November 1984
Vol. 63 No. 9

AT&T BELL LABORATORIES TECHNICAL JOURNAL

A JOURNAL OF THE AT&T COMPANIES

Queueing Networks

Word Recognition

Radio Astronomy

End Office Connection Study

AT&T BELL LABORATORIES TECHNICAL JOURNAL (ISSN0005-8580) is published ten times each year by AT&T, 550 Madison Avenue, New York, NY 10022; C. L. Brown, Chairman of the Board; T. O. Davis, Secretary. 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. 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 1H321, 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 NOVEMBER 1984 NO. 9

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 9, November 1984
Printed in U.S.A.

Open and Closed Models for Networks of Queues

By W. WHITT*

(Manuscript received April 9, 1984)

This paper investigates the relationship between open and closed models for networks of queues. In open models, jobs enter the network from outside, receive service at one or more service centers, and then depart. In closed models, jobs neither enter nor leave the network; instead, a fixed number of jobs circulate within the network. Open models are analytically more tractable, but closed models often seem more realistic. Hence, this paper investigates ways to use open models to approximate closed models. One approach is to use open models with specified expected equilibrium populations. This fixed-population-mean method is especially effective for approximately solving large closed models, where "large" may mean many nodes or many jobs. The success of these approximations is partly explained by limit theorems: Under appropriate conditions, the fixed-population-mean method is asymptotically correct. In some cases the open-model methods also yield bounds for the performance measures in the closed models.

I. INTRODUCTION AND SUMMARY

Queueing network models are now widely used to analyze communication, computing, and production systems. A relatively well-developed theory exists for the Markov Jackson network models and various extensions that have a product-form equilibrium (steady-state) distribution.[1-6] In this paper we consider both the product-form models for which exact solutions are possible and more complicated nonproduct-

* AT&T Bell Laboratories.

form models for which approximations are needed. We are primarily motivated by the desire to develop new approximations for non-product-form closed models. We discuss methods for modifying the Queueing Network Analyzer (QNA) software package[7] so that it can be used to calculate approximate congestion measures for closed models as well as open models. Our general approach to the closed models is to apply previous techniques for open models. Hence, we investigate the relationship between open and closed models.

For simplicity, we first consider a Markov Jackson network model with First-Come First-Served (FCFS) multiserver nodes (service centers) and one job class. Flow within the network is determined by stochastic routing probabilities: Each job completing service at node i goes immediately to node j with probability q_{ij}, independent of the history of the system. The individual service rates and external arrival rates, if any, are independent of the state. The service-time distributions are exponential and the external arrival processes, if any, are Poisson. It will be clear that the ideas generalize.

1.1 Open and closed models

These models can be classified as open or closed. In an open model, jobs enter the network at random from outside at a fixed rate, receive service at one or more nodes, and eventually leave the network. Thus, with an open model the total external arrival rate or throughput is an independent variable (specified as part of the model data), and the number of jobs in the system is a dependent variable (whose equilibrium distribution is described in the model solution). On the other hand, in a closed model there is a fixed population of jobs in the network. Hence, with a closed model the number of jobs in the system is an independent variable (specified as part of the model data), and the throughput (which may be defined, for example, as the departure rate from some designated node) is a dependent variable (to be calculated and described in the model solution). Since the individual service rate is part of the model data, knowing the throughput is equivalent to knowing the utilization, which is the expected proportion of the servers at the designated node that are busy in equilibrium.

Of course, there also are more complicated models, in which the simple dichotomy above is not valid. For example, Jackson introduced models in which the external arrival rate can depend on the total number of jobs in the network.[1] Then neither the external arrival rate nor the network population is fixed. There are also mixed models, which have some classes with fixed populations and other classes with fixed external arrival rates.[3,6] We will not consider these more general models, but we note that open models can be used to approximate mixed models in the same way that they can be used to approximate closed models.

It might seem that open models would be more appropriate for most applications because jobs do usually come from outside, flow through the system, and eventually depart. However, closed models are often used instead. The representation of flow through the system, i.e., the throughput, is easily handled in a closed model by assuming that a new job enters the system to replace an old one whenever the old one has received all of its required service. This can be represented in the closed model by a transition to a designated exit-entry node. At this node, arriving jobs complete all of their required service, and departures are new jobs. The rate of transitions through this node (which is both the arrival rate and the departure rate) can be regarded as the throughput. If no such exit-entry node exists originally, it is easy to add such a node. The modified network with the additional node is equivalent to the original network if all jobs at this new exit-entry node have zero service time.

Evidently, closed models are often applied because it seems natural to regard the number of jobs in the system as the independent variable and the throughput as the dependent variable. The number of jobs in the system is often subject to control; the queueing analysis is desired to determine the associated throughputs and response times. For example, in production systems, new jobs usually do not arrive at random; they are scheduled. In fact, this view was the main reason that Jackson extended queueing network theory to cover closed models:[1]

> This extension of the author's earlier work is motivated by the observation that real production systems are usually subject to influences which make for increased stability by tending, as the amount of work-in-process grows, to reduce the rate at which new work is injected or to increase the rate at which processing takes place.

Similarly, in computing systems the total number of jobs in device queues tends to be limited by resource constraints, so that it is natural to specify the number of jobs (the multiprogramming level) as a decision variable and then calculate the associated throughput (see p. 116 of Ref. 6). Also, in time-sharing systems the number of jobs is limited by the number of sources (terminals), so that the total number of jobs is not unbounded (see p. 60 of Ref. 6). Hence, even though closed models are significantly more difficult to analyze because of the normalization constant or partition function, there are good reasons for applying them.

1.2 The fixed-population-mean method

In this paper we propose and investigate a different approach that may sometimes be an attractive alternative. We propose using the open model with specified expected equilibrium population, which we refer to as the Fixed-Population-Mean (FPM) method. With the FPM method, we have the analytically more elementary open model, but the number of jobs (or, more precisely, its mean) becomes an independent variable and the throughput or total arrival rate becomes a dependent variable. Even though some of the initial modeling assumptions may not seem appropriate (e.g., unlimited population and Poisson arrivals), we believe that the approach has potential. With regard to the modeling assumptions, it is important to remember that the model solutions are describing only the equilibrium or, equivalently, long-run averages. Moreover, the closed model assumptions are often not entirely appropriate either. In many situations where closed models are applied, the total population is not nearly fixed. The FPM equilibrium solution may better describe these systems. Moreover, we can modify the FPM method in various ways to obtain a better description.

Even when a closed model is deemed appropriate, the open model with the FPM method can be useful because if often provides a convenient approximation for the more difficult closed model. Certainly the required computation is significantly reduced. In some cases, throughputs can be calculated by hand by the FPM method when some computer codes for closed models are unable to obtain any solution. Moreover, in many cases the results are very close.

The FPM method also forms the basis for one procedure to calculate approximate congestion measures for closed non-Markov networks containing multiserver FCFS nodes with nonexponential service-time distributions, using previously developed approximation procedures for open non-Markov networks such as the Queueing Network Analyzer (QNA).[7] In fact, the primary motivation for this work was the desire to modify QNA so that it can analyze closed models as well as open models. The FPM method is one way to do this. Several possible approaches for calculating approximate congestion measures for non-Markov closed networks are described in Section X.

For the basic Jackson network we are now considering, we implement the FPM method by identifying the external arrival rate that yields the specified expected equilibrium number of jobs in the system. The standard application would be to a system that was previously analyzed by a closed model. Consider such a closed model with a designated exit-entry node. In the closed model, an arrival to this node from elsewhere in the network completes its required service, and a departure represents a new job entering the system. To obtain the associated open model, cut the flow into this exit-entry node, let all internal arrivals into this node leave the system, and insert an external Poisson arrival process. The FPM throughput is the rate or intensity of the external Poisson arrival process for which the expected equilibrium population has the specified value. The approximate mean number of jobs at each node is the mean number in the open model with the FPM arrival rate.

In fact, in the FPM procedure described above, it is not necessary to have a special exit-entry node. Any node can serve as the exit-entry node. For the Jackson product-form network we are considering, the equilibrium distribution of the resulting open model is independent of the node chosen. Choosing the exit-entry node can be important, however, if we do not have a product-form network. Then it may also be appropriate to let the new external arrival process be something other than a Poisson process.

In this introductory section we give a few elementary examples to illustrate the FPM method. However, the primary motivation is the need for approximate methods to analyze more complicated models, e.g., with multiple job classes, nonexponential FCFS servers, priorities, etc. It should be clear that the FPM method is a general approach that applies to these more complicated models. We believe that the performance of the FPM method for Jackson networks indicates the performance that can be expected for more complicated models.

Example 1. Consider a closed Markov cyclic network of single-server FCFS queues with K jobs of a single class. Let there be n_1 nodes having mean service time 1 and n_2 nodes having mean service time τ, arranged in any order. As usual, cyclic means that all departures from node j go next to node $j + 1$ for $1 \leq j \leq n_1 + n_2 - 1$ and all departures from node $n_1 + n_2$ go next to node 1. To apply the FPM method, cut the flow into one node, let all original arrivals on that arc leave the system, and insert an external Poisson arrival process. We identify the external arrival rate in the associated open model, say λ, such that the expected equilibrium total population is K. Since we have a cyclic network, the arrival rate at each node is the external arrival rate. (Otherwise, we would have to solve the traffic rate equations.) Recalling that the equilibrium distribution in the open model is equivalent to independent M/M/1 queues, we solve

$$\frac{n_1\lambda}{1-\lambda} + \frac{n_2\lambda\tau}{1-\lambda\tau} = K$$

for λ, which is a quadratic equation. (If there were m different service rates, then we would have a polynomial of degree m.)

To illustrate, if $K = 20$, $n_1 = n_2 = 10$ and $\tau = 1.2$, then the approximate throughput is = 0.45. In Section IV we prove that this is a lower bound for the throughput in the original closed model. In (15) we suggest as a possible improvement $\lambda(n_1 + n_2 + K)/(n_1 + n_2 + K - 1)$, which in this case is 0.46. The actual throughput in the original closed network also turns out to be 0.46. This is easily determined using any software package for closed Markovian networks of queues; we used PANACEA.[8]

Since the mean service times at different nodes do not differ much, we could also use a quicker approximation, based on a linear equation instead of the quadratic equation, obtained by assuming that all $n_1 + n_2$ nodes have mean service time $\bar{\tau} = (n_1 + n_2\tau)/(n_1 + n_2)$, and then applying the FPM method, which yields the almost instantaneous approximation $K\bar{\tau}/(K + n_1 + n_2) = 20/44 = 0.45$ for the throughput.

This last balanced network approximation can also be used directly in the closed network, which corresponds to combining the last two suggested improvements. (See Section III.) The resulting approximate throughput by this method is $K\bar{\tau}/(K + n_1 + n_2 - 1) = 0.47$. This direct balanced network approximation for closed networks is in fact an upper bound on the throughput in the closed model, as was first shown by Zahorjan et al.[9] (also see Refs. 10 and 11).

If we use $\lambda = 0.45$ as the approximate throughput by the FPM method, then with the M/M/1 formula the mean number of jobs at each node with mean service time 1 (1.2) is 0.82 (1.18); for the closed model, it is 0.83 (1.17). The expected sojourn time at each of these nodes by the FPM method is 1.82 (2.62); for the closed model, it is 1.79 (2.53). For practical purposes, the standard congestion measures calculated by the FPM method agree with those for the closed model in this example.

Note that the FPM solution does not change if we multiply the population and the number of nodes of each type by a common constant. It turns out that the quality of the approximation improves as the network grows in this way. On the one hand, this means that the FPM method does not perform well when there are few nodes, e.g., when $n_1 = n_2 = 1$ here. On the other hand, the FPM method tends to perform well for the large models that are more difficult for closed network algorithms. In fact, in Section V we prove that the FPM method is asymptotically correct for such growing closed networks. This asymptotic property of large closed networks was apparently first observed by Gordon and Newell.[2] We contribute by providing a rigorous proof based on the local central limit theorem for sums of independent and identically distributed random vectors.[12] Also, we stress the significance of the FPM method in this asymptotic analysis. Algorithms for closed models have difficulty as the number of nodes increases. Evidently, no existing closed-network algorithm is able to handle the case of 200 nodes and 200 customers for this numerical example. With the aid of new asymptotic theory,[13,14] PANACEA[8] is able to solve much larger networks, but the asymptotic theory does not apply to this example because it requires a decoupling infinite-server node (see Section 1.4).

In general, when we apply the FPM method, we do not get a quadratic equation. However, the expected equilibrium population in the newly created open network is an increasing function of the external arrival rate, so that it is not difficult to identify the external arrival rate which yields the desired fixed population mean by a search procedure. In fact, it is usually possible to quickly obtain the desired throughput with a programmable hand calculator that can find the roots of an equation. (At the expense of some added complexity, this same general approach can be used for multiple job classes. We give an efficient iterative algorithm for special cases involving infinite-server nodes in Section VI.)

However, it is usually not necessary to carry out such a special inversion procedure. As is standard for closed models, we usually want to determine the throughput as a function of the (expected) network population. Hence, we simply solve the open model for a range of possible external arrival rates and express the expected equilibrium population as a function of the external arrival rate. It is then easy to invert the function if desired. Moreover, we can also describe the resulting population variability in the open model as a function of the external arrival rate. Thus, the FPM method consists of little more than using open models in situations where closed models were used before. Our object, then, is to better understand the relationships between these two kinds of models. We propose some algorithms and obtain some insight about when they will work well and when they will not.

When both closed and open models are available, the appropriate model might be chosen according to which better describes the population variability. We suggest using estimates of the population variance to help identify an appropriate model. It turns out that the population variability in the open model is often less than might be expected (Section II), so that the two models are often remarkably similar. For larger networks (large population or many nodes), the differences are often small relative to the quality of data typically available for modeling fitting. When this is the case, the open model is usually preferred because it is much easier to analyze.

The possible advantage of closed models over open models is also reduced if we do not restrict attention to Markov open models. For example, if we use QNA to approximately analyze a non-Markov open model, then we have an additional degree of freedom in modeling the variability because we can select variability parameters for each service-time distribution and each arrival process. If the actual arrivals are scheduled, as in many production systems, then it is natural to use clocked arrivals in QNA, i.e., deterministic interarrival times, which is achieved by setting the variability parameter for the external arrival process equal to zero. With an open model, we are not forced to have a Poisson external arrival process. From direct modeling considerations, the open model with clocked external arrival processes is often more realistic than the closed model.

Furthermore, given that we are actually interested in a closed model, the variability parameters offer the possibility of improved approximations by open models. Since the population constraint in the closed model tends to reduce the variability (see Section IV), a promising heuristic approximation procedure with the FPM method (suggested by H. Heffes) is to reduce the variability parameters in the approximating open model. For example, with a Jackson network of single-server queues, we might treat each node as a D/M/1 queue instead of an M/M/1 queue, but the actual procedure would have to be more sophisticated. The general approach using QNA for non-Markov closed models is to cut the flow into one node and replace it with an external arrival process. First, as described in Section X, we can let the variability parameter of the external arrival process be such that it agrees with the variability parameter of the departure process from the network. We use QNA to calculate approximate variability parameters for the arrival process to each node. Afterwards, to improve the approximation of the closed model, we can systematically reduce all these variability parameters. The reduction should depend on the network parameters, with the variability parameters evidently being reduced less as the number of nodes or the number of jobs increases. We briefly investigate this possibility, but we have just begun studying refined approximation procedures of this kind.

1.3 The finite-waiting-room refinement

We also propose a refinement of the FPM method for approximating closed models, which is especially useful for small models. We apply the network population constraint given for the closed model to each node separately in the open model. When there are K jobs in the closed model, we allow at most K jobs at each node in the open model. However, we implement this Finite-Waiting-Room (FWR) approximation within the product-form equilibrium distribution of the open model. We act as if there is capacity K at each node in the open model, but we do not analyze the modified open model exactly. Instead, we keep the product-form equilibrium distribution in the open model, and modify the distribution of the number of jobs at each node.

If N_i^o is the equilibrium number of jobs at node i in the open model without the refinement, then we use the conditional distribution of N_i^o given that $N_i^o \leq K$. Since N_i^o has the distribution of a birth-and-death process in a Jackson network, this conditional distribution obtained simply by truncating the original distribution at K and renormalizing is tantamount to imposing a finite waiting room at that node in isolation. [This conditioning can also be used as an approximation for more general models (see Refs. 15 and 16).]

Let $\bar{N}_i^o$ be the equilibrium number of jobs at node i by the FWR method; then

$$P(\bar{N}_i^o = k) = P(N_i^o = k \mid N_i^o \leq K) \;= P(N_i^o = k)/P(N_i^o \leq K). \quad (1)$$

For an M/M/1 queue, the mean is $EN_i^o = \rho_i/(1 - \rho_i)$ and the utilization is $u_i^o = P(N_i^o > 0) = \rho_i$, where $\rho_i = \lambda_i/\mu_i$ is the traffic intensity at node i, based on the net arrival rate λ_i and service rate μ_i. The corresponding quantities $E\bar{N}_i^o$ and $\bar{u}_i^o$ with the FWR method are

$$\begin{aligned} E\bar{N}_i^o &= w(\rho_i, K)EN_i^o \\ w(\rho_i, K) &= (1 - (K + 1)\rho_i^K + K\rho_i^{K+1})/(1 - \rho_i^{K+1}) \\ \bar{u}_i^o &= P(\bar{N}_i^o > 0) = (\rho_i - \rho_i^{K+1})/(1 - \rho_i^{K+1}), \end{aligned} \quad (2)$$

provided that $\rho_i \neq 1$ (see Section 2.5 of Ref. 17). Obviously, $E\bar{N}_i^o < EN_i^o$ and $\bar{u}_i^o < u_i^o$. If we let $\bar{u}_i^o$ be free, fix u_i^o, and let $E\bar{N}_i^o = EN_i^o$, then $\bar{u}_i^o > u_i^o$ (see Section IV), which is a refinement in the right direction. Moreover, if u_i^c is the utilization of node i and N_i^c is the equilibrium number of jobs at node i in the closed model, then $\bar{u}_i^o \leq u_i^c$ when $E\bar{N}_i^o \leq EN_i^c$. Since the ratio of the utilizations at any two nodes in the closed model is the same as in the open model,[5,6] we thus obtain a valid lower bound on u_i^c by this procedure, namely,

$$u_i^c \geq \bar{u}_i^* \equiv \min_j\{(u_i^o/u_j^o)\bar{u}_j^o\}\,. \quad (3)$$

We need (3) to obtain the valid lower bound because the property $u_i^o/u_j^o = u_i^c/u_j^c$ for all i and j does not hold for $\bar{u}_i^o/\bar{u}_j^o$. [See (15) and Section IV.]

Example 2. Consider a closed Markov network with n single-server nodes and K jobs. Let the service rates and net arrival rates be identical, so that the equilibrium distribution is symmetric. When $n = 4$ and $K = 2$, the server utilizations by direct analysis of the closed model, the FPM method, and the FPM/FWR method are, respectively, 0.400, 0.333, and 0.384. By using the FWR refinement, the error is reduced from 42 percent to 4 percent.

Using the FWR refinement to the FPM method can yield external arrival rates for which $\rho_i \geq 1$ at some nodes. Limited numerical experience indicates that the quality of the approximation often deteriorates in this case.

1.4 Decoupling infinite-server nodes

There need not be many nodes for the FPM method to be effective. The FPM method is particularly appealing to approximately solve closed Markovian networks with few nodes but a large population and an Infinite-Server (IS) node with relatively low service rate. For these models, the FPM method extends easily to multiple job classes. However, the need for help with these difficult models is much less now because an efficient algorithm for them has recently been developed by McKenna, Mitra, and Ramakrishnan,[13,14] which is implemented in their PANACEA software package.[8]

The PANACEA algorithm exploits integral representations and asymptotic expansions to reduce the original large closed network to many much smaller closed networks. Under appropriate conditions, the difficult partition function of the original closed model and related quantities such as the utilization of a particular class at a particular node can be represented by asymptotic expansions in which the coefficients are constructed from the partition functions of the smaller closed networks (the pseudonetworks in Ref. 14, which typically involve at most three classes and a total of seven customers). Moreover, the asymptotic expansions permit a thorough analysis of the truncation errors: The truncation error is less in absolute value than the first neglected term and has the same sign.

The asymptotic expansions underlying the new capabilities in PANACEA are based on several assumptions. First, it is assumed that each class visits an IS node [see (27) of Ref. 14]. Second, it is assumed that the population of each class is large [see (17) of Ref. 14]. Third, it is assumed that the individual service rates at the IS node are significantly lower than the service rates in the rest of the network [see (18) of Ref. 14]. Finally, it is assumed that utilizations of the non-IS nodes are not close to their critical values, i.e., they are not in heavy traffic [see (29) through (31) of Ref. 14]. It is worth noting that these assumptions are often realistic—e.g., in computing systems where the IS nodes correspond to "think times" at terminals.

It turns out that the FPM method tends to work well under these same conditions. Unlike PANACEA, however, the FPM method is an approximation. (The asymptotic expansions in PANACEA also can be regarded as approximations, but of a different kind; they are a numerical method that can achieve any degree of accuracy given enough computation. On the other hand, the FPM method changes the model, so that the answers are good only if the two model solutions are close.) The FPM method in this situation can be derived by a procedure that at first seems to be different from the FPM method. This alternate procedure is motivated by the observation that under the stated conditions the departure processes from the special IS nodes tend to behave much like Poisson processes. Moreover, the subnetwork without the IS nodes tends to behave much like an open network with an external Poisson arrival process. This is partly substantiated by previous work[18] in which we showed that, under appropriate conditions, the departure process from an IS node with a fixed general stationary arrival process and a general service-time distribution approaches a Poisson process that is independent of the arrival process as the individual service rate at the IS node decreases. Reference 18 does not directly apply here because the arrival processes at the IS node are changing too, but Ref. 18 suggests that the departure processes from the IS nodes are approximately Poisson processes that are independent of the rest of the network under the stated assumptions. Corresponding limit theorems for the situation here are contained in Section VIII.

The key to our procedure for these models, as with the asymptotic expansions underlying PANACEA, is a large population and the presence of IS nodes with relatively low service rates. The FPM method can be used more generally, but there is stronger supporting logic with the IS nodes. The FPM method works well if there are several IS nodes, as long as each class visits one of them, but for simplicity we assume that there is a single IS node visited by all classes. Also, each class can visit more than one IS node, but we assume that only one IS node has relatively low service rate, so that jobs tend to accumulate there. We use this IS node to decouple the network. We let its departure process leave the system and replace it by an external arrival process. The external arrival process is a set of independent Poisson processes, with one Poisson process for each class. Equivalently, there is a simple Poisson external arrival process and fixed probabilities that each arrival belongs to one of the classes. We approximately solve the original closed network by identifying the appropriate external arrival rates for the associated open model. We use the special IS node to determine what rates are appropriate. We do this by simply equating the arrival and the departure rates for each class at the IS node. It turns out that this procedure is equivalent to the FPM method discussed above (see Section VI).

Example 3. To illustrate the FPM method with an IS node having relatively low service rate, we consider a central processor model treated by McKenna, Mitra, and Ramakrishnan.[13] This is a closed cyclic network with only two nodes. The first node is the CPU, which has a single server, where service is provided according to the processor-sharing discipline. The second node is a "think" node, which is an IS node, representing independent delays at terminals before a job is next sent to the CPU. (Because of insensitivity properties,[5,6,19,20] only the means of the service-time distributions matter for the equilibrium distribution. We can also equivalently regard the CPU as an FCFS node with an exponential service-time distribution.) We shall consider the case of one job class, which is test problem I described in Table I

Table I—A comparison of throughputs using closed and open models for the two-node single-class network model of a central processor in Example 3 of Section 1.4

	Throughput or Utilization of CPU							
	For the Closed Model			For the Open Model Using FPM Method				
	From Ref. 13						Taylor's Series (39)	
Number of Jobs	By CADS	By PANACEA	Version 2.1 of PANACEA	First Upper Bound (30)	First Lower Bound (31)	FPM Solution (28) and (38)	Two Terms	Three Terms
10	0.0417	0.0414	0.0415	0.0417	0.0415	0.0415	0.0415	0.0415
50	0.207	0.207	0.2073	0.2083	0.2072	0.2072	0.2074	0.2072
100	Breakdown	0.413	0.4138	0.4167	0.4137	0.4137	0.4150	0.4143
200	Breakdown	0.819	0.8204	0.8333	0.8123	0.8150	0.8298	0.8269

of Ref. 13. The mean service times at the two nodes are 1 and 240, respectively, so indeed the IS node has relatively low individual service rate.

This example was significant in Ref. 13 because it demonstrated the advantage of PANACEA over previous closed network algorithms, in particular CADS.[21] In several cases CADS was unable to obtain a solution. This example is also significant here because for it the FPM method is both easy and accurate. The FPM method only requires the solution of a quadratic equation [see (38) in Section VII]. A quick upper bound on the CPU utilization can be found by simply multiplying the population times the IS individual service rate [see (30) in Section VI]. Both of these methods perform remarkably well. The throughput results for several population sizes are described in Table I. The CPU node has a processor-sharing service discipline with mean processing time of $\mu_1^{-1} = 1$. Hence, the throughput of jobs at the CPU is equal to the utilization of the CPU. Node 2 is an infinite-server delay node representing the think time of users at terminals. The mean think time is $\mu_2^{-1} = 240$.

For the cases in Table I it is apparent that even the trivial upper bound is adequate for practical purposes. Moreover, as we have remarked, the FPM throughput itself is a lower bound (see Section IV), so that from the FPM method alone we can determine that the quality of the approximation is satisfactory. Since the FPM throughput is a lower bound, from Table I we see that in some cases the FPM throughput is actually slightly more accurate than the published values in Ref. 13, but of course the differences are not significant for practical purposes. In the difficult case of 200 jobs, Version 2.1 of PANACEA terminated with lower and upper bounds of 0.8129 and 0.8204, based on four terms of the asymptotic expansion. In the other cases the two bounds coincide for the specified accuracy. The main point is that essentially the same answers can be obtained quickly by hand. (See Sections VI and VII for additional discussion.)

With the FPM method we avoid closed networks and the associated partition functions entirely. Instead, we approximately solve the original closed network by solving a related open model. In the case of multiple job classes, we iteratively solve a sequence of associated open networks. By working with open networks, we never calculate the complete distribution of the number of jobs of each class at each node. With open networks it suffices to work with expected values. By exploiting simple monotonicity properties, we are also able to give upper and lower bounds on the desired approximate solution at each iteration. Finally, we are able to treat very general networks; e.g., the subnetwork can have multiserver nodes. In fact, the approximation procedure is ideal for the closed-model analogs of the non-Markov networks analyzed by QNA,[7] in which there are FCFS nodes with nonexponential service-time distributions. For the first step of the analysis, in which we replace the departure process from the IS node by an external Poisson arrival process, the FPM method is still asymptotically correct. Hence, for these closed non-Markov models with decoupling IS nodes, it appears that the FPM method with QNA should perform about as well as the original QNA approximation for the corresponding open non-Markov model.

1.5 The rest of this paper

The rest of this paper is organized as follows. In Section II we discuss population variability in open networks, and show that it tends to be relatively small in large networks, which supports using the FPM method to approximate large closed models. We also suggest measuring population variability to help decide which model to use. In Section III we compare the throughputs in closed models and open models (using the FPM method) for a special class of balanced networks, and we show that the differences are, for practical purposes, negligible when the network is large (but not when the network is small). This example provides a convenient, simple, quantitative characterization of the difference between the models, as far as throughputs are concerned. The explicit balanced network results also suggest possible refinements of the FPM method for unbalanced networks.

In Section IV we present theoretical results about the closed network and the FPM approximation. We prove that the utilizations and throughput calculated by the FPM method are always lower bounds for the corresponding quantities in the closed network (see Theorem 1). For the special case of single-server and infinite-server nodes, this result can also be deduced from Zahorjan.[22] We not only treat general nodes such as multiserver nodes, but we treat models with several job classes. To make our comparison and establish other properties of the closed model, we exploit the log-concavity[23] of the distribution of the number of jobs at each node in the associated open model. In various ways we show that the distribution of jobs is less variable in the closed network than in the associated open network. In particular, given ordered means at any node, we establish increasing concave stochastic order (see Theorem 2).[24] To do this, we introduce and apply the notion of one distribution being log-concave relative to another (see Definition 1).

It is intuitively clear that the population constraint should introduce negative dependence among the queue lengths at the different nodes. In Section IV we also show that recently developed concepts of multivariate negative dependence, such as reverse-rule distributions[25,26] and negative association,[27] are ideally suited to make this idea precise. Indeed, the multivariate distribution of jobs at the different nodes has all of these properties. The closed Markovian network model can be regarded as a canonical example of negative dependence.

In Section V we present some additional theory to show that the FPM method tends to perform well for large networks. We prove that the FPM method is asymptotically correct as the number of nodes in the closed model increases with the number of customers per node held fixed (see Theorem 7). For this result, we can let the network grow in a cyclic manner so that the connectivity does not increase. To obtain a rigorous proof, we apply the local central limit theorem for partical sums of random vectors.[12] This result applies to Example 1 as a special case.

In Section VI we present the variant of FPM method to approximate closed networks with a large population and a decoupling IS node with relatively low service rate. We exploit monotonicity to obtain an efficient algorithm for multiple job classes. In Section VII we illustrate the FPM method in this context by considering the central processor model in Example 3. We consider cases involving two job classes as well as one. This example is taken from Ref. 13, so that we can conveniently compare the FPM method to numerical results for PANACEA and the CADS algorithm for closed models.[8,21]

In Section VIII we present theoretical results to support the FPM method in the context of Sections VI and VII. We show that the vector-valued queue-length process in the subnetwork of the closed model without the IS node converges in distribution to the corresponding stochastic process in the approximating open model with a Poisson external arrival process as the populations increase and the individual service rates at the IS node decrease appropriately. We establish convergence in distribution (weak convergence[28,29]) of both the stochastic processes (see Theorem 8) and the equilibrium distributions (see Theorem 9). Convergence of the departure process from the IS node to a Poisson process is established as in Refs. 18 and 30; convergence of the associated vector-valued queue-length process is established by model continuity.[31,32] As a consequence, the FPM method is asymptotically correct for the closed model under these conditions (see Theorem 12). In Section VIII we also make stochastic comparisons between the stochastic processes in the open and closed models, exploiting couplings or almost-surely ordered constructions as in Refs. 33 and 34. The stochastic comparisons are interesting in their own right, but they also play a role in establishing the convergence. We show that a first upper bound for the FPM method is also an upper bound for the closed model in terms of both transient and equilibrium throughputs and queue lengths (see Corollaries to Theorems 10 and 11).

In Section IX we discuss similar approximations for closed networks with a bottleneck node which is not an IS code. We propose a different approximation for closed networks with a bottleneck node. We delete the bottleneck node from the closed network, but we do not use the FPM method. Instead, the proposed approximation is to simply use the open model obtained by deleting the bottleneck node and replacing its departure process by an external arrival process generated by the service times at the bottleneck node. The difference between the total population and the expected population in the open subnetwork is the suggested approximation for the expected population at the bottleneck node. This approximation method is also asymptotically correct as the population grows. The vector-valued queue length process in the subnetwork of the closed network without the bottleneck node converges in distribution to the corresponding process in the open network as the population grows. This phenomenon is of course quite well known,[35,36] but some of the supporting theory here seems to be new.

In Section X we discuss methods for approximately solving non-Markov closed models. We indicate how existing procedures for non-Markov open networks such as the Queueing Network Analyzer (QNA) can be modified for this purpose. In particular, we describe in detail the changes in Ref. 7 to implement the FPM method.

In Section XI we provide some additional motivation for considering special algorithms to analyze non-Markov closed networks. It is sometimes claimed that Markov models with exponential service-time distributions adequately describe throughputs for single-server FCFS nodes with nonexponential service-time distributions with the same mean, but we show that this is not always the case. We use tight lower bounds on the throughput in closed models with FCFS single-server nodes and general service-time distributions identified by Arthurs and Stuck.[11] For highly variable distributions, the actual throughput can be much less than predicted by the Markov model. In fact, the Markov model can be arbitrarily bad. The true throughput can be arbitrarily close to zero, while the Markov model throughput is arbitrarily close to one.

In Section XII we make additional numerical comparisons that help put the different models and approximation procedures in perspective. We draw some conclusions in Section XIII.

This paper contains diverse material, ranging from heuristic algorithms and examples to theorems and proofs. These are intended to complement each other, but the primary algorithm sections (Sections VI and X) and mathematics sections (Sections IV, V, and VIII) can be read independently.

1.6 Other bounds and approximations

In this paper we introduce several approximation procedures and establish several bounds for networks of queues. Of course, many other approximations and bounds have already been developed by others. In addition to the previously mentioned balanced network bounds and others in Refs. 9 through 11, there are useful bounds in Refs. 37 through 39. There is great potential for combining them in new ways. We focus on the basic Markov network models and natural non-Markov extensions obtained by allowing nonexponential service-time distributions and non-Poisson arrival processes. However, the results also have relevance for more complicated models and other approximation procedures.

For example, open-model representations such as the FPM method can be applied in conjunction with aggregation-decomposition approximation methods for closed networks, as suggested by Zahorjan.[22] The basic approach is to replace a subnetwork of a closed network by a single "composite" node with a state-dependent service rate (see pp. 165 through 172 of Ref. 6). For the product-form models, by Norton's theorem, the aggregation step is exact if when there are m_j jobs of class j in the subnetwork, the composite node service rate of class j is precisely the throughput rate for class j for the subnetwork in isolation (as a closed model with that population vector) (see pp. 100 and 106 of Ref. 6). The FPM method can be used as an approximation here to calculate approximate throughputs for the subnetworks.

The FPM method is a very natural idea, so no doubt it has been considered before. In fact, we have indicated that it appears in the asymptotic analysis in Ref. 2. The FPM method is also intimately related to another approximation procedure, which is called the Approximate Infinite Source (AIS) method and was proposed independently by Fredericks.[40] The idea in the AIS method is to replace a finite source by an infinite source, in particular a Poisson process, so that Little's formula[6,19] relating the throughput, expected population, and expected sojourn time remains valid. However, since the original population with a finite source is fixed, it is easy to see that this constraint is equivalent to making the expected equilibrium population in the open model coincide with the actual population in the closed model (with the finite source). Hence, aside from additional refinements, the AIS and FPM methods coincide. Fredericks illustrates the effectiveness of this approach with other examples, including a two-class priority service system with separate finite sources.

II. POPULATION VARIABILITY IN AN OPEN NETWORK

At first glance, it might seem that the open model with fixed mean number of jobs would always differ dramatically from the associated closed model with fixed actual number of jobs. It might seem that the population variability in the open network would necessarily be much greater than in the closed network (for which there is no population variability at all). Indeed, for networks with few nodes there typically is a dramatic difference in the variability, but it turns out that the population variability of an open network tends to decrease as the number of nodes increases. This suggests that the FPM method should work well for large networks.

In one sense the variability in an open network increases as the number of nodes increases. Since the equilibrium numbers of jobs at the different nodes in an open model are independent, the variance of the equilibrium population is the sum of the variances at the nodes. So, roughly speaking, the population variance tends to grow as the network grows (assuming that the marginal distributions at the individual nodes do not change).

2.1 The population squared coefficient of variation

However, we believe that the squared coefficient of variation (the variance divided by the square of the mean) is usually a better measure of the relevant variability than the variance. It describes the variability relative to the mean. Suppose that in the open network under consideration there are n single-server nodes with the traffic intensity at node i being ρ_i. Since the equilibrium number, N_i^o, of jobs at node i has a geometric distribution, the mean, variance, and squared coefficient of variation of N_i^o are

$$\begin{aligned} E(N_i^o) &= \rho_i/(1-\rho_i), \\ \mathrm{Var}(N_i^o) &= \rho_i/(1-\rho_i)^2, \qquad \text{and} \quad c^2(N_i^o) = 1/\rho_i. \end{aligned}$$

Obviously, $c^2(N_i^o)$ can be arbitrarily large, so that we cannot expect the variability to be small with one single-server node.

The associated parameters of the equilibrium total number, N^o, of jobs in the entire network are

$$\begin{aligned} E(N^o) &= E(N_1^o) + \cdots + E(N_n^o), \\ \mathrm{Var}(N^o) &= \mathrm{Var}(N_1^o) + \cdots + \mathrm{Var}(N_n^o), \\ c^2(N^o) &= \mathrm{Var}(N^o)/E(N^o)^2. \end{aligned} \tag{4}$$

When the traffic intensities are all equal, i.e., $\rho_i = \rho$ for all i, $c^2(N^o) = 1/n\rho$, so that $c^2(N^o)$ tends to decrease rapidly as the number of nodes increases. There is a law of large numbers effect when there are many nodes.[41] This is also true as n increases when the traffic intensities are unequal, provided that $E(N_i^o)$ is asymptotically negligible compared to $\sum_{i=1}^n E(N_i^o)$. By the central limit theorem,[41] the distribution of N^o tends to be approximately normally distributed with the mean and variance in (4).

It is easy to see that $c^2(N^o)$ can be very large if there are relatively few nodes all in light traffic. If there is a single bottleneck node in heavy traffic, then $c^2(N^o) \approx 1$, which may also be regarded as significantly different from zero. This indicates that the FPM method might not be desirable with a bottleneck node (see Section IX). However, if there is no bottleneck node and if there are several nodes with at least moderate traffic intensity (and any number of other nodes), then $c^2(N^o)$ will not be large. For example, if there are six nodes with $\rho_i = 2/3$ for each i, then $c^2(N^o) = 0.25$.

2.2 Several servers

It is also worth noting that the equilibrium distribution at a node usually tends to become less variable as the number of servers increases, so that the single-server case we have just considered tends to be the worst case for variability. This is perhaps not true for IS nodes, which often are "delay" nodes. If λ_i is the arrival rate, μ_i is the individual service rate, and $\alpha_i = \lambda_i/\mu_i$ at node i, then since IS nodes have a Poisson distribution,

$$E(N_i^o) = \alpha_i, \qquad \mathrm{Var}(N_i^o) = \alpha_i \quad \text{and} \quad c^2(N_i^o) = 1/\alpha_i.$$

If λ_i and μ_i are the same as in the single-server case, then so is $c^2(N_i^o)$. If λ_i/μ_i tends to increase as the number of servers increases, then $c^2(N_i^o)$ decreases as well. In an M/M/s queue, it is possible to show that $c^2(N_i^o)$ decreases as s increases and converges to 0 as $s \rightarrow \infty$, provided that we fix the probability of delay (see Ref. 42). With $\rho_i = \lambda_i/\mu_i s_i$, this is tantamount to having $(1 - \rho_i)\sqrt{s_i} \rightarrow \beta_i$, $0 < \beta_i < 1$, as s_i increases. In other words, if we only increase s_i, then ρ_i decreases and λ_i/μ_i remains unchanged, but if we adjust ρ_i as we change s_i to reflect the corresponding congestion, then the distribution tends to concentrate. In particular, we then have $E(N_i^o) \rightarrow \infty$, $\mathrm{Var}(N_i^o) \rightarrow \infty$, and $c^2(N_i^o) \rightarrow 0$ as $s \rightarrow \infty$.

2.3 Practical implications

The rather informal analysis in this section indicates that the population variability measured by $c^2(N^o)$ in an open network will often be surprisingly small if (1) the network has quite a few nodes, (2) the network is not in light traffic, and (3) the network is not dominated by one or two bottleneck nodes. Perhaps the most important idea is the possibility of using $c^2(N^o)$ to help determine whether an open model or a closed model is more appropriate. In an application it seems appropriate to measure the real system and estimate the population mean and variance. Then estimate $c^2(N^o)$ for the open model to judge the quality of the fit.

2.4 Reducing the variability of the arrival processes

As mentioned in Section I, we might try to improve the open-model approximation of a closed model by artificially reducing the variability of the arrival processes in the open model. For example, we might replace each M/M/1 queue by an $E_k/M/1$ queue, where E_k is an Erlang distribution with the same mean. Of particular interest is the limiting case as $k \longrightarrow \infty$, D/M/1. It is significant that indeed both $\mathrm{Var}(N_i^\circ)$ and $c^2(N_i^\circ)$ decrease as we increase k; in fact, $\mathrm{Var}(N_i^\circ)$ and $c^2(N_i^\circ)$ for D/M/1 are the least possible values among all GI/M/1 queues with the same arrival rate and service rate. To see this, recall that for a GI/M/1 queue

$$\begin{aligned} \mathrm{Var}(N_i^\circ) &= \rho_i(1 - \rho_i + \sigma_i)/(1 - \sigma_i)^2 \quad \text{and} \\ c^2(N_i^\circ) &= (1 - \rho_i + \sigma_i)/\rho_i, \end{aligned} \tag{5}$$

where σ_i is the probability of delay, which is the root of the equation

$$\phi_i(\mu_i(1 - \sigma_i)) = \sigma_i \tag{6}$$

for

$$\phi_i(s) = \int_0^\infty e^{-st} dF_i(t) \tag{7}$$

with F_i the interarrival-time cdf of the GI/M/1 model for node i (see II.3 of Ref. 43).

The relationship is appropriately expressed in terms of stochastic orderings. We say that one random variable X_1 is less than or equal to another X_2 in the sense of *stochastic order* (denoted by $X_1 \leq_{\mathrm{st}} X_2$), *increasing convex order* (denoted by $X_1 \leq_{\mathrm{ic}} X_2$), and *convex order* (denoted by $X_1 \leq_c X_2$), respectively, if $Eg(X_1) \leq Eg(X_2)$ for all nondecreasing, nondecreasing convex, and convex real-valued functions g for which the expectations are well defined (see Sections 1.3 through 1.5 of Ref. 24). Since $g(x) = x$ and $g(x) = -x$ are both convex, convex order implies equal means. With equal means, convex order is equivalent to increasing convex order. It is significant that $D \leq_c E_{k+1} \leq_c E_k \leq_c M \leq_c H_2$ for random variables with a common mean. (H_2 is a hyperexponential distribution, the mixture of two exponential distributions.)

Let W be the equilibrium waiting time before beginning service. Stoyan and Stoyan showed that $W_1 \leq_{\mathrm{ic}} W_2$ in two GI/G/1 queues with common service-time distribution when $X_1 \leq_c X_2$, where X_i is the generic interarrival time in system i.[44] For the special case of GI/M/1, Rolski and Stoyan showed that $W_1 \leq_{\mathrm{st}} W_2$ under the same condition.[45] Since $\sigma_i = P(W_i > 0)$, $\sigma_1 \leq \sigma_2$ and, by (5), $\mathrm{Var}(N_1) \leq \mathrm{Var}(N_2)$ and $c^2(N_1) \leq c^2(N_2)$. Since $EX \leq_c X$ for any X, these quantities are minimized among GI/M/1 queues by the D/M/1 case.

Table II shows how the variability is reduced by comparing the

Table II—A comparison of congestion measures for the M/M/1 and D/M/1 queues[67]

Traffic Intensity, ρ_i	M/M/1		D/M/1		
	EN_i^o	$c^2(N_i^o)$	Delay Probability, σ_i	EN_i^o	$c^2(N_i^o)$
0.10	0.11	10.00	0.000	0.10	9.00
0.20	0.25	5.00	0.007	0.20	4.04
0.30	0.43	3.33	0.041	0.31	2.47
0.40	0.67	2.50	0.107	0.45	1.77
0.50	1.00	2.00	0.203	0.63	1.41
0.60	1.50	1.67	0.324	0.89	1.21
0.70	2.33	1.43	0.467	1.31	1.10
0.80	4.00	1.25	0.629	2.16	1.04
0.90	9.00	1.11	0.807	4.66	1.01
0.95	19.00	1.05	0.902	9.69	1.002
0.98	49.90	1.02	0.960	24.50	1.000

principal congestion measures for the M/M/1 and D/M/1 queues. The variability is reduced the most for traffic intensities near 0.5; e.g., for $\rho_i = 0.5$ the D/M/1 value is 70 percent of the M/M/1 value. [From heavy traffic theory,[42] we know that, as $\rho_i \rightarrow 1$, $c^2(N_i^o)$ approaches 1 for all GI/M/1 queues and, for D/M/1, EN_i^o approaches one-half the M/M/1 value.] This analysis shows that the proposed technique for refining the approximations by artificially reducing the variability parameters of the arrival processes would indeed reduce the variability of the number of jobs at each node. It also indicates by how much. Since the means would also decrease, this device would also increase the throughput with the FPM method. However, it remains to determine how much to reduce the variability of arrival processes and whether this will produce a good general approximation procedure for network models.

III. COMPARING THROUGHPUTS IN BALANCED NETWORKS

3.1 The closed model with single-server nodes

In this section, we compare the throughput in a closed model with the throughput in the associated open model using the FPM method. We still consider the Markov Jackson models, but for simplicity we restrict attention to single-server nodes. Given a closed model containing n single-server nodes and a fixed population, K, construct the associated open model by removing the arrival process to one node, say the first node, and replacing it by an external Poisson arrival process with sufficiently low arrival rate to have stability. Let the original arrivals to this entry node in the closed model leave the system. Then solve the traffic equations to obtain the arrival rates λ_i and traffic intensities ρ_i for each node i. Note that $n - 1$ of the n

traffic-rate equations for the original closed model are the same as for the new open model.[6] Since there is one degree of freedom in the traffic-rate equations for the closed model, the arrival rates λ_i calculated for the open model are legitimate relative traffic rates for the closed model; i.e., the ratio of the arrival rates at any two nodes thus is identical in the closed and open models.

Let N_i^o and N^o be the equilibrium number of jobs at node i and in the entire system for the open model, and let N_i^c and N^c be the corresponding quantities for the closed model. Obviously, $N^c = K$. For the open network with the given external arrival rate, the expected equilibrium number of jobs is

$$E(N^o) = \sum_{i=1}^{n} (\rho_i/(1 - \rho_i)). \tag{8}$$

It is somewhat remarkable that the equilibrium distribution of the number of jobs at each node in the original closed model can be found by considering the associated open model we have constructed, even though the arrival process we have removed is not a Poisson process. The equilibrium distribution of the numbers of jobs at each node in the original closed model can be expressed exactly in terms of the solution for the open model constructed above;[1-6] it is

$$P(N_i^c = k_i,\ 1 \le i \le n) = \frac{P(N_i^o = k_i,\ 1 \le i \le n)}{P(N^o = K)} = G \prod_{i=1}^{n} \rho_i^{k_i}, \tag{9}$$

where G is the normalization constant or partition function chosen so that the probabilities sum to one over the set of n-tuples $(k_1, k_2, \cdots, k_n)$ such that $k_1 + k_2 + \cdots + k_n = K$. [Of course, this is partly explained by the fact that ratio of arrival rates at any two nodes are the same in the open and closed models. The one remaining degree of freedom, the arbitrary arrival rate in the open model, cancels in the division in (9).]

The associated throughput in the closed model, say θ^c, is then the flow through the designated node, i.e., the utilization u_1^c times the service rate μ_1:

$$\theta^c = u_1^c\mu_1 = P(N_1^c > 0)\mu_1. \tag{10}$$

As usual, the throughput can be obtained by calculating the normalization constant recursively over smaller populations and subsets of nodes (see Section 5.5 of Ref. 6).

3.2 The special case of a balanced network

From (8) through (10), it is clear that the relation between the throughput and the total population is much more elementary for open models; for open models, we only need to know the expected total population, not the detailed distribution at the nodes. We make an interesting explicit comparison, by considering the special case in which the traffic intensities at all the nodes are identical, say ρ. For the open model, (8) reduces to $EN^o = n\rho/(1 - \rho)$, so that if we set $EN^o = K$, we obtain $\rho = K/(K + n)$. Thus, the utilization u_1^o and throughput θ^o in the open model are

$$u_1^o = K/(K + n) \quad \text{and} \quad \theta^o = \lambda_1 = K\mu_1/(K + n) \tag{11}$$

because all external arrivals but no internal arrivals go to the designated first node.

On the other hand, for the closed model,

$$1/G\rho^K = A_{K,n} = \binom{n + K - 1}{K}, \tag{12}$$

the number of ways K indistinguishable objects can be placed into n cells (p. 38 of Ref. 41). Similarly,

$$P(N_1^c = 0) = A_{K,n-1}/A_{K,n} = (n - 1)/(n + K - 1), \tag{13}$$

so that the utilization and throughput in the closed network are

$$u_1^c = K/(K + n - 1) \quad \text{and} \quad \theta^c = K\mu_1/(n + K - 1). \tag{14}$$

From (11) and (14), we see that the two throughputs are very similar. Moreover, $\theta^c > \theta^o$, so that θ^o is a conservative estimate of θ^c. (This is always true; see Section IV.)

It is significant that a good approximation for the throughput θ^c in the closed model immediately provides a good approximation for the utilizations of all the nodes. As we have indicated above, the closed and open models are linked together in a very important way: The ratio of any two utilizations is always identical in both models; i.e.,

$$u_i^o/u_j^o = u_i^c/u_j^c \tag{15}$$

for all i and j.

The difference between the two utilizations, say Δ, which for the balanced model is the difference between the throughputs in (11) and (14) normalized by dividing by the service rate μ_1, is

$$\Delta \equiv u_1^c - u_1^o = (\theta^c - \theta^o)/\mu_1 = K/(n + K)(n + K - 1). \tag{16}$$

Note that $\theta^c = \mu_1$, $\theta^o = \mu_1/2$, and $\Delta = 1/2$ when $K = n = 1$. We conjecture that $\Delta \geq 1/2$, in general. The difference Δ is small if either n or K (especially n) is large but not if n and K are both small. Representative values of n, K, θ^c, θ^o, and Δ are given in Table III. In Table III the service rate at the entry node is $\mu_1 = 1$. We also describe the population variability in the open model using (4) and (11), from

Table III—A comparison of throughputs for the network of single-server nodes with common relative traffic intensities considered in Section III

Network Parameters		Throughput Measures			
Nodes n	Jobs K	Closed θ^c	Open θ^o	Difference in Throughputs Δ (16)	Population Variability $c^2(N^o)$
2	2	0.67	0.50	0.17	1.00
2	5	0.83	0.71	0.12	0.70
2	20	0.95	0.91	0.04	0.55
5	2	0.33	0.29	0.04	0.70
5	5	0.56	0.50	0.06	0.40
5	20	0.83	0.80	0.03	0.25
20	2	0.10	0.09	0.01	0.55
20	5	0.21	0.20	0.01	0.25
20	20	0.51	0.50	0.01	0.10

which $c^2\,(N) = (K + n)/Kn$. The difference between the two models as described by both Δ and $c^2(N^o)$ decreases as n and K increase. Table III quantifies the differences.

Formulas (11) through (16) indicate that the FPM approximation for the throughput will be good if either the population, K, or the number of nodes, n, is large, but with single-server nodes it seems much better to have n large. The quality of the approximate queue-length distributions computed by the FPM method often deteriorates when there are nodes with high utilizations and few servers. Example 1 in Section I is ideal for the FPM method; both K and n are large ($K = n = 20$), but the utilizations are not. The finite-waiting-room refinement in Section 1.3 is useful for the small models.

3.3 Simple approximations for unbalanced networks

We can use the results for balanced networks to obtain simple approximations for unbalanced networks. A simple rough approximation, say θ^c_{approx}, for the throughput in a closed network with K jobs and n single-server nodes with unequal (but not too different) utilizations based on (11) and (14) is

$$\theta^c_{\text{approx}} = \theta^o(n + K)/(n + K - 1), \qquad (17)$$

where θ^o is obtained from the associated open model using (8), e.g., by simple search. We would not expect (17) to be good if there is a severe bottleneck node; we would be in serious trouble if we had six nodes, five having relative utilization 1 and the other having relative utilization 3. We also would not count nodes in relatively light traffic; if we had nine nodes, three with relative utilizations 1 and 6 with relative utilization 3, then it would be better to use $n = 6$ in (17).

We can also directly approximate θ^o by replacing the traffic intensities at each node with the average traffic intensity over all nodes. This yields $\theta^o_{\text{approx}} = K\mu_1/(K + n)$ as in (11). Due to the convexity of EN^o_i, $\theta^o_{\text{approx}} \geq \theta^o$, which is a modification in the correct direction if we wish to approximate θ^c.

Finally, we could combine these two approximations to obtain (14) for unbalanced closed networks, but it appears that this would tend to overestimate θ^c. In fact, (14) has already been shown to be an upper bound for the closed model.[9-11] The simple approximation (17) worked very nicely in Example 1 in Section 1.2.

3.4 Reducing the variability of the arrival processes

As in Section 2.4, we can consider approximations for the closed model obtained by reducing the variability in the open model. Since $EN^o_i = \rho_i/(1 - \sigma_i)$ in the GI/M/1 model, EN^o_i also decreases as σ_i decreases, so reducing the variability of the arrival process at each node increases the throughput θ^o in the open model. (Typical values of EN^o_i for the D/M/1 queue are given in Table II.) If there are n identical GI/M/1 nodes, then instead of (11) we have

$$\theta^o = K\mu_1/(n + xK), \tag{18}$$

where $x = \sigma/\rho$. Since σ depends on ρ via (6), (18) is harder to solve. Moreover, a direct application of the D/M/1 model need not yield good results because (18) can be much greater than (11). For example, if $K = 10$, $n = 16$, and $\mu_1 = 1$, then $\theta^c = 0.40$, while $\theta^o = 0.385$ and 0.50 via (11) and (18), respectively. It remains to determine how to exploit this approach.

3.5 Several servers

It is also interesting to consider networks of n identical multiserver nodes (back with the Markov models). When there are s servers with $1 < s < \infty$, the formulas are rather complicated, but the situation simplifies greatly for $s = \infty$. Then $EN^c_i = EN^o_i = K/n$ and $\theta^c = \theta^o = K\mu_1/n$, so that there is no difference at all. We conjecture that $(\theta^c - \theta^o)/\mu$, decreases in s, which would mean that the single-server case we have examined gives the worst approximation.

For the open model in which each node has s servers and a common traffic intensity, $\rho = \lambda_1/s\mu_1$, θ^o can be approximated by solving

$$n[\rho s + \delta\rho/(1 - \rho)] = K, \tag{19}$$

where δ is the probability of delay at node 1 (which also depends on λ_1) (see Ref. 42). A possible procedure is to approximate δ first and then solve the resulting quadratic equation for λ. One could then iterate, recalculating δ, etc.

IV. SUPPORTING THEORY FOR COMPARING THE MODELS

4.1 *A lower bound for the closed model*

For the special example in Section 3.2, we saw from (11) and (14) that $\theta^c > \theta^o$. In general, we should expect the throughput to be greater in the closed model because it is intuitively obvious that N_i^c is less variable than N_i^o. Given the same mean, N_i^o is evidently more likely to assume both very large values and very small values, so that we should have $P(N_i^o = 0) > P(N_i^c = 0)$. Of course, we need not actually have $EN_i^o = EN_i^c$ when $EN^o = N^c$, but this is the idea.

In this section we justify this reasoning. We assume that the open model is constructed from the closed model as described in Section 3.1. We consider the Markov Jackson network with one job class and multiserver nodes as specified in Section I, but it is significant that the throughput comparisons extend to Markov networks with multiple job classes and more general state-dependent service rates at the nodes. Some of these comparisons also apply to the finite-waiting-room approximation introduced in Section 1.3. To avoid complicated notation, we only discuss these extensions in remarks after Theorem 4.

Theorem 1: If $EN^o \le N^c$, then $\theta^o \le \theta^c$ and $u_i^o \le u_i^c$ for all i.

For the special case in which all nodes are either single-server or IS nodes, Zahorjan[22] has given a nice proof of Theorem 1. We give a different argument that allows us to treat more general nodes, e.g., multiserver nodes, and obtain some interesting additional results along the way. To establish Theorem 1, we use notions of *concave ordering*, which are closely related to the convex orderings introduced in Section II (see Section 1.4 of Ref 24). One random variable X_1 is less than or equal to another X_2 in concave (increasing concave) ordering, denoted by $X_1 \le_{\text{cv}} X_2$ ($X_1 \le_{\text{icv}} X_2$), if $Eg(X_1) \le Eg(X_2)$ for all concave (increasing concave) real-valued functions g for which the expectations are defined. The connection to the convex orderings is simple: $X_1 \le_{\text{cv}} X_2$ if and only if $X_1 \ge_c X_2$; $X_1 \le_{\text{icv}} X_2$ if and only if $-X_1 \ge_{\text{ic}} -X_2$. The following basic characterization for random variables with values in the nonnegative integers is useful: $X_1 \le_{\text{icv}} X_2$ if and only if

$$\sum_{k=0}^{n} P(X_1 \le k) \ge \sum_{k=0}^{n} P(X_2 \le k) \tag{20}$$

for all n (see Sections 1.3 through 1.5 of Ref. 24).

As a basis for Theorem 1, we establish the following result.

Theorem 2: If $EN_i^o \le EN_i^c$ for any node i, then $N_i^o \le_{\text{icv}} N_i^c$.

In fact, Theorem 2 directly implies Theorem 1 given (15). Let u_i^c and u_i^o be the utilizations of node i in the closed and open models, respectively. They are both defined as the expected number of busy servers; e.g., for the open model,

$$u_i^o = E(\min\{N_i^o, s_i\}) = \rho_i s_i = \lambda_i/\mu_i, \tag{21}$$

where, of course, λ_i is the net arrival rate determined by the traffic rate equations plus the external arrival rate, which is, in turn, determined by the FPM requirement that $EN^o = N^c$. Formula (21) is also valid for the closed model.

To prove Theorem 1, we use the following consequence of Theorem 2.

Corollary to Theorem 2: If $EN_i^o \leq EN_i^c$, then $u_i^o \leq u_i^c$.

Proof: Apply Theorem 2 observing that the function in (21) is increasing and concave. □

Proof of Theorem 1: If $EN^o \leq N^c$, then $EN_i^o \leq EN_i^c$ for some i because $\sum_{i=1}^n EN_i^c = EN^c = N^c$. For one such i, $u_i^o \leq u_i^c$ by Theorem 2 and its corollary. By (15), $u_i^o \leq u_i^c$ for all i. Since $\theta^c = u_1^c\, \mu_1$ and $\theta^o = u_1^o\, \mu_1$, $\theta^c > \theta^o$ too. □

To prove Theorem 2, we use notions of log-concavity (see p. 70 of Ref. 23). A probability mass function $\{p_k, k \geq 0\}$ is *log-concave* if

$$p_k^2 \geq p_{k+1}p_{k-1}, \qquad k \geq 1. \tag{22}$$

A log-concave distribution is unimodal; moreover, it is strongly unimodal, i.e., the convolution with any unimodal distribution is also unimodal. In fact, for discrete distributions log-concavity, strong unimodality, and the PF_2 (Polya frequency function) property are all equivalent.[23] The equilibrium distribution of any birth-and-death process is log-concave if the birth rates are nonincreasing and the death rates are nondecreasing (see example 5.7F in Ref. 23). Moreover, log-concavity is preserved under convolution. Hence, for each i and m the distributions of N_i^o and $N_1^o + \cdots + N_m^o$ are log-concave. (By example 5.7F in Ref. 23 referred to above, it suffices for the service rate at each node to be a nondecreasing function of the number of jobs present.) It turns out that this is also true for the more complicated distributions in the closed network.

Theorem 3: Let the service rate at each node be a nondecreasing function of the number of jobs present. For any m the distribution of $N_1^c + \cdots + N_m^c$ is log-concave.

Proof: Consider $m = 1$. Since log-concavity is preserved under convolution,[23] the distribution of $N_2^o + \cdots + N_n^o$ is log-concave. Then note that

$$\frac{P(N_1^c = k+1)}{P(N_1^c = k)} = \frac{P(N_1^o = k+1)P(N_2^o + \cdots + N_n^o = K - k - 1)}{P(N_1^o = k)P(N_2^o + \cdots + N_n^o = K - k)}, \tag{23}$$

with the right-hand side being the product of two ratios, both decreasing in k. A similar argument applies to $m > 1$. □

From (23), we see that in some sense the distribution of N_i^c is more log-concave than the distribution of N_i^o. We now formalize this notion.

Definition 1: One probability mass function $\{p_k^1, k \geq 0\}$ is said to be log-concave *relative* to another $\{p_k^2, k \geq 0\}$ if $(p_{k+1}^1 p_k^2)/(p_k^1 p_{k+1}^2)$ is nonincreasing in k.

From (23) it is obvious that N_i^c is log-concave relative to N_i^o. We now show that this supplies what we need for Theorem 2. The key property is that relative log-concavity implies that the ratio p_k^1/p_k^2 is unimodal.[23]

Theorem 4: If the distribution of a random variable X_2 is log-concave relative to the distribution of another random variable X_1 and $EX_1 \leq EX_2$, then $X_1 \leq_{\text{icv}} X_2$.

Proof: Our goal is to verify (20). We first show that $P(X_1 = 0) \geq P(X_2 = 0)$. If not, then the relative log-concavity implies that there is a k_0 such that $P(X_1 = k) < P(X_2 = k)$ for all $k \leq k_0$ and no $k > k_0$. This would make X_1 stochastically larger than X_2, implying that $EX_1 > EX_2$, which contradicts an assumption. Hence, $P(X_1 = 0) \geq P(X_2 = 0)$. Next let k_1 be the first k, if any, such that $P(X_1 \leq k_1) \leq P(X_2 \leq k_1)$. By the relative log-concavity, we must have $P(X_1 \leq k) \leq P(X_2 \leq k)$ for all $k \geq k_0$. Since $EX_i = \sum_{k=0}^{\infty} P(X_i \geq k)$,

$$EX_2 - EX_1 = \sum_{k=0}^{\infty} [P(X_1 \leq k) - P(X_2 \leq k)] > 0$$

and

$$\sum_{k=0}^{n} P(X_1 \leq k) \geq \sum_{k=0}^{n} P(X_2 \leq k), \qquad n \geq 0,$$

which establishes (20). □

Proof of Theorem 2: By (23), the distribution of N_i^c is log-concave relative to the distribution of N_i^o according to Definition 1. By Theorem 4, $N_i^o \leq_{\text{icv}} N_i^c$. □

Remarks: 1. In a network made up entirely of infinite-server nodes, we have $u_i^o = u_i^c$ for all i and $\theta^o = \theta^c$, so that we cannot have strict inequality in Theorems 1 and 2.

2. Theorems 2 through 4 apply to the FPM/FWR method introduced in Section 1.3. Let $\bar{N}_i^o$ be the equilibrium number at node i by this method. It is easy to see that N_i^c is log-concave relative to $\bar{N}_i^o$, which in turn is log-concave relative to N_i^o. Hence, if $EN_i^o \leq E\bar{N}_i^o$, then $N_i^o \leq_{\text{icv}} \bar{N}_i^o$; if $E\bar{N}_i^o \leq EN_i^c$, then $\bar{N}_i^o \leq_{\text{icv}} N_i^c$. However, this does not yield a proof of the analog of Theorem 1 because the relationship (15) is lost. We do obtain the lower bound (3), though.

3. As indicated at the beginning of this section, Theorems 1 through 4 extend to multiple job classes. There are many different ways to define the class structure, but we shall use only basic properties that have been established for the Markov models.[5,6] For the open model, the vector of jobs at the nodes without identifying the classes has the same equilibrium distribution as when there is only a single class, and given any number of jobs at node i in equilibrium, each job is of class j with some probability p_{ij}, independently of the other jobs. In other words, if N_i^o is the total number of jobs at node i and N_{ij}^o is the number of class j jobs at node i, then N_{ij}^o is obtained from N_i^o Bernoulli trials with probability p_{ij}:

$$P(N_{ij}^o = k) = \sum_{n=k}^{\infty} P(N_i^o = n) \binom{n}{k} p_{ij}^k (1 - p_{ij})^{n-k}. \tag{24}$$

The key property is that the distribution of N_{ij}^o in (24) is log-concave whenever the distribution of N_i^o is log-concave. This result is intuitively reasonable, but not so easy to prove. The result is established in Theorem 2 of Ref. 46. Given that N_{ij}^o has a log-concave distribution and that $N_{i_1 j}^o$ is independent of $N_{i_2 j}^o$ when $i_1 \neq i_2$, it is easy to extend all of the previous results in this section to multiple job classes. For example, the extension of Theorem 1 states that the utilizations of each class at each node are ordered, i.e., $u_{ij}^o \leq u_{ij}^c$ for all i and j, if the expected class populations are ordered, i.e., $\sum_{i=1}^{n} EN_{ij}^o \leq \sum_{i=1}^{n} N_{ij}^c = K_j$ for all j.

4. We have indicated that p^1 being log-concave relative to p^2 implies that p_k^1/p_k^2 is unimodal in k. We call this relationship Uniform Conditional Variability Order (UCVO), provided that p^1 and p^2 are not stochastically ordered, because all conditional distributions, conditioning on a common subset, are either ordered again by UCVO or are ordered by ordinary stochastic order. This property parallels uniform conditional stochastic order,[47,48] and is studied further elsewhere.[49]

4.2 Dependence in the closed model

So far in this section (in Theorem 2 and Remark 4 above), we have shown how to express the idea that the distribution of the number of jobs at each node is less variable in the closed model than in the open model, but we have yet to describe the joint distribution at several nodes. Unlike in the open model, where the marginal distributions are independent, in the closed model the marginal distributions are dependent. If there are more jobs in one subset of nodes, then there should be fewer jobs at another disjoint subset of nodes. The population constraint obviously should make the populations at different nodes negatively correlated. We can make these ideas precise using recently developed concepts of negative dependence.

One concept of negative dependence is the Multivariate Reverse

Rule distribution (MRR_2), which was introduced by Karlin and Rinott.[25] Let p be a multivariate probability mass function on the n-fold product of the nonnegative integers. The distribution p is said to be MRR_2 if

$$p(x \vee y)p(x \wedge y) \leqslant p(x)p(y) \qquad (25)$$

for all $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ in $\{0, 1, \cdots\}^n$, where

$$x \vee y = (\max\{x_1, y_1\}, \cdots, \max\{x_n, y_n\}) \quad \text{and}$$

$$x \wedge y = (\min\{x_1, y_1\}, \cdots, \min\{x_n, y_n\}).$$

In contrast, if p satisfies (25) with the inequality reversed, p is said to be Multivariate Totally Positive (MTP_2).[50] In both cases it suffices to check (25) for x and y differing in only two components.

Unlike MTP_2 distributions,[50] the marginal distributions of an MRR_2 distribution need not be MRR_2. Moreover, even having the marginal distributions all MRR_2 is not strong enough to deduce some of the desired multivariate inequalities. Karlin and Rinott[25] proposed one way to cope with this difficulty, by introducing a special subclass called the strongly MRR_2 ($SMRR_2$) distributions. An *n-dimensional* probability mass function p is $SMRR_2$ if the $(n - m)$-dimensional function $\sum p(x_1, \cdots, x_n)\ \phi_1(x_{j_1}) \cdots \phi_m(x_{j_m})$ is MRR_2 for all m and all m-tuples of indices $(j_1, \cdots, j_m)$, where the sum is over all $(x_{j1}, \cdots, \mathrm{x}_{jm})$ and ϕ_i is log-concave (PF_2) for each i.

Block, Savits, and Shaked[26] introduced a convenient structural condition (condition N) that implies $SMRR_2$. A random vector $(X_1, \cdots, X_n)$ satisfies condition N if there is a vector of $n + 1$ independent random variables $(Y_0, Y_1, \cdots, Y_n)$ each with a PF_2 density (or mass function) such that $(X_1, \cdots, X_n)$ is distributed the same as $[(Y_1, \cdots, Y_n) \mid Y_0 + \cdots + Y_n = s]$ for some s. It is easy to see that condition N applies to the closed models as a special case; just set $Y_0 = 0$ and $s = K$.

Another concept of negative dependence was proposed by Joag-Dev and Proschan.[27] They call random variables $X_1, \cdots, X_n$ and their joint distribution Negatively Associated (NA) if, for every pair of disjoint subsets A_1 and A_2 of the index set $\{1, 2, \cdots, n\}$, the covariance

$$\mathrm{cov}(f((X_i, i\epsilon A_1)), g((X_j, j\epsilon A_2))) \leq 0 \qquad (26)$$

for all nondecreasing real-valued functions f and g defined on R^{k_1} and R^{k_2}, where k_i is the cardinality of A_i. Their Theorem 8 directly implies that $(N_1^c, \cdots, N_n^c)$ is negatively associated. We collect these properties in Theorem 5.

Theorem 5: The vector $(N_1^c, \cdots, N_n^c)$ is negatively associated, satisfies condition N, is $SMRR_2$, and has all marginal distributions MRR_2.

Proof: Since $(N_1^c, \cdots, N_n^c)$ is distributed as $(N_1^o, \cdots, N_n^o | N_1^o + \cdots + N_n^o = K)$, condition N in Ref. 26 and the sufficient condition for the NA property in Theorem 2.8 of Ref. 27 are immediate. As noted in the Remark at the end of Section IV of Ref. 26, condition N implies $SMRR_2$, which in turn implies that all marginals are MRR_2. □

Remark: It is elementary to directly verify that all marginals are MRR_2.

Many important consequences of Theorem 5 are described in Refs. 25 through 27. We give some illustrative examples.

Corollary 1 to Theorem 5: Suppose that ϕ_i are all nondecreasing or all nonincreasing functions on the nonnegative integers. Then, for any k, $1 \le k \le n$,

$$E\{\phi_1(N_1^c)\phi_2(N_2^c) \cdots \phi_n(N_n^c)\}$$
$$\le E\{\phi_1(N_1^c) \cdots \phi_k(N_k^c)\} \times E\{\phi_{k+1}(N_{k+1}^c) \cdots \phi_n(N_n^c)\}.$$

Proof: Apply (1.5) of Ref. 25, noting that the PF_2 property is not used there.

Corollary 2 to Theorem 5: For any $m \le n$ and any m-tuple $(k_1, \cdots, k_m)$,

$$P(N_i^c \le k_i, 1 \le i \le m) \le \prod_{i=1}^{m} P(N_i^c \le k_i)$$

and

$$P(N_i^c \ge k_i, 1 \le i \le m) \le \prod_{i=1}^{m} P(N_i^c \ge k_i).$$

Proof: Apply Corollary 1. □

Remark: Theorem 5 and its corollaries describe multivariate dependence for a single job class. The multiple-job-class closed Markov network suggests a natural generalization of condition N in Ref. 26, which we call condition QN. A random vector $X \equiv (X_{ij}: 1 \le i \le n, 1 \le j \le m)$ in R^{mn} satisfies condition QN if it is appropriately related to another random vector. The other random vector is $Y \equiv (Y_{ij}: 0 \le i \le n, 1 \le j \le m)$ in $R^{m(n+1)}$ such that (1) the subvectors $(Y_{0j}: 1 \le j \le m)$, $(Y_{1j}: 1 \le j \le m)$, $\cdots$, $(Y_{nj}: 1 \le j \le m)$ are mutually independent, (2) the random variables $\sum_{j=1}^{m} Y_{ij}$ have a PF_2 density or mass function for each i, and (3) given $\sum_{j=1}^{m} Y_{ij}$, $(Y_{i1}, \cdots, Y_{im})$ has a multinomial distribution for each i. We say that X satisfies condition QN if X is distributed as $(Y_{ij}: 1 \le i \le n, 1 \le j \le m)$ conditional on $\sum_{i=0}^{n} Y_{ij} = s_j$, $1 \le j \le m$, for some m-tuple $(s_1, \cdots, s_m)$. For $m = 1$, of course condition QN reduces to condition N. Condition QN is being investigated; a discussion of its properties is intended for a future paper.

4.3 Changing the population

With closely related stochastic comparison concepts, we can also describe what happens when we increase the population in a closed network. Let $(N_1^c(K), \cdots, N_n^c(K))$ be the equilibrium vector of jobs at each node as a function of the total population K. Naturally, we expect it to be increasing in K in some sense. In fact, this is true in a very strong sense. Following Karlin and Rinott,[50] we say that one multivariate probability mass function p_1 is less than or equal to another p_2 in the sense of multivariate Monotone Likelihood Ratio (MLR), and we write $p_1 \leq_{\text{lr}} p_2$, if

$$p_1(x)p_2(y) \leqslant p_1(x \wedge y)p_2(x \vee y) \tag{27}$$

for all $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ in $\{0, 1, \cdots\}^n$. The *MLR* order is a generalization of MTP_2 because $p \leq_{\text{lr}} p$ if and only if p is *MTP*.[50] *MLR* order implies stochastic order for the original distributions (i.e., $p_1 \leq_{\text{st}} p_2$) and also for all conditional distributions conditioning on sublattices.[48]

The probability mass functions of the full vector $[N_1^c(K), \cdots, N_n^c(K)]$ are not usefully compared by (27) for different K because they have disjoint support sets. However, we can usefully compare the marginal distributions. There is a further complication, however, because (27) will obviously fail when x and y are in the support of both distributions but $x \vee y$ is not. However, the ordering (27) does hold over every sublattice of the support.

Theorem 6: Given a closed model with n nodes, $[N_1^c(K), \cdots, N_{n-1}^c(K)]$ *is nondecreasing in K in the MLR ordering in the sense that the support set* $\{(k_1, \cdots, k_{n-1}) \colon k_1 + \cdots + k_{n-1} \leq K\}$ *in* $\{0, 1, \cdots\}^{n-1}$ *is nondecreasing in K and* (27) *holds for K and* $K + 1$ *as an equality whenever the sum of the components of* $x \vee y$ *is less than or equal to* $K + 1$.

Proof: It suffices to establish (27) for x and y differing by one in just two indices, say 1 and 2. Let $x = (k_1 + 1, k_2, k_3, \cdots, k_{n-1})$ and $y = (k_1, k_2 + 1, k_3, \cdots, k_{n-1})$. Let p_K be the probability mass function of $[N_1^c(K), \cdots, N_n^c\ (K)]$. Then (27) holds as an equality provided that $k_1 + k_2 + k_{n-1} + 2 \leq K + 1$ because

$p_K\,(x)\,p_{k+1}\,(y)/p_K\,(x \wedge y)\,p_{K+1}\,(x \vee y)$

$$= \frac{P\left(N_n^o = K - \sum_{j=1}^{n-1} k_j - 1\right) P\left(N_n^o = (K+1) - \sum_{j=1}^{n-1} k_j - 1\right)}{P\left(N_n^o = K - \sum_{j=1}^{n-1} k_j\right) P\left(N_n^o = (K+1) - \sum_{j=1}^{n-1} k_j - 2\right)}$$

$= 1.$ □

Corollary 1 to Theorem 6: For each i and K, $N_i^c(K) \leq_{\text{lr}} N_i^c(K + 1)$.

Corollary 2 to Theorem 6: The utilizations $u_i^c(K)$ are increasing in K.

Proof: Apply Corollary 1 with the increasing function in (20). □

Corollary 3 to Theorem 6: The conditional distribution of $[N_1^c(K), \cdots, N_{n-1}^c(K)]$ given any sublattice of $\{0, 1, \cdots, m\}^{n-1}$ with maximal element $(k_1, \cdots, k_{n-1})$ is independent of K for $K \geq \sum_{j=1}^{n-1} k_j$.

Corollary 4 to Theorem 6: $P(N_i^c(K) \geq m \,|\, a_j \leq N_j^c(K) \leq b_j, 1 \leq j \leq n-1)$ is independent of K for $K \geq \sum_{j=1}^{n-1} b_j$.

Proof: Apply Corollary 3. □

V. LIMITS FOR GROWING NETWORKS

In this section we provide some additional theory to show that the FPM method tends to perform well as an approximation for closed models when the network is large. For particular kinds of growing networks we prove that the FPM method is asymptotically correct as the number of nodes increases. We assume that the population is nK when there are n nodes. As we indicated in Section 1.2, the basic idea here is due to Gordon and Newell (see pp. 261 through 265 of Ref. 2), but we formulate and prove a limit theorem.

To be precise, we must specify how the network topology and other model parameters change as the closed network grows. If the connectivity increases as the closed network grows, so that the departure processes are split into many components and the arrival processes are superpositions of many components, then it is usually possible to show that any fixed finite subset of nodes in the closed network behaves asymptotically as mutually independent queues (with mutually independent Poisson arrival processes). This is an even stronger form of independence than in the open model because it applies to the time-dependent stochastic processes as well as the equilibrium distribution. A simple example is the n-node network with routing of departures from every node to all other nodes with equal probability. Growing networks with increasing connectivity can be treated by classical limit theorems for superposition and thinning.[51,52]

Motivated by Example 1 in Section 1.2, we formulate a limit theorem in which the connectivity does not grow with n. We define our sequence of closed models as follows. We start with a general open Markov product-form network having q nodes, p job classes, and a p-tuple of independent Poisson processes determining the arrivals (one for each class). We then replicate this network n times, letting the departures from network k be the arrivals to network $k + 1$. We let the routing probabilities at each subnetwork be identical. Finally, we make it a closed model by replacing the external arrival process to network 1 by the departure process from network n and stipulating that there are nK_j customers of class j, $1 \leq j \leq p$. A symmetric closed cyclic network is the special case in which the initial building-block network has one node. Example 1 in Section 1.2 can be regarded as the special case in which the initial building-block network is a network with two single-server nodes in series, one with mean service time 1.0 and the other with mean service time 1.2. The numerical calculations were for $n = 10$.

The following result shows that the FPM method is asymptotically correct for such cyclically growing networks as $n \rightarrow \infty$. In Remark 1 following the proof, we indicate that Theorem 7 also applies to growing networks with increasing connectivity. Let $N^c_{ijk}(n)$ be the number of class j jobs at node i of the kth subnetwork in the nth closed model; let $N^o_{ijk}(\lambda)$ be the number of class j jobs at node i of the kth subnetwork in the associated open model having independent Poisson arrival processes with rate vector λ at the first subnetwork, with all departures from the nth subnetwork leaving the system.

Theorem 7: For any k_0 and (qpk_0)-tuple $(m_{111}, \cdots, m_{qpk_0})$,

$$\lim_{n\to\infty} P(N^c_{ijk}(n) = m_{ijk}\colon 1 \le i \le q,\ 1 \le j \le p,\ 1 \le k \le k_0)$$

$$= \prod_{k=1}^{k_0} \prod_{i=1}^{q} P(N^o_{ij1}(\lambda) = m_{ijk}\colon 1 \le j \le p) \equiv Z,$$

where $\lambda \equiv (\lambda_1, \cdots, \lambda_p)$ is the FPM solution for the original q-node open building-block network.

Proof: Let $M_j = \sum_{k=1}^{k_0} \sum_{i=1}^{q} m_{ijk}$ for $1 \le j \le p$. As in (9),

$$P(N^c_{ijk}(n) = m_{ijk}\colon 1 \le i \le q,\ 1 \le j \le p,\ 1 \le k \le k_o)$$

$$= \frac{Z \cdot P\left(\sum\limits_{k=k_0+1}^{n} \sum\limits_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j - M_j\colon 1 \le j \le p\right)}{P\left(\sum\limits_{k=1}^{n} \sum\limits_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j\colon 1 \le j \le p\right)}$$

for any vector of external arrival rates λ for which there is stability. (Z is defined in the statement of Theorem 7.) For the special case in which λ is the FPM vector, we can apply the local central limit theorem, pp. 75 through 79 of Spitzer,[12] to obtain our desired result. Suppose that λ is such that $E \sum_{i=1}^{q} N^o_{ij1}(\lambda) = K_j$ for each j. Since $\sum_{k=1}^{n} (\sum_{i=1}^{q} N^o_{ijk}(\lambda) - K_j)$ is the jth component of the nth partial sum of i.i.d. random p-tuples where each component has mean 0, there exists α, $0 < \alpha < \infty$, such that

$$\lim_{n\to\infty} n^{p/2} P\left(\sum_{k=1}^{n} \sum_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j + a_j\colon 1 \le j \le p\right) = \alpha$$

for any p-tuple $(a_1, \cdots, a_p)$. (It is easy to see that the aperiodicity requirement in the local central limit theorem is satisfied.) We thus obtain the desired result by multiplying both the numerator and denominator by $n^{p/2}$ and letting $n \rightarrow \infty$. □

Remarks: 1. Theorem 7 and its proof also apply to other kinds of growing networks in which the connectivity does increase. For example, consider the symmetric n-node network in which each node has the same external Poisson arrival process with rate λ_{oj} and probability r_j of departures leaving the system for class j, $1 \leq j \leq p$, independent of n. Also, let departures staying within the network be routed to all other nodes with equal probability. Then the arrival rate of class j at each node is $\lambda_{oj}/(1 - r_j)$, independent of n. Then the equilibrium vectors of jobs at any node in the open model are the same for all nodes and are independent of n, so that Theorem 7 and its proof applies with $q = 1$.

A more complex symmetric growing model with increasing connectivity that can be treated the same way is obtained by replacing each node in this example by a q-node subnetwork. The departures from each q-node subnetwork staying within the system would be routed to each possible q-node subnetwork with equal probability. Each q-node subnetwork would also have its own external arrival processes. Then the equilibrium vector of jobs of each class at each node in the open model is the same for all q-node subnetworks and is independent of n.

2. Gordon and Newell propose a refinement to the FPM approximation for large networks, (27) in Ref. 2, which is obtained by approximating the probabilities involving the large partial sums in the numerator and denominator of by the normal density function. This is justified by the Remark on p. 77 of Ref. 12.

VI. THE FPM METHOD WITH A DECOUPLING INFINITE-SERVER NODE

We now consider the special case of a closed network with an IS node. As in Section V, we allow p different job classes. We introduce this added complexity here because our algorithm is particularly well suited to cope with it. Let each class have its own population and routing probabilities. Let there be $q + 1$ nodes with node $q + 1$ being an IS node and assume that it is visited by every class. (It would suffice to have different IS nodes visited by different classes. The other nodes visited by any class might include IS nodes too; the designated IS node has especially low service rates.) Let μ_j be the individual service rate of class j at node $q + 1$. Let K_j be the given fixed population of class j, $1 \leq j \leq p$. (We are now in the setting of Ref. 14, except that we are allowing multiserver nodes.)

We now modify the original closed model by removing the departure processes for the p classes from the designated IS node and replacing them by p independent external Poisson arrival processes to the remaining q nodes. Let λ_j be the rate of the Poisson process for class j and let $\underset{\sim}{\lambda} = (\lambda_1, \cdots, \lambda_p)$.

Let $N_{ij}^o(\underset{\sim}{\lambda})$ be the equilibrium steady number of customers of class j at node i in the q-node open network without the IS node based on the Poisson external arrival processes with rate vector $\underset{\sim}{\lambda}$. (We have changed the notation somewhat to emphasize the dependence on $\underset{\sim}{\lambda}$.) We use the designated IS node to determine the appropriate arrival-rate vector $\underset{\sim}{\lambda}$. Since the arrival rates equal the departure rates in the open network, the departure rate of class j from the q-node open subnetwork is also λ_j. Since these departures all leave from the designated IS node, we use the IS node to enforce consistency. In particular, we require that the arrival rate be equal to the departure rate for each class at the IS node, i.e.,

$$\lambda_j = \left(K_j - \sum_{i=1}^{q} EN_{ij}^o(\underset{\sim}{\lambda})\right) \mu_j \tag{28}$$

for each j.

Since the expected equilibrium number of customers in a G/GI/∞ model (with general stationary arrival process) is just the arrival rate divided by the individual service rate [e.g., see (4.2.3) of Ref. 19], eq. (28) would be valid in the original closed model if (1) $\underset{\sim}{\lambda}$ were the vector of arrival rates to the IS node and (2) $\sum_{i=1}^{q} EN_{ij}^o(\underset{\sim}{\lambda})$ were the correct mean number of class j customers in the q-node subnetwork. However, $\sum_{i=1}^{q} EN_{ij}^o(\underset{\sim}{\lambda})$ is in fact an approximation based on both $\underset{\sim}{\lambda}$ and the Poisson assumption. Even if $\underset{\sim}{\lambda}$ were correct, $EN_{ij}^o(\underset{\sim}{\lambda})$ would be an approximation. In Section VIII we show that both conditions are satisfied asymptotically if we let $K_j \rightarrow \infty$, $\mu_j \rightarrow 0$, and $K_j\mu_j \rightarrow \lambda_j$ for each j. (This is completely established only for the case of one job class, but we conjecture that the convergence is valid for multiple job classes too.) Hence, there is reason to expect that the procedure will perform well as an approximation for the closed model under certain conditions. Interestingly, as indicated in Section I, these conditions are the same as those in Ref. 14.

Equation (28) also coincides with FPM method. With the FPM method we approximately solve the original closed model by finding the external arrival rate in the associated $(q + 1)$-node open network that makes the expected equilibrium population precisely K_j for class j. (We now regard the IS node as part of the open network.) However, the expected population of class j customers in the IS node is λ_j/μ_j, so that (28) is equivalent to

$$K_j = \sum_{i=1}^{q+1} EN_{ij}^{o}(\lambda) \tag{29}$$

for each j.

To complete the specification of the FPM procedure here, we must identify the vector λ satisfying (28) for all j. This can usually be done iteratively. The key is to recognize that the vector $\{EN_{ij}^{o}(\lambda), 1 \leq i \leq q, 1 \leq j \leq p\}$ is a strictly increasing continuous function of $(\lambda_1, \cdots, \lambda_p)$. We first bound λ_j above by

$$U_j^{(1)} = K_j\mu_j, \; 1 \leq \; j \leq p. \tag{30}$$

Then we bound λ_j below and above successively by

$$\begin{aligned} L_j^{(k)} &= \left(K_j - \sum_{i=1}^{q} EN_{ij}^{o}(U_1^{(k)}, \cdots, U_p^{(k)})\right)\mu_j \\ U_j^{(k+1)} &= \left(K_j - \sum_{i=1}^{q} EN_{ij}^{o}(L_1^{(k)}, \cdots, L_p^{(k)})\right)\mu_j \end{aligned} \tag{31}$$

for $k \geq 1$, $1 \leq j \leq p$. It is easy to see that

$$L_j^{(k)} < L_j^{(k+1)} < \lambda_j < U_j^{(k+1)} < U_j^{(k)}, \tag{32}$$

$L_j^{(k)} \rightarrow \lambda_j$, and $U_j^{(k)} \rightarrow \lambda_j$ for λ_j satisfying (28). [Use the fact that $EN_{ij}^{o}(\lambda)$ is a continuous strictly increasing function of λ.]

To properly initialize the procedure, we must of course have (30) be feasible arrival rates; i.e., we need stability:

$$EN_{ij}^{o}(U_1^{(1)}, \cdots, U_p^{(1)}) < \infty \tag{33}$$

for all i and j. Moreover, we need

$$\sum_{i=1}^{q} EN_{ij}^{o}(U_1^{(1)}, \cdots, U_p^{(1)}) < K_j \tag{34}$$

for each j to have (31) be feasible rates. Conditions (33) and (34) should hold if indeed K_j is large, μ_j is small, and $K_j\mu_j$ is not too large. If conditions (33) and (34) were violated, we could search for initial conditions satisfying the appropriate monotonicity.

In fact, such elaborate analysis as we have just described is often unnecessary. If, indeed, K_j is large and μ_j is small for each j, with no node in the q-node network in heavy traffic, then it often suffices to use the simple formula (30) as the approximation for the external arrival rate λ_j. As indicated in (32), the approximation (30) yields upper bounds for λ_j and $EN_{ij}^{o}(\lambda)$ for all i and j. Alternatively, the first lower bound in (32) is often a good approximation, see Section VII.

It is intuitively obvious that the first upper bound (30) for λ in (28) is also an upper bound for the vector of throughputs in the original closed model, and we prove this in Section VIII. It also seems plausible that the equilibrium queue length vector $\{N_{ij}^{o}(\lambda)\}$ in the open model with λ in (30) would be a stochastic upper bound for the corresponding random vector in the closed model, and we also prove this for the case of a single job class. (The more general case of multiple job classes remains a conjecture.) Theorem 1 in Section IV establishes for a single job class that the FPM throughput in (28) is a lower bound for the throughput in the original closed model. This extends to multiple job classes by the remark following Theorem 4. Hence, for each job class j, we have

$$L_j^{(1)} \leq L_j^{(k)} \leq \lambda_j \equiv \theta_j^o \leq \theta_j^c \leq U_j^{(1)}. \qquad (35)$$

VII. EXAMPLES WITH A DECOUPLING INFINITE-SERVER NODE

In this section, we illustrate the FPM method in Section VI by returning to Example 3 in Section 1.4, which is the central processor model treated by McKenna, Mitra, and Ramakrishnan.[13] This is a closed cyclic product-form network with two nodes. The first node is the CPU, where service is provided according to the processor-sharing discipline. Equivalently for our purposes because of insensitivity properties,[5,6,19,20] the service discipline can be FCFS with an exponential service-time distribution. The second node is a think node, which is an IS node, representing independent delays at terminals before a job is next sent to the CPU. Again, because of insensitivity properties, only the mean of the service-time distribution at the IS node matters. We shall first consider the case of one job class, which is test problem 2 described in Table I of Ref. 13. Then we consider two job classes and finally we consider the special case of a population of size one to show that the FPM method can perform poorly when the approximating conditions do not nearly hold.

7.1 One job class

The specified model in Ref. 13 is closed with a fixed population size (also referred to as degree of multiprogramming). We shall consider the associated open model obtained by cutting the arrivals to node 1 and inserting an external Poisson arrival process with rate λ. This open model has a very simple solution. To express it, let μ_1^{-1} be the mean processing time for each job at the CPU and μ_2^{-1} the mean think time (individual service time at node 2). Let $\rho_1 = \lambda/\mu_1$ and $\alpha_2 = \lambda/\mu_2$ and assume that $\rho_1 < 1$ to guarantee stability. The equilibrium distribution in the open model has independent marginal distributions with the marginal being geometric at node 1 and Poisson at node 2:

$$P(N_1^o = k_1, N_2^o = k_2) = (1 - \rho_1)\rho_1^{k_1} e^{-\alpha_2}\alpha_2^{k_2}/k_2! \qquad (36)$$

With the FPM method, we set the expected equilibrium total population equal to K; i.e.,

$$EN^o = \frac{\rho_1}{1 - \rho_1} + \alpha_2 = \frac{\lambda}{\mu_1 - \lambda} + \frac{\lambda}{\mu_2} = K. \tag{37}$$

Hence, we obtain the following formula for the external arrival rate λ:

$$2\lambda\mu_1^{-1} = 1 + x(1 + K) - \sqrt{[1 + (1 + K)x]^2 - 4Kx}\,, \tag{38}$$

where $x = \mu_2/\mu_1$. By using a Taylor series expansion in powers of x, we see that

$$\lambda/\mu_1 = Kx - Kx^2 - K(K - 1)x^3 + 0(x^3), \tag{39}$$

so that $\lambda \approx K\mu_2$ when μ_2 and $\mu_2 K$ are sufficiently small compared to μ_1.

In contrast, for the closed model we combine (9) and (36). This is conceptually simple, but the calculation can be complicated for large population sizes. McKenna, Mitra, and Ramakrishnan[13] used this example to illustrate the advantage of PANACEA over previous convolution algorithms for the closed model, such as are contained in the software package CADS.[21] For large population sizes, CADS was unable to obtain a solution, while PANACEA obtained a solution easily. Moreover, for all population sizes, the throughputs calculated by the two methods agree closely.

It is significant that comparable results can be obtained for this example by the FPM method by hand. We do not even need to use (38); we can simply use (30) to obtain $\lambda = K\mu_2$. A comparison of the throughput calculations appears in Table I. In this example we have small μ_2 ($\mu_2 = 1/240$ and $\mu_1 = 1$) and large K ($K = 10$, 50, 100, and 200). Since the FPM method throughput provides a lower bound on the closed network throughput, the FPM answer is essentially exact for $K \leq 100$. As in Refs. 13 and 14, the FPM procedure works best here if μ_2 is small, K is large, and $K\mu_2$ is not too close to μ_1. Under heavy loads, the open-network M/M/1 formula at node 1 keeps the throughput down with the FPM method.

The last two columns of Table I contain the first two and first three terms of the Taylor series expansion in (39); the first term of course corresponds to the first upper bound in (30), which appears earlier in the table. Evidently the algorithm in Section VI converges faster than the Taylor series for larger values of K.

7.2 Two job classes

We now give additional details for test problem 2 in Ref. 13, which differs from Test Problem 1 only by having two job classes. Node 1 (the CPU) is again a processor-sharing node and node 2 is the IS node.

The mean service times for the two classes are 1 and 1.5 at the CPU and 450 and 150 at the IS node, respectively. Let μ_{ij} be the service rate of class j at node i and let K_j be the population of class j. The first upper bounds for the approximate arrival rate, obtained from (30) are

$$U_j^{(1)} = K_j\mu_{2j}. \tag{40}$$

The associated approximate CPU utilization of class j, say ρ_j, is thus $\rho_j \approx K_j\mu_{2j}/\mu_{1j}$ and the associated approximate total utilization of the CPU, say ρ, is $\rho \approx \rho_1 + \rho_2$. The first lower bounds on the approximate arrival rates, obtained from (31), are

$$L_j^{(1)} = (K_j - \rho_j/(1 - \rho))\mu_{2j}. \tag{41}$$

The associated approximation for the total CPU utilization is thus

$$\rho \approx L_1^{(1)}/\mu_{11} + L_2^{(1)}/\mu_{12}. \tag{42}$$

In this case we only compute the first upper and lower bounds. Table IV shows that the approximation procedure works very well. We are able to produce results very close to those given in Ref. 13 by hand in a few minutes. As suggested in Section VI, the first lower bound in (31) and (41) seems to provide a good approximation, even though it is a lower bound. From (35), the first upper and lower bounds from (40) and (41) are upper and lower bounds on the throughput in the closed model.

7.3 Where the FPM method performs poorly

When the populations are not large, the FPM method can perform poorly. This is easily and dramatically demonstrated with the same two-node closed network in which there is a single job. Let node 1 have one exponential server at rate 1 and let node 2 be the designated IS node with individual service rate x.

The exact equilibrium distribution has the job at node 1 with probability $x/(1 + x)$, which is also the associated long-run flow rate

Table IV—A comparison of the approximation method with exact results for the two-class example in Section 7.2

Number of Jobs (Degree of Multiprogramming) Class 1/Class 2	Total Utilization of CPU				
	For the Closed Model			New Approximation	
	From Ref. 13		Version 2.1 of PANACEA	First Upper Bound (40)	First Lower Bound (41)
	CADS	PANACEA			
10/10	0.118	0.119	0.121	0.122	0.121
50/50	0.593	0.60	0.599	0.611	0.598
100/50	Breakdown	0.69	0.706	0.722	0.704
200/10	Breakdown	0.54	0.540	0.544	0.540

out of node 2. However, the actual arrival process to node 1 is a renewal process in which the renewal interval is the sum of two independent exponential variables, one with mean 1 and the other with mean $1/x$. Moreover, the arrival rate depends dramatically on the state.

When we carry out the approximation procedure, we treat node 1 as an M/M/1 queue, so that (28) becomes

$$\lambda = [1 - \lambda/(1 - \lambda)]x, \tag{43}$$

which requires $\lambda < 0.5$ to have a solution. As $x \rightarrow \infty$, $\lambda(x) \rightarrow 1/2$. Obviously, the approximation does not work well in this case. The approximate throughput approaches one-half, while the true value in the closed model approaches 1 as $x \rightarrow \infty$. The normalized difference $\Delta = (\theta^c - \theta^o)/\mu_1$ approaches one-half, which in Section III we conjectured was the lower bound.

VIII. SUPPORTING THEORY WITH AN INFINITE-SERVER NODE

In this section we establish some theoretical results that help explain why and when the FPM algorithm in Section VI approximates the closed models well. As in Section VI, we assume that there is an IS node visited by all classes. We show that the subnetwork of the closed Markov network without the IS node approaches an open Markov network as the populations increase and the service rates at the IS node decrease appropriately (see Theorems 8 and 9). As a consequence, we show that the FPM method is asymptotically correct for the closed model under these conditions (see Theorem 12).

8.1 A sequence of closed models

As in Section VI, there are p job classes and $q + 1$ nodes with node $q + 1$ being the IS node that is visited by every class. We consider a sequence of systems indexed by the superscript n. Let μ_j^n be the individual service rate of class j at node $q + 1$ in the nth system. Let K_j^n be the fixed customer population of class j in system n. As with Poisson approximations for the binomial distribution[41] and as in Ref. 18, the idea is to let $K_j^n \rightarrow \infty$ and $\mu_j^n \rightarrow 0$ in such a way that $K_j^n \mu_j^n \rightarrow \lambda_j$ for each j as $n \rightarrow \infty$.

We let the remaining network structure and parameters be fixed, independent of n; neither the total numbers of nodes $q + 1$ and classes p, nor the parameters of the q-node subnetwork change with n. We still assume the basic Markov Jackson network structure specified in Section I, modified to allow multiple classes, but many of the results extend to more general models (see subsequent remarks).

Let p_{ji} be the probability that a departure of class j from the IS node goes next to node i in the q-node subnetwork. (There could be immediate feedback to the IS node, which occurs for class j with probability $1 - \sum_{i=1}^{q} p_{ji}$.)

Let $A_{ji}^{cn}(t)$ be the counting process in the nth closed system representing the number of departures of class j from the node $q + 1$ in the interval $[0, t]$ that go next to node i. Let $N_{ij}^{cn}(t)$ represent the number of class j customers at node i at time t in the nth closed system. Let $\underset{\sim}{A}_{ji}^{cn}$, $\underset{\sim}{A}^{cn}$, $\underset{\sim}{N}_{ij}^{cn}$, and $\underset{\sim}{N}^{cn}$ represent the associated stochastic processes, i.e.,

$$\begin{aligned}
\underset{\sim}{A}_{ji}^{cn} &\equiv \{A_{ji}^{cn}(t),\ t \geq 0\} \\
\underset{\sim}{A}^{cn} &\equiv \{\underset{\sim}{A}_{ji}^{cn};\ 1 \leq j \leq p,\ 1 \leq i \leq q\} \\
\underset{\sim}{N}_{ij}^{cn} &\equiv \{N_{ij}^{cn}(t),\ t \geq 0\} \\
\underset{\sim}{N}^{cn} &\equiv \{\underset{\sim}{N}_{ij}^{cn};\ 1 \leq i \leq q,\ 1 \leq j \leq p\}. \qquad (44)
\end{aligned}$$

We can initialize the closed networks at time 0 in various ways. For example, we could assume that all $K_1^n + \cdots + K_p^n$ customers initially are at node $q + 1$. We will later simply assume that the initial distributions converge to a proper limit, which includes this situation as a special case.

Let $\Pi_{ji}(\lambda_{ji}) \equiv \Pi_{ji}(\lambda_{ji}, t) \equiv \{\Pi_{ji}(\lambda_{ji}, t),\ t \geq 0\}$ be a Poisson counting process with intensity λ_{ji}, and let $\underset{\sim}{\Pi} \equiv \underset{\sim}{\Pi}\ (\underset{\sim}{\lambda})$ be a pq-dimensional vector of independent Poisson processes with intensities $\underset{\sim}{\lambda} \equiv (\lambda_{ji};\ 1 \leq j \leq p,\ 1 \leq i \leq q)$, i.e.,

$$\underset{\sim}{\Pi}(\underset{\sim}{\lambda}) \equiv \{\Pi_{ji}(\lambda_{ji}),\ 1 \leq j \leq p,\ 1 \leq i \leq q\}. \qquad (45)$$

Let $N_{ij}^{o}(t)$ represent the number of class j customers at node i at time t in the q-node open network obtained by deleting the IS node and replacing its departure process with the external Poisson arrival process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$. By "external" we mean that we have the standard open model in which future arrivals are independent of the network state and history; i.e., $\{\Pi_{ji}(\lambda_{ji}, t + u) - \Pi_{ji}(\lambda_{ji}, t),\ u \geq 0,\ 1 \leq i \leq q,\ 1 \leq j \leq p\}$ is independent of $\{N_{ij}^{o}(s),\ s \leq t,\ 1 \leq i \leq q,\ 1 \leq j \leq p\}$ for each t. In both the open and closed models, successive service times and routings are mutually independent and independent of the history of the network prior to their generation. (At this point we are not using the FPM method in the open model; the arrival rates are simply specified as $\underset{\sim}{\lambda}$.) Let $\underset{\sim}{N}^{o}$ be the associated vector-valued stochastic process.

The following theorem expresses how the q-node subnetwork of the $(q + 1)$-node closed network without the IS node approaches a q-node open network as $n \to \infty$. The convergence of stochastic processes described below is convergence in distribution (weak convergence), which we denote by $\Rightarrow$ (see Refs. 28 and 29 and references there). The stochastic processes are random elements of the function space $D[0, \infty) \equiv D([0, \infty), R^{pq})$.

Theorem 8: Let $\lambda_{ji} = \lambda_j p_{ji}$ for each j and i. If $K_j^n \to \infty$, $\mu_j^n \to 0$, $K_j^n \mu_j^n \to \lambda_j$ for each j, and $\underset{\sim}{N}^{cn}(0) \Rightarrow \underset{\sim}{N}^{o}(0)$ in R^{pq} as $n \to \infty$, where $\underset{\sim}{N}^{o}(0)$ is a proper random vector $[P(N_{ij}^{o}(0) < \infty) = 1$ for all i and j], then as $n \to \infty$

$$(a)\ \underset{\sim}{A}^{cn} \Rightarrow \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$$

and

$$(b)\ \underset{\sim}{N}^{cn} \Rightarrow \underset{\sim}{N}^{o}.$$

(c) If, in addition, $\{[N_{ij}^{cn}(0)]^k\}$ is uniformly integrable, then

$$E[N_{ij}^{cn}(t)]^k \to E[N_{ij}^{o}(t)]^k$$

for each i, j, and t.

In the proof, as in Section IV, we use the notion of stochastic order. One random element (random element of R, R^k, $D[0, \infty)$, etc.) X_1 is stochastically less than or equal to another X_2, denoted by$X_1 \leq_{\text{st}} X_2$, if $Eh(X_1) \leq Eh(X_2)$ for all nondecreasing real-valued functions h for which the expectations are well defined.[33] For this a partial ordering must be defined on the sample space, which we take to be the usual one; e.g., $(x_1, \cdots, x_k) \leq (y_1, \cdots, y_k)$ in R^k if $x_i \leq y_i$ for each i and $\{x(t), t \geq 0\} \leq \{y(t), t \geq 0\}$ in $D[0, \infty)$ if $x(t) \leq y(t)$ for each t.

Proof: (a) The proof follows Ref. 18, which establishes convergence to a Poisson process for the departure process of certain G/GI/∞ queues under similar conditions. The result is not already contained in Ref. 18 because the arrival process to the IS node here is changing with n. However, by Corollary 1 to Theorem 1 in Ref. 18, it suffices to show that $\underset{\sim}{\Lambda}^{cn} \Rightarrow \underset{\sim}{\lambda}\ \omega$, where $\omega(t) = 1$, $t \geq 0$, as in (3.1) of Ref. 18, and $\underset{\sim}{\Lambda}^{cn}$ is the stochastic intensity of the counting process $\underset{\sim}{A}^{cn}$, defined by

$$\Lambda_{ji}^{cn}(t) = [K_j^n - N_{ij}^{cn}(t)]\mu_j^n p_{ji}, \qquad t \geq 0. \tag{46}$$

(For related theory, see Ref. 30 and references there.) The desired weak convergence of $\underset{\sim}{\Lambda}^{cn}$ follows easily from (46) because, for any T,

$$\sup_{0 \leq t \leq T} N_{ij}^{cn}(t) \leq_{\text{st}} N_{ij}^{cn}(0) + \Pi_{ji}(K_j^n \mu_j^n p_{ji}, T), \tag{47}$$

where $\leq_{\text{st}}$ denotes stochastic order defined above and the two quantities on the right are independent. Since we have assumed that $\underset{\sim}{N}^{cn}(0)$ converges and that $N_{ij}^{o}(0)$ is proper and since $K_j^n \mu_j^n p_{ji} \to \lambda_{ji}$ as $n \to \infty$, (47) implies that the sequence $\{\underset{\sim}{N}^{cn}\}$ is uniformly tight (see p. 37 of Ref. 28). This implies that $\underset{\sim}{N}_{ij}^{cn} \mu_j^n \Rightarrow 0\omega$ as $n \to \infty$ and the desired conclusion.

(b) Convergence in distribution of $\underset{\sim}{N}^{cn}$ follows by model continuity as in Refs. 31 and 32. In particular, given part (a), we can construct versions of $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ on the same sample space so that there is convergence of the sample paths, using the Skorohod embedding theorem.[29] Using the same service times and routing in all systems, we obtain convergence of the sample paths of $\underset{\sim}{N}^{cn}$ to $\underset{\sim}{N}^{o}$ with probability one on the specially constructed space. (Since $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ has no fixed jump points, simultaneous transitions need not be considered.) This implies convergence in distribution of the separate stochastic processes.

(c) The stochastic dominance used in part (a) and the new condition imply that the random variables $\{[N_{ij}^{cn}(t)]^k, n \geq 1\}$ are uniformly integrable (see p. 32 of Ref. 28). Part (b) implies that $N_{ij}^{cn}(t) \Rightarrow N_{ij}^{o}(t)$ as $n \rightarrow \infty$ for each i, j, and t. Theorem 5.4 of Ref. 28 thus implies convergence of the moments. □

Remarks: 1. All the conditions on $N_{ij}^{cn}(0)$ hold trivially if all $K_1^n + \cdots + K_p^n$ jobs are initially at the IS node for each n.

2. The conditions in Theorem 8 can be relaxed. The q-node subnetwork of the closed model can be quite general; e.g., the service-time distributions can be nonexponential with FCFS nodes. Our proof only exploits the fact that the service-time distribution at the IS node is exponential. The service-time distribution at the IS node could be made general too, as in Ref. 18, but then we would have to be careful with the initial conditions. If the initial residual service-time distributions at the servers are independent stationary-excess distributions of a service-time distribution that has no mass at zero, then part (a) holds by virtue of the limit theorem for the superposition of independent and identically distributed (i.i.d.) stationary renewal processes.[51] Of course, if the service-time distribution has positive mass at zero, then the limit process is, instead, batch Poisson with geometric batches. □

Theorem 8 implies that the random vectors $\underset{\sim}{A}^{cn}(t)$ and $\underset{\sim}{N}^{cn}(t)$ in R^{pq} converge in distribution as $n \rightarrow \infty$ for each t, but Theorem 8 says nothing about the equilibrium distributions. In fact, we have not yet ruled out the possibility that the open network is unstable; i.e., we could have $N_{ij}^{o}(t) \Rightarrow \infty$ as $t \rightarrow \infty$. Indeed, Theorem 8 is still valid in this case, but now we consider the equilibrium distributions. Let $\underset{\sim}{N}^{cn}(\infty)$ and $\underset{\sim}{N}^{o}(\infty)$ be random vectors with the equilibrium or limiting distributions as $t \rightarrow \infty$. (For the continuous-time Markov chains, they are necessarily unique.) We assume that the limiting Poisson intensities λ_{ij} are small enough so that the equilibrium or limiting distribution for $\underset{\sim}{N}^{o}(t)$ exists.

Theorem 9: Assume that a proper equilibrium distribution exists for $\underset{\sim}{N}^{o}$. Also, assume that either (1) there is a single job class or (2) the *sequence* $\{\underset{\sim}{N}^{cn}(\infty), n \geq 1\}$ *is uniformly tight. Then, under the conditions of Theorem 8,* $\underset{\sim}{N}^{cn}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ *in* R^{pq} *as* $n \longrightarrow \infty$.

We defer the proof of Theorem 9 until we develop some stochastic comparison tools, which are interesting in their own right. We are able to establish the desired stochastic comparison result (see Theorem 11) only when there is a single job class, which explains the second assumption in Theorem 9. We conjecture that the required tightness in Theorem 9 can be proved from the other assumptions for multiple classes.

8.2 Stochastic comparisons

For the counting processes $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$, we use the notion of stochastic order based on conditional failure rates or stochastic intensities, introduced in Ref. 34. The stochastic intensity of the vector-valued stochastic process $\underset{\sim}{A}^{cn}(t) \equiv \{A_{ji}^{cn}(t)\}$ is defined in (46). Of course, the stochastic intensity of the Poisson process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ is the deterministic function $\underset{\sim}{\lambda}\omega$. Following Ref. 34, the counting process $\underset{\sim}{A}^{cn}$ is said to be stochastically less than or equal to the Poisson process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ in the sense of conditional failure rates, here denoted by $\underset{\sim}{A}^{cn} \leq_f \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$, if

$$\Lambda_{ji}^{cn}(t) \leq \lambda_{ji} \tag{48}$$

with probability 1 for all j, i, and t ($\leq_1$ is used in Ref. 34). From (46), it is easy to see that indeed (48) is satisfied. Hence, trivially we have Theorem 10.

Theorem 10: *If* $K_j^n \mu_j^n \leq \lambda_j$ *for each* j, *then* $\underset{\sim}{A}^{cn} \leq_f \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$.

Corollary to Theorem 9: *In the setting of Section VI,* $K_j \mu_j$ *is an upper bound for the expected average throughput for class* j *over any time interval. Hence, the first upper bound for the FPM method in* (*30*) *yields an upper bound for the long-run throughput of each class in the closed method.*

We now establish a general stochastic comparison between $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$. We exploit a coupling or special, almost surely ordered construction, as in Refs. 33 and 34. To establish a general comparison result for $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$, we assume that there is a single customer class. We thus drop the j subscript. We also exploit the fact that the processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ are continuous-time Markov processes, but now the service rate at node i when there are k customers present can be a general nondecreasing function, say $\mu_i(k)$, for $1 \leq i \leq q$.

Theorem 11: *Suppose that there is a single job class with* $K^n \mu^n \leq \lambda$. *Let the processes* $\underset{\sim}{N}^{cn}$ *and* $\underset{\sim}{N}^{o}$ *be Markov with the service rate functions* $\mu_i(k)$ *nondecreasing in* k *for each* i, $1 \leq i \leq q$.

(*a*) *If* $\underset{\sim}{N}^{cn}(0) \leq_{\text{st}} \underset{\sim}{N}^{o}(0)$ *in* R^q, *then* $\underset{\sim}{N}^{cn} \leq_{\text{st}} \underset{\sim}{N}^{o}$ *in* $D[0, \infty)$.

(*b*) *If, in addition, the equilibrium distribution for the open network exists, then also* $\underset{\sim}{N}^{cn}(\infty) \le_{\mathrm{st}} \underset{\sim}{N}^{o}(\infty)$ *in* R^q.

Proof: (a) The argument parallels that of Theorems 6, 7, and 10 in Ref. 34. For more details on the method, see Sonderman.[53] First, Theorem 10 implies that versions of the arrival processors $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ can be constructed on the same probability space so that the points of $A_{ji}^{cn}(t)$ form a subsequence of the points in $\Pi_{ji}(\lambda_{ji}, t)$ for each j ($j = 1$ here) and i (the ordering $\le_2$ in Ref. 34). Next we can construct the service completions for $\underset{\sim}{N}^{cn}$ using the service completions of $\underset{\sim}{N}^{o}$. If there is a service completion at node i in process N^o at time t, then we let there be a corresponding service completion at node i in $\underset{\sim}{N}^{cn}$ with probability $\mu_i(N_i^{cn}(t))/\mu_i(N_i^o(t))$. When there are service completions in both processes, we let the routing be identical. By using induction on the transition epochs, we see that this special construction keeps the sample paths ordered and the distributions of the individual stochastic processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ unchanged.

(b) The stochastic order for each t as a consequence of part (a) is preserved in the limit as $t \to \infty$ (see Proposition 3 of Ref. 33). □

Remarks: 1. It is not difficult to see that Theorem 11(a) is not true for multiple job classes. For example, consider a network with two nodes plus the IS node and three job classes. Let class j jobs go from the IS node to node j and then back to the IS node for $j = 1, 2$. Let class 3 jobs go from the IS node to node 2, then node 1 and then back to the IS node. Let all service rates be identical at nodes 1 and 2. Let K_1^n and K_3^n be large and μ_1^n and μ_3^n be small so that the arrival processes of classes 1 and 3 are both nearly Poisson in the closed model. On the other hand, let $K_2^n = 1$, so that $\underset{\sim}{A}_{22}^{cn}$ is considerably smaller (stochastically) than the Poisson process associated with the open model. Let nodes 1 and 2 be initially empty. For some relatively short initial time interval, say $[0, t]$, in the open model there are more arrivals of class 2 to node 2, with negligible change for classes 1 and 3. These class 2 jobs at node 2 tend to impede the class 3 jobs at node 2, so that the class 3 jobs come to node 1 more slowly in the open model. Hence, the class 1 jobs can get through node 1 more easily; thus, we can have $EN_{11}^o(t) \le EN_{11}^{cn}(t)$ even though $K_1\mu_1 \le \lambda_1$.

2. Even though Theorem 11(a) does not extend to multiple job classes, we conjecture that Theorem 11(b) does. That would be sufficient to eliminate conditions (1) and (2) in Theorem 9.

Proof of Theorem 9: Since $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ are continuous-time Markov processes with the given equilibrium distributions, we can apply Theorem 8(b) here and Lemma 1 of Ref. 31. This implies that the desired convergence $\underset{\sim}{N}^{cn}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ holds provided that $\{\underset{\sim}{N}^{cn}(\infty), n \ge 1\}$ is uniformly tight. We use the fact that $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ have unique equilibrium distributions. For the case of a single job class, the sequence

$\{\underset{\sim}{N}^{cn}(\infty), n \geq 1\}$ is uniformly tight by Theorem 11(b). The stochastic dominance implies the desired uniform tightness because each individual probability measure is tight (see Theorem 1.4 of Ref. 28). □

Remarks: 1. Of course, Theorem 9 applies to other non-Markov product-form models that have the same equilibrium distributions by virtue of insensitivity properties.[3-6,19,20]

2. Theorem 9 also holds for more general service-time distributions in the q-node subnetwork provided that we can establish the uniform tightness. The original processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ can be made Markov by appending supplementary variables.

3. If the service-time distribution for class j at the IS node is phase type instead of exponential, then Theorem 10 remains valid with $\lambda_{ji}^{*} \geq K_j^n \mu_j^{*n}$, where μ_j^{*n} is the maximum phase service rate for class j. If the open network process $\underset{\sim}{N}^{o}$ is stable with the high intensities $\underset{\sim}{\lambda}^{*}$, then we can apply the analog of Theorem 11 to obtain the tightness needed in Theorem 9 (again for a single job class).

We now show that the first lower bound for the FPM method in (31) is a lower bound for the throughput in the closed network. As stated, this follows from (32) and Theorem 1, but we make stronger comparisons using the stochastic intensity $\underset{\sim}{\Lambda}^{cn}$ of the arrival process $\underset{\sim}{A}^{cn}$, defined in (46).

Corollary to Theorem 11: Suppose that there is a single job class with $K^n \mu^n \leq \lambda$. Then, for each i and t,

$$(a)\ \Lambda_{1i}^{cn}(t) \geq_{\mathrm{st}} [K^n - N^o(t)]\mu^n p_{1i}$$

and

$$(b)\ E\Lambda_{1i}^{cn}(t) \geq E[K^n - N^o(t)]\mu^n p_{ij}.$$

If, in addition, both systems are in equilibrium, then

$$(c)\ E\left\{t^{-1}\int_s^{s+t} \Lambda_{1i}^{cn}(u)du\right\} \geq L_1^{(1)} \quad \textit{for all } s \textit{ and } t$$

and

$$(d)\ \theta^c \geq L_1^{(1)}.$$

8.3 The FPM method is asymptotically correct

We now apply Theorems 8 and 9 to deduce that the FPM method in Section VI is asymptotically correct. Due to the second assumption in Theorem 9, we only completely treat the case of one job class. Let $\underset{\sim}{A}^{on}$ and $\underset{\sim}{N}^{on}$ be the vector-valued arrival process and queue length process obtained by using the FPM method with the nth closed model.

Theorem 12: Under the conditions of Theorem 8, as $n \rightarrow \infty$,

$$(a)\ \underset{\sim}{A}^{on} \Rightarrow \underset{\sim}{\Pi}(\underset{\sim}{\lambda}),$$

$$(b)\ \underset{\sim}{N}^{on} \Rightarrow \underset{\sim}{N}^{o},$$

and

$$(c)\ E(N_{ij}^{on}(t))^k \to E(N_{ij}^{o}(t))^k$$

for each i, j, k, and t.

(d) Under the conditions of Theorem 9, for sufficiently large n, $\underset{\sim}{N}^{on}(\infty)$ exists as a proper random vector and $\underset{\sim}{N}^{on}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ in R^{pq}.

Proof: (a) Since $\underset{\sim}{A}^{on}$ is a Poisson process for each n, it suffices to show that the associated arrival rates converge. For this, it suffices to show that the difference between the first lower bound in (31) and the upper bound in (30) is asymptotically negligible, which is immediate under the conditions of Theorem 8. Parts (b) and (c) follow exactly as in Theorem 8. Theorems 10 and 11 extend easily when $\underset{\sim}{A}^{on}$ and $\underset{\sim}{N}^{on}$ replace $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{N}^{cn}$ since the limiting system is the first upper bound for the FPM method. Finally, part (d) follows exactly as in Theorem 9. □

IX. A BOTTLENECK NODE WITH A LARGE POPULATION

9.1 A different approximation procedure

In this section we observe that the methods and results of Sections VI through VIII also apply, after appropriate modification, to closed networks with a bottleneck non-IS node. We first consider the case of one job class. For large populations, all servers at the bottleneck node will usually be busy, so that we can approximately analyze the original closed model by using the bottleneck node to decouple the network just as we used the IS node in Section VI. We remove the bottleneck node and replace its departure process by an external arrival process. We then solve, exactly or approximately, the resulting open network. If there are s servers at the bottleneck node, then the external arrival process would be the superposition of s i.i.d. renewal processes each having the bottleneck service-time distribution as the renewal-interval distribution. The routing of the external arrivals is just the original routing from the bottleneck node. When the service-time distribution at the bottleneck node is exponential, the approximating external arrival process is thus Poisson.[51] Otherwise, we would approximately characterize the external superposition arrival process as in Ref. 7 and apply the algorithm there to approximately analyze the resulting non-Markov open network.

In this setting the bottleneck node is easy to identify. As in Section III, we begin by replacing one internal arrival process by the external arrival process with a rate sufficiently small to ensure stability. Let λ_i be the net arrival rate to node i obtained from solving the traffic rate equations with the given external arrival rate, say λ_o. Then calculate the traffic intensity at node i as $\rho_i = \lambda_i / s_i \mu_i$ given that node i is a FCFS node with s_i servers, each working at rate μ_i. The node with the highest traffic intensity is the bottleneck node; call it node $q + 1$. We assume that there are no ties. The capacity of the network is thus $s_{q+1}\mu_{q+1}$. We can achieve any throughput less than $s_{q+1}\mu_{q+1}$ in the open model. The traffic intensity becomes 1 at node $q + 1$ at the capacity, which makes the system unstable. It of course is well known that $s_{q+1}\mu_{q+1}$ is an upper bound on the throughput even in non-Markov networks (see Ref. 11 and references there).

The proposed approximation procedure for the closed model with a large population is to solve the traffic rate equations for the associated open network and find the bottleneck node, which we denote by node $q + 1$. Then solve the open model obtained by deleting node $q + 1$ from the closed network and inserting an external arrival process with rate $s_{q+1}\mu_{q+1}$. However, unlike Sections III and VI, we do not use the FPM method for the full $(q + 1)$-node network; we do not require a consistency condition such as (28). We simply let the approximate number of jobs at node $q + 1$ in the original closed network be

$$EN^c_{q+1}(\infty) \approx K - \sum_{i=1}^{q} EN^o_i(s_{q+1}\mu_{q+1}). \tag{49}$$

If there are quite a few nodes but not a large population, we will use the original FPM method, but as the population grows with the number of nodes fixed, the effect of the bottleneck node becomes more pronounced.

9.2 Limit theorems

The approximation procedure just described is evidently quite well known. Supporting limit theorems are discussed by Whittle[35] and Brown and Pollett.[36] The methods and results of Section VIII provide a convenient way to prove that the approximation procedure is asymptotically correct for the q-node subnetwork excluding the bottleneck node as the population grows. Since the results and methods are similar to those in Section VIII, we only give a brief account. The analog of Theorem 8 is for one customer class. We let $K_1^n \to \infty$ as before, but now we fix μ_1^n, the individual service rate at node $q + 1$. When the service-time distribution at the bottleneck node is exponential, we can use the obvious modification of the proof of Theorem 8(a). With general service-time distributions, it is easy to show that the probability that all servers are busy at the bottleneck node throughout any interval $[0, t]$ converges to 1 as $n \to \infty$. The rest of Section VIII applies in a straightforward manner, with essentially the same remarks about generalizations.

9.3 *Multiple job classes*

We now consider multiple job classes with a special bottleneck node. We assume that there is a single-server processor-sharing bottleneck node with fixed total service rate μ whenever any customers are present. Let the service requirements of class j at the bottleneck node be exponentially distributed with mean μ_j^{-1}. Let the population of class j in the network be K_j.

Again the approximation is obtained by replacing the bottleneck node by an external Poisson process with rate μ. Each of these external arrivals is from class j with probability

$$\gamma_j = K_j\mu_j/(K_1\mu_1 + \cdots + K_p\mu_p). \tag{50}$$

Consequently, as in Section VI, there is a pq-dimensional vector of independent Poisson processes with the intensity class j going to node i being $\lambda_{ji} = \mu\gamma_j p_{ji}$. The limit theorems in Section VIII also apply here. As before, Theorem 11(a) does not hold for multiple job classes.

9.4 *Another stochastic comparison*

We now make a stochastic comparison between the closed model and the open model resulting from the bottleneck approximation. We consider the case of one job class. We compare the q-dimensional equilibrium distribution of the subnetwork of the closed model without the bottleneck node to the q-dimensional equilibrium distribution in the q-node open model with external arrival rate $\mu_{q+1}s_{q+1}$. We show that the equilibrium distribution based on the bottleneck approximation is larger in a very strong sense, namely, in the MLR ordering used in Section 4.3.

Let $(N_1^c, \cdots, N_q^c)$ be the equilibrium random vector in the closed model with population K without the bottleneck node, defined in terms of an associated $(q + 1)$-dimensional open-model equilibrium random vector $(N_1^o, \cdots, N_{q+1}^o)$ by

$$P(N_1^c = k_1, \cdots, N_q^c = k_q) = \frac{P(N_1^o = k_1) \cdots P(N_q^o = k_q)P(N_{q+1}^o = K - \sum_{j=0}^{q} k_j)}{P(N^o = K)} \tag{51}$$

for $(k_1, \cdots, k_q)$ such that $k_1 + \cdots + k_q \le K$.

Let $(N_1^b, \cdots, N_q^b)$ be the open-model equilibrium random vector with utilization at node i of u_i^o/u_{q+1}^o, where u_i^o is the utilization of node i in $(N_1^o, \cdots, N_{q+1}^o)$ in (51). We assume that $u_i^o < u_{q+1}^o$ for all i. This is tantamount to having an external Poisson arrival process with rate $u_{q+1}s_{q+1}$ in the q-node open network.

Theorem 13: $(N_1^c, \cdots, N_q^c) \le_{\mathrm{lr}} (N_1^b, \cdots, N_q^b)$.

Proof: It is immediate that the distribution of $(N_1^b, \cdots, N_q^b)$ is MTP_2 because the marginals are independent (see Proposition 3.5 of Ref. 50). Consequently, by Theorem 3 of Ref. 48, it suffices to show that $p_1(y)p_2(x) \leq p_1(x)p_2(y)$ for all $x \leq y$, where p_1 and p_2 are the associated probability mass functions. Moreover, it suffices to consider y differing from x by 1 in only one place, e.g., $x = (k_1, \cdots, k_q)$ and $y = (k_1 + 1, k_2, \cdots, k_q)$. We verify this as follows:

$$\frac{P(N_1^c = k_1 + 1, \cdots, N_2^c = k_2)\; P(N_1^b = k_1, \cdots, N_2^b = k_2)}{P(N_1^c = k_1, \cdots, N_2^c = k_2)\; P(N_1^b = k_1 + 1, \cdots, N_2^b = k_2)}$$

$$= \frac{P(N_1^o = k_1 + 1) P\left(N_{q+1}^o = K - \sum_{j=1}^{q} k_j - 1\right) P(N_1^b = k_1)}{P(N_1^o = k_1) P\left(N_{q+1}^o = K - \sum_{j=1}^{q} k_j\right) P(N_1^b = k_1 + 1)} \leq 1$$

because, for all j,

$$P(N_1^b = j + 1)/P(N_1^b = j) = u_{q+1}^o P(N_1^o = j + 1)/P(N_1^o = j)$$

and

$$P(N_{q+1}^o = j + 1)/P(N_{q+1}^o = j) \geq u_{q+1}^o. \tag{52}$$

To verify (52), recall that N_{q+1}^o has the equilibrium distribution of a birth-and-death process, so that

$$\hat{\lambda}_j P(N_{q+1}^o = j) = \hat{\mu}_{j+1} P(N_{q+1}^o = j + 1),$$

where $\hat{\lambda}_j$ is the arrival rate when $N_{q+1}^o = j$, which is independent of j, and $\hat{\mu}_{j+1}$ is the service rate when $N_{q+1}^o = j + 1$. When there is one server, $u_{q+1}^o = \hat{\lambda}_j/\hat{\mu}_{j+1}$ for all j, but in general, $\hat{\mu}_{j+1} = \mu_{q+1} \min\{j + 1, s\}$, so that $u_{q+1}^o \leq \hat{\lambda}_j/\hat{\mu}_{j+1}$. □

Remarks: 1. Theorem 13 has corollaries like those for Theorem 6. For example,

$$P(N_i^c \geq k_i \mid a_j \leq N_j^c \leq b_j) \leq P(N_i^b \geq k \mid a_j \leq N_j^b \leq b_j) \tag{53}$$

for all i, j, k_i, a_j, and b_j. Inequality (53) is interesting both when $i = j$ and $i \neq j$. Of course, when $i \neq j$, the right-hand side of (53) reduces to $P(N_i^b \geq k_i)$.

2. As in Section VIII, we can obtain results for the equilibrium distribution associated with other queue disciplines by invoking insensitivity properties.[3-6,19,20]

3. Algorithms for identifying bottleneck nodes and treating them are described by Schweitzer[54] and Goodman and Massey.[39] Stochastic bounds for open networks of single-server nodes are contained in

Massey.[38] These bounds apply to closed networks too by combining them with the comparison results in this paper.

4. As the population increases, the closed network can be said to be in heavy traffic. However, only the bottleneck node accumulates jobs in the limit. The number of jobs at the nonbottleneck nodes is asymptotically negligible compared to the number at the bottleneck node. In fact, by the analog of Theorem 9, the number of jobs at all nonbottleneck nodes, unnormalized, converges to a proper limit, as the population grows. Instead of the complicated multidimensional diffusion process approximations for networks of queues described in Reiman,[55] we have significant accumulation of customers only at the bottleneck node alone. The situation here is an example of the diffusion approximations with state space collapse discussed by Reiman.[56] However, because we are considering a closed network, the number of customers at the bottleneck node is best described by $K - \sum_{j=1}^{q} N_j^b$. Indeed, as a trivial corollary to the analog of Theorem 9, we have

$$(N_{q+1}^{cn} - K^n) \Rightarrow \sum_{j=1}^{q} N_j^b \tag{54}$$

as $n \longrightarrow \infty$. Unless there are ties for the maximum traffic intensity, only one node will be a bottleneck node for both closed and open networks. Moreover, because of the geometric tails of the queue length equilibrium distributions in Markov networks, slight differences in traffic intensities will rapidly lead to large differences in the queue lengths as the population grows. Consequently, the case of a single bottleneck node treated here seems most relevant for applications.

X. APPROXIMATIONS FOR NON-MARKOV CLOSED NETWORKS

10.1 Several possible approximation procedures

Suppose, as in Ref. 7, that the Markov property is lost because we are considering FCFS nodes with nonexponential service-time distributions. There are two natural procedures for calculating approximate congestion measures for such non-Markov closed networks based on previously developed approximations for non-Markov open networks. Just as we can use Markov open models to analyze Markov closed models, we can use the approximate solution for an associated non-Markov open model to generate an approximate solution for the given non-Markov closed model.

The first procedure for non-Markov closed models starts with the approximate equilibrium distribution of the number of customers at each node in the associated open model, as described in Section III. Then the corresponding equilibrium distribution for the closed model can be obtained by conditioning as in (9). For the open model, the standard approximation procedure is to use a product-form solution (an equilibrium distribution with independent marginal distributions). This is the procedure first suggested by Reiser and Kobayashi.[57] The Extended-Product-Form (EPF) method of Shum and Buzen[58,59] and the Generalized-Product-Form (GPF) method of Tripathi[60] are also variants of this approach. A complete approximation thus is determined by specifying the equilibrium distribution of the number of customers at each node in the open model. For example, with QNA[7] this can be done by fitting a discrete distribution to the quantities $P(N_i^o = 0)$, $E(N_i^o)$, and $\mathrm{Var}(N_i^o)$, which are currently provided in the model solution. In fact, in Ref. 7 an approximation for the waiting-time distribution at each node is obtained in this way. For single-server nodes, it is natural to use mixtures and convolutions of geometric distributions for the conditional distribution of the number of customers at each node, given that the server is busy. Such an approximation procedure based on QNA is currently being investigated.

There are some difficulties with this first procedure, however. We must do the same extensive calculation to find the normalization constant G as we do with the Markov closed model, so that we obtain no reduction in computation working with approximations. We can of course use many of the same algorithms now being used for Markov closed networks.[6]

The second procedure is to use the open model directly, as with the FPM method. We believe that this method can be expected to work about as well as it does for closed Markov models. Now, in the setting of Ref. 7 we also have variability parameters. In particular, we must specify a variability parameter as well as an arrival rate for the special new external arrival process.

There are three different situations. First, with a decoupling IS node containing most of the customers (under the conditions of Section VI), it is natural to use the FPM method and approximate the external arrival process by a Poisson process, so that there is no problem selecting the variability parameter; set it equal to 1. However, now it is important that the external Poisson arrival process replace the departure process from the special IS node. This process will be approximately Poisson, even with nonexponential service-time distributions.

Second, with a bottleneck node having s servers as in Section IX, it is natural to regard the arrival process as the superposition of s independent renewal processes each with the bottleneck service-time distribution as the renewal-interval distribution. When $s = 1$, the procedure is clear: use the squared coefficient of variation of the bottleneck service-time distribution. When $s > 1$, we can use approximations for superposition processes as in Section 4.3 of Ref. 7. As described in Section IX, we would not use the FPM method, but instead the open model with the bottleneck node removed.

The third situation is where the FPM method is appropriate but the variability parameter needs to be determined. In Section 10.2 we discuss this case in detail.

There are of course many other procedures for approximately analyzing non-Markov closed networks with nonexponential FCFS nodes,[6,11,22,61-62] but we do not discuss them here.

10.2 The FPM method for non-Markov models

A simple procedure for the FPM method more generally, in the case of a single job class, is to first specify an external arrival rate λ_0 and then, for that specified arrival rate, solve a system of linear equations to obtain the variability parameter $c_0^2(\lambda_0)$ that makes the variability parameter of the departure process from the network equal to the variability parameter of the external arrival process. (The reason for doing this, of course, is that in the closed network these two processes are actually the same process.) We then solve the open model for a range of possible external arrival rates, associating $c_0^2(\lambda_0)$ with λ_0 each time. As before, the throughput when the population is K is the value of λ_0 such that $EN^o = K$.

We now describe in detail a modification of the QNA algorithm in Ref. 7 that has been developed to approximately analyze a closed non-Markov network of queues with one job class by the FPM method. The initial model is just as in Ref. 7 but without external arrival processes. In particular, the nodes have the FCFS discipline, several servers, and general service-time distributions. We assume that the reader is familiar with Ref. 7, and we use the same notation here.

The model input is a minor modification of the standard input in Section 2.1 of Ref. 7; we just omit the data for the external arrival processes. For each network we specify the following:

n = number of nodes in the network
m_j = number of servers at node j
τ_j = mean service time at node j
c_{sj}^2 = squared coefficient of variation of the service-time distribution at node j
q_{ij} = proportion of those customers completing service at node i that go next to node j.

To apply the FPM method, we introduce an external arrival process to one node, which we stipulate is node 1. To see how the expected network population depends on the external arrival rate, we specify a set of external arrival rates, which are understood to apply to node 1.

The set is specified by the following numbers:

L = lower bound for external arrival rate to node 1
U = upper bound for external arrival rate to node 1
C = number of different arrival rates.

Given the triple (L, U, C), the network will be analyzed C times with the following external arrival rates to node 1:

$$\lambda_{01} = L + k(U - L)/(C - 1) \tag{55}$$

for $k = 0, 1, \cdots, C - 1$. The external arrival rates to all other nodes are zero. To obtain the open model for the FPM method in each case, we insert an external arrival process to node 1 with one of the rates specified in (55) and we eliminate all internal arrivals to node 1. This is done with the algorithm by setting $q_{i1} = 0$ for all i.

We begin by solving the traffic-rate equations, given the external arrival rate λ_{01}, exactly as in Section 4.1 of Ref. 7. This provides the traffic intensities at the nodes, needed for the traffic variability equations.

Next we solve the traffic variability equations. The algorithm also determines the variability parameter c_{01}^2 for the external arrival process to node 1. As indicated above, the idea is to have the variability parameter of the external arrival process agree with the variability parameter of the total departure process from the network (which would have been the arrival process to node 1 in the closed model). The equations in (24) of Ref. 7 are valid for $j = 2, \cdots, n$; i.e., we have

$$c_{aj}^2 = a_j + \sum_{i=1}^{n} c_{ai}^2 b_{ij}, \qquad 2 \le j \le n, \tag{56}$$

with a_j and b_{ij} in (25) and (26) of Ref. 7. Since $q_{i1} = 0$ for all i, $c_{a1}^2 = c_{01}^2$. The variability parameters are solved by replacing the first equation in (24) of Ref. 7 with

$$c_{a1}^2 = a_1^* + \sum_{i=1}^{n} c_{ai}^2 b_{i1}^*, \tag{57}$$

where

$$a_1^* = 1 + w_1^* \left\{ - 1 + \sum_{i=1}^{n} (d_i/d) \left[\sum_{j=2}^{n} q_{ij} + \left(1 - \sum_{j=2}^{n} q_{ij}\right) \rho_i^2 \left(1 + \frac{c_{si}^2 - 1}{\underset{i}{=}}\right)\right]\right\}, \tag{58}$$

$$b_{i1}^* = w_1^*(d_i/d) \left(1 - \sum_{j=2}^{n} q_{ij}\right) (1 - \rho_i^2), \tag{59}$$

d_i is the departure rate from the network at node i, $d = d_1 + \cdots + d_n$ as in (23) of Ref. 7, and w_1^* is the superposition weighting function in (29) of Ref. 7 with p_{i1} in (30) there replaced by d_i/d.

We derive (57) through (59) as follows. First, the departure process from the whole network is the superposition of the departure processes (leaving the network) from the separate nodes. Hence, by Section 4.3 of Ref. 7,

$$c_{a1}^2 = w_1^* \left(\sum_{i=1}^{n} (d_i/d) c_{di}^{*2} \right) + 1 - w_1^* , \tag{60}$$

where c_{di}^{*2} is the variability parameter for the departure process from the network at node i; i.e.,

$$c_{di}^{*2} = \left(1 - \sum_{j=2}^{n} q_{ij} \right) c_{di}^2 + \sum_{j=2}^{n} q_{ij} \tag{61}$$

and

$$c_{di}^2 = 1 + (1 - \rho_i^2)(c_{ai}^2 - 1) + \frac{\rho_i^2}{\sqrt{m_i}} (c_{si}^2 - 1), \tag{62}$$

using first the splitting formula (36) and then the departure formula (39) from Ref. 7.

The rest of the modified QNA algorithm is just as in Ref. 7. We next calculate the congestion measures at the nodes using the traffic rate and variability parameters already determined. By running the algorithm a few times with various (L, U, C) triples, the user can easily select a set of external arrival rates to node 1 via (55) to yield a desired range of expected network populations in the open model. The algorithm also can automatically find the external arrival rate yielding a specified expected equilibrium network population.

We can also use the finite-waiting-room refinement introduced in Section 1.3 to calculate the congestion measures at the nodes in the open model. For single-server nodes, we use the modifications in (2) and (3), even if the nodes do not correspond to M/M/1 models.

Remarks. 1. Our procedure above replaces an internal arrival process to node 1 by an external arrival process. Instead, we could have replaced the internal departure process from node 1 by an external arrival process. It would then have been immediately split according to the routing probabilities q_{1i}. The original departures from node 1 would then be removed.

2. As mentioned in Section I, it may be desirable to artificially deflate the variability parameters of the arrival processes. It is natural to do this after the traffic variability equations have been solved as described above.

3. The procedure can easily be extended to multiple job classes in various ways. For example, we can let the variability parameters of the external arrival processes for each individual class be unspecified. We then can apply the procedure in (56) through (59) in this section to specify the variability parameter for the overall external arrival process in the aggregated single-class network obtained from Section 2.3 of Ref. 7. The only remaining complication is that instead of the external arrival rates λ_{01} determined by the single triple (L, U, C), we now have a vector of external arrival rates determined by such a triple for each job class. Automatic search obviously becomes desirable in this setting.

4. The approximate solution using the FPM method can be fruitfully combined with the exact solution of the corresponding Markov model to obtain improved approximations for the closed non-Markov model. For example, we can solve the closed Markov model to obtain u_i^{cM} as the utilization of node i when the service-time distributions are all exponential. We can also apply the FPM method twice, once with general service-time distributions and once with exponential service-time distributions, to obtain corresponding utilizations u_i^{oG} and u_i^{oM}. We can then approximate u_i^{cG}, the utilization at node i in the closed network with nonexponential service-time distributions, by

$$u_i^{cG} = u_i^{cM} u_i^{oG} / u_i^{oM}. \tag{63}$$

Since (65) can lead to inconsistencies such as $u_i^{cG} > 1$, it is natural to use

$$(1 - u_i^{cG}) = (1 - u_i^{cM})(1 - u_i^{oG})/(1 - u_i^{oM}) \tag{64}$$

for the node i with the largest utilization. We then calculate u_i^{cG} for the other nodes using (15), which is justified for non-Markov models as well as Markov models, e.g., by Little's formula, (4.2.3) of Ref. 19. At least, the ratios u_i^{oG}/u_i^{oM} and $(1 - u_i^{oG})/(1 - u_i^{oM})$ can give a rough idea of how much the nonexponential service-time distributions matter.

XI. THROUGHPUT BOUNDS IN NON-MARKOV CLOSED NETWORKS

It is sometimes claimed that closed Markov models are suitable even when the service-time distributions are not exponential. In particular, it is sometimes claimed that the utilizations and throughputs, at least, do not depend critically on aspects of the service-time distributions beyond their means. Of course, this is trivially true for certain special service disciplines such as processor sharing, for which there are insensitivity results,[3-6,19,20] but with FCFS nodes the service-time distribution matters. Even with FCFS nodes, there is significant justification for this view if the service-time distributions do not depart too drastically from the exponential distribution. However, in general, throughputs obtained with the Markov model can be very bad approximations, as we show in this section. For example, for a cyclic network with n single-server nodes, equal mean service times, and K customers, we show that the set of possible utilizations for each server is the interval $(n^{-1}, 1]$ for all $K \geq n$, whereas the utilization is $K/(n + K - 1)$ in the Markov model by (14). For large n and K and arbitrarily unfavorable service-time distributions with given means, the Markov approximation can be arbitrarily bad. The true value can be arbitrarily close to 0, while the Markov approximation is arbitrarily close to 1.

We consider the same non-Markov closed model as in Section X, containing FCFS nodes with general service-time distributions, but we restrict attention to single-server nodes. All the service times are assumed to be mutually independent and the service times at any given node are identically distributed. There is a single job class with K jobs. A job completing service at node i is routed immediately to node j with probability q_{ij}, independent of the history. The matrix $Q \equiv (q_{ij})$ is a Markov chain transition matrix, which we assume is irreducible. Consequently, there is a unique equilibrium distribution associated with Q, defined by

$$\lambda_j = \sum_{j=1}^{n} \lambda_i q_{ij}, \qquad 1 \leq j \leq n, \tag{65}$$

with $\lambda_1 + \cdots + \lambda_n = 1$. By the law of large numbers, λ_j is the long-run fraction of transitions that each customer spends in node j. The system of equations (65) is also the basic traffic-rate equations for the network of queues. The throughput or equilibrium flow rate through node i, say θ_i^c, is proportional to λ_i, i.e., $\theta_i^c = \gamma\lambda_i$ for some constant γ.

Let τ_i be the mean service time (which we assume is finite and strictly positive) and let u_i^c be the utilization (long run fraction of time that the server is busy) at node i. By Little's law, (4.2.3) of Ref. 19, or by the law of large numbers again, we know the ratio of the utilizations, i.e.,

$$u_i^c/u_j^c = \lambda_i\tau_i/\lambda_j\tau_j \tag{66}$$

for any i and j just as for the Markov model in (15).

We exhibit the infimum of the server utilizations possible for service-time distributions with the given means. As will soon be clear, the infimum is approached by quite unusual service-time distributions, so that we do not rule out the possibility that the Markov model can provide good throughput approximations for typical nonexponential service-time distributions. The idea for minimizing utilizations is really quite simple. For our model, at any time at least one server must be busy. Hence, the sum of the server utilizations must exceed unity:

$$\sum_{i=1}^{n} u_i^c \geq 1. \tag{67}$$

A lower bound on the server utilizations is the case in which there is no concurrency, i.e., no two servers are ever busy at the same time. This lower bound is obviously valid in much greater generality. It is also attained in the case $K = 1$. It is somewhat remarkable that this lower bound is actually approached for any K with the general independent service times allowed here. This observation was apparently first made by Arthurs and Stuck.[11]

It is, in fact, not difficult to attain this lower bound asymptotically by considering special sequences of service-time distributions with a common mean that get successively more variable. In particular, for $m \geq 1$, let X_m be a random variable distributed as

$$P(X_m = m) = 1 - P(X_m = 0) = m^{-1}. \tag{68}$$

[Alternatively, $(X_m \mid X_m > 0)$ could have some other distribution with mean m, such as exponential.] Then let the service time at node i be distributed as $\tau_i X_m$.

Theorem 14: (*a*) *The infimum of the possible utilizations of server i for this closed network model over all service-times distributions with specified means is*

$$\inf u_i = \lambda_i \tau_i \bigg/ \sum_{j=1}^{n} \lambda_j \tau_j.$$

(*b*) *If the service times are not all deterministic, then the infimum is not attained for* $K > 1$, *but is approached asymptotically for all nodes simultaneously as* $m \rightarrow \infty$ *using the service-time distributions of* $\tau_i X_m$ *described above.*

Proof: (a) We informally sketch the proof. For very large m, occasionally (among all service times generated) a long service time occurs at some node. With high probability, thereafter all the other customers instantaneously fly around the network until they arrive at this node, where they all wait together in queue. (There is only one server at each node.) The only other possibility, which occurs with asymptotically negligible probability as $m \rightarrow \infty$, is that one of the other customers encounters another nonzero service time before all of the customers are gathered together at the same node. This event yielding the concurrency has asymptotically negligible probability because the distribution of the number of transitions for any job to go from any node i to any other node j does not change with m. Hence, for each of the $K - 1$ customers there is a fixed random number of trials (with finite mean and variance) to generate a new nonzero service time, but the probability of doing so on each trial is m^{-1}. Hence, the proportion of time during which two or more servers are simultaneously busy converges to zero as $m \longrightarrow \infty$.

(b) On the other hand, it is trivial that concurrency cannot be ruled out altogether when $\mathrm{K} > 1$ and there is some randomness. For any model with strictly positive expected service times, at least one nondeterministic distribution, and an irreducible routing matrix, concurrency occurs with positive probability. The limiting case above is not legitimate because X_m converges in distribution to the random variable X with $\mathrm{P}(X = 0) = 1$. □

Remarks: 1. It is not necessary to have all service-time distributions be of the special form (68). It suffices to have all but one. The other one can be arbitrary. At this designated node there will be a succession of ordinary service times after which the customer usually returns immediately to the end of the queue. When a customer does get a nonzero service time elsewhere, the others get there relatively quickly with high probability when they complete service. It is again not difficult to show that the proportion of time that there is concurrency is asymptotically negligible.

2. A next step would be to obtain tighter bounds under extra conditions as, for example, in Refs. 63 through 65 and references there. However, as noted above, the special service-time distributions can be H_2 (hyperexponential: a mixture of two exponential distributions), so that does not help. It would obviously help to fix the variance though. We conjecture that the X_m distributions would yield the minimum then.

3. The infimum decreases rapidly as the number of nodes increases. The possible server utilizations are not so great with two nodes. For example, suppose that $q_{12} = q_{21} = 1$, which makes the network cyclic. Then $\lambda_1 = \lambda_2 = 1/2$ and $\inf u_1 = \tau_1/(\tau_1 + \tau_2)$. In this case the maximum is 1, which is attained with the deterministic service-time distributions for $K \geq 2$. As before, the infimum corresponds to the case $\mathrm{K} = 1$. In the case of balanced loads, the utilization of each server must lie between one-half and one. It is useful to recall that this two-server cyclic network is equivalent to an M/G/1/K−1 queueing model when one of the two service-time distributions is exponential (see Ref. 66, p. 33 of Ref. 67, and Ref. 2). The M/G/1/K−1 model means that we have an external Poisson arrival process, a single queue with one server and an additional waiting room of size $K - 1$. Since we approach the infimum if all but one service-time distribution is of the special kind, the infimum is also valid for M/G/1/K−1 queueing model.

XII. MORE NUMERICAL COMPARISONS

We have indicated that the approximation methods should perform better for larger closed networks, but it is nevertheless useful to compare their performance for smaller ones. It is certainly important to realize the limitations of these procedures. They often perform poorly for small networks.

It is particularly convenient to consider two-node closed networks because these networks are equivalent to special single-node models that have been studied extensively and for which there are tables of exact values.

12.1 Two single-server nodes

As we noted in Section XI, the closed model with K jobs (all of one class) and two single-server nodes, one of which has an exponential service-time distribution, is equivalent to an M/G/1 model with a finite waiting room of size $K-1$. Similarly, the two-node closed model with K jobs and one IS node having exponential service-time distributions is equivalent to finite-source M/G/1 model with K sources. Tables for M/G/1 models with finite waiting room and finite sources are contained in Ref. 67, for example.

Table V here displays the exact values and various approximations for the throughput and the expected equilibrium number of jobs present in an M/G/1 model having a waiting room of size 10. This

Table V—A comparison of exact throughput and mean number in system in the M/G/1/10 model having a finite waiting room of size 10 with approximations based on the bottleneck method and the FPM method for $G = M, D$, and H_2

Arrival	Exact Values		Bottleneck Method		FPM Method		Predicted θ
Rate	θ^c	EN_1	θ^o	EN_1	θ^o	EN_1	With (64)
			(a) Exponential Service Times (M)				
0.50	0.4999	1.00	0.500	1.00	0.455	0.84	
0.75	0.7418	2.61	0.750	3.00	0.674	2.07	
1.00	0.9167	5.50	1.000	∞	0.846	5.50	
1.40	0.9928	8.72	1.000	8.50	0.902	9.19	
2.00	0.9998	10.00	1.000	10.00	0.910	10.16	
			(b) Deterministic Service Times (D)				
0.50	0.5000	0.75	0.500	0.75	0.461	0.65	0.500
0.75	0.7494	1.85	0.750	1.88	0.695	1.43	0.744
1.00	0.9538	5.57	1.000	∞	0.912	4.93	0.952
1.40	0.9998	9.61	1.000	9.39 (9.69)	0.973	9.68	0.998
2.00	1.0000	10.37	1.000	10.25 (10.61)	0.987	10.38	1.000
			(c) Hyperexponential Service Times With $c_s^2 = 2.25$ and Balanced Means				
0.50	0.4983	1.27	0.500	1.31	0.449	1.05	0.500
0.75	0.7248	3.01	0.750	4.41	0.651	2.71	0.739
1.00	0.8829	5.34	1.000	∞	0.787	6.00	0.885
1.40	0.9791	8.15	1.000	7.38	0.833	9.06	0.987
2.00	0.9985	9.75	1.000	9.69	0.840	10.10	1.000

corresponds to a closed network with a population of 11 and two single-server nodes. Three service-time distributions are considered: exponential, deterministic, and hyperexponential (mixture of two exponentials). The hyperexponential distribution has squared coefficient of variation $c_s^2 = 2.25$ and balanced means (see p. 8 of Ref. 67 or Section 3 of Ref. 68). The service time is set equal to 1 and five arrival rates are considered: 0.50, 0.75, 1.00, 1.40, and 2.00. The arrival rate of 2.00 (1.40) corresponds to a traffic intensity of 0.50 (0.71) when the nodes are switched. In each case, the FPM and bottleneck approximations are displayed in addition to the exact values. For the nonexponential service times, the refinement in (64) is also displayed.

The exact values come from Tables 5.1.6, 5.2.6, and 5.4.12 in Section II.5 of Ref. 67, using the FIFO or FCFS discipline. The approximations are obtained using the GI/G/1 formulas (47) and (44) in Ref. 5 with g in (45) of Ref. 5 set equal to 1. The approximate values using the Krämer-and-Langenbach-Belz correction term in (45) of Ref. 7 are given in parentheses to the right of the other values in Table V (b).

There are several important conclusions to draw from Table V. First, as we should expect from Section III, the FPM method performs poorly, much worse than the bottleneck method. However, it is important to remember that this small network tends to be a worst case for the FPM method. It is also significant that the refinement suggested in (64) produces quite accurate results. With this job population (waiting room size), the bottleneck method seems to work reasonably well as long as the utilization of the bottleneck node is no more than about 0.75.

It is also useful, to consider the Finite-Writing-Room (FWR) refinement introduced in Section 1.3 in the context of Table Va. When combined with the bottleneck procedure, the FWR refinement obviously makes the approximation exact when $n = 2$. The FPM method with the FWR refinement is also exact in the special case of equal service rates. When $n = 2$, we must have $E\bar{N}_i^o = EN_i^o = K/2$ by the FPM method. With the FWR method, this implies that $\rho_i = 1$ and $P(\bar{N}_i^o = K) = 1/(K + 1)$, so that $\bar{n}_i^o = u_i^c = K/(K + 1) > K/(K + 2) = u_i^o$, using (11) and (14). However, more generally the FPM/FWR method does not perform well, at least for the mean numbers at each node, when $n = 2$. For example, when the arrival rate is 0.50, the FPM/FWR approximations are $\theta^o = 0.69$ and $EN_1 = 2.17$. However, applying (3), we obtain $\overline{\theta}^* = 0.495$ as a lower bound on the throughput.

From Table Vb we see that the bottleneck method continues to perform well for the deterministic service-time distribution. In fact, the bottleneck approximation is clearly much better than using the exact M/M/1 values in Table Va as an approximation, which is often what is done in practice.

However, from Table Vc we see that the quality of the bottleneck approximation deteriorates when we consider the more variable hyperexponential service-time distribution. Of course, the throughputs are always close and the mean queue lengths are good when the traffic intensity at the bottleneck node is 0.5, but when the traffic intensity is 0.70 or 0.75, the open-network view exaggerates the impact of the greater variability. The fixed population in the closed model tends to damp the effect.

We can also see what happens as we change the service-time variability in Table V. From the exact values, we see that the throughput decreases in every case, but the throughputs are never near the lower bound in Theorem 14. We also see that the expected number of jobs at that node decreases when the arrival rate is greater than 1.00. This phenomenon was observed by Bondi[61] and is further discussed in Bondi and Whitt.[62] Briefly, the explanation is that under moderate to heavy loads increased variability in the service-time distribution often has a greater impact on other nodes via their arrival processes than on the congestion at the given node. It is significant that this qualitative behavior is captured by both the bottleneck and FPM methods using QNA. However, the FPM method fails to capture this bottleneck phenomenon in other cases. As noted in Refs. 61 and 62, the bottleneck phenomenon is useful to test procedures for approximately solving non-Markov closed networks.

Approximately characterizing the variability of arrival processes in a tightly coupled closed network such as the two-node model being discussed is difficult because of the constraint on the total population. If the utilization of a server is high, then the interdeparture times are distributed approximately the same as the service times, but the population constraint tends to induce negative correlations among the interdeparture times: several long (short) times are more likely to be followed by a short (long) one. Hence, the effective variability of an arrival process, e.g., as described by the asymptotic method in Ref. 68, is likely to be considerably less in a closed network than in an open one. This is the reason for developing heuristic procedures to reduce the variability parameters of the arrival processes in the approximation method.

12.2 One single-server node and one infinite-server node

Table VI displays exact and approximate results for a two-node network containing an IS node with exponential service-time distributions. In this case, the service-time distribution at the single-server node is always exponential. It is easy to apply the FPM approximation to other cases, but we had no convenient tables. The exact values from Table 2.10.7 of Ref. 67 are obtained by specifying the population

Table VI—A comparison of exact throughput and mean number in system in the M/G/1 queue having a finite source of size 10 with approximations based on the FPM method

	Exact Values	FPM	First Upper Bound
(a) Total Population 10 and Utilization 0.50			
Arrival rate per idle source	0.0547	0.0547 given	0.0547 given
Throughput or utilization	0.500 given	0.494	0.547
Expected number in system, EN_1	0.86	0.98	1.21
(b) Total Population 10 and Utilization 0.75			
Arrival rate per idle source	0.0927	0.0927 given	0.0927 given
Throughput or utilization	0.750 given	0.705	0.927
Expected number in system, EN_1	1.91	2.39	12.70

(number of sources) and the throughput. Paralleling Table V, we consider two cases: a population of 10 and throughputs of 0.50 and 0.75. As in Table V the service time is set equal to 1. The population and throughput determines the arrival rate per idle source (individual service rate at the IS node). This is the starting point for the FPM approximation, which is obtained from (28) or (38). The conditions are clearly much less favorable in Table VI than in Tables I and IV; the ratio of service rates (IS/other) was much less before. Nevertheless, the FPM method works quite well, at least in the case of utilization 0.50. From Tables V and VI, we see that the FPM method does indeed perform better with the IS node. Related numerical comparisons are contained in Ref. 69. The overall performance here based on the FPM method is somewhat better than the performance in Ref. 69, which is based on matching the server utilization. The performance in Tables I and IV is much better than we might at first expect from Ref. 69, but recall that in Tables I and IV, as the total population decreases the server utilization decreases because the service rate at the IS node is held fixed. Nevertheless, the tables in Ref. 69 help assess how well the FPM method will perform for a small network with an IS node.

XIII. CONCLUSIONS

In this paper we identified and investigated three situations in which open queueing network models should provide good approximations for more difficult closed queueing network models:

1. When the closed network has many nodes (Sections II through V, X),

2. When the closed network contains a "decoupling" infinite-server (IS) node with a relatively low service rate (see Sections VI through VIII),

3. When the closed network contains a non-IS bottleneck node under a fairly heavy load (Section IX).

The suggested approximation procedures in these situations are not the same, however. In Case 3 we remove the bottleneck server and replace its departure process by an external arrival process, which is determined solely by the number of servers and the service-time distribution at the bottleneck queue. The arrival rate is the maximum possible service rate from the bottleneck node; we do not use the FPM method. In contrast, in Case 1 no nodes are removed from the closed network. As described in Section X, an entry node is selected and the external arrival process there depends on the entire network in a rather complicated way. The arrival rate is determined by the FPM method. The variability parameter of the arrival process, using QNA,[7] is chosen so that the variability parameter of the external arrival process agrees with the variability parameter of the departure process from the network.

It is interesting that the suggested procedure for Case 2 can be regarded as a variation of either the procedure for Case 1 or the procedure for Case 3. On the one hand, the procedure for Case 1 can be applied without change to Case 2. As described in Section VI, the suggested procedure coincides with the FPM method. However, in Case 2 we know that the departure process from the IS node is approximately a Poisson process. Hence, it is natural to implement the FPM method for Case 2 by replacing the departure process from the decoupling IS node by an external Poisson arrival process. We then use the FPM method to determine the appropriate external arrival rate, but we do not have to worry about the variability parameter; we just set it equal to 1. If instead we used the FPM method as described for Case 1 and we selected an entry node for an external arrival process, then we would need to specify the variability parameter of the external arrival process there. In general, in Case 2 the arrival processes to other nodes need not be approximately Poisson. However, if we apply the standard FPM method for Case 1 to Case 2, then the results should be very similar because in Case 2 the FPM method will make the variability parameter of the departure process from the decoupling IS node nearly 1.

We can also think of the procedure for Case 2 as a modification of the procedure for Case 3. The decoupling IS node also acts as a bottleneck queue. Hence, as described in Section VI, we can analyze Case 2 by removing the IS node from the closed network and replacing its department process by an external arrival process. Because of the nature of this particular bottleneck queue, i.e., because there are many servers each with low service rate, it is appropriate to make the external arrival process a Poisson process. Incidentally, we would do this in Case 3 too if there were many servers, but finitely many, each with low service rate.

If we apply the procedure for Case 3 directly to Case 2, then we let the arrival rate of the external Poisson process be the maximal possible service rate from the IS node, which corresponds to the first upper bound described in Section VI. The suggested modification is to let the arrival rate of the external Poisson arrival be such that this arrival rate would equal the departure rate at the IS node if it were included in the network. As indicated in Section VI, this modification turns out to coincide with the FPM method.

It is important to recognize that the three situations above do not nearly cover all possibilities. As indicated in Section I, in some cases an open model might also be reasonable from direct modeling considerations; often the closed model is not entirely appropriate. However, it is clear from the analysis and examples here that open models do not always produce reasonable, let alone good, approximations for closed models. For closed networks with few nodes, few servers per node and few jobs, the open-model approximations for closed models here tend to perform poorly. Further experimentation is needed to better understand the appropriate regions for each procedure. As with any approximation tool, it is very helpful in applications to make a few initial benchmark comparisons with simulations to determine the actual quality of the approximations in that context.

The specific models discussed in this paper have been relatively elementary. Many of the theorems only relate open and closed Markov Jackson networks. The major complexity considered was nonexponential FCFS servers and the associated network model treated by QNA. It is important to realize, however, that the ideas apply much more broadly. As discussed by Zahorjan,[22] these open-model approximations for closed models can be used as modules or subroutines in more complicated approximation procedures, e.g., based on network decomposition. As illustrated by Fredericks,[40] the ideas also apply directly to closed models with other complicating features such as priorities.

XIV. ACKNOWLEDGMENTS

I am grateful to A. T. Seery for writing the QNA software and to A. B. Bondi, A. A. Fredericks, H. Heffes, M. Segal, and D. R. Smith for their interest and suggestions.

REFERENCES

1. J. R. Jackson, "Jobshop-like Queueing Systems," Manage. Sci., *10,* No. 1 (October 1963), pp. 131–42.

2. W. J. Gordon and G. F. Newell, "Closed Queueing Systems With Exponential Servers," Oper. Res., *15,* No. 2 (March-April 1967), pp. 254-65.
3. F. Baskett, M. Chandy, R. Muntz, and J. Palacios, "Open, Closed and Mixed Networks of Queues With Different Classes of Customers," J.A.C.M., *22,* No. 2 (April 1975), pp. 248-60.
4. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications,* New York: Wiley Interscience, 1976.
5. F. P. Kelly, *Reversibility and Stochastic Networks,* New York: Wiley, 1979.
6. C. H. Sauer and K. M. Chandy, *Computer Systems Performance Modeling,* Englewood Cliffs, NJ: Prentice-Hall, 1981.
7. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62,* No. 9 (November 1983), pp. 2779-815.
8. K. G. Ramakrishnan and D. Mitra, "An Overview of PANACEA, A Software Package for Analyzing Markovian Queueing Networks," B.S.T.J., *61,* No. 10 (December 1982), pp. 2849–72.
9. J. Zahorjan et al., "Balanced Job Bound Analysis of Queueing Networks," Commun. A.C.M., *25,* No. 2 (February 1982), pp. 134–41.
10. D. L. Eager and K. C. Sevcik, "Performance Bound Hierarchies for Queueing Networks," A.C.M. Trans. Comput. Syst., *1,* No. 2 (May 1983), pp. 99–115.
11. E. Arthurs and B. W. Stuck, "Upper and Lower Bounds on Mean Throughput Rate and Mean Delay in Memory-Constrained Queueing Networks," B.S.T.J., *62,* No. 2 (February 1983), pp. 541–81.
12. F. Spitzer, *Principles of Random Walk,* Princeton, NJ: Van Nostrand, 1964.
13. J. McKenna, D. Mitra, and K. Ramakrishnan, "A Class of Closed Markovian Queueing Networks: Integral Representation, Asymptotic Expansions, Generalizations," B.S.T.J., *60,* No. 5 (May-June 1981), pp. 599–641.
14. J. McKenna and D. Mitra, "Integral Representations and Asymptotic Expansions for Closed Markovian Queueing Networks: Normal Usage," B.S.T.J., *61,* No. 5 (May-June 1982), pp. 661–83.
15. D. J. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54,* No. 3 (March 1975), pp. 625–61.
16. W. Whitt, "Heavy-Traffic Approximations for Service Systems With Blocking," AT&T Bell Lab. Tech. J., *63,* No. 5 (May 1984), pp. 689–708.
17. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory,* New York: Wiley, 1974.
18. W. Whitt, unpublished work.
19. P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes,* Berlin: Akademie-Verlag, 1981.
20. D. Y. Burman, "Insensitivity in Queueing Systems," Advance. Appl. Probab., *13,* No. 4 (December 1981), pp. 846–59.
21. Information Research Associates, "User's Manual for CADS," Austin, TX, 1978.
22. J. Zahorjan, "Workload Representations in Queueing Models of Computer Systems," *Proc. ACM Sigmetrics Conf.,* Minneapolis, August 1983, pp. 70–81.
23. J. Keilson, *Markov Chain Models—Rarity and Exponentiality,* New York: Springer-Verlag, 1979.
24. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models,* New York: Wiley, 1983.
25. S. Karlin and Y. Rinott, "Classes of Orderings of Measures and Related Correlation Inequalities. II. Multivariate Reverse Rule Distributions," J. Multivar. Anal., *10,* No. 4 (December 1980), pp. 499–516.
26. H. W. Block, T. H. Savits, and M. Shaked, "Some Concepts of Negative Dependence," Ann. Probab., *10,* No. 3 (August 1982), pp. 765–72.
27. K. Joag-Dev and F. Proschan, "Negative Association of Random Variables, With Applications," Ann. Statist., *11,* No. 1 (March 1983), pp. 286–95.
28. P. Billingsley, *Convergence of Probability Measures,* New York: Wiley, 1968.
29. W. Whitt, "Some Useful Functions for Functional Limit Theorems," Math. Oper. Res., *5,* No. 1 (February 1980), pp. 67–85.
30. T. C. Brown, "Some Poisson Approximations Using Compensators," Ann. Probab., *11,* No. 3 (August 1983), pp. 726–44.
31. W. Whitt, "Continuity of Generalized Semi-Markov Processes," Math. Oper. Res., *5,* No. 4 (November 1980), pp. 494–501.
32. W. Whitt, "The Continuity of Queues," Advance. Appl. Probab., *6,* No. 1 (March 1974), pp. 175–83.
33. T. Kamae, U. Krengel, and G. L. O'Brien, "Stochastic Inequalities on Partially Ordered Spaces," Ann. Probab., *5,* No. 6 (December 1977), pp. 899–912.
34. W. Whitt, "Comparing Counting Processes and Queues," Advance. Appl. Probab.,

13, No. 1 (March 1981), pp. 207–20.

35. P. Whittle, "Equilibrium Distributions for an Open Migration Process," J. Appl. Probab., *5,* No. 3 (December 1968), pp. 567–71.
36. T. C. Brown and P. K. Pollett, "Some Distributional Approximations in Markovian Queueing Networks," Advance. Appl. Probab., *14,* No. 3 (September 1982), pp. 654–71.
37. B. Hajek, "The Proof of a Folk Theorem on Queueing Delay With Applications to Routing in Networks," J.A.C.M. *30,* No. 4 (October 1983), pp. 834–51.
38. W. A. Massey, "An Operator Analytic Approach to the Jackson Network," J. Appl. Probab., *21,* No. 2 (June 1984), pp. 379–93.
39. J. B. Goodman and W. A. Massey, unpublished work.
40. A. A. Fredericks, unpublished work.
41. W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol. I, Third Ed., New York: Wiley, 1968.
42. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues With Many Exponential Servers," Oper. Res., *29,* No. 3 (May-June 1981), pp. 567–88.
43. J. W. Cohen, *The Single Server Queue,* Amsterdam: North-Holland, 1969.
44. D. Stoyan and H. Stoyan, "Monotonieeigenschaften der Kundenwartezeiten im Model GI/G/1," Z. Angew. Math., *49,* No. 12 (1969), pp. 729–34.
45. T. Rolski and D. Stoyan, "On the Comparison of Waiting Times in GI/G/1 Queues," Oper. Res., *24,* No. 1 (January-February 1976), pp. 197–200.
46. J. Keilson, "A Threshold for Log-Concavity for Probability Generating Functions and Associated Moment Inequalities," Ann. Math. Statist. *43,* No. 5 (September 1972), pp. 1702–8.
47. W. Whitt, "Uniform Conditional Stochastic Order," J. Appl. Probab., *17,* No. 1 (March 1980), pp. 112–23.
48. W. Whitt, "Multivariate Monotone Likelihood Ratio and Uniform Conditional Stochastic Order," J. Appl. Probab., *19,* No. 3 (September 1982), pp. 695–701.
49. W. Whitt, unpublished work.
50. S. Karlin and Y. Rinott, "Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Total Positivity," J. Multivar. Anal., *10,* No. 4 (December 1980), pp. 476–98.
51. E. Cinlar, "Superposition of Point Processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications,* P. A. W. Lewis, ed., New York: Wiley, 1972, pp. 549–606.
52. F. Böker and R. Serfozo, "Ordered Thinnings of Point Processes and Random Measures," Stoch. Proc. Appl., *15,* No. 2 (July 1983), pp. 113–32.
53. D. Sonderman, "Comparing Semi-Markov Processes," Math. Oper. Res., *5,* No. 1 (February 1980), pp. 110–9.
54. P. J. Schweitzer, "Bottleneck Determination in Networks of Queues," *Applied Probability—Computer Science: The Interface,* Vol. I, R. L. Disney and T. J. Ott, eds., Boston: Brikhäuser, pp. 471–85.
55. M. I. Reiman, "Open Queueing Networks in Heavy Traffic," Math. Oper. Res., *9,* No. 3 (August 1984), pp. 441–58.
56. M. I. Reiman, "Some Diffusion Approximation With State Space Collapse," *Proc. Int. Seminar on Modeling and Performance Evaluation Methodology,* New York: Springer-Verlag, 1983.
57. M. Reiser and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Queueing Systems," IBM J. Res. Develop., *18* (March 1974), pp. 110–24.
58. A. W. Shum and J. Buzen, "A Method for Obtaining Approximate Solutions to Closed Queueing Networks With General Service Times," in *Modeling and Performance Evaluation of Computer Systems,* H. Beilner and E. Gelenbe, eds., Amsterdam: North-Holland, 1978.
59. A. W. Shum, *Queueing Models for Computer Systems with General Service Time Distributions,* New York: Garland, 1980.
60. S. K. Tripathi, "On Approximate Solution Techniques for Queueing Network Models of Computer Systems," Ph.D. dissertation, Department of Computer Sciences, University of Toronto, 1981.
61. A. B. Bondi, "Incorporating Open Queueing Models Into Closed Queueing Network Algorithms," Ph.D. dissertation, Department of Computer Sciences, Purdue University, 1984.
62. A. B. Bondi and W. Whitt, unpublished work.
63. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," AT&T Bell Lab. Tech. J., *63,* No. 1 (January 1984), pp. 115–38.
64. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Con-

straints," AT&T Bell Lab. Tech. J., *63,* No. 1 (January 1984), pp. 139-61.

65. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," AT&T Bell Lab. Tech. J., *63,* No. 1 (January 1984), pp. 163-75.
66. M. Reiser and H. Kobayashi, "The Effects of Service Time Distributions on System Performance," *Information Processing 74,* Amsterdam: North-Holland, 1974, pp. 230-4.
67. P. Kühn, *Tables on Delay Systems,* Stuttgart: Institute of Switching and Data Technics, University of Stuttgart, 1976.
68. 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.
69. J. P. Buzen and P. S. Goldberg, "Guidelines for the Use of Infinite Source Queueing Models in the Analysis of Computer System Performance," Proc. AFIPS Nat. Comput. Conf., *43,* 1974, pp. 371-4.

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. 9, November 1984
Printed in U.S.A.

On the Application of Energy Contours to the Recognition of Connected Word Sequences

By L. R. RABINER*

(Manuscript received May 10, 1984)

It has recently been shown that small but consistent improvements in isolated word recognition accuracy can be obtained by supplementing the Linear Predictive Coding (LPC) features for each frame of a word by a normalized energy value for that frame. The key idea in using energy is to normalize the frame energy by the local energy maximum in time (i.e., relative to the peak energy of the spoken word). If we want to extend the concept of using frame energy as a supplement to the LPC feature set for connected word recognition, we must provide a dynamic method of energy normalization so that the peak energy within strings can closely approximate the energy contours of individual words strung together. In this paper such a dynamic energy normalization is proposed, and it is shown to provide improvements in connected word recognition applications. The normalization consists of determining a continuous *peak energy* contour for the speech, where the peak energy is determined over periods of time essentially corresponding to a syllable, and then modifying the actual energy contour with the peak energy contour so that absolute energy maxima occur about once per syllable. In this manner, the dynamically normalized, temporal energy contour of the word string effectively provides a set of temporal markers of high-energy events (content words) that aid the recognition of connected word sequences.

I. INTRODUCTION

The effectiveness of supplementing standard spectral features with an energy measurement (suitably normalized) for isolated word rec-

* AT&T Bell Laboratories.

ognition applications has recently been demonstrated by several researchers.[1-3] The basic idea of these schemes is to define an enhanced feature set (for each frame of speech within the word to be recognized) consisting of a pth-order Linear Predictive Coding (LPC) vector, **a**, concatenated with a normalized frame log energy, $\hat{E}_T$, where the normalization is with respect to the peak energy within the entire word. In this manner, the frame energy value is relative to the peak energy within the word, and is therefore relatively insensitive to gain variations in transmission and/or recording.

For connected word recognition applications, the concept of how to provide proper energy normalization across a sentence-length utterance is one that is potentially open to a great deal of controversy. There is no exactly correct mechanism for handling the energy variations that occur naturally when words are strung together and spoken at various rates of articulation. However, it seems reasonable, and intuitively appealing, that some type of syllabic rate normalization should be able to highlight and identify a large fraction of the words (especially so-called content words) in a spoken sentence. In this manner, the increase and decrease in the overall energy level would be naturally compensated by the Automatic Gain Control (AGC) action of the normalization scheme.

The major obstacle to implementing a syllabic rate, energy normalization procedure for use with connected strings is that it is almost impossible to design such an algorithm unless the rate of articulation is known. Unfortunately, for most practical situations, we do not know the rate of articulation of the speech; hence we are forced to choose a set of implementation parameters that represent a compromise over those that are optimum for the particular spoken string, and those that are optimum for a wide class of talkers, strings, etc. The design and implementation of the syllabic rate, energy normalization procedure is discussed in Section II. In Section III we present results of an experimental evaluation of the energy normalization scheme on both connected digit strings and on sets of airlines words for use in the AT&T Bell Laboratories airlines information and reservation system.[4,5] Finally, in Section IV we discuss the results and their implications for further research.

II. ENERGY NORMALIZATION FOR CONNECTED WORD STRINGS

We define the log energy contour, $E(m)$, of the connected word string as

$$E(m) = 10 \log_{10}[V_m(0)], \qquad m = 1, 2, \cdots, M, \tag{1}$$

where $V_m(0)$ is the zeroth-order autocorrelation of the speech, i.e.,

$$V_m(0) = \sum_{n=0}^{N-1} s[n + (m - 1)L]^2, \tag{2}$$

where M, L, and N are the number of frames in the string, the number of samples shifted between frames, and the frame size, respectively, and where $s(n)$ is the speech signal. Typically, for telephone recordings, we use a sampling rate of 6.67 kHz on the speech, and use values of $N = 300$ samples (45-ms frames), and $L = 100$ samples (15-ms shift).

For isolated word sequences, the normalization of the log energy contour is straightforward, and consists of locating the peak log energy across the word, $E_{\max}$ as,

$$E_{\max} = \max_{1 \le m \le M} [E(m)] \tag{3}$$

and normalizing the energy contour by subtracting $E_{\max}$ from each frame, i.e.,

$$\hat{E}(m) = E(m) - E_{\max}. \tag{4}$$

In this manner the log energy values are constrained to have a peak value of 0 dB, and the stressed vowels for a word are essentially guaranteed to have log energy values close to 0 dB.

Based on the above normalization procedure, reference- and test-word energy contours can be compared using a simple nonlinear, energy distance metric, which is then added to the standard LPC-shape distance to give an overall distance between test and reference frames.

For connected word strings a more sophisticated energy normalization scheme is required. The idea of the normalization is to make the local energy maximum for each content word in the string as close to 0 dB as possible. By content words we mean words with distinct stressed vowels (i.e., all digits in strings), as opposed to function words (e.g., "to", "and", "the", "a") in which there is often no stressed vowel in connected speech. Basically, what is required for performing such a normalization is a syllable detector. Although several approaches to syllable detection have been described in the literature,[6–10] we chose to implement a simple, signal processing approach to normalization, which is felt to be more appropriate to the problem at hand than other alternatives.

A block diagram of the log-energy-normalization algorithm for connected word strings is given in Fig. 1. The log energy contour, $E(m)$, $m = 1, 2, \cdots, M$, of the speech signal, $s(n)$, is first computed according to eq. (1). A "syllabic rate" energy envelope, $V(m)$, is computed as

$$V(m) = \max_{\max\left[1, m-\frac{NW}{2}\right] \le q \le \min\left[M, m+\frac{NW}{2}\right]} [E(q)], \qquad (5)$$

where the parameter *NW* is the number of frames over which the energy envelope maximum is computed. (We have considered values of *NW* from 15 to 35, i.e., five to two syllables per second.)

The syllabic rate, energy envelope contour, $V(m)$, is next smoothed by a median smoother[11] with a smoother duration of *NM* frames, where *NM* is typically chosen to be about half the size of *NW*, i.e., 10 to 20 frames. The median smoother eliminates "sharp" dips in the syllabic rate, energy envelope contour between syllables.

The final step in the process is to modify the log energy contour, $E(m)$, by the median-smoothed, syllabic rate energy envelope, $\hat{V}(m)$, to give

$$\hat{E}(m) = E(m) - \hat{V}(m), \qquad (6)$$

which is the final, normalized, log energy envelope.

Figures 2 through 5 illustrate the algorithm for four sets of word strings. In each of these figures, the upper plot shows $E(m)$ (normalized so that its global peak across the string is set to 0 dB), and $\hat{E}(m)$ (dashed line) superimposed; the lower plot shows $V(m)$ and $\hat{V}(m)$ (most of the time they are identical).

Figure 2 shows results for the connected digit string /54110/ spoken at a fairly deliberate rate (2-1/2 digits per second or 150 digits per minute). Figure 2b shows that the energy envelope exhibits approximately a 7.6-dB variation from the first digit peak to the fourth digit peak. After peak energy normalization, each of the five digits in the string is clearly marked and each attains a 0-dB energy peak during the stressed vowel.

The example of Fig. 3 is for the digit string /5820/ spoken fairly rapidly (175 digits per minute). For this case the digit 2 is not properly normalized, since the median smoother misses the energy envelope by about 2 dB. However, each of the four digits in the string is more distinct in the normalized energy contour than in the original energy contour.

The example of Fig. 4 is for the sentence, "I want to make a

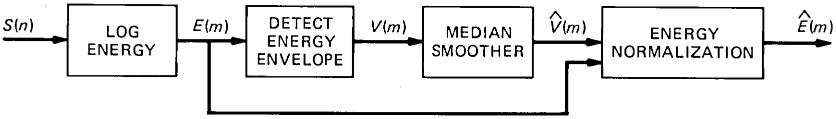


Fig. 1—Block diagram of dynamic energy normalization scheme for connected word strings.

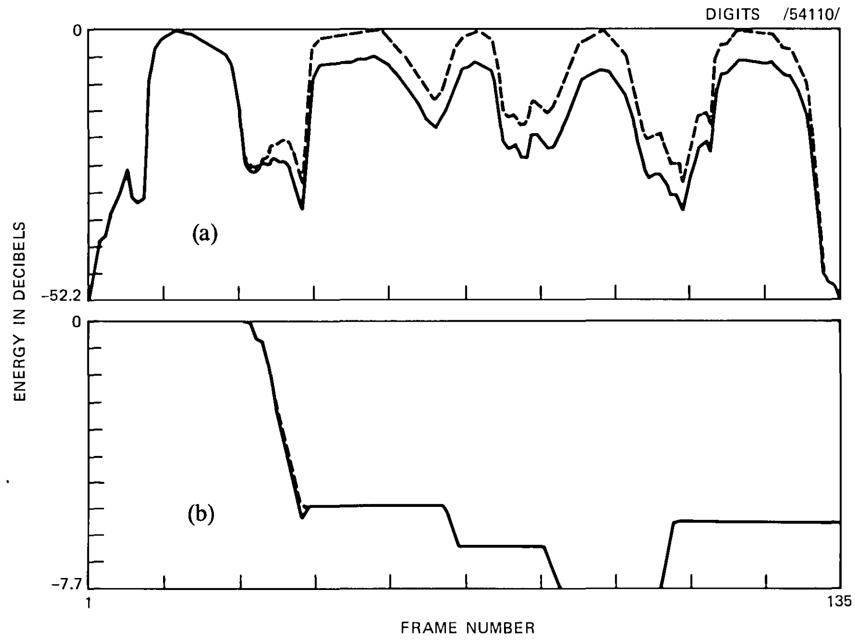


Fig. 2—(a) Log energy contours (original plus normalized) and (b) peak energy envelope contours (original plus median smoothed) for the digit string /54110/.

reservation", spoken at a rate of 221 words per minute. The energy normalization does a good job for the content words. "I", "want", and "make", but is not able to handle the brief, unstressed words "to" and "a", and actually provides a double normalization for the word "reservation", because of the presence of two stressed vowels in the four-syllable word. The inability of the algorithm to handle the very short function words in continuous speech is inherently unalterable, and the recognition algorithm, which ultimately uses the normalized energy contour, must still work reliably in the face of this type of shortcoming. Similarly, the detection of multiple stressed vowels with a single polysyllabic word is a natural result of the detection process, and must be properly handled by the recognizer. We will discuss these points further in Section III.

The final example of this group, Fig. 5, shows results for the 12-word sentence, "I would like to return on Wednesday afternoon the one three October", spoken at a rate of 172 words per minute. For this sentence a large range of energy values for the individual words is exhibited (i.e., 15.3 dB on the lower plot), and even this large a range is not quite enough to handle each of the content words in the sentence. The only word that was not properly normalized was "would", which was highly reduced. The words "Wednesday" and "October" both had

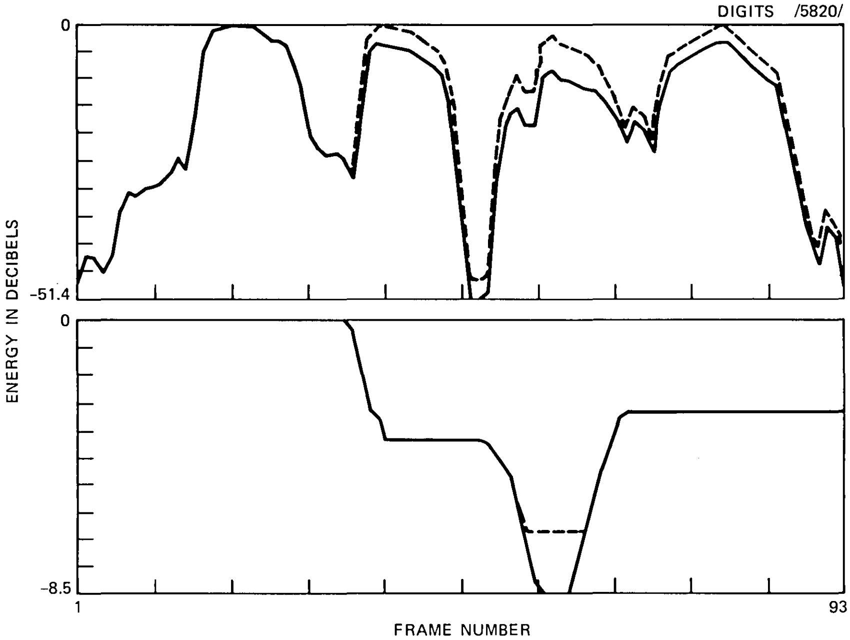


Fig. 3—Log energy contours (original plus normalized) and peak energy envelope contours (original plus median smoothed) for the digit string /5820/.

two stressed vowels and hence were normalized to 0 dB at two places within the word.

III. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of the energy normalization algorithm for connected word strings, a series of three experiments were run. For the first experiment, we performed a recognition test on 1520 connected digit strings from 19 talkers. All recordings were made over local dialed-up telephone lines and all recognition tests were run using the level-building, Dynamic Time Warping (DTW) algorithm[12] in a speaker-independent mode using word templates extracted from a different set of talkers.[13-14] Details of the way in which the reference set were extracted are given in Ref. 14.

The second recognition experiment used a vocabulary of 129 airlines terms and a deterministic language model (i.e., a grammar) to specify allowable sentences in the language. For this experiment, a syntax-directed, level-building, DTW algorithm[15] was used as the recognizer. There were six test talkers, each of whom spoke a balanced set of 51 sentences from the language. (The set was balanced in terms of usage of words in the vocabulary and in terms of covering all major paths in

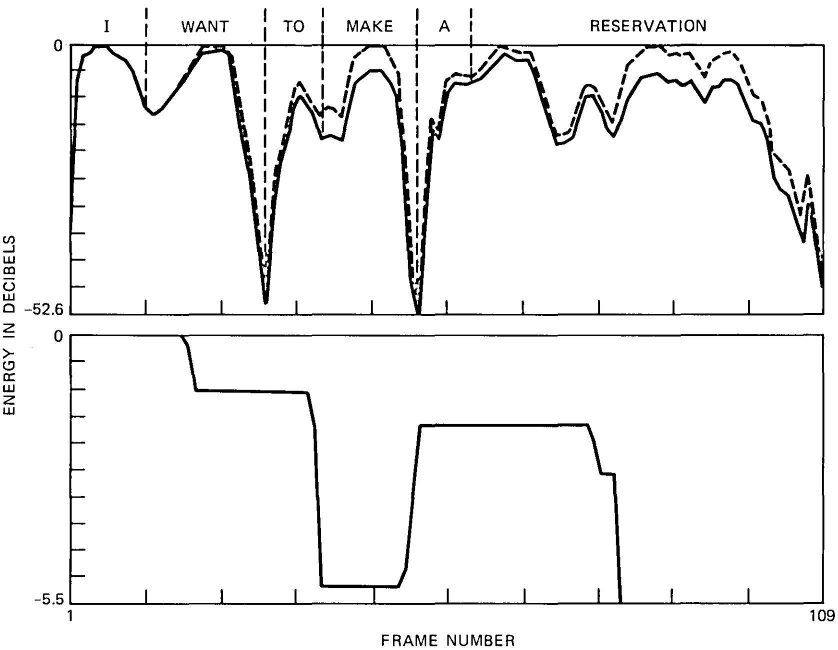


Fig. 4—Log energy contours (original plus normalized) and peak energy envelope contours (original plus median smoothed) for the sentence "I want to make a reservation". Each word in the sentence is demarked (approximately) by vertical dashed lines.

the grammar.) The list of 51 sentences used in this experiment is given in Table I. A total of 438 words occurred in the 51 sentences; hence the average sentence duration was somewhat over eight words. Four of the six test talkers provided a set of isolated-word training patterns for the 129-word vocabulary using the robust training procedure of Rabiner and Wilpon.[16] For these four talkers we ran both speaker-dependent and speaker-independent recognition tests; for the other two talkers only speaker-independent recognition tests were run. The speaker-independent runs used a speaker-independent, isolated-word reference set obtained by means of a clustering analysis of the word tokens of 100 different talkers (50 male, 50 female).[17]

The third recognition experiment again used the 129-word airlines vocabulary, but substituted a level-building, Hidden Markov Model (HMM) for the DTW recognizer.[18] Single-word HMMs were designed for each of the 129 words in the vocabulary, based on the same training set from which the speaker-independent word templates were created. (No speaker-dependent models were used in this experiment.) Word models were concatenated, according to the language model (the deterministic grammar) using the level-building concept to link ends of one model to the beginnings of the next model. The individual word

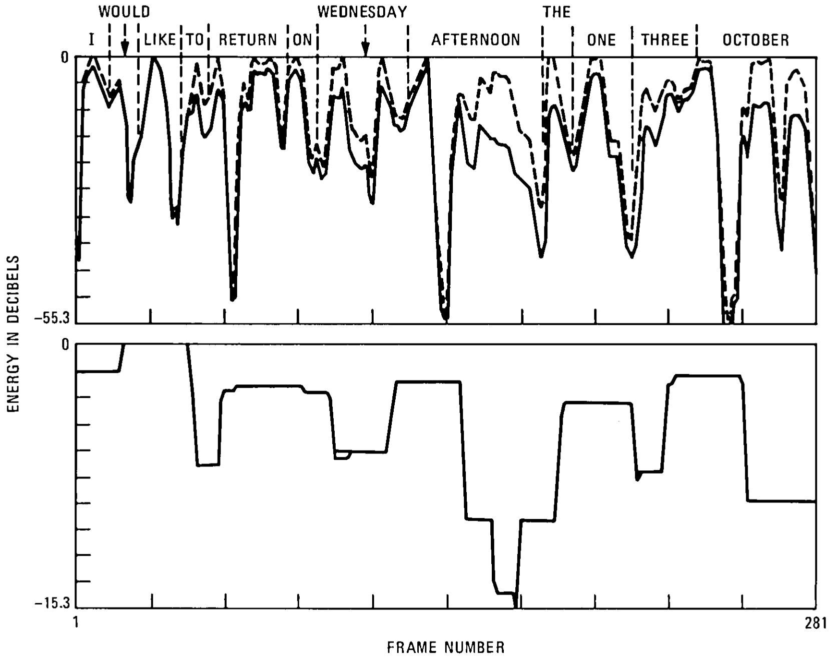


Fig. 5—Log energy contours (original plus normalized) and peak energy envelope contours (original plus median smoothed) for the sentence "I would like to return on Wednesday afternoon the one three October". Each word in the sentence is demarked (approximately) by vertical dashed lines.

models each had ten states, and used an energy-based Vector Quantizer (VQ)[2] with 128 code-book entries.

3.1 Results of experiment 1—connected digits

The results of the connected digits runs are given in Table I and Fig. 6. The 1520 strings were divided into two groups of 760 strings each; the first group was spoken at deliberate rates (about 135 digits per minute), whereas the second group was spoken at normal rates (about 170 digits per minute). String error rates were measured for the top β candidates (i.e., the probability that the correct string was not in the β best strings) with string length unknown, and for the top candidate for known string lengths. A value of β of five was used for these tests.

Table II shows the $\beta = 1$ results for a recognizer without energy (i.e., using I.PC vectors alone); a recognizer using energy, where only a global peak normalization (similar to the algorithm for isolated words) is used; a recognizer with energy, using the dynamic normalization procedure of Section II; a recognizer with a shape VQ with 128 code-book entries; and a recognizer with an energy VQ with 128 entries

Table I—Sentences used to evaluate the airline recognition system

1 I want to make a reservation.
2 I would like some information please.
3 I want to go from New York to Los Angeles on Tuesday morning.
4 I would like to return on Wednesday afternoon the one three October.
5 I would like a nonstop flight.
6 When do flights leave Philadelphia for Detroit on Monday afternoon?
7 I want to go at twelve o'clock.
8 I would like to depart at night.
9 I want to leave in the morning.
10 I want to depart from Boston on the evening of the oh nine November.
11 How many flights are there from Washington to Denver on Thursday night?
12 How many flights go from Seattle to Miami on the two eight February?
13 What plane is on flight two six to Chicago?
14 How many stops are there on the flight?
15 I would like flight number four one.
16 I will take flight five three.
17 I would like a first class seat.
18 I need three seats.
19 I want one coach seat.
20 What is the flight time from Boston to Chicago.
21 Is a meal served on the flight to Denver?
22 How much is the fare?
23 What is the fare from Detroit to Philadelphia on Sunday night?
24 When does flight number two from Los Angeles arrive?
25 At what time does flight seven one to Seattle depart?
26 My home phone number is area code two oh one six two four one two four six.
27 My office phone number is five three six two one five two.
28 Please repeat the arrival times.
29 Please repeat the departure time.
30 I will pay by credit card.
31 I prefer the Lockheed ten eleven.
32 I prefer the Boeing seven four seven.
33 I prefer the D.C. nine.
34 I prefer the Douglas D.C. ten.
35 I prefer the B.A.C. ten.
36 I will pay by Master Charge.
37 I will pay by cash.
38 I will pay by Diners Club.
39 I will pay by American Express.
40 I want to go at eleven a.m.
41 I want to go at six p.m.
42 I want to return to Chicago on the three oh December.
43 I would like to depart on Friday evening.
44 I would like one first class seat on flight number four four to Los Angeles.
45 I want to return on the oh nine March.
46 I want to go to Washington on the two four April.
47 I would like to return to New York on the oh one May.
48 I want to leave for Los Angeles on the morning of the one four June.
49 I want to go from Boston to Philadelphia on Tuesday morning the oh four July.
50 I would like to return on the oh seven August.
51 At what time do flights leave Boston for Denver on the two seven September?

and dynamic energy normalization. Figure 6 shows the string error rate, as a function of β, for the five recognizers described above. Based on the results of Table II and Fig. 6, the following observations can be made:

1. For connected digit strings, there is essentially no advantage to using energy in addition to LPC shape. The only case in which energy provided a significant performance improvement was for deliberately

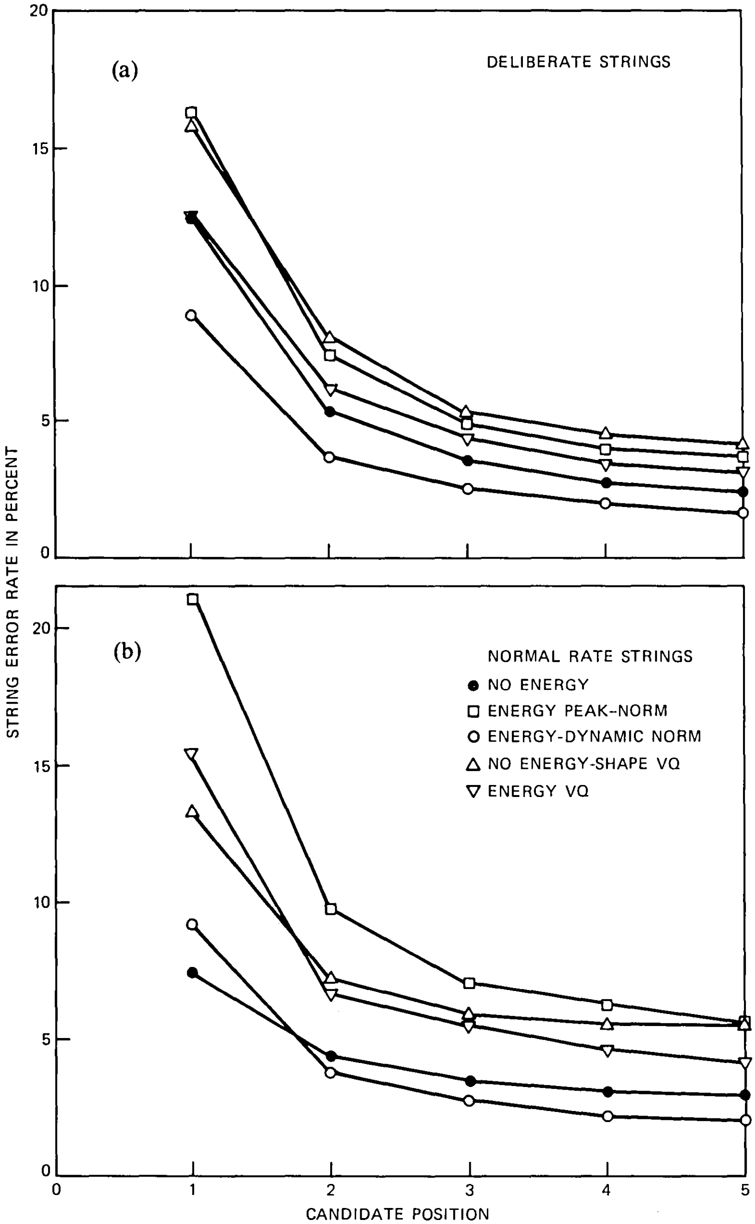


Fig. 6—String error rates as a function of candidate position for (a) deliberate strings and (b) normal rate strings for five recognition conditions.

Table II—String error rates in percent for connected digit strings

	Deliberate Strings		Normal Rate Strings	
Condition	Length Unknown	Length Known	Length Unknown	Length Known
No energy	12.4	4.9	7.4	5.3
Energy-peak norm	16.3	14.3	21.3	19.2
Energy-dynamic norm	8.8	5.9	9.2	7.2
No energy-shape VQ	15.9	9.2	13.2	11.1
Energy VQ	12.4	8.6	15.3	12.8

spoken digit strings whose length was unknown. For all other cases there was a small loss in performance when energy was incorporated into the recognizer.

2. Improper normalization of the energy contour leads to significant degradation in performance on connected digit strings. This result shows that the dynamic normalization procedure is indeed providing a better model for the energy contours of individual words than those obtained from just using the original energy contour of the utterance.

3. The small performance degradation for normal rate strings of unknown length is essentially only for the top recognition candidate. As seen in Fig. 6, for candidate positions 2 through 5 the performance with dynamic normalization of energy is indeed slightly better than without energy.

3.2 Results of experiment 2—airlines sentences using DTW

The results of the recognition runs using the airlines vocabulary and grammar, and using the DTW level-building recognizer are given in Table IIIa. This table shows average string and word error rates for both the speaker-dependent and speaker-independent runs for two conditions, namely, the recognizer without energy (i.e., using only LPC in the distance) and the recognizer with the dynamic energy normalization.

The results of Table IIIa show that in the speaker-dependent mode, the improvement in both sentence and word accuracy is dramatic (7.4 percent and 1.7 percent, respectively). In the speaker-independent mode there is an improvement in performance of 1.1 percent in string error rate when using energy, but the word error rate is essentially the same for both conditions. Presumably this result is due to the diversity of patterns and energy contours in the 12-template-per-word reference set; hence the reliance on energy to provide marker points during the word string is considerably less than for the speaker-dependent runs.

3.3 Results of experiment 3—airlines sentences using HMM

The results of the recognition runs using the airlines vocabulary and grammar, and using the HMM level-building recognizer are given

Table III—A comparison of string and word error rates for airline sentences using a DTW level-building algorithm and an HMM level-building algorithm

(a) String and Word Error Rates in Percent for Airlines Sentences Using DTW Level-Building Algorithm

	Speaker Dependent		Speaker Independent	
Condition	String Error Rate	Word Error Rate	String Error Rate	Word Error Rate
No Energy	20.6	5.5	26.9	7.4
Energy-Dynamic Norm	13.2	3.8	25.8	7.5

(b) Speaker-Independent String and Word Error Rates in Percent for Airlines Sentences Using an HMM Level-Building Algorithm

Condition	String Error Rate	Word Error Rate
Energy-Peak Norm	34.0	8.6
Energy-Dynamic Norm	25.1	6.7

in Table IIIb. This table shows average string and word error rates for two conditions, namely, using energy with only global peak normalization, and using energy with dynamic normalization. (A partial run was made without energy, but the string error rates were on the order of 95 percent! Hence, for the HMM recognizer, the use of energy, in some form, is mandatory.)

The results of Table IIIb again show a dramatic reduction in both string and word error rates when the dynamic energy normalization is used (i.e., 8.9 percent and 1.9 percent, respectively). Comparing the results to those given in Table IIIa it can be seen that the HMM level-building recognizer (which uses a 128-codeword VQ) actually outperforms a 12-template-per-word, DTW, level-building recognizer *without* VQ.

IV. DISCUSSION

The results presented in this paper on the use of energy along with LPC for recognition of connected word strings indicate the following:

1. Simple application of the peak energy normalization scheme appropriate for isolated words leads to poor performance for connected word systems.

2. Improved performance can be obtained by using a dynamic energy normalization, which essentially adjusts the energy contour according to the local maximum over a time duration roughly corresponding to a syllable.

3. For relatively simple vocabularies, such as the digits, the information contained in the energy contour is, at best, only marginally useful for improving recognizer performance. The condition under

which it performs the best is in reducing digit insertions for rates of articulation that are fairly low. For normally spoken connected digit strings, there is actually a small degradation in performance when the energy contour is used.

4. For more complex vocabularies, such as the set of airlines terms, the information contained in the energy contour can and does improve the performance of the recognizer on connected word strings; in some cases the improvements are quite dramatic. The reason for this improvement in performance is that the energy contour, when properly normalized, essentially highlights the content words in the sentence and provides a boost to the alignment of words from the grammar.

There are several issues concerning the implementation of the energy normalization that should be discussed here. First of all, it should be clear that this, and any other proposed energy normalization scheme, is essentially an ad hoc procedure for highlighting words in connected strings. There is no exactly correct method for performing the appropriate normalization; at best, we can hope that the proposed method has some desirable properties and performs well in some typical applications.

A second point concerns the variable parameters, NW and NM, of the implementation of the energy normalization algorithm. We have experimented with values of $10 \leq NW \leq 35$ and $10 \leq NM \leq 25$, and have found that the performance results are relatively insensitive over a wide range of values of NW and NM. This is a highly desirable result in that a fixed set of values can be chosen and used in all circumstances. However, it should be clear that, in individual cases, when the rate of articulation is high (e.g., over 200 words per minute), values of NW and NM near the lower limits will give better performance than those near the upper limits. Conversely, for strings articulated at low rates (near 100 to 130 words per minute), values of NW and NM near the upper limits will give the best recognition performance.

Finally, the issue arises as to how to handle polysyllabic words with more than one stressed vowel. For our runs we have made no attempt to do anything special for such cases, since the energy contours of the isolated word tokens, in these cases, naturally exhibit two strong (almost equal level) energy peaks. The result indicates no special problems with such polysyllabic words. We did do one check in which the isolated word reference energy patterns themselves were passed through the dynamic energy, normalization procedure and then used in the DTW recognizer. The results were one-for-one identical with those obtained without this reference energy correction procedure. Hence we conclude that multistressed, polysyllabic words present no real problems for the dynamic energy, normalization algorithm.

V. SUMMARY

In this paper we have proposed one approach to dynamically normalizing the energy contour of a connected word string so that energy can be used along with LPC spectral shape in the recognition of connected word strings. We have shown the approach to be reasonable from the point of view of finding content words in the string and bringing their energy levels to be local peaks of essentially fixed level in the string.

Recognition results indicate that energy is primarily useful for complex word vocabularies but is at best marginal for simple (monosyllabic) word vocabularies such as the digits. In all cases we have shown that the proposed dynamic energy normalization outperforms the simple peak energy normalization procedure that was shown to be suitable for isolated word sequences.

REFERENCES

1. M. K. Brown and L. R. Rabiner, "On the Use of Energy in LPC-Based Recognition of Isolated Words," B.S.T.J., *61*, No. 10 (December 1982), pp. 2971–87.
2. L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "A Vector Quantizer Combining Energy and LPC Parameters and Its Application to Isolated Word Recognition," AT&T Bell Lab. Tech. J., *63*, No. 5 (May–June 1984), pp. 721–35.
3. L. R. Rabiner, K. C. Pan, and F .K. Soong, "On the Performance of Isolated Word Speech Recognizers Using Vector Quantization and Temporal Energy Contours," AT&T Bell Lab. Tech. J., *63*, No. 7 (September 1984), pp. 1245–60.
4. S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a Word Recognition System Using Syntax Analysis," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 1619–26.
5. S. E. Levinson and K. L. Shipley, "A Conversational Mode Airline Information and Reservation System Using Speech Input and Output," B.S.T.J., *59*, No. 1 (January 1980), pp. 119–37.
6. O. Fujimura, "Syllable as a Unit of Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP*-23, No. 1 (February 1975), pp. 79–82.
7. P. Mermelstein, "Automatic Segmentation of Speech Into Syllabic Units," J. Acoust. Soc. Amer., *58*, No. 4 (October 1975), pp. 880–3.
8. D. C. Sargent, K. P. Li, and K. S. Fu, "Syllable Detection in Continuous Speech," J. Acoust. Soc. Amer., *45* (1974), p. 410(A).
9. A. N. Stowe, "Segmentation of Speech Into Syllables," J. Acoust. Soc. Amer., *25* (1963), p. 806(A).
10. D. Kahn, "A Syllable Parsing Algorithm for Telephone Quality Speech," J. Acoust. Soc. Amer., Sup. 1, *72* (1982), p. 530(A).
11. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 6 (December 1975), pp. 552–7.
12. C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-29*, No. 2 (April 1981), pp. 284–97.
13. L. R. Rabiner, A. Bergh, and J. G. Wilpon, "An Improved Training Procedure for Connected-Digit Recognition," B.S.T.J., *61*, No. 6 (July–August 1982), pp. 981–1001.
14. L. R. Rabiner, J. G. Wilpon, A. M. Quinn, and S. G. Terrace, "On the Application of Embedded Digit Training to Speaker Independent Connected Digit Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-32*, No. 2 (1984), pp. 272–80.
15. C. S. Myers and S. E. Levinson, "Speaker Independent Connected Word Recognition Using a Syntax-Directed Dynamic Programming Procedure," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-30*, No. 4 (August 1982), pp. 561–5.

16. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68*, No. 5 (November 1980), pp. 1271–6.
17. J. G. Wilpon, L. R. Rabiner, and A. Bergh, "Speaker-Independent Isolated Word Recognition Using a 129 Word Airline Vocabulary," J. Acoust. Soc. Amer., *72*, No. 2 (August 1982), pp. 390–6.
18. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Use of Hidden Markov Models for Speaker Independent Recognition of Isolated Words From a Medium-Size Vocabulary," AT&T Bell Lab. Tech. J., *63*, No. 4 (April 1984), pp. 627–42.

AUTHOR

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1962—. Presently, Mr. Rabiner is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Member, National Academy of Engineering, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 9, November 1984
Printed in U.S.A.

Spatial Filtering Radio Astronomical Data: One-Dimensional Case

By H. E. ROWE*

(Manuscript received May 26, 1983)

Radio astronomical measurements of radio brightness are made by pointing an antenna at a regular array of points in the sky and measuring the received noise power at each point. In the absence of receiver noise, the measured brightness is the convolution of the true brightness distribution with the antenna effective area (i.e., receiving power pattern), evaluated at the point of observation. Front-end noise in the radiometer receiver adds fluctuations inversely proportional to the observing time at each measured point. From such data, we calculate optimum mean-square estimates for two quantities: measured brightness between observations, and true brightness at and between observations. The first is interpolation; the second, called restoration, partially deconvolves the antenna pattern from the measured data. We determine the errors associated with each, as functions of: (1) receiving antenna pattern, (2) separation between observations, and (3) radiometer output signal-to-noise ratio. These results permit the construction of maps of measured and true brightness, with known mean-square errors. In this paper we study the one-dimensional version of this problem, assuming a large number of measured points. We find that measured points should be separated by about half the (full) 3-dB beamwidth for conventional antennas. Restoration is more costly than interpolation.

I. INTRODUCTION

Radio astronomical measurements of radio brightness are made by pointing an antenna at a regular array of points in the sky and

* AT&T Bell Laboratories.

measuring the power of the random signal at the antenna output with a radiometer receiver. The measured points may lie in a square array; alternatively, we may wish to consider a hexagonal array, or irregular arrays of measured points.

We denote the antenna output power for each measured point as the measured brightness; the measured brightness is the convolution of the true brightness distribution with the antenna receiving power pattern, evaluated for the particular point under study. The radiometer receiver has a "signal" output proportional to the measured brightness, plus a "noise" output, due principally to the receiver front-end noise (which far exceeds the received power at the antenna output). The receiver "noise" output has mean-square value proportional to the receiver noise temperature squared and inversely proportional to the observation (or integration) time (see the Appendix of Ref. 1).

We wish to construct maps of radio brightness from such discrete measurements. First, let us estimate the measured brightness between observations, i.e., what we would have observed if we had pointed the antenna in between the actual measured points. This has been called "interpolation".[2]

However, the antenna does not have an infinitely narrow beam, and hence the measured brightness is a smoothed version of the true brightness. We wish to deconvolve the antenna pattern so as to determine the true brightness distribution as closely as possible. This has been called "restoration".[2]

For a given antenna, two factors that limit accuracy for either interpolation or restoration are the separation between measurements and receiver noise.

Suppose first that receiver noise is absent. Measurements that are too widely separated will provide little information about the brightness at points far from the observations. However, measurements that are too close yield redundant information. Receiver noise limits the accuracy for both interpolation and restoration; however, restoration is much more severely affected since it involves deconvolution of the antenna pattern.

Receiver noise at each point is inversely proportional to the observation time at that point, as noted above. Therefore, if we allot a given total amount of time for a particular region of the sky, we can trade off separation between observations against signal-to-noise ratio. That is, we may choose few widely spaced observations, with long observing times and hence small noise, or alternatively many closely spaced observations with short observing times and hence large noise.

We require the optimum mean-square estimates for interpolation and restoration. These results will yield the best angular separation between observations, as well as the relative costs of restoration and interpolation. The optimum estimates depend on the statistical model assumed for the brightness distribution; we assume the brightness varies rapidly compared to the antenna beamwidth.

This paper explores the one-dimensional version of this problem, an antenna with a strip aperture and a fan beam. We assume an infinite number of equally spaced observed points. The mean-square interpolation and restoration errors are determined as functions of the separation between observations, antenna width, and signal-to-noise ratio for two (one-dimensional) antenna illumination functions:

1. Uniform illumination (maximum gain antenna)
2. Truncated Gaussian illumination with a 15-db taper.

For this one-dimensional problem we find that observations should be separated by about half the (full) 3-db beamwidth in the case of truncated Gaussian illumination, which corresponds to a normal antenna. For a given error, restoration is much more expensive than interpolation in observing time.

The present results serve as a guide for similar studies of the real two-dimensional case, i.e., an antenna with a circular aperture and a pencil beam.

II. SAMPLED-DATA MODEL FOR MEASUREMENT OF A ONE-DIMENSIONAL INCOHERENT FIELD BY A FAN-BEAM ANTENNA

Consider a strip antenna with its aperture located in the x-y plane, as we see in Fig. 1. The aperture has width W along the x axis, and is centered about the y-z plane; the aperture extends to $\pm\infty$ along the y axis. All field quantities are assumed to be independent of y; therefore, the antenna has a fan beam, with gain and effective width (as transmitting and receiving antenna, respectively) that depend on only the angular coordinate t, measured in the x-z plane, of a cylindrical coordinate system shown in Fig. 1. The angle t is not measured in radians, but rather is suitably normalized to simplify the following relations; the details of this normalization are unnecessary for our present purposes. Suppose for the present that the antenna beam points along the z axis; denote the effective width by $A(t)$, with Fourier transform $\mathscr{A}(f)$.

Let the aperture electric field (in the x-y plane) be polarized along the x direction, and denoted by $E(x)$. Then for narrow-beam antennas $\mathscr{A}(f) \propto E(f) \circledast E^*(-f)$, where we use the symbol $\circledast$ to denote convolution throughout. Since $E(x)$ is zero outside of the aperture, i.e., for $|x| > W/2$, it follows that $\mathscr{A}(f)$, the Fourier transform of the effective width $A(t)$, is strictly bandlimited to $|f| < W$.

We measure a one-dimensional, incoherent field, with radio brightness $x(t)$, by pointing this antenna in direction t, denoting the power out of the antenna feed as $x_o(t)$, which we call the measured brightness.

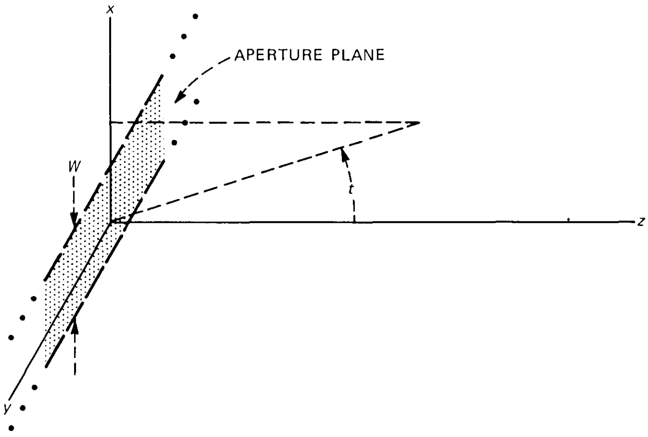


Fig. 1—Geometry of strip antenna.

The convolution of $x(t)$ with $A(t)$ is $x_o(t) = x(t) \circledast A(t)$. Such measurements are repeated at equally spaced angles $t = kT, \cdots, -1, 0, 1, \cdots$, providing measured brightness samples $x_o(kT)$. Independent noise is added by the receiver to these samples; and from these noisy samples we wish the optimum linear estimate of the brightness $x(t)$ (restoration), or of the measured brightness $x_o(t)$ (interpolation). Figure 2a shows these measurements.

The symbols in Fig. 2 have the following definitions for the one-dimensional antenna problem:

t normalized angular coordinate.

f normalized spatial frequency.

$x(t)$ radio brightness (one-dimensional).

W antenna width.

$E(x)$ aperture electric field; zero for $|x| > W/2$.

$A(t)$ effective width (as a receiving antenna).

$\mathscr{A}(f)$ Fourier transform of $A(t)$; approximately proportional to $E(f) \circledast E^* (-f)$ in narrow-beam approximation, strictly bandlimited to $|f| < W$.

kT observing angles.

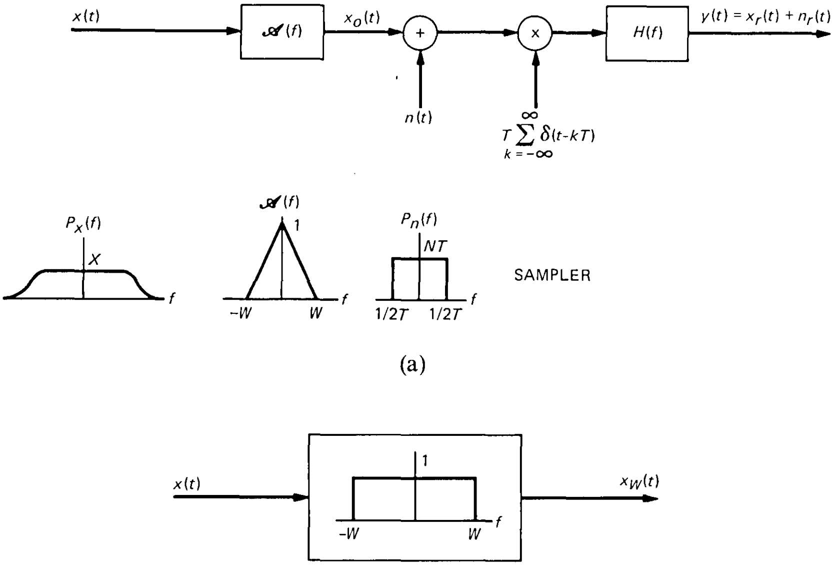
$x_o(kT)$ receiver signal samples.

$n(t)$ random function; $n(kT)$ = receiver noise samples, independent for different k.

N $= \langle n^2(kT) \rangle$, expected sample noise power; proportional to receiver noise temperature squared and inversely proportional to integration time.

$h(t)$ weight function for data samples $x_o(kT) + n(kT)$.

$H(f)$ Fourier transform of $h(t)$; transfer function used as spatial filter on data samples.



(a)

(b)

Fig. 2—Sampling and reconstruction of a stationary random function. (a) Wideband input—general input and reconstruction filters. (b) Bandlimited version of input.

The effective width $A(t)$ and its Fourier transform $\mathscr{A}(f)$ satisfy the following relations in the narrow-beam approximation, with suitable normalization of the angular coordinate t:

$$0 \leq A(t) \leq W$$

$$\mathscr{A}(f) = \frac{E(f) \circledast E^*(-f)}{\int_{-\frac{W}{2}}^{\frac{W}{2}} |E(x)|^2 dx} = \frac{\int_{\max\left(-\frac{W}{2},\, -\frac{W}{2}+f\right)}^{\min\left(\frac{W}{2},\, \frac{W}{2}+f\right)} E(x)E^*(x-f)dx}{\int_{-\frac{W}{2}}^{\frac{W}{2}} |E(x)|^2 dx}$$

$$\mathscr{A}(0) = \int_{-\infty}^{\infty} A(t)dt = 1; \qquad \mathscr{A}(f) = 0,\ |f| > W. \qquad (1)$$

We do not consider super-gain antennas in the narrow-beam approximation. Maximum effective width, $A(0) = W$, is attained for uniform illumination of the antenna aperture with zero phase error, $E(x) = 1$ for $|x| < W/2$, i.e., for an antenna with maximum on-axis gain. For this case $\mathscr{A}(f)$ is the triangular function illustrated in Fig. 2a:

$$\mathscr{A}(f) = \begin{matrix} 1 - |f/W|, & |f| \leq W \\ 0, & |f| \geq W. \end{matrix} \tag{2}$$

Alternatively, the block diagram of Fig. 2a may be considered to represent a sampled-data system. An input time function $x(t)$ is convolved with $A(t)$ or equivalently filtered by its Fourier transform $\mathscr{A}(f)$, producing $x_o(t)$. Noise $n(t)$ is added to $x_o(t)$, and their sum is sampled to produce noisy samples $x_o(kT) + n(kT)$, with different noise samples independent. The noisy samples are filtered by $H(f)$ to produce optimum linear estimates of $x(t)$ or of $x_o(t)$. Much of the following discussion will be carried out for this sampled-data system, which is equivalent to the measurement of a one-dimensional incoherent field by a one-dimensional antenna.

It remains to specify the spectra of the input signal $x(t)$ and the noise $n(t)$ in Fig. 2a. We have assumed the radio brightness $x(t)$ varies rapidly with respect to the antenna beamwidth, i.e., with respect to the width of $A(t)$; this would arise, for example, from a random distribution of point sources, dense compared to the antenna beamwidth. Since power is positive, brightness is also positive, and consequently, $\langle x(t) \rangle \geq 0$. We assume that the dc component of $x(t)$ is deterministic (see Appendix A), and we treat it separately. The ac component of $x(t)$ has power spectrum $P_x(f)$ essentially white, i.e., wide compared to the bandwidth W of $\mathscr{A}(f)$; we denote the spectral density of $P_x(f)$ within this band $|f| < W$ by X.

Finally, the noise spectrum $P_n(f)$ of Fig. 2a—white, bandlimited, with spectral density NT in the band $|f| < 0.5/T$—will yield independent noise samples $n(kT)$ with power N, as we assumed above.

Consider the block diagram of Fig. 2a as a sampled-data system. The stationary input $x(t)$ is filtered by $\mathscr{A}(f)$, which is bandlimited to W as indicated, producing the quantity $x_o(t)$. Since $P_x(f)$, the power spectrum of $x(t)$, is white with spectral density X within this band, $x_o(t)$ at the output of $\mathscr{A}(f)$ will have power spectrum

$$P_{x_o}(f) = |\mathscr{A}(f)|^2 P_x(f) \approx X \cdot |\mathscr{A}(f)|^2. \tag{3}$$

Noise is added, and the noisy filtered output is sampled at interval T; the noise samples are independent, with power N. Finally, a reconstruction filter $H(f)$ yields an output signal $x_r(t)$ and noise $n_r(t)$.

All frequency components of the original input $x(t)$ outside the band W, $|f| > W$ have been lost. Therefore, we can only estimate the bandlimited version of $x(t)$ (Fig. 2b), i.e.:

$$x_W(t) = x(t) \circledast 2W \frac{\sin 2\pi Wt}{2\pi W_t}. \tag{4}$$

Alternatively, we might wish to estimate $x_o(t)$, at the output of the filter $\mathscr{A}(f)$ of Fig. 2a. Consequently, we define the following two errors:

$$e_W(t) \equiv x_r(t) - x_W(t).$$
$$e_o(t) \equiv x_r(t) - x_o(t). \tag{5}$$

The quantities $e(t)$, which represent the errors in the absence of noise, consist of linear distortion plus aliasing. Noting that $x(t)$ and $n(t)$, and hence $x_r(t)$ and $n_r(t)$, are independent, the output error $e(t)$ is independent of the output noise $n_r(t)$. Consequently, the total mean-square deviation between desired and actual output is given in each of the two cases as follows:

$$\overline{\langle d_W^2(t)\rangle} \equiv \overline{\langle [y(t) - x_W(t)]^2\rangle} = \overline{\langle e_W^2(t)\rangle} + \overline{\langle n_r^2(t)\rangle}.$$
$$\overline{\langle d_o^2(t)\rangle} \equiv \overline{\langle [y(t) - x_o(t)]^2\rangle} = \overline{\langle e_o^2(t)\rangle} + \overline{\langle n_r^2(t)\rangle}. \tag{6}$$

Here the symbols $\langle \quad \rangle$ indicate an ensemble average and the symbol $\overline{\qquad\quad}$ indicates a time average over one sampling interval T, in the sampled-data model of Fig. 2. The quantities in (6) will depend on $H(f)$, the transfer function of the reconstruction filter. We might want to minimize either of them; the transfer functions that do so are denoted $H_W(f)$ and $H_o(f)$, respectively. By analogy to the terminology used for the two-dimensional antenna problem in Section I,[2] we call the estimation of $x_o(t)$ "interpolation", and the estimation of $x_W(t)$ "restoration" in the present one-dimensional problem.

We distinguish two cases:

$$WT < 0.5; \quad \text{oversampled}$$
$$WT > 0.5; \quad \text{undersampled.}$$

In the undersampled case the $\overline{\langle e^2(t)\rangle}$ of (6) comprise both linear distortion and aliasing; in the oversampled case aliasing is absent. The $\overline{\langle n_r^2(t)\rangle}$ of (6) arise from noise in both cases.

The oversampled case is the simplest. Here aliasing is absent, as we noted above. Both the linear distortion and, as we see below, the output noise are stationary. Consequently, we may drop the time-averaging symbols throughout (6).

In the undersampled case both the aliasing and the output noise are nonstationary. While the time-averaged quantities in (6) are easy to compute, we need in addition these quantities as functions of time, i.e., with the time-averaging symbols in (6) removed.

The optimum filters $H_W(f)$ and $H_o(f)$, which minimize the mean-square deviations in (6), are Wiener least-square-error filters. We show for the undersampled case that Wiener filters minimize not only the time-averaged mean-square deviations, but also the time-dependent mean-square deviations. The spectra of linear distortion, aliasing, and noise are all simply additive.

Such filters operate on all the data samples $x_o(kT) + n(kT)$, $-\infty < k < \infty$. In practice we will estimate $x_W(t)$ or $x_o(t)$ from a finite number of data samples. Here it is no longer possible to separate the contributions of linear distortion and aliasing; the errors are nonstationary in both the undersampled and oversampled cases. We defer treatment of this problem to a future paper.

We treat these various cases below as an introduction to the two-dimensional antenna case. We take the following quantities as given:

X	$\lim_{f \to 0} P_x(f)$, $f \neq 0$, low-frequency limit of continuous component of spectral density of $x(t)$.
$\langle x(t) \rangle$	expected value of $x(t)$.
T	sampling interval.
N	mean-square sample noise, independent for different samples.
$\mathscr{A}(f)$	input filter transfer function: $\mathscr{A}(0) = 1$; $\mathscr{A}(f) = 0$, $\lvert f \rvert \geq W$.

We assume

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} P_x(f) df = \langle x(t) \rangle^2. \tag{7}$$

III. GENERAL FILTERS

The signal and noise outputs in Fig. 2a are given in terms of the measured sample values as follows:

$$x_r(t) = T \sum_{k=-\infty}^{\infty} x_o(kT) h(t - kT)$$

$$n_r(t) = T \sum_{k=-\infty}^{\infty} n(kT) h(t - kT). \tag{8}$$

The weight function, $h(t)$, is the impulse response of the reconstruction filter of Fig. 2, i.e., the Fourier transform of $H(f)$. The output $y(t)$ may be intended as an estimate either of $x_W(t)$ or of $x_o(t)$ of Fig. 2, with errors $e_W(t)$ or $e_o(t)$, respectively, (5), and noise $n_r(t)$.

The power spectra of these errors are given, respectively, as follows:

$$P_{e_W}(f) = \lvert H(f) \rvert^2 \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} P_{x_o}\left(f - \frac{n}{T}\right) + \lvert 1 - H(f)\mathscr{A}(f) \rvert^2 P_{x_W}(f)$$

$$P_{e_o}(f) = \lvert H(f) \rvert^2 \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} P_{x_o}\left(f - \frac{n}{T}\right) + \lvert 1 - H(f) \rvert^2 P_{x_o}(f). \tag{9}$$

Here $P_{x_o}(f)$ and $P_{xW}(f)$ are related by (3) and (4):

$$P_{x_o}(f) = |\mathscr{A}(f)|^2 \cdot P_{xW}(f); \qquad P_{xW}(f) = \begin{matrix} X, & |f| < W \\ 0, & |f| \geq W. \end{matrix} \tag{10}$$

The first terms of (9) represent aliasing; the second terms represent linear distortion of the signal. The mean-square errors are[3,4]

$$\overline{\langle e^2(t) \rangle} = \langle \overline{e^2(t)} \rangle = \int_{-\infty}^{\infty} P_e(f) df, \tag{11}$$

where $e(t)$ represents either $e_W(t)$ or $e_o(t)$. The time average, indicated by $\overline{}$ in (11), may be taken over any integral number of sampling intervals T, in view of the stationarity of $x(t)$.

The output noise power spectrum is

$$P_{n_r}(f) = NT \cdot |H(f)|^2. \tag{12}$$

The mean-square output noise is

$$\overline{\langle n_r^2(t) \rangle} = \langle \overline{n_r^2(t)} \rangle = NT \int_{-\infty}^{\infty} |H(f)|^2 df, \tag{13}$$

where again the time average $\overline{}$ may be taken over any integral number of sampling intervals T.

$\mathscr{A}(f)$ is strictly bandlimited to $|f| < W$; then from (3) or (10)

$$P_{x_o}(f) = 0, \qquad |f| \geq W. \tag{14}$$

The output $y(t)$ contains only alias and noise components outside the band $|f| \geq W$; the reconstruction filter transfer function should certainly be zero there. Consequently, we assume

$$H(f) = 0, \qquad |f| \geq W \tag{15}$$

throughout the remainder of this paper.

By (14) if the sampling is fast enough, $WT < 0.5$, alias and signal components are separated in (9). Since $H(f)$ satisfies (15), aliasing is absent in the output; both $x_r(t)$ and $n_r(t)$ are stationary, and we may drop the time averages $\overline{}$ in (11) and (13). Thus in the oversampled case:

$$\left.\begin{aligned} \langle e_W^2(t) \rangle &= \int_{-W}^{W} |1 - H(f)\mathscr{A}(f)|^2 P_x(f) df \\ \langle e_o^2(t) &= \int_{-W}^{W} |1 - H(f)|^2 P_{x_o}(f) df \\ \langle n_r^2(t) \rangle &= NT \int_{-W}^{W} |H(f)|^2 df \end{aligned}\right\} W < \frac{1}{2T}. \tag{16}$$

In the undersampled case the expected values on the left-hand side of (16) become periodic functions of time. We make the additional ad hoc assumption that

$$W < \frac{1}{T}; \tag{17}$$

then from (14) no more than two terms overlap in (9). This assumption is appropriate for the antenna problem described in Sections I and II. Then from Appendix B:

$$\begin{aligned}
\langle e_W^2(t)\rangle &= 2\int_0^W \left| e^{-j2\pi\frac{t}{T}} H\left(f - \frac{1}{T}\right)\mathscr{A}(f) - 1 + H(f)\mathscr{A}(f)\right|^2 \\
&\quad \cdot P_{x_W}(f)df \\
\langle e_o^2(t)\rangle &= 2\int_0^W \left| e^{-j2\pi\frac{t}{T}} H\left(f - \frac{1}{T}\right) - 1 + H(f)\right|^2 P_{x_o}(f)df \\
\langle n_r^2(t)\rangle &= 2NT\int_0^{\frac{1}{2T}} \left| e^{-j2\pi\frac{t}{T}} H\left(f - \frac{1}{T}\right) + H(f)\right|^2 df.
\end{aligned} \tag{18}$$

Time averaging (18) subject to (15) yields directly the results obtained from (9) and (11) and from (13). This is readily seen by substituting

$$\overline{|e^{-j2\pi\frac{t}{T}}P + Q|^2} = |P|^2 + |Q|^2 \tag{19}$$

into (18), with appropriate choices for P and Q. Finally, the results (18) simplify when $H(f)$ and $\mathscr{A}(f)$ are real. This special case is significant because, as we show in Section IV, the optimum filter $H_o(f)$ is always real, and the optimum filter $H_W(f)$ is real for real $\mathscr{A}(f)$, i.e., for antenna illumination with zero phase error. These simplifications are obtained by substituting

$$|e^{-j2\pi\frac{t}{T}}P + Q|^2 = P^2 + Q^2 + 2PQ\cos 2\pi\frac{t}{T}, \quad P \text{ and } Q \text{ real}, \tag{20}$$

into (18).

IV. OPTIMUM FILTERS

The results of optimum linear mean-square filter theory are summarized as follows. Figure 3 shows an input signal $x_i(t)$ filtered by an input filter with transfer function $\mathscr{A}(f)$, to yield a measured signal $x_o(t)$. Noise $\nu(t)$ is added to $x_o(t)$. We obtain linear least-mean-square

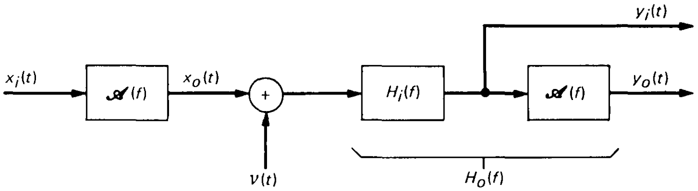


Fig. 3—Optimum linear mean-square estimation.

estimates for the input signal $x_i(t)$ and the measured signal $x_o(t)$ by filtering $x_o(t) + \nu(t)$ by $H_i(f)$ and $H_o(f)$, respectively, as shown in Fig. 3. The transfer functions of these Wiener filters are given as follows:[5]

$$H_i(f) = \frac{1}{\mathscr{A}(f) + \dfrac{P_\nu(f)}{\mathscr{A}^*(f)P_{x_i}(f)}} \tag{21}$$

$$H_o(f) = \frac{P_{x_o}(f)}{P_{x_o}(f) + P_\nu(f)} = \mathscr{A}(f)H_i(f). \tag{22}$$

Here $P_{x_i}(f)$, $P_{x_o}(f)$, and $P_\nu(f)$ are the power spectra of the input signal $x_i(t)$, the measured signal $x_o(t)$ at the output of the filter $\mathscr{A}(f)$, and of the additive noise $\nu(t)$, respectively. Note that

$$P_{x_o}(f) = |\mathscr{A}(f)|^2 P_{x_i}(f) \tag{23}$$

has been used to obtain the right-hand relation of (22).

Define the deviations of the estimates $y_i(t)$ and $y_o(t)$ of Fig. 3 from their desired values as follows:

$$\begin{aligned} d_i(t) &\equiv y_i(t) - x_i(t) \\ d_o(t) &\equiv y_o(t) - x_o(t). \end{aligned} \tag{24}$$

The power spectra of these deviations are minimized by the optimum filters of (21) and (22), as follows:

$$\begin{aligned} P_{d_i}(f) &= \frac{P_{x_i}(f)P_\nu(f)}{|\mathscr{A}(f)|^2 P_{x_i}(f) + P_\nu(f)} \\ P_{d_o}(f) &= \frac{P_{x_o}(f)P_\nu(f)}{P_{x_o}(f) + P_\nu(f)}. \end{aligned} \tag{25}$$

The corresponding minimum mean-square deviations are obtained by integrating these spectra.

These results are applied to minimize the mean-square deviations (6) by the following substitutions in (21) through (25):

$$
\begin{aligned}
x_i(t) &\rightarrow x_W(t)\\
d_i(t) &\rightarrow d_W(t)\\
P_{x_i}(f) &\rightarrow P_{x_W}(f)\\
H_i(f) &\rightarrow H_W(f)\\
P_\nu(f) &\rightarrow \sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty} P_{x_o}\left(f-\frac{n}{T}\right)+NT.
\end{aligned}
\tag{26}
$$

Thus the noise $\nu(t)$ of Fig. 3 is replaced by the noise plus alias spectra of Fig. 2.[3,6] Recall that $\mathscr{A}(f) = 0$, $|f| \geq W$, and that $P_{x_o}(f)$ and $P_{x_W}(f)$ satisfy (10). Consequently, $H_W(f)$ and $H_o(f)$ both satisfy (15). The additional condition (17) eliminates all but the $n = \pm 1$ terms in the summation of the last line of (26). We summarize these results for interpolation:

$$
H_o(f) = \frac{|\mathscr{A}(f)|^2}{|\mathscr{A}(f)|^2 + \left|\mathscr{A}\left(f-\frac{1}{T}\right)\right|^2 + \left|\mathscr{A}\left(f+\frac{1}{T}\right)\right|^2 + \frac{NT}{X}}
\tag{27}
$$

$$
P_{d_o}(f) = H_o(f)\cdot X\left[\left|\mathscr{A}\left(f-\frac{1}{T}\right)\right|^2 + \left|\mathscr{A}\left(f+\frac{1}{T}\right)\right|^2 + \frac{NT}{X}\right]
\tag{28}
$$

and for restoration:

$$
H_W(f) = \frac{1}{\mathscr{A}(f)} H_o(f)
\tag{29}
$$

$$
\begin{aligned}
P_{d_W}(f) &= \frac{P_{d_o}(f)}{|\mathscr{A}(f)|^2}\\
&= \frac{X\left[\left|\mathscr{A}\left(f-\frac{1}{T}\right)\right|^2 + \left|\mathscr{A}\left(f+\frac{1}{T}\right)\right|^2 + \frac{NT}{X}\right]}{|\mathscr{A}(f)|^2 + \left|\mathscr{A}\left(f-\frac{1}{T}\right)\right|^2 + \left|\mathscr{A}\left(f+\frac{1}{T}\right)\right|^2 + \frac{NT}{X}}.
\end{aligned}
\tag{30}
$$

By (17), the terms $\left|\mathscr{A}\left(f-\frac{1}{T}\right)\right|^2$ and $\left|\mathscr{A}\left(f+\frac{1}{T}\right)\right|^2$ in (27), (28), and (30) never overlap. As $|f| \rightarrow W$, since $\mathscr{A}(f) \rightarrow 0$, then $H_o(f)$, $H_W(f)$, and $P_{d_o}(f)$ all → 0 but $P_{d_W}(f) \rightarrow X$.

V. UNIFORM ILLUMINATION (MAXIMUM-GAIN ANTENNA)

From (2)

$$
\mathscr{A}(f) = \begin{array}{ll} 1 - |f/W|, & |f| \leq W \\ 0, & |f| \geq W. \end{array}
\tag{31}
$$

We distinguish two cases:

$$\begin{aligned} &0 < WT < 0.5; &&\text{oversampled} \\ &0.5 < WT < 1; &&\text{undersampled.} \end{aligned} \tag{32}$$

It is convenient to introduce an auxiliary parameter $S = \langle x_o^2(t) \rangle$ representing the total power of the quantity $x_o(t)$ at the output of the filter $\mathscr{A}(f)$ (Fig. 2); from (10)

$$S = X \int_{-W}^{W} |\mathscr{A}(f)|^2 df. \tag{33}$$

For the present maximum-gain antenna, (31) substituted into (33) yields

$$S = \frac{2}{3} XW. \tag{34}$$

Then with the substitution of (34), (27) through (30) yield, for interpolation:

$$\begin{aligned} \overline{\langle d_o^2(t) \rangle} = {} & 2S \frac{WT}{S/N} \int_0^{\min\left(1, \frac{1}{WT} - 1\right)} \frac{(1-y)^2}{(1-y)^2 + \frac{2}{3} \frac{WT}{S/N}} dy \\ & + 3S \int_{\min\left(1, \frac{1}{WT} - 1\right)}^{1} \frac{(1-y)^2 \left[\left(1 + y - \frac{1}{WT}\right)^2 + \frac{2}{3} \frac{WT}{S/N} \right]}{(1-y)^2 + \left(1 + y - \frac{1}{WT}\right)^2 + \frac{2}{3} \frac{WT}{S/N}} dy, \end{aligned} \tag{35}$$

and for restoration:

$$\begin{aligned} \overline{\langle d_W^2(t) \rangle} = {} & 2S \frac{WT}{S/N} \int_0^{\min\left(1, \frac{1}{WT} - 1\right)} \frac{1}{(1-y)^2 + \frac{2}{3} \frac{WT}{S/N}} dy \\ & + 3S \int_{\min\left(1, \frac{1}{WT} - 1\right)}^{1} \frac{\left(1 + y - \frac{1}{WT}\right)^2 + \frac{2}{3} \frac{WT}{S/N}}{(1-y)^2 + \left(1 + y - \frac{1}{WT}\right)^2 + \frac{2}{3} \frac{WT}{S/N}} dy. \end{aligned} \tag{36}$$

Here we combine the two cases of (32). S/N is the observed signal-to-noise power of the samples in Fig. 2, and WT is the parameter bounded by (32), indicating the relative degree of over- or undersampling and consequent aliasing. The first terms of (35) and (36) consist of noise and distortion. In the oversampled case the second terms are zero; in the undersampled case the second terms contain aliasing as well.

While it is possible to evaluate (35) and (36) in closed form, the results are messy. Moreover, such evaluations for other more realistic antenna patterns than the present uniform illumination (e.g., the truncated Gaussian illumination considered in the following section) will be worse in this regard. Consequently, these integrals have been evaluated numerically, with results given in Figs. 4 and 5, showing the average deviation power versus sampling parameter for interpolation and for restoration, respectively, with observed signal-to-noise ratio as a parameter. We note that the deviation is worse for restoration than for interpolation, because in the former case we attempt to equalize the input filter $\mathscr{A}(f)$, thereby enhancing the noise. Alternatively, Figs. 4 and 5 may be obtained by direct numerical integration of (18), using (6), (19), (27), and (29). Analytical expressions for these results in the oversampled case, $0 < WT < 0.5$, are given in eqs. (77) and (78) of Appendix C.

For the undersampled case, $WT > 0.5$, we require in addition the ac component of the deviation powers. We write

$$\underset{W}{\langle d_o^2(t)\rangle} \equiv \overline{\underset{W}{\langle d_o^2(t)\rangle}} - \underset{W}{D_o} \cos 2\pi \frac{t}{T}, \tag{37}$$

where D_o and D_W are the amplitudes of the ac components of the deviation powers for interpolation and for restoration, respectively.

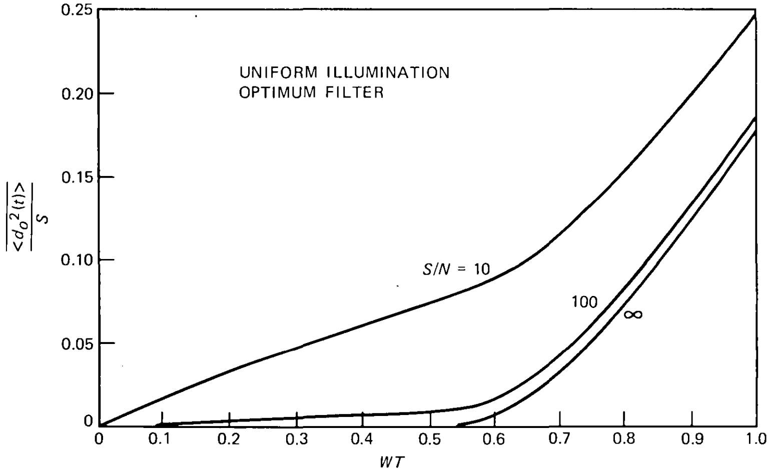


Fig. 4—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: uniform illumination and optimum filter.

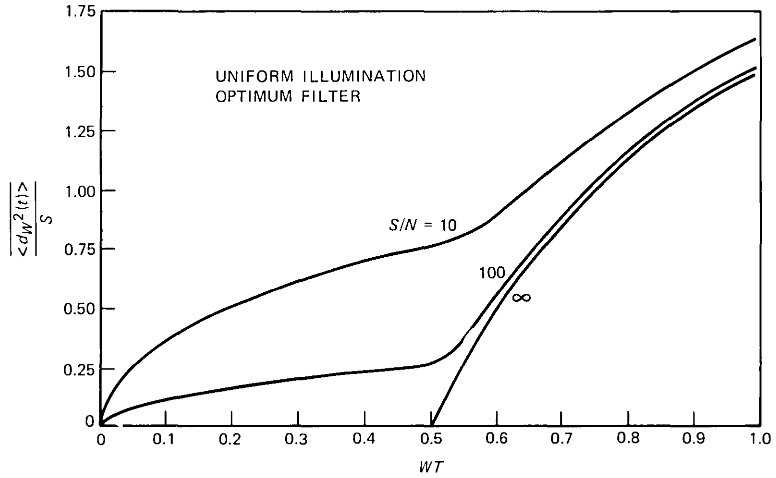


Fig. 5—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for restoration of a bandlimited function: uniform illumination and optimum filter.

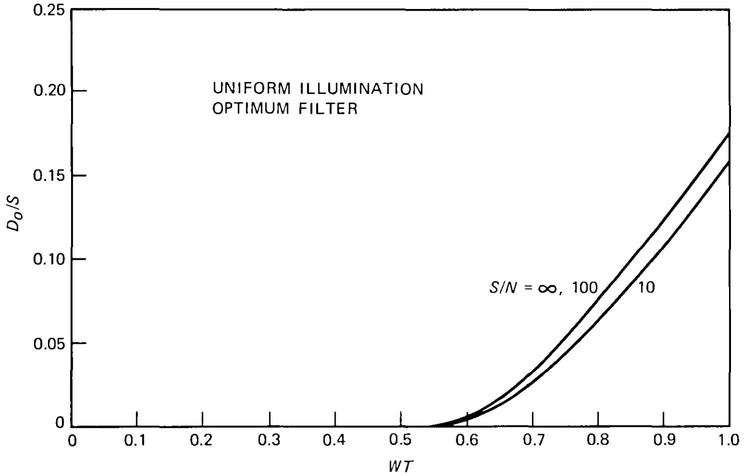


Fig. 6—Diagram of ac deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: uniform illumination and optimum filter.

We determine D_o and D_W from this definition (37) by combining (6), (18) through (20), (27), and (29). Numerical integration of the resulting expressions yields the results shown in Figs. 6 and 7.

A partial check on these results is obtained by observing that for zero noise optimum interpolation must recover the filtered input with zero error at the sample points, i.e., $x_o(nT)$ must be reconstructed

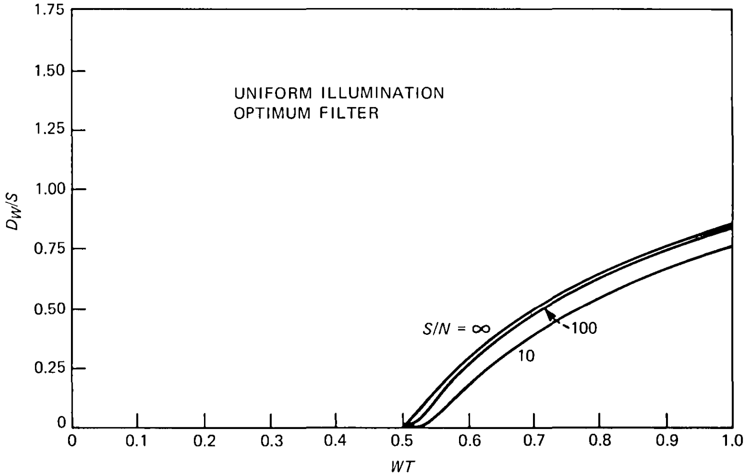


Fig. 7—Diagram of ac deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for restoration of a bandlimited function: uniform illumination and optimum filter.

without error for $N = 0$. Consequently, the $S/N = \infty$ curves of Fig. 4 and Fig. 6 must coincide, and they do.

Observe that while $H_o(f)$ of (27), the optimum filter for interpolation, decreases monotonically as f goes from 0 to W, the same is not true for $H_W(f)$. In contrast $H_W(f)$, the optimum filter for restoration, initially increases to a maximum before finally dropping to zero at $f = W$. For example, consider the oversampled case, $WT < 0.5$. Then

$$H_W(f) = \frac{1 - \frac{f}{W}}{\left(1 - \frac{f}{W}\right)^2 + \frac{2}{3}\frac{WT}{S/N}}. \tag{38}$$

Obviously,

$$H_W(0) = \frac{1}{1 + \frac{2}{3}\frac{WT}{S/N}}; \qquad H_W(W) = 0. \tag{39}$$

The peak is given by

$$H_W(f)\,|_{\max} = 0.5\sqrt{\frac{3}{2}\frac{S/N}{WT}} \quad \text{at} \quad \frac{f}{W} = 1 - \sqrt{\frac{2}{3}\frac{WT}{S/N}}. \tag{40}$$

For large S/N the peak of $H_W(f)$ is very large, and very close to $f = W$. However, the results of Fig. 5 and 7 remain finite as $S/N \rightarrow \infty$.

D_o and D_W of (37) contain contributions from the second and first equations of (18), respectively, and from the third equation of (18), used together with (20). The former consist of linear distortion and aliasing, the latter of noise. All such contributions (to the ac coefficients D) are 0 for the oversampled case, $WT < 0.5$. For the undersampled case, $0.5 < WT < 1$, distortion and aliasing make positive contributions to D_o and D_W, while noise makes negative contributions to these quantities. Taking note of the negative sign on the last term in (37), the deviation due to distortion and aliasing is worst between the sample points, while the deviation due to noise is worst at the sample points.

The deviation powers in Figs. 4 through 7 have been normalized to the signal power at the output of $\mathscr{A}(f)$ in Fig. 2a, $S \equiv \langle x_o^2(t) \rangle$, given for the present case by (34). This is perfectly appropriate for Figs. 4 and 6 ("interpolation"), where we wish to reconstruct the quantity $x_o(t)$. It is less appropriate for restoration, where we wish to reconstruct $x_W(t)$, the bandlimited version of the input $x(t)$ in Fig. 2a, given by (4). Here a more appropriate normalization might be the total power of $x_W(t)$. For the present case (31), (10) and (34) yield

$$\langle x_W^2(t) \rangle = 2XW = 3S. \tag{41}$$

To normalize Figs. 5 and 7 ("restoration") in this way, simply divide the numbers on the vertical axes by 3 and relabel these axes accordingly.

The parameter S/N of Figs. 4 through 7 is appropriate for both interpolation and restoration, since it is the signal-to-noise ratio observed at the sampled output of a radiometer used to measure incoherent fields. However, observe that S/N defined here, and used throughout the remainder of this paper, is different than the conventional signal-to-noise ratio at the output of a radiometer receiver. In the present work S is proportional to the *fluctuation* in the radiometer signal output as the antenna is scanned across the sky; while the conventional radiometer signal output is taken as the average signal output as the antenna is scanned across the sky. As we noted in Section II, we assume the average radio brightness is deterministic, and we treat it separately (see Appendix A).

VI. GAUSSIAN ILLUMINATION

Let a one-dimensional antenna of width W have an aperture field that is Gaussian:

$$E(x) = \begin{cases} e^{-0.2 \ln 10 \cdot d \cdot \left(\frac{x}{W}\right)^2}, & |x| < \dfrac{W}{2} \\ 0, & |x| > \dfrac{W}{2}. \end{cases} \tag{42}$$

The field at the edge of the aperture is d dB down from the maximum field (at the center of the aperture);

$$d = -20 \log_{10} E\left(\frac{W}{2}\right). \tag{43}$$

The symbol d represents the aperture "taper".

$\mathscr{A}(f)$ of (3) and Fig. 2, the Fourier transform of the effective width $A(t)$ of the antenna, is proportional to the convolution of $E(f)$ of (42) with itself; by (1)

$$\mathscr{A}(f) = \frac{\displaystyle\int_{\max\left(-\frac{W}{2},\, -\frac{W}{2}+f\right)}^{\min\left(\frac{W}{2},\, \frac{W}{2}+f\right)} E(x)E(x-f)dx}{\displaystyle\int_{-\frac{W}{2}}^{\frac{W}{2}} E^2(x)dx}. \tag{44}$$

For $d = 0$, (42) substituted into (44) yields (31) of the preceding section for uniform illumination.

Figures 8 through 11 give the average and ac deviation powers versus sampling parameter with signal-to-noise ratio as a parameter, for interpolation and for restoration, for Gaussian aperture illumination with taper d = 15 dB. These results may be compared with Figs. 4 through 7, respectively, for uniform illumination, treated in the preceding section. As in this prior case, numerical integration seems

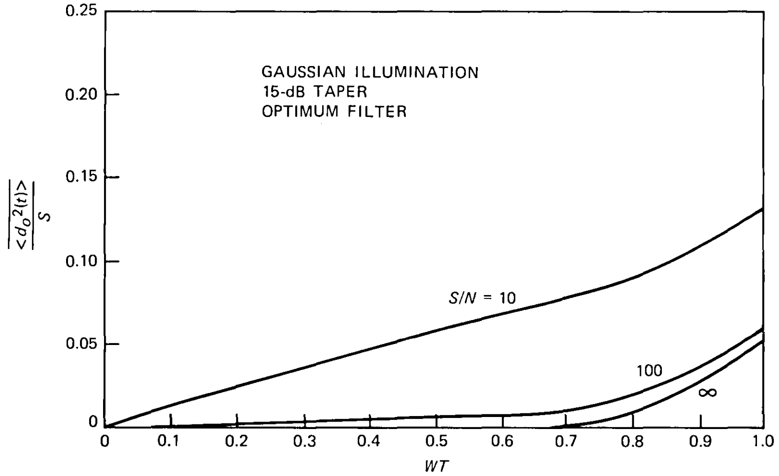


Fig. 8—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: Gaussian illumination with 15-dB taper, optimum filter.

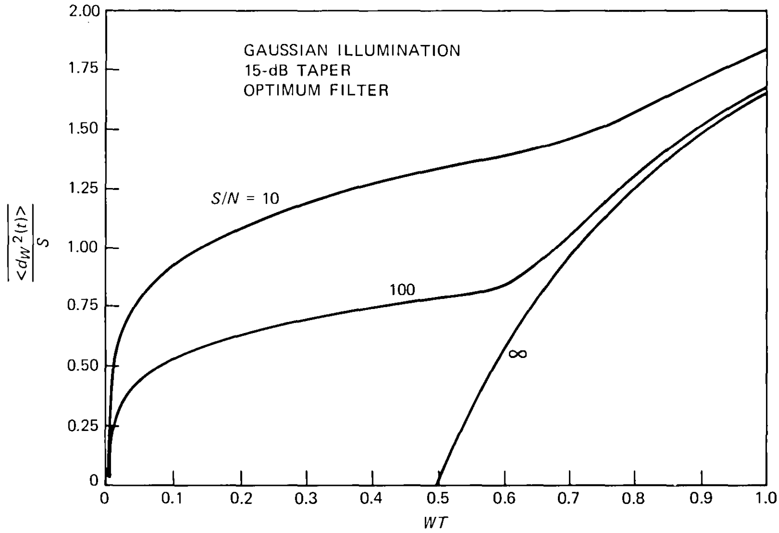


Fig. 9—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for restoration of a bandlimited function: Gaussian illumination with 15-dB taper, optimum filter.

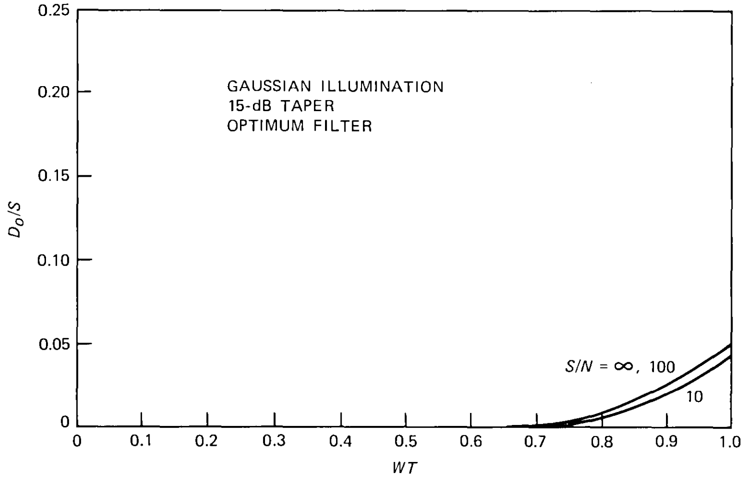


Fig. 10—Diagram of ac deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: Gaussian illumination with 15-dB taper, optimum filter.

preferable to analytical treatment. Equation (44) is evaluated using (42), the results substituted into (27), (29), and (18), and finally (6) and (37) are evaluated using (19) and (20).

Much of the discussion of the preceding section for uniform illumination applies also to the present Gaussian case. For zero noise the

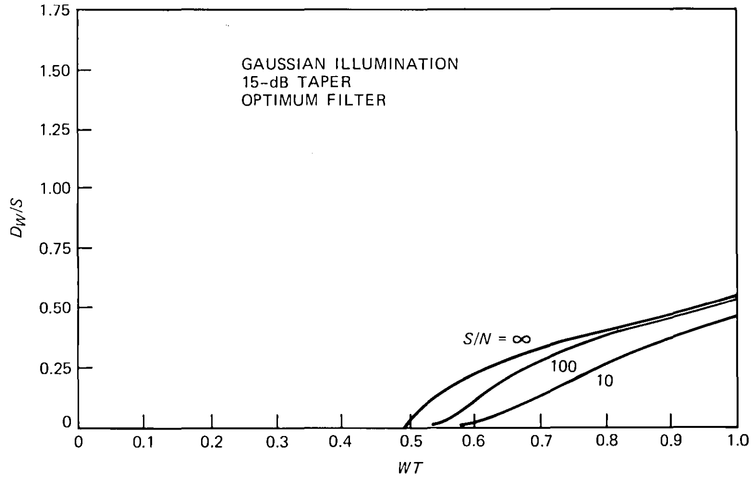


Fig. 11—Diagram of ac deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for restoration of a bandlimited function: Gaussian illumination with 15-dB taper, optimum filter.

$S/N = \infty$ curves of Figs. 8 and 10 again coincide. To normalize the "restoration" results to the total power of $x_W(t)$ (rather than to $S \equiv \langle x_o^2(t) \rangle$), (10) and (42), (44) yield

$$\langle x_W^2(t) \rangle = 2xW = 3.281S, \qquad d = 15 \text{ dB}; \tag{45}$$

simply divide the numbers on the vertical axes of Figs. 9 and 11 by 3.281 and relabel these axes accordingly.

An important difference between the results for uniform illumination (discussed in the preceding section) and the present results for the more practical Gaussian illumination with 15-dB taper appears in the "restoration" results of Figs. 5 and 9 for these two cases, respectively. Restoration is much more difficult in the present case because the input filter $\mathscr{A}(f)$ in Fig. 2 falls off much faster with f, requiring greater equalization by the output filter and hence yielding more output noise.

VII. SUBOPTIMUM FILTERS

We have so far considered only optimum reconstruction filters, according to Section IV, for both interpolation and for restoration. Such filters must be changed for each different signal-to-noise ratio S/N. We now explore the use of *fixed* filters, for interpolation and for restoration, which are independent of S/N.

Consider interpolation first. The optimum filter $H_o(f)$ of (27) is well behaved and, in particular, changes only slightly as the noise power N

increases from zero. It is natural to use $H_o(f)$ for $N = 0$ as a suboptimum filter for finite but small N, i.e., for large but finite S/N.

The situation for restoration is quite different. The optimum filter $H_W(f)$ of (29) is badly behaved, as we discussed in connection with (38) through (40); small changes in S/N can produce large changes in $H_W(f)$ near its peak. For the oversampled case, $0 < WT < 0.5$, $H_W(f)$ for $N = 0$ has a pole at $f = W$, and this filter therefore yields infinite output noise for finite S/N. Consequently, use of $H_W(f)$ for $N = 0$ for finite S/N is restricted to the undersampled case, $0.5 < WT < 1$.

Figures 12 through 15 show the average deviation power for interpolation and for restoration, for uniform illumination and for Gaussian illumination with a 15-dB taper. In all cases the suboptimum filter used is the optimum filter for $S/N = \infty$. For restoration, only the undersampled case, $0.5 < WT < 1$, is shown, as we discussed above.

Observe that the curves of Figs. 12 and 14 (suboptimum interpolation, for uniform and for Gaussian illumination, respectively) are identical in the oversampled case, $0 < WT < 0.5$. Here the reconstruction filter is simply

$$H(f) = \begin{matrix} 1, & |f| < W \\ 0, & |f| > W. \end{matrix} \tag{46}$$

Thus, since aliasing is absent, in the absence of noise the suboptimum

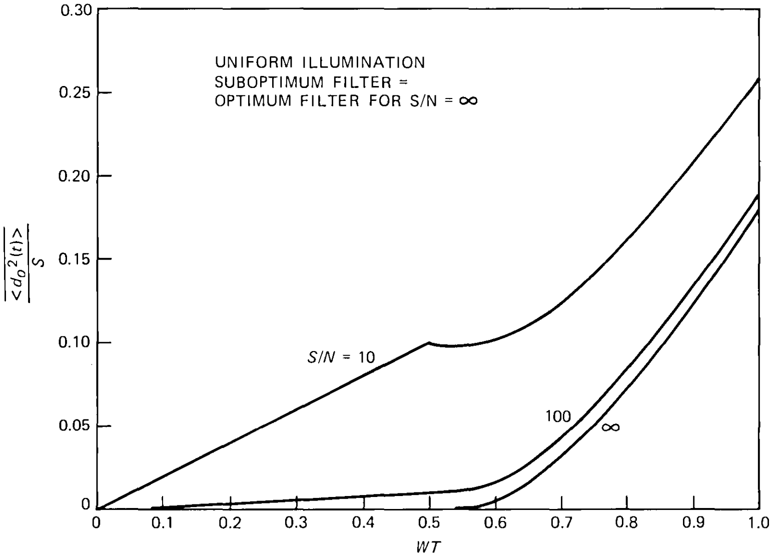


Fig. 12—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: uniform illumination, suboptimum filter equal to optimum filters for $S/N = \infty$.

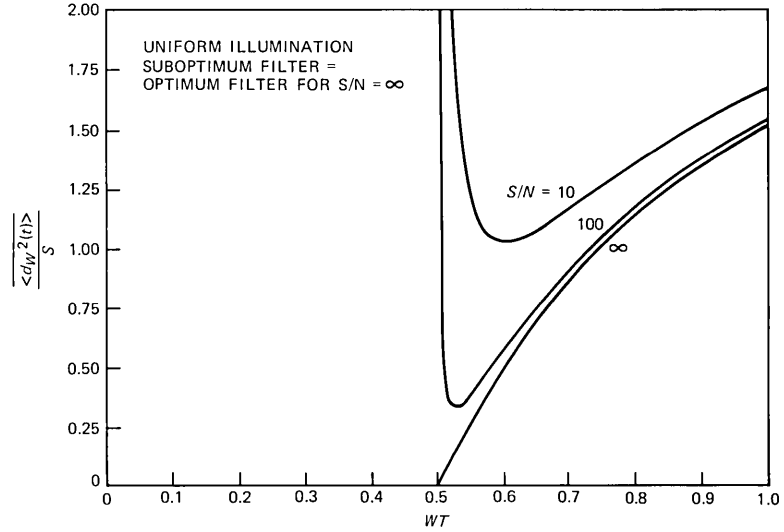


Fig. 13—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for restoration of a bandlimited function: uniform illumination, suboptimum filter equal to the optimum filter for $S/N = \infty$.

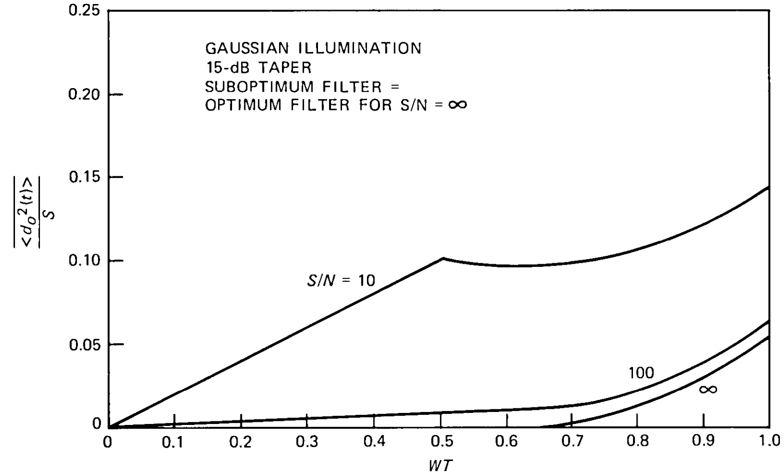


Fig. 14—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter, for interpolation of a bandlimited function: Gaussian illumination with 15-dB taper, suboptimum filter equal to optimum filter for $S/N = \infty$.

(i.e., zero noise) reconstruction filter simply passes the input without distortion. The output deviation then results from noise alone. A simple result is given in (79), Appendix C, for this case. We can observe further comparing (77) and (79) that the $S/N = 10$ curves in Figs. 4 and 12 start out with identical slope for small WT, but that as

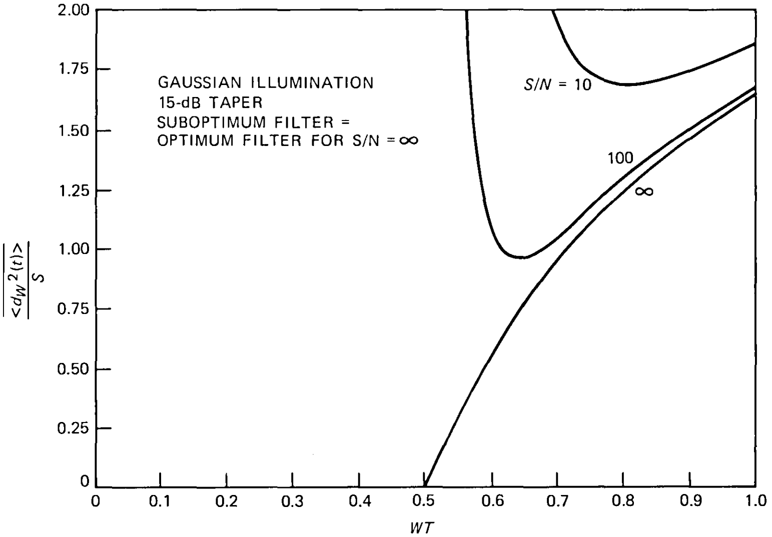


Fig. 15—Average deviation power versus sampling parameter, with signal-to-noise ratio as a parameter for restoration of a bandlimited function: Gaussian illumination with 15-dB taper, suboptimum filter equal to optimum filter for $S/N = \infty$.

WT increases the curve of Fig. 4, for the optimum interpolation filter, falls below the curve of Fig. 12, for the suboptimum interpolation filter.

We conclude that for interpolation with large S/N, we may use the reconstruction filter designed for $S/N = \infty$ with little loss. This is *not* true for restoration. These results are to be expected from the discussion of the optimum filters given above in these two cases. The ac deviation powers for interpolation do not change much from those of Figs. 6 and 10, so we omit plots of them here.

For interpolation with smaller S/N, the greatest penalty for the suboptimum filter occurs near critical sampling, $WT = 0.5$. For example, compare Figs. 4 and 8 with Figs. 12 and 14, respectively, for $S/N = 10$; the suboptimum filter is about 35 percent worse for uniform illumination and about 67 percent worse for Gaussian illumination with a 15-dB taper than the optimum filters for these cases.

VIII. DISCUSSION

The present results show how to process data obtained by measurement of a one-dimensional incoherent field with a one-dimensional antenna, and what the resulting errors will be. We assume the present results provide some indication for the real, two-dimensional case.

In the model of Fig. 2, T is the sampling interval and N is the noise power. By the definitions in Section II, T corresponds to the angular separation between antenna observations, and N to the mean-square error in receiver output due to receiver noise. The noise of a radiometer receiver is inversely proportional to the observation time;[1] moreover, if a given time is allotted to measure a given region of the sky, the time per observation is inversely proportional to the number of observations, or directly proportional to the angular separation between observations. Consequently,

$$NT = \frac{\text{Constant}}{\text{Observing Time per Unit Angle of Sky}}, \tag{47}$$

where the constant in (47) is independent of the antenna.

Let us examine the data of Figs. 4, 5, 8, 9, 12, or 14 for fixed observing time per unit angle, i.e., from (47) for NT = constant. Consider as a specific example Fig. 4. Take the point given by $S/N = 100$, $WT = 1$; then $\overline{\langle d_o^2(t)\rangle}/S \approx 0.187$. The measurement parameters $S/N = 10$, $WT = 0.1$ yield the same value of NT, and hence by (47) the same observing time per unit angle of sky; for these parameters $\overline{\langle d_o^2(t)\rangle}/S \approx 0.017$. In this example, sampling ten times as often with 1/10 the signal-to-noise ratio has reduced the mean-square deviation by a factor of 0.187/0.017 ≈ 11. More generally, for average deviation in any of the above figures, compare:

1. Any $S/N = 10$ curve.
2. The corresponding $S/N = 100$ curve with its horizontal scale compressed by a factor of 10, i.e., each point x, y replotted to $x/10$, y.

In every case the compressed $S/N = 100$ curve will lie above the $S/N = 10$ curve for $WT > 0.05$. For $0 \leq WT \leq 0.05$ the two curves coincide precisely, as we see in the second paragraph of Appendix C.

From this, we conclude that undersampling is *always* bad; it will always be better to reduce T and increase N, i.e., to take more closely spaced observations each with poorer signal-to-noise ratio, to avoid operating in the undersampled region. Moreover, in the preferred oversampled region the family of curves in each of these figures can be combined into a single curve; i.e., for $0 \leq WT \leq 0.5$ the mean-square deviation $\overline{\langle d^2(t)\rangle}/S$ is a function of the single normalized variable $[(S/N)/(2WT)]$. These results are in agreement with a prior observation of R. W. Wilson.

Recalling from Section VII that Figs. 12 and 14 are identical for $0 \leq WT \leq 0.5$, only five curves are required to summarize the data of Figs. 4, 5, 8, 9, 12, and 14 in the oversampled region. This is done in Figs. 16 and 17, for interpolation and for restoration, respectively. Note that the vertical axis of Fig. 16 has been normalized to $S \equiv \langle x_o^2(t)\rangle$ of (33), the same as Figs. 4, 8, 12, and 14. However, the vertical axis of Fig. 17 has rather been normalized to $\langle x_W^2(t)\rangle$, as discussed in

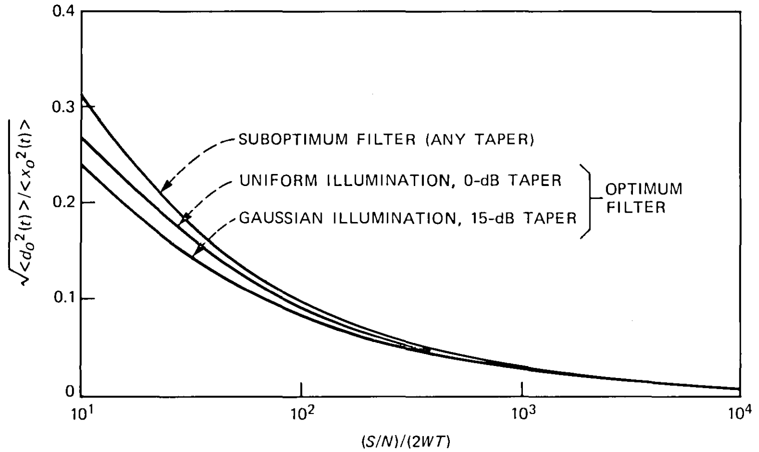


Fig. 16—Root-mean-square deviation for interpolation versus signal-to-noise ratio in the oversampled case, $0 \leq WT \leq 0.5$.

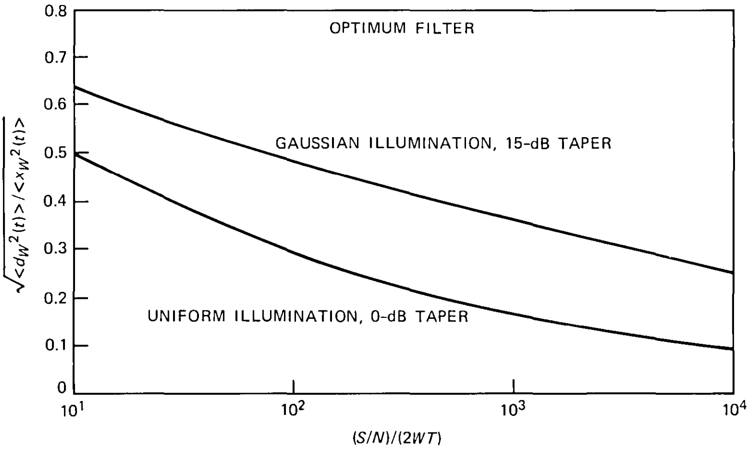


Fig. 17—Root-mean-square deviation for restoration versus signal-to-noise ratio in the oversampled case, $0 \leq WT \leq 0.5$.

(41) and (45), i.e., different than the normalization for Figs. 5 and 9. For a given antenna and a given receiver noise temperature, the total observing time for a given area of sky is proportional to the product of the area observed and the horizontal axis variable $[(S/N)/(2WT)]$ of Figs. 16 and 17; however, observe from (33) and (47) that the constant of proportionality depends on the antenna illumination, and hence differs for the uniform and for the Gaussian cases.

Finally, we observe that while undersampling ($WT > 0.5$) is bad, oversampling ($WT < 0.5$) offers no advantage over critical sampling

($WT = 0.5$). Reducing WT below 0.5 requires more data storage, with no reduction in error.

Let us examine the implications of these results for ordinary antennas. Consider the antenna of Section VI, with Gaussian illumination (42) and significant taper (e.g., $d = 15$ dB, as in the examples). To determine the gross structure of the main lobe we may neglect the truncation of the aperture field in (42), i.e., assume $E(x)$ is given by the top expression of (42) for all x. Then the effective width is the Fourier transform of (44) with infinite limits on the integrals:

$$A(t) \approx W\sqrt{\frac{10\pi}{d \ln 10}}\, e^{-\frac{10}{d \ln 10}(\pi Wt)^2}, \qquad Wt < \sqrt{\frac{d}{10}}; \qquad d \gg 1. \qquad (48)$$

Observations are frequently taken at an angular separation of one full 3-dB beamwidth; i.e., the receiving power patterns at two adjacent observations overlap at their 3-dB points. In our present model this corresponds to a sampling interval

$$T = 2t_{3\text{-dB}}, \qquad (49)$$

where $t_{3\text{-dB}}$ is the 3-dB half-width of the antenna pattern (48);

$$A(t_{3\text{-dB}}) = \frac{1}{2}\cdot A(0). \qquad (50)$$

From (48) through (50)

$$W\cdot 2t_{3\text{-dB}} = \frac{1}{\pi}\sqrt{\frac{d \ln 10 \ln 2}{10}}. \qquad (51)$$

For the 15-dB taper chosen for the examples, (51) yields

$$W\cdot 2t_{3\text{-dB}} = 0.985, \qquad d = 15\text{-dB taper}. \qquad (52)$$

Thus, measurements separated by a full 3-dB beamwidth with a 15-dB antenna taper will be undersampled by about a factor of 2, with corresponding penalties indicated in Figs. 8 and 9.

Table I illustrates the above discussion. The same antenna size and

Table I—Numerical examples

				Normalized rms Deviations					
				Interpolation			Restoration		
WT	$10 \log_{10} S/N$ (dB)	d (dB)	Normalized Observing Time	Avg.	Max.	Min.	Avg.	Max.	Min.
1.0	20	15	1	0.24	0.33	0.095	0.72	0.83	0.60
0.5	17	15	1	0.12	0.12	0.12	0.53	0.53	0.53
0.5	57	15	10,000				0.12	0.12	0.12
0.5	36	0	73				0.12	0.12	0.12

receiver noise figure are assumed for the four cases shown. A typical antenna, with Gaussian illumination and a 15-dB taper, is used for the first three examples. The final example uses a maximum-gain antenna, with uniform illumination (no taper). The root-mean-square (rms) deviations have been normalized to $\langle x_o^2(t) \rangle = S$ for interpolation (33) and to $\langle x_W^2(t) \rangle = 2XW$ for restoration (41), (45). In each case, from (47)

$$\text{Observing Time} = \frac{\text{Constant}}{NT} = \text{Constant} \cdot \frac{W}{S} \cdot \frac{S/N}{WT}, \qquad (53)$$

with the third factor determined from the first and second columns of the above table, and S in the denominator of the second factor given by (45) for the first three examples and by (41) for the last example. Finally, the observing times are normalized such that the first two examples have normalized observing times equal to unity.

In the top row observations are taken at twice critical separation (i.e., observations separated by about a full 3-dB beamwidth). Since the data are undersampled, the rms deviations vary, being minimum at the observation points and maximum half-way in between. The normalized deviation is much larger for restoration than for interpolation.

The second row shows the same antenna and receiver with critical sampling. Twice as many observations are made, each for half the time, with the signal-to-noise ratio reduced by 3 dB; hence, the total observing time for a given area of the sky is the same as for the first row. The rms deviations are now independent of position with respect to the observation points. The rms interpolation deviation is about half as large, and the rms restoration deviation is about three-fourths as large, as the corresponding average deviations for the undersampled case, in row 1. The deviation for restoration remains much larger than that for interpolation.

The third row shows the same antenna and receiver as the first two rows, with a 40-dB higher signal-to-noise ratio than the second row; consequently, the total observing time is increased by a factor of 10,000. The deviations are greatly reduced, that for restoration now being equal to the interpolation deviation for the preceding case of row 2.

Finally, the fourth row shows that a maximum-gain antenna, with uniform illumination (no taper), and the same receiver as that of row 3, will attain the same rms restoration deviation with a total observing time of only 73, as compared to 10,000 for an antenna with Gaussian illumination and a 15-dB taper (row 3). Of course, the observing time is still large compared to that of row 2; i.e., to obtain the same rms deviation for restoration with an antenna with uniform illumination as for interpolation with an antenna with a 15-dB taper takes 73 times as long.

The above discussion makes undersampling appear unattractive. Nevertheless, much existing data have been taken in this way; the present results show the best way to process such data, and the resulting errors. The optimum filters (27) and (29) in the undersampled case minimize the time-average mean-square deviations $\overline{\langle d^2(t) \rangle}$ in the present model (6). It is shown in Appendix D that they also minimize the time-dependent mean-square deviations $\langle d^2(t) \rangle$. In the corresponding antenna problem the optimum data spatial filters minimize the mean-square error everywhere in the region containing the observed points.

IX. CONCLUSIONS

Consider astronomical measurements of radio brightness that has white spatial variation (i.e., that varies rapidly compared to the beamwidth of the antenna used to make the measurement), made at a regular array of points in the sky. Optimum mean-square estimates for the measured and true brightness at any point in the sky are called "interpolation" and "restoration", respectively. Idealize this problem to one dimension. Then:

1. Data points should be separated by about half the (full) 3-dB beamwidth for a normal antenna, having a tapered aperture illumination; i.e., the receiving power patterns at two adjacent observations should overlap at their 0.75-dB points. This is often not done.

2. Interpolation can be accomplished with reasonable accuracy and reasonable observation time.

3. Restoration with reasonable accuracy requires much longer observation time than interpolation.

4. Optimum spatial filters depend on the signal-to-noise ratio. For interpolation this dependence is very weak. The optimum filter for zero noise works fairly well for interpolation of finite signal-to-noise ratios; this is not true for restoration.

5. A maximum-gain antenna, with uniform aperture illumination, is better for restoration than a conventional antenna with tapered aperture illumination.

The following additional studies are suggested by the present work:

1. Interpolation and restoration with a finite number of data points, perhaps not regularly spaced (e.g., edge effects).

2. Tolerances in the spatial filters applied to the measured data, and in the antenna illumination.

3. Interpolation and restoration with reduced resolution, by additional spatial filtering.

4. Nonwhite sky brightness statistics, e.g., strong isolated point sources embedded in white brightness.

5. Treatment of the real two-dimensional problem. Greater variety is evident, e.g., square, hexagonal, and irregular sampling patterns are of interest in two dimensions.

X. ACKNOWLEDGMENT

I would like to thank R. W. Wilson for suggesting this problem and for many helpful discussions, T. S. Chu for suggesting the study of restoration with reduced resolution, L. J. Greenstein for helpful discussions of the equivalent sampled-data system and Ref. 3, R. E. Hills for suggesting the study of the maximum-gain antenna, and M. J. Gans, M. Kavehrad, and J. Minkoff for careful reviews of this paper.

REFERENCES

1. H. E. Rowe, "Processing Channel-Bank Spectrometer Data," AT&T Bell Lab. Tech. J., *63*, No. 4 (April 1984), pp. 565–85.
2. R. N. Bracewell, "Two-Dimensional Aerial Smoothing in Radio Astronomy," Australian J. Phys., *9* (September 1956), pp. 297–314.
3. K. W. Cattermole, *Principles of Pulse Code Modulation*, New York: American Elsevier, 1969, pp. 44–64.
4. H. E. Rowe, *Signals and Noise in Communications Systems*, New York: Van Nostrand Reinhold, 1965, p. 244.
5. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965, pp. 400–7.
6. J. W. Brault and O. R. White, "The Analysis and Restoration of Astronomical Data via the Fast Fourier Transform," Astron. Astrophys., *13* (July 1971), pp. 169–89 (see pp. 183–6).
7. Ref. 4, pp. 41–3.
8. Ref. 5, pp. 390–400.

APPENDIX A

Condition for DC Component To Be Deterministic

A real wide-sense stationary random process has mean $\langle x(t) \rangle$ and covariance $\phi_x(\tau) \equiv \langle x(t+\tau)x(t) \rangle$, both independent of t. We assume

$$\lim_{|\tau|\to\infty} \phi_x(\tau) = a^2. \tag{54}$$

Then define $\phi_{x_c}(\tau)$ by the relation

$$\phi_x(\tau) \equiv a^2 + \phi_{x_c}(\tau). \tag{55}$$

Thus,

$$\lim_{|\tau|\to\infty} \phi_{x_c}(\tau) = 0. \tag{56}$$

The spectral density $P_x(f)$, the Fourier transform of $\phi_x(\tau)$, is by (55)

$$P_x(f) = a^2\delta(f) + P_{x_c}(f), \tag{57}$$

where $P_{x_c}(f)$ is the Fourier transform of $\phi_{x_c}(\tau)$ of (55). By (56), $P_{x_c}(f)$ contains no component proportional to $\delta(f)$, i.e., $P_{x_c}(f)$ contains no delta function at the origin. The dc power of $x(t)$ is thus a^2.

Now define

$$\overline{x(t)} \equiv \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt \tag{58}$$

as the dc component of an individual noise wave $x(t)$. We have

$$\langle \overline{x(t)} \rangle = \langle x(t) \rangle. \tag{59}$$

Next,

$$\langle \overline{x(t)}^2 \rangle = \lim_{T\to\infty} \frac{1}{(2T)^2} \int_{-T}^{T} \int_{-T}^{T} \phi_x(t-s)dt\, ds = a^2. \tag{60}$$

Define $x_{\mathrm{ac}}(t)$ by the relation

$$x(t) \equiv \overline{x(t)} + x_{\mathrm{ac}}(t). \tag{61}$$

Then

$$\langle x_{\mathrm{ac}}(t) \rangle = 0, \tag{62}$$

$$\phi_x(\tau) = a^2 + \phi_{x_{\mathrm{ac}}}(\tau). \tag{63}$$

If we compare (55) and (63),

$$\phi_{x_{\mathrm{ac}}}(\tau) = \phi_{x_c}(\tau). \tag{64}$$

We investigate the conditions under which $\overline{x(t)} = \langle x(t) \rangle$ with probability 1, i.e., for which almost every $x(t)$ has the same dc component. Define

$$y \equiv \overline{x(t)} - \langle x(t) \rangle. \tag{65}$$

Obviously $\langle y \rangle = 0$. Now,

$$\langle y^2 \rangle = \langle \overline{x(t)}^2 \rangle - \langle x(t) \rangle^2 = a^2 - \langle x(t) \rangle^2, \tag{66}$$

the last step following from (60).

Therefore, if

$$\lim_{|\tau|\to\infty} \phi_x(\tau) = \langle x(t) \rangle^2, \tag{67}$$

then with probability 1

$$\overline{x(t)} \equiv \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt = \langle x(t) \rangle. \tag{68}$$

The dc component may then be treated as a fixed quantity, and the ac component [whose spectrum is the second term of (57)] estimated separately. A stationary shot noise, for example, satisfies (67). Note that in (67)

$$\lim_{|\tau|\to\infty} \phi_x(\tau) = \lim_{\epsilon\to o} \int_{-\epsilon}^{\epsilon} P_x(f)df. \qquad (69)$$

APPENDIX B

Derivation Of Time-Varying Deviations

The time average of the expected error powers (5) and noise power are given immediately in terms of their power spectra, (9) and (12), by (11) and (13). However, the time-varying error and noise powers (18) are not so easily obtained. All three quantities of (18) are derived in a similar manner; we choose the middle one, $\langle e_o^2(t)\rangle$, as representative.

Use the Fourier series representation for the random process $x_o(t)$, as follows:[7]

$$x_o(t) = \sum_{|n|<N} x_{on}e^{jn2\pi f_0 t} \qquad (70)$$

$$N \equiv \frac{W}{f_0} \qquad (71)$$

$$\lim_{f_0\to 0} \frac{1}{f_0} \langle x_{on}x_{om}^*\rangle = 0, \qquad n \neq m. \qquad (72)$$

$$\lim_{f_0\to 0} \frac{1}{f_0} \langle |x_{on}|^2\rangle = P_{x_o}(nf_0). \qquad (73)$$

The restriction on the summation in (70), with N defined in (71), arises from (14). From Fig. 2, the error $e_o(t)$ of (5) is given by

$$\begin{aligned} e_o(t) = &\sum_{|n|<N} (H_n - 1)x_{on}e^{jn2\pi f_0 t} \\ &+ e^{j2\pi\frac{t}{T}} \sum_{-N<n<0} H_{n+K}x_{on}e^{jn2\pi f_0 t} \\ &+ e^{-j2\pi\frac{t}{T}} \sum_{0<n<N} H_{n-K}x_{on}e^{jn2\pi f_0 t}, \end{aligned} \qquad (74)$$

where

$$H_n \equiv H(nf_0), \qquad K \equiv \frac{1}{f_0T}. \qquad (75)$$

The restrictions on the summations in (74) arise from (14), (15), and

(17). The first line of (74) represents linear distortion; the second and third lines represent aliasing. If we combine terms in (74),

$$e_o(t) = \sum_{-N<n<0} [e^{j2\pi \frac{t}{T}} H_{n+K} - 1 + H_n] x_{on} e^{jn2\pi f_0 t}$$

$$+ \sum_{0<n<N} [e^{-j2\pi \frac{t}{T}} H_{n-K} - 1 + H_n] x_{on} e^{jn2\pi f_0 t}, \qquad (76)$$

where we ignore the $n = 0$ term by the assumption that any dc component is deterministic and treated separately. Calculating $\langle e_o^2(t) \rangle$ using (72) and (73), we obtain the middle relation of (18). The other two relations are similarly obtained.

APPENDIX C

Some Analytical Results

Consider the results (35) and (36) for a one-dimensional antenna with uniform illumination in the oversampled case, $0 < WT < 0.5$. Here min $(1, 1/(WT) - 1) = 1$; aliasing is absent, the upper limit of the first terms in (35) and (36) is 1, and the second terms in these relations are absent. Then these results are readily evaluated by partial fraction expansions of the integrands to yield the following results for optimum reconstruction filters, for interpolation and for restoration, respectively:

$$\langle d_o^2(t) \rangle = 2S \frac{WT}{S/N} \left[1 - \sqrt{\frac{2}{3} \frac{WT}{S/N}} \tan^{-1} \sqrt{\frac{3}{2} \frac{S/N}{WT}} \right] \qquad (77)$$

and

$$\langle d_W^2(t) \rangle = 3S \sqrt{\frac{2}{3} \frac{WT}{S/N}} \tan^{-1} \sqrt{\frac{3}{2} \frac{S/N}{WT}}. \qquad (78)$$

These relations give the curves of Figs. 4 and 5 for $0 < WT < 0.5$. We omit similar, but messier, results for $0.5 < WT < 1$. Observe that the normalized deviations (77) and (78) are functions of only $(WT)/(S/N)$. In the oversampled case we can reduce the sampling interval and signal-to-noise ratio proportionately; i.e., in Figs. 4 and 5 for $0 < WT < 0.5$, the points with equal y-coordinates on the $S/N = 100$ and $S/N = 10$ curves have x-coordinates WT whose ratio is precisely 10. This simple property does not hold if either x-coordinate $WT > 0.5$.

We have no such analytic results for Gaussian illumination, Figs. 8 and 9. However, R. W. Wilson has pointed out that in the absence of aliasing, halving the sampling interval and doubling the sample noise power must leave the final result unaltered. As a result, in the oversampled case, $0 < WT < 0.5$, the normalized deviations must be functions of only $(WT)/(S/N)$; in particular, the $S/N = 100$ and $S/N = 10$ curves of Figs. 8 and 9 for $0 < WT < 0.5$ must scale in the same way as described above for Figs. 4 and 5. This is most readily seen from present results by observing that for $0 < WT < 0.5$ all terms $\mathscr{A}(f \pm 1/T)$ in (27) through (30) disappear, and the noise and sampling interval appear only as the product NT. Thus the deviations remain unchanged as long as NT is fixed, i.e., if the number of samples and the signal-to-noise ratio are multiplied by the same factor.

Finally, we have an extremely simple result for the suboptimum filter for interpolation in the oversampled case, for both uniform and Gaussian illumination. Here $H(f)$ is given by (46). There is neither aliasing nor linear distortion, and substituting (46) into the third relation of (18) yields

$$\langle d_o^2(t) \rangle = S \cdot \frac{2WT}{S/N}. \tag{79}$$

The linear relation (80) yields the curves of Figs. 12 and 14 for $0 < WT < 0.5$. Again the normalized deviation depends only on $(WT)/(S/N)$.

APPENDIX D

Optimum Filters Minimize Time-Dependent Mean-Square Deviations

The optimum linear estimator for $x_o(t)$ of Fig. 2 (i.e., "interpolation") may be written as[8]

$$y(t) = \sum_k a_k(t)[x_o(kT) + n(kT)], \tag{80}$$

where $x_o(kT)$ and $n(kT)$ are the signal and noise samples of (8); the coefficients a_k, necessarily functions of time, are selected to minimize the mean-square deviation $d_o^2(t)$, where

$$d_o(t) \equiv y(t) - x_o(t). \tag{81}$$

The $a_k(t)$ satisfy[8]

$$\langle \{x_o(t) - \sum_k a_k(t)[x_o(kT) + n(kT)]\} \cdot \{x_o(iT) + n(iT)\} \rangle = 0. \tag{82}$$

Equation (82) yields

$$\phi_{x_o}(t - iT) - \sum_k a_k(t) \phi_{x_o}((k - i)T) + N \cdot a_i(t) = 0, \tag{83}$$

where N is the power of the (independent) noise samples (Fig. 2) and $\phi_{x_o}(\tau)$ is the covariance of $x_o(t)$, i.e., the Fourier transform of $P_{x_o}(f)$ of (1).

The index i and the summation index k in eqs. (82) and (83) range over the set of samples used to estimate $x_o(t)$. While this set may be finite, the present work assumes an infinite number of samples are used. Then it is immediately obvious by the stationarity of $x_o(t)$ that

$$a_k(t) = a_0(t - kT) \equiv a(t - kT), \tag{84}$$

where we drop the subscript 0 as unnecessary. Substituting (84) into (80), the optimum estimator for $x_o(t)$ is

$$y(t) = \sum_{k=-\infty}^{\infty} [x_o(kT) + n(kT)]a(t - kT), \tag{85}$$

where $a(t)$ is given by

$$\phi_{x_o}(t) = \sum_{k=-\infty}^{\infty} \phi_{x_o}(kT)a(t - kT) + N\cdot a(t). \tag{86}$$

If we Fourier transform (86), use the Poisson sum formula, and solve for the transform of $a(t)$,

$$A(f) = \frac{P_{x_o}(f)}{\dfrac{1}{T}\displaystyle\sum_{k=-\infty}^{\infty} P_{x_o}\left(f - \frac{k}{T}\right) + N}. \tag{87}$$

Then, if we substitute (3) into (87), comparing (8) with (80), we identify $a(t)$ with $T\cdot h(t)$, and hence replace $A(f)$ by T times $H_o(f)$ of (27), to yield

$$H_o(f) = \frac{|\mathscr{A}(f)|^2}{\displaystyle\sum_{k=-\infty}^{\infty} \left|\mathscr{A}\left(f - \frac{k}{T}\right)\right|^2 + \frac{NT}{X}}. \tag{88}$$

Finally, imposing the constraint (17) and recalling that $\mathscr{A}(f)$ is strictly bandlimited to $|f| < W$ [Fig. 2 or (3) and (14)], (88) becomes identical to (27).

We recall that in the undersampled case ($0.5 < WT < 1$) the optimum interpolation filter $H_0(f)$ of (27) minimized the time-average mean-square deviation $\overline{\langle d_0^2(t)\rangle}$ of (6). The present appendix shows that $H_o(f)$ also minimizes the time-dependent mean-square deviation $\langle d_o^2(t)\rangle$ for all t.

A similar discussion may be given for restoration; the optimum filter $H_W(t)$ of (29), which minimizes $\overline{\langle d_W^2(t)\rangle}$, also minimizes $\langle d_W^2(t)\rangle$ for all t.

AUTHOR

Harrison E. Rowe, B.S., M.S., Sc.D. (Electrical Engineering), The Massachusetts Institute of Technology, 1948, 1950, 1952; U.S. Navy, 1945–1946, AT&T Bell Laboratories, 1952–1984; Stevens Institute of Technology, 1984—.

Mr. Rowe was a Member of Technical Staff in the Radio Research Laboratory prior to his retirement in 1984. Presently he is Anson Wood Burchard Professor of Electrical Engineering at Stevens. His publications include numerous papers and one textbook, spanning a variety of fields including parametric amplifiers, noise and communication theory, propagation in random media, and related problems in waveguide, radio, and optical communication systems. He is the joint author of five patents. Fellow, IEEE; corecipient, 1977 David Sarnoff Award and 1972 Microwave Prize; member, Commission C of URSI, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 9, November 1984
Printed in U.S.A.

1982/83 End Office Connection Study: ASPEN Data Acquisition System and Sampling Plan

By J. D. HEALY,* M. LAMPELL,* D. G. LEEPER,* T. C. REDMAN,* and E. J. VLACICH*

(Manuscript received December 19, 1983)

To characterize the transmission performance of the public switched network, the Bell System conducted an intensive field measurement study of end-office-to-end-office connections from October 1982 to January 1983. A special multistage sampling plan and ASPEN, a flexible, automatic data acquisition system based on the *UNIX*™ operating system, were developed to evaluate the more than 10,000 transmission paths included in the study. This paper describes both the ASPEN measurement system and the sampling plan. A companion paper in this issue of the *AT&T Bell Laboratories Technical Journal* describes the network transmission performance results.

I. INTRODUCTION AND SUMMARY

The Bell System traditionally conducted field measurement studies of network transmission performance to evaluate existing administrative and maintenance procedures and to set objectives for new transmission systems and equipment.[1-11] The 1982/83 End Office Connection Study (EOCS) had additional goals arising from Computer Inquiry II and the 1984 divestiture of the Bell operating companies from AT&T. Specifically, the EOCS was intended to support allocation of performance objectives across segments of the network; to aid planners of new telecommunications systems, services, and terminal equipment;

* AT&T Bell Laboratories; present affiliation Bell Communications Research, Inc.

and to serve as a benchmark against which new system arrangements and services might be compared. The EOCS and allied studies now provide a database to support work on both the traditional and divestiture-related goals.

To conduct the EOCS, a software-controlled field measurement system called ASPEN (Automatic System for Performance Evaluation of the Network) was developed, a special sampling plan was prepared, and software tools were adapted to allow for rapid screening and analysis of the measurement data. This section presents an overview of ASPEN and the EOCS sampling plan. The accompanying paper describes the study results in detail, along with a description of the data analysis methods.[12]

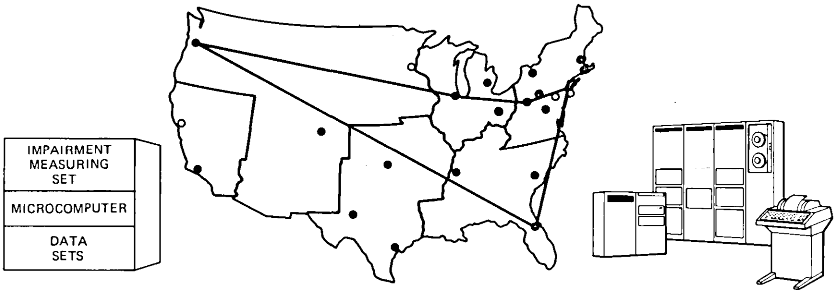
1.1 The ASPEN Data Acquisition System

The most recent field measurement study comparable to the EOCS is the 1969/70 Connection Survey.[8] In that study, manually operated measurement equipment was carried from office to office for a period of one year, and a total of about 600 transmission paths were evaluated. (A given *test connection* offers two directions of transmission for testing, hence two *transmission paths.* In the 1969/70 connection survey, measurements were made on one transmission path per test connection.) Measurement data were entered manually on a terminal connected to a central computer. While this procedure worked well, a highly automated data acquisition system was deemed desirable for the EOCS for two fundamental reasons.

First, setting performance objectives requires accurate information about the tails of performance distributions, including distributions conditioned on end-office switch type, time of day, mileage band, and other criteria. It was estimated that a sample size of 10,000 transmission paths was needed to meet the study goals. The time and cost of manual data acquisition would have been prohibitive.

Second, a modifiable and reusable measurement tool appeared to be the most efficient way to meet future demand for field performance measurements of inter- and intraexchange network segments, as well as a host of planned new services. Progress during the 1970s in automatic measurement equipment and mini/microcomputer hardware and software had made the development of such a tool possible.

As Fig. 1 shows, the ASPEN application to the EOCS consisted of 20 remotely controlled instrumentation packages (called Remote Test Units or RTUs) connected to the line side of mainframes in selected Bell System end offices. Under the control of a host computer, the RTUs called one another over the public switched network just as actual customers would. Once a connection between two RTUs was



REMOTE NODE
CONNECTION SUITABILITY MEASUREMENTS
LOSS, NOISE, ATTENUATION DISTORTION
ENVELOPE DELAY DISTORTION, PHASE JITTER
HITS, DROPOUTS
BIT AND BLOCK ERROR RATE
CONNECTION AVAILABILITY MEASUREMENTS
NETWORK COMPLETION RATE
CUTOFF PROBABILITY

CENTRAL CONTROL/PROC
MEASUREMENT SCHEDULING
NODE CONTROL
DATA COLLECTION/MANAGEMENT
DATA ANALYSIS

Fig. 1—ASPEN System: End Office Connection Study.

established, automatic measurement equipment in the RTUs evaluated more than 25 transmission parameters on that connection (see Table I).

The ASPEN structure offered several advantages over the 1969/70 approach, the most notable of which was speed. Evaluating more than 120 transmission paths per day, the equivalent of the 1969/70 data was gathered in five days as opposed to a full year. In all, over a period extending from October 1982 to January 1983, parameters were measured on more than 10,000 paths from over 7,000 test connections. (In the EOCS, approximately 3000 of the 7000 test connections were evaluated in both directions, while the remaining connections were evaluated in one direction only. Thus the total number of transmission paths evaluated was approximately 2(3000) + 4000 = 10,000.)

A second advantage offered by the ASPEN structure was flexibility. As described in Sections II and III, the sequence of operations at an RTU was controlled by a microprocessor whose software was down loaded from the host computer. The measurement sequence could be, and was, changed as needed.

Finally, because the data screening and preliminary analysis ran concurrently with data acquisition, information was available to make changes in the data acquisition process while the study was still in progress. The analysis tools also made it possible to obtain final results quickly, despite the much larger volume of data.

A particularly challenging aspect of ASPEN development was the preparation of host computer software that could handle fault conditions gracefully. The ASPEN host computer for the EOCS was a

Table I—1982/83 End office connection study parameters measured

Voice and Voiceband Data	Insertion loss (1 kHz)
	Loss vs. frequency (30 freqs)
	C-message noise
	Call cutoffs
	Propagation delay
Voiceband Data	Signal-to-noise ratio
	C-notched noise
	3-kHz flat noise
	3-kHz notched noise
	Noise-to-ground
	Envelope delay vs. frequency (30 freqs)
	Peak-to-average ratio (P/AR)
	2nd-order intermodulation distortion
	3rd-order intermodulation distortion
	Phase jitter (2-300 Hz)
	Phase jitter (20-300 Hz)
	Amplitude jitter (2-300 Hz)
	Amplitude jitter (20-300 Hz)
	Impulse noise (6 thresholds)
	Gain hits (3 thresholds)
	Phase hits (3 thresholds)
	Dropouts
	Frequency shift
	1200-b/s bit/block error rate
	4800-b/s bit/block error rate

*VAX-11/780** computer running under the *UNIX* operating system. As described in Section III, ASPEN's host computer control software consisted of three distinct layers responsible for call management, connection supervision, and overall connection scheduling. This structure successfully handled the occasional dropped connection, intermittent RTU failure, or host computer service interruption (both scheduled and unscheduled). As a result, ASPEN could run virtually unattended on a daily basis.

In addition to providing a friendly environment for program development, the *UNIX* operating system offered data storage and analysis tools to permit data screening and analysis to run concurrently with data acquisition. A relational database package and the *S* statistical analysis package (created at AT&T Bell Laboratories) were a key part of the ASPEN system and the EOCS. Together with special C language data screening programs, these tools were used to ensure data integrity, to store the data in a compact, logical structure, and to allow rapid exploratory analysis of results. Because the data screening programs accepted data directly from the RTUs, manual handling of measurement data was completely eliminated.

1.2 The end office connection study sampling plan

The sampling plan included both RTU location (spatial) and test

* Trademark of Digital Equipment Corporation.

connection scheduling (temporal) components. As described in Section IV, RTU locations were selected to yield performance data representative of different mileage bands, end office switch technologies (*ESS*™, crossbar, and step-by-step switching equipment), and other strata. A unique aspect of the EOCS spatial sampling plan was that the sampled units, end offices, were quite different from the units being evaluated, namely, telephone calls.

With 20 RTUs deployed, there were 380 possible originating/terminating pairs over which calls could be placed. The EOCS temporal sampling plan, operating through the ASPEN Scheduler software, managed the selection of the RTU pairs. A unique feature of the temporal sampling plan algorithm was its ability to enhance the number of busy-hour connections and to handle gracefully the additions or deletions of RTUs caused by equipment problems.

1.3 Summary

The ASPEN approach to acquisition of network performance data proved highly successful; data were gathered more than 20 times as fast as by previous manually oriented methods. The ASPEN host computer software structure shows how a fault-tolerant automatic performance characterization system may be implemented, and the ASPEN spatial and temporal sampling plans show how locations may be chosen and measurements scheduled to evaluate a network or other telecommunications service.

Sections II and III provide a detailed description of the hardware and software design of ASPEN as it was applied to the EOCS. While the components used in ASPEN are explicitly listed, the hardware description is a generic one because there are many possible choices for RTU and host computer components. The software description in Section III is more explicit. While the actual code is not presented, the structure is described in some detail, with emphasis on fault-handling capabilities.

Section IV describes how the a priori study requirements and the opportunities offered by the fully automated ASPEN approach combined to create special sampling challenges. Also described is an iterative preselection step that made it possible to meet the need for stratification of results by mileage band and end-office switch technology (*ESS*, crossbar, or step-by-step switching equipment).

II. ASPEN HARDWARE SUBSYSTEMS

The ASPEN data acquisition system is composed of two major hardware subsystems: the remote test unit and the host computer. This section describes the hardware components used in each of these subsystems.

2.1 Remote test unit hardware

The Remote Test Unit (RTU) subsystem has five major hardware components: a microcomputer that controls the operation of the RTU, a remotely controlled transmission Impairment Measuring Set (IMS), "test" data sets used in conjunction with a bit and block error rate testing capability, and a switching matrix. Table II lists the specific equipment used for the EOCS, and Fig. 2 shows a complete RTU, ready for shipment and installation.

Figure 3 is a schematic representation of the ASPEN data acquisition system. The host computer and RTU microprocessor communicate by means of 1200-b/s full-duplex data sets operating over an ordinary dial-up connection through the public switched network. Using the switching matrix, the microprocessor connects the line under test, IMS, test data sets, Bit Error Rate Receiver (BERR), and Bit Error Rate Transmitter (BERT) in the various configurations required to measure the parameters listed in Table I.

The IMS measures all the parameters in Table I except bit/block error rate and propagation delay. Selection and testing of the IMS, an especially critical phase of the ASPEN project, was guided by AT&T PUB 41009, a technical reference on the evaluation of transmission impairment measurement equipment. The instrument chosen for ASPEN (see Table II) struck the compromise among performance, cost, and availability that was most appropriate for the EOCS. The

Table II—ASPEN remote test unit components

Transmission impairment measuring set
Hekimian Laboratories Inc. Model 3701 Communications Test System with EIA RS232C Remote Control Option

Microprocessor
Colorado Data Systems 53A Smart Hardware System*
Card Complement:
1. Zilog Z80 microprocessor, two RS232 input/output ports with buffered input and output
2. Three relay cards (ten relays per card)
3. Bit error rate receiver card
4. Bit error rate transmitter card
5. Counter card (four counters per card)

Data modems
1. For host-RTU communications:
   Western Electric 212AR
2. For bit and block error rate measurements:
   Western Electric 212AR
   Racal Vadic 3450
   Western Electric 208B-L1B
   Codex 5208R

Equipment case
Environmental Container Systems, Inc. fibercase enclosure (2 ft × 2 ft × 4 ft) with shock-mounted rack (see Fig. 2).

* Trademark of Colorado Data Systems.

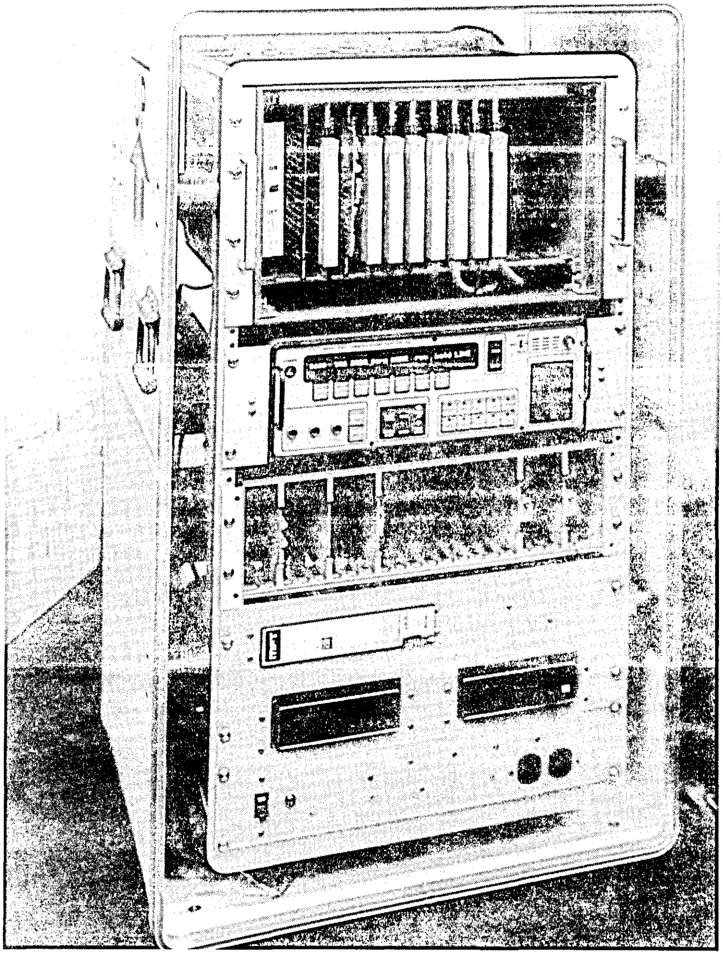


Fig. 2—ASPEN RTU.

RTU instruments were remotely self-checked for accuracy throughout the data-gathering period.

As is common practice, ASPEN measures bit and block error rates by transmitting and receiving a repeated pseudorandom bit stream with the BERT and BERR equipment. The data sets used in the EOCS (see Table II) were widely used in the network and readily available at the time of the study. All data sets were pretested in the laboratory to ensure that none exhibited performance aberrations absent in other sets of the same type.

ASPEN uses an extension of the error rate measurement technique

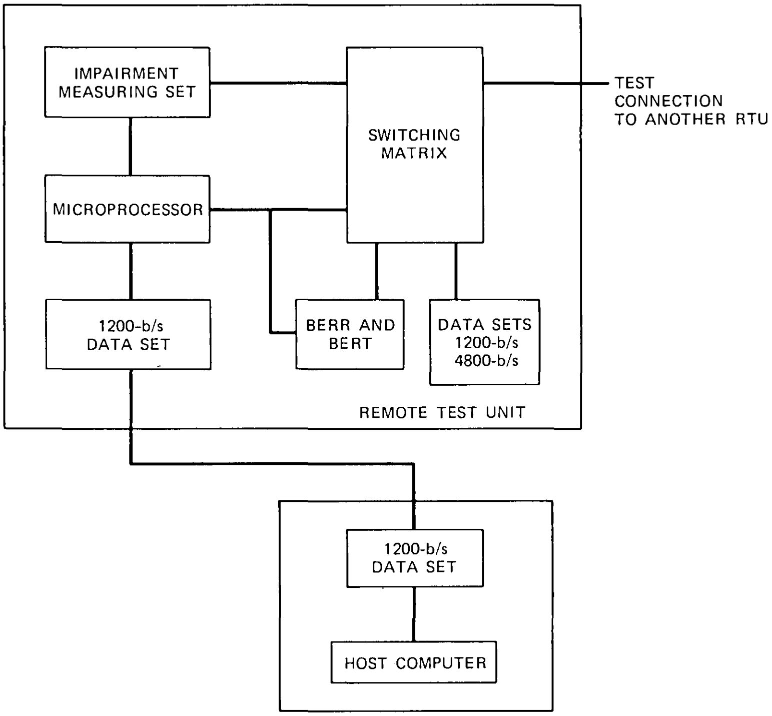


Fig. 3—Schematic representation of the ASPEN instrumentation.

to measure round trip propagation delay. As Fig. 4 shows, an error is deliberately injected into the transmitted bit stream at the near-end BERT. The bit stream containing the error is transmitted by a 1200-b/s full-duplex data set, and the error is detected by the far-end BERR. This event is used to trigger injection of an error into the far-end BERT, which is transmitted by the data sets back to the near end BERR. The elapsed time between near-end error injection at the BERT and detection at the BERR (minus a predetermined constant to allow for data set delays) is the round trip propagation delay.

2.2 Host computer hardware

The ASPEN system host computer consists of a mini- or microcomputer capable of running the *UNIX* operating system. Table III lists the host computer hardware used in the EOCS. The number of input/output (I/O) ports and amount of primary memory (i.e., random access) required depend on how many RTUs are to be controlled simultaneously. In addition, enough secondary memory (i.e., disk subsystems) is required to support the installation of a database large enough to store the data to be collected. If the RTUs are not connected

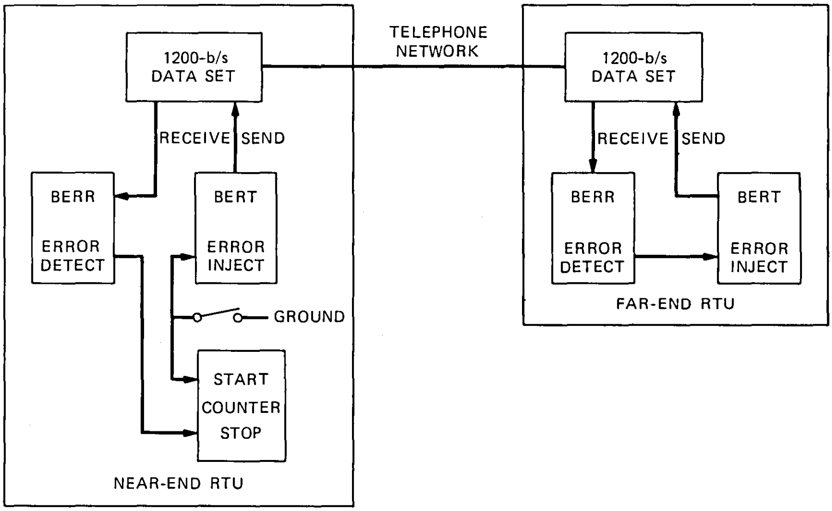


Fig. 4—Round-trip propagation delay measurement.

Table III—ASPEN/end office connection study host hardware

Components	Equipment Used
CPU	*VAX-11/780* with battery backup and *DEC* floating point accelerator
Primary memory	3.75 M-bytes random access memory
Secondary memory	Three *DEC* RM05 removable disk drives, 256M bytes each
Tape backup	One *DEC* TU77 high-speed tape drive
I/O ports	Four *DEC* DZ11 8-channel asynchronous multiplexers, providing 32 ports, assisted by two *DEC* KMC11B auxiliary microprocessors
Dial-out	Four *DEC* DN11-DA Automatic Calling Unit (ACU) Controllers, and four WE 557A ACU-sharing arrangements.

directly to the host computer, Automatic Calling Units (ACUs) or their equivalent must be present to dial up all the RTUs.

III. ASPEN SOFTWARE SUBSYSTEMS

In the early planning stages, a decision was made to utilize the *UNIX* operating system for all aspects of the study. The *UNIX* system provided an excellent software environment for such assorted tasks as formulation and testing of statistical plans, development of ASPEN system software, collection and analysis of data, and generation of reports.

3.1 Remote test unit software subsystem

The RTU software communicates with the host computer while it simultaneously controls the RTU hardware. The RTU contains an operating system that accepts down-loaded programs and commands from the host computer, executes subroutines in these programs, and exchanges data with the host.

The RTU software locally controls the various hardware functions by dividing the program into subroutines. Each subroutine contains an option that allows two or more subroutines to be linked.

In the EOCS the RTUs measured performance parameters of Direct Distance Dialing (DDD) connections. This involved the use of the RTUs in pairs. Because the RTU memory is limited, a separate program was needed for the originating end and the terminating end of the connection. Therefore, two RTUs testing a particular DDD connection had complementary subroutines that were initiated by “start” commands from the host computer. The subroutine timing was designed to be robust enough to accommodate a plus-or-minus 3-second error in the individual starting times of the complementary subroutines.

The complementary subroutine functions used in a typical EOCS sequence are shown in Table IV.

3.2 RTU-host interface

The interaction between the host computer and the RTU is based on the concept of a software module. A *module* is defined as a set of instructions at the host computer that causes execution of a subroutine at the RTU, and passes the RTU information necessary to execute the subroutine. There exists a one-to-one correspondence between

Table IV—Subroutine functions from typical measurement sequence

Originating RTU	Terminating RTU
Dial RTU test connection to far end	Answer
Measure analog parameters	Send test tones
Send data to host	Send data to host
Send test tones	Measure analog parameters
Send data to host	Send data to host
Dial far-end RTU reference connection	Answer
Measure envelope delay	Send test tones
Send test tones	Measure envelope delay
Send data to host	Send data to host
Connect 1200-b/s data set	Connect 1200-b/s data set
Measure propagation delay	Establish error return loop
Send data to host	—
Measure bit/block error rate	Measure bit/block error rate
Send data to host	Send data to host
Connect 4800-b/s data set	Connect 4800-b/s data set
Transmit 4800-b/s data	Measure bit/block error rate
—	Send data to host
Measure bit/block error rate	Transmit 4800-b/s data
Send data to host	—

host computer modules and RTU subroutines. The concept of modules will be covered more thoroughly in a later section.

From the point of view of the host computer, *executing* a module means conversing with the RTU microprocessor, exchanging information with it, and starting the execution of a subroutine at the RTU. Related tasks include sending and retrieving data from the RTU and, in general, executing any of the RTU built-in operating system commands. The RTU microprocessor operating system gives the host immediate feedback about the disposition of a command at the RTU. This feedback consists both of echo checks to ensure the integrity of commands transmitted to the RTU, as well as status checks to make sure that the subroutines have been executed correctly. Whenever a subroutine finishes executing, the RTU generates a status symbol, which verifies that the subroutine is completed.

The host computer also synchronizes the two RTUs engaged in testing a transmission path. By commanding execution of modules at both RTUs simultaneously, any critical timing relationships between corresponding subroutines at the different RTUs can be maintained.

Another mechanism used in the host-RTU interaction is the generation of checksums. Whenever a program is down loaded from the host computer to the RTU, the RTU generates a unique checksum, which is used by the host computer to verify the integrity of the program. In addition, transmission of data from the RTU is implemented in an error-free fashion by the use of parity checks and character counts, with retransmissions in case of errors.

The simplest way for the host and the RTU to communicate is by maintaining a communication link open between them for the duration of a measurement sequence (e.g., a dial-up connection). However, with more sophisticated programs available for the RTUs, the need for this communication link diminishes, and it suffices to poll the RTUs periodically to retrieve data from them or to reinitialize them. This significantly reduces transmission costs.

3.3 Host computer software

The remainder of this section contains detailed descriptions of the structure and the major components of the ASPEN host software. Wherever necessary, specific references to the EOCS implementation of the host software are made, but the software structure is presented in a generic form. For ease of readability, program names are shown in typewriter face and file references are presented in italics. The index *n* represents the number of physical RTUs in the ASPEN study. As Fig. 5 shows, there are three principal layers of ASPEN control and communication software:

1. Layer 1 contains *n* `call` programs, each providing a connection

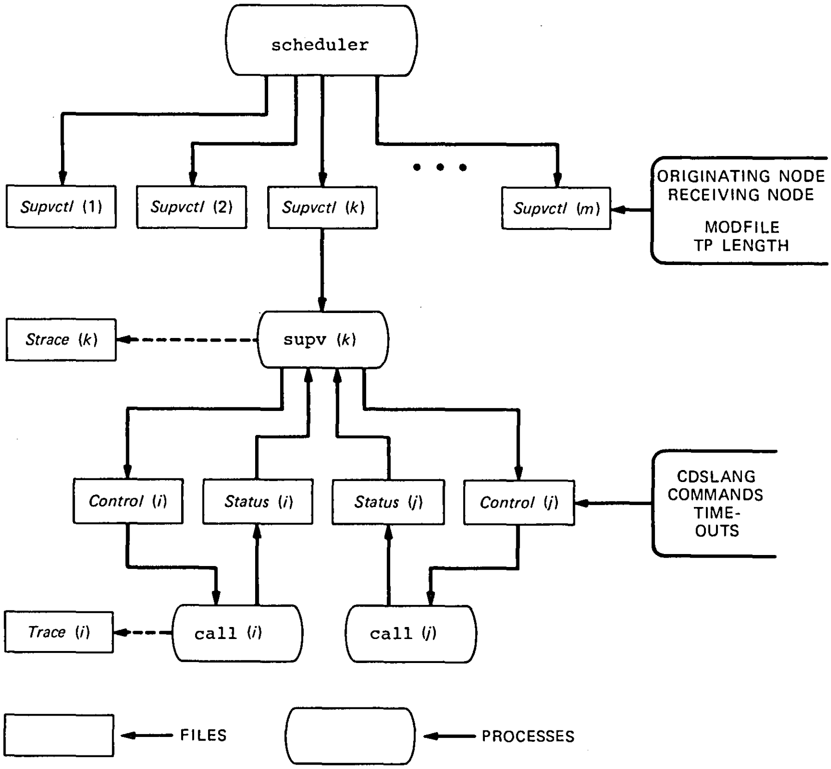


Fig. 5—ASPEN host software configuration.

from the host computer to one RTU, as well as a library of I/O functions used to interact with it.

2. Layer 2 contains $n/2$ `supv` programs, each responsible for controlling and synchronizing one pair of RTUs, by interacting with the subordinate pair of `call` programs.

3. Layer 3 consists of a single `scheduler` program, responsible for implementing the testing schedule for the study, as well as for dealing with the real-time constraints of the host computer system.

3.3.1 Call

`call` is the base software layer and backbone of the ASPEN system. It provides communication with the RTU in a robust, error-tolerant fashion. The `call` program is based on a standard *UNIX* system utility, the `cu` (call *UNIX*) command. The `cu` command is a general-purpose software tool that enables the user to establish a telephone connection from the host computer to a remote computer system (based on the *UNIX* system or some other system). The `cu` command provides a full-duplex environment to the user by splitting itself into two parts, one part for each direction of transmission. It manages an interactive conversation with the remote system and allows file transfers in either direction. The `call` program differs from `cu` in that it operates in half-duplex mode, requiring only a single process to run.* It is also designed specifically to converse with an RTU,† and it is not interactive, but takes its commands from a command file. A benefit of a system such as the *UNIX* system is that resources spent developing a single program can be exploited in multiple invocations of the same program. For the EOCS, 20 invocations of the `call` program ran simultaneously.

One `call` process exists for each working RTU in the ASPEN system. Upon its execution, `call` establishes a control link between the host and the RTU.‡ It then checks that the connection quality exceeds a minimum standard, and enters a quiescent state in which it waits for a signal to proceed. Each `call` process is assigned an index, which is used thereafter to identify it to various files with which it interacts. For example, process `call`(*i*)notifies the operator of its status by depositing information in the file *Trace*(*i*). Similarly, `call`(*i*) will reads its instructions from the file *Control*(*i*). There are four I/O files each `call` process interacts with: *Control, Data, Status* and *Trace.*

The *Control* file contains the instructions to be executed at any given time. In the case of the EOCS, these individual instructions are written in a format denoted CDSLANG, or CDS Language, an intermediate interpretive language based on the built-in commands of the CDS 53A hardware controller, the microprocessor-based controller used in each RTU. However, the `call` program isolates this language in such a way that other formats could be designed to control different RTUs. A collection of CDSLANG instructions, bound together to perform a discrete function, is denoted a *module.* Modules, which are stored as simple *UNIX* system files, are named according to the function they cause the RTU to perform. For example, a module named *Reset* might logically accomplish the software resetting of a component of the RTU.

* The *UNIX* operating system is a multitasking operating system, wherein a process is an image of a software program, loaded in core memory, with its own data and stack segments and a possibly shared text segment. Multiple invocations of a single program can result in multiple processes resident in core memory. Since the system operates in half-duplex mode, fewer processes reside in core memory, and a substantial reduction in CPU load can be achieved.

† For the EOCS, the `call` program contained code specific to the CDS 53A hardware system controller.

‡ The control link for the EOCS consisted of a telephone connection using 1200-b/s modems.

To execute a module, that is, get a `call` process to execute the instructions comprising the module, it must be copied into the *Control* file, and the `call` program must be signaled to begin. The `call` program will wake up and begin executing the CDSLANG instructions until it successfully completes all of them, or until one fails. A failure may be due to transmission difficulties between the host and the RTU, or to RTU hardware problems. The outcome of the module execution is reported to both the *Trace* file and the *Status* file. More than 50 modules were developed for the EOCS, using as building blocks more than 30 CDSLANG instructions.

A *Trace* file exists to store the current status of each `call` process whenever it changes. By utilizing an appropriate display program, an operator may monitor RTU actions.

The *Data* file is a transient template file used to deposit, on a temporary basis, data that are retrieved from the RTU. From this file data are transferred to other, more permanent, files that are time-stamped to reflect the date and time of acquisition.

The *Status* file is the I/O channel between the `call` program and the next higher layer, the `supv` program. This file holds information on the success/failure of modules treated as whole entities. No information is available as to which instruction within a module failed, merely that the module as a whole has failed. In addition, the *Status* file is used to report problems associated with the control link between the RTU and the host (e.g., a dropped telephone line). The `call` program itself is designed to take care of such situations and redial dropped connections, but the `supv` program makes sure that the second `call` program is suspended until the control link is reestablished.

3.3.2 Supv

The `supv` program constitutes the intermediate software layer in the ASPEN system. Its role is to supervise the operation of the two `call` processes beneath it, guiding them throughout the execution of a prearranged sequence of modules. The `supv` follows instructions contained in the *modfile*, a file specifying an ordered sequence of modules plus logical rules to follow in case the predefined sequence does not proceed normally. The program is able to monitor the actions of the two RTUs by analyzing the status information provided by `call` processes. The success of the `supv` may be quantified by the number of modules it guides the `call` processes through, within a specified time frame.

In the case of the EOCS, a typical modfile specified a collection of 30 modules to be executed sequentially, which would:

1. Down load programs into each of two RTUs

2. Establish a test connection between the two RTUs
3. Make a sequence of analog measurements on the test connection
4. Perform bit and block error rate measurements on the test connection
5. Transfer measurement data to the host computer
6. Reset the RTU hardware and release the test connection.

Upon invocation, the `supv` processes, one for each pair of `call` processes, would begin to execute the modules named in the modfile. One by one, a module would be copied into the *Control* files corresponding to the two `call` processes it was supervising, and the `call` processes would be awakened. `Supv` would then wait for the status of the module execution to be appended to the *Status* files. If execution was successful, `supv` would move on to the next module prescribed in the modfile. Otherwise, a set of logical rules specified in the modfile would direct `supv` through a sequence of actions ranging from repeating the module to executing an alternate set of modules, or aborting the whole modfile. To prevent endless loops, each module has a maximum allowable execution time, or "time-out" constraint. For the EOCS, eight principal modfiles controlled the measurement sequencing. New modfiles are easily developed, a flexibility which makes it possible to change the direction of a measurement study or to generate special-purpose substudies. When operating all RTUs, 10 `supv` processes supervised 20 `call` processes simultaneously, with data collection taking place on 10 test connections.

The `supv` processes take control of the `call` processes at the beginning of each new time period, defined in the next section. Each `supv` process is also assigned an index upon invocation, and only then finds out which two `call` processes it is to supervise. This information is conveyed by means of the *Supvctl* files. The *Strace* files also exist to store diagnostics from the `supv`.

Much as the success of the `supv` program is based on the outcome of individual modules, the progress of the `scheduler` program is based on the success or failure of whole modfiles. In the example above, a successful execution of the modfile would mean that the two RTUs involved successfully executed all the modules specified by the modfile on a test connection. A failed modfile execution would imply that a connection between the two particular RTUs needs to be rescheduled by the `scheduler`, since it has failed.

3.3.3 Scheduler

The `scheduler` program keeps track of the status of the entire measurement system. In contrast to `call` and `supv`, the `scheduler` program is not well defined; its implementation will vary depending on the criteria specified by a statistical sampling plan. It can range from a simple control program, which repeatedly schedules a connection between the same set of RTUs, to a complex set of procedures, which govern a large number of RTUs, assuring that certain combinations of tests occur at certain times of the day, with special consideration for a subset of important RTUs. For the EOCS, the `scheduler` implemented a scheduling algorithm in accordance with the EOCS sampling plan. The aim was to ensure a uniform distribution of the number of measurements of each of the 380 possible RTU pair transmission paths (with 20 RTUs), while maximizing the number of RTU pair connections established during the busy hour(s) at the originating RTU location (nominally 10:00 to 11:00 a.m. and 2:00 to 3:00 p.m., local time).

In addition, the `scheduler` is responsible for making sure the study can accommodate whatever maintenance procedures exist for the host computer. The `scheduler` design provided that RTUs would not be active during a one-hour period per day scheduled a week in advance for maintenance of the host computer and stored in a file readable by the `scheduler`. By agreement, the computer servicing time occurred within a five-hour period before 5 a.m. daily.

The `scheduler` also requires robustness; it must provide adequate procedures for recovering from unexpected system downtime cleanly. For the EOCS, the `scheduler` consisted of a table- and list-driven C program that worked in combination with several smaller shell scripts. (The shell, a high-level programming language and command interpreter, is a fundamental component of the *UNIX* operating system.) For less complex studies, simpler implementations of the `scheduler` are possible entirely in shell control language.

For the EOCS, the concept of a run and a time period were developed. A *time period* consists of the time necessary for a pair of `call` processes to execute the modules specified by the modfile under the control of the `supv` process, approximately 2.5 hours for the EOCS. A *run* consists of a full cycle of time periods such that all possible RTU pair combinations are tested. For the EOCS a run was 38 time periods with 10 simultaneous RTU-RTU connections per time period. The `scheduler` used the beginning of each time period to resynchronize the scheduling algorithm, compute the connections for the next time period, manipulate lists and tables necessary to implement the busy-hour requirement of the sampling algorithm, and deal with RTUs that were out of service. As soon as the list of connections for the next time period was available, the `scheduler` would deposit the information into each of $n/2$ files, labeled *Supvctl*(1) through *Supvctl*(m), where n is the number of RTUs in the study and $m = n/2$. These files constitute the I/O channel to the intermediate `supv` level described above.

IV. END OFFICE CONNECTION STUDY SAMPLING PLAN

As we noted in the Introduction, the use of ASPEN technology has engendered new statistical problems involving sampling and data analysis. In this section we examine sampling issues somewhat generally. For the EOCS, sampling involves two components: choice of RTU locations and scheduling of calls in time. They are treated separately in this paper since they involve spatial and temporal considerations, respectively. Furthermore, the preselection method and the scheduling algorithm used to resolve these issues are general tools that may be used separately in other applications. Data analysis issues are examined in the companion paper.[12]

The problem involving selection of RTU locations arises because sampling methods required to support the new measurement methodology do not fit the classical sampling context. In the classical context experimental units are items on which measurements are made, sampling units are items that could be included in the sample, and sampling and experimental units are the same. For the EOCS, experimental units are connections, defined by ordered pairs of end office switching machines. The sampling units are end office buildings and they are *not the same* as experimental units. The situation was further complicated since study goals required adequate representation of certain strata in the sample, and stratification variables are defined in terms of both experimental and sampling units. For example, strata defined by airline mileage (a property of a pair of end office switching machines and hence defined in terms of experimental units) and by originating switch type (defined in terms of sampling units) were required.

A new method of sampling, herein called the preselection method, provides a general method for sampling networks under representation constraints. The method, an extension of the method of snowball sampling,[13] is the subject of Section 4.1.

As Section 4.2 describes, the scheduling problem involves determining when to place calls between the various RTU pairs. An algorithm was developed that provides two outputs: a synchronous, clocked schedule of RTU pairs to be in conversation at any time, and a list of the end office switches to which RTUs are to be connected. (Where end offices contained more than one type of switch, the EOCS sampling plan specified connections using each switch type.) It also provides for stratification of calls by busy versus nonbusy hour.

4.1 Remote test unit location sampling

With 20 RTUs allocated to the EOCS* at least 380 basic experi-

* Examination of data from Ref. 7 suggests that one deployment of 20 RTUs would be sufficient to achieve the goals of the study.

Table V—Sample representation requirements for the end office connection study sampling plan

Dimension	Stratum	Requirement
Mileage	0–360 miles*	At least 50 RTU pairs
	360–720 miles	At least 80 RTU pairs
	720–1320 miles	At least 100 RTU pairs
	1320 + miles	At least 110 RTU pairs
Originating switch type	*ESS* switching equipment	At least 6 switches
	Crossbar	At least 6 switches
	Step-by-step	At least 6 switches, 2 of which are Community Dial Offices (CDOs)
Facility	Satellite	At least 12 RTU pairs
	Terrestrial	Remainder
Region	Long Lines region†	3 RTUs per Long Lines region plus 2 in New York City

* Airline miles

† At the time of the EOCS sampling, Long Lines (now AT&T Communications) was divided into six administrative regions.

mental units are defined by originating-terminating pairs.* These experimental units cannot correspond to sampling units since obtaining a sample of 380 experimental units by selecting 380 office pairs would lead to the use of (up to) 760 RTUs. Thus, sampling units must be buildings that house end office switches, and the experimental and sampling units do not correspond.

As we noted, the sample was required to adequately represent various strata. Strata definitions and representation requirements were derived using 1969/70 Connection Survey data.[7] The most important strata variables are airline mileage and originating switch type. Airline mileage serves as an easily determined surrogate for route mileage, which is known to affect many parameters and is not easy to measure. Impulse noise is associated with step-by-step switches. In addition, strata were defined by geographic regions and by whether telephone connections between two locations could be carried by a satellite. Strata definitions and representation requirements are given in Table V.

The preselection method, used in drawing a sample that meets the representation requirements noted above, is diagramed in Fig. 6. It features three steps:

1. Creation of clusters of experimental units and definition of selection probabilities.

2. The preselection step, the output of which is a set of clusters, guaranteed to produce an acceptable sample.

* Many Bell operating company buildings house more than one switch, and the RTUs were designed for connection to up to three separate switches. Since experimental units are defined using switches rather than buildings, there were more than 380 such units in the EOCS. Multiple-switch buildings play a key role in sampling, as we will discuss.

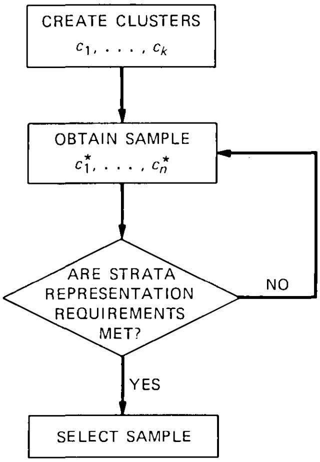


Fig. 6—The preselection method.

3. The selection step, in which the actual sample of experimental units is drawn.

The preselection method is quite general. Creation of clusters admits great latitude. Selection probabilities need not be equal, and at each step sampling can be with or without replacement, with or without stratification, and so on.

For the EOCS, clusters were created through definition of subregions, areas of about 10,000 square miles each, in each Long Lines (now AT&T Communications) administrative region. Clusters included all Bell System buildings within a given subregion. Selection probabilities were based on the number of subscriber lines served by a building and the type(s) of switches it housed. Three weights, one per switch type, were defined for each building:

$$w(b, s) = \begin{cases} T_b = \text{number of lines served by building } b, \text{ if switch type } s \text{ is housed in building } b, \\ 0, \text{ if } s \text{ is not in } b, \text{ where } s = \\ \quad \mathit{ESS}, \text{ crossbar, step-by-step switching equipment.} \end{cases}$$

Selection probabilities were defined through normalization of weights throughout. They favor inclusion of buildings that house more than one switch type, because such buildings lead to comparisons of switch type performance that are not confounded with other variables. Subregion weights were calculated by summing over all buildings located within the subregion.

In preselection, subregions were sampled according to the following algorithm:

1. Twenty artificial units, six each labeled *ESS*, crossbar, and step switching equipment, and two unlabeled, were defined. Starting in the Northeast, an artificial unit was drawn without replacement and a subregion sampled, again without replacement, using the subregion selection probabilities that correspond to the (switch) label of the artificial unit.

2. Sampling proceeded in this manner until the appropriate number of subregions had been sampled in each region. There was one exception to this rule: The two subregions in which Chicago and Orlando are located were included in the sample with probability one to help meet the satellite representation criterion (Table V). The label associated with the first artificial unit drawn in the Midwest and Southern regions was assigned to the Chicago and Orlando subregions, respectively.

3. Except for the mileage criteria, the sampling scheme ensures that all representation requirements are met. These criteria were checked by assuming that the selected buildings would be located at the center of their respective subregions and by computing the resultant distances. If a sample of subregions did not satisfy these representation criteria, that sample would be discarded and steps 1 through 3 repeated.

Once the sample of subregions was accepted by the preselection step, the selection step was used to produce an actual sample of buildings. Building selection probabilities corresponding to the switch types assigned each region were used in this step.

Locations of sampled buildings are shown in Fig. 7.

4.2 Test connection scheduling algorithm

When RTUs are in place and operating, the ASPEN system is capable of collecting data nearly continuously. Calls placed between RTUs simulate calls that could be placed by customers; they are not sampled from that population, however. This has implications for data analysis.

For the EOCS the following goals were adopted:

1. Minimize measurement equipment idle time

2. Maximize the number of busy-hour calls placed and ensure that both busy- and nonbusy-hour calls are made between all RTU pairs

3. Place and receive calls through all switches connected to each RTU with equal frequency

4. Be robust to RTU failure.

The scheduling algorithm described herein works whenever n, the number of RTUs, is even, though it is described here as used in the EOCS ($n = 20$).

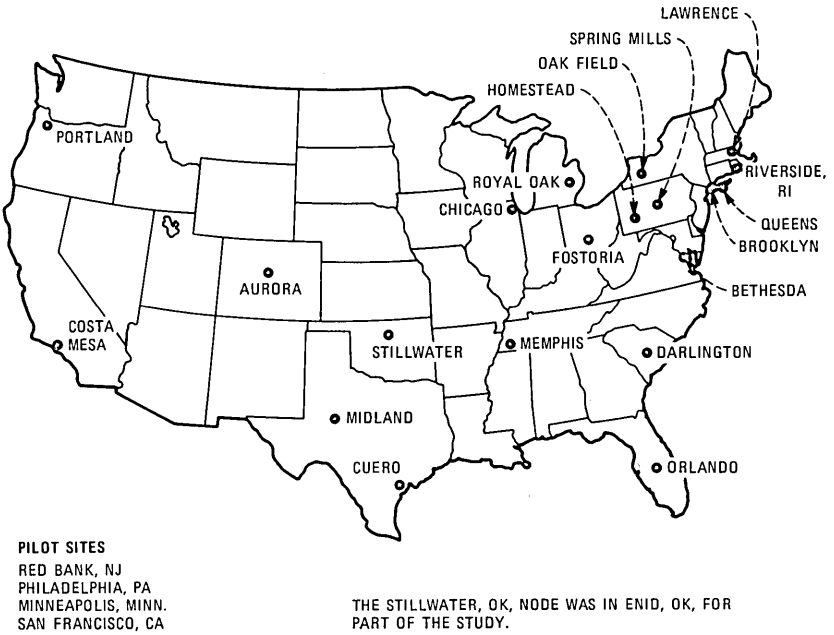


Fig. 7—Measurement equipment locations for end office connection study.

The algorithm specifies a synchronous, clocked schedule. That is, at the beginning of a measurement sequence, ten connections (of the 380 possible) are established and a predetermined span of time (the time period) is allotted for completion of test sequences. Thirty-eight time periods define a run, during which a call on each ordered RTU pair is established once and only once. The basic elements of the algorithm are:

1. The RTU pair table—This table consists of all 380 possible ordered pairs of 20 symbols, grouped in 38 sets, with each symbol used once and only once in each set. When the scheduling algorithm is implemented, symbols will be assigned to RTUs and sets to time periods. Construction of this table is discussed below.

2. The busy-hour table—This table gives the busy hour for each RTU and, as the experiment proceeds, the number of busy-hour connections established to date.

3. The switch definition table—This table specifies the switches through which calls are established for each RTU pair based on the run number.

At the start of each run:

1. RTUs are randomly assigned to symbols in the RTU pair table to provide 38 sets of 10 RTU pairs each.

2. Each of the 38 sets is assigned a time period. First, when the busy-hour table is being used, sets associated with RTU pairs on which few busy-hour calls have been placed are preferentially assigned time periods that correspond to the busy-hour for those pairs. Eventually, such assignments can no longer be made and the remaining sets are randomly assigned time periods.

3. Finally, the switches through which RTU pairs are to make each connection are determined by consulting the switch definition table.

Thus far, the algorithm provides a complete schedule of calls, but provides no method of protecting itself from intermittent RTU failure. To do so, the algorithm specifies that at the start of each time period lists be made of out-of-service RTUs and idled RTUs (mates of RTUs out of service). Connections that cannot be established, including designated switch pairs, are then added to a "calls missed" list. Finally, idled RTUs are used to establish previously missed connections.

The *RTU pair table* is the cornerstone of the scheduling algorithm since it establishes an efficient mechanism for specifying study test connections among the various RTU pairs. Creation of such a table is nontrivial, and so we describe the method used for the EOCS. This method starts with a Latin square,[14] but is not completely general in that Latin squares exist that do not yield an RTU pair table. The authors speculate that restriction to a special class of Latin squares would lead to a method of full generality.

The method is described below for the general n-RTU case, n even, and is illustrated using $n = 4$ throughout.

1. Start with a Latin square of size n, using symbols 0, ..., n-1. Attach row and column letters in such a way that zeroes are (i, i) entries.

EXAMPLE: For a four-RTU experiment, there are twelve RTU pairs and two pairs can be connected in each time period. Six time periods will be required to connect all RTUs in all possible pair combinations. An augmented Latin square is:

	a	b	c	d
a	0	2	3	1
d	1	3	2	0
b	2	0	1	3
c	3	1	0	2

2. Group the (row, column) ordered pairs that correspond to each nonzero entry of the Latin square.

EXAMPLE (continued): For the entry 1 we obtain:

(a, d), (d, a), (b, c), and (c, b).

3. Split each group into two subgroups such that each symbol

appears in each subgroup once and only once. This yields the RTU pair table.

EXAMPLE (continued): For the group above we obtain:

$$\{(a, d), (b, c)\} \text{ and } \{(d, a), (c, b)\}.$$

Each bracketed set corresponds to RTU pairs over which measurements will be made during one of the six time periods in each run of the four-RTU experiment. Note that these groups need not be unique.

V. ACKNOWLEDGMENTS

Many individuals contributed to the success of both the ASPEN Data Acquisition System and the End-Office Connection Study. We would especially like to thank Alan Furtado and Jon Palmer for their tireless efforts in installing and maintaining the remote test units, Jay MacMaster for his elegant physical design and construction of the RTUs, Louis Iacona for his vigilant management of the Study database, Karel Ehrlich for his efficient design of the ASPEN scheduler software, and Bob Sturges for his vital assistance in design and construction of RTU support circuitry.

Michele Carey, Han-Tee Chen, Fred Descloux, Jim Ingle, and Kun Park, the authors of the companion paper, provided the screening algorithms and rapid analyses of pilot and full study data that permitted the high level of data integrity achieved in this study. We were especially fortunate to have Jim Ingle take charge of the testing and acceptance of transmission impairment measuring sets for ASPEN, an invaluable contribution to a critical phase of the project.

We also wish to thank Paul Auer for his help in shaping the ASPEN system and in designing the transient parameter measurement procedures, Blan Godfrey for his far-sighted support of the original ASPEN concept, and Pete Lopiparo for his steady confidence in and support of the ASPEN project team. We gratefully acknowledge the cooperation and contributions of Bell operating company central office personnel and the many valuable consultations with our colleagues at AT&T Bell Laboratories. This was in every sense a Bell System study.

REFERENCES

1. A. A. Alexander, R. M. Gryb, and D. W. Nast, "Capabilities of the Telephone Network for Data Transmission," B.S.T.J., *39,* No. 3 (May 1960), pp. 431–76.
2. I. Nasell, "The 1962 Survey of Noise and Loss on Toll Connections," B.S.T.J., *43,* No. 2 (March 1964), pp. 697–718.
3. J. H. Fennick and I. Nasell, "The 1963 Survey of Impulse Noise on Bell System Carrier Facilities," IEEE Trans. Commun. Technology, *COM-14,* No. 4 (August 1966), pp. 520–4.
4. I. Nasell, "Some Transmission Characteristics of Bell System Toll Connections," B.S.T.J., *47,* No. 6 (July–August 1968), pp. 1001–18.
5. I. Nasell, C. R. Ellison, Jr., and R. Holmstrom, "The Transmission Performance of Bell System Intertoll Trunks," B.S.T.J., *47,* No. 8 (October 1968), pp. 1561–1613.

6. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48,* No. 10 (December 1969), pp. 3337–86.
7. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50,* No. 4 (April 1971), pp. 1311–47.
8. M. D. Balkovic, H. W. Klancer, S. W. Klare, and W. G. McGruther, "High-Speed Voiceband Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50,* No. 4 (April 1971), pp. 1349–84.
9. H. C. Fleming and R. M. Hutchinson, Jr., "Low-Speed Voiceband Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50,* No. 4 (April 1971), pp. 1385–1405.
10. J. E. Kessler, "The Transmission Performance of Bell System Toll Connecting Trunks," B.S.T.J., *50,* No. 8 (October 1971), pp. 2741–77.
11. F. P. Duffy, G. K. McNees, I. Nasell, and T. W. Thatcher, Jr., "Echo Performance of Toll Telephone Connections in the United States," B.S.T.J., *54,* No. 2 (February 1975), pp. 209–43.
12. M. B. Carey, H.-T. Chen, A. D. Descloux, J. F. Ingle, and K. I. Park, "1982/83 End-Office Connection Study, Analog and Voiceband Data Performance on the Public Switched Network," AT&T Bell Lab. Tech. J., this issue.
13. L. A. Goodman, "Snowball Sampling," Ann. Math. Statist., *32* (March 1961), pp. 148–70.
14. O. Kempthorne, *The Design and Analysis of Experiments,* Huntington, NY: Robert E. Krieger Publishing Co., Inc. 1952.

AUTHORS

John Healy, B.S. (Mathematics), 1971, St. Bonaventure University; Ph.D. (Mathematical Statistics), 1976, Purdue University, AT&T Bell Laboratories, 1976–1983. From 1976 through 1983 at Bell Laboratories, he held various technical and supervisory positions in the areas of statistics, reliability, and network measurements. His research has been published in the Reliability and Maintainability Symposium Proceedings, Bell System Technical Journal, the Journal of the American Statistical Association, Psychometrika, and the Journal of Multivariate Analysis. He is currently the District Manager of the Reliability and Maintainability Methods District in the Quality Assurance Technology Center at Bell Communications Research.

Maurice Lampell, B.A. (Physics) and B.S. (Electrical Engineering), 1979, Boston University; M.S. (Electrical Engineering), 1980, Stanford University, AT&T Bell Laboratories, 1980–1983. Present affiliation Bell Communications Research, Inc. Mr. Lampell has been involved with the development of automated data acquisition systems running under the *UNIX* operating system. He is currently working on network characterization planning, and is developing a framework for the transfer of data acquisition expertise to the Bell operating companies.

David G. Leeper, B.S. (Electrical Engineering), 1969, Washington University; M. Eng. (Electrical Engineering), 1970, Cornell University; Ph.D. (Electrical Engineering), 1977, University of Pennsylvania. AT&T Bell Laboratories, 1969–1983. Present affiliation Bell Communications Research, Inc. From 1969 to 1978 Mr. Leeper was involved in design, engineering, and performance measurements on new and existing digital transmission systems. From 1978 to 1983 he supervised design and construction of computerized performance measurement systems and their application in network performance field measurement studies. He is currently Division Manager, Exchange Network Services Planning.

Thomas C. Redman, B.S. (Mathematics) 1976, Northwestern University; M.S., Ph.D. (Statistics), Florida State University, in 1978 and 1980, respectively; AT&T Bell Laboratories, 1980–1983. Present affiliation Bell Communications Research, Inc. Mr. Redman has worked on the planning, execution, and statistical aspects of (telephone) network characterization studies. He is now with Bell Communications Research and continues to work on similar problems. Member, ASA.

Edward J. Vlacich, B.S. (Electrical Engineering), 1964, Fairleigh Dickinson University; M.S. (Statistics), 1970, Rutgers University; AT&T Bell Laboratories, 1961–1983. Present affiliation Bell Communications Research, Inc. Mr. Vlacich has been involved in the field instrumentation systems used for characterizations of the telephone network. He helped design ASPEN, which is used to characterize transmission impairments associated with data transmission. He is currently working to incorporate state-of-the-art measurement techniques into new survey equipment.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 9, November 1984
Printed in U.S.A.

1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network

By M. B. CAREY,* H.-T. CHEN,* A. DESCLOUX,* J. F. INGLE,* and K. I. PARK*

(Manuscript received December 19, 1983)

A comprehensive systemwide field study, referred to as the 1982/83 End Office Connection Study (EOCS), was undertaken by Bell Laboratories from October 1982 through January 1983 to characterize the transmission performance of the predivestiture Bell System public switched telecommunications network. Analog voice and voiceband data transmission parameters were measured on about 6500 direct-distance-dialing connections among 20 end office buildings located throughout the continental United States. The analog parameters measured on the connections included loss; noise; frequency response; envelope delay distortion; intermodulation distortion; phase jitter; amplitude jitter; peak-to-average ratio; frequency shift; propagation delay; transient phenomena such as impulse noise, gain hits, phase hits, and dropouts; and error rates of 1200-b/s full-duplex and 4800-b/s half-duplex data sets. This paper presents the results of the EOCS data analysis; a companion paper describes the measurement equipment and the sampling plan. The performance characterization information presented in this paper updates the similar information provided by a survey conducted in 1969/70. The results represent the last predivestiture Bell System network performance characterization and may serve as a benchmark for the end-to-end performance in the post-divestiture environment.

* AT&T Bell Laboratories; present affiliation Bell Communications Research, Inc.

I. INTRODUCTION

Bell Laboratories undertook a comprehensive systemwide field study from October 1982 through January 1983 to characterize the transmission performance of the predivestiture Bell System public switched telecommunications network. This study, hereinafter referred to as the 1982/83 End Office Connection Study (EOCS), employed special measurement equipment, referred to as ASPEN (Automatic System for Performance Evaluation of the Network). Analog voice and voiceband data transmission parameters were measured on 6141 Direct-Distance-Dailing (DDD) connections among 20 end office buildings located throughout the continental United States; in addition, another 395 connections were measured among four pilot study locations at the start of the study. (Several end office buildings visited in the EOCS had multiple end office switches. When multiple switches were measured in the same building, they were individually identified in the EOCS database.) Measurements were typically made in one direction on each test connection. However, measurements were often made in both directions and sometimes repeated for stationarity studies, resulting in over 9000 1004-Hz loss measurements, for example.

The ASPEN equipment consisted of 20 Remote Test Units (RTUs) (one per sampled end office building) under the control of a computer located at Holmdel, New Jersey. Each RTU was connected to the line side of the main distributing frame in its central office and contained a microprocessor, a transmission impairment measuring set, and modems for communication with the central computer and for data performance tests. The transmission impairment measuring set (HLI 3701 Communications Test Set) was designed to meet the requirements specified in AT&T PUB 41009[1] for measuring the impairments described in AT&T PUB 41008.[2]

Analog parameters measured on the connections included loss, noise, frequency response, envelope delay distortion, intermodulation distortion, phase jitter, amplitude jitter, Peak-to-Average Ratio (P/AR), frequency shift, propagation delay, and transient phenomena such as impulse noise, gain and phase hits, and dropouts. Error rates of voiceband data sets were also measured for 1200-b/s full-duplex and 4800-b/s half-duplex transmission. Table I contains a complete list of the EOCS measurement parameters.

This paper presents the results of the EOCS data analysis; a companion paper[3] describes the ASPEN measurement equipment and the sampling plan. The performance characterization information presented in this paper updates similar information provided by the 1969/70 Connection Survey.[4] Furthermore, with the 1984 divestiture of the Bell System, the results represent the last predivestiture Bell System

Table I—Parameter coverage

Classification	Measured Parameter
Voice and voiceband data transmission	1004-Hz insertion loss Frequency response (30 frequencies) C-message noise Propagation delay Call cutoffs
Voiceband data transmission	Signal-to-C-notched-noise ratio C-notched noise 3-kHz flat noise 3-kHz flat notched noise 3-kHz noise-to-ground Envelope delay distortion (30 frequencies) P/AR Second-order intermodulation distortion Third-order intermodulation distortion Phase jitter (2 to 300 Hz) Phase jitter (20 to 300 Hz) Amplitude jitter (2 to 300 Hz) Amplitude jitter (20 to 300 Hz) Frequency shift Impulse noise (six thresholds) Phase hits (three thresholds) Gain hits (three thresholds) Dropouts
Data set error performance	1200-b/s bit error rate (two modems) 4800-b/s bit error rate (two modems) 1200-b/s block error rate (two modems) 4800-b/s block error rate (one modem)

network performance characterization and may serve as a benchmark for the end-to-end performance in the post-divestiture environment.

II. THE DATA ANALYSIS METHOD

The primary goal of data analysis in the EOCS was to provide estimates of the distributions of the parameters for various strata of interest and for the overall network. Estimation of the distribution was selected, since other measures such as means, variances, and quantiles are obtained as a function of the distribution.

The sampling design of the EOCS involved a stratification according to two variables:[3] connection airline mileage and type of switch at the end office. An analysis of covariance (with switch type as factor and mileage as covariate)[5] was carried out on each parameter to determine the effects of these two variables. Although the effects of mileage and switch type were often both significant at the 5-percent level, one effect usually dominated the other for the value of its F-statistic in the covariance model. Results are displayed as a function of the most relevant effect.

When a display as a function of airline mileage is deemed necessary, the results are split into three categories: short (0 to 180 miles), medium (181 to 720 miles), and long (>720 miles). These particular mileage bands were chosen so that the EOCS results could be compared with similar results from the 1969/70 Connection Survey.[4]

The number of measurements varies with the measured parameter. To give an idea of the number of measurements taken for different strata, the number of 1004-Hz loss measurements made in the EOCS is shown in Fig. 1 as a function of airline mileage in 100-mile blocks, where the abscissa shows the midpoints of the 100-mile blocks (e.g., 5 represents the 451- to 550-mile block). Table II gives the number of 1004-Hz loss measurements in the EOCS for the short, medium, and long mileage categories, together with the number of different pairs of end office buildings evaluated in the EOCS. The estimated percentage of toll traffic[6] in each of the three mileage bands is also listed. Table III shows the number of 1004-Hz loss measurements made in the EOCS for each of the three types of switch—electronic switching system (digital switch), crossbar, and step-by-step switches—in the measuring end office, as well as the percentage of toll traffic estimated in each of the three switch strata.[7]

Table II—Stratification of data by connection airline mileage

Designation	Airline Mileage	Estimated Route Mileage	Number of Loss (1004-Hz) Measurements	Number of Nonordered Pairs of Sites	Percent of Toll Traffic
Short	0–180	0–341	776 8.1%	11 5%	77.8
Medium	181–720	342–1064	3717 38.8%	76 34.2%	13.4
Long	721–2576	1065–3378	5088 53.1%	135 60.8%	8.8
Total			9581	222	

Table III—Stratification of data by type of switch in the office where measurements were made

Designation	Number of Loss (1004-Hz) Measurements	Percent of Toll Traffic
Digital switch	4154 43.3%	58
Crossbar	4146 43.3%	30
Step-by-step	1281 13.4%	12
Total	9581	

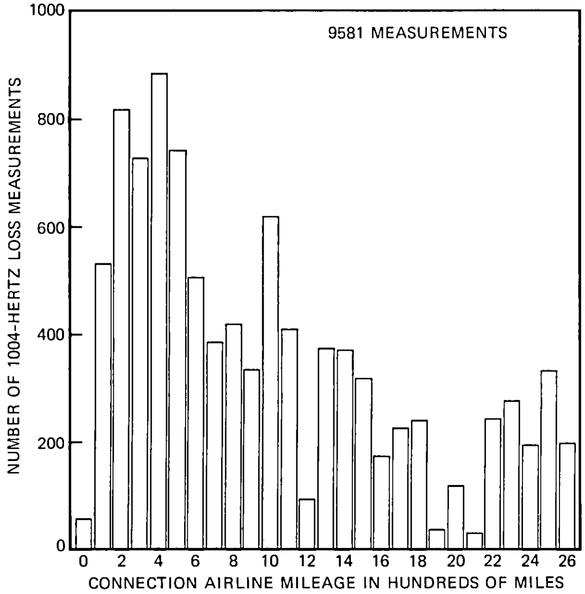


Fig. 1—Number of 1004-Hz loss measurements versus connection airline mileage in 100-mile blocks.

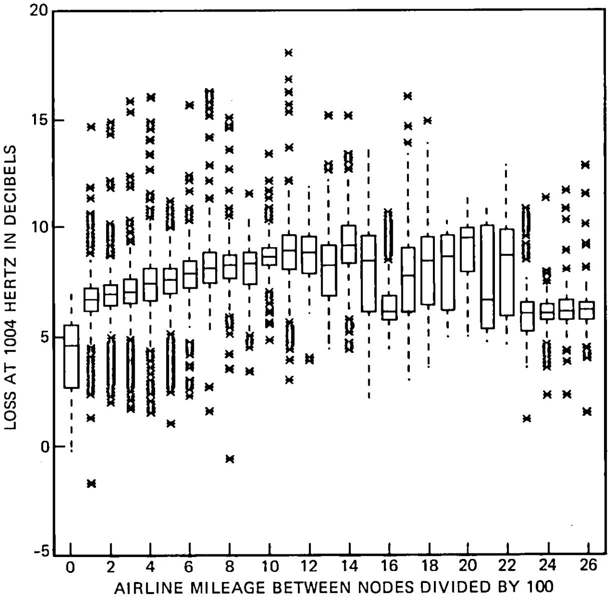


Fig. 2—The 1004-Hz loss versus connection airline mileage in 100-mile blocks.

Within each stratum, the measurements were considered self-weighted. Weights proportional to the traffic occurring in each stratum were used for calculation of results relevant to the whole network. A statistical model of components of variance was considered within each stratum for each parameter. The model led to the estimation of the variability of a parameter on a given pair of end office buildings, and of the variability between different pairs of end office buildings. These two variabilities, together with the means, are displayed in Table IV for all EOCS parameters other than the transient parameters. The variabilities are taken into consideration in the evaluation of the 90-percent confidence interval around the mean in each of the three mileage band strata, also shown in Table IV.

One goal of this paper is to characterize, where possible, the customer-premises-to-customer-premises performance, i.e., the end-office-to-end-office connection plus two loops. However, the impairment contributions from customer loops to the overall connection are usually negligible for most of the parameters discussed here. Therefore, the characterization information presented in this paper is, for the most part, based on the information available for end-office-to-end-office connections only, i.e., the EOCS results. However, for 1004-Hz loss, frequency response, and C-message noise, for which loops are significant, loop impairment contributions are concatenated to the end office to end office values, using a common source of information for the loop impairments, the 1980 Loop Survey.[8]

To concatenate the loop effects for loss and noise, distributions for 1224 predivestiture Bell System representative loops for 1004-Hz loss (measured at the main distributing frame) and C-message noise (measured at the customer premises) were extracted from the 1980 Loop Survey. Loop loss values were available at only three frequencies from that survey, so loop frequency responses were calculated from the transfer characteristics of loops obtained from the record survey associated with the 1980 Loop Survey. The distribution of a parameter for customer-premises-to-customer-premises connections was then obtained by analytically concatenating (using discrete convolution techniques) the distributions of the parameter on the loops with the distribution of the same parameter on the end-office-to-end-office connections. For frequency response, this concatenation process was repeated at each frequency.

III. RESULTS

All test signals were transmitted from the RTUs at −12 dBm.

3.1 1004-Hz loss

A 1004-Hz tone at −12 dBm was applied to one end of the connection, and the received level at the other end was measured, with the difference in levels being defined as the loss of the connection. (If a tone of exactly 1000 Hz [an integer submultiple of the 8000-Hz sampling rate in time-division multiplex systems] were used, the loss readings would "bobble" by +/− 0.25 dB, the signal-to-C-notched-noise ratio readings would vary by +/− 5 dB, and the jitter readings would be erratic. For this reason, a 1004-Hz tone was used in the EOCS as the test tone for loss and as the holding tone for C-notched noise, jitter, and transients. [This is the frequency used in modern transmission maintenance systems.]) The means and standard deviations of the losses are summarized in Table V for the three mileage bands. For comparison, the loss results of the 1969/70 Connection Survey[4] are also shown in the same table. The results from both surveys show that the mean loss increases with mileage, as would be expected from the Via Net Loss (VNL) transmission plan. The VNL design of trunks calls for increasing trunk loss with distance to offset the subjective effect of the increasing echo path delay, until a mileage is reached where an echo suppressor or echo canceler is applied to the trunk. Comparison of the two surveys shows that the mean loss is about the same in both instances, but the standard deviation observed in the EOCS is substantially smaller than in the 1969/70 Connection Survey.

Figure 2 shows "boxplots" of the loss versus airline mileage between the end offices in 100-mile blocks, with the same abscissa as that of Fig. 1. (See Fig. 1 for the number of measurements in each 100-mile block.) For each mileage block on the abscissa, the corresponding boxplot shows the variability of the measurements falling in that mileage block. The upper and lower boundaries of the "box" indicate the 75th and 25th percentiles of the distribution, and the line inside the box indicates the median. Therefore, the height of the box, referred to as the interquartile range, is a measure of variability in the distribution, and the deviation of the median line from the center of the box shows skewness of the distribution. The distance from the median line to the tip of the "whisker" is equal to 1.5 times the interquartile range if there are points falling outside the whisker. If all points fall within the range of 1.5 times the interquartile range on either side of the box, the tip of the lower (or upper) whisker coincides with the minimum (or maximum) value.

Figure 2 shows an overall tendency of increasing loss with mileage of up to about 1200 airline miles, and a decline thereafter. This loss decline can be attributed to the application of echo control devices on trunks, for which the intertoll trunk losses are set to zero. There seems to be a substantial variability in loss for distances between 1200 and 2200 miles. In this mileage region, connections with and without echo control devices are likely to be encountered, which could account for

Table IV—Summary of means and variability measures for analog parameters on end-to-end toll connections

Parameter	Short Mileage Band			Medium Mileage Band			Long Mileage Band			All (Weighted)
	Mean	Standard Deviation Between Pairs Of End Offices	Standard Deviation Within Pairs Of End Offices	Mean	Standard Deviation Between Pairs Of End Offices	Standard Deviation Within Pairs Of End Offices	Mean	Standard Deviation Between Pairs Of End Offices	Standard Deviation Within Pairs Of End Offices	Mean
Loss										
404 Hz	7.86 ± 0.65 dB	1.07	1.37	8.75 ± 0.20	0.93	1.27	9.19 ± 0.28	1.64	1.48	8.10 ± 0.51
1004 Hz	6.53 ± 0.57 dB	0.93	1.34	7.29 ± 0.17	0.80	1.18	7.84 ± 0.21	1.24	1.29	6.75 ± 0.44
2804 Hz	7.88 ± 1.05 dB	1.72	2.00	8.09 ± 0.25	1.17	1.57	8.60 ± 0.25	1.49	1.55	7.97 ± 0.82
EDD										
604 Hz	853 ± 198 μs	316	155	1150 ± 87	393	259	1337 ± 101	571	224	935 ± 155
2804 Hz	599 ± 111 μs	175	149	701 ± 45	203	185	727 ± 39	219	161	624 ± 86
P/AR	87.5 ± 3.0	4.7	4.0	84.5 ± 1.2	5.3	5.5	84.2 ± 1.3	7.8	4.4	86.8 ± 2.3
Delay round trip										
No satellite	9.6 ± 2.0 ms	3.6	2.2	16.0 ± 0.8	3.8	2.4	32.6 ± 1.4	8.6	2.5	12.5 ± 1.6
Only satellite							520.5 ± 3.4	4.9	3.1	

Noise										
C-message	24.6 ± 2.2 dBrnC	3.6	3.4	28.5 ± 0.5	2.2	3.6	31.4 ± 0.5	2.7	2.8	25.7 ± 1.7
C-notch	36.5 ± 1.1 dBrnC	1.8	3.2	37.4 ± 0.3	1.5	2.3	37.9 ± 0.3	1.7	2.1	36.7 ± 0.8
s/n	34.9 ± 1.1 dB	1.7	2.	33.3 ± 0.3	1.3	2.3	32.3 ± 0.2	1.2	2.	34.4 ± 0.8
Flat	43.9 ± 2.6 dBrn	4.2	6.7	44.0 ± 0.5	2.3	6.8	45.4 ± 0.4	2.5	6.4	44.0 ± 2.0
Flat notch	46.2 ± 1.4 dBrn	2.3	5.2	46.0 ± 0.4	1.5	5.1	47.0 ± 0.3	1.9	5.1	46.2 ± 1.1
Noise-to-ground	52.0 ± 3.6 dBrn	6.0	7.9	49.6 ± 0.8	3.5	6.2	50.8 ± 0.6	3.4	5.6	51.6 ± 2.8
Second-order IMD	53.3 ± 1.8 dB	2.8	6.0	52.2 ± 0.6	2.8	5.5	51.6 ± 0.6	3.8	4.8	53.0 ± 1.4
Third-order IMD	53.4 ± 2.5 dB	4.1	5.5	51.0 ± 0.9	4.0	5.9	50.2 ± 0.8	4.9	4.8	52.8 ± 2.0
Jitter										
Amplitude 2–300 Hz	2.8 ± 0.5%	0.8	1.1	3.2 ± 0.1	0.4	1.3	3.6 ± 0.1	0.4	1.5	2.9 ± 0.4
Amplitude 20–300 Hz	2.6 ± 0.4%	0.7	0.8	2.8 ± 0.1	0.3	1.0	3.2 ± 0.1	0.4	1.1	2.7. ± 0.3
Phase 2–300 Hz	5.5 ± 1.9 deg.	3.0	2.2	6.4 ± 0.5	2.1	2.8	7.6 ± 0.4	2.5	2.8	5.8 ± 1.5
Phase 20–300 Hz	3.1 ± 0.5 deg.	0.8	1.2	3.9 ± 0.2	1.1	1.3	4.9 ± 0.2	1.3	1.6	3.3 ± 0.4

Table V—Comparison of 1004-Hz loss from 1982/83 EOCS and 1969/70 Connection Survey

	1969/70 Survey				EOCS Survey	
	Primary		Secondary		DDD	
Connection Length (Airline Miles)	Mean (dB)	Standard Deviation (dB)	Mean (dB)	Standard Deviation (dB)	Mean (dB)	Standard Deviation (dB)
All	6.7 ± 0.6	2.1	6.6 ± 0.3	2.1	6.7 ± 0.4	1.3
0–180	6.5 ± 0.7	2.0	6.4 ± 0.4	2.1	6.5 ± 0.6	1.6
180–720	7.3 ± 0.4	2.3	7.1 ± 0.6	2.1	7.3 ± 0.2	1.4
721–2900	7.7 ± 0.5	2.5	7.4 ± 0.3	2.0	7.8 ± 0.2	1.8

the variability. Above about 2200 miles, the median loss is nearly constant at 6 dB and the variability is small, suggesting that most of the connections in that mileage region have an echo control device. The constant 6-dB loss observed is consistent with a picture of toll connections with one toll-connecting trunk at each end with about 3-dB loss, plus a long, zero-loss intertoll trunk with an echo control device.

Figure 3 gives the Cumulative Distribution Functions (CDFs)* of loss for the three mileage bands. Most of the losses—75 to 90 percent, depending on the mileage category—are greater than 6 dB. This observation is consistent with typical toll connections consisting of the two toll-connecting trunks described above, plus zero to seven (but rarely more than two) intertoll trunks with VNL design losses ranging from 0.5 to 2.9 dB. The losses at the lower tail for the short and medium categories could have come from the connections with no intertoll trunks or with intertoll trunks with negligible VNL loss. The lower tail losses for the long category could have come from the connections with zero-loss intertoll trunks with echo control devices.

Figure 4 shows the CDFs of 1004-Hz loss for customer-premises-to-customer-premises connections. These CDFs were derived by analytically concatenating the 1004-Hz loss of the 1980 Loop Survey[8] to the end office to end office 1004-Hz loss. Figure 5 shows the CDF of 1004-Hz loss for the loops used in the concatenation.

3.2 Frequency response

Frequency response, also referred to as loss-versus-frequency characteristic or attenuation distortion, is a measure of loss variation over the frequency band of a communications channel. It can be measured

* All CDFs in this paper are plotted with an ordinate having the normal probability scale. A normal CDF will show up as a straight line on such a plot, and the vertical scale near the tails of the distribution will be expanded for greater readability.

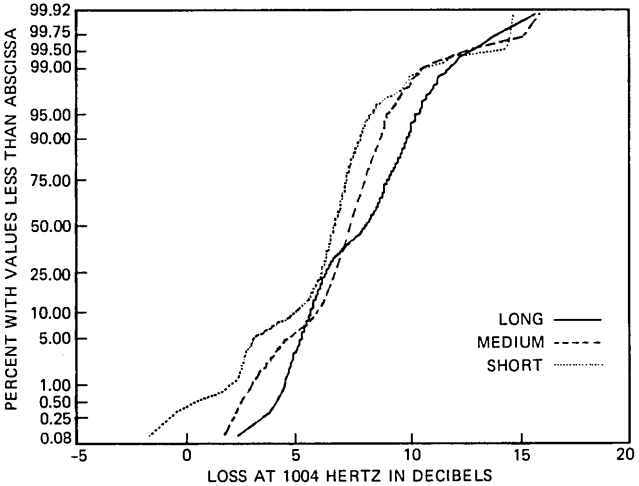


Fig. 3—CDFs of 1004-Hz loss for the short, medium, and long mileage bands.

with sinusoidal test tones or with the 50-percent amplitude-modulated test signal employed to measure the Envelope Delay Distortion (EDD) characteristic. The response of an averaging detector (specified for level measurement by Ref. 1) is the *same* for a sinusoidal tone as for a 50-percent amplitude-modulated signal.

To characterize frequency response, loss was measured from 204 to 3504 Hz. Losses at 204, 254, and 3504 Hz were obtained from a sinusoidal test tone, while the losses at the other frequencies were

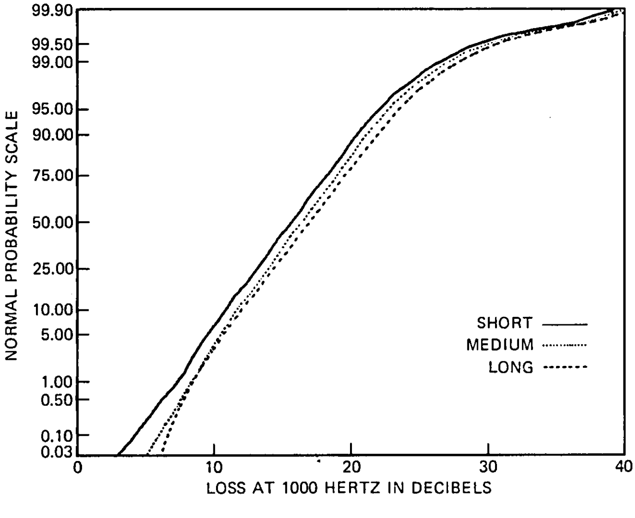


Fig. 4—CDFs of customer premises-to-customer premises 1004-Hz loss.

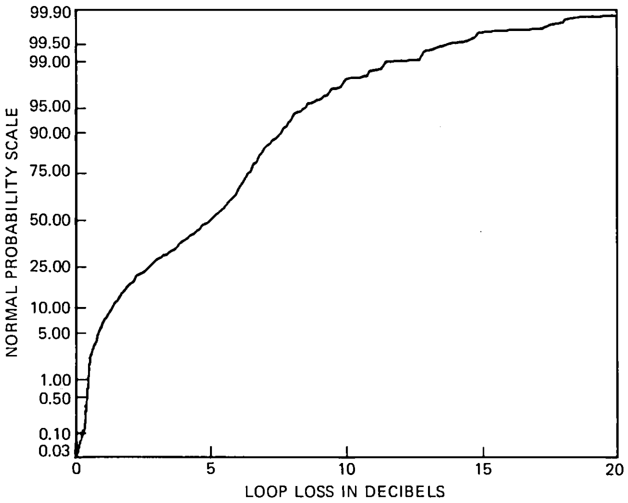


Fig. 5—CDF of 1004-Hz loss on customer loops.

obtained from level measurements of the envelope delay distortion test signal.

Tables VI through VIII show the means, standard deviations, and selected percentiles of loss versus frequency relative to 1004 Hz for

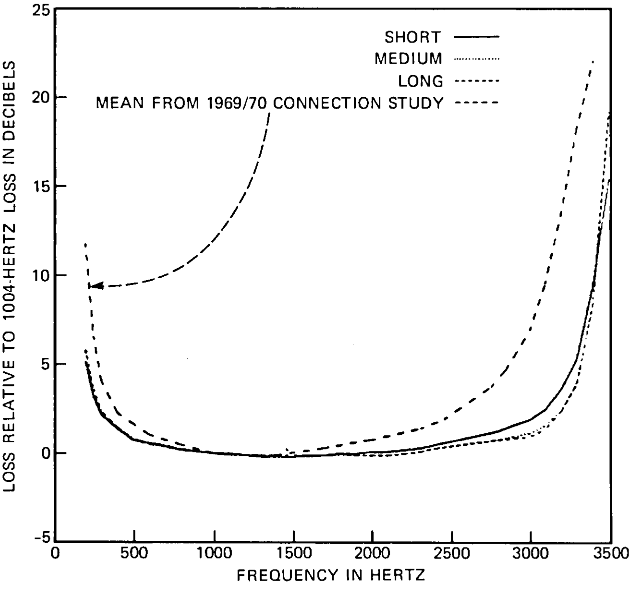


Fig. 6—Man attenuation distortion relative to 1004 Hz for short, medium, and long mileage bands.

Table VI—Attenuation distortion relative to 1004 Hz: short connections

Frequency		Standard	Quantiles				
in Hz	Mean	Deviation	1%	10%	50%	90%	99%
204	5.1 ± 2.9	2.8	2.0	2.5	3.8	10.0	13.2
254	3.3 ± 1.9	1.9	0.9	1.4	2.8	6.4	7.4
304	1.8 ± 1.1	1.2	0.2	0.5	1.6	3.7	4.7
404	1.1 ± 0.4	0.7	−0.8	0.4	1.1	2.0	2.9
504	0.7 ± 0.2	0.4	0.1	0.2	0.7	1.3	2.0
604	0.4 ± 0.1	0.3	−0.6	0.0	0.4	0.9	1.3
704	0.3 ± 0.0	0.3	−0.4	0.0	0.3	0.6	1.2
804	0.2 ± 0.1	0.2	−0.5	0.0	0.2	0.4	0.6
904	0.1 ± 0.0	0.3	−1.2	0.0	0.1	0.2	1.1
1004	0.0 ± 0.0	0.0	0.0	0.0	0.0	0.0	0.0
1104	−0.1 ± 0.1	0.4	−2.3	−0.3	−0.1	0.1	1.3
1204	−0.1 ± 0.1	0.2	−0.6	−0.4	−0.1	0.0	0.4
1304	−0.2 ± 0.2	0.3	−1.7	−0.5	−0.1	0.1	0.7
1404	−0.2 ± 0.2	0.4	−1.2	−0.6	−0.2	0.1	0.6
1504	−0.2 ± 0.2	0.4	−1.3	−0.6	−0.1	0.2	0.8
1604	−0.1 ± 0.1	0.4	−1.2	−0.5	−0.1	0.2	1.0
1704	0.0 ± 0.1	0.4	−1.3	−0.4	0.0	0.3	1.0
1804	0.0 ± 0.2	0.4	−1.5	−0.3	0.0	0.4	1.3
1904	0.1 ± 0.3	0.4	−1.2	−0.3	0.0	0.6	1.2
2004	0.1 ± 0.5	0.6	−1.4	−0.4	−0.1	0.9	1.5
2104	0.2 ± 0.7	0.7	−1.4	−0.4	−0.1	1.0	2.5
2204	0.3 ± 0.8	0.9	−1.4	−0.4	−0.1	1.4	3.3
2304	0.4 ± 0.9	1.0	−1.2	−0.4	0.0	1.9	3.4
2404	0.6 ± 1.1	1.1	−0.9	−0.3	0.1	2.3	3.5
2804	1.7 ± 2.7	2.6	−0.6	−0.1	0.3	5.8	10.0
2904	2.0 ± 3.2	3.0	−0.7	−0.2	0.4	6.8	11.4
3004	2.3 ± 3.5	3.3	−0.6	−0.2	0.5	7.3	12.3
3104	3.1 ± 4.2	4.0	−0.4	0.0	0.9	8.8	15.3
3204	4.1 ± 4.9	4.6	0.2	0.5	1.5	11.5	18.2
3304	5.3 ± 5.3	5.0	1.1	1.5	2.3	13.8	17.7
3404	7.4 ± 4.6	4.2	3.2	5.0	5.9	14.4	23.8
3504	16.5 ± 7.5	6.0	4.2	5.2	19.2	21.2	22.2

short, medium, and long connections. In the tables, the measured frequencies above 304 Hz are at intervals of 10 Hz, except 2504, 2604, and 2704 Hz were omitted. Because of the wide use of the 2600-Hz idle-circuit tone for signaling, the presence of a signal with energy concentrated near 2600 Hz could inadvertently cause disconnection.

Figure 6 presents the mean losses of Tables VI through VIII as a function of frequency for short, medium, and long connections. The losses between 2404 and 2804 Hz (not measured in the EOCS) were obtained by linear interpolation.

For comparison, the mean losses relative to 1004 Hz from the 1969/70 Connection Survey are also shown in Fig. 6. As we can see in the figure, the mean frequency response has improved since that survey. This improvement can be attributed to the increasing use of T-carrier to replace voice-frequency cable facilities in the toll-connecting portion

Table VII—Attenuation distortion relative to 1004 Hz: medium connections

Frequency in Hz	Mean	Standard Deviation	Quantiles 1%	10%	50%	90%	99%
204	5.7 ± 0.3	2.1	2.4	3.6	5.3	8.3	13.1
254	3.4 ± 0.2	1.4	1.0	1.9	3.2	5.2	8.2
304	2.1 ± 0.1	1.0	0.3	1.0	2.0	3.2	5.5
404	1.4 ± 0.1	0.7	0.0	0.7	1.4	2.2	3.3
504	0.9 ± 0.1	0.5	−0.1	0.4	0.9	1.5	2.1
604	0.6 ± 0.0	0.3	0.0	0.3	0.6	1.1	1.6
704	0.5 ± 0.0	0.3	−0.1	0.2	0.5	0.8	1.3
804	0.3 ± 0.0	0.2	−0.1	0.1	0.3	0.5	0.9
904	0.1 ± 0.0	0.1	−0.2	0.0	0.1	0.2	0.5
1004	0.0 ± 0.0	0.0	0.0	0.0	0.0	0.0	0.0
1104	0.0 ± 0.0	0.1	−0.4	−0.2	0.0	0.1	0.2
1204	−0.1 ± 0.0	0.2	−0.6	−0.3	−0.1	0.1	0.3
1304	−0.2 ± 0.0	0.2	−0.8	−0.4	−0.1	0.1	0.3
1404	−0.2 ± 0.0	0.2	−0.9	−0.5	−0.2	0.0	0.3
1504	−0.2 ± 0.0	0.3	−0.9	−0.5	−0.2	0.1	0.3
1604	−0.2 ± 0.0	0.3	−1.0	−0.5	−0.2	0.1	0.4
1704	−0.1 ± 0.0	0.4	−0.9	−0.5	−0.1	0.2	0.6
1804	−0.1 ± 0.0	0.3	−1.0	−0.5	−0.1	0.3	0.7
1904	−0.1 ± 0.0	0.4	−1.0	−0.5	−0.1	0.3	0.8
2004	−0.1 ± 0.1	0.6	−1.1	−0.5	−0.1	0.4	1.0
2104	−0.1 ± 0.1	0.6	−1.1	−0.6	−0.1	0.5	1.1
2204	0.0 ± 0.1	0.6	−1.1	−0.5	0.0	0.6	1.4
2304	0.1 ± 0.1	0.7	−1.0	−0.5	0.0	0.7	1.7
2404	0.3 ± 0.1	0.9	−1.0	−0.4	0.2	0.9	2.7
2804	0.8 ± 0.2	1.2	−0.7	−0.1	0.6	2.0	5.4
2904	1.0 ± 0.2	1.1	−0.6	−0.0	0.7	2.5	4.4
3004	1.2 ± 0.3	1.3	−0.6	0.0	0.8	3.3	5.4
3104	1.6 ± 0.3	1.6	−0.5	0.2	1.1	4.5	6.8
3204	2.3 ± 0.4	2.0	0.2	0.8	1.8	5.9	8.9
3304	3.8 ± 0.5	2.5	1.3	1.8	3.0	8.4	12.5
3404	7.5 ± 0.5	2.5	4.1	5.4	6.9	10.1	16.4
3504	19.2 ± 0.5	2.5	12.7	16.0	19.3	22.2	24.3

of the network. This is corroborated by the observation that digital channel banks of the D3 and D4 type, which provide the interface between the analog voice-frequency signals and the 1.544-Mb/s digital bit stream of the T-carrier system, have a frequency response similar to that of the current survey, shown in Fig. 6.

The gain slope at 404 or 2804 Hz is defined as the loss at that frequency minus the loss at 1004 Hz. Figure 7 plots the CDFs of the larger of the gain slopes (per connection) at 404 and 2804 Hz for short, medium, and long connections. Figures 8 through 10 show the CDFs of loss difference for the frequency pairs of 2804 and 604 Hz, 2404 and 804 Hz, and 2104 and 1304 Hz, respectively. As the figures show, there is little mileage dependence for these loss differences except for a tendency for short connections (possibly on VF cables) to have somewhat higher loss difference.

Table VIII—Attenuation distortion relative to 1004 Hz: long connections

Frequency in Hz	Mean	Standard Deviation	Quantiles				
			1%	10%	50%	90%	99%
204	5.9 ± 0.4	2.8	2.2	3.1	5.0	10.1	14.1
254	3.7 ± 0.3	2.0	0.7	1.7	3.2	6.6	10.4
304	2.1 ± 0.2	1.5	−0.2	0.5	1.8	4.2	7.0
404	1.4 ± 0.1	0.9	−0.2	0.4	1.2	2.5	4.3
504	0.8 ± 0.1	0.5	−0.2	0.2	0.8	1.5	2.5
604	0.5 ± 0.0	0.4	−0.2	0.0	0.5	1.0	1.6
704	0.4 ± 0.0	0.3	−0.3	0.0	0.4	0.8	1.3
804	0.2 ± 0.0	0.2	−0.2	0.0	0.2	0.5	0.9
904	0.1 ± 0.0	0.1	−0.2	0.0	0.1	0.2	0.5
1004	0.0 ± 0.0	0.0	0.0	0.0	0.0	0.0	0.0
1104	0.0 ± 0.0	0.2	−0.5	−0.2	0.0	0.1	0.3
1204	−0.1 ± 0.0	0.2	−0.7	−0.3	−0.1	0.1	0.4
1304	−0.1 ± 0.0	0.2	−0.8	−0.4	−0.1	0.1	0.4
1404	−0.1 ± 0.0	0.4	−0.9	−0.4	−0.1	0.1	0.4
1504	−0.1 ± 0.0	0.4	−1.0	−0.5	−0.1	0.1	0.5
1604	−0.1 ± 0.0	0.4	−1.1	−0.5	−0.1	0.2	0.6
1704	−0.1 ± 0.0	0.4	−1.2	−0.5	0.0	0.3	0.7
1804	0.0 ± 0.0	0.5	−1.2	−0.5	0.0	0.3	0.8
1904	−0.1 ± 0.0	0.5	−1.2	−0.5	−0.1	0.4	0.9
2004	−0.1 ± 0.0	0.5	−1.3	−0.6	−0.1	0.4	1.1
2104	−0.1 ± 0.1	0.6	−1.3	−0.6	−0.1	0.4	1.3
2204	0.0 ± 0.1	0.6	−1.2	−0.6	−0.1	0.5	1.6
2304	0.1 ± 0.1	0.7	−1.1	−0.5	0.0	0.7	1.8
2404	0.3 ± 0.1	0.8	−1.0	−0.4	0.2	1.0	2.5
2804	0.8 ± 0.1	1.1	−0.8	−0.1	0.6	1.9	4.7
2904	0.9 ± 0.2	1.2	−0.8	−0.2	0.6	2.3	5.0
3004	1.0 ± 0.2	1.3	−0.7	−0.1	0.7	2.6	5.7
3104	1.4 ± 0.2	1.6	−0.6	0.1	1.0	3.2	7.4
3204	2.2 ± 0.3	1.9	0.0	0.6	1.7	4.2	10.2
3304	3.6 ± 0.4	2.4	1.0	1.6	2.9	6.2	13.4
3404	7.5 ± 0.3	2.1	4.2	5.4	6.9	10.4	14.3
3504	19.2 ± 0.6	2.8	9.5	16.4	19.5	22.0	25.4

Figure 11 shows the same type of information that is in Fig. 6 for customer-premises-to-customer-premises connections, obtained by concatenating the frequency response of the loops to the frequency response of the EOCS connections. The concatenation was done, using the 1980 Loop Survey data, using the same technique described in the previous section. The mean frequency response of the loops (two per connection) used in the concatenation is also shown in Fig. 11. Loop effects clearly dominate trunk effects in end-to-end frequency response.

3.3 Envelope delay distortion

Envelope delay is defined as the negative of the derivative of the phase of the received signal with respect to frequency. *Envelope Delay Distortion* (EDD) at a given frequency is defined as the difference

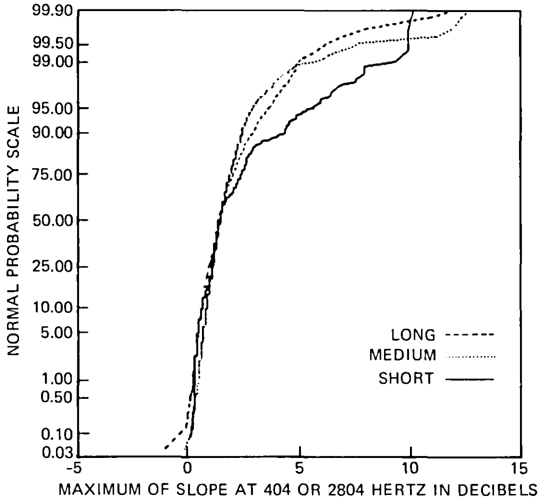


Fig. 7—CDFs of the maxium of the two gain slopes at 404 and 2804 Hz relative to 1004 Hz for short, medium, and long mileage bands.

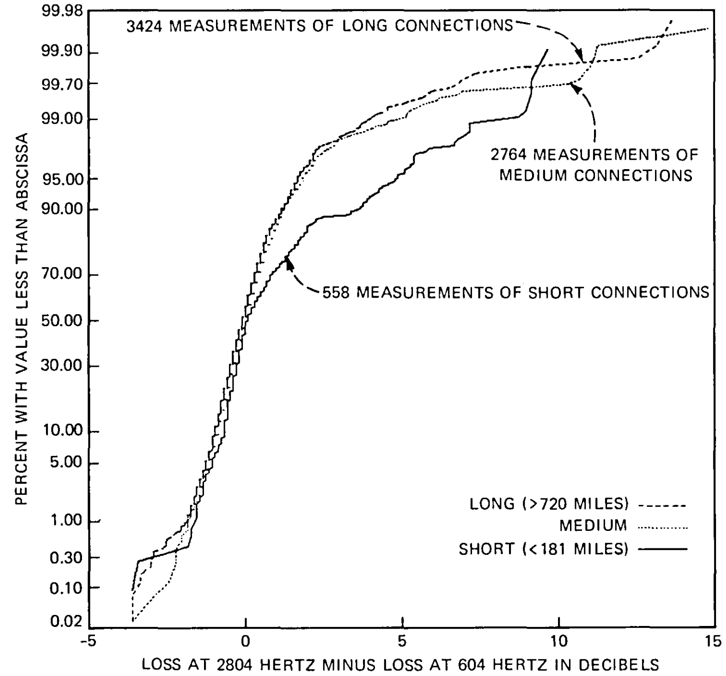


Fig. 8—CDFs of loss at 2804 Hz minus loss at 604 Hz for short, medium, and long mileage bands.

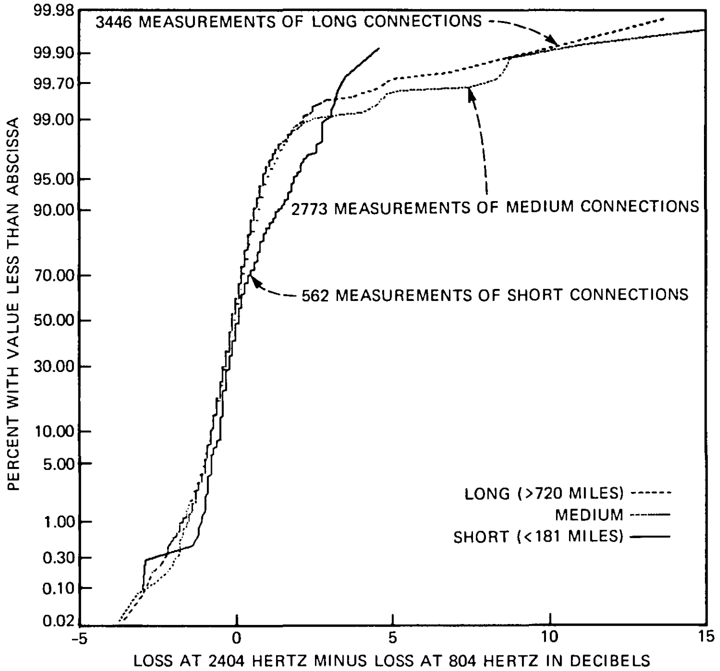


Fig. 9—CDFs of loss at 2404 Hz minus loss at 804 Hz for short, medium, and long mileage bands.

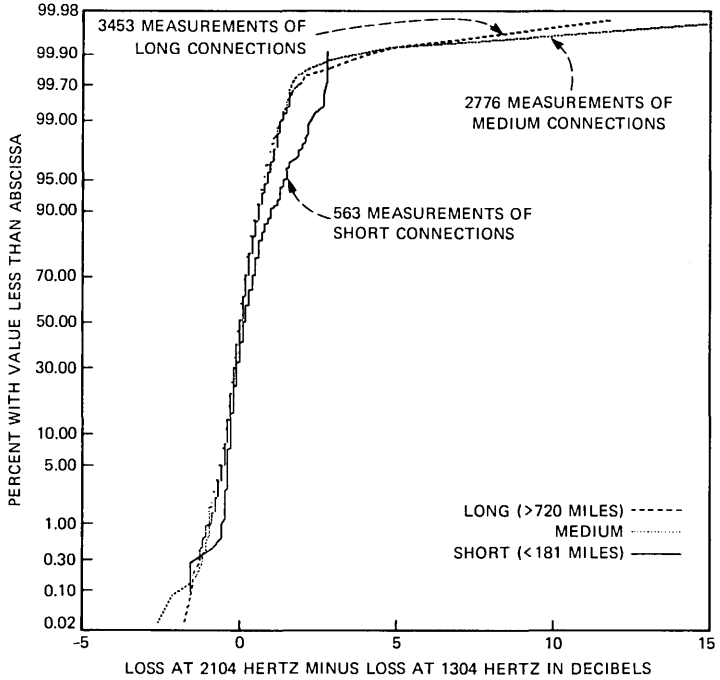


Fig. 10—CDFs of loss at 2104 Hz minus loss at 1304 Hz for short, medium, and long mileage bands.

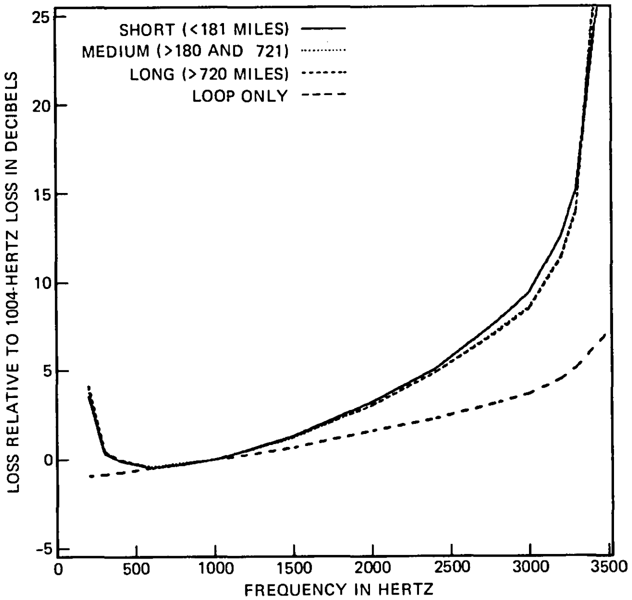


Fig. 11—Mean customer-premises-to-customer-premises attenuation distortion relative to 1004 Hz obtained by analytically concatenating the EOCS result and the 1980 Loop Survey result. The 1980 Loop Survey result is also shown.

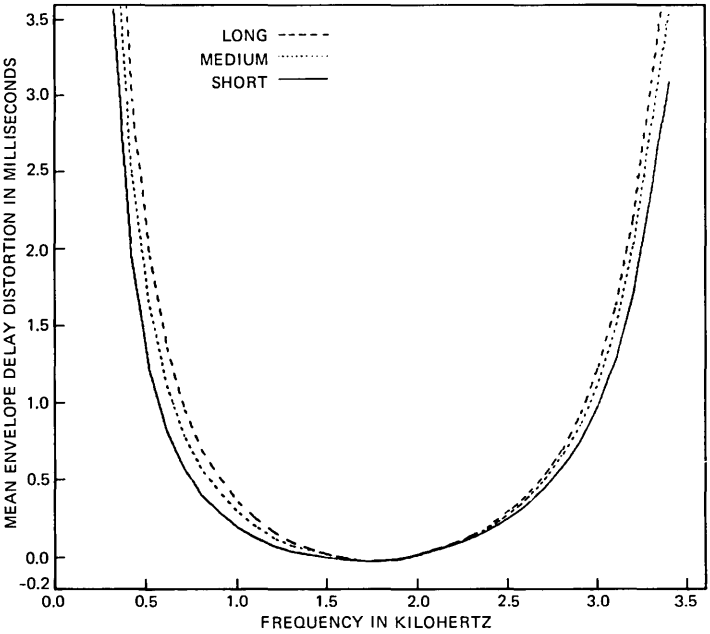


Fig. 12—Mean EDD for short, medium, and long mileage bands.

between the envelope delay at that frequency and the envelope delay at the reference frequency (usually between 1600 and 1800 Hz), where the delay is near the minimum value. The reference frequency used in the EOCS was 1704 Hz.

The test signal used for envelope delay distortion measurement in the United States, which is also used in the EOCS, is a voiceband carrier frequency amplitude modulated (50 percent) by an 83-1/3 Hz tone. A return reference path is required to establish the phase reference so that the phase of the transmitted 83-1/3 Hz envelope can be compared to the phase of the received 83-1/3 Hz envelope. Since the frequency aperture of the modulated test signal remains fixed at twice the 83-1/3 Hz as the carrier frequency of the modulated test signal is varied, the recovered phase changes give estimates of the slope of the phase-versus-frequency curve (EDD). The more desirable direct measurement of phase versus frequency is not made because of the possibility of frequency shift on the facility.

The effect of a nonlinear phase-versus-frequency characteristic (as measured by EDD) on a data signal is such that the different frequency components of the signal have different transit times, which results in distortion in the received signal. The effect of EDD on data signals can be compensated by employing equalizers in the data set receiver. The effectiveness of an equalizer depends on the equalization scheme used—fixed or adaptive—and the complexity (e.g., number of taps) in the equalizer.

Tables IX through XI show the means, standard deviations, and selected percentiles of EDD versus frequency for the short, medium, and long connections. As in the 1969/70 Connection Survey, 1704 Hz was selected as the reference frequency for EDD measurements in the EOCS. Figure 12 shows the mean EDD versus frequency for the short, medium, and long mileage categories. Although not shown in Fig. 12, the mean EDD for the 1969/70 Connection Survey is practically coincident with the short mileage category.

Figure 13 shows the CDFs of the larger of the EDD values at 604 and 2804 Hz for the short, medium, and long mileage categories. Figure 14 is a scatter plot of EDD at 604 Hz compared to that at 2804 Hz, which shows that there is little dependence (correlation coefficient of 0.53) between the EDDs at the two frequencies. Figures 15 and 16 show that there is even less correlation between EDD and loss at these frequencies (correlation coefficient of 0.32 at 604 Hz and 0.37 at 2804 Hz).

3.4 Peak-to-average ratio

Peak-to-Average Ratio (P/AR) measurements are made on a straightaway basis with a transmitter and a receiver attached at

Table IX—Envelope delay distortion relative to 1704 Hz (all statistics expressed in μs): short connections

Frequency in Hz	Mean	Standard Deviation	Quantiles				
			1%	25%	50%	75%	99%
304	3580	1221	1103	3029	3324	4226	6725
404	1959	684	492	1695	1876	2185	3885
504	1242	460	108	1114	1209	1365	2492
604	839	334	15	748	819	922	1749
704	577	258	−71	506	556	637	1278
804	409	190	−35	356	398	456	922
904	298	148	−71	254	293	333	700
1004	205	121	−137	161	201	239	526
1104	136	92	−125	95	128	165	375
1204	87	79	−121	44	77	111	273
1304	54	53	−67	20	48	76	199
1404	36	47	−119	12	32	52	145
1504	20	37	−150	5	19	32	91
1604	7	36	−85	−1	7	14	66
1704	0	0	0	0	0	0	0
1804	6	32	−78	−2	4	11	69
1904	21	32	−92	12	21	29	105
2004	50	39	−58	35	49	61	146
2104	82	47	−39	63	84	98	216
2204	119	57	−27	92	122	142	259
2304	166	65	8	136	165	197	329
2404	220	88	36	177	212	261	493
2804	609	217	102	541	584	727	1021
2904	787	248	136	704	766	909	1243
3004	1017	312	159	949	1008	1176	1567
3104	1312	405	198	1248	1307	1466	1967
3204	1734	551	226	1651	1744	1951	2606
3304	2410	772	253	2304	2470	2625	3739
3404	3132	1102	295	3062	3256	3443	5189

opposite ends of a connection. The transmitter generates a precisely controlled complex waveform of known peak-to-average ratio. The energy in the waveform is dispersed in time by the bandwidth reduction and envelope delay distortion encountered on the connection in a way that may be directly related to intersymbol interference (eye closing) of data signals.[9] The P/AR receiver measures the peak and full-wave average values of the waveform and displays their ratio on a zero-suppressed scale. A P/AR value of 100 suggests no pulse degradation.

The P/AR signal is largely insensitive to noise, phase jitter, and intermodulation distortion, and is unaffected by frequency shift or transient phenomena. P/AR does not produce unambiguous diagnostic information, so there are no externally published requirements for P/AR. Since P/AR ignores transients, P/AR readings cannot predict data set error performance on connections where the transients dominate data set performance. P/AR values may be the same for different EDD shapes occurring on real connections, and with the addition of

Table X—Envelope delay distortion relative to 1704 Hz (all statistics expressed in μs): medium connections

Frequency in Hz	Mean	Standard Deviation	Quantiles				
			1%	25%	50%	75%	99%
304	4290	1292	2787	3261	3755	4951	7819
404	2497	936	1579	1830	2101	2959	5465
504	1633	631	1016	1191	1340	2009	3706
604	1129	457	661	809	912	1403	2654
704	794	339	418	557	638	977	1904
804	574	249	275	401	456	735	1371
904	427	189	190	299	339	553	1034
1004	306	152	109	207	240	402	775
1104	208	112	34	132	161	279	543
1204	135	97	−18	77	106	183	378
1304	86	69	−40	46	70	122	265
1404	61	78	−31	30	48	83	182
1504	37	56	−29	18	30	50	115
1604	16	46	−31	6	12	22	55
1704	0	0	0	0	0	0	0
1804	1	71	−45	−9	−1	7	35
1904	15	27	−53	3	16	26	79
2004	45	46	−45	26	44	59	153
2104	85	65	−33	59	80	98	252
2204	126	77	−17	93	115	143	353
2304	175	109	3	130	155	197	458
2404	235	153	24	174	204	261	585
2804	692	265	264	557	596	768	1588
2904	888	308	383	723	772	977	1950
3004	1156	377	551	955	1007	1261	2437
3104	1523	481	757	1268	1321	1646	3175
3204	2048	638	1079	1697	1784	2178	4242
3304	2828	843	1694	2352	2507	2964	5558
3404	3585	877	2277	3093	3285	3724	6320

sophisticated EDD adaptive equalizers in data sets, the utility of P/AR has diminished. P/AR serves as a quick straightaway measure of the relative bandwidth reduction and EDD on a significant percentage of connections.

Figure 17 shows that there is almost no relationship between P/AR and connection mileage except that there are almost no short connections with a P/AR rating below 70. The figure also shows rare P/AR values above 100, which usually suggest connections where the loss at the band edges is less than that at the center of the band. Such connections increase the peak value of the received P/AR signal relative to the average value. Such a connection in tandem with a "normal" connection (which has higher loss at the band edges) will improve the P/AR value over that of the normal connection alone.

Figures 18 through 20 are scatter plots of P/AR versus the maximum of EDDs at 604 and 2804 Hz, which are test frequencies for network performance objectives. These frequencies have no special relationship

Table XI—Envelope delay distortion relative to 1704 Hz (all statistics expressed in μs): long connections

Frequency		Standard	Quantiles				
in Hz	Mean	Deviation	1%	25%	50%	75%	99%
304	5032	1427	2928	3616	4881	6055	8194
404	2997	1245	1633	2028	2398	3624	6714
504	1959	881	1031	1280	1498	2356	4534
604	1366	643	680	875	1035	1640	3222
704	966	469	438	610	729	1165	2292
804	698	350	280	435	521	853	1685
904	515	266	184	317	386	641	1255
1004	371	201	105	224	275	470	921
1104	258	146	53	153	191	328	659
1204	173	112	16	101	133	217	459
1304	111	75	−12	64	89	143	305
1404	73	72	−26	40	57	96	216
1504	42	62	−38	20	32	56	134
1604	16	29	−31	4	12	23	66
1704	0	0	0	0	0	0	0
1804	2	28	−49	−7	2	9	39
1904	20	29	−64	6	21	31	80
2004	51	46	−65	32	51	66	141
2104	90	59	−72	67	88	108	219
2204	132	73	−66	102	125	154	304
2304	181	115	−54	140	166	206	407
2404	239	129	−16	186	214	275	515
2804	726	275	285	567	604	807	1573
2904	954	347	412	745	797	1051	2073
3004	1258	452	588	976	1039	1393	2765
3104	1659	586	836	1277	1374	1857	3657
3204	2224	761	1182	1721	1852	2509	4788
3304	3105	989	1750	2457	2623	3502	6249
3404	3912	1018	2260	3247	3447	4496	6804

to the P/AR signal. These figures show that there is a reasonably strong correlation (particularly for longer connection mileages) between EDD at 604 or 2804 Hz (whichever is higher) and P/AR, in that when EDD* is higher, P/AR is lower. The two straight lines (labeled 1 and 2) in these figures are the regression lines of P/AR versus EDD and EDD versus P/AR. The degree to which line 1 differs from line 2 is directly related to the coefficient of correlation: the two lines would coincide if P/AR and EDD were perfectly correlated.

The ellipses of concentration shown in these figures have five parameters: two of them determine the center position; one, the angular orientation; and two, the lengths of the major/minor axes. These parameters were determined so that a uniform, elliptical mass of data points would have the same means, standard deviations, and

* EDD in the remainder of this subsection refers to the maximum (per connection) of EDDs at 604 and 2804 Hz.

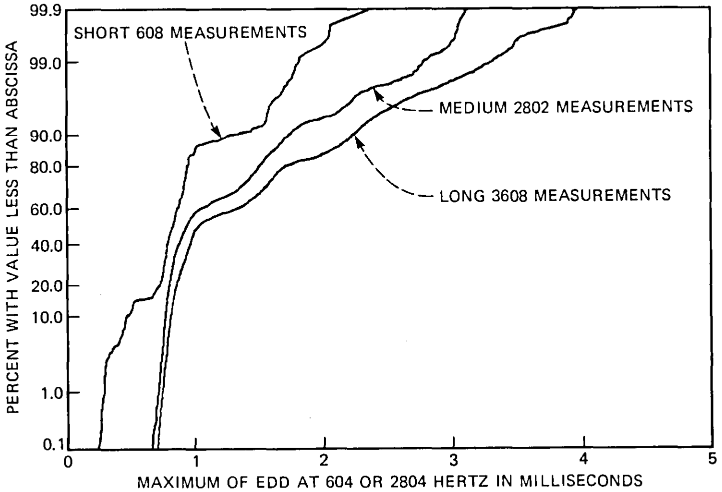


Fig. 13—CDFs of the maximum of the two EDDs at 604 and 2804 Hz relative to 1704 Hz for short, medium, and long mileage bands.

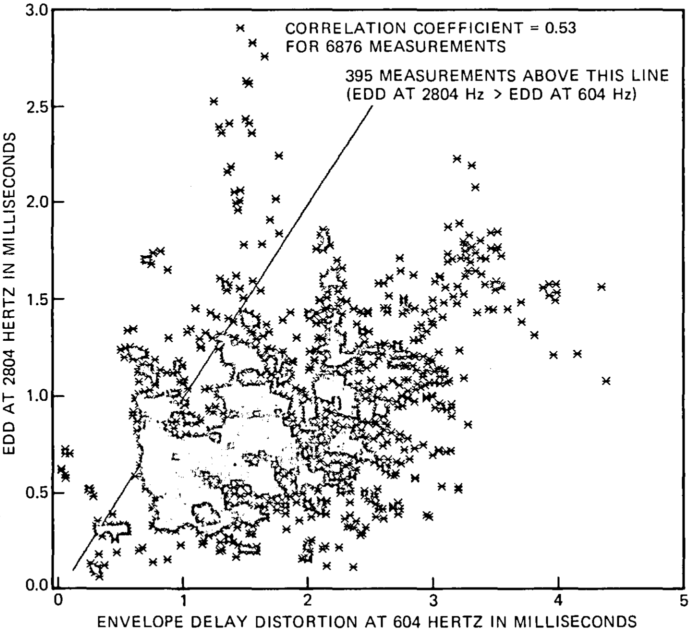


Fig. 14—EDD (relative to 1704 Hz) at 2804 Hz versus EDD (relative to 1704 Hz) at 604 Hz.

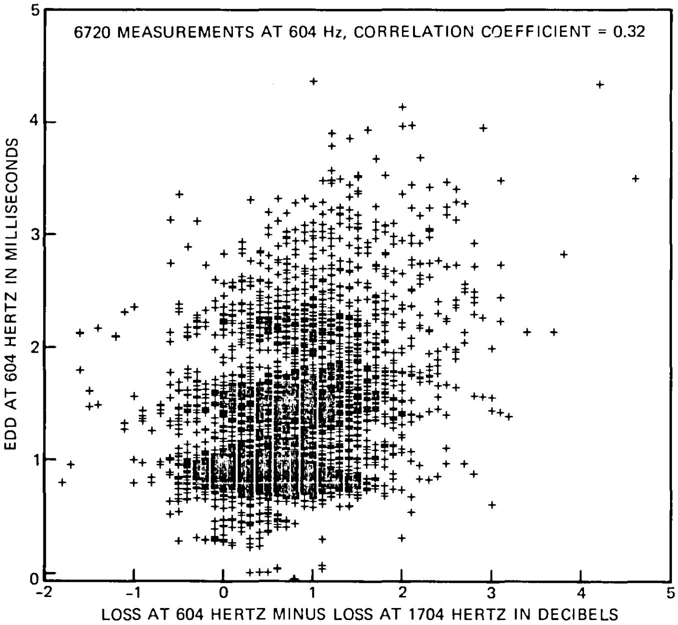


Fig. 15—EDD versus attenuation distortion at 604 Hz, both relative to 1704 Hz.

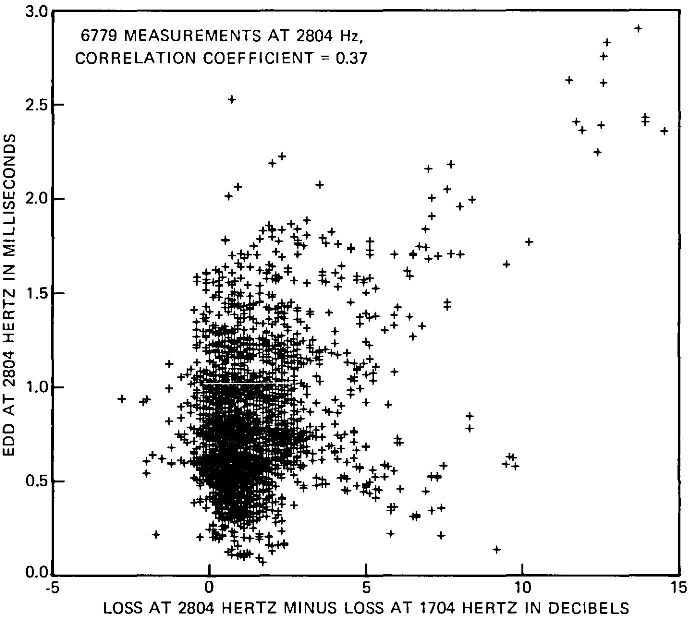


Fig. 16—EDD versus attenuation distortion at 2804 Hz, both relative to 1704 Hz.

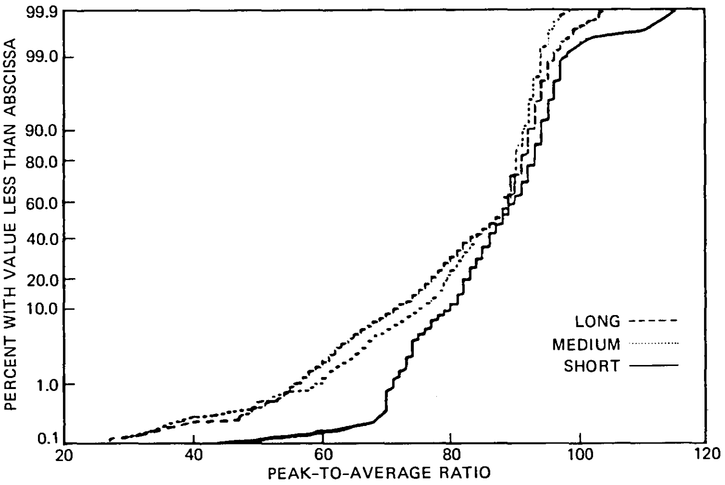


Fig. 17—CDFs of P/AR for short, medium, and long mileage bands.

correlation coefficient as the original data. The ellipse of concentration has two horizontal and two vertical tangents passing through the points where, respectively, line 1 and line 2 intersect the ellipse. The distance between the two horizontal tangents is equal to four times the standard deviation of P/AR. Similarly, the distance between the

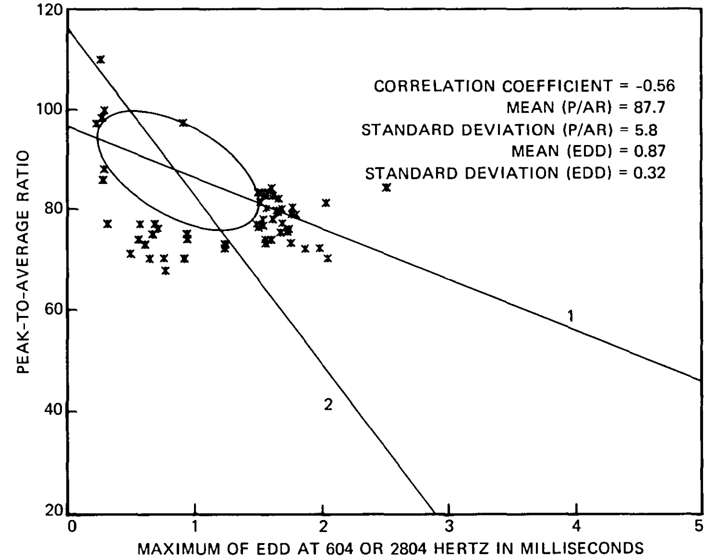


Fig. 18—P/AR versus the maximum of the EDDs at 604 and 2804 Hz for the short mileage band (563 measurements). Eighty-six percent of the points fall within the ellipse and are not plotted.

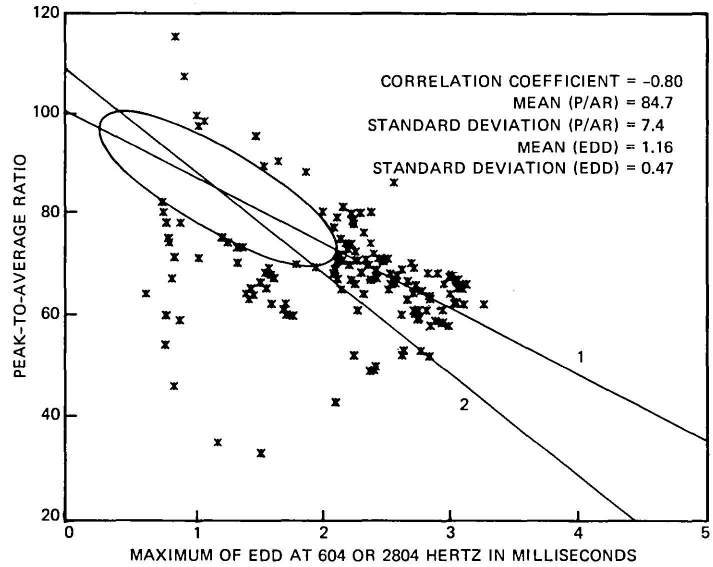


Fig. 19—P/AR versus the maximum of the EDDs at 604 and 2804 Hz for the medium mileage band (2586 measurements). Ninety percent of the points fall within the ellipse and are not plotted.

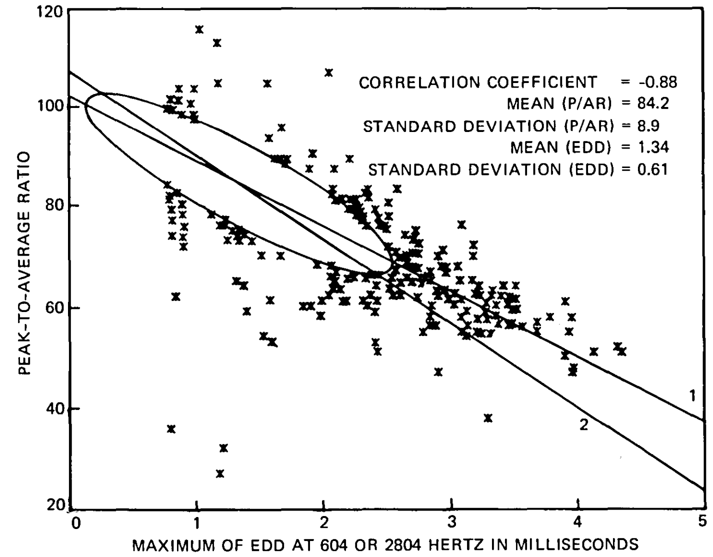


Fig. 20—P/AR versus the maximum of the EDDs at 604 and 2804 Hz for the long mileage band (3218 measurements). Ninety percent of the points fall within the ellipse and are not plotted.

two vertical tangents is equal to four times the standard deviation of EDD. Almost 90 percent of the measurements fall within the ellipses.

3.5 Propagation delay

Round-trip propagation delay measurements were made in the EOCS by using full-duplex 1200-b/s data sets (to be described in 3.13) under the control of the microprocessor in the ASPEN RTU. An error was introduced in a continuous repetition of a 511-bit pseudorandom word transmitted by the near-end RTU microprocessor bit error rate generator to the low-band modulator of the full-duplex 1200-b/s data set. When the far-end RTU microprocessor recognized the error from the low-band demodulator of its 1200-b/s data set, it immediately introduced an error in the continuous repetition of the same 511-bit pseudorandom word being transmitted back to the near end by the high-band 1200-b/s data set modulator. When the near-end RTU microprocessor recognized the forced error, it corrected for the known (fixed) processing delay to get a first estimate of round-trip propagation delay. This sequence was repeated nine more times to obtain enough valid measurements to reject those affected by random errors on the connection.

Figure 21 shows boxplots of round-trip propagation delay versus mileage (with the same abscissa as that of Fig. 1), except for the 52 measurements taken on satellite connections.* Round-trip delay measurements on these satellite connections ranged from 508 to 539 ms. The variability in round-trip delay for the same end office building pairs can be attributed to alternate trunk facilities as well as alternate routing. Figure 22 shows CDFs of round-trip delay for the three mileage bands. In Figs. 21 and 22, measurements on satellite connections were excluded from these CDFs with the 52 measurements on satellites excluded. The boxplots and the CDFs show a strong dependence of round-trip delay on airline mileage.

3.6 Message circuit noise and signal-to-C-notched noise ratio

Message circuit noise was measured both with and without a 1004-Hz holding tone, with both the C-message and 3-kHz flat weighting filters, as Ref. 1 specifies. The C-message filter weighting characteristic was derived in 1957 from tests made with subjects assessing the interfering effects of single frequency interference as heard over an ordinary telephone handset. This weighting is also appropriate for high-speed data transmission because most high-speed modems concentrate their transmitted energy in approximately the same band of sensitivity as the C-message filter. The ac power-line hum at 60 Hz

* Propagation delay is the only parameter in this paper for which measurements on satellite connections are treated separately.

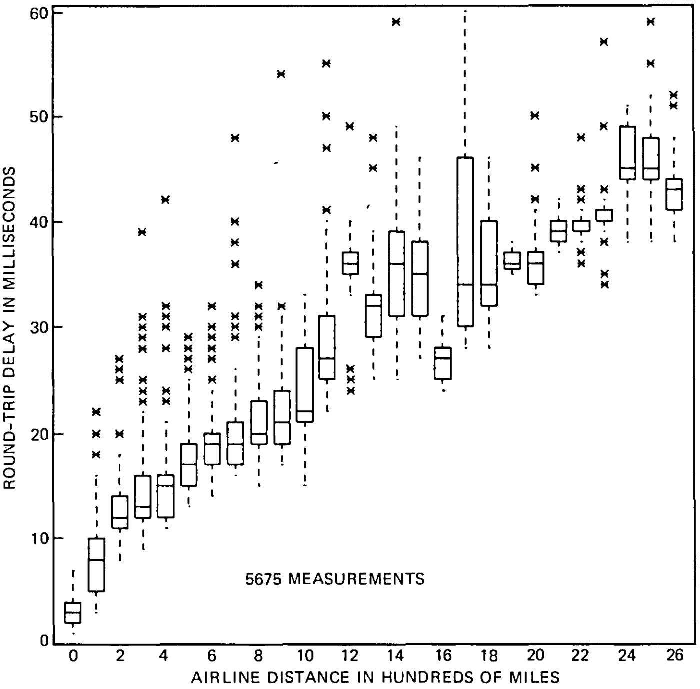


Fig. 21—Round-trip propagation delay versus connection airline mileage in 100-mile blocks.

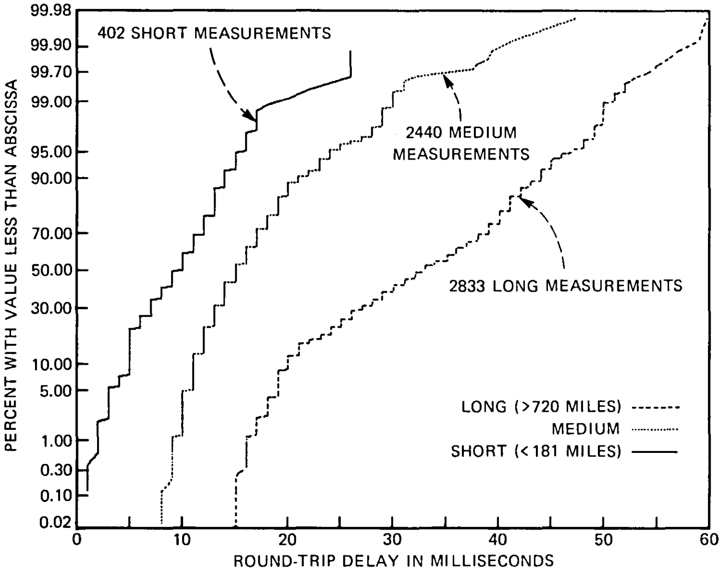


Fig. 22—The cumulative distributions of round-trip propagation delay.

(and odd harmonics of 60 Hz), frequently encountered in the loop plant, is attenuated by the C-message filter but is included in the noise measurements made with the 3-kHz flat filter. All noise measurements were reported as they were measured, without corrections for losses in the office in which the measurements were made, which permits direct comparison with the 1969/70 Connection Survey,[4] and concatenation with the loop plant.[8]

If a connection has facilities with digital channel banks or compandors, noise on that connection can be substantially different, depending on whether it is measured with or without a holding tone. For data transmission, therefore, C-notched noise is more relevant than C-message noise. C-notched noise is obtained by filtering out the 1004-Hz holding tone with a deep (50-dB) notch filter and then measuring C-message noise. Signal-to-C-notched-noise ratio (s/n) is the ratio of the received 1004-Hz holding tone power to the C-notched noise power.

The s/n is an essential performance measure for digital channel banks. As Fig. 23 shows, digital channel banks have an approximately logarithmic (μ255) encoder/decoder that maintains a nearly constant s/n over a reasonable range of signal levels. Because most of the connections measured in the EOCS have at least one T-carrier link, the s/n results from the EOCS are dominated by this characteristic. This can be observed in the figures that follow.

Figure 24 shows the CDFs of s/n for the three mileage bands. The CDFs confirm mileage dependence of s/n, particularly in the region of "good" s/n or the upper tails. Comparison of these CDFs with similar CDFs obtained from the 1969/70 Connection Survey shows that the variability of s/n on short connections has been reduced substantially since the last survey. This tighter s/n distribution for short connections in the EOCS is caused by the introduction of T-carrier systems in the network since the 1969/70 Connection Survey. The same comparison suggests that the percentage of short connections with s/n better than 40 dB—the s/n ceiling for digital channel banks observed in Fig. 23—was greater in the 1969/70 Connection Survey than in the EOCS. However, this s/n degradation, which can also be attributed to T-carrier facilities, is largely inconsequential to the performance of data sets in the region where it occurs, i.e., the upper tail of the CDF. Listeners are usually unable to discern differences in s/n ratios above 40 dB.

Figure 25 shows the Probability Density Functions (PDFs) of s/n for the three mileage bands. The PDF for the short mileage category shows bimodality, whereas the PDFs for the medium and long mileage categories are unimodal. The two peaks for the short category occur at 34 and 38 dB. The peak at 38 dB is consistent with the noise expected on a connection with one T-carrier system with one digital

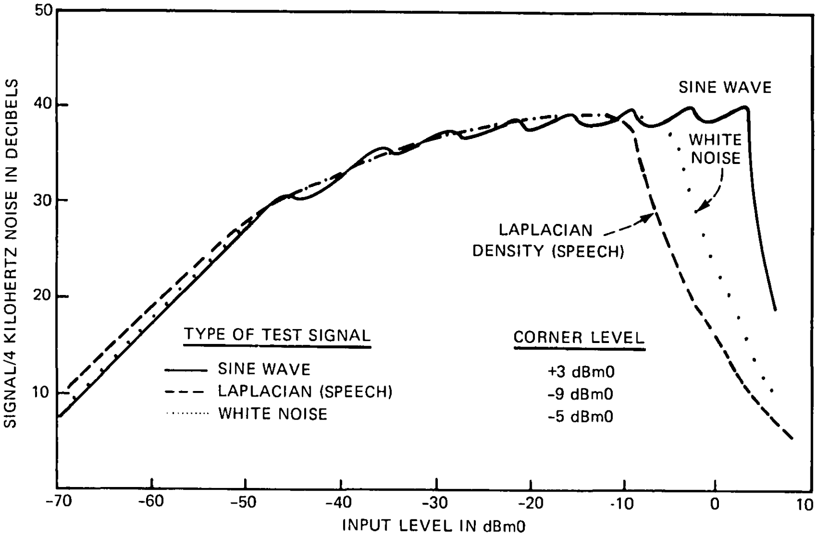


Fig. 23—Signal-to-distortion performance of 8-bit μ=255 coder-decoder (15-segment approximation, 8 segments +, 8 −).

channel bank at each end—one analog-to-digital (A/D) and one digital-to-analog (D/A) conversion, i.e., an end-to-end digital connection or an analog connection with exactly one digital facility or digital switch; see Fig. 23. The lower peak is consistent with occurrences of two digital links in tandem.

Figure 26 shows a scatter plot of C-notched noise versus C-message

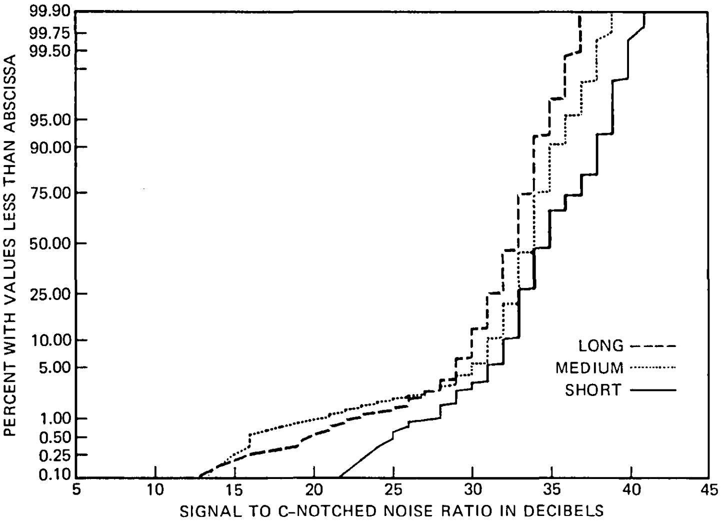


Fig. 24—CDFs of signal-to-C-notched-noise ratio for short, medium, and long mileage bands.

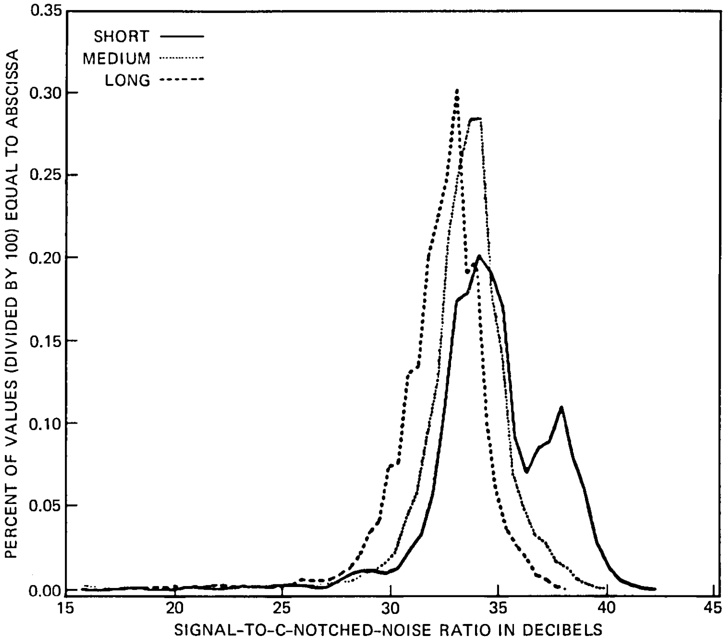


Fig. 25—PDFs of signal-to-C-notched-noise ratio for short, medium, and long mileage bands.

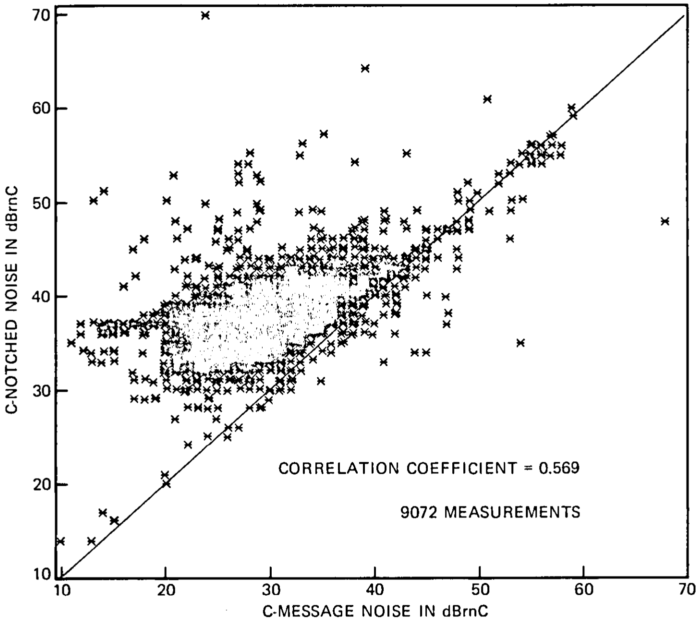


Fig. 26—C-notched noise versus C-message noise.

noise. Also shown in the figure is a straight line on which C-notched noise is equal to C-message noise. The large cluster of points above the line, corresponding to the connections with higher C-notched noise than C-message noise, show the effects of compandors, quantizing noise, and harmonic distortion on the holding tone. The points scattered far above the line may correspond to the connections where bad coders were encountered or where the C-notched noise measurements were affected by impulse noise. The points along the line represent the connections where the tone had no effect on noise. Small deviations from the line can be expected, considering possible time variation of noise between the two types of noise measurement. The points scattered far below the line indicate the possible effects of impulse noise during the C-message noise measurements.

Figure 27 shows the CDFs of C-message noise for the three mileage bands. The mileage dependence of C-message noise can be observed in the figure, as would be expected for the analog carrier facilities normally encountered on longer trunks. The airline-mileage effect on the C-message noise can also be seen on the boxplots of Fig. 28 (see Fig. 2 for the explanation of the abscissa).

Figure 29 presents the CDFs of C-notched noise for the three mileage bands. Mileage dependence is less apparent with C-notched noise than with C-message noise. In particular, Fig. 29 shows virtually no difference in the CDF of C-notched noise between the medium and long mileage categories. It appears that most connections measured for these mileage categories had T-carrier facilities on the toll-connecting trunks, one at each end, which dominated the connection C-notched noise. The dominance of the toll connecting trunk noise reduces the dependence of noise on connection mileage for the longer connections. Digital switches in tandem with T-carrier links do not degrade the C-notched noise.

It appears that the C-notched noise measurements for the short mileage category consist of measurements from two groups of connections, as evidenced by the bimodality of the s/n PDF in Fig. 25: one group containing two pairs of digital channel banks, and the other containing one pair of digital channel banks. As Fig. 29 shows, the upper half of the CDF for the short mileage category is almost the same as the CDFs for the other two categories, suggesting that it is made up of the measurements from the connections with two pairs of digital channel banks. The lower half of the CDF for the short mileage category, however, is significantly different from those for the other two mileage categories, suggesting that the measurements are from the connections with one pair of digital channel or from Voice Frequency (VF) cable.

Figure 30 shows the CDFs of the 3-kHz flat noise for the three

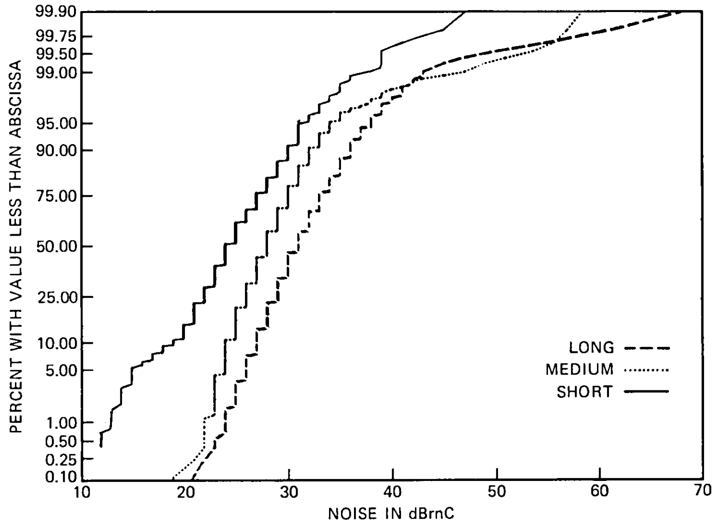


Fig. 27—CDFs of C-message noise for short, medium, and long mileage bands.

mileage bands. As we can see in the figure, the 3-kHz flat noise shows almost no dependence on mileage and is much higher than the C-message noise. This shows that this type of noise is dominated by sources outside the range of the C-message filter, primarily multiples

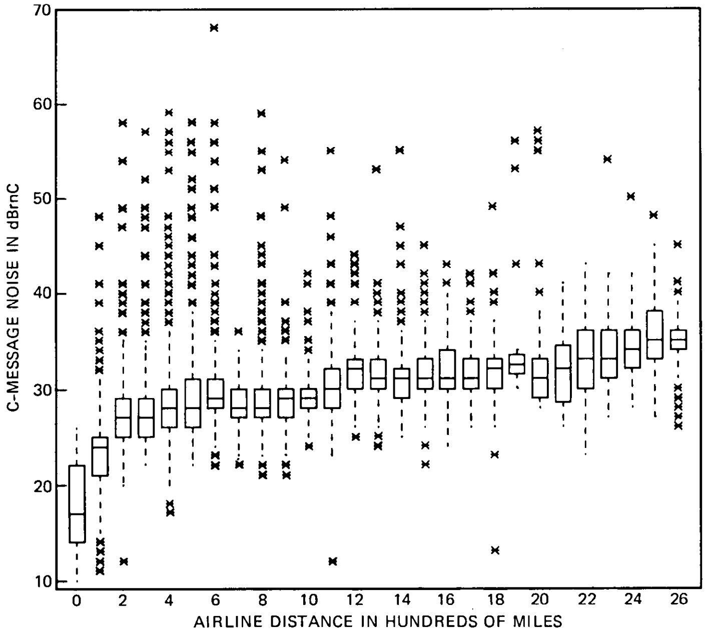


Fig. 28—C-message noise versus airline mileage in 100-mile blocks.

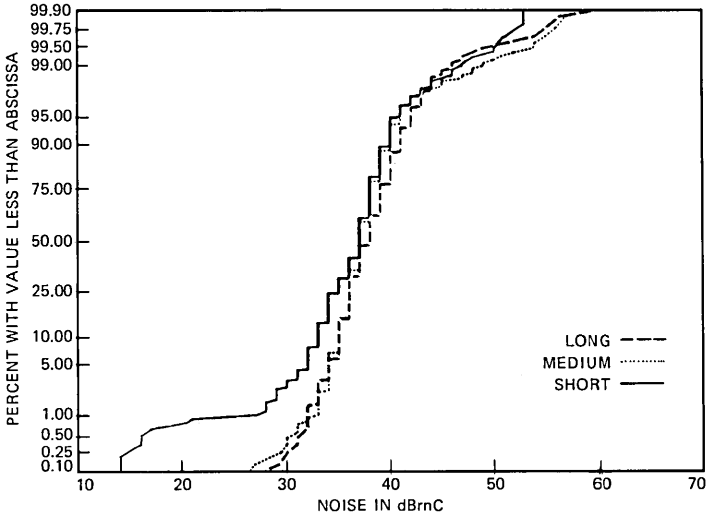


Fig. 29—CDFs of C-notched noise for short, medium, and long mileage bands.

of 60 Hz from the end office line-circuit battery feed. The same remarks made for the 3-kHz flat noise hold for the 3-kHz flat notched noise. Figure 31 shows the CDFs of the 3-kHz flat notched noise for the three mileage bands.

Figure 32 shows the CDFs of the 3-kHz flat noise-to-ground per environment of the end office: urban, suburban, and rural. An urban

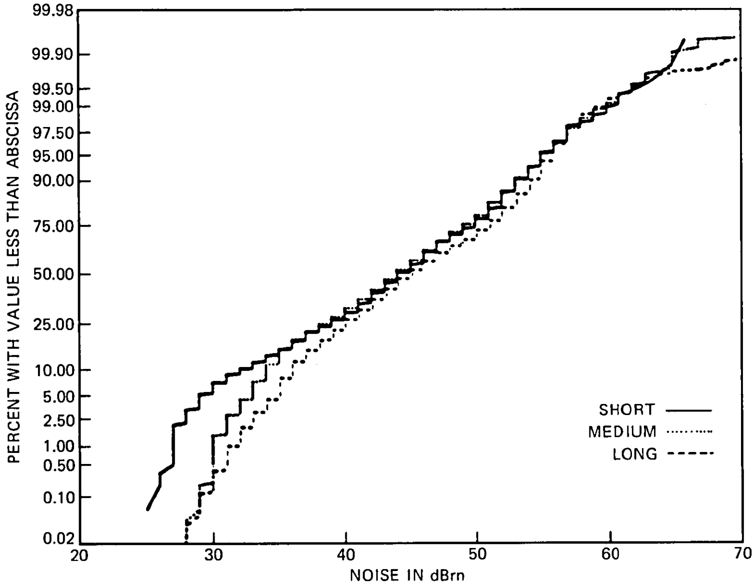


Fig. 30—CDFs of 3-kHz flat noise for short, medium, and long mileage bands.

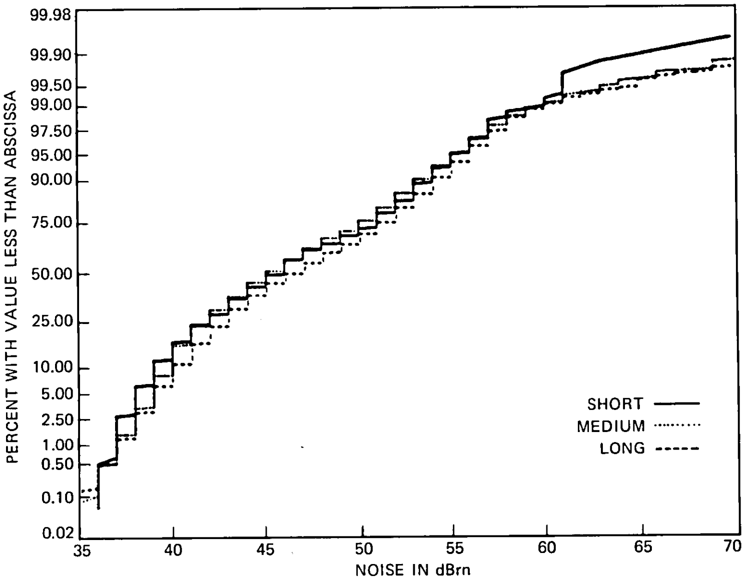


Fig. 31—CDFs of 3-kHz flat-notched noise for short, medium, and long mileage bands.

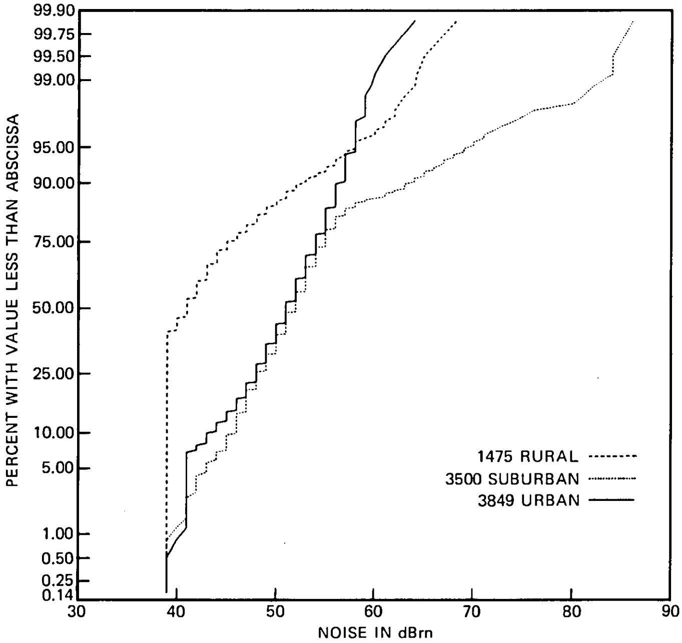


Fig. 32—CDFs of 3-kHz flat noise-to-ground in central offices classified as rural, suburban, and urban.

end office was defined in Ref. 8 as one serving more than 20,000 assigned pairs; a rural end office as one serving fewer than 5000 assigned pairs; and a suburban end office as one serving between 5000 and 20,000 assigned pairs. The bottom of the measuring range for the test equipment for noise-to-ground was 40 dBrn, so measured values below this value were conservatively estimated at 39 dBrn, causing the truncation in the CDFs at that value. It should be recalled that no loops were connected for EOCS measurements. These measurements confirmed that the medians for 3-kHz flat noise-to-ground for trunks in the central office with no loops connected were well below the median contribution for trunks connected to the loop plant[8] (by 12-dB for urban, 20 dB for suburban, and 40 dB for rural central offices).

Figure 33 shows the CDFs of C-message noise on customer-premises-to-customer-premises connections. The calculated customer C-message noise power of a connection was obtained by power-summing the noise power of the connection measured at the end office attenuated by the loss of the loop, and the C-message noise on the loop (customer end). Figure 33 was obtained by analytically concatenating (using discrete convolution techniques) the loop noise from the 1980 Loop Survey to the end office to end office C-messsage noise attenuated by the loop loss. Also included in the same figure for comparison is the CDF of the end office to end office C-message noise for the EOCS medium connection length category. The distribution of the C-message noise can be seen to improve with the addition of the loop. The contribution of the loop loss (which attenuates the noise from the end office) is apparently more important than the effect of the noise on the loop.

Figure 34 shows the CDF of C-message weighted metallic noise for Bell System loops from the 1980 Loop Survey used in the concatenation for Figure 33. (The CDF of 1004-Hz loss for the Bell System loops used in the concatenation appears on Fig. 5.)

3.7 Intermodulation distortion

In the past, a harmonic distortion measurement was used to characterize nonlinearities by applying a single tone to the connection and measuring the received power at the second and third harmonics of the test frequency with a selective detector. However, this type of measurements did not properly characterize nonlinearities as they affected data transmission. The harmonics of the sine wave could cancel one another on connections with multiple nonlinearities, and the PDF of noise-like high-speed data signals is markedly different from that of a single tone.

In the EOCS, nonlinearities were evaluated by an intermodulation distortion measurement using the four-tone method.[1] The four-tone

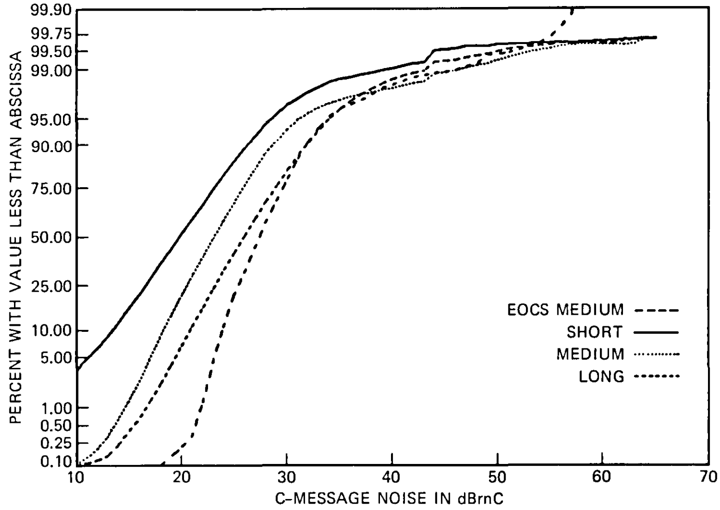


Fig. 33—CDFs of the customer-premises-to-customer-premises C-message noise for short, medium, and long mileage bands.

method is not subject to the cancellation effect, and the PDF of its test signal is a much better approximation to the PDF of a high-speed data modem signal than is a sinusoidal (one-tone) test signal.

The intermodulation distortion measured with the four-tone test signal may contain components caused by background or quantizing

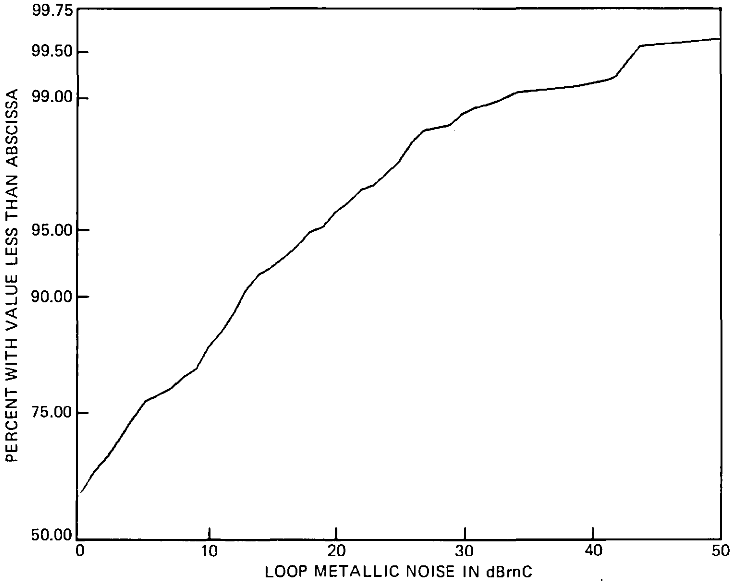


Fig. 34—CDF of C-message weighted metallic noise on customer loops.

noise. To correct for this, the noise component was measured by removing two of the four tones and measuring the energy in the narrow bands where the four-tone intermodulation distortion would fall. The corrected intermodulation distortion was then calculated by power subtraction of the noise component from the original measurement, as outlined in Ref. 1. (For example, if the power measured with four tones was 1 dB larger than that measured with two tones, the true intermodulation distortion would be 7 dB below the value measured with four tones. If the levels measured with four tones and with two tones were the same, the conservative value of 8 dB was subtracted from the four-tone intermodulation distortion measurement.)

Figures 35 and 36 show the CDFs of second- and third-order intermodulation distortion for the three mileage bands. The abscissa shows intermodulation distortion expressed as a signal-to-distortion ratio in decibels, and thus higher values on the abscissa represent better performance. As we can see in the figure, there is little dependence on mileage, particularly for the medium and long mileage categories. Intermodulation distortion on connections in these mileage categories could have been contributed mostly by central office equipment, such as multiplexors, whose appearance is only weakly correlated with mileage.

The scatter plot of third-order versus second-order intermodulation distortion of Fig. 37 shows moderate correlation between the two parameters.

3.8 Phase and amplitude jitter

Phase jitter is the deviation or "jitter" of zero-crossings of a 1004-Hz tone from their nominal position in time. Phase jitter was measured by comparing the average phase of the signal (determined by a phase-locked loop) and the instantaneous phase of the received signal. The normal bandwidth for the measurement of (demodulated) phase jitter is 20 to 300 Hz. In the EOCS, phase jitter was measured in two bands: 20 to 300 Hz and 2 to 300 Hz. The bandwidth of the phase jitter detector in the transmission test set used in the EOCS extends below the recommended 4-Hz corner of Ref. 1.

Amplitude jitter is the deviation or "jitter" of the peak of a 1004-Hz tone from its nominal value. Amplitude jitter was measured with the same two frequency bands as the phase jitter, and the phase and amplitude jitter circuits used the same post-detection filter and peak detector.

Figures 38 and 39 show the CDFs of phase jitter for the 20- to 300-Hz band and for the 2- to 300-Hz band, respectively, for the three mileage bands. The figures show dependence of phase jitter on mileage.

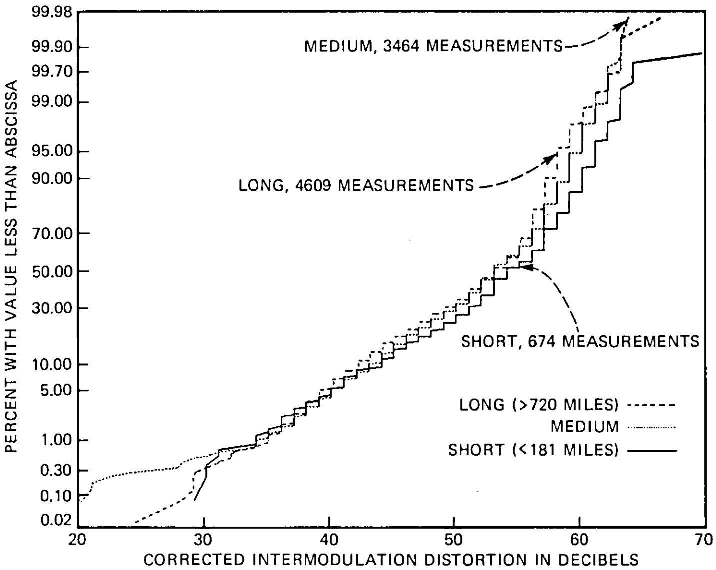


Fig. 35—CDFs of second-order intermodulation distortion for short, medium, and long mileage bands.

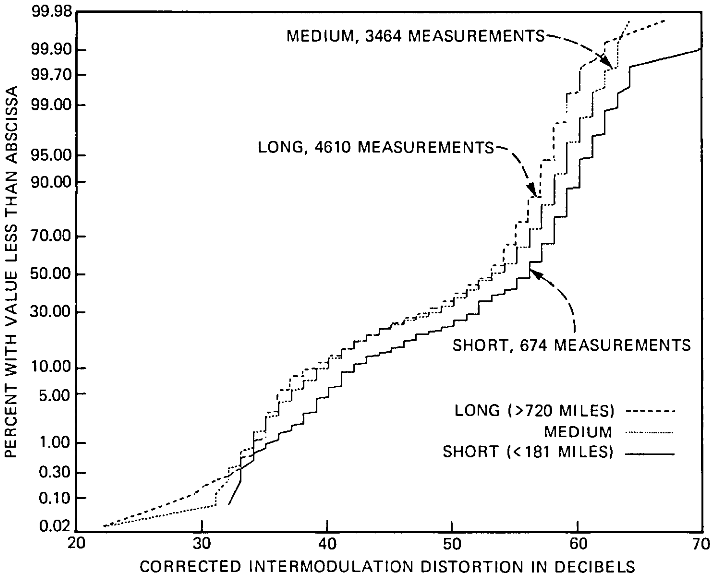


Fig. 36—CDFs of third-order intermodulation distortion for short, medium, and long mileage bands.

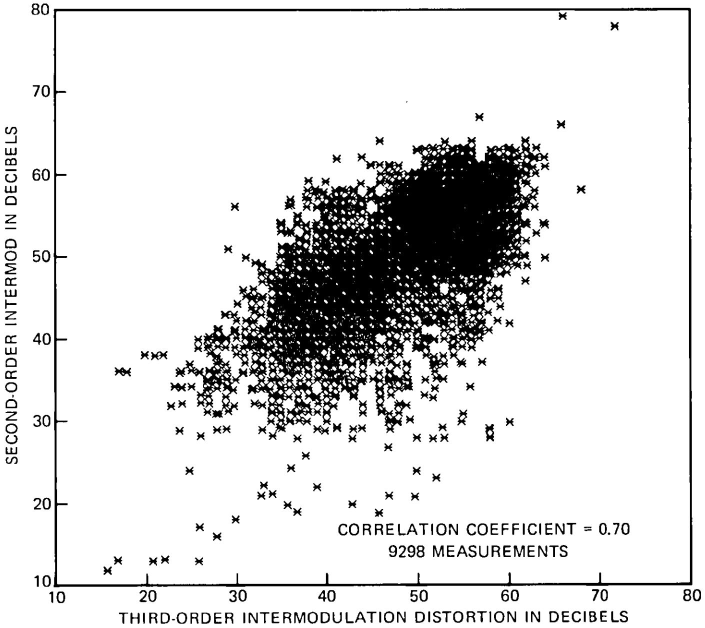


Fig. 37—Second-order versus third-order intermodulation distortion.

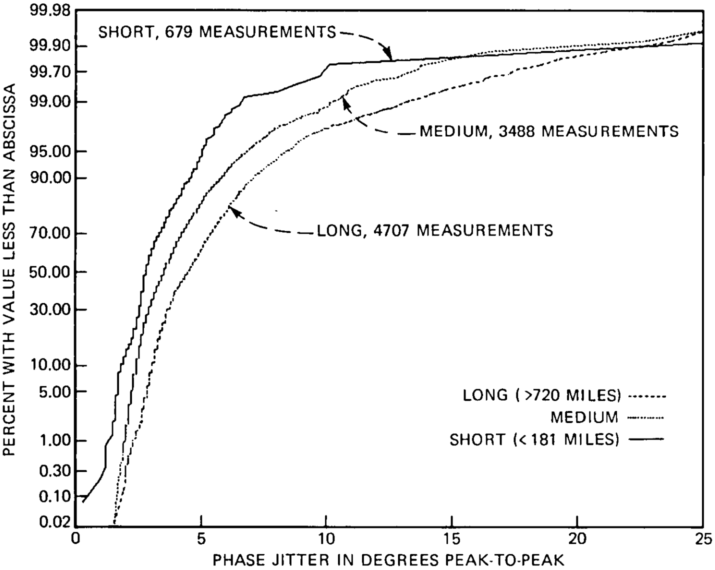


Fig. 38—CDFs of 20- to 300-Hz phase jitter for short, medium, and long mileage bands.

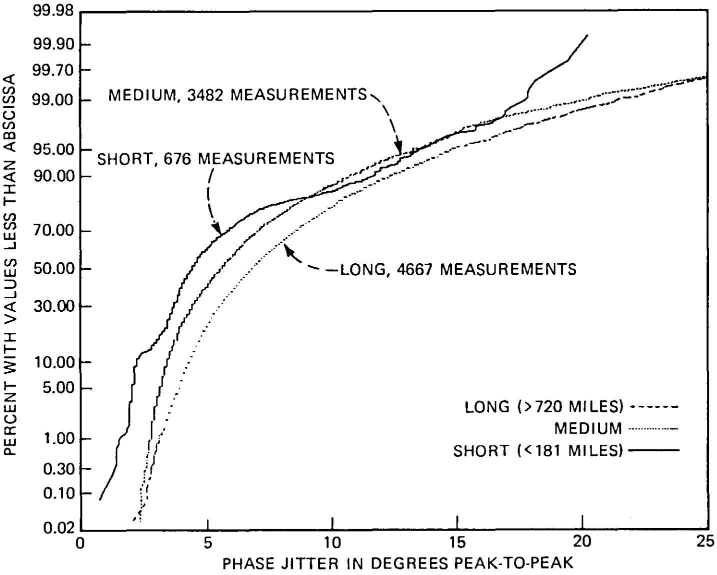


Fig. 39—CDFs of 2- to 300-Hz phase jitter for short, medium, and long mileage bands.

Figures 40 and 41 present the CDFs of amplitude jitter for the three mileage categories, for the two frequency bands, respectively.

Phase jitter can be caused by phase modulation as well as by noise. Measurement of phase jitter is appropriate to predict high-speed data set performance, but only as an indirect measure of phase modulation. However, the phase jitter measuring set alone cannot distinguish phase jitter caused by noise from that caused by phase modulation. The primary purpose of the amplitude jitter measurement is to separate phase jitter caused by the two sources.

Although both noise and amplitude modulation can cause amplitude jitter, a signal is unlikely to encounter amplitude modulation sources in the network, leaving noise as the sole source of amplitude jitter. On the other hand, the network has both sources for phase jitter, namely, noise and phase modulation. Therefore, the *amplitude* jitter measurements can be compared with *phase* jitter measurements to distinguish phase jitter caused by the two sources. For example, a high phase jitter measurement accompanied by a low amplitude jitter measurement on a connection is an indication that the phase jitter on that connection is not caused by noise, but most likely is caused by phase modulation.

Figure 42 is a scatter plot of the 20- to 300-Hz amplitude jitter versus the 20- to 300-Hz phase jitter. Also shown in the figure are two demarcation lines (labeled A and B) experimentally obtained by testing the phase and amplitude jitter measurement equipment used in the

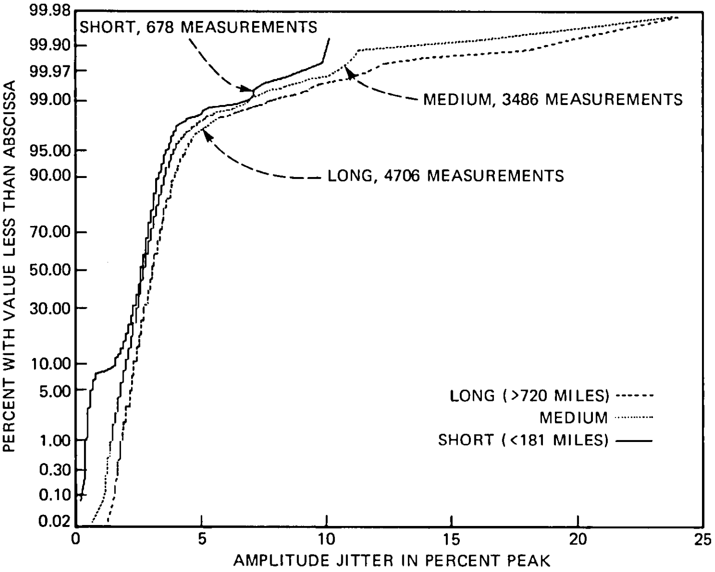


Fig. 40—CDFs of 20- to 300-Hz amplitude jitter for short, medium, and long mileage bands.

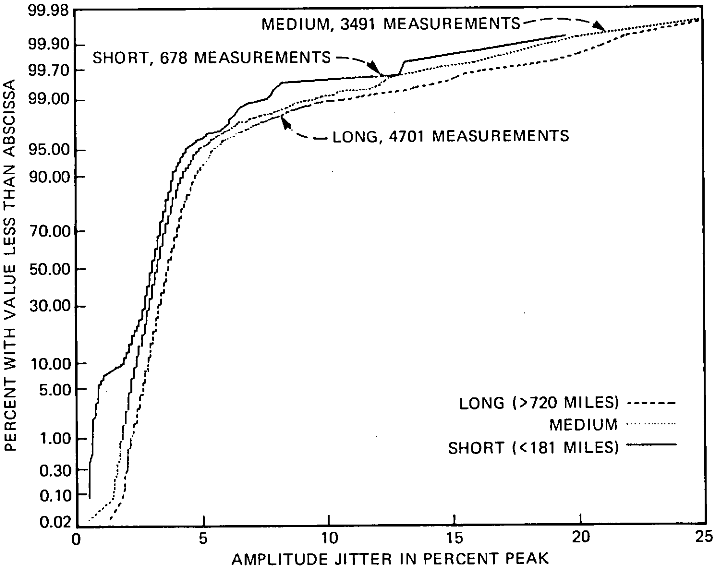


Fig. 41—CDFs of 2- to 300-Hz amplitude jitter for short, medium, and long mileage bands.

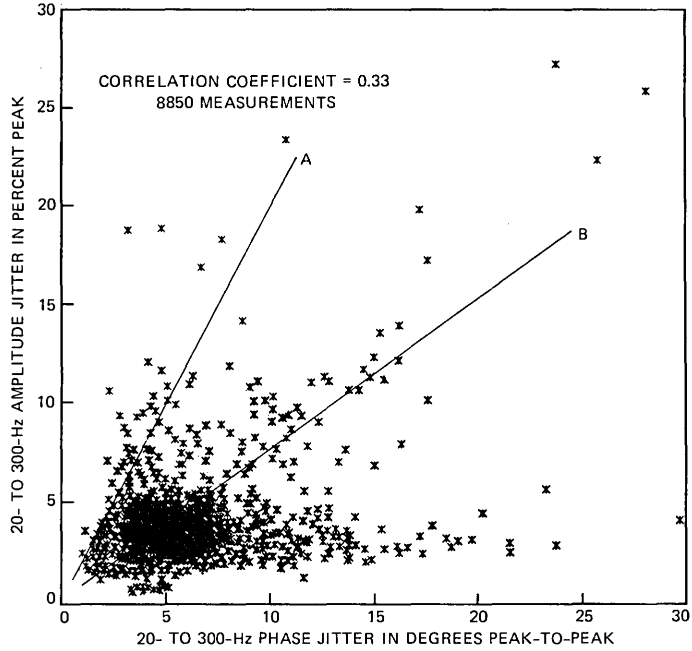


Fig. 42—Amplitude versus phase jitter in the 20- to 300-Hz band.

EOCS. Points between lines A and B indicate connections where phase and amplitude jitter show good correlation. Therefore, these points almost all correspond to phase jitter solely caused by noise. Points above line A show connections with high amplitude jitter but little phase jitter, and they are attributed to the effect of impulse noise on the amplitude jitter detector. (The requirements for all jitter detectors mandate peak detectors, which also respond to momentary increases from impulse noise.) Points below line B, showing connections with high phase jitter but low amplitude jitter, suggest that phase jitter on those connections is largely caused by phase modulation, and is occasionally caused by impulse noise.

3.9 Frequency shift

Frequency shift, or absolute frequency offset, is a critical parameter, for example, for the proper functioning of echo cancelers. Echo cancelers continue to adapt during voice calls and can track frequency shifts below 1 Hz. Depending on the magnitude of the frequency shift, brief transient echoes may be heard after conversational pauses, since adaptation only occurs when just one party is talking. Since echo cancelers freeze on calls where a continuous data set signal is present, frequency shift will cause echoes whose magnitudes change at the frequency shift rate.

Frequency shift was measured in the 1969/70 Connection Survey by transmitting at −12 dBm a 1200-Hz tone whose frequency was known to 0.1 Hz. The frequency of the received tone was measured to a precision of 0.1 Hz at the far end of the connection leading to an overall accuracy of approximately 0.2 Hz. The difference between the two frequencies was the frequency shift of the connection. The measured frequency shifts were not normally distributed. An offset of greater than 3 Hz was observed on two of the 600 measurements.

In the EOCS, the transmitted frequencies were known to a precision of 0.1 Hz. The received frequencies were measured to a resolution of 1 Hz with a frequency counter that could be momentarily driven upward by impulse noise, and could momentarily be driven downward by transient power line harmonics. Connections with poor s/n caused a positive 1-Hz offset in the frequency counter output. Taking only those connections for which multiple, stable frequency shifts were observed:

1. A positive frequency shift of 2 Hz was observed on eight of 4222 connections (0.19 percent).
2. A negative frequency shift of 1 Hz was observed on ten of 4222 connections (0.24 percent).

All but one of the stable frequency shifts observed were for connections to a single end office that had N3 (frequency division multiplex) carrier toll-connecting trunks. The precision of the frequency measurements in the EOCS is not sufficient to draw conclusions about the performance of the network for frequency shift, particularly since the observed shifts were associated with the toll-connecting trunks for a single office.

3.10 Impulse noise

Impulse noise was measured through a C-notched filter by counting the number of times the noise exceeded a given threshold. The impulse noise measurement consisted of three five-minute measurements in sequence, with three different thresholds for each five-minute interval.

The thresholds for impulse noise counts were set based on the received rms level of the 1004-Hz holding tone. For the first five minutes, thresholds were set at −12, −8, and −4 dB relative to the received holding tone level; for the second five minutes, at −12, −4, and +4 dB; for the last five minutes, at −8, 0, and +8 dB. Since measurements at the −12, −8, and −4 dB thresholds were made twice for five minutes on each connection, while measurements at the 0, +4, and +8 dB thresholds were made only once for five minutes per connection, all figures and results in this section are based on twice as many observations for the −12, −8, and −4 dB thresholds than for the other three higher thresholds.

Impulse noise counters with multiple thresholds have the characteristic that a single impulse exceeding the highest threshold must also register on all the lower thresholds. This means that, in any given five-minute interval, the counter for the lowest threshold will have the same as or higher count than the counter for a higher threshold. As one would expect, the CDF of impulse noise counts at the −12 dB threshold falls to the right of the CDF at the −8 dB threshold. Since the thresholds were changed for each of the three five-minute transient measuring intervals, it is possile (but unlikely) that on any given connection, there could be a smaller count for the 0-dB threshold than for the 4-dB threshold, for example.

Statistically, the effect of connection airline mileage was found not to be significant on impulse noise count. The type of switch at either end of the connection was found to be a significant factor. Figures 43 through 45 show the CDFs of impulse noise counts per five-minute interval at various thresholds for different switch types (digital, crossbar, and step-by-step switches) at the measuring end office. The type of switch in the end office from which the tone was sent was taken into consideration through a weighting process based on predivestiture Bell System traffic statistics. A connection with a digital switch at the measuring end is expected to have fewer impulse noise counts than a connection with an electromechanical switch (crossbar, or step-by-step) in which the operation and release of adjacent relays can sometimes cause impulse noise.

The limiting bend in the CDF for the −12 dB threshold at approximately 2000 counts per five-minute interval in Fig. 45 might have been caused by power-line-hum pickup in the small rural step-by-step offices, which caused continuous impulse counts at the maximum counting rate of 420 counts per minute.

Figures 46 and 47 show the effect of switch type at both ends of the connection. Three types of switch were evaluated in the EOCS, leading to six nonordered pairs of switches when both end switches were taken into consideration. Figure 46 shows the percentage of impulse noise counts per five-minute interval for the −12 dB threshold for each of six different pairs of switches. The bars are divided, reading from the bottom up, as 0 counts, 1 to 10 counts, 11 to 20 counts, 21 to 50 counts, and more than 50 counts. The digital switch performs better than the crossbar, which is superior to the step-by-step, and this holds for either end of the connection. Figure 47 is similar to Fig. 46 except for the −4 dB threshold.

Figure 48 shows the impulse noise counts per five-minute interval for the −12 dB threshold at the step-by-step measuring end office switch plotted as a function of the time of day. It indicates an increase in the impulse noise counts in step-by-step offices during the busy

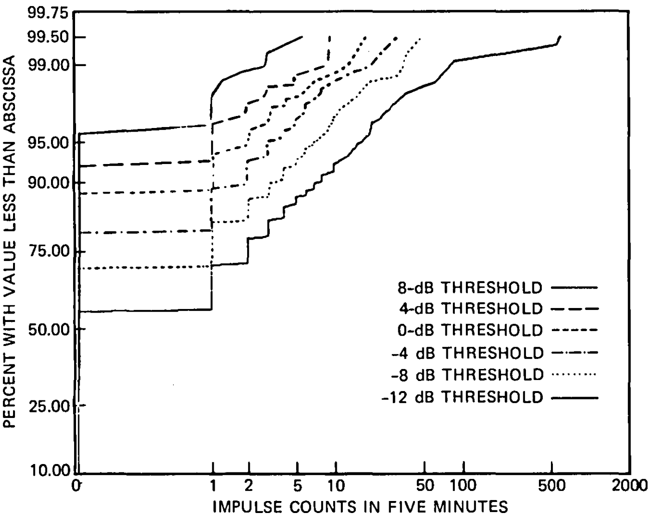


Fig. 43—CDFs of the impulse noise counts for the different thresholds as measured at electronic switching system end offices.

hours of a day, as one would expect when step-by-step switches in adjacent bays operate and release as other customers start and finish calls. A similar effect, with a smaller amplitude, was observed for the other thresholds and for the other types of switch.

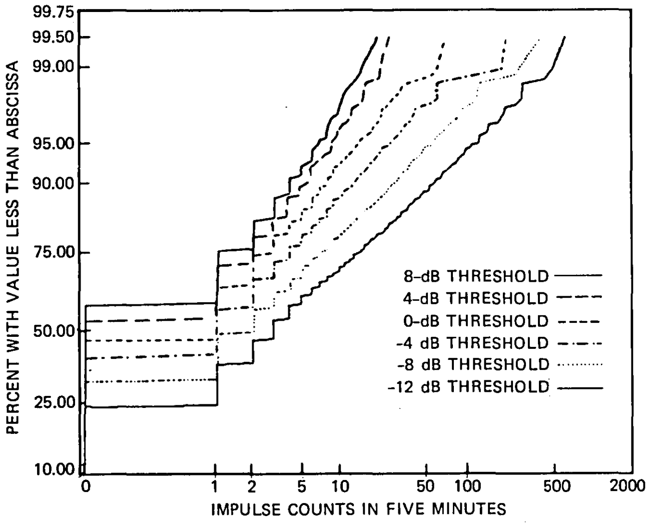


Fig. 44—CDFs of the impulse noise counts for the different thresholds as measured at crossbar end offices.

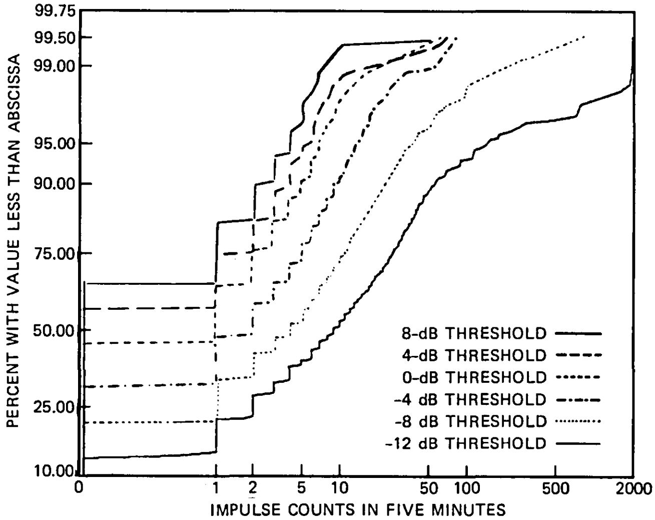


Fig. 45—CDFs of the impulse noise counts for the different thresholds as measured at step-by-step end offices.

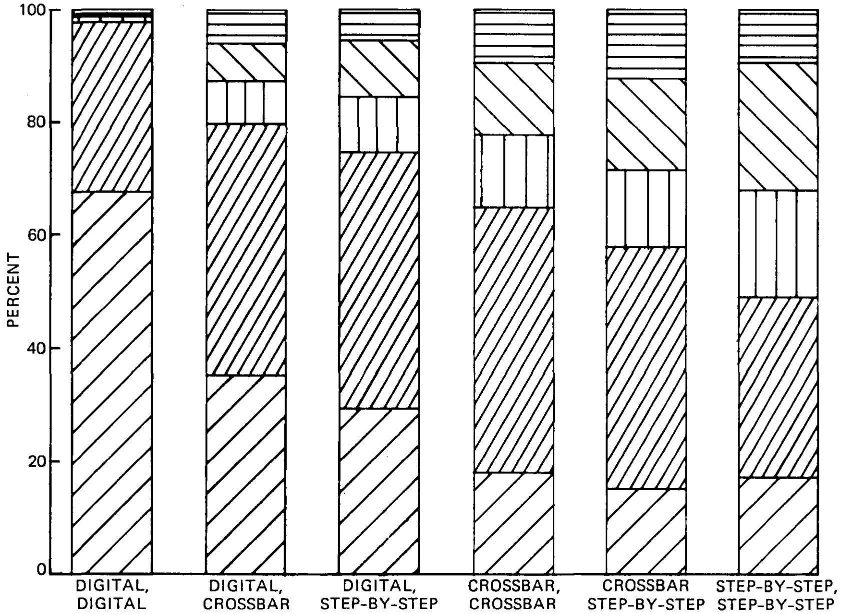


Fig. 46—Impulse noise counts for five-minute interval for the threshold of 12 dB below the received signal and for the six different pairs of end offices.

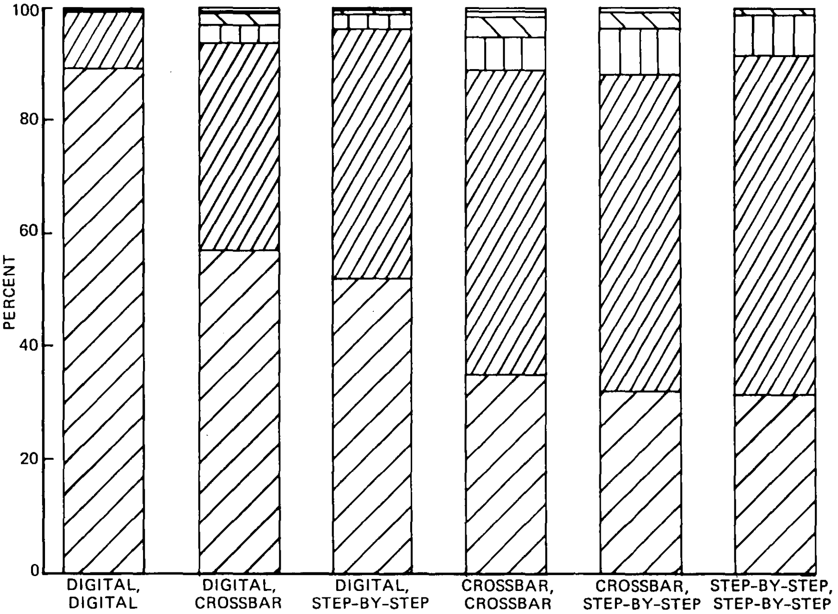


Fig. 47—Impulse noise counts per five-minute interval for the threshold of 4 dB below the received signal and for the six different pairs of end offices.

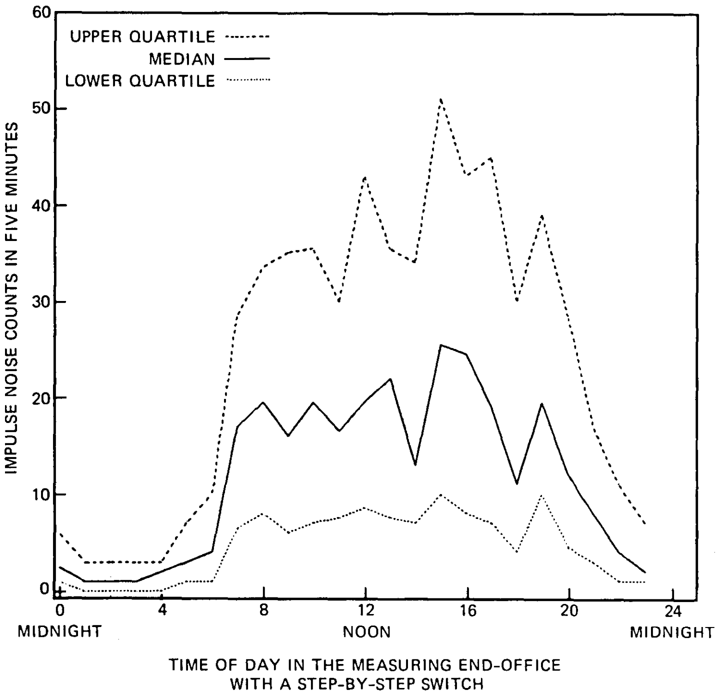


Fig. 48—Impulse noise counts for −12 dB threshold measured at step-by-step switches shown as a function of time of day.

3.11 Phase and gain hits

A phase hit is an abrupt change in the nominal phase of the received 1004-Hz holding tone lasting at least 4 ms. A gain hit is an abrupt change in the nominal level of the received 1004-Hz holding tone lasting at least 4 ms. The precision of the phase- and gain-hit measurement, the tracking rate for the phase-locked loop for the phase-hit counter, and the rate of change for the automatic gain control for the gain-hit counter are given in Ref. 1.

Phase and gain hits were measured simultaneously with impulse noise during the three five-minute transient measurement periods discussed in the previous section. During the first five minutes, the phase-hit threshold was set at 10 degrees and the gain hit threshold was set at 2 dB; during the second five minutes, the phase- and gain-hit thresholds were set at 15 degrees and 3 dB, respectively; finally, during the last five minutes, the phase- and gain-hit thresholds were set at 20 degrees and 6 dB, respectively. The phase- and gain-hit counting circuits respond to both positive and negative hits.

Figure 49 shows the CDFs of phase-hit counts per five-minute interval for the thresholds of 10, 15, and 20 degrees. The figure shows little difference between the CDFs corresponding to the two higher thresholds, 15 and 20 degrees. However, the CDF with 10-degree threshold is clearly worse than the other two CDFs. This may be caused by low-frequency phase modulation which can trigger the phase-hit counter at the 10-degree threshold.

Figure 50 shows the CDFs of gain-hit counts per five-minute interval for the thresholds of 2, 3, and 6 dB. There is a distinct reduction in the counts as the threshold is increased.

3.12 Dropouts

A dropout is defined as a 12-dB reduction of received signal level, as measured at the start of the 15-minute transient measurement interval, lasting for at least 4 ms. The dropout counter circuit has no automatic gain control circuit in contrast with the gain-hit counter. Figure 51 shows the CDF of dropout counts per 15-minute interval.

3.13 Bit and block error rates

The error rate performance of two widely used 1200-b/s data sets and two widely used 4800-b/s data sets was measured on the same connections on which the analog transmission impairments were measured. To simulate the environment in which the data sets would normally be operating, an artificial loop was placed in front of the data set in each ASPEN RTU. The artificial loop simulates 6000 feet of 26-gauge cable in series with 6000 feet of 24-gauge cable to achieve the mean loop loss of 5.3 dB determined from the 1980 Loop Survey. The data signal level at the data set was approximately −9 dBm.

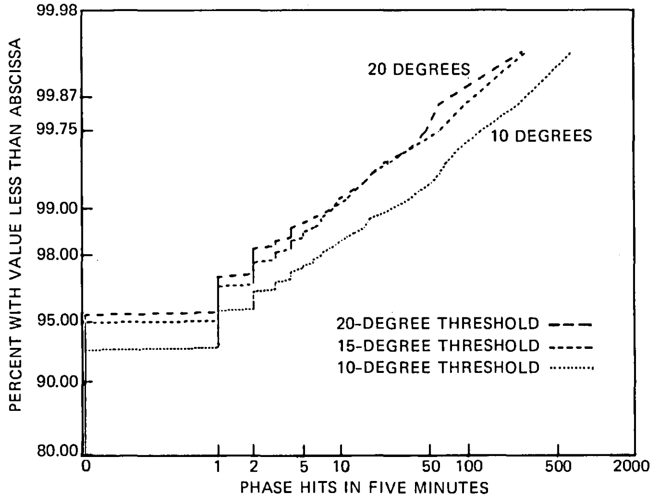


Fig. 49—CDFs of the phase hit counts for the thresholds of 10, 15, and 20 degrees.

The two 1200-b/s data sets used in the EOCS were full-duplex, four-phase, Differential Phase-Shift Keyed (DPSK) data sets. The data sets transmitted 1200-b/s (600-baud) synchronous binary serial data simultaneously in both directions by splitting the voiceband into a low band (carrier frequency of 1200 Hz) and high band (carrier frequency of 2400 Hz) of frequencies. The energy in the low-band line signal extended from 720 to 1680 Hz, and the high band, from 1920 to 2880 Hz. These data sets employed scramblers to prevent steady marking

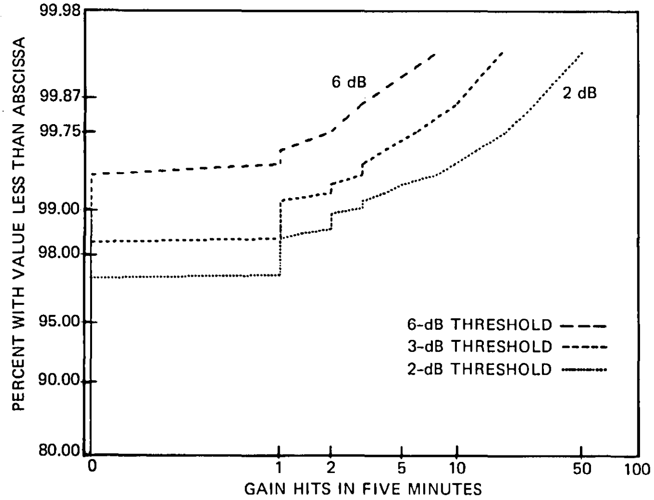


Fig. 50—CDFs of the gain-hit counts for the thresholds of 2, 3, and 6 dB.

or spacing to the transmitter causing a continuous stream of zero-degree phase shifts on the line (which would block the receiver timing-recovery circuit reference phase extraction from the incoming signal). These data sets had no adaptive equalizers to mitigate the effects of bandwidth reduction or poor EDD on the connection.

The two 4800-b/s data sets used in the EOCS were half-duplex, eight-phase DPSK data sets. They transmitted a 1600-baud (3 bits per symbol) signal, and 23-stage, multiple-tap scramblers were employed for the same reason as the scramblers in the 1200-b/s data sets. Most of the energy in the line spectrum was between 800 and 2800 Hz, and neither data set had a secondary channel. The receivers had adaptive equalizers to reduce the effects of connection bandwidth reduction and EDD on intersymbol interference.

The solid and dotted curves of Figs. 52, 53, and 54 show the CDFs of the bit error rates for the two 1200-b/s data sets for short, medium, and long connections. For these three figures, the low and high-band error rates were combined. The 1200-b/s data set bit error performance is poorer for longer connections. In Fig. 55, the bit error rates for long connections were separated into low and high band for the two 1200-b/s data sets. This figure demonstrates how the selection of the compromise equalizer in the data set can affect the relative performance of the two bands.

Since about one million bits were transmitted, there are no data points on the CDFs between "No Errors" and 1 × 10^{-6} bit error rate. The bit error rate counter in the ASPEN RTU attempted to resynchronize when 99 errors were received, and therefore the bit error rate plots were truncated at the value corresponding to 99 errors.

In the middle of the data collection period, the ASPEN RTUs were modified to permit both bit and block error rate measurement. For block error rate measurements, one thousand 1000-bit blocks were transmitted, and the restriction of no more than 99 errors in a single block was removed. All bit error rate figures include data collected over the entire EOCS data collection period, while the block error rate figures come from only the later part of the collection period. The CDFs for the block error rate performance for the two 1200-b/s data sets for short, medium, and long connections are shown in Figs. 56, 57, and 58. The block error rate performance of the 1200-b/s data sets is also poorer for the longer connections.

Figures 59, 60, and 61 show the bit error rate performance of the two 4800-b/s data sets for short, medium, and long connections.* As

* All error rate data presented here were taken with continuous carrier mode data set operation. Continuous carrier mode is the normal operation mode for the 1200-b/s full-duplex data sets. For the 4800-b/s half-duplex data sets, however, switched carrier mode is the typical mode of operation. In general, error performance with switched carrier mode operation is poorer than that with continuous carrier mode operation.

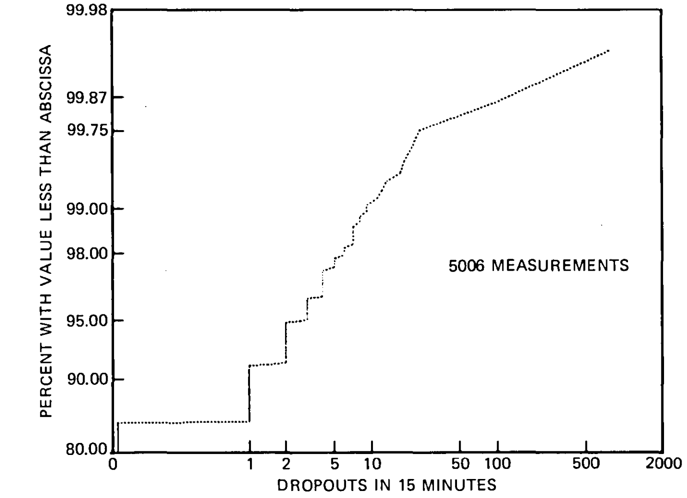


Fig. 51—CDF of dropout counts.

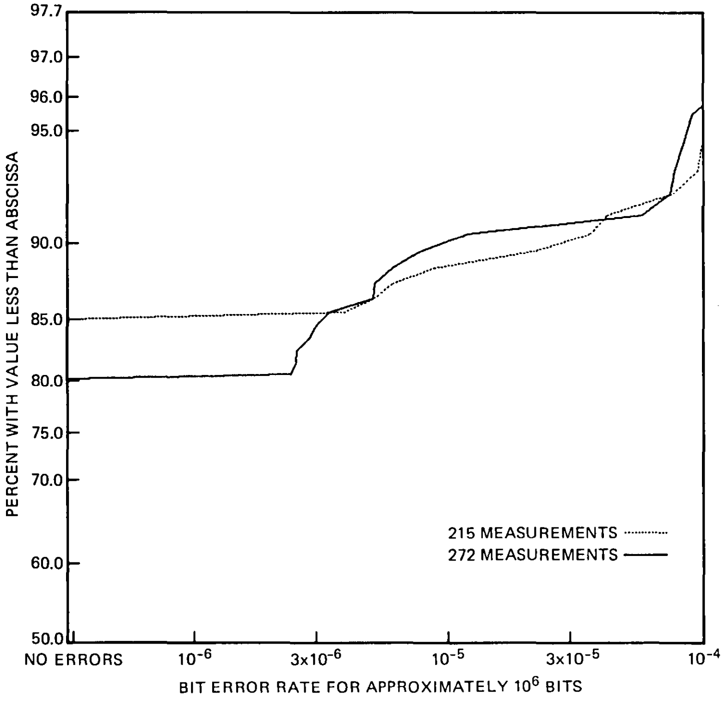


Fig. 52—CDFs of bit error rates for two 1200-b/s data sets for short connections.

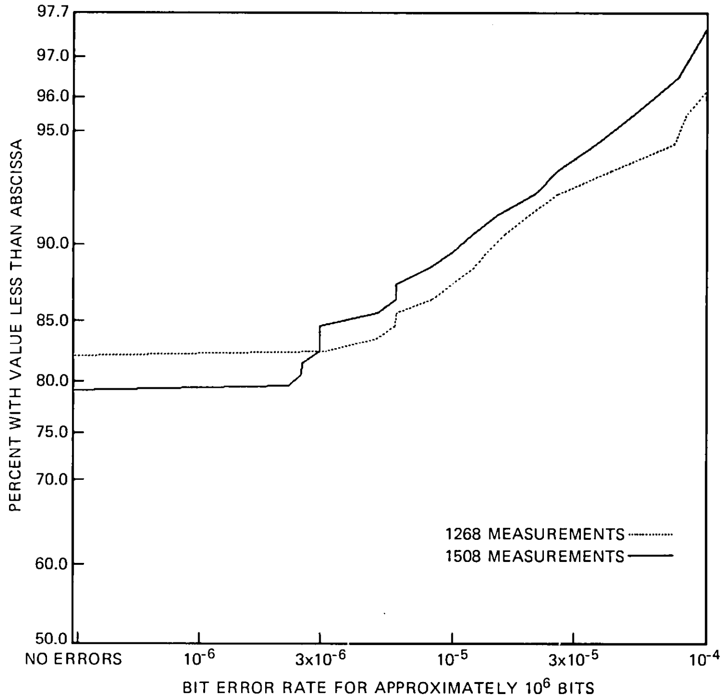


Fig. 53—CDFs of bit error rates for two 1200-b/s data sets for medium connections.

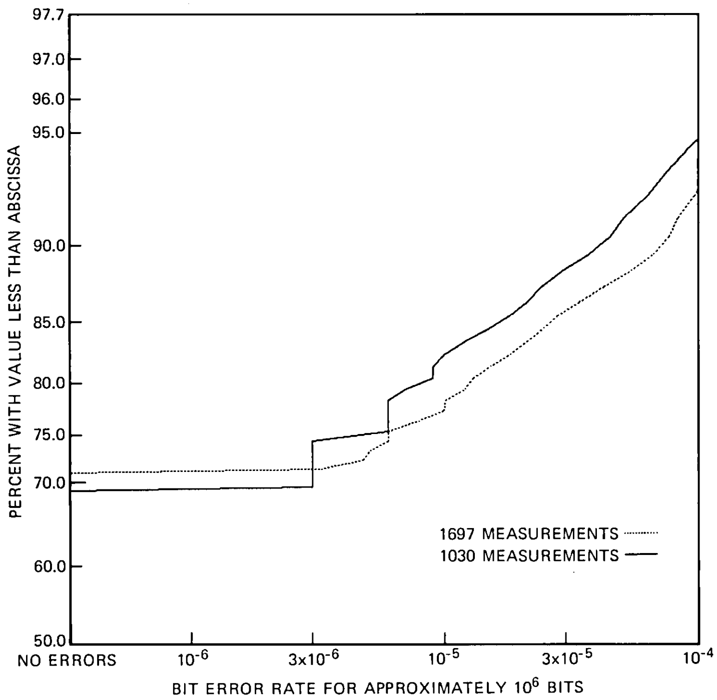


Fig. 54—CDFs of bit error rates for two 1200-b/s data sets for long connections.

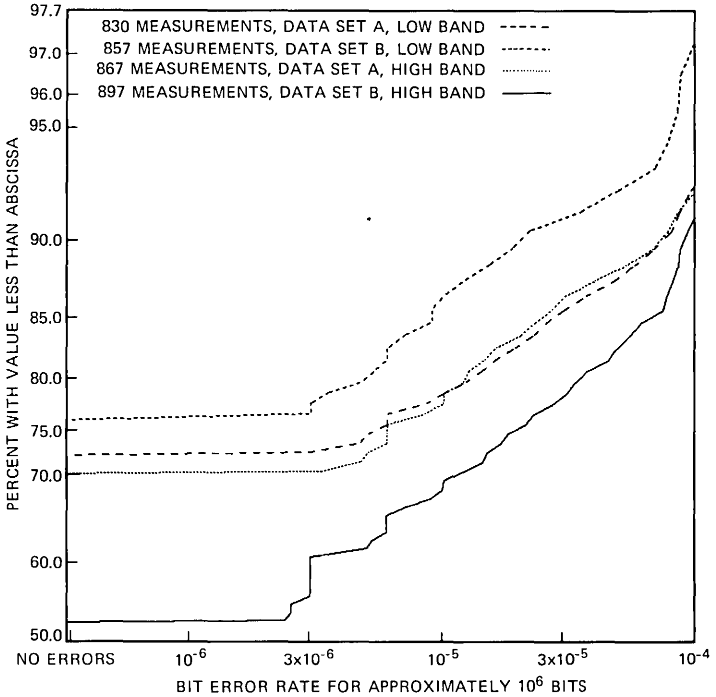


Fig. 55—CDFs for bit error rates for low and high bands for two 1200-b/s data sets for long connections.

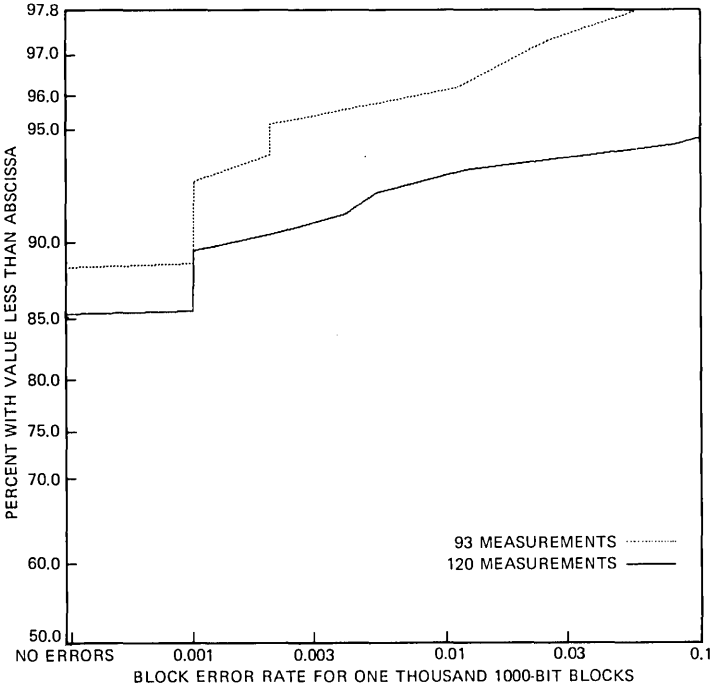


Fig. 56—CDFs of block error rates for two 1200-b/s data sets for short connections.

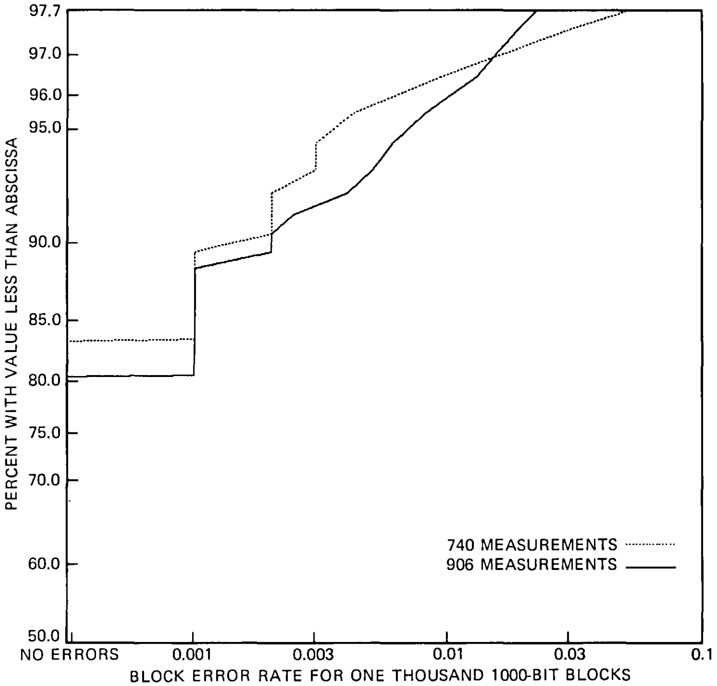


Fig. 57—CDFs of block error rates for two 1200-b/s data sets for medium connections.

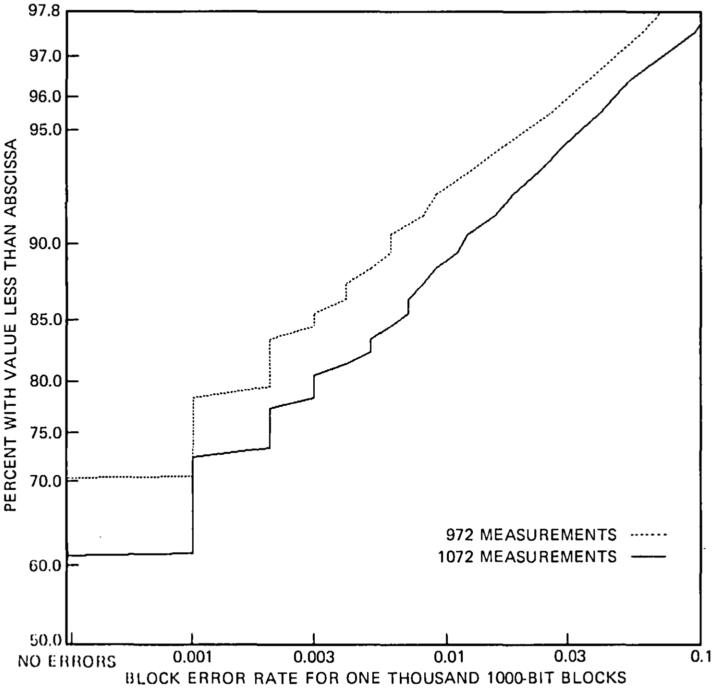


Fig. 58—CDFs of block error rates for two 1200-b/s data sets for long connections.

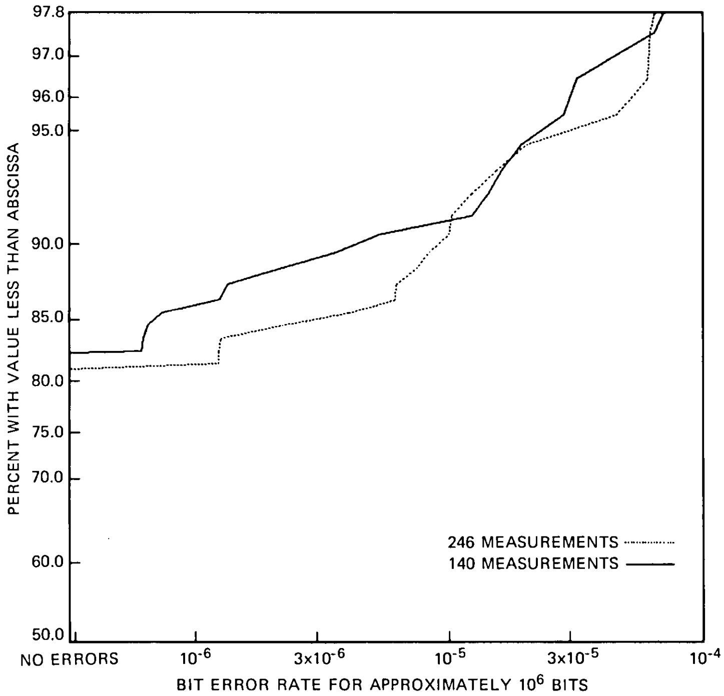


Fig. 59—CDFs of bit error rates for two 4800-b/s data sets for short connections.

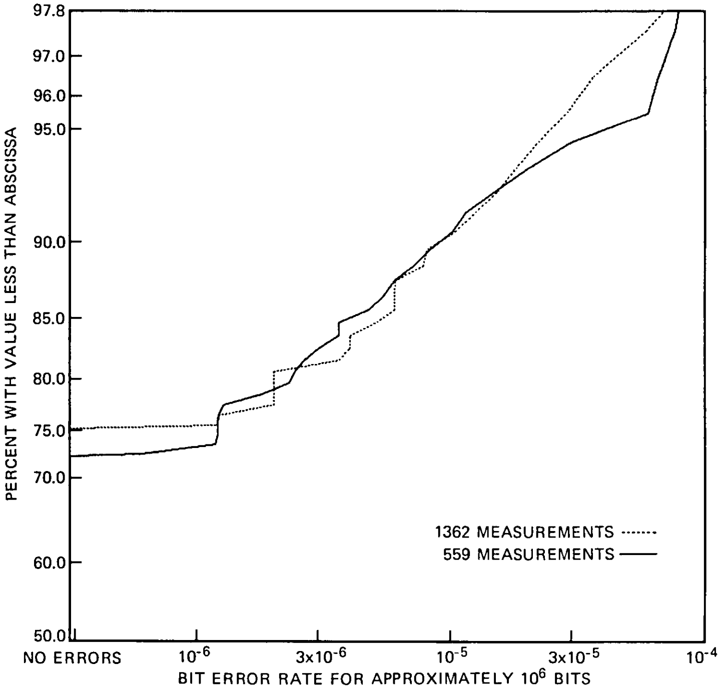


Fig. 60—CDFs of bit error rates for two 4800-b/s data sets for medium connections.

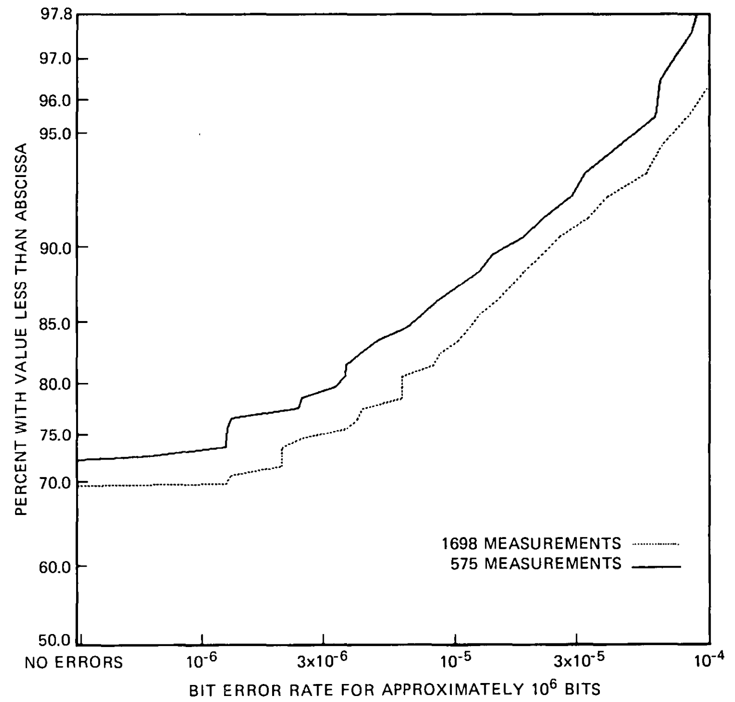


Fig. 61—CDFs of bit error rates for two 4800-b/s data sets for long connections.

we can see from the comparison of these figures and Figs. 52, 53, and 54, the bit error rates for the 1200-, and 4800-b/s data sets are not markedly different. This could be expected, considering that the much more expensive 4800-b/s data sets used in the EOCS had adaptive equalizers, and that the 1200-b/s, full-duplex data transmission is really two independent 1200-b/s data transmissions on the same connection.

The block error rate performance of one of the 4800-b/s data sets for short, medium, and long connections is shown in Fig. 62, once again demonstrating the poorer performance for long connections.

Figure 63 is a scatter plot of bit error rate versus block error rate for both of the 1200-b/s data sets, with a small amount of dither added in the horizontal axis to show where multiple points occur. As we can see from the line where errors occur in only one of the one thousand blocks (0.001), there is a preponderance of bit error counts of three, six, nine, and twelve. Figure 64 is a similar plot for one of the 4800-b/s data sets, which shows a tendency for an even number of error counts.

The figures that compare the two 1200-b/s data sets (or the two 4800-b/s data sets) show that there are differences in their perform-

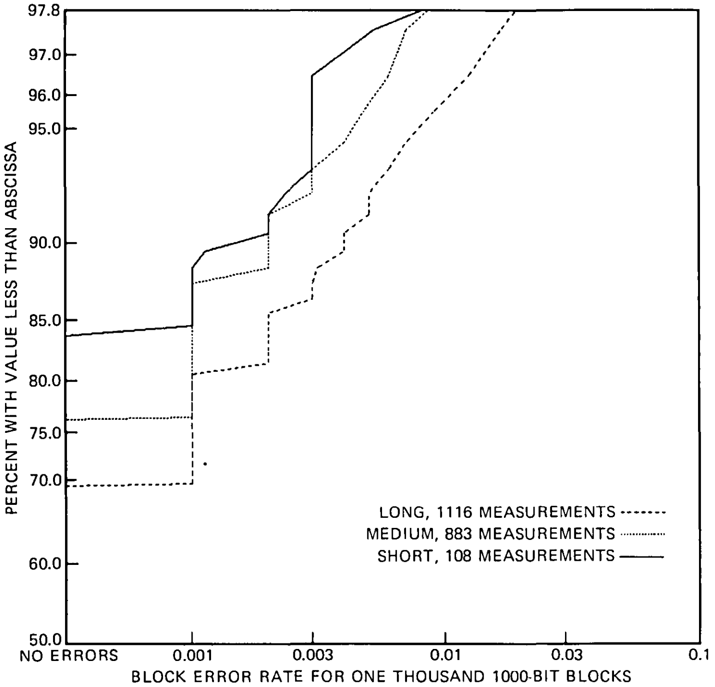


Fig. 62—CDFs for block error rate for a 4800-b/s data set for short, medium, and long connections.

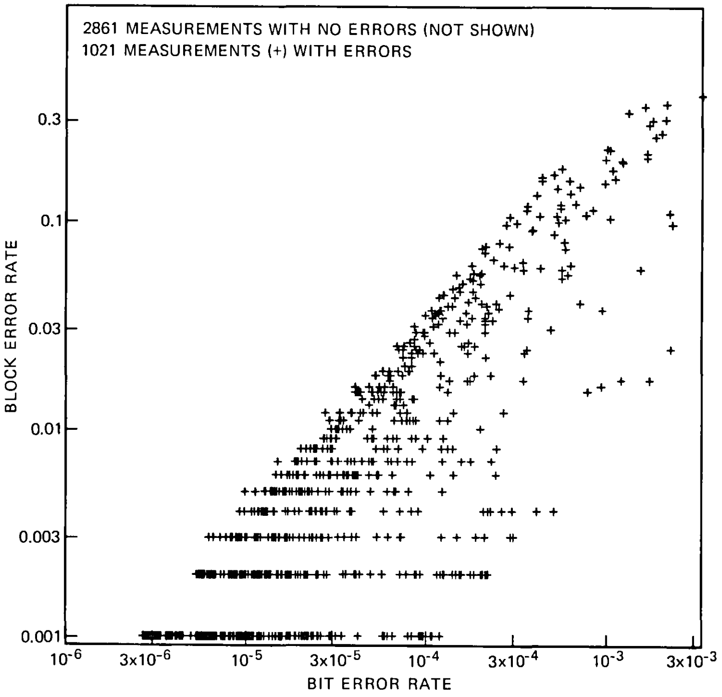


Fig. 63—Block error rate versus bit error rate for two 1200-b/s data sets for one thousand 1000-bit blocks.

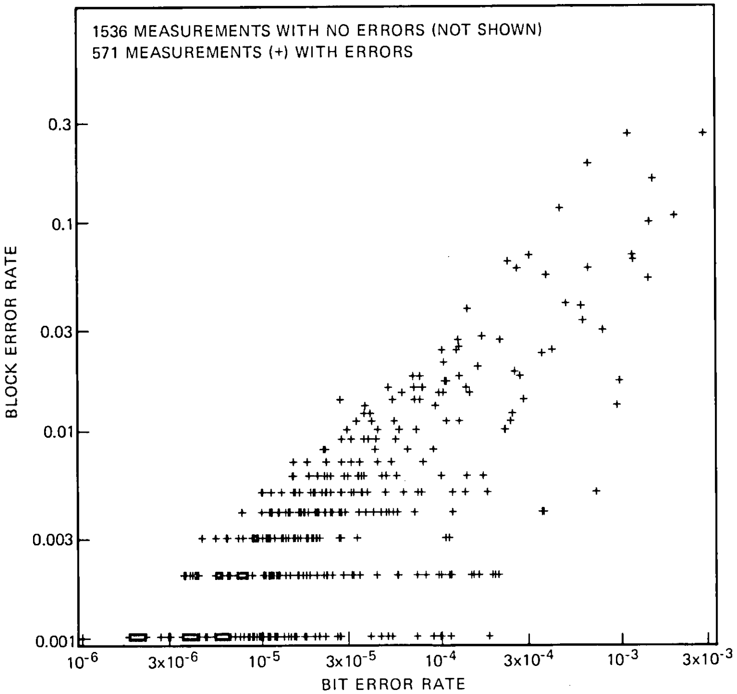


Fig. 64—Block error rate versus bit error rate for a 4800-b/s data set for one thousand 1000-bit blocks.

ance. Prediction of the performance of other types of data sets by extrapolation from these results is not warranted.

3.14 Cutoffs

The only nontransmission parameter estimated from the EOCS measurements is the call cutoff rate. A cutoff is a connection dropped prematurely; it can occur primarily because of failure in the communications path caused by transmission or switching problems. Failures resulting in carrier group alarms on T-carrier systems and talk off of in-band channel signaling are sources of transmission-caused cutoffs. Cutoffs caused by switching systems are the result of hardware failures, software failures (e.g., in digital switches), and procedural errors. The cutoff rate can be measured by examining a sample of representative connections through the network and observing the number of calls disconnected prematurely.

Four thousand toll connections were checked in the EOCS sample for possible cutoffs during the 15-minute intervals of impulse noise measurements discussed in Section 3.10. A cutoff was tallied if a call was disconnected at any time during the 15 minutes, given that the call was up at the beginning of the same 15-minute interval. Twenty-two cutoffs occurred during the 60,000 call-minute sample (4000 × 15), leading to a nonweighted average cutoff rate of 2.2×10^{-3} for a six-minute holding time. The 90-percent confidence interval was calculated to be between 1.4×10^{-3} and 3.0×10^{-3}.

IV. ACKNOWLEDGMENTS

Many individuals contributed to the success of the ASPEN/End Office Connection Study (EOCS): Dave Leeper led the group that developed ASPEN and conducted the data acquisition phase of the EOCS; Jay MacMaster, Ed Vlacich, Al Furtado, and Jon Palmer designed, constructed, tested, and installed the ASPEN Remote Test Units (RTUs); Maurice Lampell, Louis Iacona, and Karel Ehrlich wrote and monitored the ASPEN central computer programs; Paul Auer designed the Transient Recording System and provided help in shaping the ASPEN system; Bob Sturges provided help in the design and construction of the RTU support circuitry; John Healy and Tom Redman developed the EOCS sampling plan; Blan Godfrey provided support for the original ASPEN concept; and Pete Lopiparo provided continued support throughout the completion of the ASPEN development, and the EOCS data acquisition and analysis. Finally, the cooperation and contributions of Bell Operating Company central office personnel are gratefully acknowledged.

REFERENCES

1. AT&T PUB 41009, *Transmission Parameters Affecting Voiceband Data Transmission—Measuring Techniques,* May 1975.
2. AT&T PUB 41008, *Transmission Parameters Affecting Voiceband Data Transmission—Description of Parameters,* July 1974.
3. J. D. Healy, M. Lampell, D. G. Leeper, T. C. Redman, and E. J. Vlacich, "1982/83 End Office Connection Study: ASPEN Data Acquisition System and Sampling Plan," AT&T Bell Lab. Tech. J., this issue.
4. F. P. Duffy and T. W. Thatcher, Jr., "1969–70 Connection Survey: Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50,* No. 4 (April 1971), pp. 1311–47.
5. S. R. Searle, *Linear Models,* New York: Wiley, 1971, pp. 340–60.
6. R. Vernon, private communication.
7. T. C. Redman, private communication.
8. D. V. Batorsky and M. E. Burke, "1980 Bell System Noise Survey of the Loop Plant," AT&T Bell Lab. Tech. J., *63,* No. 5 (May-June 1984), pp. 775–818.
9. T. C. Anderson and D. L. Favin, *The P/AR Meter,* 1964 IEEE Int. Conv. Rec., *12,* Part 5, pp..155–65.

AUTHORS

Michele B. Carey, B.S. (Mathematics), 1968; M.S. (Statistics), 1970, University of Sciences of Jussieu, Paris, France; Ph.D. (Applied Mathematical Sciences), 1982, University of Rhode Island; Bell Laboratories, 1982–1983. Present affiliation Bell Communications Research, Inc. Ms. Carey was involved in the data analysis of field measurements of the performance of the predivestiture Bell System telephone network. She is now involved in the design and analysis of field studies aiming at the characterization of the transmission performance of the telephone network.

Han-Tee Chen, B.S. (Chemistry), 1970, Tsing Hua University, Taiwan; M.S. (Physical Chemistry), 1973, Worcester Polytechnic Institute, MA. Bell Laboratories, 1980–1983. Present affiliation Bell Communications Research, Inc. Ms. Chen designed and developed databases for network performance surveys and analyzed the results from those surveys. She is currently working on the software design and development of an automated reliability prediction system in the Methods and Information Division of Bell Communications Research.

Alfred Descloux, Diploma (Mathematics and physics), 1948, Swiss Federal Institute of Technology, Zurich, Switzerland; Ph.D. (Mathematics Statistics), 1961, University of North Carolina; Bell Laboratories, 1956–1983. Present affiliation Bell Communications Research, Inc. Mr. Descloux has been concerned with the probabilistic aspects of telephone traffic. In particular, his past work included new analytical results in the area of traffic measurements. Mr. Descloux is presently studying the performance of switching networks with integrated voice and data traffic. Member, American Mathematical Society, Institute of Mathematical Statistics.

James F. Ingle, B.S. (Electrical Engineering), 1955, Rensselaer Polytechnic Institute; M.S. (Electrical Engineering), 1961, New York University; Bell Laboratories, 1955–1983. Present affiliation Bell Communications Research, Inc. Mr. Ingle designed transmission test equipment for assessing the quality of video and voice transmission facilities. He holds seven patents in the test equipment field, and he wrote the Bell System technical reference for voiceband transmission test equipment (AT&T PUB 41009). Mr. Ingle is presently interpreting survey measurement results. He is a member of the IEEE subcommittee working on voiceband transmission test equipment requirements.

Kun I. Park, B.S. (Electrical Engineering), 1966, Seoul National University; M.S. (Electrical Engineering), 1968, Ph.D. (Electrical Engineering), 1972, University of Pennsylvania; Bell Laboratories, 1973–1983. Present affiliation Bell Communications Research, Inc. At Bell Laboratories, Mr. Park was Supervisor of the Data Communications Performance Characterization group, and he worked on data communications performance objectives and characterizations. He is now District Manager of the Data Services Performance District at Bell Communications Research, and he is responsible for performance planning for data services. His current projects include ISDN performance planning, field performance characterizations, construction of digital and voiceband data performance laboratories, and development of computer tools for circuit- and packet-switched data performance analysis. Member, IEEE, Sigma Xi.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Agrawala A. K., Coffman E. G., Garey M. R., Tripathi S. K., **A Stochastic Optimization Algorithm Minimizing Expected Flow Times on Uniform Processors.** IEEE Comput 33(4):351–356, 1984.

Bentley J. L., **Programming Pearls—How to Sort.** Comm ACM 27(4):287–291, 1984.

Bentley J., **Squeezing Space.** Comm ACM 27(5):416–421, 1984.

Carroll J. D., Arabie P., **Indclus—An Individual-Differences Generalization of the Adclus Model and the Mapclus Algorithm.** Psychometri 48(2):157–169, 1983.

Chung F. R. K., **The Number of Different Distances Determined by *N*-Points in the Plane.** J Comb Th A 36(3):342–354, May 1984.

Coffman E. G. et al., **A Note on Expected Makespans for Largest-First Sequences of Independent Tasks on Two Processors.** Math Oper R 9(2):260–266, May 1984.

Cuykendall R., Domic A., Joyner W. H., Johnson S. C., Kelem S., McBride D., Mostow J., Savage J. E., Saucier G., **Design Synthesis in VLSI and Software Engineering.** J Syst Soft 4(1):7–12, 1984.

Desarbo W. S., Carroll J. D., Clark L. A., Green P. E., **Synthesized Clustering—A Method for Amalgamating Alternative Clustering Bases With Differential Weighting of Variables.** Psychometri 49(1):57–78, Mar 1984.

Desarbo W. S., Nahajan V., **Constrained Classification—The Use of A Priori Information in Cluster-Analysis.** Psychometri 49(2):187–215, Jun 1984.

Desoete G., Carroll J. D., **A Maximum-Likelihood Method for Fitting the Wandering Vector Model.** Psychometri 48(4):553–566, Dec 1983.

Fishburn P. C., **A Generalization of Comparative Probability on Finite Sets.** J Math Psyc 27(3):298–310, 1983.

Fishburn P. C., **On the Sphericity and Cubicity of Graphs.** J Comb Th B 35(3):309–318, 1983.

Fishburn P. C., **Discrete Mathematics in Voting and Group Choice.** Siam J Alg 5(2):263–275, Jun 1984.

Foregger M. F., **The Skew Chromatic Index of a Graph.** Discr Math 49(1):27–39, 1984.

Gehani N. H., Cargill T. A., **Concurrent Programming in the Ada Language—The Polling Bias.** Software 14(5):413–427, May 1984.

Harel D., Tarjan R. E., **Fast Algorithms for Finding Nearest Common Ancestors.** SIAM J Comp 13(2):338–355, 1984.

Hausman J. A., Taylor W. E., **Identification in Linear Simultaneous-Equations Models With Covariance Restrictions—An Instrumental Variables Interpretation.** Econometric 51(5):1527–1549, 1983.

Hawkins D. T., **The Literature of Online Information-Retrieval—An Update.** Online Rev 8(2):153–164, 1984.

Hilinski E. F., Huppert D., Kelley D. F., Milton S. V., Rentzepis P. M., **Photodissociation of Haloaromatics—Detection, Kinetics, and Mechanism of Arylmethyl Radical Formation.** J Am Chem S 106(7):1951–1957, 1984

Hwang F. K., **Three Versions of a Group-Testing Game.** Siam J Alg 5(2):145–153, Jun 1984.

Lee J. C., Tan W. Y., **On the Degree of Polynomial in a General Linear Model.** Comm St-The 13(6):781–790, 1984.

Liebson L. S., Cheek T. B., **Prepress Systems Become Cost Effective—Basic System Elements Make the Difference.** Comput Gr W 7(4):27–28, 1984.

Luecke G. R., Hickey K. R., **A Note on the Filtered Least-Squares Minimal Norm Solution of First-Kind Equations.** J Math Anal 100(2):635–641, May 1984.

Mallows C. L., Sloane N. J. A., **Designing an Auditing Procedure, or How to Keep Bank Managers on Their Toes.** Math Mag 57(3):142–151, 1984.

Manna Z., Wolper P., **Synthesis of Communicating Processes From Temporal Logic Specifications.** ACM T Progr 6(1):68–93, Jan 1984.

Martin O., Odlyzko A. M., Wolfram S., **Algebraic Properties of Cellular Automata.** Comm Math P 93(2):219–258, 1984.

Massey W. A., **An Operator-Analytic Approach to the Jackson Network.** J Appl Prob 21(2):379–393, Jun 1984.

McIlroy M. D., **Best Approximate Circles on Integer Grids.** ACM T Graph 2(4):237–263, 1983.

McKenna J., Mitra D., **Asymptotic Expansions and Integral-Representations of Moments of Queue Lengths in Closed Markovian Networks.** J ACM 31(2):346–360, 1984.

Miller J. F., Skerry H. B., **Regular and Mercerian Generalized Lototsky Method.** Can Math B 27(1):65–71, 1984.

Papadimitriou C. H., Yannakakis M., **The Complexity of Facets (and Some Facets of Complexity).** J Comput Sy 28(2):244–259, Apr 1984.

Ruskai M. B., Stillinger F. H., **Binding Limit in the Hartree Approximation.** J Math Phys 25(6):2099–2103, Jun 1984.

Sethi R., **Control Flow Aspects of Semantics—Directed Compiling.** Acm T Progr 5(4):554–595, Oct 1983.

Steffen J. L., **Experience With a Portable Debugging Tool.** Software 14(4):323–324, 1984.

Stevenson D. E., Warner D. D., Brown T. R., **Biobell—A Simulation System for Biochemistry and Biophysics.** Comput Biol 14(1):35–46, 1984.

Suurballe J. W., Tarjan R. E., **A Quick Method for Finding Shortest Pairs of Disjoint Paths.** Networks 14(2):325–336, 1984.

Tarjan R. E., **A Simple Version of Karzanov Blocking Flow Algorithm.** Oper Res L 2(6):265–268, 1984.

Tarjan R. E. et al., **Worst-Case Analysis of Set Union Algorithms.** J ACM 31(2):245–281, 1984.

Weinberger P. J., **Finding the Number of Factors of a Polynomial.** J Algorithm 5(2):180–186, Jun 1984.

Whitt W., **Comparison Conjectures About the M/G/S Queue.** Oper Res L 2(5):203–209, 1983.

Yannakakis M., **Serializability by Locking.** J ACM 31(2):227–244, 1984.

ENGINEERING

Agrawal G. P., **Line Narrowing in a Single-Mode Injection-Laser Due to External Optical Feedback (Letter).** IEEE J Q El 20(5):468–471, May 1984.

Alfernes R. C., Korotky S.K., Buhl L. L., Divino M. D., **High-Speed Low-Loss Low-Drive-Power Traveling-Wave Optical Modulator for Lambda = 1.32 μm.** Electr Lett 20(8):354–355, 1984.

Alfernes R. C., Korotky S. K., Marcatil E. A., **Velocity-Matching Techniques for Integrated Optic Traveling-Wave Switch Modulators.** IEEE J Q El 20(3):301–309, 1984.

Amitay N., Greenstein L. J., **Multipath Outage Performance of Digital Radio Receivers Using Finite-Tap Adaptive Equalizers.** IEEE Commun 32(5):597–608, May 1984.

Anderson W. T., Lenahan T. A., **Length Dependence of the Effective Cutoff Wavelength in Single-Mode Fibers.** J Lightw T 2(3):238–242, Jun 1984.

Antler M., **Fretting Corrosion of Solder-Coated Electrical Contacts.** IEEE Compon 7(1):129–138, Mar 1984.

Beni G., Hackwood S., Rin L., **Dynamic Sensing for Robots—An Analysis and Implementation.** P Soc Photo 449(P2):589–595, 1984.

Besomi P., Wilson R. B., Brown R. L., Dutta N. K., Wright P. D., Nelson R. J., **High-Temperature Operation of 1.55-μm InGaAsP Double-Channel Buried-Heterostructure Lasers Grown by LPE.** Electr Lett 20(10):417–419, May 10 1984.

Bosch M. A., Herbst D., Tewksbury S. K., **The Influence of Light on the Properties of NMOS Transistors in Laser μ-Zoned Crystallized Silicon Layers.** IEEE Elec D 5(6):204–206, Jun 1984.

Chemla D. S., Miller D. A. B., Smith P. W., Gossard A. C., Wiegmann W., **Room-Temperature Excitonic Nonlinear Absorption and Refraction in GaAs/AlGaAs Multiple Quantum-Well Structures.** IEEE J Q El 20(3):265–275, 1984.

Cheng C. L., Liao A. S. H., Chang T. Y., Leheny R. F., Coldren L. A., Lalevic B., **Submicrometer Self-Aligned Recessed Gate InGaAs MISFET Exhibiting Very High Transconductance.** IEEE Elec D 5(5):169–171, 1984

Coldren L. A., Koch T. L., **Analysis and Design of Coupled-Cavity Lasers. 1. Threshold Gain Analysis and Design Guidelines.** IEEE J Q El 20(6):659–670, Jun 1984.

Coldren L. A., Koch T. L., **Analysis and Design of Coupled-Cavity Lasers. 2. Transient Analysis.** IEEE J Q El 20(6):671–682, Jun 1984.

Coldren L. A., Koch T. L., Burrus C. A., Swartz R. G., **Intercavity Coupling Gap Width Dependence in Coupled-Cavity Lasers.** Electr Lett 20(8):350–352, 1984.

Coldren L. A., Swartz R. G., **Generation of Constant-Power Curves for Three-Terminal Coupled-Cavity Lasers.** Electr Lett 20(8):329–330, 1984.

Corbett J. W., Frisch H. L., Snyder L. C., **On the Thermal Donors in Silicon.** Mater Lett 2(3):209–210, 1984.

Cox D. C., Arnold H. W., **Comparison of Measured Cross-Polarization Isolation and Discrimination for Rain and Ice on a 19-GHz Space-Earth Path.** Radio Sci 19(2):617–628, 1984.

Demarest K., Plourde J. K., **An Optimum Combination of Power Levels and Combiner Weighting for Generating 64-QAM From Three 4-PSK Signals (Letter).** IEEE Commun 32(3):320–322, 1984.

Demerdash N. A., Nehl T.W., Fouad F. A., Arkadan A. A., **Analysis of the Magnetic Field in Rotating Armature Electronically Commutated DC Machines by Finite-Elements.** IEEE Power 103(7):1829–1836, Jul 1984

Dutta N. K., Wilt D. P., Nelson R. J., **Analysis of Leakage Currents in 1.3-μm InGaAsP Real-Index-Guided Lasers.** J Lightw T 2(3):201–208, Jun 1984.

Dyer C. K., **Effect of Geosynchronous Altitude Radiation on Performance of Ni/H_2 Cells.** J Power Sou 12(3–4):323–334, July–Aug 1984.

Felder R. J., Hauser J. J., **Amorphous Cu-Zr Foils.** Mater Lett 2(3):232–233, 1984.

Fork R. L., Martinez O. E., Gordon J. P., **Negative Dispersion Using Pairs of Prisms.** Optics Lett 9(5):150–152, 1984.

Goksel A. K., Fields J. A., Larocca F. D., Lu P. M., Troutman W. W., Wong K. N., **A VLSI Memory Management Chip—Design Considerations and Experience.** IEEE J Soli 19(3):325–328, Jun 1984.

Gordon J. P., Fork R. L., **Optical-Resonator With Negative Dispersion.** Optics Lett 9(5):153–155, 1984.

Hackwood S., Beni G., Nelson T. J., **Torque Sensitive Tactile Array for Robotics.** P Soc Photo 449(P2):602–608, 1984.

Haque C. A., Uhrig T. A., **Dynamic Contact Resistance and Film Thickness of Relay Contacts Coated With Fluoropolymer Thin Films.** IEEE Compon 7(1):76–80, Mar 1984.

Hedinger R. A. et al., **On the Relationship Between Geostationary Orbit Capacity and the Interference Allowance.** IEEE Commun 32(5):627–634, May 1984.

Henry C. H., Henry P. S., Lax M., **Partition Fluctuations in Nearly Single-Longitudinal-Mode Lasers.** J Lightw T 2(3):209–216, Jun 1984.

Henry C. H., Logan R. A., Merritt F. R., Bethea C. G., **Radiative and Nonradiative Lifetimes in N-Type and P-Type 1.6-μm InGaAs.** Electr Lett 20(9):358–359, 1984.

Hornak L. A., Hackwood S., Beni G., **Reentrant-Loop Magnetic-Effect Proximity Sensor for Robotics.** P Soc Photo 449(P1):323-327, 1984.

Korotky S. K., Alferness R. C., Buhl L. L., Joyner C. H., Marcatili E. A., **High-Speed Pulse Generation Using a Sinusoidally Driven Ti-$LiNbO_3$ Directional Coupler Traveling-Wave Optical Modulator.** Electr Lett 20(9):384–386, 1984.

Kuhn M., **Telecommunications in Canada—A Century of Symbiotic Development.** IEEE Comm M 22(5):104–114, 1984.

Lee, T. P., Burrus C. A., Liu P. L., Sessa W. B., Logan P.A., **An Investigation of the Frequency Stability and Temperature Characteristics of 1.5-μm Coupled-Cavity Injection Lasers.** IEEE J Q El 20(4):374–384, Apr 1984.

Lee T. P., Burrus C. A., Sessa W. B., Tsang W. T., **$InAs_xP_{1-x}/InP$ Photodiodes Prepared by Molecular-Beam Epitaxy.** Electr Lett 20(9):363–364, 1984.

Levine B. F., Bethea C. G., **Detection of Single 1.3 μm Photons at 45 Mb/s.** Electr Lett 20(6):269–271, 1984.

Linke R. A., **Transient Chirping in Single-Frequency Lasers—Lightwave Systems Consequences.** Electr Lett 20(11):472–474, May 24, 1984.

Linke R. A., Kasper B. L., Campbell J. C., Dentai A. G., Kaminow I. P., **120 km Lightwave Transmission Experiment at 1 Gb/s Using a New Long-Wavelength Avalanche Photodetector.** Electr Lett 20(12):498–499, Jun 7 1984.

Luryi S., Kastalsky A., Gossard A. C., Hendel R. H., **Charge Injection Transistor Based on Real-Space Hot-Electron Transfer.** IEEE Device 31(6):832–839, Jun 1984.

Mahajan S., Chin A. K., Zipfel C. L., Brasen D., Chin B. H., Tung R. T., Nakahara S., **The Origin of Dark Spot Defects in InP/InGaAsP Aged Light-Emitting Diodes.** Mater Lett 2(3):184–188, 1984.

Martinez O. E., Fork R. L., Gordon J. P., **Theory of Passively Mode-Locked Lasers Including Self-Phase Modulation and Group-Velocity Dispersion.** Optics Lett 9(5):156–158, 1984.

Mercer M. R., Agrawal V. D., **A Novel Clocking Technique for VLSI Circuit Testability.** IEEE J Soli 19(2):207–212, 1984.

Messina F. D., **Fluorescence Detection of Thin-Film Electrical Contact Lubricants.** IEEE Compon 7(1):47–55, Mar 1984.

Miller D. A. B., Gossard A. C., Wiegmann W., **Optical Bistability Due to Increasing Absorption.** Optics Lett 9(5):162–164, 1984.

Mottine J. J., Reagor B. T., **Investigation of Fretting Corrosion at Dissimilar Metal Interfaces in Socketed IC Device Applications.** IEEE Compon 7(1):61–68, Mar 1984.

Munson D. C., Strickland J. H., Walker T. P., **Maximum Amplitude Zero-Input Limit Cycles in Digital Filters.** IEEE Circ S 31(3):266–275, 1984.

Obenchain R. L., **Maximum-Likelihood Ridge Displays.** Comm St-The 13(2):227–240, 1984.

Ogorman L. et al., **The Converging Squares Algorithm—An Efficient Method for Locating Peaks in Multidimensions.** IEEE Patt A 6(3):280–288, 1984.

Olsson N. A., Dutta N. K., **Effect of External Optical Feedback on the Spectral Properties of Cleaved-Coupled-Cavity Semiconductor Lasers.** Appl Phys L 44(9):840–842, 1984.

Olsson N. A., Dutta N. K., Logan R. A., Besomi P., **Fiber-Dispersion and Propagation-Delay Measurements With Frequency-Modulated and Amplitude-Modulated Cleaved-Coupled-Cavity Semiconductor Lasers.** Optics Lett 9(5):180–182, 1984.

Olsson N. A., Tsang W. T., **An Optical Switching and Routing System Using Frequency Tunable Cleaved-Coupled-Cavity Semiconductor-Lasers (Letter).** IEEE J Q El 20(4):332–334, Apr 1984.

Ong E., Hu E. L., **Multilayer Resists for Fine Line Optical Lithography.** Sol St Tech 27(6):155–160, Jun 1984.

Ota Y., **Accelerated Ion Doping in Si MBE.** J Vac Sci A 2(2):393–400, Apr–Jun 1984.

Panock R., Forrest S. R., Kohl P. A., Dewinter J. C., Nahory R. E., Yanowski E. D., **An Experimental Low-Loss Single-Wavelength Bidirectional Lightwave Link.** J Lightw T 2(3):300–305, Jun 1984.

Pedersen O., Bonderup E., Golovchenko J., **The Influence of Multiple-Scattering on Channeling Radiation From GeV Positrons.** Nucl Inst B 230(1–3):83–89, 1984.

Phillips J. M., Yashinovitz C. J., **Epitaxial Relations in Alkaline-Earth Fluoride Semiconductor Systems.** J Vac Sci A 2(2):415–417, Apr–Jun 1984.

Schick G., **Galvanic Corrosion of Metals Coupled to Carbon Black Filler Polyethylene.** Mater Perf 23(4):47–53, 1984.

Segen J., **Locating Randomly Oriented Objects From Partial View.** P Soc Photo 449(P2):676–684, 1984.

Taylor G. N., **Guidelines for Publication of High-Resolution Resist Parameters.** Sol St Tech 27(6):105–110, Jun 1984.

Taylor G. N., **X-Ray Resist Trends.** Sol St Tech 27(6):124–131, Jun 1984.

Tsang W. T., **The Preparation of Materials for Optoelectronic Applications by Molecular-Beam Epitaxy.** J Vac Sci A 2(2):409–414, Apr–Jun 1984.

Tucker R. S., Lin C., Burrus C. A., Besomi P., Nelson R. J., **High-Frequency Small-Signal Modulation Characteristics of Short-Cavity InGaAsP Lasers.** Electr Lett 20(10):393–394, May 10 1984.

Van Der Ziel J. P., Mikulyak R. M., **Single-Mode Operation of 1.3 μm InGaAsP/InP Buried Crescent Lasers Using a Short External Optical Cavity.** IEEE J Q El 20(3):223–229, 1984.

Wang C. C., Moeller R. P., Burns W. K., Kaminow I. P., **Fibre-Optic Recirculating Analog Delay-Line.** Electr Lett 20(12):486–488, Jun 7 1984.

Wei V. K., **An Error-Trapping Decoder for Nonbinary Cyclic Codes (Letter).** IEEE Info T 30(3):538–541, May 1984.

White J. C., **Upconversion of Excimer Lasers Via Stimulated Anti-Stokes Raman Scattering (Letter).** IEEE J Q El 20(3):185–187, 1984.

White J. C., Henderson D., **An Indium Anti-Stokes Raman Laser (Letter).** IEEE J Q El 20(5):462–464, May 1984.

Whitt W., **Queue Tests for Renewal Processes.** Oper Res L 2(1):7–12, 1983.

Yeh Y. S., Schwartz S. C., **Outage Probability in Mobile Telephony Due to Multiple Log-Normal Interferers.** IEEE Commun 32(4):380–388, 1984.

MANAGEMENT/ECONOMICS

Dansby R. E., Conrad C., **Commodity Bundling.** Am Econ Rev 74(2):377–381, 1984.

Fishburn P. C., **Foundations of Risk Measurement. 1. Risk as Probable Loss.** Manag Sci 30(4):396–406, Apr 1984.

Goldman M. B., Leland H. E., Sibley D. S., **Optimal Nonuniform Prices.** Rev Econ S 51(2):305–319, 1984.

Majumdar M., Radner R., **Stationary Optimal Policies With Discounting in a Stochastic Activity Analysis Model.** Econometric 51(6):1821–1837, 1983.

Perry M. K., **Scale Economies, Imperfect Competition, and Public Policy.** J Ind Econ 32(3):313–333, 1984.

PHYSICAL SCIENCES

Agrawal G. P., **Level-Degeneracy Effects in Resonant Nonlinear Phenomena—Three-Level Atomic Model.** Pramana 22(3–4):293–301, 1984.

Ahlers G., Hohenberg P. C., Lucke M., **Externally Modulated Rayleigh-Benard Convection—Experiment and Theory.** Phys Rev L 53(1):48–51, Jul 2 1984.

Alpar M. A., Anderson P. W., Pines D., Shaham J., **Vortex Creep and the Internal Temperature of Neutron Stars. 2. Vela Pulsar.** Astrophys J 278(2):791–805, 1984.

Arwin H., Aspnes D. E., **Unambiguous Determination of Thickness and Dielectric Function of Thin-Films by Spectroscopic Ellipsometry.** Thin Sol Fi 113(2):101–113, 1984.

Aspnes D. E., Kelso S. M., **Fourier-Analysis of Optical Spectra—Application to $Al_xGa_{1-x}As$ and $GaAs_{1-x}Px$.** P Soc Photo 452.

Barry S. R., Gelperin A., **Acetylcholine Turnover in an Autoactive Molluscan Neuron.** Cell Mol N 4(1):15–29, Mar 1984.

Bean J. C., Feldman L. C., Fiory A. T., Nakahara S., Robinson I. K., **Ge_xSi_{1-x}/Si Strained-Layer Superlattice Grown by Molecular-Beam Epitaxy.** J Vac Sci A 2(2):436–440, Apr–Jun 1984.

Bertz S. H., Dabbagh G., Cook J. M., Honkan V., **Organocuprate Reactions With Cyclopropanes—Evidence for Three Types of Mechanism.** J Org Chem 49(10):1739–1743, May 18 1984.

Bohn P. W., Harris T. D., Bhat R., Cox H. M., **Selectively Excited Luminescence in GaAs.** Appl Spectr 38(3):417–422, May–Jun 1984.

Bohrer M. P., Patterson G. D., Carroll P. J., **Hindered Diffusion of Dextran and Ficoll in Microporous Membranes.** Macromolec 17(6):1170–1173, Jun 1984.

Boring J. W., Garrett J. W., Cummings T. A., Johnson R. E., Brown W. L., **Sputtering of Solid SO_2.** Nucl Inst B 229(2–3):321–326, 1984.

Bovey F. A., **NMR and Macromolecules.** ACS Symp S (247):3–17, 1984.

Bowers J. E., Bjorkholm J.E., Burrus C. A., Coldren L. A., Hemenway B. R., Wilt D. P., **Cleaved-Coupled-Cavity Lasers With Large Cavity Length Ratios for Enhanced Stability.** Appl Phys L 44(9):821–823, 1984.

Brown W. L., Augustyniak W. M., Marcantonio K. J., Simmons E. H., Boring J. W., Johnson R. E., Reimann C. T., **Electronic Sputtering of Low-Temperature Molecular-Solids.** Nucl Inst B 229(2–3):307–314, 1984.

Bruch M. D., Bovey F. A., **Proton-Resonance Assignments in Copolymer Spectra by Two-Dimensional NMR (Letter).** Macromolec 17(4):978–981, 1984.

Bruinsma R., Aeppli G., **Interface Motion and Nonequilibrium Properties of the Random-Field Ising Model.** Phys Rev L 52(17):1547–1550, 1984.

Brus L. E., **Electron Electron and Electron-Hole Interactions in Small Semiconductor Crystallites—The Size Dependence of the Lowest Excited Electronic State.** J Chem Phys 80(9):4403–4409, 1984.

Cais R. E., Kometani J. M., **The Synthesis of Novel Regioregular Polyvinyl Fluorides and Their Characterization by High-Resolution NMR.** ACS Symp S (247):153–165, 1984.

Capasso F., Kasper B., Alavi K., Cho A. Y., Parsey J. M., **New Low Dark Current, High-Speed $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ Avalanche Photodiode by Molecular-Beam Epitaxy for Long Wavelength Fiber Optic Communication Systems.** Appl Phys L 44(11):1027–1029, Jun 1 1984.

Cava R. J., Murphy D. W., Zahurak S., Santoro A., Roth R. S., **The Crystal Structures of the Lithium-Inserted Metal-Oxides $Li_{0.5}TiO_2$ Anatase. $LiTi_2O_4$ Spinel, and $Li_2Ti_2O_4$.** J Sol St Ch 53(1):64–75, Jun 1984.

Chen C. Y., Kasper B. L., Cox H. M., **High-Sensitivity $Ga_{0.47}In_{0.53}As$ Photoconductive Detectors Prepared by Vapor-Phase Epitaxy.** Appl Phys L 44(12):1142–1144, Jun 15 1984.

Chen C. Y., Pang Y. M., Cho A. Y., Alavi K., Garbinski P. A., **Modulation-Doped $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ Photodetector Prepared by Molecular-Beam Epitaxy.** J Vac Sci B 2(2):262–264, Apr–Jun 1984.

Chen M. C., Lang D. V., Dautremont-Smith, W. C., Sergent A. M., Harbison J. P., **Effects of Leakage Current on Deep Level Transient Spectroscopy.** Appl Phys L 44(8):790–792, 1984.

Chin B. H., Frahm R. E., Sheng T. T., Bonner W. A., **Carrier Saturation in Tin-Doped InP Films Grown by Liquid-Phase Epitaxy.** J Elchem So 131(6):1373–1374, Jun 1984.

Chu S., Mills A. P., Hall J. L., **Measurement of the Positronium 13S1-23S1 Interval by Doppler-Free Two-Photon Spectroscopy.** Phys Rev L 52(19):1689–1692, 1984.

Cladis P. E., Brand H. R., **Novel Liquid-Crystal Phase-Diagram—An Inverted Cholesteric Phase Surrounded by Different Types of Smectic-A Phases.** Phys Rev L 52(25):2261–2264, Jun 18 1984.

Comin F., Rowe J. E., Citrin P. H. **SEXAFS Studies of Nickel Silicide Nucleation on Si(111).** P Soc Photo 447:107–116, 1984.

Connor J. A., Ahmed Z., **Diffusion of Ions and Indicator Dyes in Neural Cytoplasm.** Cell Mol N 4(1):53–66, Mar 1984.

Connor J., Alkon D. L., **Light-Dependent and Voltage-Dependent Increases of Calcium-Ion Concentration in Molluscan Photoreceptors.** J Neurphysl 51(4):745–752, 1984.

Craft D. C., Dutta N. K., Wagner W. R., **Anomalous Polarization Characteristics of 1.3-μm InGaAsP Buried Heterostructure Lasers.** Appl Phys L 44(9):823–825, 1984.

Craighead H. G. et al., **Characterization and Optical Properties of Arrays of Small Gold Particles.** Appl Phys L 44(12):1134–1136, Jun 15 1984.

Craighead H. G., **Ten-NM Resolution Electron-Beam Lithography.** J Appl Phys 55(12):4430–4435, Jun 15 1984.

Crochet M. J. et al., **Numerical-Simulation of the Horizontal Bridgman Growth of a Gallium-Arsenide Crystal.** J Cryst Gr 65(1–3):166–172, 1983.

Crochet M. J., Wouters P. J., Geyling F. T., Jordan A. S., **Finite-Element Simulation of Czochralski Bulk Flow.** J Cryst Gr 65(1–3):153–165, 1983.

Daniels P. G., Simpkins P. G., **The Flow Induced by a Heated Vertical Wall in a Porous Medium.** Q J Mech Ap 37(May):339–354, 1984.

Darcie T. E., Whalen M. S., **Determination of Optical Constants Using Pseudo-Brewster Angle and Normal Incidence Reflectance Measurements (Letter).** Appl Optics 23(8):1130–1131, 1984.

Davies J. H., Lee P. A., Rice T. M., **Properties of the Electron Glass.** Phys Rev B 29(8):4260–4271, 1984.

Donnelly V. M., Geva M., Long J., Karlicek R. F., **Excimer Laser-Induced Deposition of InP and Indium-Oxide Films.** Appl Phys L 44(10):951–953, 1984.

Dubois L. H., Nuzzo R. G., **Reactivity of Intermetallic Thin-Films Formed by the Surface Mediated Decomposition of Main Group Organometallic Compounds.** J Vac Sci A 2(2):441–445, Apr–Jun 1984.

Dutta N. K., Craft D. C., **Effect of Stress on the Polarization of Stimulated-Emission From Injection Lasers.** J Appl Phys 56(1):65–70, Jul 1 1984.

Dutta N. K., Olsson N. A., Heritage J. P., Liu P. L., **Temperature-Dependence of Threshold Current of Injection Lasers for Short Pulse Excitation.** Appl Phys L 44(10):943–944, 1984.

Ebeling K. J., Coldren L. A., **Optoelectronic Properties of Coupled Cavity Semiconductor Lasers.** Appl Phys L 44(8):735–737, 1984.

Eibschutz M., Lines M. E., Van Uitert L. G., Guggenheim H. J., Zydzik G. J., **Mossbauer Study of Amorphous FeF_3.** Phys Rev B 29(7):3843–3851, 1984.

Eisinger J., Flores J., Bookchin R. M., **The Cytosol-Membrane Interface of Normal and Sickle Erythrocytes—Effect of Hemoglobin Deoxygenation and Sickling.** J Biol Chem 259(11):7169–7177, Jun 10 1984.

Eizenberg M., Brener R., Murarka S. P., **Thermal-Stability of the Aluminum Titanium Carbide Silicon Contact System.** J Appl Phys 55(10):3799–3803, 1984.

Elman B. S. et al., **Observation of Two-Dimensional Ordering in Ion-Damaged Graphite During Post-Implantation Annealing.** Phys Rev B 29(8):4703–4708, 1984.

Fiory A. T., Hebard A. F., **Electron-Mobility, Conductivity, and Superconductivity Near the Metal-Insulator Transition.** Phys Rev L 52(23):2057–2060, Jun 4, 1984.

Fisher D. S., Frohlich J., Spencer T., **The Ising Model in a Random Magnetic-Field.** J Stat Phys 34(5–6):863–870, Mar 1984.

Frankenthal R. P., Siconolfi D. J. **Effect of Ion Mass and Ion Energy on the Surface-Composition of Two Sputtered Tin Lead Alloys.** J Vac Sci A 2(2):1089–1092, Apr–Jun 1984.

Glarum S. H., Marshall J. H., **An Admittance Study of the Lead Electrode.** J Elchem So 131(4):691–701, 1984.

Glass A. M., Johnson A. M., Olson D. H., Simpson W., Ballman A. A., **Four-Wave Mixing in Semi-Insulating InP and GaAs Using the Photorefractive Effect.** Appl Phys L 44(10):948–950, 1984.

Gooden R., Winslow F. H., Hutton R. S., Hellman M. Y. **The Solid-State Photochemistry of Poly(ethylene-Carbon Monoxide)—Structural Characterization of Photodegradation and Photooxidation.** Polym Prepr 25(1):50–51, 1984.

Gossmann H. J., Bean J. C., Feldman L. C., Gibson W. M., **Observation of a (5 × 5) Leed Pattern From Ge_xSi_{1-x}(111) Alloys (Letter).** Surf Sci 138(2–3):L175–L180, 1984.

Graebner J. E., Allen L. C., **Thermal-Conductivity of Amorphous-Germanium at Low-Temperatures.** Phys Rev B 29(10):5626–5633, May 15 1984.

Graedel T. E., Franey J. P., Kammlott G. W., **Ozone-Enhanced and Photon-Enhanced Atmospheric Sulfidation of Copper.** Science 224(4649):599–601, 1984.

Green M. L., Levy R. A., Nuzzo R. G., Coleman E., **Aluminum Films Prepared by**

Metal Organic Low-Pressure Chemical Vapor-Deposition. Thin Sol FI 114(4):367–377, Apr 27 1984.

Greene B. I., Scott T. W., **Time-Resolved Multiphoton Ionization in the Organic Condensed Phase—Picosecond Conformational Dynamics of Cis-Stilbene and Tetraphenylethylene.** Chem P Lett 106(5):399–402, May 4 1984.

Gresillon D., Olivain J., Truc A., Lehner T., Surko C. M., **The Ion Feature in a Laboratory Plasma—Theory and Experiment Using CO_2-Laser Light Scattering.** Phys Fluids 27(4):1030–1040, 1984.

Grier B. H., Shapiro S. M., Cava R. J., **Inelastic Neutron-Scattering Measurements of the Diffusion in Beta-Ag_2S.** Phys Rev B 29(7):3810–3814, 1984.

Haight R., Feldman L. C., Buck T. M., Gibson W. M., **Neutralization of Energetic He Ions Scattered From Clean Two × One Si (100).** Nucl Inst B 230(1–3):501–504, 1984.

Hannon J. J., Cook J. M., **Oxidative Removal of Photoresist by Oxygen Freon-116 Discharge Products.** J Elchem SO 131(5):1164–1169, 1984.

Hasegawa A. et al. **Study of Reversed Field Pinch With Surface Current.** J Phys Jpn 53(4):1316–1325, Apr 1984.

Hauser J. J., **Evidence for Tunnelling Relaxation in the AC Conductivity of Chalcogenide Glasses.** Sol St Comm 50(7):623–626, 1984.

Hebard A. F., Blonder G. E., Suh S. Y., **Optical-Recording Applications of Reactive Ion-Beam Sputter Deposited Thin-Film Composites.** Appl Phys L 44(11):1023–1025, Jun 1 1984.

Heimann P. A., Schutz R. J., **Optical Etch-Rate Monitoring—Computer-Simulation of Reflectance.** J Elchem So 131(4):881–885, 1984.

Henderson J. B., Verma Y. P., Tant M. R., Moore G. R., **Measurement of the Thermal-Conductivity of Polymer Composites to High-Temperatures Using the Line-Source Technique.** High Temp T 2(2):107–112, May 1984.

Hensel J. C., Tung R. T., Poate J. M., Unterwald F. C., **Electrical Transport Properties of $CoSi_2$ and $NiSi_2$ Thin Films.** Appl Phys L 44(9):913–915, 1984.

Heyward I. P., Chan M. G., Ludwick A. G., **The Effect of Antioxidants on the Thermo-Oxidative Stabilities of UV-Cured Coatings.** Polym Prepr 25(1):44–45, 1984.

Holmes R. J., Smyth D.M., **Titanium Diffusion Into $LiNbO_3$ as a Function of Stoichiometry.** J Appl Phys 55(10):3531–3535, 1984.

Hopfield J. J., **Neurons With Graded Response Have Collective Computational Properties Like Those of Two-State Neurons.** P Nas Biol 81(10):3088–3092, May 1984.

Huber D. L., Broer M. M., Golding B., **Low-Temperature Optical Dephasing of Rare-Earth Ions in Glass.** Phys Rev L 52(25):2281–2284, Jun 18 1984.

Huse D. A., **Melting of a Physisorbed Commensurate Phase.** Phys Rev B 29(9):5031–5038, 1984.

Hwang J. C. M., Kastalsky A., Stormer H.L., Keramidas V. G., **Transport Properties of Selectively Doped GaAs-(AlGa)As Heterostructures Grown by Molecular-Beam Epitaxy.** Appl Phys L 44(8):802–804, 1984.

Ibbotson D. E., Flamm D. L., Mucha J. A., Donnelly V. M., **Comparison of XeF_2 and F-Atom Reactions With Si and SiO_2.** Appl Phys L 44(12):1129–1131, Jun 15 1984.

Jelinski L. W., Dumais J. J., Cholli A. L., **Characterization of Polymer Motions Using Solid-State Deuterium NMR Spectroscopy.** Polym Prepr 25(1):348–350, 1984.

Johnson A. M., Stolen R. H., Simpson W. M., **80X Single-Stage Compression of Frequency Doubled Nd-Yttrium Aluminum Garnet Laser Pulses.** Appl Phys L 44(8):729–731, 1984.

Jordan A. S., Nikolakopoulou G. A., **A Numerical Study of Manganese Redistribution in GaAs Employing an Interstitial-Substitutional Model.** J Appl Phys 55(12):4194–4207, Jun 15 1984.

Kaminow I. P., Wiesenfeld J. M., Choy D. S. J., **Argon-Laser Disintegration of Thrombus and Atherosclerotic Plaque (Letter).** Appl Optics 23(9):1301–1302, 1984.

Kevan S. D., **Photoelectron Diffraction—Present Applications and Future Prospects.** P Soc Photo 447:74–81, 1984.

Khorami J. et al., **Interpretation of EGA and DTG Analyses of Chrysotile Asbestos.** Thermoc Act 76(1–2):87–96, May 15 1984.

Klauder J. R., **Coherent-State Langevin Equations for Canonical Quantum Systems With Applications to the Quantized Hall Effect.** Phys Rev A 29(4):2036–2047, 1984.

Kuk Y., Feldman L. C., Robinson I. K., **Atomic Displacements in the Au(110)-(1 × 2) Surface (Letter).** Surf Sci 138(2–3):L168–L174, 1984.

Kuo C. Y., Patel C. K. N., **Direct Measurement of Optoacoustic Induced Ultrasonic Waves.** Appl Phys L 44(8):752–754, 1984.

Kuo C. Y., Vieira M. M. F., Patel C. K. N., **Transient Optoacoustic Pulse Generation and Detection.** J Appl Phys 55(9):3333–3336, 1984.

Lake G., **Windows on a New Cosmology.** Science 224(4650):675–681, 1984.

Lake G., Schommer R. A., **Mass-to-Light Ratios for Binary Pairs of Dwarf Irregular Galaxies.** Astrophys J 279(1):L19–L22, 1984.

Laudise R. A., **Crystal-Growth Progress in Response to the Needs for Optical Communications.** J Cryst Gr 65(1–3):3–23, 1983.

Levy L. P. et al., **ColPrices.** Rev Econ S 51(2):305–319, 1984.

Lin C. L., Burrus C. A., Eisenstein G., Tucker R. S., Besomi P., Nelson R. J., **11.2 GHz Picosecond Optical Pulse Generation in Gain-Switched Short-Cavity InGaAsP Injection-Lasers by High-Frequency Direct Modulation.** Electr Lett 20(6):238–240, 1984.

Lines M. E., **Scattering Losses in Optic Fiber Materials. 1. A New Parametrization.** J Appl Phys 55(11):4052–4057, Jun 1 1984.

Lines M. E., **Scattering Losses in Optic Fiber Materials. 2. Numerical Estimates.** J Appl Phys 55(11):4058–4063, Jun 1 1984.

Lu N. C. C., Lu C. Y., Lee M. K., Shih C. C., Wang C. S., Reuter W., Sheng T. T., **The Effect of Film Thickness on the Electrical Properties of LPCVD Polysilicon Films.** J Elchem So 131(4):897–902, 1984.

Luongo J. P., **IR Study of Amorphous-Silicon Nitride Films.** Appl Spectr 38(2):195–199, 1984.

Macrande A. T., Schwartz B., Focht M. W., **Deep and Shallow Levels in N-Type Indium-Phosphide Irradiated With 200-keV Deuterons.** J Appl Phys 55(10):3595–3602, 1984.

Martin P. et al., **Observation of Exceptional Temperature Humidity Stability in Multilayer Filter Coatings (Letter).** Appl Optics 23(9):1307–1308, 1984.

Mattheiss L. F., Hamann D. R., **Electronic-Structure of the Tungsten (001) Surface.** Phys Rev B 29(10):5372–5381, May 15 1984.

McBrierty V.J., Douglass D. C., Furukawa T., **A Nuclear Magnetic-Resonance Study of Poled Vinylidene Fluoride Trifluoroethylene Copolymer.** Macromolec 17(6):1136–1139, Jun 1984.

Mims W.B., Davis J. L., Peisach J., **The Accessibility of Type-I Cu(II) Centers in Laccase, Azurin, and Stellacyanin to Exchangeable Hydrogen and Ambient Water.** Biophys J 45(4):755–766, 1984.

Murray C. A., Bodoff S., **Depolarization Effects in Raman-Scattering From Monolayers on Surfaces—The Classical Microscopic Local Field.** Phys Rev L 52(25):2273–2276, Jun 18 1984.

Nakahara S., Kinsbron E., **Room-Temperature Interdiffusion Study of Au/Ga Thin-Film Couples.** Thin Sol Fi 113(1):15–26, 1984.

Nash D. L., Simpson J. R., Wood, D. L., **Quantitative Emission Spectrographic Determination of Ge in MCVD Deposits.** Am Ceram S 63(5):726–729, May 1984.

Novembre A., Bowmer T. N., **A Novel Technique for Determining Radiation Chemical Yields of Negative Electron-Beam Resists.** Polym Prepr 25(1):324–325, 1984.

Nuzzo R. G., Smolinski G., **Preparation and Characterization of Functionalized Polyethylene Surfaces.** Macromolec 17(5):1013–1019, May 1984.

Orenstein J., Etemad S., Baker G. L., **Photoinduced Absorption in a Polydiacetylene. (Letter).** J Phys C 17(10):L297–L300, 1984.

Paalanen M. A., Tsui D. C., Gossard A. C., Hwang J. C. M., **Disorder and the Fractional Quantum Hall Effect.** Sol ST Comm 50(9):841–844, 1984.

Paoletti A. et al., **Nonlinearity Effects in the Current-Voltage Characteristic of P-Type Yttrium Iron-Garnet Epitaxial-Films.** J Appl Phys 55(10):3699–3701, 1984

Patterson G. D., Carroll P. J., Stevens J. R., Wilson W., Bair H. E., **Hypersonic Attenuation in Poly(dimethylsiloxane) as a Function of Temperature and Pressure.** Macromolec 17(4):885–888, 1984.

Pearson D. S., Helfand E., **Viscoelastic Properties of Star-Shaped Polymers.** Macromolec 17(4):888–895, 1984.

Peterson G. E., Kurkjian C. R., Carnevale A., **Summarization of Glassy Spectral Data With Particular Reference to the Boron Anomaly.** J Am Ceram 67(5):319–324, May 1984.

Phillips J. C., **Microscopic Origin of Anomalously Narrow Raman Lines in Network Glasses.** J Non-Cryst 63(3):347–355, 1984.

Phillips J. C., **Microscopic Theory of Covalent-Ionic Transition of Amorphizability of Nonmetallic Solids.** Phys Rev B 29(10):5683–5686, May 15 1984.

Phillips J. C., **Physics of Ag Photodoping in Chalcogenide Alloy Glasses.** J Non-Cryst 64(1–2):81–85, 1984.

Pinczuk A., Shah J., Miller R. C., Gossard A. C., Wiegmann W., **Optical Processes of Two-Dimensional Electron-Plasma in GaAs-(AlGa) as Heterostructures.** Sol St Comm 50(8):735–739, 1984.

Pirronello V. et al., **Formaldehyde Formation in Cometary Nuclei.** Astron Astr 134(2):204–206, May 1984.

Pokor J. et al., **Generation of 35-nm Coherent Radiation.** P Soc Photo 461.

Pryde C. A., **Weathering of Polycarbonates—A Survey of the Variables Involved.** Polym Prepr 25(1):52–53, 1984.

Raghavachari K. et al., **Theoretical Study of Sn_2 Reactions Involving Cationic Substrates.** J Am Chem S 106(11):3124–3128, May 30 1984.

Ravaine D., Nassau K., Glass A. M., **The Ionic-Conductivity of Rapidly Quenched Tungstate and Molybdate Glasses Containing Lithium Halides.** Sol St Ion 13(1):15–20, Apr 1984.

Rentzepis P. M., Bondybey V. E., **Large Molecule Relaxation—Spectroscopy, Structure, and Vibrational-Energy Redistribution in Naphthazarin.** J Chem Phys 80(10):4727–4737, May 15 1984.

Rice C. E., Jackel J. L., **Structural Changes With Composition and Temperature in Rhombohedral $Li_{1-x}H_xNbO_3$.** Mater Res B 19(5):591–597, May 1984.

Rocklin S. M., Kashper A., Varvaloucas G. C., **Capacity Expansion Contraction of a Facility With Demand Augmentation Dynamics.** Operat Res 32(1):133–147, Jan–Feb 1984.

Rongved L. **Ray-Propagation in an Inviscid Fluid.** Nuov Cim B 80(1):109–120, 1984.

Rossetti R., Ellison J. L., Gibson J. M., Brus L. E., **Size Effects in the Excited Electronic States of Small Colloidal CDS Crystallites.** J Chem Phys 80(9):4464–4469, 1984.

Rousseau D. L., Tan S. L., Ondrias M. R., Ogawa S., Noble R. W., **Absence of Cooperative Energy at the Heme in Liganded Hemoglobins.** Biochem 23(13):2857–2865, Jun 19 1984.

Royer W. A., Smith N. V., **Two-Stage Electron-Energy Analyzer for Angle-Resolved Photoemission Spectroscopy.** Rev Sci Ins 55(6):909–911, Jun 1984.

Rust R. D., Rhodes R. J., Parker A. A., **Uniform Plasma Etching of Printed Circuit Boards.** Sol St Tech 27(4):270–275, 1984.

Schaefer J., Sefcik M. D., Stejskal E. O., McKay R. A., Dixon W. T., Cais R. E., **Molecular Motion in Glassy Polystyrenes.** ACS Symp S (247):43–54, 1984.

Schaefer J., Sefcik M. D., Stejskal E. O., McKay R. A., Dixon W. T., Cais R. E., **Molecular Motion in Glassy Polystyrenes.** Macromolec 17(6):1107–1118, Jun 1984.

Schneemeyer L. F., Spengler S. E., DiSalvo F. J., Waszczak J. V., **Preparation and Properties of Reduced Bismuth and Antimony Molybdenum Oxides.** Mater Res B 19(4):525–529, 1984.

Siconolfi D. J., Frankenthal R. P., **The Determination of Oxide Film Thickness**

and Composition of Indium and Chromium by Decomposition of Auger-Electron Spectra. Corros Sci 24(2):137–144, 1984.

Simard A. J. et al., **A General Procedure for Sampling and Analyzing Wildland Fire Spread.** Forest Sci 30(1):51–64, Mar 1984.

Snyder L. C., **Abinitio Quantum Chemical Study of the Dimerization of Silicon Monoxide.** J Chem Phys 80(10):5076–5079, May 15, 1984.

Snyder L. C., **Modified Milk Stool on Wurtzite Layer Model for Si(111) 7 × 7 Surface Reconstruction.** Surf Sci 140(1):101–107, May 1984.

Stall R. A., Swaminathan V., Schumaker N., **Reduction of Arsenic Oxide Contamination in Molecular-Beam Epitaxy Vacuum Chambers.** J Vac Sci B 2(2):148–150, Apr–Jun 1984.

Stark A. A., Carlson E. R., **The Molecular Halo of M82.** Astrophys J 279(1):122–124, 1984.

Stillinger F. H., Weber T. A., **Inherent Pair Correlation in Simple Liquids.** J Chem Phys 80(9):4434–4437, 1984.

Swaminathan V., Dautremont-Smith W. C., Anthony P. J., **Photoluminescence Changes Under High-Intensity Excitation of Al_2O_3-Coated GaAs and Their Relevance to Evaluation of Facet Coating Procedures.** Mater Lett 2(3):179–183, 1984.

Tarascon R., Hartney M., Bowden M. J., **Synthesis, Characterization and Lithographic Evaluation of Chlorinated Polymethylstyrene.** Polym Prepr 25(1):289–290, 1984.

Teo B. K., **New Topological Electron-Counting Theory.** Inorg Chem 23(9):1251–1257, 1984.

Teo B. K., Longoni G., Chung F. R. K., **Applications of Topological Electron-Counting Theory to Polyhedral Metal Clusters.** Inorg Chem 23(9):1257–1266, 1984.

Thompson M. O., Galvin G. J., Mayer J. W., Peercy P. S., Poate J. M., Jacobson D. C., Cullis A. G., Chew N. G., **Melting Temperature and Explosive Crystallization of Amorphous Silicon During Pulsed Laser Irradiation.** Phys Rev L 52(26):2360–2363, Jun 25 1984.

Thurston R. N., **Equilibrium Distributions of Electric-Field in a Cell With Adsorbed Charge at the Surfaces.** J Appl Phys 55(12):4154–4161, Jun 15 1984.

Thurston R. N., Boyd G. D., Senft D. C., **Thickness Dependence of the Switching Time of the Bistable Boundary-Layer Liquid-Crystal Display.** Appl Phys L 44(8):813–815, 1984.

Thurston R. N., Boyd G. D., Senft D. C., **Experiments in Switching the Bistable Boundary-Layer Liquid-Crystal Display by the Application of DC.** J Appl Phys 55(10):3846–3855, 1984.

Tolk N. H. et al., **Desorption Induced by Electronic Transitions.** Nucl Inst B 230(1–3):457–460, 1984.

Tolk N. H. et al., **Resonant Neutralization of He Ions Into Excited-States at Cu(110) and Ni(110) Surfaces.** Nucl Inst B 230(1–3):488–490, 1984.

Tomlinson W. J., Stolen R. H., Shank C. V., **Compression of Optical Pulses Chirped by Self-Phase Modulation in Fibers.** J Opt Soc B 1(2):139–149, 1984.

Tsang W. T., **Growth of Bright (300K) Luminescence $InAs_xP_{1-x}$ X(Lambda = 1.7–2.1-μ) on InP Substrates by Molecular-Beam Epitaxy.** J Appl Phys 55(8):2901–2903, 1984.

Vaidya S., Schutz R. J., Sinha A. K., **Shallow Junction Cobalt Silicide Contacts With Enhanced Electromigration Resistance.** J Appl Phys 55(10):3514–3517, 1984.

Vandenberg J. M., Temkin H., **An In situ X-Ray Study of Gold Barrier-Metal Interactions With InGaAsP/InP Layers.** J Appl Phys 55(10):3676–3681, 1984.

Vansaarloos W., Kurtze D. A., **Location of Zeros in the Complex Temperature Plane—Absence of Lee-Yang Theorem.** J Phys A 17(6):1301–1311, 1984.

Vasile M. J., **Velocity Dependence of Secondary-Ion Emission.** Phys Rev B 29(7):3785–3794, 1984.

Venkatesan T., Dynes R. C., Wilkens B., White A. E., Gibson J. M., Hamm R., **Comparison of Conductivity Produced in Polymers and Carbon Films by Pyrolysis and High-Energy Ion Irradiation.** Nucl Inst B 229(2–3): 599–604, 1984.

Venkatesan T., Edelson D., Brown W. L., **Ionization Induced Decomposition and Diffusion in Thin Polymer Films.** Nucl Inst B 229(2–3):286–290, 1984.

Vonseggern H., West J. E., Kubli R. A., **Determination of Charge Centroids in Two-Side Metallized Electrets.** Rev Sci Ins 55(6):964–967, Jun 1984.

Weeks J. D., **Scaling Relations Between Correlations in the Liquid-Vapor Interface and the Interface Width.** Phys Rev L 52(24):2160–2163, Jun 11 1984.

Weeks J. D., Bedeaux D., Zielinski B. J., **Anisotropic Vanderwaals Model of the Liquid Vapor Interface.** J Chem Phys 80(8):3790–3800, 1984.

Weinberger B. R., Roxlo C. B., Etemad S., Baker G. L., Orenstein J., **Optical-Absorption in Polyacetylene—A Direct Measurement Using Photothermal Deflection Spectroscopy.** Phys Rev L 53(1):86–89, Jul 2 1984.

Wertheim G. K., Citrin P. H., **Surface Atom Core Level Shifts in the Noble Metals.** P Soc Photo 447:47–51, 1984.

Wilson T., Veselka J. J., Jackel J. L., **A Liquid-Crystal Cutoffs Modulator for Multimode Optical Waveguides.** Optik 67(1):37–41, 1984.

Wood O. R., Macklin J. J., Silfvast W. T., **Single-Ion Recombination Lasers in CO_2 Laser-Vaporized Target Material.** Appl Phys L 44(12):1123–1125, Jun 15 1984.

Yen R., Downey P. M., Shank C. V., Auston D. H., **Low Jitter Streak Camera Triggered by Subpicosecond Laser Pulses.** Appl Phys L 44(8):718–720, 1984.

SOCIAL AND LIFE SCIENCES

Lucky R. W., **Entertain Me—Television's Omnipresence.** IEEE Spectr 21(6):85–89, Jun 1984.

Sternberg S., **Stage Models of Mental Processing and the Additive-Factor Method.** Behav Brain 7(1):82–84, 1984.

SPEECH/ACOUSTICS

Jain V. K., Crochiere R. E., **Quadrature Mirror Filter Design in the Time Domain.** IEEE Acoust 32(2):353–361, 1984.

Rabiner L. R., Wilpon J. G., Quinn A. M., Terrace S. G., **On the Application of Embedded Digit Training to Speaker Independent Connected Digit Recognition.** IEEE Acoust 32(2):272–280, 1984.

Schroeder M R., **Progress in Architectural Acoustics and Artificial Reverberation—Concert Hall Acoustics and Number Theory.** J Aud Eng S 32(4):194–203, 1984.

Vysotsky G. J., **A Speaker-Independent Discrete Utterance Recognition System, Combining Deterministic and Probabilistic Strategies.** IEEE Acoust 32(3):489–499, Jun 1984.

CONTENTS, DECEMBER 1984

Part 1

Part 2

The *SLC*-96 Subscriber Loop Carrier System

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