Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/5521--95-7744

Novel Techniques for the Analysis of Wireless Integrated Voice/Data Networks

JEFFREY E. WIESELTHIER
CRAIG M. BARNHART

Communication Systems Branch
Information Technology Division

ANTHONY EPHREMIDES

Kaman Sciences Corporation
Alexandria, Virginia

and

University of Maryland
College Park, Maryland

July 24, 1995

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 24, 1995	Interim Report 6/93-5/95

4. TITLE AND SUBTITLE

Novel Techniques for the Analysis of Wireless Integrated Voice/Data Networks

5. FUNDING NUMBERS

PE - 61153N
PR - RR015-09-41
WU - DN159-036

6. AUTHOR(S)

Jeffrey E. Wieselthier, Craig M. Barnhart, and
Anthony Ephremides*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/5521--95-7744

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Arlington, VA 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

*University of Maryland and Kaman Sciences Corporation

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this report, we consider the evaluation of data-packet delay in wireless integrated voice/data networks. In networks that support circuit-switched voice, the voice occupancy process satisfies a product-form solution under reasonable modeling assumptions. Although this product-form solution provides an accurate characterization of equilibrium voice-traffic behavior, it does not directly provide a method to evaluate data-packet delay. However, examination of each link separately in a manner that incorporates interaction with the rest of the network permits us to take advantage of the wireless nature of the network and obtain a three-flow characterization of each link, which also satisfies a product-form solution and is hence termed a "mini-product-form" solution. By matching the values of these flows to the average values obtained from the product-form solution of the entire network, we obtain a three-dimensional Markov chain characterization of the voice occupancy state on the link, which permits a simpler evaluation of data-packet delay. A further reduction is possible by converting the three-dimensional chain to a single-dimensional one. Performance results demonstrate that these models provide satisfactory delay estimates that also appear to be upper bounds on delay.

14. SUBJECT TERMS

Communications network
Voice/data integration
Radio network
Performance evaluation
Product form

15. NUMBER OF PAGES

37

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribe by ANSI Std 239-18
298-102

CONTENTS

Novel Techniques for the Analysis of Wireless Integrated Voice/Data Networks

1. Introduction

The accurate prediction of average data delay in a multihop network is a notoriously difficult problem that has been extensively addressed in the literature [1, 2]. This problem is especially difficult for integrated voice/data networks in which the quantity of resources available for data traffic (channels in the case of wireline networks and transceivers in the case of wireless networks) depends on the current state of the voice traffic. Furthermore, for a wireless network, the difficulty is compounded by the interaction amongst the transmissions of neighboring nodes.

In this report we address the development of models for data-packet delay evaluation in multihop wireless integrated networks. A major obstacle to the development of accurate delay models has been the lack of an accurate characterization of the time-varying nature of voice traffic in such networks. Such a characterization is necessary because models based simply on the expected quantity of resources available for data, or even on the equilibrium distribution of the quantity of resources available for data, can be highly inaccurate. Based on a standard model for circuit-switched voice networks, our starting point is the use of the product-form solution to characterize the state of voice-call traffic. Although such models are normally used to characterize equilibrium, rather than time-varying, behavior, we show that the product-form solution can be used in a novel way to develop an approximate Markovian model that offers a practical and useful characterization of the voice occupancy of any link in the network, which in turn provides the time-varying "residual capacity" available for data, and hence serves as the basis for the evaluation of data-packet delay. We compare our model to one based on the well-known reduced-load approximation, and find significant improvement in accuracy.

1.1 Earlier studies and their relation to this report

This report is part of our continuing study of voice/data integration in wireless networks, in which we have addressed several different aspects of design and performance analysis. Some of our most relevant publications are summarized below. For example, in [3-6] we developed

Manuscript approved May 10, 1995

and evaluated the WIMA channel-access protocol for integrated networks. The Markovian system model developed for that protocol has served as the basis for the multidimensional Markovian model for system evolution in integrated wireless networks that we use in this report for data-packet delay evaluation. In [7-9] we addressed the problem of voice-call admission control in integrated wireless networks, taking into consideration the product-form of the equilibrium voice-call state that holds under reasonable modeling assumptions. We developed both a recursive search procedure and a descent-search technique to speed up the determination of the optimal admission-control policy. Our use of the product-form solution to characterize equilibrium behavior in those studies has served as the basis for its use in dynamic system models in the current report. Also of major significance are our studies of ordinal optimization methods for the determination of good admission-control policies in integrated wireless networks. In [10-13] we demonstrated that crude analytical or simulation models with low complexity can provide nearly optimal performance, although the delay estimates they provide are typically poor. In the present report, we focus on the development of models for data-packet delay evaluation in integrated wireless networks. Portions of this work have been presented recently at IEEE INFOCOM'95 [14]. Many of our recent papers are available by anonymous ftp from mars.itd.nrl.navy.mil, in the directory "nrlpapers/code_5521/W.E.B.", or on the WWW at http://tang.itd.nrl.navy.mil/WEBindex.html.

1.2 Outline of the Report

In Section 2 we discuss our integrated voice/data multihop wireless network model, and address the differences between wireless and wired networks. In Section 3 we discuss issues in the development of delay models for integrated networks, including the need to characterize the time-varying behavior of the voice-call process (and hence the residual capacity) over individual network links. In Section 4 we develop our approximate "mini-product-form" model, which provides a three-dimensional first-order Markov chain description of the voice-call process in the system; this provides a considerable reduction in complexity over the exact Markov chain description. Then in Section 5 we demonstrate how the voice-call description can be reduced to an approximate one-dimensional first-order Markov chain, with little loss of accuracy, thus increasing significantly the size of the problems that can be handled with the mini-product-form network approach. In Section 6 we present performance results that demonstrate the accuracy of our models, and in Section 7 we present our conclusions.

2. Integrated wireless network model

In wireless networks it is possible to establish a link between a pair of neighboring nodes, provided that each has a transceiver available for this purpose. Thus, unlike the case of wired networks, the set of network links and their capacities is not determined a priori, but can adapt to changing traffic demands. This is a crucial difference that eventually forms the basis of what we propose in this report. We assume that each node has several transceivers, thereby permitting flexibility in link establishment, while, unfortunately, also complicating performance evaluation. In this report we do not address the protocol issues that are associated with call setup or with link activation, but instead focus on the development of a mathematical system model that will serve as the basis for data-packet delay evaluation for a given protocol of link activation.

We consider multihop wireless networks in which circuit switching is used for voice traffic and packet switching is used for data traffic. After describing the voice and data traffic models, we address in more detail the special issues associated with performance evaluation in wireless networks.

2.1 Voice traffic

We consider multihop wireless networks in which voice-call service is provided by the establishment of a circuit over a predetermined path between the originator of a call and its destination node. This requires the commitment of a transceiver at every node along the path for the duration of the call. Let T_i denote the number of transceivers at node i. We refer to calls that use circuit j $(1 \le j \le J)$ as calls of type j; they arrive to circuit j according to a Poisson process with rate λ_j^V, and their lengths are exponentially distributed with parameter μ_j^V. The offered load to circuit j is then $\rho_j^V = \lambda_j^V / \mu_j^V$.[1] The state of the system is described by the vector $\boldsymbol{x} = \{x_1, x_2, \ldots, x_J\}$, where x_j is the number of calls currently active on circuit j. This is a commonly used circuit-switched model for voice networks (see [8, 9, 15, 16]).

A central controller makes the decisions on whether or not to accept calls based on perfect knowledge of the number of calls of each type that are currently in progress (i.e., the system state $\boldsymbol{x}$), and hence the set of resources that are available for new calls. The transceivers needed to establish a circuit are acquired simultaneously when the call arrives, and are released simultaneously when the call is completed. Calls are blocked when one or more nodes along the path do not have a transceiver available, or when a decision is made not to accept a call despite the availability of transceivers.

[1] It is reasonable to have values of ρ_j^V greater than 1 because there are generally several transceivers at each node.

If we restrict the voice admission-control policy to the class of policies known as *coordinate convex* [17], the Markovian nature of the voice traffic, coupled with its independence from data traffic, results in a *product-form* solution for the equilibrium voice state of the system [18, 19]. Coordinate-convex policies are a reasonable, easily implemented class of policies [8, 18]. Each such policy is specified in terms of the set of admissible states Ω; resources (transceivers) freed by completed calls are immediately available for use by new calls, ongoing calls do not get rerouted, and a new call is admitted if and only if the state to be entered is in the admissible region (otherwise, it is blocked and lost from the system). Although the optimal admission-control policy is not generally coordinate convex, it was demonstrated in [18, 19] for several example problems that the best coordinate-convex policy provides nearly optimal performance. The simplest example of a coordinate-convex policy is the example of an "uncontrolled" system, i.e., a network in which calls are admitted as long as resources are available to support them.

The equilibrium state probability $\pi(\cdot)$ is given by

$$\pi(\boldsymbol{x}) \;=\; \pi(0)\prod_{j=1}^{J}\frac{\rho_j^{x_j}}{x_j!}, \qquad \boldsymbol{x}\in\Omega$$

where $\pi(0)=\left(\sum_{\boldsymbol{x}\in\Omega}\prod_{j=1}^{J}\rho_j^{x_j}/x_j!\right)^{-1}$ is the normalization constant associated with the admissible region Ω. For any coordinate-convex admission-control policy, it is straightforward (though possibly time consuming) to evaluate $\pi(0)$, which in turn permits the evaluation of performance measures such as throughput and blocking probability.

2.2 Data traffic

Data traffic consists of single fixed-length packets with Poisson arrival statistics (at rate λ_k^d at queue k). At each node a separate queue is formed for packets intended for each of its neighbors. We consider only the single-hop data case, i.e., relaying is not addressed.[2] As in the case of voice traffic, a transceiver is needed at both the source and destination nodes throughout the duration of a data-packet transmittal. Voice traffic has priority over data traffic in the sense that the decision to accept a voice call depends only on the number of voice calls (of each type)

[2] The restriction to single-hop data traffic is not really a major limitation. For a multihop case, we would derive the data-packet arrival rate at each queue along a given path from the end-to-end arrival rates (i.e., the arrival rates of each data-stream type) and the parameters of a given routing algorithm. Then delay over multihop data paths, for the case of fixed routing, can be obtained by concatenating the delays over the links along the paths. Of course, the implicit assumption of independence along the concatenated links makes this computation an approximate one.

currently in progress in the system, and is independent of the size of data queues. At any time instant, the transceivers not being used to support voice traffic are available to support data traffic. Thus, the network resources available for data depends on the voice-call statistics, and on the voice-call admission-control policy.

2.3 Data-packet delay evaluation considerations in wireless networks

The evaluation of data-packet delay is complicated by the flexible manner in which each node's transceivers are allocated to voice and data traffic in wireless networks. Let us consider two neighboring nodes a and b. Typically, at an arbitrary time instant, some of each node's transceivers will be supporting voice calls. The remainder are available for data as in the movable-boundary scheme that is commonly used for integrated multiplexing and multiple access [3, 20]. As a first step toward developing a delay model, we assume that any transceivers not occupied by voice traffic at nodes a and b are available to support data communication between these two nodes. We thus define the *residual capacity* C_{ab} available for data on "link (a,b)" as

$$C_{ab} = \min_{i=a,b} \{\text{number of transceivers not occupied by voice at node } i\} .$$

In general, not all of C_{ab} will be available to support link (a,b), since some transceivers at nodes a or b may form links for data traffic with the node's other neighbors. Typically, a link activation [21-23], frequency management, or other adaptive demand-based protocol is used to partition the use of the node's transceivers among all the links that the node may form with its neighbors; thus, in a data-only system, our link (a,b) would receive a certain fraction (say θ) of the maximum data rate that can be supported between nodes a and b. In the integrated network, one possible approach would be for link (a,b) to receive the same fraction θ of C_{ab}. To avoid the need to specify a particular link-activation schedule, in this report we simply assume that the entire residual capacity of the nodes (measured in terms of transceivers) is available to each link of interest. It would be straightforward to apply our approach to a system operating under any particular link-activation scheme[3] by using this reinterpretation of residual capacity.

In Section 3.2 we discuss a crude model for delay based on the average residual capacity, and in Sections 4 and 5 we develop more-accurate models that take into account the time varying nature of the residual capacity.

[3] This is true as long as the link-activation scheme does not depend on the data-packet queue sizes in the network.

3. On the development of delay models

Although performance measures such as voice blocking probability and throughput can be exactly calculated by means of the product-form solution, there are no accurate models available for calculating data-packet delay. Accurate models require a characterization of the time-varying behavior of the voice state. Since the voice state of the system is Markovian, it is possible, in principle, to trace the distribution of the voice state as a function of time, and thereby to obtain its time-varying behavior. The expected data-packet queue size (for the present case of single-hop data traffic with Poisson arrivals) can then be studied by viewing each link in the system as a variable rate queueing system with service rate dependent on the voice process. Once the expected data-packet queue size is obtained, expected delay can be determined by means of Little's formula:

$$E\{\text{time data packet spends in system}\} = \frac{E\{\text{data packet queue size}\}}{\lambda^d}$$

where λ^d is the data-packet arrival rate. Unfortunately, however, such a Markov model for voice traffic would be practical only for very small systems (i.e., small number of call types and small maximum number of calls of each type). In addition, the study of the behavior of a queueing system with variable rate of service is highly nontrivial. Thus approximate models are needed.

In Section 3.1 we discuss an exact Markov model, which is impractical because of its dimensionality, but which serves as the basis for the approximate "mini-product-form" model that is the focus of this report. We then discuss a very crude M/D/1 approximation in Section 3.2, which provides remarkably accurate ordinal rankings of control policies, but poor estimates of actual delay.

3.1 An exact (but impractical) Markov model

In this section we outline the development of an exact Markov model for the numerical performance evaluation of integrated networks of the type described in Section 2. We make use of the assumption made in Section 2.3 that the entire residual capacity of any pair of nodes is available to the link that connects them. Although the data-packet delay over any link is independent of the data arrival process at any other link (because we are considering single-hop data traffic), the delay actually depends strongly on the time-varying residual capacity available for data, which is determined by the voice process throughout the network. Therefore, a characterization of the system state over the link of interest must incorporate the following quantities:

$V(k) = \{V_1(k), V_2(k), ..., V_J(k)\}$, where $V_j(k)$ is the number of calls of type j that are in the system at time slot k. $V(k)$ determines the total number of calls of all types that are present at each node at time slot k, and thus determines the residual capacity of any given link at time slot k.

$R(k) =$ the number of data packets in queue at the link of interest at time slot k.[4]

The duration of a time slot is the time required for a transceiver to transmit one fixed-length data packet (all transceivers transmit at the same rate).

The analysis of our integrated network is similar to that developed for the movable-boundary Wireless Integrated Multiple Access (WIMA) protocol [3], and can proceed, in principle, as follows. The behavior of the link of interest can be characterized by the Markov chain $\{R(k), V(k)\}$, which has transition probabilities $\Pr\{R(k+1), V(k+1) \mid R(k), V(k)\}$. The steady-state distribution of the system state can then be obtained by repeated iteration of the transition probability matrix until equilibrium is reached. In the numerical evaluation it is necessary to truncate the data-packet queue size (which is inherently unbounded) at a finite value, which should be chosen sufficiently large so that truncation effects on the resulting equilibrium probability distribution are negligible; no such truncation is needed for the voice-call process, which is finite because it is limited by the number of transceivers at the nodes or possibly by an admission-control policy. The expected data-packet delay at this link can be evaluated by determining the expected data-packet queue size in equilibrium. The numerical evaluation outlined here is repeated at each link for which expected delay values are desired. It is a complex and cumbersome evaluation, and is presented here just to sketch out the process by which an exact evaluation could be carried out if the size of the network were small.

To reduce the dimensionality of the transition-probability matrix (which may be prohibitively large), we observe as in [3] that the data traffic depends on the voice state, whereas voice is independent of data. Therefore, we have

$$\Pr\{R(k+1), V(k+1) \mid R(k), V(k)\} = \Pr\{R(k+1) \mid R(k), V(k)\} \Pr\{V(k+1) \mid V(k)\} .$$

The transition from $R(k)$ to $R(k+1)$ depends on $V(k)$ (because $V(k)$ determines the residual capacity), but not on $V(k+1)$. The transition from $V(k)$ to $V(k+1)$ does not depend on $R(k)$ or $R(k+1)$.

---

[4] A single conceptual queue at link (a,b) is used to represent all data traffic in either direction between nodes a and b. This entails no loss of generality since we assume that a full transceiver is required at both the transmitting and receiving nodes of a data link.

We observe that the transitions from slot k to slot k+1 can be modeled as a two-step process. Data transitions are considered first. Given $\Pr\{R(k), V(k)\}$, we first determine $\Pr\{R(k+1), V(k)\}$. The evolution of the data-packet process between slots k and k+1 can be described as follows:

$$R(k+1) = [R(k) - c(k)]^+ + A(k),$$

where $c(k)$ is the residual capacity of the link of interest in slot k (determined directly from $V(k)$), $A(k)$ is the number of data-packet arrivals in slot k (these packets cannot be transmitted before slot k+1), and $[x]^+ = \max\{x, 0\}$. We assume that data-packet arrivals form a Bernoulli process with at most one packet per slot[5] with an arrival probability of λ_d; thus

$$\Pr\{A(k)=1\} = \lambda_d \quad \text{and} \quad \Pr\{A(k)=0\} = 1-\lambda_d.$$

Therefore, the data-packet transition probabilities (conditioned on the residual capacity) can be directly written as:

$$\Pr\left\{R(k+1) = [R(k) - c(k)]^+ \mid c(k)\right\} = 1-\lambda_d,$$

and

$$\Pr\left\{R(k+1) = [R(k) - c(k)]^+ + 1 \mid c(k)\right\} = \lambda_d,$$

from which we obtain $\Pr\{R(k+1) = n \mid R(k) = i, V(k) = v\}$. The joint distribution $\Pr\{R(k+1), V(k)\}$ is then determined as follows:

$$\Pr\{R(k+1) = n, V(k) = v\} = \sum_{i=n-1}^{n+T} \Pr\{R(k+1) = n \mid R(k) = i, V(k) = v\} \Pr\{R(k) = i, V(k) = v\},$$

where T is the maximum residual-capacity value that the link of interest can have (namely the maximum of the number of transceivers at the two nodes forming this link).

Next, the voice transitions are considered. Given $\Pr\{R(k+1), V(k)\}$ we determine $\Pr\{R(k+1), V(k+1)\}$. Since the voice transitions are independent of the data traffic, the same transition probability matrix is used for all values of $R(k+1)$. Thus, the following is evaluated:

---

[5] Other distributions for A_k could certainly be considered as well, e.g., a Poisson arrival distribution (truncated at some finite value to facilitate computation).

$$\Pr\{R(k+1)=j, V(k+1)=w\} \;=\; \sum_{v=\bar{0}}^{V_{max}} \Pr\{V(k+1)=w \mid V(k)=v\}\;\Pr\{R(k+1)=j, V(k)=v\},$$

where the summation is actually a J-dimensional summation over all possible voice states, which range from the empty voice state ($v = \bar{0}$) to voice states at the boundary of the admissible region (denoted loosely as V_{max}). The equilibrium distribution of the system state is determined by repeating this two-step iteration until convergence is achieved, i.e., until there is negligible change in the distribution.[6]

Although this model is exact (except for the assumption that the entire residual capacity is available for the link of interest), its use is impractical because of the high dimensionality, which results from the need to model the J-dimensional voice state.

3.1.1 A quasi-static approximation

To reduce the complexity of this evaluation, it is possible to consider an approximate quasi-static model that exploits the fact that the voice-call process changes much more slowly than the data-packet process; e.g., in practical systems the voice-call duration may typically be several orders of magnitude greater than the data-packet time slot. First, we assume that $V(k)$ is constant for all time. For each possible value of $V(k)$ we can then determine the equilibrium distribution of $R(k)$ by numerical iteration of the transition probability matrix corresponding to the value of residual capacity associated with $V(k)$. From this equilibrium distribution we can determine the expected queue size, and hence expected delay, associated with each voice-call state $V(k)$. Finally, we can average the distribution of $R(k)$ over all possible values of $V(k)$, whose distribution can be determined exactly by means of the product-form distribution without requiring the iterative approach discussed here. However, unless the instantaneous residual capacity in all voice states is greater than the offered data load, the resulting expected delay will be infinite; e.g., if any state with nonzero probability provides a residual capacity of zero on the link of interest, any finite data rate on that link will produce infinite expected delay.[7] Our mini-product-form model, which incorporates dynamic system behavior, only requires that the offered data traffic be less than the expected residual capacity, and does not require that any transceivers be reserved for the exclusive use of data.

---

[6] In our studies of the mini-product-form model, we have stopped the iteration process when the change in expected queue size between iterations n and $n + 10$ was less than 10^{-3} percent.

[7] This approach was used to evaluate data-packet delay for the WIMA protocol in [5] for "movable-boundary" examples in which some resources were reserved for exclusive use by data traffic.

3.2 A crude data-traffic model

For each voice-call admission-control policy, we can compute the expected residual capacity at every link by means of the product-form solution. If we make the obviously inaccurate approximation that the data packets at link (a,b) are served at the constant rate of the average residual capacity of the link, C_{ab}, the resulting system is an M/D/1 queueing system in which the delay (not including service time) is

$$D = \frac{\rho}{2\mu(1-\rho)}.$$

Here, $\rho = \lambda/\mu$, where λ is the Poisson data-packet arrival rate and μ = 1/(time-slot duration) = C_{ab}.

This delay model is deficient in several ways. Most significant is the implicit assumption that μ, the data-packet service rate at link (a,b), is constant at the expected residual-capacity value C_{ab} (the fact that C_{ab} is not, in general, integer-valued is a minor point here), whereas the number of transceivers available in the real system varies, based on the voice state. Thus data-packet queue sizes may increase rapidly when the voice-call occupancy is high and decrease rapidly when it is low. This model also ignores the duration of voice calls[8] and hence the statistics of the amount of time spent in each voice-occupancy system state. Despite the poor predictive quality of this delay estimate (see Section 6.2), we have found that this model performs surprisingly well as a method for *ordinal optimization*, i.e., for the ranking of the voice-call admission-control policies with respect to the achieved average data delay computed under this model; in fact, the resultant ranking was in almost perfect agreement with the ranking found by extensive accurate simulations (which do not make the M/D/1 approximation) [12]. However, our objective in the present report is to improve the quality of the actual delay estimate.

The high complexity of the exact Markovian model and the inaccuracy of the approximate models discussed in this section point to the need for the development of accurate and practical models for delay performance. Such models must address the time-varying nature of the residual capacity, without being too complex for feasible computations. In Section 4 we discuss the development of such a model.

[8] It uses only the ratio of the voice arrival rate to its service rate, and is independent of the actual arrival rate and service rate values.

4. Improved approximate models

The use of the exact Markovian model discussed in Section 3.1 is impractical because of its high dimensionality, which results from the need to track the voice-call state throughout the network while evaluating data-packet delay at a single link of interest. The system state at the link of interest is given by $\{R(k), \boldsymbol{V}(k)\}$, in which the voice state $\boldsymbol{V}(k)$ (for a system with J voice circuits) is a J-dimensional first-order Markov chain. In this section we discuss an approximate model in which a three-dimensional first-order Markov chain, which we refer to as the "mini-product-form" model, is used to describe the voice-call state at any link in the network, resulting in a practical methodology for performance evaluation. Our model is similar in principle to the well-known reduced-load approximate models [15, 16, 24-28], which we discuss in Section 4.2. However, we demonstrate in Section 6 that our mini-product-form model is generally more accurate than a time-varying model based on the reduced-load approximation.

4.1 The mini-product-form model

Our objective is to find a description of the voice-call process on the link of interest that is simpler than that provided by the J-dimensional $\boldsymbol{V}(k)$, but yet maintains a high degree of accuracy. In doing so, we have developed an approximate model in which we consider each link separately, in terms of both voice and data traffic. The objective is to characterize the time-varying capacity available for data traffic over each such link, while taking into account the effects of voice traffic throughout the entire network. In particular, the voice traffic carried by the two nodes that form the link of interest (nodes a and b) must be modeled. The traffic supported by these nodes is the superposition of the traffic carried by all circuits passing through them. The throughput actually delivered by each of these circuits is generally less than the offered traffic level because of blocking, both at the nodes of interest and at other nodes. In recent years a number of "reduced-load" approximations have been developed to account for blocking at other links or nodes in the network. These are based on the assumption that blocking occurs independently at each node, an assumption that is asymptotically accurate for large networks with a large number of transceivers per node and a large number of call types, where the ratio of traffic to capacity is kept fixed. We discuss the application of reduced-load approximations to our problem in Section 4.2. But first we develop our mini-product-form model, which is the main contribution of this report.

Consider a link defined by node pair ab, as shown in Fig. 1, where nodes a and b have T_a and T_b transceivers, respectively. The flexibility of wireless communications, in the sense that a transceiver can form a link with any of its neighbors that also have an available transceiver,

eliminates the need to assign a priori any particular neighbor or call type to a transceiver. We define three voice-traffic flows that involve link (a,b) as follows:

f_{ab} = flow passing through both nodes a and b (i.e., the sum of the expected number of calls on all circuits traversing link (a,b), whether terminating or starting at either node a or b or whether continuing through either node in the in-bound or out-bound direction),

f_a = flow passing through node a but not node b (i.e., the sum of the expected number of calls on all circuits traversing node a, whether terminating, starting, or continuing through that node but not involving node b),

f_b = flow passing through node b but not node a.

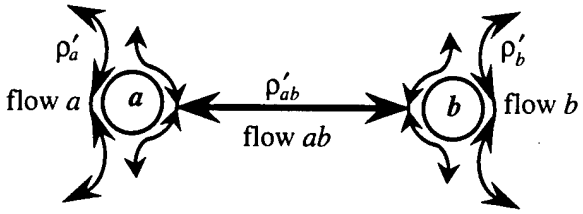


Fig. 1 — The "mini-product-form" network.

These flows would be artificial and useless in a wired network. They are very real and useful, however, in the context of a wireless network because they provide a way to distinguish between the time-average number of transceivers at each node dedicated to service link (a,b) and those assigned to service calls that simply go through nodes a or b without traversing link (a,b). Note that the total flow through node a is then $f_{ab} + f_a$, and the expected residual capacity at node a is $T_a - (f_{ab} + f_a)$. Thus we can consider a three-dimensional voice state description of these two nodes at any time instant given by $\boldsymbol{x} = \{x_a, x_b, x_{ab}\}$, where

x_{ab} = number of voice calls currently passing through both nodes a and b,

x_a = number of voice calls currently passing through node a but not node b,

x_b = number of voice calls currently passing through node b but not node a.

Note that the total number of calls currently passing through node a is $x_{ab} + x_a$, and the instantaneous residual capacity of node a is $T_a - (x_{ab} + x_a)$.

Unfortunately, the random process $\boldsymbol{x}$ is not Markovian. The arrival processes associated with the three flows are not Poisson because of blocking in the network. However, we make the approximation that these flows are characterized by Poisson arrival processes, at rates chosen to match the traffic levels in the overall network. In doing so we are, in effect, approximating the

distribution of the interarrival times for the voice traffic on a given link by its maximum entropy estimate [29]. As in the overall network, calls may be blocked because of a lack of resources or because of the control policy. This two-node three-flow system (under the Poisson arrival assumption) is also a product-form system, and we refer to it as our "mini-product-form" model. By focusing on the immediate neighborhood of two nodes, we attempt to capture some of the interactions that complicate the network behavior.

We conjecture that the behavior of the mini-product-form network associated with nodes a and b closely approximates the behavior of nodes a and b in the overall network, provided that appropriate system parameters are chosen. In particular, our objective is to find the offered traffic levels $\boldsymbol{\rho}' = \{\rho'_a, \rho'_b, \rho'_{ab}\}$ for the mini-product-form network that produce the same expected traffic flows ($\boldsymbol{f} = \{f_a, f_b, f_{ab}\}$) as those of the product-form solution for the network as a whole operating under the offered load $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_J\}$. Here we discuss the simple iterative procedure we have developed to determine the offered load $\boldsymbol{\rho}'$, and in Section 6 we demonstrate that the mini-product-form does, indeed, provide a good characterization of system behavior.

Flow-Matching Algorithm:

Guess initial values for $\boldsymbol{\rho}'$.

Repeat until ($error < \varepsilon$):

{Calculate via product form the flows f'_a, f'_b, and f'_{ab} in the mini-product-form network with offered load $\boldsymbol{\rho}'$;

$error = (f_a - f'_a + f_b - f'_b + f_{ab} - f'_{ab})/3$;

adjust $\boldsymbol{\rho}'$ proportional to individual errors: $f_a - f'_a$, $f_b - f'_b$, and $f_{ab} - f'_{ab}$, i.e., $\rho'_a := \rho'_a + \delta(f_a - f'_a)$, $\rho'_b := \rho'_b + \delta(f_b - f'_b)$, $\rho'_{ab} := \rho'_{ab} + \delta(f_{ab} - f'_{ab})$. }

The primed flows f'_a, f'_b, and f'_{ab} denote the flows in the mini-product-form network operating under the offered load ρ'_a, ρ'_b, and ρ'_{ab}. We have found that reasonable initial values for ρ'_a, ρ'_b, and ρ'_{ab} are the desired flows f_a, f_b, and f_{ab}. In our experiments, we have used a proportionality coefficient of $\delta = 1.0$ in adjusting $\boldsymbol{\rho}'$, and have typically used an error limit value of $\varepsilon = 10^{-6}$.[9]

This algorithm converges rapidly to provide accurate values for $\boldsymbol{\rho}'$. For example, Figure 2 shows the number of iterations that were required for our iterative procedure to converge; the example considered is that of link (5,13) in the network of Fig. 3, which will be discussed in Section 6. The figure shows that the number of iterations required to converge to within ε of the

[9]Use of $\varepsilon = 10^{-5}$ results in values of $\boldsymbol{\rho}'$ that differ from those for $\varepsilon = 10^{-6}$ only in the third or fourth decimal place.

desired flows increases with offered load, but is generally not prohibitive. Although the number of iterations depends on the offered load, the time required to perform a single iteration of this procedure, i.e., to find the product-form solution for a single parameter set and adjust ρ', depends primarily on the number of transceivers, which in turn determine the number of states that must be evaluated. For example, with four transceivers per node there are 55 states, and a single iteration of this procedure take about 5.5 ms; with eight transceivers per node there are 285 states, and an iteration takes about 14.4 ms; and with 16 transceivers per node each iteration requires nearly 81 ms to evaluate the 1785 states associated with the mini-product-form model. The rate of growth in the time required to perform an iteration reflects the exponential growth of the state space with increasing numbers of transceivers per node. In Section 5, we show that the number of states in the mini-product-form model is proportional to the cube of the number of transceivers at a node (assuming both nodes have an equal number of transceivers). Hence, this approach is feasible for the moderately sized examples used in this report. However, for larger problems an approximation, such as the reduced-load approximation (see Section 4.2), will have to be used.

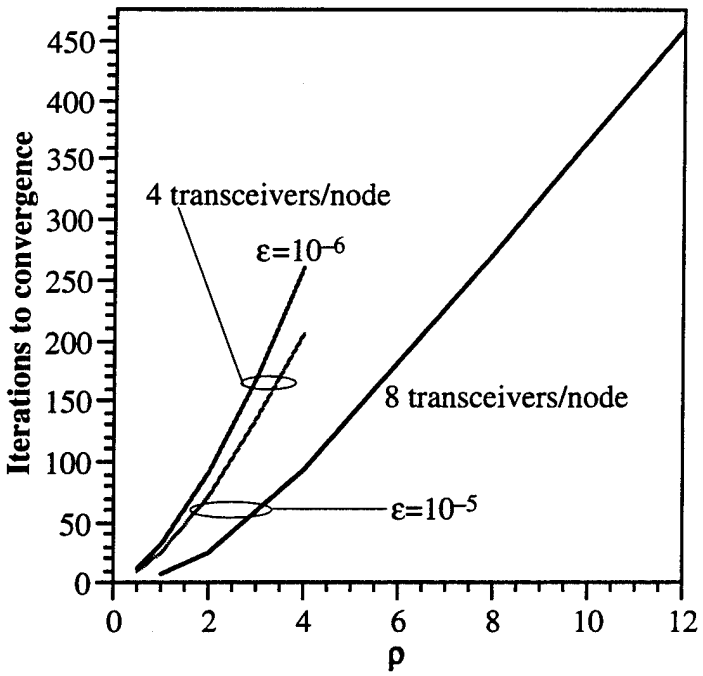


Fig. 2 — Iterative procedure convergence vs offered load ρ .

Thus although the product-form solution is normally used to characterize equilibrium behavior, we use it as the basis for our dynamic time-varying model. By matching the traffic flows, we attempt to capture some of the effects of blocking at remote nodes. Once the values of ρ'_a, ρ'_b, and ρ'_{ab} are determined, the time-varying behavior of the voice-call link occupancy can

be determined by numerical solution of the resulting three-dimensional Markov chain for the mini-product-form network by following the procedure outlined in Section 3.1 for the overall network. Doing so is considerably easier than solving the J-dimensional Markov chain for the overall network, particularly when the number of call types is large.

A potential limitation of the mini-product-form model is that its use requires the evaluation of the normalization constant associated with the product-form solution, which is computationally intensive for large networks. However, it is anticipated that an acceptable estimate of the voice-call state, from which the parameters for our mini-product-form model can be obtained, can be evaluated by using a reduced-load approximation to obtain the equilibrium state description. In Section 4.2 we discuss the application of the reduced-load approximation in a somewhat unconventional manner to obtain an alternative approximate time-varying model for our network, but show that its performance is inferior to that of our model. In Section 6 we discuss performance results based on the mini-product-form model, and demonstrate its accuracy over a wide range of situations.

4.1.1 Wired vs wireless networks

In developing the mini-product-form model for wireless networks we have been able to exploit the flexibility of wireless communication, which permits a transceiver to form a link with any of its neighboring nodes. This flexibility has permitted the characterization of the voice traffic over the link of interest in terms of a three-flow process. In contrast, the development of accurate models for wired networks is more difficult because of the fixed nature of the network link capacities, as we now explain.

In an attempt to obtain an accurate model for voice-call traffic over individual links in product-form wired networks, we investigated the use of an M/M/m/m queueing model,[10] which is commonly used to model voice systems. Here, the parameter m is simply the number of channels supported by the link of interest. The throughput of an M/M/m/m queue is

$$\text{Throughput}_{\text{M/M/}m\text{/}m} = \rho\left(1-\frac{\rho^m/m!}{\sum_{n=0}^{m}\rho^n/n!}\right).$$

In a manner analogous to that used in our wireless network model, we modeled an arbitrary link

[10] An M/M/m/m queue is one in which both the arrival and service times are exponential, the number of servers is m, and the total number of customers permitted in the system is also m. Thus this is a loss system in which queueing is not permitted; customers are rejected if they arrive when all servers are busy.

as an M/M/*m*/*m* queue, and we determine the offered load ρ' (a single parameter in this case) so as to reflect the interactions of the voice traffic on the link of interest with the rest of the network. Specifically, ρ' was chosen such that the expected voice-call throughput supported by the link (in isolation, using the M/M/*m*/*m* model) was equal to that which is generated by the overall product-form solution, which characterizes global network performance. Although this model performed well in some cases, errors were typically much larger than those obtained by using the mini-product-form solution for wireless networks.

It is not surprising that the mini-product-form solution for wireless networks performs better than the M/M/m/m model for wired networks, because the mini-product-form solution matches not only the flow across the link of interest, but also the adjacent flows. To match flows in a wired network would require a separate flow for each link adjacent to the link of interest. In some cases this might require more than J (the number of voice circuits) flows, thus resulting in a model with complexity greater than that of the Markovian model for the overall network (which has one flow per circuit type). The feature of the wireless network that permits a reasonably accurate system description using three flows is the fact that the transceivers at a node are not a priori assigned to specific links, but may form links with any of their neighboring nodes. Hence, a single flow can incorporate the effects of all flows adjacent to the link of interest.

4.2 A reduced-load model

The "reduced-load" model has been proposed for the evaluation of traffic statistics at individual links or nodes in multihop circuit-switched product-form networks [15, 16, 24-28]. If there were no blocking in the network, the average load offered to a node would be simply the sum of the offered load on each circuit that uses the node. That is, $\rho'_a = \sum_{i \in \beta_a} \rho_i$, where β_a = {circuits that use node a}, $\rho_j = \lambda_j/\mu_j$ is the offered load on circuit j and ρ'_a is the average load offered to node a. However, as a result of resource limitations (e.g., a finite number of transceivers) or admission-control policies, the load offered to a node can be "thinned" or reduced by blocking at other nodes. The reduced-load approximation estimates the thinned average offered load at a node in such a network by making the assumption that blocking occurs independently at each node so that the estimated offered load at node a is

$$\rho'_a = \sum_{j \in \beta_a} \rho_j \prod_{l \in j-a} \left(1 - P_b^n(l)\right), \qquad (1)$$

where the product is taken over all nodes l that are elements of circuit j except node a, and $P_b^n(l)$ is the probability of blocking at node l. Thus, the product term represents the probability of call

acceptance. The remaining difficulty is calculating $P_b^n(l)$. Typically a method of repeated substitution is used to solve the Erlang loss formula

$$P_b^n(l) = \frac{(\rho_l')^{T_l} / T_l!}{\sum_{n=0}^{T_l} (\rho_l')^n / n!},$$

where T_l is the number of transceivers at node l.

For comparison purposes, we have used the reduced-load approximation to estimate the load ρ_a' that should be offered to a node in isolation to yield the same statistics as would have been seen at node a operating as an element of the whole network. For relatively small networks, we can calculate the product-form solution for the network to obtain exact values of circuit blocking probability (denoted $P_b^c(j)$ for circuit j) and node blocking probability (denoted $P_b^n(a)$ for node a), thus eliminating the need to use repeated substitution.[11] Based on the independence assumption used in this method, we observe that

$$\prod_{b \in j} \left(1 - P_b^n(b)\right) = 1 - P_b^c(j) \quad \Rightarrow \quad \prod_{b \in j-a} \left(1 - P_b^n(b)\right) = \frac{1 - P_b^c(j)}{1 - P_b^n(a)}.$$

We then use Eq. (1) to obtain

$$\rho_a' = \frac{1}{\left(1 - P_b^n(a)\right)} \sum_{j \in \beta_a} \rho_j \left(1 - P_b^c(j)\right).$$

Following that, we can compute the state occupancy distribution $\pi(y_a)$ for node a in isolation since it has the product-form given by

$$\pi(y_a) = \frac{(\rho_a')^{y_a} / y_a!}{\sum_{n=0}^{T_a} (\rho_a')^n / n!}, \qquad y_a = 0,1,\ldots,T_a,$$

where y_a, the total number of calls currently passing through node a, is equal to $x_a + x_{ab}$ in the mini-product-form model, as defined earlier.[12]

---

[11] For examples too large to compute the normalization constant (in which case the reduced-load approximation generally performs well), the standard method of repeated substitution would be necessary. In fact, this method could be used to obtain the parameters for our mini-product-form model in large examples.

[12] Our use of the reduced-load model is unconventional in that we assume that the normalization constant is known. Usually the reduced-load model is used to approximate the normalization constant, as discussed in Section 4.1.

To obtain our reduced-load estimate of the statistics of link (a,b) in a wireless network, we make the further approximation that the state occupancy distributions $\pi(y_a)$ and $\pi(y_b)$ are independent, i.e., that $\pi(y_a, y_b) = \pi(y_a)\ \pi(y_b)$.[13] The residual capacity available for data on link (a,b) is $C_{ab} = \min\{(T_a - y_a), (T_b - y_b)\}$, and its distribution can be obtained from the approximate joint node occupancy distribution $\pi(y_a, y_b)$.

In summary, in our reduced-load approximation the voice traffic over link (a,b) is modeled as the superposition of two voice flows at nodes a and b, which are conditionally independent, given the offered load to the network. By contrast, in the mini-product-form model the correlation in the flows through nodes a and b is reflected by the flow f_{ab}. Taking this correlation into consideration results in a much better estimate of system performance, as shown in Section 6.

5. A further simplification of the model

Although the three-dimensional mini-product-form model for voice traffic results in a substantial reduction in model complexity as compared with the exact J-dimensional model, the complexity remains high. In this section we discuss the use of a one-dimensional Markov model for voice traffic over a link, whose use introduces only small errors as compared to the three-dimensional mini-product-form model.

Let us again consider the mini-product-form network shown in Fig. 1, and in particular the case in which $T_a = T_b = T$ (i.e., the number of transceivers at both nodes a and b is T). The number of voice states can be evaluated as follows. For a given value of x_{ab}, the number of possible values of x_a is $(T - x_{ab} + 1)$ because the number of transceivers occupied by voice at node a can range from 0 to T; the same is true for the number of possible values of x_b. Thus the number of possible $\{x_a, x_b\}$ pairs for the specified value of x_{ab} is $(T - x_{ab} + 1)^2$. Summing over all possible values of x_{ab} to find the total number of states of the form $\{x_a, x_b, x_{ab}\}$ yields:

$$\text{number of states in mini-product-form model} = \sum_{n=1}^{T+1} n^2$$

$$= \frac{(T+1)(T+2)(2T+3)}{6} = \frac{T^3}{3} + \frac{3T^2}{2} + \frac{13T}{6} + 1 \ .$$

Overall, the total number of states possible for this link is this number multiplied by $D_{max} + 1$,

[13] This independence approximation can result in errors in the joint distribution that are significantly larger than those based on the reduced-load approximation at the individual nodes.

where D_{max} is the value at which the data-packet queue size is truncated in the numerical evaluation. At high throughput levels we have used values of D_{max} as great as 1000 to maintain the accuracy of the calculations. This dimensionality limits the applicability of the mini-product-form solution to relatively small problems.

We make the observation that the data-packet queueing process on link (a,b) depends on the voice-call process only through the residual capacity C_{ab} available for data, which at any given time is

$$C_{ab} = \min\{[T_a - (x_a + x_{ab})], [T_b - (x_b + x_{ab})]\}, \quad (2)$$

where the time dependence of C_{ab}, x_a, x_b, and x_{ab} has been suppressed to simplify the notation. The advantage of describing the voice-call state in terms of the residual capacity is that the number of states is reduced to $T + 1$.[14] Unfortunately, C_{ab} is not Markovian; if it were, a complete description of the state of the link could then be given by the pair $\{R, C_{ab}\}$, i.e., a pair consisting of the number of data packets in the queue and the residual capacity. However, in our approximate model, we make the assumption that indeed C_{ab} is first-order Markovian. Under this assumption, the transition probability matrix for C_{ab} can be obtained numerically from the three-dimensional voice-state description $\boldsymbol{x} = \{x_a, x_b, x_{ab}\}$, which we approximated to be first-order Markovian in our derivation of the mini-product-form model (by assuming that the three flows are Poisson processes). The derivation proceeds as follows.

Each element of the transition probability matrix for C_{ab} can be written as

$$\Pr(c' = C' \mid c = C) = \frac{\Pr(c' = C', c = C)}{\Pr(c = C)},$$

where c represents C_{ab} at slot k and c' represents C_{ab} at slot k+1. The denominator, $\Pr(c = C)$, is obtained from the equilibrium distribution for $\boldsymbol{x}$ (i.e., $\Pr(\boldsymbol{x} = \boldsymbol{X})$) by summing the probabilities of each three-dimensional voice state that corresponds to the value $c = C$. The equilibrium distribution for $\boldsymbol{x}$ is obtained by raising the transition probability matrix for $\boldsymbol{x}$ to a sufficiently high power. The numerator is obtained from the joint equilibrium distribution of $\boldsymbol{x}$ at two consecutive time slots, where

$$\Pr(\boldsymbol{x}' = \boldsymbol{X}', \boldsymbol{x} = \boldsymbol{X}) = \Pr(\boldsymbol{x} = \boldsymbol{X}) \Pr(\boldsymbol{x}' = \boldsymbol{X}' \mid \boldsymbol{x} = \boldsymbol{X}).$$

[14] For example, for $T = 4$, the number of voice states is reduced from 55 to 5, and for $T = 8$ it is reduced from 285 to 9.

Numerical results, discussed in Section 6, show that this approximate model is slightly optimistic. At low throughput rates it produces expected queue sizes approximately 2% lower than that of the mini-product-form model. At high throughputs (high enough to produce expected queue sizes of 30 packets) the discrepancy is typically less than 10%. Furthermore, if the highly accurate results supplied by the three-dimensional model are needed, the speed of the calculation may be increased by using the one-dimensional model to obtain an initial condition for the much-slower three-dimensional model's calculation. Data-packet delay results based on this model are discussed in Section 6.2.

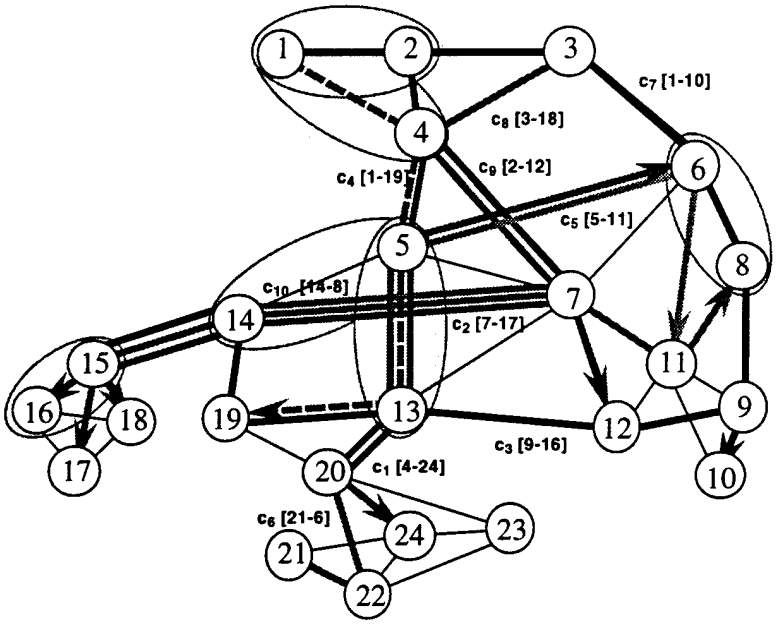


Fig. 3 — Example 24-node network with 10 source-destination pairs.

6. Performance results

We have applied the mini-product-form model to examples such as the 24-node network shown in Fig. 3. The shaded lines in the figure show the $J = 10$ voice circuits that we considered.[15] In this section we discuss the accuracy of the mini-product-form model in characterizing both the voice process and the data-packet delay. In all examples, we assume the use of an "uncontrolled" system, i.e., voice calls are admitted as long as transceivers are available to support them.

[15] Although the mini-product-form model has been developed to improve the accuracy of data-delay estimates, we only need to consider the voice portion of the integrated network to evaluate the accuracy of the mini-product-form model in predicting the time-varying behavior of the residual capacity.

6.1 Accuracy of the mini-product-form model characterization of the voice process

Here we evaluate the mini-product-form characterization of the voice process on the circled links in Fig. 3. Clearly, data-packet delay depends strongly on the residual capacity C_{ab}. For the case of $T_a = T_b = T$, we consider the complement of residual capacity, namely,

$$x = T - C_{ab}.$$

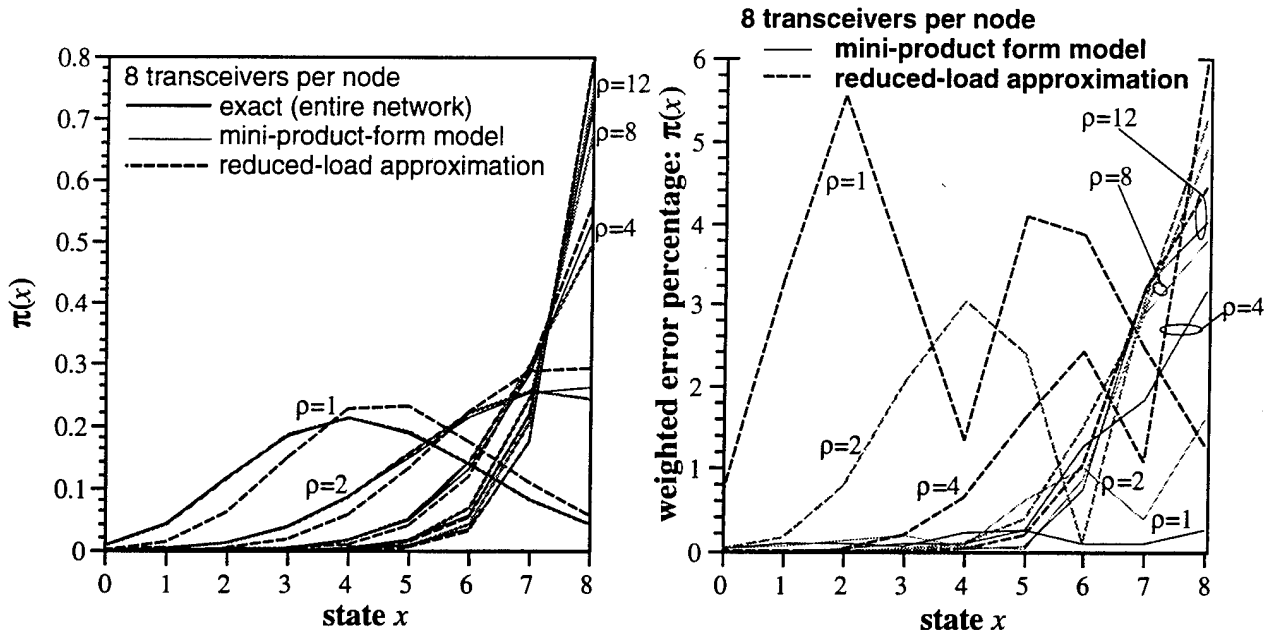
Based on Eq. (2), we can rewrite this expression as

$$x = x_{ab} + \max\{x_a, x_b\},$$

and we can interpret x as the effective number of voice calls that are currently being supported by link (a,b). In the remainder of this report we refer to x as the "state" since it is closely related to the voice state $\boldsymbol{x} = \{x_a, x_b, x_{ab}\}$ defined earlier. First, we use the product-form solution to obtain $\pi(x)$, the equilibrium distribution of x. We also evaluate the first and second moments of the "state-visit duration," $E\{t(x)\}$ and $E\{t^2(x)\}$ respectively, i.e., the time spent in each state x. These quantities are evaluated by means of simulations of the entire network, the mini-product-form network, and the reduced-load approximation discussed in Section 4.2. Although such timing estimates are not used explicitly in our delay evaluation, it appears clear that an accurate characterization of the time-varying behavior of the voice process is a necessary component of a good data-packet delay model.

In Fig. 4,[16] we compare the exact state-occupancy distribution of link (5,13) of our example 24-node network (computed by means of the product-form solution) to those computed by using the reduced-load approximation and the mini-product-form model estimate. Link (5,13) is centrally located and serves five voice circuits. We considered values of ρ ranging from 1 to 12 ($\rho_j^V = \rho$, $j = 1, \ldots, 10$) for the case of $T = 8$ transceivers per node. In Fig. 4(a), the plot of probability mass functions (pmf's) shows that the mini-product-form model yields a curve that is very close to the exact one; the two curves can be distinguished only for larger values of ρ at state $x = 8$ (at which point $C_{5,13} = 0$). The probability mass from the reduced-load approximation is generally shifted to the right of the exact result, thus predicting lower residual capacity.

[16] In our plots we use continuous curves to represent discrete quantities to allow easy comparison.



(a) Distribution

(b) Weighted error

Fig. 4 — State-occupancy distribution on centrally located link (5,13) for various ρ values.

To evaluate the accuracy of the pmf, we consider the error percentage and weighted error percentage as follows:

$$\pi(x) \text{ error percentage } = 100 \frac{|\tilde{\pi}(x) - \pi(x)|}{\pi(x)},$$

and

$$\pi(x) \text{ weighted error percentage } = 100\ \pi(x) \frac{|\tilde{\pi}(x) - \pi(x)|}{\pi(x)} = 100\ |\tilde{\pi}(x) - \pi(x)|,$$

where $\tilde{\pi}(x)$ is the estimated (either by the mini-product-form model or by the reduced load approximation) probability of being in state x. We limit our discussion of performance results to the case of the weighted measure because relatively high percentage errors in infrequently visited states are not expected to have a significant impact on overall system performance. Figure 4(b) shows that the mini-product-form model delivers a more accurate estimate than the reduced-load approximation. With the mini-product-form model, the weighted error percentage generally increases with offered load (ρ). Apparently this is so because contention for resources increases with increasing load, thus causing blocking that cannot be anticipated by the mini-product-form model. Although there is not perfect agreement, the mini-product-form model does provide a

good estimate of the state occupancy distribution on link (5,13); the weighted error percentage shows that the states where there is a large error are, in fact, infrequently visited.

Figure 5 shows the voice-call blocking probability on link (5,13) and the average voice-call blocking probability for the overall network (where the average is taken over each call type). For the present case of $T = 8$ transceivers at each node, the blocking probability is about 0.5 or greater for $\rho \geq 4$. Thus our results for $\rho \geq 4$ represent operation in a regime of high blocking probability.

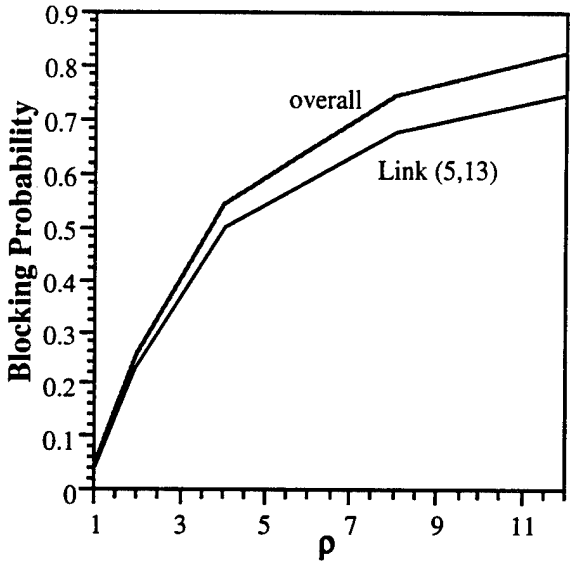
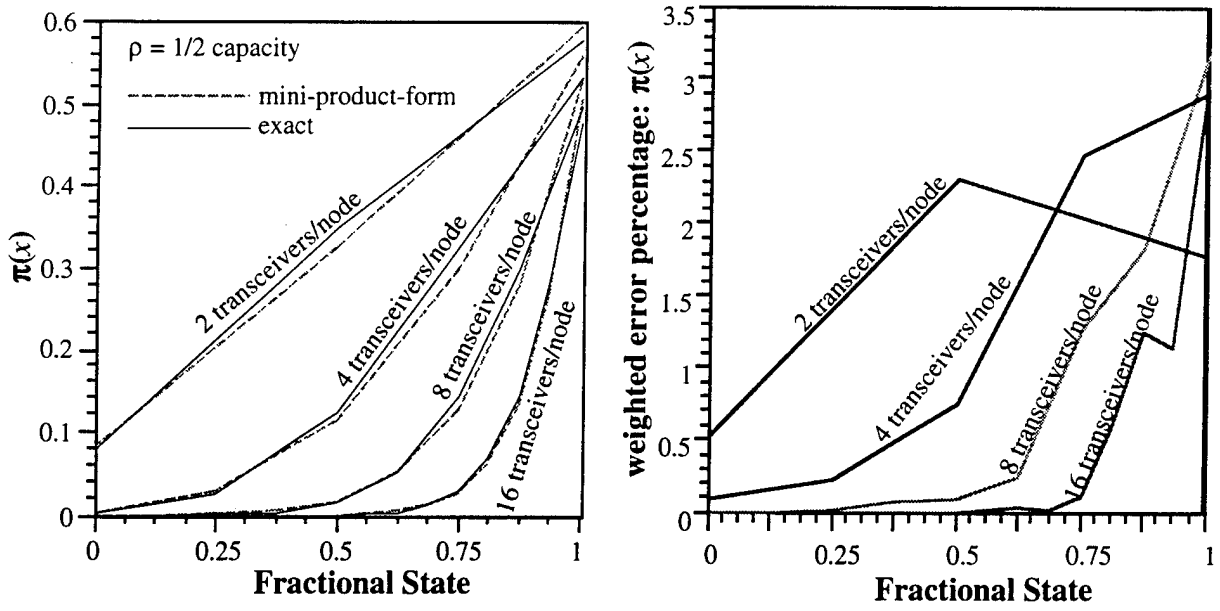


Fig. 5 — Blocking probability in the 24-node network and on link (5,13); 8 transceivers per node.

We continue our evaluation of the mini-product-form model of link (5,13) in Fig. 6, where we plot the state-occupancy distribution (Fig. 6(a)) and the weighted error percentage (Fig. 6(b)) obtained from four experiments. The experiments differed in the number of transceivers per node, as shown in the figure. The value of ρ was fixed at one half the number of transceivers per node. For example, in the experiment with four transceivers per node, we set $\rho_j^V = \rho = 2, j = 1, \ldots, 10$. In the figures, the x axis is the "fractional state," defined as the real state divided by the number of transceivers per node. The figure shows that the mini-product-form model tends to yield more-accurate results as the number of transceivers per node increases, although reasonably accurate results are achieved even for $T = 2$. This is consistent with reduced-load approximations that are asymptotically accurate in the limit of heavy traffic and large link capacities (for a fixed ratio of traffic/capacity) [15, 16, 24, 25].



(a) distribution (b) weighted error

Fig. 6 — State-occupancy distribution on centrally located link (5,13) for various numbers of transceivers per node.

In Fig. 7 we compare the first and second moments ($E\{t(x)\}$ and $E\{t^2(x)\}$ respectively) of the state-visit duration obtained from simulations of the entire network (with eight transceivers per node, and various ρ values), of the mini-product-form network, and of a "reduced-load network" (i.e., a two node, two flow network, where the loads offered to the two flows are at the rate given by the reduced-load approximation). Here again we see that the mini-product-form delivers more-accurate estimates than the reduced-load approximation.[17]

To evaluate the accuracy of the mini-product-form and reduced-load models we have evaluated mean error percentages and mean weighted error percentages for state-visit durations as follows:

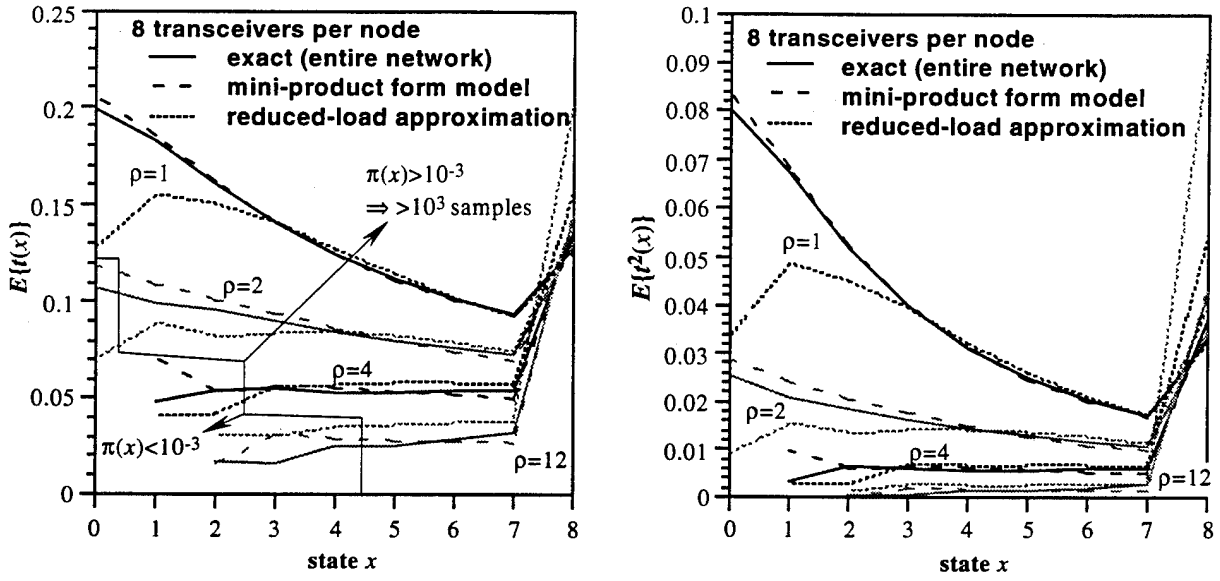
$$\text{mean } E\{t(x)\} \text{ error percentage} = \frac{1}{T+1}\sum_{x=0}^{T}\frac{100\ \left|\tilde{E}\{t(x)\} - E\{t(x)\}\right|}{E\{t(x)\}},$$

and

$$\text{mean } E\{t(x)\} \text{ weighted error percentage} = \frac{1}{T+1}\sum_{x=0}^{T}\frac{100\ \pi(x)\ \left|\tilde{E}\{t(x)\} - E\{t(x)\}\right|}{E\{t(x)\}},$$

[17] For values of ρ of 4 or greater, some of the low occupancy states are infrequently visited, thus resulting in potentially large percentage errors in the simulation results. To reduce the possibility of artificially inflating the probability of visiting these states, the initial state in the simulations was two calls per circuit (rather than starting from the empty state). The simulations were run until 10^6 voice-call arrivals had occurred. In Fig. 7(a) the plot is divided by a stairstep line on $\pi(x) = 10^{-3}$; estimates on the left side of this line are based on fewer than 10^3 samples, and are therefore subject to larger sampling error.

with a similar definition for $E\{t^2(x)\}$. The quantities with the tildes are the approximate values (based on the mini-product-form or reduced load approximations), and those without are the "exact" values as determined numerically ($\pi(x)$) or by simulation ($E\{t(x)\}$ and $E\{t^2(x)\}$).



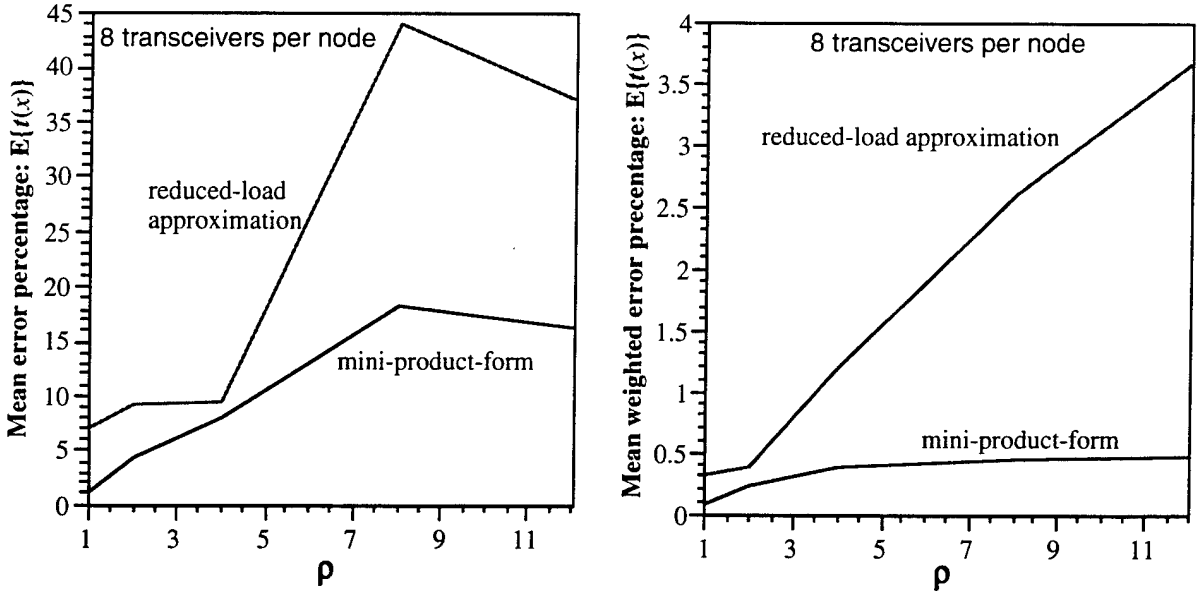
(a) first moment

(b) second moment

Fig. 7 — First and second moments of state-visit duration on centrally located link (5,13).

In Fig. 8, we summarize the accuracy of the mini-product-form model and the reduced-load-approximation estimates of the expected state-visit duration by plotting the mean error percentage (Fig. 8 (a)) and the mean weighted error percentage (Fig. 8 (b)) as ρ varies from 1 to 12. The figure shows that the mini-product-form model delivers significantly more accurate estimates than the reduced-load approximation. Note that, in terms of mean weighted error percentage, the mini-product-form model is relatively insensitive to the network load, whereas the accuracy of the reduced-load approximation decreases rapidly with increasing load.

We have examined in a similar manner what we feel is a representative set of links in the 24-node network of Fig. 3, and compiled the results for easy comparison in Figs. 9 and 10. The links we examined in addition to link (5,13) are link (15,16), a peripheral link with flow $f_b = 0$, three circuits deliver the traffic for flows f_{ab} and f_a; link (1,2), a peripheral link that serves three circuits; link (6,8), a peripheral link that serves four circuits; link (1,4), a link that serves five circuits; and link (5,14), a centrally located link that serves eight circuits, but has $f_{ab} = 0$. In these figures we show the mean weighted error percentage for the state-occupancy distribution (Fig. 9), and the first and second moments of the expected state-visit duration (Fig. 10).



(a) mean error (b) mean weighted error

Fig. 8 — Mean error percentage of state-visit duration on centrally located link (5,13).

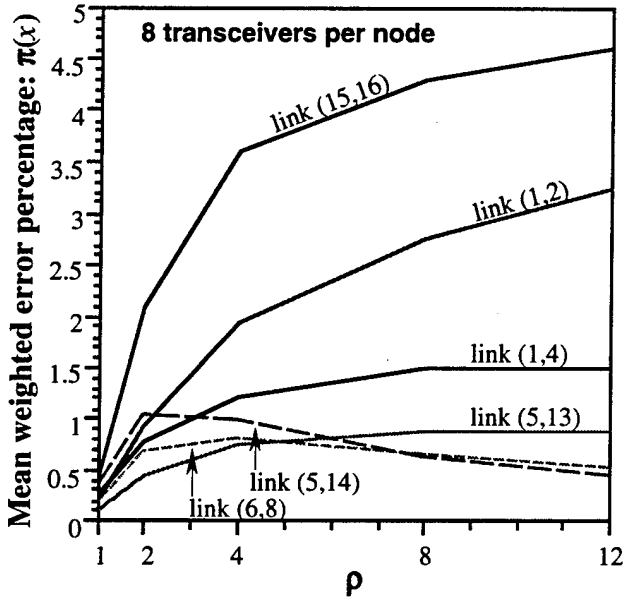
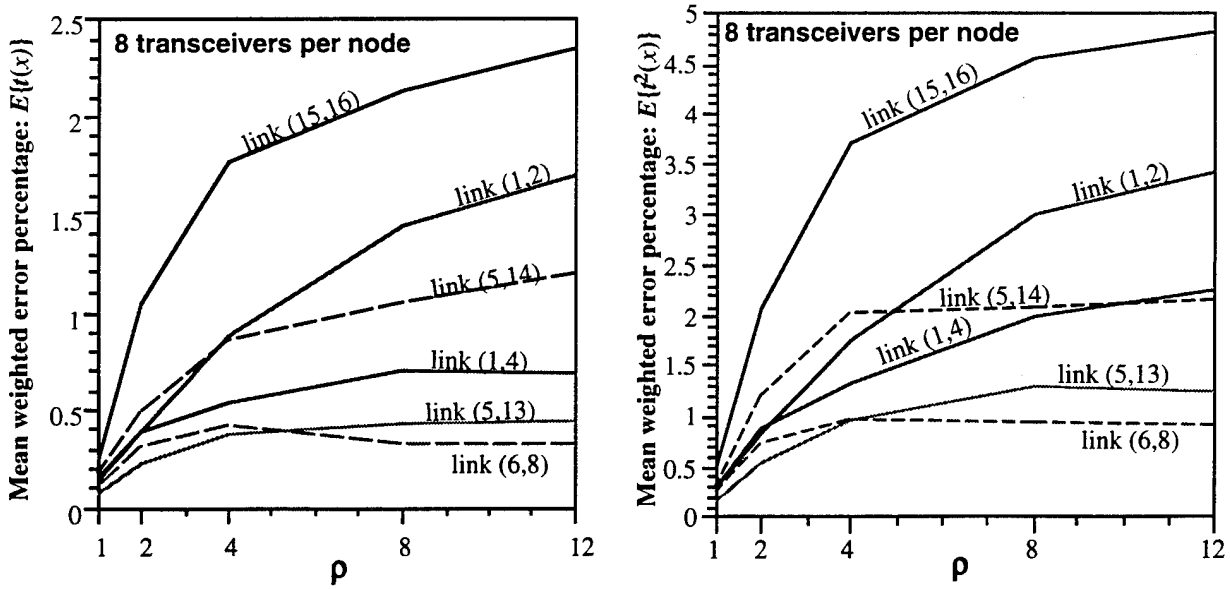


Fig. 9 — Mean weighted error percentage in state-occupancy distribution for various links.

The results shown in Figs. 9 and 10 suggest that the mini-product-form model is more accurate when applied to links that serve many voice circuits. For example, the mini-product-form characterization of the statistics for links (15,16) and (1,2) is generally less accurate than that of the other links. These support only three circuits each. In addition, flow f_b on the mini-product-form model of link (15,16) is zero. The results also show that, as we noted in our detailed study of link (5,13), the accuracy tends to decrease as the offered load is increased.

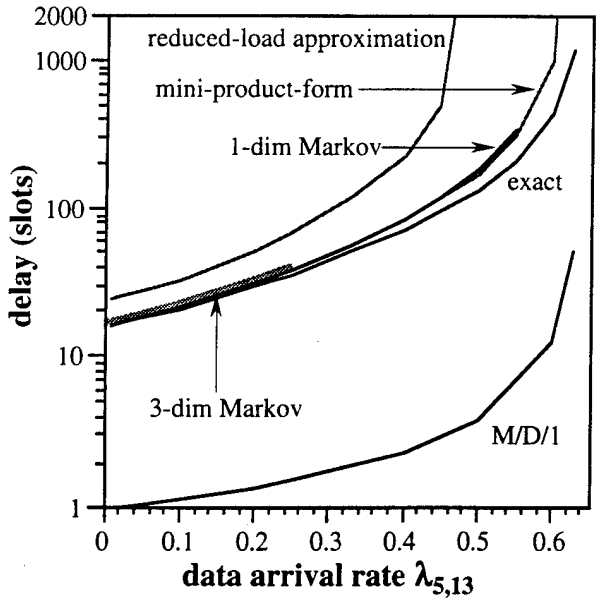


(a) E{$t(x)$} (b) E{$t^2(x)$}

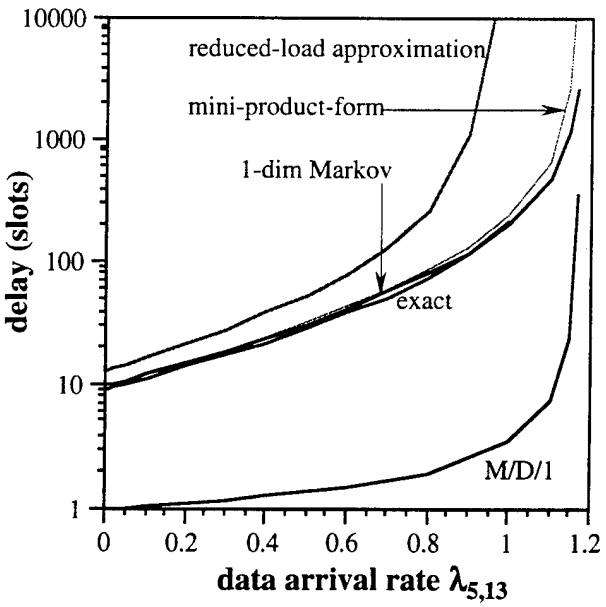
Fig. 10 — Mean weighted error percentage in first two moments of expected state-visit duration using the mini-product-form model.

6.2 Data-packet delay evaluation

Data-packet delay has been evaluated using the models discussed in this report. As discussed in Section 3, the expected delay is obtained from the expected queue size by means of Little's formula. Figures 11 and 12 show data-packet delay as a function of data-packet arrival rate for two selected links of the network of Fig. 3, namely links (5,13) and (1,2), for the case of $T = 4$ transceivers at each node and for voice call loadings of $\rho_j^V = \rho^V = 1$ and 2, $j = 1, 2, ..., 10$. These loadings result in overall blocking probabilities (i.e., fraction of calls arriving to entire network that are blocked) of 0.352 and 0.591, respectively; thus both cases represent a significant amount of blocking, and hence a significant amount of interaction between the links of interest and the rest of the network. The blocking probabilities for the links of interest ($P_{B(a,b)}$) are shown with each figure, as are the expected residual capacities C_{ab}. The expected voice-call duration ($1/\mu^V$) is 100 data packets in all cases. The data-packet arrival rate λ_{ab} is the superposition of the traffic rate from node a to node b and from node b to node a. The "utilization" of the residual capacity of link (a,b) by data traffic is then λ_{ab}/C_{ab}, and the delay over link (a,b) approaches infinity as λ_{ab} approaches C_{ab}.

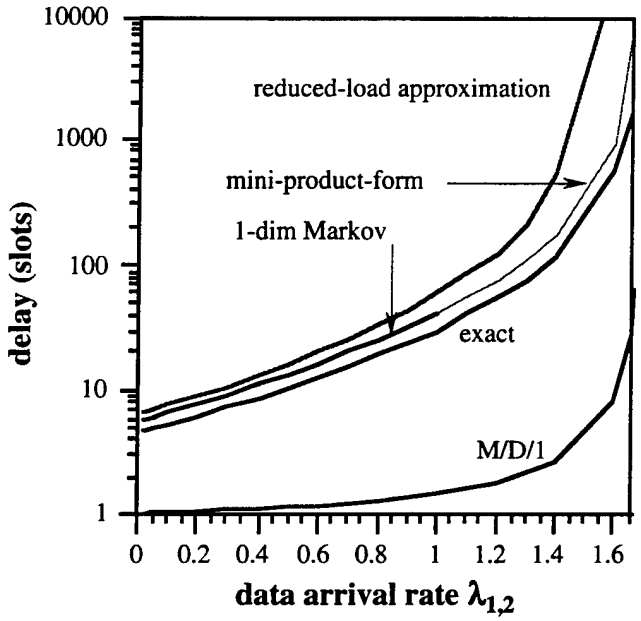


a) $\rho^V = 2$ ($C_{5,13} = 0.64$; $P_{B(5,13)} = 0.533$)

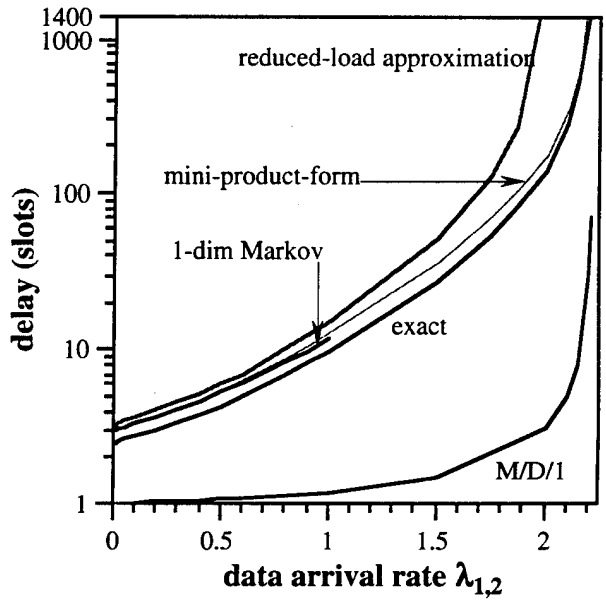


b) $\rho^V = 1$ ($C_{5,13} = 1.17$; $P_{B(5,13)} = 0.311$)

Fig. 11 — Data-packet delay over link (5,13) ($T = 4$ transceivers per node).



a) $\rho^V = 2$ ($C_{1,2} = 1.67$; $P_{B(1,2)} = 0.138$)



b) $\rho^V = 1$ ($C_{1,2} = 2.22$; $P_{B(1,2)} = 0.0575$)

Fig. 12 — Data-packet delay over link (1,2) ($T = 4$ transceivers per node).

In Figs. 11 and 12 "exact" refers to results obtained from simulations of the entire network, "mini-product-form" refers to simulations of the three-flow network, and "reduced-load approximation" refers to simulations of the two-flow model discussed in Section 4.2. All simulations were run for a duration of 10^7 data-packet arrivals. The term "3-dim Markov" (results currently available only for link (5,13) for $\lambda_{5,13} \leq 0.25$ and $\rho^V = 2$) refers to the numerical

results obtained from the Markovian model based on the three-flow mini-product-form model, and "1-dim Markov" refers to the simplified model of Section 5 in which the residual capacity is assumed to be Markovian. The M/D/1 model is based on a constant data-packet service rate equal to the average residual capacity of the link.

Figure 11 shows that the mini-product-form model accurately predicts data-packet delay over link (5,13) at low-through-moderate utilization ranges. Although the accuracy is not as good at high utilization, the mini-product-form model appears to provide a relatively close upper bound on delay, which becomes looser as λ_{ab} approaches C_{ab}. Especially significant is the fact that the results for the one-dimensional Markov model are virtually indistinguishable from those of the simulated mini-product-form model, thus demonstrating that the reduction in complexity does not come at the expense of a significant loss in accuracy.[18] By contrast, the reduced-load approximation results in large errors, severely overestimating delay at moderate to high utilization rates.[19] As expected, the M/D/1 model provides a very poor (and extremely optimistic) estimate of delay. The poor performance of the M/D/1 model is easily explained; since the expected voice-call duration is much greater than the data-packet length, the system can remain in a state with low instantaneous residual capacity for a significant period of time, during which the data-packet queue can grow to large values; such behavior is not predicted by the M/D/1 model.

Figure 12 shows that the mini-product-form model overestimates the delay for link (1,2). Although the accuracy of this estimate is not as good as that obtained for link (5,13), it is much better than that provided by either the M/D/1 or reduced-load models. The fact that the mini-product-form model is more accurate for link (5,13) than for link (1,2) is consistent with the discussion pertaining to Figs. 9 and 10.

6.2.1 Data-packet queue-size distribution

As noted earlier, our delay values have been obtained by means of Little's formula, which relates the expected packet delay to expected queue size. We have also examined the distribution

---

[18] Although results for the one-dimensional Markov model are not currently available for data-packet arrival rates greater than 1.0, the simulated mini-product-form results demonstrate the accuracy of the mini-product-form approach for higher data rates as well.

[19] These results are representative in the sense that the mini-product-form model generally performs better than the reduced-load model. Although the reduced-load model does provide smaller errors for some voice-state occupancies for some links under particular loading considerations (see Fig. 4(b)), in all of the examples we have studied, the mini-product-form model produces a more-accurate estimate of data-packet delay over the entire range of data-traffic arrival rate.

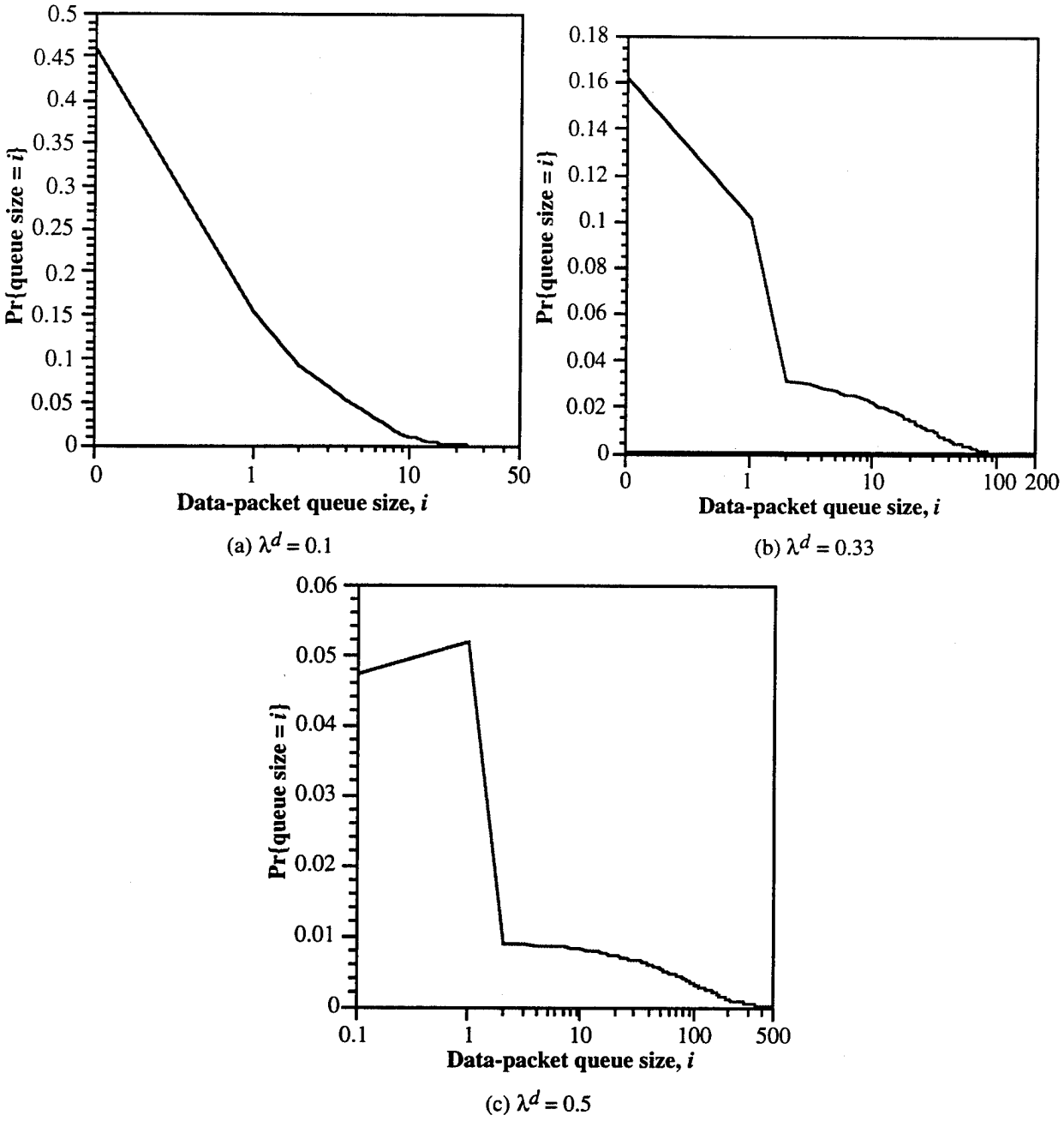
of queue size (rather than just its first moment) in an effort to understand better the delay performance. Figure 13 shows the distribution of queue size[20] for the example corresponding to Fig. 11(a), namely the case of link (5,13) with $\rho^V = 2$ and an expected voice-call duration of 100 times that of data packets, for data-packet arrival rates of $\lambda^d = 0.1$, 0.33, and 0.5 packets per time slot. These results have been obtained numerically using the one-dimensional Markovian model. A modified logarithmic scale has been used for the horizontal axis to illustrate better the behavior for small values of queue size.

Figure 13(a) shows that for $\lambda^d = 0.1$, the queue is empty almost half of the time and contains at most one packet about 60% of the time. However, the probability mass at higher values is sufficient to cause significant delay, as compared to the results predicted by the M/D/1 model. As λ^d increases, the probability mass shifts away from the empty queue. However, it is interesting to note that the two most likely queue sizes are still 0 and 1 for data rates at least as high as 0.5 (a relatively high data rate based on the expected residual capacity of 0.64). In all three curves, there is significant probability mass located at much greater queue sizes than the expected value, which results in a large variance of queue size. In fact, the standard deviation of the queue length is greater than its expected value, as shown in Table 1. The large values of the first two moments of queue size can be attributed principally to the fact that the expected voice-call duration is much greater than the data packet duration, resulting in intervals during which the data-packet queue size can grow to large values, as discussed earlier. For this reason, the data-packet array size used in the numerical evaluation of the Markov chain must be sufficiently large so that significant probability mass is not lost.

Table 1 — First two moments of queue size for the example in Fig. 11(a).

λ^d	E{queue size}	Std. Dev.{queue size}
0.1	2.190	3.386
0.33	18.587	22.970
0.5	90.866	94.171

[20] We define the "queue size" at any time to include the packet being served at that time.



(a) $\lambda^d = 0.1$

(b) $\lambda^d = 0.33$

(c) $\lambda^d = 0.5$

Fig. 13 — Probability mass functions of data-packet queue size for the example in Fig. 11(a).

7. Conclusions

A major challenge in the development of networking models is the attainment of a sufficient degree of accuracy under the constraint of acceptable levels of complexity. For example, an exact model for the evaluation of data-packet delay in integrated wireless networks (under the assumptions made in this report) requires a J+1 dimensional, first-order Markov chain, where J is the number of distinct voice circuits. The resulting combinatorial explosion in

the size of the state space makes this exact approach infeasible, except for extremely small problems. On the other hand, our simplest model for delay, the M/D/1 model, is based on the assumption that the residual capacity available for voice is constant at its expected value. Our studies have demonstrated that, not surprisingly, such a model predicts unacceptably optimistic estimates of delay (by several orders of magnitude for typical system parameters). We have also addressed the possible use of a quasi-static model, which is based on the fact that the voice state changes much more slowly than the data state because the expected voice-call duration is typically several orders of magnitude greater than the data-packet duration. However, we have noted that such models predict infinite delay whenever the offered data-packet load is greater than the residual capacity in any voice state with nonzero probability. Thus such models are of no practical use in estimating delay unless sufficient capacity is reserved for the exclusive use of data.

In this report we have introduced a "mini-product-form" model for the evaluation of expected data-packet delay in integrated multihop wireless networks. In this model, an approximate three-dimensional Markovian model is used to characterize the time-varying voice process at any link of interest, resulting in significant reduction in complexity over the exact system model. In addition, we have developed a simplified one-dimensional characterization of the voice process, which very closely approximates the three-dimensional model. Most importantly, we have demonstrated, although preliminarily, that both of these models can provide accurate delay estimates that also happen to be bounds, at least for certain networking examples. Further work is needed to establish the applicability of our models to more-general scenarios.

Both the mini-product-form model and its simplified one-dimensional version represent a compromise between the complexity and accuracy of the exact Markovian model and the simplicity but inaccuracy of the M/D/1 model. They provide better accuracy than the alternative model that is based on the reduced-load approximation, which makes the inaccurate assumption that the voice state at a node is independent of the voice state at its neighboring nodes. Our models are certainly not the final word on the development of delay models. These efforts represent a modest step toward the development of a universally applicable methodology for delay evaluation. It is hoped that they will serve as the basis for the development of simpler, but acceptably accurate, models that can facilitate the evaluation of system performance in larger networks and in more general scenarios.

References

[1] L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, New York: McGraw-Hill, 1964.

[2] D. Bertsekas and R. Gallager, *Data Networks*, Englewood Cliffs: Prentice Hall, pp. 382-403, 1987.

[3] J. E. Wieselthier and A. Ephremides, "Fixed- and Movable-Boundary Channel-Access Schemes for Integrated Voice/Data Networks," *IEEE Transactions on Communications*, **43** pp. 64-74, January 1995.

[4] J. E. Wieselthier and A. Ephremides, "Performance Analysis of Fixed- and Movable-Boundary Channel-Access Schemes for Integrated Voice/Data Wireless Networks," *Proceedings of IEEE INFOCOM'93*, San Francisco, CA, pp. 1204-1213, March 1993.

[5] J. E. Wieselthier and A. Ephremides, "A Movable-Boundary Channel-Access Scheme for Integrated Voice/Data Networks," *Proceedings of IEEE INFOCOM'91*, Bal Harbour, FL, pp. 721-731, April 1991.

[6] J. E. Wieselthier and A. Ephremides, "A Study of Channel-Access Schemes for Integrated Voice/Data Radio Networks," NRL Report 9359, Naval Research Laboratory, November 29 1991.

[7] C. M. Barnhart, J. E. Wieselthier, and E. A, "Admission Control Policies for Integrated Communication Networks," *submitted to ACM Journal on Wireless Networks*, 1994.

[8] C. M. Barnhart, J. E. Wieselthier, and A. Ephremides, "An Approach to Voice Admission Control in Multihop Wireless Networks," *Proceedings of IEEE INFOCOM'93*, San Francisco, CA, pp. 246-255, March 1993.

[9] C. M. Barnhart, J. E. Wieselthier, and A. Ephremides, "Admission Control in Integrated Voice/Data Multihop Radio Networks," NRL/MR/5521--93-7196, Naval Research Laboratory, January 18, 1993.

[10] J. E. Wieselthier, C. M. Barnhart, and A. Ephremides, "Standard Clock Simulation and Ordinal Optimization Applied to Admission Control in Integrated Communication Networks," *Journal of Discrete Event Dynamic Systems: Theory and Applications*, **5** pp. 243-279, 1995.

[11] J. E. Wieselthier, C. M. Barnhart, and A. Ephremides, "Ordinal Optimization of Admission Control in Wireless Multihop Integrated Networks via Standard Clock Simulation," NRL Formal Report NRL/FR/5521—95-9781, Naval Research Laboratory, 1995.

[12] J. E. Wieselthier, C. M. Barnhart, and A. Ephremides, "Ordinal Optimization of Admission Control in Wireless Multihop Voice/Data Networks via Standard Clock Simulation," *Proceedings of IEEE INFOCOM'94*, Toronto, Ontario, Canada, pp. 29-38, June 1994.

[13] C. M. Barnhart, J. E. Wieselthier, and A. Ephremides, "Ordinal Optimization by means of Standard Clock Simulation and Crude Analytical Models," *Proceedings of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, FL, pp. 2645-2647, December 1994.

[14] J. E. Wieselthier, C. M. Barnhart, and A. Ephremides, "A Mini-Product-Form-Based Solution to Data-Delay Evaluation in Wireless Integrated Voice/Data Networks," *Proceedings of IEEE INFOCOM'95*, Boston, MA, pp. 1044-1052, April 1995.

[15] F. P. Kelly, "Blocking Probabilities in Large Circuit-Switched Networks," *Advances in Applied Probability*, **18** pp. 473-505, 1986.

[16] K. W. Ross and D. Tsang, "Teletraffic Engineering for Product-Form Circuit-Switched Networks," *Advances in Applied Probability*, **22** pp. 657-675, 1990.

[17] J. M. Aein, "A Multi-User-Class, Blocked-Calls-Cleared, Demand Access Model," *IEEE Transactions on Communications*, **COM-26**-No. 3 pp. 378-385, March 1978.

[18] S. Jordan and P. Varaiya, "Control of Multiple Service, Multiple Resource Communication Networks," *IEEE Transactions on Communications*, **42**-11 pp. 2979-2988, November 1994.

[19] S. Jordan and P. Varaiya, "Throughput in Multiple Service, Multiple Resource Communication Networks," *IEEE Transactions on Communications*, **39**-No. 8 pp. 1216-1222, August 1991.

[20] G. J. Coviello and P. A. Vena, "Integration of Circuit/Packet Switching by a SENET (Slotted Envelope Network) Concept," *Conference Record of National Telecommunications Conference*, pp. 42.12-42.17, 1975.

[21] D. J. Baker, A. Ephremides, and J. A. Flynn, "The Design and Simulation of a Mobile Radio Network with Distributed Control," *IEEE Journal on Select Areas in Communications*, **SAC-2**-No. 1 pp. 226-237, January 1984.

[22] A. Ephremides, J. E. Wieselthier, and D. J. Baker, "A Design Concept for Reliable Mobile Radio Networks with Frequency Hopping Signaling (Invited)," *Proceedings of the IEEE*, **75**-No. 1 pp. 56-73, January 1987.

[23] M. J. Post, P. E. Sarachik, and A. S. Kershenbaum, "A 'Biased Greedy' Algorithm for Scheduling Multi-Hop Radio Networks," *Proceedings of the '85 Conference on Information Science and Systems*, pp. 564 - 572, March 1985.

[24] F. P. Kelly, "Routing in Circuit-Switched Networks: Optimization, Shadow Prices and Decentralization," *Advances in Applied Probability*, **20** pp. 112-144, 1988.

[25] K. W. Ross and D. H. K. Tsang, "The Stochastic Knapsack Problem," *IEEE Transactions on Communications*, **37**-No. 7 pp. 740 - 747, July 1989.

[26] S.-P. Chung and K. W. Ross, "Reduced Load Approximations for Multirate Loss Networks," *preprint*, February 1991.

[27] D. Mitra, "Asymptotic Analysis and Computational Methods for a Class of Simple, Circuit-Switched Networks with Blocking," *Advances in Applied Probability*, **19** pp. 219-239, 1987.

[28] E. Geraniotis and I.-H. Lin, "Control of Integrated Voice/Data Multihop Radio Networks Via Reduced-Load Approximations," NRL/MR/5521--93-7390, Naval Research Laboratory, September 1993.

[29] J. E. Shore, "Information Theoretic Approximations for M/G/1 and G/G/1 Queueing Systems," *Acta Informatica*, **17** pp. 43-61, 1982.