January 1984
Vol. 63 No. 1

# AT&T BELL LABORATORIES TECHNICAL JOURNAL

A JOURNAL OF THE AT&T COMPANIES

AT&T BELL LABORATORIES TECHNICAL JOURNAL (ISSN0005-8580) is published by AT&T, 550 Madison Avenue, New York, NY 10022; 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 ten times each year. The Computing Science and Systems section and the special issues are included as they become available. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. A subscription to the Computing Science and Systems section only is $10 ($12 foreign). Single copies of most issues 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 Technical Journal and sent to AT&T Bell Laboratories, Circulation Dept., Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, AT&T Bell Laboratories Technical Journal, Room 1J319, 101 J. F. Kennedy Pky, Short Hills, NJ 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 AT&T. 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, NJ 07078 and additional mailing offices. Postmaster: Send address changes to the AT&T Bell Laboratories Technical Journal, Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

AT&T BELL LABORATORIES

# Technical Journal

Vol. 63 JANUARY 1984 No. 1

# A New Beginning

This issue of the *Journal* continues a publishing venture that began in 1922. Behind us we have sixty-two years of publishing the results of telecommunications research and development in the AT&T companies. This month—January 1984—we begin a new era, with new goals and a restructured company. And the *Journal* has a new name, the *AT&T Bell Laboratories Technical Journal.*

In the 1922 inaugural issue, the editor stated the rationale for establishing the *Journal*: to bring together in one place significant papers on electrical communications. From the perspective of today's level of technical sophistication, the term "electrical communication" may seem somewhat archaic, but first-class research and engineering were done in those early years, resulting in innovations that laid the groundwork for the communications explosion we are witnessing today.

Creative research and development arise out of the needs of a culture and the talents of its human resources and, most importantly, out of the freedom of intellectual exchange that we value so highly. As participants in this very human process, the *Journal* staff and its advisers have been rewarded by a sense of satisfaction in being part of the process.

Now we anticipate the great scientific and technological challenges that confront us in this new era. And we intend that the *Journal*—representing all of the AT&T companies—will continue to publish papers on the variety of technology being investigated to meet these challenges.

Editor

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# A Probabilistic Model for the Performance of Word Recognizers

By A. E. ROSENBERG*

(Manuscript received February 7, 1983)

This paper develops a probabilistic model to account for the error-rate behavior of isolated-word speech-recognition systems. It examines two kinds of errors, confusion error, an a priori characterization of a recognizer, which measures differences between words, and recognition (rank) error, an a posteriori characterization, which, in addition to taking into account differences between words, accounts for differences between different tokens of the same word. It is shown that these kinds of errors can be modeled by describing recognition trials as Bernoulli trials. Good models of error-rate behavior as a function of vocabulary size can be obtained if the distributions of confusion and recognition (rank) number are considered to be mixtures of binomial distributions. The data obtained from a recent experiment in isolated-word recognition with a large vocabulary (1109 words) are used to evaluate the model. Experimental error-rate functions obtained from each of six talkers and three partitions of the vocabulary are fit by means of an optimization algorithm to model functions based on mixture distributions. The results indicate that two-way mixture distributions account quite well for the experimental performance results.

## I. INTRODUCTION

A critical concern in the study and development of automatic speech-recognition systems is specification of their performance. Performance is typically specified by recognition error rate, which is the

* AT&T Bell Laboratories.

fraction of trials in a test of the system in which incorrect decisions are obtained. This specification should be accompanied by a description of the test vocabulary, the talker population, the talking environment, and other pertinent conditions relating to both the training and testing of the system. The interaction among these factors and their effect on performance are not well understood. Indeed, altering a variable associated with any of these factors can change the performance of a system in often unpredictable and drastic ways.

A more general specifier of recognizer performance is the rate at which the best $n$ choices offered by the recognizer contain the correct word. More recently, specifiers measuring "complexity"[1] and efficiency[2] have been introduced. The relationship among specifiers is another aspect of recognizer performance that is not well understood.

It is the purpose of this paper to examine and establish probabilistic models to describe the performance of isolated-word speech-recognition systems and to relate various performance specifiers. The distinction will be made between performance specifiers that characterize systems through the training phases of the system and those that characterize the overall behavior in test use of the system. We will focus on modeling performance behavior as a function of vocabulary size for a given recognizer, over a small population of talkers, and three types of vocabularies. Some speculation will be offered on the relation between model parameters and the recognizer, talker, and vocabulary.

The paper is organized as follows. In Section II, performance measures are defined and the probabilistic models, which form the basis for describing the behavior of these measures as a function of vocabulary size, are introduced. In Section III we make use of data obtained in an experiment with a speaker-dependent isolated-word recognizer using a large vocabulary to illustrate the behavior of some of the performance measures and evaluate how well the probabilistic models account for the behavior. Section IV presents a discussion that offers some speculation regarding the significance of the parameters that specify the models. Section V presents some conclusions.

## II. DEFINITIONS AND PROBABILISTIC MODELS

### 2.1 *Bernoulli trials as the basis for confusion and rank*

Suppose we have a vocabulary of $N$ words, $V = \{v_1, v_2, \cdots, v_N\}$. Let $d_{ij}$ be a distance measure between a token of word $v_i$ and a token of word $v_j$. The source for this distance might be some perceptual experiment, a phonetic or linguistic measurement, or the output of an automatic recognizer. In what follows the distance is considered to be the output of a recognizer. As the output of a recognizer it will normally be assumed that the first index, $i$, refers to an input test word while the second index, $j$, refers to a (single) prototype word.

The experiment underlying the formulations that follow is the comparison of a token of an input word $v_I$ with the prototype for each of the remaining $N - 1$ words in the vocabulary.

Suppose we are concerned with a particular word, $v_I$, in the vocabulary. Consider two events

$$d_{Ij} < T, \qquad j \neq I \tag{1}$$

and

$$d_{Ij} < d_{II}, \qquad j \neq I, \tag{2}$$

where $T$ is some preassigned distance threshold and $d_{II}$ is a "self-distance". Note that self-distance, $d_{II}$, generally represents the distance between two different tokens of a word, $v_I$, and therefore, must be greater than zero.

Now consider the number of occurrences of these events in the underlying experiment, defined as follows:

$$q_I(T) = |\{j \neq I\colon d_{Ij} < T\}| \tag{3}$$

and

$$r_I = |\{j \neq I\colon d_{Ij} < d_{II}\}|, \tag{4}$$

where $|\{\quad\}|$ is the cardinality or count of the events in the brackets. Note that $q_I(T)$ is the basis for the notion of confusability or complexity introduced in Rabiner et al.[1] whereas $r_I + 1$ is the rank of the correct word input to a recognizer. When $r_I = 0$, the best matching reference prototype corresponds to the correct word $v_I$.

If, in eq. (3), both $I$ and $j$ represent reference word tokens, $q_I(T)$ can be considered to characterize a recognizer through the training phase of the system, in other words, an a priori characterization. In eq. (4), however, the self-distance, $d_{II}$, specifically represents the distance between a test word input and the reference prototype for that word. Thus $r_I$ characterizes a system in its test or use phase, and is therefore an a posteriori characterization.

Consider now a probabilistic formulation that can be applied to either of the events defined in (1) and (2). Define a random variable $X_{Ij}$ such that

$$X_{Ij} = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

with

$$\text{Prob}\{X_{Ij} = 1\} = p_I \tag{6}$$

and

$$\text{Prob}\{X_{Ij} = 0\} = 1 - p_I. \tag{7}$$

Each event [(1) or (2)] is thus considered to be a Bernoulli trial. We assume that the Bernoulli probability is independent of the reference word $j$. When referring to event (1), $p_I$ depends on $T$, but this is omitted to keep the notation simple. The counts defined in eqs. (3) and (4) are thus sums over $j$ of the random variable $X_{Ij}$. We denote this sum generically as $s_I$, it being understood that $s_I = q_I(T)$ when referring to events of type (1), and $s_I = r_I$ when referring to events of type (2). We can see that

$$s_I = \sum_{j \neq I} X_{Ij}, \tag{8}$$

from which it follows that $0 \leq s_I \leq N - 1$. Given eqs. (6), (7), and (8) we obtain

$$\mathscr{E}\{s_I\} = (N - 1)p_I \tag{9}$$

and

$$\text{Var}\{s_I\} = (N - 1)p_I(1 - p_I). \tag{10}$$

Also, the probability that $s_I$ assumes a particular value $k$, $0 \leq k \leq N - 1$, obeys the binomial probability law

$$\text{Prob}\{s_I = k\} = \binom{N-1}{k} p_I^k (1 - p_I)^{N-1-k}, \tag{11}$$

where $\binom{N-1}{k}$ is a binomial coefficient.

Two kinds of error measures are introduced. An error measure is a monotonically increasing function of $s_I$ that equals 0 when $s_I$ equals 0, and approaches or equals 1 when $s_I$ equals $N - 1$.

The first error measure, $e_I$, is defined as

$$e_I = 1 - \frac{1}{1 + s_I}, \tag{12}$$

from which it follows that $0 \leq e_I \leq 1 - 1/N$. When $s_I$ is understood to be $q_I(T)$, $e_I$ is similar to confusability or complexity error as defined in Rabiner et al.[1] When $s_I = r_I$, $e_I$ is related to the notion of "efficiency" introduced by Smith and Erman[2] to characterize recognizer performance. Using eqs. (11) and (12) it can be shown that

$$\mathscr{E}\{e_I\} = \sum_{k=0}^{N-1} P\{s_I = k\}\left(1 - \frac{1}{1+k}\right)$$

$$= 1 - \frac{1 - (1 - p_I)^N}{p_I N}. \tag{13}$$

The second error measure is associated with the occurrence of any nonzero value of $s_I$, that is, any confusion, $q_I(T)$, at all, or any rank, $r_I + 1$, other than 1. Define

$$E_I = \begin{cases} 1 & \text{if} \quad s_I > 0 \\ 0 & \text{if} \quad s_I = 0. \end{cases} \tag{14}$$

This is the conventional or standard recognition error generally used to characterize the performance of automatic recognizers. From eq. (11) we have

$$\text{Prob}\{s_I > 0\} = 1 - \text{Prob}\{s_I = 0\} = 1 - (1 - p_I)^{N-1}. \tag{15}$$

Then it follows that

$$\mathscr{E}\{E_I\} = 1 - (1 - p_I)^{N-1}. \tag{16}$$

Often recognition-error rates are calculated to provide the frequency with which the correct word is included among the top $c$ choices provided by the recognizer $s_I = 0, 1, 2, \cdots, c - 1$. This represents a generalization of the preceding definition for $E_I$ which is expressed by

$$E_I(c) = \begin{cases} 1 & \text{if} \quad s_I \geq c \\ 0 & \text{if} \quad s_I < c. \end{cases} \tag{17}$$

Since

$$\text{Prob}\{s_I \geq c\} = \sum_{k=c}^{N-1} p_I^k (1 - p_I)^{N-1-k} \tag{18}$$

we have

$$\mathscr{E}\{E_I(c)\} = \sum_{k=c}^{N-1} p_I^k (1 - p_I)^{N-1-k}. \tag{19}$$

Although the models which are derived in this paper could easily include this generalization, we restrict our attention to the case for which $c = 1$, referred to as standard error.

### *2.2 Mixture models*

The foregoing formulas pertain to a single word $v_I$ in a vocabulary $V$. Our object is to model behavior of confusability or rank over an entire vocabulary $V$. It is therefore necessary to make some assumptions about the behavior of $p_I$ over all the words in $V$. The simplest possible assumption is that $p_I$ is constant over $V$, i.e., $p_I = p$ for $I = 1, 2, \cdots, N$. It will be shown in the following sections that this assumption leads to very poor models of the actual experimental behavior.

A more general assumption is the following. Assume that the Bernoulli probability defined in eq. (6) is itself a random variable, $p_V$, which may assume different values from word to word in a vocabulary, or indeed, from trial to trial of the same word. Suppose there are $M$ values $p_V$ can assume, $p_m$, $m = 1, 2, \cdots, M$,* such that

$$\text{Prob}\{p_V = p_m\} = h_m, \qquad m = 1, 2, \cdots, M \tag{20}$$

with

$$\sum_{m=1}^{M} h_m = 1, \tag{21}$$

where $h_m$ is the probability that $p_V$ assumes the value $p_m$. (It is possible to generalize still further by assuming $p_V$ to be continuously distributed.) It is now possible to generalize $s_I$ to $s_V$ over the entire vocabulary $V$. From eqs. (11) and (16) we obtain

$$\text{Prob}\{s_V = k\} = \sum_{m=1}^{M} h_m \binom{N-1}{k} p_m^k (1 - p_m)^{N-1-k}. \tag{22}$$

This expression represents a so-called compound binomial distribution or mixture of binomial distributions.[3,4] With this interpretation, each time we perform the underlying experiment represented by (1) or (2) the probability assumed one of the $M$ values $p_m$ over all the $N - 1$ comparisons with the words in $V$. (This is in contrast to the situation in which the probability may assume different values for each comparison in the underlying experiment.) Using eq. (22), general expressions can be obtained for the mean and variance of $s_V$ and for the two generalized error formulations $e_V$ and $E_V$. All of these have the same form as eq. (22), that is, $\mathscr{E}\{\cdot\} = \sum_{m=1}^{M} h_m \mathscr{E}\{\cdot \mid m\}$. Thus,

$$\mathscr{E}\{s_V\} = (N - 1)\bar{p}_V \tag{23}$$

and

$$\text{Var}\{s_V\} = (N-1) \sum_{m=1}^{M} h_m p_m (1 - p_m) + (N-1)^2 \sum_{m=1}^{M} h_m (p_m - \bar{p}_V)^2, \tag{24}$$

* Note that the index $m$ on $p$ no longer refers in general to individual words in the vocabulary as in eq. (6).

where

$$\bar{p}_V = \sum_{m=1}^{M} h_m p_m. \tag{25}$$

Also,

$$\begin{aligned}\mathscr{E}\{\mathrm{ev}\} &= \sum_{m=1}^{M} h_m \left(1 - \frac{1-(1-p_m)^N}{p_m N}\right) \\ &= 1 - \sum_{m=1}^{M} h_m \left(\frac{1-(1-p_m)^N}{p_m N}\right),\end{aligned} \tag{26}$$

and

$$\mathscr{E}\{E_V\} = 1 - \sum_{m=1}^{M} h_m (1-p_m)^{N-1}. \tag{27}$$

With $M$ set to 1, these expressions revert to the form of the earlier expressions for a single word $I$.

## III. EXPERIMENTAL EVALUATION

The experimental data used in this study were obtained using the AT&T Bell Laboratories Linear Predictive Coefficient (LPC) based isolated-word recognition system.[5,6] The vocabulary was the 1109-word so-called Basic English vocabulary of Ogden.[7] The recognizer was used in a speaker-dependent mode. Six native American talkers, three male, three female, participated in the experiment. Each talker trained the system using the robust training procedure of Rabiner and Wilpon,[8] giving a single reference prototype for each word in the vocabulary. In addition, four sets of test utterances were obtained from each talker over a four-week period. Both the training and test utterances were collected over dialed-up telephone lines using an ordinary telephone handset with the talker seated in a sound booth. For each talker, each test word was input to the recognizer and compared with every reference prototype word for that talker. For each such comparison the recognizer provides a distance figure measuring how closely the test word matches a prototype word. In a typical recognition trial the word recognized is associated with the best matching prototype word, that is, the one with the smallest distance. The raw experimental data consist of four sets of 1109 × 1109 distance matrices for each of the six talkers.

The large size of the vocabulary provides an opportunity to investigate recognition performance over a variety of experimental conditions related to vocabulary size and content by choosing appropriate subsets of the whole vocabulary. A series of such experiments using this experimental database has been described in a previous report.[1]

In the present experiment we focus on three partitions of the 1109-word vocabulary, the 605 monosyllabic words contained in the vocabulary, the remaining 504 polysyllabic words, and the entire vocabularly itself. For each of these, randomly selected subsets of various sizes are chosen. The subset sizes chosen are

$$N = 10,\ 20,\ 50,\ 100,\ 200,\ 400,\ (800),\ PARTSIZ,$$

where $PARTSIZ = 605, 504$, or $1109$ for the monosyllabic, polysyllabic, and whole vocabulary partitions, respectively. For each subset size $N$, a total of $MT = \min[50, PARTSIZ/N]$ subsets of words selected at random without replacement from each partition are specified. The same subsets are specified over all test sets and talkers. Thus, in the aggregate, for each subset size $N$, results are obtained over $N^*MT$ different words, where $500 \le N^*MT \le PARTSIZ$.

The experimental performance data that are presented in this paper are generally given as functions of subset size for each talker and vocabulary type, and represent an average over all the talker's four test sets and vocabulary subsets for each subset size.

### *3.1 Experimental performance measures*

This section presents experimental examples of the performance measures introduced in Section II. To recapitulate, confusion number and rank number are defined as follows:

1. $q_I(T)$: confusion number for a given word $v_I \epsilon V$, is the number of words (other than $v_I$) in a given vocabulary subset $V$ whose distance to the given word is less than some threshold $T$ [from eq. (3)].

2. $r_I$: rank number for a given word $v_I \epsilon V$, is the number of words (other than $v_I$) in a given vocabulary subset $V$ whose distance is less than the self-distance for $v_I$ [from eq. (4)].

Experimental averages are obtained as follows. Suppose word $v_I$ is included in $V_m(N)$, where $V_m(N)$ is the $m$th vocabulary subset of size $N$, $m = 1, 2, \cdots, MT$, and $MT$ is the total number of subsets of size $N$. The words in each subset $V_m(N)$ are selected at random without replacement from a vocabulary $V$ of total size $Q \ge N$. Then, given the confusion number and rank number, $q_{I,V_m(N),s,t}(T)$ and $r_{I,V_m(N),s,t}$, respectively, for word $v_I$ from subset $V_m(N)$, test set $t$, and talker $s$, the experimental averages are

$$\bar{q}_{s,V}(N, T) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} q_{I,V_m(N),s,t}(T) \tag{28}$$

and

$$\bar{r}_{s,V}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} r_{I,V_m(N),s,t}\,, \tag{29}$$

respectively. Similarly, for the two error measures that were introduced in Section II, the experimental averages are

$$\bar{e}_{q,s,V}(N, T) = 1 - \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)}\frac{1}{1 + q_{I,V_m(N),s,t}(T)}, \tag{30}$$

and

$$\bar{e}_{r,s,V}(N) = 1 - \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)}\frac{1}{1 + r_{I,V_m(N),s,t}} \tag{31}$$

for the efficiency errors of confusion and rank, respectively, and

$$\begin{aligned}&\bar{E}_{q,s,V}(N, T)\\ &\quad = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\,|\{I\text{: } I \epsilon V_m(N),\ q_{I,V_m(N),s,t}(T) \geq 0\}|\end{aligned} \tag{32}$$

and

$$\bar{E}_{r,s,V}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\,|\{I\text{: } I \epsilon V_m(N),\ r_{I,V_m(N),s,t} \geq 0\}| \tag{33}$$

for the standard errors of confusion and rank, respectively.

Note that both experimental confusion and rank or recognition data are obtained by averaging over four test utterances. In the previous section we noted that confusion could be considered an a priori, training characterization of system performance if it is based on distances between prototypes. Since this is not the case here, the description of confusion does not strictly hold. However, we do not expect distances between test utterances of different words to be significantly different from distances between prototypes of the same words. This point is discussed again in Section IV.

Shown plotted in Fig. 1a as a function of $N$ are experimental averages for confusion number and rank number for talker 3 and the whole vocabulary $V_W$, $\bar{q}_{3,V_W}(N, T)$ and $\bar{r}_{3,V_W}(N)$. Average confusion number is plotted for five threshold values, $T = 0.20, 0.25, 0.30, 0.35$, and $0.40$. For each of these plots, straight-line fits are obtained by least-squares regression. It can be seen that straight-line fits are quite good, each one having a correlation coefficient of better than 0.9995 with the data, with the exception of $\bar{q}_{3,V_W}(N, 0.20)$ and $\bar{r}_{3,V_W}(N)$, whose coefficients of fit are both 0.998. The linear trend is predicted by the model as expressed in eq. (23). The interpretation of the linearity is quite natural. Simply, as the size of the vocabulary grows, the number of confusable words, or the number of words better than the input word (the rank number of the input word), grows proportionately.

For the same talker and vocabulary, and for the same set of thresholds, experimental averages for efficiency error, $\bar{e}_{q,3,V_W}(N, T)$ and

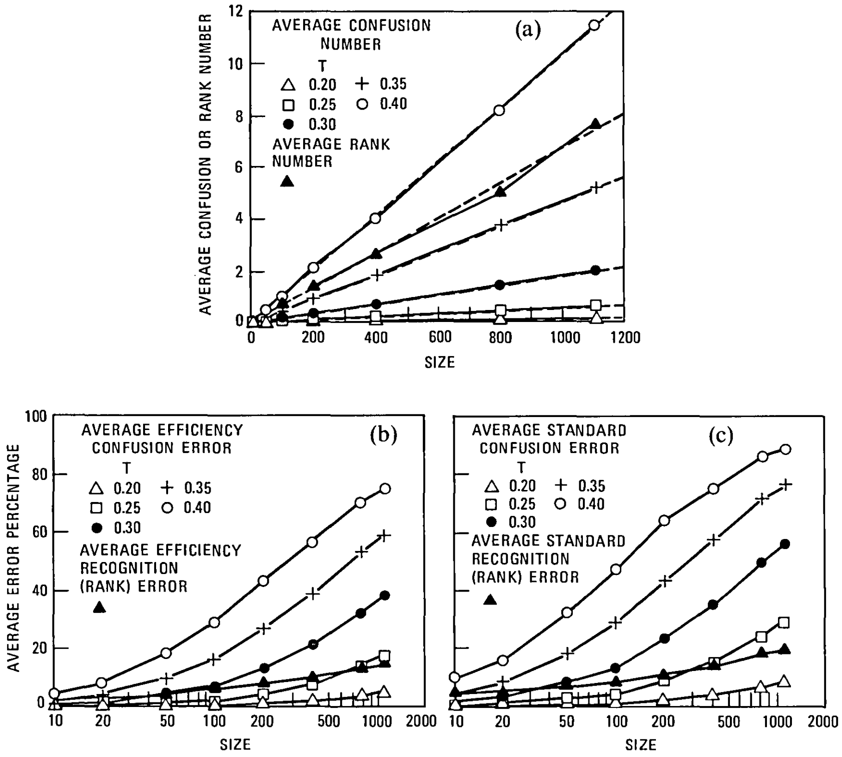


Fig. 1—(a) Average confusion and rank number, (b) average efficiency confusion and rank error, and (c) average standard confusion and rank error as functions of vocabulary size, for talker 3, vocabulary type $V_W$, and five values of threshold, $T$ (for confusion).

$\bar{e}_{r,3,V_W}(N)$, are plotted in Fig. 1b and standard error $\bar{E}_{q,3,V_W}(N, T)$ and $\bar{E}_{r,3,V_W}(N)$ are plotted in Fig. 1c, both as a function of $N$ scaled logarithmically. Both the efficiency- and standard-error curves assume the same trends as a function of vocabulary size, increasing monotonically and approaching one asymptotically. The efficiency-error curves assume uniformly smaller values than their standard error counterparts for each value of $N$. For small $N$ each efficiency-error curve has approximately half the value of its standard-error counterpart. For confusion error, error increases monotonically as the distance threshold, $T$, is increased or relaxed.

The standard recognition- or rank-error curve plotted in Figure 1c is representative of results presented in the earlier report.[1] In the present study, additional data values are presented that extend the curve to vocabulary sizes less than 100. Standard error rate for this talker is approximately 4 percent for 10-word vocabularies, 9 percent for size 100, and 20 percent for the full vocabulary of 1109 words. (The same curve is shown with an expanded error scale in Fig. 2.) In the earlier report it was suggested that for vocabulary sizes greater than 100, doubling the size increases error by a constant amount, a linearly increasing trend with $N$ scaled logarithmically. It can be seen here that with the extension to smaller vocabulary sizes the linear trend is restricted and approximate.

### *3.2 The relation between recognition and confusion error*

The difference in form between the rank- or recognition-error curves and the confusion-error curves, for any threshold value, is quite marked. The relation between recognition error and confusion error as a function of vocabulary size is rather complex.

The following are two hypotheses for relating recognition and confusion error. First, we might examine average confusion number as a function of distance threshold $T$ for any given vocabulary size and find that value of $T$ for which average confusion number is equal to average rank number. For the results shown in Fig. 1a for talker 3 and vocabulary $V_W$, this value of $T$ lies between 0.35 and 0.40. We could reason that confusion-error rates ought to be the same as recognition-error rates for a value of $T$, which on the average includes as many confusable words as the rank of the correct word. However, examining the error rates in Figs. 1b and 1c we find that this hypothesis holds only for the very smallest vocabulary size, $N = 10$. The threshold suggested by this hypothesis leads to confusion-error rates much greater than the recognition-error rates actually observed for larger values of $N$.

The second hypothesis suggests that the appropriate threshold for which confusion-error rates ought to be the same as recognition-error rates can be found by associating the threshold with the average self-distance. We have carried out the calculation of average self-distance for this talker and vocabulary in the same way as the other calculated experimental averages,

$$\overline{DSLF}_{3,V_W}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} d_{II}. \qquad (34)$$

We obtain $\overline{DSLF}_{3,V_W}(N) \approx 0.235$ for all values of $N$. Examining the confusion-error rates in Figs. 1b and 1c, only for the very largest value of $N$, where recognition error is bracketed by confusion error rates for $T = 0.20$ and $T = 0.25$, do we find agreement with this hypothesis.

It is not surprising that average self-distance is independent of $N$, since the similarity between two tokens of the same word is not dependent on the size of the vocabulary from which the words are taken. Nor is it surprising, as we have seen earlier, that the rate of increase in average rank number is independent of $N$. But neither of the associated thresholds is adequate to relate confusion and recognition for all values of $N$. The explanation lies in the following observations. Referring back to the basic definitions expressed in (1) and (2), if the individual self-distance, $d_{II}$, were absolutely constant from trial to trial and from word to word, there would be a threshold equal to this constant for which confusion-error rates would be equal to recognition-error rates. However, even though average self-distance is constant for all $N$, individual self-distances fluctuate widely from trial to trial resulting in rank number distributions which are also quite wide. Thus for small subset sizes, these fluctuations produce an average rank number that is significantly greater than the average confusion number associated with a threshold equal to the average self-distance. However, when vocabulary size grows, so does word density, that is, the average distance between different words decreases. As this occurs the self-distance fluctuations become less important compared with errors attributed to the increasing density. Thus for large vocabularies a threshold equal to average self-distance relates confusion error to recognition error.

The interpretation of the relation between confusion and recognition in the light of model parameters will be brought up in the discussion, Section IV.

### 3.3 Estimation of model parameters and fits to experimental results

It has already been shown that average confusion number and average rank number grow linearly with subset size in agreement with the model as expressed in eq. (23). The slope of this linear function is an estimate of $\bar{p}_V$ given in eq. (25), the average model Bernoulli probability of an error or confusion. Estimates of $\bar{p}_V$ are obtained as the slope estimates for the regression-line fits shown in Fig. 1a. Table I shows these estimates along with the coefficients of fit (correlation coefficients). Since these are linear relations, they present no information concerning individual mixture probabilities, nor, indeed, whether there are any mixtures at all.

The effect of mixtures becomes apparent when we attempt to model

Table I—Linear growth of average confusion and average rank numbers

| $T$ Values | $\bar{p}_V$ Estimates | $r$ Coefficients |
|---|---|---|
| | Average Confusion Number | |
| 0.20 | $1.18 \times 10^{-4}$ | 0.99784 |
| 0.25 | $5.43 \times 10^{-4}$ | 0.99980 |
| 0.30 | $1.77 \times 10^{-3}$ | 0.99995 |
| 0.35 | $4.63 \times 10^{-3}$ | 0.99993 |
| 0.40 | $1.03 \times 10^{-2}$ | 0.99994 |
| | Average Rank Number | |
| — | $6.67 \times 10^{-3}$ | 0.99801 |

error-rate behavior as a function of vocabulary size. To illustrate the effect, the standard-rank error-rate curve shown in Fig. 1b is displayed once more on an expanded error-rate scale in Fig. 2a. (We refer to rank and recognition error rate interchangeably.) Along with this curve, we have plotted the function for expected standard error rate given in eq. (27) with $M$ set to 1 for four different values of $p$. These

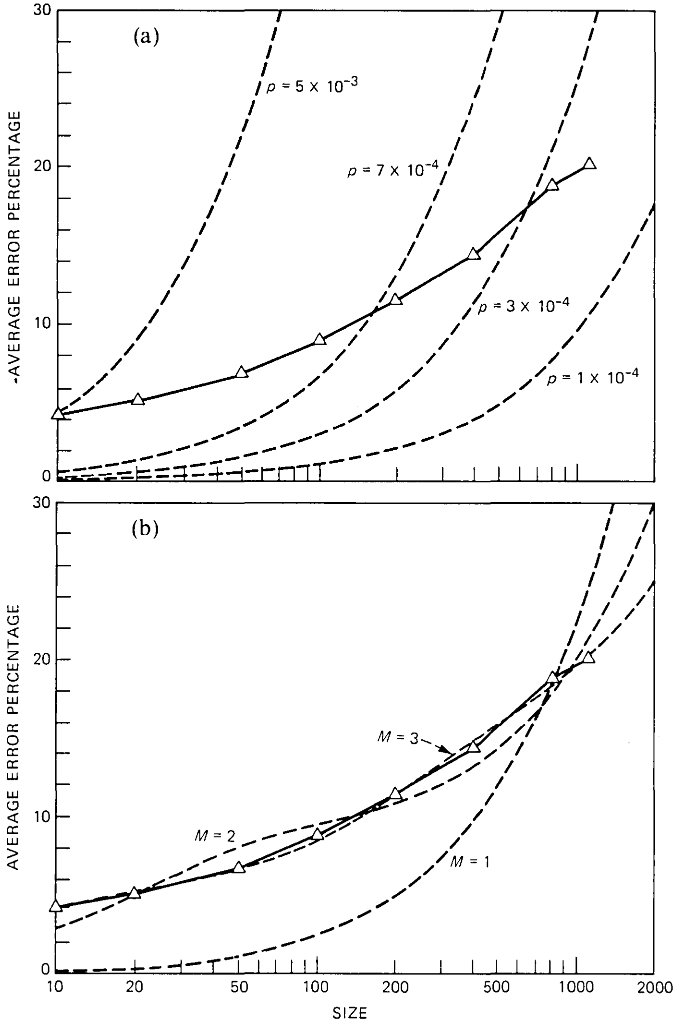


Fig. 2—Average standard recognition (rank) error as a function of vocabulary size for talker 3, and vocabulary type $V_W$ with (a) four different simple (one-way) models of standard error [eq. (27) with $M = 1$], and (b) one-, two-, and three-way mixture-model fits of standard error.

four values of $p$ bracket the range of the values of $\bar{p}_V$ given in Table I. It is quite evident that the simple Bernoulli trial model, obtained with $M$ set to 1 in eq. (27), cannot provide a good fit to the experimental results.

The recognition error-rate curve is plotted once again in Fig. 2b along with the function of eq. (27) for three values of $M$, M = 1, 2, and 3. Table II shows the $h$ and $p$ parameters selected for these functions along with coefficients of fit. The parameters are obtained by using an optimization routine[9] to provide a best fit to the experimental data. We employ the convention that $p_1 \geq p_2 \geq , \cdots , \geq p_M$ and $h_M = 1 - \sum_{m=1}^{M-1} h_m$ to satisfy eq. (21). Note that the fits obtained improve progressively with increasing $M$ and the values of $\bar{p}_V$ obtained for $M = 2$ and $M = 3$ bracket the observed value for the slope estimate for $\bar{r}_{3,V_W}(N)$ given in Table I.

Although it is clear that the three-way mixture model, $M = 3$, provides a superior fit, hereafter we will provide only two-way mixture model fits, $M = 2$. It seems reasonable to expect that models with larger $M$'s should provide better fits because it is reasonable to expect such models to accommodate and discriminate better among all the effects that contribute to the error-rate functions. However, it is also true that there are only eight data points, which is a small number of points to support the number of parameters associated with such models. In addition, there is a certain appeal of parsimony in using two-way models, since it may lead to simpler or more direct interpretations of the parameters. Some suggestions for interpretations are discussed in Section IV. Also, although the two-way model fit is somewhat deficient for the case shown here, in most of the other experimental results that are presented the fits are quite adequate.

For the two-way model, we refer to $p_1$ and $p_2$, $p_1 \geq p_2$, as the type 1 and type 2 population probabilities, respectively, and $h$ (dropping the subscript), as the mixing coefficient for the two populations.

The optimization-fitting procedure is described briefly. The function that is minimized is the sum over the subset sizes of the squared differences between the observed values and the calculated model function value. This function, the gradient of the function, initial values for the parameters, and some convergence constants are supplied to the optimization routine. Usually the routine is run several

Table II—Model parameter estimates for average standard recognition error

| $M$ | $h_1, h_2, \ldots h_{M-1}$ | $p_1, p_2, \ldots p_M$ | $\bar{p}_V$ | $r$ |
|---|---|---|---|---|
| 1 | — | $2.61 \times 10^{-4}$ | $2.61 \times 10^{-4}$ | 0.9607 |
| 2 | 0.085 | $4.32 \times 10^{-2}$, $1.37 \times 10^{-4}$ | $3.82 \times 10^{-3}$ | 0.9856 |
| 3 | 0.046, 0.093 | $1.71 \times 10^{-1}$, $4.72 \times 10^{-3}$, $7.26 \times 10^{-5}$ | $8.35 \times 10^{-3}$ | 0.9990 |

times for each set of experimental points with different sets of initial values to ensure that the optimized parameters represent a global rather than a local minimization. In some cases, particularly for overall low error rates, the minimization is relatively insensitive to variation of the $h$ parameter.

### *3.4 Two-way mixture model fits to experimental data*

This section presents a variety of experimental confusion- and recognition-error results as a function of vocabulary size, together with two-way mixture model fits. The object is to demonstrate that the model represents the error-rate behavior as a function of vocabulary size quite well, and to point out the effects of talker, vocabulary type, etc., on the parameter estimates obtained from the model.

#### *3.4.1 Model fits to recognition error as a function of talker and vocabulary type*

Recognition error rate results as a function of vocabulary size, both efficiency and standard error, are displayed in Figs. 3 and 4, together with model fits for each example. Figure 3 shows results for the three vocabulary types, the whole vocabulary, $V_W$, monosyllables, $V_M$, and polysyllables, $V_P$, for three talkers. Figures 3a through 3c show efficiency-error results for the three talkers while Figs. 3d through 3f show standard-error results. Figure 4 presents results for all six talkers for a single vocabulary type, $V_W$. Fig. 4a presents efficiency-error results and Fig. 4b presents standard-error results. The three talkers selected for Fig. 3 are associated with median performances in Fig. 4. The performance trends of these three talkers for the three vocabulary types in Fig. 3 are representative of all six talkers.

Recall once again the distinction between standard error and efficiency error. Standard error is based on a count of trials with nonzero rank, while efficiency error accounts for the distribution of all rank numbers over all the trials, and is therefore, in some sense, a finer characterization of error performance. The differences between the two are generally predictable, as pointed out in the previous section. Both kinds of error results are shown, principally, to compare the parameter estimates obtained for each model.

The trend in error-rate performance as a function of vocabulary type for individual talkers presented in Fig. 3 is a familiar one. That is, performance degrades for any vocabulary size from the more redundant to the less redundant vocabulary types, from $V_P$ to $V_W$ to $V_M$.

The performance of individual talkers for a single vocabulary type presented in Fig. 4 shows considerable variability. The performance of one talker, talker 4, is distinctly poorer than the others. The best

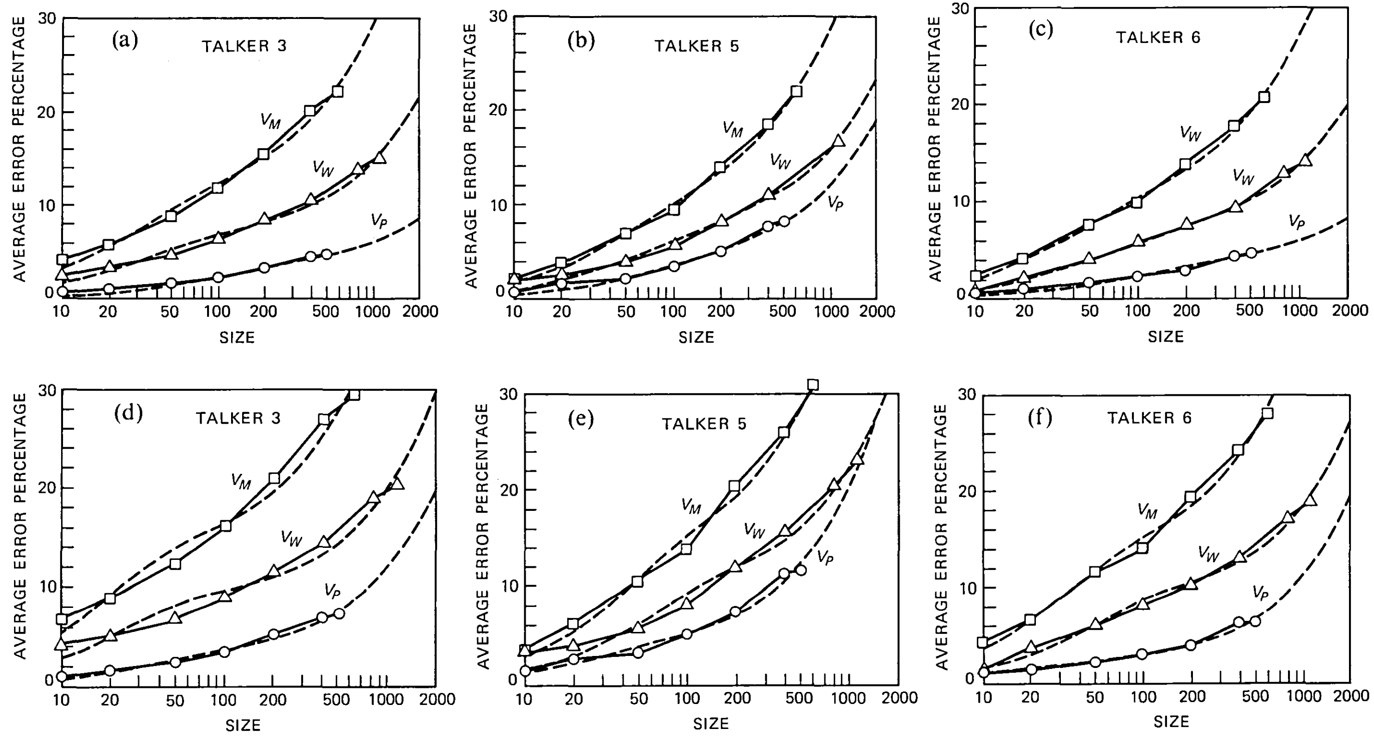


Fig. 3—(a), (b), (c) Average efficiency recognition (rank) error, and (d), (e), (f) average standard recognition (rank) error as a function of vocabulary size, with two-way mixture model fits, for three vocabulary types.

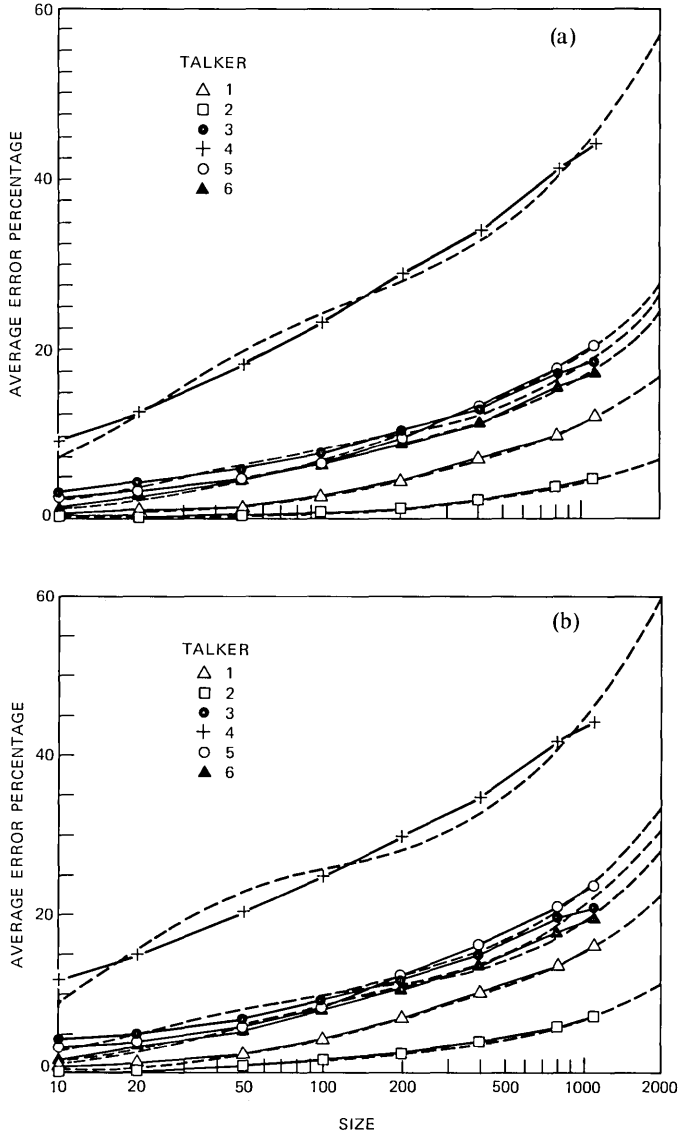


Fig. 4—(a) Average efficiency recognition (rank) error, and (b) average standard recognition (rank) error as a function of vocabulary size, with two-way mixture model fits, for six talkers and vocabulary type $V_W$.

performances are obtained for talkers 1 and 2, while the remaining three are grouped together in an intermediate range of performance.

Model fits have been carried out, as described previously, for both the efficiency and standard-error results for each of the six talkers

Table III—Model parameter estimates for average efficiency recognition (rank) error

| | $V_P$ | | | $V_W$ | | | $V_M$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Talker | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ |
| 1 | 0.037 | $7.85 \times 10^{-3}$ | $4.92 \times 10^{-5}$ | 0.055 | $9.49 \times 10^{-3}$ | $9.80 \times 10^{-5}$ | 0.105 | $8.29 \times 10^{-3}$ | $1.96 \times 10^{-4}$ |
| 2 | 0.020 | $9.99 \times 10^{-4}$ | $2.53 \times 10^{-5}$ | 0.020 | $5.24 \times 10^{-3}$ | $4.17 \times 10^{-5}$ | 0.040 | $9.04 \times 10^{-3}$ | $9.31 \times 10^{-5}$ |
| 3 | 0.040 | $1.99 \times 10^{-2}$ | $5.24 \times 10^{-5}$ | 0.076 | $5.47 \times 10^{-2}$ | $1.68 \times 10^{-4}$ | 0.126 | $6.22 \times 10^{-2}$ | $4.36 \times 10^{-4}$ |
| 4 | 0.119 | $1.09 \times 10^{-1}$ | $5.07 \times 10^{-4}$ | 0.208 | $7.36 \times 10^{-2}$ | $4.21 \times 10^{-4}$ | 0.277 | $8.16 \times 10^{-2}$ | $9.40 \times 10^{-4}$ |
| 5 | 0.046 | $2.23 \times 10^{-2}$ | $1.74 \times 10^{-4}$ | 0.076 | $3.04 \times 10^{-2}$ | $1.99 \times 10^{-4}$ | 0.109 | $3.71 \times 10^{-2}$ | $4.79 \times 10^{-4}$ |
| 6 | 0.038 | $1.69 \times 10^{-2}$ | $5.07 \times 10^{-5}$ | 0.072 | $3.28 \times 10^{-2}$ | $1.52 \times 10^{-4}$ | 0.109 | $4.66 \times 10^{-2}$ | $4.16 \times 10^{-4}$ |
| Mean | 0.050 | $2.95 \times 10^{-2}$ | $1.43 \times 10^{-4}$ | 0.085 | $3.44 \times 10^{-2}$ | $1.75 \times 10^{-4}$ | 0.128 | $4.08 \times 10^{-2}$ | $4.27 \times 10^{-4}$ |
| Standard deviation | 0.035 | $3.98 \times 10^{-2}$ | $1.86 \times 10^{-4}$ | 0.064 | $2.62 \times 10^{-2}$ | $1.30 \times 10^{-4}$ | 0.079 | $2.91 \times 10^{-2}$ | $2.94 \times 10^{-4}$ |
| | | | | | | Without Talker 4 | | | |
| Mean | 0.036 | $1.36 \times 10^{-2}$ | $7.03 \times 10^{-5}$ | 0.060 | $2.65 \times 10^{-2}$ | $1.32 \times 10^{-4}$ | 0.098 | $3.26 \times 10^{-2}$ | $3.24 \times 10^{-4}$ |
| Standard deviation | 0.010 | $8.92 \times 10^{-3}$ | $5.90 \times 10^{-5}$ | 0.024 | $1.99 \times 10^{-2}$ | $6.22 \times 10^{-5}$ | 0.033 | $2.37 \times 10^{-2}$ | $1.69 \times 10^{-4}$ |

| | Mean | | | Standard Deviation | | |
|---|---|---|---|---|---|---|
| Talker | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ |
| 1 | 0.066 | $8.54 \times 10^{-3}$ | $1.14 \times 10^{-4}$ | 0.035 | $8.49 \times 10^{-4}$ | $7.48 \times 10^{-5}$ |
| 2 | 0.027 | $5.09 \times 10^{-3}$ | $5.34 \times 10^{-5}$ | 0.012 | $4.02 \times 10^{-3}$ | $3.54 \times 10^{-5}$ |
| 3 | 0.081 | $4.56 \times 10^{-2}$ | $2.19 \times 10^{-4}$ | 0.043 | $2.26 \times 10^{-2}$ | $1.97 \times 10^{-4}$ |
| 4 | 0.201 | $8.81 \times 10^{-2}$ | $6.23 \times 10^{-4}$ | 0.079 | $1.86 \times 10^{-2}$ | $2.78 \times 10^{-4}$ |
| 5 | 0.077 | $2.99 \times 10^{-2}$ | $2.84 \times 10^{-4}$ | 0.032 | $7.41 \times 10^{-3}$ | $1.69 \times 10^{-4}$ |
| 6 | 0.073 | $3.21 \times 10^{-2}$ | $2.06 \times 10^{-4}$ | 0.036 | $1.49 \times 10^{-2}$ | $1.89 \times 10^{-4}$ |

and each of the three vocabulary types. This includes all the results shown in Figs. 3 and 4 plus the $V_M$ and $V_P$ results for talkers 1, 2, and 4. These are two-way mixture fits obtained by setting $M$ to 2 in eq. (26) for efficiency error and eq. (27) for standard error. Model parameter estimates for efficiency error are presented in Table III. The parameter estimates for standard error are not shown, but comparing them to the efficiency-error estimates indicates general, if not necessarily close, agreement. This agreement reinforces our assumption that the models developed to account for both kinds of error functions are substantially correct since the same experimental data underlay both performance measures.

Table IV compares model fits for average rank, efficiency recognition error, and standard recognition error for each of the six talkers for $V_W$. The table presents estimates for $\bar{p}_V$ and coefficients of fit, $r$. As in Table I, the estimates of $\bar{p}_V$ for average rank are obtained from slope estimates for least-squares regression lines. For efficiency and standard error, the $\bar{p}_V$ estimates are reconstructed using eq. (25). The $\bar{p}_V$ estimates obtained from efficiency and standard-error parameters are in fairly good agreement with each other, but are generally less than half the values of the estimates obtained for average rank. This discrepancy was pointed out in the previous section, where it was implied that it is related to the extent that two-way mixtures model the data compared with models with higher specified values of $M$. An examination of the model function fits in Figs. 3 and 4 and the coefficients of fit in Table IV indicates generally close agreement with the experimental results. The possible exceptions are associated with high error-rate performances (for example, for talkers 3 and 4). For these cases the fits are poorer for the standard-error functions than for the corresponding efficiency-error functions.

As an aid to improve interpretations for the model parameters, it would be useful in an examination of the parameter estimates to detect significant trends associated with the variation of experimental con-

Table IV—Comparison of model fits for average rank, efficiency recognition error, and standard recognition error

| | Average Rank | | Efficiency Recognition Error | | Standard Recognition Error | |
|---|---|---|---|---|---|---|
| Talker | $\bar{p}_V$ | $r$ | $\bar{p}_V$ | $r$ | $\bar{p}_V$ | $r$ |
| 1 | $1.13 \times 10^{-3}$ | 0.99705 | $6.11 \times 10^{-4}$ | 0.99905 | $5.54 \times 10^{-4}$ | 0.99942 |
| 2 | $3.49 \times 10^{-4}$ | 0.97985 | $1.46 \times 10^{-4}$ | 0.99877 | $2.49 \times 10^{-4}$ | 0.99891 |
| 3 | $6.67 \times 10^{-3}$ | 0.99801 | $4.31 \times 10^{-3}$ | 0.99454 | $3.28 \times 10^{-3}$ | 0.98563 |
| 4 | $2.69 \times 10^{-2}$ | 0.99994 | $1.57 \times 10^{-2}$ | 0.99522 | $1.26 \times 10^{-2}$ | 0.98681 |
| 5 | $6.05 \times 10^{-3}$ | 0.99905 | $2.49 \times 10^{-3}$ | 0.99675 | $1.79 \times 10^{-3}$ | 0.99350 |
| 6 | $4.86 \times 10^{-3}$ | 0.99997 | $2.49 \times 10^{-3}$ | 0.99881 | $1.97 \times 10^{-3}$ | 0.99742 |

ditions. In particular, it would be useful to identify those parameters that remain more or less constant over a particular set of conditions. General trends are apparent. As performance degrades, either from one talker to another or from one vocabulary type to another, the estimates of each of the parameters, $h$, $p_1$, and $p_2$, generally increase. Differential trends are harder to detect. No definite conclusions are provided in this set of estimates, but some of it is suggestive.

Table III provides means and standard deviations for each parameter estimate across vocabulary types for each talker and across talkers for each vocabulary type. Since there are only three vocabulary types, caution should be exercised with respect to statistics over this variable. If we use the ratio of the standard deviation to the mean for each parameter as an indicator of variability, we find that $p_1$, across vocabulary types, has consistently less variability than $h$ or $p_2$, with ratios generally less than 0.5 for both efficiency and standard error. Across talkers, suggestions are somewhat vaguer, chiefly because of the especially large variability provided by talker 4. If we disregard the estimates for talker 4, which may be justified by the fact that the two-way mixture fits are rather poor for this talker, then low variability is indicated for the $h$ parameter, and to a lesser extent, for $p_2$, as shown by the second set of means and standard deviations in the table.

Another general observation that can be made is that the ratio of $p_1$ to $p_2$ is of the order of 100 and generally decreases across vocabulary types from $V_P$ to $V_W$ to $V_M$.

### *3.4.2 Model fits to confusion error as a function of threshold, talker, and vocabulary type*

We turn now to two-way mixture models of confusion error and estimates of the model parameters. Experimental confusion error results are shown plotted as a function of vocabulary size in Fig. 5 for talker 6, vocabulary type $V_W$, and seven threshold values. Efficiency error is plotted in Fig. 5a and standard error in Fig. 5b. Accompanying each curve is a two-way mixture model fit based on eqs. (26) and (27). Parameter estimates for efficiency error are presented in Table V. As threshold increases so does confusion error, as well as all the model parameters, $h$, $p_1$, and $p_2$. As with recognition error, parameter estimates for standard error are omitted. However, there is reasonable agreement between the parameter estimates obtained from efficiency error and standard error with the exception of the lowest threshold value, where the data are too sparse for reliable estimation. Above the lowest threshold the ratio of $p_1$ to $p_2$ remains fairly constant at approximately nine.

Estimates of $\bar{p}_V$ derived from average confusion number data and

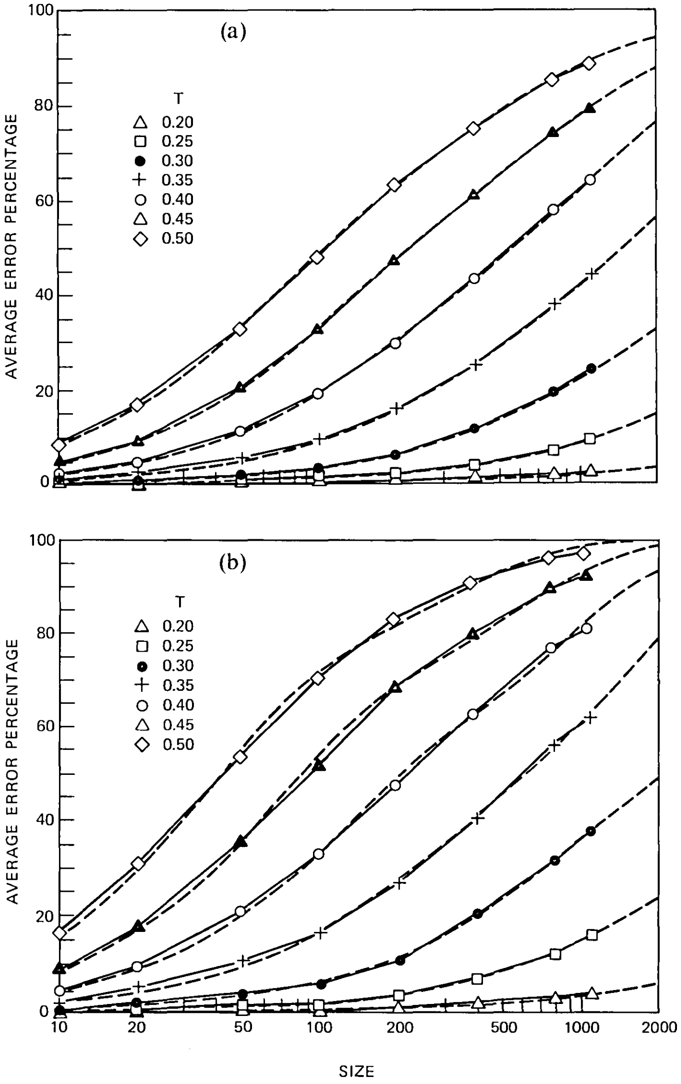


Fig. 5—(a) Average efficiency confusion error, and (b) average standard confusion error as a function of vocabulary size, with two-way mixture model fits for talker 6, vocabulary type $V_W$, and seven values of threshold, $T$.

from estimates of $h$, $p_1$, and $p_2$ for efficiency and standard confusion error are presented in Table VI together with coefficients of fit. Note that compared to recognition error results shown in Table IV, the coefficients of fit indicate better fits for the confusion models, sustaining a subjective impression gained by examining the figures. In addition there is much better agreement in estimates of $\bar{p}_V$ between

Table V—Model parameter estimates for average efficiency confusion error

| $T$ | $h$ | $p_1$ | $p_2$ |
|---|---|---|---|
| 0.20 | 0.050 | $4.62 \times 10^{-4}$ | $1.46 \times 10^{-5}$ |
| 0.25 | 0.250 | $5.70 \times 10^{-4}$ | $6.04 \times 10^{-5}$ |
| 0.30 | 0.350 | $1.73 \times 10^{-3}$ | $1.14 \times 10^{-4}$ |
| 0.35 | 0.310 | $5.27 \times 10^{-3}$ | $5.67 \times 10^{-4}$ |
| 0.40 | 0.379 | $1.15 \times 10^{-2}$ | $1.26 \times 10^{-3}$ |
| 0.45 | 0.527 | $1.84 \times 10^{-2}$ | $2.02 \times 10^{-3}$ |
| 0.50 | 0.620 | $3.08 \times 10^{-2}$ | $3.47 \times 10^{-3}$ |

Table VI—Comparison of model fits for average confusion number, efficiency confusion number, and standard confusion number

|  | Average Confusion Number |  | Efficiency Confusion Error |  | Standard Confusion Error |  |
|---|---|---|---|---|---|---|
| $T$ | $\bar{p}_V$ | $r$ | $\bar{p}_V$ | $r$ | $\bar{p}_V$ | $r$ |
| 0.20 | $3.73 \times 10^{-5}$ | 0.99521 | $3.70 \times 10^{-5}$ | 0.99539 | $4.67 \times 10^{-5}$ | 0.99670 |
| 0.25 | $2.03 \times 10^{-4}$ | 0.99952 | $1.88 \times 10^{-4}$ | 0.99910 | $1.97 \times 10^{-4}$ | 0.99895 |
| 0.30 | $7.17 \times 10^{-4}$ | 0.99984 | $6.81 \times 10^{-4}$ | 0.99971 | $6.62 \times 10^{-4}$ | 0.99950 |
| 0.35 | $2.13 \times 10^{-3}$ | 0.99994 | $2.02 \times 10^{-3}$ | 0.99981 | $2.05 \times 10^{-3}$ | 0.99951 |
| 0.40 | $5.26 \times 10^{-3}$ | 0.99999 | $5.15 \times 10^{-3}$ | 0.99989 | $5.00 \times 10^{-3}$ | 0.99956 |
| 0.45 | $1.14 \times 10^{-2}$ | 0.99999 | $1.06 \times 10^{-2}$ | 0.99995 | $1.01 \times 10^{-2}$ | 0.99964 |
| 0.50 | $2.24 \times 10^{-2}$ | 0.99998 | $2.04 \times 10^{-2}$ | 0.99986 | $1.97 \times 10^{-2}$ | 0.99930 |

efficiency and standard error, and between these estimates and the estimates obtained from the regression line fits for average confusion-number results. This improved agreement is attributed to the fact that the ratio of $p_1$ to $p_2$ is much smaller for confusion than for recognition results. The size of this ratio is a good indicator of the disparity among the underlying populations that we are attempting to model with two-way mixtures. The smaller the disparity, the better is the model. In the limit when $p_1$ equals $p_2$, indicating a uniform population, a simple model containing no mixtures is appropriate.

Figures 6 and 7 and Table VII present some additional aspects of confusion-error models. Figure 6 shows confusion error results for the three vocabulary types as a function of vocabulary size, together with model fits, with the threshold, $T$, set to 0.3. Results are shown individually for each of three talkers. Efficiency error results are shown in Fig. 6a though 6c and standard error results in Fig. 6d through 6f. The usual degradation in performance is found passing from $V_P$ to $V_W$ to $V_M$. Model parameter estimates for efficiency error for all six talkers and three vocabulary types are presented in Table VII. It can be noted that the increase in error rate across these vocabulary types is not consistently accompanied by an increase in the value of the parameter estimates, as was obtained for recognition error.

Figure 7 shows confusion error results with model fits for the six

Table VII—Model parameter estimates for average efficiency confusion error

| | $V_P$ | | | $V_W$ | | | $V_M$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Talker | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ |
| 1 | 0.350 | $3.59 \times 10^{-3}$ | $1.16 \times 10^{-4}$ | 0.209 | $8.34 \times 10^{-3}$ | $7.29 \times 10^{-4}$ | 0.723 | $6.65 \times 10^{-3}$ | $3.63 \times 10^{-5}$ |
| 2 | 0.204 | $2.84 \times 10^{-3}$ | $4.02 \times 10^{-5}$ | 0.200 | $4.00 \times 10^{-3}$ | $3.64 \times 10^{-4}$ | 0.661 | $4.34 \times 10^{-3}$ | $3.11 \times 10^{-5}$ |
| 3 | 0.450 | $1.97 \times 10^{-3}$ | $1.10 \times 10^{-5}$ | 0.200 | $6.45 \times 10^{-3}$ | $5.90 \times 10^{-4}$ | 0.700 | $5.74 \times 10^{-3}$ | $9.55 \times 10^{-5}$ |
| 4 | 0.060 | $1.94 \times 10^{-3}$ | $1.12 \times 10^{-4}$ | 0.300 | $9.87 \times 10^{-4}$ | $4.89 \times 10^{-5}$ | 0.353 | $2.60 \times 10^{-3}$ | $6.52 \times 10^{-5}$ |
| 5 | 0.161 | $2.50 \times 10^{-3}$ | $1.67 \times 10^{-4}$ | 0.400 | $1.96 \times 10^{-3}$ | $3.07 \times 10^{-5}$ | 0.551 | $3.66 \times 10^{-3}$ | $8.43 \times 10^{-5}$ |
| 6 | 0.070 | $4.18 \times 10^{-3}$ | $2.41 \times 10^{-4}$ | 0.350 | $1.73 \times 10^{-3}$ | $1.14 \times 10^{-4}$ | 0.450 | $3.68 \times 10^{-3}$ | $3.00 \times 10^{-4}$ |
| Mean | 0.216 | $2.84 \times 10^{-3}$ | $1.15 \times 10^{-4}$ | 0.277 | $3.91 \times 10^{-3}$ | $3.13 \times 10^{-4}$ | 0.573 | $4.45 \times 10^{-3}$ | $1.02 \times 1^{-4}$ |
| Standard deviation | 0.156 | $8.99 \times 10^{-4}$ | $8.73 \times 10^{-5}$ | 0.087 | $2.94 \times 10^{-3}$ | $2.97 \times 10^{-4}$ | 0.149 | $1.49 \times 10^{-3}$ | $1.00 \times 10^{-4}$ |

| | Mean | | | Standard Deviation | | |
|---|---|---|---|---|---|---|
| Talker | $h$ | $p_1$ | $p_2$ | $h$ | $p_1$ | $p_2$ |
| 1 | 0.427 | $6.19 \times 10^{-3}$ | $2.94 \times 10^{-4}$ | 0.266 | $2.41 \times 10^{-3}$ | $3.79 \times 10^{-4}$ |
| 2 | 0.355 | $3.73 \times 10^{-3}$ | $1.45 \times 10^{-4}$ | 0.265 | $7.87 \times 10^{-4}$ | $1.90 \times 10^{-4}$ |
| 3 | 0.450 | $4.72 \times 10^{-3}$ | $2.32 \times 10^{-4}$ | 0.250 | $2.41 \times 10^{-3}$ | $2.13 \times 10^{-4}$ |
| 4 | 0.238 | $1.84 \times 10^{-3}$ | $7.54 \times 10^{-5}$ | 0.156 | $8.11 \times 10^{-4}$ | $3.28 \times 10^{-5}$ |
| 5 | 0.371 | $2.71 \times 10^{-3}$ | $9.40 \times 10^{-5}$ | 0.197 | $8.69 \times 10^{-4}$ | $6.87 \times 10^{-5}$ |
| 6 | 0.290 | $3.20 \times 10^{-3}$ | $2.18 \times 10^{-4}$ | 0.197 | $1.29 \times 10^{-3}$ | $9.50 \times 10^{-5}$ |

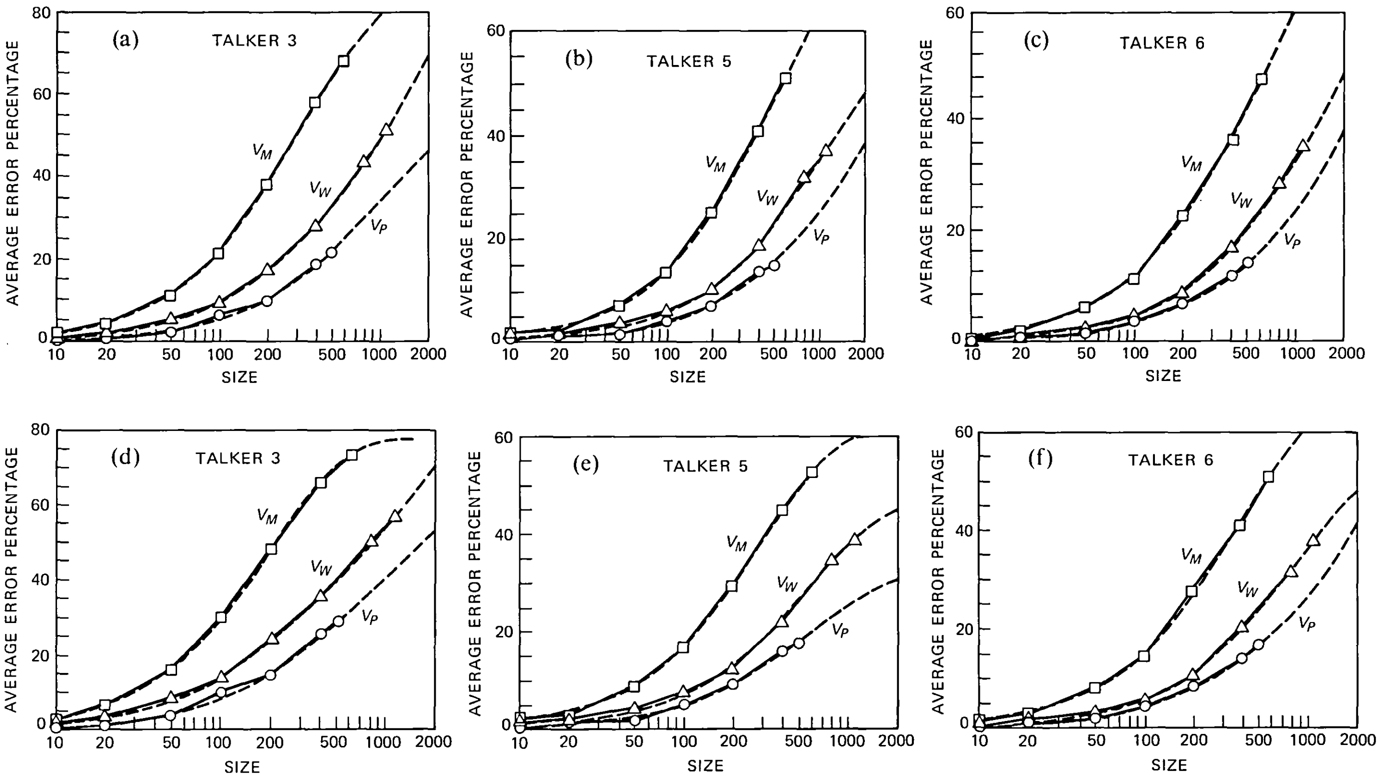


Fig. 6—(a), (b), (c) Average efficiency confusion error, and (d), (e), (f) average standard confusion error as a function of vocabulary size, with two-way mixture model fits, for three vocabulary types.

talkers and the single vocabulary type $V_W$ with the threshold, $T$, set at 0.3. Efficiency error results are shown in Fig. 7a and standard error results in Fig. 7b. As with recognition error there is considerable variability across talkers, although for confusion error there are no prominent extreme individual performances. There is also an apparent greater tendency for the error rates to converge for small vocabulary

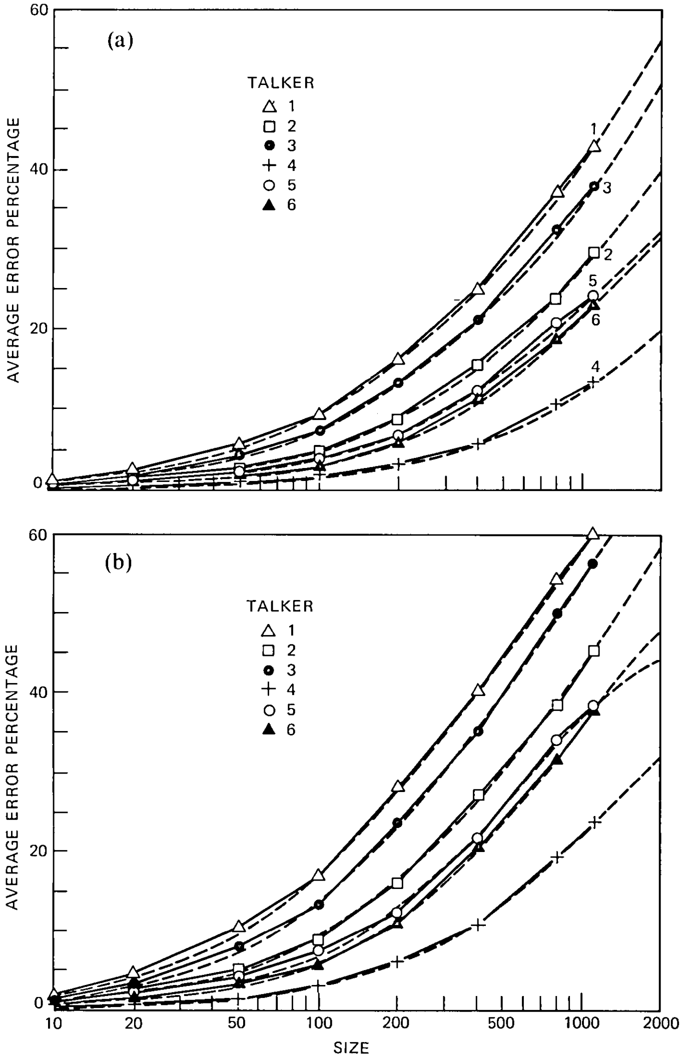


Fig. 7—(a) Average efficiency confusion error, and (b) average standard confusion error as a function of vocabulary size, with two-way mixture model fits, for six talkers, vocabulary type $V_W$, and threshold value, $T = 0.30$.

sizes. From Table VII, it is apparent that there is a fairly consistent tendency for the parameters $p_1$ and $p_2$, but not $h$, to increase with increasing error rate. Means and standard deviations are calculated across talkers and across vocabulary types in the tables. Using these as indicators of variability, as was done for recognition-error parameters, it appears that estimates of $p_1$ across vocabulary types, and $h$ across talkers, have relatively small variability, the same as for recognition error. There is also some suggestion that $p_1$ across talkers also has low variability.

The foregoing observations, together with similar ones made for recognition error, will be discussed in the following section in connection with interpretation of the parameters.

## IV. DISCUSSION

In the preceding section we have shown that both the confusion- and recognition-error performance of a recognition system can be modeled quite closely by assuming that there is a mixture of types of recognition or confusion trials, each type associated with a distinct probability for the occurrence of a recognition or confusion error. We have shown that, in most cases, assuming a mixture of two population types is quite adequate to represent the experimental behavior that has been presented, although more than two types might very well underlie this behavior.

A substantial dichotomy of population types is evidenced by a large ratio of $p_1$ to $p_2$, the probability estimates of the two populations. It has been found that large ratios, of the order of 100 or 200, are generally associated with recognition error, while smaller ratios, from 10 to 30, generally characterize confusion error.

Where substantial dichotomies exist, it would be most interesting and useful to relate the different population types to actual phenomena associated with the speech-recognition process. Unfortunately, the experimental results incorporate and merge various sources of recognition and confusion error, including the talker, the talking environment, various aspects of the vocabulary, and the recognizer itself. It is difficult, if not impossible, to uniquely characterize these sources in the model parameters. Moreover, in these experimental results, trials are averaged over all four repetitions of each word and all the words in a subset [see eqs. (28) to (33)]. It is therefore not possible to uniquely attribute the different types of behavior to phenomena associated with repeated utterances of the same word on one hand, or to different word types in a subset, on the other. It is reasonable to believe, however, that a dichotomy of types for repeated utterances of the same word would be more significant for rank or recognition performance than for confusion performance. This is because an atypical pronunciation of a word should perturb the self-distance distribution, and consequently the rank distribution, far greater than the distribution of distances to other words in the vocabulary. In fact, in some results not presented here, confusion error was calculated from the distances between reference prototypes alone. Only small differences were obtained with the confusion error results calculated from distances between test utterances and prototypes, which have been presented in the previous section. We are therefore led to believe that differences in populations of trials for confusion error are mostly associated with different types of words, rather than different types of pronunciations of words. So we speculate that for confusion error there are two (or more) populations of words with confusion probabilities differing by ratios of 10 to 30; whereas, for recognition error there are two (or more) populations of trials with recognition probabilities differing by ratios of 100 to 200, where large discrepancies in self-distance distributions for repeated utterances of words are superimposed on population differences among words.

There are two other possibilities for making educated speculations associating model parameters with various phenomena in the recognition process. First, we can examine trends as an experimental parameter such as vocabulary type or distance threshold varies. For example, in the previous section note was taken of which model parameters vary little over the range of some experimental variables. Second, some clues might be obtained if small or large subset approximations of the model error formulations isolate one model parameter from the others. In anticipation of this possibility, some derivations of small and large subset approximations are provided.

### *4.1 Small and large subset size approximations*

Small subset size approximations for the error formulations are obtained by assuming that for small $p$ and $N$, $(1 - p)^N$ can be represented by the first few terms in the binomial expansion. Rewriting eq. (26) for $M = 2$ we obtain

$$\mathscr{E}\{e_V\} = 1 - h\,\frac{1 - (1 - p_1)^N}{p_1 N} - (1 - h)\,\frac{1 - (1 - p_2)^N}{p_2 N}\,, \qquad (35)$$

where, as in Section 3.3, we assume $p_1 > p_2$. Using the approximation

$$(1 - p)^N \approx 1 - Np + \frac{N(N - 1)}{2}\,p^2 \qquad (36)$$

for small $N$, we obtain

$$\mathscr{E}\{e_V\} \approx \frac{N - 1}{2}\,(hp_1 + (1 - h)p_2) = \frac{N - 1}{2}\,p_V. \qquad (37)$$

Similarly, eq. (27) rewritten for $M = 2$ is given by

$$\mathscr{E}\{E_V\} = 1 - h(1 - p_1)^{N-1} - (1 - h)(1 - p_2)^{N-1}. \tag{38}$$

Approximating $(1 - p)^{N-1}$ by $1 - (N - 1)p$, we obtain

$$\mathscr{E}\{E_V\} \approx (N - 1)[hp_1 + (1 - h)p_2] = (N - 1)p_V. \tag{39}$$

Thus, for small $N$, expected error grows linearly with $N$ just as expected rank or confusion number does for all $N$ [see eq. (23)]. In fact, the approximation for standard error is identical to eq. (23). Also, for these small $N$ formulations, the expected value of efficiency error is just one half the expected value of standard error. This can be observed in the experimental results as pointed out in Section 3.1. It is easily verified from the basic expressions for efficiency and standard error found in eqs. (12) and (14) for $N = 2$.

For large subset size approximations we might assume that $(1 - p)^N \approx 0$ for large $N$. Then we can approximate the efficiency error formulation, eq. (35), by

$$\mathscr{E}\{e_V\} \approx 1 - \frac{1}{N}\left(\frac{h}{p_1} + \frac{1 - h}{p_2}\right). \tag{40}$$

A comparable approximation does not exist for standard error. However, it is possible to approximate $(1 - p)^N$ by $e^{-pN}$ for small $p$ and moderate $pN$. This can be introduced in eq. (39) to obtain

$$\mathscr{E}\{E_V\} \approx 1 - he^{-(N-1)p_1} - (1 - h)e^{-(N-1)p_2}, \tag{41}$$

which is a potentially useful approximation.

For these small and large vocabulary size approximations to be useful in providing interpretations for the model parameters, conditions must exist for one or the other of the population types to dominate. For small $N$, since we have assumed $p_1 > p_2$, for type 1 populations to dominate in eqs. (37) and (39) we should have

$$\frac{h}{1 - h}\frac{p_1}{p_2} \gg 1. \tag{42}$$

Conversely, for type 2 populations to dominate in eq. (40), we should have

$$\frac{1 - h}{h}\frac{p_1}{p_2} \gg 1. \tag{43}$$

### *4.2 Small and large subset approximations and the relation between rank and confusion error*

These small and large $N$ formulations for error coincide with our earlier discussion in Section 3.1 on the relation between confusion and rank or recognition error. There we found that for small $N$, recognition error is approximately equal to confusion error at a threshold for which average confusion number equals average rank number. For large $N$ we found that recognition error approximates confusion error at a threshold equal to average self-distance. Examining these relationships once more with respect to model parameters, we see that for small $N$, from eq. (37) or (39), the threshold for equality should be set such that $\bar{p}_{q_V(T)} = \bar{p}_{r_V}$. (The subscripts are used to differentiate between confusion and rank.) From eq. (23) we know that $\bar{p}_V$ is the slope coefficient for expected rank or confusion number, so our earlier hypothesis for small $N$ is confirmed. For the example of talker 3 and vocabulary $V_W$ used in Section III, $\bar{p}_{r_{V_W}}$ is $6.67 \times 10^{-3}$ from Table I. From the same table, we see that for $\bar{p}_{q_{V_W}(T)}$ to have this value, $T$ should be between 0.35 and 0.40.

For large $N$, from eq. (40), assuming eq. (43) holds, equal errors should be obtained if $1 - h/p_2$ is the same for both confusion and rank. Again for the same example, for rank or recognition, $1 - h/p_2 \approx 6 \times 10^3$. Although confusion model parameter estimates are not shown for talker 3 as a function of threshold, for a threshold of 0.235, corresponding to average self-distance, they are approximately 0.04, $5.3 \times 10^{-3}$, and $1.8 \times 10^{-4}$, for $h$, $p_1$, and $p_2$, respectively. Thus, $1 - h/p_2 \approx 5.3 \times 10^3$, which agrees well with the value obtained for rank. Thus, the model formulations and parameter estimates support the original hypothesis for large $N$ as well. For large $N$ the density of words for a given vocabulary type is great enough so that even though the distribution of rank numbers and confusion numbers for a threshold set to average self-distance is not the same, the proportion of zero and nonzero rank and confusion numbers, which correlates well with both kinds of error, is about the same. In the discussion that follows, we will conjecture that the parameters of the large $N$ formulation, $h$ and $p_2$, are largely associated with vocabulary type which should be the major factor controlling density.

### *4.3 Small and large subset size approximations and dominance of population types*

Now let us examine some of the experimental results to determine to what extent eq. (42) or (43) holds. For recognition error, in Table III, we find that generally high ratios of $p_1$ to $p_2$ are to a large extent offset by small values of $h/1 - h$. Thus, although the value of the expression in eq. (42) is nearly always greater than one, it is not consistently greater than 10, a value for which we could say unequivocally that type 1 populations dominate small vocabulary size behavior. Those instances in which the expression assumes values less than 10 are associated with low error rates, for example, for talkers 1 and 2. Quite the opposite is true for large size behavior, since in eq. (43), $1 - h/h$ is always greater than one and the ratio of $p_1$ to $p_2$ remains large. Thus large size behavior is consistently dominated by type 2 populations in eq. (40), and also in eq. (41), as is easily verified.

For confusion error, with the threshold fixed at 0.3, the results in Table VII indicate that although the ratio of $p_1$ to $p_2$ is smaller than for recognition error, the value of $h/1 - h$ is generally greater. Consequently, overall, the expression in eq. (42) assumes about the same range of values as for recognition error. Similar observations are made for large size behavior in both confusion error and recognition error.

To the extent that type 1 populations control small vocabulary behavior and type 2 populations control large vocabulary behavior, it is natural to associate type 1 populations with trials or words that are chronically "bad" in some sense, and type 2 populations with the "natural" density of a particular vocabulary type. Thus, type 1 errors persist when alternate choices are few and the vocabulary size is small, while the natural density of words in the vocabulary must be important when the vocabulary size is large. By this hypothesis we should expect that good performance associated with low error rates should have only a weak dominance of type 1 trials or words, since there should be fewer bad words or trials. This is, in fact, what is observed. The hypothesis is also in agreement with the observations made earlier in this section on the dichotomy of populations for recognition and confusion error.

### *4.4 Experimental variability of model parameter estimates and parameter origins*

We can now stretch further our speculations on the origins of model parameters by recalling our earlier observations of their relative variability across the experimental variables we observed. We assume three sources of error, the talker, the vocabulary, and everything else which we lump into the recognition system. For confusion error, $p_1$ was observed to have low variability across vocabulary types, and to a lesser extent, across talkers. Therefore, we could associate $p_1$, the type 1 probability, largely with the system. For recognition or rank performance, $p_1$ has low variability only across vocabulary types, and is therefore associated with both talker and the system. This reflects the effect of self-distance distribution, which is clearly talker dependent. The type 2 probability, $p_2$, was observed to have low variability across talkers for recognition or rank performance (except one talker). Although a similar observation was not made for confusion performance, it is natural to associate $p_2$ with vocabulary type and the system. This hypothesis is compatible with the vocabulary density role associated with $p_2$ earlier in our discussion. Finally, $h$, the mixing coefficient for the two types of populations, was observed to have low variability across talkers for both confusion and recognition performance. We are therefore led to believe that $h$ is largely a function of vocabulary type, with the role of the system unclear.

## V. CONCLUSION

The data extracted from a series of isolated word recognition experiments with large vocabularies have enabled us to hypothesize and verify a simple probabilistic model underlying performance of recognizers. Essentially, we have attempted to model the distributions of confusion number, an a priori characterization of a recognizer, and rank number, an a posteriori characterization. Expressions have been derived for three confusion or rank number functions, average confusion or rank number, and two error functions, standard error and efficiency error. Models have been evaluated and interpreted using experimental values of these functions. The difference between standard error and efficiency error has been described and an attempt has been made to describe and interpret the difference between confusion and rank performance.

It is significant that good models for performance are obtained only by assuming a mixture of probability distributions as the basis. The reduction of the performance of a recognition system over a large range of vocabulary sizes to as little as three parameters enhances our understanding of the processes involved and has some potential practical utility in the evaluation of systems. Over the range of experimental variables available in this series of experiments we have been able to speculate on associations of the model parameters with variables in the recognition process. To place these suggestions on firmer footing will require additional experimental data. For example, useful data can be obtained from a large number of repeated utterances for a given talker and vocabulary in order to attribute behavior differences uniquely to the different words in a vocabulary or to repetitions of the same words. Examining results obtained by passing the same utterances through different recognizers, or systematic variations of the same recognizer or recording environment, will also be revealing. The use of a larger number and more sharply distinct vocabulary types will also provide useful information. In addition it is important to devise experiments to evaluate the predictive power of the models. Thus, once the parameters of a model have been estimated, new experimental data obtained with controlled variation of experimental variables should be consistent with the model.

## VI. ACKNOWLEDGMENTS

The author wishes to thank J. G. Wilpon for maintaining the experimental database, L. R. Rabiner, S. E. Levinson, J. L. Hall, and M. M. Sondhi for stimulating discussions, and L. A. Shepp for valuable suggestions on the use of probability mixtures.

## REFERENCES

1. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61,* No. 10 (December 1982), pp. 2989–3005.
2. A. R. Smith and L. D. Erman, "Noah—A Bottom-up Word Hypothesizer for Large-Vocabulary Speech Understanding System," IEEE Trans. on Pattern Analysis and Machine Intelligence, *PAMI-3* (January 1981), pp. 41–51.
3. W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. II,* New York: Wiley, pp. 52–7.
4. W. L. Johnson and S. Kotz, *Distributions in Statistics: Discrete Distributions,* Boston: Houghton-Mifflin, 1969, pp. 76–9.
5. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoust., Speech, and Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
6. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. Commun., *COM-29* (May 1981), pp. 621–59.
7. C. K. Ogden, *Basic English: International Second Language,* New York: Harcourt, Brace and World, Inc., 1968.
8. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68* (November 1980), pp. 1271–6.
9. J. E. Dennis, Jr. and H. H. W. Mei, "An Unconstrained Optimization Algorithm Which Uses Function and Gradient Values," unpublished report based on M. J. D. Powell's, "A New Algorithm for Unconstrained Optimization," contained in J. B. Rosen et al., eds., *First Symposium on Nonlinear Programming,* New York: Academic Press, 1970.

## AUTHOR

**Aaron E. Rosenberg,** S.B. and S.M. (Electrical Engineering), 1960, Massachusetts Institute of Technology; Ph.D. (Electrical Engineering), 1964, University of Pennsylvania; AT&T Bell Laboratories, 1964—. Mr. Rosenberg is a Member of Technical Staff in the Acoustics Research Department. Since joining AT&T Bell Laboratories his research interests have included auditory psychophysics, speech perception, and currently, speech and speaker recognition. He has authored or coauthored over 35 papers in these fields. Former chairman, IEEE Acoustics, Speech, and Signal Processing Society Technical Committee on Speech Communication; secretary, ASSP Conference Board; associate editor for speech communication, ASSP Transactions. Member, IEEE, Sigma Xi. Fellow, Acoustical Society of America.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# A Simulation-Based Comparison of Voice Transmission on CSMA/CD Networks and on Token Buses

By J. D. DeTREVILLE*

(Manuscript received March 14, 1983)

Digitized speech can be transmitted over a variety of digital media. An interesting choice is the use of a Local-Area Network (LAN), for which digitized speech is packetized at the transmitter and depacketized at the receiver. Many local-area networks exhibit good throughput but poor delay characteristics; variable or excessive transmission delay can become noticeable and objectionable to the users of such a voice system. A number of simulations were performed to assess the delay characteristics of a Carrier Sense Multiple Access/Collision Detection (CSMA/CD) LAN and of a similar token bus LAN. A comparison of the results shows that the token bus performs somewhat better. The CSMA/CD LAN's performance was characterized by carrying voice well until a point of collapse is reached; the token bus's performance degraded more continuously. In either case, throughput close to the theoretical capacity of the LAN was found achievable with appropriate techniques.

## I. CHARACTERISTICS OF DIGITAL VOICE

Human speech of telephone quality can be easily encoded into a 64-kb/s bit-stream containing 8000 8-bit speech samples per second; although much more efficient encodings are possible, this synchronous 64-kb/s speech encoding is assumed throughout this paper.† Speech

* AT&T Bell Laboratories.

† As a simple improvement, the use of delta-modulation to transmit only the differences between successive samples could produce savings of approximately 2:1. More extensive processing could result in more extensive savings by taking further advantage of the regular properties of human speech. Improvements in the encoding bit rate will

Table I—Speech sample delay breakdown

| Type of Delay | | Consisting of |
|---|---|---|
| Fixed | The (nominal) temporal length of the packet: the packet size measured by its acquisition time | The delay after the packet is acquired until the packet is completed and transmitted (the temporal length of the portion of the packet following this sample) |
| | | Plus the delay after the packet is received until the sample is played back (the temporal length of the portion of the packet preceding this sample) |
| | Plus much smaller fixed delays (e.g., the transmission time) | |
| Variable | The delay in transmitting the packet | The delay in obtaining the transmission medium |
| | Plus (typically) smaller variable delays | |

consists of *talkspurts* separated by *silences:* a speaker in a typical conversation talks about 40 percent of the time and is silent for the remainder, and an approach that transmits speech only during talkspurts can therefore be desirable. Silences, of course, are relative. Ideally, no speech should be lost by being considered silence, and no extraneous background sounds should intrude during silences. This ideal can be approached through the use of cutoff levels with memory.

Transmitting digital speech over a shared packet network entails packetizing the digital signal at the transmitter, transporting it over the network, and depacketizing it at the receiver; these operations can introduce delay. Table I gives a high-level breakdown of the delay in the transmission. The delay includes a fixed component and a variable component. Because of the variable component, if the receiving station begins playing a packet as soon as it is received, this can introduce artificial silences at some points (when a packet is delayed more than the one before it) and lost speech at others (when a packet is delayed less than the one before it), ultimately producing effects audible to the users. This variability of performance can be partially overcome by artificially delaying packets at the receiver, such that only those packets whose variable delay is greater than some threshold will cause anomalies; since speech is inherently real-time, arbitrary queueing of packets at the transmitter or receiver is not possible.

The user-level model of speech used in this paper is that of a typical two-way conversation, in which real-time constraints exist at both

result in improvements in the performance figures presented in this paper, but these performance improvements will typically not be linear, since a reduction in the bit rate will make other factors relatively more important. Similarly, although variable bit-rate encodings can produce further savings over fixed bit-rate encodings, they can lose many of the advantages shown for fixed bit rates in this paper, and will again have less of a total impact than might otherwise be expected.

ends. If either side of the voice conversation were to be a computer or similar device, knowledge of this fact could be used to ease the constraints somewhat, although this optimization is not considered in this paper. If both ends were known to be computers, speech could then be transmitted as a nonreal-time data transfer.

If packets are artificially delayed, the one-way voice sample delay from the transmitter to the receiver during successful transmission will roughly equal the packet size plus the threshold delay. Increasing the packet size will increase the effective bandwidth of the system (by reducing per-packet overhead); increasing the artificial delay will reduce the incidence of anomalies (by reducing the probability that a packet will have been delayed for longer than the threshold). Reducing the traffic speeds access to the shared network; reducing anomalies postpones the onset of overload. On the other hand, increasing these values increases the delay through the system, which will eventually become perceptible to the user; this suggests a compromise between the extremes. For example, the one-way delay on a single-hop synchronous-orbit satellite voice circuit is 270 ms, which many users view as disruptive; the double-hop delay of 540 ms is considered much worse. Considering that a system built on one LAN may frequently communicate with another system on another LAN (thereby at least doubling the end-to-end delay), this suggests that the one-way delay on a given LAN should be kept well below 270/2 = 135 ms. An alternative might be to treat inter-LAN connections differently from intra-LAN connections; this possibility is not considered here. In any case, the delay cannot be allowed to grow without bound. It should be noted that echo is perceived as being much more disruptive than simple delay, with the audible threshold occurring much earlier, but echos, where they might occur, can be controlled through the use of echo cancelers. The exact nature of this compromise depends upon the precise psychoacoustic characteristics of the importance of this delay compared to, say, the effect of the anomalies caused by variable delay; this trade-off is not well understood.

## II. A TYPICAL CSMA/CD LAN

*Ethernet** is a typical Carrier-Sense Multiple Access/Collision Detection (CSMA/CD) LAN.[1] Data packets are transmitted bidirectionally over a coaxial cable with an acyclic branching topology. Access to the net is distributed ("multiple access") and statistical. A station wishing to transmit first listens to determine whether the net is in use ("carrier sense"); if it is, the station defers until the current user has finished transmitting its packet. If the net is not in use, the station

* *Ethernet* is a trademark of Xerox Corporation.

begins to transmit. Due to race conditions, two stations could begin to transmit simultaneously; when one station notices another transmitting ("collision detection"), it aborts its transmission, jams the net to ensure that other stations also notice the collision and abort their transmissions, and retries after a random amount of time, thereby statistically avoiding recollision.

An *Ethernet* CSMA/CD network is bit serial and runs at 10 Mb/s ; a bit-time is thus 0.1 $\mu$s. Assuming 64-kb/s speech, complete utilization of the bandwidth would result in carrying up to 195.3 two-person conversations (in which each person spoke 40 percent of the time). Such efficiency, however, can never be achieved in practice.

One reason is simple per-packet overhead. A transmission on an *Ethernet* CSMA/CD network begins with 64 sync bits, followed by the packet. A packet contains 112 bits of header, a 368- to 12,000-bit data field (thus between 5.75 ms and 187.5 ms of 64 kb/s speech) and a 32-bit CRC field. A station may begin to transmit when it has seen the net idle for 96 bit-times. Assuming (arbitrarily) that voice stations are uniformly distributed along a maximum-length linear CSMA/CD network, computations based on the *Ethernet* propagation delay budget give a worst-case mean one-way propagation time of about 10.06 $\mu$s; we can expect an arbitrary station to see the net go idle 100.6 bit-times after the arbitrary preceding station actually ceased to transmit. A linear CSMA/CD network is in ways a "best case," since the mean distance between stations will be less than in a more general topology. However, the limiting case in complex topologies is extremely unlikely. Similarly, uniform distribution is a "best case," but a more accurate characterization seems difficult to achieve.

Taking per-packet overhead into account, we see that, at the smallest packet size, the speech samples can occupy only 47.6 percent of the bandwidth, allowing a maximum of 93.0 conversations; at the largest packet size, 96.7 percent of the bandwidth can be speech samples, allowing 188.9 conversations.

Studies of the *Ethernet* specifications under varying load conditions have typically shown *Ethernet* CSMA/CD networks to have very desirable throughput characteristics (e.g., see Ref. 2). Throughput tends to rise linearly with offered load until saturation is approached, and then levels off, with an asymptotic throughput within a few percent of maximum for large packets and within several percent for small packets. (For an experimental *Ethernet* CSMA/CD network described in Ref. 2, whose numerical parameters differed significantly from those of the specifications discussed here, measured throughput reached 96 percent for maximum-size packets, and 83 percent for minimum-size packets. The traffic in this study, as in the case considered here, was produced by a number of stations each offering a fraction of the total load.) Throughput may decrease under certain cases of extreme overload (for example, two stations each attempting to offer 100-percent load to the net would ultimately transmit less data together than either would individually, due to their contention), although this decrease evidently does not become pathological.

On the other hand, individual packets may experience significantly greater delays under heavy loads than under light loads. The nature of this increase in the delay has not been well characterized in past studies of data traffic, which is not as badly affected by variable delays as is real-time voice. Although we can be reasonably certain that the voice samples will make the journey from the transmitter to the receiver, almost up to the physical transport limits of the network, this might be inadequate if they require excessive time to do so.

## III. A SIMULATION STUDY

To determine the performance characteristics of voice traffic on a CSMA/CD network, a computer simulation was prepared based on *Ethernet* specifications. The stations were assumed to be uniformly distributed over a maximum-length network. Both voice and data traffic were modeled.

The voice stations modeled typical two-person conversations. The stations were therefore paired, with an appropriate distribution and correlation of talkspurts and silences (adapted from Brady, as discussed in Ref. 3). Brady's study included filtering out very short silences and very short talkspurts, thereby increasing the mean length of silences and talkspurts and otherwise modifying their distribution. The exact type of voice filtering best suited for transmission over an LAN is still uncertain. Voice packets were not transmitted during silences. The simulations began with one conversation, after which an additional conversation was added every 0.5 second of simulated time, until the system had passed saturation. This staged introduction of conversations helps to eliminate anomalies associated with the start-up of several conversations at once. Although it is possible that the monotonically increasing number of conversations could produce history artifacts in the simulation results, none were observed in the CSMA/CD simulations; these did occur in the token bus studies outlined in Section IX.

Figure 1 shows the actual number of speakers over time, as a function of the number of conversations, for one particular voice traffic pattern; this pattern and another like it were used throughout the simulations to control the effect of differing traffic patterns in separate simulations. (As it turned out, the effect of a particular voice traffic pattern on observed behavior was less than anticipated, and is

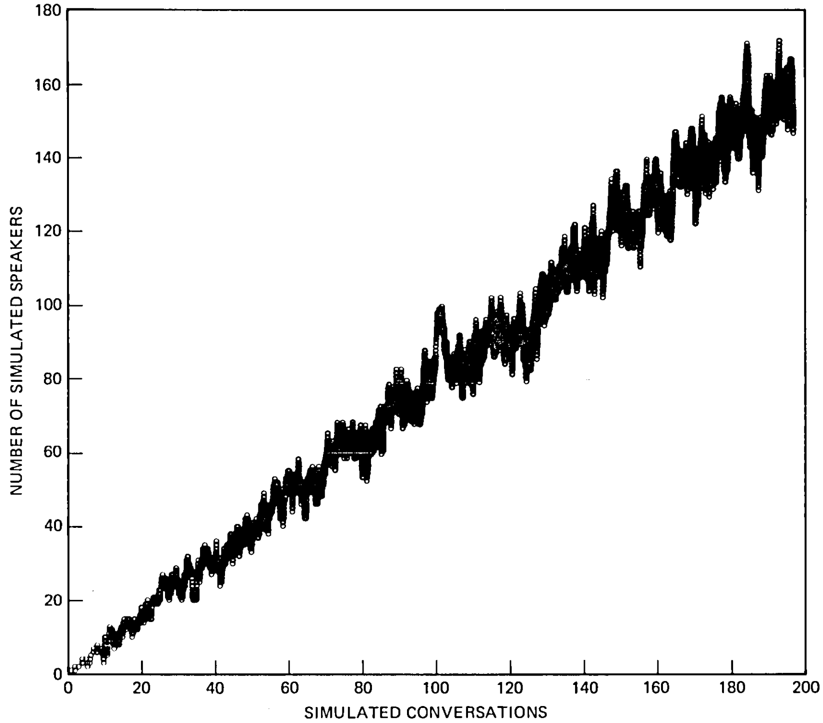


Fig. 1—Number of simulated speakers in a voice traffic pattern. This graph presents a voice traffic pattern used in most of the simulations presented in this paper, plotting the number of instantaneous speakers as a function of the number of conversations. One simulated two-person conversation is added every 0.5 second of simulated time; there are 0.8 expected instantaneous speakers per conversation. Speakers divide their time between talkspurts and silences. This simulation uses an empirically derived distribution of the lengths of talkspurts and silences and of the correlation between the states of the two potential speakers in a conversation.

easily compensated for. Thus, the use of the same voice traffic patterns throughout the simulations seems to have been unnecessary.)

The data stations presented a bimodal distribution of packet lengths, typical of data traffic on real nets, with 80-percent minimum-size packets and 20-percent maximum-size packets (giving approximately the opposite distribution when weighted by length). The packet arrivals were modeled by a Poisson process: the traffic generated by a Poisson process is not as bursty as real data traffic, but the difference was expected to be relatively unimportant in determining the effect of the data traffic upon the voice traffic. The simulations included between 0- and 10-percent steady data loading of the system, the latter value being well beyond the measured steady loadings of current *Ethernet* CSMA/CD networks.

## IV. VOICE TRANSPORT ALGORITHMS

The simplest algorithm for transmitting voice would be to packetize the digital speech, dropping packets containing only silences, and to send them to an autonomous network interface to be transmitted asynchronously. The simplest algorithm for receiving voice would be to receive packets asynchronously from an autonomous network interface, and begin to play back the first packet of a talkspurt after some artificial delay, with subsequent packets of the talkspurt each immediately following its predecessor.

An important improvement on the transmission algorithm at the source deals with the case when packet transmission must be delayed until the net can be acquired. If, while the packet is waiting to be transmitted, more speech samples are being buffered, these can be appended to the old packet before it is transmitted instead of being used to start a new packet. This approach has three advantages:

1. It tends to transmit fewer packets under a heavy load, thereby applying a degree of negative feedback.

2. The varying length of a packet serves as a sort of time-stamp. Since the last speech sample in the packet was collected just before the packet was successfully transmitted, this allows the receiver to determine the exact age of the first speech sample, allowing more precise control over packet playback.

3. It produces an adaptive effect. In the simplest case, each station will begin to attempt to transmit a new packet one packet-time after beginning to attempt to transmit the previous packet; here, though, this will occur one packet-time after the last packet was successfully transmitted. In the first case, if two stations happen to collide with each other once, they will then collide with each other every packet-time afterwards until one of them begins a silence; in the second case, a collision once resolved creates a phase shift that persists thereafter.

At the receiver, we buffer packets to cope with their variable delay. If the speech samples are implicitly time-stamped by the variable packet size, it is possible to correct for the delay that the first packet of a talkspurt has already experienced in transmission.

The receiver implementation can be quite simple. The voice path is implemented as a first in first out (FIFO) buffer: packets are inserted as they are received while samples are extracted synchronously. The first packet of a talkspurt is preceded in the FIFO by the appropriate amount of artificial silence; the beginning of a talkspurt can be detected by the FIFO being empty. This scheme is easily extended to the case of connections with more than one other speaker, with multiple independent speech sources being merged together; each speaker is assigned a separate FIFO and summing is performed on the outputs of the FIFOs. The FIFOs can be implemented in hardware or software.

Samples that are too late are discarded; they will have been preceded by an artificial silence. Excessive delays can result in packets that are longer than the FIFO, in which case part of the packet can be discarded. For every amount of artificial silence we accidentally introduce, we lose an equivalent amount of speech, except when the last samples of a talkspurt are delayed excessively, in which case the artificial silence before playing them back is matched by losing part of the real silence elsewhere. With proper matching between transmitter and receiver, it is possible for the transmitter to predict which samples the receiver would discard, and simply not transmit these in the first place, thereby reducing net traffic under heavy load and avoiding a potential instability.

## V. BASIC CSMA/CD PERFORMANCE

Simulations were performed to measure the voice capacity and related characteristics of CSMA/CD LANs. For a simulation in which voice stations used (nominally) minimum-size voice packets (5.75 ms), and in which there was no data traffic, Fig. 2 shows the transmission delays that voice packets experienced. Note that the delay is essentially zero (i.e., less than the quantizing sample time of 125 $\mu$s) until the equivalent of approximately 60 conversations is reached, at which point the delay rises roughly linearly. [While the expected number of speakers at an arbitrary point in the simulation is 0.8 times the number of conversations, the actual number of speakers will vary from this, depending on the details of the traffic pattern. We define the *effective* number of conversations at a point in time as the actual number of speakers divided by 0.8 (the expected value of the effective number of conversations is the actual number of conversations). It was found that much smoother graphs were obtained by plotting transmission performance using the effective number of conversations rather than the actual number, and that the curves were thereby made much more similar across different traffic patterns. Most of the graphs in this paper are based on effective conversations rather than actual conversations; they may be converted to actual conversations by the addition of appropriate axial randomness.] Note that the standard deviation is several times the mean, due to the long tail of the distribution of delays; this is illustrated in Fig. 3, which shows the distribution of delays at a 50-conversations loading.

If we set a threshold of an additional 5.75-ms artificial delay of voice samples, we can bring the total delay through the system to 11.5 ms. (Given a desired total delay of 11.5 ms, it would be possible to allocate less of the total to variable delay and more to packet size; the reverse would also be possible. An extreme position in either direction can be

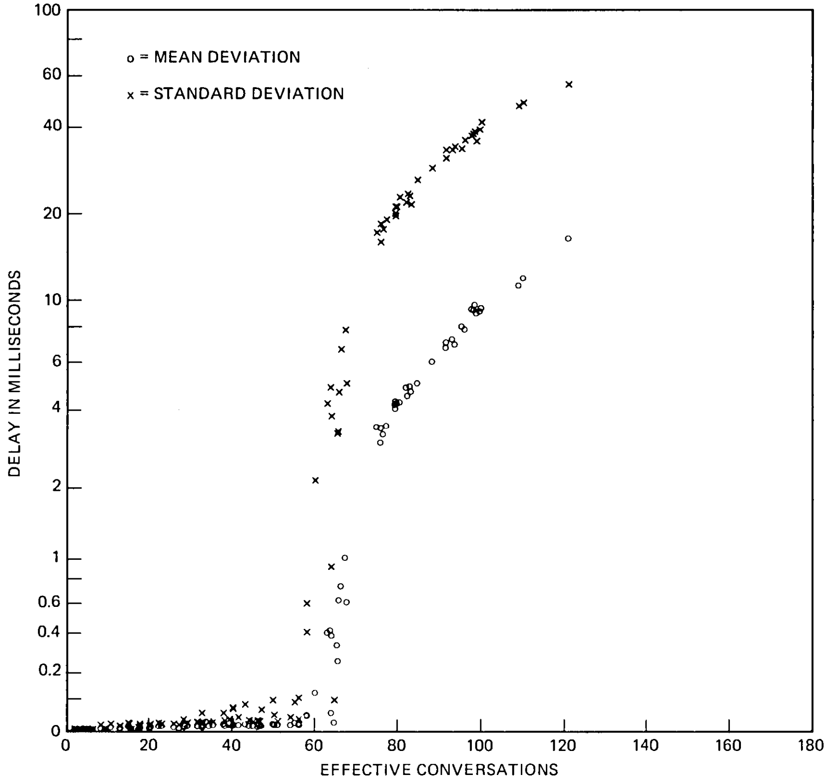


Fig. 2—CSMA/CD delay for 5.75-ms voice packets in the absence of data. This graph shows the mean and standard deviation of the delay experienced in the transmission of (nominally) 5.75-ms (i.e., minimum-size) voice packets in the absence of data traffic, as a function of the number of effective conversations. Note that both the mean and the standard deviation are essentially zero (i.e., less than the quantizing sample time, 125 $\mu$s) until about 60 effective conversations are reached, at which point they grow roughly linearly (distorted here by the logarithmic vertical scale) and become quite large; the standard deviation far exceeds the mean.

counterproductive, so an equal division is not totally unreasonable. However, as will be shown later in this paper, it seems more optimal, for CSMA/CD networks, to allocate significantly more delay to packet size than to variable delay.) At this delay we can expect to transmit up to about 60 conversations well, and to lose some speech samples past that point, as shown in Fig. 4. Here, the vertical axis measures the percentage of speech samples lost, which roughly models the degradation of the channel. Further study is needed to determine the effect of other parameters of the artificial silences and loss speech upon human users: for example, the number and length of these anomalies are probably important.

As a test of validity, the results of these simulations (as well as the

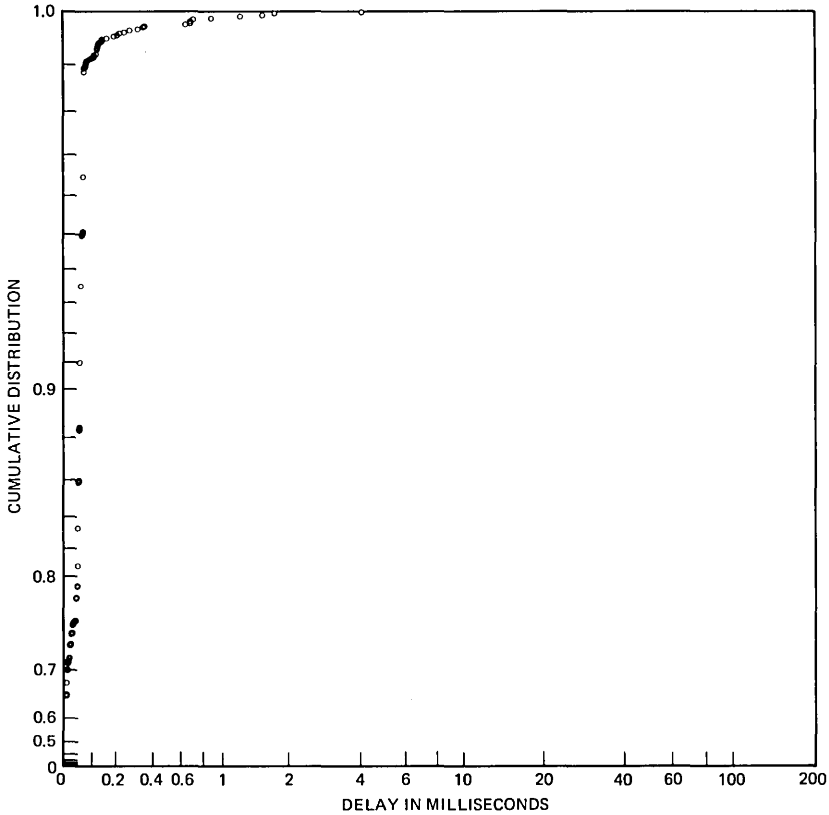


Fig. 3—CSMA/CD cumulative distribution of voice packet transmission delays for a loading of 50 conversations and 5.75-ms packets. This graph shows the cumulative distribution of the variable transmission delays experienced over a period of 0.5 second when the simulated CSMA/CD network was loaded with 50 conversations and no data traffic. The vertical axis is exponential; the horizontal axis is logarithmic. We see that about 65 percent of the packets experienced no delay, that over 99 percent were transmitted in less than 125 μs (the quantum phase shift possible using the adaptive algorithm), and one took over 4 ms. The shape of this curve causes the standard deviation to exceed the mean, as shown in Fig. 2.

ones following) were compared to a previous study of voice transmission on CSMA/CD networks[4]; the results were found to correspond closely.

As an example of the importance of the adaptive nature of the variable packet-size algorithm, Fig. 5 shows the effect of using fixed packet sizes; we see that the channel degrades much sooner.

## VI. EFFECT OF DATA TRAFFIC ON CSMA/CD CAPACITY

As we have seen, the natural synchronous nature of voice traffic in conjunction with an adaptive algorithm enables the transmitters, in

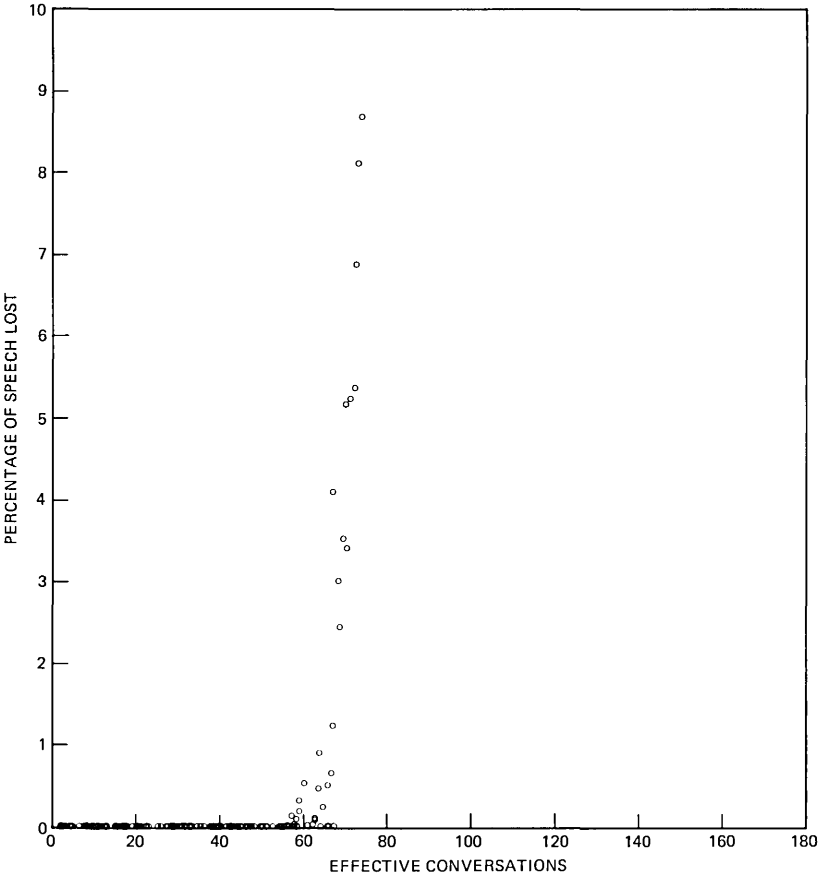


Fig. 4—CSMA/CD voice channel degradation with increasing load with 5.75-ms packets and 5.75-ms artificial delay. This graph shows the voice signal degradation experienced on a simulated CSMA/CD network with 5.75-ms packets and an additional 5.75-ms artificial delay at the receiver. Two simulations with different voice traffic patterns were performed and their results superimposed. Degradation is measured as the percentage of voice samples that are discarded (here at the transmitter). There is no degradation until about 58 effective conversations, soon after which the degradation rises roughly vertically: the network is saturated and each new conversation causes a conversation's worth of speech samples to be lost.

effect, to slot themselves and thereby interfere only minimally with each other. As the traffic increases, though, talkspurts begin to arrive faster than they can settle in and this structure begins to disintegrate. It is therefore to be expected that the addition of data traffic, with its inherent asynchronous nature, will interfere with the voice traffic more than its share, so that the addition of some amount of data traffic will eliminate more than an equivalent amount of voice traffic capacity.

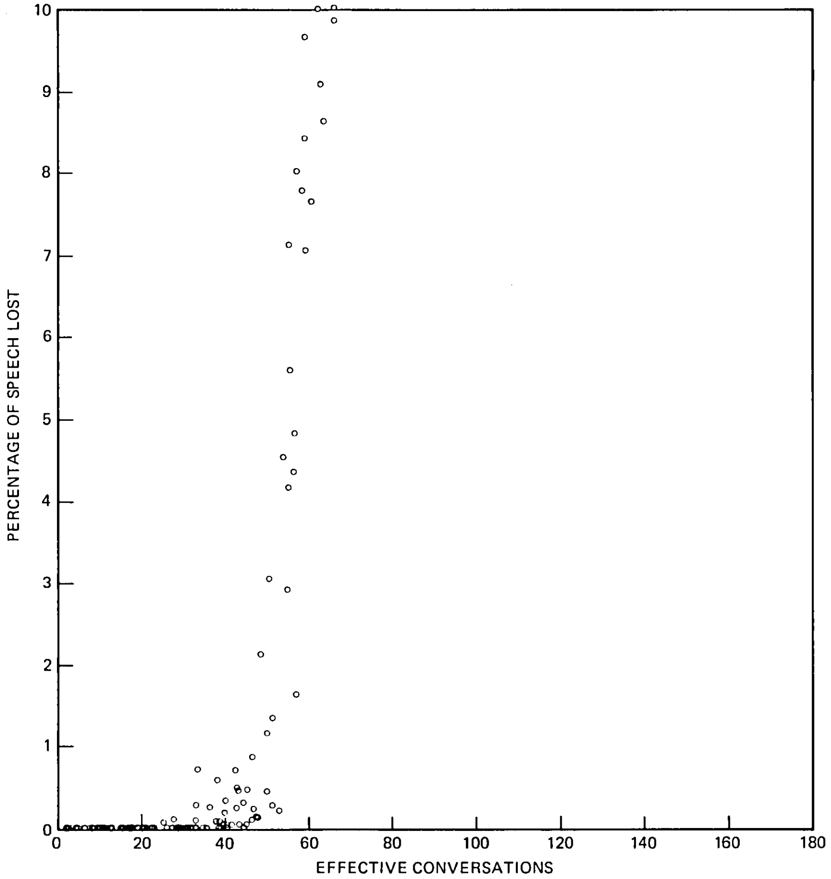


Fig. 5—CSMA/CD voice channel degradation with increasing load with 5.75-ms packets and 5.75-ms artificial delay, with a nonadaptive algorithm. This graph shows the same relation between simulated network load and voice channel degradation as does Fig. 4, except that it uses a simple nonadaptive transmission algorithm. As we see, the expected performance of such a system is significantly poorer than one with the adaptive algorithm, in regard both to the point at which degradation begins and the point at which the curve becomes essentially vertical.

This phenomenon does in fact occur. Figure 6 shows the delay experienced by voice packets on an CSMA/CD system with 5-percent data loading. Note that there is no longer any region of essentially zero delay, as in Fig. 2 without data loading, and that the knees of the curves, although significantly less well-defined here, certainly occur more than 5 percent sooner than earlier. Figure 7 shows the channel degradation allowing 5.75-ms buffering at the receiver.

Additional simulation results, not shown here, were obtained for 10-percent data loading; they basically extend this trend.

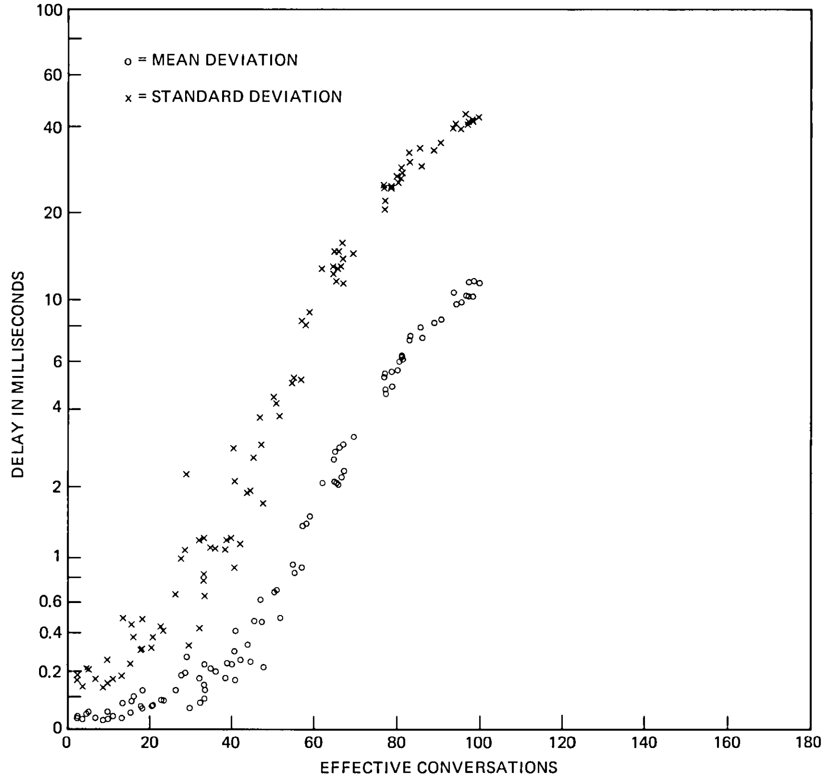


Fig. 6—CSMA/CD delay for 5.75-ms voice packets with 5-percent data loading. This graph shows the effect of a 5-percent data traffic loading on the mean and the standard deviation of the voice packet delay on the simulated CSMA/CD network; it should be compared with Fig. 2, where no data loading was assumed. Notice that there is no longer any large region of essentially zero delay, and that the 5-percent data loading has shifted the curves to the left far more than 5 percent.

## VII. INCREASING THE DELAY IN A CSMA/CD SYSTEM

We can increase the effective bandwidth of the system by increasing the packet size at the transmitter or by increasing the variable delay threshold at the receiver. If we choose a relatively large value of 50 ms for each, resulting in a 100-ms total delay through the system, we find that, as shown in Fig. 8, about 150 effective conversations can take place on the *Ethernet* CSMA/CD network in the absence of data. Assuming 5-percent data loading reduces this number to about 125 effective conversations, as shown in Fig. 9.

It seems likely that the point of diminishing returns has been reached at the 50-ms level; further increases in the packet size or receiver delay cannot produce any great increase in the capacity of the network, but they could subjectively degrade the channel by increasing its delay.

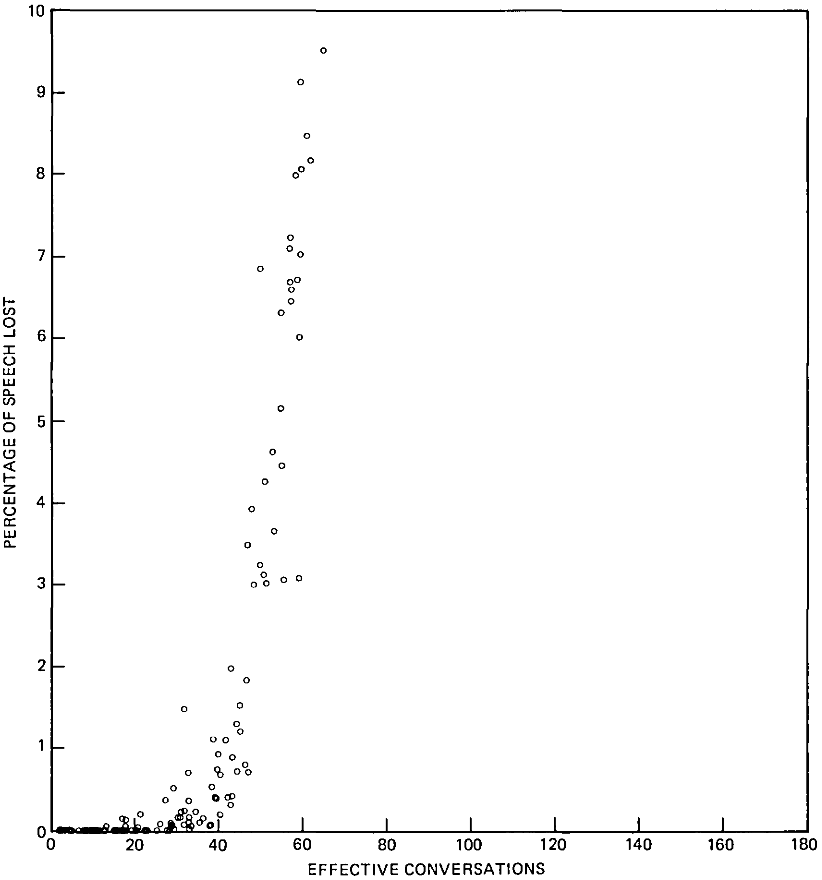


Fig. 7—CSMA/CD voice channel degradation with increasing load with 5.75-ms packets and 5.75-ms artificial delay, with 5-percent data loading. This graph shows the signal degradation on simulated voice channels over a CSMA/CD network in the presence of 5-percent data loading; it should be compared with Fig. 4, in which there was no data loading. Again, two simulations were performed. Degradation rises significantly earlier than with no data loading; there is more than a 5-percent degradation in the effective bandwidth. The effect of a 5-percent data loading on a system with a nonadaptive fixed packet size (not shown here) is comparatively less, since it does not take advantage of the synchronous nature of the voice packets.

## VIII. A TOKEN BUS

An additional simulation study was performed to determine the suitability of a *token-passing* LAN for carrying voice. In a token-passing LAN, contention is resolved through use of a conceptual circulating token. A station may transmit only if it has possession of the token, and must then pass the token to the next station in logical sequence. A *token ring* is a token-passing LAN with a physical ring

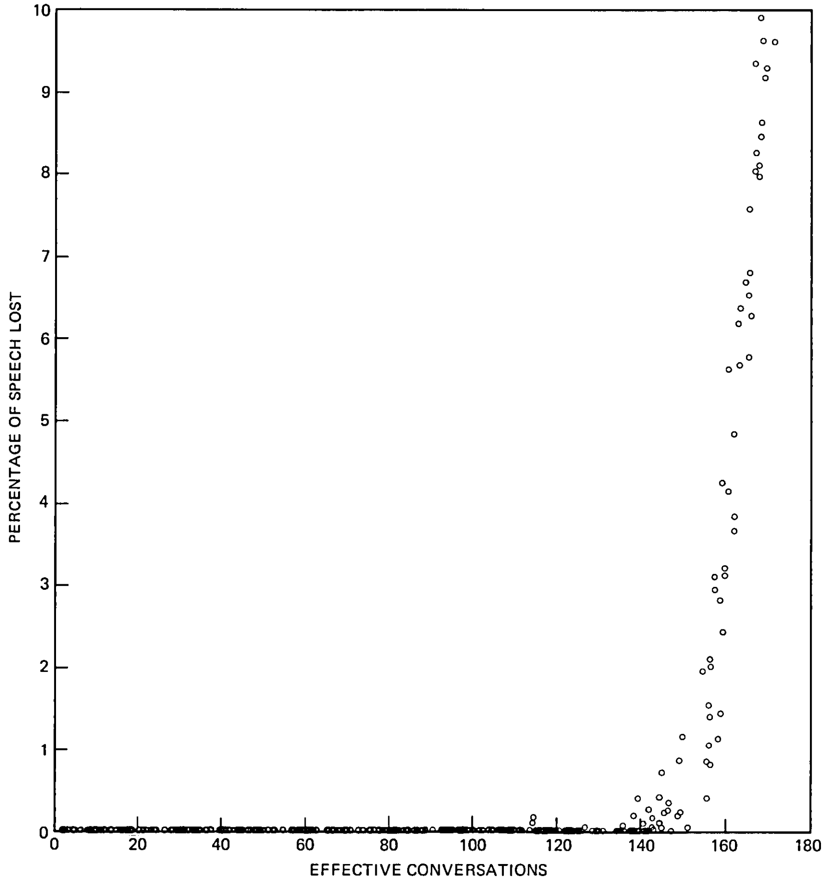


Fig. 8—CSMA/CD voice channel degradation with increasing load with 50-ms packets and 50-ms artificial delay. This graph shows the signal degradation experienced on simulated voice channels over a CSMA/CD network with 50-ms packets and an additional 50-ms artificial packet delay at the receiving station; it should be compared with Fig. 4, which assumes smaller numerical values. As in Fig. 4, two simulations were performed and their results superimposed. We see that a large increase in the delay through the system can produce a significant increase in its effective bandwidth.

topology; the logical sequence is typically the same as the physical sequence of stations on the ring. A *token bus* is a token-passing LAN with a physical bus topology (linear or acyclic branching); the logical sequence can often be arbitrary but is most efficient if it corresponds to the physical sequence.

A token bus was chosen as the token-passing LAN most directly comparable with a CSMA/CD LAN, and the numerical parameters of the token bus were chosen to be as similar as possible to those of the *Ethernet* specifications and the choices of the above CSMA/CD sim-

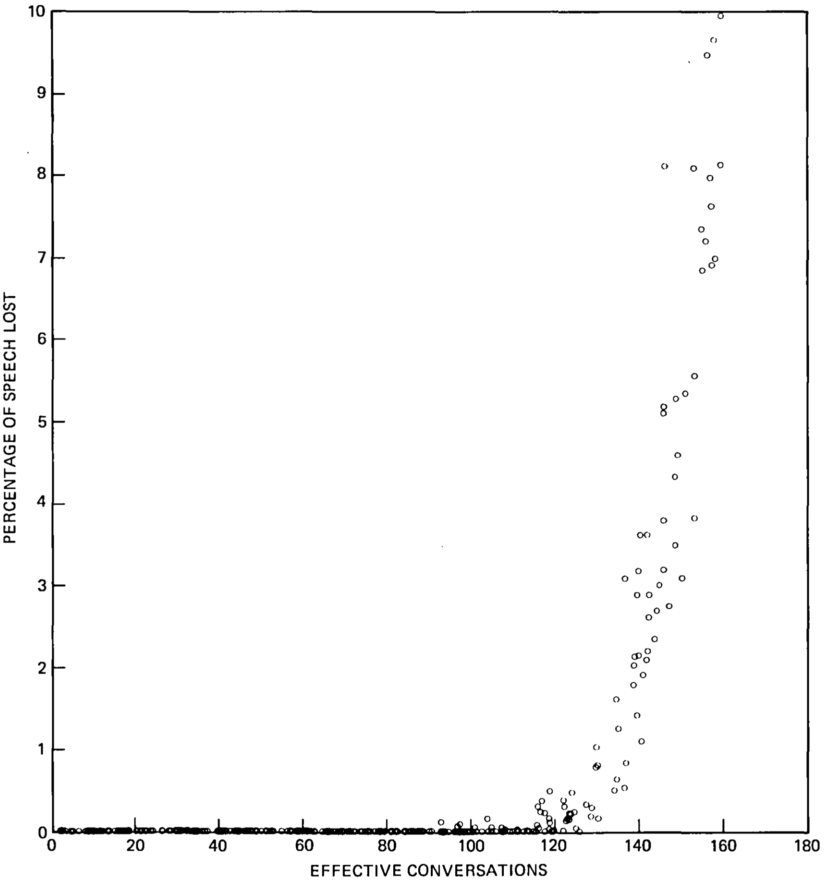


Fig. 9—CSMA/CD voice channel degradation with increasing load with 50-ms packets and 50-ms artificial delay, with 5-percent data loading. This graph shows the signal degradation experienced on simulated voice channels over a CSMA/CD network with a long delay through the system, in the presence of 5-percent data loading; it should be compared with Fig. 8, in which there was no data loading. Again, two simulations were performed and their results superimposed. Note that the proportionate drop in effective bandwidth caused by the data traffic is much less than when small delays were considered, as in the difference between Figs. 4 and 7.

ulations. These choices are quite possibly far from optimal for a token-passing LAN, but they allow for a simple comparison with the CSMA/CD results; there is no typical design for token-passing systems that corresponds to *Ethernet* among CSMA/CD systems.

The simulated token bus LAN has a single token circulating; when a station receives the token, it either transmits a packet, which implicitly passes the token to the next station in logical sequence, or it transmits an abbreviated packet, containing only a header, to explicitly pass the token. The bus is linear, of maximum *Ethernet* length, with stations uniformly distributed. No attempt is made to match the token-passing sequence to the physical sequence of stations on the bus, or to model the (small) control traffic needed to expand the sequence when new conversations, and their associated stations, are added.

The voice packet delays in the token bus simulation are shown in Fig. 10. [To understand the strange shape of the curves in Figure 10, consider a simplified case. We transmit (nominally) 5-ms packets. There are enough stations in the ring for the token to require 2 ms to circulate in the absence of any transmissions. Assume that for each

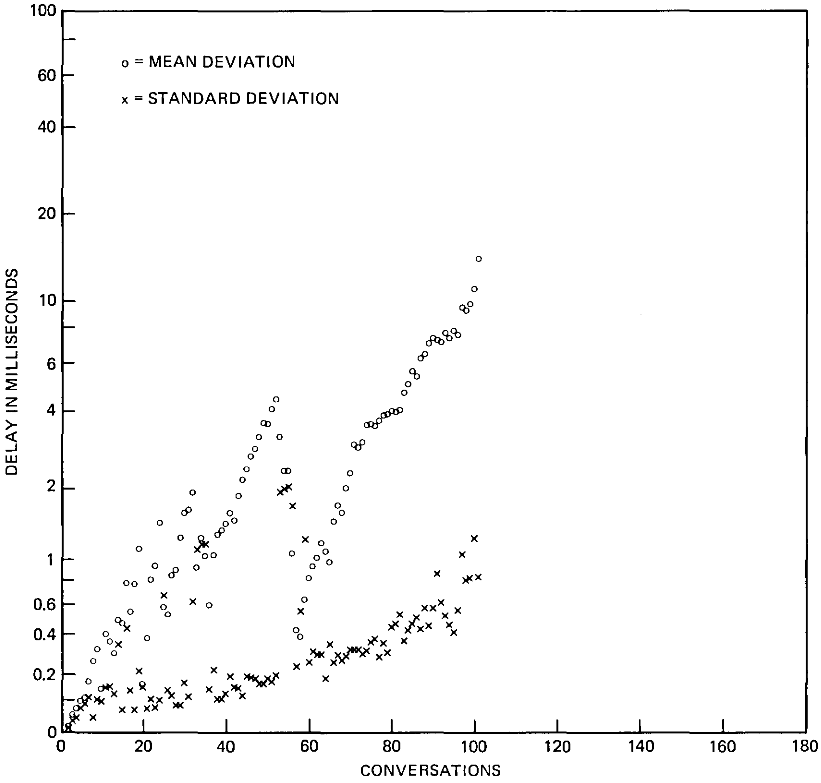


Fig. 10—Token bus delay for 5.75-ms voice packets in the absence of data. This graph shows the mean and standard deviation of the transmission delay on a simulated token bus with no data; it should be compared with Fig. 2, which shows the corresponding delay for a CSMA/CD network. The mean delay is never less than for the CSMA/CD case; the standard deviation under heavy load is much less than for the CSMA/CD case but greater under light load. The sawtooth shape of the curves show that these figures are nonunique and depend on the transmission history. The horizontal axis measures actual conversations instead of effective conversations since even silent stations take part in token circulation.

active station (one associated with an active speaker) to transmit its 5-ms packet each time around would require an additional 4 ms total. An active station will be ready to transmit 5 ms after it has last transmitted, but if every station transmits every time around, the token will take (over) 6 ms to circulate, and so every station will experience the same (over) 1-ms delay; this can remain as constant as the load on the net. On the other hand, if the token were circulating faster, so that it needed only 4 ms for its transit, then a station would transmit only every other time around and experience a delay of 3 (4+4-5) ms. If stations transmitted only every other time around, the time needed under the original assumptions for a token cycle will be 2+4/2 = 4 ms, as assumed; this shows that the performance of a

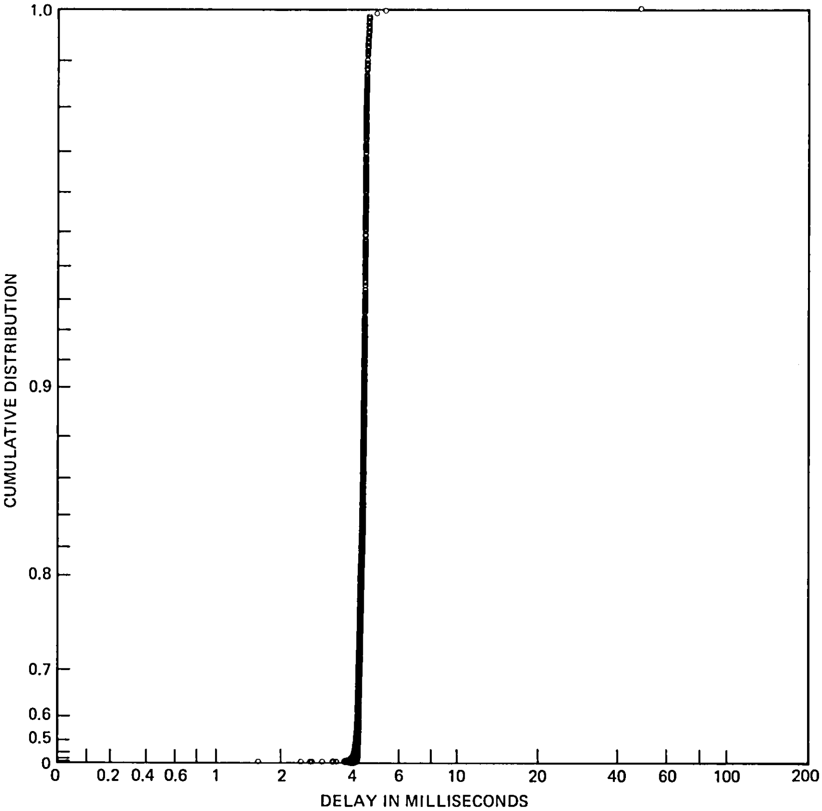


Fig. 11—Token bus cumulative distribution of voice packet transmission delays for a loading of 50 conversations and 5.75-ms packets. This graph shows the cumulative distribution of the variable transmission delays experienced on a simulated token bus loaded with 50 conversations and no data traffic; it should be compared with Fig. 3, which shows the equivalent case for a simulated CSMA/CD network. We note that almost all packets are delayed essentially the same amount of time, which reflects an essentially constant token circulation rate during this period.

token-passing system can be nonuniquely determinable from the load, and can therefore depend upon history.] We note that the mean delay is never less than the mean delay for the corresponding CSMA/CD case shown in Fig. 2. However, the standard deviation for a token ring under sufficient load is much smaller than for the CSMA/CD network: all packets experience very similar delays, as shown in Fig. 11.

Adding 5-percent data loading to a token bus increases the delays to those shown in Fig. 12. Again, the mean is never less than the mean for the CSMA/CD case shown in Fig. 6, but the standard deviation is much smaller under heavy load.

To allow a more direct comparison, Fig. 13 shows the capacity of a CSMA/CD network as a function of the amount of buffering at the receiver, with 5.75-ms packets and no data loading, and allowing 1-

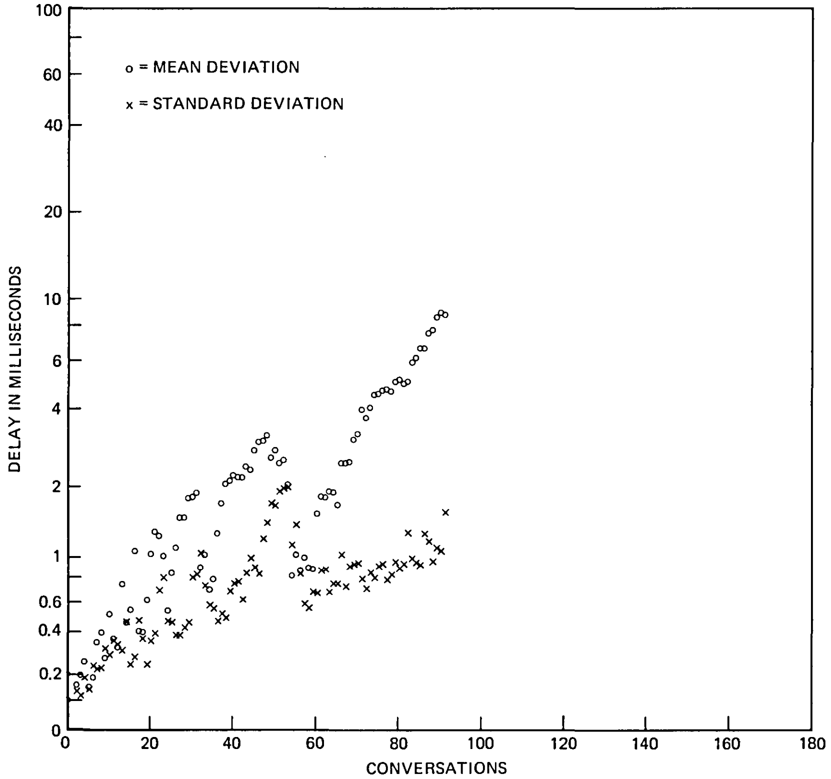


Fig. 12—Token bus delay for 5.75-ms voice packets with 5-percent data loading. This graph shows the mean and the standard deviation of the delay experienced in the transmission of voice packets on a simulated token bus with 5-percent data loading; it should be compared with Fig. 6, which shows the corresponding delay for a CSMA/CD network. Note that the mean delay is never less than for the CSMA/CD case; the standard deviations for a token bus are significantly less than for CSMA/CD under heavy load, although they are greater under light load.

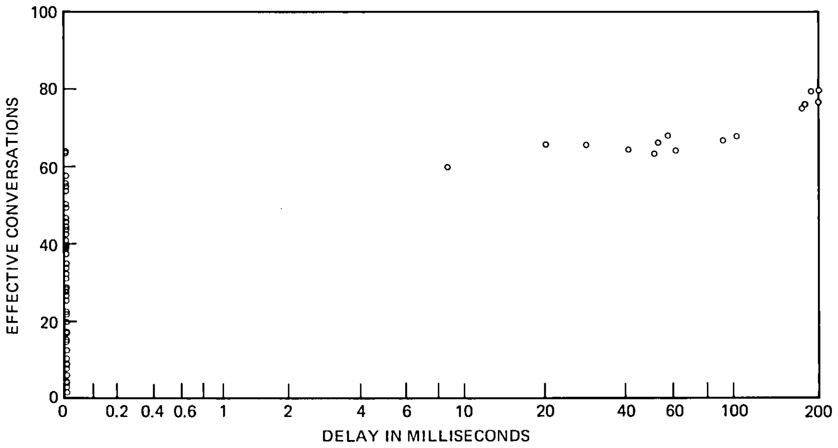


Fig. 13—CSMA/CD capacity as a function of receiver buffering delay, with 5.75-ms packets, 1-percent sample loss, and no data loading. This graph shows the capacity, measured in effective conversations, of a simulated CSMA/CD network as a function of the buffering delay at the receiving station, with (nominally) 5.75-ms packets and allowing up to 1-percent of the speech samples to be lost (at the receiver), in the absence of data. We note that the capacity depends very little on the buffering at the receiver; this suggests that, of some total allowable delay through the system, more delay should be allocated to packet length than to receiver buffering.

percent speech sample loss. No compensation at the transmitter for the buffering at the receiver, in the form of locally discarding samples that would otherwise simply be discarded remotely, was performed in these simulations. Figure 14 shows CSMA/CD capacity with 5-percent data loading. By contrast, Figs. 15 and 16 show the corresponding relations for the token bus with no data loading and with 5-percent data loading, respectively. Figures 13 through 16 show the token bus to offer significantly more capacity than the CSMA/CD network, suggesting that a token-passing network is superior to an CSMA/CD network for voice transmission. We can see that the CSMA/CD LAN's performance is much less dependent on the receiver delay than is that of the token ring, suggesting that the increase in the overall delay caused by increasing the receiver buffering would be better spent in increasing the packet length while keeping the receiver delay relatively small. On the other hand, a token bus can efficiently keep a relatively small nominal packet size and benefit directly from an increase in receiver buffering, as shown.

## IX. CONCLUSIONS

It is possible to transmit a large number of voice conversations on either a CSMA/CD LAN or a token-passing LAN in the presence of

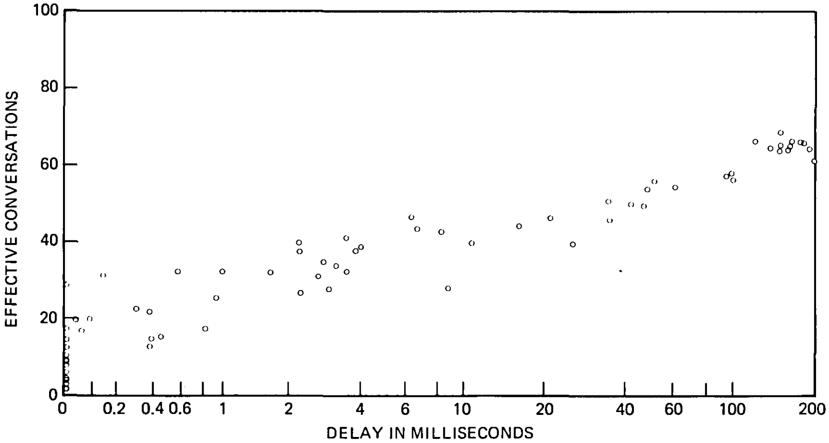


Fig. 14—CSMA/CD capacity as a function of receiver buffering delay, with 5.75-ms packets, 1-percent sample loss, and 5-percent data loading. This graph shows the capacity, measured in effective conversations, of a simulated CSMA/CD network as a function of the buffering delay at the receiving station, with (nominally) 5.75-ms packets and allowing up to 1-percent of the speech samples to be lost (at the receiver), with 5-percent data loading. As in Fig. 13, which considered the corresponding case with no data loading, the capacity depends fairly little on the amount of buffering at the receiver, again suggesting that receiver buffering should be kept fairly small and its share of the overall delay used in allowing the packet size to grow.

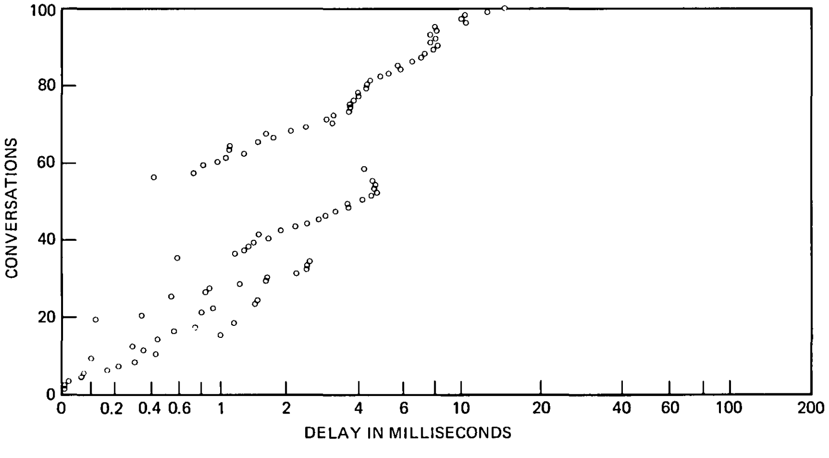


Fig. 15—Token bus capacity as a function of receiver buffering delay, with 5.75-ms packets, 1-percent sample loss, and no data loading. This graph shows the capacity, measured in actual conversations, of a simulated token bus as a function of the buffering delay at the receiving station, with (nominally) 5.75-ms packets and allowing up to 1-percent of the speech samples to be lost (at the receiver), in the absence of data. The significant increase in capacity with increased receiver buffering, plus some reasoning on the nature of token-passing, suggest that the total delay through a token bus system should be allocated predominantly to receiver buffering, with relatively small nominal packet sizes.

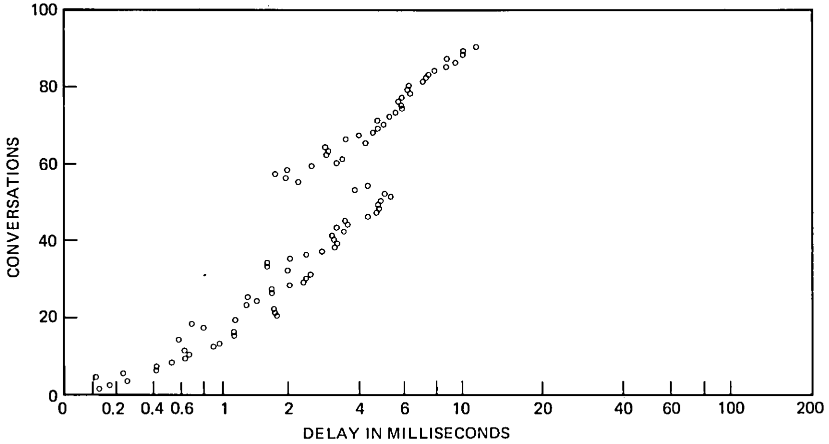


Fig. 16—Token bus capacity as a function of receiver buffering delay, with 5.75-ms packets, 1-percent sample loss, and 5-percent data loading. This graph shows the capacity, measured in actual conversations, of a simulated token bus as a function of the buffering delay at the receiving station, with (nominally) 5.75-ms packets and allowing up to 1-percent of the speech samples to be lost (at the receiver), with 5-percent data loading. As in Fig. 15, which considered the corresponding case with no data loading, the capacity depends significantly on the amount of buffering at the receiver, again suggesting that receiver buffering in a token bus system should be kept fairly large and the nominal packet size fairly small.

reasonable data loading. The performance of token-passing seems superior to that of CSMA/CD.

There are even better mechanisms for transmitting digital voice: time-division multiplexing schemes, for example, can do an excellent job for voice, but are not exceptional for carrying data because of their inherently synchronous nature. Similarly, CSMA/CD LANs can be superior to token-passing for many data applications. It is still a research problem to find an LAN that can carry both voice and data "optimally", or to identify more exactly the appropriate trade-offs.

One significant unanswered question is the potential of such a system for serving large numbers of users; there is an inherent limit of the number of users on one LAN. It is uncertain to what extent internetworking can help, since internetworking would increase the mean and standard deviation of the delays.

## REFERENCES

1. *The* Ethernet/*A Local Area Network/Data Link Layer and Physical Layer Specifications,* Digital Equipment Corporation, Intel Corporation and Xerox Corporation, Version 1.0, September 30, 1980.
2. J. F. Shoch and J. A. Hupp, "Measured Performance of an Ethernet Local Network," Comm. ACM, *23,* 9 (December 1980), pp. 711–21.
3. P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., *47,* 1 (January 1968), pp. 73–91.
4. G. J. Nutt and D. L. Bayer, "Performance of CSMA/CD Networks Under Combined

Voice and Data Loads," IEEE Trans. Commun., *COM-30,* 1 (January 1982), pp. 6–11.

## AUTHOR

**John D. DeTreville,** B.S. (Mathematics), 1970, University of South Carolina; S.M. (Computer Science), 1972, Massachusetts Institute of Technology; Ph.D. (Computer Science), 1978, Massachusetts Institute of Technology; AT&T Bell Laboratories, 1978—. Mr. DeTreville's first work at AT&T Bell Laboratories was on 5*ESS*™; he moved to Murray Hill in 1980, where his work has incorporated topics in distributed systems and program synthesis. Member, Association for Computing Machinery.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# Trunk Implementation Plan for Hierarchical Networks

By A. N. KASHPER* and G. C. VARVALOUCAS*

(Manuscript received March 29, 1983)

The Trunk Implementation Plan (TIP) is a multiyear schedule of planned trunk augments and disconnects that minimizes the impact of varying demand and forecast uncertainties on the cost of implementing a network meeting objective service criteria. This paper presents a theoretical development of the TIP algorithm for hierarchical networks that accounts for modularity, facility, and demand servicing constraints. First, we solve the only-route TIP problem analytically using stochastic dynamic programming techniques. We show that our analytical solution yields a numerically efficient algorithm for calculating a multiyear, minimum-cost policy. Second, we present the results of our analysis showing that, in the presence of forecast uncertainty, a near-optimal traffic network is obtained by introducing reserve capacity on the final trunk groups only. Based on this result, we construct the TIP algorithm for hierarchical networks by combining conventional network engineering principles with an optimal disconnect policy for high-usage trunk groups and the only-route TIP sizing procedure for final groups. Using this algorithm we obtain an economical multiyear schedule of trunk augments and disconnects for hierarchical networks.

## I. INTRODUCTION

### *1.1 Background and motivation*

In the Bell operating companies and AT&T Communications, the trunk forecasting process consists of: (1) traffic measurement and offered load estimation, (2) projection of future traffic demands, and

* AT&T Bell Laboratories.

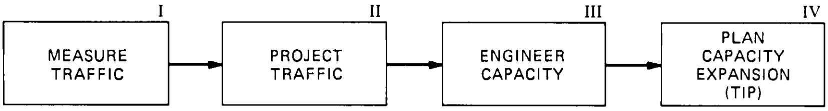


Fig. 1—Trunk network forecasting process.

(3) determination of the trunk group sizes for each of five or more future years. Currently, the engineering procedure utilized in (3) is based on independent, single-year network designs, each of which minimizes the cost of satisfying anticipated demands for a given future year.

Although substantial work has been done to improve the quality of the trunk forecasting process,[1-3] the existing methods do not account for several important implementation considerations, namely, the existing trunk network, the variation of trunk demand from year to year, uncertainty of demand forecasts, economics of maintaining or rearranging trunks, and facility constraints. Consequently, in practice, the output of (3) is modified by trunk forecasters to make the final multiyear schedule of planned trunk augments and disconnects feasible and economically sensible. The adjustments to the mechanized forecasting process are based on heuristic trunk disconnect guidelines and engineering judgment. However, no quantitative attempt is made to find an optimal multiyear trunk provisioning plan.

Accordingly, we identified the need for a mechanized system that would compute an economical capacity expansion plan for the trunk network. As Fig. 1 illustrates, the capacity expansion planning can be regarded as the fourth major function of the trunk forecasting process. The new mechanized system, called the Trunk Implementation Plan (TIP), will provide a multiyear schedule of trunk augments and disconnects that accounts for facility constraints while minimizing the impact of forecast uncertainty and demand dynamics on the cost of implementing a network meeting objecting service criteria.

In this work, we present a theoretical development of the TIP algorithm and generalize the mathematical model of Ref. 4 to reflect modularity and facility constraints. The problem formulation in Ref. 4 assumes a nonmodular engineering environment and no facility constraints; i.e., the multiyear schedule of trunk augments and disconnects is cost-effective under the assumption that the sizes of the trunk groups are nonnegative real numbers and that the cost of a trunk group is proportional to the number of trunks in the group.

However, the assumptions of nonmodular engineering do not hold, for example, for a new generation of digital terminals such as the Digital Carrier Trunk (DCT), used for the 1A *ESS*,* and the Digital

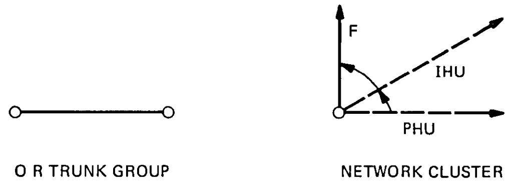
*Trademark of AT&T Technologies, Inc.

Interface Frame (DIF), used for 4*ESS*. The DCT and DIF require that digital carriers [equivalent to 24 voice frequency (VF) circuits] terminate directly on the switch. The dedication of network facilities by destination implies that the cost of a trunk group has a per-module component in addition to the per-circuit component. Therefore, in Ref. 1, W. Elsner concluded that trunk groups terminating on such facilities should be modularly engineered. In particular, Ref. 1 shows that an engineering procedure that assumes only modular sizes (24 trunks for two-way groups) for high-usage and certain final trunk groups provides significant economic benefits; see Fig. 2 for definitions of trunking terminology.

In this paper, we replace the continuous TIP model of Ref. 4 by a discrete TIP formulation that incorporates modularity constraints. Then we derive an optimal modular network expansion policy for high-usage and final trunk groups. In addition, we show how to modify the TIP multiyear trunk-sizing procedures to reflect facility constraints.

### *1.2 Overview*

Section II defines the notation and describes the mathematical model for the only-route TIP problem. In Section III we present a complete solution to the only-route TIP problem that accounts for modularity, facility, and demand servicing constraints. Section IV shows how to combine conventional network engineering principles



HU - HIGH-USAGE GROUP—DESIGNED TO OVERFLOW TO AN ALTERNATE ROUTE

PHU - PRIMARY HU GROUP—HU THAT DOES NOT RECEIVE OVERFLOW TRAFFIC

IHU - INTERMEDIATE HU GROUP—HU THAT RECEIVES OVERFLOW TRAFFIC

F - FINAL GROUP—LAST-CHOICE GROUP THAT IS ENGINEERED TO A SPECIFIC BLOCKING OBJECTIVE

O R - ONLY-ROUTE GROUP—F GROUP THAT RECEIVES ONLY FIRST ROUTE TRAFFIC

Fig. 2—Trunking terminology.

with the only-route TIP results to design an economical multiyear hierarchical network.

## II. TIP MODEL FOR ONLY-ROUTE TRUNK GROUPS

We start our derivation of the TIP algorithm by considering the multiyear engineering problem for only-route trunk groups (see Fig. 2 for a definition of only-route trunk groups). As we discuss in Section 4.3, our solution of the only-route problem will be utilized to plan multiyear capacity expansion for hierarchical networks.

### *2.1 Mathematical model*

#### *2.1.1 Notation*

First, we define the notation used in our mathematical model.

$T(k)$—number of trunks in service at the beginning of the $k$th year

$u(k)$—number of planned trunk augments/disconnects at the beginning of the $k$th year

$d(k)$—the maximum number of trunks (peak demand) in trunks during year $k$ to guarantee the engineered blocking level

$F_k$—the distribution of the peak trunk demand during year $k$

$c_1^k(u(k))$—capital cost during year $k$

$c_2^k(u(k))$—labor cost during year $k$

$c_3^k(T(k), u(k))$—maintenance cost during year $k$

$c_4^k(d(k), T(k), u(k))$—underprovision cost during year $k$

$N$—number of years in the forecast horizon

We assume that the number of trunks in service, $T(k)$; the planned trunk level, $T(k) + u(k)$; and the peak demand, $d(k)$; are expressed in modules of $m$ trunks. In the message network, $m$ is equal to 1, 12, or 24. Consequently, $F_k$ is a discrete distribution function, defined in accordance with the established rounding rules for engineering modular final groups.[1] That is, if $F_k'$ is a continuous distribution of the peak demand, then $F_k$ is obtained by

$$F_k(m\ell) = F_k'(m\ell + r_\ell),$$

where $\ell$ is a nonnegative integer, and $r_\ell$ is a rounding threshold, $0 < r_\ell < m$.

#### *2.1.2 Trunk group dynamics*

According to the AT&T practice, if the blocking objective on an only-route or alternate final trunk group is violated significantly, then the trunk group is augmented during the year on an emergency basis (demand servicing) to restore the engineered blocking level. Therefore, the number of trunks in service at the beginning of the $(k + 1)$th year is the sum of the planned trunk level for year $k$ and the demand servicing augmentation, if any, during year $k$. Thus, the trunk group dynamics that reflect the planned and demand servicing components of the trunk provisioning process are modeled by

$$\begin{aligned} T(k+1) &= [T(k) + u(k)] + \max[0, d(k) - (T(k) + u(k))] \\ &= \max[y(k), d(k)], \end{aligned} \tag{1}$$

where $y(k) = T(k) + u(k)$ represents the planned trunk level at year $k$.

### *2.1.3 Objective function*

The goal of the only-route TIP is to minimize the expected present worth of trunk provisioning costs. If we denote the present worth of the total cost for year $k$ by $g_k(d(k), T(k), u(k))$, then the TIP objective function can be expressed as

$$\min_{\mathbf{u}} J(\mathbf{u}) = \min_{\mathbf{u}} E \left\{ \sum_{i=0}^{N-1} g_i[d(i), T(i), u(i)] \right\}, \tag{2}$$

where $\mathbf{u} = (u(0), \cdots, u(N-1))$ and the expected value is taken over the demands $d(0), d(1), \cdots, d(N-1)$.

The present worth of trunk provisioning costs at year $k$ is equal to

$$\begin{aligned} g_k(d(k), T(k), u(k)) &= \rho^k[c_1^k(u(k)) + c_2^k(u(k)) \\ &\quad + c_3^k(T(k), u(k)) + c_4^k(d(k), T(k), u(k))], \end{aligned} \tag{3}$$

where $\rho$ is the discount factor ($\rho < 1$) that measures the worth of the next year's dollars in terms of present dollars.

The capital, labor, maintenance, and underprovision costs are assumed to be piecewise linear with respect to modules of $m$ trunks and are defined for $k = 0, \cdots, N-1$ by

$$c_1^k(u(k)) = \begin{cases} a_1^k u(k) & u(k) \geq 0 \\ b_1^k u(k) & u(k) < 0, \end{cases} \tag{4}$$

$$c_2^k(u(k)) = \begin{cases} a_2^k u(k) & u(k) \geq 0 \\ -b_2^k u(k) & u(k) < 0 \end{cases} \tag{5}$$

$$c_3^k(T(k), u(k)) = a_3^k y(k), \tag{6}$$

$$c_4^k(d(k), T(k), u(k)) = a_4^k \max(0, d(k) - y(k)), \tag{7}$$

where $y(k) = T(k) + u(k)$ is a planned trunk level at year $k$.

Equation (7) states that if the peak demand $d(k)$ exceeds the planned trunk level $y(k)$ during year $k$, then the cost of providing the additional trunks is proportional to the trunk shortage; if $d(k)$ does not exceed the planned level, then no cost is incurred. Thus, the underprovision cost reflects the demand servicing policy as described by the trunk-group dynamics equation (1). The assumption that the underprovision cost is linear with respect to the trunk shortage will be discussed in Section 2.1.4.

We assume that in (4) through (7) the per-trunk costs $a_1^k, \cdots, a_4^k$ and $b_1^k$, $b_2^k$ are nonnegative and satisfy the conditions

$$a_1^k + a_2^k > b_1^k - b_2^k, \tag{8a}$$

$$b_1^k - b_2^k > 0, \tag{8b}$$

$$a_1^k + a_2^k > \rho(a_1^{k+1} + a_2^{k+1}), \tag{9a}$$

$$b_1^k - b_2^k > \rho(b_1^{k+1} - b_2^{k+1}), \tag{9b}$$

$$a_4^k > a_1^k + a_2^k + a_3^k. \tag{10}$$

Inequalities (8a) and (8b) state: first, that the cost of buying and installing a trunk module is always greater than its salvage value minus the disconnect expense; and second, that there is always an incentive for disconnecting a trunk module. Inequalities (9a) and (9b) show that it is uneconomical to augment a trunk group if not necessary and also uneconomical to delay the disconnect decision. Finally, (10) reflects the fact that it is always more costly to augment a trunk group on an emergency, rather than on a planned, basis.

### *2.1.4 Demand servicing constraints*

In general, the underprovision cost or, equivalently, the unsatisfied demand penalty cost, involves all of the costs of planned servicing: capital, labor, and maintenance, plus a penalty due to the fact that demand servicing cannot be carried out with the normal planning intervals and orderly procedures associated with planned servicing. It is important to note that, in practice, the incremental underprovision cost depends on a variety of factors such as switch, trunk, and/or personnel availability. Typically, if the existing personnel and facilities can satisfy the emergency servicing need, then the cost of underprovisioning is marginally higher than the cost of planned trunk augmentation and can be computed by (7). However, if there is a shortage of personnel or facilities, demand servicing becomes much more expensive (than planned servicing) and highly undesirable.

Thus, to complete the problem formulation we need to specify that the feasible solutions in the TIP model correspond to a level of demand servicing not greater than an allowable threshold. In particular, we require that the expected amount of demand servicing at year $k$ not exceed a specified level

$$E\{\max(0, d(k) - T(k) - u(k))\} \leq \beta_k Ed(k),$$

where $k = 0, \cdots, N - 1$ and $\beta_k$ is a given constant.

### *2.1.5 Mathematical model*

The TIP problem can be viewed as a sequential stochastic decision process. The state of the system, the number of trunks in service, varies according to the trunk group dynamics given in eq. (1). At each state of the process the cost function $g_k(d(k), T(k), u(k))$ is defined via (4) through (7). The problem then is: given the initial trunk level $T(0)$ and future peak demand distributions $F_0, F_1, \cdots, F_{N-1}$, find a set of decisions (augments, disconnects), $\mathbf{u}^* = \{u^*(0), \cdots, u^*(N-1)\}$, the optimal policy, that minimizes the total expected trunk provisioning cost over $N$ years

$$J(\mathbf{u}^*) = \min_{\mathbf{u}} E\left\{\sum_{k=0}^{N-1} g_k(d(k), T(k), u(k))\right\} \tag{11}$$

subject to possible capacity limitation conditions

$$0 \le T(k) + u(k) \le \gamma(k) \tag{12}$$

and demand servicing constraints

$$E\{\max(0, d(k) - T(k) - u(k))\} \le \beta_k E d(k), \tag{13}$$

where $k = 0, 1, \cdots, N-1$ and $\gamma(k)$ are given modular thresholds that represent facility constraints.

## III. SOLUTION FOR ONLY-ROUTE TRUNK GROUPS

Throughout the rest of this paper, we assume that the random variables $d(k)$ are statistically independent. We shall prove that under this assumption the optimal control law to the only-route TIP problem of Section 2.1.5 is defined by $N$ pairs of scalars $(\underset{\sim}{S}(0), \tilde{S}(0)), \cdots, (\underset{\sim}{S}(N-1), \tilde{S}(N-1))$. The pair $(\underset{\sim}{S}(k), \tilde{S}(k))$ provides two critical levels for year $k$. Specifically, at year $k$, the optimal control law is to augment the number of trunks up to the level $\underset{\sim}{S}(k)$, to maintain the trunk level if $\underset{\sim}{S}(k) \le T(k) < \tilde{S}(k)$, or to disconnect down to the level $\tilde{S}(k)$; that is, for $k = 0, 1, \cdots, N-1$ the optimal decision $u_k^* = u_k^*(T(k))$ is given by

$$u_k^*(T(k)) = \begin{cases} \underset{\sim}{S}(k) - T(k) & \text{if} \quad T(k) < \underset{\sim}{S}(k) \\ 0 & \text{if} \quad \underset{\sim}{S}(k) \le T(k) < \tilde{S}(k) \\ \tilde{S}(k) - T(k) & \text{if} \quad T(k) \ge \tilde{S}(k). \end{cases} \tag{14}$$

We note that the independence assumption for future demands is critical for obtaining an analytical solution to the only-route TIP problem. However, our numerical experience shows that cost-effectiveness of the TIP solution as compared to currently used heuristic trunk provisioning strategies is not sensitive to this assumption.

We start our derivation by considering the TIP formulation that ignores demand servicing constraints. Specifically, Sections 3.1 and 3.2 present a complete solution to the modular only-route TIP problem described by (11) and (12). A similar nonmodular capacity management problem is considered in Ref. 5. In Section 3.3, we generalize this solution and obtain a numerically efficient algorithm that computes an optimal policy under the constraint on the level of demand servicing.

### *3.1 Optimality conditions*

Since the demands are assumed to be independent, the TIP objective function (2) can be transformed to the form[6]

$$\min_{\mathbf{u}} J(\mathbf{u}) = \min_{u(0)} \mathop{E}_{d(0)} \{g_o(d(0), T(0), u(0)) + \min_{u(1)} \mathop{E}_{d(1)} \{\cdots + \min_{u(N-1)} \mathop{E}_{d(N-1)} \{g_{N-1}(d(N-1), T(N-1), u(N-1))\} \cdots\}\}.$$

To convert the right-hand side into a recursive form, we introduce the optimal cost-to-go function for the year $k$, $V_k(T(k))$, that is, the minimum expected cost over all possible strategies for years $k$, $k+1, \cdots, N-1$ that assume $T(k)$ trunks in service at the beginning of year $k$. Then, by Bellman's principle of optimality,[6] the optimal policy in year $k$ is obtained by solving the backward dynamic programming recursion:

$$V_k(T(k)) = \min_{u(k)} \mathop{E}_{d(k)} \{g_k(d(k), T(k), u(k)) + \rho V_{k+1}(\max(d(k), T(k) + u(k)))\}, \tag{15}$$

where $V_N(T(N)) = 0$ and the minimum is taken over all modular $u(k)$ such that the planned trunk level at year $k$ satisfies the condition

$$0 \le T(k) + u(k) \le \gamma(k). \tag{16}$$

The sum of trunk provisioning costs for year $k$ is composed of two functions $g_{k,1}(\cdot)$ and $g_{k,2}(\cdot)$ defined by

$$\begin{aligned} & g_k(d(k), T(k), u(k)) \\ &= \begin{cases} g_{k,1}(d(k), T(k), u(k)), & u(k) \ge 0 \\ g_{k,2}(d(k), T(k), u(k)), & u(k) < 0 \end{cases} \\ &= \begin{cases} a_1^k u(k) + a_2^k u(k) + a_3^k(T(k) + u(k)) \\ \quad + a_4^k \max(d(k), T(k) + u(k)), & u(k) \ge 0 \\ b_1^k u(k) - b_2^k u(k) + a_3^k(T(k) + u(k)) \\ \quad + a_4^k \max(d(k), T(k) + u(k)), & u(k) < 0. \end{cases} \end{aligned} \tag{17}$$

In what follows, we consider $g_k(\cdot)$ as a function of $d(k)$, $T(k)$, and

$y(k) = T(k) + u(k)$. Then, from the optimality principle and (17) the year $k$ cost-to-go function is given by

$$J_k(T(k), y(k)) = \begin{cases} J_{k,1}(T(k), y(k)), & y(k) \geq T(k) \\ J_{k,2}(T(k), y(k)), & 0 \leq y(k) < T(k) \end{cases}$$

$$= \begin{cases} \underset{d(k)}{E}\,[g_{k,1}(d(k),\ T(k),\ y(k)) \\ \quad + \rho V_{k+1}(\max(d(k),\ y(k)))], \\ & y(k) \geq T(k) \\ \underset{d(k)}{E}\,[g_{k,2}(d(k),\ T(k),\ y(k)) \\ \quad + \rho V_{k+1}(\max(d(k),\ y(k)))], \\ & 0 \leq y(k) < T(k). \end{cases} \qquad (18)$$

Note that from (17) $J_{k,1}(T, y)$ and $J_{k,2}(T, y)$ are well-defined functions for all $y = 0, m, 2m, \cdots$.

To prove that a solution of (18) is given by (14) we first develop sufficient conditions for the optimality of an $(\underset{\sim}{S}, \tilde{S})$ policy without capacity constraints ($\gamma = \infty$). We observe that $J_{k,1}(T, T) = J_{k,2}(T, T)$ for all $T \geq 0$. Accordingly, to prove the optimality of (14) it is sufficient to show that the solutions of the minimization problems

$$\min_{y \geq T \geq 0} J_1(T, y) \qquad (19)$$

and

$$\min_{T \geq y \geq 0} J_2(T, y) \qquad (20)$$

are defined by

$$y_1^* = \begin{cases} \underset{\sim}{S}, & T < \underset{\sim}{S} \\ T, & T \geq \underset{\sim}{S} \end{cases} \qquad (21)$$

and

$$y_2^* = \begin{cases} \tilde{S}, & T > \tilde{S} \\ T, & T \leq \tilde{S}, \end{cases} \qquad (22)$$

respectively, for some numbers $\underset{\sim}{S}$ and $\tilde{S}$ such that

$$0 \leq \underset{\sim}{S} \leq \tilde{S} < \infty.$$

To simplify the notation, we have dropped the index $k$ in (19) through (22) and in the formulas in the rest of this section.

Consequently, the optimality of (21) and (22) is evident if there exist a pair of numbers $(\underset{\sim}{S}, \tilde{S})$, $\underset{\sim}{S} \leq \tilde{S}$ such that for any $T \geq 0$

$$\begin{cases} J_1(T, x) \geq J_1(T, y), & 0 \leq x \leq y < \underset{\sim}{S} \\ J_1(T, x) \leq J_1(T, y), & \underset{\sim}{S} \leq x \leq y \leq \infty \end{cases} \tag{23}$$

and

$$\begin{cases} J_2(T, x) \leq J_2(T, y), & \tilde{S} \leq x \leq y < \infty \\ J_2(T, x) \geq J_2(T, y), & 0 \leq x \leq y \leq \tilde{S}. \end{cases} \tag{24}$$

To demonstrate (23) and (24) and to construct $\underset{\sim}{S}$ and $\tilde{S}$ we shall prove an even stronger statement. Specifically, we consider the first differences of $J_1$ and $J_2$, defined by

$$\Delta J_1(T, y) = J_1(T, y + m) - J_1(T, y)$$

and

$$\Delta J_2(T, y) = J_2(T, y + m) - J_2(T, y),$$

and show that $\Delta J_1$ and $\Delta J_2$ satisfy the conditions:

1. $\Delta J_1(T, y) = \Delta J_1(y)$ and $\Delta J_2(T, y) = \Delta J_2(y)$,
2. $\Delta J_1(y)$ and $\Delta J_2(y)$ are nondecreasing, and
3. $0 \leq \underset{\sim}{S} \leq \tilde{S} < \infty$,

where

$$\underset{\sim}{S} = \min\{y \mid \Delta J_1(y) \geq 0, y = 0, m, \cdots\}$$

and

$$\tilde{S} = \min\{y \mid \Delta J_2(y) \geq 0, y = 0, m, \cdots\}. \tag{25}$$

Note that (25) defines the minimum points of $J_1$ and $J_2$, respectively. Furthermore, conditions (1) through (3) not only imply the optimality of (14) but also suggest that we can find the critical thresholds efficiently by applying a modular version of the bisection methods to the first differences of $J_1$ and $J_2$. The proof that $J_{k,1}$ and $J_{k,2}$ [defined by (18)] satisfy these conditions is given in Appendix A.

The generalization to the constrained case follows immediately. Because of the monotonicity of $\Delta J_1(y)$ and $\Delta J_2(y)$ the optimal solution under capacity constraints is given by

$$\underset{\sim}{S}^* = \min[\underset{\sim}{S}, \gamma]$$

and

$$\tilde{S}^* = \min[\tilde{S}, \gamma]. \tag{26}$$

### 3.2 Computational procedure

As we prove in Appendix A, an optimal modular TIP policy is described by $N$ pairs of critical threshold levels $\{(\underset{\sim}{S}(0), \tilde{S}(0)), \cdots,$

$(\underset{\sim}{S}(N-1), \tilde{S}(N-1))\}$ and each pair $(\underset{\sim}{S}(k), \tilde{S}(k))$ can be computed by (25).

Consequently, to obtain an algorithm for calculating the optimal policy we shall derive backward recursions for $\Delta J_{k,1}(T(k) + u(k))$ and $\Delta J_{k,2}(T(k) + u(k))$. In Appendix A we show that

$$\Delta J_{k,1}(y(k)) = [a_1^k + a_2^k + a_3^k - a_4^k(1 - F_k(y(k)))]m \\ + \rho\Delta \underset{d(k)}{E}\, V_{k+1}(\max(d(k), y(k))) \quad (27)$$

and

$$\Delta J_{k,2}(y(k)) = [b_1^k - b_2^k + a_3^k - a_4^k(1 - F_k(y(k)))]m \\ + \rho\Delta \underset{d(k)}{E}\, V_{k+1}(\max(d(k), y(k))), \quad (28)$$

where $y(k) = T(k) + u(k)$.

Thus, we can confine our effort to the derivation of the backward recursion for the first difference of the expected optimal cost-to-go function defined by

$$\Delta \underset{d(k)}{E}\, V_{k+1}[\max(d(k), y(k))] \\ = \underset{d(k)}{E}\, V_{k+1}[\max(d(k), y(k) + m)] - \underset{d(k)}{E}\, V_{k+1}[\max(d(k), y(k))]. \quad (29)$$

The details of our derivation are presented in Appendix B; we demonstrate

$$\Delta \underset{d(k)}{E}\, V_{k+1}[\max(d(k), y(k)]$$

$$= F_k(y)\cdot \begin{cases} -(a_1^{k+1} + a_2^{k+1})m & \text{if} \quad y(k) < \underset{\sim}{S}(k+1) \\ [a_3^{k+1} - a_4^{k+1}(1 - F_{k+1}(y(k))]m \\ \quad + \rho\Delta \underset{d(k+1)}{E}\, V_{k+2}\{\max[d(k+1), y(k)]\} \\ \quad \text{if} \quad \underset{\sim}{S}(k+1) \le y(k) < \tilde{S}(k+1) \\ -(b_1^{k+1} - b_2^{k+1})m & \text{if} \quad y(k) \ge \tilde{S}(k+1). \end{cases} \quad (30)$$

Consequently, the backward recursions (27), (28), (30), and formulas (25) define an algorithm that yields the complete set of optimal threshold levels $\{\underset{\sim}{S}(N-1, \tilde{S}(N-1), \cdots (\tilde{S}(0), \underset{\sim}{S}(0))\}$ for the problem described by (15) and (16).

### *3.3 Final solution of the only-route TIP problem*

Now we apply the Lagrangean relaxation approach to show that the computational procedure of Section 3.2 can be used to find an optimal solution to the original TIP formulation (11) through (13) that includes demand servicing constraints. Indeed, let us consider the functional

$$H(\mathbf{u}, \lambda) = J(\mathbf{u}) + \sum_{k=0}^{N=1} \lambda_k \rho^k E \max(0, d(k) - T(k) - u(k)), \quad (31)$$

where $\lambda = (\lambda_0, \cdots, \lambda_{N-1})$.

Clearly, for any given $\lambda \geq \mathbf{0}$, minimizing $H(\mathbf{u}, \lambda)$ with respect to $\mathbf{u}$ subject to (16) is equivalent to solving the problem described by (11) and (12) with the incremental underprovision cost $a_4^k$ replaced by

$$\hat{a}_4^k = a_4^k + \lambda_k.$$

To demonstrate that a solution to the problem (12) and (31) (with appropriately fixed $\lambda$) is, in fact, a solution to the original TIP problem, (11) through (13), we need to prove the following general proposition:

Let $\mathbf{u}^*$ be a minimum of the functional

$$J(\mathbf{u}) + \sum_{k=0}^{N-1} \lambda_k E_k(\mathbf{u})$$

for all $\mathbf{u} \in U$, where $J(\mathbf{u})$, $E_k(\mathbf{u})$ are arbitrary real-valued functions, $U$ is some set of admissible controls, and $\lambda \geq \mathbf{0}$. Then, $\mathbf{u}^*$ is a solution to the problem:

$$\min_{\mathbf{u} \in U} J(\mathbf{u})$$

subject to

$$E_k(\mathbf{u}) \leq E_k(\mathbf{u}^*) \quad (32)$$

for all $k$ such that $\lambda_k > 0$.

The proof of this proposition is simple. By our hypothesis, for all $\mathbf{u} \in U$ we have

$$J(\mathbf{u}^*) + \sum_{k=0}^{N-1} \lambda_k E_k(\mathbf{u}^*) \leq J(\mathbf{u}) + \sum_{k=0}^{N-1} \lambda_k E_k(\mathbf{u}),$$

or

$$J(\mathbf{u}^*) \leq J(\mathbf{u}) + \sum_{k=0}^{N-1} \lambda_k [E_k(\mathbf{u}) - E_k(\mathbf{u}^*)]. \quad (33)$$

Since $\lambda$ is a nonnegative vector the second term of the right-hand side of (33) is nonpositive for any $\mathbf{u} \in U$ such that conditions (32) are satisfied. Therefore,

$$J(\mathbf{u}^*) \leq J(\mathbf{u})$$

for all admissible controls $\mathbf{u}$ such that $E_k(\mathbf{u}) \leq E_k(\mathbf{u}^*)$ when $\lambda_k > 0$. Q.E.D.

This result implies that the optimal solution of the original TIP problem is also of the $(\underset{\sim}{S}, \tilde{S})$ type. Moreover, the proposition shows that if we can find an optimal solution to the TIP problem (11) and (12) with some incremental underprovision costs $\hat{a}_4^k = a_4^k + \lambda_k$, $\lambda_k > 0$, and if this solution results in expected demand servicing equal to $100\beta_k$ percent, then it is an optimal solution to the problem described by (11) through (13).

To utilize this result we need to derive formulas for computing the expected demand servicing level for a given policy $\pi = \{(\underset{\sim}{S}(0), \tilde{S}(0)), \cdots, (\underset{\sim}{S}(N-1), \tilde{S}(N-1))\}$. The derivation is given in Appendix C.

Now we can describe an algorithm to solve the TIP problem described by (11) through (13). A numerical procedure for obtaining the TIP solution can be outlined as follows:

*Step 1*—Set the initial vector of the Lagrange multipliers:

$$\lambda = \mathbf{0} \tag{34}$$

and identify the set K of years $k$ for which the demand servicing constraint level would be exceeded unless a positive value were set for $\lambda_k$.

*Step 2*—Using the computational procedure described in Section 3.2, determine

$$\pi = \{(\underset{\sim}{S}(0), \tilde{S}(0)), \cdots, (\underset{\sim}{S}(N-1), \tilde{S}(N-1))\}$$

that minimizes

$$H(\mathbf{u}, \lambda). \tag{35}$$

*Step 3*—Using the Step 2 solution $\pi$, for all years $k$, $k \in$ K, calculate

$$\eta_k(\lambda) = \{E \max(0, d(k) - T(k))\} - \beta_k Ed(k) \tag{36}$$

and determine whether $|\eta_k| < \epsilon_k$, where $\epsilon_k > 0$ is some tolerance level. If for every $k$ the tolerance level $\epsilon_k$ is satisfied, stop; otherwise go to Step 4.

*Step 4*—To reduce $|\eta_k|$, go back to Step 2 with $\lambda$ replaced by

$$\lambda + \omega\eta(\lambda), \tag{37}$$

where $\omega$ is an appropriate positive constant. (The constant $\omega$ can be adjusted by trial and error to speed up the iteration procedure.)

In general, the Lagrange multipliers vary from year to year depending on the demand characteristics, constraint levels, and cost coefficients. Our numerical experience revealed, however, that under reasonable assumptions for TIP application [growing (declining) demand pattern, $\beta_k = \beta$, and $a_i^k = a_i$] the vector of Lagrange multipliers $\lambda^*$ can be approximated by $\lambda' = (\lambda', \cdots, \lambda')$, where the constant $\lambda'$ depends only on a given level of demand servicing $\beta$ and the coefficient of variation of the demand.

Using the approach outlined in Appendix C, we verified that the approximation of $\lambda^*$ by $\lambda'$ does not result in a significant cost penalty, i.e.,

$$\min_{\mathbf{u}} H(\mathbf{u}, \lambda^*) \approx \min_{\mathbf{u}} H(\mathbf{u}, \lambda').$$

Consequently, to obtain a numerically efficient TIP algorithm and to facilitate TIP implementation, we developed a conversion table (as illustrated on Fig. 3) that defines $\lambda'$ as a function of the demand servicing constraint level $\beta$.

## IV. NETWORK TIP

This section extends the only-route TIP to determine a multiyear schedule of trunk augments and disconnects for a hierarchical network that minimizes the present worth of the expected cost of planned and demand servicing subject to capacity and demand servicing constraints.

We note that there is a fundamental difference between the only-route and the hierarchical network TIP problems. The only-route TIP problem answers two basic questions: first, how much extra capacity is needed on a trunk group to hedge against forecast uncertainty and/or to satisfy the constraint on the amount of demand servicing; second,

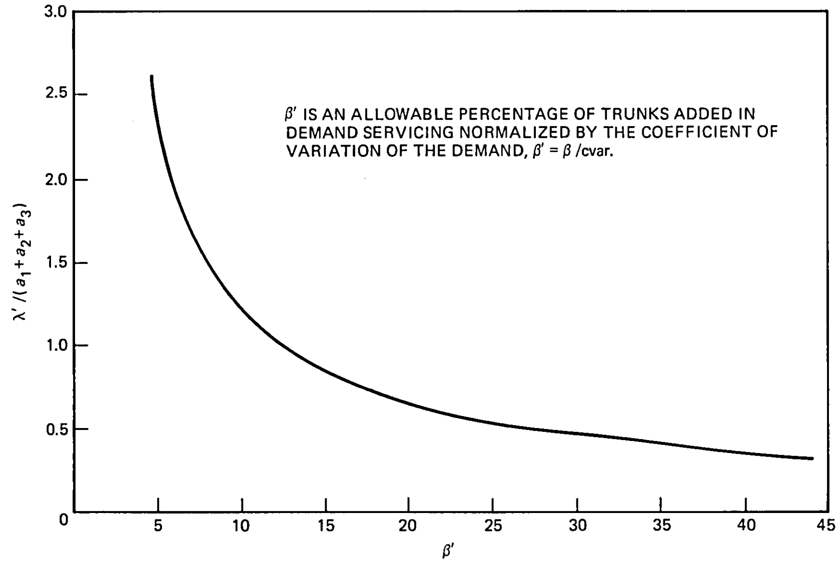


Fig. 3—Approximation of Lagrange multipliers.

how should the year-by-year trunk requirements be smoothed to obtain the optimal balance between the costs of maintaining and rearranging trunks over a planning horizon. In the network case the major additional question is to determine *where* to provide extra capacity in the network, that is, should this additional capacity be provided on all of the trunk groups in the network or on specific trunk groups only?

### 4.1 Overview of network TIP solution

To develop a network TIP algorithm, we exploit the heuristic principles used in conventional trunk group sizing procedures. In particular, we decouple the network TIP problem into individual cluster TIP problems, where a cluster is defined by a final trunk group and all subtending high-usage groups that overflow to that final group (Fig. 2 illustrates trunking terminology). To simplify the analysis we shall assume that the demand servicing policy is to augment only the final group when the blocking objective is violated. Our numerical experience shows that if an unbiased traffic load forecasting algorithm is used (such as in Ref. 3), then our demand servicing policy assumption is not critical to the optimality of the final TIP solution. Consequently, in this section we present a cluster-sizing procedure that minimizes the expected present worth of planned servicing expenditures on each high-usage (HU) trunk group plus the planned and demand servicing expenditures on the corresponding final trunk group subject to facility and demand servicing constraints.

The key idea of our solution is based on a heuristic argument that suggests that a near optimal solution can be obtained by accounting for forecast uncertainty and demand servicing constraints on the final trunk group only. We draw this conclusion from our analysis in Section 4.2, where we consider a single-year TIP problem. Extending this to a multiyear case we then assume that Truitt's engineering procedure, the ECCS rule,[7] can be used to find single-year, initial HU trunk requirements that then will be adjusted to eliminate uneconomical trunk group rearrangements.

Accordingly, in Section 4.3 we derive an optimal disconnect policy for HU trunk groups, that is, we show how to satisfy the ECCS trunk requirements for primary HU trunk groups while minimizing the present worth of the planned servicing cost over a planning horizon. As Fig. 4 shows, when the initial five-year TIP solution on primary HU groups is obtained, the TIP algorithm proceeds by calculating overflow traffic and sizing the intermediate HU trunk groups using, again, the Truitt's engineering procedure and an optimal disconnect policy. After all subtending HU groups have been sized, the final trunk group becomes the last and only choice for the remaining traffic to reach its destination. Consequently, the capacity expansion planning

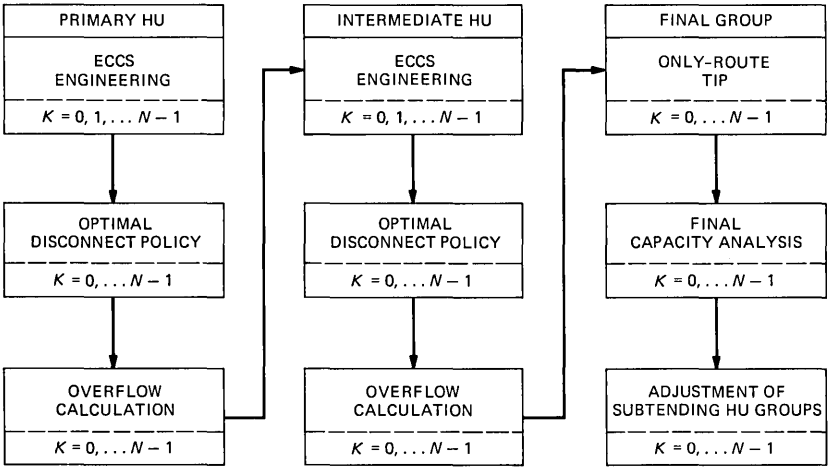


Fig. 4—Network TIP algorithm.

problem for the final trunk group reduces to an only-route TIP problem (11) through (13).

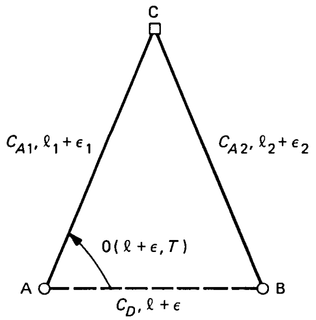
In Section 4.4 we show that under certain circumstances the initial cluster TIP solution may provide less (more) trunk capacity on the final trunk group than that necessary to satisfy the blocking objective and the demand servicing constraint. In that case we show how to improve the initial TIP solution by increasing (reducing) the sizes of the subtending HU groups in an economical fashion.

### *4.2 Alternate routing under uncertainty*

To analyze the impact of forecast error on the optimal design of hierarchical networks we first formulate a single-year TIP problem for a Truitt alternate routing triangle.[7] Referring to Fig. 5, we assume that load $(\mathscr{A} + \epsilon)$ is offered to the direct (HU) group, and background loads $(\mathscr{A}_1 + \epsilon_1)$ and $(\mathscr{A}_2 + \epsilon_2)$ are offered to the first and second legs of the alternate route, respectively, where $\epsilon$, $\epsilon_1$, $\epsilon_2$ are the errors in the load forecast. The trunk group sizes on the direct and alternate routes are denoted by $T$, $T_1$, and $T_2$, respectively. Then the problem of determining $T$, $T_1$, $T_2$ to minimize the expected cost of trunk provisioning activities during the year is given by

$$\min_{T,T_1,T_2} E[C_D T + C_{A1} T_1 + C_{A2} T_2 + C_{S1} \max(0, d_1 - T_1) + C_{S2} \max(0, d_2 - T_2)], \quad (38)$$

where $d_1$ and $d_2$ are the number of trunks required on the alternate route to satisfy the network service objective; $C_D$ is the incremental



| ROUTE | COST | OFFERED LOAD | TRUNKS IN SERVICE | TRUNKS REQUIRED |
|---|---|---|---|---|
| AB | $C_D$ | $\ell + \epsilon$ | $T$ | – |
| AC | $C_{A1}$ | $\ell_1 + \epsilon_1 + O(\ell + \epsilon, T)$ | $T_1$ | $d_1$ |
| CB | $C_{A2}$ | $\ell_2 + \epsilon_2 + O(\ell + \epsilon, T)$ | $T_2$ | $d_2$ |

Fig. 5—Alternate routing under uncertainty.

cost of adding a trunk to the direct route on a planned basis; $C_{A1}$, $C_{A2}$ and $C_{S1}$, $C_{S2}$ are the incremental costs of planned and demand servicing on the alternate route legs. Also, in (38) we assume that when the calculated number of trunks required exceeds the number of trunks in service the demand servicing augmentation is performed on the final group only.

Our goal is to minimize (38) under the demand servicing constraints given by

$$E \max(0, d_1 - T_1) \leq \beta E d_1$$

and

$$E \max(0, d_2 - T_2) \leq \beta E d_2, \tag{39}$$

where the expected value is taken with respect to $\epsilon$, $\epsilon_1$ and $\epsilon$, $\epsilon_1$, $\epsilon_2$, respectively.

In our numerical studies we investigated the cases when 10 to 100 erlangs of traffic are offered to the direct group and 10 to 500 erlangs are offered to the alternate route. We also assumed that the coefficient of variation of the demand forecast on the HU group, $CV_D$, and each leg of the alternate route, $CV_{A1}$ and $CV_{A2}$, vary between 0.0 to 0.25, and that the demand servicing threshold, $\beta$, is in the range from 5 to 30 percent. Finally, we assumed that the forecast errors $\epsilon$, $\epsilon_1$, and $\epsilon_2$ are statistically independent.

### 4.2.1 Major conclusions

We compared the optimal single-year TIP solution with the solution that sizes the direct route by Truitt's formula and then sizes the final to satisfy the blocking objective and demand servicing constraint at minimum cost. Our sensitivity analysis revealed that the optimal trunk requirement on the HU trunk group does not change significantly with changes in the coefficient of variation of the forecast, i.e., the optimal solution accounts for uncertainty by providing extra capacity, mainly on the final trunk group. More importantly, the cost difference between the optimal and the ECCS-based solutions is less than 1 percent. Thus, we conclude that the expected trunk provisioning cost function is very flat in the neighborhood of a solution point and the Truitt's HU solution is relatively close to the optimal HU trunk size.

To exploit the single-year ECCS design as a basis for a five-year trunk plan on HU groups, we next address the question of how to adjust the single-year trunk requirements to obtain an economical trunk implementation schedule for a given planning horizon.

## 4.3 Optimal disconnect policy for high-usage groups

We shall use the notions and notation of Section 2.1, except that $d(k)$ now represents the deterministic (rather than random) ECCS trunk requirement at year $k$.

As explained in Section 3.3, we omit the demand servicing constraint while sizing HU trunk groups. Consequently, for HU trunk groups the objective is to fulfill the ECCS trunk requirements at minimum cost, i.e., to minimize

$$\sum_{k=0}^{N-1} \rho^k[c_1^k(u(k)) + c_3^k(u(k)) + c_3^k(T(k), u(k))] \tag{40}$$

subject to the ECCS constraints

$$T(k) + u(k) \geq d(k),$$

where $T(0)$ is given and $T(k+1)$ is defined by

$$T(k+1) = T(k) + u(k). \tag{41}$$

Under conditions (8) and (9) we will show that the optimal decision, $u^*(k)$, has the following form:

$$u^*(k) = \begin{cases} d(k) - T(k), & \text{if} \quad d(k) \geq T(k) \\ \min\left\{\max\limits_{i=0,\cdots j^*} d(k+i) - T(k), 0\right\}, & \text{if} \quad d(k) < T(k), \end{cases} \tag{42}$$

where $j^*$ is the largest integer $j$ such that

$$b_1^k - b_2^k + \sum_{i=0}^{j^*-1} \rho^i a_3^{k+i} < \rho^{j^*}(a_1^{k+j^*} + a_2^{k+j^*}). \tag{43}$$

Note that in (43) the cost coefficients have superscripts, while only $\rho$ is raised to a power.

Condition (43) states that the present worth of money recovered by disconnecting a trunk module in year $k$ and maintaining $T(k+i) - m$ trunks is less than the present worth of a trunk module purchase and installation in year $k + j^*$ but is greater than this cost in year $k + j^* + 1$.

The second half of the policy (42) dictates that if $T(k)$ or more trunks are required in some year $k, k+1, \cdots, k+j^*$, no trunks are disconnected. Otherwise, trunks are disconnected to the lowest possible level not requiring any reconnections prior to year $k + j^* + 1$.

To demonstrate that $u^*(k) = d(k) - T(k)$ if $d(k) \geq T(k)$, we need to show only that if $u(k) > d(k) - T(k)$, then the corresponding control strategy $\mathbf{u}^1$ and $(u(0), \cdots, u(N-1))$ can be improved by the strategy $\mathbf{u}^2 = (u(0), \cdots, u(k) - m, u(k+1) + m, \cdots, u(N-1))$, where $\mathbf{u}^2$ is necessarily feasible. To show this we consider the two possible scenarios:

$$u(k+1) > 0 \quad \text{and} \quad u(k+1) \leq 0.$$

For $u(k+1) > 0$, the cost difference, $L(\mathbf{u}^1) - L(\mathbf{u}^2)$, of the two strategies $\mathbf{u}^1$, $\mathbf{u}^2$ is

$$L(\mathbf{u}^1) - L(\mathbf{u}^2) = \rho^k m(a_1^k + a_2^k + a_3^k) - \rho^{k+1} m(a_1^{k+1} + a_2^{k+1}).$$

Then from (27) and the positivity of $a_3^k$, $L(\mathbf{u}^1) > L(\mathbf{u}^2)$. Similarly, when $u(k+1) \leq 0$ we obtain

$$L(\mathbf{u}^1) - L(\mathbf{u}^2) = \rho^k m(a_1^k + a_2^k + a_3^k) - \rho^{k+1} m(b_1^{k+1} - b_2^{k+1}), \quad (44)$$

or, from (8) and (9), $L(\mathbf{u}^1) > L(\mathbf{u}^2)$.

To prove the second part of statement (42) we consider two cases. First, let us show that if we have a control strategy $\mathbf{u}^1 = (u(0), \cdots, u(N-1))$ that disconnects fewer trunk modules in year $k$ than suggested by (42), i.e.,

$$\max_{i=0,\cdots,j^*} \{d(k+i)\} - T(k) < u(k) \leq 0, \quad (45)$$

then $\mathbf{u}^1$ can be improved.

First assume that $u(k+j) < 0$ for some $j$ such that $1 \leq j \leq j^*$. Let $k+j$ be the first such year. Then a better feasible policy is given by $\mathbf{u}^2 = (u(0), \cdots, u(k) - m, \cdots, u(k+j), \cdots)$. Indeed, for the difference in planned servicing cost we get

$$L(\mathbf{u}^1) = L(\mathbf{u}^2) = \rho^k m \left(b_1^k - b_2^k + \sum_{i=0}^{j-1} \rho^i a_3^{k+i}\right)$$
$$- \rho^{k+j} m(b_1^{k+j} - b_2^{k+j}), \quad (46)$$

and from (9),

$$L(\mathbf{u}^1) - L(\mathbf{u}^2) > 0.$$

If there is no year $k + j$ $(1 \le j \le j^*)$ for which $u(k + j) < 0$, then from (43) a less expensive feasible solution is presented by $\mathbf{u}^2 = (u(0), \cdots, u(k) - m, \cdots, u(k + j^* + 1) - m, \cdots)$.

Now, we consider the second case and demonstrate that if we disconnect more trunks than specified by (42), i.e., $u^1(k) < u^*(k) = \max\{d(k + 1)\} - T(k)$, then the solution $\mathbf{u}^1$ can be improved by $\mathbf{u}^2 = (u(0), \cdots, u(k) + m, \cdots, u(k + j) - m, \cdots)$, where $j$ is such that $d(k + j) = \max\{d(k + i)\}$. By the first part of (42) we add only as many trunks as needed. Therefore, if $u^1(k) < u^*(k)$ then we can assume that $u^1(k + j) \ge 0$. Consequently, we get

$$L(\mathbf{u}^1) - L(\mathbf{u}^2) = -m\rho^k \left[ b_1^k - b_2^k + \sum_{i=0}^{j-1} \rho^i a_3^{k+i} - \rho^j (a_1^{k+j} + a_2^{k+j}) \right]$$

and from (43) we conclude

$$L(\mathbf{u}^1) - L(\mathbf{u}^2) > 0.$$

The proof is complete.

Using the intuitively appealing solution described by (42) we can construct a simple, numerically efficient scheme that evaluates the optimal policy $\mathbf{u}^*$:

*Step 1*—If $d(k) - T(k)$ is positive, set $u^*(k) = d(k) - T(k)$ and go to Step 3.

*Step 2*—If $d(k) - T(k) \le 0$, find maximum $j$ for which (43) is satisfied; that is, find

$$j^* = \max \left\{ j \,|\, 1 \le j \le N - 1 - k, \quad b_1^k - b_2^k + \sum_{i=0}^{j-1} \rho^i a_3^{k+i} < \rho^j (a_1^{k+j} + a_2^{k+j}) \right\}.$$

Then, compute $d^* = \max_{0 \le i \le j^*} d(k + i)$. If $d^* \ge T(k)$, set $u^*(k) = 0$ and go to Step 3; if $d^* < T(k)$, set $u^*(k) = d^* - T(k)$ and go to Step 3.

*Step 3*—If $k = N - 1$, stop. Otherwise, set $T(k + 1) = T(k) + u^*(k)$, replace $k$ by $k + 1$, and go to Step 1.

If $d(k) > T(k)$, the first step of algorithm simply augments the trunk group up to the ECCS requirement at year $k$. The second step determines how many trunks to disconnect if the current requirement is less than the number of trunks in service.

### *4.4 Final solution for HU trunk groups*

The TIP algorithm described in Section 4.1 constructs a near optimal schedule of trunk augments and disconnects by smoothing the

ECCS high-usage trunk group requirements and by accounting for forecast uncertainty on final trunk groups only. As we stated in Section 3.1, the TIP solution on final groups is defined by $N$ pairs of critical thresholds $(\underset{\sim}{S}^*(k), \tilde{S}^*(k))$, $k = 0, \cdots, N-1$. In practice, it is quite possible that because of the condition

$$T(k) + u(k) \leq \gamma(k),$$

the final group cannot be augmented to satisfy the blocking objective and the demand servicing constraint at year $k$. In that case we show how to adjust the sizes of subtending HU groups to reduce the load on the final.

Thus, in the development to follow we assume that $\underset{\sim}{S}(k)$ found by (25) is greater than the constraint level at year $k$,

$$\underset{\sim}{S}(k) > \gamma(k),$$

and, therefore, from eq. (26)

$$\underset{\sim}{S}^*(k) = \tilde{S}^*(k) = \gamma(k).$$

We note that the lower optimal threshold $\underset{\sim}{S}(k)$ represents the minimum number of trunks required to satisfy the blocking objective and demand servicing constraint. Consequently, the difference $\underset{\sim}{S}(k) - \underset{\sim}{S}^*(k)$ indicates the deficit in final trunks due to the facility constraints. To account for this deficit economically, we formulate the problem of augmenting high-usage groups to relieve the final. First, let us introduce the notation:

$z(k)$ is the deficit in final trunks at year $k$, $z(k) = \underset{\sim}{S}(k) - \underset{\sim}{S}^*(k)$;

$z_j(k)$ is the portion of the deficit (number of final trunks) covered by augmenting subtending high-usage group $j$ at year $k$;

$d_j(k)$ is the trunk requirement on the subtending group $j$ at year $k$;

$\beta_j(k)$ is the additional number of trunks on the subtending group $j$ at year $k$ that compensates for one final trunk;

$\delta_j(k)$ is the maximum reduction in the final trunk requirement that can be obtained by augmenting trunk group $j$ at year $k$;

$M$ is the total number of subtending trunk groups.

Also, we shall use the notation introduced in Section II for the cost functions, $c_1^{kj}(\cdot)$, $c_2^{kj}(\cdot)$, $c_3^{kj}(\cdot)$, controls, $u_j(k)$, and number of trunks in service at the beginning of the year, $T_j(k)$, where the index $j$ identifies the subtending trunk groups.

Given the initial trunk levels for high-usage groups, $T_1(0), \cdots, T_M(0)$, we wish to find a multiyear schedule of trunk augments and disconnects that minimizes the present worth of the planned servicing costs,

$$\min L = \min_{\mathbf{u}_1,\cdots,\mathbf{u}_M} \sum_{k=0}^{N-1} \sum_{j=1}^{M} \rho^k[c_1^{kj}(u_j(k)) + c_2^{kj}(u_j(k)) + c_3^{kj}(T_j(k), u_j(k))] \quad (47)$$

subject to the following conditions:

1. The number of trunks in service on high-usage group $j$ at year $k$ must be greater than the new, inflated trunk requirement, that is,

$$T_j(k) \geq d_j(k) + \beta_j(k)z_j(k). \quad (48)$$

2. The total trunk requirements on the subtending high-usage groups at year $k$ [given by (48)] must be sufficient to cover the deficit in final trunks, that is,

$$\sum_{j=1}^{M} z_j(k) \geq z(k). \quad (49)$$

3. Since we assumed that the demand servicing augmentation is performed on the final group only, the trunk group dynamics equation for the subtending high-usage groups is described by

$$T_j(k + 1) = T_j(k) + u_j(k). \quad (50)$$

4. The unknown variables $z_j(k)$ must satisfy feasibility constraints

$$0 \leq z_j(k) \leq \delta_j(k). \quad (51)$$

To solve the nonlinear optimization problem (47) through (51) note that if all the nonnegative variables $z_j(k)$ are fixed, then the minimization problem can be decomposed as follows:

$$\begin{aligned} \min_{\mathbf{u}_1,\cdots,\mathbf{u}_M} L &= \sum_{j=1}^{M} \min_{\mathbf{u}_j} L_j(z_j(0), \cdots, z_j(N - 1)) \\ &= \sum_{j=1}^{M} \min_{\mathbf{u}_j} \sum_{k=0}^{N-1} \rho^k[c_1^{kj}(u_j(k)) + c_2^{kj}(u_j(k)) + c_3^{kj}(T_j(k), u_j(k))]. \end{aligned}$$

Also, when $z_j(k)$ are fixed, the minimum of the cost functional $L_j$, $L_j^*(z_j(0), \cdots, z_j(N-1))$ can be determined by the algorithm presented in Section 4.3. Therefore, the trunk capacity allocation problem is described by

$$\min_{z_j(k)} \sum_{j=1}^{M} L_j^*(z_j(0), \cdots, z_j(N - 1)), \quad (52)$$

subject to

$$\sum_{j=1}^{M} z_j(k) \geq z(k), \qquad k = 0, \cdots, N - 1,$$

where $z_j(k)$ satisfy feasibility constraints and $L_j^*(\cdot)$ is computed by the algorithm of Section 4.3.

We start by considering the case in which there is only one year, $k = k'$, such that $z(k) > 0$. Then since the unknown trunk quantities, $z_j(k')$, are integers we can use a forward dynamic programming recursion for solving (47), i.e.,

$$f_j(y_j) = \min_{z_j(k')}[L_j^*(z_j(k')) + f_{j-1}(y_j - z_j(k'))], \qquad j = 1, \cdots, M \quad (53)$$

where $f_o(y) = 0$, $0 \le z_j(k') \le y_j$; $y_M = z(k')$, and $L_j^*(z_j(k')) = L_j^*(0, \cdots, z_j(k'), \cdots, 0)$ is calculated by the algorithm of Section 4.3.

Note that modularity constraints on HU groups can be easily incorporated into the discrete dynamic programming formulation (53) to reduce the computational burden. Furthermore, the method outlined by (53) can be used sequentially for each year for which $z(k) > 0$. There is no guarantee, however, that this "one-year-at-a-time" procedure will terminate at a global optimum. Various refinements of the "one-year-at-a-time" method are considered in Ref. 8. In general, these refinements increase a chance to reach an optimum but require significantly more computation.

Finally, we add that the same mathemathical approach can be used to decrease the number of trunks on subtending high-usage groups economically when there is an extra capacity on the final. Recall that $\underset{\sim}{S}^*(k)$ represents the minimum trunk requirement at year $k$ to satisfy the blocking and demand servicing constraints. Consequently, if $T(k) > \underset{\sim}{S}^*(k)$, then the difference between the planned trunk level and the $\underset{\sim}{S}^*(k)$ defines the amount of extra capacity on the final group at year $k$ that can be used to reduce the planned servicing cost on subtending high-usage groups.

## V. FINAL REMARKS

We have described a theoretical development of a new capacity expansion planning process, TIP, that provides a cost-effective multiyear schedule for trunk augments and disconnects for hierarchical networks. In contrast to the existing traffic engineering procedures, our solution accounts for forecast uncertainty, demand dynamics, trunk implementation costs, facility constraints, and demand servicing constraints. As we have shown in Sections III and IV, the dynamic programming approach to the stochastic capacity expansion problem yields a numerically efficient TIP algorithm that is easy to implement.

New, mechanized forecasting systems based on the TIP algorithm have been recommended for implementation in the operating companies and AT&T Communications.

## VI. ACKNOWLEDGMENTS

We would like to acknowledge that C. D. Pack and S. M. Rocklin made substantial contributions on the first stages of the TIP project.

Also, we express our sincere appreciation to J. P. Moreland and W. L. Roach, Jr. Their insightful comments helped us to crystallize our ideas and to improve the quality of the TIP solution.

## REFERENCES

1. W. B. Elsner, "Dimensioning Trunk Groups for Digital Networks," B.S.T.J., *59,* No. 7 (September 1980), pp. 1089–122.
2. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability Engineered Group," B.S.T.J., *55,* No. 7 (September 1976), pp. 831–42.
3. J. P. Moreland, "A Robust Sequential Projection Algorithm for Traffic Load Forecasting," B.S.T.J., *61,* No. 1 (January 1982), pp. 15–38.
4. A. Kashper, C. D. Pack, and G. C. Varvaloucas, "Minimum-Cost Multiyear Trunk Provisioning," Proc. of 10th Int. Teletraffic Cong., Session 1.2, Paper 5, Montreal, 1983.
5. S. M. Rocklin, A. Kashper, and G. C. Varvaloucas, "Capacity Expansion, Contraction of a Facility with Demand Augmentation Dynamics," approved for publication in *Oper. Res.*
6. D. P. Bertsekas, *Dynamic Programming and Stochastic Control,* New York: Academic Press, 1976.
7. C. J. Truitt, "Traffic Engineering for Determining Trunk Requirements in Alternate Routing Networks," B.S.T.J., *33,* No. 2 (March 1954), pp. 277–302.
8. D. J. Wilde, *Optimum Seeking Methods,* Englewood Cliffs, NJ: Prentice-Hall, 1964.
9. R. T. Rockafellar, *Convex Analysis,* Princeton: Princeton University Press, 1970.

## APPENDIX A

### *Proof of ($\underset{\sim}{S}$, $\tilde{S}$) Optimality*

We start our inductive proof by showing that an ($\underset{\sim}{S}$, $\tilde{S}$)-type policy is optimal and by constructing the critical thresholds if $N$ is equal to one, i.e., the last year, $N - 1$, is, in fact, the only year of the planning horizon.

### *A.1 The single-stage problem*

For economy of notation, we shall drop the index, $N - 1$, from our equations. From (15) we wish to find an optimal policy, $u^* = u^*(T)$, that satisfies

$$\underset{d}{E}\{g(d,\, T,\, u^*)\} = \min_u \underset{d}{E}\{g(d,\, T,\, u)\}. \tag{54}$$

Equivalently, from (4) through (7), assuming $T$ trunks in service at the beginning of the year, we seek a planned trunk level $y^*(T) = T + u^*(T)$ that minimizes the single-year cost functional:

$$J(T, y) = \begin{cases} a_1(y - T) + a_2(y - T) + a_3 y & \\ \quad + a_4 E \max(0, d - y), & y \geq T \\ b_1(y - T) - b_2(y - T) + a_3 y & \\ \quad + a_4 E \max(0, d - y), & y \leq T. \end{cases} \tag{55}$$

Note that when $u = 0$, the two branches of (55) are identical.

As we stressed in Section 3.1, to find an optimal solution of the single-stage problem described by (55), we shall show that the first differences of $J_1(T, T + u)$ and $J_2(T, T + u)$ satisfy conditions (1) through (3) of Section 3.1. Thus,

$$\Delta J_1(y) = (a_1 + a_2 + a_3 - a_4)m$$
$$+ a_4E[\max(d, y + m) - \max(d, y)]$$
$$\Delta J_2(y) = (b_1 - b_2 + a_3 - a_4)m$$
$$+ a_4E[\max(d, y + m) - \max(d, y)]. \qquad (56)$$

The second term in (56) represents the expected savings in demand servicing if one additional trunk module is planned. This savings will occur with probability $1 - F(y)$. Thus,

$$\Delta J_1(y) = (a_1 + a_2 + a_3)m - a_4(1 - F(y))m,$$
$$\Delta J_2(y) = (b_1 - b_2 + a_3)m - a_4(1 - F(y))m. \qquad (57)$$

From (57), condition (1) and (2) of Section 3.1 are satisfied. Therefore, we obtain the minimum points of $J_1(T, x)$ and $J_2(T, x)$ for all modular $x$ by applying (25). From (57) $\underset{\sim}{S}$ and $\tilde{S}$ are the smallest values of $x$ on the discrete set M such that

$$F(x) \geq 1 - \frac{a_1 + a_2 + a_3}{a_4}$$

and

$$F(x) \geq 1 - \frac{b_1 - b_2 + a_3}{a_4}. \qquad (58)$$

Then, from inequalities (8) through (10), $\underset{\sim}{S}$ and $\tilde{S}$ satisfy (3). Consequently, the optimal single-stage decision for augmenting or disconnecting trunks in the unconstrained case is given by (3.1) and $\underset{\sim}{S}$ and $\tilde{S}$ are defined by (58). In the presence of capacity constraints, we need to modify $\underset{\sim}{S}$ and $\tilde{S}$ only by (26).

### *A.2 The multistage problem*

In this section, we prove by induction the optimality of an $(\underset{\sim}{S}, \tilde{S})$-type policy for stage $k$ by showing that the two branches of the cost-to-go function $J_k(T(k), y(k))$ satisfy conditions (1) to (3) of Section 3.1.

To simplify the recursion for $J_k(T(k), y(k))$ we introduce the auxiliary function $W_k(y)$:

$$W_k(y) = a_4^k y + \rho V_{k+1}(y). \qquad (59)$$

From the optimality principle and (59), the cost-to-go function from state $T(k)$ can be expressed by

$$J(T(k), y(k)) = \begin{cases} (a_1^k + a_2^k + a_3^k - a_4^k)y(k) - (a_1^k + a_2^k)T(k) \\ \quad + \mathop{E}_{d(k)} W_k(\max(d(k), y(k))), \\ \quad y(k) \geq T(k) \\ (b_1^k - b_2^k + a_3^k - a_4^k)y(k) - (b_1^k - b_2^k)T(k) \\ \quad + \mathop{E}_{d(k)} W_k(\max(d(k), y(k))), \\ \quad y(k) \leq T(k). \end{cases} \tag{60}$$

From (15) and (59), $W_{k-1}(y)$ satisfies the recursion

$$W_{k-1}(y) = \min_{u(k)} \mathop{E}_{d(k)} \{a_4^{k-1}y + \rho[c_1^k(u(k)) + c_2^k(u(k))$$

$$+ a_3^k(y + u(k)) - a_4^k(y + u(k)) + W_k(\max(d(k), y + u(k)))]\}, \tag{61}$$

where $k = N - 1, \cdots, 0$.

As in Section 3.1, we consider the two branches of $J_k(T(k), y(k))$, that is, $J_{k,1}(T(k), y(k))$ and $J_{k,2}(T(k), y(k))$. Then, we need to show that $J_{k,1}$ and $J_{k,2}$ satisfy conditions (1) to (3).

From the definition of $J_{k,1}$ and $J_{k,2}$ and (60), condition (1) is trivial. To demonstrate (2) we have to prove that for any $k$

$$\Delta H_k(x) = \mathop{E}_{d(k)} W_k(\max(d(k), x + m)) - \mathop{E}_{d(k)} W_k(\max(d(k), x))$$

is a nondecreasing function of $x$. First, we consider the case $k = N - 1$.

We shall approach (2) by studying the properties of $W_{N-1}(x)$ and applying standard convexity results. In particular, since $W_{N-1}(x)$ and $\max(d, x)$ are monotonically increasing functions in $x$ with nondecreasing first differences, the composite function $W_{N-1}(\max(d, x))$ must also be an increasing function in $x$ with nondecreasing first differences.[9] In addition, the monotonicity of the functions $W_{N-1}(\max(d, x))$ and $\Delta W_{N-1}(\max(d, x))$ is preserved by the expected value operation. Thus, condition (2) is satisfied for $k = N - 1$.

To prove condition (2) inductively for an arbitrary $k$ we have to show that recursion (61) preserves monotonicity of $W_k(y)$ and $\Delta W_k(y)$.

First, assuming the monotonicity of $W_k(y)$ and $\Delta W_k(y)$ we obtain (in a fashion similar to that for the case $k = N - 1$) that the composite function $W_k[\max(d(k), y + u(k))]$ has the same properties in $y$. Because of linearity in $y$ of the remaining part of the right-hand side of (61) and the cost-of-money assumption, $\rho < 1$, the expression under the expected value sign in (61) is the sum of increasing functions in $y$ with nondecreasing first differences. Thus

$$W_{k-1}(y) = \min_{u(k)} \mathop{E}_{d(k)} f(y, d(k), u(k)), \tag{62}$$

where $f(y, d(k), u(k))$ is increasing in $y$ and $\Delta f(y, d(k), u(k))$ is nondecreasing in $y$.

Second, the monotonicity of $f(\cdot)$ and $\Delta f(\cdot)$ in $y$ implies the monotonicity of the expected value. Thus, from D. Dantizig's convexity preservation result, it follows that the minimum of the expected value of $f(\cdot)$ is also a convex sequence in $y$, that is, the first differences are nondecreasing in $y$. Also, the monotonicity of $f(\cdot)$ is preserved by the minimum (infimum) operation. Indeed, if an arbitrary function $f(y, z)$ is increasing in $y$ for each $z$, then for any $y_1 < y_2$, and $z$, we have

$$\inf_z f(y_1, z) \le f(y_1, z) \le f(y_2, z)$$

and, therefore,

$$\inf_z f(y_1, z) \le \inf_z f(y_2, z).$$

For completion of the proof, we need to demonstrate (3), i.e., that in the case with no capacity constraints, $\gamma(k) = \infty$, the minimum points of $J_{k,1}$ and $J_{k,2}$, $\underset{\sim}{S}(k)$ and $\tilde{S}(k)$, respectively, are finite and satisfy the condition

$$0 \le \underset{\sim}{S}(k) \le \tilde{S}(k) < \infty. \tag{63}$$

Calculating the first differences of $J_{k,1}$ and $J_{k,2}$, we obtain

$$\begin{aligned} \Delta J_{k,1}(y(k)) &= [a_1^k + a_2^k + a_3^k - a_4^k(1 - F_k(y(k))]m \\ &\qquad + \rho \Delta \underset{d(k)}{E}\, V_{k+1}(\max(d(k), y(k))), \\ \Delta J_{k,2}(y(k)) &= [b_1^k - b_2^k + a_3^k - a_4^k(1 - F_k(y(k)))]m \\ &\qquad + \rho \Delta \underset{d(k)}{E}\, V_{k+1}(\max(d(k), y(k))), \end{aligned} \tag{64}$$

where the first differences $\Delta E V_{k+1}$ are taken with respect to $y(k)$. Since $F_k(x) \longrightarrow 1$ as $x \longrightarrow \infty$, it follows from (64) that

$$\begin{aligned} \lim_{x \to \infty} \Delta J_{k,1}(x) &= [a_1^k + a_2^k + a_3^k]m - \rho[b_1^{k+1} - b_2^{k+1}]m, \\ \lim_{x \to \infty} \Delta J_{k,2}(x) &= [b_1^k - b_2^k + a_3^k]m - \rho[b_1^{k+1} - b_2^{k+1}]m. \end{aligned} \tag{65}$$

Relationship (65) shows that for sufficiently large $x$, both expressions are necessarily positive. Therefore, $\underset{\sim}{S}(k)$ and $\tilde{S}(k)$, which denote the smallest elements of the discrete set $M = \{0, m, 2m, \cdots\}$, such that

$$\Delta J_{k,1}(x) \ge 0 \quad \text{and} \quad \Delta J_{k,2}(x) \ge 0 \tag{66}$$

for (64), are finite. In addition, since $a_1^k + a_2^k > b_1^k - b_2^k$, it follows that $\Delta J_{k,1}(x) \geq J_{k,2}(x)$ for all $x$; hence,

$$\underset{\sim}{S}(k) \leq \tilde{S}(k).$$

Thus, if $\gamma(k)$ is equal to infinity, then an $(\underset{\sim}{S}, \tilde{S})$-type policy is optimal for the year $k$. We note that convexity preservation arguments hold for any region on which our sequences are defined. Consequently, in the constrained case, an optimal solution is given by (26).

The proof for the multistage problem is complete.

## APPENDIX B

### *Derivation of the Recursion*

To obtain explicit formulas for calculating $\Delta J_{k,1}$ and $\Delta J_{k,2}$, we shall use several recursions derived in Appendix A. In particular, since the optimal policy for stage $k + 1$ is described by $(\underset{\sim}{S}(k + 1), \tilde{S}(k + 1))$, applying $u^*(k + 1)$ in (60) we can rewrite eq. (15) as

$$V_{k+1}(T(k+1)) = \begin{cases} (a_1^{k+1} + a_2^{k+1} + a_3^{k+1} - a_4^{k+1})\underset{\sim}{S}(k + 1) \\ \quad - (a_1^{k+1} + a_2^{k+1})T(k + 1) \\ \quad + W_{k+1}(\max(d(k + 1), \underset{\sim}{S}(k + 1))) \\ \quad \text{if} \quad T(k + 1) < \underset{\sim}{S}(k + 1) \\ a_3^{k+1}T(k + 1) - a_4^{k+1}T(k + 1) \\ \quad + W_{k+1}(\max(d(k+1), T(k+1))) \\ \quad \text{if} \quad \underset{\sim}{S}(k + 1) \leq T(k + 1) < \tilde{S}(k + 1) \\ (b_1^{k+1} + b_2^{k+1} + a_3^{k+1} - a_4^{k+1})\tilde{S}(k + 1) \\ \quad - (b_1^{k+1} - b_2^{k+1})T(k + 1) \\ \quad + W_{k+1}(\max(d(k + 1), \tilde{S}(k + 1))) \\ \quad \text{if} \quad T(k + 1) \geq \tilde{S}(k + 1), \end{cases} \tag{67}$$

where the expected value is taken with respect to $F_{k+1}$, the demand distribution at year $k + 1$. Note that at the boundary ($\underset{\sim}{S}(k + 1)$ and $\tilde{S}(k + 1)$) the two corresponding branches of (67) are identical. Now, we shall carry (67) one step backward. Thus, we replace $T(k + 1)$ by the trunk group dynamics equation and consider that $T(k)$ [rather than $T(k + 1)$] is fixed. To simplify the notation we replace $T(k) + u(k)$ by $y$. In the rest of the section our objective is to obtain a recursion for

$$\Delta \underset{d(k)}{E} V_{k+1}(\max(d(k), y))$$

$$= \underset{d(k)}{E} V_{k+1}[\max(d(k), y + m)] - \underset{d(k)}{E} V_{k+1}[\max(d(k), y)].$$

From (67) we obtain

$$\Delta \underset{d(k)}{E} V_{k+1}(\max(d(k),y)) = \begin{cases} -(a_1^{k+1}+a_2^{k+1})mF_k(y) \\ \quad \text{if} \quad y<\underset{\sim}{S}(k+1) \\ [a_3^{k+1}-a_4^{k+1}+a_4^{k+1}F_{k+1}(y)]mF_k(y) \\ \quad +\Delta \underset{d(k)}{E} \Big\{\rho \underset{d(k+1)}{E} V_{k+2}\{\max \\ \quad \cdot[d(k+1),\max(d(k),y)]\}\Big\} \\ \quad \text{if} \quad \underset{\sim}{S}(k+1)\leq y\leq \widetilde{S}(k+1) \\ -(b_1^{k+1}-b_2^{k+1})mF_k(y) \\ \quad \text{if} \quad y\geq \underset{\sim}{S}(k+1). \end{cases} \tag{68}$$

First, we note that

$$\Delta V_{k+1}(\max(d(k),\, y)) = \begin{cases} \Delta V_{k+1}(y), & d(k) \leq y \\ 0, & d(k) > y. \end{cases} \tag{69}$$

From (69) it follows that

$$\Delta \underset{d(k)}{E} V_{k+1}(\max(d(k),\, y)) = F_k(y)\Delta V_{k+1}(y). \tag{70}$$

Second, we set $R = \max[d(k),\, d(k+1)]$. Then

$$\Delta V_{k+2}\{\max[d(k+1),\, \max(d(k),\, y)]\}$$

$$= \Delta V_{k+2}\{\max(y,\, R)\} = \begin{cases} \Delta V_{k+2}(y), & R \leq y \\ 0, & R > y. \end{cases} \tag{71}$$

Taking expectation with respect to the demand distribution $F_{k+1}$, we obtain

$$\Delta \underset{d(k+1)}{E} V_{k+2}\{\max(y,\, R)\} = \begin{cases} F_{k+1}(y)\Delta V_{k+2}(y), & d(k) \leq y \\ 0, & d(k) > y. \end{cases} \tag{72}$$

Consequently,

$$\Delta \underset{d(k)}{E} \left\{ \underset{d(k+1)}{E} V_{k+2}\{\max(y,\, R)\} \right\} = F_k(y)F_{k+1}(y)\Delta V_{k+2}(y). \tag{73}$$

Applying (70) and (73), we arrive at

$$\Delta V_{k+1}(y) = \begin{cases} -(a_1^{k+1} + a_2^{k+1})m \quad \text{if} \quad y < \underset{\sim}{S}(k+1) \\ [a_3^{k+1} - a_4^{k+1}(1 - F_{k+1}(y))]m \\ \quad + \rho F_{k+1}(y)\Delta V_{k+2}(y) \\ \quad \text{if} \quad \underset{\sim}{S}(k+1) \leq y < \widetilde{S}(k+1) \\ -(b_1^{k+1} - b_2^{k+1})m \quad \text{if} \quad y \geq \underset{\sim}{S}(k+1). \end{cases} \tag{74}$$

Finally, we note that because of (70), (74) gives the desired recursion:

$$\Delta \underset{d(k)}{E} V_{k+1}(\max(d(k), y))$$

$$= F_k(y) \cdot \begin{cases} -(a_1^{k+1} + a_2^{k+1})m \\ \quad \text{if} \quad y < \underset{\sim}{S}(k+1) \\ [a_3^{k+1} - a_4^{k+1}(1 - F_{k+1}(y))]m \\ \quad + \rho\Delta \underset{d(k+1)}{E} V_{k+2}\{\max[d(k+1), y]\} \\ \quad \text{if} \quad \underset{\sim}{S}(k+1) \le y < \tilde{S}(k+1) \\ -(b_1^{k+1} - b_2^{k+1})m \\ \quad \text{if} \quad y \ge \tilde{S}(k+1). \end{cases}$$

## APPENDIX C

### *Computing Demand Servicing Level*

As we discussed in Section 3.3, to solve the TIP problem described by (11) through (13) we need to learn how to compute the expected level of demand servicing for a given ($\underset{\sim}{S}$, $\tilde{S}$)-type policy $\pi$.

For a given $\pi$, the planned trunk level $y(k) = T(k) + u(k)$ is a random variable that depends only on the previous demands $d(0), \cdots, d(k-1)$ and is independent of the future demands $d(k), \cdots, d(N-1)$. Accordingly, for the expected level of demand servicing corresponding to $\pi$, we get

$$\begin{aligned} &E\{\max(0, d(k) - y(k))\} \\ &\quad = E_{y(k)}\{E\{\max(0, d(k) - y(k)) \mid y(k)\}\} \\ &\quad = \int_0^\infty \left[\int_x^\infty (y - x) dF_k(y)\right] dG_k(x), \end{aligned} \tag{75}$$

where $G_k(x)$ is the distribution function of the planned trunk level $y(k)$.

By the definition of the optimal policy $u^*$, the distribution function $G_k$ of $T(k) + u^*(k)$ is

$$G_k(x) = P(y(k) \le x) = P(T(k) + u^*(k) \le x)$$

$$= \begin{cases} 0, & \text{if} \quad x \underset{\sim}{S} < (k) \\ P(T(k) \le x), & \text{if} \quad \underset{\sim}{S}(k) \le x < \tilde{S}(k) \\ 1, & \text{if} \quad x \ge \tilde{S}(k) \end{cases} \tag{76}$$

for $k = 0, 1, \cdots, N-1$.

Using the fact that, for a given policy $\pi$, the planned trunk level $y(k)$ is independent of the demand at that year, $d(k)$, and our assumption that $d(k-1)$ is independent of $d(0), d(1), \cdots, d(k-2)$, we can derive a simple recursive formula for $G_k(x)$ by calculating $P(T(k) \leq x))$:

$$
\begin{aligned}
H_k(x) &= P(T(k) \leq x) \\
&= P\{\max(d(k-1), y(k-1)) \leq x\} \\
&= P(d(k-1) \leq x) \cdot P(y(k-1) \leq x) \\
&= F_{k-1}(x) \cdot G_{k-1}(x) \\
&= F_{k-1}(x) \cdot \begin{cases} 0, & \text{if } x < \underset{\sim}{S}(k-1) \\ H_{k-1}(x), & \text{if } \underset{\sim}{S}(k-1) \leq x < \tilde{S}(k-1) \\ 1, & \text{if } x \geq \tilde{S}(k-1), \end{cases} \qquad (77)
\end{aligned}
$$

for $k = 1, 2, \cdots, N-1$.

Recalling that, at the beginning of the first year, the number of trunks in service, $T(0)$, is specified, we have

$$
H_0(x) = \begin{cases} 0, & x < T(0) \\ 1, & x \geq T(0). \end{cases} \qquad (78)
$$

Formulas (76) to (78) together with expression (27) specify the forward recursion for calculating the expected demand servicing level at year $k$ ($k = 0, \cdots, N-1$) associated with the policy $\pi$.

Note that using similar independence arguments we can derive forward recursions for calculating other quantities of practical interest such as the total expected cost of trunk provisioning over the planning horizon, the difference in capital cost for two competing ($\underset{\sim}{S}$, $\tilde{S}$)-type policies, or the probability of demand servicing. The last quantity is another important measure of demand servicing activity, since it allows us to predict the portion of only-route groups that will require emergency servicing in a given year.

To calculate the probability of demand servicing, for example, we observe that for a given policy $\pi$ the planned trunk level $y(k)$ depends only on the previous demands $d(0), \cdots, d(k-1)$. Thus we obtain

$$
P\{d(k) > y(k)\} = \int_0^\infty [1 - F_k(x)] dG_k(x),
$$

where $G_k(\cdot)$ is the distribution function of $y(k)$. Integrating by parts and replacing $G_k$ by $H_k$, we arrive at

$$
P(d(k) > y(k)) = \int_{\underset{\sim}{S}(k)}^{\tilde{S}(k)} H_k(x) dF_k(x) + 1 - F_k(\tilde{S}(k)),
$$

$$
k = 0, 1, \cdots, N-1.
$$

## AUTHORS

**Arik N. Kashper,** M.Sc. (Mathematics), 1969, Leningrad University, USSR; Ph.D (Systems Engineering), 1979, University of Arizona, Tucson; AT&T Bell Laboratories, 1979—. At the University of Arizona, Mr. Kashper worked in the area of system identification and parameter estimation. At AT&T Bell Laboratories, he is concerned with various aspects of traffic network planning. Member, SIAM.

**G. C. Varvaloucas,** B.S. (Electrical Engineering), 1972, National Technical University of Athens; M.S., 1975, and Ph.D., 1978 (Systems Engineering), Case Western Reserve University; AT&T Bell Laboratories, 1977—. Mr. Varvaloucas has done work in the general area of telephone network planning. He is presently Supervisor of the Interexchange Architecture Planning Group, with responsibility for defining certain aspects of the interexchange network planning process in the post-divestiture environment. Member, IEEE.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# Analysis of a Demand Assignment TDMA Blocking System

By S. M. BARTA* and M. L. HONIG†

(Manuscript received May 6, 1983)

This paper presents an analysis of a multichannel Time Division Multiple Access (TDMA) blocking system. Such a system is of interest for real-time voice-traffic applications. The effects of different traffic-assignment algorithms, traffic loads, number of channels, number of time slots, and number of traffic nodes on system performance are studied, where performance is measured by the probability that an incoming message will be blocked. An approximate analytical solution is found, the results of which compare exceedingly well with results obtained from computer simulation. Also derived is a rigorous lower bound on the blocking probability. Collectively, these results indicate that, for most systems of interest, blocking probability is insensitive to the assignment algorithm used. The performance of an assignment algorithm that is simplest to implement is therefore nearly optimal.

## I. INTRODUCTION

A multichannel Time Division Multiple Access (TDMA) protocol provides an efficient means of sharing a high-capacity communication channel among a network of users. In a multichannel TDMA system, the aggregate channel capacity is partitioned in both the frequency and time domains. Each of several channels has some fraction of total bandwidth and consists of a series of time slots. A fixed number of channel time slots are combined to form the TDMA frame. There is an extensive literature describing and analyzing this protocol, espe-

*AT&T Information Systems.

†AT&T Bell Laboratories. Now a member of the Central Services Organization of the Regional Bell Operating Companies.

cially in the context of scanning beam communication satellite systems.[1-6]

The sequence of slot-by-slot switching configurations, which describes the origin and destination nodes of the traffic links assigned to each channel and time slot, is called a traffic assignment. Much attention has been focused on the problem of designing efficient traffic-assignment algorithms for the static case, i.e., where the assignment schedule does not change from frame to frame. However, because messages originate at random times and are of random duration, such a static assignment can be wasteful of bandwidth, since a time slot is unused during idle periods.

To overcome the inefficiency of a static assignment, a network controller can allocate channels and time slots to the traffic nodes according to instantaneous traffic needs. In this scheme, the switching configuration may change from frame to frame. This dynamic assignment of channel capacity is called Demand Assignment TDMA (DA/TDMA).

This paper presents an analysis of a multichannel DA/TDMA protocol. We consider a blocking system in which incoming traffic that cannot be immediately serviced is blocked (i.e., turned away). Only one type of incoming traffic is considered. Thus, this model is appropriate for a voice-traffic system.

We compare the blocking probability obtained using an optimal-assignment algorithm, which allows a complete reconfiguration of the switching pattern in each frame, with the blocking probability obtained using a fixed-assignment algorithm, which allows no rearrangement of existing traffic; i.e., in the fixed-assignment case, a message that requires more than one frame for transmission occupies the same channel and time slot in each frame. Notice that both the optimal- and fixed-assignment algorithms are more general than a static reservation scheme in which the switching configuration is the same in each frame. A tight lower bound on blocking probability obtained using an optimal-assignment algorithm is derived in addition to an accurate approximation for the blocking probability resulting from a type of fixed assignment called random assignment. Based upon our analytical results, we conclude that, for systems of moderate size, the blocking probabilities obtained using optimal- and fixed-assignment schemes are nearly identical. This is a significant finding since it implies that the complexity of optimal assignment is usually unnecessary.

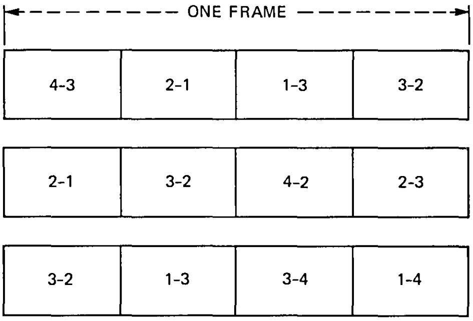
The next section describes the multichannel DA/TDMA protocol and the traffic assignment problem. Section III contains a description of the network mathematical model used for the analysis, and a derivation of the equilibrium-state equations for the associated Markov process. Section IV computes the probability that an incoming-traffic request is blocked, and Section V presents comparisons of analytical with computer simulation results.

## II. THE MULTICHANNEL ASSIGNMENT PROBLEM

This section describes the multichannel traffic-assignment problem. We start with a network consisting of communicating traffic nodes. The channel capacity in this case is partitioned both in the frequency and time domains. In particular, the bandwidth, $B$, of the transmission medium is divided into $m$ channels, each having bandwidth $B/m$, and each channel consists of a series of time slots. A prespecified number of channel time slots are combined to form a transmission frame that continually repeats itself. The reservation multichannel TDMA protocol under consideration assumes a network controller that assigns time slots to incoming-traffic requests on a noninterfering basis.

Figure 1 shows one frame of a multichannel TDMA scheme with three channels and four time slots per frame. Each slot shows a traffic link assigned to that interval. The configuration shown in Fig. 1 will henceforth be referred to as a channel-time slot matrix. Denoting the number of channels by $m$ and the number of time slots by $c$, this matrix will in general have $m$ rows and $c$ columns, and each entry will be a two-dimensional vector consisting of the transmitting and receiving nodes. Notice that this multichannel technique assumes that each node can transmit and/or receive on any (single) channel during a given time slot, and that the channel on which a node is transmitting or receiving can change over successive time slots. Furthermore, we also assume that the assigned channel and time slot for a given traffic link may change from frame to frame. In particular, the traffic link 1-2 shown in Fig. 1 may be reassigned from channel 1/time slot 2 to some other channel and time slot in subsequent frames.

The assignment of incoming traffic to available time slots is gov-



| | | | |
|---|---|---|---|
| 4-3 | 2-1 | 1-3 | 3-2 |
| 2-1 | 3-2 | 4-2 | 2-3 |
| 3-2 | 1-3 | 3-4 | 1-4 |

Fig. 1—One three-channel time division multiple access frame.

erned by the following *fundamental constraint*: one node may not transmit or receive on two different channels during the same time slot. In contrast to single-channel TDMA, in the multichannel case the network controller must not only determine whether a slot is being used, but must also determine whether a given traffic request can be assigned to that slot without violating the fundamental constraint. It therefore may not be possible to assign a given traffic request even though empty time slots exist.

Given unlimited computational capability, it may be desirable to rearrange traffic already assigned to the channel-time slot matrix in order to accommodate a new traffic arrival. It is therefore of interest to know under what conditions a set of traffic requests can be assigned. Denoting the number of traffic nodes by $n$, we define the $n \times n$ traffic matrix $\mathbf{T}$ as the matrix whose $(i, j)$th element contains the number of traffic-units node $i$ is transmitting to node $j$ (each traffic unit represents one packet and is assigned to one time slot). Given a traffic matrix, $\mathbf{T}$, and an empty channel-time slot matrix, all of the traffic can be assigned without violating the fundamental constraint if and only if the following *matrix constraints* are satisfied:[7]

$$\mathbf{T}\underline{1} \leqslant c\underline{1}, \tag{1a}$$

$$\underline{1}'\mathbf{T} \leqslant c\underline{1}' \tag{1b}$$

and

$$\underline{1}'\mathbf{T}\underline{1} \leqslant mc, \tag{1c}$$

where $\underline{1}$ is an $n$-dimensional vector whose elements are all unity, $m$ denotes the number of channels, $c$ denotes the number of time slots, and prime denotes transpose. These equations imply that no transmitting node requires more than $c$ time slots, no receiving node requires more than $c$ time slots, and that the total traffic demand is less than the total number ($mc$) of available time slots. To check whether it is possible to assign a new traffic request, one therefore need only check the matrix constraint set (1) where $\mathbf{T}$ contains traffic already assigned in addition to the new traffic request.

Because traffic requests are time varying, of ultimate interest is how to assign incoming traffic dynamically so as to minimize the number of times a new traffic request cannot be assigned without rearranging assigned traffic. This problem is quite difficult and has not yet been addressed in the open literature. The difficulty arises from the fact that future traffic requests are unknown, and hence, given any assignment rule, it is always possible to receive a sequence of traffic requests such that one can be assigned only if existing traffic is rearranged. Given a matrix $\mathbf{T}$ such that the constraint set (1) is satisfied, however, the static problem of efficiently assigning all of the traffic in $\mathbf{T}$ to an empty channel-time slot matrix has been addressed in Refs. 6 and 8 through 10. We also point out that these methods do not result in a unique assignment.

If we assume that a total rearrangement of existing traffic is allowed at the time each traffic request is made, then an optimal traffic assignment scheme can be inferred from the matrix constraint set (1). ("Optimal" in this case means that the probability of not being able to assign a new traffic request is minimized.) In particular, each time a new traffic request is made, the traffic matrix constraint set (1) is checked. If they are satisfied, then a "brute force" method for assigning the new traffic request would be to "empty out" the existing channel-time slot matrix and reassign all of this traffic along with the new traffic request via one of the methods suggested in Refs. 6 and 8 through 10. Certainly, this scheme requires much more computational power than necessary. If a new traffic request cannot be assigned to existing empty time slots, it is likely that very few (i.e., one or two) existing traffic assignments would have to be rearranged in order to assign the new traffic request.

If traffic assignments are to be made real time, as in a satellite system, the complexity of the assignment scheme becomes a crucial issue. The brute force optimal assignment scheme previously described would optimize system performance; however, under moderate to heavy loads, this scheme is likely to be impractical so that simpler assignment schemes,. which yield suboptimal performance (i.e., a higher blocking probability), must be used. This raises the question of how much performance degradation is caused by using simpler assignment schemes rather than an optimal assignment scheme that permits an unlimited number of rearrangements.

The approach taken in this paper is to compare analytically the blocking probability obtained for a given system using an optimal assignment algorithm with the blocking probability obtained using fixed-assignment algorithms, which allow no rearrangement. Under a fixed-assignment algorithm, if a new traffic request cannot be assigned to a given channel-time slot matrix, it is blocked. Notice that to determine whether incoming traffic can be assigned when using an optimal-assignment algorithm, the traffic matrix must be examined, whereas when using a fixed-assignment algorithm, the channel-time slot matrix must be examined.

## III. MATHEMATICAL FORMULATION

### *3.1 Traffic model*

The purpose of this subsection is to specify the mathematical model used to generate the analytical results in the following sections. The incoming traffic is modeled as the sum of independent Poisson processes flowing between each pair of traffic nodes. We therefore have an arrival rate matrix, $\Lambda$, whose $(i, j)$th element is the Poisson rate at which messages are transmitted from node $i$ to node $j$. The flow rates for traffic between each pair of nodes are assumed to be identical, i.e., $\Lambda = \lambda' \underline{1}\underline{1}'$ where $\lambda'$ is a constant. The total traffic arrival rate, $\lambda$, is therefore the sum of the rates between each pair of nodes.

Because the traffic out of each node is assumed to be independent of the traffic out of all other nodes, given a new traffic request, the probability that the request originated from a specific node $i$ is $1/n$ and, similarly, the probability that the destination is node $j$ is also $1/n$. Because the transmission processes between each pair of nodes are independent, the probability that a given traffic request originates from node $i$ *and* is sent to node $j$ is $1/n^2$.

If a traffic request can be assigned, it occupies one slot of the channel-time slot matrix for a random amount of time, depending on the message length. This "service" time is assumed to be exponentially distributed. Associated with the incoming traffic is therefore the exponential service rate, $\mu$. For analytical convenience, we assume that this distribution is continuous in the sense that departures from the channel-time slot matrix can occur at any time instant. Notice, however, that for a real system, departures occur only at the end of a given frame. The corresponding service time distributions must therefore be "discretized" since service times must be an integer numbers of frames. This effect becomes negligible, however, if the average service time consists of a large number of frames (i.e., >100).

### 3.2 Derivation of equilibrium-state equations for optimal assignment

The derivations of the state equations for the Markov processes associated with optimal and fixed assignments are virtually identical. We present, therefore, only the details of the derivation for optimal assignment and then indicate how to modify that derivation to obtain the state equations for fixed assignment.

When an optimal assignment algorithm is used, the traffic matrix **T** determines whether an incoming traffic request is blocked. Because the traffic arrival process is Poisson and the service times are exponential, the evolution of the traffic matrix **T** is described by a Markov process. The set of states for the Markov process associated with optimal assignment is, therefore, the set of $n \times n$ matrices **T** with integer elements that satisfy the constraint set (1). The following notation is helpful in defining the transition rates for the process:

$$S \equiv \{\mathbf{T} \,|\, \mathbf{T} \text{ satisfies (1a) through (1c)}\} \tag{2a}$$

$$S_k \equiv \{\mathbf{T} \,|\, \mathbf{T} \in S,\ \underline{1}'\mathbf{T}\underline{1} = k\} \tag{2b}$$

$$S_{\mathbf{T}}^{-} \equiv \{\mathbf{T}^* \,|\, \mathbf{T}^* \in S,\ \exists(i, j) \text{ such that } \mathbf{T}^* = \mathbf{T} - \underline{e}_i\underline{e}_j'\} \tag{2c}$$

$$S^+_{\mathbf{T}} \equiv \{\mathbf{T}^* \,|\, \mathbf{T}^* \in S,\ \exists(i, j) \text{ such that } \mathbf{T}^* = \mathbf{T} + \underline{e}_i\underline{e}_j'\}, \tag{2d}$$

where $\underline{e}_i$ is the $n \times 1$ vector with 1 in the $i$th place and 0 elsewhere. The set $S_k$ is the set of all states such that there are $k$ messages in the system; $S^-_{\mathbf{T}}$ and $S^+_{\mathbf{T}}$ are the sets of states into which the process makes transitions via departures and arrivals, respectively, given that the process is in state $\mathbf{T}$. Notice that if $\mathbf{T}^* \in S^-_{\mathbf{T}}$, then $\mathbf{T} \in S^+_{\mathbf{T}^*}$ and, similarly, if $\mathbf{T}^* \in S^+_{\mathbf{T}}$, then $\mathbf{T} \in S^-_{\mathbf{T}^*}$.

Let $r(\mathbf{T}^1, \mathbf{T}^2)$ be the transition rate[†] between any two states $\mathbf{T}^1$ and $\mathbf{T}^2$. Notice that a transition from state $\mathbf{T}^1$ to a state $\mathbf{T}^2$ can occur only via a single arrival or a single departure. The rate $r(\mathbf{T}^1, \mathbf{T}^2)$ is, therefore, nonzero if and only if $\mathbf{T}^1 = \mathbf{T}^2$, or $\mathbf{T}^1$ and $\mathbf{T}^2$ differ from each other in exactly one element and that difference is one. Thus,

$$r(\mathbf{T}^1, \mathbf{T}^2) = \begin{cases} t^1_{ij}\mu, & \text{if } \ \mathbf{T}^2 \in S^-_{\mathbf{T}^1} \\ \lambda/n^2, & \text{if } \ \mathbf{T}^2 \in S^+_{\mathbf{T}^1} \\ 1 - \sum\limits_{i,j} t^1_{ij}\mu - \dfrac{\lambda}{n^2}\,|S^+_{\mathbf{T}^1}|, & \text{if } \ \mathbf{T}^1 = \mathbf{T}^2 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{T}^1$ and $\mathbf{T}^2$ differ in the $(i, j)$th element, $t^1_{ij}$ is the $(i, j)$th element of $\mathbf{T}^1$, $1/\mu$ is the mean service time for messages, $\lambda/n^2$ is the expected rate at which messages between nodes $i$ and $j$ are generated, and $|A|$ denotes the number of elements in the set $A$.

Because the performance of a TDMA assignment algorithm is measured in terms of steady-state network behavior, we focus on the limiting steady-state distribution of the state variable $\mathbf{T}$. To prove that the steady-state distribution exists, note that the state space consists of matrices with integer elements that satisfy the constraint set (1), and it is therefore finite. The transition rates given in (3) do not depend explicitly on time, so the process is time homogeneous. Finally, it can be shown the process is irreducible and contains no periodicities. Therefore, the steady-state distribution exists. Although we obtain a closed-form expression for the steady-state distribution in Appendix A, it can be explicitly computed only in very few cases. We therefore develop an alternative method of analyzing the system, which does not require the explicit formula. This approach, moreover, provides valuable insight into our problem and allows us to provide a unified treatment of both optimal- and fixed-assignment algorithms.

The next step is to derive the equilibrium equations that the steady-state probability distribution $p(\mathbf{T})$ must satisfy. Using the notation introduced earlier, we have the flow equation[11]

[†] The function $r(\cdot\,, \cdot)$ is technically the infinitesimal generator of the Markov process associated with optimal assignment.

$$\sum_{\substack{\mathbf{T}'\in S\\ \mathbf{T}'\neq T}} p(\mathbf{T})r(\mathbf{T}, \mathbf{T}') = \sum_{\substack{\mathbf{T}'\in S\\ \mathbf{T}'\neq \mathbf{T}}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}), \qquad \mathbf{T} \in S. \tag{4}$$

The intuitive meaning of (4) is that for every state **T**, the probability flow rate out of **T** must equal the probability flow rate into **T**.

The following lemma may be used to derive an alternative form for (4). Let $|A|$ represent the number of elements in the set A and let $p(f\,|\,\mathbf{T})$ denote the probability, conditioned on state **T**, that an arrival, chosen from the uniform distribution of origin-destination pairs, does not violate the set of matrix constraints; i.e., $p(f\,|\,\mathbf{T})$ is the probability that an arrival "fits" (can be assigned). Then we have

*Lemma 1: The probability flow out of state* $\mathbf{T} \in S$ *satisfies*

$$\sum_{\substack{\mathbf{T}'\neq \mathbf{T}\\ \mathbf{T}'\in S}} p(\mathbf{T})r(\mathbf{T}, \mathbf{T}') = \sum_{\mathbf{T}'\in S_{\mathbf{T}}^{-}} p(\mathbf{T})r(\mathbf{T}, \mathbf{T}') + \sum_{\mathbf{T}'\in S_{\mathbf{T}}^{+}} p(\mathbf{T})r(\mathbf{T}, \mathbf{T}') \tag{5a}$$

$$= \mu k p(\mathbf{T}) + \lambda p(f\,|\,\mathbf{T})p(\mathbf{T}), \tag{5b}$$

*where* $\underline{1}'\mathbf{T}\underline{1} = k$, *which equals the number of occupied time slots, and*

$$p(f\,|\,\mathbf{T}) = \frac{|S_{\mathbf{T}}^{+}|}{n^2}. \tag{6}$$

The proof of this lemma appears in Appendix B.

This lemma shows that the flow out of any state **T** is composed of two terms: the first term in (5b), $\mu k p(\mathbf{T})$, reflects transitions that occur because there is a departure, while the second term, $\lambda p(f\,|\,\mathbf{T})p(\mathbf{T})$, reflects transitions that occur because an admissible request is generated.

Applying Lemma 1 to the left side of (4) and noting that, if $\mathbf{T}' \neq \mathbf{T}$, $r(\mathbf{T}', \mathbf{T}) \neq 0$ if and only if $\mathbf{T}' \in S_{\mathbf{T}}^{-}$ or $\mathbf{T}' \in S_{\mathbf{T}}^{+}$, we obtain the flow equation

$$\mu k p(\mathbf{T}) + \lambda p(f\,|\,\mathbf{T})p(\mathbf{T}) = \sum_{\mathbf{T}'\in S_{\mathbf{T}}^{-}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}) + \sum_{\mathbf{T}'\in S_{\mathbf{T}}^{+}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}), \qquad \mathbf{T} \in S_k. \tag{7}$$

Equation (7) describes the probabilistic flows into and out of each state. In principle, (7) can be solved to yield a closed form solution for the steady-state distribution of the state variable, $p(\mathbf{T})$ (see Appendix A). Unfortunately, it is very difficult to evaluate this expression for most cases of interest. In order to calculate the system blocking probability, $P_B$, however, it is unnecessary to evaluate the steady-state distribution of **T**. It is shown in the next section that it is enough to know only certain aggregate quantities depending only on the number of messages in the system, $k$. Notice that the "state" $k$ is an aggregate state composed of all traffic matrices $\mathbf{T} \in S_k$. We therefore are interested in deriving an analogous equation to (7) that describes the steady-state flows between the sets of states $S_k$, $k = 0, 1, 2, \cdots$. A direct technique for obtaining this equation is to sum both sides of (7) over the set $\mathbf{T} \in S_k$.

Let $p_O(k)$ be the steady-state probability that the state of the system is contained in $S_k$, which is the probability that there are $k$ occupied time slots; and let $p_O(f \mid k)$ be, in analogy with $p(f \mid \mathbf{T})$, the probability of a fit given that the system contains $k$ messages, i.e.,

$$p_O(k) \equiv \sum_{\mathbf{T} \in S_k} p(\mathbf{T}) \tag{8a}$$

$$p_O(f \mid k) \equiv \frac{\sum_{\mathbf{T} \in S_k} p(f \mid \mathbf{T}) p(\mathbf{T})}{p_O(k)}, \tag{8b}$$

where the $O$ subscript indicates that an optimal assignment algorithm is assumed. The following theorem gives the flow equations for the distribution $p_O(k)$.

*Theorem 1: The probabilities* $\{p_O(k)\}$ *satisfy the equations*

$$[\mu k + \lambda p_O(f \mid k)] p_O(k) = \mu(k+1) p_O(k+1) + \lambda p_O(f \mid k-1) p_O(k-1), \tag{9}$$

*where* $p_O(k) = 0$ for $k < 0$ or $k > mc$.

The proof of this theorem is given in Appendix C.

Equation (9) expresses the equality of probability flows between the sets $S_k$. Given that the system is in state $\mathbf{T}$, the next transition can only be into a state $\mathbf{T}' \in S_{\mathbf{T}}^-$ or $\mathbf{T}' \in S_{\mathbf{T}}^+$. Suppose, for example, that $\mathbf{T} \in S_k$; then $\mathbf{T}' \in S_{k-1}$ or $\mathbf{T}' \in S_{k+1}$; i.e., transitions out of any state in $S_k$ are always into a state in $S_{k-1}$ or $S_{k+1}$. Note that (9) is a set of "birth-death" equations, which we will solve in Section IV to give an expression for $\{p_O(k)\}$ in terms of $\lambda$, $\mu$, and $\{p_O(f \mid k)\}$. (These equations do not, however, imply that the $k$ process is Markov.) In contrast to the solution of (7) we show in Section IV that it is possible to obtain a close approximation to the solution of (9) that can be easily evaluated.

### 3.3 Equilibrium state equations for fixed assignment

To derive the flow equations [corresponding to (5), (7), and (9)] for the Markov process defined by a fixed-assignment algorithm, we redefine the set of states as the set of channel-time slot matrices $\mathbf{M}$ defined in Section II. The following notation is analogous to (2):

$$L \equiv \{\mathbf{M} \mid \mathbf{M} \text{ is an admissible channel-time slot matrix}\} \tag{10a}$$

$$L_k \equiv \{\mathbf{M} \mid \mathbf{M} \in L, \mathbf{M} \text{ contains } k \text{ units of traffic}\} \tag{10b}$$

$$L_{\mathbf{M}}^{-} \equiv \{\mathbf{M}^* \,|\, \mathbf{M}^* \in L,\ \exists(i-j) \text{ and } (m^*, c^*) \text{ such that } \mathbf{M}^* = \mathbf{M} - (i-j)\underline{e}_{m^*}\underline{e}_{c^*}'\} \tag{10c}$$

$$L_{\mathbf{M}}^{+} \equiv \{\mathbf{M}^* \,|\, \mathbf{M}^* \in L,\ \exists(i-j) \text{ and } (m^*, c^*) \text{ such that } \mathbf{M}^* = \mathbf{M} + (i-j)\underline{e}_{m^*}\underline{e}_{c^*}'\}, \tag{10d}$$

where $(i - j)$ represents one unit of traffic from node $i$ to node $j$, $\underline{e}_{m^*}$ is the $m \times 1$ vector with 1 in the $m^*$th place and 0 elsewhere, $e_{c^*}$ is the $c \times 1$ vector with 1 in the $c^*$th place and 0 elsewhere, and the notation $\mathbf{M}^* = \mathbf{M} + (i-j)\underline{e}_{m^*}\underline{e}_{c^*}'$ means that the channel-time slot matrices $\mathbf{M}^*$ and $\mathbf{M}$ differ from each other only in the $(m^*, c^*)$ position. A + sign indicates that $\mathbf{M}^*$ contains the traffic pair $(i - j)$ in the $(m^*, c^*)$ position, whereas the $(m^*, c^*)$ position in $\mathbf{M}$ is empty; the − sign indicates that $\mathbf{M}$ contains the traffic pair $(i - j)$ in the $(m^*, c^*)$ position, whereas the $(m^*, c^*)$ position in $\mathbf{M}^*$ is empty.

Let $r(\mathbf{M}^1, \mathbf{M}^2)$ be the transition rate between any two states $\mathbf{M}^1$ and $\mathbf{M}^2$.[†] Then we have

$$\mathrm{r}(\mathbf{M}^1, \mathbf{M}^2) = \begin{cases} \mu, & \text{if } \mathbf{M}^2 \in L_{\mathbf{M}^1}^{-} \\ \lambda(\mathbf{M}^1, \mathbf{M}^2), & \text{if } \mathbf{M}^2 \in L_{\mathbf{M}^1}^{+} \\ 1 - |\mathbf{M}^1|\mu - \sum\limits_{\mathbf{M}^2 \in L_{\mathbf{M}^1}^{+}} \lambda(\mathbf{M}^1, \mathbf{M}^2), & \text{if } \mathbf{M}^1 = \mathbf{M}^2 \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where $|\mathbf{M}^1|$ is the total number of traffic pairs in $\mathbf{M}^1$. The transition rate $\lambda(\mathbf{M}^1, \mathbf{M}^2)$ is the rate at which a new arrival $(i - j)$ is assigned to $\mathbf{M}^1$, thereby changing $\mathbf{M}^1$ to $\mathbf{M}^2$. Let $a_{\mathbf{M}^1}^{(i-j)}$ denote the number of slots in $\mathbf{M}^1$ into which the arrival $(i - j)$ can be assigned. If assignments are made by selecting one out of the $a_{\mathbf{M}^1}^{(i-j)}$ available slots from a uniform distribution, then

$$\lambda(\mathbf{M}^1, \mathbf{M}^2) = \frac{\lambda}{n^2 a_{\mathbf{M}^1}^{(i-j)}},$$

where $\lambda$ is the total traffic-arrival rate. This fixed-assignment scheme, where traffic arrivals are randomly assigned to available time slots according to a uniform distribution, will henceforth be referred to as random assignment. It is not necessary, however, to assume random assignment in order to derive the flow equation that follows.

We denote the steady-state probability distribution of the Markov process for fixed assignment as $p(\mathbf{M})$. The existence proof for $p(\mathbf{M})$ is the same as that for $p(\mathbf{T})$. Also, let $p(f\,|\,\mathbf{M})$ denote the conditional probability, given that the state is $\mathbf{M}$, that an arrival chosen from the

[†] The function $r(\cdot\,, \cdot)$ is the infinitesimal generator for the associated fixed assignment process.

uniform distribution of origin-destination pairs can be assigned without reassigning any existing traffic, i.e., the probability of a fit under the fixed-assignment rule. Finally, we define

$$A_{\mathbf{M}} \equiv \{(i-j) \mid a_{\mathbf{M}}^{(i-j)} > 0\} \tag{12}$$

as the set of all traffic pairs which can be assigned to **M**.

Proceeding as in the case of optimal assignment, we can derive the flow equations for the fixed-assignment case by merely changing notation. The probability of a fit, $p(f \mid \mathbf{M})$, in this case is simply the number of traffic pair arrivals that can be assigned under the fixed assignment rule divided by the total number of possible pairs. Thus, in analogy with (6) we have

$$p(f \mid \mathbf{M}) = \frac{|\mathrm{A}_{\mathbf{M}}|}{n^2}, \tag{13}$$

where $A_{\mathbf{M}}$ is defined in (12). In analogy with the flow equations (7) and (9) for optimal assignment, the fixed-assignment flow equations are

$$\mu k p(\mathbf{M}) + \lambda p(f \mid \mathbf{M}) p(\mathbf{M}) = \sum_{\mathbf{M}' \in L_{\mathbf{M}}^{-}} p(\mathbf{M}') r(\mathbf{M}', \mathbf{M}) + \sum_{\mathbf{M}' \in L_{\mathbf{M}}^{+}} p(\mathbf{M}') r(\mathbf{M}', \mathbf{M}), \qquad \mathbf{M} \in L_k, \quad k = 1, \cdots, mc, \tag{14}$$

and

$$[\mu k + \lambda p_F(f \mid k)] p_F(k) = \mu(k+1) p_F(k+1) + \lambda p_F(f \mid k-1) p_F(k-1), \qquad k = 1, \cdots, mc, \tag{15}$$

where

$$p_F(k) \equiv \sum_{\mathbf{M} \in L_k} p(\mathbf{M}) \tag{16a}$$

$$p_F(f \mid k) \equiv \frac{\sum_{\mathbf{M} \in L_k} p(f \mid \mathbf{M}) p(\mathbf{M})}{p_F(k)}, \tag{16b}$$

and the $F$ subscript indicates that a fixed-assignment algorithm is assumed. Note that (15) is, like (9), a system of birth-death equations.

## IV. PERFORMANCE ANALYSIS

### *4.1 Properties of blocking probabilities*

The performance of a multichannel TDMA blocking system is measured in terms of the steady-state probability that an arrival is blocked or lost. Of course, this blocking probability depends on the assignment algorithm. Conditioned on the state of the system, the blocking probability is simply $1 - p(f\,|\,\mathbf{T})$ or $1 - p(f\,|\,\mathbf{M})$, for optimal or fixed assignment, respectively. The unconditional blocking probability is given by

$$1 - P_B^O = \sum_{\mathbf{T}\in S} p(f\,|\,\mathbf{T})p(\mathbf{T}) \tag{17}$$

or

$$1 - P_B^F = \sum_{\mathbf{M}\in L} p(f\,|\,\mathbf{M})p(\mathbf{M}), \tag{18}$$

where $P_B^O$ and $P_B^F$ are blocking probabilities using respectively optimal- and fixed-assignment algorithms. As discussed in Section III and Appendix A, it is extremely difficult to compute $p(\mathbf{T})$ for most systems of interest, and hence (17) and (18) cannot be used directly to compute $P_B^O$ and $P_B^F$. In this section, however, we derive bounds and approximations for (17) and (18) by using the aggregate-state equations (9) and (15). This provides a means of quantitively comparing assignment algorithms and estimating the performance of the system as parameters such as $n$, $m$, and $c$ vary.

We obtain equivalent expressions for the blocking probabilities as follows:

$$\begin{aligned} 1 - P_B^O &= \sum_{\mathbf{T}\in S} p(f\,|\,\mathbf{T})p(\mathbf{T}) \\ &= \sum_{k=1}^{mc} \sum_{\mathbf{T}\in S_k} p(f\,|\,\mathbf{T})p(\mathbf{T}) \\ &= \sum_{k=1}^{mc} p_O(f\,|\,k)p_O(k), \end{aligned} \tag{19}$$

where $p_O(f\,|\,k)$ and $p_O(k)$ satisfy (8); similarly $P_B^F$ satisfies

$$1 - P_B^F = \sum_{k=1}^{mc} p_F(f\,|\,k)p_F(k), \tag{20}$$

where $p_F(f\,|\,k)$ and $p_F(f)$ satisfy (16).

Equations (19) and (20) are particular cases of the following general expression for blocking probability;

$$1 - P_B = \sum_{k=1}^{mc} \phi_k p(k), \tag{21}$$

where $\phi_k$ and $p(k)$ satisfy the birth-death system of equations

$$[\mu k + \lambda\phi_k]p(k) = \mu(k+1)p(k+1) + \lambda\phi_{k-1}p(k-1), \tag{22}$$

subject to the boundary conditions $\phi_{mc} = 0$ and $p(k) = 0$ for $k < 0$ or $k > mc$; and $\Sigma_k\, p(k) = 1$. Solving (22) and substituting into (21) yields an expression for the blocking probability in terms of the $\phi_k$'s; i.e.,

$$p(k) = \frac{\frac{1}{k!} \rho^k \prod_{j=0}^{k-1} \phi_j}{\sum_{i=0}^{mc} \frac{1}{i!} \rho^i \prod_{j=0}^{i-1} \phi_j}, \tag{23}$$

and from (21)

$$1 - P_B = \frac{\sum_{k=0}^{mc-1} \frac{1}{k!} \rho^k \prod_{j=0}^{k} \phi_j}{\sum_{i=0}^{mc} \frac{1}{i!} \rho^i \prod_{j=0}^{i-1} \phi_j}, \tag{24}$$

where $\rho \equiv \lambda/\mu$. If $\phi_k = p_O(f \mid k)$ or $\phi_k = p_F(f \mid k)$, then $P_B = P_B^O$ or $P_B = P_B^F$, respectively. Another case of particular interest is where $\phi_k = 1$ if $0 \leqslant k < mc$ and $\phi_k = 0$ otherwise. We denote this set of probabilities, which corresponds to the $mc$ server Erlang-loss system, as $p_E(f \mid k)$. The expression for blocking probability, $P_B^E$, is called Erlang's loss or B formula. The Erlang formula applies to a single-channel TDMA system where only the constraint (1c) in (1) must hold. An important property of $P_B$ stated in the following lemma, which may be used to derive bounds on assignment algorithm performance, is that $P_B$ is a monotonically nonincreasing function of $\phi_k$ for $k = 0, 1, \cdots, mc$.

*Lemma 2: The blocking probability $P_B \equiv P_B(\phi_0, \cdots, \phi_{mc})$ satisfies*

$$\frac{\partial P_B}{\partial \phi_k} \leqslant 0, \qquad k = 0, \cdots, mc. \tag{25}$$

The proof of this lemma appears in Appendix D.

It is intuitively reasonable that the blocking probability should increase as more system constraints are added. This is indeed the case, as the next theorem shows.

*Theorem 2: The optimal assignment, fixed assignment, and Erlang blocking probabilities satisfy*

$$P_B^E \leqslant P_B^O \leqslant P_B^F. \tag{26}$$

*Proof:* If the system is in state $k$, then an arrival can be assigned by a fixed-assignment algorithm only if it can be assigned by an optimal algorithm and so, for all $k$, we have $p_F(f \mid k) \leqslant p_O(f \mid k)$. It is clear that $p_O(f \mid k) \leqslant p_E(f \mid k)$, for all $k$. Therefore, (25) implies (26). Q.E.D.

As the ratio of the number of nodes $n$ to the number of channels $m$ increases, it becomes less likely that an arrival will match any of the traffic pairs already assigned to a given time slot. Consequently, for large $n/m$, we expect that if there is an open slot, it should be relatively easy to assign a new traffic request. To be more precise, as $n/m$ increases, the system performance approaches that of an Erlang system.

*Theorem 3:*
$$\lim_{\frac{n}{m}\to\infty} P_B^F = \lim_{\frac{n}{m}\to\infty} P_B^O = P_B^E. \qquad (27)$$

*Proof:* Theorem 2 implies that it is sufficient to show that $P_B^F \to P_B^E$. For all $k < (mc - 1)$ we have $p_F(f|k) \geq p_F(f|mc - 1)$. Because the nodes are symmetric, the probability of a fit given one empty slot is:

$$p_F(f|mc - 1) = \left(\frac{n - m + 1}{n}\right)^2,$$

which is the probability that the origin and destination nodes of an arrival chosen from a uniform distribution do not match $(m - 1)$ randomly chosen origin-destination pairs. Thus, we have

$$\lim_{\frac{n}{m}\to\infty} p_F(f|k) = \begin{bmatrix} 1, & k < mc \\ 0, & k = mc, \end{bmatrix}$$

which is $p_E(f|k)$. Because (24) is a continuous function of the $\phi_k$'s, we have $\lim_{n/m\to\infty} P_B^F = P_B^E$. Q.E.D.

Although it is easy to compute the Erlang lower bound $P_B^E$, it is extremely difficult to calculate $P_B^O$ or $P_B^F$. The difficulty lies in the calculation of $p_F(f|k)$ and $p_O(f|k)$ as functions of $k$. In particular, notice from (8b) and (16b) that the state probabilities $p(\mathbf{T} \in S_k)$ and $p(\mathbf{M} \in L_k)$ must be known. The aggregate flow eqs. (9) and (15) have therefore not made the exact computation of $P_B^O$ and $P_B^F$ any easier. However, the advantage in using these equations is that we can derive an accurate approximation for $p_F(f|k)$, thereby enabling the calculation of an approximation for the corresponding blocking probability.

### *4.2 Approximate calculation of blocking probability with random assignment*

In order to approximate the blocking probability resulting from fixed assignment, we first approximate $p_F(f|k)$ and subsequently substitute the resulting expression for $\phi_k$ in (24). Let $\underline{k} \equiv (k_1, \cdots, k_c)$ be the vector of time slot occupancy; i.e., there are $k_i$ units of traffic in the $i$th column of the channel-time slot matrix. The probability of a fit can be expressed as

$$p_F(f|k) = \sum_{\underline{k}\in\Omega_k} p_F(f|\underline{k})p_F(\underline{k}|k), \qquad (28)$$

where $\Omega_k$ is the set of all occupancy vectors $\underline{k}$ such that $\Sigma_i k_i = k$, $p_F(f|\underline{k})$ is the conditional fit probability given that the occupancy vector is $\underline{k}$, and $p_F(\underline{k} \mid k)$ is the conditional probability that the occupancy vector is $\underline{k}$ given there are $k$ units of traffic in the system.

To calculate $p_F(f \mid \underline{k})$, note that: (1) the $k_i$ traffic units in time slot $i$ are characterized by $k_i$ pairs of integers between 1 and $n$, which satisfy the fundamental constraint, and because the traffic between nodes is assumed to be symmetric, these integer pairs are equally likely; (2) if $k_i = m$, then any new requests cannot be assigned to the $i$th time slot. We also use the following assumption; and (3) information about traffic in time slot $i$ provides no information about traffic in a different time slot $j$ (traffic in different slots is independent). This assumption is certainly very accurate for a large number of nodes; however, it is quite difficult to prove or disprove in general. An arriving unit of traffic can be assigned to time slot $i$ if and only if its origin does not match any of the $k_i$ origins and its destination does not match any of the $k_i$ destinations of traffic already assigned to column $i$. Observation (1) implies that this event has probability $[(n - k_i)/n]^2$. A unit of traffic cannot be assigned, i.e., does not fit, if and only if it does not fit into any time slot. Thus, using observations (2) and (3), we have that the probability of no fit is the product

$$\prod_{\{i \mid k_i < m\}} \left[ 1 - \left( \frac{n - k_i}{n} \right)^2 \right],$$

and hence, the probability of a fit, given $\underline{k}$, is

$$p_F(f \mid \underline{k}) = 1 - \prod_{\{i \mid k_i < m\}} \left[ 1 - \left( \frac{n - k_i}{n} \right)^2 \right], \tag{29}$$

where the product is assumed to be one if $\{i \mid k_i < m\}$ is the empty set.

Intuition suggests that given $k$ units of traffic in the system, the occupied slots, when random assignment is used, are uniformly distributed throughout the channel-time slot matrix. That is,

$$p_F(\underline{k} \mid k) = \frac{\prod_{i=1}^{c} \begin{bmatrix} m \\ k_i \end{bmatrix}}{\begin{bmatrix} mc \\ k \end{bmatrix}}, \tag{30}$$

where the numerator is the number of ways to have $k_i$ units of traffic in slot $i$ ($i = 1, \cdots, c$) and the denominator is the total number of ways to have $k$ units of traffic in the system. The following example shows, however, that this intuition is misleading and (30) is not the correct distribution.

Consider a system with $n = 3$, $m = 2$, $c = 2$. Figure 2 is a transition diagram illustrating transition rates into the states corresponding to $k = 2$. The ordered pairs, e.g., (1, 1), represent the occupancy vectors

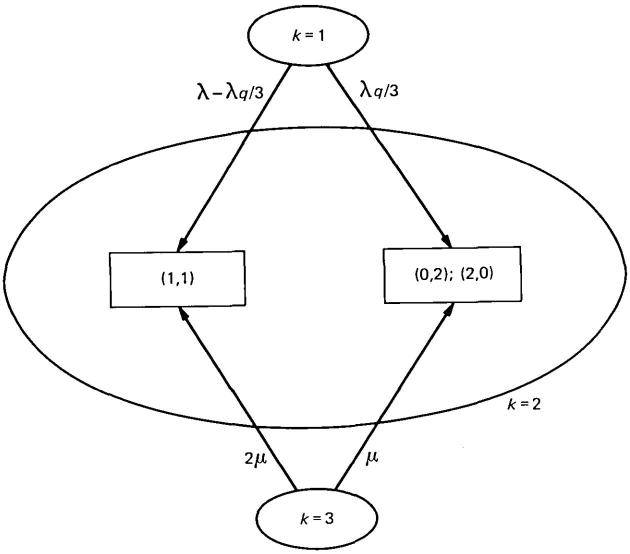


Fig. 2—Transition diagram for example with two channels, two time slots, and three nodes.

$\underline{k}$ for channel-time slot matrices containing two units of traffic, and $q$ is the probability that an arriving unit of traffic can be assigned to the column containing the single unit already assigned. Using (30), it is easily verified that the uniformly distributed assumption implies that

$$p_F[(1, 1) \mid k = 2] = 2\{p_F[(0, 2) \mid k = 2] + p_F[(2, 0) \mid k = 2]\}. \quad (31)$$

It is easy to show by writing the flow equation and using the fact $p_F(\underline{k}) = p_F(\underline{k} \mid k)p(k)$ that (31) can be satisfied only if the rate from $k = 1$ into (1, 1) is twice the rate into the pair of states (0, 2) and (2, 0), i.e., only if $\lambda - \lambda q/3 = 2\lambda q/3$. But the preceding equation can be satisfied only if $q = 1$, which is clearly not true for this system. Therefore, (30) is not correct and the assumption of uniformly distributed occupancy is not strictly true. We may, however, approximate $p_F(\underline{k} \mid k)$ by (30), which when combined with (28) and the expression for $p_F(f \mid \underline{k})$, (29), gives an approximation to $p_F(f \mid k)$ when random assignment is used. That approximation may, in turn, be used in (24) to provide an approximation to $P_B^F$.

Intuitively, one would expect that the approximation for $p_F(f \mid k)$ given by (28) through (30) is accurate for the case of random assignment. In this case (30) becomes more accurate as the ratio of the number of nodes to the number of channels $n/m$ increases. This is because the probability that a new arrival can be assigned to a randomly picked empty time slot increases with $n/m$. The selection of empty slots in the channel-time slot matrix to which new arrivals are assigned therefore becomes less biased. The simulations in Section V indicate that in fact our approximation is extremely accurate for relatively small systems (i.e., four channels, five time slots, and ten nodes). Fixed-assignment schemes other than random assignment are possible, however, where the approximation (30) may not be accurate. For example, it may be desirable to pack the assigned traffic as closely as possible to the left or to the right of the channel-time slot matrix to increase the probability that some column has a relatively large number of empty slots. For this case, the distribution of slot patterns $p(\underline{k}|k)$ may be significantly different from the distribution resulting from random assignment.

An upper bound on the blocking probability obtained using any fixed-assignment scheme can be derived in principle by calculating a lower bound on the fit probability, $p_F(f|k)$. From (28),

$$\begin{aligned} p_F(f|k) &= \sum_{\underline{k}\in\Omega_k} p_F(f|\underline{k})p_F(\underline{k}|k) \\ &\geq \min_{\underline{k}\in\Omega_k} p_F(f|\underline{k}) \equiv p_F^{MIN}(f|k). \end{aligned} \tag{32}$$

A lower bound on $p_F(f|k)$ is therefore obtained by assuming that traffic already assigned is arranged in the configuration which minimizes the probability that a new arrival can be assigned. From Lemma 2 and Theorem 2 we have that

$$P_B^O \leqslant P_B^F \leqslant P_B[p_F^{MIN}(f|1), \cdots, p_F^{MIN}(f|mc)], \tag{33}$$

where the blocking probability, $P_B$, as a function of the fit probabilities is given by (24). Unfortunately, the expression for $p_F(f|\underline{k})$, given by (29), relies upon an independence assumption which has not been proven. Consequently, combining (24), (29), and (33) may not constitute a rigorous upper bound on the blocking probability. The derivation of a tight upper bound on $p_F(f|\underline{k})$, and hence on the blocking probability, $P_B^F$, therefore remains an open problem.

This completes the presentation of analytical results that can be used to evaluate multichannel TDMA performance using either optimal- or fixed-assignment schemes. To summarize, we have obtained an approximation for the blocking probability resulting from random assignment [given by (24), (28), (29), and (30)], and a lower bound on the blocking probability using optimal assignment (i.e., the Erlang blocking probability, $P_B^E$). These quantities can be easily evaluated with the aid of a computer. In the next section we compare these analytical results with computer simulation results.

## V. NUMERICAL RESULTS

The analytical results of the last two sections are now illustrated via some specific examples. Figure 3 shows plots of the probability

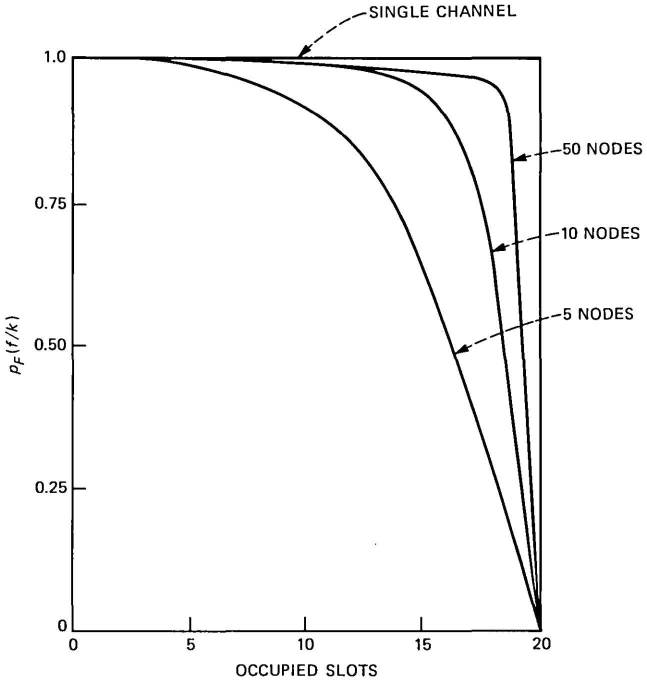


Fig. 3—Probability of a fit vs. number of occupied slots using random assignment for a system with four channels and five time slots.

that a new traffic arrival can be assigned, given that there are $k$ units of traffic already present in the channel-time slot matrix [$p_F(f|k)$ given by (28) through (30)], vs. $k$ for a system with four channels and five time slots per channel. Curves are shown for a system with 5 nodes, 10 nodes, and 50 nodes. These curves approximate the probability that an incoming traffic arrival can be assigned using random assignment. As the number of nodes increases, the curves converge rapidly to the single-channel (step function) case with $nm$ time slots. The same set of curves computed for a system with 4 channels and 10 time slots per channel were nearly identical to those shown in Fig. 3 and are therefore omitted.

It is reasonable to expect that if the fit probabilities shown in Fig. 3 are close to the single-channel case, then the corresponding system blocking probabilities should also be close to the analogous single-channel blocking probability. This is indeed the case as illustrated in Figs. 4 through 7. In each case plots of blocking probabilities vs. normalized load for the single-channel case (Erlang B formula with $nm$ servers), and for the multichannel case using random and optimal assignment are shown. The optimal-assignment curves were obtained by computer simulation. The random-assignment curves were ob-

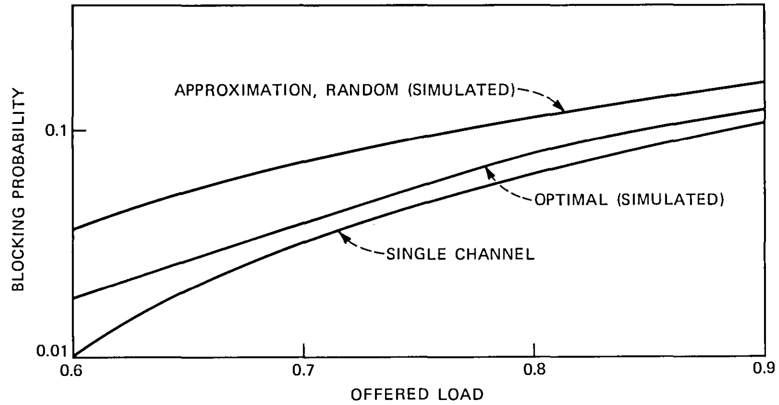


Fig. 4—Blocking probability vs. offered load for a system with 4 channels, 5 time slots, and 10 nodes.

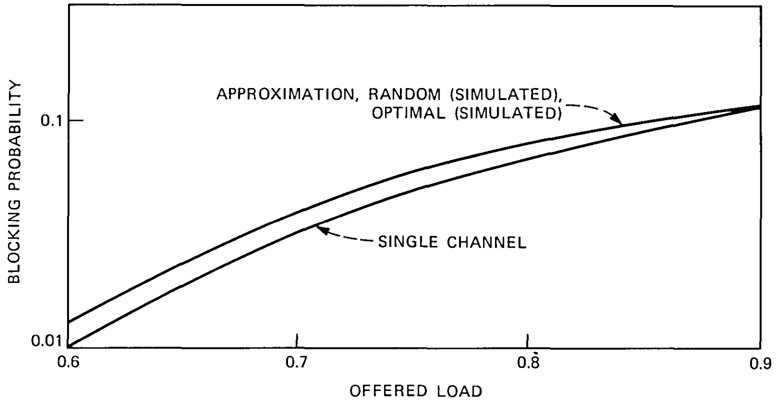


Fig. 5—Blocking probability vs. offered load for a system with 4 channels, 5 time slots, and 50 nodes.

tained both analytically via the approximation described in the last section [(24) and (28) through (30)], and by computer simulation. In all cases the approximate analytical curves are nearly identical to the corresponding simulated curves. Figures 4 and 5 show plots for a system with 4 channels, 5 time slots per channel, and 10 nodes and 50 nodes, respectively. Figures 6 and 7 show analogous plots for systems with 4 channels and 10 time slots per channel.

Figures 4 through 7 indicate that the differences between the simulation results, the analytical approximation, and the lower (single-channel) bound on multichannel blocking probabilities are significant only for systems with a relatively small number of nodes. For the cases shown here, the single-channel system exhibits at most moderate

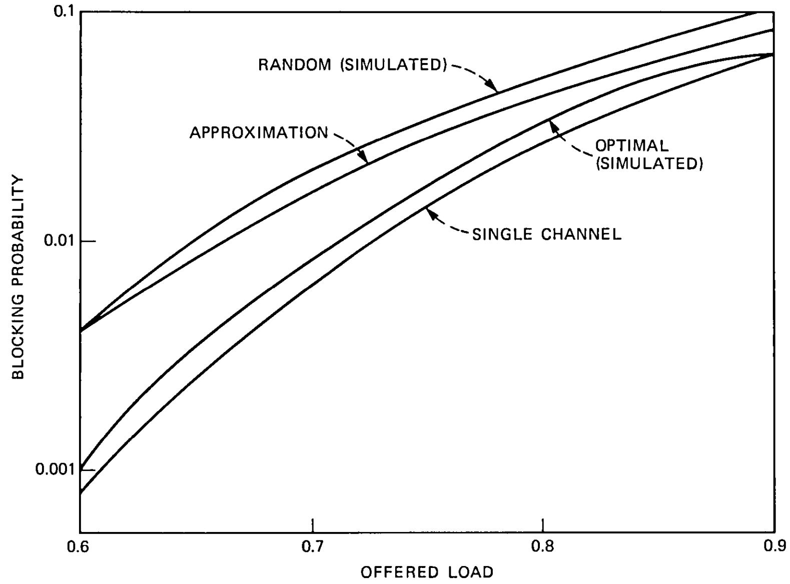


Fig. 6—Blocking probability vs. offered load for a system with 4 channels, 10 time slots, and 10 nodes.

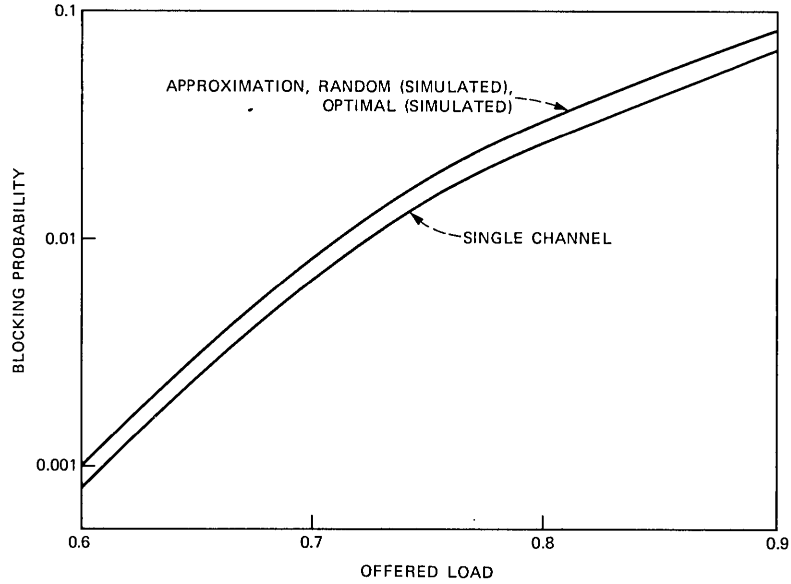


Fig. 7—Blocking probability vs. offered load for a system with 4 channels, 10 time slots, and 50 nodes.

performance improvements over the multichannel random-assignment case. Results obtained for additional cases indicate that if the ratio of nodes to channels ($n/m$) is greater than 10, the difference between blocking probabilities obtained using a multichannel system with a fixed-assignment algorithm and an analogous single-channel system is negligible. This condition is likely to be satisfied in many satellite systems, and hence we reach the important conclusion that, for practical systems, the simplest of assignment schemes will perform nearly as well as an optimal-assignment scheme.

We point out that the traffic model used here must be modified in order to study duplex voice traffic. In this case each traffic request from node $i$ to node $j$ also generates a simultaneous traffic request from node $j$ to node $i$. This alternative traffic model does not require major changes in any of our previous arguments, and hence results obtained from using this model should correspond with those given here.

## VI. CONCLUSIONS

This paper has provided tools with which to evaluate the performance of multichannel TDMA blocking systems. For any multichannel assignment scheme (fixed or optimal), a lower bound on system blocking probability has been obtained along with an accurate approximation for the blocking probability resulting from random assignment.

The numerical results in Section V indicate that multichannel blocking probability is relatively insensitive to the assignment algorithm used when a moderate number of nodes are present. If the ratio of the number of nodes to number of channels is 10 or greater, the difference between blocking probabilities obtained using a multichannel system with a fixed-assignment algorithm and an analogous single-channel system is negligible. This conclusion is fortunate since it implies that the performance of an assignment algorithm, which is simplest to implement, will be nearly optimal.

The results in this paper pertain to networks which handle voice traffic only. Of equal interest are analogous results which apply to networks handling data traffic. Specifically, it would be useful to know whether the performance of a multichannel TDMA queueing system is also insensitive to the particular assignment algorithm used. This issue requires further investigation.

## VII. ACKNOWLEDGMENTS

The authors thank A. S. Acampora and B. E. Simon for many valuable discussions and comments concerning this work.

## REFERENCES

1. R. Cooperman and W. G. Schmidt, "A Satellite Switched SDMA/TDMA System for Wideband Multibeam Satellites," ICC Conf. Rec., June 1973.
2. T. Muratani, "Satellite-Switched Time Domain Multiplex Access," EASCON, 1974.
3. D. O. Reudink and Y. S. Yeh, "A Scanning Spot Beam Satellite System," B.S.T.J., *56,* No. 8 (October 1977), pp. 1549–60.
4. D. O. Reudink, "Spot Beams Promise Satellite Communications Breakthrough," IEEE Spectrum, *15,* No. 9 (September 1978), pp. 36–42.
5. A. S. Acampora, D. O. Reudink, and Y. S. Yeh, "The Transmission Capacity of Multibeam Communication Satellites," Proc. IEEE, *69,* No. 2 (February 1981), pp. 209–25.
6. I. Gopal, D. Coppersmith, and C. K. Wong, "Minimizing Packet Waiting Time in a Multibeam Satellite System," IEEE Trans. Commun., *COM-30,* No. 2 (February 1982), pp. 305–6.
7. A. S. Acampora and B. R. Davis, "Efficient Utilization of Satellite Transponders via Time-Division Multibeam Scanning," B.S.T.J., *57,* No. 8 (October 1978), pp. 2901–14.
8. G. Bongiovanni et al., "An Optimum Time Slot Assignment Algorithm for an SS/TDMA System with Variable Number of Transponders," IEEE Trans. Commun., *COM-29,* No. 5 (May 1981), pp. 721–6.
9. Y. Ito, "Analysis of a Switch Matrix for an SS/TDMA System," Proc. IEEE, *65,* No. 3 (March 1977), pp. 411–9.
10. T. Inukai, "An Efficient SS/TDMA Time Slot Assignment Algorithm," IEEE Trans. Commun., *COM-27,* No. 10 (October 1979), pp. 1449–55.
11. L. Kleinrock, *Queueing Systems,* Vol. 1, New York: John Wiley, 1975, Chapters 2 and 4.
12. F. P. Kelly, *Reversibility and Stochastic Networks,* New York: John Wiley, 1979, Chapter 1.
13. A. S. Acampora, unpublished work.
14. E. J. Messerli, "Proof of a Convexity Property of the Erlang B Formula," B.S.T.J., *51,* No. 4 (April 1972), pp. 951–3.

## APPENDIX A

### *Steady-State Distribution of the Markov Process Associated with Optimal Assignment*

To derive the steady-state distribution of the Markov process associated with optimal assignment, consider a set of $n^2$ independent $M/M/\infty$ queues, labeled by the indices $(i, j)$, with arrival rates $\lambda_{ij}$ and mean service times $\mu_{ij}^{-1}$. Define the Markov process $q_{ij}$ as the number of customers in the queue $(i, j)$. Then each process $q_{ij}$ is a birth-death process with a steady-state distribution given by

$$p_{ij}(q_{ij}) = \frac{1}{q_{ij}!}\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^{q_{ij}} e^{-\lambda_{ij}/\mu_{ij}}, \qquad q_{ij} \geqslant 0.$$

Thus, the matrix process $\mathbf{Q} \equiv [q_{ij}]$ has the steady-state distribution given by

$$p(\mathbf{Q}) = \prod_{i,j} \frac{1}{q_{ij}!}\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^{q_{ij}} e^{-\lambda_{ij}/\mu_{ij}}, \qquad q_{ij} \geqslant 0.$$

Now restrict the process $\mathbf{Q}$ to the set of states such that the matrix constraints (1) are satisfied, and set $\lambda_{ij} = \lambda/n^2$ and $\mu_{ij} = \mu$. The resulting process is the Markov process $\mathbf{T}$ defined in Section III. By

Corollary 1.10 in Kelly,[12] the steady-state distribution of $\mathbf{T}$ is

$$p(\mathbf{T}) = \frac{\prod\limits_{i,j} \frac{1}{t_{ij}!}\left(\frac{\lambda}{n^2\mu}\right)^{t_{ij}} e^{-\lambda/n^2\mu}}{\sum\limits_{\mathbf{T}\in S}\prod\limits_{i,j}\frac{1}{t_{ij}!}\left(\frac{\lambda}{n^2\mu}\right)^{t_{ij}} e^{-\lambda/n^2\mu}}, \qquad \mathbf{T}\in S, \tag{34}$$

where $S$ is defined by (2a).[†] Note that (34) is the conditional probability distribution of the $\mathbf{Q}$ process, given that the state is contained in $S$, i.e., $p(\mathbf{Q}\,|\,\mathbf{Q}\in S)$. Evaluating the denominator in (34) requires an enumeration off all the states in $S$, a formidable task for moderately sized systems.

## APPENDIX B

### *Proof of Lemma 1*

This appendix contains the proof of Lemma 1. From (4), we have

$$\sum_{\mathbf{T}'\neq\mathbf{T}} p(\mathbf{T})r(\mathbf{T},\mathbf{T}') = \sum_{\mathbf{T}'\in S_{\mathbf{T}}^-} p(\mathbf{T})r(\mathbf{T},\mathbf{T}') + \sum_{\mathbf{T}'\in S_{\mathbf{T}}^+} p(\mathbf{T})r(\mathbf{T},\mathbf{T}'). \tag{35}$$

Substituting (3) for $r(\mathbf{T},\mathbf{T}')$ in the first term on the right yields

$$\begin{aligned}\sum_{\mathbf{T}'\in S_{\mathbf{T}}^-} p(\mathbf{T})r(\mathbf{T},\mathbf{T}') &= p(\mathbf{T}) \sum_{\mathbf{T}'\in S_{\mathbf{T}}^-} r(\mathbf{T},\mathbf{T}') \\ &= p(\mathbf{T}) \sum_{i,j} t_{ij}\mu = p(\mathbf{T})\mu k,\end{aligned} \tag{36}$$

where $\Sigma_{i,j}\, t_{ij} = k$ is the total traffic in $\mathbf{T}$. Similarly, for the second term on the right,

$$\begin{aligned}\sum_{\mathbf{T}'\in S_{\mathbf{T}^+}} p(\mathbf{T})r(\mathbf{T},\mathbf{T}') &= p(\mathbf{T}) \sum_{\mathbf{T}'\in S_{\mathbf{T}^+}} r(\mathbf{T},\mathbf{T}') \\ &= p(\mathbf{T}) \sum_{\mathbf{T}'\in S_{\mathbf{T}^+}} \lambda/n^2 \\ &= p(\mathbf{T})\lambda\,|S_{\mathbf{T}^+}|/n^2.\end{aligned} \tag{37}$$

An arrival is modeled as a selection from $n^2$ equally likely possibilities, of which $|S_{\mathbf{T}^+}|$ can be assigned, given that the state is $\mathbf{T}$. Therefore, the conditional probability that an arrival can be assigned is

$$p(f\,|\,\mathbf{T}) = |S_{\mathbf{T}}^+|/n^2. \tag{38}$$

Combining (35) through (38) gives (5).

[†] This distribution was also derived in Ref. 13.

## APPENDIX C

### *Proof of Theorem 1*

To prove Theorem 1, we sum the flow equations in (7) over the set of states with $k$ units of traffic, i.e.,

$$\sum_{\mathbf{T}\in S_k} [\mu k p(\mathbf{T}) + \lambda p(f\,|\,\mathbf{T})p(\mathbf{T})]$$

$$= \sum_{\mathbf{T}\in S_k} \sum_{\mathbf{T}'\in S^{+}_{\mathbf{T}}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}) + \sum_{\mathbf{T}\in S_k} \sum_{\mathbf{T}'\in S^{-}_{\mathbf{T}}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}). \qquad (39)$$

The left-hand term is

$$\sum_{\mathbf{T}\in S_k} [\mu k p(\mathbf{T}) + \lambda p(f\,|\,\mathbf{T})p(\mathbf{T})] = \mu k \sum_{\mathbf{T}\in S_k} p(\mathbf{T}) + \lambda \sum_{\mathbf{T}\in S_k} p(f\,|\,\mathbf{T}) \cdot p(\mathbf{T})$$

$$= \mu k p_O(k) + \lambda p_O(f\,|\,k)p_O(k). \qquad (40)$$

We evaluate the terms on the right side of (39) by interchanging the order of summation, which is permissible since the sums are always finite. Because $S_{k+1} = \cup_{\mathbf{T}\in S_k} S^{+}_{\mathbf{T}}$, and for $\mathbf{T}' \in S_{k+1}$, $r(\mathbf{T}', \mathbf{T}) \neq 0$ only if $\mathbf{T}' \in S^{+}_{\mathbf{T}}$, interchanging the sums in the first term on the right yields

$$\sum_{\mathbf{T}\in S_k} \sum_{\mathbf{T}'\in S^{+}_{\mathbf{T}}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}) = \sum_{\mathbf{T}'\in S_{k+1}} \sum_{\mathbf{T}\in S_k} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}).$$

But if $\mathbf{T}' \in S_{k+1}$ and $\mathbf{T} \in S_k$, then $r(\mathbf{T}', \mathbf{T}) \neq 0$ only if $\mathbf{T} \in S^{-}_{\mathbf{T}'}$. Consequently,

$$\sum_{\mathbf{T}'\in S_{k+1}} \sum_{\mathbf{T}\in S_k} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}) = \sum_{\mathbf{T}'\in S_{k+1}} \sum_{\mathbf{T}\in S^{-}_{\mathbf{T}'}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T})$$

$$= \sum_{\mathbf{T}'\in S_{k+1}} p(\mathbf{T}')\mu(k + 1)$$

$$= \mu(k + 1)p_O(k + 1), \qquad (41)$$

where the next to the last step follows from Lemma 1. A similar argument shows that the second term on the right is

$$\sum_{\mathbf{T}\in S_k} \sum_{\mathbf{T}'\in S^{-}_{\mathbf{T}}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T}) = \sum_{\mathbf{T}'\in S_{k-1}} \sum_{\mathbf{T}\in S_k} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T})$$

$$= \sum_{\mathbf{T}'\in S_{k-1}} \sum_{\mathbf{T}\in S^{+}_{\mathbf{T}'}} p(\mathbf{T}')r(\mathbf{T}', \mathbf{T})$$

$$= \sum_{\mathbf{T}'\in S_{k-1}} p(\mathbf{T}')\lambda p(f\,|\,\mathbf{T}')$$

$$= \lambda p_O(f\,|\,k - 1)p_O(k - 1). \qquad (42)$$

Combining (40) through (42) yields (9).

## APPENDIX D

### *Proof of Lemma 2*

Lemma 2 is stated in the literature and although some unpublished proofs are referenced in Ref. 14, there does not seem to be a published proof.

We prove this theorem by calculating the partial derivatives and showing that they are nonnegative in the region of interest. Differentiating (24) gives

$$\frac{\partial}{\partial \phi_l}\left[\frac{\sum_{k=0}^{mc-1} \frac{1}{k!} \rho^k \prod_{j=0}^{k} \phi_j}{\sum_{k=0}^{mc} \frac{1}{k!} \rho^k \prod_{j=0}^{k-1} \phi_j}\right] = \frac{b_1(l)a_2(l) - a_1(l)b_2(l)}{[b_1(l) + b_2(l)\phi_l]^2}, \tag{43}$$

where

$$a_1(l) \equiv \sum_{k=0}^{l-1} \frac{1}{k!} \rho^k \alpha(k) \qquad b_1(l) \equiv \sum_{k=0}^{l} \frac{1}{k!} \rho^k \alpha(k-1)$$

$$a_2(l) \equiv \sum_{k=l}^{mc-1} \frac{1}{k!} \rho^k \beta(k) \qquad b_2(l) \equiv \sum_{k=l+1}^{mc} \frac{1}{k!} \rho^k \beta(k-1)$$

and

$$\alpha(k) \equiv \prod_{j=0}^{k} \phi_j \qquad \beta(k) \equiv \prod_{\substack{j=0 \\ j\neq l}}^{k} \phi_j$$

$$\alpha(-1) = \beta(-1) = 1.$$

The derivative (43) is nonnegative if and only if

$$b_1(l)a_2(l) - a_1(l)b_2(l) \geqslant 0. \tag{44}$$

The expression (44) is a polynomial in $\rho$ with powers ranging from $l$ to $mc + l - 1$. We will show that the coefficient of each power is nonnegative. Let $mc + l - 1 \geqslant s \geqslant l$ and define the sets

$$\Omega_1 \equiv \{(k_1, k_2) \mid k_1 + k_2 = s,\ 0 \leqslant k_1 \leqslant l,\ l \leqslant k_2 \leqslant mc - 1\}$$

$$\Omega_2 \equiv \{(k_3, k_4) \mid k_3 + k_4 = s,\ 0 \leqslant k_3 \leqslant l - 1,\ l + 1 \leqslant k_4 \leqslant mc\}.$$

Now the coefficient of $\rho^s$ in (44) may be written as

$$\sum_{\Omega_1} \frac{1}{k_1!k_2!}\, \alpha(k_1 - 1)\beta(k_2) - \sum_{\Omega_2} \frac{1}{k_3!k_4!}\, \alpha(k_3)\beta(k_4 - 1). \tag{45}$$

Define the set

$$\Omega_{12} = \{(k_1, k_2) \mid (k_1, k_2) \in \Omega_1,\ k_1 \geqslant 1\}$$

and note that $(k_3, k_4) \in \Omega_2$ if and only if $k_3 = k_1 - 1$, $k_4 = k_2 + 1$, where $(k_1, k_2) \in \Omega_{12} \subset \Omega_1$. This implies that (45) is equal to $\omega_1 + \omega_2$, where

$$\omega_1 \equiv \sum_{\Omega_1 - \Omega_{12}} \frac{1}{k_1!k_2!} \alpha(k_1 - 1)\beta(k_2) \geqslant 0 \tag{46a}$$

$$\omega_2 \equiv \sum_{\Omega_{12}} \left[\frac{1}{k_1!k_2!} - \frac{1}{(k_1 - 1)!(k_2 + 1)!}\right] \alpha(k_1 - 1)\beta(k_2). \tag{46b}$$

A term in the sum $\omega_2$ is negative if and only if $(k_2 + 1)/k_1 < 1$. But this inequality and the definition of $\Omega_{12}$ imply that $l + 1 \leqslant k_2 + 1 < k_1$, which is impossible because $k_1 \leqslant l$ for $(k_1, k_2) \in \Omega_{12}$. Consequently, we have $\omega_2 \geqslant 0$. This $\omega_1 \geqslant 0$ implies that (44) and (45) are nonnegative; therefore, each partial derivative (43) is nonnegative. Q.E.D.

## AUTHORS

**Michael L. Honig,** B.S. (Electrical Engineering), 1977, Stanford University; M.S., Ph.D. (Electrical Engineering), University of California, Berkeley, in 1978 and 1981, respectively; AT&T Bell Laboratories, 1981–1982; AT&T Information Systems, 1983—. At AT&T Bell Laboratories and AT&T Information Systems, Mr. Honig has worked on echo cancellation of voiceband data signals, performance evaluation of convolutional codes, and performance analysis of local area networks. He is currently working on office information networks. Member, IEEE, Tau Beta Pi, Phi Beta Kappa.

**Steven M. Barta,** B.S. (Mathematics and Physics), 1973, Yale University; S.M., Ph.D. (Electrical Engineering), Massachusetts Institute of Technology, in 1976 and 1978, respectively; AT&T Bell Laboratories, 1978–1982; AT&T Information Systems, 1983—. As a Member of the Technical Staff at AT&T Bell Laboratories and AT&T Information Systems, Mr. Barta worked on economic studies, performance evaluation of convolutional codes, and analysis and simulation of traffic flow in data communication networks. He is currently developing a real-time traffic simulator for local area networks. Member, Phi Beta Kappa, Sigma Xi, IEEE.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# On Approximations for Queues, I: Extremal Distributions

By W. WHITT*

(Manuscript received May 4, 1983)

Many approximations for queueing characteristics such as the mean equilibrium queue length are based on two moments of the interarrival and service times. To evaluate these approximations, we suggest looking at the set of all possible values of the queueing characteristics given the specified moment parameters. This set-valued function is useful for evaluating the accuracy of approximations. For several models, such as the GI/M/1 queue, the set of possible values for the mean queue length given limited-moment information can be conveniently described by simple extremal distributions. Here we calculate the set of possible values for the mean queue length in a GI/M/1 queue and show how it depends on the traffic intensity and the second moment. We also use extremal distributions to compare alternative parameters for approximations. The results provide useful insights about approximations for non-Markov networks of queues and other complex queueing systems. The general procedure is widely applicable to investigate the accuracy of approximations.

## I. INTRODUCTION AND SUMMARY

Queueing models are important tools for studying the performance of complex systems, but despite the substantial queueing theory literature, it is often necessary to use approximations. The purpose of this series of papers is to help develop a theory for evaluating queueing

* AT&T Bell Laboratories.

approximations. Devising appropriate queueing approximations no doubt will continue to be largely an art, but we believe that there is a need and a real possibility for more supporting theory.

In this series of papers we examine the accuracy of queueing approximations that are based on a few parameters partially characterizing the arrival process and the service-time distribution. We use an approach originally introduced by Holtzman[1] and Eckberg[2] at Bell Laboratories and Rolski[3–5] in Poland. Since the approximations apply to all arrival processes and all service-time distributions with the same parameters, we propose evaluating the approximations by examining the set of all possible values of the congestion measure consistent with the specified parameters. To be specific, consider the GI/G/1 queue, which has a single server, unlimited waiting room, the first-come first-served discipline, and a renewal arrival process independent of iid (independent and identically distributed) service times. Many approximations for the equilibrium mean queue length in the GI/G/1 queue are based on the first two moments of the interarrival-time and service-time distributions; see Shanthikumar and Buzacott[6] and Whitt.[7] In this context we suggest considering the set-valued function that maps the four moment parameters into the set of possible values of the mean queue length.

It should be clear that we are in an excellent position to develop and evaluate approximations if we can identify such set-valued functions. We can see if a candidate approximation is an element of this set for all parameters of interest; then there always is a system for which the approximation is exact. We can also see if an approximation is in the middle of this set; then large errors are avoided and the approximation usually corresponds to a typical system value.

There is also much to be learned without considering any specific approximation. The range of values indicates the possible accuracy of any approximation. We can investigate how this range depends on the parameters to determine how the possible accuracy depends on the parameters. We can see how the range is reduced by incorporating additional information, e.g., another moment. We can also compare different parameter specifications by comparing the different set-valued functions.

This approach has wide applicability in queueing and elsewhere, provided that we can indeed identify the desired set-valued functions. As one would expect, this task is usually difficult, but there is an emerging methodology for attacking this problem. It is sometimes possible to identify relatively simple extremal distributions that yield the maximum and minimum values of the congestion measure given the parameters. A major tool for this purpose is the theory of complete Tchebycheff systems in Karlin and Studden.[8] The idea of applying complete Tchebycheff systems and extremal distributions to congestion models is due to Holtzman[1] and Rolski.[3] Eckberg[2] first used this approach to compare alternate parameter specifications, primarily the peakedness versus the variance as a second parameter in addition to the mean in GI/M/s loss systems. Other relevant references are Bergmann et al.,[9] Daley and Rolski,[10] Karr,[11] Stoyan,[12] and Whitt.[13,14]

The principal focus in the papers here is the GI/M/1 queue, which has an exponential service-time distribution. (We also have results for more general GI/G/1 queues; see Section V of this paper and Sections VI and VII of Part III, a subsequent paper in this issue of the *Journal.*) In Part I, we describe the set of all possible values of the mean queue length in the GI/M/1 model given the service rate and various parameters partially characterizing the interarrival-time distribution, especially the first two moments. We obtain useful descriptions of the way this set depends on the parameters (see Section II). For example, the maximum relative error [defined in (4)] in the mean queue length given the first two moments of the interarrival time turns out to be precisely the squared coefficient of variation (variance divided by the square of the mean) of the interarrival time; see Corollary 1. We also evaluate alternate parameter specifications (see Sections III and IV).

We must emphasize that we are not actually interested in the GI/M/1 model itself. Given a GI/M/1 model, it is obviously not difficult to calculate the mean queue length exactly. We are actually interested in more general models in which exact solutions are not possible. Where GI/M/1 models arise, they arise as approximations, e.g., the arrival process is approximated by a renewal process partially characterized by the first two moments of the renewal interval.[15-18] Then there is no corresponding renewal-interval distribution for exact analysis.

We became motivated to conduct this study while developing the software package QNA (Queueing Network Analyzer),[17,18] which calculates approximate congestion measures for non-Markovian networks of queues, i.e., with non-Poisson arrival processes and nonexponential service-time distributions. The procedure in QNA is, first, to approximate each arrival process by a renewal process partially characterized by the first two moments of the renewal interval and, second, for each node to apply approximation formulas for the congestion measures in a GI/G/m queue partially characterized by the first two moments of the interarrival-time and service-time distributions. It is natural to study these two steps separately. The first step is studied in Whitt[15] and Albin.[16] The second step is studied here.

For the network of queues and other applications, we would actually like to treat the more general GI/G/m model, but we are not yet able to do this. Nevertheless, we believe that the GI/M/1 results here are important. They indicate what happens more generally. While the exponential distribution is exceptional in its analytic simplicity, it is rather typical in its degree of variability (in between deterministic and highly variable). Moreover, the sharp analytic results available for the GI/M/1 model will be useful theoretical reference points for other cases that require relatively complicated numerical methods or simulation. Even if an extremal distribution is identified for other GI/G/m queues, it may be a nontrivial task to calculate the mean queue length.

We emphasize that the relevance of the extremal distributions for the GI/M/1 model was established before.[1-5] Here we apply this theory to examine in detail the implications for queueing approximations. We determine which parameters are best, how the quality of approximations depends on the parameters, and how much additional information helps.

As an important part of our results, we display the extremal distributions yielding the extreme values of the mean queue length. These extremal distributions are of interest beyond the GI/M/1 queue considered here because they are also extremal in many other settings. (This will be evident from Sections II and V.) Moreover, in settings such as the GI/G/m queue in which the actual extremal distributions are still unknown, the GI/M/1 extremal distributions can be used in numerical methods and simulations to get an approximate range of possible congestion values.

To describe the situation for the GI/M/1 queue, let $u$ be an interarrival time, $v$ a service time, $\rho$ the traffic intensity ($\rho = Ev/Eu$), $c^2$ the squared coefficient of variation of an interarrival time, and $L$ the expected equilibrium queue length (number in system) at an arbitrary time. For the GI/M/1 queue,[19]

$$L = \rho/(1 - \sigma), \tag{1}$$

where $\sigma$ is the unique root in the open interval (0, 1) of the equation

$$\phi[\mu(1 - \sigma)] = \sigma, \tag{2}$$

with $\mu = 1/Ev$ and $\phi(s)$ the Laplace-Stieltjes transform of the interarrival-time cdf, say $F$,

$$\phi(s) = \int_0^\infty e^{-st} dF(t). \tag{3}$$

The root $\sigma$ in (2) is also of interest itself because it is the probability that a customer will have to wait before beginning service. It is clear from (1) and (2) that $\sigma$ and $L$ depend on the entire cdf $F$, not just its first two moments.

So, what about the range of possible values for $\sigma$ and $L$ in the GI/M/1 queue? Unfortunately, the range can be very wide. For example, let $Eu = 2$, $Eu^2 = 12$ (so that $\mathrm{Var}(u) = 8$ and $c^2 = 2$), and $Ev = 4/3$ (so that $\rho = 2/3$). The possible values of $\sigma$ range from 0.417 to 0.806 and the possible values of $L$ range from 1.14 to 3.44, giving a maximum relative error of 200 percent (Table IV).

This wide range naturally causes us to question the value of the various two-moment approximations. However, the particular distributions yielding the extreme values of $L$ suggest an explanation. These extremal distributions are two-point distributions, so they are obviously very unusual. We would hope that for typical (nice) distributions $\sigma$ and $L$ would not vary much among interarrival-time distributions with the same moments. In Parts II and III,[20,21] we investigate how much the range is reduced by imposing various shape constraints on the interarrival-time distribution. Part II by Klincewicz and Whitt[20] presents a new approach. Since the theory of complete Tchebycheff systems no longer applies with shape constraints, Part II uses nonlinear programming to identify the extreme values of $L$ and the associated extremal interarrival-time distributions given various shape constraints. We believe that Part II is the first investigation of extremal distributions in the presence of shape constraints.

The numerical results in Part II are strikingly similar to the theoretical results in Part I, suggesting that a theory corresponding to Part I can be developed for many kinds of shape constraints. Part III shows how this can be done in one important special case. Part III shows that the theory of complete Tchebycheff systems can be applied again for one important kind of shape constraint: assuming that the distribution is a mixture of exponential distributions.

Overall, this study indicates that two-moment approximations can perform poorly, but if the distribution is not too irregular then they should perform reasonably well. At any rate, numbers are provided so that we can reach our own conclusions, which may depend on the circumstances.

Here is how the rest of this paper is organized. In Section II we study the extremal distributions with the first two moments fixed. In Section III we do a similar analysis with the mean and the peakedness (the transform evaluated at the service rate) fixed. In Section IV we investigate other parameter specifications, including the first three moments. Finally, in Section V we briefly discuss extremal distributions in other models such as the GI/G/1 queue and the GI/M/1 loss system. It is significant that the theory of extremal distributions is not limited to the GI/M/1 model.

## II. EXTREMAL DISTRIBUTIONS GIVEN THE FIRST TWO MOMENTS

Consider the set of all probability distributions on the interval $[0, bm_1]$, $b \leqslant \infty$, having first two moments $m_1$ and $m_2$ (and no mass at infinity). This is a convex set depending on the three parameters $b$, $m_1$ and $c^2$, where $c^2$ is the squared coefficient of variation: $c^2 = (m_2 - m_1^2)/m_1^2$. The set is nonempty provided that $b \geqslant 1 + c^2$. Two distributions in this set are of particular interest; we call them the upper and lower bounds because they yield the maximum and minimum mean queue lengths, respectively, among interarrival-time distribution in this set. The *upper bound* is the two-point distribution with mass $c^2/(1 + c^2)$ on 0 and mass $1/(1 + c^2)$ on $m_1(1 + c^2)$, having cdf denote by $F_u$, and the *lower bound* is the two-point distribution with mass $c^2/[c^2 + (b - 1)^2]$ on $bm_1$ and mass $(b - 1)^2/[c^2 + (b - 1)^2]$ on $m_1[1 - c^2/(b - 1)]$, having cdf denoted by $F_\ell$. As $b \to \infty$, the lower bound approaches (converges in law) to the *limiting lower bound*, which is the one-point distribution with mass 1 on $m_1$, having cdf denoted by $F_{\hat{\ell}}$.

Note that the limiting lower bound is not actually in the reference set because it has zero variance. These distributions are especially useful because they are minimal and maximal elements for a partial ordering of the distributions based on the Laplace-Stieltjes transforms.

*Definition 1:* $F_1 \leqslant_L F_2$ for two cdf's on $[0, \infty)$ if $\phi_1(s) \leqslant \phi_2(s)$ for all $s \geqslant 0$, where $\phi_i$ is the Laplace-Stieltjes transform of $F_i$ defined in (3).

Since the transform $\phi(s)$ is the expectation of a decreasing function, the smaller cdf in the ordering $\leqslant_L$ tends to have what we would normally think of as the stochastically larger distribution; in fact, in Section 1.8 of Stoyan,[12] $F_1 \leqslant_L F_2$ is said to hold if $\phi_1(s) \geqslant \phi_2(s)$ for all $s \geqslant 0$. However, smaller interarrival times mean more arrivals and more congestion. We use this definition because the upper-(lower-) bound distribution yields the maximum (minimum) mean queue length.

Let $\mathbf{F} \equiv \mathbf{F}(m_1, c^2, b)$ be the set of all cdf's with parameters $m_1$, $c^2$, and $b$. Let $F_u$ and $F_\ell$ be the cdf's in $\mathbf{F}$ associated with the special extremal distributions, and let $F_{\hat{\ell}}$ be the associated limiting lower-bound cdf. The following proposition is just a restatement of 2.1.1 of Eckberg,[2] which in turn is an elementary consequence of the theory of complete Tchebycheff systems.[8]

*Proposition* 1: *For all* $F \in \mathbf{F}$, $F_{\hat{\ell}} \leqslant_L F_\ell \leqslant_L F \leqslant_L F_u$.

It is a simple matter to check the following property.

*Proposition 2:* $F_\ell$ *decreases in* $\leqslant_L$ *as* $b$ *increases and* $\phi_\ell(s) \to \phi_{\hat{\ell}}(s)$ *for all* $s$ *as* $b \to \infty$.

As noted by Holtzman,[1] Rolski,[3-5] and Eckberg,[2] the ordering $\leqslant_L$ and the extremal distributions have immediate application to queues. Consider the GI/M/1 queue with fixed service rate $\mu$ and interarrival-time distributions in $\mathbf{F}$. Without loss of generality, assume $m_1 = 1$. Now it is natural to work with the three parameters $\rho$, $c^2$, and $b$. Let $L$ and $\sigma$ in (1) and (2) be indexed to indicate the extremal interarrival-time distributions. As an immediate consequence of Proposition 1 and (2), we have

*Proposition* 3: *For all* $F \in \boldsymbol{F}(\rho, c^2, b)$, $\sigma_{\hat{\ell}} \leqslant \sigma_{\ell} \leqslant \sigma \leqslant \sigma_{u}$ *and* $L_{\hat{\ell}} \leqslant L_{\ell} \leqslant L \leqslant L_{u}$.

*Remark 1:* More generally, if $F_1 \leqslant_L F_2$ for two interarrival-time cdf's, then $\sigma_1 \leqslant \sigma_2$ in the associated GI/M/1 queue with common service rate. This in turn implies not only that $L_1 \leqslant L_2$ but also that the associated steady-state queue-length distributions are stochastically ordered; see Theorem 5.2.3b of Stoyan.[12]

For approximations, it is interesting to know about the maximum relative error (*MRE*) in $L$, defined by

$$MRE \equiv MRE(\rho, c^2, b) \equiv (L_{u} - L_{\ell})/L_{\ell}. \tag{4}$$

From (1), we see that $MRE = (\sigma_{u} - \sigma_{\ell})/(1 - \sigma_{u})$.

Now we show how the extremal queue characteristics ($\sigma_{\ell}$, $L_{\ell}$, etc.) and *MRE* depend on the parameters $\rho$, $c^2$, and $b$. We first describe how $\sigma_{\hat{\ell}}$ depends on $\rho$, the only relevant parameter for the limiting lower bound.

*Theorem 1: For* $0 < \rho < 1$, $\sigma_{\hat{\ell}} < \rho$ and

$$\frac{d\sigma_{\hat{\ell}}}{d\rho} = \frac{\rho^{-2}(1 - \sigma_{\hat{\ell}})e^{-(1-\sigma_{\hat{\ell}})/\rho}}{1 - \rho^{-1}e^{-(1-\sigma_{\hat{\ell}})/\rho}} > 0.$$

*Proof:* Consider eq. (2) for $F_{\hat{\ell}}$. The function

$$f(x) = x - e^{-(1-x)/\rho} \tag{5}$$

is positive for $0 < x < \sigma_{\hat{\ell}}$ and negative for $\sigma_{\hat{\ell}} < x < 1$, so to show that $\sigma_{\hat{\ell}} < \rho$ it suffices to show that $f(\rho) = \rho - e^{-(1-\rho)/\rho}$ is negative for $0 < \rho < 1$. Make the change of variables $y = (1 - \rho)/\rho$ to obtain $f(y) = (1 + y)^{-1} - e^{-y}$, which is clearly positive for all $y > 0$. To verify the inequality for the derivative, differentiate $f(x)$ in (5). Use $\sigma_{\hat{\ell}} < \rho$ to show that the denominator is always positive:

$$\rho^{-1}e^{-(1-\sigma_{\hat{\ell}})/\rho} \leqslant \rho^{-1}e^{-(1-\rho)/\rho} < 1.$$

We now show that all results for $\sigma_{\hat{\ell}}$ immediately imply results for $\sigma_{u}$.

*Theorem 2:* $\sigma_{u} = 1 - (1 - \sigma_{\hat{\ell}})/(1 + c^2)$.

*Proof:* For the upper bound, eq. (2) is

$$\frac{c^2}{c^2 + 1} + \frac{1}{c^2 + 1}\, e^{-(1-\sigma_{u})(1+c^2)/\rho} = \sigma_{u}.$$

Multiply both sides by $c^2 + 1$, subtract $c^2$ from both sides, and then make the change of variables $1 - \sigma_{\hat{\ell}} = (1 - \sigma_{u})(1 + c^2)$ to obtain eq. (2) for the limiting lower bound.

*Corollary 1:* $L_u = L_{\hat{\ell}}\,(1 + c^2)$ *and* $MRE(\rho, c^2, \infty) = c^2$.

*Remark:* Theorems 1 and 2 together with (1) imply that $\sigma_u$ and $L_u$ are increasing in $\rho$.

We now turn to the lower bound when there is a bound on the distribution ($b < \infty$). Straightforward but tedious calculations (differentiation) verify the expected monotonicity properties:

*Theorem 3:* (*a*) *The lower-bound characteristics* $\sigma_\ell$ *and* $L_\ell$ *are increasing in* $\rho$ *and* $c^2$ *and decreasing in b.* (*b*) $MRE(\rho, c^2, b)$ *is increasing in b.*

Combining Theorems 2 and 3b, we obtain

*Corollary 2:* $MRE(\rho, c^2, b) \leqslant c^2$.

Numerical evaluation of $MRE(\rho, c^2, b)$ for 14 values of $\rho$, 4 values of $c^2$, and 5 values of $b$ support the following conjecture.

*Conjecture 1:* $MRE(\rho, c^2, b)$ *is decreasing in* $\rho$.

In Table I we display $MRE(\rho, c^2, b)$ for three values of $\rho$, four values of $c^2$, and four values of $b$. These specific cases show that $MRE(\rho, c^2, b)$ is strongly affected by each of the parameters $\rho$, $c^2$, and $b$. The bound $b$ can make a big difference, especially for larger $\rho$ and $c^2$; see the case $\rho = 0.9$ and $c^2 = 4$. These specific cases demonstrate that $MRE(\rho, c^2, b)$ is not monotone in $c^2$. In fact, when $c^2$ increases with $b$ fixed, the lower-bound distribution $F_\ell$ eventually coincides with the upper-bound distribution $F_u$, becoming the two-point distribution with mass $b^{-1}$ on $b$ and mass $1 - b^{-1}$ on 0 ($m_1 = 1$). Of course, as $c^2 \to 0$, $F_\ell$ and $F_u$ both approach $F_{\hat{\ell}}$, so that $MRE(\rho, c^2, b) \to 0$ too as $c^2 \to 0$. The numerical results also support the following conjecture:

*Conjecture 2:* $MRE(\rho, c^2, b)$ *is unimodal in* $c^2$.

We now investigate how the extremal queue characteristics and

Table I—Values of MRE($\rho$, $c^2$, $b$) for the GI/M/1 queue*

| Traffic Intensity, $\rho$ | Squared Coefficient of Variation, $c^2$ | Bound on Interarrival-Time Distribution in Multiples of the Mean | | | |
|---|---|---|---|---|---|
| | | $b = 5$ | $b = 10$ | $b = 20$ | $b = 40$ |
| 0.5 | 0.5 | 0.373 | 0.442 | 0.472 | 0.487 |
| | 1.0 | 0.604 | 0.833 | 0.924 | 0.964 |
| | 2.0 | 0.527 | 1.40 | 1.75 | 1.89 |
| | 4.0 | 0.000 | 1.28 | 3.01 | 3.59 |
| 0.7 | 0.5 | 0.231 | 0.350 | 0.424 | 0.462 |
| | 1.0 | 0.290 | 0.583 | 0.791 | 0.897 |
| | 2.0 | 0.185 | 0.699 | 1.33 | 1.68 |
| | 4.0 | 0.000 | 0.349 | 1.56 | 2.86 |
| 0.9 | 0.5 | 0.070 | 0.143 | 0.248 | 0.353 |
| | 1.0 | 0.072 | 0.174 | 0.365 | 0.610 |
| | 2.0 | 0.043 | 0.143 | 0.374 | 0.858 |
| | 4.0 | 0.000 | 0.071 | 0.232 | 0.712 |

* The maximum relative error in the steady-state mean queue length $L$ given the traffic intensity $\rho$, the interarrival-time squared coefficient of variation $c^2$, and the bound on the interarrival-time distribution $b$ (in multiples of the mean); see Section IV.

$MRE(\rho, c^2, b)$ behave in light and heavy traffic, i.e., as $\rho \rightarrow 0$ and $\rho \rightarrow 1$. As an easy consequence of (2), we obtain

*Theorem 4: As* $\rho \rightarrow 0$,

$$\sigma_u \rightarrow c^2/(1 + c^2), \qquad \sigma_\ell \rightarrow 0, \qquad \textit{and} \quad MRE(\rho, c^2, b) \rightarrow c^2.$$

We describe the behavior as $\rho \rightarrow 1$ for $b < \infty$ in more detail. The following result provides an interesting refinement to the classical heavy-traffic limit theorem,[7] from which we can deduce that $(1 - \rho)L \rightarrow (1 + c^2)/2$ as $\mu$ approaches $\lambda$ from above for any fixed renewal arrival process.

*Theorem 5: For all b,*

$$1 - \sigma_u = \frac{2(1-\rho)}{1+c^2} - \frac{4(1-\rho)^2}{3(1+c^2)} + 0(1-\rho)^3 \tag{6}$$

*and, for* $b < \infty$,

$$1 - \sigma_\ell = \frac{2(1-\rho)}{1+c^2} - \frac{4(1-\rho)^2}{3(1+c^2)}\frac{m_3}{(1+c^2)^2} + 0(1-\rho)^3, \tag{7}$$

*where*

$$m_3 = \frac{c^2b^3}{c^2 + (b-1)^2} + \frac{(b-1-c^2)^3}{(b-1)(c^2+(b-1)^2)}, \tag{8}$$

*so that, for* $b < \infty$,

$$\lim_{\rho \rightarrow 1} \frac{MRE(\rho, c^2, b)}{1-\rho} = \frac{4}{3}\left(\frac{m_3}{(1+c^2)^2} - 1\right). \tag{9}$$

*Proof:* Let $x = (1 - \sigma_\ell)/\rho$. To find the derivative of $x$ with respect to $\rho$, differentiate with respect to $\rho$ in eq. (2), i.e.,

$$1 - \rho x = e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + 0(x^4)$$

or

$$-\rho = \frac{x}{2} - \frac{x^2}{6} + 0(x^3).$$

After successive differentiation with L'Hospital's rule, this yields $x'(1) = -2$ and $x''(1) = -8/3$. From Taylor's theorem and Theorem 1, we obtain (6). The calculation for the lower bound in (7) is similar.

*Remarks:* It is possible to check the consistency of (6) and (7) because they must agree as $b \rightarrow 1 + c^2$. It is not possible to do a consistency check as $b \rightarrow \infty$ because the two iterated limits involving $b \rightarrow \infty$ and $\rho \rightarrow 1$ are not equal.

We conclude this section by displaying in Table II the extremal

Table II—The extremal GI/M/1 characteristics for fixed traffic intensity, $\rho$, squared coefficient of variation, $c^2$, and bound on the distribution $b$: Case of $c^2 = 2.0$

| Traffic Intensity, $\rho$ | Upper-Bound Characteristics | Bound on Interarrival-Time Distribution in Multiples of the Mean | | | |
|---|---|---|---|---|---|
| | | $b = 5$ | $b = 10$ | $b = 20$ | $b = 40$ |
| 0.2 | $\sigma_u = 0.669$ | $\sigma_\ell = 0.092$ | $\sigma_\ell = 0.022$ | $\sigma_\ell = 0.012$ | $\sigma_\ell = 0.009$ |
| | $L_u = 0.604$ | $L_\ell = 0.220$ | $L_\ell = 0.204$ | $L_\ell = 0.202$ | $L_\ell = 0.202$ |
| 0.5 | $\sigma_u = 0.734$ | $\sigma_\ell = 0.594$ | $\sigma_\ell = 0.361$ | $\sigma_\ell = 0.269$ | $\sigma_\ell = 0.233$ |
| | $L_u = 1.88$ | $L_\ell = 1.23$ | $L_\ell = 0.783$ | $L_\ell = 0.684$ | $L_\ell = 0.652$ |
| 0.7 | $\sigma_u = 0.822$ | $\sigma_\ell = 0.790$ | $\sigma_\ell = 0.698$ | $\sigma_\ell = 0.585$ | $\sigma_\ell = 0.524$ |
| | $L_u = 3.94$ | $L_\ell = 3.32$ | $L_\ell = 2.32$ | $L_\ell = 1.69$ | $L_\ell = 1.47$ |
| 0.9 | $\sigma_u = 0.936$ | $\sigma_\ell = 0.933$ | $\sigma_\ell = 0.926$ | $\sigma_\ell = 0.912$ | $\sigma_\ell = 0.880$ |
| | $L_u = 13.98$ | $L_\ell = 13.41$ | $L_\ell = 12.23$ | $L_\ell = 10.17$ | $L_\ell = 7.52$ |

characteristics $\sigma_\ell$, $L_\ell$ and $\sigma_u$, and $L_u$ for the cases in Table I with $c^2 = 2$. The associated maximum relative errors for $\rho = 0.5$, 0.7, and 0.9 are given in Table I. These will be compared with other parameter specifications in the following sections.

## III. THE SECOND PARAMETER: VARIANCE VERSUS PEAKEDNESS

The first two moments are natural parameters if two parameters are to be used to partially characterize an interarrival-time or a service-time distribution, but it is not clear that these are the best two parameters. Of course, the chosen parameters should be easy to estimate and easy to use in approximations for queues. Also, the parameters should have power determining descriptive queue characteristics; i.e., there should be a small *MRE* or a small range of possible values of $L$. In this regard, Eckberg[2] has shown that the peakedness of a renewal arrival process is a much better second parameter in addition to the mean than the variance for GI/M/k loss systems and also, to some extent, for GI/M/k delay systems. The peakedness is the ratio of the variance to the mean of the steady-state number of busy servers in an associated GI/M/∞ system; see Holtzman,[1] Eckberg,[22] and references there. Knowing the peakedness of a renewal process, say $z$, is equivalent to knowing $\phi(\mu)$, the transform evaluated at the service rate $\mu$:

$$\phi(\mu) = 1 - (z + \lambda/\mu)^{-1}. \tag{10}$$

The peakedness is an important parameter to consider because it is often available as an approximate characterization of overflow processes via the equivalent random method.[1,22] Since Eckberg's results[2] suggest that the mean and the parameter $\phi(\mu)$ might be much better than the mean and variance, we investigate this new parameter pair here.

However, before examining this new parameter pair, we explain why the variance might be a better second parameter for single-server delay systems. Knowing the mean and variance (i.e., $c^2$) is equivalent to knowing the first two derivatives of the transform $\phi(s)$ at 0. It is intuitively reasonable that we might pin down the transform $\phi(s)$ better by fixing the value at $\mu$, $\phi(\mu)$ than by fixing the second derivative at 0, $\phi''(0)$. However, this depends on the way the queue characteristics depend on the transform. For the GI/M/k loss system, the relevant parameters are $\phi(j\mu)$ for $j = 1, 2, \cdots, k$, with the parameters tending to be of less importance as $j$ increases. These parameters are values of the transform $\phi(s)$ evaluated at points $s$ such that $s \geqslant \mu$. For approximations, it is clearly better to specify $\phi(\mu)$ and $\phi'(0)$ than $\phi''(0)$ and $\phi'(0)$.

For the GI/M/1 delay system the key parameter in (2) is the transform value $\phi[\mu(1 - \sigma)]$. Of course, we do not know $\sigma$ in advance, but the argument is always less than $\mu$. Since $\sigma$ tends to be near $\rho$, the argument tends to be near $\mu(1 - \rho)$. Clearly, for large $\rho$, knowing $\phi''(0)$ should be better than knowing $\phi(\mu)$. On the other hand, for small $\rho$, knowing $\phi(\mu)$ should be better than knowing $\phi''(0)$.

Our results substantiate this intuitive reasoning. In marked contrast to GI/M/k loss systems, for GI/M/1 delay systems the parameter $\phi(\mu)$ is not uniformly better than the variance as a second parameter. Which second parameter is better depends on the traffic intensity, with the variance improving as $\rho$ increases. Consistent with the intuitive discussion above, we shall show that asymptotic behavior of the maximum relative error as $\rho$ approaches 0 and 1 is strikingly different given $\phi(\mu)$ instead of $c^2$. Moreover, the variance does better for the upper bound, whereas $\phi(\mu)$ does better for the lower bound.

It is also appropriate to mention that we are considering the peakedness of the renewal arrival process as a single parameter, which by (10) can be represented as the transform $\phi$ evaluated at the service rate $\mu$. If, instead, we knew the peakedness as a function of the service rate as in Eckberg,[22] then we would know the entire transform, which is equivalent to knowing the entire interarrival-time distribution. Moreover, if we could choose one argument of the transform, then we obviously could do better by picking a value less than $\mu$. For example, there would be no error for the GI/M/1 queue if we could guess $\sigma$ and make the argument $\mu(1 - \sigma)$. If we could choose one argument given only the arrival rate and service rate, then a natural choice would be $\mu(1 - \rho)$. (This parameter is considered here in Section IV.) In applications, however, we typically have no choice. Then the arrival process (which may not be renewal) may be partially characterized (by the equivalent random method and related techniques) by rate and peakedness. Moreover, the given peakedness might be with respect to a different service rate (or even a different service-time distribution[22]). In the context of the GI/M/1 queue, this peakedness parameter will lead to better approximations if the argument of the transform, after using (10), is close to $\mu(1 - \sigma)$. The parameter $\phi(\mu)$ considered here should give some idea about what will happen in general.

The new parameter pair involving $\phi(\mu)$ leads to new two-point extremal distributions and a new partial ordering of the distributions. Now consider the set of all probability distributions on the interval $[0, bm_1]$, $b \leqslant \infty$, having first moment $m_1$ and transform $\phi(\mu)$ at $s = \mu$ (and no mass at infinity). This is a convex set depending on the parameters b, $m_1$, and $\phi(\mu)$. The extremal distributions here are the *upper bound*, which is the two-point distribution with mass $p = (b - 1)/(b - x)$ on $xm_1$ and mass $1 - p$ on $bm_1$, where $x$ satisfies

$$pe^{-x/\rho} + (1 - p)e^{-b/\rho} = \phi(\mu); \tag{11}$$

and lower bound, which is the two-point distribution with mass $1 - x^{-1}$ on 0 and mass $x^{-1}$ on $xm_1$, where

$$x = (1 - \rho^{-x/\rho})/(1 - \phi(1/\rho)). \tag{12}$$

Unlike Section II, the upper bound here depends on $b$ while the lower bound does not. As $b \longrightarrow \infty$, the upper bound converges in law to a *limiting upper bound*, which is the one-point distribution with mass 1 on $-(\log \phi(\mu))/\mu$. Note that the limiting upper bound is not actually in the reference set because the mean is not $m_1$. These distributions are minimal and maximal elements for another partial ordering of the distributions based on the transform.

*Definition 2:* *$F_1 \leqslant_\mu F_2$ for two cdf's on $[0, \infty)$ if*

$$\phi_1(s) \leqslant \phi_2(s), \qquad s \leqslant \mu, \qquad \textit{and} \quad \phi_1(s) \geqslant \phi_s(s), \qquad s \geqslant \mu.$$

Let $\mathbf{G} = \mathbf{G}(m_1, \mu, \phi(\mu), b)$ be the set of all cdf's with parameters $m_1$, $\mu$, $\phi(\mu)$, and b. Without loss of generality, let $m_1 = 1$. Let $G_{\mathscr{u}}$, $G_{\ell}$, and $G_{\hat{\mathscr{u}}}$ be the cdf's associated with the special extremal distributions. From Section 2.2.3 and (5) of Eckberg,[2] we obtain

*Proposition 4:* *For all $G \in \mathbf{G}$, $G_{\ell} \leqslant_\mu G \leqslant_\mu G_{\mathscr{u}} \leqslant_\mu G_{\hat{\mu}}$.*

It is easy to see the effect of changing b:

*Proposition 5:* *$G_{\mathscr{u}}$ increases in $\leqslant_\mu$ as b increases and $\phi_{\mathscr{u}}(s) \longrightarrow \phi_{\hat{\mathscr{u}}}(s)$ for each s as $b \longrightarrow \infty$.*

Here are the implications for the GI/M/1 queue. A tilde is used to indicate that the extremal distributions are from this section (because we want to relate them to those in Section II).

*Proposition 6:* *For all $G \in \mathbf{G}$,*

$$\tilde{\sigma}_{\ell} \leqslant \sigma \leqslant \tilde{\sigma}_{\mathscr{u}} \leqslant \tilde{\sigma}_{\hat{\mathscr{u}}} \quad \textit{and} \quad \tilde{L}_{\ell} \leqslant L \leqslant \tilde{L}_{\mathscr{u}} \leqslant \tilde{L}_{\hat{\mathscr{u}}}\,.$$

Using the same change of variables argument as in Theorem 2, we can express $\tilde{\sigma}_{\ell}$ in terms of $\sigma_{\hat{\ell}}$.

*Theorem 6:* $\tilde{\sigma}_{\ell} = 1 - (1 - \sigma_{\hat{\ell}})/x$ *for* $x$ *in* (12).

*Remarks:* As a consequence of Theorem 6, $\tilde{L}_{\ell} = xL_{\hat{\ell}}$ for $x$ in (12). Since $x^{-1}$ is a probability, $\sigma_{\hat{\ell}} \leqslant \tilde{\sigma}$ and $L_{\hat{\ell}} \leqslant L_{\ell}$. Moreover, $\tilde{\sigma}_{\ell}$ and $\tilde{L}_{\ell}$ are decreasing in $\phi(\mu)$ for fixed $\mu$ and $\rho$. Finally, we can combine Theorems 2 and 6 to obtain $\tilde{\sigma}_{\ell} \leqslant \sigma_{u}$ and $\tilde{L}_{\ell} \leqslant L_{u}$; use the fact that $x^{-1} \leqslant 1 \leqslant 1 + c^2$.

We now consider the upper-bound characteristic $\tilde{\sigma}_{u}$. Let $\sigma_{\tilde{\ell}}(\rho)$ be the limiting lower bound in Section II as a function of $\rho$.

*Theorem 7: If* $\phi(\mu) \geqslant e^{-1}$, *then the GI/M/*1 *queue based on* $G_{\hat{\mu}}$ *is unstable and* $\tilde{\sigma}_{\hat{\mu}} = 1$ *is the only root. If.*$\phi(\mu) < e^{-1}$, *then*

$$\tilde{\sigma}_{\hat{\mu}} = \sigma_{\hat{\ell}}\,(-1/\mathit{log}\ \phi(\mu)). \tag{13}$$

*Remarks:* As a consequence of Theorems 1 and 7, if $\phi(\mu) < e^{-1}$, then $\tilde{\sigma}_{\hat{\mu}} < -\log \phi(\mu)$ and $\tilde{\sigma}_{\hat{\mu}}$ is increasing in $\phi(\mu)$.

Paralleling Theorem 7, we have (omitting the proof)

*Theorem 8:* (*a*) *The characteristics* $\tilde{\sigma}_{u}$ *and* $\tilde{L}_{u}$ *are decreasing in* $\phi(\mu)$ *and increasing in b.* (*b*) *MRE*$(\rho, \mu, \phi(\mu), b)$ *is increasing in b.*

We now consider limits as the traffic intensity $\rho$ approaches 0 and 1. Here we assume the transform is based on a fixed interarrival-time cdf and that $\rho$ changes by changing $\mu$.

*Theorem 9: As* $\rho \longrightarrow 1$ $(\mu \longrightarrow 1)$, $\tilde{\sigma}_{\mu} \longrightarrow 1$ *and* $\tilde{\sigma}_{\ell} \longrightarrow \tilde{\sigma}_{\ell}(1) < 1$, *where* $\tilde{\sigma}_{\ell}(1)$ *is the root* $\sigma$ *in* (0, 1) *of*

$$1 - 1/x + (1/x)e^{-(1-\sigma)x} = \sigma \tag{14}$$

*and*

$$x = (1 - e^{-x})/(1 - \phi(1)). \tag{15}$$

*Proof:* For $\tilde{\sigma}_{u}$, use Theorem 5 and the fact that $\sigma_{\ell} \leqslant \tilde{\sigma}_{u}$. For the lower bound, note that $\mu \longrightarrow 1$ and $\phi(\mu) \longrightarrow \phi(1)$ as $\rho \longrightarrow 1$, so that $x$ in (12) approaches (15) and eq. (2) approaches (14).

*Corollary 3: As* $\rho \longrightarrow 1$, *MRE*$(\rho, \mu, \phi(\mu), b) \longrightarrow \infty$.

We have not yet been able to treat all cases when $\rho \longrightarrow 0$. Several possibilities are covered by the next theorem.

*Theorem 10: If* $\rho \longrightarrow 0$ $(\mu \longrightarrow \infty)$, *then* (*a*) $\tilde{\sigma}_{\ell} \longrightarrow 0$; (*b*) $\tilde{\sigma}_{u} \longrightarrow 0$ *when* $F(\epsilon) = 0$ *for some* $\epsilon > 0$; (*c*) $\tilde{\sigma}_{u} \longrightarrow \tilde{\sigma}_{u}(0)$ *when* $F(0) > 0$, *where* $\tilde{\sigma}_{u}(0)$ *is the root* $\sigma$ *in* (0, 1) *of*

$$((b - 1)/b)^{\sigma} F(0)^{1-\sigma} = \sigma. \tag{16}$$

*Proof:* (a) Use Theorems 6 and 4. Note that $x \longrightarrow 1$ as $\mu \longrightarrow \infty$ for $x$ in (14). (b) Note that $\phi(\mu) \leqslant e^{-\mu\epsilon}$ so $x \geqslant \epsilon/\lambda$ for sufficiently large $\mu$. Hence, from (2), $\tilde{\sigma}_{u} \longrightarrow 0$. (c) Note that $\phi(\mu) \longrightarrow F(0)$ and $e^{-\mu b} \longrightarrow 0$ as $\mu \longrightarrow \infty$, so that $x \to 0$ for $x$ satisfying (11), $x/\rho \to -\log[bF(0)/(b-1)]$ and $\tilde{\sigma}_{u} \to \tilde{\sigma}_{u}(0)$ as claimed.

*Corollary 4: If $\rho \to 0$, then $MRE(\rho, \mu, \phi(\mu), b) \to 0$, when $F(\epsilon) = 0$ for some $\epsilon > 0$ and $MRE(\rho, \mu, \phi(\mu), b) \to a$ for some constant $a > 0$ when $F(0) > 0$.*

In Table III we display the extremal characteristics $\tilde{\sigma}_{\ell}$ and $\tilde{\sigma}_{u}$ and $MRE(\rho, \mu, \phi(\mu), b)$ for four values of $\rho$ and four values of $b$. In each case, the given transform values $\phi(\mu)$, which are also displayed in Table III, are calculated for the prototype distribution used in Part II with $m_1 = 2$ and $c^2 = 2$. Since the mean interarrival time is 2, $\mu = 1/2\rho$.

It is interesting to compare Table III with Table II and the $c^2 = 2$ case of Table I. The main conclusion from Tables I and III is that the $MRE$ is always smaller with $c^2$ than with $\phi(\mu)$. For $\rho = 0.9$ it is smaller by a factor of ten.

From Tables II and III, we see that $\sigma_{u} \leq \tilde{\sigma}_{u}$ in all cases except $\rho = 0.2$ and $b = 5$. Also $\sigma_{\ell}$ tends to be better (bigger) than $\tilde{\sigma}_{\ell}$ as $\rho$ increases and $b$ decreases, but neither characteristic is uniformly better.

From Table III and additional cases, it is apparent that the $MRE$ is quite insensitive to changes in $\rho$, varying very little from $\rho = 0.2$ to $\rho = 0.9$. Table III also shows that $MRE(\rho, \mu, \phi(\mu), b)$ is not monotone in $\rho$. The data suggest the following conjecture.

*Conjecture 3: $MRE(\rho, \mu, \phi(\mu), b)$ is unimodal as a function of $\rho$ with a maximum that increases with $b$ (assuming $\phi(\mu)$ is calculated for a fixed interarrival-time distribution).*

Finally, note that $\phi(\mu) > e^{-1} = 0.3678$ for each $\rho$ in Table III, so the queue based on $G_{\hat{u}}$ is unstable, $\tilde{\sigma}_{u} \to 1$, and $MRE(\rho, \mu, \phi(\mu), b) \to \infty$ as $b \to \infty$.

Table III—The extremal GI/M/1 characteristics and maximum relative error MRE($\rho$, $\mu$ $\phi(\mu)$, $b$) for fixed mean and transform value $\phi(\mu)$ based on the prototype distribution in Part II having mean 2 and $c^2 = 2$ (so that $\mu = 1/2\rho$)

| Traffic Intensity, $\rho$ | Transform Value and Lower Bound | Bound on Interarrival-Time Distribution in Multiples of the Mean | | | |
|---|---|---|---|---|---|
| | | $b = 5$ | $b = 10$ | $b = 20$ | $b = 40$ |
| 0.2 | $\phi(\mu) = 0.377$<br>$\tilde{\sigma}_{\ell} = 0.381$ | $\tilde{\sigma}_{u} = 0.607$<br>$MRE = 0.573$ | $\tilde{\sigma}_{u} = 0.705$<br>$MRE = 1.10$ | $\tilde{\sigma}_{u} = 0.783$<br>$MRE = 1.85$ | $\tilde{\sigma}_{u} = 0.844$<br>$MRE = 6.81$ |
| 0.5 | $\phi(\mu) = 0.466$<br>$\tilde{\sigma}_{\ell} = 0.563$ | $\tilde{\sigma}_{u} = 0.737$<br>$MRE = 0.664$ | $\tilde{\sigma}_{u} = 0.832$<br>$MRE = 1.61$ | $\tilde{\sigma}_{u} = 0.900$<br>$MRE = 3.38$ | $\tilde{\sigma}_{u} = 0.944$<br>$MRE = 6.81$ |
| 0.7 | $\phi(\mu) = 0.518$<br>$\tilde{\sigma}_{\ell} = 0.648$ | $\tilde{\sigma}_{u} = 0.834$<br>$MRE = 0.648$ | $\tilde{\sigma}_{u} = 0.898$<br>$MRE = 1.68$ | $\tilde{\sigma}_{u} = 0.942$<br>$MRE = 3.74$ | $\tilde{\sigma}_{u} = 0.969$<br>$MRE = 7.72$ |
| 0.9 | $\phi(\mu) = 0.562$<br>$\tilde{\sigma}_{\ell} = 0.905$ | $\tilde{\sigma}_{u} = 0.942$<br>$MRE = 0.625$ | $\tilde{\sigma}_{u} = 0.965$<br>$MRE = 1.69$ | $\tilde{\sigma}_{u} = 0.981$<br>$MRE = 3.84$ | $\tilde{\sigma}_{u} = 0.990$<br>$MRE = 8.15$ |

## IV. ADDITIONAL PARAMETER SPECIFICATIONS

We now consider several other parameters in addition to the first two moments $[m_1, m_2]$ and the mean and the transform value $[m_1, \phi(\mu)]$. We consider two different three-parameter specifications: the first three moments $[m_1, m_2, m_3]$, and the first two moments and the transform value $[m_1, m_2, \phi(\mu)]$. We also consider two-parameter specifications involving the transform value $\phi(\mu(1-\rho))$, combining it with the mean and $\phi(\mu)$. Each parameter specification is considered with and without an upper bound on the distribution.

In each case the extremal distributions can be obtained from the theory of complete Tchebycheff systems by solving systems of equations. The general formulas for the extremal distributions are either displayed explicitly in Eckberg[2] or can easily be obtained from the theory there.

To obtain the parameter values themselves, we use the two prototype distributions described in Section II of Part II.[20] Prototype I is more variable with $c^2 = 2.0$ and Prototype II is less variable with $c^2 = 0.8$. We also consider two values of the traffic intensity; $\rho = 2/3$ and $\rho = 9/10$. Finally, we consider both an upper bound of 20 on the distribution and no upper bound. Since the means for Prototypes I and II are 2.0 and 4.0, respectively, the upper bounds are $b = 10$ and $b = 5$ times the mean, respectively. The value 20 was chosen for the bound to be consistent with the prototype distributions. All the prototype parameter values are given in Tables IV and V. The extremal probability distributions themselves are displayed in Tables VI through IX. These are probability mass functions with all mass on one, two, or three points. The points are often the distribution boundary points 0 and 20. In the case of two transform values $\{\phi(\mu), \phi[\mu(1-\rho)]\}$ the distribution is defective (positive mass at infinity) in the lower bound for Prototype I and the upper bound for Prototype II.

The following is a list of conclusions drawn from the numerical results in Tables IV through IX. These conclusions represent clear tendencies indicated by these (and other) data, but they are not theorems. For example, with respect to the results in Section II, the first conclusion is supported in part by Corollary 1, but is limited by the observation before Conjecture 2.

1. For all parameter specifications, the *MRE* is much less with less variability; it is much less in Table V with $c^2 = 0.8$ than in Table IV with $c^2 = 2.0$.

2. As noted in Section II, two moments and a bound on the distribution are sufficient for approximations with high traffic intensities (here $MRE \leqslant 8$ percent for $\rho = 0.9$), but not for all traffic intensities.

3. An extra moment helps significantly. Three moments and a bound are good enough for approximations in all cases ($MRE \leqslant 10$

Table IV—Extremal characteristics and maximum relative errors for the GI/M/1 queue with various parameter specifications: Case of Prototype I (mean = 2, $c^2 = 2$)

| Given Parameter Values | $\rho = 0.667$ | | | | | | $\rho = 0.900$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $b = \infty$ | | | $b = 10$ (times the mean) | | | $b = \infty$ | | | $b = 10$ (times the mean) | | |
| | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* |
| $m_1, \phi(\mu)$ | 0.698 | 1.000 | ∞ | 0.698 | 0.887 | 1.67 | 0.900 | 1.000 | ∞ | 0.900 | 0.965 | 1.86 |
| $m_1, \phi(\mu(1-\rho))$ | 0.754 | 1.000 | ∞ | 0.754 | 0.802 | 0.242 | 0.931 | 1.000 | ∞ | 0.931 | 0.935 | 0.062 |
| $\phi(\mu), \phi(\mu(1-\rho))$ | 0.730 | 0.793 | 0.304 | 0.747 | 0.793 | 0.222 | 0.908 | 0.945 | 0.673 | 0.918 | 0.945 | 0.491 |
| $m_1, m_2$ | 0.417 | 0.806 | 2.00 | 0.645 | 0.806 | 0.830 | 0.807 | 0.936 | 2.00 | 0.926 | 0.936 | 0.063 |
| $m_1, m_2, m_3$ | 0.754 | 0.806 | 0.268 | 0.754 | 0.776 | 0.098 | 0.932 | 0.936 | 0.063 | 0.932 | 0.933 | 0.015 |
| $m_1, m_2, \phi(\mu)$ | 0.698 | 0.787 | 0.418 | 0.747 | 0.787 | 0.187 | 0.900 | 0.934 | 0.515 | 0.931 | 0.934 | 0.046 |

Characteristics of Prototype I: $m_1 = 2.00$, $m_2 = 12.00$, $c^2 = 2.00$, $m_3 = 119.01$; the upper bound on the distribution $b$ is in multiples of the mean. $\rho = 0.667$: $\mu = 0.750$, $\phi(\mu) = 0.5098$, $\phi(\mu(1-\rho)) = 0.7073$, $\sigma = 0.7676$; $\rho = 0.900$: $\mu = 0.555$, $\phi(\mu) = 0.5615$, $\phi(\mu(1-\rho)) = 0.9046$, $\sigma = 0.9324$.

Table V—Extremal characteristics and maximum relative errors for the GI/M/1 queue with various parameter specifications: Case of Prototype II (mean = 4, $c^2 = 0.8$)

| Given Parameter Values | $\rho = 0.667$ | | | | | | $\rho = 0.900$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $b = \infty$ | | | $b = 5$ (times the mean) | | | $b = \infty$ | | | $b = 5$ (times the mean) | | |
| | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* | $\sigma_\ell$ | $\sigma_u$ | *MRE* |
| $m_1, \phi(\mu)$ | 0.602 | 1.000 | ∞ | 0.602 | 0.733 | 0.491 | 0.868 | 1.000 | ∞ | 0.868 | 0.920 | 0.650 |
| $m_1, \phi(\mu(1-\rho))$ | 0.638 | 1.000 | ∞ | 0.638 | 0.645 | 0.020 | 0.889 | 1.000 | ∞ | 0.889 | 0.890 | 0.009 |
| $\phi(\mu), \phi(\mu(1-\rho))$ | 0.640 | 0.650 | 0.031 | 0.640 | 0.648 | 0.023 | 0.888 | 0.898 | 0.098 | 0.888 | 0.893 | 0.047 |
| $m_1, m_2$ | 0.417 | 0.676 | 0.799 | 0.571 | 0.676 | 0.324 | 0.807 | 0.893 | 0.804 | 0.885 | 0.893 | 0.075 |
| $m_1, m_2, m_3$ | 0.637 | 0.676 | 0.120 | 0.637 | 0.650 | 0.037 | 0.890 | 0.893 | 0.028 | 0.890 | 0.890 | ≈0 |
| $m_1, m_2, \phi(\mu)$ | 0.602 | 0.651 | 0.140 | 0.631 | 0.651 | 0.057 | 0.868 | 0.891 | 0.211 | 0.888 | 0.891 | 0.019 |

Characteristics of Prototype II: $m_1 = 4.00$, $m_2 = 28.80$, $c^2 = 0.80$, $m_3 = 279.83$; $\rho = 0.667$; $\mu = 0.375$, $\phi(\mu) = 0.3881$, $\phi(\mu(1-\rho)) = 0.6589$, $\sigma = 0.6429$; $\rho = 0.900$: $\mu = 0.278$, $\phi(\mu) = 0.4613$, $\phi(\mu(1-\rho)) = 0.8991$, $\sigma = 0.8901$.

Table VI—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype I ($m_1 = 2$, $c^2 = 2$) with $\rho = 2/3$

| Given Parameter Values | Extremal Characteristics | Extremal Probability Mass Function, Mass $p_k$ on $x_k$ | | | | | |
|---|---|---|---|---|---|---|---|
| Upper Bounds | $\sigma_u$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, \phi(\mu)]$ | 1.000 | 1.000 | 0.90 | — | — | — | — |
| $[m_1, \phi(\mu), b]$ | 0.887 | 0.9381 | 0.81 | 0.0619 | 20.00 | — | — |
| $[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$ | 0.806 | 0.6667 | 0.00 | 0.3333 | 6.00 | — | — |
| $[m_1, \phi(\mu(1-\rho)), b]$ | 0.802 | 0.9583 | 1.22 | 0.0417 | 20.00 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.793 | 0.4565 | 0.00 | 0.5435 | 3.09 | — | — |
| $[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$ | 0.787 | 0.8060 | 0.61 | 0.1940 | 7.77 | — | — |
| $[m_1, m_2, m_3, b]$ | 0.776 | 0.5760 | 0.00 | 0.4132 | 4.32 | 0.0107 | 20.00 |
| Lower Bounds | $\sigma_\ell$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$ | 0.754 | 0.906 | 1.09 | 0.094 | 10.79 | — | — |
| $[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$ | 0.754 | 0.5787 | 0.00 | 0.4213 | 4.75 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.747 | 0.8311 | 0.59 | 0.1689 | 20.00 | — | — |
| $[m_1, m_2, \phi(\mu), b]$ | 0.747 | 0.4628 | 0.00 | 0.5208 | 3.20 | 0.0167 | 20.00 |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ | 0.730 | 0.8330 | 0.65 | — | — | — | — |
| $[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$ | 0.698 | 0.4810 | 0.00 | 0.5190 | 3.85 | — | — |
| $[m_1, m_2, b]$ | 0.645 | 0.9759 | 1.56 | 0.0241 | 20.00 | — | — |
| $[m_1, m_2]$ | 0.417 | 1.000 | 2.00 | — | — | — | — |

percent). The third moment reduces the *MRE* by approximately a factor of 10.

4. The upper bound on the distribution can make a big difference. It matters when the extremal distribution has mass on the upper bound, which occurs for either the upper or lower extremal distribution but not for both.

5. As noted in Section III, overall the second moment is better than the transform value $\phi(\mu)$ as a second parameter in addition to the mean. However, for lower traffic intensities and no bound on the distribution, $\phi(\mu)$ is better for the lower bound. The second moment is always better for the upper bound. Similarly, the third moment is always better than the transform value $\phi(\mu)$ as a third parameter in addition to the first two moments. However, for lower traffic intensities and no bound on the distribution, $\phi(\mu)$ is better for the upper bound.

6. The transform value $\phi(\mu(1 - \rho))$ is always better than the transform value $\phi(\mu)$ since $\phi(\mu(1 - \rho))$ is closer to $\mu(1 - \sigma)$. Even

Table VII—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype I ($m_1 = 2$, $c^2 = 2$) with $\rho = 0.9$

| Given Parameter Values | Extremal Characteristics | Extremal Probability Mass Function, Mass $p_k$ on $x_k$ | | | | | |
|---|---|---|---|---|---|---|---|
| Upper Bounds | $\sigma_u$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[\mathrm{m}_1, \phi(\mu)]$ | 1.000 | 1.000 | 1.04 | — | — | — | — |
| $[\mathrm{m}_1, \phi(\mu), b]$ | 0.965 | 0.9442 | 0.94 | 0.0558 | 20.00 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.945 | 0.5024 | 0.00 | 0.4976 | 3.83 | — | — |
| $[m_1, m_2]$, $[m_1, m_2, m_3]$ and $[m_1, m_2, b]$ | 0.936 | 0.6667 | 0.00 | 0.3333 | 6.00 | — | — |
| $[m_1, \phi(\mu(1-\rho)), b]$ | 0.935 | 0.9716 | 1.47 | 0.0284 | 20.00 | — | — |
| $[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$ | 0.934 | 0.8060 | 0.61 | 0.1940 | 7.76 | — | — |
| $[m_1, m_2, m_3, b]$ | 0.933 | 0.5760 | 0.00 | 0.4132 | 4.32 | 0.0107 | 20.00 |
| Lower Bounds | $\sigma_\ell$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$ | 0.932 | 0.906 | 1.09 | 0.094 | 10.79 | — | — |
| $[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$ | 0.931 | 0.6438 | 0.00 | 0.3562 | 5.62 | — | — |
| $[m_1, m_2, \phi(\mu), b]$ | 0.931 | 0.484 | 0.00 | 0.500 | 3.37 | 0.016 | 20.00 |
| $[m_1, m_2, b]$ | 0.926 | 0.9759 | 1.56 | 0.0241 | 20.00 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.918 | 0.9248 | 0.90 | 0.0752 | 20.00 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ | 0.908 | 0.9538 | 0.95 | — | — | — | — |
| $[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$ | 0.900 | 0.4812 | 0.00 | 0.5188 | 3.855 | — | — |
| $[m_1, m_2]$ | 0.807 | 1.000 | 2.00 | — | — | — | — |

better is $\phi(\mu(1 - \hat{\sigma})$, where $\hat{\sigma}$ is the Kraemer and Langenbach-Belz[23] approximation for the root $\sigma$. The parameters $\phi(\mu(1 - \rho)$ and $\phi(\mu(1 - \hat{\sigma}))$ do not appear very useful, however, because if it is possible to calculate them, it should also be easy to calculate the root $\sigma$ itself. On the other hand, an approximation for $\phi(\mu)$ might be available from the peakedness without knowing the distribution or even without actually having a renewal process. Given the peakedness $z$, we obtain $\phi(\mu)$ for a renewal process from (10).

7. For each parameter specification, one bound (either the upper or the lower) is "soft" and the other is "hard"; the soft bound can be greatly improved by adding an additional parameter, while the hard bound cannot. The hard bound also tends to be much better than the soft bound. For example, consider the parameter pair $[m_1, m_2]$. The lower bound is soft because it can be improved substantially by specifying $b$ or $m_3$. On the other hand, the upper bound is hard because no improvement is obtained by specifying $b$ or $m_3$. Moreover, the hard upper bound is clearly much better than the soft lower bound (as

Table VIII—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype II ($m_1 = 4$, $c^2 = 0.8$) with $\rho = 2/3$

| Given Parameter Values | Extremal Characteristics | Extremal Probability Mass Functions, Mass $p_k$ and $x_k$ | | | | | |
|---|---|---|---|---|---|---|---|
| Upper Bounds | $\sigma_u$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, \phi(\mu)]$ | 1.000 | 1.000 | 2.52 | — | — | — | — |
| $[m_1, \phi(\mu), b]$ | 0.733 | 0.9012 | 2.25 | 0.988 | 20.00 | — | — |
| $[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$ | 0.676 | 0.444 | 0.00 | 0.556 | 7.20 | — | — |
| $[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$ | 0.651 | 0.6886 | 1.60 | 0.3114 | 9.32 | — | — |
| $[m_1, m_2, m_3, b]$ | 0.650 | 0.3571 | 0.00 | 0.6229 | 5.78 | 0.0199 | 20.00 |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ | 0.650 | 0.8585 | 2.12 | — | — | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.648 | 0.8317 | 2.03 | 0.1683 | 20.00 | — | — |
| $[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$ | 0.645 | 0.3908 | 0.00 | 0.6092 | 6.57 | — | — |
| Lower Bounds | $\sigma_\ell$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.640 | 0.2869 | 0.00 | 0.7131 | 5.21 | — | — |
| $[m_1, \phi(\mu(1-\rho)), b]$ | 0.638 | 0.9329 | 2.85 | 0.0671 | 20.00 | — | — |
| $[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$ | 0.637 | 0.7811 | 2.11 | 0.2189 | 10.75 | — | — |
| $[m_1, m_2, \phi(\mu), b]$ | 0.631 | 0.2783 | 0.00 | 0.6913 | 4.91 | 0.0304 | 20.00 |
| $[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$, and $[m_1, m_2, \phi(\mu)]$ | 0.602 | 0.3094 | 0.00 | 0.6906 | 5.79 | — | — |
| $[m_1, m_2, b]$ | 0.571 | 0.9524 | 3.20 | 0.0476 | 20.00 | — | — |
| $[m_1, m_2]$ | 0.417 | 1.000 | 4.00 | — | — | — | — |

measured by the distance from the actual value of the prototype distribution). Similarly, for the pair $[m_1, \phi(\mu)]$, the upper bound is soft and the lower bound is hard. Of course, all these bounds are tight: they can either be attained for a given distribution or, for any $\epsilon > 0$, the bound can be attained within $\epsilon$ by a given distribution. This notion of limiting tightness is needed, for example, for the lower bound when specifying $[m_1, m_2]$.

## V. OTHER MODELS

We have used the GI/M/1 model to study extremal distributions because the model is analytically tractable and because we believe that similar results will hold for more complicated systems. For example, Bergmann et al.[9] have shown that the variance and higher cumulants of the equilibrium delay in a GI/G/1 system, given the first two moments of the interarrival time and service time, are maximized and minimized using the extremal distributions in Section II for the

Table IX—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype II ($m_1 = 4$, $c^2 = 0.8$) with $\rho = 0.9$

| Given Parameter Values | Extremal Characteristics | Extremal Probability Mass Function, Mass $p_k$ and $x_k$ | | | | | |
|---|---|---|---|---|---|---|---|
| Upper Bounds | $\sigma_{\mathcal{U}}$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, \phi(\mu)]$ | 1.000 | 1.000 | 2.78 | — | — | — | — |
| $[m_1, \phi(\mu), b]$ | 0.920 | 0.9120 | 2.45 | 0.0880 | 20.00 | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ | 0.898 | 0.9683 | 2.67 | — | — | — | — |
| $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.893 | 0.8999 | 2.41 | 0.1001 | 20.00 | — | — |
| $[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$ | 0.893 | 0.4440 | 0.00 | 0.5560 | 7.20 | — | — |
| $[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$ | 0.891 | 0.6886 | 1.60 | 0.3114 | 9.32 | — | — |
| $[m_1, \phi(\mu(1-\rho)), b]$ | 0.890 | 0.4319 | 0.00 | 0.5681 | 7.04 | — | — |
| $[m_1, m_2, m_3, b]$ | 0.890 | 0.3571 | 0.00 | 0.6229 | 5.78 | 0.0199 | 20.00 |
| Lower Bounds | $\sigma_{\mathcal{L}}$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| $[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$ | 0.890 | 0.7810 | 2.11 | 0.2190 | 10.75 | — | — |
| $[m_1, \phi(\mu(1-\rho)), b]$ | 0.889 | 0.9480 | 3.12 | 0.0520 | 20.00 | — | — |
| $[m_1, m_2, \phi(\mu), b]$ | 0.888 | 0.2978 | 0.00 | 0.6740 | 5.09 | 0.0282 | 20.00 |
| $[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$ | 0.888 | 0.3307 | 0.00 | 0.6693 | 5.88 | — | — |
| $[m_1, m_2, b]$ | 0.885 | 0.9524 | 3.20 | 0.0476 | 20.00 | — | — |
| $[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$ | 0.868 | 0.3163 | 0.00 | 0.6837 | 5.85 | — | — |
| $[m_1, m_2]$ | 0.807 | 1.000 | 4.00 | — | — | — | — |

interarrival times and service times. Using $F_{\mathcal{U}}$ for the interarrival time and $F_{\mathcal{L}}$(actually the limit as $b \to \infty$) for the service time yields the maximum, while the reverse yields the minimum. As a consequence, Daley conjectured that related extremal properties held for the mean delay (or, equivalently, the mean queue length); see Bergmann et al.,[9] Open Problem 5.2.4 at the end of Section V in Stoyan,[12] and Daley and Trengove.[24] In particular, Daley conjectured that for GI/G/1 queues with the first two moments of the interarrival and service times given, the steady-state mean queue length $L$ would be maximized and minimized using the extremal distributions in Section II for the interarrival-time and service-time distributions. Moveover, these extremal properties should still hold if only one of the distributions is allowed to vary, and the other is fixed arbitrarily.

Unfortunately, we now know that neither part of this conjecture is correct in general, but the principle does apply for some systems. Of course, the GI/M/1 results in Section II are consistent with the conjecture. Daley and Trengove[24] showed that the limiting extremal distribution $F_{\ell}$ for the interarrival time yields the minimum mean queue length for *all* service-time distributions. Another system consistent with the conjecture is the $K_2$/G/1 queue, which has an interarrival-time distribution with a rational Laplace-Stieltjes transform with a denominator of degree 2; see p. 329 of Cohen.[19] As with the GI/M/1 queue, $L$ depends on a single root of an equation involving the transform in addition to the specified parameters; see (5.205) on p. 330 of Ref. 19. Paralleling the GI/M/1 case, we have

*Theorem 11: For any $K_2/G/1$ queue with fixed interarrival-time distribution and service-time distribution partially specified by the first two moments, L is maximized and minimized by using the extremal distributions in Section II for the service-time distribution.*

We do not give the proof of Theorem 11; related results for $K_2$/G/1 queues are obtained in Whitt[14] and discussed in Part III.[21] However, the analysis there also disproves the part of Daley's conjecture claiming that the same extremal service-time distributions should yield the maximum (minimum) mean queue length for all fixed interarrival-time distributions. The analysis in Whitt[14] shows that the extremal distribution maximizing $L$ depends on the interarrival-time distribution. For example, if the interarrival-time distribution is the convolution of two exponential distributions, then $L$ is minimized by letting the service-time distribution be the upper-bound two-point distribution with mass $c^2/(1 + c^2)$ on 0. On the other hand, if the interarrival-time distribution is the mixture of two exponential distributions, then $L$ is maximized by letting service-time distribution be this upper-bound two-point distribution. (See Section VII of Part III[21] for further discussion.)

We also succeeded in disproving the first part of the conjecture by identifying a service-time distribution that produces a smaller mean queue length than either extremal distribution in Section II for the D/G/1 queue. Since the D arrival process obtained via the limiting extremal distribution $F_{\hat{\ell}}$ was shown by Daley and Trengove[24] to yield the minimum given any service-time distribution, this counterexample applies to the global minimum as well as the minimum given a fixed interarrival distribution. The particular service-time distribution we used for our numerical example had all mass on multiples of the constant interarrival time. Daley (private communication) subsequently observed that recent results of Ott[25] for the D/G/1 queue imply that these special service-time distributions are in fact extremal for the D/G/1 queue.

The extremal distributions for the different parameter specifications in this paper should also be useful to give an indication of the range of possibilities in more complicated models. Even if the extremal distributions here are not actually extreme for the descriptive char-

Table X—The extreme values for the blocking probability in a GI/M/1 loss system, which is the transform value $\phi(\mu)$, given the service rate, $\mu$, and the moments of the interarrival time

| Given Parameter Values | Prototype Distribution I, $c^2 = 2.0$ | | Prototype Distribution II, $c^2 = 0.8$ | |
|---|---|---|---|---|
| Upper Bounds | $\rho = 2/3$ | $\rho = 9/10$ | $\rho = 2/3$ | $\rho = 9/10$ |
| $[m_1, m_2]$, $[m_1, m_2, b]$ and $[m_1, m_2, m_3]$ | 0.670 | 0.678 | 0.481 | 0.519 |
| $[m_1, m_2, m_3, b]$ | 0.592 | 0.613 | 0.428 | 0.482 |
| The actual blocking probability | 0.510 | 0.562 | 0.388 | 0.461 |
| Lower Bounds | $\rho = 2/3$ | $\rho = 9/10$ | $\rho = 2/3$ | $\rho = 9/10$ |
| $[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$ | 0.400 | 0.495 | 0.358 | 0.446 |
| $[m_1, m_2, b]$ | 0.304 | 0.411 | 0.287 | 0.392 |
| $[m_1, m_2]$ | 0.223 | 0.329 | 0.223 | 0.329 |

acteristics of the more complicated model, these distributions should give a good idea of the range for the given parameters.

It should be remembered, however, that the model affects which parameters are most useful. For a central-server closed network of queues, Lazowska[26] found percentiles much better than moments. Our GI/M/1 delay system results are also very different from Eckberg's[2] GI/M/k loss system results. The GI/M/k blocking probability depends on the $k$ parameters $\phi(j\mu)$, $j = 1, 2, \cdots, k$. Hence, all the extremal distributions are extreme for this descriptive characteristic given the various parameter sets. However, $\phi(\mu)$ strongly dominated $m_2$ as a second parameter in addition to the mean.

To make a specific comparison, we consider the GI/M/1 loss system (no waiting room). For this system the blocking probability is just the transform value $\phi(\mu)$. By Proposition 1, the extremal distributions in Sections II through IV are extreme for $\phi(\mu)$. In Table X we display the extreme values of the blocking probability given the first two and first three moments, with and without the upper bound on the distribution. It is evident that the absolute and relative errors for $\phi(\mu)$ are much greater than for $\sigma$ and $L$.

## VI. ACKNOWLEDGMENT

I am grateful to John Klincewicz for writing programs to obtain the data in Tables I through III and for many helpful discussions.

## REFERENCES

1. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," B.S.T.J., *52*, No. 9 (November 1973), pp. 1673–9.
2. A. E. Eckberg, Jr., "Sharp Bounds on Laplace-Stieltjes Transforms, with Applications to Various Queueing Problems," Math. Oper. Res., *2*, No. 2 (May 1977), pp. 135–42.

3. T. Rolski, "Some Inequalities for GI/M/n Queues," Zast. Mat., *13,* No. 1 (1972), pp. 43–7.
4. T. Rolski, "Some Inequalities in Queueing Theory," Colloquia Math. Soc. Janos Bolyai, *9* (1974), pp. 653–9.
5. T. Rolski, "Order Relations in the Set of Probability Distribution Functions and Their Applications in Queueing Theory," *Dissertationes Mathematicae,* Polish Scientific Publishers, Warsaw, 1976.
6. J. G. Shanthikumar and J. A. Buzacott, "On the Approximations to the Single Server Queue," Int. J. Prod. Res., *18,* No. 6 (1980), pp. 761–73.
7. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Letters, *1,* No. 5 (November 1982), pp. 165–9.
8. S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics,* New York: John Wiley and Sons, 1966.
9. R. Bergmann, D. J. Daley, T. Rolski, and D. Stoyan, "Bounds for Cumulants of Waiting-Times in GI/GI/1 Queues," Math. Operationsforsch. Statist., Ser. Optimization, *10,* No. 2 (1979), pp. 257–63.
10. D. J. Daley and T. Rolski, "A Light Traffic Approximation for a Single-Server Queue," Math. Oper. Res., *9* (1984), to be published.
11. A. F. Karr, "Extreme Points of Certain Sets of Probability Measures, with Applications," Math. Oper. Res., *8,* No. 1 (February 1983), pp. 74–85.
12. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models,* New York: John Wiley and Sons, to be published. (English translation edited by D. J. Daley of Qualitative Abschäzungen Stochastischer Modelle, 1977.)
13. W. Whitt, "Untold Horrors of the Waiting Room: What the Equilibrium Distribution Will Never Tell about the Queue Length Process," Management Sci., *29,* No. 4 (April 1983), pp. 395–408.
14. W. Whitt, "Minimizing Delays in a GI/G/1 Queue," Oper. Res., *32* (1984), to be published.
15. W. Whitt, "Approximating a Point Process by a Renewal Process: Two Basic Methods," Oper. Res., *30,* No. 1 (January-Feburary 1982), pp. 125–47.
16. S. L. Albin, *Approximating Queues with Superposition Arrival Process,* Ph.D. dissertation, School of Engineering Science, Columbia University, 1981.
17. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62,* No. 9, Part 1 (November 1983), pp. 2779–2815.
18. W. Whitt, "Performance of the Queueing Network Analyzer," B.S.T.J., *62,* No. 9, Part 1 (November 1983), pp. 2817–43.
19. J. W. Cohen, *The Single Server Queue,* Amsterdam: North-Holland, 1969.
20. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," AT&T Bell Lab. Tech. J., this issue.
21. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," AT&T Bell Lab. Tech. J., this issue.
22. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," Tenth Int. Teletraffic Cong., Montreal, 1983, 4.4b.3.
23. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," Eighth Int. Teletraffic Cong., Melbourne, 1976, pp. 235-1-8.
24. D. J. Daley and C. D. Trengove, "Bounds for Mean Waiting Times in Single-Server Queues: A Survey," Department of Statistics, the Austrialian National University, 1977.
25. T. J. Ott, unpublished work.
26. E. D. Lazowska, "The Use of Percentiles in Modeling CPU Service Time Distributions," *Computer Performance,* eds. K. M. Chandy and M. Reiser, New York: North-Holland, 1977, pp. 53–66.

## AUTHOR

**Ward Whitt,** A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and congestion models.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# On Approximations for Queues, II: Shape Constraints

By J. G. KLINCEWICZ* and W. WHITT*

(Manuscript received April 22, 1983)

This paper continues the investigation begun in Part I of approximations for queues that are based on a few parameters partially characterizing the arrival process and the service-time distribution. Part I provides insight into approximations for intractable systems by considering the set of all possible values of the mean queue length in the GI/M/1 queue given the service rate and the first two moments of the interarrival-time distribution. The distributions yielding the maximum and minimum values of the mean queue length turn out to be quite unusual, e.g., two-point distributions. This paper shows that the range of possible values can be reduced dramatically by imposing realistic shape constraints on the interarrival-time distribution with given first two moments. We found extremal distributions in the presence of shape constraints by restricting our attention to discrete distributions with all mass on a fixed finite set of points and solving nonlinear programs. The results strongly support the use of two-moment approximations in general queueing systems when the interarrival-time and service-time distributions are not too irregular.

## I. INTRODUCTION AND SUMMARY

This paper continues the investigation begun in Part I[1] of the set of possible values of the mean queue length $L$ (number in system) in a GI/M/1 queue given the service rate, $\mu$, and various parameters partially characterizing the interarrival time cdf $F$ (e.g., the first two moments $m_1$ and $m_2$). As explained in Part I, we are not primarily

* AT&T Bell Laboratories.

interested in the GI/M/1 model itself; we wish to provide a basis for evaluating approximations for more complex queueing models such as the nodes in a non-Markov network of queues.[2] For such complex models, the arrival process may be approximated by a renewal process, partially characterized by the first two moments of the renewal interval. Then the GI/M/1 model arises as an approximation and there is no complete interarrival-time distribution for an exact solution. We examine the GI/M/1 queue because it is tractable and because we believe it is indicative of what happens more generally.

For the GI/M/1 queue,[3]

$$L = \rho/(1 - \sigma), \tag{1}$$

where $\rho$ is the traffic intensity ($\rho = 1/\mu m_1$) and $\sigma$ is the unique root in the open interval (0, 1) of the equation

$$\phi[\mu(1 - \sigma)] = \sigma, \tag{2}$$

with $\phi(s)$ the Laplace-Stieltjes transform of the interarrival-time cdf $F$:

$$\phi(s) = \int_0^\infty e^{-st} dF(t). \tag{3}$$

Unfortunately, given $m_1$, $m_2$, and $\mu$, the range of possible values of $L$ can be very wide. (See the example in Section I of Part I.) This wide range naturally raises doubts about the value of two-moment approximations, but the particular distributions yielding the extreme values of $L$ suggest that the approximations may still be useful. As we indicated in Part I, these extremal distributions are discrete probability distributions with positive probability on just two points. These two-point distributions are obviously very unusual. We would hope that for typical (nice) distributions $L$ would not vary much among interarrival-time distributions with the same moments. In this paper, we investigate how much the range is reduced by imposing regularity conditions on the interarrival-time distribution. The regularity conditions we consider are shape constraints such as unimodality and log-convexity (a natural smoothness condition; see Chapter 5 of Keilson[4] and Section II).

A major contribution here, we believe, is the method. To study the effect of the shape constraints, we restrict attention to discrete distributions with all mass on a fixed finite set of points. We then find the range of the mean queue length $L$ by means of nonlinear programming.

Since typical interarrival-time distributions are smooth (have densities), some may distrust results based on discrete distributions. However, continuity theorems show that there is no loss of generality, at least in principle, in considering distributions concentrating on a fixed finite set of points; see Section 11 of Borovkov.[5] With enough points, such discrete distributions can be used to approximate an arbitrary interarrival-time distribution arbitrarily well (in the usual sense of convergence in distribution and convergence of moments). In turn, the queue-length distribution and the mean queue length $L$ associated with finite-valued probability mass functions can be used to approximate the queue-length distribution and the mean queue length $L$ associated with the arbitrary interarrival-time distribution.

The point is that we need not worry about the local behavior of the interarrival-time distribution. For sufficiently small positive $\epsilon$, if we change an interarrival-time density, say $f(t)$, only on the interval $[t_0, t_0 + \epsilon]$, for example, by making

$$f_n(t) = \begin{cases} f(t) & t \notin [t_0, t_0 + \epsilon] \\ nf[t_0 + n(t - t_0)], & t_0 \leq t \leq t_0 + \epsilon/n \\ 0 & t_0 + \epsilon/n < t \leq t_0 + \epsilon, \end{cases}$$

then the new density $f_n(t)$ will be very different from the density $f(t)$ on $[t_0, t_0 + \epsilon]$ for large $n$, but the associated cdf's will be close and the behavior of the associated queueing systems will be virtually indistinguishable.

While there is no loss in generality in restricting attention to discrete distributions, it is not clear how many points are enough and where they should be located. We have not made a systematic investigation of this question, but we believe that we have used enough points in our study. It is important to recognize that extra points are not free because the nonlinear programs typically become harder to solve.

Throughout this paper, we use 21 points on the integers $\{0, 1, 2, \cdots, 20\}$. By comparing the programming results without shape constraints here with the theoretical results based on the complete Tchebycheff systems in Part I, we can see the effect of the discreteness. This effect can be seen in Tables II and V. The upper bound 20 on the support of the distribution (which might not be regarded as an essential aspect of the discreteness) can have a significant impact, but otherwise the discreteness matters little.

Do the shape constraints help? For the GI/M/1 example, with $\rho = 2/3$ and $c^2 = 2$, assuming a log-convex probability mass function reduces the maximal possible error in $L$ from 200 percent to 8 percent. If the third moment is fixed as well, the maximal possible error is less than 1 percent.

These results indicate that two-moment approximations can be very useful, provided the interarrival-time distribution is actually not unusually irregular. In this paper we only study the GI/M/1 queue, but we believe the results are indicative of what happens in GI/G/1 queues and more general systems.

On the other hand, even for the GI/M/1 queue, the results do not imply that the two-moment approximations will work well in all circumstances or that they should be used blindly. If it is known that the interarrival-time distribution has an unusual shape, then the approximation should probably be modified. If additional information is known that would permit working with a third parameter such as the third moment or the peakedness, then better results can be expected. As noted by Kuczura[6] in a related context, a third parameter seems to offer the possibility of significant improvement, but additional parameters are rarely worth the effort.

This paper is organized as follows. In Section II, we define prototype distributions, introduce the shape constraints, and formulate the mathematical programs. The prototype distributions are intended to be typical interarrival-time distributions, which we use to generate parameter values and the "exact" queue characteristics $\sigma$ and $L$. In Section III, we discuss the computational results for shape constraints with the first two moments fixed. In Section IV, we discuss the results for shape constraints with other parameters fixed. In Section V, we compare our results to other bounds and approximations. Finally, in Section VI, we discuss mathematical programming issues. It turns out that solving the nonlinear programs was not routine. These queueing problems may be interesting test problems for nonlinear programming codes.

We conclude this introduction by mentioning an interesting outcome of our experiments. Unlike Part I, the approach here is primarily numerical, being based on nonlinear programs, but the extremal distributions yielding the minimum and maximum values of $L$ obtained from the nonlinear programs exhibit regularity that suggests the possibility of an analytic treatment similar to Part I. The extremal distributions we obtain on the set $\{0, 1, \cdots, 20\}$ have special structure and evidently do not depend on the traffic intensity. Hence, it may be possible to obtain analytic characterizations; this is a promising direction of research. (In fact, an analytic approach to shape constraints is carried out in Part III,[7] but not for discrete distributions and not for the shape constraints considered here.) Also, the robustness of the extremal distributions suggests that, just as with the extremal distributions in Part I, they should be useful in other contexts, e.g., to study the quality of two-moment approximations for inventory and reliability models as well as other queues.

## II. PROTOTYPE DISTRIBUTIONS, SHAPE CONSTRAINTS, AND NONLINEAR PROGRAMS

### *2.1 Prototype distributions*

To compare alternate parameter specifications and shape constraints in a consistent and meaningful way, we introduce two "prototype" distributions. The specified parameter values, e.g., the moments, will be the parameter values of one of the prototype distributions. The specified shape constraints also will be satisfied by one of the prototype distributions. In this way, we guarantee that there is at least one reasonable probability mass function satisfying all the conditions.

Since mixtures and convolutions of two exponential distributions are frequently used in queueing, we use the discrete analogues: mixtures and convolutions of two geometric distributions. The mixture of two geometric distributions has probability mass function

$$p_k = \gamma(1 - \alpha)\alpha^k + (1 - \gamma)(1 - \beta)\beta^k, \qquad k \geq 0, \tag{4}$$

for probabilities $\alpha$, $\beta$, and $\gamma$. As is often done with mixtures of exponential distributions,[8] we assume balanced means; i.e., we assume that $\gamma\alpha/(1 - \alpha) = (1 - \gamma)\beta/(1 - \beta)$. The convolution of two geometric distributions, on the other hand, has probability mass function

$$p_k = \sum_{j=0}^{k} (1 - \alpha)\alpha^j(1 - \beta)\beta^{k-j} \tag{5}$$

for probabilities $\alpha$ and $\beta$.

To have finite support, we truncate the distributions, and work with the conditional distribution given that the upper bound is not exceeded. We truncate at 20, so that the support is the set of 21 integers $\{0, 1, 2, \cdots, 20\}$. In each case the upper bound 20 is at least 5 standard deviations above the mean.

Mixtures of exponential and geometric distributions are relatively more variable with squared coefficient of variation $c^2 > 1$, while convolutions are relatively less variable with $c^2 < 1$. Hence, we consider one prototype distribution of each type. Prototype I is a truncated mixture of two geometric distributions, having $c^2 = 2.0$; Prototype II is a truncated convolution of two geometric distributions, having $c^2 = 0.8$.

To obtain the specific prototype distributions, we start with the first two moments. For Prototype I, $m_1 = 2.0$ and $m_2 = 12.0$ ($c^2 = 2.0$) and, for Prototype II, $m_1 = 4.0$ and $m_2 = 28.8$ ($c^2 = 0.8$). To obtain a Prototype I distribution with the chosen values of $m_1$ and $m_2$, we numerically solve a system of three nonlinear equations in the three unknowns $\alpha$, $\beta$, and $\gamma$. Two of these equations are the formulas for the moments $m_1$ and $m_2$ of the truncated distribution; the third is the "balanced means" equation. To obtain a Prototype II distribution with the chosen values of $m_1$ and $m_2$, we solve the system of two nonlinear equations in $\alpha$ and $\beta$ corresponding to the moments $m_1$ and $m_2$ of the truncated distribution. The two prototype distributions are displayed in Table I. Additional parameters (the third moment, $m_3$, and trans-

Table I—The two prototype distributions: probability mass functions with $p_k$ on $k$

| | Prototype I | | Prototype II | |
|---|---|---|---|---|
| $k$ | $p_k$ | $p_k/p_{k+1}$ | $p_k$ | $p_k/p_{k+1}$ |
| 0 | 0.3572 | 1.58 | 0.1215 | 0.79 |
| 1 | 0.2262 | 1.58 | 0.1536 | 1.04 |
| 2 | 0.1435 | 1.57 | 0.1475 | 1.16 |
| 3 | 0.0912 | 1.57 | 0.1272 | 1.22 |
| 4 | 0.0583 | 1.56 | 0.1040 | 1.26 |
| 5 | 0.0374 | 1.54 | 0.0825 | 1.28 |
| 6 | 0.0243 | 1.52 | 0.0642 | 1.30 |
| 7 | 0.0160 | 1.49 | 0.0494 | 1.31 |
| 8 | 0.0107 | 1.45 | 0.0377 | 1.32 |
| 9 | 0.0074 | 1.40 | 0.0286 | 1.32 |
| 10 | 0.0053 | 1.34 | 0.0217 | 1.32 |
| 11 | 0.0040 | 1.27 | 0.0163 | 1.33 |
| 12 | 0.0031 | 1.21 | 0.0123 | 1.33 |
| 13 | 0.0026 | 1.15 | 0.0093 | 1.33 |
| 14 | 0.0022 | 1.11 | 0.0070 | 1.33 |
| 15 | 0.0020 | 1.08 | 0.0053 | 1.33 |
| 16 | 0.0019 | 1.05 | 0.0040 | 1.33 |
| 17 | 0.0018 | 1.04 | 0.0030 | 1.33 |
| 18 | 0.0017 | 1.03 | 0.0022 | 1.33 |
| 19 | 0.0017 | 1.02 | 0.0017 | 1.33 |
| 20 | 0.0016 | — | 0.0013 | — |
| | mean $m_1$ | 2.00 | mean $m_1$ | 4.00 |
| | $c^2$ | 2.00 | $c^2$ | 0.80 |

form values, e.g., evaluated at the service rate $\mu$) are given in Tables IV and V of Part I.

### *2.2 Shape constraints*

Mixtures and convolutions of exponential and geometric distributions have many nice properties; see Chapter 5 of Keilson.[4] Mixtures of exponential and geometric distributions are log-convex and thus are DFR, i.e., have decreasing failure rate. For discrete distributions with probability mass functions $p_k$ on the nonnegative integers, *log-convexity* means

$$p_k^2 \le p_{k-1}p_{k+1}, \qquad k \ge 1. \tag{6}$$

Since the ratios $p_k/p_{k-1}$ are nondecreasing with log-convexity, the distribution changes smoothly. The *failure rate* is

$$r_k = p_k \bigg/ \sum_{j=k}^{\infty} p_j, \qquad k \ge 0. \tag{7}$$

Decreasing failure rate of course implies that the probability mass function is decreasing. For log-convex distributions, $c^2 \ge 1$ and $m_3 \ge (3/\sqrt{2})m_2^{3/2}$ (see p. 69 of Keilson[4]).

Convolutions of exponential and geometric distributions are *log-concave*, i.e., the inequality (6) is reversed. Log-concavity is equivalent to *strong unimodality.* A probability mass function $p_k$ on the non-negative integers is *unimodal* if there is an integer $k_0$ such that

$$p_k \geq p_{k-1} \quad \text{for} \quad k \leq k_0$$

and

$$p_k \geq p_{k+1} \quad \text{for} \quad k \geq k_0. \tag{8}$$

A probability mass function $p_k$ is *strongly unimodal* if the convolution with any unimodal probability mass function remains unimodal. In addition to being strongly unimodal, log-concave distributions are IFR, i.e., have increasing failure rate. For log-concave distributions, $c^2 \leq 1$ and $m_3 \leq (3/\sqrt{2})m_2^{3/2}$ (see p. 69 of Keilson[4]).

Of course, truncation and conditioning alter some of these properties. For example, the failure rates are changed significantly. For Prototype I, the failure rate is decreasing for the first thirteen values but is increasing after that. The failure rate remains increasing for Prototype II. The mass function ratios $p_k/p_{k+1}$ are unchanged by the truncation, however. Also the unimodality properties are unchanged: Prototype I is decreasing and Prototype II is unimodal with a mode at 1.

In our study, we focus on the shape constraints unaffected by the truncation, namely, log-convexity or log-concavity and unimodality. We also consider additional parameters such as the third moment, transform values, and constraints on the cdf $F$.

### 2.3 The nonlinear programs

From (1), we see that the mean queue length $L$ depends only on the traffic intensity $\rho$ and the root $\sigma$ of (2). Since $L$ is an increasing function of $\sigma$, the maximum and minimum values of $L$ are attained by the maximum and minimum values of $\sigma$. For interarrival-time distributions with probability mass functions $\{p_k\}$ on the set $\{0, 1, 2, \cdots, 20\}$, (2) becomes

$$\sum_{k=0}^{20} e^{-\mu(1-\sigma)k} p_k = \sigma. \tag{9}$$

To find the maximum and minimum values of $\sigma$, we solve nonlinear programs. The variables are $\sigma$ and the probability masses $p_k$, $k = 0, 1, \cdots, 20$. The constraints specify that $\{p_k\}$ is a proper probability distribution with the specified properties and that (9) holds.

Given the two interarrival-time moments $m_1$ and $m_2$, the upper bound $b = 20$ on the support of the interarrival-time distribution, the service rate $\mu$, and no shape constraints, we have a nonlinear program (NLP) for the maximum of the form:

$$\text{(NLP)} \qquad \max \sigma, \tag{10a}$$

subject to:

$$\sum_{k=0}^{20} e^{-\mu(1-\sigma)k} p_k = \sigma, \tag{10b}$$

$$\sum_{k=0}^{20} p_k = 1, \tag{10c}$$

$$\sum_{k=0}^{20} k p_k = m_1, \tag{10d}$$

$$\sum_{k=0}^{20} k^2 p_k = m_2, \tag{10e}$$

$$p_k \geq 0 \quad \text{for all} \quad k, \tag{10f}$$

$$0 \leq \sigma \leq 1 - \epsilon, \tag{10g}$$

where $0 < \epsilon < 1$. For any probability mass function $\{p_k\}$, the queue is stable if and only if $\rho = 1/\mu m_1 < 1$, in which case $\sigma$ is the unique solution to (10b) in the open interval (0, 1). Since $\sigma = 1$ is also a solution to (10b), we rule it out by bounding $\sigma$ above in (10g).

Of course, we obtain a corresponding NLP for the minimum value of $\sigma$ by changing (10a) from a maximum to a minimum. If there is a mode at $k_0$, then we add the constraints (8) to (10). In the nonlinear program for $c^2 = 2.0$, we assumed that the mode is at 0; in the nonlinear program for $c^2 = 0.8$, we assumed that the mode is at 1. This is consistent with the location of the modes in the prototype distributions. If we were to assume only unimodality without specifying where the mode is, then we would have to solve a program for each possible mode location, and then optimize over the solutions.

When log-convexity is assumed, we add the constraints (6) for $k = 1, \cdots, 19$, to (10). For log-concavity, we add the constraints (6) with the inequality reversed.

## III. SHAPE CONSTRAINTS WITH TWO MOMENTS FIXED

In this section, we give the minimum and maximum values of the root $\sigma$, denoted by $\sigma_\ell$ and $\sigma_u$, respectively, and the interarrival-time distributions yielding these extreme values of $\sigma$. From (1), we obtain the extreme values of $L$, denoted by $L_\ell$ and $L_u$. We also give the maximum relative error (MRE) in $L$, which is computed as

$$\text{MRE} = \frac{L_{u} - L_{\ell}}{L_{\ell}} = \frac{\sigma_{u} - \sigma_{\ell}}{1 - \sigma_{u}}. \tag{11}$$

Table II gives the extremal characteristics and the MRE for the two prototype distributions ($c^2 = 2.0$ and 0.8), two values of the traffic intensity ($\rho = 2/3$ and 9/10), and five constraint cases:

1. Two moments fixed only
2. Plus an upper bound $b = 20$ on the support of the distribution
3. Plus discrete, all mass on $\{0, 1, \cdots, 20\}$
4. Plus unimodal
5. Plus log-convex ($c^2 = 2.0$) or log-concave ($c^2 = 0.8$).

The results in the first two cases, before discreteness is imposed, come from Tables IV and V of Part I. The last three cases are the solutions to the nonlinear programs described in Section 2.3. Tables III through V give the associated extremal probability mass functions. Notice that these extremal distributions are the same for both values of $\rho$.

Each successive case adds an additional constraint to the one before, so the subsets of feasible interarrival-time distributions are nested, and the extremal characteristics get closer to the values for the prototype distributions.

The main conclusion is that with fairly strong but reasonable shape constraints the maximum relative error given two moments is dramatically reduced, becoming small enough to justify two-moment approximations. In particular, with log-convexity or log-concavity the MRE is always less than 8 percent, with the average MRE over the four cases being 3.8 percent. Unimodality helps, but is not good in the case $c^2 = 2.0$ and $\rho = 2/3$, yielding a 33.7-percent MRE. However, from Tables III and IV it is apparent that the unimodal extremal distributions are still quite irregular.

As in Part I, we see that the MRE gets smaller as $\rho$ increases and $c^2$ decreases. From Table II, it is evident that this property holds with shape constraints as well as without. We also see that the upper bound of 20 on the support of the interarrival-time distribution strongly affects the minimum characteristic $\sigma_{\ell}$ but does not change the maximum characteristic $\sigma_{u}$ at all. The discreteness either has no effect (for $\sigma_{u}$ when $c^2 = 2.0$) or only a very small effect.

As we indicated above, there is another significant conclusion. The extremal probability distributions on the set of integers $\{0, 1, 2, \cdots, 20\}$ obtained from the nonlinear programs evidently share an important property with the extremal distributions on $[0, 20]$ given fixed parameters, treated in Part I: The extremal distributions computed by the nonlinear programs are evidently independent of the traffic intensity $\rho$.

Consider the case of no shape constraints (Table V). The extremal distributions on the set $\{0, 1, \cdots, 20\}$ computed by the nonlinear

Table II—The extremal characteristics, $\sigma_\ell$ and $\sigma_u$, and maximum relative errors (MRE) for the GI/M/1 queue: the cases of Prototype Distributions I and II, traffic intensities 2/3 and 9/10, and different shape constraints

| Constraints on the Inter-arrival-Time Distribution | Prototype Distribution I, $c^2 = 2.0$ | | | | | | Prototype Distribution II, $c^2 = 0.8$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho = 2/3$ | | | $\rho = 9/10$ | | | $\rho = 2/3$ | | | $\rho = 9/10$ | | |
| | $\sigma_\ell$ | $\sigma_u$ | MRE | $\sigma_\ell$ | $\sigma_u$ | MRE | $\sigma_\ell$ | $\sigma_u$ | MRE | $\sigma_\ell$ | $\sigma_u$ | MRE |
| Two moments $m_1$ and $m_2$ | 0.417 | 0.806 | 2.00 | 0.807 | 0.936 | 2.00 | 0.417 | 0.676 | 0.80 | 0.807 | 0.893 | 0.808 |
| Plus upper bound at 20 | 0.645 | 0.806 | 0.83 | 0.926 | 0.936 | 0.16 | 0.571 | 0.676 | 0.32 | 0.885 | 0.893 | 0.075 |
| Plus discrete, all mass on (0, 1, 2, . . . , 20) | 0.660 | 0.806 | 0.75 | 0.9268 | 0.9356 | 0.14 | 0.5732 | 0.6760 | 0.32 | 0.8848 | 0.8926 | 0.073 |
| Plus unimodal | 0.730 | 0.798 | 0.34 | 0.9306 | 0.9350 | 0.07 | 0.6145 | 0.6608 | 0.14 | 0.8878 | 0.8915 | 0.034 |
| Plus log-convex (I) or log-concave (II) | 0.762 | 0.779 | 0.08 | 0.9320 | 0.9336 | 0.02 | 0.6387 | 0.6533 | 0.04 | 0.8897 | 0.8909 | 0.001 |
| The prototype distribution | 0.7676 | 0.7676 | 0.00 | 0.9324 | 0.9324 | 0.00 | 0.6429 | 0.6429 | 0.00 | 0.8901 | 0.8901 | 0.000 |

Table III—The distributions minimizing the GI/M/1 mean queue length $L$ given two moments and the shape constraints

| | Prototype Distribution I, $c^2 = 2.0$ | | Prototype Distribution II, $c^2 = 0.8$ | |
|---|---|---|---|---|
| | Unimodal | Log-Convex | Unimodal | Log-Concave |
| | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ |
| 0 | 0.2460 | 0.3486 | 0.0000 | 0.0619 |
| 1 | 0.2460 | 0.2260 | 0.1810 | 0.2155 |
| 2 | 0.2460 | 0.1465 | 0.1810 | 0.1663 |
| 3 | 0.2090 | 0.0950 | 0.1810 | 0.1283 |
| 4 | 0.0031 | 0.0616 | 0.1810 | 0.0990 |
| 5 | 0.0031 | 0.0399 | 0.1748 | 0.0764 |
| 6 | 0.0031 | 0.0259 | 0.0068 | 0.0589 |
| 7 | 0.0031 | 0.0168 | 0.0068 | 0.0455 |
| 8 | 0.0031 | 0.0109 | 0.0068 | 0.0351 |
| 9 | 0.0031 | 0.0071 | 0.0068 | 0.0271 |
| 10 | 0.0031 | 0.0046 | 0.0068 | 0.0209 |
| 11 | 0.0031 | 0.0030 | 0.0068 | 0.0161 |
| 12 | 0.0031 | 0.0019 | 0.0068 | 0.0124 |
| 13 | 0.0031 | 0.0012 | 0.0068 | 0.0096 |
| 14 | 0.0031 | 0.0008 | 0.0068 | 0.0074 |
| 15 | 0.0031 | 0.0005 | 0.0068 | 0.0057 |
| 16 | 0.0031 | 0.0003 | 0.0068 | 0.0044 |
| 17 | 0.0031 | 0.0002 | 0.0068 | 0.0034 |
| 18 | 0.0031 | 0.0001 | 0.0068 | 0.0026 |
| 19 | 0.0031 | 0.0001 | 0.0068 | 0.0020 |
| 20 | 0.0031 | 0.0088 | 0.0068 | 0.0016 |

programs are related to the analytic extremal distributions on the interval [0, b], derived in Section I of Part I. In Part I, the distribution yielding the upper limit of $L$ is a two-point distribution with positive probability mass on 0 and another point $x_u$. The extremal distribution yielding the lower bound is also a two-point distribution with mass on a point $x_\ell$ and on $b$. (The points $x_u$ and $x_\ell$ are determined by the requirement that the distributions have moments $m_1$ and $m_2$.) Our results support the following conjecture:

*Conjecture 1: The extremal distributions on the set of integers* $\{0, 1, 2, \cdots b\}$ *given the same two moments* $m_1$ *and* $m_2$ *have as mass points the triples* $(0, \lfloor x_u \rfloor, \lceil x_u \rceil)$ *and* $(\lfloor x_\ell \rfloor, \lceil x_\ell \rceil, b)$*, respectively, where* $\lfloor x \rfloor$ *is the greatest integer less than x and* $\lceil x \rceil$ *is the least integer greater than x. If* $x_u$ *or* $x_\ell$ *is an integer, then the three-point extremal distribution reduces to a two-point distribution.*

With no additional shape constraints, we can show that the solution to the NLP (10) has at most three nonzero values of $p_k$. To see this, consider the situation where an extreme value of $\sigma$ in (10) is known for particular values of $m_1$ and $m_2$. Then, we can combine (10a) and (10b) to form a linear objective function in the remaining variables $p_k$. With this linear objective, the three linear constraints (10c), (10d),

Table IV—The distributions maximizing the GI/M/1 mean queue length $L$ given two moments and the shape constraints

| | Prototype Distribution I, $c^2 = 2.0$ | | Prototype Distribution II, $c^2 = 0.8$ | |
|---|---|---|---|---|
| | Unimodal | Log-Convex | Unimodal | Log-Concave |
| | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ | $\rho = 2/3$ and $9/10$ |
| 0 | 0.5778 | 0.4377 | 0.1985 | 0.1719 |
| 1 | 0.0500 | 0.1571 | 0.1985 | 0.1450 |
| 2 | 0.0500 | 0.1133 | 0.0629 | 0.1223 |
| 3 | 0.0500 | 0.0817 | 0.0629 | 0.1031 |
| 4 | 0.0500 | 0.0589 | 0.0629 | 0.0870 |
| 5 | 0.0500 | 0.0425 | 0.0629 | 0.0734 |
| 6 | 0.0500 | 0.0306 | 0.0629 | 0.0619 |
| 7 | 0.0500 | 0.0221 | 0.0629 | 0.0522 |
| 8 | 0.0500 | 0.0159 | 0.0629 | 0.0440 |
| 9 | 0.0222 | 0.0115 | 0.0629 | 0.0371 |
| 10 | 0.0000 | 0.0083 | 0.0629 | 0.0312 |
| 11 | 0.0000 | 0.0060 | 0.0367 | 0.0262 |
| 12 | 0.0000 | 0.0043 | 0.0000 | 0.0221 |
| 13 | 0.0000 | 0.0031 | 0.0000 | 0.0181 |
| 14 | 0.0000 | 0.0022 | 0.0000 | 0.0043 |
| 15 | 0.0000 | 0.0016 | 0.0000 | 0.0001 |
| 16 | 0.0000 | 0.0012 | 0.0000 | 0.0000 |
| 17 | 0.0000 | 0.0008 | 0.0000 | 0.0000 |
| 18 | 0.0000 | 0.0006 | 0.0000 | 0.0000 |
| 19 | 0.0000 | 0.0004 | 0.0000 | 0.0000 |
| 20 | 0.0000 | 0.0003 | 0.0000 | 0.0000 |

and (10e), and the bounding constraints (10f), we can determine the values for the $p_k$ by solving a linear program for which only three variables will be in the basis. Hence, to establish Conjecture 1, it suffices to verify that the special three-point distributions are optimal among all feasible three-point distributions for all these objective functions, i.e., for all arguments of the transform. Of course, if the extremal mass points $x_{\ell}$ and $x_{u}$ for the distributions on $[0, b]$ are integer, then these extremal distributions on $[0, b]$ are feasible for the smaller set $\{0, 1, \cdots, b\}$ and are thus still optimal. This happens here for the upper bound with $c^2 = 2.0$.

The following conjecture for the cases with shape constraints is also supported by our experiments (we solved the programs for traffic intensities ranging from 0.01 to 0.9):

*Conjecture 2: For each kind of shape constraint considered, the extremal interarrival-time distributions on* $\{0, 1, 2, \cdots, b\}$ *for the GI/M/1 queue, given the first two moments of the interarrival-time distribution, are independent of the traffic intensity,* $\rho$.

Moreover, there is an obvious regularity in the extremal unimodal

Table V—The extremal GI/M/1 interarrival-time distributions without shape constraints: the effect of discreteness and an upper bound on the support of the distribution

| | Prototype Distribution I, $c^2 = 2.0$ | | | | | |
|---|---|---|---|---|---|---|
| Upper Bounds, $\sigma_u$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| Cases 1, 2, and 3 | 0.6667 | 0.000 | 0.3333 | 6.00 | — | — |
| Lower Bounds, $\sigma_\ell$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| Case 3 | 0.4211 | 1.000 | 0.5555 | 2.00 | 0.0234 | 20.00 |
| Case 2 | 0.9759 | 1.556 | 0.0241 | 20.00 | — | — |
| Case 1 | 1.0000 | 2.000 | — | — | — | — |
| | Prototype Distribution II, $c^2 = 0.8$ | | | | | |
| Upper Bounds, $\sigma_u$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| Cases 1 and 2 | 0.444 | 0.000 | 0.556 | 7.20 | — | — |
| Case 3 | 0.4429 | 0.000 | 0.4571 | 7.00 | 0.1000 | 8.00 |
| Lower Bounds, $\sigma_\ell$ | $p_1$ | $x_1$ | $p_2$ | $x_2$ | $p_3$ | $x_3$ |
| Case 3 | 0.7529 | 3.000 | 0.2000 | 4.00 | 0.0471 | 20.00 |
| Case 2 | 0.9524 | 3.200 | 0.0476 | 20.00 | — | — |
| Case 1 | 1.000 | 4.000 | — | — | — | — |

Note: The cases are described at the beginning of Section II of this paper. $x_i$ is the $i$th point with positive probability mass; $p_i$ is the probability mass at point $x_i$.

distributions: they have only a few points of mass change. This can be explained by making a change of variables. For unimodal distributions on $\{0, 1, \cdots, b\}$ with a mode at 0, we can make the change of variables

$$q_k = (k + 1)(p_k - p_{k+1}), \qquad k \geq 0, \tag{12}$$

with $p_{b+1} = 0$. Then $q_k \geq 0$ for all $k$, and the constraints for $p_k$ become the following constraints for $q_k$:

$$\sum_{k=0}^{b} q_k = 1, \qquad \sum_{k=0}^{b} kq_k = 2m_1 \tag{13}$$

and

$$\sum_{k=0}^{b} k^2 q_k = 3m_2 - m_1/2. \tag{14}$$

Moreover, the linear objective function $\sum_{j=0}^{b} e^{-sj} p_j$ is transformed into the linear objective function

$$\sum_{k=0}^{b} q_k (k + 1)^{-1} \sum_{j=0}^{k} e^{-sj}.$$

Solving the transformed linear program yields three-point solutions. Hence, the extreme points for the decreasing distributions with unimodal constraints have at most three points of decrease after 0. For decreasing probability mass functions, we thus make the following conjecture.

*Conjecture 3*: *Let* $(0, x_u)$ *and* $(x_\ell, b)$ *be the pairs of mass points for the extremal distributions on* $[0, b]$ *given the first two moments* $2m_1$ *and* $3m_2 - m_1/2$, *obtained from Section II of Part I. Then, for the GI/M/*1 *queue characteristics, the extremal* decreasing *probability mass functions on the set of integers* $\{0, 1, 2, \cdots, b\}$ *given the first two moments* $m_1$ *and* $m_2$ *have as points of decrease the triples* $(0, \lfloor x_u \rfloor, \lceil x_u \rceil)$ *and* $(\lfloor x_\ell \rfloor, \lceil x_\ell \rceil, b)$, *respectively. (This completely determines the extremal probability distributions.)*

For other modes, say $k_0$, we can do a similar change of variables, namely,

$$q_{k_0+j} = \begin{cases} \dfrac{(2j+1)}{2}(p_{k_0+j} - p_{k_0+j+1}), & 0 \le j \le b - k_0, \\ \\ \dfrac{-(1+2j)}{2}(p_{k_0+j+1} - p_{k_0+j}), & -k_0 - 1 \le j \le -1, \end{cases} \tag{15}$$

with $p_{-1} = p_{b+1} = 0$, so that $q_k$ is a probability mass function on $\{-1, 0, 1, \cdots, 20\}$ for which

$$p_j = \sum_{i=0}^{i=j} a_{ij} q_{i-1}, \qquad 0 \le j \le b, \tag{16}$$

where $a_{ij}$ are appropriate constants determined by (15). As before, the two linear moment constraints for $p_k$ become linear moment constraints for $q_k$. In addition, there is an extra linear constraint on the $q_k$ when $k_0 > 0$ since the support of $q_k$ has one more point, i.e., is $\{-1, 0, 1, \cdots, b\}$ instead of $\{0, 1, \cdots, b\}$. In particular, from (15) it is easy to see that

$$\sum_{j=0}^{b-k_0} \left(\frac{2}{2j+1}\right) q_{k_0+j} + \sum_{j=1}^{k_0+1} \left(\frac{2}{2j-1}\right) q_{k_0-j} = 0. \tag{17}$$

Therefore, solving the transformed linear program would require the inclusion of (17) as a fourth linear constraint. This results in four positive values among the $q_k$. Therefore, extremal unimodal distributions with mode $k_0 > 0$ must have at most four points of mass change, including any positive mass at 0 and 20. The unimodal extremal distributions for Prototype II with $k_0 = 1$ obtained from the nonlinear programs have this property; see Tables III and IV. For Prototype II, we also found that the extremal characteristics $\sigma_\ell$ and $\sigma_u$ both decreased as the mode $k_0$ was changed from 0 to 1 to 2, e.g., $\sigma_\ell$ changed from 0.633 to 0.614 to 0.603.

The numerical results also show that the extremal distributions for the log-convex and log-concave constraints have special regularity that can be seen by looking at the successive ratios $p_k/p_{k+1}$.

*Conjecture 4: In the log-convex case, the upper-(lower-) bound distribution has constant ratios* $p_k/p_{k+1}$ *for* $k = 1, 2, \cdots, b$ ($k = 0, 1, \cdots, b - 1$) *with an extra mass at* 0 ($b$). (*This determines the interarrival-time distribution.*)

Conjecture 4 is supported by Tables II and IV. With log-convex constraints, the upper-(lower-) bound ratios are $p_1/p_2 \approx 1.387$ ($p_1/p_2 \approx 1.542$).

## IV. SHAPE CONSTRAINTS WITH OTHER PARAMETER SPECIFICATIONS

In this section, we investigate different parameter specifications, both with and without shape constraints. However, attention is focused on the cases with shape constraints because alternate parameter specifications without shape constraints were considered in Sections III and IV of Part I. We consider only Prototype Distribution I with the traffic intensity $\rho = 2/3$. This is the difficult case in Section III, yielding the largest maximum relative errors.

Table VI contains the major results. It gives the maximum relative errors in $L$ for various combinations of two and three parameters with no shape constraints and with log-convex shape constraints. Of course, we still consider the first two moments $m_1$ and $m_2$. The additional parameters that we consider are: the third moment, $m_3$, the Laplace-Stieltjes transform evaluated at the service rate, $\phi(\mu)$, and the interarrival time cdf evaluated at $k$, $F(k)$, i.e., $F(k) = p_0 + p_1 + \cdots + p_k$. These parameters are fixed at the values satisfied by Prototype I. In particular, we use $m_3 = 119.01$, $\phi(\mu) = 0.5098$, $F(0) = 0.35724$, $F(2) = 0.72692$, and $F(7) = 0.95409$. Combinations of two parameters are

Table VI—A comparison of alternate second- and third-parameter specifications: the maximum relative error (MRE) in the mean queue length $L$ in a GI/M/1 queue, based on Prototype I with $\rho = 2/3$

| | The Second Parameter in Addition to $m_1$ | | | | |
|---|---|---|---|---|---|
| | $m_2$ | $\phi(\mu)$ | $F(0)$ | $F(2)$ | $F(7)$ |
| No shape constraint | 0.75 | 1.53 | 3.10 | 5.09 | 3.81 |
| Plus log-convexity | 0.077 | 0.083 | 0.096 | 0.370 | 0.555 |
| | The Third Parameter in Addition to $m_1$ and $m_2$ | | | | |
| | $m_3$ | $\phi(\mu)$ | $F(0)$ | $F(2)$ | $F(7)$ |
| No shape constraint | 0.069 | 0.175 | 0.331 | 0.604 | 0.609 |
| Plus log-convexity | 0.009 | 0.012 | 0.019 | 0.049 | 0.020 |

Note: $m_k$ is the $k$th moment, $\phi(\mu)$ is the Laplace-Stieltjes transform evaluated at the service rate $\mu$, and $F(k)$ is the cdf evaluated at $k$, i.e., $F(k) = p_0 + p_1 + \ldots + p_k$, of Prototype Distribution I. These values are $m_1 = 2.00$, $m_2 = 12.00$, $m_3 = 119.01$, $\phi(\mu) = 0.5098$, $F(0) = 0.35724$, $F(2) = 0.72692$, and $F(7) = 0.95409$. The distribution itself appears in Table I. All these results are obtained from the nonlinear programs.

formed by specifying each of the additional parameters together with the first moment $m_1$. Combinations of three parameters are formed by specifying each of the additional parameters together with the first two moments $m_1$ and $m_2$.

The first conclusion is that, with log-convexity, the third moment or almost any other third parameter in addition to the first two moments makes the maximum relative error negligible. For the third moment, the maximum relative error is less than one percent and for all but one of the other third parameters it is less than two percent. This suggests that with nice distributions three-moment approximations ought to work very well for more general models.

The second conclusion is that the next higher moment is the best additional parameter in all cases. However, the advantage of the moment over the transform value decreases dramatically with log-convexity. Although the cdf constraints certainly reduce the MRE, the next higher moment and the transform value perform better as additional parameters.

In order to have the maximum relative error small enough to justify approximations, say less than 10 percent, it appears that three constraints are enough. It suffices to specify either three moments without shape constraints (6.9-percent MRE) or specify two moments with log-convex shape constraints (7.7-percent MRE). We can think of log-convexity as being roughly equivalent to another moment parameter.

The values of $\sigma$, the GI/M/1 probability of delay, for the various parameter specifications and shape constraints are given in Tables VII and VIII. From Table VII we see an interesting reversal in form with log-convexity. Without shape constraints, the next higher moment is better than the transform value $\phi(\mu)$ as an additional parameter for the upper bound but not as the second parameter for the lower bound. With log-convexity, these orderings are reversed. From Table VIII, we see that $F(0)$ is significantly better than the other two cdf values as an additional parameter for the upper bound with log-convexity and for the lower bound with no shape constraints, but not in the other cases.

We also tabulated the extremal interarrival-time distributions for the different combinations of parameters and shape constraints, but they have been omitted to save space. As in Section III, these extremal distributions have important regularity properties. With $k$ parameters and no shape constraints, the extremal distributions have at most $k$ + 2 positive mass points; with $k$ parameters and a decreasing mass function, the extremal distribution have at most $k + 2$ points of mass change after 0. As in Section III, there is also regularity in the extremal distributions in the log-convex case, which can be seen by looking at the successive ratios $p_k/p_{k+1}$. There appear to be only a few points

Table VII—The GI/M/1 extremal characteristics $\sigma$ (the probability of delay) given $c^2 = 2.0$, $\rho = 2/3$, and the different shape constraints

| | | Parameters Specified in Addition to the Mean | | | | |
|---|---|---|---|---|---|---|
| | | $m_2$ | $\phi(\mu)$ | $\phi(\mu(1-\rho))$ | $m_2, m_3$ | $m_2, \phi(\mu)$ |
| | No shape constraints | 0.8057 | 0.8811 | 0.7994 | 0.7768 | 0.7844 |
| $\sigma_u$ | Unimodal | 0.7983 | 0.8203 | 0.7777 | 0.7708 | 0.7777 |
| | Log-convex | 0.7790 | 0.7736 | 0.7690 | 0.7694 | 0.7680 |
| | Prototype Distribution | 0.7675 | 0.7675 | 0.7675 | 0.7675 | 0.7675 |
| | Log-convex | 0.7621 | 0.7542 | 0.7638 | 0.7673 | 0.7652 |
| $\sigma_\ell$ | Unimodal | 0.7374 | 0.7192 | 0.7574 | 0.7615 | 0.7606 |
| | No shape constraints | 0.6601 | 0.6986 | 0.7545 | 0.7548 | 0.7466 |

Table VIII—The extremal characteristics $\sigma$ (the probability of delay) for the GI/M/1 queue given values of the cumulative distribution function $F$ in addition to the first moment, $m_1$, or the first two moments, $m_1$ and $m_2$: the case of Prototype Distribution I with traffic intensity $\rho = 2/3$

| | | Additional Parameter with $m_1$ | | | Additional Parameter with $m_1$ and $m_2$ | | |
|---|---|---|---|---|---|---|---|
| | | $F(0)$ | $F(2)$ | $F(7)$ | $F(0)$ | $F(2)$ | $F(7)$ |
| $\sigma_u$ | No shape constraint | 0.9093 | 0.9135 | 0.9034 | 0.7927 | 0.8019 | 0.8042 |
| | Unimodal | 0.8523 | 0.8558 | 0.8408 | 0.7867 | 0.7940 | 0.7959 |
| | Log-convex | 0.7760 | 0.8278 | 0.8385 | 0.7684 | 0.7781 | 0.7706 |
| $\sigma_\ell$ | Log-convex | 0.7546 | 0.7641 | 0.7489 | 0.7639 | 0.7670 | 0.7659 |
| | Unimodal | 0.6849 | 0.6683 | 0.6859 | 0.7544 | 0.7388 | 0.7427 |
| | No shape constraint | 0.6280 | 0.4736 | 0.5351 | 0.7241 | 0.6823 | 0.6849 |

where these ratios change. Including the final mass point, for $k$ parameters there appear to be $k$ points where the ratios change. Given $m_2$ and $m_3$, the ratios $p_{k-1}/p_k$ change for the lower bound at $k \in \{10, 11\}$ and for the upper bound at $k \in \{2, 20\}$. Given only $m_2$, the ratio changes for the lower bound at $k = 20$ and for the upper bound at $k = 2$.

Although we report results for only a single value of the traffic intensity $\rho$, it also appears that the extremal distributions do not change with $\rho$, i.e., the numerical solutions to the nonlinear programs were indistinguishable for a range of $\rho$ values tested from 0.01 to 0.9. There are natural extensions for Conjectures 1 through 4 to other parameter specifications.

We also found the extreme values of the transform values $\phi(\mu)$, which are the blocking probabilities for the associated GI/M/1 loss system, for given moments and shape constraints. The numerical solutions for the extremal distributions appear to be the same as those in which $\sigma$ is the objective. The extremal blocking probabilities are given in Table IX. As in Section V and Table X of Part I, the constraints pin down the delay probability $\sigma$ better than the associated blocking probability $\phi(\mu)$.

## V. OTHER BOUNDS AND APPROXIMATIONS FOR GI/G/1 QUEUES

Having obtained the extreme values of the GI/M/1 mean queue length $L$ given two moments and various shape constraints, we note how these results compare with other bounds and approximations for the GI/G/1 queue that depend only on the first two moments of the interarrival times and service times. Several of these other bounds and approximations are defined and compared in Shanthikumar and Buzacott.[9] These bounds are stated for the mean waiting time, but they are easily translated into the mean queue length by Little's formula. Among the bounds and approximations treated there is the Kingman[10] upper bound and the Marchal[11] approximation based on it. Recently, Daley[12] obtained a better upper bound, (1.5) there, which can be used to produce an approximation by scaling to make the M/G/1 case exact, just as Marchal did for the Kingman bound. We call this new approximation Marchal (D) and the original Marchal approximation Marchal (K). Shanthikumar and Buzacott also discuss the Kraemer and Langenbach-Belz[13] approximation and a modification of Page's[14] approximation based on it, formula (8) there, which we call Modified-Page. They also discuss an approximation by Sakasegawa[15] and Yu,[16] which coincides with the monotone-failure-rate approximation in Whitt.[17] Another natural two-moment approximation is to fit a hyperexponential distribution with balanced means to the two moments, provided $c^2 \geq 1$ (see Section III of Whitt[7]) and solve the resulting $H_2^b/H_2^b/1$ queue via a vector-state Markov process. When a distribution is exponential, $H_2^b$ becomes M, so for the setting of the GI/M/1 queue based on Prototype I we obtain the $H_2^b$/M/1 queue. Finally, the crudest

Table IX—The extremal blocking probabilities for the associated GI/M/1 loss system (the transform values $\phi(\mu)$) with given moments and shape constraints: case of Prototype I with $m_1 = 2$, $m_2 = 12$, $m_3 = 119$, and $\mu = 4/3$ ($\rho = 2/3$)

| | The Moment Parameters | |
|---|---|---|
| The shape constraints | $m_1, m_2$ | $m_1, m_2, m_3$ |
| Max $\phi(\mu)$, no shape constraints | 0.6641 | 0.5902 |
| Max $\phi(\mu)$, unimodal | 0.6225 | 0.5414 |
| Max $\phi(\mu)$, log-convex | 0.5499 | 0.5147 |
| Prototype I | 0.5098 | 0.5098 |
| Min $\phi(\mu)$, log-convex | 0.5026 | 0.5087 |
| Min $\phi(\mu)$, unimodal | 0.4395 | 0.4713 |
| Min $\phi(\mu)$, no shape constraints | 0.3279 | 0.4049 |

approximation is obtained by ignoring the second moments and using the M/M/1 formula $L = \rho/(1 - \rho)$. There is also a related collection of approximations arising from diffusion approximations that we will not consider here; see Whitt[18] and references there.

We also include bounds for GI/G/1 queue in which the interarrival-time distribution is IFR or DFR.[17] Marshall[19] obtained a lower bound for IFR/G/1 queues and an upper bound for DFR/G/1 queues. Stoyan and Stoyan[20] also obtained an upper bound for IFR/G/1 queues and a lower bound for DFR/G/1 queues, which is just the M/G/1 queue with the given arrival rate. (In fact, the interarrival-time distribution is only required to be NBUE or NWUE, i.e., new better or worse than used in expectation.) The DFR bounds, but not the IFR bounds, are tight.[17]

In Table X these various bounds and approximations are compared with the extreme values of $L$ for $c^2 = 2.0$ and 0.8 (the two prototype distributions), and $\rho = 2/3$ and 9/10. When interpreting these results, note that none of the other bounds and approximations use the fact that the service-time distribution is exponential. Also, the DFR and the IFR bounds are based on interarrival-time distributions having densities with support on the entire positive half line, whereas the bounds obtained here in Section II are based on interarrival-time distributions with support $\{0, 1, \cdots, 20\}$. The upper bound $b = 20$ on the support of the interarrival-time distribution has a significant impact on the lower bound mean queue length, $L_{\ell}$, when the interarrival-time distribution in DFR ($c^2 > 1$) and on the upper bound mean queue length, $L_{u}$, when the interarrival-time distribution is IFR ($c^2 < 1$).

The first conclusion is that all the approximations, with the exception of the M/M/1 approximation when $c^2 = 2$, appear to be within the range of reasonable values for actual GI/M/1 systems. However, for $c^2 = 2.0$ and $\rho = 2/3$, the Modified-Page approximation seems a bit high. The Kraemer and Langenbach-Belz approximations for $c^2 = 2.0$ seem low compared to the log-convex discrete lower bounds (Case 5), but note that the Kraemer and Langenbach-Belz approximations are close to the $H_2^b$/M/1 values.

The second conclusion is that the D/M/1 lower bound and the Kingman and Daley upper bounds are not close enough to be good approximations. Of course, the upper bounds are asymptotically tight in heavy traffic, so they are not too bad when $\rho = 0.9$.

We believe that the shape constraints play a very useful role. They narrow down the range of possible values for $L$, so it is reasonable to consider approximations based on two moments only. Instead of concluding that it is not possible to obtain a good approximation when $c^2 > 1$ (p. 765 of Shanthikumar and Buzacott[9]), we conclude that it is

Table X—A comparison of the GI/M/1 extreme values of the mean queue length, $L$, with other bounds and approximations for $L$ in GI/G/1 queues that depend on the first two moments of the interarrival times and service times

| | Prototype Distribution I, $c^2 = 2.0$ | | Prototype Distribution II, $c^2 = 0.8$ | |
|---|---|---|---|---|
| GI/G/1 upper bounds | $\rho = 2/3$ | $\rho = 9/10$ | $\rho = 2/3$ | $\rho = 9/10$ |
| Kingman | 4.33 | 14.95 | 2.53 | 8.95 |
| Daley | 4.00 | 14.85 | 2.40 | 8.91 |
| DFR or IFR | 3.00 | 13.50 | 2.00 | 9.00 |
| GI/M/1 upper bounds | | | | |
| Case 1, two moments only | 3.44 | 14.06 | 2.06 | 8.41 |
| Case 2, bound on support | 3.44 | 14.06 | 2.06 | 8.41 |
| Case 3, discrete | 3.44 | 13.98 | 2.06 | 8.38 |
| Case 4, unimodal | 3.30 | 13.85 | 1.97 | 8.29 |
| Case 5, log-concave or log-convex | 3.02 | 13.55 | 1.92 | 8.25 |
| Approximations | | | | |
| Marchal (K) | 2.92 | 13.48 | 1.82 | 8.11 |
| Marchal (D) | 2.89 | 13.46 | 1.82 | 8.11 |
| Kraemer and L-B | 2.56 | 12.85 | 1.86 | 8.18 |
| Modified-Page | 3.28 | 13.34 | 1.83 | 8.13 |
| Sakasegawa and Yu | 2.67 | 13.05 | 1.87 | 8.19 |
| $H_2^b$/M/1 | 2.64 | 13.03 | — | — |
| M/M/1 | 2.00 | 9.00 | 2.00 | 9.00 |
| GI/M/1 lower bounds | | | | |
| Case 5, log-concave or log-convex | 2.80 | 13.24 | 1.85 | 8.16 |
| Case 4, unimodal | 2.47 | 12.97 | 1.73 | 8.02 |
| Case 3, discrete | 1.96 | 12.30 | 1.56 | 7.83 |
| Case 2, bound on support | 1.88 | 12.16 | 1.55 | 7.81 |
| Case 1, two moments only | 1.14 | 4.66 | 1.14 | 4.66 |
| GI/G/1 lower bound | | | | |
| DFR or IFR | 2.00 | 9.00 | 1.80 | 8.00 |

Note: The actual values of $L$ for the prototype distributions are 2.87, 13.31, 1.87, and 8.19, respectively.

possible to consider approximations based on two moments, with the caveat that the distributions should not be too irregular.

## VI. MATHEMATICAL PROGRAMMING ISSUES

Solving the nonlinear programs turned out to be quite complicated, especially when the shape constraints were included. The programs involve 22 variables and up to 46 constraints. This is a reasonably large problem for most general-purpose nonlinear programming codes. In addition, when the nonlinear constraints (6) are present, the problems apparently become ill-conditioned and poorly scaled, causing numerical difficulties and, often, nonconvergence of standard nonlinear programming algorithms.

The numerical results reported in this paper were obtained using two nonlinear programming codes from the Harwell Subroutine Library, compiled by the Numerical Analysis Group at the United Kingdom Atomic Energy Authority. These codes were VF01AD, an augmented Lagrangian code described in Fletcher,[21] and VF02AD, a quadratic approximation code due to Powell.[22] They were run in double precision on an Amdahl 470/V6 computer operating under multiple virtual storage. Both codes are included in a recent performance comparison of available state-of-the-art computer codes compiled by Schittkowski.[23]

Although numerical problems were experienced with both codes, they were far more prevalent with the augmented Lagrangian code VF01AD. The augmented Lagrangian code solves a sequence of unconstrained optimization subproblems. Unfortunately, some of the problems (especially for $\rho = 0.9$ and Prototype II) resulted in ill-conditioned subproblems and, occasionally, subproblems with an unbounded optimum, in which case, the augmented Lagrangian code did not converge. Our experience bears out the experience of Schittkowski, who reported that the performance of this code deteriorates drastically for ill-conditioned problems and is highly sensitive to slight variations of the problem. Certain standard measures, however, were able to overcome the numerical difficulties in most instances. For example, in some runs, the default settings for certain penalty parameters were overridden according to rules of thumb suggested by Gill, Murray, and Wright (see pp. 295–6 of Ref. 24).

Fortunately, for those experiments for which code VF01AD did not obtain a solution, code VF02AD did. This supports Schittkowski's conclusion that code VF02AD is one of the most robust and reliable codes available. Even though several individual runs experienced numerical overflows and underflows and eventual nonconvergence, it was always possible eventually to obtain convergence with this quadratic approximation code using some starting point. In particular, problems with $\rho = 0.9$ and Prototype II were solved with less difficulty using VF02AD.

All runs were tried with a variety of starting points. These starting points included the prototype distributions, the uniform distribution, solutions obtained for other parameter settings, and an initial all-zero solution.

In summary, then, although this nonlinear programming method for analyzing the quality of queueing approximations provides considerable insight and potential for future applications, great care must be exercised in the solution of the nonlinear programs. Our experience indicates that the computer code, the parameter settings, the starting points, and the scaling of the variables must be chosen judiciously in order to obtain useful results.

## VII. ACKNOWLEDGMENT

We thank Hanan Luss and Moshe Segal for many suggestions on the presentation.

## REFERENCES

1. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," AT&T Bell Lab. Tech. J., this issue.
2. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62*, No. 9 (November 1983), pp. 2779–2815.
3. J. W. Cohen, *The Single Server Queue*, Amsterdam: North Holland, 1969.
4. J. Keilson, *Markov Chain Models–Rarity and Exponentiality*, New York: Springer-Verlag, 1979.
5. A. A. Borovkov, *Stochastic Processes in Queueing Theory*, New York: Springer-Verlag, 1976.
6. A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *53*, No. 3 (March 1973), pp. 437–48.
7. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," AT&T Bell Lab. Tech. J., this issue.
8. W. Whitt, "Approximating a Point Process by a Renewal Process: Two Basic Methods," Oper. Res., *30*, No. 1 (January–February 1982), pp. 125–47.
9. J. G. Shanthikumar and J. A. Buzacott, "On the Approximations to the Single Server Queue," Int. J. Prod. Res., *18* (1980), pp. 761–73.
10. J. F. C. Kingman, "Some Inequalities for the Queue GI/G/1," Biometrika, *49*, No. 3 (1962), pp. 315–24.
11. W. G. Marchal, "An Approximate Formula for Waiting Time in Single Server Queues," AIIE Trans. (1976), pp. 473–4.
12. D. J. Daley, "Inequalities for Moments of Tails of Random Variables, with a Queueing Application," Z. Wahrscheinlichkeitstheorie verw. Gebiete, *41* (1977), pp. 139–43.
13. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," Eighth Int. Teletraffic Cong., Melbourne, 1976, pp. 235-1–8.
14. E. Page, *Queueing Theory in OR*, New York: Crane Russak and Co., 1972.
15. H. Sakasegawa, "An Approximation Formula $L_q \simeq \alpha\rho^\beta/(1-\rho)$," Ann. Inst. Statist. Math., *29A* (1977), pp. 67–75.
16. P. S. Yu, "On Accuracy Improvement and Applicability Conditions of Diffusion Approximation with Application to Modelling of Computer Systems," Digital Systems Laboratory, *TR-129*, Stanford University, 1977.
17. W. Whitt, "The Marshall and Stoyan Bounds for IMRL/G/1 Queues are Tight," Oper. Res. Lett., *1*, No. 6 (December 1982), pp. 209–13.
18. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Lett., *1*, No. 5 (November 1982), pp. 165–9.
19. K. T. Marshall, "Some Inequalities in Queueing," Oper. Res., *16*, No. 3 (May–June 1968), pp. 651–65.
20. H. Stoyan and D. Stoyan, "Monotonicity Properties of Waiting Times in the GI/G/1 Queue," Zeit. angew Math. Mech., *49*, No. 12 (1969), pp. 729–34 (in German).
21. R. Fletcher, "An Ideal Penalty Function for Constrained Optimization," Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., New York: Academic Press, 1975.
22. M. J. D. Powell, "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," *Proceedings of the 1977 Dundee Conference on Numerical Analysis*, Lecture Notes in Mathematics, New York: Springer-Verlag, 1978.
23. K. Schittkowski, *Nonlinear Programming Codes: Information, Tests, Performance*, Lecture Notes in Economics and Mathematical Systems, New York: Springer-Verlag, 1980.

24. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, New York: Academic Press, 1981.

## AUTHORS

**John G. Klincewicz,** S.B. (Mathematics), 1975, Massachusetts Institute of Technology; M.A., 1977, and Ph.D. (Operations Research), 1979, Yale University; AT&T Bell Laboratories, 1979—. At AT&T Bell Laboratories, Mr. Klincewicz is a member of the Operations Research Department. His research interests include applications of mathematical programming and the development of algorithms for network flow problems and facility location problems. Member, Operations Research Society, Mathematical Programming Society.

**Ward Whitt,** A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and congestion models.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# On Approximations for Queues, III: Mixtures of Exponential Distributions

By W. WHITT*

(Manuscript received May 3, 1983)

To evaluate queueing approximations based on a few parameters (e.g., the first two moments) of the interarrival-time and service-time distributions, we examine the set of all possible values of the mean queue length given this partial information. In general, the range of possible values given such partial information can be large, but if in addition shape constraints are imposed on the distributions, then the range can be significantly reduced. The effect of shape constraints on the interarrival-time distribution in a GI/M/1 queue was investigated in Part II (see "On Approximations for Queues, II: Shape Constraints," this issue) by restricting attention to discrete probability distributions with probability on a fixed finite set of points and then solving nonlinear programs. In this paper we show how one kind of shape constraint—assuming that the distribution is a mixture of exponential distributions—can be examined analytically. By considering GI/G/1 queues in which both the interarrival-time and service-time distributions are mixtures of exponential distributions with specified first two moments, we show that additional information about the distributions is more important for the interarrival time than for the service time.

## I. INTRODUCTION AND SUMMARY

Many approximations for the mean steady-state queue length in the GI/G/1 queue are based on the first two moments of the general interarrival-time and service-time distributions. To evaluate these approximations, it is natural to compare the approximations with the set of possible values of the mean queue length given this limited

* AT&T Bell Laboratories.

moment information. For several special cases, the minimum and maximum values of the mean queue length are attained by simple two-point extremal distributions. In Part I the extremal distributions were used to calculate the extreme values of the mean queue length in the GI/M/1 queue and show how they depend on the traffic intensity, the second moment of the interarrival-time distribution, and an upper bound on the distribution.[1] Extremal distributions were also used to compare different parameters for approximations.

Unfortunately, the range of possible values of the mean queue length in the GI/M/1 queue given this limited moment information can be very wide. However, since the extremal interarrival-time distributions are quite unusual, this still leaves the possibility that the range would not be too wide for typical distributions. Part II showed that the range of possible values for the mean queue length in the GI/M/1 queue can indeed be reduced dramatically by imposing shape constraints such as unimodality and log-convexity on the interarrival-time distributions with given first two moments.[2] This was done by restricting attention to discrete distributions with all mass on a fixed finite set of points and solving nonlinear programs.

Unlike Part I, the approach in Part II was computational, based on nonlinear programs. However, the extremal distributions obtained from the nonlinear programs exhibit regularity that suggests the possibility of an analytic treatment similar to Part I. This paper sets out to treat analytically one kind of shape constraint. We show that the theory underlying Part I also applies to mixtures of exponential distributions. Within this class of distributions there are extremal distributions with respect to the same partial orderings based on Laplace transforms used in Part I. The extremal distributions in this class of mixtures are obtained by using the extremal distributions of Part I as the mixing distributions. These extremal distributions yield the minimum and maximum mean queue length as interarrival-time distributions in the GI/M/I queue and as service-time distributions in the $K_2$/G/1 queue (with interarrival-time distributions having a rational Laplace-Stieltjes transform with a denominator of degree 2, see Section V of Part I and Section VII here).

The rest of this paper is organized as follows. In Section II, we briefly review the theory yielding distributions that minimize or maximize the Laplace-Stieltjes transforms for all arguments. In Section III, we show that this theory applies to mixtures of exponential distributions, and in Section IV, we apply the results to the H/M/1 queue, having interarrival-time distributions that are mixtures of exponential distributions. In Section V, we examine the case of $H_2$ interarrival-time distributions (mixtures of two exponential distributions) in more detail. In Section VI, we indicate how some of the results for H/M/1 queues extend to GI/G/1 queues with interarrival-time having increasing mean residual life (the service-time distribution is general instead of exponential and the interarrival-time distribution can be more general than a mixture of exponentials). Finally, in Section VII, we indicate how the ordering of transforms can be applied to compare different service-time distributions in $K_2$/H/1 queues. There, Table III gives a good picture of the way the mean queue length can vary in a large class of GI/G/1 queues with the first two moments of the interarrival time and the service time specified.

## II. EXTREME VALUES OF THE LAPLACE-STIELTJES TRANSFORM

As in Eckberg[3] and references there, we obtain the extremal distributions for queues with specified moments for the interarrival times and service times from extremal distributions for the Laplace-Stieltjes transform. For the transform, the object is to find a cdf (cumulative distribution function) $F$ with support on the interval $[0, bm_1]$, $b \leq \infty$, to minimize or maximize the transform $\phi(s)$, defined by

$$\phi(s) = \int_0^\infty e^{-st} dF(t), \qquad s \geq 0, \tag{1}$$

subject to moment constraints

$$m_j = \int_0^\infty t^j dF(t), \tag{2}$$

for $j = 1, 2, \cdots, n$. The key idea is to apply the theory of Tchebycheff systems in Karlin and Studden,[4] which implies that the optimization problem involving (1) and (2) has a very nice solution. First, the minimizing and maximizing cdf's are independent of the variable $s$ in the transform $\phi(s)$. Second, the extremal distributions are discrete distributions with positive mass on at most $(n + 2)/2$ mass points. Finally, the points with positive mass and the associated probability masses are obtained simply by solving a system of linear equations. (See Section 2 of Eckberg[3] and Section II of Part I[1] for more discussion.)

## III. MIXTURES OF EXPONENTIAL DISTRIBUTIONS

Now we consider the optimization problem in Section II for distributions that are mixtures of exponential distributions. It turns out that the theory of Tchebycheff systems can be applied again because the extremal distributions in this class of mixtures can be obtained by using extremal mixing distributions.

A cdf $F$ is a mixture of exponential distributions if it satisfies

$$1 - F(x) = \int_0^\infty \mathrm{e}^{-x/t} dG(t), \qquad x \ge 0, \tag{3}$$

for some mixing cdf $G$. Densities of mixtures of exponential distributions are also called completely monotone; see Section 5.4 of Keilson.[5] A density $f$ has this property if and only if it has derivatives $f^{(n)}$ of all orders $n$ and $(-1)^n f^{(n)}(x) \ge 0$ for all $x$ and $n$. Mixtures of exponentials are log-convex (see Part II) and thus are DFR (have decreasing failure rate).

It turns out that the moments and transform of $F$ are easily expressed via $G$:

$$m_k(F) = \int_0^\infty t^k dF(t) = k! \int_0^\infty t^k dG(t) = k! m_k(G) \tag{4}$$

and

$$\phi(s) = \int_0^\infty \mathrm{e}^{-sx} dF(x) = \int_0^\infty (1 + st)^{-1} dG(t). \tag{5}$$

Moreover, the functions $1, t, 2t^2, \cdots, (k!)t^k, (1 + st)^{-1}$ form a complete Tchebycheff system, so extremal distributions $F$ within the class of mixtures are obtained by using the associated extremal mixing cdf's $G$. If the first $n$ moments of $F$ are specified as $m_1, m_2, \cdots, m_n$, then the first $n$ moments of $G$ are $m_1, m_2/2, \cdots, m_n/n!$

First suppose that the two moments of $F$ are specified as $m_1$ and $m_2$. Let $c^2$ be the squared coefficient of variation of $F$, i.e., $c^2 = (m_2 - m_1^2)/m_1^2$. Also require that the mixing cdf $G$ has support on the interval $[0, bm_1]$, $b < \infty$. Then the extremal distributions are:

1. Upper bound—the two-point mixture with mass $(c^2 - 1)/(c^2 + 1)$ on 0 and mass $2/(c^2 + 1)$ on the exponential distribution with mean $m_1(1 + c^2)/2$, which has cdf

$$F_{\mathscr{U}}(x) = 1 - [2/(1 + c^2)]\mathrm{e}^{-2x/m_1(1+c^2)}, \qquad x \ge 0, \tag{6}$$

and

2. Lower bound—the mixture of two exponential distributions, one having mean $bm_1$ with probability $(c^2 - 1)/(c^2 - 1 + 2(b - 1)^2)$ and the other having mean $m_1[1 - (c^2 - 1)/2(b - 1)]$ with probability $2(b - 1)^2/(c^2 - 1 + 2(b - 1)^2)$; the cdf is

$$\begin{aligned} F_{\hat{\mathscr{L}}}(x) = 1 - [c^2 - 1 + 2(b - 1)^2]^{-1}\{&(c^2 - 1)\mathrm{e}^{-bm_1x} \\ &+ 2(b - 1)^2\mathrm{e}^{-m_1[1 - (c^2 - 1)/2(b - 1)]x}\}, \quad x \ge 0. \end{aligned} \tag{7}$$

As $b \longrightarrow \infty$, the lower bound approaches (converges in law) to

3. Limiting lower bound—the exponential distribution with mean $m_1$, having cdf $F_{\hat{\mathscr{L}}}(x) = 1 - \mathrm{e}^{-m_1x}$, $x \ge 0$.

The upper bound cdf $F_{\mathscr{u}}$ may not be considered a mixture of exponential distributions because of the atom at 0, but the atom at 0 can be thought of as an exponential distribution having mean 0. Alternatively, $F_{\mathscr{u}}$ can be realized as the limit in distribution of mixtures of two exponential distributions having means $\lambda_1^{-1}$ and $\lambda_2^{-2}$ and proper moments where $\lambda_1^{-1} \rightarrow 0$ and $\lambda_2^{-1} \rightarrow m_1(1 + c^2)/2$.

Let $\phi_{\hat{\ell}}(s)$, $\phi_{\ell}(s)$, and $\phi_{\mathscr{u}}(s)$ be the transforms of the extremal cdf's $F_{\hat{\ell}}$, $F_{\ell}$, and $F_{\mathscr{u}}$, respectively. The theory of Tchebycheff systems implies that

$$\phi_{\hat{\ell}}(s) \leq \phi_{\ell}(s) \leq \phi(s) \leq \phi_u(s) \tag{8}$$

for all $s$ and the transforms $\phi$ of cdf's $F$ of the form (3) having first two moments $m_1$ and $m_2$.

*Remark*: It is no doubt possible to study extremal distributions for other kinds of mixtures, but we have not. Mixtures of exponentials seem particularly appropriate for the queueing application.

## IV. THE H/M/1 QUEUE

The results of Section III apply immediately to GI/M/1 queues in which the interarrival-time distribution is a mixture of exponential distributions; see Section II of Part I. Since the mixture of $k$ exponential distributions is called hyperexponential and is denoted by $H_k$, we use $H$ to refer to interarrival-time distributions that are general mixtures of exponentials.

Note that the upper bound cdf $F_{\mathscr{u}}$ in (6) as an interarrival-time distribution corresponds to a batch Poisson arrival process with geometrically distributed batches having mean $m_B = (1 + c^2)/2$ and squared coefficient of variation $c_B^2 = (m_B - 1)/m_B$. Let $M^B$ represent a batch Poisson arrival process. Of course, the limiting lower bound corresponds to a Poisson arrival process with intensity $1/m_1$. What we obtain is the ordering

$$\mathrm{M/M/1} \leq \mathrm{H/M/1} \leq \mathrm{M^B/M/1}, \tag{9}$$

which means that the mean queue lengths (expected number in the system, including any in service) are ordered and in fact the entire steady-state queue-length distributions are stochastically ordered as in (9), provided the traffic intensity $\rho$ and the squared coefficient of variation of the interarrival-time distribution, $c^2$, are fixed. We obtain these orderings because in the case of exponential service-time distributions the entire steady-state queue-length distribution depends only on the traffic intensity $\rho$, which is fixed, and the root $\sigma$ in the interval (0, 1) of the equation

$$\phi[\mu(1 - \sigma)] = \sigma. \tag{10}$$

It is easy to see that the queue-length distributions $P(Q_i \leq k)$ are stochastically ordered, i.e.,

$$P(Q_1 \geq k) \leq P(Q_2 \geq k) \quad \text{for all} \quad k \geq 0 \tag{11}$$

if the roots satisfy $\sigma_1 < \sigma_2$. Moreover, it is easy to see that the roots are ordered if the transforms are ordered in the sense (8).

Let $\sigma_u$ and $L_u$ be the probability of delay and mean queue length in the H/M/1 queue with interarrival-time distribution $F_u$, and similarly for $F_\ell$ and $F_{\hat{\ell}}$. Here are the main results:

*Theorem 1: For an H/M/1 queue with traffic intensity $\rho$ and interarrival-time squared coefficient of variation $c^2$,*

$$\sigma_{\hat{\ell}} = \rho \quad \textit{and} \quad \sigma_u = 1 - 2(1 - \rho)/(1 + c^2), \tag{12}$$

*so that*

$$L_{\hat{\ell}} = \rho/(1 - \rho), \qquad L_u = L_{\hat{\ell}}(1 + c^2)/2 \tag{13}$$

*and the maximum relative error (MRE) is*

$$MRE \equiv (L_u - L_{\hat{\ell}})/L_{\hat{\ell}} = (\sigma_u - \sigma_{\hat{\ell}})/(1 - \sigma_u) = (c^2 - 1)/2. \tag{14}$$

*Proof:* Since $\sigma = \rho$ for an M/M/1 queue, $\sigma_{\hat{\ell}} = \rho$. For $\sigma_u$, follow the proof of Theorem 2 in Part I, making the change of variables $(1 - \sigma_{\hat{\ell}}) = (1 - \sigma_u)(1 + c^2)/2$.

From Corollary 1 of Part I and Theorem 1, we see that the shape constraint reduces the maximum relative error from $c^2$ to $(c^2 - 1)/2$. If $c^2$ is near its lower limit 1 for mixtures of exponentials, then of course the MRE is very small.

Given the first two moments, the upper bound is hard and the lower bound is soft: The upper bound depends on $c^2$; the lower bound does not. The upper bound is not improved by specifying the third moment; the lower bound is. From Section IV of Part I, we see that the extremal distributions given three moments are two-point mixtures of exponentials:

*Theorem 2: For H/M/1 queues, specifying the third moment of the interarrival-time distribution in addition to the first two does not change the upper bound cdf $F_u$ and makes the lower bound cdf $F_{\hat{\ell}}$ the unique $H_2$ distribution (two-point mixing distribution) specified by these three parameters.*

The formula for calculating $H_2$ parameters given the first three moments is given in (3.5) and (3.6) of Ref. 6.

*Example 1:* Consider an interarrival-time distribution with moments $m_1 = 2.00$, $m_2 = 12.00$, and $m_3 = 119.01$, which are the moments of Prototype Distribution I in Part II. With mixtures of exponential distributions, the upper bond cdf is

$$F_{u}(x) = 1 - 0.6667e^{-.3333x}, \qquad x \geq 0,$$

and the lower bound cdf is

$$F_{\ell}(x) = 1 - 0.5146e^{-0.2964x} + 0.4854e^{-1.8386x}, \qquad x \geq 0.$$

From Theorem 1, given just the first two moments, $\sigma_{\ell} = 0.6667$ and $\sigma_{u} = 0.7778$ for $\rho = 0.6667$ and $\sigma_{\ell} = 0.9000$ and $\sigma_{u} = 0.9333$ for $\rho = 0.9000$. From Theorem 2, also specifying the third moment changes the lower bound to $\sigma_{\ell} = 0.76705$ for $\rho = 0.6667$ and $\sigma_{\ell} = 0.93259$ for $\rho = 0.9000$. To get these, we solved the appropriate $H_2$/M/1 queue. Imposing the shape constraint in addition to the first two moments reduced the MRE from $c^2 = 2.0$ to $(c^2 - 1)/2 = 0.50$. Also specifying the third moment further reduces the MRE to 0.048 when $\rho = 2/3$ and 0.011 when $\rho = 9/10$.

## V. MIXTURES OF TWO EXPONENTIALS: $H_2$ DISTRIBUTIONS

Mixtures of two exponential distributions, i.e., $H_2$ distributions, play a key role in many approximations. This is a three-parameter distribution with density

$$f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}, \qquad x > 0, \tag{15}$$

where $p_2 = 1 - p_1$. Instead of the three parameters $p_1$, $\lambda_1$, and $\lambda_2$, one may choose to work with the first three moments $m_1$, $m_2$, and $m_3$ or the mean $m_1$, the squared coefficient of variation $c^2$, and the proportion of the total mean in the component with the smaller mean $r$, defined by

$$r = \frac{p_1/\lambda_1}{(p_1/\lambda_1) + (p_2/\lambda_2)}, \tag{16}$$

where $\lambda_1 > \lambda_2$. Given the parameters $p_1$, $\lambda_1$, and $\lambda_2$, it is easy to calculate any of the other parameters. The formulas for $p_1$, $\lambda_1$, and $\lambda_2$ given the first three moments appear in (3.5) and (3.6) of Ref. 6. Given $m_1$, $c^2$, and $r$, $m_2 = m_1^2(c^2 + 1)$, $p_1 = rm_1\lambda_1$, $\lambda_2 = (1 - rm_1\lambda_1)/(1 - r)m_1$ and

$$\lambda_1 = (-B + \sqrt{B^2 - 4AC})/2A, \tag{17}$$

where $A = rm_1m_2/2$, $-B = (m_2/2) + (rm_1)^2 - (1 - r)^2m_1^2$, and $C = rm_1$.

For two-moment approximations based on $H_2$ distribution, one of the three parameters is often eliminated by setting $r = 1/2$; see Section 3.1 of Ref. 6. The range of all possible values given the first two moments is indicated in Section IV since both the upper and lower bounds are $H_2$ distributions. Since this range is pretty wide, it is natural to ask how the distribution and the GI/M/1 queue characteristics vary with the third parameter—either $r$ or $m_3$. For what values of $r$ is the approximation by $r = 1/2$ reasonable?

In order to answer this question, we have calculated the third moment $m_3$ and the queue characteristics $\sigma$ and $L$ for two values of $c^2$ (2 and 12), three values of $\rho$ (0.3, 0.7, and 0.9), and thirteen values of $r$ (ranging from 0.001 to 0.999). The results appear in Tables I and II.

For $c^2 = 2.0$, the approximation by $r = 1/2$ appears quite robust. For $r$ in the interval [0.2, 0.8], the maximum relative error is 15.8

Table I—The possible third parameters and queue characteristics for an $H_2/M/1$ queue given $c^2 = 2.0$ with $\rho = 0.3$, 0.7, and 0.9

| Proportion of Total Mean in Component With Smaller Mean, $r$ | Skewness, Third Moment $m_3/m_1^3$ | Key Root, Probability of Delay $\sigma$ | | | Mean Queue Length, $L$ | | |
|---|---|---|---|---|---|---|---|
| | | $\rho = 0.3$ | $\rho = 0.7$ | $\rho = 0.9$ | $\rho = 0.3$ | $\rho = 0.7$ | $\rho = 0.9$ |
| Upper bound | 13.5 | 0.5333 | 0.8000 | 0.9333 | 0.643 | 3.500 | 13.500 |
| 0.001 | 13.5 | 0.5323 | 0.7999 | 0.9333 | 0.641 | 3.499 | 13.499 |
| 0.01 | 13.6 | 0.5230 | 0.7992 | 0.9333 | 0.629 | 3.486 | 13.486 |
| 0.10 | 14.6 | 0.4627 | 0.7933 | 0.9327 | 0.558 | 3.386 | 13.381 |
| 0.20 | 15.4 | 0.4280 | 0.7885 | 0.9323 | 0.525 | 3.309 | 13.291 |
| 0.30 | 16.2 | 0.4059 | 0.7842 | 0.9319 | 0.505 | 3.244 | 13.210 |
| 0.40 | 17.1 | 0.3894 | 0.7801 | 0.9314 | 0.491 | 3.183 | 13.127 |
| 0.50 | 18.0 | 0.3757 | 0.7757 | 0.9310 | 0.481 | 3.121 | 13.036 |
| 0.60 | 19.2 | 0.3633 | 0.7707 | 0.9304 | 0.471 | 3.053 | 12.924 |
| 0.70 | 20.9 | 0.3512 | 0.7643 | 0.9295 | 0.462 | 2.970 | 12.771 |
| 0.80 | 23.9 | 0.3382 | 0.7552 | 0.9281 | 0.453 | 2.860 | 12.522 |
| 0.90 | 32.1 | 0.3226 | 0.7394 | 0.9248 | 0.443 | 2.686 | 11.966 |
| 0.99 | 167.9 | 0.3029 | 0.7065 | 0.9074 | 0.430 | 2.385 | 9.715 |
| 0.999 | 1518.0 | 0.3003 | 0.7007 | 0.9009 | 0.429 | 2.339 | 9.080 |
| Lower bound | ∞ | 0.3000 | 0.7000 | 0.9000 | 0.429 | 2.333 | 9.000 |

Table II—The possible third parameters and queue characteristics for an $H_2/M/1$ queue given $c^2 = 12.0$ with $\rho = 0.3$, 0.7, and 0.9

| Proportion of Total Mean in Component With Smaller Mean, $r$ | Skewness, Third Moment $m_3/m_1^3$ | Key Root, Probability of Delay $\sigma$ | | | Mean Queue Length, $L$ | | |
|---|---|---|---|---|---|---|---|
| | | $\rho = 0.3$ | $\rho = 0.7$ | $\rho = 0.9$ | $\rho = 0.3$ | $\rho = 0.7$ | $\rho = 0.9$ |
| Upper bound | 253.5 | 0.8923 | 0.9539 | 0.9846 | 2.789 | 15.17 | 58.50 |
| 0.001 | 253.8 | 0.8921 | 0.9538 | 0.9846 | 2.779 | 15.16 | 58.49 |
| 0.01 | 256.1 | 0.8897 | 0.9536 | 0.9846 | 2.721 | 15.10 | 58.43 |
| 0.10 | 280.7 | 0.8590 | 0.9516 | 0.9844 | 2.128 | 14.48 | 57.80 |
| 0.20 | 312.6 | 0.8006 | 0.9488 | 0.9842 | 1.505 | 13.68 | 56.99 |
| 0.30 | 351.6 | 0.7114 | 0.9451 | 0.9839 | 1.040 | 12.74 | 56.01 |
| 0.40 | 401.2 | 0.6142 | 0.9396 | 0.9836 | 0.778 | 11.59 | 54.76 |
| 0.50 | 468.0 | 0.5311 | 0.9311 | 0.9831 | 0.640 | 10.16 | 53.10 |
| 0.60 | 565.1 | 0.4650 | 0.9163 | 0.9823 | 0.561 | 8.36 | 50.73 |
| 0.70 | 722.7 | 0.4120 | 0.8881 | 0.9808 | 0.510 | 6.25 | 47.00 |
| 0.80 | 1031.7 | 0.3686 | 0.8380 | 0.9776 | 0.475 | 4.32 | 40.20 |
| 0.90 | 1946.0 | 0.3319 | 0.7708 | 0.9646 | 0.449 | 3.05 | 25.39 |
| 0.99 | 18,287. | 0.3030 | 0.7070 | 0.9089 | 0.430 | 2.39 | 9.88 |
| 0.999 | 181,638. | 0.3003 | 0.7007 | 0.9009 | 0.429 | 2.34 | 9.08 |
| Lower bound | ∞ | 0.3000 | 0.7000 | 0.9000 | 0.429 | 2.33 | 9.00 |

percent, 15.7 percent, and 6.1 percent for $\rho = 0.3$, 0.7, and 0.9. Very large values of $r$ greatly extend the range.

On the other hand, for very large values of $c^2$ such as 12, the approximation by $r = 1/2$ is not robust: two moments do not pin down the $H_2$ distribution well. Using $r = 1/2$ as an approximation works better as $\rho$ increases and $c^2$ decreases. Of course, by Theorem 1, $\rho$ plays no role in the MRE over all $r$, but if we bound $r$, then $\rho$ plays a role. We interpret these results as providing support for $H_2$ approximation based on $r = 1/2$, but large values of $m_2$ or $m_3$ are clear danger signals.

*Example 2:* Example 1 was based on Prototype Distribution I from Part II. Since Prototype I is a discrete probability mass function it is not a mixture of exponential distributions, and is thus not entirely satisfactory. Suppose we use the $H_2$ density with balanced means ($r = 0.5$) as a prototype instead. With $m_1 = 1$ and $c^2 = 2.0$, the prototype $H_2$ density is

$$f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}, \qquad x \geq 0,$$

where

$$p_1 = 0.78867,\ \lambda_1 = 1.577, \text{ and } \lambda_2 = 0.42265.$$

Given the first two moments with $\rho = 0.7$ and 0.9, $\sigma_{u}$ can be obtained from Table II of Part I and $\sigma_{\hat{\ell}}$ can be obtained from Theorem 2 there. The values are $\sigma_{\hat{\ell}} = 0.466$ and $\sigma_{\hat{u}} = 0.822$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.808$ and $\sigma_{\hat{u}} = 0.936$ for $\rho = 0.9$. The corresponding extremal characteristics among $H_2$ densities are $\sigma_{\hat{\ell}} = 0.700$ and $\sigma_{u} = 0.800$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.900$ and $\sigma_{u} = 0.933$ for $\rho = 0.9$.

The third moment 18.0 (see Table I) pins down the $H_2$ distribution, but among all H densities it is a lower bound. Among H densities with $m_3 = 18.0$, $\sigma_{\ell} = 0.7757$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.9310$ for $\rho = 0.9$. The MRE given only two moments is 200 percent for $\rho = 0.7$ and 0.9. Working with mixtures of exponentials reduces the MRE to 50 percent. Specifying the third moment too reduces the MRE to 12 percent for $\rho = 0.7$ and 3 percent for $\rho = 0.9$.

## VI. THE H/G/1 QUEUE

The assumption of exponential service-time distributions played a crucial role in Section IV. With exponential service-time distributions, the mean queue length $L$ depends on the transform of the interarrival-time distribution, so that we can apply the ordering in (8). However, it turns out that the ordering in (9) also applies for the mean queue length with general service-time distributions, i.e., we have

$$M/G/1 \leq H/G/1 \leq M^B/G/1, \tag{18}$$

by which we mean that $L_{\hat{\ell}} \leq L \leq L_{\hat{u}}$ (but not the more general stochastic order) for all systems with common service-time distribution and given first two moments of the interarrival-time distribution.

To obtain (18), it suffices to observe that known formulas for $L$ in the $M^B/G/1$ and M/G/1 systems agree with previously established lower and upper bounds for $L$ in GI/G/1 queues having interarrival-time distributions with increasing mean residual life and with the first two moments of the interarrival times and service times specified. (See Ref. 7 for more details.) This result dramatically demonstrates that these papers have applicability beyond the special case of the GI/M/1 model.

## VII. THE $K_2$/H/1 QUEUE

Whenever the interarrival-time distribution or the service-time distribution in a GI/G/1 queue has a Laplace-Stieltjes transform that is a rational function, then the steady-state distribution can be characterized in terms of the roots of an equation involving the transforms of the interarrival-time and service-time distributions; see II.5.10,11 of Cohen.[8] When the interarrival-time distribution has a rational transform with a denominator of degree 2, denoted by $K_2$, the mean queue length and the probability of delay depend on the service-time distribution only through its first two moments and a single root of an equation involving the transforms of the interarrival-time and service-time distributions; see p. 330 of Cohen,[8] Section V of Part I,[1] and Ref. 9.

Hence, for $K_2$/G/1 queues it is possible to find extremal service-time distributions using the ordering of transforms in (8). Let $GE_2$ denote the convolution of two exponential distributions (an Erlang, $E_2$, is a special case), which is $K_2$. An $H_2$ distribution is also $K_2$. Paralleling Section V of Part I, we obtain from the analysis in Ref. 9 that

$$GE_2/M^B/1 \leq GE_2/H/1 \leq GE_2/\hat{M}/1 \tag{19}$$

and

$$H_2/\hat{M}/1 \leq H_2/H/1 \leq H_2/M^B/1, \tag{20}$$

by which we mean that the mean queue lengths are ordered as indicated. A significant feature of (19) and (20) is that the maximizing distributions are different for the different $K_2$ interarrival-time distributions. (This is explained in Ref. 9.) By $\hat{M}$, we mean the extremal service-time distribution $F_{\hat{\ell}}$ for large $b$. As $b \rightarrow \infty$, the distribution approaches the exponential distribution, but the fixed variance of $F_{\hat{\ell}}$ is lost in the limit. As $b \rightarrow \infty$, the key root in the equation for the $K_2$/G/1 queue approaches the root for the $K_2$/M/1 queue, but the mean queue length also depends on the variance of $F_{\hat{\ell}}$. The mean queue length in the $K_2/\hat{M}/1$ system is the limit as $b \longrightarrow \infty$ of the mean queue length in the $K_2/G/1$ system with service-time distributions $F_{\hat{\ell}}$. This limiting mean queue length can be computed by using the fixed service-time variance together with the root for the $K_2/M/1$ system.[8,9]

As in Section V, if we specify three service-time moments instead of two, the $M^B$ bound is unchanged, but the $\hat{M}$ bound is replaced by the $H_2$ distribution uniquely determined by the three moments, i.e., with the interarrival-time distribution and *three moments* of the service time specified, we get

$$GE_2/M^B/1 \le GE_2/H/1 \le GE_2/H_2/1 \qquad (21)$$

and

$$H_2/H_2/1 \le H_2/H/1 \le H_2/M^B/1. \qquad (22)$$

We conclude by exhibiting the mean queue length, $L$, for several $K_2/H_2/1$ queues. We consider five different $H_2$ service-time distributions with a common mean 0.7 and a common squared coefficient of variation $c_s^2 = 2.0$. (We use subscripts "s" and "a" to indicate that parameters are associated with the service-time distribution or the interarrival-time distribution.) As in Section V, the $H_2$ distributions are characterized by the parameter $r_s$. We consider distributions close to the two extremal distributions $M^B$ ($r_s = 0.01$) and $\hat{M}$ ($r_s = 0.99$), as well as the intermediate values $r_s = 0.1$, 0.5, and 0.9. The case $r_s = 1.0$ differs from the exponential distribution because the small mass at a large value, necessary to have $c^2 = 2.0$ instead of 1.0, still has an effect. (This is not the case for the $H_2$ interarrival-time distributions.)

We consider six interarrival-time distributions: the same five $H_2$ distributions and the Erlang ($E_2$) distribution. All the interarrival-time distributions have mean 1.0, so that the traffic intensity is always $\rho = 0.7$. As with the service-time distributions, the $H_2$ interarrival-time distributions have squared coefficient of variation $c_a^2 = 2.0$.

The results for the 30 cases are displayed in Table III. For the extremal H interarrival-time distributions, $M^B$ and M, the mean queue length, $L$, does not depend on $r_s$ because $L$ depends on the service-time distribution only through its first two moments.[7] The range of $L$ values over $r_s$ increases for $H_2$ interarrival-time distributions as $r_a$ moves away from the endpoints 0.0 and 1.0. The range is bigger for $c_a^2 = 2.0$ ($H_2$) than for $c_a^2 = 0.5$ ($E_2$) when $r_a = 0.5$, but obviously not for all $r_a$.

Table III gives an indication of the quality of two-moment approximations for GI/G/1 queues when $c_a^2 = c_s^2 = 2.0$ and $\rho = 0.7$. A natural two-moment approximation would be based on the $H_2/H_2/1$ queue

Table III—The mean queue length, $L$, in several $K_2/H_2/1$ systems with traffic intensity $\rho = 0.7$

| | | | Service-Time Distribution | | | | |
|---|---|---|---|---|---|---|---|
| | | | $(M^B)$ | Hyperexponential $(H_2)$ | | | $(\hat{M})$ |
| | | | $r_s = 0.01$ | $r_s = 0.1$ | $r_s = 0.5$ | $r_s = 0.9$ | $r_s = 0.99$ |
| Interarrival-time distribution | $E_2$ | | 2.61 | 2.62 | 2.63 | 2.63 | 2.63 |
| | (M) | $r_a = 1.0$ | 3.15 | 3.15 | 3.15 | 3.15 | 3.15 |
| | Hyperexponential $(H_2)$ | $r_a = 0.9$ | 3.60 | 3.60 | 3.59 | 3.56 | 3.52 |
| | | $r_a = 0.5$ | 4.02 | 4.01 | 3.99 | 3.96 | 3.94 |
| | | $r_a = 0.1$ | 4.23 | 4.23 | 4.21 | 4.21 | 4.20 |
| | $(M^B)$ | $r_a = 0.0$ | 4.32 | 4.32 | 4.32 | 4.32 | 4.32 |

Notes: 1. The hyperexponential $(H_2)$ distributions all have squared coefficient of variation $c^2 = 2.00$. 2. The Erlang $(E_2)$ distribution has squared coefficient of variation $c^2 = 0.5$. 3. The $\hat{M}$ service-time distribution differs from an exponential distribution because of the small mass at a very large value. This causes the $H_2/\hat{M}/1$ values of $L$ to differ from the $H_2/M/1$ values of $L$ in Table I.

with $c_a^2 = c_s^2 = 2.0$ and $r_a = r_s = 0.5$. The range of $H_2/H_2/1$ values as $r_a$ and/or $r_s$ varies indicates the possible deviations from the approximations when the distributions are required to be mixtures of exponential distributions. The maximum relative error is (4.32-3.15)/3.15 or 37 percent, but would be much less if we restricted $r_a$ and $r_s$ to some reasonable interval, e.g., [0.2, 0.8].

Table III enables us to compare the effect of additional information about the interarrival-time and service-time distributions. Table III shows that, given two moments, other properties of the distribution are much more important for the interarrival-time distribution than for the service-time distribution in determining the mean queue length. This phenomenon was previously noted by Sahin and Perrakis.[10]

The program for calculating the mean queue length and the probability of delay in a $K_2/G/1$ queue used to obtain Table III is being used as part of a three-parameter procedure for approximating general G/G/1 queues with bursty, possibly nonrenewal arrival processes.[11] The general bursty arrival process is approximated by a renewal process with an $H_2$ interarrival-time distribution, which is characterized completely by the first three moments of the renewal interval.[6] Then the expected queue length and probability of delay are calculated exactly for the resulting $H_2/G/1$ model. Additional descriptions of the $H_2/G/1$ queue, such as an entire waiting-time distribution, are obtained using approximations similar to the ones in the software package QNA (see Section 5.1 of Ref. 12). This approach is part of a new three-parameter algorithm for QNA.

## VIII. ACKNOWLEDGMENT

This research was motivated by the results in Part II obtained jointly with John Klincewicz. I am grateful to him for his assistance. I am also grateful to Richard Stubing for writing programs to obtain the data in Tables I through III.

## REFERENCES

1. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," AT&T Bell Lab. Tech. J., this issue.
2. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," AT&T Bell Lab. Tech. J., this issue.
3. A. E. Eckberg, Jr., "Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. Math. Oper. Res., *2,* No. 2 (May 1977), pp. 135–42.
4. S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, New York: John Wiley and Sons, 1966.
5. J. Keilson, *Markov Chain Models—Rarity and Exponentiality,* New York: Springer-Verlag, 1979.
6. W. Whitt, "Approximating a point process by a renewal process, I: two basic methods," Oper. Res., *30,* No. 1 (January–February 1982), pp. 125–47.
7. W. Whitt, "The Marshall and Stoyan bounds for IMRL/G/1 queues are tight," Oper. Res. Letters, *1,* No. 6 (December 1982), pp. 209–13.
8. J. W. Cohen, *The Single Server Queue,* Amsterdam: North-Holland, 1969.
9. W. Whitt, "Minimizing delays in the GI/G/1 queue," Oper. Res., to be published.
10. I. Sahin and S. Perrakis, "Moment inequalities for a class of single server queues," INFOR, *14,* No. 2 (June 1976), pp. 144–52.
11. W. Whitt, unpublished work.
12. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62,* No. 9 (November 1983), pp. 2779–815.

## AUTHOR

**Ward Whitt,** A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and congestion models.

AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

# Computing Inductive Noise of Chip Packages

By A. J. RAINAL*

(Manuscript received July 30, 1982)

Inductive noise limits the physical design of high-speed, high pin-out chip packages. This paper presents the derivation of some basic equations that are useful for computing inductive noise of various chip packages, and also presents simple asymptotic and limiting results that reduce to some useful approximate results proposed by others. These results are helpful for computing inductive noise in arrays of wire bonds, solder balls, dual in-line package leads, package pins, and connector pins. Computed results agreed well with measured results. We present two simple rules for minimizing inductive noise and also discuss the inductive noise of power and ground planes.

## I. INTRODUCTION

If $n$ drivers each switch current at $\dot{I} = 20$ mA/ns, the inductive noise across a common ground lead inductance, $L_g$, is approximately $nL_g\dot{I}$. For a 32-bit processor and $L_g = 1$ nh, this inductive noise component is about (32)(1 nh) 20mA/ns = 640 mV. It is known that a 50-mil-long wire bond used as an input/output (I/O) lead of an integrated circuit chip has a self-inductance of about 1 nh. Thus, many such leads must be connected in parallel to reduce this inductive noise component to tolerable levels. This is necessary because present large-scale integrated (LSI) circuits have a total noise margin of only a few hundred millivolts. This inductive noise problem has been recognized and discussed by C. W. Deisch.[1]

This paper derives some general equations that are useful for computing the inductive noise of high-speed, high pin-out chip packages.

* AT&T Bell Laboratories.

A basic role is played by the mutual inductance between two parallel conductors.

## II. MUTUAL INDUCTANCE OF TWO PARALLEL CONDUCTORS

Consider the two parallel conductors shown in Fig. 1. By applying the Biot-Savart law,[2] a current $I$ flowing in the $y$ direction produces a magnetic flux density, $B(x, y)$, given by

$$B(x, y) = \frac{\mu I}{4\pi} \int_{-\ell/2}^{\ell/2} \frac{\sin\theta}{r^2}\, dy_0 = \frac{\mu I}{4\pi} \int_{-\ell/2}^{\ell/2} \frac{x}{r^3}\, dy_0 \tag{1}$$

$$= \frac{\mu I}{4\pi x}\left[\frac{(y + \ell/2)}{\sqrt{(y + \ell/2)^2 + x^2}} - \frac{(y - \ell/2)}{\sqrt{(y - \ell/2)^2 + x^2}}\right], \tag{2}$$

where $\mu$ = permeability of the medium. The total flux, $\Lambda$, linking the idle conductor is then given by

$$\Lambda \equiv \int_d^{\infty} dx \int_{-\ell/2}^{\ell/2} B(x, y)dy$$

$$= \frac{\mu I \ell}{2\pi}\left[ln\left(\frac{\ell}{d} + \sqrt{1 + \left(\frac{\ell}{d}\right)^2}\right) - \sqrt{1 + \left(\frac{d}{\ell}\right)^2} + \frac{d}{\ell}\right]. \tag{3}$$

If the medium is a vacuum or air, then $\mu = \mu_0 = (4\pi)10^{-7}$ h/m and the mutual inductance, $M$, is given by

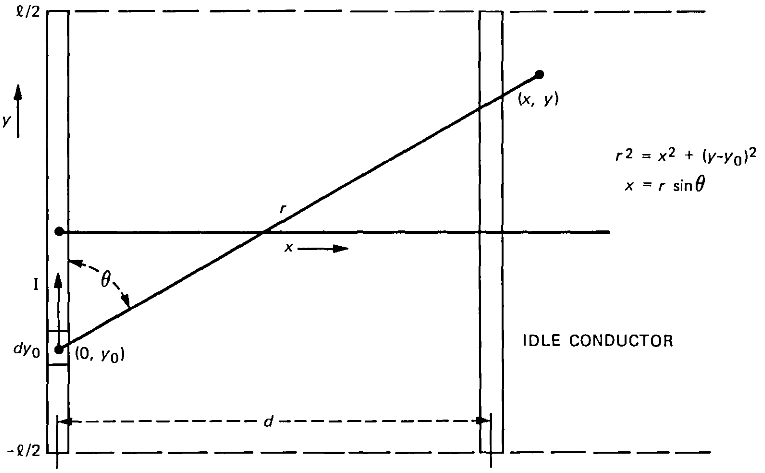


Fig. 1—Coordinate system for derivation of mutual inductance.

$$M \equiv \frac{\Lambda}{I} \doteq 5\ell \left[ \ln\left(\frac{\ell}{d} + \sqrt{1 + \left(\frac{\ell}{d}\right)^2}\right) - \sqrt{1 + \left(\frac{d}{\ell}\right)^2} + \frac{d}{\ell} \right] \text{nh}, \qquad (4)$$

where

$\mu_0/2\pi \doteq 5$ nh/in

$\ell$ = length in inches

$d$ = separation in inches.

A useful asymptotic result for small $d/\ell$ is given by

$$M \sim 5\ell \left[ \ln\left(\frac{2\ell}{d}\right) - 1 + \frac{d}{\ell} - \left(\frac{d}{2\ell}\right)^2 \right] \text{nh}. \qquad (5)$$

Equation (4) can also be derived by evaluating the Neumann inductance integral. Equations (4) and (5) agree with eqs. (1) and (3) of Ref. 3.

The inductances discussed in this paper are more precisely known as partial self and mutual inductances. However, we shall follow Ref. 3 and refer to them as merely self and mutual inductances.

## III. SELF-INDUCTANCE OF A STRAIGHT CONDUCTOR

### *3.1 Self-inductance resulting from the internal field*

Consider the current element shown in Fig. 2. A basic definition of self-inductance, $L$, is

$$L \equiv \frac{N\phi}{I} = \text{total number of flux linkage per ampere.} \qquad (6)$$

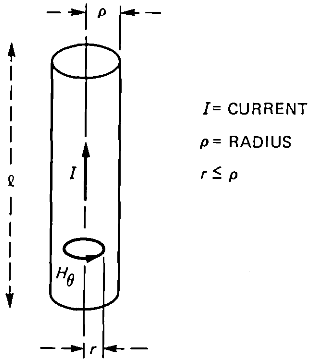


Fig. 2—Notation for derivation of self-inductance.

From Maxwell's equation[2] (i.e., Ampere's law), the magnetic intensity, $H_\theta$, internal to the conductor is given by

$$\oint \vec{H}\cdot\vec{ds} = \left(\frac{I}{\pi\rho^2}\right)(\pi r^2) = I\left(\frac{r}{\rho}\right)^2 \qquad 0 \le r \le \rho. \tag{7}$$

Equation (7) assumes that the current density, $I/(\pi\rho^2)$, is uniform in the conductor. (Skin effect is neglected. References 2, 3, and 4 show that skin effect tends to reduce the internal self-inductance, $L_i$.) From eq. (7),

$$H_\theta \equiv |\vec{H}| = \frac{I}{2\pi r}\left(\frac{r}{\rho}\right)^2 \qquad 0 \le r \le \rho. \tag{8}$$

The flux density, $B_\theta$, internal to the conductor is then

$$B_\theta \equiv \mu H_\theta = \frac{\mu I}{2\pi r}\left(\frac{r}{\rho}\right)^2 \qquad 0 \le r \le \rho. \tag{9}$$

As Ref. 4 shows, a given flux line of radius $r \le \rho$ encloses a fraction $(r/\rho)^2$ of the total current $I$. Thus, from eqs. (6) and (9) the self-inductance, $L_i$, resulting from the internal magnetic field is

$$L_i = \frac{\ell}{I}\int_0^\rho B_\theta\left(\frac{r}{\rho}\right)^2 dr = \frac{\mu\ell}{8\pi}. \tag{10}$$

Equation (10) can also be derived from internal energy considerations. If $\mu = \mu_0$, and $\ell$ is in inches,

$$L_i = \left(\frac{\mu_0}{2\pi}\right)\left(\frac{\ell}{4}\right) \doteq 5\left(\frac{\ell}{4}\right) \text{nh}. \tag{11}$$

### *3.2 Total self-inductance*

The total self-inductance, $L_s$, of a straight conductor is obtained by adding the contributions from the external and internal magnetic fields. Thus, from eqs. (4) and (11) we have

$$L_s = M|_{d=\rho} + L_i = M|_{d=\rho} + 5\left(\frac{\ell}{4}\right) \text{nh}. \tag{12}$$

Also, for small $\rho/\ell$,

$$L_s \sim 5\ell\left[ln\left(\frac{2\ell}{\rho}\right) - \frac{3}{4}\right] \text{nh}. \tag{13}$$

Equation (13) agrees with eq. (7) of Ref. 3.

When the cross section of the conductor is rectangular, Ref. 3 shows that the self-inductance is given by

$$L_s \doteq 5\ell \left[ \ln \left( \frac{4\ell}{p} \right) + \frac{1}{2} \right] \text{nh}, \tag{14}$$

where $p$ = perimeter of cross section in inches.

## IV. INDUCTANCE OF POWER AND GROUND PLANES

Consider the power and ground (P/G) planes shown in Fig. 3. Assume that both planes carry equal, thin sheets of current, $I$, in opposite directions. Again, using Ampere's law,

$$\oint \vec{H} \cdot \vec{ds} = I. \tag{15}$$

The magnetic field is more intense and approximately uniform in the space between the P/G planes. The magnetic field outside the planes is assumed to be negligible because of field cancellation. Thus,

$$|\vec{H}|\, W \equiv HW = I, \tag{16}$$

$$B \equiv \mu H = \frac{\mu I}{W}, \tag{17}$$

and

$$L \equiv \frac{N\phi}{I} = \frac{B\ell h}{I} = \frac{\mu \ell h}{W}. \tag{18}$$

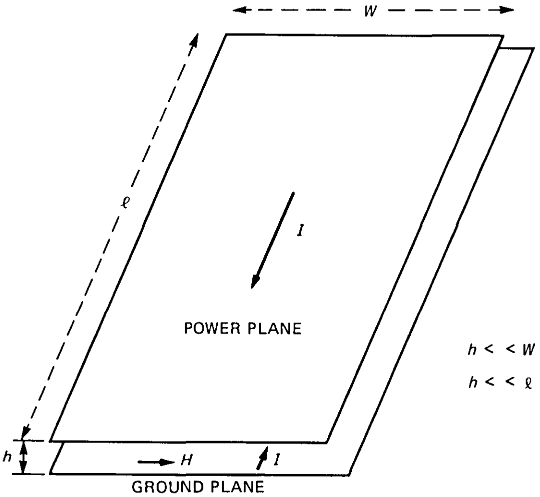


Fig. 3—Notation for derivation of inductance of P/G planes.

If $\mu = \mu_0$; and $\ell$ is in inches,

$$L = \left(\frac{\mu_0}{2\pi}\right) 2\pi\ell \left(\frac{h}{W}\right) = 10\pi\ell \left(\frac{h}{W}\right) \text{ nh.} \tag{19}$$

One half of this $L$ is associated with the ground plane and the other half is associated with the power plane. Thus, the inductance of the ground plane, $L_g$, and the inductance of the power plane, $L_p$, are given by

$$L_g = L_p = 5\pi\ell\left(\frac{h}{W}\right) \text{ nh.} \tag{20}$$

## V. COMPUTING INDUCTIVE NOISE

### *5.1 Pair of conductors*

Consider the pair of conductors shown in Fig. 4. The noise voltage, $v_n$, is a result of the self and mutual inductances of the conductors. Thus, using eqs. (12) and (4),

$$v_n = L_s\dot{I} - M\dot{I} = (L_s - M)\dot{I} \text{ mV,} \tag{21}$$

where $\dot{I} = dI/dt$ = time rate of change of current, mA/ns.

For small $d/\ell$, eqs. (5), (13), and (21) yield the asymptotic result

$$v_n \sim 5\ell\dot{I}\left[\ln\left(\frac{d}{\rho}\right) + \frac{1}{4} - \left(\frac{d}{\ell}\right) + \left(\frac{d}{2\ell}\right)^2\right] \text{mV,} \tag{22}$$

where $\ell$, $d$, $\rho$ are expressed in inches and $\dot{I}$ is expressed in mA/ns. As $d/\ell \rightarrow 0$, eq. (22) agrees with eq. (6-26) of Ref. 4 and eq. (16) of Ref. 3. Also, the first term of eq. (22), or

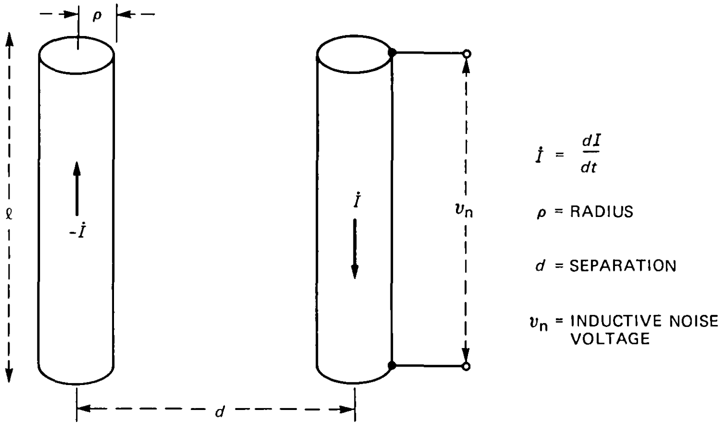


Fig. 4—Notation for derivation of inductive noise voltage, $v_n$.

$$v_{\mathrm{n}} \doteq 5\ell\dot{I}\ln\left(\frac{d}{\rho}\right)\ \mathrm{mV} \tag{23}$$

was proposed as an approximation in the work reported in Ref. 1.

Table I shows some numerical values of inductive noise voltages for a pair of conductors such that $\dot{I} = 20$ mA/ns, $\ell = 0.2$ inches, $\rho = 0.01$ inches, and $d = 0.1$, 0.2, 0.3, and 1.0 inch. These numerical results show that the "exact" inductive noise voltage is somewhat less than that given by the approximate eq. (23). However, for small $d/\ell$, the results obtained from eqs. (23) or (22) are suitable.

In general, the duration of the inductive noise voltages is approximately equal to the signal rise time.

## 5.2 Array of conductors

### 5.2.1 General equations

Consider an array of $N + 1$ conductors having equal lengths, $\ell$, equal radii, $\rho$, and separations, $d_i$, as shown in Fig. 5. Let us compute the inductive noise voltage, $v_{\mathrm{n}}$, induced in a particular conductor located

Table I—Inductive noise voltages for a pair of conductors

| $d$ | $d/\dot{I}$ | $\dot{I} = 0.2$ in.,<br>Approximate<br>(eq. 23) | $\rho = 0.01$ in.,<br>Exact<br>(eq. 21) | $\dot{I} = 20$ mA/ns,<br>Asymptotic<br>(eq. 22) |
|---|---|---|---|---|
| 0.1 in. | 0.5 | 46.0 mV | 43.25 mV | 42.3 mV |
| 0.2 | 1.0 | 60.0 | 50.4 | 49.9 |
| 0.3 | 1.5 | 68.0 | 53.3 | 54.3 |
| 1.0 | 5 | 92.1 | 57.8 | 122.1 |

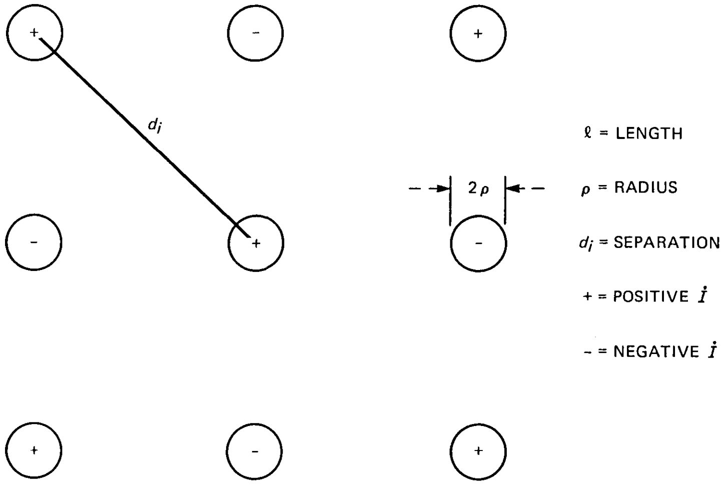


Fig. 5—Array of conductors.

at the center for convenience. Let us denote the rate of current change by $\dot{I}_i$(mA/ns), where $i = 0, \cdots, N$ with $i = 0$ denoting the particular conductor of interest.

One component of the inductive noise voltage, $v_{\mathrm{m}}$, is a result of the mutual inductances (i.e., inductive crosstalk) and is given by

$$v_{\mathrm{m}} = \sum_{i=1}^{N} M_i \dot{I}_i \text{ mV}, \tag{24}$$

where $M_i$ is given by eq. (4) with

$$M_i \equiv M|_{d=d_i}.$$

The other component of the inductive noise voltage, $v_{\mathrm{s}}$, is a result of the self-inductance of the particular conductor and is given by

$$v_{\mathrm{s}} = L_{\mathrm{s}} \dot{I}_0 \text{ mV}, \tag{25}$$

where $L_{\mathrm{s}}$ is given by eq. (12). Thus, the total inductive noise voltage, $v_{\mathrm{n}}$, induced in the particular conductor is given by

$$v_{\mathrm{n}} = v_{\mathrm{s}} + v_{\mathrm{m}} = L_{\mathrm{s}} \dot{I}_0 + \sum_{i=1}^{N} M_i \dot{I}_i \text{ mV}. \tag{26}$$

Equation (26) is the most general equation for computing the inductive noise of an array of conductors.

If the $\dot{I}_i$ in the array of conductors are constrained so that they satisfy the subsidiary condition (i.e., Kirchhoff's current law),

$$\sum_{i=0}^{N} \dot{I}_i = 0, \tag{27}$$

then eq. (26) can be written as

$$v_{\mathrm{n}} = -\sum_{i=1}^{N} (L_{\mathrm{s}} - M_i) \dot{I}_i \text{ mV}. \tag{28}$$

This equation is a generalization of eq. (21). Also, if eq. (27) holds, then for small $d_i/\ell$, eqs. (5), (13), and (26) yield the asymptotic result

$$v_{\mathrm{n}} \sim 5\ell \left\{ -\dot{I}_0 \left( \ln(\rho) - \frac{1}{4} \right) - \sum_{i=1}^{N} \dot{I}_i \left[ \ln(d_i) - \left( \frac{d_i}{\ell} \right) + \left( \frac{d_i}{2\ell} \right)^2 \right] \right\} \text{ mV}, \tag{29}$$

where $\ell$, $d_i$, $\rho$ are expressed in inches and the $\dot{I}$'s are expressed in mA/ns.

If all $d_i/\ell \to 0$, eq. (29) reduces to eq. (6-27) in Ref. 4. If all $d_i/\ell \to 0$, and $|\ln(\rho)| \gg 1/4$, eq. (29) reduces to eq. (2) of Ref. 1, namely,

$$v_{\mathrm{n}} \doteq 5\ell \left\{ -\dot{I}_0 \ln(\rho) - \sum_{i=1}^{N} \dot{I}_i \ln(d_i) \right\} \mathrm{mV}. \tag{30}$$

If all conductor separations $d_i \longrightarrow \infty$, only the self-inductance of the conductor introduces inductive noise and eq. (26) reduces to

$$v_{\mathrm{n}} = L_{\mathrm{s}} \dot{I}_0. \tag{31}$$

In contrast, eqs. (29) and (30) are not applicable if any of the $d_i \longrightarrow \infty$.

### *5.2.2 Grounded wire bonds*

As an example of computing inductive noise voltage across common ground leads in an array of conductors, consider the particular array of signal, power, and ground wire bonds on an integrated circuit chip shown in Fig. 6. Let us suppose each of the 32 signal bits switch current, simultaneously, at a rate of $\dot{I}$ mA/ns through the signal (S) wire bonds, as indicated in Fig. 6. What is the induced noise voltage, $v_{\mathrm{n}}$, across the common ground (G) leads? We shall assume that when the G leads switch, the P leads are idle. This is a property shared by many chip driver circuits. We shall also assume that the magnetic fields associated with wire bonds on different sides of the chip do not interact significantly. Finally, the chip driver circuits are assumed to be, approximately, uniformly loaded.

From Kirchhoff's current law,

$$2\dot{I}_1 + \dot{I}_2 + 8\dot{I} = 0. \tag{32}$$

From eq. (26), the voltage $v_1$, at the corner grounds is

$$v_1 = L_{\mathrm{s}}\dot{I}_1 + \dot{I}m_1 + \dot{I}m_2 + \dot{I}_2 M_{10\Delta} + \dot{I}_1 M_{20\Delta}, \tag{33}$$

where $L_{\mathrm{s}}$ is given by eq. (12),

$$m_1 = M_{2\Delta} + M_{3\Delta} + M_{7\Delta} + M_{8\Delta}$$

$$m_2 = M_{12\Delta} + M_{13\Delta} + M_{17\Delta} + M_{18\Delta},$$

and $M_{i\Delta}$ is given by eq. (4) with

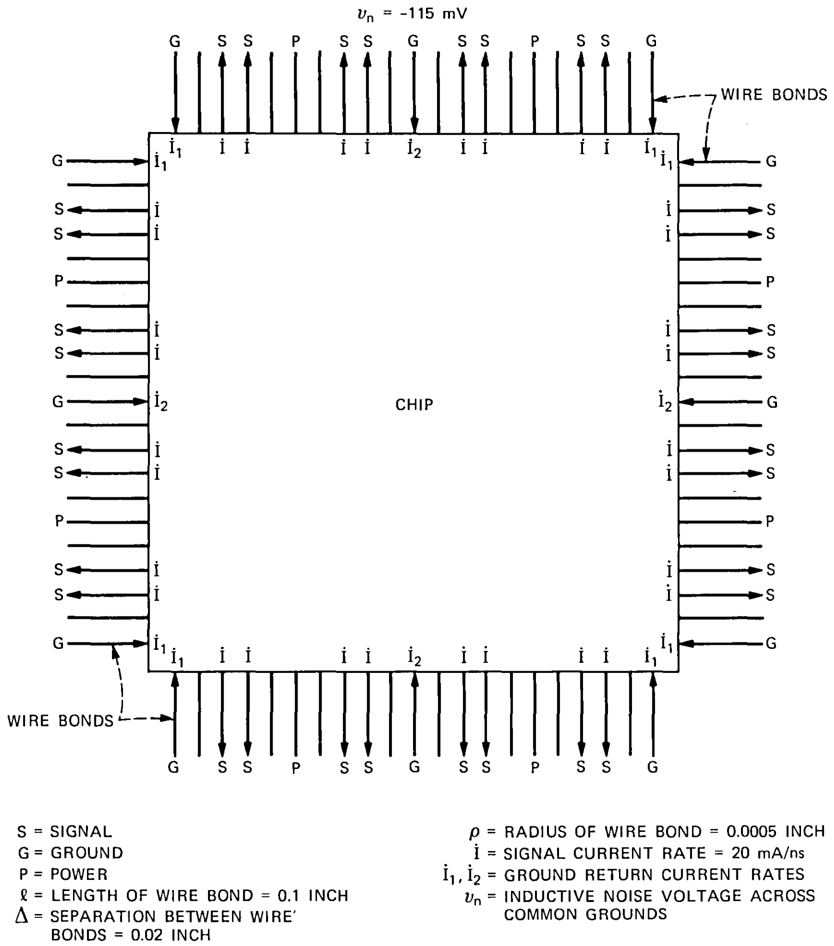
$$M_{i\Delta} \equiv M|_{d=i\Delta}.$$

Similarly, the voltage, $v_2$, at the center grounds is

$$v_2 = L_{\mathrm{s}}\dot{I}_2 + 2\dot{I}m_1 + 2\dot{I}_1 M_{10\Delta}. \tag{34}$$

By equating $v_1 = v_2$, the common ground condition, and using eq. (32), there result two simultaneous equations:

$$A_1\dot{I}_1 + B_1\dot{I}_2 = C_1\dot{I} \tag{35}$$



$v_n$ = -115 mV
G S S P S S G S S P S S G
WIRE BONDS
CHIP
$\dot{I}_1$
$\dot{I}$
$\dot{I}_2$
WIRE BONDS
G S S P S S G S S P S S G
S = SIGNAL
G = GROUND
P = POWER
ℓ = LENGTH OF WIRE BOND = 0.1 INCH
Δ = SEPARATION BETWEEN WIRE' BONDS = 0.02 INCH
ρ = RADIUS OF WIRE BOND = 0.0005 INCH
$\dot{I}$ = SIGNAL CURRENT RATE = 20 mA/ns
$\dot{I}_1, \dot{I}_2$ = GROUND RETURN CURRENT RATES
$v_n$ = INDUCTIVE NOISE VOLTAGE ACROSS COMMON GROUNDS

Fig. 6—Driver circuits on integrated circuit chip switching 32 signal bits simultaneously, with $v_n$ = –115 mV.

$$2\dot{I}_1 + \dot{I}_2 = -8\dot{I}, \tag{36}$$

where

$$A_1 = L_s + M_{20\Delta} - 2M_{10\Delta}$$

$$B_1 = M_{10\Delta} - L_s$$

$$C_1 = m_1 - m_2.$$

The solution of eqs. (35) and (36) is

$$\frac{\dot{I}_1}{\dot{I}} = \frac{C_1 + 8B_1}{A_1 - 2B_1} \tag{37}$$

$$\frac{\dot{I}_2}{\dot{I}} = \frac{-2[4A_1 + C_1]}{A_1 - 2B_1}. \qquad (38)$$

From eq. (34), the inductive noise voltage, $v_n$, across the grounded wire bonds becomes

$$v_n = v_2 = v_1 = \left[L_s\left(\frac{\dot{I}_2}{\dot{I}}\right) + 2m_1 + 2\left(\frac{\dot{I}_1}{\dot{I}}\right) M_{10\Delta}\right] \dot{I} \text{ mV}. \qquad (39)$$

If $\ell = 0.1$ inch, $\Delta = 0.02$ inch, $\rho = 0.0005$ inch and $\dot{I} = 20$mA/ns, the results are

$$\begin{aligned} L_s &= 2.623 \text{ nh} \\ \dot{I}_1 &= -2.577\dot{I} \text{ mA/ns} \\ \dot{I}_2 &= -2.846\dot{I} \text{ mA/ns} \\ m_1 &= 1.1675 \text{ nh} \\ M_{10\Delta} &= 0.1226 \text{ nh} \\ v_n &= -115.2 \text{ mV}. \qquad (40) \end{aligned}$$

As an approximation, one can assume a uniform distribution of return current rates and apply eq. (30). The results are

$$\begin{aligned} \dot{I}_1 = \dot{I}_2 &= -8\dot{I}/3 = -2.667\dot{I} \text{ mA/ns} \\ v_1 &= -122.9 \text{ mV} \\ v_2 &= -91.91 \text{ mV}. \qquad (41) \end{aligned}$$

The average of $v_1$, $v_1$ and $v_2$ is $-112.57$ mV, which is approximately equal to $v_n$ of eq. (40). This averaging method was used in Ref. 1, and it can also be used with the exact eqs. (26) or (28) to obtain approximate results.

As a nonsymmetrical example, consider the particular array of wire bonds shown in Fig. 7. To simplify the analysis, we shall neglect the mutuals on the nonsignal sides of the chip. By eq. (27), we have

$$2[\dot{I}_3 + 2\dot{I}_1 + \dot{I}_2] + 32\dot{I} = 0. \qquad (42)$$

Again, by using eqs. (26) one can write equations for $v_1$, the voltage across the corner grounds, $v_2$, the voltage across the center grounds, and

$$v_3 = L_s\left(\frac{\dot{I}_3}{3}\right). \qquad (43)$$

By equating $v_1 = v_2 = v_3$, the common ground condition, and using eq. (42), there result three independent equations for $\dot{I}_1$, $\dot{I}_2$, and $\dot{I}_3$. By

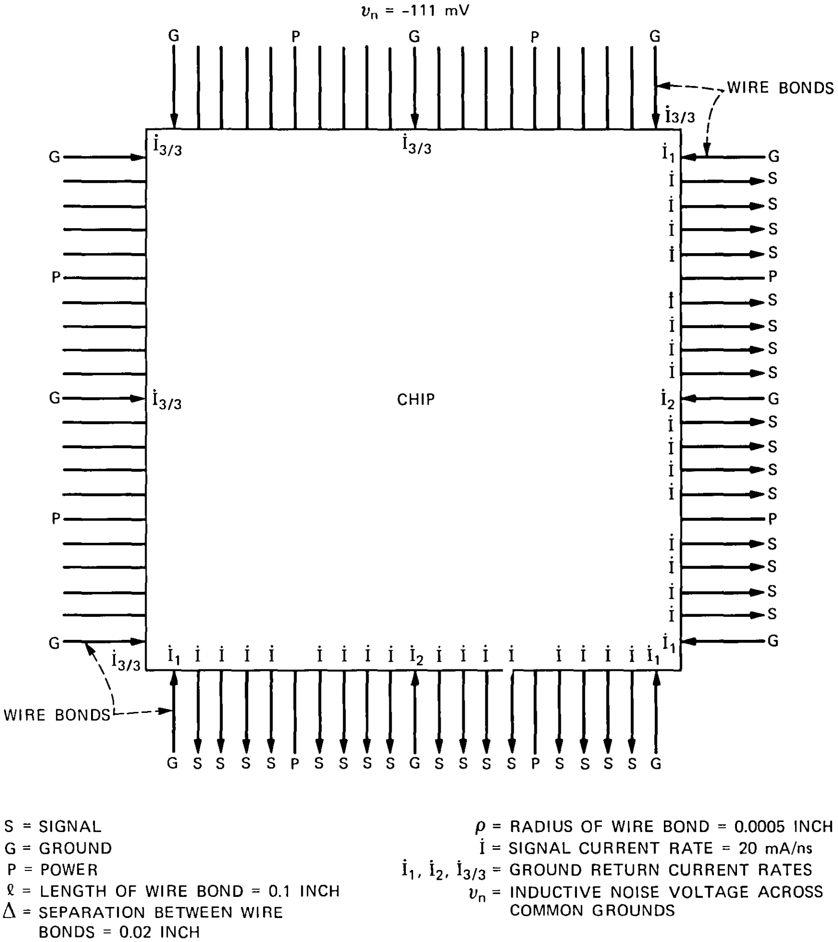

$v_n$ = -111 mV
G P G P G
WIRE BONDS
$\dot{I}_{3/3}$
$\dot{I}_1$
$\dot{I}_2$
$\dot{I}$
CHIP
G S S S S P S S S S G S S S S P S S S S G
WIRE BONDS
S = SIGNAL
G = GROUND
P = POWER
ℓ = LENGTH OF WIRE BOND = 0.1 INCH
Δ = SEPARATION BETWEEN WIRE BONDS = 0.02 INCH
ρ = RADIUS OF WIRE BOND = 0.0005 INCH
$\dot{I}$ = SIGNAL CURRENT RATE = 20 mA/ns
$\dot{I}_1$, $\dot{I}_2$, $\dot{I}_{3/3}$ = GROUND RETURN CURRENT RATES
$v_n$ = INDUCTIVE NOISE VOLTAGE ACROSS COMMON GROUNDS

Fig. 7—Driver circuits on integrated circuit chip switching 32 signal bits simultaneously, with $v_n = -111$ mV.

solving these three simultaneous equations, we can determine $\dot{I}_3$ and from eq. (43) we can determine the inductive noise voltage, $v_n$, across the grounded wire bonds. If $\ell = 0.1$ inch, $\Delta = 0.02$ inch, $\rho = 0.0005$ inch, and $\dot{I} = 20$ mA/ns, the results are

$$L_s = 2.623 \text{ nh}$$

$$\dot{I}_3 = -6.368\dot{I} \text{ mA/ns}$$

$$v_n = -111.4 \text{ mV}. \tag{44}$$

Thus, from the inductive noise point of view, the configurations shown in Figs. 6 and 7 are comparable.

In a similar manner, one can compute the inductive noise voltage across the common power or ground leads of an arbitrary array of conductors.

In general, the inductive noise voltage, $v_n$, is linear in the switching current rate, $\dot{I}$. Thus, if $\dot{I} = 10$ mA/ns, $v_n$ as given by eqs. (40) and (44) would decrease by a factor of two. Accordingly, it is very important to keep $\dot{I}$ as small as is necessary for proper circuit operation.

Also, eq. (39) shows that the self-inductance, $L_s$, of the wire bonds is a major contributor to the inductive noise, $v_n$. Equation (12) shows that $L_s$ can be reduced by reducing the length, $\ell$, of the wire bonds or increasing its radius, $\rho$.

Notice that if all the mutual inductances were to vanish, the inductive noise voltages across the grounded wire bonds of Fig. 6 and 7 would increase in magnitude to

$$v_n = -\frac{L_s(32\dot{I})}{N_g} = -\frac{L_s(32\dot{I})}{12} = -140.0 \text{ mV}, \tag{45}$$

where $N_g$ = number of chip grounds.

Thus, in these cases, the mutual inductances serve to reduce the magnitude of inductive noise by about 20 percent.

### 5.2.3 *Minimization of inductive noise*

To help minimize the magnitude of inductive noise, two general rules are now apparent:

1. Separate the P leads, and separate the G leads. Attempt to locate them symmetrically. This serves to minimize the buildup of flux linkages produced by current flow in the same direction and provides symmetry for the P/G leads.

2. Locate the signal leads as close as possible to the P/G leads. This serves to reduce flux linkages produced by current flow in opposite directions.

Similar rules were also given by C. W. Deisch.[1]

These two rules were applied to determine the P/G/S lead assignments for Fig. 8, which shows min $|v_n \leq 110$ mV. In contrast, the rules were violated drastically in Fig. 9 with the result that max $|v_n| \geq 196$ mV. By comparison with eq. (45), we see that the mutual inductances have now increased the magnitude of the inductive noise by at least 40 percent.

The two simple rules are useful to help minimize inductive noise resulting from general arrays of coupled conductors.

### 5.2.4 *Comparison with experiment*

Values of $v_m$, $v_n$ as given by eqs. (24) and (26) were found to agree well with experimental values of inductive noise voltage measured on

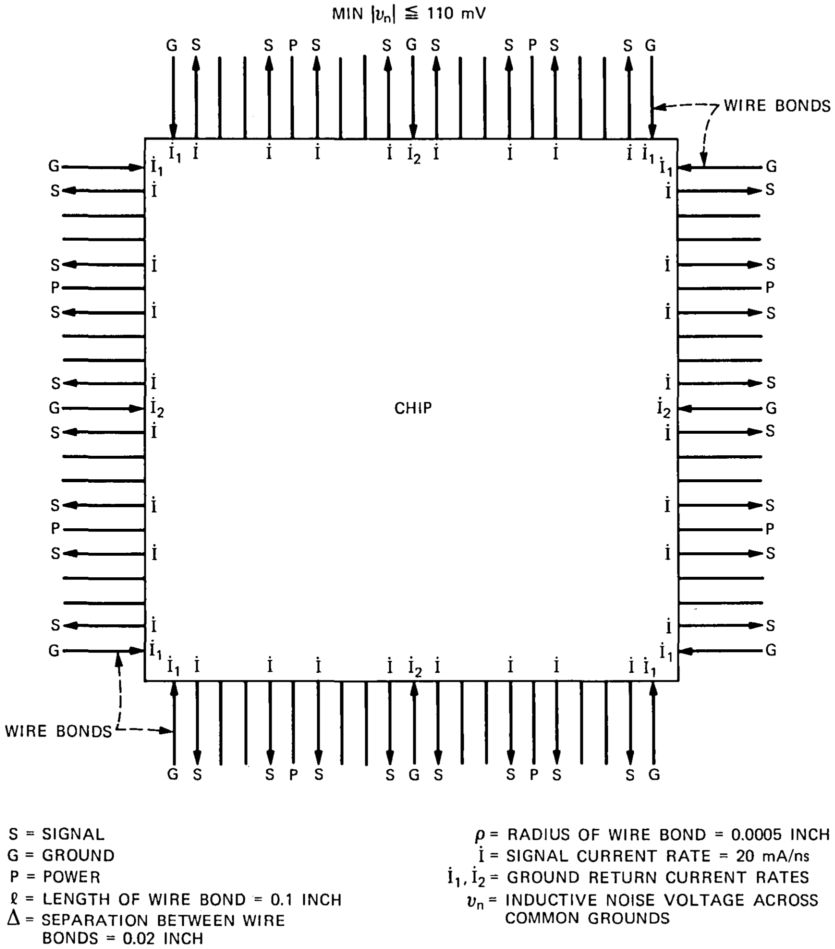

MIN $|v_n| \leq$ 110 mV
G S S P S S G S S P S S G
WIRE BONDS
G
S
S
P
S
S
G
S
S
P
S
S
G
CHIP
$i_1$ $i$ $i_2$
WIRE BONDS
G S S P S S G S S P S S G
S = SIGNAL
G = GROUND
P = POWER
$\ell$ = LENGTH OF WIRE BOND = 0.1 INCH
$\Delta$ = SEPARATION BETWEEN WIRE BONDS = 0.02 INCH
$\rho$ = RADIUS OF WIRE BOND = 0.0005 INCH
$\dot{i}$ = SIGNAL CURRENT RATE = 20 mA/ns
$\dot{i}_1, \dot{i}_2$ = GROUND RETURN CURRENT RATES
$v_n$ = INDUCTIVE NOISE VOLTAGE ACROSS COMMON GROUNDS

Fig. 8—Driver circuits on integrated circuit chip switching 32 signal bits simultaneously, with min$|v_n| \leq 110$ mV.

arrays of conductors in a 4 by 10 section of a circuit-pack connector.[5,6] Details are presented in the Appendix.

### *5.2.5 Some generalizations*

When the array of conductors contains nonparallel conductors, eq. (52) of Ref. 3 can be used to generalize eq. (4) above. Also, for more general configurations of parallel conductors, eq. (28) of Ref. 3 generalizes eq. (4) above. These generalizations, along with the associated self-inductances, can also be used in eq. (26) or (28) to compute inductive noise voltage.

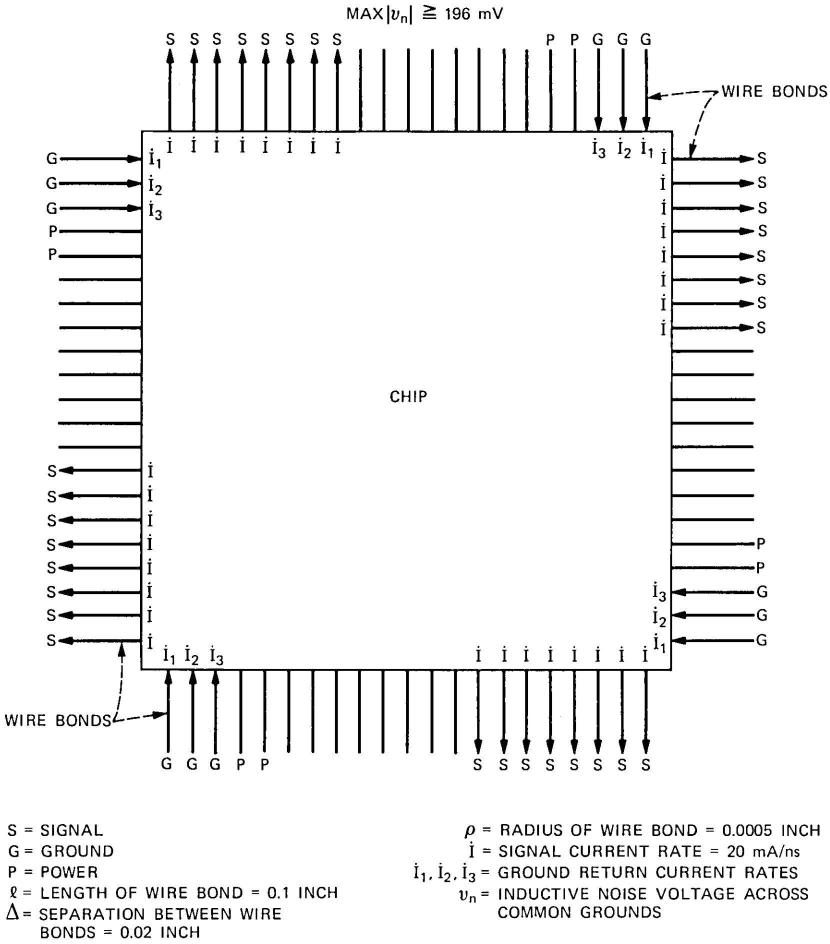

MAX$|v_n|$ ≧ 196 mV
S S S S S S S S
P P G G G
WIRE BONDS
$\dot{i}$ $\dot{i}$ $\dot{i}$ $\dot{i}$ $\dot{i}$ $\dot{i}$ $\dot{i}$ $\dot{i}$
$\dot{i}_3$ $\dot{i}_2$ $\dot{i}_1$
G G G P P
$\dot{i}_1$ $\dot{i}_2$ $\dot{i}_3$
S S S S S S S S
CHIP
S S S S S S S S
P P G G G
$\dot{i}_3$ $\dot{i}_2$ $\dot{i}_1$
WIRE BONDS
G G G P P
S S S S S S S S
S = SIGNAL
G = GROUND
P = POWER
ℓ = LENGTH OF WIRE BOND = 0.1 INCH
Δ = SEPARATION BETWEEN WIRE BONDS = 0.02 INCH
ρ = RADIUS OF WIRE BOND = 0.0005 INCH
$\dot{i}$ = SIGNAL CURRENT RATE = 20 mA/ns
$\dot{i}_1, \dot{i}_2, \dot{i}_3$ = GROUND RETURN CURRENT RATES
$v_n$ = INDUCTIVE NOISE VOLTAGE ACROSS COMMON GROUNDS

Fig. 9—Driver circuits on integrated circuit chip switching 32 signal bits simultaneously, with max$|v_n| \geq$ 196 mV.

### *5.3 Power and ground planes*

The inductive noise voltage, $v_n$, in a power or ground plane can be computed from eq. (20). If the time rate of change of the current flowing in the power and ground plane is $\dot{I}_0$ mA/ns, the inductive noise voltage in either the power or ground plane is given by

$$v_n = \dot{I}_0 L_g = 5\pi \dot{I}_0 \ell \left(\frac{h}{W}\right) \text{mV}, \tag{46}$$

where $\ell$ is expressed in inches. For example, if $\ell = 1$ inch, $W = 1$ inch, $h = 0.005$ inch, and $\dot{I}_0 = 200$ mA/ns, then $v_n = 15.7$ mV.

## VI. IMPEDANCE MATCH OF CHIP PACKAGES AND CIRCUIT PACKS

The computation of inductive noise as discussed in this paper applies when the array of conductors is considered as lumped electrical elements. This is the case for the electrically short power and ground leads and electrically short segments of signal leads. However, for electrically long signal leads, a transmission line point of view is more appropriate. In this case, an important consideration is the design of chip packages having signal leads that are impedance matched to the signal leads in a circuit pack. The impedance matching of chip packages to circuit packs was treated in Ref. 7.

## VII. CONCLUSIONS

Inductive noise limits the physical design of high-speed, high pin-out chip packages. The general eqs. (26) and (28) derived in this paper are useful for computing the inductive noise resulting from the interconnections in high-speed, high pin-out chip packages. When the distances between conductors are small relative to conductor lengths, the general equations reduce to the approximate equations given as eqs. (29) and (30). The equations are useful for computing inductive noise in general arrays of wire bonds, solder balls, DIP leads, package pins, and connector pins. Computed results were found to agree well with measured results. Two simple rules are presented for minimizing inductive noise. The inductive noise of P/G planes can also be computed.

## VIII. ACKNOWLEDGMENTS

The author is indebted to C. W. Deisch and F. Huff for clarifying system requirements concerning inductive noise. The author also wishes to thank A. P. Irwin, F. A. Scaglione, J. N. Arnold, and J. C. Shank for valuable assistance.

## REFERENCES

1. C. W. Deisch, unpublished work.
2. J. A. Stratton, *Electromagnetic Theory*, New York and London: McGraw-Hill, 1941.
3. F. W. Grover, *Inductance Calculations Working Formulas and Tables*, New York: Dover Publications, Inc., 1962.
4. H. H. Skilling, *Electric Transmission Lines*, Huntington, NY: R. E. Krieger Publishing Co., 1979.
5. W. L. Harrod and A. G. Lubowe, "The *BELLPAC*® Modular Electronic Packaging System," B.S.T.J., *58*, No. 10 (December 1979), pp. 2271–88.
6. C. L. Winings, "The 963-Type Printed-Circuit-Board Connector Family for *BELLPAC*®," IEEE Trans. Components, Hybrids, Manufacturing Technology, *CHMT-3*, No. 4 (December 1980), pp. 601–9.
7. A. J. Rainal, unpublished work.

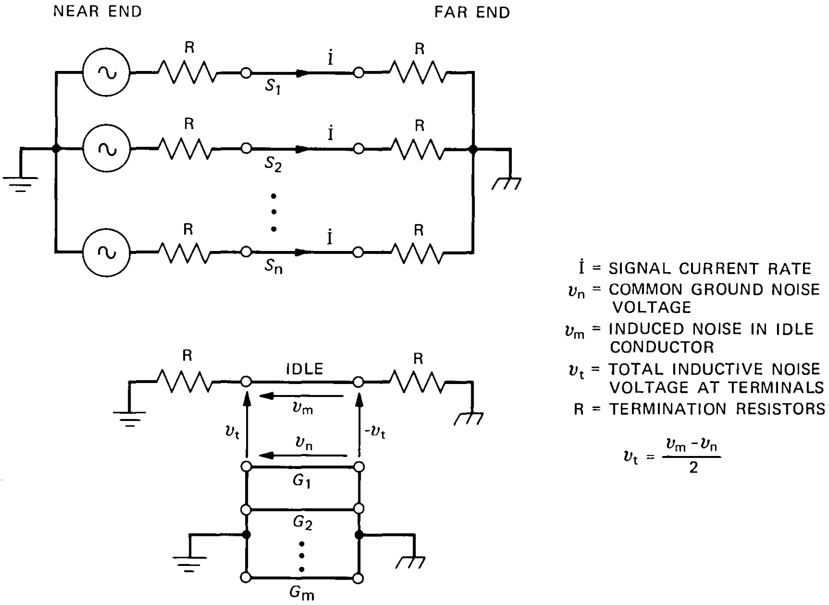

NEAR END
FAR END
R
R
$\dot{I}$
$S_1$
$S_2$
$S_n$
IDLE
$v_m$
$v_n$
$v_t$
$-v_t$
$G_1$
$G_2$
$G_m$
$\dot{I}$ = SIGNAL CURRENT RATE
$v_n$ = COMMON GROUND NOISE VOLTAGE
$v_m$ = INDUCED NOISE IN IDLE CONDUCTOR
$v_t$ = TOTAL INDUCTIVE NOISE VOLTAGE AT TERMINALS
R = TERMINATION RESISTORS
$v_t = \frac{v_m - v_n}{2}$

Fig. 10—Inductive noise model for an array of signal ($S_i$), ground ($G_i$), and idle conductors.

## APPENDIX

### *Experimental and Computed Inductive Noise of Interconnections*

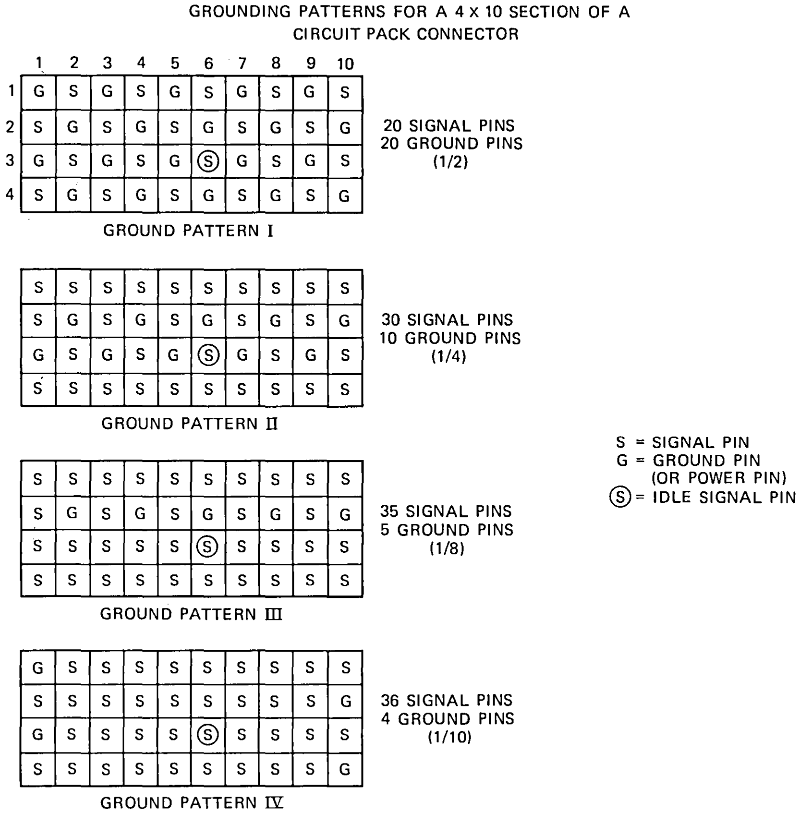
To compare computed results with experimental results, some inductive noise measurements were made on arrays of conductors in a 4 by 10 section of a circuit-pack connector.[5,6]

A general electrical model for an array of signal ($S_i$), ground ($G_i$), and idle conductors is shown in Fig. 10. The signal leads ($S_i$) are assumed to carry current rates $\dot{I}$, which occur simultaneously. The voltages $v_t$, $v_n$, and $v_m$ are used to characterize the inductive noise induced in the closed circuit loops. The two different ground potentials represent two equipotential surfaces (i.e., two copper ground planes).

The measurements were performed for the four grounding patterns shown in Fig. 11. The percentage of grounds varies from 50 percent for ground pattern I to 10 percent for ground pattern IV.

The experimental results for $v_t$ are presented in Table II, along with the corresponding computed results for the case of a signal rise time of 6 ns and termination resistors of 100Ω. The physical dimensions used were obtained by measurements on the circuit-pack connector.[5,6]

The entries labeled single refer to the case when the average radius (with respect to length) is taken as 0.0234 inch and the total conductor length is taken as 0.790 inch.



GROUNDING PATTERNS FOR A 4 X 10 SECTION OF A CIRCUIT PACK CONNECTOR

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | G | S | G | S | G | S | G | S | G | S |
| 2 | S | G | S | G | S | G | S | G | S | G |
| 3 | G | S | G | S | G | Ⓢ | G | S | G | S |
| 4 | S | G | S | G | S | G | S | G | S | G |

20 SIGNAL PINS
20 GROUND PINS
(1/2)

GROUND PATTERN I

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| S | S | S | S | S | S | S | S | S | S |
| S | G | S | G | S | G | S | G | S | G |
| G | S | G | S | G | Ⓢ | G | S | G | S |
| S | S | S | S | S | S | S | S | S | S |

30 SIGNAL PINS
10 GROUND PINS
(1/4)

GROUND PATTERN II

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| S | S | S | S | S | S | S | S | S | S |
| S | G | S | G | S | G | S | G | S | G |
| S | S | S | S | S | Ⓢ | S | S | S | S |
| S | S | S | S | S | S | S | S | S | S |

35 SIGNAL PINS
5 GROUND PINS
(1/8)

GROUND PATTERN III

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| G | S | S | S | S | S | S | S | S | S |
| S | S | S | S | S | S | S | S | S | G |
| G | S | S | S | S | Ⓢ | S | S | S | S |
| S | S | S | S | S | S | S | S | S | G |

36 SIGNAL PINS
4 GROUND PINS
(1/10)

GROUND PATTERN IV

S = SIGNAL PIN
G = GROUND PIN
(OR POWER PIN)
Ⓢ = IDLE SIGNAL PIN

Fig. 11—Grounding patterns I, II, III, and IV.

Table II—Comparison of experimental and computed inductive noise with $T_R = 6$ ns, $R = 100\Omega$

| Pattern | Percent $v_t$ (Experimental) | Percent $v_n$ | Percent $v_m$ | Percent $v_t$ (Computed) | |
|---|---|---|---|---|---|
| I | 0.3 | −0.889 | −0.481 | 0.204 | Single |
| | | −0.917 | −0.429 | 0.244 | Cascade |
| | | −0.976 | −0.492 | 0.242 | Joint |
| II | 0.9 | −4.46 | −2.34 | 1.06 | Single |
| | | −4.21 | −1.91 | 1.15 | Cascade |
| | | −4.73 | −2.36 | 1.19 | Joint |
| III | 4.0 | −9.88 | −1.18 | 4.35 | Single |
| | | −9.46 | −0.943 | 4.26 | Cascade |
| | | −10.51 | −1.19 | 4.66 | Joint |
| IV | 9.0 | −13.83 | +9.24 | 11.54 | Single |
| | | −13.23 | +6.63 | 9.93 | Cascade |
| | | −14.64 | +9.24 | 11.94 | Joint |

The entries labeled "cascade" refer to the case when the $v_t$'s for two subsections of each conductor were added. The first subsection represents a radius of 0.015 inch and a length of 0.5 inch. The remaining subsection was of radius 0.038 inch and of length 0.290 inch.

Finally, the entries labeled "joint" refer to the case when a single $v_t$ was evaluated for each conductor having the radii and lengths given in the previous paragraph.

By comparing the experimental values of $v_t$ in Table II with the corresponding triplet of computed values, we see that there is indeed good agreement in all cases.

## AUTHOR

**Attilio J. Rainal,** University of Alaska; University of Dayton, 1950–52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel University; D. Eng., 1963, Johns Hopkins University; AT&T Bell Laboratories, 1964—. Mr. Rainal's early work involved research on noise theory with application to detection, estimation, radiometry, and radar theory. He has also been engaged in the analysis of FM communication systems. His more recent work includes studies of crosstalk on multilayer boards, voltage breakdown, and current-carrying capacity of printed wiring, electrical aspects of chip packages, and electromagnetic compatibility. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

# PAPERS BY AT&T BELL LABORATORIES AUTHORS

## COMPUTING/MATHEMATICS

Aumann R. J. et al., **Approximate Purification of Mixed Strategies.** Math Oper R 8(3):327-341, 1983.

Baker B. S., Calderbank A. R., Coffman E. G., Lagarias J. C., **Approximation Algorithms for Maximizing the Number of Squares Packed into a Rectangle.** SIAM J Alg 4(3):383-397, 1983.

Baker B. S., Schwarz J. S., **Shelf Algorithms for Two-Dimensional Packing Problems.** SIAM J Comp 12(3):508-525, 1983.

Chen R., Shepp L. A., **On the Sum of Symmetric Random-Variables.** Am Statistn 37(3):237, 1983.

Conway J. H., Sloane N. J. A., **The Coxeter-Todd Lattice, the Mitchell Group, and Related Sphere Packings.** Math Proc C 93(May):421-440, 1983.

Fishburn P. C., **Threshold-Bounded Interval Orders and a Theory of Picycles.** SIAM J Alg 4(3):290-305, 1983.

Fishburn P. C., Brams S. J., **Paradoxes of Preferential Voting.** Math Mag 56(4):207-214, 1983.

Fowlkes E. B., Mallows C. L., **A Method for Comparing Two Hierarchical Clusterings.** J Am Stat A 78(383):553-569, 1983.

Fowlkes E. B., Mallows C. L., **A Method For Comparing Two Hierarchical Clusterings—Rejoinder.** J Am Stat A 78(383):584, 1983.

Gilstein C. Z., **On the Joint Asymptotic-Distribution of Extreme Midranges.** Ann Statist 11(3):913-920, 1983.

Goldfarb D., Idnani A., **A Numerically Stable Dual Method for Solving Strictly Convex Quadratic Programs.** Math Progr 27(1):1-33, 1983.

Landau H. J., **The Inverse Problem for the Vocal-Tract and the Moment Problem.** SIAM J Math 14(5):1019-1035, 1983.

Sandberg I. W., **The Mathematical Foundations of Associated Expansions for Mildly Non-Linear Systems.** IEEE Circ S 30(7):441-455, 1983.

## ENGINEERING

Arnold H. W., Bodtmann W. F., **Interfade Interval Statistics of a Rayleigh-Distributed Wave (Letter).** IEEE Commun 31(9):1114-1116, 1983.

Brews J. R., **Rapid Interface Parameterization Using a Single MOS Conductance Curve.** Sol St Elec 26(8):711-716, 1983.

Buhl L. L., **Optical Losses in Metal/$SiO_2$-Clad Ti-$LiNbO_3$ Wave-Guides.** Electr Lett 19(17):659-660, 1983.

Campbell J. C., Burrus C. A., Copeland J. A., Dentai A. G., **Wavelength-Discriminating Photodetector for Lightwave Systems.** Electr Lett 19(17):672-674, 1983.

Chaban E. E., Chabal Y. J., **Sample Manipulator for Operation Between 20-K and 2000-K in Ultrahigh-Vacuum.** Rev Sci Ins 54(8):1031-1033, 1983.

Chraplyvy A. R., Henry P. S., **Performance Degradation Due to Stimulated Raman-Scattering in Wavelength-Division-Multiplexed Optical-Fibre Systems.** Electr Lett 19(16):641-643, 1983.

Cheng J., Thurston R. N., Boyd G. D., Meyer R. B., **A New Low-Voltage Electrically-Addressed Bistable Nematic Liquid-Crystal Boundary-Layer Display.** P Sid 24(2):180-185, 1983.

Dalal S. R., **Exact Simultaneous Confidence Bands for Random Intercept Regression.** Technomet 25(3):263-269, 1983.

Dudderar T. D. et al., **Application of Fiber Optics to Speckle Metrology—A Feasibility Study.** Exp Mech 23(3):289-297, 1983.

Dudderar T. D., Thibault L. R., **A Noncontacting System for Measuring Effective Stress-Strain Curves of Ultrathin X-Ray-Mask Materials.** Exp Mech 23(3):322-328, 1983.

Dutta N. K., Hartman R. L., Tsang W. T., **Gain and Carrier Lifetime Measurements in AlGaAs Single Quantum Well Lasers.** IEEE J Q El 19(8):1243–1246, 1983.

Ehrlich N., **Advanced Mobile Phone Service Using Cellular Technology.** Microwave J. 26(8):119+, 1983.

Gehrlein W. V., Fishburn P. C., **Scoring Rule Sensitivity to Weight Selection.** Publ Choice 40(3):249–261, 1983.

Glance B., Greenstein L. J., **Frequency-Selective Fading Effects in Digital Mobile Radio With Diversity Combining.** IEEE Commun 31(9):1085–1094, 1983.

Herb G. K., Caffrey R. E., Eckroth E. T., Jarrett Q. T., Fraust C. L., Fulton J. A., **Plasma Processing—Some Safety, Health and Engineering Considerations.** Sol St Tech 26(8):185–194, 1983.

Herbst D., Bosch M. A., Tewksbury S. R., **Island-Edge Effects of Transistors Fabricated in Large-Area Laser Micro-Zone Crystallized Si on Insulator.** IEEE Elec D 4(8):280–282, 1983.

Herbst D., Bosch M. A., Tewksbury S. K., **Substrate Influence on NMOS Transistors in Large-Area Laser Crystallized Isolated Si Layers.** IEEE Elec D 4(7):205–207, 1983.

Jamshidi M., Malek-Zavarei M., **A Hierarchical Control for Large-Scale Time-Delay Systems.** Large Scale 4(2):149–163, 1983.

Kadota T. T., Seery J. B., **Probability-Distributions of Randomly Moving-Objects on a Plane (Letter).** IEEE Info T. 29(5):756–761, 1983.

Kaminow I. P., Stulz L. W., Ko J. S., Dentai A. G., Nahory R. E., Dewinter J. C., Hartman R. L., **Low-Threshold InGaAsP Ridge Wave-Guide Lasers At 1.3-μm.** IEEE J Q El 19(8):1312–1319, 1983.

Kaufman S., Williams J. L., Smith E. E., Przybyla L. J., **Large-Scale Fire Tests of Building Riser Cables.** J Fire Sci 1(1):54–65, 1983.

Keramidas V. G., King W. C., **High-Performance Monolithic GaAlAs/GaAs Photo-Diode Arrays for Voltage-Controlled Semiconductor Switches.** IEEE Device 30(8):883–886, 1983.

Khanna S. K., Congdon W. I., **Engineering and Molding Properties of Poly(Vinyl-Chloride), Acrylonitrile-Butadiene-Styrene and Polyester Blends.** Polym Eng S 23(11):627–631, 1983.

Lifshitz N., Luryi S., **Influence of a Resistive Sublayer at the Polysilicon Silicon Dioxide Interface on MOS Properties.** IEEE Device 30(7):833–836, 1983.

Liou M. L., Kuo Y. L., Lee C. F., **A Tutorial on Computer-Aided Analysis of Switched-Capacitor Circuits.** P IEEE 71(8):987–1005, 1983.

Luss H., **A Multifacility Capacity Expansion Model With Joint Expansion Set-Up Costs.** Nav Res Log 30(1):97–111, 1983.

Marcuse D., **Classical Derivation of the Laser Rate-Equation.** IEEE J Q El 19(8):1228–1231, 1983.

Meyer R. B., Thurston R. N., **Discovery of dc Switching of a Bistable Boundary-Layer Liquid-Crystal Display.** Appl Phys L 43(4):342–344, 1983.

Miydshi T., Omori H., Maeda G., **Reduction of Magnetic-Flux Leakage From an Induction-Heating Range.** IEEE Ind AP 19(4):491–497, 1983.

Ng K. K., Taylor G. W., **Effects of Hot-Carrier Trapping in N-Channel and P-Channel MOSFETs.** IEEE Device 30(8):871–876, 1983.

Oikonomou K. N., Kain R. Y., **Abstractions for Node Level Passive Fault-Detection in Distributed Systems.** IEEE Comput 32(6):543–550, 1983.

Paalanen M. A., Thomas G. A., **Experimental Tests of Localization in Semiconductors.** Helv Phys A 56(1–3):27–34, 1983.

Pearsall T. P., Logan R. A., Bethea C. G., **GaInAs/InP Large Bandwidth (Greater-Than-2-GHz) Pin Detectors.** Electr Lett 19(16):611–612, 1983.

Temes G. C., Tsividis Y., **The Special Section on Switched-Capacitor Circuits (Editorial).** P IEEE 71(8):915–916, 1983.

Tsang W. T., Ditzenberger J. A., Olsson N. A., **Improvement of Photo-Luminescence of Molecular-Beam Epitaxially Grown $Ga_xAl_yIn_{1-x-y}As$ by Using an $As_2$ Molecular-Beam.** IEEE Elec D 4(8):275–277, 1983.

Tsang W. T., Olsson N. A., Logan R. A., **Mode-Locked Semiconductor-Lasers with**

**Gateable Output and Electrically Controllable Optical Absorber.** Appl Phys L 43(4):339–341, 1983.

Vanderzi J. P., Temkin H., Logan R. A., **Wavelength Multiplexing of 1.31-$\mu$m InGaAsP Buried Crescent Laser Arrays.** Appl Phys L 43(5):401–403, 1983.

Witschorik C. A., **The Real-Time Debugging Monitor for the Bell System-1A Processor.** Software 13(8):727–743, 1983.

## MANAGEMENT/ECONOMICS

Brown S. J., Weinstein M. I., **A New Approach to Testing Asset Pricing-Models—The Bilinear Paradigm.** J Finance 38(3):711–743, 1983.

Fishburn P. C., **A New Characterization of Simple Majority.** Econ Lett 13(1):31–35, 1983.

Francis A. A., McCalla C., **Optimal Rates of Extraction of Exhaustible Resource With an Analysis of the Optimal Time Problem.** Social Econ 29(1):90–100, 1980.

Greenwald B. C., **A General-Analysis of Bias in the Estimated Standard Errors of Least-Squares Coefficients.** J Economet 22(3):323–338, 1983.

Greenwald B. C. et al., **Adverse Selection in the Market for Slaves—New-Orleans, 1830–1860.** Q J Econ 98(3):479–499, 1983.

Laitinen K., Theil H., Raparla T., **A Generalization of Workings Model.** Econ Lett 13(1):97–100, 1983.

## PHYSICAL SCIENCES

Aeppli G., Bruinsma R., **Linear Response Theory and the One-Dimensional Ising Ferromagnet in a Random Field.** Phys Lett A 97(3):117–120, 1983.

Ahlers F. J., Lohse F., Spaeth J. M., Mollenauer L. F., **Magneto-Optical Studies of Atomic Thallium Centers In KCl-Magnetic Circular-Dichroism Tagged by Spin-Resonance.** Phys Rev B 28(3):1249–1255, 1983.

Altarelli M., Bachelet G.B., Bouche V., Delsole R., **Electron Core-Hole Interactions at Surfaces—An Exactly Soluble Model.** Surf Sci 129(2–3):447–481, 1983.

Anthony L. J., Prescott B. E., **Simultaneous Determination of Small Amounts of Hydrochloric and Hydrobromic Acids by Derivatization with Ethylene-Oxide and Gas-Chromatography.** J Chromat 264(3):405–413, 1983.

Arovas D., Bhatt R. N., Shapiro B., **Anisotropic Bond Percolation in Two Dimensions.** Phys Rev B 28(3):1433–1437, 1983.

Aspnes D. E., Heller A., **Barrier Height and Leakage Reduction in N-GaAs-Platinum Group Metal Schottky Barriers Upon Exposure to Hydrogen.** J Vac Sci B 1(3):602–607, 1983.

Bachelet G. B., Schluter M., **Bonding Geometries of Cl on Si(111) and Ge(111).** J Vac Sci B 1(3):726–728, 1983.

Bally J., Snell R. L., Predmore R., **Radio Images of the Bipolar H-II Region $SiO_6$.** Astrophys J 272(1):154–162, 1983.

Ballman A. A., Glass A. M., Nahory R. E., Brown H., **Double Doped Low Dislocation Density InP With Low Optical-Absorption.** Inst Phys C1983(65):15–21, 1983.

Banavar J. R., Cieplak M., **Scaling Stiffness of Spin-Glasses (Letter).** J Phys C 16(21):L755–L759, 1983.

Bevk J., **Ultrafine Filamentary Composites (Review).** Ann R Mater 13:319–338, 1983.

Beck J. W., **Analysis of a Camera Based Spect System.** Nucl Instru 213(2–3):415–436, 1983.

Bowmer T. N., Vroom W. I., Hellman M. Y., **The Radiation Crosslinking of Poly(Vinyl-Chloride) with Trimethylolpropanetrimethacrylate. 3. Effect of Diundecyl Phthalate—Chemical-Kinetics of a Three-Component System.** J Appl Poly 28(8):2553–2565, 1983.

Cava R. J., Santoro A., Murphy D. W., Zahurak S., Roth R. S., **The Structures of Lithium Inserted Metal-Oxides-$Li_2FeV_{308}$.** J Sol St Ch 48(3):309–317, 1983.

Chemla D. S., **Quasi-Two-Dimensional Excitons in GaAs/$Al_xGa_{1-x}As$ Semiconductor Multiple Quantum Well Structures.** Helv Phys A 56(1-3):607-637, 1983.

Chin A. K., Camlibel I., Sheng T. T., Bonner W. A., **Extremely Rapid Out Diffusion of Sulfur in InP.** Appl Phys L 43(5):495-497, 1983.

Choudhury A. N. et al., **Ion-Implantation of Si and Be in $Al_{0.48}In_{0.52}As$.** J Appl Phys 54(8):4374-4377, 1983.

Chylek P., Ramaswami V., Ashkin A., Dziedzic J. M., **Simultaneous Determination of Refractive-Index and Size of Spherical Dielectric Particles from Light-Scattering Data.** Appl Optics 22(15):2302-2307, 1983.

Cohen R. L., West K. W., **Characterization of Metals and Alloys by Electrical-Resistivity Measurements.** Mater Eval 41(9):1074-1077, 1983.

Coldren L. A., Furuya K., Miller B. I., **On the Formation of Planar-Etched Facets in GaInAsP InP Double Heterostructures.** J Elchem So 130(9):1918-1926, 1983.

Coppersmith S. N., Fisher D. S., **Pinning Transition of the Discrete Sine-Gordon Equation.** Phys Rev B 28(5):2566-2581, 1983.

Cox D. E., Shapiro S. M., Nelmes R. J., Ryan T. W., Bleif H. J., Cowley R. A., Eibschutz M., Guggenheim H. J., **X-Ray-Diffraction and Neutron-Diffraction Measurements on $BaMnF_4$ (Letter).** Phys Rev B 29(3):1640-1643, 1983.

Cox H. M., Prior A. S., Keramidas V. G., **High-Throughput $AsCl_3/Ga/H_2$ Vapor-Phase Epitaxial System for Growth of Extremely Uniform Multilayer GaAs Structures.** Inst Phys C1983(65):133-140, 1983.

Dautremontsmith W. C., Feldman L. C., **Surface Structural Damage Produced in InP(100) by RF Plasma or Sputter Deposition.** Thin Sol Fi 105(2):187-196, 1983.

Doak R. B., Harten U., Toennies J. P., **Anomalous Surface Phonon-Dispersion Relations for Ag(111) Measured by Inelastic-Scattering of He-Atoms.** Phys Rev L 51(7):578-581, 1983.

Friedman J. M. et al., **The Iron-Proximal Histidine Linkage and Protein Control of Oxygen Binding in Hemoglobin—A Transient Raman Study.** J Biol Chem 258(17):564-572, 1983.

Gooden R., **Photo-Oxidation of Trichlorosilane in Silicon Tetrachloride.** Inorg Chem 22(16):2272-2275, 1983.

Hagen M. et al., **Random-Fields and Three-Dimensional Ising-Models-$Co_xZn_{1-x}F_2$.** Phys Rev B 28(5):2602-2613, 1983.

Hasegawa A., **A Test of Self-Organization Hypothesis in Jovian and Saturnian Wind Systems.** J Phys Jpn 52(6):1930-1934, 1983.

Hasegawa A., Tsui K. H., Assis A. S., **A Theory of Long Period Magnetic Pulsations. 3. Local Field Line Oscillations.** Geophys R L 10(8):765-767, 1983.

Helfand E., Pearson D. S., **Statistics of the Entanglement of Polymers—Unentangled Loops and Primitive Paths.** J Chem Phys 79(4):2054-2059, 1983.

Heller A., Leamy H. J., Miller B., Johnston W. D., **Chemical Passivation of Carrier Recombination at Acid Interfaces and Grain-Boundaries of P-InP.** J Phys Chem 87(17):3239-3244, 1983.

Hilinski E. F., Milton S. V., Rentzepis P. M., **Transient Species in Electron-Transfer—Reactions of Chloranil with Donor Aromatic-Compounds.** J Am Chem S 105(16):5193-5196, 1983.

Hilinski E. F., Rentzepis P. M., **Picosecond Spectroscopy—Methods and Recent Applications (Review).** Analyt Chem 55(11):1121+, 1983.

Ihm J. et al., **Study of High-Order Reconstructions of the Sp(100) Surface.** J Vac Sci B 1(3):705-708, 1983.

Inoue A., Chen H. S., Krause J. T., Masumoto T., Hagiwara M., **Youngs Modulus of Fe-Based, Co-Based, Po-Based and Pt-Based Amorphous Wires Produced by the In-Rotating-Water Spinning Method.** J Mater Sci 18(9):2743-2751, 1983.

Johnson P. D., Woodruff D. P., Farrell H. H., Smith N. V., Traum M. M., **Photoelectron Diffraction From Te on Ni(100) and Cu(100).** Surf Sci 129(2-3):366-374, 1983.

Johnson R. E., Boring J. W., Reimann C. T., Barton L. A., Sieveka E. M., Garrett J. W., Farmer K. R., Brown W. L., Lanzerotti L. J., **Plasma Ion-Induced Molecular Ejection on the Galilean Satellites—Energies of Ejected Molecules.** Geophys R L 10(9):892-895, 1983.

Kim O. K., Bonner W. A., **Infrared Reflectance and Absorption of N-Type InP.** J Elec Mat 12(5):827–836, 1983.

Levin R. M., Sheng T. T., **Oxide Isolation for Double-Polysilicon VLSI Devices.** J Elchem So 130(9):1894–1897, 1983.

Lin C., Burrus C. A., Linke R. A., Kaminow I. P., Ko J. S., Dentai A. G., Logan R. A., Miller B. I., **Short-Coupled-Cavity (SCC) InGaAsP Injection-Lasers for CW and High-Speed Single-Longitudinal-Mode Operation.** Electr Lett 19(15):561–562, 1983.

Lin P. S. D., Marcus R. B., Sheng T. T., **Leakage and Breakdown in Thin Oxide Capacitors—Correlation With Decorated Stacking-Faults.** J Elchem So 130(9):1878–1883, 1983.

Ling H. C., Yan M. F., **Second Phase Development in Sr-Doped $TiO_2$.** J Mater Sci 18(9):2688–2696, 1983.

Lyon S. A., Worlock J. M., **Role of Electromagnetic Resonances in the Surface-Enhanced Raman Effect.** Phys Rev L 51(7):593–596, 1983.

MacCallum C. J., Leventhal M., **Observations of Annihilation Radiation From the Galactic-Center Region.** AIP Conf PR1983(101):211–229, 1983.

Meisner G. P., Ku H. C., Barz H., **Superconducting Equiatomic Ternary Transition-Metal Arsenides.** Mater Res B 18(8):983–991, 1983.

Mills A. P., Pfeiffer L., Platzman P. M., **Positronium Velocity Spectroscopy of the Electronic Density of States at a Metal-Surface.** Phys Rev L 51(12):1085–1088, 1983.

Murarka S. P., **Transition-Metal Silicides (Review).** Ann R Mater 13:117–137, 1983.

Nahory R. E., Ballman A. A., Brown H., Wilson M. R., **LEC Growth of InP(Mn) for P-Type and Semi-Insulating Materials.** Inst Phys C1983(65):7–14, 1983.

Nishikawa K. I., Okuda H., Hasegawa A., **Heating of Heavy-Ions on Auroral Field Lines.** Geophys R L 10(7):553–556, 1983.

Odagaki T., Lax M., Puri A., **Hopping Conduction in the D-Dimensional Lattice Bond Percolation Problem.** Phys Rev B 28(5):2755–2765, 1983.

Ogielski A. T., **Monte-Carlo Study of Scale-Covariant Field-Theories.** Phys Rev D., 28(6):1461–1472, 1983.

Olego D., Chang T. Y., Silberg E., Caridi E. A., Pinczuk A., **Compositional Dependence of Band-Gap Energy, Conduction-Band Effective Mass and Lattice-Vibrations of $In_{1-x-y}Ga_xAl_yAs$ Lattice Matched to InP.** Inst Phys C1983(65):195–202, 1983.

Patterson G. D., **Light-Scattering From Bulk Polymers (Review).** Ann R Mater 13:219–245, 1983.

Phillips J. C., **Unexplained Glass Flow—Reply (Letter).** Phys Today 36(8):87,1983.

Pian T. R., Tolk N. H., Traum M. M., Kraus J., Collins W. E., **Energy-Dependence of Electron Stimulated Desorption of Excited Neutral Alkalis from Alkali-Halides.** Surf Sci 129(2–3):573–580, 1983.

Remke R. L., Vonseggern H., **Modeling of Thermally Stimulated Currents in Polytetrafluoroethylene.** J Appl Phys 54(9):5262–5266, 1983.

Schluter M., Varma C. M., **Configuration Mixing in the Ground-State of Ce.** Helv Phys A 56(1–3):147–161, 1983.

Schultz H. J., **Phase-Transitions in Monolayers Adsorbed on Uniaxial Substrates.** Phys Rev B 28(5):2746–2749, 1983.

Shah J., Etienne B., Leheny R. F., Nahory R. E., **Hot Carrier Relaxation Processes in 1.3-$\mu$m Quaternary InGaAsP—Analysis and Application to Temperature-Dependence of Laser Threshold.** Inst Phys C1983(65):303–310, 1983.

Shank C. V. et al., **Femtosecond-Time-Resolved Surface Structural Dynamics of Optically-Excited Silicon.** Phys Rev L 51(10):900–902, 1983.

Shank C. V. et al., **New Experiments in Femtosecond Condensed Matter Spectroscopy.** Helv Phys A 56(1–3):373–381, 1983.

Shank C. V., Fork R. L., Yen R., Shah J., Greene B. I., Gossard A. C., Wiesbuch C., **Picosecond Dynamics of Hot Carrier Relaxation in Highly Excited Multi-Quantum Well Structures.** Sol St Comm 47(12):981–983, 1983.

Silberg E., Chang T. Y., Caridi E. A., Evans C. A., Hitzman C. J., **Spatially Correlated Redistribution of Mn and Ge in $In_{1-x}Ga_xAs$ MBE Layers.** Inst Phys C1983(65):187–194, 1983.

Sinha S. K., Varma C. M., **Possibility of Acoustic Plasmons in Mixed-Valence Metals and Their Interaction with Phonons.** Phys Rev B 28(4):1663–1666, 1983.

Slater N. J. et al., **Ion-Implantation Doping of InGaAs and InAlAs.** Inst Phys C1983(65):627–634, 1983.

Sternheim M., Kinsbron E., Alspectror J., Heimann P. A., **Properties of Thermal Oxides Grown on Phosphorus Insitu Doped Polysilicon.** J Elchem So 130(8):1735–1740, 1983.

Stillinger F. H., **Molecular Theory of Capillarity—Rowlinson, J. S., Widom, B. (Book Review).** Nature 304(5928):760,1983.

Stillinger F. H., Weber T. A., **Inherent Structure in Water.** J Phys Chem 87(15):2833–2840, 1983.

Surko C. M., Slusher R. E., **Waves and Turbulence in a Tokamak Fusion Plasma.** Science 221(4613):817–822, 1983.

Tarascon J. M., Waszczak J. V., Hull G. W., DiSalvo F. J., Blitzer L. D., **Synthesis and Physical-Properties of New Superconducting Chevrel Phases $Hg_xMo_6S_8$.** Sol St Comm 47(12):973–979, 1983.

Thurston R. N., **Exact-Solutions for Liquid-Crystal Configurations and an Improved Boundary-Layer Model.** J Appl Phys 54(9):4966–4988, 1983.

Tompkins H. G., Bennett J. E., Augis J. A., Paskowski T. M., **Oxidation of $Cu_3Sn$ and $Cu_6Sn_5$ Films in Room Air From 175-Degrees-C to 250-Degrees-C.** J Elchem So 130(8):1758-1762, 1983.

Valdes F., Tyson J. A., Jarvis J. F., **Alignment of Faint Galaxy Images—Cosmological Distortion and Rotation.** Astrophys J 271(2):431–441, 1983.

Vansaarloos W., Weeks J. D., **Surface Undulations in Explosive Crystallization—A Thermal-Instability.** Phys Rev L 51(12):1046–1049, 1983.

Vanuitert L. G., Gallagher P. K., Singh S., Zydzik G. J., **Time and Temperature Dependences of Phosphorus Evolution from InP.** J Vac Sci B 1(3):825–826, 1983.

Varma C. M., **Conditions for Stimulated Annihilation in a Degenerate E--E+ Fluid at the Surface of Pulsars.** AIP Conf PR1983(101):421–427, 1983.

Venkatesan T., Edelson D., Brown W. L., **Pulsed Ion-Beam Technique for Measuring Diffusion-Coefficient of a Slow Diffusant in Polymers.** Appl Phys L 43(4):364–366, 1983.

Vogel E. M., **Method of Characterizing the Optical-Quality of Glass (Letter).** Appl Optics 22(15):2241–2242, 1983.

Wang T. T., Von Seggern H., **High Electric-Field Poling of Electroded Poly(Vinylidene Fluoride) at Room-Temperature.** J Appl Phys 54(8):4602–4604, 1983.

Weber T. A., Helfand E., **Time-Correlation Functions From Computer-Simulations of Polymers.** J Phys Chem 87(15):2881–2889, 1983.

## SOCIAL AND LIFE SCIENCES

Cassidy M. F., Knowlton J. Q., **Visual Literacy—A Failed Metaphor.** Ect J 31(2):67–90, 1983.

Devereux R. B. et al., **Left-Ventricular Hypertrophy in Patients With Hypertension—Importance of Blood-Pressure Response to Regularly Recurring Stress.** Circulation 68(3):470–476, 1983.

Goetz E. T. et al., **Reading in Perspective—What Real Cops and Pretend Burglars Look for in a Story.** J Educ Psyc 75(4):500–510, 1983.

Littlerbishop S. et al., **Sexual Harassment in the Workplace as a Function of Initiators Status—The Case of Airline Personnel.** J Soc Issue 38(4):137–148, 1982.

Tartter V. C., **The Effects of Symmetric and Asymmetric Dyadic Visual Access on Attribution During Communication.** Lang Commun 3(1):1–10, 1983.

## SPEECH/ACOUSTICS

Allen R. B., **Composition and Editing of Spoken Letters.** Int J Man M 19(2):181–193, 1983.

Broek H. W., **Concurrent Underwater Acoustic Wavefront Fluctuations at Two Frequencies.** J Acoust So 74(2):559–563, 1983.

Karmakar S. B., **Stability-Criteria for Class of Non-Linear Feedback-Systems.** J Sound Vib 89(1):1–5, 1983.

Landauer T. K., Galotti K. M., Hartwell S., **Natural Command Names and Initial Learning—A Study of Text-Editing Terms.** Comm ACM 26(7):495–503, 1983.

Tartter V. C., Kat D., Samuel A. G., Repp B. H., **Perception of Intervocalic Stop Consonants—The Contributions of Closure Duration and Formant Transitions.** J Acoust So 74(3):715–725, 1983.

West J. E. et al., **Foil Electret Transducer for Blood-Pressure Monitoring.** J Acoust So 74(3):680–686, 1983.

# CONTENTS, FEBRUARY 1984

**AT&T BELL LABORATORIES 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.

AT&T
Bell Laboratories