutative, and associative, by the formula $x \leq y$ iff $x = x \cap y$. Here for $x \cap y$ we can simply use literal set intersections: $x \cap y$ is exactly the state consisting of those calls and their respective routes that are common to x and y .

An *assignment* is a specification of what inlets are to be connected to what outlets. The set A of assignments can be represented as the set of all fixed-point-free correspondences from subsets of I to Ω . The assignments form a semilattice in the same way that the states do, and A is related to S as follows: call two states x, y in S equivalent as to assignment, written $x \sim y$, iff all and only those inlets $u \in I$ are connected in x to outlets $v \in \Omega$, which are connected to the same v in y , though possibly by different routes. The realizable assignments can then be identified with the equivalence classes of states under \sim , and there is a natural map $\gamma: S \rightarrow A$, the projection that carries each state x into the assignment $\gamma(x)$ it realizes, i.e., the equivalence class it belongs to under \sim .

With x and y states such that $x \geq y$, it is convenient to use $x - y$ to mean the state resulting from x by removing from x all the calls in y . Similarly, with a and b assignments such that $a \geq b$, we use $a - b$ to mean the assignment resulting from a by dropping all the connections

intended in b . Note that here $x - y$, $a - b$ have their usual set-theoretic meaning.

It can now be seen that the map γ is a semilattice homomorphism of S into A , with the properties:

$$\begin{aligned} x \geq y & \Rightarrow \gamma(x) \geq \gamma(y) \\ x \geq y & \Rightarrow \gamma(x - y) = \gamma(x) - \gamma(y) \\ \gamma(x \cap y) & \leq \gamma(x) \cap \gamma(y) \\ \gamma(x) = \phi & \Rightarrow x = 0 = \text{zero state, with no calls up.} \end{aligned}$$

Not every assignment need be realizable by some state of S . Indeed, it is common for practical networks to realize only a vanishing fraction of the possible assignments, and the networks that do realize every assignment, the so-called *rearrangeable* networks, have been the objects of substantial theoretical study. Thus, the image set $\gamma(S)$ of realizable assignments is typically much smaller than the set A in which it is embedded. A *unit* assignment is, naturally, one that assigns exactly one outlet to some inlet, and it corresponds to having just one call in progress. It is convenient to identify calls c and unit assignments, and to write $\gamma(x) \cup c$ for the larger assignment consisting of $\gamma(x)$ and the call c together, with the understanding, of course, that c is "new in x " in the sense that neither of its terminals is busy in x .

We denote by A_x the set of states that are immediately above x in the partial ordering \leq of S , and by B_x the set of those that are immediately below. Thus,

$$\begin{aligned} a_x &= \{\text{states reachable from } x \text{ by adding a call}\} \\ B_x &= \{\text{states reachable from } x \text{ by hangup}\}. \end{aligned}$$

For c new in x , let $A_{cx} = A_x \cap \gamma^{-1}[\gamma(x) \cup c]$; A_{cx} is the subset of states of A_x that could result from x by putting up the call c , because $\gamma^{-1}\gamma(y)$ is precisely the equivalence class of y under \sim . If A_{cx} is empty then we say c is blocked in x : there is no $y \in A_x$ that realizes the larger assignment $\gamma(x) \cup c$. It can be seen that with F_x the set of new calls of x that are not blocked, the family $\{A_{cx}, c \in F_x\}$ forms the partition of A_x induced by equivalence \sim .

XI. ROUTING OF CALLS

We shall use a routing matrix $R = (r_{xy})$ as a convenient formal description of how routes are chosen for calls. The class of routing matrices, R , can be described thus: for each $x \in S$ let Π_x be the partition of A_x induced by the relation \sim of "having the same calls up", or satisfying the same assignment of inlets to outlets; it can be seen that Π_x consists of exactly the sets A_{cx} for c free and not blocked

in x ; for $Y \in \Pi_x$, r_{xy} for $y \in Y$ is to be a probability distribution over Y , that is $r_{sx} \geq 0$ and $\sum_{y \in Y} r_{xy} = 1$; r_{xy} is to be 0 in all other cases.

The interpretation of the routing matrix as a method of choice is to be this: any $Y \in \Pi_x$ represents all the ways in which a particular call c (free and not blocked in x) *could* be completed when the network is in state x ; for $y \in Y$, r_{xy} is the chance (or fraction of times) that if call c arises in state x it will be completed by being routed in the network so as to take the system to state y . The distribution $\{r_{xy}, y \in Y\}$ indicates how the calling rate owing to c is to be spread over the possible ways of putting up this call. Evidently, such a description of routing could be made time-dependent, and extended to cover refusal of unblocked calls as an option; we do not consider these possibilities here. The problem of choosing an optimal routing matrix R has been worked on at some length.

XII. STOCHASTIC MODEL

We now recall² a stochastic model for the traffic offered to a network. A Markov stochastic process x_t taking values on S can be based on these simple probabilistic and operational assumptions:

(i) Holding times of calls are mutually independent variates, each with the negative exponential distribution of unit mean.

(ii) If u is an inlet idle in state $x \in S$, and $v \neq u$ is any outlet, there is a conditional probability $\lambda h + o(h)$, $\lambda > 0$, as $h \rightarrow 0$, that u attempts a call to v in $(t, t + h)$ if $x_t = x$.

(iii) A routing matrix $R = (r_{xy})$ is used to choose routes, as follows: If $c = \{(u, v)\}$ is a call free and not blocked in x , then the fraction of times that the system passes from x to $y \in A_{cx}$ if c arises when $x_t = x$ is just r_{xy} .

(iv) Blocked calls and calls to busy terminals are declined, with no change of state.

It is convenient to collect these assumptions into a transition rate matrix $Q = (q_{xy})$, the generator of x_t ; this matrix is given by

$$q_{xy} = \begin{cases} 1 & \text{if } y \in B_x \\ \lambda r_{xy} & \text{if } y \in A_x \\ -|x| - \lambda s(x) & \text{if } y = x, \text{ with } s(x) = |F_x| \\ 0 & \text{otherwise,} \end{cases}$$

and the associated statistical equilibrium (or state) equations take the simple form

$$[|x| + \lambda s(x)]p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y r_{yx} \quad x \in S,$$

where $\{p_x, x \in S\}$ is the asymptotic distribution of x_t . Here $|x|$ denotes the number of calls in progress in x , and $s(x)$ is the number of

unblocked idle inlet-outlet pairs in x , the possible “successes” in x ; note that $s(x) = |F_x|$.

XIII. PARAMETERS OF INTEREST FOR DESIGN AND ENGINEERING

We shall frequently use about a dozen basic parameters, characteristic of the operating network, and important for these reasons: They describe load, cost, and performance, or they can be measured readily, or they arise naturally in the associated traffic theory and are convenient for calculations and asymptotic analyses. For two-sided networks ν , these parameters are the following:

- λ = calling rate per idle inlet-outlet pair
- N = number of terminals (inlets, or outlets) on each side
- bl = probability of blocking
- m = carried load = expected number of calls in progress
- p = usage = m/N = erlangs carried per terminal
- σ = standard deviation of number of calls in progress
- X = total number of crosspoints
- s = number of stages (if ν is arranged in stages)
- $\alpha = \lambda N^2$ = total offered load when everyone is idle
- $w = \max_{x \in S} |x|$ = maximum possible number of calls in progress
- n = “neck size,” defined for ν arranged in stages separating junctor groups T_1, T_2, \dots, T_{s+1} , as the cardinality of the smallest T_i .

Remark: The parameter $\alpha = \lambda N^2$ is a convenient abbreviation for total offered load, especially for certain weak “large network” asymptotics for which λN^2 is held constant as $\lambda \rightarrow 0$ and $N \rightarrow \infty$.

Remark: The ratios w/N and n/N are rough *global* measures of concentration, global because there could be, for example, remote local concentrators with a concentration ratio different from each of these. Clearly, $w \leq n$, when both w and n are defined.

Notation: We write $X(\nu)$, $p(\nu)$, etc, whenever it is necessary to express the dependence of a parameter on the network ν .

XIV. THE PARAMETER SURFACE

In the early¹ applications of traffic theory to trunking problems, a central role was played by Erlang’s loss formula, which depended on two parameters, the load α , and the number c of trunks. For connecting network studies, though, to take into account at least the size of the network and the “finite source” effect, if not other network features, a modification of Erlang’s formula is more suitable. (The finite source effect is a recognition that busy terminals generate no traffic.) Such a formula has been derived in earlier work.² We shall exhibit many

useful results that follow from it, or from it together with reasonable but special hypotheses, such as the property of a network that it is made of square switches, or is arranged in stages, or provides full access.

For a two-sided network with N terminals on a side, the load, loss, load deviation, and rate parameter, λ , are related by the following formula:

Generalization of Erlang's formula to networks:

$$1 - bl = \frac{1}{\lambda} \frac{m}{(N - m)^2 + \sigma^2} \quad (1)$$

$$\alpha(1 - bl) = \frac{m}{(1 - p)^2 + (\sigma/N)^2}.$$

Proof: $1 - bl$ is the fraction of attempted calls that are not blocked. By the law of large numbers, this fraction is the rate of successful attempts divided by the total rate of attempts, both in equilibrium. For a state x let $s(x)$ be the number of inlet-outlet pairs (u, v) that are not blocked in x ; " $s(x)$ " stands for "successes in x ." The success rate is then $\lambda \sum_{x \in S} p_x s(x)$ and the total attempt rate $\lambda \sum_{x \in S} p_x (N - |x|)^2$.

However, in equilibrium, the rate in equals the rate out, so

$$\lambda \sum_{x \in S} p_x s(x) = \sum_{x \in S} p_x |x| = m.$$

Evidently the total attempt rate can be written as $\lambda[(N - m)^2 + \sigma^2]$, and the general Erlang formula is proved.

Remarks: The gist of the Erlang formula is that six of the parameters of interest cannot assume arbitrary values but must lie on a surface described by a simple rational function. It is apparent that similar but more complex formulas can be proved for one-sided networks, or two-sided ones with different numbers of terminals on each side; we shall not consider these, because the basic ideas are the same. Note that blocking, a complicated quantity in the model, is determined solely by the first two moments of the number of calls in progress. Also, if N and λ vary so that $(1 - p)^2 + (\sigma/N)^2$ approaches 1, the formula approaches the exact form $m = \alpha(1 - bl)$ it has in the "true" Erlang case.¹

XV. SOME TECHNICAL RESULTS

The generalized Erlang formula has many useful consequences. Some of them are summarized in the next few results, all of which are additional relationships among the engineering parameters.

Lemma: $\frac{\sigma^2}{N^2} < p - p^2$. (2)

Proof: Clearly, $E|x_t|^2 = \sum_{x \in S} p_x |x|^2 \geq Nm$. Hence,

$$\sigma^2 = E|x_t|^2 - (E|x_t|)^2 \leq Nm - m^2.$$

Lemma: $\frac{m}{a(1 - bl)} \leq 1 - p$. (3)

Proof: Lemma (2) implies that

$$p^2 - 2p + 1 + \frac{\sigma^2}{N^2} < 1 - p.$$

By (1) the left-hand side is $m/a (1 - bl)$.

Representation of N :

$$N = \frac{m + \sqrt{m^2 - \left[1 - \frac{m}{a(1 - bl)}\right] (m^2 + \sigma^2)}}{1 - \frac{m}{a(1 - bl)}} = \frac{m(1 + \sqrt{1 - p})}{1 - \frac{m}{a(1 - bl)}}, \quad (4)$$

where

$$p = \left[1 - \frac{m}{a(1 - bl)}\right] \left[1 + \frac{\sigma^2}{m^2}\right]$$

and

$$0 < p < 1.$$

Proof: N is one of the quadratic roots

$$\frac{m(1 \pm \sqrt{1 - p})}{1 - \frac{m}{a(1 - bl)}}. \quad (5)$$

To show that the plus branch is the right one we note that the quadratic (in y) function

$$y^2 \left[1 - \frac{m}{a(1 - bl)}\right] - 2my + m^2 + \sigma^2 \quad (6)$$

equals $m^2 + \sigma^2$ at $y = 0$, and has a negative minimum at

$$y^* = \frac{m}{1 - \frac{m}{a(1 - bl)}}$$

This value is a minimum because the second derivative is

$$2 \left[1 - \frac{m}{a(1 - bl)} \right] > 2p$$

by Lemma (3). This value is negative because it is

$$m^2 + \sigma^2 - \frac{m^2}{1 - \frac{m}{a(1 - bl)}} \quad (7)$$

and the Erlang formula (1) implies that

$$\sigma^2 + m^2 = 2mN - N^2 \left[1 - \frac{m}{a(1 - bl)} \right]. \quad (8)$$

Substituting this into (7) we see that the value is

$$-N^2 \left[1 - \frac{m}{a(1 - bl)} \right] + 2mN - \frac{m^2}{1 - \frac{m}{a(1 - bl)}}$$

clearly negative. Thus y^* separates the two real roots of (6). However, Lemma (3) implies that $N > y^*$, so N must be given by the plus sign in (5). It is obvious that p is positive; to show $p < 1$ we divide (8) by m^2

and multiply by $1 - \frac{m}{a(1 - bl)}$ to get

$$p = 1 - \left\{ 1 - \frac{N}{m} \left[1 - \frac{m}{a(1 - bl)} \right] \right\}^2 < 1,$$

which incidentally strengthens Lemma (3) to

$$p < 1 - \frac{m}{a(1 - bl)} < 2p.$$

Lemma: If v is such that, at most, w calls can be in progress, then for $w < N$

$$\begin{aligned} (i) \quad & bl \geq 1 - \frac{m}{\lambda(N - w)^2} \\ (ii) \quad & w \geq N(1 - \sqrt{1 - p}) \\ (iii) \quad & p \leq \frac{2w}{N} - \left\{ \frac{w}{N} \right\}^2. \end{aligned} \quad (9)$$

Proof: In this situation the average calling rate is greater than or equal

to $(N - w)^2\lambda$, so (i) follows from the generalized Erlang formula (1). Thus also

$$\frac{m}{\lambda N^2[(1-p)^2 + \sigma^2 N^{-2}]} \leq \frac{m}{\lambda(N-w)^2} (N-w)^2 \leq N^2[(1-p)^2 + \sigma^2 N^{-2}] = N^2 \frac{m}{a(1-bl)}.$$

But by Lemma (3),

$$\frac{m}{a(1-bl)} \leq 1-p$$

and thus (ii) and (iii) via

$$N-w \leq N \sqrt{1-p}.$$

The same argument proves

$$w \geq N[1 - \sqrt{(1-p)^2 + \sigma^2 N^{-2}}].$$

XVI. NONCONCENTRATING NETWORKS GROW LIKE $N \log N$ EVEN IN THE WEAK LIMIT

One way to display the value of concentrating traffic is to show how bad things are when you do not do it. We shall look at a large class of practical networks that have no concentration, and show that to all intents and purposes, these networks grow like $N \log N$ for N terminals. In particular, without concentration, these networks have no way of trading off nonzero or even substantial blocking (up to some $\epsilon > 0$) for slow growth, i.e., slower than $N \log N$, or constant $\times N \log N$ with a constant that depends on usage p . Especially, these networks have $N \log N$ growth even in the weak limit $a = \lambda N^2 = \text{constant}$; we show later that, in this special case, linear growth is achievable if concentration is used, even with blocking kept less than ϵ .

Now the prototypical network without concentration is one that is made of *square* switches arranged in stages, and these constitute the class we consider. Since engineers are primarily interested in networks with a high degree of symmetry, which "look the same" from any terminal within a junctor group, or from any inlet, or outlet, we shall restrict attention to networks ν with these properties:

(i) ν has $s \geq 1$ stages

(ii) The switches in each stage are square, and identical.

Note that the number of stages is not fixed, and that different stages may have different numbers of switches; also, the interconnection patterns between the stages (in old terminology, the *cross-connect fields*) can represent arbitrary N -permutations.

Theorem: Let n_1, n_2, \dots, n_s be s divisors of N , possibly with repetitions, and consider a network ν with N/n_i $n_i \times n_i$ switches in its i th stage, and blocking $bl < \epsilon$. Then its crosspoint count $X(\nu)$ satisfies

$$X(\nu) \geq N \sum_{i=1}^s n_i \geq N \log N + N(s + \log(1 - \epsilon)) - a. \quad (10)$$

Remark: The number of outlets reachable from an inlet is at most $\prod_{i=1}^s n_i$. If every inlet can reach every outlet, then this product exceeds $N - 1$, and in this "usual" case

$$\sum_{i=1}^s n_i \geq s + \log \prod_{i=1}^s n_i \geq s + \log N \quad (11)$$

and the conclusion of the theorem follows. In the opposite "unusual" case, $\prod_{i=1}^s n_i < N$, some calls are permanently blocked, and the network does not provide full access. Thus, one point of the theorem is that even in this combinatorially poor situation, blocking cannot save more than a linearly growing number of crosspoints when $a = \lambda N^2$ is constant or grows at most linearly, so that $X(\nu) = O(N \log N)$.

Proof: Only the "unusual" case $N > \prod_{i=1}^s n_i$ need be considered. Let $s(x)$ be the number of possible "successes" in x , i.e., of inlet-outlet pairs idle and not blocked in state x , so that

$$(N - |x|)^2 - s(x)$$

is the number of idle inlet-outlet pairs that cannot be connected in x . In the unusual case an idle inlet cannot reach at least $N - \prod_{i=1}^s n_i$ outlets. In state x , at most $|x|$ of these unreachable outlets are busy; thus there are at least

$$N - \prod_{i=1}^s n_i - |x|$$

idle outlets with no path to our test inlet. This being true for each idle inlet, and there being $N - |x|$ idle inlets, the number β_x of blocked idle inlet-outlet pairs in state x satisfies

$$\beta_x \geq (N - |x|) \left(N - \prod_{i=1}^s n_i - |x| \right),$$

or since $(N - |x|)^2 = s(x) + \beta_x$,

$$s(x) \leq (N - |x|) \prod_{i=1}^s n_i.$$

Averaging this inequality with respect to the equilibrium state probabilities $\{p_x, x \in S\}$, and noting that $m = \sum_{x \in S} |x| p_x = \lambda \sum_{x \in S} s(x) p_x$, we find

$$\frac{m}{\lambda} \leq (N - m) \prod_{i=1}^s n_i$$

$$\prod_{i=1}^s n_i \geq \frac{Nm}{a(1 - p)}.$$

Therefore,

$$X(\nu) = N \sum_{i=1}^s n_i$$

$$\geq N \left(s + \log \prod_{i=1}^s n_i \right)$$

$$\geq Ns + N \log N + N \log \frac{m}{a(1 - p)}.$$

To find a suitable lower bound on the last term we use the basic generalization (1) of Erlang's formula:

$$1 - bl = \frac{m}{\lambda(N - m)^2 + \lambda\sigma^2} = \frac{m}{a[(1 - p)^2 + \sigma^2 N^{-2}]}$$

$$\leq \frac{m}{a(1 - p)^2}.$$

Since the blocking bl is less than ϵ , one finds

$$\log(1 - \epsilon) \leq \log(1 - bl) \leq \log \frac{m}{a(1 - p)} - \log(1 - p)$$

$$X(\nu) \geq N \log N + sN + N \log(1 - p) + N \log(1 - \epsilon).$$

Next we argue that, as in Lemma (2),

$$\sigma^2 N^{-2} \leq p - p^2$$

$$(1 - p)^2 + \sigma^2 N^{-2} \leq 1 - p,$$

and so by the Erlang formula (1) again,

$$p = \frac{1 - bl}{N} a[(1 - p)^2 + \sigma^2 N^{-2}] \leq \frac{1 - p}{N} a \leq \frac{a}{N + a}$$

whence

$$\log(1 - p) > \log N - \log(N + a) \geq -\frac{a}{N}$$

and the proof is complete.

XVII. GROWTH OF NETWORKS ARRANGED IN STAGES

Almost all the connecting networks used in practice are *made of*

stages, or arranged in stages. This property means that the terminals of the network are partitioned into disjoint sets T_1, T_2, \dots, T_{s+1} , which are simply ordered as indicated; the first set T_1 consists of the inlets, the next $s - 2$ sets are internal junctors, and the last set T_{s+1} consists of the outlets; crosspoints are placed only from terminals in a given set T_i to terminals in the next set T_{i+1} in the ordering (see Fig. 1). Thus every call in progress is a path from an inlet to an outlet that passes through each set T_i once in the order specified. The crosspoint pattern between successive sets is called a stage of switching, and is representable as a bipartite graph. The sets T_i need not all have the same numbers of terminals; if they do not, there is expansion or concentration. The size of the smallest T_i is called the "neck size," and is of course an upper bound on the number of calls in progress:

$$|x| \leq n = \min_{1 \leq i \leq s+1} |T_i|.$$

We shall restrict attention to *symmetric* networks, in which the network looks the same to every terminal in a given T_i , $i = 1, \dots, s + 1$. A network provides *full access* if every inlet-outlet pair can be connected by a path through the network, with no doubling back allowed. Full access is a natural convenient condition that greatly simplifies arguments, but it is not necessary for proving $N \log N$ growth.

Theorem: Let ν be a symmetric network with N inlets (& outlets), with neck size n , providing full access through s stages. Then the crosspoint count $X(\nu)$ satisfies

$$X(\nu) \geq en \log N \quad e = 2.71828 \dots \quad (12)$$

Proof: Let n_i , $i = 1, \dots, s$, be the number of crosspoints in stage i connected to a junctor between stages i and $i + 1$. By symmetry this number is the same for all such junctors or terminals. If stage i is represented by a bipartite graph, n_i is the degree of each "input" vertex. Thus, if the neck size is n , then stage i has at least nn_i crosspoints, and so

$$X(\nu) \geq n \sum_{i=1}^s n_i.$$

Since ν provides full access it must be true that

$$N \leq \prod_{i=1}^s n_i;$$

for an inlet can reach no more than $\prod_{i=1}^s n_i$ outlets; so if N exceeded this product there would be some it could not reach. Hence, by the inequality linking arithmetic and geometric means,

$$\begin{aligned}
X(\nu) &\geq n \sum_{i=1}^s n_i \\
&\geq ns \left(\prod_{i=1}^s n_i \right)^{\frac{1}{s}} \\
&\geq nsN^{\frac{1}{s}}.
\end{aligned}$$

But $sN^{\frac{1}{s}}$, viewed as a function of a real variable s , has a unique minimum at $s = \log N$, so that

$$X(\nu) \geq n \log N N^{\frac{1}{\log N}} = en \log N, \quad e = 2.71828 \dots$$

Remark: If the neck size is N , as it is for networks made of square switches, then for symmetric ν providing full access

$$X(\nu) \geq e N \log N.$$

Lemma: If ν has neck size $n < N$, then

$$n \geq N \left\{ 1 - \left[\frac{m}{a(1-bl)} \right]^{1/2} \right\} \geq N \{1 - \sqrt{1-p}\}. \quad (13)$$

This result links the neck size n to the performance parameters m and bl and to the traffic parameter a . The second inequality is remarkable in involving only the line usage p ; the higher p is to be, the closer the neck size must be to N .

Proof: If ν has neck size n , then at most n calls can be in progress at a time, so that $N - |x| \geq N - n$, and by the Erlang formula (1)

$$\begin{aligned}
1 - bl &= \frac{m}{\lambda \sum p_x (N - |x|)^2} \leq \frac{m}{\lambda (N - n)^2} \\
(N - n)^2 &\leq N^2 \frac{m}{a(1-bl)}.
\end{aligned}$$

The second inequality follows from Lemma (3); it leads at once to this basic result:

Theorem: If ν is a symmetric network arranged in stages, providing full access and with each inlet carrying p erlangs, then

$$X(\nu) \geq e(1 - \sqrt{1-p})N \log N. \quad (14)$$

Proof: (12) and (13).

Discussion: This inequality says that any symmetric network providing full access has const. $N \log N$ crosspoints, where the constant depends only on the line usage, no matter what blocking is incurred. Using the

first inequality in (13) gives a larger constant, now dependent on loads (offered and carried) and blocking.

Theorem: Let ν be a network arranged in stages, not necessarily providing full access. Then

$$X(\nu) \geq e(1 - \sqrt{1 - p})[\frac{1}{2}N \log N - \frac{1}{2}N \log \lambda + N \log(1 - bl) + N \log p]. \quad (15)$$

Proof: As in the proof of Theorem (12) we find

$$X(\nu) \geq ns \left(\prod_{i=1}^s n_i \right)^{\frac{1}{s}}. \quad (16)$$

Next, the averaging argument of Theorem (14) gives

$$\prod_{i=1}^s n_i \geq \frac{Nm}{\alpha(1 - p)} = \frac{p}{\lambda(1 - p)}.$$

For $b > 0$, $sb^{\frac{1}{s}}$ assumes a unique minimum at $s = \log b$, and $(b)^{\frac{1}{\log b}} = e = 2.71828, \dots$, so by (16),

$$\begin{aligned} X(\nu) &\geq ne \log \frac{p}{\lambda(1 - p)} \\ &\geq e(1 - \sqrt{1 - p})N[\log p - \log \lambda - \log(1 - p)]. \end{aligned} \quad (17)$$

The generalized Erlang formula (1) can be put in the forms

$$\lambda N(1 - bl)[(1 - p)^2 + \sigma^2 N^{-2}] = p \quad (18)$$

$$\lambda N(1 - bl)(1 - p)^2 + (1 - p) - 1 + \lambda N(1 - bl)\sigma^2 N^{-2} = 0. \quad (19)$$

The second form is a quadratic equation for $1 - p$ whose solution, picking the plus branch, is

$$1 - p = \frac{\sqrt{1 + 4\lambda N(1 - bl)[1 - \lambda(1 - bl)\sigma^2 N^{-1}]} - 1}{2\lambda N(1 - bl)}.$$

By (18) above, the quantity (factor) $1 - \lambda(1 - bl)\sigma^2 N^{-1}$ under the square root is equal to

$$1 - p \frac{\sigma^2}{(N - m)^2 + \sigma^2}$$

and lies strictly between 0 and 1. Therefore,

$$1 - p < \frac{\sqrt{1 + 4\lambda N(1 - bl)} - 1}{2\lambda N(1 - bl)}.$$

Let $y = 2\lambda N(1 - bl)$ for short, so that

$$1 - p < \frac{\sqrt{1 + 2y} - 1}{y}.$$

Hence,

$$\begin{aligned} \log(1 - p) &< \log \frac{\sqrt{1 + 2y} - 1}{y} = \log \frac{2}{\sqrt{1 + 2y} + 1} \\ -\log(1 - p) &> \log(\sqrt{1 + 2y} + 1) - \log 2 \\ &> \frac{1}{2} \log(1 + 2y) - \log 2 \\ &> \frac{1}{2} (\log 4 + \log \lambda + \log N + \log(1 - bl)) - \log 2. \end{aligned}$$

Returning now to formula (17) we find

$$X(\nu) \geq e(1 - \sqrt{1 - p})N[\log p - \frac{1}{2} \log \lambda + \frac{1}{2} \log N + \frac{1}{2} \log(1 - bl)].$$

Remark: Theorem (15) shows that full access is not necessary for $N \log N$ growth; it just makes the constant bigger and the argument simpler.

Theorem: Let ν_N be a sequence of networks on N terminals arranged in stages, and such that

- (i) $p(\nu_N) \geq p_0 > 0$
- (ii) $bl(\nu_N) \leq \epsilon$
- (iii) $\lambda(\nu_N)$ is bounded.

Then as $N \rightarrow \infty$

$$X(\nu_N) \geq \frac{e}{2} (1 - \sqrt{1 - p_0})N \log N + O(N).$$

Proof: Immediate from (15).

Remarks: Theorems (15) and (20) of course apply also to the networks made of stages of square switches considered in Theorems (10) and (11). However, it should be noted that the possible traffic asymptotics in the two theorems are different, although they might overlap. In (11) $\alpha = \lambda N^2$ grows at most linearly, while in (20) it grows at least linearly; in (11) $\alpha = \lambda N^2$ could be identically a constant (the weak limit case), so that λ and p both go to zero, and $X(\nu)$ grows like $N \log N$ instead of linearly as it might [see Theorem (24)]; in (20), on the other hand, p is bounded away from zero, hence λN is also, so $\alpha = \lambda N^2$ increases at least linearly, and $X(\nu)$ grows like constant $\times N \log N$, with constant depending on the lower bound for p . The point is that the absence of concentration exemplified in the square switches compels $N \log N$ growth even in the weak limit ($\alpha = \lambda N^2$ constant, p vanishing), while

in general if concentration is allowed it takes the “strong condition” $p \geq p_0 > 0$ to force $N \log N$ growth.

XVIII. ANALYZABLE LARGE NETWORKS

We turn now to the study of three simple patterns or structures for networks with concentration. Interest in them arises from the fact that their loads, losses, and complexity can be calculated or rigorously bounded for arbitrarily large values of N . Most of them embody features, such as frames and concentrators, which are familiar in telephone network design, and some provide tenuous links to previous approximate blocking formulas based on independence assumptions. These formulas suggest inequalities stating that certain natural “blocking polynomials” are in fact upper bounds on the probability of blocking; their proof or disproof has eluded us so far, but they are worth mentioning nevertheless.

XIX. CENTRAL BUSES CONCEPT

A useful extreme case is a network that, like the trunk group, has no blocking states until a certain number c of calls in progress is reached, at which point *all* calls are blocked. One way to build such a network is to concentrate both the N inlets and the N outlets down to c terminals in a nonblocking way, and then to put a c -by- c nonblocking network in between, as shown in Fig. 5. A better way results when we note that the central network is superfluous; all you need are c *central buses* with “expanding” networks on each side such that any idle terminal can reach an idle bus. An arrangement of this kind is shown in Fig. 2, in which each bus has an appearance on every (inlet or outlet) subnetwork; when these are nonblocking, a theoretically useful solvable case results. We call such networks *central bus networks*.

Remark: Clearly, the central bus idea for networks springs right out of the idea that in a large network with lightly loaded lines, only a

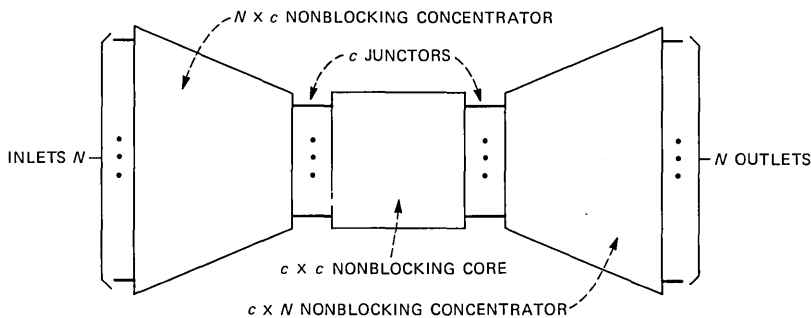


Fig. 5—Network nonblocking up to c calls.

moderate number of customers will be talking. The thought is this: if the low and moderate occupancy states have all the probability, let's use all our switching gear to make *them* as nonblocking as possible, and just ban the unlikely high occupancy states altogether. Accordingly, the designer guesses or calculates what that moderate occupancy might be, on the average, and provides some larger number c of central buses, with nonblocking access for everyone.

Remark: The central bus network is also a good candidate for the best disposition of a fixed number X of crosspoints at very low traffic λ . For it is known² that

$$\text{blocking} = \text{const. } \lambda^m + o(\lambda^m) \quad \text{as } \lambda \downarrow 0,$$

where

$$m = \min_{x \in S} \{ |x| : \text{some call is blocked in } x \}.$$

Thus, of all networks made out of X crosspoints, the ones for which m is largest will have asymptotically least blocking at low traffic, as $\lambda \rightarrow 0$.

For traffic purposes, the number of calls in progress is an adequate notion of state for central bus networks. Under our assumptions, the equilibrium probabilities p_n of n calls up satisfy

$$p_n = p_0 \frac{\lambda^n}{n!} (N - n + 1)^2 (N - n + 2)^2 \dots N^2$$

and so the blocking is

$$bl = \frac{\frac{\lambda^c}{c!} (N - c)^2 (N - c + 1)^2 \dots N^2}{N^2 + \sum_{j=1}^c \frac{\lambda^j}{j!} (N - j)^2 (N - j + 1)^2 \dots N^2}.$$

We introduce the parameter $a = \lambda N^2$ by writing this as

$$bl = \frac{\frac{a^c}{c!} \left(1 - \frac{c}{N}\right)^2 \left(1 - \frac{c-1}{N}\right)^2 \dots \left(1 - \frac{1}{N}\right)^2}{1 + \sum_{j=1}^c \frac{a^j}{j!} \left(1 - \frac{j}{N}\right)^2 \left(1 - \frac{j-1}{N}\right)^2 \dots \left(1 - \frac{1}{N}\right)^2}. \quad (21)$$

If we extend each product in the denominator all the way up to c , and replace the 1 by the product in the numerator, we increase the formula to exactly the Erlang loss function $E(c, a)$. Thus,

$$bl < E(c, a). \quad (22)$$

A similar argument shows that the traffic carried, m , satisfies

$$m < a[1 - E(c, a)]. \quad (23)$$

Thus, for central bus networks the load and loss are bounded above by the corresponding Erlang load and loss for c trunks and incoming traffic a .

XX. LINEAR GROWTH IS POSSIBLE IN THE WEAK LIMIT

The central bus concept also leads to an asymptotic estimate of what is possible in the way of network growth. We prove

Theorem: For every $\epsilon > 0$, there is an integer c , and a sequence ν_N of networks on N terminals with c central buses such that as $N \rightarrow \infty$ with $a \equiv \text{constant}$

$$\begin{aligned} (i) \quad & bl(\nu_N) < E(c, a) < \epsilon, \quad bl(\nu_N) \rightarrow E(c, a) \\ (ii) \quad & m(\nu_N) < a[1 - E(c, a)], \quad m(\nu_N) \rightarrow a[1 - E(c, a)] \\ (iii) \quad & X(\nu_N) \leq 136N \log_2 c + O(c) \\ (iv) \quad & s(\nu_N) \leq 4(1 + \log_2 c). \end{aligned} \quad (24)$$

Remarks: This result says that there exist arbitrarily large networks with specified blocking whose growth in crosspoints is linear (with slope dependent on blocking), whose complexity in number of stages is logarithmic, and whose load approaches a constant. Here the growth in "size" N is accompanied by a diminution in the offered load λ per idle pair, according to $\lambda N^2 \equiv a$, a constant, in a natural passage from finite to infinite sources. Because of the way these networks will be defined, they will be at the "combinatorially efficient, hard to control" end of the trade-off spectrum.

Proof of Theorem (24): Given $\epsilon > 0$, choose c to be the smallest integer such that $E(c, a) < \epsilon$, and construct a sequence of networks ν_N with N terminals on a side and c central buses, with nonblocking access to an idle bus from each side. Property (i) follows from (21), (22); (ii) is a result of (23) and the general Erlang formula (1); we have

$$N = \frac{m(1 + \sqrt{1 - \rho})}{1 - \frac{m}{a(1 - bl)}}, \quad (4)$$

where m is the load, $a = \lambda N^2$, bl = blocking, and $\rho \in (0, 1)$. Since $m \leq c$ we must have, by (4)

$$\frac{n}{a(1 - bl)} \rightarrow 1 \quad \text{as } N \uparrow \infty.$$

But by (i), $bl \rightarrow E(c, a)$, whence the limit in (ii).

To complete the proof we use the basic bounds on the complexity of nonblocking networks given⁵ by Bassalygo and Pinsker, according to whom each of the $c \times c$ nonblocking networks needed in Fig. 2 can be

made using at most $68c \log_2 c + O(c)$ crosspoints and $2(1 + \log_2 c)$ stages.

XXI. A DIRECT ARGUMENT

It is a property of the "direction" picked for the asymptotics in Theorem (24) that the carried loads m approach a limit, so that the line usage $p = m/N$ goes to zero. In practical terms this means that a fixed amount of traffic is being spread over more and more customers, while the load contribution from any one vanishes. Ultimately, then, this limit "direction" is suitable only for very many lightly loaded lines, and it would be more interesting to have similar or analogous results in which the carried load would increase as the networks grew, and the usage p would be bounded away from zero, with the blocking always less than some prescribed number $\epsilon > 0$.

We know from Theorems (15) and (20) that even without the constraint " $bl < \epsilon$," such a positivity constraint on p necessitates $N \log N$ growth for practical networks. It is instructive, however, to give a separate direct argument for this behavior in the case of the networks constructed in Theorem (24).

Lemma: For $1 > \epsilon > 0$, let c be the earliest integer such that $E(c, a) \leq \epsilon$. Then

$$c \geq a(1 - \epsilon). \quad (25)$$

Proof: By hypothesis, $E(c, a) \leq \epsilon < E(c - 1, a)$. As is well-known, the Erlang function satisfies the recurrence

$$E(c, a) = \frac{1}{1 + \frac{c}{aE(c-1, a)}}.$$

Thus,

$$\frac{1}{1 + \frac{c}{aE(c-1, a)}} \leq \epsilon < E(c-1, a).$$

Replacing $E(c - 1, a)$ on the left by the smaller ϵ will decrease the left, giving

$$\frac{1}{1 + \frac{c}{a\epsilon}} \leq \epsilon.$$

Proposition: Let ν be a central bus network constructed to have $bl \leq \epsilon$, as in Theorem (24), by the method⁵ by Bassalygo and Pinsker. Then

$$X(\nu) \geq 8pN \log_2 N + 4Np[\log_2(1 - \epsilon) + \log_2 \lambda]. \quad (26)$$

Proof: ν achieves usage $p = m/N$, so there must be a state $x \in S$ with $|x| \geq Np$. The method of construction implies that ν is a network having $4 + 4 \log_2 c$ stages, where c is the number of central buses, chosen to be the earliest integer c such that $E(c, a) \leq \epsilon$. Thus every call passes through $4 + 4 \log_2 c$ crosspoints. Since two calls do not pass through the same crosspoints, the state x has at least $4Np \log_2 c$ busy crosspoints, and so

$$X(\nu) \geq 4Np \log_2 c.$$

But by Lemma (25) and the choice of c , we have

$$\begin{aligned} c &\geq a(1 - \epsilon) = \lambda N^2(1 - \epsilon) \\ \log_2 c &\geq 2 \log_2 N + \log_2(1 - \epsilon) + \log_2 \lambda \\ X(\nu) &\geq 8pN \log_2 N + 4Np[\log_2(1 - \epsilon) + \log_2 \lambda]. \end{aligned} \tag{27}$$

It follows that any sequence of such networks, growing so that p is bounded away from zero, will grow like $N \log N$. For then λN is bounded below, and (27) implies

$$X(\nu) \geq 2pN \log_2 N + O(N), \quad \text{as } N \rightarrow \infty.$$

XXII. FRAME CONCEPT

The second kind of network structure we shall study is called a *frame*. It is familiar to engineers from the No. 5 and earlier crossbar systems. The idea of the frame is to mount all N terminals on a side in k groups of N/k on subnetworks that are connected pairwise by dedicated junctor groups. In the two-sided case shown in Fig. 3, these connections are described by a complete bipartite graph. We shall suppose that the inlet subnetworks are N/k by kc and identical, and are mirror images of the outlet subnetworks, with c trunks or juncctors connecting each pair of inlet-outlet subnetworks. A solvable limiting case results when we make the subnetwork nonblocking, so that loss is due always to overload of one of the groups of c juncctors between a pair of subnetworks.

Remark: It is tempting to conjecture here that in the natural "weak" limit $N \rightarrow \infty$, $\lambda \rightarrow 0$, $\lambda N^2 \equiv a$, a constant, the loss for the frame network with nonblocking concentrators will approach the Erlang loss $E(c, ak^{-2})$ for c trunks and Poisson traffic ak^{-2} .

It is not hard to see that if the subnetworks in Fig. 3 are nonblocking, then to define the transition rates that devolve from the stochastic model it is enough to know how many trunks are busy in each of the k^2 dedicated groups of size c . Therefore we can use, instead of our usual microscopic semilattice, a very much reduced³ notion of state, as follows: we describe the system by a $k \times k$ matrix x of integers (x_{ij}) ,

$0 \leq i, j \leq k$, with the interpretation that

x_{ij} = number of calls in progress from subnetwork i on the left to subnetwork j on the right.

The x_{ij} are restricted to be integers between 0 and c , the capacity of each (i, j) trunk group. It is useful to have the notations

$$\begin{aligned}
 x_i &= \sum_{j=1}^k x_{ij} = \text{number of calls in progress on subnetwork} \\
 &\quad i \text{ on left} \\
 x^j &= \sum_{i=1}^k x_{ij} = \text{number of calls in progress on subnetwork} \\
 &\quad j \text{ on right} \\
 x^+(i, j) &= \text{the state resulting from } x \text{ when a new call} \\
 &\quad \text{is added to the } (i, j) \text{ trunk group} \\
 x^-(i, j) &= \text{the state resulting from } x \text{ when a hangup occurs} \\
 &\quad \text{on the } (i, j) \text{ trunk group.}
 \end{aligned} \tag{28}$$

With p_x the stationary probability of x , the statistical equilibrium equations are

$$\begin{aligned}
 p_x &\left[\sum_{i,j=1}^k x_{ij} + \lambda \sum_{i,j=1}^k 1_{x_{ij} \neq c} (N - x_i)(N - x^j) \right] \\
 &= \sum_{i,j=1}^k [p_{x^+(i,j)}(x_{ij+1}) 1_{x_{ij} \neq c} + \lambda 1_{x_{ij} \neq 0} p_{x^-(i,j)}(N - x_i + 1)(N - x^j + 1)].
 \end{aligned}$$

Here the indicator functions for $x_{ij} \neq 0$ and $x_{ij} \neq c$ give the right equations on the "boundary" of the state space. The solution of these equations has the convenient product form

$$\begin{aligned}
 p_x &= p_0 \frac{\lambda^{|x|}}{\prod_{i,j} x_{ij}!} \prod_{i,j} x_i! x^j! \binom{N}{x_i} \binom{N}{x^j} \\
 &= p_0 \lambda^{|x|} \prod_{i,j} \binom{x_i}{x_{i1}, \dots, x_{ik}} \binom{x^j}{x_{1j}, \dots, x_{kj}} \binom{N}{x_i} \binom{N}{x^j},
 \end{aligned}$$

where p_0 is the chance that no calls are up, given as the reciprocal of the normalizing sum as

$$p_0^{-1} = \sum_x \lambda^{|x|} \prod_{i,j} \binom{x_i}{x_{i1}, \dots, x_{ik}} \binom{x^j}{x_{1j}, \dots, x_{kj}} \binom{N}{x_i} \binom{N}{x^j}.$$

The sum over the states x is over all $k \times k$ matrices whose entries are integers in $[0, c]$, including the identically 0 state, which contributes a term 1.

Under our symmetry assumptions the probability of blocking between two outer subnetworks i and j is the same as the overall loss,

and is given by

$$bl = \frac{\sum_x p_x 1_{x_j=c} (N-x_i)(N-x^j)}{\sum_x p_x (N-x_i)(N-x^j)}.$$

This formula does not depend on the normalizer p_0^{-1} . Although it is exact, it will be of interest to us primarily for the insight it gives into the asymptotic behavior of bl as $N \rightarrow \infty$, $\lambda \rightarrow 0$, with $\lambda N^2 = a$, constant. To this end it is enough to look at p_x/p_0 . Writing it in the form, for $x \neq 0$, and using $\sum_i x_i = \sum_j x^j = |x|$,

$$\begin{aligned} p_x/p_0 &= \frac{\lambda^{|x|}}{\prod_{i,j} x_{ij}} \prod_{i,j} \binom{x_i-1}{k-m} \prod_{\ell=0}^{x^j-1} \binom{N}{k-\ell} \\ &= \left(\frac{\lambda N^2}{k^2}\right)^{|x|} \\ &\quad \cdot \frac{\prod_{i,j} \left(1 - \frac{x_i-1}{N/k}\right) \cdots \left(1 - \frac{1}{N/k}\right) \left(1 - \frac{x^j-1}{N/k}\right) \cdots \left(1 - \frac{1}{n/K}\right)}{\prod_{i,j} x_{ij}!} \end{aligned}$$

(the products interpreted as 1 if x_i or x^j is 0) one can see that

$$\lim_{\substack{N \rightarrow \infty \\ \lambda \rightarrow 0 \\ \lambda N^2 = a}} p_x/p_0 = \frac{(ak^{-2})^{|x|}}{\prod_{i,j} x_{ij}!};$$

write $\alpha = ak^{-2}$ for simplicity; the probability of loss goes to

$$\frac{\frac{\alpha^c}{c!} \sum_{x_{ij}=c} \frac{\alpha^{|x|-c}}{\prod_{k,\ell \neq i,j} x_{k\ell}!}}{\sum_{n=0}^c \frac{\alpha^n}{n!} \sum_{x_{ij}=n} \frac{\alpha^{|x|-n}}{\prod_{k,\ell \neq i,j} x_{k\ell}!}}.$$

It can be verified that for the various n up to c , the $\sum_{x_{ij}=n}$ summands in the denominator are all equal, and equal to the $\sum_{x_{ij}=c}$ sum in the numerator. Thus, as conjectured,

$$\lim_{\substack{n \rightarrow \infty \\ \lambda \rightarrow 0 \\ \lambda N^2 = a}} bl = \frac{\frac{\alpha^c}{c!}}{\sum_{n=0}^c \frac{\alpha^n}{n!}} = E(c, \alpha) = E(c, ak^{-2}). \quad (29)$$

Table I—Load loss relationships

	Loss	Load
Central bus	$E(c, a)$	$a[1 - E(c, a)]$
Frame	$E(c, ak^{-2})$	$a[1 - E(c, ak^{-2})]$

As for central bus networks, one can get linear growth in the weak limit, described as follows:

Theorem: For every $\epsilon > 0$, there is an integer c and a growing sequence ν_N of two-sided frame networks on N terminals, with N/k subnetworks on a side, such that as $N \rightarrow \infty$, $\lambda \rightarrow 0$, and $a = \lambda N^2 \equiv \text{constant}$,

$$\begin{aligned}
 (i) \quad & bl(\nu_N) \rightarrow E(c, ak^{-2}) \leq \epsilon \\
 (ii) \quad & m(\nu_N) \rightarrow a[1 - E(c, ak^{-2})] \\
 (iii) \quad & X(\nu_N) \leq 2k[68N \log_2 c + O(c)] \\
 (iv) \quad & s(\nu_N) = 4(1 + \log_2 kc).
 \end{aligned} \tag{30}$$

Remark: The reader can check that it is quite proper to regard the frame network with k concentrating subnetworks on a side as a system of k^2 central bus networks: the loss is asymptotically Erlang $E(c, \cdot)$ in both cases, and the carried load for the frame is k^2 times the carried load on the corresponding central bus, as shown in Table I.

Proof of Theorem (30): This proof is very much like that of (24). We chose c to be the least integer such that $E(c, ak^{-2}) \leq \epsilon$, and construct nonblocking $N/k \times kc$ concentrators for the subnetworks, each using no more than $68(N/k)\log_2(kc) + O(kc)$ crosspoints and $2(1 + \log_2 kc)$ stages. Convergence of the loss has been proved as (29), and that of the load follows as usual from the Erlang formula (1), as in (24).

XXIII. REMOTE CONCENTRATOR CONCEPT

A third network structure worth looking at consists of concentrating subnetworks each connected only to one and the same central "core" network by a group of c high-usage trunks, as in Fig. 4. We call this structure, well-known to traffic engineers, the "remote concentrator concept;" it represents an extreme form of the advice to separate concentration and distribution in the network. We shall suppose that the subnetworks are divided into two groups, one carrying the inlets, the other the outlets, so that the whole network is still two-sided. When the outer subnetworks (concentrators) and the inner core are all nonblocking, a useful solvable case results, and we can again guess that as the right limit is taken there will be some connection with Erlang's E function. Since the success of a call attempt depends

entirely on finding first a free trunk into the core, and then one going out from it to the destination concentrator, it is tempting to guess that asymptotically the loss is given by the "blocking polynomial" $1 - (1 - b)^2$, where b is the chance that all c trunks in a group are busy. This simple guess is probably not true, because of the correlation between loads on trunk groups; it may nevertheless be a bound, although we have not been able to prove this: the question whether the blocking owing to simultaneously full groups is larger or smaller than b^2 is open. We can, however, give Erlang E bounds on both kinds of blocking, as well as exact loss formulas for finite N in terms of logarithmic derivatives of a partition function.

Let k be a divisor of N , fixed henceforth, to be interpreted as the number of concentrators on the inlet side of the network, each with N/k inlets, c trunks to the core, and nonblocking. The outlet side is similarly constituted. As a notion of state we can take the matrices $x = (x_{ij}, 0 \leq i, j \leq k)$ with the meaning

x_{ij} = number of calls in progress
from inlet concentrator i to outlet concentrator j .

These matrices are subject to the condition that both rows and columns must not sum to more than c , the trunk group size. This is the same notion of state used for frame networks, except that there the *entries* had to be at the most c , while here the row and column sums are thus bounded.

Using the same notations (28) as for the frame networks, we can put down the following equilibrium equations:

$$p_x \left[\sum_{i,j=1}^k x_{ij} + \lambda \sum_{i,j=1}^k 1_{x_i < c, x^j < c} \left(\frac{N}{k} - x^j \right) \right] \\ = \sum_{i,j=1}^k \left[p_{x^+(i,j)} (x_{ij} + 1) 1_{x_{ij} < c} + \lambda 1_{x_{ij} > 0} p_{x^-(i,j)} \left(\frac{N}{k} x_{i+1} \right) \left(\frac{N}{k} - x^j + 1 \right) \right].$$

Again, the indicator factors give the right equations at the boundary of the state space, and the solution has the same product form as for frame networks:

$$p_x = p_0 \lambda^{|x|} \prod_{i,j=1}^k \binom{x}{x_{i1}, \dots, x_{ik}} \binom{x^j}{x_{ij}, \dots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x^j},$$

where p_0 is the chance of no calls in progress, the reciprocal of the normalizer

$$\sum_{x \in S} \lambda^{|x|} \prod_{i,j=1}^k \binom{x_i}{x_{i1}, \dots, x_{ik}} \binom{x^j}{x_{ij}, \dots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x^j}.$$

The sum over the states $x \in S$ is over all $k \times k$ matrices with nonnegative integer entries, and row *and* column sums at most c .

By symmetry, the probability of blocking between two remote concentrators i and j is the same as the overall loss, and is given by

$$bl = \frac{\sum_{x \in S} p_x [1_{x_i=c} \wedge 1_{x_j=c}] \left(\frac{N}{k} - x_i\right) \left(\frac{N}{k} - x_j\right)}{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right) \left(\frac{N}{k} - x_j\right)}.$$

This formula depends only on the known ratios p_x/p_0 . We examine its asymptotic compartment as $\lambda \rightarrow 0$, $N \rightarrow \infty$, with $\lambda N^2 \equiv a$, constant. The argument for (29) gives

$$\lim_{\substack{N \rightarrow \infty \\ \lambda \rightarrow 0 \\ \lambda N^2 = a}} p_x/p_0 = \frac{\left(\frac{a}{k^2}\right)^{|x|}}{\prod_{i,j=1} x_{ij}!}.$$

XXIV. THE PARTITION FUNCTION

We have, for the remote concentrator concept,

$$\begin{aligned} p_x &= p_0 \lambda^{|x|} \prod_{i,j=1}^{|x|} \binom{x_i}{x_{i1}, \dots, x_{ik}} \binom{x_j}{x_{1j}, \dots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x_j} \\ &= p_0 \lambda^{|x|} c(x). \end{aligned}$$

Hence, introducing the generating function

$$\phi(y) = 1 + \sum_{j=1}^{kc} y^j \sum_{\substack{|x|=j \\ x \in S}} c(x),$$

the moments of the number of calls in progress can be expressed as logarithmic derivatives; especially,

$$\begin{aligned} m &= \lambda \frac{d}{d\lambda} \log \phi(\lambda) \\ \sigma^2 &= \lambda^2 \frac{d^2}{d\lambda^2} \log \phi(\lambda) + \lambda^2 \frac{d}{d\lambda} \log \phi(\lambda), \end{aligned}$$

and by the generalized Erlang formula (1), the blocking is determined via

$$\begin{aligned} 1 - bl &= \frac{1}{\lambda} \frac{m}{(N - m)^2 + \sigma^2} \\ &= \frac{\frac{d}{d\lambda} \log \phi(\lambda)}{\left[N - \lambda \frac{d}{d\lambda} \log \phi(\lambda) \right]^2 + \lambda^2 \frac{d^2}{d\lambda^2} \log \phi(\lambda) + \lambda^2 \frac{d}{d\lambda} \log \phi(\lambda)}. \end{aligned}$$

The interested reader can verify that analogous results hold in the Erlang case of c trunks offered traffic a : the generating function is

$$\phi(y) = 1 + y + \frac{y^2}{2} + \dots + \frac{y^c}{c!}$$

so that

$$1 - bl = 1 - E(c, a) = \frac{\phi'(a)}{\phi(a)}$$

$$m = a(1 - E(c, a)) = a \frac{d}{da} \log \phi(a).$$

The essential form of these relationships persists in the weak limit $\lambda \rightarrow 0, N \rightarrow \infty, \lambda N^2 = a \equiv \text{constant}$. We write

$$p_x = p_0 a^{|x|} \prod_{i,j=1}^k \binom{x_i}{x_{i1}, \dots, x_{ik}} \binom{x_j}{x_{1j}, \dots, x_{kj}} \frac{(N/k)!(N/k)!N^{-2|x|}}{\left(\frac{N}{k} - x_i\right)!x_i \left(\frac{N}{k} - x_j\right)!x_j!}$$

and take the limit as above. Stirling's formula implies that the partition function becomes

$$\phi(y) = 1 + \sum_{\ell=1}^{kc} y^\ell \sum_{\substack{|x|=\ell \\ \text{row sums} \leq c \\ \text{column sums} \leq c}} k^{-2\ell} \cdot \prod_{i,j=1}^k \frac{\binom{x_i}{x_{i1}, \dots, x_{ij}} \binom{x_j}{x_{1j}, \dots, x_{kj}}}{x_i!x_j!}. \quad (31)$$

Then since $(N - m)^2 + \sigma^2 \sim N^2$ in the weak limit, one finds

$$m \rightarrow a \frac{d}{da} \log \phi(a) \quad (32)$$

$$1 - bl \rightarrow \frac{d}{da} \log \phi(a). \quad (33)$$

Theorem: For every integer k and every $\epsilon > 0$, there is an integer c , and a sequence v_N of "remote concentrator" networks, with c trunks from each of k $N/k \times c$ concentrators on a side to a central core $kc \times kc$, such that as $\lambda \rightarrow 0, N \rightarrow \infty, \lambda N^2 = a \equiv \text{constant}$

- (i) $bl(v_N) \rightarrow 1 - \frac{d}{da} \log \phi(a) < \epsilon$
- (ii) $m(v_N) > a \frac{d}{da} \log \phi(a)$
- (iii) $X(v_N) \leq \frac{2N}{k} [68c \log_2 c + O(c)] + 68kc \log_2 kc + O(kc)$
- (iv) $s(v_N) = 6(1 + \log_2 c) + 2 \log_2 k$.

Proof: Again, this proof is like that of (24). Using (32), the blocking can be written as

$$bl = 1 - \frac{m}{\lambda N^2 \sum_{x \in S} p_x \left(1 - \frac{x_i}{N}\right) \left(1 - \frac{x^j}{N}\right)} = 1 - \frac{m}{a + o(1)}$$

$$= 1 - \frac{d}{da} \log \phi(a) + o(1).$$

So with $\lambda N^2 = a$, fixed, pick the integer c in the partition function $\phi(y)$ defined by (31) so that

$$1 - \frac{d}{da} \log \phi(a) < \epsilon.$$

This is possible because the limit blocking must decrease to 0 with increasing c . Thus, (32) proves (i) and (ii); (iii) and (iv) follow by the same kind of concentrator construction as before, using the method of Bassalygo and Pinsker.⁵

XXV. BLOCKING INEQUALITIES FOR REMOTE CONCENTRATOR CONCEPT

When the concentrating subnetworks and the core are nonblocking, it is possible to derive some interesting Erlang E bounds for the blocking in a remote concentrator structure. We first note that the contributions to blocking are of two kinds: a call is blocked between subnetworks i and j because $x_i = c$, or $x^j = c$, or both, i.e., $x_i + x^j = 2c$. Thus,

$$bl = \frac{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right) \left(\frac{N}{k} - x^j\right) (1_{x_i=c} + 1_{x^j=c} - 1_{x_i+x^j=2c})}{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right) \left(\frac{N}{k} - x^j\right)}$$

$$= \sum_{x \in S} p_x (1_{x_i=c} + 1_{x^j=c} - 1_{x_i+x^j=2c}) + o(1).$$

By symmetry the $1_{x_i=c}$ and $1_{x^j=c}$ terms contribute the same amount, so the problem of estimating the blocking reduces to estimating:

(i) the probability $Pr\{x_i = c\} = \sum_{x_i=c} p_x$ of having "all trunks busy" on concentrator i , and

(ii) the "double trouble" term $Pr\{x_i + x^j = 2c\}$.

It is convenient to define, for states $x \in S$, and $1 \leq i, j \leq k$,

$$s^{ij}(x) = (1 - 1_{x_i=c})(1 - 1_{x^j=c}) \left(\frac{N}{K} - x_i\right) \left(\frac{N}{k} - x^j\right).$$

This is the number of unblocked idle inlet-outlet pairs (u, v) with u on concentrator i and v on concentrator j .

Lemma: For integers $0 \leq t \leq w = \max_{x \in S} |x|$, the chance of t calls up on concentrator i can be represented as

$$Pr\{x_i = t\} = Pr\{x_i = 0\} \frac{\alpha^t}{t!} \prod_{\ell=0}^{t-1} \eta_\ell, \quad (35)$$

where

$$\begin{aligned} \eta_\ell &= N^{-2\ell} \sum_{x_i=\ell} \frac{P_x}{Pr\{x_i = \ell\}} \sum_{j=1}^k s^{ij}(x) \\ &= N^{-2\ell} E\{\text{number of unblocked calls on } i | \ell i\text{-trunks are busy}\} \\ &\leq k^{-1}. \end{aligned}$$

Proof: This follows from the form of the statistical equilibrium equations, which says that the rate into a set is the rate out of it. We use the sets $\{x \in S: x_i = t\}$, which partition S and communicate by pairs in a simply ordered array except at the endpoints $\ell=0$ and $\ell=c$. The result follows by iteration. In a similar way it is found that

Lemma:

$$Pr\{x_i + x^j = t\} = Pr\{x_i + x^j = 0\} \frac{\alpha^t}{t!} \prod_{\ell=0}^{t-1} \xi_\ell, \quad (36)$$

where

$$\begin{aligned} \xi_\ell &= N^{-2\ell} \sum_{x_i+x^j=\ell} \frac{P_x}{Pr\{x_i + x^j = \ell\}} \sum_{m=1}^k [s^{im}(x) + (1 - \delta_{im})s^{mj}(x)] \\ &= N^{-2\ell} E\{\text{number of unblocked calls to } i \text{ or } i | x_i + x^j = \ell\} \\ &\leq \frac{2k - 1}{k^2}. \end{aligned}$$

Theorem:

$$\begin{aligned} Pr\{x_i = c\} &\leq E(c, ak^{-2}) \\ Pr\{x_i + x^j = 2c\} &= E(2c, 2ak^{-1} - ak^{-2}). \end{aligned} \quad (37)$$

Proof: Introduce the function on the positive orthant

$$f(y_1, \dots, y_c) = \frac{y_1 y_2 \cdots y_c}{1 + y_1 + y_1 y_2 + \cdots + y_1 y_2 \cdots y_c} = \frac{N}{D};$$

this is increasing in each y_i there, since

$$\frac{\partial f}{\partial y_i} = \frac{N(1 + y_2 + y_2 y_3 + \cdots + y_1 y_2 \cdots y_{i-1})}{y_i D^2} > 0.$$

By normalization, $\sum_{i=0}^c Pr\{x_i = t\} = 1$, so Lemma (35) says that

$$\begin{aligned} Pr\{x_i = c\} &= f\left(a\eta_0, \frac{a}{2}\eta_1, \dots, \frac{a}{c}\eta_{c-1}\right) \\ &\leq f\left(\frac{a}{k}, \frac{a/k}{2}, \dots, \frac{a/k}{c}\right) = E(c, a/k). \end{aligned}$$

The proof for the “double trouble” term is analogous, from Lemma (36).

XXVI. CONCLUSIONS AND PROSPECTS

For the narrow class of probability models for telephone networks described by “finite sources, exponential holding times,” we have shown that the $N \log N$ rate of growth (of the number X of crosspoints) characteristic of nonblocking networks extends also to those with blocking. This narrow class provided easy methodological devices for carrying out the proofs. Extensions to more general statistics have been mentioned in an interesting series of papers by N. Pippenger, listed in the bibliography. However, his results are either combinatorial or restricted to Markovian models similar to ours. Since some of his principal demonstrations depend on what amounts to the old “lost calls held” convention applied to finite sources, his results are strictly not comparable to those given here. Extensions to the distribution-free context remain to be made. As Pippenger suggests, the most useful tools are likely to be the entropy concept and ergodic theory.

REFERENCES

1. E. Brockmeyer, H. L. Halström, and A. Jensen, *The life and work of A. K. Erlang*, Acta Polytechnica Scandinavica, Mathematics and Computing Machinery Series, No. 6, Copenhagen, 1960.
2. V. E. Beneš, “Markov Processes Representing Traffic in Connecting Networks,” B.S.T.J., 42, No. 6 (November 1963) pp. 2795–838.
3. V. E. Beneš, “Reduction of Network States Under Symmetries,” B.S.T.J., 57, No. 1 (January 1978), pp. 111–49.
4. C. Clos, “A Study of Nonblocking Switching Networks, B.S.T.J., 32, No. 2 (March 1953), pp. 406–24.
5. L. A. Bassalygo and M. S. Pinsker, “Complexity of an Optimal Nonblocking Switching Network Without Reconnections,” Problemy Peredachi Informatsii, 9 (1973), pp. 84–7; translated into English in Problems of Information Transmission, 9 (1974), pp. 64–6.

BIBLIOGRAPHY OF BACKGROUND READING AND RELATED WORK

- Beneš, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
- Ikeno, N., “A Limit on Crosspoint Number,” 1959 International Symposium on Circuit and Information Theory, Los Angeles, CA, June 16–18, 1959.
- Pippenger, N., “The Complexity Theory of Switching Networks,” Sci. D. Thesis, Mass. Inst. Technology, 1973.
- Pippenger, N., “Complexity of Seldom Blocking Networks,” Proc. IEEE Communications Conference, 1976, paper 7–8.
- Pippenger, N., “On the Complexity of Strictly Nonblocking Networks,” IEEE Trans. Communications, COM-22 (1974), pp. 1890–2.
- Shannon, C. E., “Memory Requirements in a Telephone Exchange, B.S.T.J., 29, No. 3 (July 1950), pp. 343–9.

Upper and Lower Bounds on Mean Throughput Rate and Mean Delay in Memory-Constrained Queueing Networks

By E. ARTHURS and B. W. STUCK

(Manuscript received April 27, 1982)

Operators use terminals to enter transactions into a system and then wait for the system to respond. The system contains serially reusable resources, and can hold a maximum number of jobs. Each job requires a total mean amount of service at each stage. We calculate upper and lower bounds on the mean throughput rate and mean delay as a function of model parameters, and present examples that show these bounds are sharp, in the sense that they are achievable given only mean values. We also present partial results for closed queueing networks where the long-term, time-averaged distribution of number of jobs at each node in the network obey so-called product form separation of variable type of probability distributions. Examples and data from actual systems illustrate the utility of the work.

I. INTRODUCTION

At present there is great interest in modeling the traffic-handling characteristics of computer and communication systems using queueing networks.¹⁻⁴ The change in cost of electronic solid-state circuitry^{5,6} and rising personnel costs^{7,8} offers strong incentive to design cost-effective digital systems.

Computer communications systems can often be modeled quite naturally by a network of queues, where a job receives service at one stage or queue and then migrates to another stage, until it is completely serviced. Examples of actual systems and associated models are presented in later sections of this report. This class of models captures several fundamental phenomena of such systems, including asynchronous and concurrent execution of different jobs and different amounts of service required at each stage of execution. To answer whether this

type of model is valid, controlled experimentation and measurement must be carried out, and goals or criteria must be set for judging goodness of fit. Finally, one would like to use these models to predict or extrapolate behavior into unknown regions of operation to guide decision making.

Here we focus on one technique for bounding the mean throughput rate and mean delay of an abstraction of a computer communications system. This is only one factor among many others, such as cost, flexibility, and reliability, that must be considered in choosing one design over another for a given application. We drop these other factors from consideration after this point in the interest of brevity.

Broadly speaking, there are two reasons for wanting to quantify the traffic-handling capabilities in a computer communication system:

- (i) Cost reduction of an existing service or product:
 - (a) In an existing system, it is often possible to modify existing scheduling policies to improve performance at an acceptable cost. An example would be to change from one memory partition per application program to a memory pool shared among all application programs.
 - (b) In a system handling a fixed set of job types, different equipment configurations can accomplish these jobs at different costs. Which should be chosen? An example would be to compare using two slow disks versus one fast disk.
- (ii) Comparisons are often desired between current operations and wholly new modes of operations. An example would be using an existing batch computer system for time sharing, using the existing time sharing system for electronic mail, or using existing word processors for voice-annotated text services.

To quantify these issues, typically two stages are involved: the first is *synthesis*, where goals are stated along with different alternatives for reaching those goals, while the second is *analysis*, where the performance (here the mean throughput rate of finishing jobs and the mean delay for each stage of job execution) is quantified. Goals may be either oriented toward the total system, such as total number of jobs of a given type that are handled during an hour, or toward an individual, such as the mean delay to handle one or more stages of a job; along with goals such as these there should be some measure of the *sensitivity* of the goals to different operating points, and so forth.

Analysis often begins by postulating a set of parameters that carry or capture specific operational aspects, drawing inferences based on these parameters (either by mathematical analysis or by discrete event simulation),⁹⁻¹¹ measuring actual or simulated operation, and then repeating this process until it is felt that additional work is no longer warranted.

Our goal here is to demonstrate how to carry out part of this process by a straightforward technique for obtaining bounds on mean throughput rate and mean delay, given only mean value information for the service required at each stage of job execution. In our opinion, there are three principal contributions:

(i) A new technique for obtaining a *lower* bound on mean throughput rate and an associated *upper* bound on mean delay. Earlier workers (e.g., Ref. 1, pp. 212–25) obtained an *upper* bound on mean throughput rate and an associated *lower* bound on mean delay. Furthermore, we present an example that shows that given *only* mean values for the amount of service required at each visit to each stage in the queueing network model, *either* bound can be approached arbitrarily close, depending upon on the amount of fluctuation present about the mean service times. This shows that these bounds are *sharp*, much as was done earlier for loss systems.^{12,13} The interested reader is referred to related works.^{14,15}

(ii) A new technique for calculating both *upper* and *lower* bounds on the mean throughput rate and mean delay for a class of closed queueing networks whose long-term, time-averaged distribution for the number of jobs in system obeys a so-called *product form* or separation of variables decomposition^{2,3} (for an application case study, see Ref. 16). The upper bound on mean throughput rate is the reciprocal of the total mean time to execute the transaction plus the *average* time spent in execution per node, while the lower bound on mean throughput rate is the reciprocal of the total mean time to execute the transaction plus the *maximum* mean time spent in execution per node. The tightness of these bounds will thus depend on how close these factors are to one another (Zahorjan et al., developed these results independently;¹⁷ our derivation is felt to be more straightforward).

(iii) Data from controlled experimentation on actual computer and computer communication systems is presented to validate the approach presented here.

The examples presented are deliberately elementary, chosen for tractability. Everything of interest can be represented by formulas. Furthermore, this approach is a natural starting point for virtually any study of traffic-handling performance, can be refined in a variety of ways, used to check and bound much more complex analyses or simulations, and can be immediately related to measurements in an actual system. Often data are simply not available to describe the arrival statistics and service required for each step of each job, such as would be needed in simulation studies; this suggests using a mean value (distribution free) analysis, rather than more stringent distributional assumptions, and then assessing performance sensitivity by

varying the mean value, rather than investing effort in simulation studies. We advocate *synthesis* via *analysis* of the performance of a given configuration. The approach adopted here is not exhaustive, but it is fundamental. The examples show that only two avenues are available for improving computer communication system performance, reducing the time to handle a given task (i.e., speedup) and handling two or more tasks simultaneously [i.e., concurrency, either real (multiplexing multiple resources) or apparent (via scheduling a single resource)].

II. A MODEL OF A PROCESSOR AND DISK SYSTEM

In this section we deal with a mathematical abstraction of an on-line transaction processing system.

2.1 Model description

Clerks at terminals spend a mean amount of time reading, thinking, and entering the transaction, and then wait for the system to respond before repeating this cycle. Each transaction requires a mean amount of processor time and disk secondary storage access time to be completely executed. The system is configured with a finite amount of memory, and hence can hold a maximum number of jobs at any one time. Figure 1 shows a hardware block diagram of the system. The cycle that a job or transaction follows can be described by a path through a network of queues. The first stage or queue is associated with operators at terminals entering each transaction. Next, each job enters a *staging* queue, where it waits if there are already the maximum number of allowable jobs in the system, and otherwise it immediately enters the system. Once inside the system, a job will receive some processing, then require accessing some data from secondary disk storage, then some processing, and so forth until it is completely executed. Finally, control will return to the operator at the terminal and the process begins anew. Figure 2 shows a queueing network block diagram for this system, consisting of four queues: one for operator jobs, one for staging, one for processors, and one for disks.

The ingredients we need are

(i) The number of clerks, C , actively submitting transactions to the system

(ii) The number of processors, P , and disks, D , connected by a common switch. (The switch is assumed to be much faster than any step of job execution involving either a processor or a disk, and is ignored from this point on.)

(iii) The maximum number of jobs allowed inside the system at any one time, M

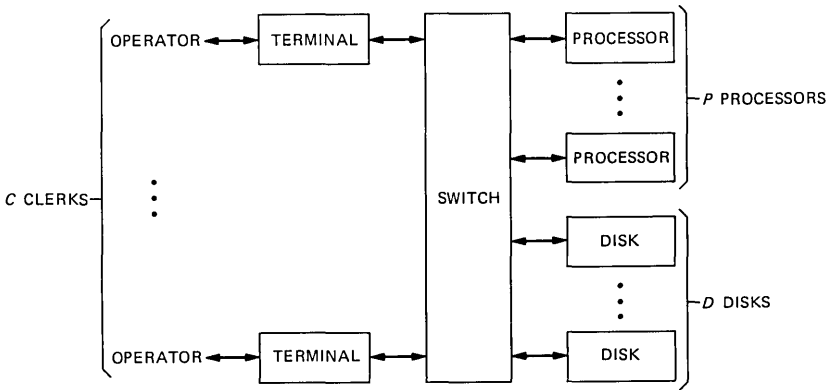


Fig. 1—Block diagram of the system hardware.

(iv) The mean time each operator spends reading, thinking, and entering each transaction, denoted T_{think}

(v) The mean processor time, $T_{\text{processor}}$, and mean disk access time, T_{disk} , per transaction.

The outputs of the analysis are the mean throughput rate of executing transactions and the mean response time (as seen by operators, including both execution time plus waiting time) as a function of model parameters. No job is assumed to be capable of executing in parallel with itself. The operating system multiplexes available jobs among available processors and disks to achieve some degree of concurrent use of resources.

From the vantage point of an operator at a terminal, we see that each transaction undergoes two stages of processing:

(i) A stage spent preparing the transaction for execution, with mean time interval, T_{think}

(ii) A stage spent waiting for the transaction to execute, with mean time interval, R .

For one operator at one terminal, the mean cycle time per transaction is simply the sum of the mean preparation time and mean delay. Hence, when C operators are active, the mean throughput rate equals simply C times the mean throughput rate for one operator:

$$\text{mean throughput rate} = \lambda = \frac{C}{T_{\text{think}} + R}.$$

If we rewrite this equation to find the mean response time, we see

$$R = \frac{C}{\text{mean throughput rate}} - T_{\text{think}}.$$

These two relationships will be fundamental in determining feasible operating regions for mean throughput rate and mean delay.

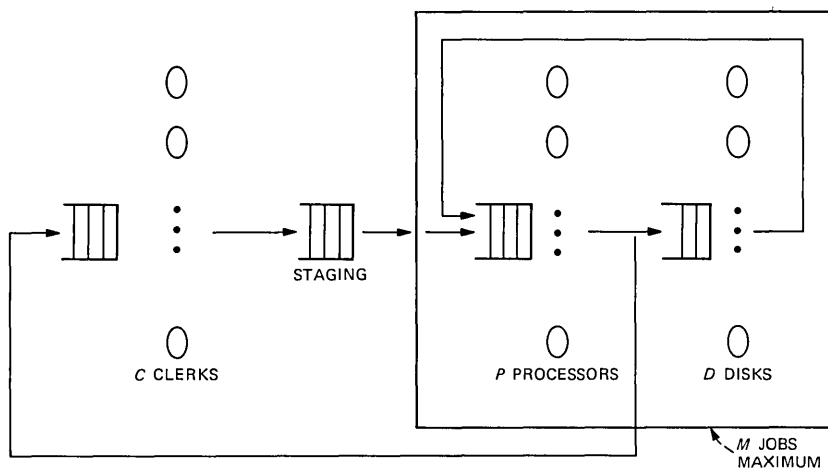


Fig. 2—Block diagram of the system queuing network.

2.2 Mathematical problem statement

The system state space is denoted by Ω , where

$$\Omega = \{(J_{\text{operator}}, J_{\text{stage}}, J_{\text{processor}}, J_{\text{disk}}) |$$

$$J_{\text{operator}} + J_{\text{stage}} + J_{\text{processor}} + J_{\text{disk}} = C; J_{\text{processor}} + J_{\text{disk}}$$

$$= \min[M, C - J_{\text{operator}}]\}.$$

At any instant of time, the system is in a state given by $(J_{\text{operator}}, J_{\text{stage}}, J_{\text{processor}}, J_{\text{disk}})$. Each component is integer valued, and refers to the number of jobs or transactions either in execution or waiting to be executed at that stage. The admissible state space is constrained because

- (i) There are at most C jobs being worked on at any one time
- (ii) The system can hold at most M jobs at any one time.

In a later section, we will show that the mean number of tasks in execution with operators, processors, and disks, averaged over a suitably long time interval, equals the mean throughput rate, λ , multiplied by the total mean execution time for that stage. This is summarized in the following equations, where $E(\cdot)$ denotes a time average of the argument:

$$E[\min(J_{\text{operator}}, C)] = \lambda T_{\text{think}}$$

$$E[\min(J_{\text{processor}}, P)] = \lambda T_{\text{processor}}$$

$$E[\min(J_{\text{disk}}, D)] = \lambda T_{\text{disk}}.$$

In a later section, we show that λ can be upper and lower bounded,

given only this information, in terms of model parameters, as follows:

$$\lambda_{\text{lower}} \leq \lambda \leq \lambda_{\text{upper}}$$

$$\lambda_{\text{lower}} = \frac{C}{T_{\text{think}} + \frac{C}{\min(C, M, P)} T_{\text{processor}} + \frac{C}{\min(C, M, D)} T_{\text{disk}}}$$

$$\lambda_{\text{upper}} = \min \left[\frac{\min(C, M, P)}{T_{\text{processor}}}, \frac{\min(C, M, D)}{T_{\text{disk}}}, \frac{\min(C, M)}{T_{\text{processor}} + T_{\text{disk}}}, \frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}} \right].$$

Each of the upper bounds on mean throughput rate has a physical interpretation, as follows:

(i) The processors are limiting the maximum mean throughput rate

$$\lambda_{\text{upper}} = \frac{\min(C, M, P)}{T_{\text{processor}}}$$

(ii) The disks are limiting the maximum mean throughput rate

$$\lambda_{\text{upper}} = \frac{\min(C, M, D)}{T_{\text{disk}}}$$

(iii) The clerks are limiting the maximum mean throughput rate

$$\lambda_{\text{upper}} = \frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}}$$

(iv) Memory is limiting the maximum mean throughput rate

$$\lambda_{\text{upper}} = \frac{\min(C, M)}{T_{\text{processor}} + T_{\text{disk}}}.$$

The lower bound on mean throughput rate has the physical interpretation of executing jobs one at a time on each processor/disk pair. The upper bound on mean throughput rate is associated with the best possible concurrency, while the lower bound on mean throughput rate is associated with the worst possible parallelism. The upper bound on mean throughput rate yields a lower bound on mean delay; the lower bound on mean throughput rate yields an upper bound on mean delay:

$$\frac{C}{\lambda_{\text{upper}}} - T_{\text{think}} \leq R \leq \frac{C}{\lambda_{\text{lower}}} - T_{\text{think}}.$$

These bounds define an admissible or feasible region of operation and are plotted in Figs. 3 and 4 for the case of one processor and one disk, versus $C = M$.

Two regimes are evident: a *lightly loaded* regime, where the number

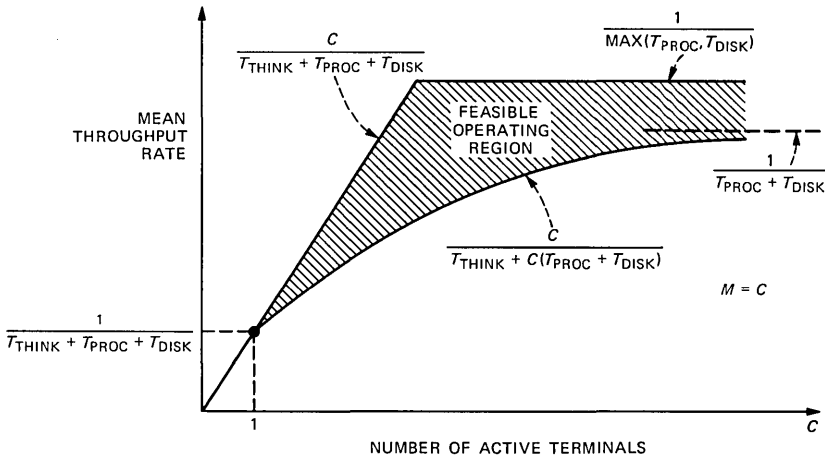


Fig. 3—Mean throughput rate versus number of active terminals.

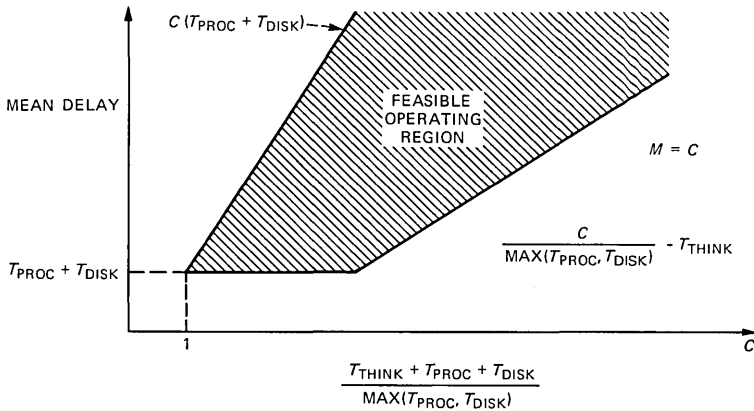


Fig. 4—Mean delay versus number of active terminals.

of *clerks* is directly proportional to the mean throughput rate and the mean delay is independent of the number of clerks, so the clerks are a bottleneck; and a *heavily loaded* regime, where the on-line computer communication system is the bottleneck, with the mean throughput rate independent of the number of clerks and the mean delay directly proportional to the number of clerks.

If we vary the number of clerks, there is a natural breakpoint between these two regimes:

$$\text{breakpoint number of operators} = \lambda_{\text{upper}}[T_{\text{processor}} + T_{\text{disk}} + T_{\text{think}}].$$

As long as the number of clerks is well below this breakpoint, the

clerks and not the system will be limiting the mean throughput rate, and the mean delay per transaction will be approximately the mean delay to execute a transaction with no contention. With the number of clerks well above this breakpoint, the system will be limiting the mean throughput rate, and the mean delay per transaction will be well in excess of that time to execute a transaction with no contention. Analysis suggests measurements to determine where these two regimes lie; synthesis would involve choosing which regime we wish to operate in (remember, we will always have *some* bottleneck!) and designing the system accordingly.

2.3 Impact of memory constraint for one processor and one disk

Here are two possible scheduling policies for a system with one processor and one disk:

(i) Only one job is allowed into the system to be executed at any one time. This is called *single-thread* scheduling, and corresponds to the maximum number of jobs in the system equal to one, $M = 1$.

(ii) More than one job is allowed in the system to be executed at any one time. This is called *multiple-thread* scheduling, and corresponds to $M > 1$. For $M = 1$ we see

$$\frac{C}{T_{\text{think}} + C(T_{\text{processor}} + T_{\text{disk}})} \leq \lambda_{\text{single thread}}$$

$$\leq \min \left(\frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}}, \frac{1}{T_{\text{processor}} + T_{\text{disk}}} \right) M = 1$$

If we allow multiplexing of the processors and disks amongst transactions, then $M > 1$ is allowed, but now one or the other of the two serially reusable resources will become completely utilized for M sufficiently large:

$$\frac{C}{T_{\text{think}} + C(T_{\text{processor}} + T_{\text{disk}})} \leq \lambda_{\text{multiple thread}}$$

$$\leq \min \left(\frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}}, \frac{1}{T_{\text{processor}}}, \frac{1}{T_{\text{disk}}} \right) M > 1$$

In either case, the lower bound on mean throughput rate is identical, but the upper bound can be different, owing to different bottlenecks:

(i) The number of clerks is a bottleneck

$$\lambda_{\text{single thread}}, \lambda_{\text{multiple thread}} \leq \frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}}$$

(ii) The processor is a bottleneck

$$\lambda_{\text{multiple thread}} \leq \frac{1}{T_{\text{processor}}}$$

(iii) The disk is a bottleneck

$$\lambda_{\text{multiple thread}} \leq \frac{1}{T_{\text{disk}}}$$

(iv) Memory is limiting the maximum mean throughput rate

$$\lambda_{\text{single thread}} \leq \frac{1}{T_{\text{processor}} + T_{\text{disk}}}$$

Provided that the clerks or operators are not limiting the maximum mean throughput rate, the ratio of the two different upper bounds is an indication of the gain owing to scheduling or allowing more than one job inside the system at any one time:

$$\frac{\lambda_{\text{multiple thread}}}{\lambda_{\text{single thread}}} = \frac{T_{\text{processor}} + T_{\text{disk}}}{\max(T_{\text{processor}}, T_{\text{disk}})}$$

For one processor and one disk, this gain owing to scheduling can be at most two, no matter what $T_{\text{processor}}$ or T_{disk} are! Moreover, this will only be achieved when $T_{\text{processor}}$ equals T_{disk} , but in general these two mean times will *not* be equal and hence the gain will *not* be as great as a factor of two; for example, if T_{disk} were ten times as great as $T_{\text{processor}}$, then the gain would be at most ten percent, and other factors ignored in this analysis may swamp this.

2.4 Asymptotics

We close with an investigation of asymptotic behavior of this system. One type of asymptotic analysis is to let all parameters be fixed except one, and the final one becomes progressively larger and larger. Here a natural candidate for such a parameter is the number of operators or jobs circulating in the system C . As the number of operators or clerks becomes large, $C \rightarrow \infty$, we see

$$\frac{1}{\frac{T_{\text{processor}}}{\min(P, M)} + \frac{T_{\text{disk}}}{\min(D, M)}} \leq \lambda \leq \min\left(\frac{P}{T_{\text{processor}}}, \frac{D}{T_{\text{disk}}}, \frac{M}{T_{\text{processor}} + T_{\text{disk}}}\right), C \rightarrow \infty.$$

This in turn will yield upper and lower bounds on mean response time:

$$\left(\frac{C}{\lambda_{\text{lower}}} - T_{\text{think}}\right) \rightarrow \infty \leq R \leq \left(\frac{C}{\lambda_{\text{upper}}} - T_{\text{think}}\right) \rightarrow \infty \quad C \rightarrow \infty.$$

In other words, the mean throughput rate lies between two *finite* bounds, while the mean response time is *infinite* (will exceed any finite threshold as we add more and more clerks).

A second type of asymptotic analysis is to fix the ratio of two

parameters, and allow them both to become progressively larger, holding all other parameters constant. Here, a natural candidate is the ratio of the number of jobs over the mean think time per operator, which we denote by α , which is a measure of the *total* offered rate of submitting jobs:

$$\alpha \equiv \frac{C}{T_{\text{think}}}, C \rightarrow \infty, T_{\text{think}} \rightarrow \infty.$$

We allow the number of jobs or terminals to become large, as well as the mean intersubmission time of jobs from each terminal, thus weakening the contribution to the total offered rate of each terminal. In the literature, an analogous procedure is called passing from the so-called *finite source* to *infinite source* arrival process (e.g., see Ref. 18, pp. 102-3), granted certain additional distribution assumptions that we do *not* make here (e.g., see Ref. 18, pp. 80-2). If we fix α while allowing $C, T_{\text{think}} \rightarrow \infty$, we see

$$\frac{\alpha}{1 + \alpha \left[\frac{T_{\text{processor}}}{\min(P, M)} + \frac{T_{\text{disk}}}{\min(D, M)} \right]} \leq \lambda \leq \min \left(\frac{P}{T_{\text{processor}}}, \frac{D}{T_{\text{disk}}}, \frac{M}{T_{\text{processor}} + T_{\text{disk}}} \right).$$

This in turn yields the following lower bound on mean delay:

$$R \geq \begin{cases} \infty & \alpha > \frac{1}{\max \left[\frac{T_{\text{processor}}}{\min(P, M)}, \frac{T_{\text{disk}}}{\min(D, M)}, \frac{M}{T_{\text{processor}} + T_{\text{disk}}} \right]} \\ T_{\text{processor}} + T_{\text{disk}} & \alpha < \frac{1}{\max \left[\frac{T_{\text{processor}}}{\min(P, M)}, \frac{T_{\text{disk}}}{\min(D, M)}, \frac{T_{\text{processor}} + T_{\text{disk}}}{M} \right]} \end{cases}$$

The remaining case, an upper bound on mean delay or mean response time, is trivial:

$$R \leq \infty, \alpha \text{ fixed}, C \rightarrow \infty, T_{\text{think}} \rightarrow \infty.$$

Additional (distributional) information must be available to allow us to handle the case where

$$\alpha = \frac{1}{\max \left[\frac{T_{\text{processor}}}{\min(P, M)}, \frac{T_{\text{disk}}}{\min(D, M)}, \frac{T_{\text{processor}} + T_{\text{disk}}}{M} \right]}.$$

Intuitively we see that if the total mean arrival rate is less than the

upper bound on the mean throughput rate, then the system is capable of having a *finite* lower bound for mean response time; when the total mean arrival rate is greater than the upper bound on the mean throughput rate, then the mean response time lower bound is *infinite*.

Note that the mean throughput rate lies between two *finite* limits, while the mean response time can lie between a *finite* and *infinite* value, given only mean value information, i.e., the mean response time is *not* well bounded given only this amount of information. This is well known in other types of queueing systems, such as the M/G/1 system (e.g., see Ref. 18, pp. 189-92), where the *mean* delay depends not only on the *first* moment of the service time distribution but also the *second* moment of the service time distribution: mean value information does *not* specify the mean delay in such systems by itself, but rather we need the actual distribution of service time to deal with this issue.

III. PROTOTYPE DIRECTORY ASSISTANCE SYSTEM CASE STUDY

Here is a case study in using these techniques. A prototype of an on-line transaction processing system was built to handle telephone number directory assistance queries. In a typical cycle of operation, a person at a terminal would

- (i) Receive a query from a customer via voice telephone
- (ii) Enter the given information into a computer terminal while talking to the customer
- (iii) Wait for the system to respond with the answer to the query
- (iv) Tell the customer over the voice telephone the reply
- (v) Close out customer interaction
- (vi) Wait to receive the next customer query.

The hardware configuration for the system consisted of C terminals, a single processor, a single disk controller, and a single disk spindle. An operating system coordinated scheduling and management of these devices, while a set of prototype application programs handled transaction processing.

Measurements on the prototype system in operation showed that

- (i) The mean time spent by a person talking, reading, and thinking, denoted by T_{think} , was twenty seconds
- (ii) The mean processor time per transaction was broken down into three sets of application programs
 - (a) The operator interface front-end programs consumed 180 milliseconds of processor time per query on the average
 - (b) The index manipulation application programs consumed 420 milliseconds of processor time per query on the average
 - (c) The data retrieval application programs consumed 330 milliseconds of processor time per query on the average
 - (d) Miscellaneous application programs that were invoked for

accounting, system administration, and other purposes consumed one hundred and forty milliseconds (140 ms) per query

Hence, the total mean processor time per query, $T_{\text{processor}}$, was 1.07 seconds

(iii) The mean number of disk accesses per query was twenty six (26), with the disk capable of making one access every twenty five milliseconds (25 ms), which results in a mean time the disk is busy per query, denoted T_{disk} , of six hundred fifty milliseconds (650 ms).

The above measurements on total mean processor time and disk access counts were based on examining the mean resources required for one hundred different transactions to the system; the measurement error on the processor time was felt to be under ten milliseconds, while the measurement error on the number of disk accesses was felt to be under one access. For this level of analysis, the upper and lower mean value bounds on mean response time are given by

$$\max \left[T_{\text{processor}} + T_{\text{disk}}, \frac{C}{\max(T_{\text{processor}}, T_{\text{disk}})} - T_{\text{think}} \right] \leq R \leq C(T_{\text{processor}} + T_{\text{disk}}),$$

while the associated upper and lower mean value bounds on mean throughput rate are given by

$$\frac{C}{T_{\text{think}} + C(T_{\text{processor}} + T_{\text{disk}})} \leq \lambda \leq \min \left[\frac{C}{T_{\text{think}} + T_{\text{processor}} + T_{\text{disk}}}, \frac{1}{\max(T_{\text{processor}}, T_{\text{disk}})} \right].$$

These bounds are plotted in Figs. 5 and 6, along with observed data gathered over an eight-hour time interval with twelve $C = 12$ operators and calculations based upon a closed queueing network model obeying product-form-type solution. The goodness of fit of the closed queueing network model to actual data was felt to be acceptable for the purposes at hand; the mean value lower bound on mean delay and upper bound on mean throughput rate were also felt to give an indication of performance limitations at an early stage of development, which the data gathering and refinement via a closed queueing network model only strengthened further. Note that the system is achieving a great deal of concurrency, because the actual mean throughput rate is much closer to the upper bound, not the single-thread lower bound. Similar observations hold for mean delay.

IV. PROTOTYPE DATA BASE ADMINISTRATION SYSTEM CASE STUDY

A transaction processing system administers the data base for a

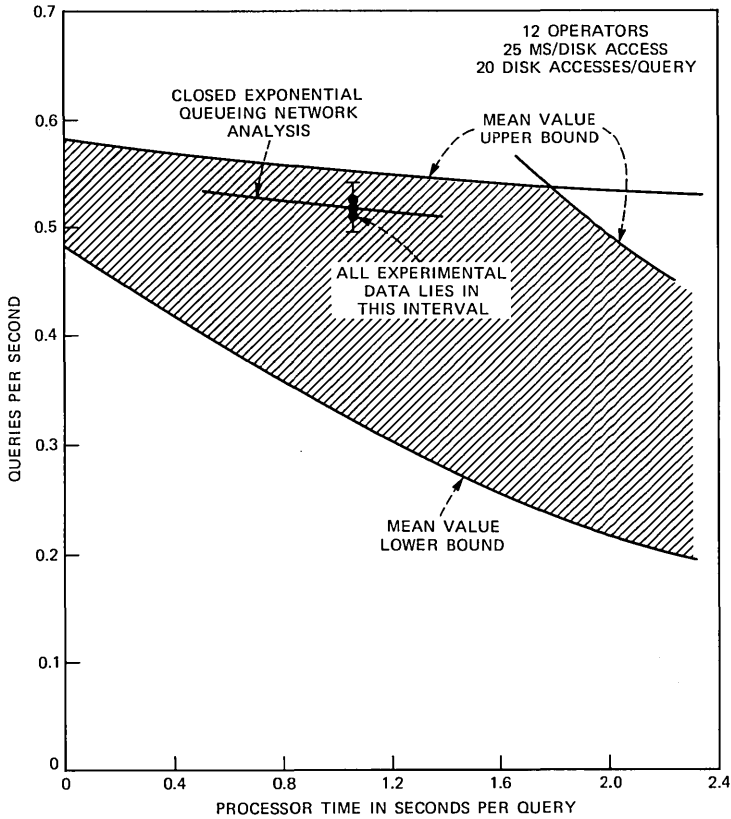


Fig. 5—Mean throughput rate (bounds and data) versus $T_{\text{processor}}$.

second system that switches telephone calls; hence this system is called a *front-end* system to the *back-end* telephone call switching system. Transactions involve additions, deletions, and changes to existing telephone numbers in the switching system files. A prototype system had a hardware configuration consisting of a single processor, a single disk controller, and a single disk spindle, with a fixed number of asynchronous terminals. This same prototype had an operating system to coordinate and schedule these resources, while application programs handled the transaction processing. The application programs were structured into a front end for handling operator terminal interactions, a data base management system, and a back-end communications system for interacting with various switching systems.

The same formulas for upper and lower mean value bounds on mean response time and mean throughput rate hold as in the previous example, except for a change in the numbers.

Two sets of measurements were gathered, one at the start of per-

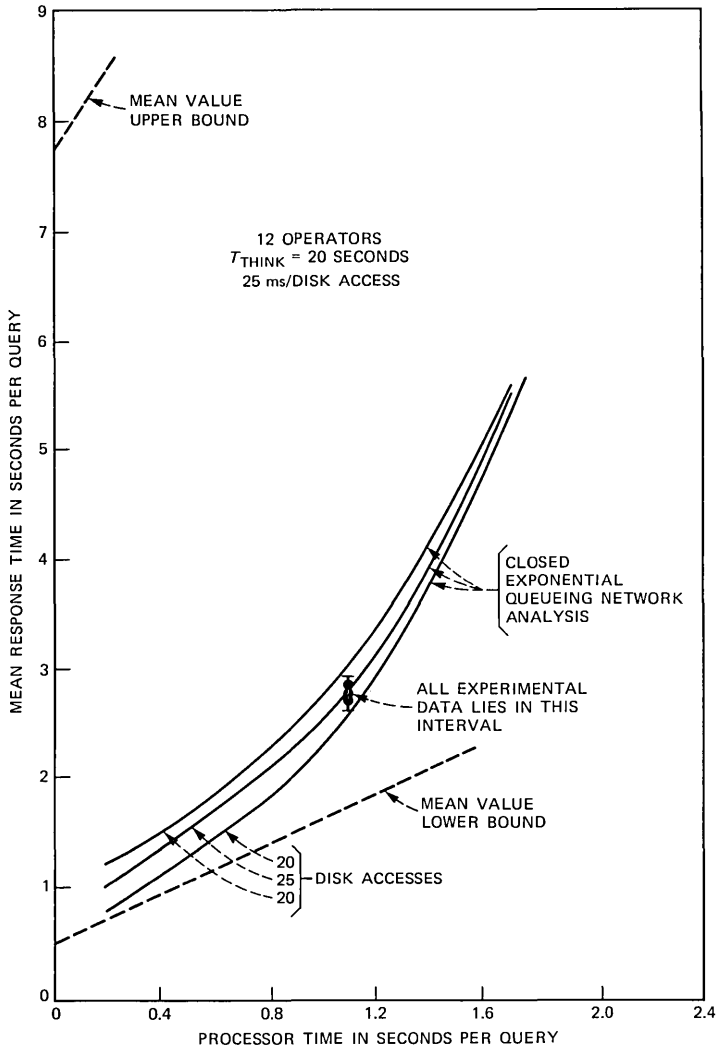


Fig. 6—Mean delay (bounds and data) versus $T_{\text{processor}}$.

formance analysis, labeled *initial* in Table I, and one after completing two months of performance analysis, which involved recoding application programs to take better advantage of operating system features, with the same hardware configuration, labeled *final* in Table I. Measurements were carried out in a controlled environment where the actual hardware, operating system, and application programs were used, but the operator behavior was simulated by a second computer. The behavior of each operator was modeled by a script, involving a time for reading, thinking, and typing, followed by a time waiting for

the system to respond. After an initial startup transient the measurements of response time were quite predictable for all operations, with the measurement error being one second at most. Each operator submitted tens of jobs, and the results were averaged over all operators and all jobs, so the final statistics were felt to be statistically reproducible, to within a fraction of a second.

Figures 7 and 8 plot the mean value upper and lower bounds as well as data from these measurements for the mean response time and mean throughput rate as a function of number of operators. The goodness of fit to mean value bounds was felt to be acceptable for the purpose. Unlike the first case study, the data here clearly shows that a great deal of fluctuation was encountered in system operation under load: for the initial system, the fluctuations were so great that the system apparently was always executing only one transaction at a time, while for the final system, as load built up, the system effectively moved from a regime of two tasks making use of both serially reusable resources to a regime where only one task at a time was in execution. This is in contrast to the other set of data, where the system is always achieving a great deal of concurrency under load. A closed exponential

Table I—Prototype system measurements

Quantity	Initial (seconds)	Final (seconds)
T_{think}	15.0	15.0
$T_{\text{processor}}$	8.2	3.5
T_{disk}	5.0	0.5

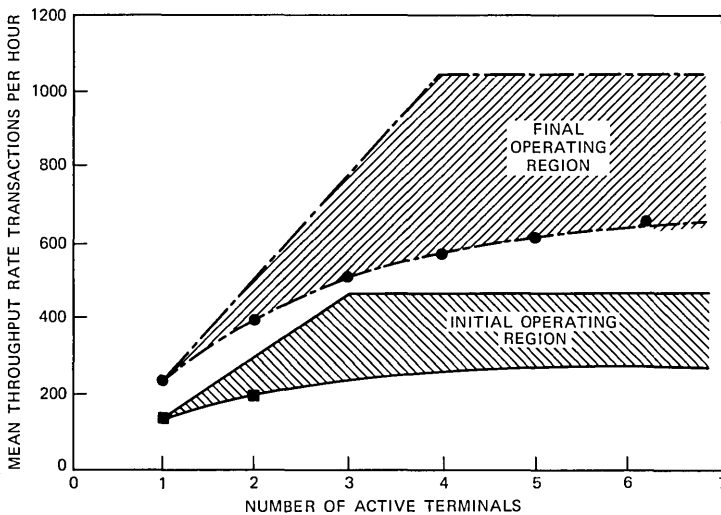


Fig. 7—Mean throughput rate (bounds and data) versus number of active clerks.

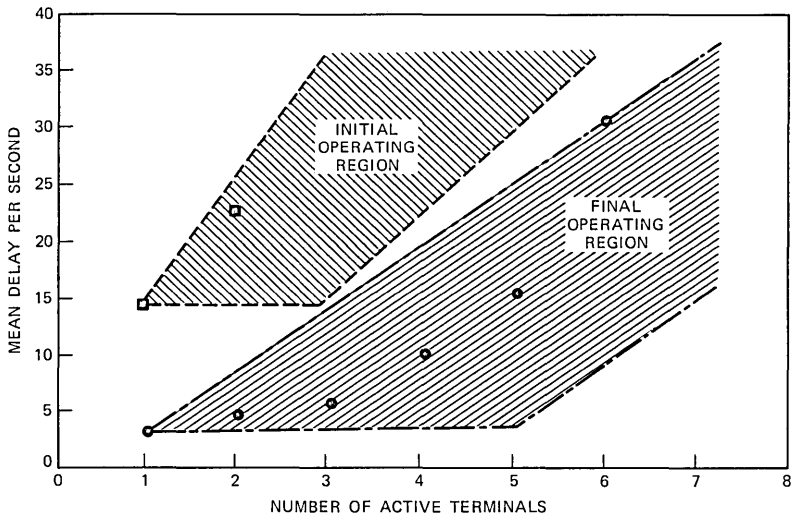


Fig. 8—Mean delay (bounds and data) versus number of active clerks.

queueing network model of this system would predict behavior that closely tracked the upper bound on mean throughput rate and lower bound on mean delay, and would simply not allow for sufficient fluctuation to drive operation into a mode of operation of executing one task at a time. In fact, this suggested a problem with *memory management* that was forcing the system into this mode of operation; an obvious test that was not carried out owing to lack of time was to add more main memory to see if more concurrency might be achieved.

V. A MODEL OF FLOW CONTROL OVER A SINGLE LINK

An on-line communications system consists of operators at terminals who send messages to one another. The system consists of a transmitter and a receiver, with communication channels connecting the transmitter and receiver. The receiver is capable of buffering only a maximum number of messages at any one time, which is a memory constraint.

5.1 Model description

A communications system is composed of a transmitter processor, a receiver processor, a set of buffers each capable of holding one message at the receiver, and a noiseless communications link. Here are the steps involved in sending a message from the transmitter to the receiver:

- (i) The transmitter processes a message. This step has a mean

duration T_{trans} at the transmitter, and it requires both the transmitter and a buffer at the receiver.

(ii) The message is sent over the link from the transmitter to the receiver. This step has a mean duration $T_{\text{trans-rec}}$.

(iii) The receiver processes the message. This step has a mean duration T_{rec} .

(iv) An acknowledgment of correct receipt of the message is sent from the receiver to the transmitter. This step has a mean duration $T_{\text{rec-trans}}$. At the start of this step, the receiver marks the buffer free.

(v) The transmitter processes the acknowledgment. This step has a mean duration of T_{ack} .

At the end of this step, the transmitter marks the buffer free.

We assume from this point on that the time required by the transmitter to process the acknowledgment is zero. Figure 9 shows a hardware block diagram of the system. Figure 10 shows a queueing network block diagram of the system. The system state is denoted by Ω where

$$\Omega = \{(J_{\text{trans}}, J_{\text{trans-rec}}, J_{\text{rec}}, J_{\text{rec-trans}}) \mid J_{\text{trans}} + J_{\text{trans-rec}} + J_{\text{rec}} + J_{\text{rec-trans}} = B\}$$

At any instant of time, the system is in a state given by a four tuple, $(J_{\text{trans}}, J_{\text{trans-rec}}, J_{\text{rec}}, J_{\text{rec-trans}})$, where each component is nonnegative and integer valued, and the state space constraint is obeyed.

The mean throughput rate is denoted by λ . The mean number of jobs in execution in the transmitter and in the receiver equals the mean throughput rate multiplied by the total mean execution time, as shown in a later section. We denote by $E(\cdot)$ the time average of the argument, and write:

$$\lambda T_{\text{trans}} = E[\min(J_{\text{trans}}, P_{\text{trans}} = 1)]$$

$$\lambda T_{\text{rec}} = E[\min(J_{\text{rec}}, P_{\text{rec}} = 1)].$$

Our goal is to find upper and lower bounds on mean throughput rate, subject to meeting state space constraints.

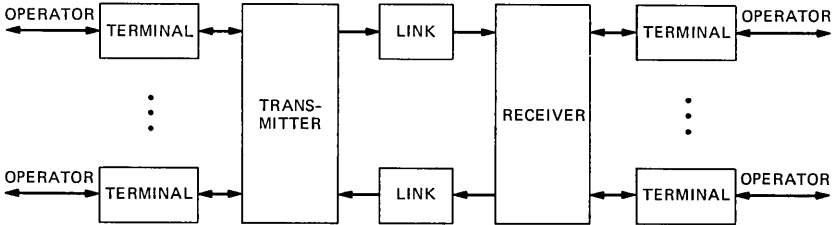


Fig. 9—Hardware block diagram.

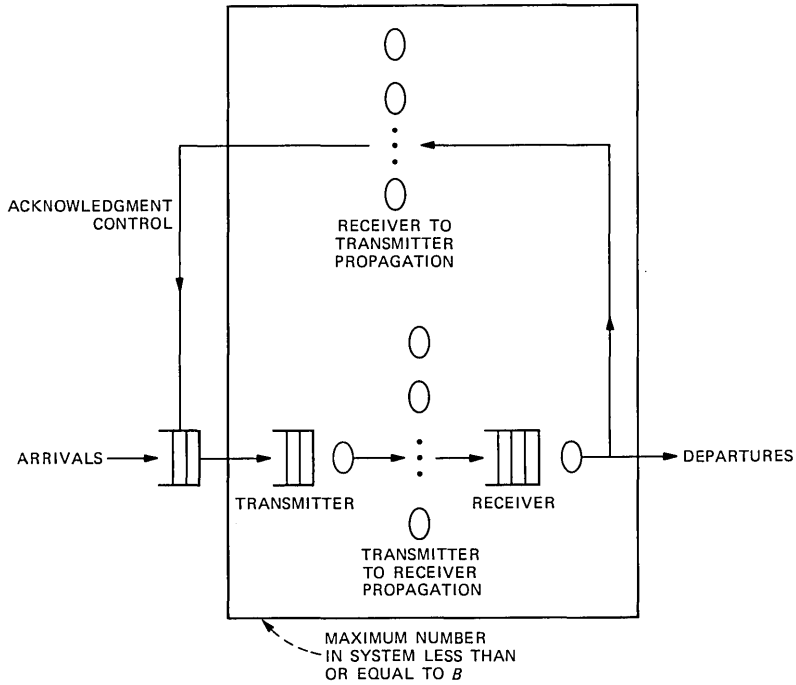


Fig. 10—Queueing network model.

In a later section, we show

$$\frac{B}{BT_{\text{trans}} + T_{\text{trans-rec}} + BT_{\text{rec}} + T_{\text{rec-trans}}} = \lambda_{\text{lower}} \leq \lambda$$

$$\lambda \leq \lambda_{\text{upper}} = \min \left(\frac{1}{T_{\text{trans}}}, \frac{1}{T_{\text{rec}}}, \frac{B}{T_{\text{trans}} + T_{\text{trans-rec}} + T_{\text{rec}} + T_{\text{rec-trans}}} \right).$$

The physical interpretation of the upper bound on mean throughput rate is as follows

(i) The transmitter is the bottleneck

$$\lambda_{\text{upper}} = \frac{1}{T_{\text{trans}}}$$

(ii) The receiver is the bottleneck

$$\lambda_{\text{upper}} = \frac{1}{T_{\text{rec}}}$$

(iii) The number of buffers is the bottleneck

$$\lambda_{\text{upper}} = \frac{B}{T_{\text{trans}} + T_{\text{trans-rec}} + T_{\text{rec}} + T_{\text{rec-trans}}}.$$

The physical interpretation of the lower bound is that at most one message at a time is being handled by the system.

Figures 11 through 13 plot these upper and lower bounds, as well as the results of an exponential queueing network analysis,¹⁻³ for the special case where

$$T_{\text{trans}} = T_{\text{rec}}, T_{\text{trans-rec}} = T_{\text{rec-trans}}$$

for three different cases, where the propagation delay is much smaller, equal, and much larger than the mean processing time at either end of the link. The fraction of time the queueing network model predicts the system to be in state J is denoted by $\pi(J)$, where

$$\pi(J) = \frac{1}{G} T_{\text{trans}}^{J_{\text{trans}}} \frac{T_{\text{trans-rec}}^{J_{\text{trans-rec}}}}{J_{\text{trans-rec}}!} T_{\text{rec}}^{J_{\text{rec}}} \frac{T_{\text{rec-trans}}^{J_{\text{rec-trans}}}}{J_{\text{rec-trans}}!}.$$

The system partition function denoted G is chosen to normalize the probability distribution:

$$\sum_{J \in \Omega_B} \pi(J) = 1.$$

5.2 Negligible link propagation delay

We now restrict attention to the special case where the propagation delay is negligible compared to the processing at either end of the link, from this point on. For one buffer, the mean throughput rate is upper bounded by

$$\lambda \leq \lambda_{\text{upper}} = \frac{1}{T_{\text{trans}} + T_{\text{trans-rec}} + T_{\text{rec}} + T_{\text{rec-trans}} + T_{\text{ack}}} = \frac{1}{T_{\text{trans}} + T_{\text{rec}}}.$$

There is no concurrency or parallel execution of messages, and the

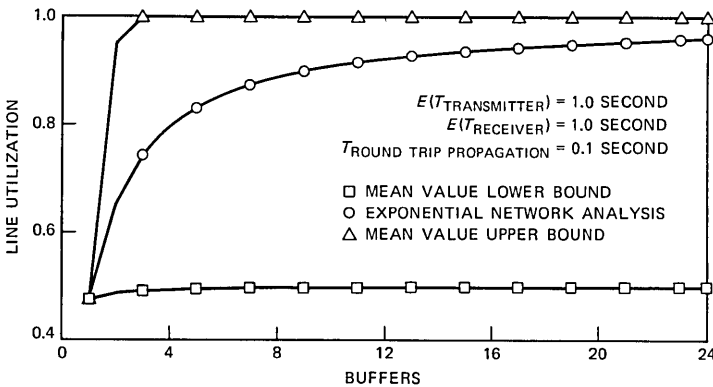


Fig. 11—Line utilization vs. number of buffers ($T_{\text{trans-rec}} = T_{\text{trans}}/10$).

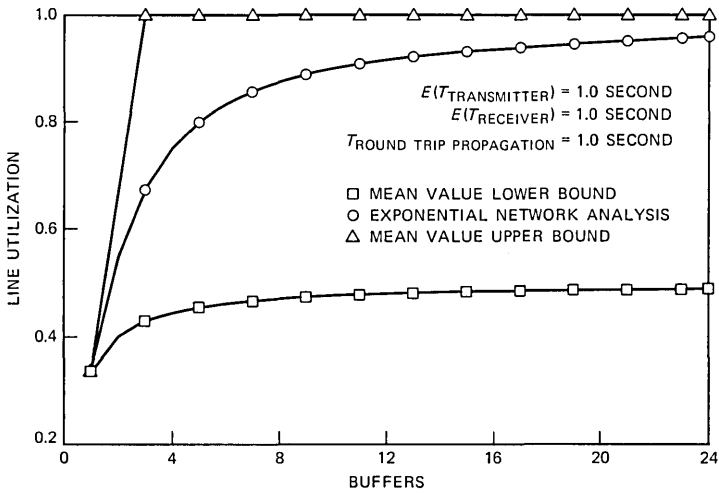


Fig. 12—Line utilization versus number of buffers ($T_{\text{trans-rec}} = T_{\text{trans}}$).

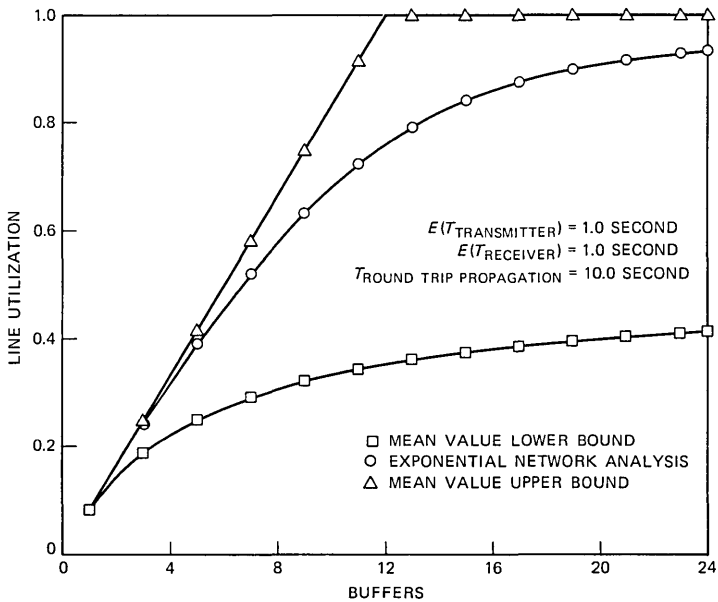


Fig. 13—Line utilization versus number of buffers ($T_{\text{trans-rec}} = 10 T_{\text{trans}}$).

total time required for message handling is the sum of the individual steps.

For more than one buffer, this will yield an upper bound on the mean throughput rate of simply B times the mean throughput rate for

one buffer:

$$\lambda \leq \lambda_{\text{upper}} = \frac{B}{T_{\text{trans}} + T_{\text{trans-rec}} + T_{\text{rec}} + T_{\text{rec-trans}} + T_{\text{ack}}} = \frac{B}{T_{\text{trans}} + T_{\text{rec}}}.$$

On the other hand, as the number of messages increases, then either the transmitter or the receiver (or both) will become completely busy, yielding different upper bounds on mean throughput rate:

(i) The transmitter is a bottleneck

$$\lambda \leq \lambda_{\text{upper}} = \frac{1}{T_{\text{trans}} + T_{\text{ack}}} = \frac{1}{T_{\text{trans}}}$$

(ii) The receiver is a bottleneck

$$\lambda \leq \lambda_{\text{upper}} = \frac{1}{T_{\text{rec}}}.$$

Combining all this, we see

$$\lambda \leq \lambda_{\text{upper}} = \min \left(\frac{1}{T_{\text{trans}} + T_{\text{ack}}}, \frac{1}{T_{\text{receiver}}}, \frac{B}{T_{\text{trans}} + T_{\text{trans-rec}} + T_{\text{rec}} + T_{\text{rec-trans}} + T_{\text{ack}}} \right)$$

$$\lambda \leq \lambda_{\text{upper}} = \min \left(\frac{1}{T_{\text{trans}}}, \frac{1}{T_{\text{rec}}}, \frac{B}{T_{\text{trans}} + T_{\text{rec}}} \right).$$

Increasing the number of buffers from one to two, $B = 1$ to $B = 2$ always increases the maximum mean throughput rate, and now we see

$$\lambda \leq \lambda_{\text{upper}} = \min \left(\frac{1}{T_{\text{trans}}}, \frac{1}{T_{\text{rec}}} \right) \quad B > 1.$$

Furthermore, this increase is maximized for $T_{\text{trans}} = T_{\text{rec}}$, and then the upper bound *doubles* in going from one buffer to more than one buffer. Why is this so? By having more than one buffer, both the transmitter and receiver can simultaneously be filling and emptying a buffer, allowing greater concurrency or parallelism compared with the single-buffer case. We also note that allowing more than two buffers, e.g., *infinite* buffers, will not increase the upper bound on the maximum mean throughput rate any further. This is because there are only two serially reusable resources, a transmitter and a receiver, so once they are concurrently busy, no further gains can be achieved.

For the lower bound on mean throughput rate, we see that

$$\lambda \geq \lambda_{\text{lower}} = \frac{B}{BT_{\text{trans}} + BT_{\text{rec}}} = \frac{1}{T_{\text{trans}} + T_{\text{rec}}},$$

which is identical to the upper bound for $B = 1$. Why is this so? There may be significant fluctuation about the mean values shown above,

and in the limit of one big swing about the mean value all of the messages will pile up at one stage in the network and nothing will be transmitted until buffers become available.

5.3 Impact of fluctuations

We now examine one special case of this problem in detail, where $T_{\text{trans-rec}} = T_{\text{rec-trans}} = T_{\text{ack}} = 0$, and we wish to study the impact of fluctuations about mean values on system performance. We assume the transmitter processing times are sequences of independent identically distributed exponential random variables with mean T_{trans} . We assume the receiver processing times are sequences of independent identically random variables with common hyperexponential distribution $G_{\text{receiver}}(X)$:

$$G_{\text{receiver}}(X) = (1 - \alpha) + \alpha(1 - e^{-X\mu_{\text{rec}}}).$$

In words, a fraction $1 - \alpha$ will require zero processing time at the receiver, while a fraction α will require an exponentially distributed amount of processing time with mean $1/\mu_{\text{rec}}$. The parameter α gives us an additional degree of freedom to model fluctuations in the receiver processing times. For this case, we choose to fix the *squared coefficient of variation* denoted by C^2 , which for the random variable X is defined as the ratio of the variance to square of the mean (the standard deviation, measured in units of mean value, squared):

$$\text{squared coefficient of variation} = \frac{\text{variance}(X)}{E^2(X)} \equiv C^2.$$

When this is zero, the variance is zero, and there is zero fluctuation about the mean. When this is one, we have an exponential distribution, where the standard deviation equals the mean. When this is greater than one, the standard deviation is greater than the mean. For this particular case, we see $0 < \alpha \leq 1$ and hence

$$C^2 = \frac{2}{\alpha} - 1 \geq 1.$$

If the mean is fixed but α is varied from one (the exponential distribution case, where the fluctuations are the order of the mean) to zero (increasing fluctuations about the mean), with most jobs taking zero time but a few taking a very long time, we can gain insight into the impact on performance. Since we have fixed the squared coefficient of variation, the mean is also fixed, since

$$T_{\text{rec}} = \frac{\alpha}{\mu_{\text{rec}}}.$$

The distribution of the number in the receiver subsystem at the

completion of processing at the receiver of a message is denoted by $F(K)$, $K = 0, \dots, B$. If none are left behind, then the mean time to the next completion epoch is $T_{\text{trans}} + T_{\text{rec}}$. If more than zero are left behind at the receiver, then the mean time to the next completion epoch is T_{rec} . The mean throughput rate is given by

$$\lambda = \frac{1}{F(0)(T_{\text{trans}} + T_{\text{rec}}) + [1 - F(0)](T_{\text{rec}})} = \frac{1}{F(0)T_{\text{trans}} + T_{\text{rec}}}.$$

Once we find the distribution of the number of messages in the system at completion epochs, we are done. However, this is a well-known result (see Ref. 18, pp. 235-40), and we merely summarize the known formulas here for the sake of completeness:

$$F(0) = \frac{Q(0)}{\sum_{K=0}^{B-1} Q(K)}.$$

The terms $Q(K)$, $K = 0, \dots, B - 1$ are given implicitly via the following moment-generating function $\zeta(X)$:

$$\zeta(X) = \sum_{K=0}^{\infty} Q(K)X^K = \frac{(1 - X)E[e^{-\lambda(1-X)T_{\text{rec}}}]}{E[e^{-\lambda(1-X)T_{\text{rec}}}] - X}.$$

Illustrative numerical results are plotted in Figs. 14 through 16 assuming the transmitter and receiver service times are independent, identically distributed, exponential random variables. We note that for the special case where $T_{\text{rec}} = T_{\text{trans}} = 1$, the mean throughput rate is given by

$$\lambda = \frac{1 + \frac{2(B + 1)}{C^2 + 1}}{2 + \frac{2(B - 1)}{C^2 + 1}} \quad C^2 \geq 1.$$

Here we see that as $C^2 \rightarrow \infty$ that the mean throughput rate approaches the lower bound of $1/2$ arbitrarily close, i.e., there is no concurrency or gain in going to more than one buffer if the fluctuations are too great. On the other hand, as $B \rightarrow \infty$ for C^2 fixed, the mean throughput rate approaches one, which is the best possible. The numerical plots show in which regime which phenomenon (the fluctuations or the buffering and concurrency) dominates the actual mean throughput rate. The impact of speed mismatch (i.e., as the transmitter and receiver mean message execution times start to differ) tends to swamp the impact of fluctuations: the greater the speed mismatch, the greater concurrency is achieved, because the exact mean throughput rate approaches the upper bound closer and closer as the speed mismatch increases be-

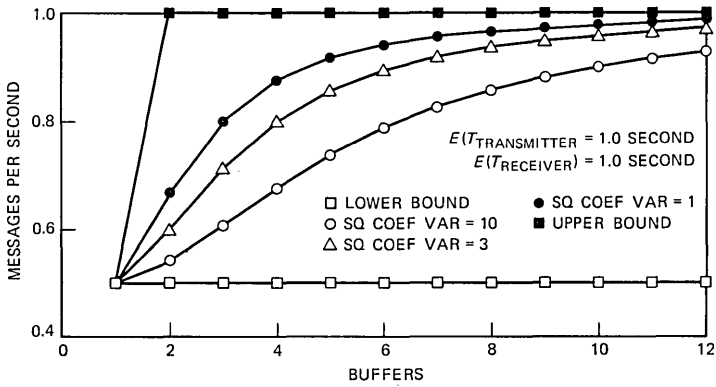


Fig. 14—Mean throughput rate vs. number of buffers for a closed queuing network model ($T_{trans} = 1.0$ and $T_{rec} = 1.0$).

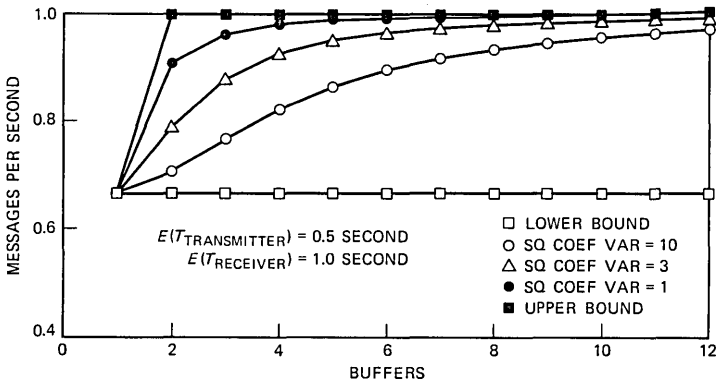


Fig. 15—Mean throughput rate vs. number of buffers for a closed queuing network model ($T_{trans} = 0.5$ and $T_{rec} = 1.0$).

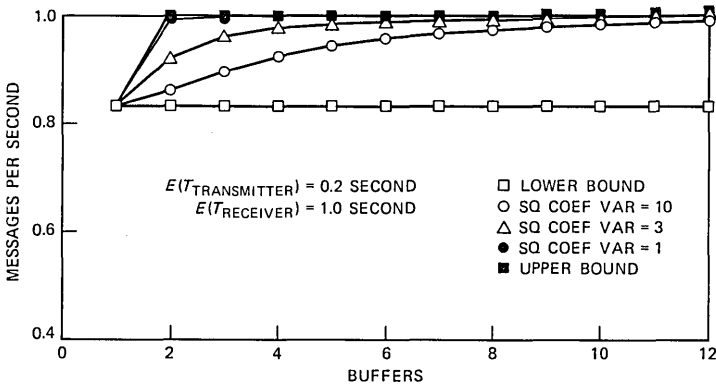


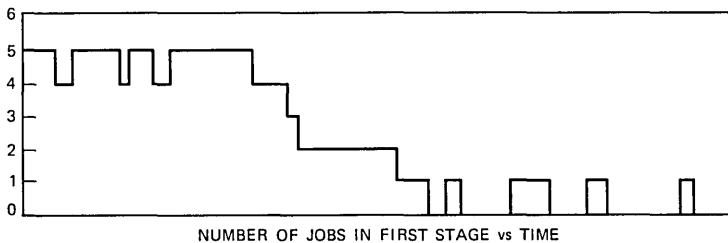
Fig. 16—Mean throughput rate vs. number of buffers for a closed queuing network model ($T_{trans} = 0.2$ and $T_{rec} = 1.0$).

tween transmitter and receiver. Note that the upper bound on mean throughput rate corresponds to a squared coefficient of variation of zero, while the lower bound corresponds to a squared coefficient of variation that becomes infinite.

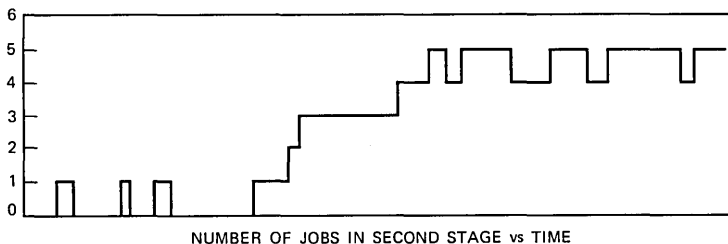
We now discuss this phenomenon in more detail, because the formulas give only one way of understanding this model. Figure 17 shows a sample path generated from a simulation of the model, for a total number of five jobs in the system. In the initial part of the simulation, the first stage fluctuates between four and five jobs, while the second stage fluctuates between zero and one job; in the final part of the simulation, the situation is reversed; after sufficiently long time, we would return to the first case. When most of the jobs are at one stage, the mean throughput rate is roughly the reciprocal of the time to execute one job from start to finish, and there is no concurrency. The other cases, where there are multiple jobs at each stage, are transient and the system spends relatively little time in these states.

The analysis developed above can make these intuitive notions more precise. For example, the mean fraction of time that there are zero jobs at the receiver is

$$\text{fraction of time zero jobs at receiver} = \frac{F(0)}{F(0) + \frac{T_{\text{rec}}}{T_{\text{trans}}}},$$



(a)



(b)

Fig. 17—Sample path generated from a simulation of the model for (a) first stage vs. time and (b) second stage vs. time.

while the fraction of time all the jobs are at the receiver is

$$\text{fraction of time all jobs at receiver} = 1 - \frac{1}{F(0) + \frac{T_{\text{rec}}}{T_{\text{trans}}}}.$$

As we allow $\alpha \rightarrow 0$, i.e., as the fluctuations and squared coefficient of variation become larger, while the mean time spent at the transmitter and receiver stay fixed, we see that the sum of these two fractions can be made *arbitrarily* close to one, which is what the simulation result in Fig. 17 shows. At the same time, we see that the mean sojourn time in the state where the receiver is empty is given by

mean sojourn time in idle receiver state

$$= \sum_{K=1}^{\infty} (1 - \alpha)^{K-1} \alpha K T_{\text{trans}} = \frac{T_{\text{trans}}}{\alpha} \rightarrow \infty \quad \alpha \rightarrow 0.$$

Put differently, if one were to measure the operation of this system, the system might be in the receiver idle state for the entire duration of the observation process, and the other state of the receiver having all jobs (which will also become successively longer and longer as $\alpha \rightarrow 0$) will never be observed, or vice versa! In Fig. 17, this would correspond to gathering data in the first part of the simulation, never in the second part, or vice versa.

5.4 Queueing network analysis for negligible propagation delay

In a later section, we show that the mean throughput rate is upper and lower bounded by

$$\lambda_{\text{lower}} \leq \lambda \leq \lambda_{\text{upper}}$$

$$\lambda_{\text{lower}} = \frac{1}{T_{\text{trans}} + T_{\text{rec}} + \max(T_{\text{trans}}, T_{\text{rec}})}$$

$$\lambda_{\text{upper}} = \frac{1}{T_{\text{trans}} + T_{\text{rec}} + 1/2(T_{\text{trans}} + T_{\text{rec}})}.$$

The mean value bounds, the queueing network upper and lower bounds, and exact queueing network analysis mean throughput calculations are all plotted in Figs. 18 through 20 for $T_{\text{rec}} = 1.0$ and $T_{\text{trans}} = 1.0, 0.5, 0.2$. The queueing network bounds are identical to the exact analysis when the transmitter and receiver execute messages at the same rate. When the transmitter becomes faster than the receiver, the bounds and exact analysis tend to track the upper bound on mean throughput rate; in other words, the speed mismatch is of greater importance than the impact of fluctuations.

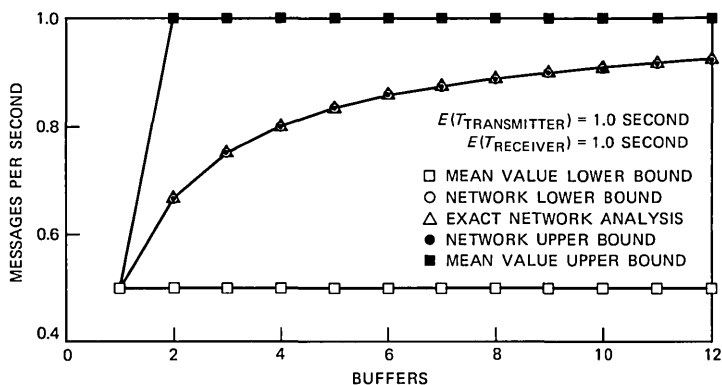


Fig. 18—Mean throughput rate vs. number of buffers for zero propagation time ($T_{trans} = 1.0$ and $T_{rec} = 1.0$).

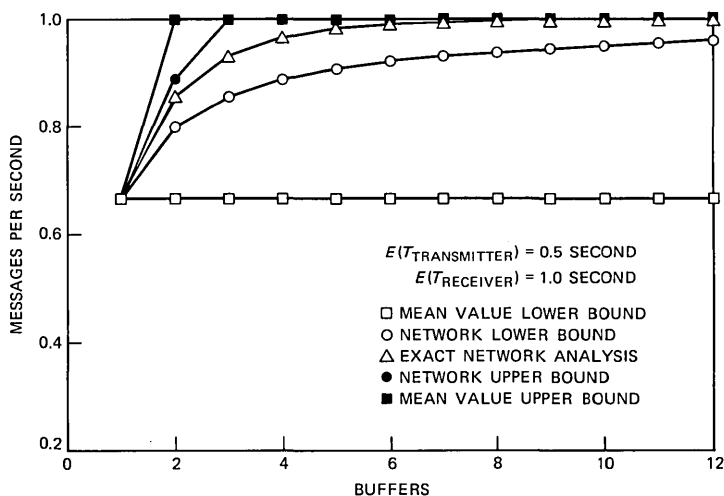


Fig. 19—Mean throughput rate vs. number of buffers for zero propagation time ($T_{trans} = 0.5$ and $T_{rec} = 1.0$).

5.5 Experimental data

To test predictions of this analysis against actual operations, a series of experiments were carried out to determine the mean maximum throughput rate of a data communications link constructed with two computers, one transmitting and one receiving, over a data link where the link propagation time was negligible compared to the data communications processing at either end of the link or the data transmission time of a packet over this link. The test described here involved sending 51,200 bytes of data over a 9600-b/s data link; similar results were found for a 1200-b/s data link. The source data were encoded

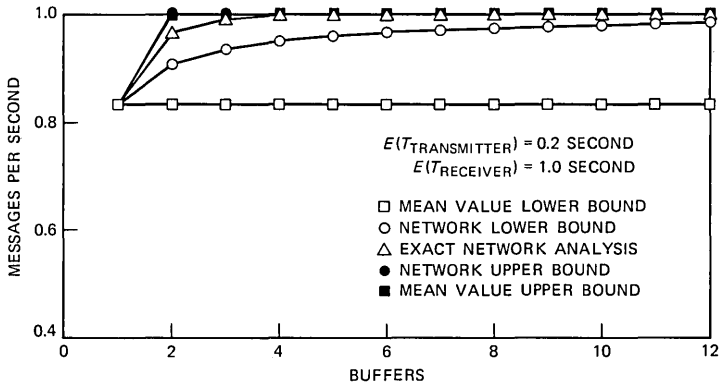


Fig. 20—Mean throughput rate vs. number of buffers for zero propagation time ($T_{trans} = 0.2$ and $T_{rec} = 1.0$).

into packets containing either 32, 64, 128, or 256 bytes (one byte equals eight bits) of data. The system and numerous other details of the experiment will be described elsewhere in a different report.

We wish to test the gain in going from start-stop or single buffering to double buffering and to greater than double buffering; our previous analysis assumes that a mean value of data communications processing time at the transmitter and receiver adequately characterizes the system performance.

The experiment involved simply measuring the time required to transmit 51,200 bytes of data over each link for each size packet. No processing was done on the data at either the transmitter or receiver other than to do the data communications processing required for correct operation. The transmitter and receiver processes resided in the same PDP 11/45 computer with a *UNIX**-like operating system environment.

Table II summarizes the results of that experiment. The time required to send each of 51,200 bytes of data plus two additional bits (for parity and control) over a serial 9600-b/s data link is 53.3 seconds; thus, the data link transmission speed and not the transmitter or receiver is limiting data flow here. This can also be seen directly by noting that the link utilization is approaching one hundred per cent in Table II. This table shows that double buffering at the receiver offers substantial improvement in mean message throughput over single buffering, and there is no apparent advantage in terms of throughput in choosing a receiver buffer larger than two (e.g., seven was tried). Finally, this suggests that for this purpose this level of analysis is

* Trademark of Bell Laboratories.

Table II—PDP 11/45 loop-around experiment—maximum mean throughput rate for transmission of 51,200 bytes over 9600-b/s data link

Number of Buffers	Packet Size (bytes)	Time (seconds)	Maximum Throughput (b/s)	Link Utilization (percent)
1	32	160.0	3200	33
1	64	125.0	4096	43
1	128	90.0	5688	59
1	256	80.0	6400	67
2	32	80.0	6400	67
2	64	58.5	8752	91
2	128	55.5	9225	96
2	256	55.0	9309	97
7	32	64.0	8000	83
7	64	58.0	8827	92
7	128	55.0	9309	97
7	256	54.5	9412	98

appropriate, i.e., that other phenomena that are present are in fact negligible for these purposes, as shown by the data.

VI. CONCLUSIONS

A performance study of an computer communication system may be carried out in at least one of three ways:

- (i) Mean value analysis, as described here^{14,15}
- (ii) Jackson queueing network analysis³
- (iii) Discrete event simulation model.⁹⁻¹¹

In this paper we have demonstrated the ability of the mean value analysis to present a clear picture of the dependence of computer communication system performance on the values of the model parameters. The mean value analysis is a simple, flexible, inexpensive approach to performance analysis and should always be used, even if it is required to supplement the analysis with one or both of the other techniques. The other approaches quantify the impact of fluctuations about mean values on performance, refining the mean value analysis.

The utility or validity of any of these approaches cannot be judged in the abstract: whichever approach or combination of methods is most appropriate must be judged in terms of the data gathering and measurements, and how the data is used to draw inferences concerning cause and effect phenomena, coupled with the spectrum of practical feasible alternatives. The mean value approach presented here is simply one tool for carrying out this complex decision-making process.

VII. ACKNOWLEDGMENT

The authors are indebted to numerous colleagues throughout Bell Laboratories for stimulating this work, people involved with the de-

velopment of over one hundred diverse systems who have shared with us their insight into design and analysis considerations over the past five years. The authors are also indebted to Bell Laboratories for providing such a stimulating work environment.

REFERENCES

1. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, Chapters Five and Six, New York: Wiley Interscience, 1976.
2. F. P. Kelly, "Networks of Queues," *Advances in Applied Probability*, 8 No. 2 (June 1976), pp. 416-32.
3. F. P. Kelly, *Reversibility and Stochastic Networks*, Chichester: Wiley, 1979.
4. C. Sauer and K. Chandy, *Computer Systems Performance Modeling*, Englewood Cliffs, NJ: Prentice Hall, 1981.
5. M. Phister, Jr. *Data Processing: Technology and Economics*, Second Edition, Bedford, MA: Digital Press, 1979.
6. A. F. Shackil, "Design Case History: Wang's Word Processor," *IEEE Spectrum*, 18, No. 8 (August 1981), pp. 29-33.
7. G. H. Engel, J. Groppuso, R. A. Lowenstein, and W. G. Traub, "An Office Communications System," *IBM Systems Journal*, 18, No. 4 (1979), pp. 402-31.
8. R. P. Uhlig, D. J. Farber, and J. H. Bair, *The Office of the Future: Communication and Computers*, Amsterdam: North-Holland, 1979.
9. G. S. Fishman, *Concepts and Methods in Discrete Event Digital Simulation*, New York: Wiley, 1973.
10. G. S. Fishman, *Principles of Discrete Event Simulation*, New York: Wiley, 1978.
11. H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, Reading, MA: Addison Wesley, 1978.
12. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," *B.S.T.J.*, 52, No. 9 (September 1973), pp. 1673-9.
13. A. E. Eckberg, "Sharp Bounds on Laplace-Stieltjes Transforms, with Applications to Various Queueing Problems," *Mathematics of Operations Research*, 2, No. 2 (1977), pp. 135-42.
14. K. Omahen, "Capacity Bounds for Multiresource Queues," *J.A.C.M.*, 24 (1977), pp. 646-63.
15. P. J. Denning and J. P. Buzen, "The Operational Analysis of Queueing Network Models," *Computing Surveys*, 10, No. 3 (1978), pp. 225-61.
16. S. S. Lam, "Queueing Networks with Population Size Constraints," *IBM J. Research and Development*, 21, No. 7 (July 1977), pp. 370-8.
17. J. Zahorjan, K. C. Sevick, D. L. Eager, and B. Galler, "Balanced Job Bound Analysis of Queueing Networks," *Communications of the ACM*, 25, No. 2 (February 1982), pp. 134-41.
18. R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, New York: North Holland, 1981.
19. J. D. C. Little, "A Proof of the Queueing Formula $L = \lambda W$," *Operations Research*, 9 (1961), pp. 383-7.
20. W. S. Jewell, "A Simple Proof of: $L = \lambda W$," *Operations Research*, 15 (1967), pp. 1109-16.
21. R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, Reading, MA: Addison-Wesley, 1967; Little's formula, pp. 18-19.
22. W. L. Smith, "Renewal Theory and Its Ramifications," *J. Royal Statistical Society (Series B)*, 20, No. 2 (February 1958), pp. 243-302.
23. G. P. Klimov, *An Ergodic Theorem for Regenerating Processes*, *Theory of Probability and Its Applications*, 21, No. 2 (June 1976), pp. 392-5.

APPENDIX A

Little's Law

Jobs enter a system, spend time within the system, and depart. The system attributes of interest here are:

- (i) $L(t)$ denotes the number of jobs in the system at time t

(ii) $C(t)$ denotes the number of completions in the time interval $(0, t]$

(iii) Every job that enters the system leaves the system.

Our goal is to relate the mean throughput rate of jobs, the mean time a job spends in the system, and the mean number of jobs in the system.

The total mean time spent by all the jobs in the system is simply the area underneath the function $L(t)$:

$$\text{total mean time in system by all jobs} = \int_0^t L(\tau) d\tau.$$

The total mean time spent in the system by any one job is given by

$$\text{mean time in system per job in } (0, t] \equiv \frac{\int_0^t L(\tau) d\tau}{C(t)}.$$

We multiply and divide by t as follows:

$$\text{mean time in system per job} = \frac{\int_0^t L(\tau) d\tau}{t} \times \frac{1}{C(t)/t}.$$

The first term is simply the mean number of jobs in the system, averaged over a time interval of duration t :

$$\text{mean number of jobs in system in } (0, t] \equiv \frac{\int_0^t L(\tau) d\tau}{t}.$$

The second term is simply the mean throughput rate:

$$\text{mean throughput rate in } (0, t] \equiv \frac{C(t)}{t}.$$

Hence, we have shown that the mean number of jobs inside the system equals the mean throughput rate multiplied by the mean time in system per job, all over an interval of duration $(0, t]$:

mean number of jobs in system in $(0, t]$

= mean throughput rate in $(0, t] \times$ mean time in system/job in $(0, t]$.

If the observation interval becomes infinite, $t \rightarrow \infty$, and the mean values defined here in fact stabilize and do not fluctuate, then we have what is called Little's Law.¹⁹⁻²¹ These other derivations rely on averaging over an ensemble of equally likely experiments, and draw on deep results from the theory of stochastic processes,^{22,23} the difficulty is in showing that the limits in fact exist in a meaningful mathematical

sense. Here we focus exclusively on time averages of quantities of interest, since these can be readily measured.

We close with an application of this result that we will use in the following section. Jobs arrive for processing by a system. Each job requires a total mean amount of service T . The system consists of a single queue feeding P identical processors. At any given instant of time, there are J jobs in the system, either running or waiting to run. The mean throughput rate of jobs is denoted by λ .

We now restrict attention to a subsystem of the total system, the subsystem of jobs in execution. Since we have P processors, the number of jobs in execution at any instant of time is $\min[J, P]$. Hence, we see that the mean number of jobs in execution, averaged over a time interval, equals the mean throughput rate multiplied by the mean time a job spends in execution:

$$\text{mean number of jobs in execution} \equiv E[\min(J, P)] = \lambda T.$$

The actual service pattern of the job is not of interest here: a job may actually consist of a series of steps with different processing at each step, and at the conclusion of each step of execution the job returns to the end of the queue (or to some point in the queue based upon the step) until it is completely executed.

APPENDIX B

Mean Value Analysis

We now present the mathematical analysis to justify assertions in earlier sections. The model we deal with is a system handling only one type of job or transaction. Each transaction consists of one or more steps; at each step, a given amount of a serially reusable resource is required for a given time interval. A resource is any entity that is required for subsequent execution of a transaction; examples of physical hardware resources are processors, memory, disk spindles, disk controllers, communication links, local backplane buses and so forth; example of logical software resources are files, tables, messages, semaphores, and so forth. Here the first step of each transaction involves entering the transaction into the system via an operator at a terminal; the second step of each transaction involves placing the transaction in a staging queue where it will wait if there are more than a given maximum number of jobs already in the system and otherwise will enter the system immediately; and it will go through one or more additional steps inside the system, where the job holds a single serially reusable resource for each step of execution and then moves on, until the job is completely executed and control returns to the operator at the terminal. For each step of each transaction, we are given the amount of each resource and the *mean* time required to hold that set

of resources. We denote by T_K the total mean time spent by a transaction holding resource type K , which we stress is the sum total execution time of all visits to that stage by a transaction.

The mathematical model consists of

(i) $N + 2$ stages of stations: station 0 is associated with operators at terminals, station 1 is the staging station, and stations 2, \dots , $N + 1$ (N total) are associated with a single, serially reusable resource

(ii) Stage $K = 0, 2, \dots, N + 1$ has P_K identical parallel servers or processors

(iii) A maximum of M jobs can be held at all stages $K = 2, \dots, N + 1$

(iv) Each job moves from station to station, and requires T_K total mean amount of service time at stage $K = 0, 2, \dots, N + 2$.

Figure 21 is a queueing network block diagram of this system.

We denote by λ the total mean throughput rate of completing jobs; R denotes the total mean response time (queueing or waiting time plus execution time) per job. The system state space is denoted by Ω . Elements in the state space are denoted by $\mathbf{J} = (J_0, \dots, J_{N+1})$. J_K , $K = 0, 2, \dots, N + 1$ denotes the number of jobs either waiting or in execution at stage K .

Feasible elements in the state space obey the following constraints:

(i) The total number of tasks in the system is fixed at P_0

$$P_0 = |\mathbf{J}| = \sum_{K=0}^{N+1} J_K.$$

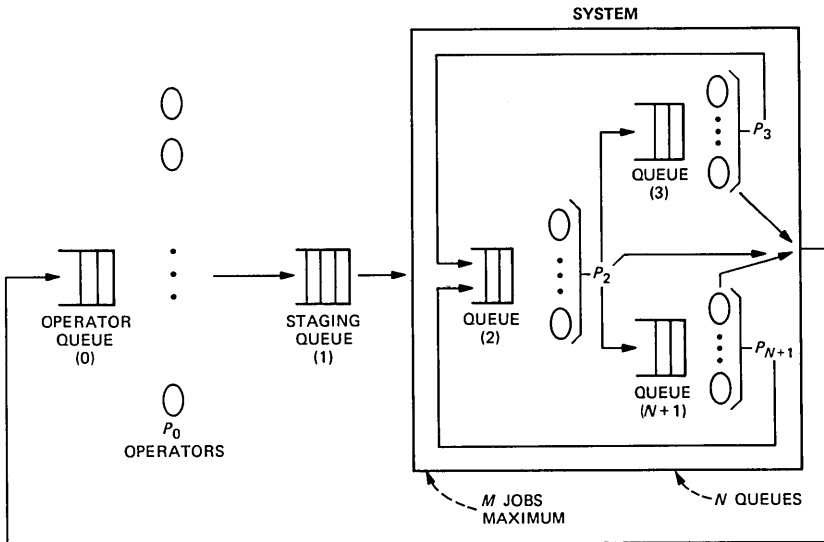


Fig. 21.—Block diagram of the queueing network model.

(ii) There can be at most a maximum of M jobs inside the system:

$$\sum_{K=2}^{N+1} J_K = \min[M, P_0 - J_0].$$

Combining all these, we see that elements \mathbf{J} in Ω are nonnegative integer-valued tuples such that

$$\mathbf{J} \in \Omega = \{V \mid V = (V_0, \dots, V_{N+1}); \quad V_N \geq 0 \quad K = 0, \dots, N + 1; \\ \sum_{K=0}^{N+1} V_K = P_0; \quad \sum_{K=2}^{N+1} V_K = \min(M, P_0 - V_0)\}.$$

The number of jobs *in execution* at stage $K = 0, 2, \dots, N + 1$ is given by $\min(J_K, P_K)$ at any given instant of time. From the previous section, Little's Law allows us to write:

mean number in execution at stage K

$$= E[\min(J_K, P_K)] = \lambda T_K \quad K = 0, 2, \dots, N + 1,$$

where $E(\cdot)$ denotes the time average of the argument. Our goal is to find upper and lower bounds on λ subject to the state space constraints on $J_K, K = 0, \dots, N + 1$.

Since mean throughput rate and mean response time or delay are related via

$$\lambda = \frac{P_0}{T_0 + R}$$

we will also obtain associated lower and upper bounds on mean delay.

B.1 Lower bound on mean throughput rate

We first divide both sides of the following equation

$$E[\min(J_0, P_0)] = \lambda T_0$$

by P_0 . In like manner, we divide both sides of the following equations

$$\lambda T_K = E[\min(J_K, P_K)] \quad K = 2, \dots, N + 1$$

by $\min(M, P_0, P_K)$. Now we add up these $N + 1$ equations:

$$\frac{E[\min(J_0, P_0)]}{P_0} + \sum_{K=2}^{N+1} \frac{E[\min(J_K, P_K)]}{\min(M, P_0, P_K)} \\ = \lambda \left[\frac{T_0}{P_0} + \sum_{K=2}^{N+1} \frac{T_K}{\min(M, P_0, P_K)} \right].$$

Now we interchange the mean value with the summation on the left-hand side:

$$E \left[\frac{\min(J_0, P_0)}{P_0} + \sum_{K=2}^{N+1} \frac{\min(J_K, P_K)}{\min(M, P_0, P_K)} \right] = \lambda \left[\frac{T_0}{P_0} + \sum_{K=2}^{N+1} \frac{T_K}{\min(M, P_0, P_K)} \right].$$

Our goal is to lower bound the left-hand side by one, which will yield a lower bound on λ .

Two cases can arise. First, there can exist one $I = 2, \dots, N + 1$ such that $P_I \leq J_I$. Since all the terms on the left-hand side are non-negative, we can lower bound the left-hand side by ignoring all of these terms except term I :

$$\begin{aligned} \frac{\min(J_0, P_0)}{P_0} + \sum_{K=2}^{N+1} \frac{\min(J_K, P_K)}{\min(M, P_0, P_K)} &\geq \frac{\min(J_I, P_I)}{\min(M, P_0, P_I)} \\ &\geq \frac{P_I}{\min(M, P_0, P_I)} \geq 1 \quad I = 2, \dots, N + 1. \end{aligned}$$

Second, for all $K = 0, 2, \dots, N + 1$, $P_K > J_K$ and hence

$$\min(J_K, P_K) = J_K \quad K = 2, \dots, N + 1.$$

Two subcases arise: if $P_0 - J_0 \leq M$ then there is no waiting by any job in the staging queue, and

$$\frac{J_0}{P_0} + \sum_{K=2}^{N+1} \frac{J_K}{\min(M, P_0, P_K)} \geq \frac{J_0}{P_0} + \frac{P_0 - J_0}{P_0} = 1.$$

The other subcase is if $P_0 - J_0 > M$ and then there is waiting in the staging queue, so

$$\frac{\min(J_0, P_0)}{P_0} + \sum_{K=2}^{N+1} \frac{J_K}{\min(M, P_0, P_K)} \geq \sum_{K=2}^{N+1} \frac{J_K}{\min(M, P_0, P_K)} = \frac{M}{M} = 1.$$

Hence, we see that

$$\lambda \left[\frac{T_0}{P_0} + \sum_{K=2}^{N+1} \frac{T_K}{\min(M, P_0, P_K)} \right] \geq 1$$

and we obtain the desired lower bound:

$$\lambda_{\text{lower}} = \frac{P_0}{T_0 + \sum_{K=2}^{N+1} \frac{P_0}{\min(M, P_0, P_K)} T_K}.$$

The total mean time to execute a job at each stage in the system has been stretched from T_K , $K = 2, \dots, N + 1$ to \tilde{T}_K , $K = 2, \dots, N + 1$, where

$$\tilde{T}_K = \frac{P_0}{\min(M, P_0, P_K)} T_K \geq T_k \quad K = 2, \dots, N+1$$

$$\lambda_{\text{lower}} = \frac{P_0}{T_0 + \sum_{K=2}^{N+1} \tilde{T}_K},$$

which is one way of quantifying the slowdown at each node owing to congestion.

B.2 Upper bound on mean throughput rate

From the definition of λ we see

$$\lambda = \frac{E[\min(J_K, P_K)]}{T_K} \leq \frac{\min(P_K, P_0, M)}{T_K} \quad K = 0, 2, \dots, N+1.$$

From this same identity, we obtain a second upper bound:

$$\begin{aligned} \lambda T_K &\leq E(J_K) \quad K = 0, 2, \dots, N+1 \\ &\rightarrow \lambda \left(T_0 + \sum_{K=2}^{N+1} T_K \right) \leq E \left(J_0 + \sum_{K=2}^{N+1} J_K \right) = P_0. \end{aligned}$$

The constraint on the maximum number of jobs inside the system can be written as

$$\sum_{K=2}^{N+1} J_K \leq \min(P_0, M).$$

If we use Little's Law, we see

$$\lambda \sum_{K=2}^{N+1} T_K = \sum_{K=2}^{N+1} E(J_K) \leq \min(M, P_0).$$

In summary, we have shown

$$\lambda \leq \min \left\{ \min_{K=0,2,\dots,N+1} \left[\frac{\min(P_0, P_K, M)}{T_K} \right], \frac{P_0}{T_0 + \sum_{K=2}^{N+1} T_K}, \frac{\min(M, P_0)}{\sum_{K=2}^{N+1} T_K} \right\}.$$

B.3 Interpretation

One intuitive explanation for these bounds is the following. To achieve the upper bound on mean throughput rate, each step of job execution has little fluctuation relative to its mean value, and jobs interleave with one another. The mean throughput rate can be upper bounded via the following mechanisms:

(i) The total number of jobs circulating in the system is limiting the mean throughput rate; in this regime, as we increase the number

of jobs, the mean throughput rate increases in roughly the same proportion

(ii) One stage is executing jobs at its maximum rate, limiting the mean throughput rate; in this regime, as we increase either the speed of each processor at the stage, or the number of processors with the same speed, the mean throughput rate increases in roughly the same proportion

(iii) The constraint on the maximum number of jobs in the system is limiting the mean throughput rate; in this regime, as we increase the allowable maximum number of jobs in the system, the mean throughput rate increases accordingly.

To achieve the lower bound on mean throughput rate, each step of job execution has large fluctuations relative to its mean value, so that all jobs in the system are congested at one node. A different way of gaining insight into this lower bound is to replace the service or processing time distribution at each node with a bimodal distribution with the same mean as the old distribution, where $(1 - \epsilon_K)$ denotes the fraction of jobs at stage K that are executed in "zero" time and ϵ_K is the fraction of jobs at stage K that are executed in time $1/\mu_K$ such that $T_K = \epsilon_K/\mu_K$. Here in normal operation two things can occur: the mean time for a job to cycle through the network will be roughly zero, since most stages will take zero time, and hence the number of jobs in circulation will limit the mean throughput rate, or one stage of execution will take a time that is much longer relative to all the other times, and hence all but one or two jobs will be congested at one node, thus limiting the mean throughput rate.

APPENDIX C

Product Form Distribution Results

The mathematical model considered in this section is a special case of that considered in the previous section:

- (i) One type of job that migrates amongst S stations or stages
- (ii) A *single* processor available to execute a job at stage $K = 1, \dots, S$
- (iii) N tasks or jobs circulate among the nodes
- (iv) T_K denotes the total mean amount of service required by a job summed over all its visits to stage $K = 1, \dots, S$.

The system state is denoted by Ω :

$$\Omega = \left\{ (J_1, \dots, J_S) \mid \sum_{K=1}^S J_K = N \right\}.$$

At any given instant of time, the system is in state $\mathbf{J} = (J_1, \dots, J_S)$,

where $J_K, K = 1, \dots, S$ denotes the number of jobs at node K (both waiting and in execution). The long-term time-averaged distribution of number of jobs at each node at an arbitrary instant of time is assumed from this point on to obey a so-called *product form* or separation of variables formula

$$\text{PROB}(J_1 = K_1, \dots, J_S = K_S) = \frac{1}{G_N} \prod_{I=1}^S T_I^{K_I} \quad (K_1, \dots, K_S) \in \Omega$$

$$G_N = G_N(T_1, \dots, T_S) = \sum_{J \in \Omega} \prod_{I=1}^S T_I^{J_I}.$$

The interested reader is referred to the literature³ for probabilistic assumptions that lead to this type of probability measure on Ω , the admissible state space. G_N is the *system partition function* chosen to normalize the product form to a probability measure.³ Granted these assumptions, we observe that the mean throughput rate of jobs making a complete cycle of the system is given by

$$\lambda = \frac{\text{PROB}(J_K > 0)}{T_K} = \frac{G_{N-1}(T_1, \dots, T_S)}{G_N(T_1, \dots, T_S)}.$$

Our goal is to obtain *tighter* upper and lower bounds on mean throughput rate and hence mean delay than we obtained in the previous section, using this additional information. We begin by observing that

$$\lambda \sum_{K=1}^S T_K = \frac{G_{N-1}(T_1, \dots, T_S) \sum_{K=1}^S T_K}{G_N(T_1, \dots, T_S)}$$

is a symmetric function of the S variables T_K , i.e., we do not change the value of the function when we interchange any two variables. This property allows us to show that this function has its maximum when all the variables are equal to one another. This follows from calculating first the gradient of the function at that point and showing that it is zero, and second showing that the Hessian, the determinant of all partial second derivatives, is negative definite at that point. An alternate way of seeing this holds is to realize that the gradient is zero (from the symmetry of the function) at the point where all coordinates are unity, so this point must either be a minimum or a maximum; we then evaluate the function at a neighboring point, say the point where all coordinates except one equal zero, and see that this is *less* than at the point where all coordinates are one, so this must be a maximum. Hence, we see

$$\lambda \sum_{K=1}^S T_K \leq \frac{SG_{N-1}(T_1 = 1, \dots, T_S = 1)}{G_N(T_1 = 1, \dots, T_S = 1)}$$

$$\leq \frac{S \binom{S+N-2}{N-1}}{\binom{S+N-1}{N}} = \frac{SN}{S+N-1}.$$

We now rearrange this upper bound to see

$$\lambda \leq \frac{N}{\sum_{K=1}^S T_K + (N-1)T_{\text{average}}}.$$

The first term in the denominator is the mean time for a job to make a complete cycle through the network:

$$T_{\text{cycle}} = \sum_{K=1}^S T_K.$$

The second term in the denominator is the mean amount of time per node spent by a job in one cycle of the network:

$$T_{\text{average}} = \frac{1}{S} \sum_{K=1}^S T_K.$$

The same method is now used to obtain a lower bound on the mean throughput rate, by observing that

$$\frac{1}{\lambda} = \frac{G_N(T_1, \dots, T_S)}{G_{N-1}(T_1, \dots, T_S)}.$$

Without loss of generality, we number the nodes such that node S is the node that will do the greatest amount of processing on a job on the average during one cycle:

$$T_S = \max_{K=1, \dots, S} T_K.$$

This can be used to rewrite the above expression:

$$\frac{1}{\lambda} = T_S + \frac{G_N(T_1, \dots, T_{S-1})}{G_{N-1}(T_1, \dots, T_S)}.$$

Since the second term is positive, this immediately gives us an *upper* bound on the mean throughput rate:

$$\lambda \leq \frac{1}{T_S}.$$

In words, node S is the *bottleneck* node, in this sense.

We now rewrite our expression for the reciprocal of the mean throughput rate:

$$\frac{1}{\lambda} = T_S + \sum_{K=1}^{S-1} T_K F(T_1, \dots, T_S)$$

$$F(T_1, \dots, T_S) = \frac{G_N(T_1, \dots, T_{S-1})}{G_{N-1}(T_1, \dots, T_S) \sum_{K=1}^{S-1} T_K}$$

We now manipulate this expression as we did above:

$$F(T^1, \dots, T^S) \leq \frac{G^N(T^1 = 1, \dots, T^{S-1} = 1)}{(S-1)G^{N-1}(T^1 = 1, \dots, T^S = 1)}$$

$$= \frac{\binom{S+N-2}{N}}{(S-1) \binom{S+N-2}{N-1}} = \frac{\frac{(S+N-2)!}{N!(S-2)!}}{\frac{(S-1)(S+N-2)!}{(S-1)!(N-1)!}} = \frac{1}{N}$$

Combining all this, we see

$$\frac{1}{\lambda} \leq T_S + \frac{1}{N} \sum_{K=1}^{S-1} T_K$$

Rearranging, we see

$$\frac{N}{\sum_{K=1}^S T_K + (N-1)T_{\max}} \leq \lambda$$

The first term in the denominator is the mean time for one job to make one complete cycle of the network:

$$T_{\text{cycle}} = \sum_{K=1}^S T_K$$

The second term is the maximum mean time a job spends at any one node in the network:

$$T_{\max} = \max_{K=1, \dots, S} T_K$$

In summary, we see

$$\frac{N}{\sum_{K=1}^S T_K + (N-1)T_{\max}} \leq \lambda \leq \min \left[\frac{1}{T_{\max}}, \frac{N}{\sum_{K=1}^S T_K + (N-1)T_{\text{average}}} \right]$$

$$\frac{N}{T_{\text{cycle}} + (N-1)T_{\max}} \leq \lambda \leq \min \left[\frac{1}{T_{\max}}, \frac{N}{T_{\text{cycle}} + (N-1)T_{\text{average}}} \right]$$

If the *average* time per node spent in execution by one job during a cycle and the *maximum* time per node per job are roughly comparable to one another, these bounds will be quite close to one another.

CONTRIBUTORS TO THIS ISSUE

E. Arthurs, Ph.D., 1957 (Electrical Engineering), Massachusetts Institute of Technology; M.I.T., 1957–1962; Bell Laboratories, 1962—. From 1957 to 1962 Mr. Arthurs was on the faculty of the M.I.T. Electrical Engineering Department. Since joining Bell Laboratories in 1962, he has worked on a variety of computer and communication network problems.

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research in traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, stochastic control, and filtering. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. In 1971 he taught stochastic processes at SUNY Buffalo, and in 1971–72 he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963), and of *Mathematical Theory of Connecting Networks and Telephone Traffic* (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, IEEE, Phi Beta Kappa.

Ronald Caruso, B.S., 1956, Rutgers University; M.S., 1964, Stevens Institute of Technology; Bell Laboratories, 1968—. Mr. Caruso is engaged in materials characterization of semiconductor crystals and epitaxial iron garnet films. Member, American Chemical Society, Phi Beta Kappa, Pi Mu Epsilon, Phi Lambda Upsilon.

Gerard J. Foschini, B.S.E.E., 1961, Newark College of Engineering, Newark, NJ; M.E.E., 1963, New York University, New York; Ph.D., 1967 (Mathematics), Steven Institute of Technology, Hoboken, NJ; Bell Laboratories, 1961—. Mr. Foschini has been with Bell Laboratories, Holmdel, NJ, since 1961. He initially worked on real-time program design. For many years he worked in the area of communication theory. In the spring of 1979 he taught at Princeton University. Mr. Foschini has supervised planning the architecture of data communications networks. Currently, he is involved with digital radio research. Member, Sigma Xi, Mathematical Association of America, IEEE, New York Academy of Sciences.

Craig A. Gaw, B.S. (with Distinction), 1970, M.S., 1974, Ph.D., 1979 (Electrical Engineering), Northwestern University; Argonne National Laboratory, 1967–1970; Bell Laboratories, 1978—. At Bell Laboratories, Mr. Gaw has been engaged in the characterization of GaAs double heterostructure injection laser material and devices. He is particularly interested in identifying parameters that affect device quality and reliability. Member, Tau Beta Pi, Sigma Nu, Sigma Xi, IEEE, EMSA.

Basil W. Hakki, B.S.E.E., 1957, M.S., 1958, Ph.D., 1960 (Electrical Engineering), University of Illinois; Bell Laboratories, 1963—. After joining Bell Laboratories, Mr. Hakki was involved in the design and analysis of GaAs two-valley microwave oscillators. He then worked with III-V light-emitting diodes and lasers. His work on GaAs double heterostructure injection lasers covered many aspects of the laser, including reliability, device design for cw and high-power operation, and device physics. He is currently involved in the study of quaternary lasers.

Walter R. Holbrook, B.S.M.E., 1969, Lafayette College; Bell Laboratories, 1959—. At Bell Laboratories Mr. Holbrook's work included the design of light-emitting diodes. Currently, he is a member of the Laser Development Department.

Andrew S. Jordan, B.S. (Metallurgy), 1959, Pennsylvania State University; Ph.D. (Metallurgy), 1965, University of Pennsylvania; Bell Laboratories, 1965—. Mr. Jordan has worked mainly in the area of compound semiconductors. He had been involved in the growth, phase equilibria, and impurity incorporation of ZnTe, CdTe, GaP, and GaAs. More recently, he has studied the degradation and reliability of GaP LEDs. Currently, he is engaged in modeling GaAs crystal growth. Member, Electrochemical Society.

Chinlon Lin, B.S.E.E., 1967, National Taiwan University; M.S., 1970, University of Illinois; Ph.D., 1973, University of California at Berkeley; Bell Laboratories, 1974—. Mr. Lin has worked on tunable dye lasers, short-pulse generation, nonlinear optics in fibers for frequency conversion, single-mode fiber dispersion and bandwidth studies, picosecond-injection-laser pulse generation and high-speed optoelectronics. At Berkeley, he received an IBM Fellowship and a Lankersheim Scholarship. He is currently in the Physical Optics and Electronics Research Department. Mr. Lin received an IEE (London) Electronics Letters Premium Paper Award in 1980 for a paper on zero-

dispersion-wavelength tailoring in single-mode fibers. Senior Member, IEEE, Topical Advisor for Fiber and Integrated Optics for the Optical Society of America.

Pao-Lo Liu, B.S., 1973 (Physics), National Taiwan University; M.S., 1976, Ph.D., 1979 (Applied Physics), Harvard University; Bell Laboratories, 1979—. At Bell Laboratories, Mr. Liu has worked on picosecond pulse generation, high-speed integrated optical modulators. He is currently a member of the Coherent Optics Research Department.

Dan L. Philen, B.S., 1968 (Chemistry), Auburn University; Ph.D., 1975 (Physical Chemistry), Texas A&M University; Georgia Institute of Technology, 1976–1979; Bell Laboratories, 1979—. Since joining Bell Laboratories Mr. Philen has been engaged in exploratory measurements on optical fiber properties. Member, American Chemical Society, Optical Society of America, Sigma Xi, Sigma Pi Sigma, Phi Lambda Upsilon.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked on remote line concentrators for the electronic switching system. Since 1968 he has supervised a group engaged in theoretical studies in data communications and is currently a member of the Communications Methods Research Department. During the academic year 1967–68, he was on leave as Professor of Electrical Engineering at the University of Florida. In Spring 1981, he was a visiting lecturer at Stanford University. Member, Sigma Xi.

B. W. Stuck, S.B., S.M., Sc.D. (Electrical Engineering), Massachusetts Institute of Technology in 1969, 1969, and 1972, respectively; Bell Laboratories, 1972—. Since joining Bell Laboratories in 1972, Mr. Stuck has worked on a variety of digital communication and computer systems. Member, MAA, SIAM, IMS, ORSA, IEEE.

Akira Tomita, B.S., 1974, M.S., 1978 (Applied Physics), Hokkaido University, Sapporo, Japan; M.S., 1979, Ph.D., 1980 (Optics), University of Arizona; Bell Laboratories, 1980—. Mr. Tomita has worked in the field of nonlinear optical phase conjugation. He is currently working on lightwave telecommunication systems.

A. R. Tynes, B.S., 1950 (Engineering Physics), Montana State University; M.S., 1953 (Physics), Ph.D., 1963 (Physics), Bell Labora-

tories, 1961—. Presently Mr. Tynes is a member of the Undersea Systems Laboratory.

Allyn R. Von Neida, B.S. (E.E.), B.S. (Metallurgy), 1954, Lehigh University; Ph.D., 1960, Yale University; Bell Laboratories, 1961—. Mr. Von Neida has worked on materials for magnetic memories and is now engaged in crystal growth and characterization of semiconductors. Member, AIME, American Physical Society, Sigma Xi.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

- Brainard R. C., Scattaglia J. V., **Programmable Test-Bed for Composite Television Coding.** *Smpte J* 91(10):906-911, 1982.
- Cleveland W. S., Devlin S. J., **Calendar Effects in Monthly Time-Series—Modelings and Adjustment.** *J Am Stat A* 77(379):520-528, 1982.
- Cleveland W. S., Harris C. S., McGill R., **Judgments of Circle Sizes on Statistical Maps.** *J Am Stat A* 77(379):541-547, 1982.
- Daubechies I., Klauder J. R., **Constructing Measures for Path-Integrals.** *J Math Phys* 23(10):1805-1822, 1982.
- Nikolakopoulou G. A., Edelson D., Schryer N. L., **Modeling Chemically Reacting Flow Systems. 2. An Adaptive Spatial Mesh Technique for Problems With Discontinuities and Steep Fronts.** *Comput Chem* 6(3):93-99, 1982.
- Willie J. S., **Covariation of a Time-Series and a Point Process.** *J Appl Prob* 19(3):609-618, 1982.
- Willie J. S., **Measuring the Association of a Time-Series and a Point Process.** *J Appl Prob* 19(3):597-608, 1982.
- Worrall B. M., Hall M. A., **The Analysis of an Inventory Control Model Using Polynomial Geometric-Programming.** *Int J Prod* 20(5):657-667, 1982.

ENGINEERING

- Alferness R. C., **Wave-Guide Electrooptic Modulators.** *IEEE Micr T* 30(8):1121-1137, 1982.
- Antler M., **Field Studies of Contact Materials—Contact Resistance Behavior of Some Base and Noble-Metals.** *IEEE Compon* 5(3):301-307, 1982.
- Chen C. Y., Bethea C. G., Cho A. Y., Garbinski P. A., **Temporal Resolution of an $Al_xGa_{1-x}As/GaAs$ Bias-Free Photodetector.** *Electr Lett* 18(20):890-891, 1982.
- Feuer A., Barmish B. R., **Instability of Optimal Aim Control—Reply (Letter).** *IEEE Auto C* 27(5):1140, 1982.
- Howard R. E., Jackel L. D., Swartz R. G., Grabbe P., Archer V. D., Epworth R. W., Hu E. L., Tennant D. M., Voshchenkov A. M., **Buried Channel MOSFETs With Gate Lengths From 2.5- μ m to 700-A.** *Elec Dev L* 3(10):322-324, 1982.
- Hwang J. C. M., Flahive P. G., Wemple S. H., **Performance of Power FETs Fabricated on MBE-Grown GaAs-Layers.** *Elec Dev L* 3(10):320-321, 1982.
- Lee T. P., Burrus C. A., Liu P. L., Dentai A. G., **High-Efficiency Short-Cavity InGaAsP Laser With One High-Reflectivity Mirror.** *Electr Lett* 18(19):805-806, 1982.
- Lemons R. A., Bosch M. A., **Electrically Amplified Optical-Recording.** *Elec Dev L* 3(9):254-255, 1982.
- Lin C., Tomita A., Tynes A. R., Glodis P. F., Philen D. L., **Picosecond Dispersionless Transmission of InGaAsP Injection-Laser Pulses at the Minimum Chromatic Dispersion Wavelength in a 27-km-Long Single-Mode Fiber.** *Electr Lett* 18(20):882-884, 1982.
- Logan R. A., Vanderziel J. P., Temkin H., Henry C. H., **InGaAsP/InP (1.3- μ m) Buried-Crescent Lasers With Separate Optical Confinement.** *Electr Lett* 18(20):895-896, 1982.
- Luryi S., Kazarinov R. F., **On the Theory of the Thermionic Emission Transistor. 2. TET as an Element of Logic-Circuits.** *Sol St Elec* 25(9):933-942, 1982.
- Luryi S., Kazarinov R. F., **Optimum Baritt Structure.** *Sol St Elec* 25(9):943-945, 1982.
- Martin R. D., Thomson D. J., **Robust-Resistant Spectrum Estimation.** *P IEEE* 70(9):1097-1115, 1982.
- Minford W. J., **Accelerated Life Testing and Reliability of High k-Multilayer Ceramic Capacitors.** *IEEE Compon* 5(3):297-300, 1982.

- Schiavone J. A., **Microwave Radio Meterology—Diurnal Fading Distributions.** *Radio Sci* 17(5):1301-1312, 1982.
- Thomson D. J., **Spectrum Estimation and Harmonic-Analysis (Review or Bibliog.).** *P IEEE* 70(9):1055-1096, 1982.
- Troxel D. E., Schreiber W. F., Burzinski N. J., Matson M. D., **Interactive Enhancement of Tone Scale.** *Opt Eng* 21(5):841-846, 1982.
- Tsang W. T., Logan R. A., Ditzenberger J. A., **Ultralow Threshold, Graded-Index Wave-Guide, Separate Confinement, CW Buried-Heterostructure Lasers.** *Electr Lett* 18(19):845-847, 1982.
- Wong D. Y., Juang B. H., Gray A. H., **An 800 bit/s Vector Quantization LPC Vocoder.** *IEEE Acoust* 30(5):770-780, 1982.
- Luss H., **Operations-Research and Capacity Expansion Problems—A Survey (Review or Bibliog.).** *Operat Res* 30(5):907-947, 1982.

MANAGEMENT/ECONOMICS

- Guasch J. L., Weiss A., **An Equilibrium-Analysis of Wage Productivity Gaps.** *Rev Econ S* 49(4):485-497, 1982.
- Hakansson N. H., **Changes in the Financial Market—Welfare and Price Effects and the Basic Theorems of Value Conservation.** *J Finance* 37(4):977-1004, 1982.

PHYSICAL SCIENCES

- Ackerman J. R., Kohler B. E., Huppert D., Rentzepis P. M., **Radiationless Decay of 1,3,5,7-Octatetraene.** *J Chem Phys* 77(8):3967-3973, 1982.
- Alferness R. C., Buhl L. L., **High-Speed Wave-Guide Electrooptic Polarization Modulator.** *Optics Lett* 7(10):500-502, 1982.
- Allara D. L., **Analysis of Surfaces and Thin-Films By IR, Raman, and Optical Spectroscopy (Review or Bibliog.).** *ACS Symp S* 1982(199):33-47, 1982.
- Auston D. H., Smith P. R., **Picosecond Optical Electronic Sampling—Characterization of High-Speed Photodetectors.** *Appl Phys L* 41(7):599-601, 1982.
- Bean J. C., Rozgonyi G. A., **Patterned Silicon Molecular-Beam Epitaxy With Sub-Micron Lateral Resolution.** *Appl Phys L* 41(8):752-755, 1982.
- Bertz S. H., Dabbach G., **Factors Governing the Thermal-Stability of Organocopper Reagents—Two New Classes of Heterocuprates With Greatly Improved Thermal-Stability.** *J Chem S CH*1982(18):1030-1032, 1982.
- Bhat R., Oconnor P., Temkin H., Dingle R., Keramidas V. G., **Acceptor Incorporation in High-Purity OMCVD Grown GaAs Using Trimethyl and Triethyl Gallium Sources.** *Inst Phys C*1982(63):101-106, 1982.
- Bhat R., Keramidas V. G., **Comparative-Study of GaAs Grown by Organo-Metallic Chemical Vapor-Deposition (OMCVD) Using Trimethyl and Triethyl Gallium Sources.** *P Soc Photo* 323:104-109, 1982.
- Chabal Y. J., Rowe J. E., Poate J. M., Franciose A., Weaver J. H., **Stoichiometry and Structural Disorder Effects on the Electronic-Structure of Ni and Pd Silicides.** *Phys Rev B* 26(6):2748-2758, 1982.
- Chin A. K., Zipfel G. L., Mahajan S., Ermanis F., DiGiuseppe M. A., **Cathodoluminescence Evaluation of Dark Spot Defects in InP/InGaAsP Light-Emitting-Diodes.** *Appl Phys L* 41(6):555-557, 1982.
- Cone R. L., Ender D. A., Otteson M. S., Fisher P. L., Friedman J. M., Guggenheim H. J., **Multiresonant 2-Photon-Absorption-Induced 4-Wave Mixing in Crystalline Rare-Earth Insulators.** *AIP Conf PR*1982(90):471-477, 1982.
- Daniels J. M., Cladis P. E., Finn P. L., Powers L. S., Smith J. C., Filas R. W., Goodby J. W., Leslie T. M., **Polarization Absorption-Spectroscopy—Determination of the Direction and Degree of Orientation of Absorption Transitions.** *J Appl Phys* 53(9):6127-6136, 1982.
- Dentai A. G., Burrus C. A., Lee T. P., Campbell J. C., Copeland J. A., Oliver J. D., **LPE InGaAs/InP With N_D-N_A Greater-Than $5 \times 10^{14} \text{ cm}^{-3}$ for Photosensitive Devices.** *Inst Phys C*1982(63):467-471, 1982.
- Donnelly V. M., Flamm D. L., Collins G., **Laser Diagnostics of Plasma-Etching—Measurement of Cl_2^+ in a Chlorine Discharge.** *J Vac Sci T* 21(3):817-823, 1982.

- Donnelly V. M., Karlicek R. F., **Development of Laser Diagnostic Probes for Chemical Vapor-Deposition of InP/InGaAsP Epitaxial Layers.** *J Appl Phys* 53(9):6399-6407, 1982.
- Dudderar T. D., Gilbert J. A., **Fiber optic Measurement of the Deformation Field on a Remote Surface Using Numerically Processed White-Light Speckle.** *Appl Optics* 21(19):3520-3527, 1982.
- Dupuis R. D., Lynch R. T., Thurmond C. D., Bonner W. A., **Growth of InP by Metalorganic Chemical Vapor-Deposition (MOCVD).** *P Soc Photo* 323:131-136, 1982.
- Eisenstein G., Vitello D., **Chemically Etched Conical Microlenses for Coupling Single-Mode Lasers into Single-Mode Fibers.** *Appl Optics* 21(19):3470-3474, 1982.
- Fleury P. A., **(RS) Phase-Transitions, Critical Phenomena and Instabilities (Review or Bibliog.).** *Usp Fiz Nau* 138(1):129-145, 1982.
- Fork R. L., **Physics of Optical Switching.** *Phys Rev A* 26(4):2049-2064, 1982.
- Forrest S. R., Kaplan M. L., Schmidt P. H., Venkatesan T., Lovinger A. J., **Large Conductivity Changes in Ion-Beam Irradiated Organic Thin-Films.** *Appl Phys L* 41(8):708-710, 1982.
- Gallagher P. K., Gyorgy E. M., Jones W. R., **An Evolved Gas-Analysis Study of the Reduction of Nickel-Oxide by Hydrogen.** *J Therm Ana* 23(1-2):185-192, 1982.
- Gedanken A., Robin M. B., Kuebler N. A., **Non-Linear Photochemistry in Organic, Inorganic, and Organo-Metallic Systems.** *J Phys Chem* 86(21):4096-4107, 1982.
- Gibbs H. M., Jewell J. L., Moloney J. V., Tarnig S. S., Tai K., Watson E. A., Gossard A. C., McCall S. L., Passner A., Venkatesan T. N., **Switching of a GaAs Bistable Etalon—External Switching On and Off, Regenerative Pulsations, Transverse Effects, and Lasing.** *P Soc Photo* 321:67-74, 1982.
- Gibson J. M., Bean J. C., Poate J. M., Tung R. T., **The Effects of Nucleation and Growth on Epitaxy in the CoSi₂/Si System.** *Thin Sol Fi* 93(1-2):99-108, 1982.
- Ginsberg A. P., Lindsell W. E., Sprinkle C. R., West K. W., Cohen R. L., **Disulfur and Diselenium Complexes of Rhodium and Iridium.** *Inorg Chem* 21(10):3666-3681, 1982.
- Harrison T. R., Johnson A. M., Tien P. K., Dayem A. H., **NiSi₂-Si Infrared Schottky Photodetectors Grown by Molecular-Beam Epitaxy.** *Appl Phys L* 41(8):734-736, 1982.
- Heaven M. C., Clyne M. A. A., **Interpretation of the Spontaneous Predissociation of Cl₂[B³π(0_u)].** *J Chem S F2* 78(P8):1339-1344, 1982.
- Heaven M., Miller T. A., English J. H., Bondyby V. E., **Laser-Induced Fluorescence-Spectra of YAG-Laser Vaporized-SE₂.** *Chem P Lett* 91(4):251-257, 1982.
- Heimann P. A., Murarka S. P., Sheng T. T., **Electrical-Conduction and Breakdown in Oxides of Polycrystalline Silicon and Their Correlation With Interface Texture.** *J Appl Phys* 53(9):6240-6245, 1982.
- Huse D. A., **Tricriticality of Interacting Hard Squares—Some Exact Results.** *Phys Rev L* 49(16):1121-1124, 1982.
- Jackel J. L., Rice C. E., Veselka J. J., **Proton-Exchange for High-Index Wave-Guides in LiNbO₃.** *Appl Phys L* 41(7):607-608, 1982.
- Jackel J. L., Rice C. E., **Variation in Wave-Guides Fabricated by Immersion of LiNbO₃ in AgNO₃ and TiNO₃—The Role of Hydrogen.** *Appl. Phys L* 41(6):508-510, 1982.
- Jayarama A., Batlogg B., Maines R. G., Bach H., **Effective Ionic Charge and Bulk Modulus Scaling in Rocksalt-Structured Rare-Earth Compounds.** *Phys Rev B* 26(6):3347-3351, 1982.
- Jin B. J., Lin T. H., Chu C. W., Shen Y. S., **Magnetic Phase-Transitions in CrBe₁₂ Under High-Pressures.** *Phys Rev B* 26(7):3878-3881, 1982.
- Karlicek R. F., Donnelly V. M., Johnston W. O., **Laser Spectroscopic Monitoring of a Hydride Transport Vapor-Phase Epitaxy (VPE) Reactor.** *P Soc Photo* 323:62-66, 1982.
- Lax M. et al., **Non-Adiabatic Formulation for Radiationless Transitions Induced by Classical Lattice-Vibrations.** *Phys Rev B* 26(7):3547-3558, 1982.
- Leung S. Y., Schumake N. E., **Slider Induced Convection in Horizontal Liquid-Phase Epitaxy (LPE) System.** *P Soc Photo* 323:156-163, 1982.
- Levy N., **Analysis of an Epoxy Curing Reaction by Differential Scanning Calorimetry (Review or Bibliog.).** *Acs Symp S* 1982(197):313-327, 1982.

- Liao P. F., Stern M. B., **Surface-Enhanced Raman-Scattering on Gold and Aluminum Particle Arrays.** *Optics Lett* 7(10):483-485, 1982.
- Liminga R. et al., **Gamma-Lithium Iodate Structure at 515 K and the α -LiIO₃ to γ -LiIO₃ to β -LiIO₃ Phase-Transitions.** *J Chem Phys* 77(8):4222-4226, 1982.
- Liu P. L., **Bandwidth, Field Distribution, and Optimal Electrode Design for Wave-Guide Modulators.** *J Appl Phys* 53(10):6681-6686, 1982.
- Lloyd J. R., Nakahara S., **Formation and Growth of Voids and or Gas-Bubbles in Thin-Films.** *Thin Sol Fi* 93(3-4):281-286, 1982.
- Logan R. A., **Optical-Device Structures Grown by Liquid-Phase Epitaxy (LPE).** *P Soc Photo* 323:150-155, 1982.
- Lourenco J. A., **The Effect of LPE Growth Atmosphere on Thermal-Decomposition of (100) S-InP Substrates.** *J Cryst Gr* 59(3):563-566, 1982.
- Marra W. C. et al., **X-Ray-Diffraction Studies—Melting of Pb Monolayers on Cu(110) Surfaces.** *Phys Rev L* 49(16):1169-1172, 1982.
- McIlrath T. J., Freeman R. R., **Laser Techniques for Extreme Ultraviolet Spectroscopy (Editorial).** *AIP Conf Pr* 1982(90):1-8, 1982.
- Miller D. A. B., Chemla D. S., Eilenberger D. J., Smith P. W., Gossard A. C., Tsang W. T., **Large Room-Temperature Optical Nonlinearity in GaAs/Ga_{1-x}Al_xAs Multiple Quantum Well Structures.** *Appl Phys L* 41(8):679-681, 1982.
- Mills A. P., **Surface-Analysis and Atomic Physics With Slow Positron Beams.** *Science* 218(4570):335-340, 1982.
- Nakahara S., **Phenomenological Description of Thin-Film Inter-Diffusion.** *P Soc Photo* 346:39-46, 1982.
- Orenstein J., Baker G. L., **Photogenerated GaP States in Polyacetylene.** *Phys Rev L* 49(14):1043-1046, 1982.
- Oron M., Tamari N., Shirikman H., Burrus C. A., **Lasing Properties of InGaAsP Buried Heterojunction Lasers Grown on a Mesa Substrate.** *Appl Phys L* 41(7):609-611, 1982.
- Paek U. C., Peterson G. E., Carnevale A., **Effects of Depressed Cladding on the Transmission Characteristics of Single-Mode Fibers With Graded-Index Profiles.** *Appl Optics* 21(19):3430-3436, 1982.
- Pearsall T. P., Hermann C., **Direct Measurement of Alloy Disorder in Lattice-Matched Ga_xIn_{1-x}AsP_{1-y} Alloys ($y \approx 2.2x$).** *Inst Phys C* 1982(63):269-274, 1982.
- Petroff P. M., Gossard A. C., Logan R. A., Wiegmann W., **Toward Quantum Well Wires—Fabrication and Optical-Properties.** *Appl Phys L* 41(7):635-638, 1982.
- Pfeiffer L., Walstedt R. E., Bell R. F., Kovacs T., **Temperature-Dependence of the Orthorhombic Charge-Density-Wave Parameters in 2H-TaSe₂ by Se-77 Nuclear Magnetic-Resonance.** *Phys Rev L* 49(16):1162-1165, 1982.
- Phillips T. G., Dolan G. J., **SiS Mixers.** *Physica B&C* 110(1-3):2010-2019, 1982.
- Portal J. C., Nicholas R. J., Brummell M. A., Cho A. Y., Cheng K. Y., Pearsall T. P., **Quantum Transport in GaInAs-AlInAs Heterojunctions, and the Influence of Intersubband Scattering.** *Sol St Comm* 43(12):907-911, 1982.
- Powers L., **X-Ray Absorption-Spectroscopy Application to Biological Molecules (Review or Bibliog.).** *Bioc Biop A* 683(1):1-38, 1982.
- Presby H. M., **Geometric Measurement of Preform Rods and Starting Tubes.** *Appl Optics* 21(19):3528-3530, 1982.
- Revesz A. G., Wemple S. H., **The Optical-Properties of Noncrystalline Silicon and Si_{1-x}H_x Films.** *Phys St S-A* 72(2):721-729, 1982.
- Reynolds A. H., Straub K. D., Rentzepis P. M., **Picosecond Spectroscopy of Cu(II) Cytochrome-C.** *Biophys J* 40(1):27-31, 1982.
- Reynolds C. L., Tamargo M. C., Gaw C. A., **Influence of Cooling Rate on the Short-Time LPE Growth of the Active Layer in (AlGa)As DH Lasers.** *J Cryst Gr* 59(3):525-530, 1982.
- Rogovin D., Nagel J., **Quantum-Theory of the DC Josephson Effect—Static Tunneling Characteristics of Ultra-Small Josephson-Junctions.** *Phys Rev B* 26(7):3698-3732, 1982.
- Rosenbaum T. F., Rupp L. W., Thomas G. A., Chen H. S., Banavar J. R., Varma C. M., **Observation of Electron-Spin-Resonance Nonlinearities in a Spin-Glass (Letter).** *J Phys C* 15(27):L975-L979, 1982.
- Schneemeyer L. F., Miller B., **InP and CdS Photo-Anodes in Concentrated Aqueous Iodide Electrolytes.** *J Elchem So* 129(9):1977-1981, 1982.

- Schwartz G. P., Sunder W. A., Griffiths J. E., Gualtieri G. J., **Condensed Phase-Diagram for the In-As-O System.** *Thin Sol Fi* 94(3):205-212, 1982.
- Schwartz G. P., Griffiths J. E., Gualtieri G. J., **Thermal-Oxidation and Native Oxide Substrate Reactions on InAs and InXGa_{1-X}As.** *Thin Sol Fi* 94(3):213-222, 1982.
- Shay J. L., Schiavone L. M., Epworth R. W., Taylor D. W., **Performance of a 7-Segment Iridium Oxide Electrochromic Display.** *J Appl Phys* 53(9):6004-6006, 1982.
- Silfvast W. T., Wood O. R., **Recombination Lasers in the Vacuum Ultra-violet.** *AIP Conf PR1982(90)*:128-136, 1982.
- Skocpol W. J., Jackel L. D., Hu E. L., Howard R. E., Fetter L. A., **One-Dimensional Localization and Interaction Effects in Narrow (0.1 μm) Silicon Inversion-Layers.** *Phys Rev L* 49(13):951-955, 1982.
- Skocpol W. J., Voshchenkov A. M., Howard R. E., Hu E. L., Jackel L. D., Epworth R. W., Fetter L. A., Grabbe P., Tennant D. M., **Preliminary Observation of 1-D Effects in Narrow Si MOSFET Structures.** *Physica B&C* 110(1-3):2105-2107, 1982.
- Stolen R. H., Botineau J., Ashkin A., **Intensity Discrimination of Optical Pulses With Birefringent Fibers.** *Optics Lett* 7(10):512-514, 1982.
- Temkin H., Joyce W. B., Chin A. K., DiGiuseppe M. A., Ermanis F., **Effect of p-n-Junction Position on the Performance of InGaAsP Light-Emitting Diodes.** *Appl Phys L* 41(8):745-747, 1982.
- Tenan M. A., Hackwood S., Beni G., **Friction in Capillary Systems.** *J Appl Phys* 53(10):6687-6692, 1982.
- Unguris J., Seiler A., Celotta R. J., Pierce D. T., Johnson P. D., Smith N. V., **Spin-Polarized Inverse Photo-Electron Spectroscopy of Solid-Surfaces—Ni(110).** *Phys Rev L* 49(14):1047-1050, 1982.
- Vanderzi J. P., Tsang W. T., **Integrated Multilayer GaAs-Lasers Separated by Tunnel-Junctions.** *Appl Phys L* 41(6):499-501, 1982.
- Vandover R. B., Bacon D. D., Sinclair W. R., **Superconductive Tunneling Into NbN Deposited Near Room-Temperature.** *Appl Phys L* 41(8):764-766, 1982.
- Wagner W. R., Cho A. Y., **Al_{0.3}Ga_{0.7}P_{0.01}As_{0.99} GaAs-Laser Heterostructures Grown by Molecular-Beam Epitaxy.** *J Appl Phys* 53(9):6032-6036, 1982.
- Wang T. T., West J. E., **Polarization of Poly(Vinylidene Fluoride) by Application of Breakdown Fields.** *J Appl Phys* 53(10):6552-6556, 1982.
- Weber T. A., Stillinger F. H., **Dynamical Study of the H₅O₂⁺ + H₃O₂⁻ Neutralization Reaction Using the Polarization Model.** *J Chem Phys* 77(8):4150-4155, 1982.
- White J. C., Henderson D., **Anti-Stokes Raman Lasers.** *AIP Conf PR1982(90)*:117-127, 1982.
- Woodruff D. P., Smith N. V., Johnson P. D., Royer W. A., **K-Resolved Inverse Photo-Electron Spectroscopy and its Application to Cu(001), Ni(001), and Ni(110).** *Phys Rev B* 26(6):2943-2955, 1982.

SOCIAL AND LIFE SCIENCES

- Chance B., Moore J., Powers L., Ching Y., **A Redox Equilibrator for the Preparation of Cytochrome-Oxidase of Mixed-Valence States and Intermediate Compounds for X-Ray Synchrotron Studies.** *Analyt Bioc* 124(2):239-247, 1982.
- Chance B. et al., **Synchrotron X-Ray Studies of Biological Preparations at Low-Temperatures With Optical Monitoring of Sample Integrity.** *Analyt Bioc* 124(2):248-257, 1982.
- Day M. C., Stone C. A., **Developmental and Individual-Differences in the Use of the Control-of-Variables Strategy.** *J Educ Psych* 74(5):749-760, 1982.
- Friedman J. M., Stepnoski R. A., Noble R. W., **Time-Resolved Resonance Raman Studies of Carp Hemoglobin.** *FEBS Letter* 146(2):278-282, 1982.
- Krauskopf J., Williams D. R., Heeley D. W., **Cardinal Directions of Color Space.** *Vision Res* 22(9):1123-1131, 1982.
- Weisstein N., Williams M. C., Harris G. S., **Depth, Connectedness, and Structural Relevance in the Object-Superiority Effect—Line Segments Are Harder to See in Flatter Patterns.** *Perception* 11(1):5-17, 1982.

CONTENTS, MARCH 1983

Part 1

Measurements of Selective Near-In Sidelobe Reduction of a Pyramidal, Horn-Reflector Antenna

R. A. Semplak

An Experimental Study of Atmospheric Optical Transmission

B. G. King, P. J. Fitzgerald, and H. A. Stein

Star Network With Collision-Avoidance Circuits

A. Albanese

Pressure-Volume-Temperature Behavior in the System $\text{H}_2\text{O-NaOH-SiO}_2$ and Its Relationship to the Hydrothermal Growth of Quartz

E. D. Kolb, P. L. Key, R. A. Laudise, and E. E. Simpson

Variable Rate ADPCM Systems Based on Explicit Noise Coding

N. S. Jayant

Random Processes With Specified Spectral Density and First-Order Probability Density

M. M. Sondhi

A Method to Characterize the Mechanical Properties of Undersea Cables

T. C. Chu

Performance of a Fast Algorithm for FIR System Identification Using Least Squares Analysis

S. L. Marple, Jr., and L. R. Rabiner

Part 3

TRAFFIC SERVICE POSITION SYSTEM NO. 1B

Overview and Objectives

R. E. Staehler and J. I. Cochrane

System Description

N. X. DeLessio and N. A. Martellotto

Real-Time Architecture Utilizing the DMERT Operating System

R. J. Gill, G. J. Kujawinski, and E. H. Stredde

Hardware Configuration

G. T. Clark, H. A. Hilsinger, J. H. Tendick, and R. A. Weber

Software Development System

T. G. Hack, T. Huang, and L. C. Stecher

Integration and System Testing

R. Ahmari, R. S. DiPietro, S. C. Reed, and J. R. Williams

Retrofitting the Processor

J. C. Dalby, Jr., D. Van Haften, and L. A. Weber

Capacity and Reliability Evaluation

B. A. Crane and D. S. Suk

Switching Control Center System Interface

J. J. Bodnar, J. R. Daino, and K. A. VanderMeulen

TSPS No. 1/1B Long-Range Planning Tools

P. L. Bastien and B. R. Wycherley

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.



Bell System