Investigating Dependence in Packet Queues with the Index of Dispersion for Work

Kerry W. Fendick, Vikram R. Saksena, *Senior Member, IEEE,* and Ward Whitt

Abstract— **This paper continues an investigation of the way diverse traffic from different data applications affects the performance of packet queues. This traffic often exhibits significant dependence among successive interarrival times, among successive service times, and between interarrival times and service times, which can cause a significant degradation of performance under heavy loads (and often even under moderate loads). This dependence and its effect on performance (specifically, the mean steady-state workload) is partially characterized here by the cumulative correlations in the total input process of work, which we refer to as the index of dispersion for work (IDW). This paper evaluates approximations for the mean steady-state workload based on the IDW by making comparisons with computer simulations. The packet queue model has a single server with unlimited waiting space and a work-conserving discipline plus multiple classes of traffic with class-dependent arrival processes and service requirements, as is appropriate for packet networks serving different data applications with variable packet lengths.**

I. Introduction

THIS paper is the second part of an investigation into the way diverse traffic from different applications affects the performance of packet queues. In [1], we showed that the traffic to packet queues often exhibits three different kinds of dependence: 1) among successive interarrival times, 2) among successive service times, and 3) between interarrival times and service times (the last two occurring because of variable packet lengths). We also provided methods for quantifying and measuring these three kinds of dependence in the traffic, and predicting how the three kinds of dependence affect the performance of a heavily loaded queue. (The queue is assumed to have a single server, unlimited waiting room and the first-in–first-out service discipline, without any windows or other flow control mechanisms; the particular performance measure considered is the mean steady-state workload.) From our simulation experiments and from previous studies of queues with arrival process that are not renewal processes [2]–[12], we know that dependence in the traffic affects performance much more when a queue is heavily loaded (has a high traffic intensity) than when it is lightly loaded. An $M/G/1$ model (or more general $GI/G/1$ model plus approximations [13]), ignoring the three kinds of dependence mentioned above, may accurately approximate the mean steady-state workload when the traffic intensity is low, but may underestimate the mean steady-state workload (often by a factor of ten or more) when the traffic intensity is high. Thus, the heavy-traffic approximations in [1] and previous papers are important to predict degradation in performance of packet queues resulting from increased loads.

While the heavy-traffic approximations are useful to understand system behavior under heavy loads, we typically want to predict system performance under more realistic design loads. Unfortunately, it often happens that neither a light-traffic approximation, which usually counts *none* of the correlations in the traffic nor a heavy-traffic approximation, which usually counts *all* the correlations in the traffic, provides an adequate description of performance under realistic design loads. (The reason that a heavy-traffic approximation counts all the correlations is explained in [1].) *A major goal of this study is to determine approximate performance measures for realistic design loads.*

We attack this problem by trying to estimate how much of the total correlation in the offered traffic actually has a significant impact on the performance of the queue as a function of the traffic intensity of the queue. Our method is described in [14]. It is similar in spirit to the hybrid approximations based on the stationary-interval and asymptotic approximations in [4], [5], [9] and the interpolations based on light and heavy-traffic limit theorems in [6], [15], [16], but our method has a somewhat different focus, being more closely connected to traffic measurements. In particular, we propose measurements to describe the dependence in the traffic, and we develop associated approximations based on these measurements to predict the performance of queues to which this traffic is offered. We also develop approximations based on these concepts that depend on model parameters instead of measurements. As in [1] and [9], we exploit indexes of dispersion for this purpose. Indexes of dispersion describe the cumulative correlations in a stochastic process; indexes of dispersion are also known as variance-time curves; Fourier transforms of them and their derivatives are known as the spectral measure and the spectral density; see [14, Section III], [17, p. 71], and [18]. Of course, these indexes of dispersion are only partial characterizations of the dependence in the offered traffic, but we believe useful partial characterizations.

In [1] we proposed a new three-dimensional index of dispersion for intervals (3-D–IDI) in order to capture the three kinds of dependence in the traffic mentioned above. In [14] we proposed yet another index of dispersion, the *index of dispersion for work* (IDW), which has the advantage of being one dimensional. Our purpose now is to investigate

Paper approved by the Editor for Queueing/Buffer Management of the IEEE Communications Society. Manuscript received February 7, 1989; revised February 18, 1990.

K.W. Fendick and V.R. Saksena are with AT&T Bell Laboratories, Homdel, NJ 07733.

W. Whitt is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9100832.

0090–6778/91/0800–1231$01.00 © 1991 IEEE

how well these (and other) indexes of dispersion describe the offered traffic to packet queues from the perspective of the mean steady-state workload. By graphically comparing how the indexes of dispersion and the mean steady-state workload vary with model parameters, we identify the IDW as the most promising index of dispersion from this perspective. We then evaluate the accuracy of approximations from [14] for the mean workload that use the IDW as the *sole* description of dependence in the offered traffic. Our analysis indicates that the IDW describes much of the effect of the dependence in the offered traffic on the mean workload of packet queues, so that the approximations from [14] may suffice for many engineering purposes. On the other hand, our analysis shows that we would need to use additional information, beyond the IDW alone, to achieve a high degree of accuracy. Given the partial success here of approximations that use only the IDW, further work on approximations, based on the IDW but not exclusively so, would appear to be fruitful.

In this paper we focus on only one performance measure in an idealized model: the mean steady-state workload in a single-server queue with unlimited waiting space and the first-in–first-out discipline (or any other work-conserving discipline) without windows or other flow control mechanisms. It would be desirable to consider more complicated models, and in the future we intend to do so. However, to understand the predictive power of the IDW, it is useful to consider first a relatively simple model.

The workload at time t is the total remaining service time of all packets waiting to be served at time t, including the packet in service. (The workload is often called the virtual waiting time.) The mean workload obviously is not the only performance measure worth considering; we focus on the mean workload primarily because it seems to be the measure that we have the best chance of accurately describing with the IDW. We find that the IDW is relatively effective for describing the effect of the dependence on the mean workload, but it may not capture other effects of the dependence. For example, Ramaswami [19] describes relatively high probabilities of successive lost packets from the same source even under moderate loads in the model of [9], [10].

Besides being of interest in its own right, the mean steady-state workload is often used as a surrogate for the mean wait in queue. For systems with Poisson arrival processes, the mean steady-state workload and the mean steady-state waiting time coincide at all traffic intensities [20]. Moreover, under very general conditions [1], [2], the mean steady-state workload and the mean wait coincide asymptotically as the traffic intensity of the queue approaches 1. However, for many packet queues where individual applications offer packets to the queue in bursts, the mean steady-state workload and the mean wait differ substantially. Thus, we regard the analysis here as being primarily for the mean workload. In the future, we intend to study approximations for the mean wait experienced by each application.

Without loss of generality, we can express the mean steady-state workload EZ as

$$EZ = \frac{\tau \rho c_Z^2(\rho)}{2(1-\rho)} \tag{1}$$

where τ is the mean service time, ρ is the traffic intensity, and $c_Z^2(\rho)$ is some function of ρ, which we regard as a measure of the variability in the offered traffic relevant at traffic intensity ρ, see [14, Section I-B]. We call $c_Z^2(\rho)$ the *normalized mean workload.* In an $M/G/1$ queue, $c_Z^2(\rho) = c_s^2 + 1$ where c_s^2 is the squared coefficient of variation (variance divided by the square of the mean) of the service-time distribution. Thus, in an $M/D/1$ queue, $c_Z^2(\rho) = 1$. Consequently, $c_Z^2(\rho) = EZ/E(Z; M/D/1)$, i.e., it is the ratio of the mean workload in the given model to what it would be in an $M/D/1$ queue at the same traffic intensity.

As in [14], we intend to approximate the normalized mean workload using the IDW. The IDW is defined by

$$I_w(t) = \frac{\text{Var } X(t)}{\tau EX(t)}, \quad t \geq 0, \tag{2}$$

if $EX(t) > 0$ and $I_w(t) = 0$ if $EX(t) = 0$ where $X(t)$ represents the total input of work in the time interval $[0, t]$ and τ represents the average service time; see [14, Sections II-A, III-C]. Since the IDW $\{I_w(t) : t \geq 0\}$ is a function of time, we must relate time t to the traffic intensity ρ. Roughly speaking, we think of $c_Z^2(\rho) = I_w(t(\rho))$ where $t(\rho)$ is an increasing function of ρ. The traffic intensity ρ should determine the relevant time scale for the queue, i.e., how much of the variability measured by the IDW plays a significant role; see [14, Section I].

The rest of this paper is organized as follows. In Section II, we describe our multiclass packet queue model, which allows for diverse characteristics of traffic from different applications. In Section III, we use simulations of the packet queue model to see how each index of dispersion partially characterizes the dependence in the offered traffic and enables us to predict the mean workload. These experiments support using the IDW as a basis for approximating the mean workload. In Section IV, we evaluate four different approximations for the mean workload from [14] by comparing them to simulations. None of these four emerges as a clear winner, but all four are far superior to approximations such as those in QNA [21], [22] that consider only one of the three kinds of dependence. Finally, in Section V we draw conclusions. The Appendix contains a heavy-traffic limit theorem for the workload process which proves that $c_Z^2(1) = I_w(\infty)$ in considerable generality.

II. The Packet Queue Model

In this section we describe a model of a multiclass packet queue, which is a special case of the idealized model from [1, Section VI-C] in which the flow of packets from each virtual circuit is unconstrained by windows. (As indicated by [14], the IDW and approximations based on it are not restricted to this model). Here, k classes, each corresponding to a virtual circuit carrying a particular kind of data, share a transmission facility. We assume that the transmission facility has unlimited waiting room. Packets from each class arrive in batches (e.g., messages) with space between successive packets of the same batch due to constraints on the throughput rate. The space between packet arrivals within a batch represents the time between the arrival of the first bit of one packet and the arrival of the first bit of the next packet. After the space that follows the arrival of the last packet of a batch, there is an idle period before the arrival of the first packet of the next batch. The idle period is typically much longer than the space between successive packets within the same batch. For each class, successive service times, idle periods, batches sizes, and spaces between packets of the same batch are independent sequences of i.i.d. (independent and identically distributed) random variables. This model differs from models in [5] and [7]–[12]; there all classes share the same parameters and, in particular, the same mean service times.

Most importantly, the distributions of service times, idle periods, and spaces are each general and class dependent. For class i, service times have mean τ_i and squared coefficient of variation c_{si}^2; idle periods have mean ω_i and squared coefficient of variation c_{Ii}^2; and spaces between packets of the same batch have mean ξ_i and squared coefficient of variation c_{xi}^2. Batch sizes are also class dependent, but are assumed to be geometrically distributed, so that *the packet arrival process from each class is a renewal process.* (This assumption is essential to some, but not all, of the approximations in Section IV.) For class i, batch sizes have mean m_i and squared coefficient of variation c_{bi}^2; because batch sizes are geometrically distributed, $c_{bi}^2 = (m_i - 1)/m_i$. The packet interarrival-time distribution is thus a mixture as follows. With probability $(m_i - 1)/m_i$, the interarrival time corresponds to a single space; with probability $1/m_i$, the interarrival time is the sum of a space plus an idle period. The interarrival-time distribution function for class i is denoted by $F_i(t)$ and its kth moment by μ_{ki}. Then, $c_{ai}^2 = \mu_{2i}/\mu_{1i}^2 - 1$ is the squared coefficient of variation for a class i interarrival time.

Let λ be the total arrival rate of batches and let λp_i be the arrival rate of batches from class i where $p_1 + \cdots + p_k = 1$. (Note that $\lambda = \sum_{i=1}^k \lambda p_i$ and $\lambda p_i = (\omega_i + \xi_i m_i)^{-1}$.) Then $\tau = \sum_{i=1}^k p_i m_i \tau_i / \sum_{i=1}^k p_i m_i$ is the mean service time for all packets, $\overline{\lambda} = \sum_{i=1}^k \lambda p_i m_i$ is the overall packet arrival rate, and $\rho = \lambda \sum_{i=1}^k p_i m_i \tau_i = \overline{\lambda}\tau$ is the traffic intensity. In addition, $q_i = p_i m_i / \sum_{i=1}^k p_i m_i$ is the proportion of all packet arrivals that are of class i, $r_i = \tau_i/\tau$ is the normalized service time for class i, and $\beta_i = m_i \xi_i/(m_i \xi_i + \omega_i)$ is the proportion of busy time in each busy-idle cycle for class i.

We introduce this traffic model as an attempt to capture the burstiness of the traffic. Of course, other related models have also been introduced for this purpose, e.g., see Jain and Routhier [23], Descloux [24], and Li and Mark [25]. We aim to quantify how the clustering of arrivals for each class affects queue performance, in particular, the steady-state mean workload. For a given batch arrival rate (λp_i) and a given batch size m_i for each class, we especially want to describe the behavior of a queue as a function of the mean idle period ω_i and mean space between packets within a batch ξ_i, with all other parameters held fixed. In other words, we want to develop approximations that work well both when packets from the same batch arrive at nearly the same instant and when packets from the same batch arrive spaced far apart. As we demonstrate with simulation results, changes in spacing between packets of the same batch *greatly* affect the mean workload, so understanding the relationship between the clustering of packets and queue performance is important. (Similar observations are made by Descloux [24], whose model incorporates finite buffers, but requires all packets to have the same service time.)

TABLE I
COMMON PARAMETERS FOR THE FOUR EXAMPLES OF MULTICLASS PACKET QUEUES. NOTE THAT THE TRAFFIC INTENSITY ρ IS VARIED BY CHANGING THE MEAN SERVICE TIME τ_i FOR EACH CLASS

class	$(\lambda p_i)^{-1}$	m_i	$\tau_i = r_i$
1–20	121.904	2.0	1.385ρ
21–40	121.904	2.0	0.139ρ
41–45	872.727	30.0	2.771ρ
46–50	872.727	30.0	0.139ρ

Throughout this paper, we consider examples of packet queues that share the parameters given in Table I. For all examples, the idle periods between batches are exponentially distributed, so that $c_{Ii}^2 = 1$, and packet lengths and spaces for each class are assumed to be constant, so that $c_{si}^2 = c_{xi}^2 = 0$. With these parameters, $F_i(t) = 0$ for all $t < \xi_i$. As a consequence, the density of $F_i(t)$, denoted $f_i(t)$, exists in a neighborhood of 0, and $f_i(0) = 0$. (This is relevant for the light-traffic behavior.) Expressing the first three moments of the interarrival-time distribution in terms of the basic model parameters, we find that $\mu_{1i} = (\omega_i + m_i\xi_i)/m_i$, $\mu_{2i} = (2\omega_i^2 + 2\omega_i\xi_i + m_i\xi_i^2)/m_i$ and $\mu_{3i} = (6\omega_i^3 + 6\omega_i^2\xi_i + 3\omega_i\xi_i^2 + m_i\xi_i^3)/m_i$. Classes 1–20 represent virtual circuits carrying interactive data, while classes 41–45 represent virtual circuits carrying file transfer data. Classes 21–40, and classes 46–50 represent virtual circuits carrying acknowledgments for interactive data and file-transfer data, respectively. (As an approximation, we assume that the acknowledgment classes are independent of the data classes.)

The examples differ in the way packets arrive within a batch, i.e., the examples are completely specified except for the mean spacing between packets within the same batch, ξ_i. In Example 1, all packets from a batch arrive at the same instant, i.e., for all i, $\xi_i = \beta_i = 0$ and $\omega_i = 1/\lambda p_i$. Thus, Example 1 is a multiclass batch-Poisson $\left(\sum \left(M^{B_i}/G_i\right)/1\right)$ queue of [1]. In Example 2, a more realistic example of a packet queue, packets from the same batch arrive separated from one another by deterministic, class-dependent spaces, $\xi_1 = \cdots \xi_{40} = 0.227$ and $\xi_{41} = \cdots = \xi_{50} = 0.554$. For each class $i, \omega_i = (1/\lambda p_i) - m_i\xi_i$.

Fig. 1 displays results for the normalized mean workload, $c_Z^2(\rho)$, for Examples 1 and 2. [Recall that $c_Z^2(\rho) = 2(1-\rho)EZ/\tau\rho$ by (1).] For these examples, and all others introduced later, we vary the traffic intensity by varying the mean service time for all classes by a common proportion. In this way, we observe how the performance of a queue with a fixed set of arrival characteristics changes as a function of the traffic intensity. For ease of exposition, we normalize all arrival and service parameters so that the overall packet arrival rate $\overline{\lambda}$ is equal to 1, and the overall mean packet service time τ is equal to the traffic intensity, ρ. It is important to note that when we consider model behavior as a function of ρ we are simply

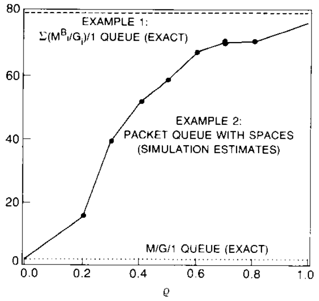


Fig. 1. The normalized mean workload for Examples 1 and 2.

multiplying all service times for the case $\rho = 1$ by ρ, i.e., the arrival process remains unchanged as we change ρ. This is not a restrictive assumption, but a convenient convention (see [14, Sections I-E, II-D, III-D]), but it is different from what is done in most performance analysis studies; typically we increase the load by adding sources and, thus, altering the arrival process, while leaving the service times unchanged.

For Example 1 (the batch-Poisson case), we compute the exact value EZ as given in [1, eq. (2)] and obtain

$$c_Z^2(\rho) = c_A^2 + c_S^2 - 2c_{AS}^2 = 79.9 \quad \text{for all } \rho \tag{3}$$

where c_A^2, c_S^2, and c_{AS}^2 are the asymptotic variability parameters discussed in [1, Section V-B] and [14, Section III].) For Example 2 (with spaces), we use estimates $\hat{Z}(\rho)$ for EZ obtained for several values of ρ via computer simulation to obtain the estimates

$$\hat{c}_Z^2(\rho) = 2(1-\rho)\hat{Z}(\rho)/\tau\rho. \tag{4}$$

Each simulation estimate $\hat{Z}(\rho)$ is the average of 545 000 samples of the workload. The space between samples was exponentially distributed with a mean approximately equal to 6.93 packet interarrival times. Thus, each simulation estimate was based on more than 3 750 000 packet arrivals. A 95% confidence interval for each estimate with a relative width of about 10% was found by dividing the simulation run into 40 intervals of equal length and assuming that the averages for the 40 intervals were distributed according to a t distribution. For each example, the values at $\rho = 0$ and $\rho = 1$ represent the exact asymptotic limits for $c_Z^2(\rho)$, as described in [14, Section IV].

For Example 2, (4) represents what $c_Z^2(\rho)$ must be to match simulation results. In Fig. 1, we also give the corresponding results for an $M/G/1$ queue with the same squared coefficient of variation for a packet service time as in the two packet queue examples, i.e., we let $c_Z^2(\rho) = 1 + c_s^2$, where $c_s^2 \equiv \bar{c}_{S1}^2 = 0.96$ is the exact squared coefficient of variation of a service time computed via (10) of [1], which does not represent any dependence. For the $M/G/1$ model, $c_Z^2(\rho)$ does not depend on ρ.

Fig. 1 dramatically depicts the problem we are facing. Except at very low traffic intensities, a simple $M/G/1$ model seriously underestimates the normalized mean workload in Example 2. The heavy-traffic limit for Example 2 and the batch-Poisson model in Example 1 both describe the normalized mean workload in Example 2 well at high-traffic intensities, but certainly not at lower traffic intensities where the $M/G/1$ model is more appropriate. Assuming that the batch-Poisson model provides an upper bound for the normalized mean workload, the function $\{c_Z^2(\rho) : 0 \leq \rho \leq 1\}$ representing the normalized mean workload as a function of ρ for the packet-queue examples (all with the common parameters in Table I) satisfies $0 \leq c_Z^2(\rho) \leq 79.9$ for all ρ. However, the precise form of the function depends on the relative sizes of ξ_i and ω_i. What we would like to do is obtain an estimate for the function $c_Z^2(\rho)$ directly from the model parameters that matches the rising curve in Fig. 1. From our experiments, we conclude that the indexes of dispersion can serve as the basis for such estimates.

III. The Indexes of Dispersion for Examples 1 and 2

In this section we use simulation to estimate the different indexes of dispersion for Examples 1 and 2. (In some cases, we could also calculate these indexes of dispersion analytically, e.g., by Laplace transform inversion; see [14, Section III-G].) The indexes of dispersion are dimensionless functions of time or the customer index that describe the cumulative correlations in the traffic process; see [17, p. 71], [9], [1, Section V-A -B] and [14, Section III]. The first two moments (as a function of time or the customer index) appearing in the indexes of dispersion are estimated from the simulated offered traffic by sample means. We give more details about the estimation procedures in Section III-B below.

A. Comparing the Indexes of Dispersion to the Normalized Mean Workload

Fig. 2 displays the IDI's (index of dispersion for intervals, see [9, eq. (1)] or [14, eq. (45)], estimated from simulation, for the arrival processes in the two examples from Section II. *The two IDI's are similar and thus do not provide immediate insight into the differing behavior of the normalized mean workload for the two examples, as shown in Fig. 1.* We note that the full IDI is more closely related to packet *waiting-times* than the workload because the IDI and the waiting time are functions of the customer index as opposed to time. To see the limitations of the IDI, consider a single-server, single-class batch-Poisson queue with fixed arrival rate. Then the distribution of the number of packets in a batch completely determines the queue's IDI, while the total amount of work arriving in a batch determines the distribution of the queue's workload; see [1, Section IV]).

Fig. 3 displays the IDC's (index of dispersion for counts, see [9, eq. (3)] or [14, eq. (46)], for the same arrival processes and examples (exact for Example 1 and simulation estimate for Example 2). Unlike the IDI's there is a strong correspondence for each example between the *shape* of the IDC and the *shape* of $c_Z^2(\rho)$ in Fig. 1. First, the $\sum \left(M^{B_i}/G_i\right)/1$ queue, which has a constant normalized mean steady-state workload $c_Z^2(\rho)$, see [1, eq. (2)], also has a constant IDC. (This follows because the arrival process of packets to this queue is a compound-Poisson process, so that there is a zero covariance between the number of packets arriving in disjoint intervals.) Second, the shapes

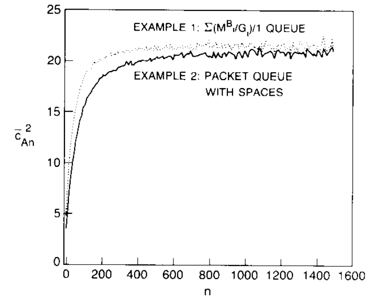


Fig. 2. The one-dimensional index of dispersion for intervals (IDI) for the arrival processes of Examples 1 and 2, estimated by simulation.

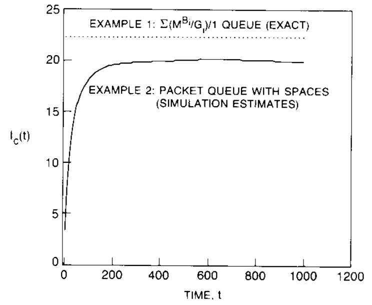


Fig. 3. The index of dispersion for counts (IDC) for the arrival processes of Examples 1 and 2.

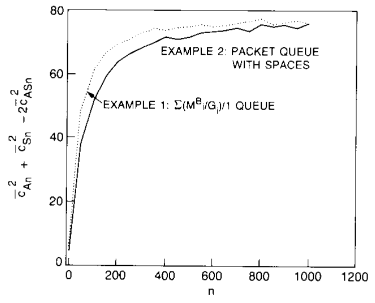


Fig. 4. A composite index of dispersion from the three-dimensional index of dispersion for intervals (3-D–IDI) for Examples 1 and 2, estimated from simulation.

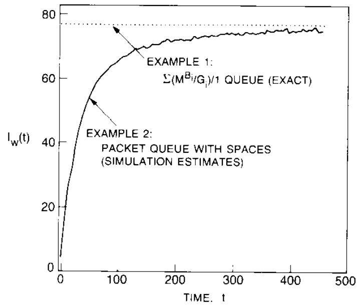


Fig. 5. The index of dispersion for work (IDW) for Examples 1 and 2.

of the curves for Example 2 are also similar. Figs. 1 and 3 together strongly suggest that the IDC provides information for Examples 1 and 2 not only about the total correlation in the arrival process, but also about how much of this total correlation is relevant at a given traffic intensity. However, note that the *values* of $I_c(t)$ in Fig. 3 are substantially less than the values of $c_Z^2(\rho)$ in Fig. 1. From [1], we know that this is because the IDC captures only one of the three kinds of dependence.

Perhaps the main contribution of [1] was to point out the importance of treating the interarrival times and service times together in multiclass queues with class-dependent service times and non-Poisson arrival processes. In [1] we focused on the three-dimensional index of dispersion for intervals (3-D–IDI), which is the sequence $\{(\bar{c}^2_{An}, \bar{c}^2_{Sn}, \bar{c}^2_{Asn}) : 1 \leq n \leq \infty\}$ where $\bar{c}^2_{An}$ is the ordinary IDI for the interarrival times, $\bar{c}^2_{Sn}$ is the corresponding IDI for the service times, and $\bar{c}^2_{ASn}$ is an analogous sequence for the interarrival times and service times together (with $c^2_A = c^2_{A\infty}, c^2_S = c^2_{S\infty}$ and $c^2_{AS} = c^2_{AS\infty}$). Fig. 4 displays the composite sequence $\{\bar{c}^2_{An} + \bar{c}^2_{Sn} - 2\bar{c}^2_{Asn}\}$ estimated from simulation for each of the two examples. We know that the limit of this sequence represents the relevant variability for the queue in heavy traffic. Each of the sequences $\{\bar{c}^2_{An},\}, \{\bar{c}^2_{Sn}\}$, and $\{\bar{c}^2_{Asn}\}$ contributes significantly to the composite sequence: for the $\sum (M^{B_i}/G_i)/1$ queue, $c^2_{A\infty} = 22.3, c^2_{S\infty} = 40.19$, and $c^2_{AS\infty} = -8.8$, while for the queue with spaces $c^2_{A\infty} = 21.5$, $c^2_{S\infty} = 38.7$, and $c^2_{AS\infty} = -8.4$. Thus, the IDI and IDC which capture only the correlations in the arrival process, do not describe the total correlation for these queues well. The 3-D–IDI, on the other hand suffers from the same limitations as the 1-D–IDI in describing the workload under light loads. Fig. 4 shows that the composite sequences for the two examples are similar, so that they do not explain the markedly different behavior of the two queues under light loads seen in Fig. 1.

Fig. 5 displays the IDW's for Examples 1 and 2. The IDW for the $\sum (M^{B_i}/G_i)/1$ queue is constant because the total input process of work is a compound Poisson process and thus has a zero covariance between the work arriving in disjoint intervals; see [14, Section IV-C]. The IDW for Example 2 is estimated from simulation. The general shape of each IDW is similar to that of the corresponding IDC in Fig. 3, and, as with the IDC, corresponds to the shape of the normalized workload $c_Z^2(\rho)$ in Fig. 1. However, unlike the IDC in Fig. 3, the numerical values in Figs. 1 and 5 are very close. Indeed, for the $\sum (M^{B_i}/G_i)/1$ queue, the constant value of the IDW *is equal* to the constant value of $c_Z^2(\rho)$; see [14, eq. (123)]. Moreover, from [14, eqs. (8), (12), and Section IV], we know that in great generality $c_Z^2(1) = I_w(\infty)$ and $c_Z^2(0) = I_w(0)$. The Appendix here contains the supporting heavy-traffic limit theorem cited in [14, Section IV-B1].

After comparing the various indexes of dispersion in Figs. 2–5 to the normalized mean workload in Fig. 1, we conclude that the IDW is most promising for developing approximations for the mean workload; see [14] for further discussion. However, the other indexes of dispersion may be useful for other purposes.

B. The Estimation Procedure

Methods for estimating the indexes of dispersion are discussed in [17]. Here, we briefly describe what we did. For the IDI in Fig. 2, we estimated the first two moments as a function of the customer index n by using samples based on disjoint subsets of interarrival times (for a given n), i.e., the estimate for moment k is

$$\overline{m}_k(n) = m^{-1}\sum_{i=0}^{m-1}\left[\sum_{j=1}^{n} u_{in+j}\right]^k. \tag{5}$$

An attractive alternative to (5) is

$$\hat{m}_k(n) = (mn)^{-1}\sum_{j=1}^{mn}\left[\sum_{i=j}^{i=j+n-1} u_i\right]^k \tag{6}$$

which uses overlapping interarrival times. The estimate (6) is typically more reliable (has lower variance) than (5) for large n using the same number (mn) of interarrival times. However, it should be noted that the estimate (6) based on overlapping interarrival times produces highly correlated estimates of the values of the IDI at successive customer indexes, so that the estimated curves can be smooth even when the statistical precision is not great.

For the indexes of dispersion in Fig. 3–5, we used methods in [17] based on overlapping interarrival times for the IDI and overlapping intervals for the IDC. One method for estimating the IDI in [17] involves estimating, for each i, the covariance C_i between two interarrival times separated by $i - 1$ other interarrival times (see [17, Section 5.2, eq. (11)] and then estimating the full IDI (see [17, Section 4.4, eqs. (10), (11)]. We estimated the 3-D–IDI in Fig. 4 in this way, after replacing the interarrival times in the equations by the values $\hat{d}_n$ equal to the difference between the nth service time, normalized by its expected value, and the nth interarrival time, normalized by its expected value; see [1, Section VII-B].

We estimated the IDC in Fig. 3 as in [17, Section 5.4, eqs. (3), (12)] and [17, Section 4.5, eq. (3)]. This requires dividing the time axis into disjoint time intervals of length δ and estimating, for each i, the covariance $C_i(\delta)$ between the numbers of arrivals in two intervals separated by $i - 1$ other intervals. As suggested in [17], we chose the interval δ equal to less than one half the smallest time value for which we want an estimate of the IDC. The smallest time value we considered was about 1.5 mean service times; subsequent values were about 4 mean services times apart, so that there were 250 data points in Table III.

Finally, we estimated the IDW in Fig. 5 in essentially the same way as the IDC, replacing arriving customers per interval in the equations with arriving work per interval.

IV. Approximations for the Mean Workload

In [14, Sections I-D, and V] we proposed four ways to approximate the normalized mean workload $c_Z^2(\rho)$ based on the IDW $I_w(t)$. Our purpose here is to see how these approximations work for packet queues. The first two approximations are based on the asymptotic behavior of $I_w(t)$ as $t \to 0$ and $t \to \infty$, including the derivatives as well as the limits, which can be expressed directly in terms of the packet queue model parameters. The first procedure is a $GI/G/1$ model approximation, obtained by finding a $GI/G/1$ model with an IDW that matches the asymptotic behavior of the given IDW, see [14, Section V-C]. For packet queues, this procedure typically yields an $H_2/G/1$ queue (hyperexponential interarrival times, a mixture of two exponentials), which can easily be solved exactly by solving for a root of an equation; [26, p. 329]. (In fact, here it yields an $H_2/M/1$ queue.) In Section IV-C below, we show that the $GI/G/1$ model approximation yields a remarkably good fit to the IDW's. Its approximation for the normalized mean workload $c_Z^2(\rho)$ is reasonably good, but not spectacular. Thus, these examples show limitations in the extent to which $I_w(t)$ determines $c_Z^2(\rho)$.

The second procedure is the light-traffic and heavy-traffic interpolation approximation from [16] based on the limits $c_Z^2(0)$ and $c_Z^2(1)$ and the derivatives $\dot{c}_Z^2(0)$ and $\dot{c}_Z^2(1)$, which are determined or approximated by the asymptotic behavior of $I_w(t)$; see [14, eqs. (17), (18), and Section V-B]. The interpolation approximation seems to work as well as any of the other approximations, suggesting that it may not be worthwhile determining the full approximating $GI/G/1$ model or the full IDW.

The third and fourth approximations require the full IDW. They are both based on the idea that $c_Z^2(\rho)$ can be approximated by a time-transformation of $I_w(t)$, i.e.,

$$c_Z^2(\rho) \approx I_w(t(\rho)), \quad 0 \le \rho \le 1 \tag{7}$$

where $t(\rho)$ is an increasing function of ρ with $t(0) = 0$ and $t(\rho) \to \infty$ as $\rho \to 1$. The third approximation is (7) with

$$t(\rho) = \frac{\rho I_w(\infty)}{2(1-\rho)^2}, \quad 0 \le \rho \le 1; \tag{8}$$

the fourth is (7) with $t(\rho)$ being the fixed point solution of the equation

$$t(\rho) = \frac{\rho I_w(t(\rho))}{1-\rho}; \quad 0 \le \rho \le 1; \tag{9}$$

see [14, Sections I-C, D, V-A].

A. The Four Packet-Queue Examples

In this section we introduce two new examples in addition to the two in Section II. All four examples share the parameters in Table I. For all four examples, the idle periods between batches are exponentially distributed, so that $c_{Ii}^2 = 1$, and packet lengths and spaces for each class are assumed constant, so that $c_{si}^2 = 0$ and $c_{\xi i}^2 = 0$. Hence, the four examples differ only by the size ξ_i of the space between the arrival of successive packets in the same batch and the resulting mean idle periods $\omega_i = (\lambda p_i)^{-1} - m_i\xi_i$. The space parameters for the four examples are given in Table II. In each case, the spaces are twice as long for the file transfer packets as for the interactive packets. The spaces are increased in each succeeding example, so that each succeeding example becomes less bursty.

TABLE II

THE SPACE BETWEEN PACKETS IN THE SAME BATCH ξ_i; THE PROPORTION OF THE TOTAL PACKET ARRIVAL RATE DUE TO CLASS i, q_i; THE PROPORTION OF BUSY TIME PER BUSY CYCLE FOR CLASS i, β_i; AND THE ASYMPTOTIC ARRIVAL-PROCESS VARIABILITY PARAMETER c^2_{Ai} FOR CLASS i IN THE FOUR PACKET-QUEUE EXAMPLES. ALSO GIVEN ARE THE OVERALL ASYMPTOTIC VARIABILITY PARAMETERS

Variable	Class	Example 1	Example 2	Example 3	Example 4
ξ_i	1–20	0.0	0.277	0.831	6.926
	21–40	0.0	0.277	0.831	6.926
	41–45	0.0	0.554	1.662	13.853
	46–50	0.0	0.554	1.662	13.853
q_i	1–20	0.0164	0.0164	0.0164	0.0164
	21–40	0.0164	0.0164	0.0164	0.0164
	41–45	0.0343	0.0343	0.0343	0.0343
	46–50	0.0343	0.0343	0.0343	0.0343
β_i	1–20	0.0	0.0045	0.0136	0.114
	21–40	0.0	0.0045	0.0136	0.114
	41–45	0.0	0.0190	0.0571	0.476
	46–50	0.0	0.0190	0.0571	0.476
c^2_{Ai}	1–20	3.0	2.97	2.92	2.35
	21–40	3.0	2.97	2.92	2.35
	41–45	59.0	56.8	52.5	16.2
	46–50	59.0	56.8	52.5	16.2
c^2_A		22.2	21.4	19.9	7.1
c^2_S		40.0	38.6	35.7	11.4
c^2_{AS}		−8.8	−8.5	−7.8	−2.2
$c^2_Z(1)$		79.9	77.0	71.2	22.9

Also given in Table II are associated variables calculated from the basic parameter four-tuples $(\lambda p_i, m_i, \tau_i, \xi_i)$: in particular, the proportion of the total packet arrival rate due to class i, $q_i = p_i m_i / \sum_{i=1}^{50} p_i m_i$; the proportion of busy time per class-i busy cycle, $\beta_i = m_i \xi_i/(m_i \xi_i + \omega_i)$; and the asymptotic arrival process variability parameter for class $i, c^2_{Ai} = (1-\beta_i)^2(2m_i - 1)$, which is obtained from [1, eq. (26)]. (There, $c^2_{bi} = (m_i - 1)/m_i, c^2_{Ii} = 1$ and $c^2_{Ti} = 0$.) Moreover, the overall asymptotic variability parameters c^2_A, c^2_S, c^2_{AS}, and $c^2_Z(1) = c^2_A + c^2_S - 2c^2_{AS}$ are given, as obtained from [1, eqs. (17) and (2)] using c^2_{Ai} above.

B. The IDW in the Packet Queue Model

Now we develop expressions for our IDW and its asymptotic behavior as $t \to 0$ and $t \to \infty$ for the packet queue model, drawing on [14]. The expressions also are valid for the more general packet-queue model in [1, Section II] where the distributions (for each class) of the space between packets and the idle period between batches are both general. When comparing this analysis to [14], note that in the definition of the IDW here the total arrival rate is fixed at 1, whereas in [14] the total service rate is fixed at 1. Thus, in [14] there appears an extra time scaling of the IDW.

We assume equilibrium conditions, i.e., that the arrival counting process for each class, denoted by $A_i(t)$, and thus also the total input process for each class, denoted by $X_i(t)$, has stationary increments, so that $EX_i(t) = \rho_i t, t \geq 0$. Since different classes in the packet-queue model have independent total input processes of work, from (2) we obtain

$$I_w(t) = \frac{\operatorname{Var}\left[\sum_{i=1}^{k} X_i(t)\right]}{\tau\rho t}$$

$$= \sum_{i=1}^{k} \frac{\tau_i \rho_i}{\tau\rho} \frac{\operatorname{Var} X_i(t)}{\tau_i \rho_i t} = \sum_{i=1}^{k} \frac{\tau_i \rho_i}{\tau\rho} I_{wi}(t) \qquad (10)$$

where $I_{wi}(t)$ is the IDW for class i alone. For each class, service times are i.i.d. random variables that are independent of arrival times, so by [14, eq. (59)],

$$I_{wi}(t) = c^2_{si} + I_{ci}(t) \qquad (11)$$

where $I_{ci}(t)$ is the *index of dispersion for counts* (IDC) for class i alone, defined by $I_{ci}(t) = \operatorname{Var} A_i(t)/EA_i(t)$, $t \geq 0$. Equations (10) and (11) together show that the service-time distribution for each class enters into the IDW only through the dimensionless variability parameter c^2_{si} and the dimensionless ratios (τ_i/τ) and (ρ_i/ρ). This implies that the IDW is invariant under changes in the overall mean service rate (assuming that the different class rates are changed proportionately). Hence, when we vary the traffic intensity in each of our examples by changing the mean service rates for each class, the IDW remains unchanged.

As noted in Section II, the packet arrival process for each class is a renewal process. This allows us to calculate $V_i(t) \equiv \operatorname{Var} A_i(t)$ for each i, which characterizes the IDC because $I_{ci}(t) = \mu_i V_i(t)/t$ where μ_i is the mean interarrival time for class i. If $H_{0i}(t)$ is the renewal function for class i arrivals, i.e., the mean number of arrivals in the interval $(0, t]$ given that a new interarrival time begins at $t = 0$, then

$$V_i(t) = \frac{2}{\mu_i} \int_0^t \left[H_{0i}(u) - \frac{u}{\mu_i} + \frac{1}{2}\right] du; \qquad (12)$$

see [14, eq. (62)]. We can find the renewal function $H_{0i}(t)$ by inverting its Laplace transform,

$$\hat{H}_{0i}(s) \equiv \int_0^\infty e^{-st} H_{0i}(t)\, dt = \frac{\hat{f}_i(s)}{s\left(1 - \hat{f}_i(s)\right)} \qquad (13)$$

where $\hat{f}(s) \equiv \int_0^\infty e^{-st} dF_i(t)$ is the Laplace–Stieltjes transform of the interarrival-time distribution for class i. If we cannot obtain a closed-form expression for $H_{0i}(t)$ from (13), we may find $V_i(t)$ by numerically inverting its Laplace transform, $\hat{V}_i(s) \equiv \int_0^\infty e^{-st} V_i(t)\, dt$, which is (see [14, eq. (62)]

$$\hat{V}_i(s) = \frac{2}{\mu_i s}\left[\hat{H}_{0i}(s) - \frac{1}{\mu_i s^2} + \frac{1}{2s}\right]. \tag{14}$$

Hence, the IDC, and thus the IDW too by (11), can be computed analytically for our packet queue model.

We now describe the asymptotic behavior of the IDW. First, by [14, eq. (92)]

$$I_w(0) = \sum_{i=1}^{k} \frac{\tau_i \rho_i}{\tau \rho}\left(c_{si}^2 + 1\right) = c_s^2 + 1. \tag{15}$$

Second, by [14, eq. (99)]

$$I'_{wi}(0) = h_{0i}(0) - \frac{1}{\mu_i} = f_i(0) - \frac{1}{\mu_i} \tag{16}$$

where $h_{0i}(0)$ is the density of the renewal function and $f_i(0)$ is the density of the stationary interarrival-time distribution at 0. Of course, for Examples 2–4, $f_i(0) = 0$ so that $I'_w(0) < 0$ for these examples.

We can also characterize the large-time behavior of $I_{ci}(t)$ in simple terms, again using the fact that the arrival process for class i is renewal. If $\mu_{1i} = \tau_i/\rho_i, \mu_{2i},$ and μ_{3i} are the first three moments of a class i interarrival time, respectively, and $c_{ai}^2 = \left(\mu_{2i}/\mu_{1i}^2\right) - 1$, then,

$$I_{ci}(t) = c_{ai}^2 - \frac{1}{t}\left[\frac{\mu_{1i}}{3}\right]\left[\frac{\mu_{3i}}{\mu_{1i}^3} - 1.5\left(c_{ai}^2 + 1\right)^2\right] + o\left[\frac{1}{t}\right] \quad \text{as } t \to \infty; \tag{17}$$

see [14, eq. (115)]. Hence,

$$I_w(t) = A - \frac{B}{t} + o\left[\frac{1}{t}\right] \quad \text{as } t \to \infty \tag{18}$$

where $A = c_A^2 + c_S^2 - 2c_{AS}^2 = c_Z^2(1)$ and

$$B = \sum_{i=1}^{k} \frac{\tau_i \rho_i}{\tau \rho}\, \frac{\mu_{1i}}{3}\left[\frac{\mu_{3i}}{\mu_{1i}^3} - 1.5\left(c_{ai}^2 + 1\right)^2\right]; \tag{19}$$

see [14, eq. (116)].

C. The $GI/G/1$ Model Approximation

The standard $GI/G/1$ queue is a special case of our general packet queue model in which there is one class with $m_1 = 1$ and $c_{b1}^2 = \xi_1 = c_{x1}^2 = 0$, so that the asymptotics above also apply to the $GI/G/1$ model. Hence, for the $GI/G/1$ queue,

$$I_w(0) = c_s^2 + 1 \quad \text{and} \quad I'_w(0) = f(0) - 1 \tag{20}$$

by (15) and (16), and (18) holds with

$$A = I_w(\infty) = c_a^2 + c_s^2 \quad \text{and} \quad B = \left[\mu_3 - 1.5\left(c_a^2 + 1\right)^2\right]/3 \tag{21}$$

where 1, $\left(c_a^2 + 1\right)$ and μ_3 are the first three moments of an interarrival time and c_s^2 is the squared coefficient of variation of a service time.

Given the parameters $I_w(0), A$ and B, we thus fit a $GI/G/1$ queue by letting

$$\begin{aligned} c_s^2 &= I_w(0) - 1, \\ c_a^2 &= I_w(\infty) - c_s^2 = I_w(\infty) - I_w(0) + 1, \\ \mu_3 &= 3B + 1.5\left(c_a^2 + 1\right)^2 = 3B + 1.5(I_w(\infty) - I_w(0) + 2)^2. \end{aligned} \tag{22}$$

By [14, eqs. (8) and (12)] a $GI/G/1$ queue with these parameter will share common values for $c_Z^2(0)$ and $c_Z^2(1)$ with the packet queue in addition to a similar correlation structure, as captured by the asymptotics of the IDW. We do not use our knowledge of $I'_w(0)$ for the packet-queue model in fitting the $GI/G/1$ models here, primarily because we succeed in getting a good fit of the IDW without it.

The commonly studied distributions for interarrival times and service times of $GI/G/1$ queues have limits on the range of values assumed by their first three moments. To fit $GI/G/1$ queues to the packet-queue examples, we must use particular qualities of the packet queues in choosing appropriate distributions for the $GI/G/1$ model. We may then apply methods given in [4, Section 3] for fitting these distributions to the moments. For all examples in this paper, $c_s^2 = 0.96 \approx 1.0$, so we use an exponential service-time distribution for the $GI/G/1$ model. (Our procedure does not require this, however.) Although the examples cover a broad range of values for ω_i and ξ_i, the parameter B in (19) is positive and $I_w(\infty) > I_w(0)$ for all examples (except the $\sum\left(M^{B_i}/G_i\right)/1$ multiclass batch-Poisson example). Thus, it is sufficient to select an interarrival-time distribution for the $GI/G/1$ queue that can satisfy (22) whenever $c_a^2 > 1$ and $\mu_3 > 1.5\left(c_a^2 + 1\right)^2$. Conveniently, the conditions $c_a^2 > 1$ and $\mu_3 > 1.5\left(c_a^2 + 1\right)^2$ together are sufficient (as well as necessary) for a hyperexponential (H_2) distribution to exist with these moments; see [4, p. 136]. The density of an H_2 distribution is given by

$$f(t) = p\lambda_1 e^{-\lambda_1 t} + (1 - p)\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0. \tag{23}$$

The first three moments of the interarrival-time distribution uniquely determine p, λ_1, and λ_2.

An H_2 distribution for interarrival times has two qualities important for our analysis. First, a simple expression exists for the mean steady-state waiting time for the $H_2/G/1$ queue (although evaluating this expression includes numerically finding a root of an equation; see [26, p. 329]. Using the mean steady-state waiting time, we can thus find the mean steady-state workload by applying Brúmelle's formula [27], in which we make the substitution $E(vW) = (Ev)(EW)$ since in a $GI/G/1$ queue the waiting time W and service time v of each customer are independent. Second, we can calculate the IDW for an $H_2/G/1$ queue analytically. By [14, eq. (65)], the IDC for an H_2 arrival process is given by

$$I_c(t) = c_a^2 - \frac{2\eta}{\gamma t} + \frac{2\eta}{\gamma t} e^{-\gamma t}, \quad t \geq 0 \tag{24}$$

TABLE III

A COMPARISON OF THE FOUR APPROXIMATIONS PLUS QNA [21] WITH SIMULATION ESTIMATES OF THE NORMALIZED MEAN WORKLOAD IN EXAMPLES 2–4

example	traffic intensity ρ	simulation estimate	$H_2/M/1$ model fit	interpolation	fixed-point equation (7) + (9)	simple time transformation (7) + (8)	QNA [21]
2	0.0	2.0	2.0	2.0	2.0	2.0	2.0
	0.2	15.4	4.6	22.7	4.0	24.2	2.0
	0.3	39.5	13.3	32.7	14.0	37.6	2.1
	0.4	51.7	31.8	41.9	41.0	50.7	2.1
	0.5	58.9	46.2	49.9	57.0	61.4	2.2
	0.6	67.3	56.3	57.2	65.5	68.6	2.4
	0.7	70.7	63.6	63.5	71.0	72.9	2.9
	0.8	71.1	69.2	68.9	74.0	75.4	4.1
	1.0	77.0	77.0	77.0	77.0	77.0	22.4
3	0.0	2.0	2.0	2.0	2.0	2.0	2.0
	0.2	1.4	2.5	10.5	2.0	9.6	2.0
	0.3	–	2.9	16.4	2.0	16.0	2.1
	0.4	14.4	3.9	23.0	3.0	24.6	2.1
	0.5	–	6.6	30.1	6.0	35.8	2.2
	0.6	41.0	16.8	37.6	21.1	48.5	2.4
	0.7	–	33.9	45.5	42.2	59.6	2.8
	0.8	61.5	50.0	53.8	56.0	66.7	3.9
	1.0	71.2	71.2	71.2	71.2	71.2	29.9
4	0.0	2.0	2.0	2.0	2.0	2.0	2.0
	0.2	1.2	1.8	2.0	2.0	1.8	2.0
	0.3	–	2.0	2.0	2.0	1.6	2.0
	0.4	1.3	2.1	2.0	2.0	1.7	2.0
	0.5	–	2.1	2.0	1.8	2.3	2.1
	0.6	1.4	2.2	2.0	1.8	3.3	2.1
	0.7	–	2.2	2.0	1.7	5.4	2.3
	0.8	5.4	2.5	2.3	1.7	9.8	2.6
	1.0	22.9	22.9	22.9	22.9	22.9	8.1

where $\gamma = (1-p)\lambda_1 + p\lambda_2$ and $\eta = p(1-p)(\lambda_1 - \lambda_2)^2/\gamma^2$. From (11) and (24), we can calculate the IDW's for $H_2/G/1$ models to find how closely they match the IDW's of the packet-queue examples.

Figs. 6–8 show the IDW's for Examples 2–4 (estimated from simulation) and the IDW's of the $H_2/M/1$ queues selected to satisfy (22). The plotted IDW's of the $H_2/M/1$ model and the packet-queue model are very close to one another in each case. Figs. 9–11 and Table III show the normalized mean workload for each example along with that of the corresponding $H_2/M/1$ queue, as well as the other approximations to be discussed below. To show the importance of capturing dependence caused by class-dependent service times in these example, Table III also includes a QNA approximation for the normalized mean workload from [21, eqs. (29), (33)] and [14, eq. (7)]. (Note that the QNA approximation characterizes variability due to service times only through the squared coefficient of variation c_s^2 for all classes together.) The IDW and normalized mean workload values for the packet queue examples are estimated from simulation, while the values for the $H_2/M/1$ queue are calculated analytically. The 95% confidence intervals for Examples 3 and 4 were calculated in the same way as for Example 2. The relative width of the confidence intervals was again approximately 10%.

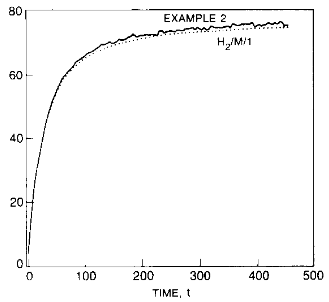


Fig. 6. The $H_2/M/1$ fit for the IDW of Example 2, based on asymptotics in Section IV-C.

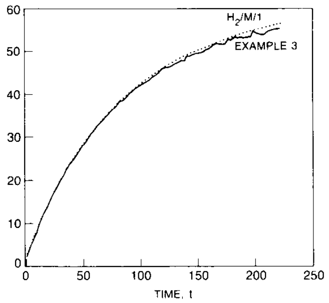


Fig. 7. The $H_2/M/1$ fit for the IDW of Example 3, based on asymptotics in Section IV-C.

For each example, we conclude that there is a strong correspondence between the curves describing the normalized workload for the packet-queue example and for the $H_2/M/1$ fit. The steady-state mean workloads for a packet queue and the corresponding $H_2/M/1$ queue have similar qualitative behavior as a function of ρ (in addition to identical light-traffic and heavy-traffic limits). For any given example, however, the relative error in an IDW-based $GI/G/1$ approximation can

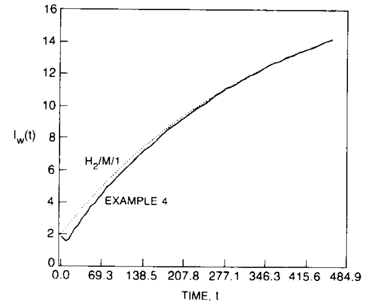


Fig. 8. The $H_2/M/1$ fit of the IDW of Example 4, based on asymptotics in Section IV-C.

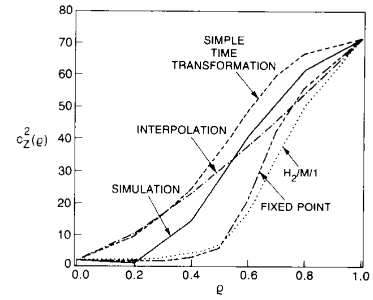


Fig. 10. A comparison of the four approximations with simulation estimates of the normalized mean workload in packet-queue Example 3.

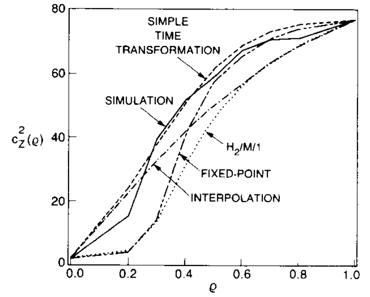


Fig. 9. A comparison of the four approximations with simulation estimates of the normalized mean workload in packet-queue Example 2.

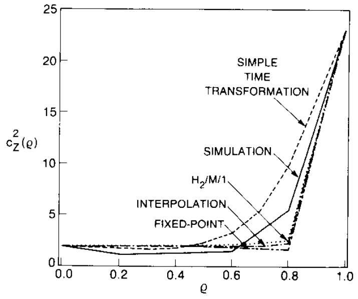


Fig. 11. A comparison of the four approximations with simulation estimates of the normalized mean workload in packet-queue Example 4.

be quite large. Indeed, Example 4 has a maximum relative error of nearly 62%, which occurs at $\rho = 0.4$. However, the value of IDW-based approximations becomes evident when we consider the wide range of behavior that results from varying the burstiness parameters (ω_i and ξ_i) for each class, while holding all other parameters fixed. While the normalized mean workload for the $GI/G/1$ approximation differs from that of Example 4 by 62% at $\rho = 0.4$, the workloads for the $GI/G/1$ model and Example 4 appear close together when compared to that of the $\sum (M^{B_i}/G_i)/1$ queue (Example 1), which shares all the parameters in Table I with Example 4 and differs only in the relative sizes of ω_i and ξ_i for each class i. The approximations successfully capture the large changes in the mean steady-state workload from example to example due to changes in the burstiness parameters ω_i and ξ_i.

The tendency of the normalized mean workload in these examples to increase suddenly at some traffic intensity from values near the light-traffic limit to values near the heavy-traffic limit make these examples difficult to approximate accurately. Even if an approximation matches the packet-queue model reasonably well overall, significant errors result if the traffic intensity at which the approximation suddenly increases does not correspond precisely to the traffic intensity at which the packet-queue model suddenly increases. Considering the difficulties inherent in approximating the workload of these examples, the $H_2/M/1$ approximations seem to do pretty well. As Table III shows, the $H_2/M/1$ fit based on the IDW is strikingly more accurate in predicting the normalized mean workload in Examples 2–4 than QNA [21], which captures only the dependence among interarrival times.

D. The Interpolation Approximation

In Section IV-C above, we fit $GI/G/1$ queues to the packet-queue examples by using the asymptotics of the IDW. We were then able to show graphically that the entire IDW's of the approximating $GI/G/1$ queues matched those of the packet queue examples, even over regions where the asymptotics does not apply. We could have generated the $GI/G/1$ approximation in the same way without having the complete IDW of the packet-queue model for comparison, but we would not have known whether the fit of the IDW was good at intermediate time values. Such an approximation, though, is not without justification. Assuming that for the packet-queue model the small- (large-) time behavior of the IDW primarily determines the light- (heavy-) traffic behavior of the mean normalized workload, such $GI/G/1$ approximations would *implicitly* interpolate between light- and heavy-traffic regions where the approximations are justified on the basis of the matching asymptotics of the IDW's.

A simple way to achieve comparable results is to apply the closed-form approximations from [16] that are obtained by *explicitly* interpolating between light- and heavy-traffic limits. These approximations use the endpoints, $c_Z^2(0)$ and $c_Z^2(1)$, as well as the derivative $\dot{c}_Z^2(0)$ and $\dot{c}_Z^2(1)$. By, [14, eqs. (8), (12)] we naturally set $c_Z^2(0) = I_w(0)$ and $c_Z^2(1) = I_w(\infty)$ in the approximations to obtain the *exact* endpoints. We obtain

TABLE IV
THE LIGHT- AND HEAVY-TRAFFIC ENDPOINTS, $c_Z^2(0)$ AND $c_Z^2(1)$, AND DERIVATIVES, $\dot{c}_Z^2(0)$ AND $\dot{c}_Z^2(1)$, FOR THE PACKET-QUEUE EXAMPLES

Example	$c_Z^2(0)$	$\dot{c}_Z^2(0)$	$c_Z^2(1)$	$\dot{c}_Z^2(1)$
1	79.9	0	79.9	0
2	1.96	−0.013	77.0	31.3
3	1.96	−0.013	71.2	90.1
4	1.96	−0.013	22.9	400.7

approximations for $\dot{c}_Z^2(0)$ and $\dot{c}_Z^2(1)$ from simple $GI/G/1$ models that we relate to the packet-queue examples through the asymptotics of their IDW's.

From [14, eqs. (104), (138)], we obtain the approximation

$$\dot{c}_Z^2(0) \approx I_w(0) I_w'(0) \tag{25}$$

From [14, eqs. (114), (116), (135)], we obtain the approximation

$$\dot{c}_Z^2(1) \approx \frac{2B}{A}. \tag{26}$$

for A and B in (18) and (19).

The function used for interpolation in [16] is of the form

$$c_Z^2(\rho) = a_0 + a_1\rho + a_2\rho^2 + a_3\rho^{b_3} + a_4(1-\rho)^{b_4}, \quad 0 \le \rho \le 1 \tag{27}$$

where the constants a_i and b_i depend on $c_Z^2(0), c_Z^2(1), \dot{c}_Z^2(0)$, and $\dot{c}_Z^2(1)$, but not ρ. To endow (27) with intuitively appealing characteristics, the interpolation recipe in [16] selects the constants in (27) to make the lowest order derivative monotone that can be made monotone, including $c_Z^2(\rho)$ (the zeroth-order derivative). For all packet-queue examples in this paper, $A > 0$ and $B > 0$. This, together with (26) implies that, for the interpolation approximation, we have $\dot{c}_Z^2(1) > 0$. Furthermore, $I_w(0) > 0$ and $I_w'(0) < 0$ for all packet-queue examples, so that, by (25), $\dot{c}_Z^2(0) < 0$ for the interpolation approximation. Clearly, it is not possible to select a monotone function $c_Z^2(\rho)$ with these asymptotic properties. Indeed, the lowest order derivative that we can make monotone is the second derivative, and we do so by applying the methods in [16]. Table IV gives the values $c_Z^2(0), \dot{c}_Z^2(0), c_Z^2(1)$, and $\dot{c}_Z^2(1)$ for Examples 1–4.

For Example 2, we have $\dot{c}_Z^2(1) \le \left[c_Z^2(1) - c_Z^2(0)\right]$, so we choose the coefficients in (27) to make $\ddot{c}_Z^2(\rho)$ a decreasing function of ρ as in [16, case 1.2.4]:

$$\begin{aligned} &a_0 = a_3 = b_3 = 0, \quad a_1 = 2c_Z^2(1) - \dot{c}_Z^2(1), \\ &a_2 = \dot{c}_Z^2(1) - c_Z^2(1), \quad a_4 = c_Z^2(0), \\ &b_4 = \left[2c_Z^2(1) - \dot{c}_Z^2(1) - \dot{c}_Z^2(0)\right]/c_Z^2(0). \end{aligned} \tag{28}$$

For Examples 3 and 4, we have $\left[c_Z^2(1) - c_Z^2(0)\right] < \dot{c}_Z^2(1), -\dot{c}_Z^2(0) \le 2c_Z^2(0)$, and $\dot{c}_Z^2(0) < 2(\dot{c}_Z^2(1) - \left[c_Z^2(1) - c_Z^2(0)\right])$, so a convex fit is achieved by choosing coefficients in (27) as in [16, case 1.2.6a]:

$$\begin{aligned} &a_0 = c_Z^2(0), \quad a_1 = \dot{c}_Z^2(0), \quad a_2 = -\frac{\dot{c}_Z^2(0)}{2}, \\ &a_3 = c_Z^2(1) - c_Z^2(0) - \frac{\dot{c}_Z^2(0)}{2}, \quad a_4 = b_4 = 0, \end{aligned}$$

$$b_3 = \frac{2\dot{c}_Z^2(1)}{2[c_Z^2(1) - c_Z^2(0)] - \dot{c}_Z^2(0)}. \tag{29}$$

Table III and Figs. 9–11 show the results of these interpolation approximations for Examples 2–4, along with the results of the $H_2/M/1$ approximations. Relative to the $H_2/M/1$ approximations, the interpolation approximations are at least as accurate, more easily applied, and more easily extended using the methods in [16], e.g., to packet queues with $I_w(\infty) < I_w(0)$. Indeed, overall, the curves from the interpolation approximation are closer to the corresponding curves from the packet-queue simulation than are the curves from the $H_2/M/1$ approximation. At first glance, this seems paradoxical because, for the $H_2/M/1$ approximation, the entire $H_2/M/1$ IDW was close to the IDW of the packet-queue examples, whereas the interpolation approximations are based only on the asymptotics of the IDW.

A possible explanation for the interpolation approximation being more accurate is as follows. By having a polynomial-like form and by satisfying the monotonicity criterion for the lowest order derivative possible, the curves from the interpolation approximation tend to rise gradually from the light-traffic limit to the heavy-traffic limit without having the sudden increases that are exhibited by the packet-queue and $H_2/M/1$ models. Thus, the interpolation approximation tends to overestimate the workload in the region before the packet queue's steep increase and underestimate the workload in the region after the increase, but its shape tends to limit the size of the maximum relative error. In other words, the interpolation approximation may serve quantitatively as a better approximation than the $H_2/M/1$ queue *because* it is qualitatively a worse approximation.

E. The Other Approximations

Finally, Table III and Figs. 9–11 also include the time-transformation approximations (7) plus (8) and (7) plus (9). These approximations have the disadvantage that they require the full IDW, but they have the advantage that they apply to an arbitrary single-server queue, i.e., they are not restricted to the packet-queue model. Of course, the first two approximations can also be applied to arbitrary single server queues, but we will not always have simple expressions in terms of model parameters. If we cannot calculate the asymptotic behavior of the IDW from model parameters, then we must estimate the asymptotic behavior of the IDW from the IDW function estimated from data.

Overall, all four approximations have comparable accuracy for Examples 2–4 and of course all four are exact for Example 1.

V. CONCLUSION

The obvious initial conclusion, which is consistent with previous work, is that packet queues can exhibit complex stochastic behavior due to the burstiness of the offered traffic. For describing standard performance measures such as the mean steady-state workload, simple models such as $M/G/1$ often are far off the mark.

In this paper we continued to develop the idea that the performance of queues with complicated offered traffic can be analyzed by focusing on the covariance structure of the offered traffic. In Section III, we presented evidence suggesting that the IDW should be better than previous indexes of dispersion for developing approximations for the mean steady-state workload. Using simulations, we showed that the IDW better matches the shape and values of the normalized mean workload for packet queue models. Four different approximations from [14] were evaluated in Section IV by making comparisons to simulations. The performance of these approximations seems good enough to justify their use for many typical engineering purposes, but the approximations achieve only a moderate degree of accuracy. As indicated in [14, Section 6.1] and Section IV-C here, matching the entire IDW very closely does not guarantee an exceptionally close approximation for the mean workload. Indeed, the light- and heavy-traffic interpolation approximation performed as well as the other more involved procedures. However, we feel that the central hypothesis of this work—that the normalized mean workload in (1) is primarily determined by the IDW in (2)—is supported by the numerical comparisons. Hopefully, even better approximations can be obtained by applying the IDW together with other model parameters.

An important direction for future research is to incorporate other relevant features into the packet queue model, such as finite buffers, windows and other flow control mechanisms. From [9, Section IV] and [28], we anticipate that such features will often significantly reduce the degradation of performance caused by dependence in the offered traffic. Nevertheless, significant dependence in the traffic seems to be an important phenomenon, which can be effectively studied using the indexes of dispersion.

Appendix
The Heavy-Traffic Limit for the Workload Process

The purpose of this Appendix is to provide additional theoretical support for the conclusion that the IDW in (2) characterizes the normalized mean workload in (1) exactly in heavy traffic, i.e., in great generality $c_Z^2(1) = I_W(\infty)$. We justify this by establishing central limit theorems for both the total input of work and the workload process. The main ideas are outlined in [14, Section IV-B1)], here we provide the details. The result here is closely related to [1, Theorem 1], which establishes a heavy-traffic limit theorem for the waiting times in a general $G/G/1$ queue. Here we show that essentially the same limit holds for the workload process. Along the way we also treat the total input of work. See [1] and references cited there for additional background.

As in [1, Section VI-A], we consider a sequence of $G/G/1$ queueing systems indexed by n (without any direct independence or stationarity assumptions). The nth queueing system is specified by the sequence of interarrival times and service times $\{(u_{nk}, \nu_{nk}): \ k \geq 1\}$ with finite long-run averages λ_n^{-1} and τ_n for each n. For each n, let $(\hat{U}_n, \hat{V}_n)$ be a random element of the function space $D \times D$ where D is the function space $D[0, \infty)$, defined by

$$[\hat{U}_n(t), \hat{V}_n(t)] = \left[n^{-1/2}[U_{n,[nt]} - nt\lambda_n^{-1}], n^{-1/2}[V_{n,[nt]} - nt\tau_n]\right], \quad t \geq 0, \tag{A.1}$$

where

$$U_{n,k} = u_{n1} + \cdots u_{nk} \quad \text{and}$$
$$V_{n,k} = \nu_{n1} + \cdots + \nu_{nk}, \quad k \geq 1,$$

and $[x]$ is the greatest integer less than or equal to x. For each n, let $\hat{A}_n$, $\hat{X}_n$, $\hat{Y}_n$, and $\hat{Z}_n$ be the random elements induced by the arrival counting process $\{A_n(t) : \ t \geq 0\}$, the total input of work process $\{X_n(t): \ t \geq 0\}$, the net input process $\{Y_n(t) : \ t \geq 0\}$ and the workload process $\{Z_n(t) : \ t \geq 0\}$ with initial workload $Z_n(0) = 0$, defined as follows:

$$\begin{aligned}\hat{A}_n(t) &= n^{-1/2}[A_n(nt) - nt\lambda_n],\\ \hat{X}_n(t) &= n^{-1/2}[X_n(nt) - nt\rho_n],\\ \hat{Y}_n(t) &= n^{-1/2}Y_n(nt),\\ \hat{Z}_n(t) &= n^{-1/2}Z_n(nt), \quad t \geq 0\end{aligned} \tag{A.2}$$

where $\rho_n = \lambda_n \tau_n < 1$; see [14, eqs. (21)–(24)] for more discussion. Let $f: \ D \rightarrow D$ be the function corresponding to the impenetrable reflecting barrier at the origin, i.e.,

$$f(x)(t) = x(t) - \inf\{x(s): \ 0 \leq s \leq t\}, \quad t \geq 0. \tag{A.3}$$

Let $e(t) = t, t \geq 0$, and let $B(t)$ be standard Brownian motion with zero drift and unit variance.

Theorem A.1: a) If $(\hat{U}_n, \hat{V}_n) \Rightarrow (\hat{U}, \hat{V})$ in $D \times D$ where $\hat{U}$ has continuous paths w.p.1, $\lambda_n \rightarrow \lambda, 0 < \lambda < \infty$, and $\tau_n \rightarrow \tau, 0 < \tau < \infty$, as $n \rightarrow \infty$, then

$$(\hat{U}_n, \hat{V}_n, \hat{A}_n, \hat{X}_n) \Rightarrow (\hat{U}, \hat{V}, \hat{A}, \hat{X}) \quad \text{in } D^4 \quad \text{as } n \rightarrow \infty$$

where

$$\hat{A} = -\lambda^{3/2}\hat{U} \quad \text{and} \quad \hat{X} = \lambda^{1/2}\hat{V} + \tau\hat{A} = \lambda^{1/2}(\hat{V} - \rho\hat{U}).$$

b) If in addition, $n^{1/2}(1 - \rho_n) \rightarrow \mu, 0 \leq \mu < \infty$, then [jointly with the other processes in a)]

$$(\hat{Y}_n, \hat{Z}_n) \Rightarrow (\hat{Y}, \hat{Z}) \quad \text{in } D^2 \quad \text{as } n \rightarrow \infty$$

where

$$\hat{Y} = \hat{X} - \mu e = \lambda^{1/2}(\hat{V} - \hat{U}) - \mu e \quad \text{and} \quad \hat{Z} = f(\hat{Y}).$$

c) If, in addition, the limit $(\hat{U}, \hat{V})$ is two-dimensional Brownian motion without drift and with covariance matrix elements $\sigma_{11}^2, \sigma_{22}^2$, and $\sigma_{12}^2 = \sigma_{21}^2$, then $\hat{U}, \hat{V}, \hat{A}$, and $\hat{X}$ are one-dimensional Brownian motions without drift, $\hat{X}$ has variance coefficient $\lambda(\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2)$, and $\hat{Z} = f(\hat{Y})$ is regulated or reflecting Brownian motion (RBM) with drift $-\mu$ and variance coefficients $\lambda(\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2)$ which has an exponential steady-state distribution with mean

$$E\hat{Z}(\infty) = \frac{\lambda(\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2)}{2\mu}. \tag{A.4}$$

Proof: To successively treat $\hat{A}_n, \hat{X}_n, \hat{Y}_n$, and $\hat{Z}_n$ in a) and b), apply [29, Theorems 7.3, 5.1, 4.1, and 6.4]. For c), first note that $\hat{V}_n - \hat{U}_n$ is one-dimensional Brownian motion with zero drift and variance $(\sigma_{11}^2+\sigma_{22}^2-2\sigma_{12}^2)$, so that $\lambda^{1/2}(\hat{V}-\hat{U})$ has zero drift and variance coefficient $\lambda(\sigma_{11}^2+\sigma_{22}^2-2\sigma_{12}^2)$. The added term μe in $\hat{Y}$ and $\hat{Z}$ gives them drift $-\mu$. ■

It is easy to connect Theorem A.1c) to the indexes of dispersion. Let $c_A^2(n), c_S^2(n)$, and $c_{AS}^2(n)$ be the asymptotic variability parameters for the nth model. For each n,

$$c_A^2(n) \approx \lambda_n^2\sigma_{11}^2, \quad c_S^2(n) \approx \tau_n^{-2}\sigma_{22}^2, \quad \text{and} \quad c_{AS}^2(n) \approx \lambda_n\tau_n^{-1}\sigma_{12}^2. \tag{A.5}$$

Since $\lambda_n \to \lambda, \tau_n \to \tau$, and $\rho_n = \lambda_n\tau_n \to 1$ as $n \to \infty$,

$$c_A^2(n) \to c_A^2 \equiv \tau^{-2}\sigma_{11}^2, \quad c_S^2(n) \to c_S^2 \equiv \tau^{-2}\sigma_{22}^2 \quad \text{and} \quad c_{AS}^2(n) \to c_{AS}^2 \equiv \tau^{-2}\sigma_{12}^2 \quad \text{as } n \to \infty. \tag{A.6}$$

To generate an approximation, let $\mu = 1$, so that n and ρ_n are related by $n^{1/2}(1-\rho_n) = 1, i.e., 1-\rho_n = n^{-1/2}$. Then the steady-state approximation becomes $\hat{Z}_n(\infty) \approx \hat{Z}(\infty)$ or

$$(1-\rho_n)^{1/2}Z_n(\infty) \approx \hat{Z}(\infty) \tag{A.7}$$

and

$$EZ_n(\infty) \approx \frac{E\hat{Z}(\infty)}{1-\rho_n} = \frac{\lambda(\sigma_{11}^2+\sigma_{22}^2-2\sigma_{12}^2)}{2(1-\rho_n)} = \frac{\tau(c_A^2+c_S^2-2c_{AS}^2)}{2(1-\rho_n)}. \tag{A.8}$$

In Theorem A.1, we have also obtained the limit for the normalized version of total input of work $\hat{X}_n$ in (A.2). Together with uniform integrability to get convergence of moments, this establishes that $I_W(\infty) = c_A^2 + c_S^2 - 2c_{AS}^2$ so that indeed $c_Z^2(1) = I_W(\infty)$ under the general assumptions of Theorem A.1. The conditions of Theorem A.1 are satisfied by the packet queue model, as discussed in [1, Sections VI-B, -C].

REFERENCES

[1] K.W. Fendick, V.R. Saksena, and W. Whitt, "Dependence in packet queues," *IEEE Trans. Commun.* vol. 37, pp. 1173–1183, 1989.

[2] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: Sequences, networks, and batches," *Adv. Appl. Prob.*, vol. 2, no. 2, pp. 355–369, 1970.

[3] H. Heffes, "A class of data traffic processes—Covariance function characterization and related queueing results," *Bell Syst. Tech. J.*, vol. 59, pp. 897–929, 1980.

[4] W. Whitt, "Approximating a point process by a renewal process, I: Two basic methods," *Oper. Res.*, vol. 30, pp. 125–147, 1982.

[5] S.L. Albin, "Approximating a point process by a renewal process, II: Superposition arrival processes to queues," *Oper. Res.*, vol. 32, pp. 1133–1162, 1984.

[6] D.Y. Burman and D.R. Smith, "Asymptotic analysis of a queueing model with bursty traffic," *Bell Syst. Tech. J.*, vol. 62, pp. 1433–1453, 1983.

[7] G.F. Newell, "Approximations for superposition arrival processes in queues," *Manage. Sci.*, vol. 30, no. 5, pp. 623–632, 1984.

[8] W. Whitt, "Queues with superposition arrival processes in heavy traffic," *Stochast. Process. Appl.*, vol. 21, pp. 81–91, 1985.

[9] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select Areas Commun.*, vol. SAC-4, pp. 833–846, 1986.

[10] H. Heffes and D.M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 856–868, 1986.

[11] I. Ide, "Superposition of interrupted Poisson processes and its application to packetized voice multiplexers," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12,* M. Bonatti, Ed. Amsterdam, The Netherlands: North-Holland, 1989, pp. 1399–1405.

[12] M.H. Rossiter, "Characterizing a random point process by a switched poisson process," Ph.D. dissertation, Dep. Math., Monash Univ., Melbourne, Australia, 1989.

[13] W. Kraemer and M. Langenbach-Belz, "Approximate formulae for the delay in the queueing system $GI/G/1$, in *Proc. Eighth Int. Teletraffic Congress,* Melbourne, Australia, 1976, pp. 235-1–235-8.

[14] K.W. Fendick and W. Whitt, "Measurements and approximations to describe the offered traffic and predict the average workload in a single server queue," *Proc. IEEE,* Y.C. Ho, Ed., vol. 77, pp. 171–194, 1989.

[15] M.I. Reiman and B. Simon, "An interpolation approximation for queueing systems with Poisson input," *Oper. Res.,* vol. 36, no. 3, pp. 454–469, 1988.

[16] W. Whitt, "An interpolation approximation for the mean workload in a $GI/G/1$ queue," *Oper. Res.,* vol. 37, no. 6, pp. 936–952, 1989.

[17] D.R. Cox and P.A.W. Lewis, *The Statistical Analysis of Series of Events.* London, England: Methuen, 1966.

[18] M.S. Bartlett, "The spectral analysis of point processes," *J. Roy. Statis. Soc.,* Ser. B, vol. 25, pp. 264–296, 1963.

[19] V. Ramaswami, "Traffic performance modeling for packet communication: Whence, where and whither," Keynote address, Third Australian Teletraff. Sem., 1989.

[20] R.W. Wolff, "Poisson arrivals see time averages," *Oper. Res.,* vol. 30, pp. 223–231, 1982.

[21] W. Whitt, "The queueing network analyzer," *Bell Syst. Tech. J.,* vol. 62, no. 9, pp. 2779–2815, 1983.

[22] M. Segal and W. Whitt, "A queueing network analyzer for manufacturing," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12,* M. Bonatti, Ed. Amsterdam, The Netherlands: North-Holland, 1989, pp. 1146–1152.

[23] R. Jain and S.A. Routhier, "Packet trains-measurments and a new model for computer network traffic," *IEEE J. Select Areas Commun.,* vol. SAC-4. pp. 986–995, 1986.

[24] A. Descloux, "Contention probabilities in packet switching networks with strung input processes," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12,* M. Bonatti, Ed. Amsterdam, The Netherlands: North-Holland, 1989, pp. 815–821.

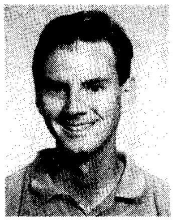
[25] S.Q. Li and J.W. Mark, "Traffic characterization for integrated services networks," *IEEE Trans. Commun.,* vol. 38, pp. 1231–1243, Aug. 1990.

[26] J.W. Cohen, *The Single Server Queue,* Amsterdam, The Netherlands: North-Holland, 1982.

[27] S.L. Brumelle, "On the relation between customer and time averages in queues," *J. Appl. Prob.,* vol. 9, pp. 508–520, 1971.

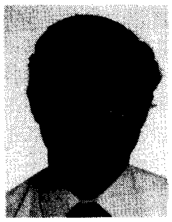
[28] K. Sriram and D.M. Lucantoni, "Traffic smoothing effects of bit dropping in a packet voice multiplexer," *IEEE Trans. Commun.,* vol. 37, pp. 703–712, 1989.

[29] W. Whitt, "Some useful functions for functional limit theorems," *Math. Oper. Res.,* vol. 5, pp. 67–85, 1980.



Kerry W. Fendick received the B.A. degree in 1982 from Colgate University, Hamilton, NY, and the M.S. degree in mathematics in 1984 from Clemson University, Clemson, SC.

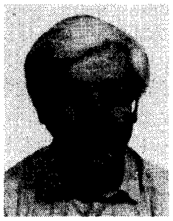
Since 1984 he has worked at AT&T Bell Laboratories, Holmdel, NJ, as a member of the Data Network Analysis Department. He primarily works on the traffic engineering and performance analysis of packet networks.



Vikram R. Saksena (S'79–M'82–SM'87) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology in 1978, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1980 and 1982, respectively.

He joined AT&T Bell Laboratories in 1982 where he worked on problems concerning traffic engineering, routing, and network design methods for AT&T's Packet Transport Network, fundamental analysis of packet network architectures, and modeling and analysis of integrated voice/data networks. He is currently the supervisor of a group responsible for projects that deal with the design and analysis of fault-tolerant data networks, design and development of traffic management systems for data networks, traffic engineering of local-area networks, wide-area networks and their interconnections, and the analysis of high-speed network architectures for multiservice integration. His overall interest is in applying systems engineering techniques to problems in modeling, analysis, and synthesis of modern communications networks.

Dr. Saksena is a Member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.



Ward Whitt received the A.B. degree in mathematics from Dartmouth College, Hanover, NH, in 1964, and the Ph.D. degree in operations research from Cornell University, Ithaca, NY, in 1969.

He taught in the Department of Operations Research at Stanford University in 1968–1969 and in the Department of Operations Research (formerly Department of Administrative Sciences) at Yale University from 1969 to 1977. Since 1977, he has been employed at AT&T Bell Laboratories. He is currently a member of the Mathematical Sciences Research Center, Murray Hill, NJ.

Dr. Whitt is a member of the Operations Research Society of America and the Institute of Mathematical Statistics.