Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue

KERRY W. FENDICK AND WARD WHITT

Invited Paper

We propose measurements and approximations to describe the variability of offered traffic to a queue (i.e., the variability of the arrival process together with the service requirements), and predict the average workload in the queue (which is assumed to have a single-server, unlimited waiting space, and a work-conserving service discipline). The principal traffic measurement considered is a normalized version of the variance of the total input of work as a function of time, which we call the index of dispersion for work (IDW). Given ample traffic data, the IDW can easily be estimated using sample averages. Given a mathematical model, such as a multiclass queue in which each class has GI/G/1 offered traffic, the IDW can often be calculated analytically, or approximated by exploiting limits as $t \to 0$ and $t \to \infty$. Our basic premise is that the average workload is primarily determined by the offered traffic, beyond the offered load (the deterministic rate work arrives), through the IDW. In this paper we provide support for this premise and indicate how the average workload can be predicted from the IDW or basic model parameters.

I. Introduction and Summary

Our purpose is to gain a better understanding of complicated queueing systems (and thus a large class of discrete event systems). We consider only one queue with a single-server, unlimited waiting space and exogenous input, but the queue might be one queue in a network. The problem is that there may be complicated offered traffic, e.g., several classes with different service requirements and non-Poisson arrival processes. To better understand the offered traffic (the arrival process together with the service requirements), and the resulting congestion in a queue to which it is offered, we propose traffic measurements and simple approximations. To a large extent, the measurements can be applied without specifying a detailed probability model. On the other hand, for a large class of models, approximations can be obtained for the principal performance measure considered here (the average workload) from basic model parameters, without measurements.

Manuscript received April 19, 1988; revised September 19, 1988.

K. W. Fendick is with AT&T Bell Laboratories, Holmdel, NJ 07733, USA.

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

IEEE Log Number 8825285.

A. Measurements to Describe the Variability of the Offered Traffic

The measurements we propose are intended to describe the *variability* of the offered traffic, because the variability of the offered traffic, appropriately defined, is a primary determinant of performance. We wish to capture the variability of individual interarrival and service times, but also the variability resulting from dependence among these variables. In the context of the growing literature on traffic measurements [28], [40], [44], our contribution is to suggest new measurements that hopefully will provide additional insight into the offered traffic and its effect on performance. We attempt to link traffic measurements more closely to performance analysis.

The main measurement we propose for the offered traffic is the *index of dispersion for work* (IDW). Let $X(t)$ be the total work in service time to enter the queue during the time interval $[0, t]$. The IDW is the function of time

$$I_w(t) = \frac{\text{var}\,[X(t)]}{\tau E[X(t)]}, \quad t \geq 0 \tag{1}$$

where τ is the average service time, var is the variance, and E is the expectation. We typically consider equilibrium conditions, under which $\{X(t): t \geq 0\}$ has stationary increments with $E[X(t)] = \rho t$, where ρ is the *offered load* or *traffic intensity*. Then the IDW is just a scaled version of var $[X(t)]$, the *variance-time curve* of $X(t)$. From ample data, the IDW is easy to estimate using sample averages; e.g., we can estimate $E[X(t)^k]$ from a sample path of the total input process by

$$\overline{m}_k(t) = n^{-1} \sum_{i=1}^{n} [X(t + s_i) - X(s_i)]^k \tag{2}$$

from n times points $0 \leq s_1 < s_2 < \cdots < s_n$ (or the associated integral) and var $[X(t)]$ by $\overline{m}_2(t) - \overline{m}_1(t)^2$. (The statistical precision of the estimate can be an important issue, but we do not discuss it here.)

When the service times are i.i.d. (independent and identically distributed) and independent of the arrival process, the IDW essentially reduces to the relatively familiar *index*

0018-9219/89/0100-0171$01.00

of dispersion for counts (IDC) of the arrival process alone; see Section III-B and (59) below. Thus, in many applications the IDW will be describing the arrival process alone. However, when the independence conditions do not hold, it can be important to consider the IDW instead of the IDC; see [24]-[26].

We suggest using the IDW and similar traffic measurements to better understand queueing networks. From the IDW we can see the variability in the offered traffic to each queue; from IDW measurements at several queues we can see how the network topology and other model features affect variability and performance.

In this paper, we provide theoretical support for IDW measurements and approximations and give a few examples. We intend to discuss actual measurements in other papers. We discuss IDW measurements and approximations for a multiclass packet queue in [25], [26]. Elsewhere, we intend to describe IDW measurements of the arrival processes at each queue of several queues in series. The IDW helps explain interesting phenomena that occur there. For example, high variability in an external arrival process may be largely removed from the arrival processes to subsequent queues after passing through a queue with low-variability service times, but may reappear at a queue several queues later that has a higher traffic intensity than the low-variability queue. From the IDWs of the offered traffic at successive queues, this phenomenon can be understood. The high variability in the external arrival process is reflected in the large-time behavior of the IDWs at *all* subsequent queues, and this large-time behavior determines the performance of any queue in heavy traffic. From the IDWs we can better understand the behavior of a fairly complex system.

It is significant that in the context of probability models the IDW can often be obtained from analytical calculations (e.g., inversion of Laplace transforms) or approximations (e.g., based on the limiting behavior as $t \to 0$ and $t \to \infty$) using basic model parameters, as well as from measurements. Thus the IDW also provides a basis for approximations without measurements. We discuss this analytical approach as well as the measurements in Section III, and illustrate it by developing approximations for a queue with a superposition arrival process, without measurements, in Section VI.

B. The Performance Measure of Interest

In this paper we consider only one performance measure—the (long-run) average workload. The workload at time t is the amount of unfinished work (in service time) in the system at time t. (The workload process is also called the virtual waiting time process.) We focus on the average workload because it is an important index of performance that is fairly robust; we contend that it is possible to develop relatively good approximations for the average workload without considering the fine structure of the model. We restrict attention to the average workload, not because it is the only performance measure worth considering (it is not), but because it seems to be a good place to start in the relatively new direction we are heading.

We actually do not focus directly on the average workload, but on a normalized version that is intended to highlight the impact of the variability in the offered traffic, separate from the measuring units and the offered load. In particular, let $E(Z_\rho)$ denote the long-run average, or mean steady-state workload as a function of the offered load ρ when the mean service time is τ. Instead of $E(Z_\rho)$, we consider the *normalized mean workload*

$$c_Z^2(\rho) = \frac{2(1-\rho)E(Z_\rho)}{\tau\rho}, \quad 0 < \rho < 1. \tag{3}$$

Even though $c_Z^2(\rho)$ is a trivial modification of $E(Z_\rho)$, we contend that $c_Z^2(\rho)$ is very useful to see the impact of the variability in the offered traffic on the average workload. The normalization in (3) is very convenient for graphical display, in the spirit of Section 5.5.2 of Allen [6] and Chapter 1 of Newell [42]. The normalized workload $c_Z^2(\rho)$ is a partial characterization of queue performance as a function of ρ, somewhat like the caudal characteristic curve introduced by Neuts [41].

As a function of τ and ρ alone, the average workload might be described by the exact M/D/1 value (assuming a Poisson arrival process and deterministic service times)

$$E(Z_\rho, \text{M/D/1}) = \frac{\tau\rho}{2(1-\rho)}. \tag{4}$$

The normalized mean workload in (3) can be interpreted as the ratio of the given average workload to what it is in the M/D/1 case, i.e., from (3) and (4),

$$c_Z^2(\rho) = \frac{E(Z_\rho)}{E(Z_\rho, \text{M/D/1})}. \tag{5}$$

(Similarly, normalization with respect to the M/M/1 queue are considered by Burman and Smith [17].) We think of the M/D/1 formula in (4) as describing the *first-order rate effect* of the offered traffic on the average workload, including the dramatic increase associated with $(1-\rho)^{-1}$ as $\rho \to 1$. We think of the normalized mean workload in (3) as describing the *second-order variability effect* of the offered traffic on the average workload. Indeed, in Section II-D we use scaling arguments to show that the normalized mean workload in (3) is invariant under a change of the deterministic rates.

Thus, the performance measure of interest is the normalized mean workload $c_Z^2(\rho)$ in (3) as a *function of* ρ. We consider $c_Z^2(\rho)$ as a function of ρ partly because we often want to consider the performance of a queue at different traffic intensities. For example, we may want to consider how much traffic to offer to the queue. Alternatively, for a given offered traffic, we may want to consider different rates at which service may be performed. We primarily think of different traffic intensities resulting from relatively simple deterministic scaling of the rate of the arrival process or the service process, but it might occur in more complicated ways. We suggest focusing on the normalized mean workload as a function of ρ, regardless of how it changes with ρ. Of course, it is important to properly represent the way the normalized mean workload does change with ρ. When there is feedback or a control mechanism, the normalized mean workload and the IDW will often change with ρ in unanticipated ways. In this paper we assume that ρ changes by the relatively simple deterministic scaling. Even if we are only interested in one system with a given traffic intensity, considering the queue as a function of ρ is useful to develop approximations based on light-traffic ($\rho \to 0$) and heavy-traffic ($\rho \to 1$) limits.

Unlike the average workload $E(Z_\rho)$, $c_Z^2(\rho)$ typically has nondegenerate limits as $\rho \to 0$, and as $\rho \to 1$. Indeed, for M/G/1 queues, the normalized mean workload is constant, in particular,

$$c_Z^2(\rho, \text{M/G/1}) = c_S^2 + 1 \tag{6}$$

where c_S^2 is the squared coefficient of variation (variance divided by the square of the mean) of a service time. (Here, it is natural to change ρ by just changing the Poisson arrival rate.) However, for other GI/G/1 queues (the interarrival times and service times come from independent sequences of i.i.d. random variables), $c_Z^2(\rho)$ changes with ρ, showing a response to the variability in the offered traffic that depends on ρ. For example, the normalized mean workloads of five GI/G/1 queues are displayed in Fig. 1. (Here, we change ρ

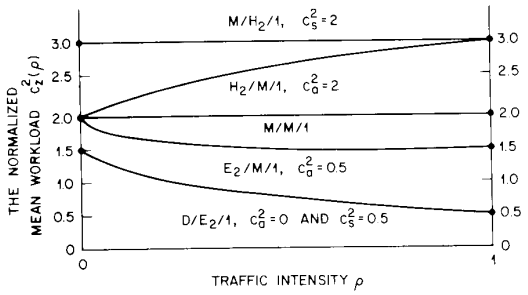


Fig. 1. The normalized mean workload $c_Z^2(\rho)$ in (3) for five GI/G/1 models (D = deterministic, E_2 = Erlang of order 2, M = exponential, H_2 = hyperexponential; c_a^2 and c_S^2 are the squared coefficients of variation of the nonexponential interarrival times and service times).

by multiplying all interarrival times by a constant.) To a large extent, the shape of the normalized mean workload curve for the GI/G/1 model is a function of the variability of the arrival process. When the arrival process is Poisson, $c_Z^2(\rho)$ is constant. When the arrival process is more variable than Poisson, as in the case of H_2 (hyperexponential, mixture of two exponentials), $c_Z^2(\rho)$ tends to increase; when the arrival process is less variable than Poisson, as in the case of E_2 (Erlang of order two, the convolution of two exponentials) or D (deterministic), $c_Z^2(\rho)$ tends to decrease. Indeed, a rough two-moment approximation for $c_Z^2(\rho)$ in GI/G/1 queues, which we develop in Section IV-A1, is

$$c_Z^2(\rho, \text{GI/G/1}) \approx c_S^2 + 1 + \rho(c_a^2 - 1) \tag{7}$$

where c_a^2 is the squared coefficient of variation of an interarrival time.

The curves in Fig. 1 are relatively tame, showing a total variation of at most 1. More interesting normalized mean workloads occur in multiclass packet queues with bursty arrival processes. Fig. 2 displays normalized mean workloads for two packet queue models based on simulation estimates from Fendick *et al.* Here, in each case there is a total variation of 40. Moreover, the behavior is not due to the arrival process alone. A similar normalized mean workload curve for a statistical multiplexer queue can be constructed from Table III of [51]. (See these references for model description.) Fig. 2 shows a dramatic increase in the average workload, relative to the M/D/1 model, as the traffic intensity increases. Moreover, in contrast to (7), the curves

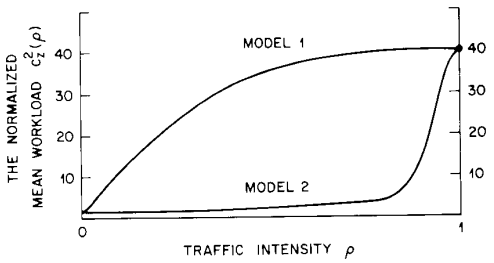


Fig. 2. The normalized mean workload $c_Z^2(\rho)$ in (3) for two multiclass packet queue models with bursty arrival processes, based on simulation estimates and heavy-traffic and light-traffic limits, from [25], [26].

in Fig. 2 are markedly nonlinear. Many different shapes are possible.

In order to predict how offered traffic will affect the average workload in a queue, it is natural to estimate the normalized mean workload directly via simulation (using either historical or randomly generated data), and indeed we suggest this approach in Section II-C. However, we also suggest obtaining information about the normalized mean workload indirectly from the IDW, because the IDW is often much easier to work with, especially with multiclass queues. We also consider the IDW in addition to the normalized mean workload because we want to better *understand* how the offered traffic affects the normalized mean workload.

C. A Special Relationship: The IDW and the Normalized Mean Workload

As a *general idea*, we suggest using traffic measurements to gain insight into the variability of offered traffic, and the performance of a queue to which this traffic is offered. To support this general idea, we primarily consider only *one* traffic measurement, the IDW in (1), and only *one* performance measure, the normalized mean workload $c_Z^2(\rho)$ in (3). As a *specific idea*, we suggest using the IDW to gain insight into the normalized mean workload $c_Z^2(\rho)$. The IDW may be useful more generally, but we contend that it is especially appropriate for predicting the normalized mean workload.

We contend that $c_Z^2(\rho)$ is primarily determined by the IDW, so that it should be possible to predict $c_Z^2(\rho)$ reasonably well, given ρ and $\{I_w(t): t \geq 0\}$. Assuming that $\{c_Z^2(\rho): 0 \leq \rho \leq 1\}$ is obtained by deterministic scaling of the interarrival or service times, we contend that the two curves $\{c_Z^2(\rho): 0 \leq \rho \leq 1\}$ and $\{I_w(t): 0 \leq t \leq \infty\}$ are very similar. For example, for two different queues, if one IDW is always higher than the other, then we predict that the normalized mean workloads are similarly ordered. If the two IDWs are nearly the same, then we predict that the normalized mean workloads are nearly the same. Moreover, the actual values should be nearly the same. (These suggested relations are *approximations*. In Section V-C we construct several models with *identical* IDWs, but *different* normalized mean workloads. Nevertheless, approximating the normalized mean workload of one of these models by the normalized mean workload in another of these models is reasonable.)

At this point it seems appropriate to explain why we might expect the two functions $\{c_Z^2(\rho): 0 \leq \rho \leq 1\}$ and $\{I_w(t): 0 \leq t \leq \infty\}$ to be related. For us, the starting point was the *heavy-*

traffic diffusion process limit (discussed here in Section IV)

$$\lim_{\rho \to 1} c_\lambda^2(\rho) \equiv c_\lambda^2(1) = I_w(\infty) \equiv \lim_{t \to \infty} I_w(t) \tag{8}$$

which follows from [24], [25], [37], [52], or [42]; i.e., it is known that $c_\lambda^2(1) = I_w(\infty)$; the heavy-traffic limit $c_\lambda^2(1)$ depends on the total input process of work *only* via the IDW, and that in this limit *all* the correlations play a role. However, it is apparent that as ρ decreases from 1, distant correlations in the total input process of work should have less impact on $c_\lambda^2(\rho)$. Roughly speaking, $I_w(t)$ for a given t has an impact upon $c_\lambda^2(\rho)$ for a given ρ only if fluctuations in the offered traffic at time s have an impact upon the workload at time $s + t$ at that ρ. If the workload process hits zero before time $s + t$, then this impact tends to be gone, but otherwise not. (This assertion is rigorously justified when the zero state is a regeneration point for the model; otherwise, it is an approximation.) In other words, correlations within busy periods should be more significant than correlations across busy periods, and higher ρ means longer busy periods.

An approximation for $c_\lambda^2(\rho)$ suggested by this heuristic analysis and (8) is a deterministic time transformation of $I_w(t)$, i.e.,

$$c_\lambda^2(\rho) \approx I_w(t(\rho)), \quad 0 \le \rho \le 1 \tag{9}$$

where $t(\rho)$ is strictly increasing, with $t(0) = 0$ and $t(1) = \infty$.

More precisely, we think of $c_\lambda^2(\rho)$ as being approximately a mixture of $I_w(t)$ determined by a mixing cdf (cumulative distribution function) $G_{w\rho}(t)$ that depends on *both* the IDW and ρ, i.e.,

$$c_\lambda^2(\rho) \approx \int_0^\infty I_w(t)\, dG_{w\rho}(t), \quad 0 \le \rho \le 1 \tag{10}$$

where $G_{w\rho}(t)$ is stochastically increasing with ρ, i.e., for any nondecreasing real-valued function h,

$$\int_0^\infty h(t)\, dG_{w\rho_1}(t) \le \int_0^\infty h(t)\, dG_{w\rho_2}(t) \tag{11}$$

whenever $\rho_1 < \rho_2$. Equation (9) is obtained from (10) by making the simplifying assumptions, first, that $G_{w\rho}$ is independent of w (the IDW) and, second, that $G_{w\rho}$ corresponds to a unit point mass at the point $t(\rho)$.

The connection between $c_\lambda^2(\rho)$ and $I_w(t)$ and approximations (9) and (10), are further supported by a *light-traffic limit*

$$\lim_{\rho \to 0} c_\lambda^2(\rho) \equiv c_\lambda^2(0) = I_w(0) \equiv \lim_{t \to 0} I_w(t) \tag{12}$$

which is also established in Section IV. Further insight can be gained from the M/G/1 queue, for which $I_w(t) = c_\lambda^2(\rho) = c_s^2 + 1$ for *all* t and ρ; see (6) and (60). More generally, using the IDW to approximate the normalized mean workload can be regarded as a consequence of a Lévy process approximation for the net input process of work; see Section IV-C. This perspective is helpful because it does not depend on heavy-traffic or light-traffic limits.

D. Approximations

The most elementary approximation for $c_\lambda^2(\rho)$ given $I_w(t)$ is (9). One possibility for the time transformation, assuming that the mean service time is 1, is

$$t(\rho) = \frac{\rho I_w(\infty)}{2(1-\rho)^2}, \quad 0 \le \rho \le 1 \tag{13}$$

yielding our first candidate approximation

$$c_\lambda^2(\rho) \approx I_w(\rho I_w(\infty)/2(1-\rho)^2), \quad 0 \le \rho \le 1. \tag{14}$$

The exponent 2 in the denominator of (13) is motivated by established heavy-traffic behavior of queues with superposition arrival processes; see [43] and [59]. In Section III-E, we show that if the offered traffic is the superposition from n i.i.d. sources, then the IDW for the superposition of n processes, say $I_{wn}(t)$, satisfies $I_{wn}(t) = I_{w1}(t/n)$, so that (13) leads to $c_\lambda^2(\rho) \approx I_{w1}(\rho I_{w1}(\infty)/2n(1-\rho)^2)$, which has the correct asymptotic behavior as $\rho \to 1$ and $n \to \infty$; i.e., $c_\lambda^2(\rho) \to I_{w1}(\infty)$ if $n(1-\rho)^2 \to 0$, and $c_\lambda^2(\rho) \to I_w(0)$ if $n(1-\rho)^2 \to \infty$. (See Section VI for examples with large n.) The time transformation (13) also constitutes an approximation for the *relaxation time* in the queue; e.g., see pp. 133 and 151 of [42], [1] and [2].

A comparison between the exact values and $I_w(t(\rho))$ in (9) for $t(\rho) = \rho/(1-\rho)$, $\rho/(1-\rho)^2$, and $\rho I_w(\infty)/2(1-\rho)^2$ for the $H_2/M/1$ and $E_2/M/1$ models appears in Table 1. (In this case

Table 1 A Comparison of the Exact Normalized Mean Workload $c_\lambda^2(\rho)$ in (3) with Three Simple Time-Scaled IDWs, $I_w(t(\rho))$ for $t(\rho) = \rho/(1-\rho)$, $\rho/(1-\rho)^2$ and $\rho I_w(\infty)/2(1-\rho)^2$ as in (13), and the Light-Traffic and Heavy-Traffic Interpolation in (17) and (18). The H_2 Distribution has Balanced Means, so that the Third Moment is $m_{33} = 18.0$.

Traffic Intensity ρ	$H_2/M/1$, $c_a^2 = 2.0$					$E_2/M/1$, $c_a^2 = 0.5$				
		$I_w(t(\rho))$ for $t(\rho) =$			Interpolation		$I_w(t(\rho))$ for $t(\rho) =$			Interpolation
	$c_\lambda^2(\rho)$	$\rho/(1-\rho)$	$\rho/(1-\rho)^2$	$\rho I_w(\infty)/2(1-\rho)^2$	(17) and (18)	$c_\lambda^2(\rho)$	$\rho/(1-\rho)$	$\rho/(1-\rho)^2$	$\rho I_w(\infty)/2(1-\rho)^2$	(17) and (18)
0.0	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
0.1	2.07	2.04	2.04	2.06	2.07	1.85	1.90	1.89	1.92	1.84
0.2	2.15	2.08	2.10	2.14	2.16	1.77	1.82	1.79	1.82	1.73
0.3	2.24	2.13	2.18	2.25	2.25	1.70	1.74	1.69	1.73	1.66
0.4	2.34	2.19	2.29	2.40	2.35	1.66	1.67	1.61	1.64	1.62
0.5	2.45	2.27	2.45	2.57	2.46	1.62	1.62	1.56	1.58	1.59
0.6	2.56	2.37	2.66	2.74	2.57	1.59	1.58	1.53	1.54	1.57
0.7	2.68	2.49	2.81	2.87	2.68	1.56	1.55	1.52	1.52	1.55
0.8	2.78	2.65	2.93	2.95	2.79	1.54	1.53	1.51	1.51	1.53
0.9	2.90	2.83	2.98	2.99	2.90	1.52	1.51	1.50	1.50	1.52
1.0	3.00	3.00	3.00	3.00	3.00	1.50	1.50	1.50	1.50	1.50
max (over ρ) % error	—	7.5%	5.4%	7.1%	0.4%	—	2.8%	3.5%	3.8%	2.2%

the H_2 distributions have balanced means: if λ_i is the exponential rate and p_i the probability of component i, then $p_1/\lambda_1 = p_2/\lambda_2$. The exact IDW formulas are given in Section III-G.) Also appearing in Table 1 is another approximation discussed below. The advantages of having $I_w(\infty)$ in (13) is not evident from Table 1; it is evident when $I_w(\infty)$ is much higher, e.g., when it is 77, as in the packet queue examples in [26].

In this paper we restrict attention to deterministic time-transformation approximations of the form (9), but we recognize that the appropriate time transformation $t(\rho)$ should ideally depend on the IDW. Thus, in Section V we develop what we call a *variability-fixed-point approximation* to obtain $t(\rho)$ from $I_w(t)$. To a large extent, this approach is motivated by the work of Newell, pp. 132–151 of [42] and [43], although the specific approximation here seems to be new.

In particular, we propose approximation (9) where $t(\rho)$ is determined by the fixed point equation

$$t(\rho) = x(\rho) I_w(t(\rho)), \quad 0 \le \rho \le 1 \tag{15}$$

where $x(\rho)$ is an increasing function of ρ, such as

$$x(\rho) = \rho/(1 - \rho). \tag{16}$$

Just as it is not clear that $t(\rho)$ in (13) is best for (9), here it is not clear that $x(\rho) = \rho/(1 - \rho)$ in (16) is best in (15). Heuristic arguments in Sections V-A and V-B also suggest $x(\rho) = \rho/2(1 - \rho)^2$, and $x(\rho) = \rho/(1 - \rho^2)$, respectively, but (16) often seems reasonable. In any case, with (15), $t(\rho)$ *and* $I_w(t(\rho))$ are obtained for each ρ by finding the intersection of the line $t/x(\rho)$ with the function $I_w(t)$, as illustrated in Fig. 3.

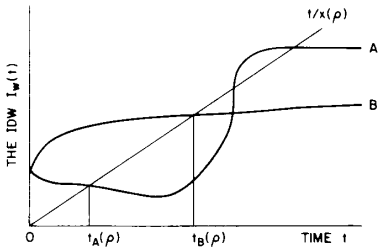


Fig. 3. The variability-fixed-point approximation applied to two different IDWs: $c_Z^2(\rho) = I_w(t(\rho))$ as in (9) where $t(\rho)$ satisfies (15) with $x(\rho)$ in (16).

In Fig. 3 the variability fixed-point approximation is applied to two different IDW curves at the same ρ. Since IDW A has lower variability for small t than IDW B, we contend that it is appropriate to have the fixed points ordered by $t_A(\rho) < t_B(\rho)$ for all sufficiently small ρ. In Fig. 3, we see that IDW A intersects the line $t/x(\rho)$ *three times*, so that (15) can have multiple solutions. We avoid ambiguity by always selecting the minimal solution, but there is a fundamental difficulty when (15) has multiple solutions; then $t(\rho)$ and $I_w(t(\rho))$ are *discontinuous* in ρ, whereas $c_Z^2(\rho)$ should be continuous. The difficulty stems from approximating the mixing cdf $G_{w\rho}(t)$ in (11) to be a unit point mass. We could refine the approximation by smoothing the curve $I_w(t(\rho))$ after applying (13), but here we just observe that multiple solutions to (15) signals a problem with the fixed-point approximation. In most applications, (15) should have a unique solution, at least for when ρ is not too small.

We also apply the fixed-point framework to develop approximations for the derivatives $\dot{c}_Z^2(0)$ and $\dot{c}_Z^2(1)$ to use with the end values in (8) and (12) in a light-traffic and heavy-traffic interpolation. A specific interpolation approximation for $c_Z^2(\rho)$ given the four values $c_Z^2(0)$, $\dot{c}_Z^2(0)$, $c_Z^2(1)$, and $\dot{c}_Z^2(1)$ is developed in [60]. The interpolation approximation has the form

$$c_Z^2(\rho) \approx a_0 + a_1\rho + a_2\rho^2 + a_3\rho^{b_3} + a_4(1 - \rho)^{b_4} \tag{17}$$

for $0 \le \rho \le 1$, where the coefficients a_i and exponents b_i depend on the four limits $c_Z^2(0)$, $\dot{c}_Z^2(0)$, $c_Z^2(1)$, and $\dot{c}_Z^2(1)$, but not ρ. This interpolation obviously avoids the discontinuity problem above. The parameters a_i and b_i are chosen so that the approximation and its first derivative are monotone whenever possible, and so that the second derivative is monotone whenever the first derivative cannot be monotone.

We express the derivatives $\dot{c}_Z^2(0)$ and $\dot{c}_Z^2(1)$ in terms of the derivatives $J'_w(0)$ and $J'_w(1)$ of $J_w(\rho) = I_w(\rho/(1 - \rho))$, which in turn are often available analytically from the asymptotic behavior of $I_w(t)$. From (15) and additional asymptotics (Section V), we obtain the approximations

$$\dot{c}_Z^2(0) \approx J_w(0)J'_w(0) \quad \text{and} \quad \dot{c}_Z^2(1) \approx \frac{2J'_w(1)}{\max\{1, J_w(1)\}} \tag{18}$$

where $J_w(1) = I_w(\infty) = c_Z^2(1)$, and $J_w(0) = I_w(0) = c_Z^2(0)$ by (8) and (12). In Section V-B we show that it is convenient to work with $J_w(\rho)$ in (18) instead of $I_w(t)$.

Equation (18) is the basis for an *approximate* light-traffic and heavy-traffic interpolation, given only the asymptotic behavior of $J_w(\rho)$ as $\rho \to 0$ and as $\rho \to 1$. In support of (18), we show that the approximation for $\dot{c}_Z^2(0)$ is correct for the GI/G/1 queue, and the approximation for $\dot{c}_Z^2(1)$ is correct for the GI/M/1 queue. Comparisons of the interpolation approximation based on (17) and (18) with exact values for the H_2/M/1 and E_2/M/1 models appear in Table 1. For H_2/M/1, (17) and (18) offer a substantial improvement. Indeed, all the approximations do remarkably well for these examples. Unfortunately, the approximations are not always nearly so good (e.g., see Section VI), but overall, our numerical experience is encouraging. On the basis of performance and simplicity, the leading candidates are the simple time transformation (9) with (13), and the interpolation approximation (17) and (18).

We became interested in new ways to approximate $E(Z_\rho)$ and $c_Z^2(\rho)$ because of difficult models, such as those having the normalized mean workloads displayed in Fig. 2. However, every model need not be so difficult. For example, we could have a queue with complicated dependence among the interarrival times and service times, but a measured IDW, as in Fig. 4. Our results lead us to predict (probably quite accurately) from such a relatively level IDW that the normalized mean workload in this case will be approximately the same as in an M/G/1 queue, with $c_s^2 + 1 = 5$; i.e., in some applications, the behavior will be better than we might expect.

E. Scaling

The functions $c_Z^2(\rho)$ and $I_w(t)$ are intended to characterize the variability of the offered traffic. It is thus natural to ask what properties such a variability measure should have. One property that seems very natural is *scale invariance*, i.e., an

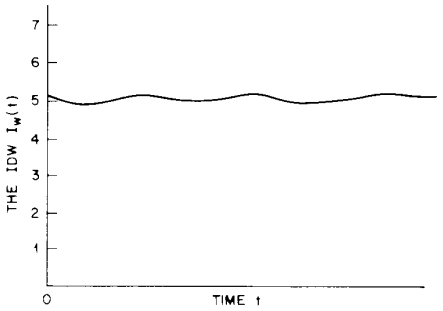


Fig. 4. A relatively level IDW, suggesting the approximation $c_Z^2(\rho) \approx 5$, $0 \le \rho \le 1$.

invariance to changes in deterministic linear rates. We develop this important idea further in Sections II-D and III-D.

F. Related Literature

This paper is a natural extension of previous work on approximations for point processes, and the performance of queues at which they serve as arrival processes, e.g., [3]–[5], [22], [34], [35], [42], [43], [51], [54]–[57], and [59]. There are two major departures from this earlier work: First, we place a greater emphasis on measurements. Second, we treat the arrival process together with the service times. Treating the arrival process together with the service times is important when the interarrival times and service times are mutually dependent. As shown in [24]–[26], such dependence can be a major factor in multiclass queues with non-Poisson arrival processes and class-dependent service-time distributions. Moreover, such multiclass queues frequently arise in models of communication networks and manufacturing systems.

For such multiclass queues, heavy-traffic limit theorems have been established in [11], [37], and [52] that yield relatively simple approximations for the mean steady-state workload, and other performance measures. Extensions to networks of queues with feedback have been established by Reiman [48], Johnson [38], Peterson [45], and Harrison [33]. Unfortunately, however, for the models motivating this work (e.g., [24]–[26], and [51]), these heavy-traffic approximations are not accurate at typical design loads, so it is desirable to develop refinements. (At typical design loads, the heavy-traffic approximation can exceed the actual mean workload by a factor of ten or more; see Fig. 2 and [51].) This paper provides refinements via the IDW, and by focusing on light-traffic and heavy-traffic asymptotics, in the spirit of [5], [17], [49], [50], [55], [59], and [60]. In the context of Fig. 2, we want to do better than a simple linear interpolation between $c_Z^2(0)$ and $c_Z^2(1)$, and thus distinguish between the two models.

G. The Rest of This Paper

In Section II, we develop a general framework for studying the average workload and its offered traffic. In Section III, we discuss ways to describe the variability of point processes and marked point processes, providing background and motivation for the IDW. In Section IV, we relate the IDW to the normalized mean workload, and describe the light-traffic and heavy-traffic behavior of both. In Section V, we discuss approximations for the normalized mean workload based on the IDW or basic model parameters. In particular, we develop (9)–(18) there, and discuss a GI/G/1 model approximation. In Section VI, we give an example illustrating the four approximations in (9) plus (13), (9) plus (15), (17) plus (18), and the GI/G/1 model approximation. Finally, we draw conclusions in Section VII.

II. A General Framework for the Workload

In this section we define three fundamental features of our queueing model: the *offered traffic*, the *workload process*, and the *normalized mean steady-state workload*, all of which we regard as functions of the traffic intensity ρ. We indicate how to efficiently estimate the normalized mean workload as a function of ρ from data. We also discuss how estimates of the normalized mean workload can be used to statistically characterize offered traffic. Finally, we discuss alternative scalings that can be used to produce the offered traffic and the workload as functions of ρ. There are three natural places to do the scaling—the arrival rate, the mean service requirement, or the rate service is performed—but we show that all three are essentially equivalent, so that it suffices to consider any one.

A. The Offered Traffic and the Workload Process

We assume that service is provided at unit rate whenever there is work to be done, according to some work-conserving discipline, e.g., FIFO (first-in first-out), LIFO (last-in first-out), or multiclass with nonpreemptive priorities; see p. 418 of [36]. Since we focus on the workload, the particular work-conserving discipline does not matter. This is one strong sense in which the workload is a robust description of performance.

The stochastic behavior of the workload process is completely determined by the offered traffic and the initial conditions. (Service is provided by a deterministic mechanism.) We model the *offered traffic* as a one-parameter family of sequences of nonnegative random variables $\{\{(\rho^{-1}u_n, v_n): n \ge 1\}, 0 < \rho < 1\}$, where $\rho^{-1}u_n$ represents the interarrival time between the $(n-1)$th and nth customer, and v_n the service time requirement of the nth customer. (The sequence $\{(\rho^{-1}u_n, v_n)\}$ plus the service discipline generates the sequence of discrete events—arrivals and departures—in the discrete event system.) We describe the steady-state or long-run average behavior, so we assume that the offered traffic is a stationary and ergodic sequence for each ρ with

$$E(u_n) = E(v_n) = 1 \tag{19}$$

and

$$E(u_n^2) = c_a^2 + 1 < \infty \quad \text{and} \quad E(v_n^2) = c_s^2 + 1 < \infty. \tag{20}$$

Without loss of generality (by choosing the measuring units for time), in (19) we let the mean service time be 1, so that the *arrival rate* (the reciprocal of the mean interarrival time) coincides with the *offered load* or *traffic intensity* ρ. We introduce the parameter ρ in order to be able to consider the workload as a function of ρ. Our definition of the offered traffic introduces ρ as a scaling of the arrival process. It might seem more natural to scale the service requirements (i.e., to consider $\{(u_n, \rho v_n): n \ge 1\}$), or to scale the rate at which

service is performed. In Section II-D, we show that all three scalings can be obtained from any one.

Our model of the offered traffic is a family of *random marked point processes* (RMPPs) as defined by Franken *et al.* [27]. (The arrival process is the basic point process, and the service times are the marks.) Our definition is the *synchronous version*, as defined on p. 30 of [27]. We have stipulated that the offered traffic be stationary and ergodic, but we also have in mind applications in which the offered traffic is only asymptotically stationary, or is even fundamentally nonstationary. Nevertheless, the stationary and ergodic case is a useful framework. From the RMPP point of view, we are interested in measurements and approximations that an engineer can use to describe an RMPP, and the congestion that occurs when it serves as the input to a queue. So far, the RMPP literature has been concerned primarily with general structural issues, such as existence, uniqueness, continuity, and insensitivity of stationary distributions, and the relation between time-stationary and customer-stationary distributions.

We now define other associated processes, leading up to the workload process itself. For each ρ, let $\boldsymbol{A}_\rho \equiv \{A_\rho(t): t \geq 0\}$ be the *arrival counting process* (point process), defined by $A_\rho(t) = A(\rho t)$ where

$$A(t) = \sup \{n \geq 1: u_1 + \cdots + u_n \leq t\} \text{ for } t \geq u_1 \geq 0 \tag{21}$$

and $A(t) = 0$ for $u_1 > t \geq 0$; let $\boldsymbol{X}_\rho \equiv \{X_\rho(t): t \geq 0\}$ be the *total input* process and $\boldsymbol{Y}_\rho \equiv \{Y_\rho(t): t \geq 0\}$ the *net input process*, defined by

$$X_\rho(t) = \sum_{i=1}^{A_\rho(t)} v_i \quad \text{and} \quad Y_\rho(t) = X_\rho(t) - t, \quad t \geq 0. \tag{22}$$

Note that $Y_\rho(t)$ represents the total input minus the total *potential* output up to time t. Also note that $X_\rho(t) = X_1(\rho t)$.

Remark (2.1): Let U_n and V_n be the partial sums

$$U_n = u_1 + \cdots + u_n \quad \text{and} \quad V_n = v_1 + \cdots + v_n, \quad n \geq 1. \tag{23}$$

Since the offered traffic is stationary and ergodic, and (19) holds, $n^{-1}U_n \to 1$, and $n^{-1}V_n \to 1$ w.p.1 (with probability 1) as $n \to \infty$. As a further consequence, $t^{-1}A_\rho(t) \to \rho$ and $t^{-1}X_\rho(t) \to \rho$ w.p.1 as $t \to \infty$. □

As a simplifying assumption we assume that the system starts out empty. The *workload process* (as a function of ρ) $\boldsymbol{Z}_\rho \equiv \{Z_\rho(t): t \geq 0\}$ can then be defined by

$$\begin{aligned} Z_\rho(t) &= f(Y_\rho)(t) = Y_\rho(t) - \inf_{0 \leq s \leq t} Y_\rho(s) \\ &= \sup_{0 \leq s \leq t} \{Y_\rho(t) - Y_\rho(s)\}, \quad t \geq 0. \end{aligned} \tag{24}$$

Note that the workload process $\boldsymbol{Z}_\rho$ depends on the offered traffic only through the net input process $\boldsymbol{Y}_\rho$, which in turn is a minor modification of the total input process $\boldsymbol{X}_\rho$. The mapping f taking $\boldsymbol{Y}_\rho$ into $\boldsymbol{Z}_\rho$ in (24) corresponds to the imposition of an impenetrable reflecting barrier at the origin. Equation (24) can be understood by identifying a sample path of $\boldsymbol{Z}_\rho$ with a sample path of $\boldsymbol{Y}_\rho$ modified by moving the origin for $\boldsymbol{Z}_\rho$ when $\boldsymbol{Y}_\rho$ becomes negative; at time t the origin is at $\inf_{0 \leq s \leq t} Y_\rho(s)$. It is easy to see that the total time in $[0, t]$ during which the server is idle is

$$I_\rho(t) = -\inf_{0 \leq s \leq t} Y_\rho(s), \quad t \geq 0 \tag{25}$$

and

$$Z_\rho(t) = Y_\rho(t) + I_\rho(t), \quad t \geq 0. \tag{26}$$

Equations (24)–(26) can be proved from "first principles" by considering the successive jump points of $\boldsymbol{X}_\rho$, $\boldsymbol{Y}_\rho$, and $\boldsymbol{Z}_\rho$, and applying mathematical induction.

It is significant that the representations (24)–(26) follow from sample-path properties, without imposing any probabilistic assumptions, such as independence or stationarity. This is a second strong sense in which the workload is a robust description of performance. Beneš [8] first explored the possibility of a general treatment of the workload process, without assuming the Markov property, including derivations of (24)–(26) from first principles. This paper is intended to contribute to the long-range goal described in Chapter 1 of [8] of obtaining useful engineering curves, tables, etc., in such a general framework.

We conclude this section by making a simplifying assumption. We assume that the arrivals occur one at a time, i.e., $P(u_n = 0) = 0$. For studying the workload process, this assumption is without loss of generality, because if batch arrivals occur in the original offered traffic, then we can remove them by considering the arrival process for batches, letting the batch service requirement be the sum of the service requirements of all customers in the batch. Obviously, the total input process and the workload process are unchanged by this modification.

B. The Normalized Mean Workload

We assume that the workload process $\boldsymbol{Z}_\rho \equiv \{Z_\rho(t): t \geq 0\}$ has a steady-state random value Z_ρ for each ρ, $0 < \rho < 1$, which may be obtained by having $Z_\rho(t) \Rightarrow Z_\rho$ as $t \to \infty$, where $\Rightarrow$ denotes convergence in distribution, or by assuming that $\{Z_\rho(t): t \geq 0\}$ has a stationary version. (General sufficient conditions can be found in Chapter 5 of [8], Section 6 of [12] and Chapter 2 of [27].) Our object is to describe the long-run average workload, which in this framework coincides with the mean (steady-state) workload $E(Z_\rho)$. However, instead of $E(Z_\rho)$ itself, we consider the *normalized mean workload*, defined by (3).

We now describe an efficient algorithm to estimate $E(Z_\rho)$, and thus $c_z^2(\rho)$, as a function of ρ from a segment of a sample path of $\{(u_n, v_n): n \geq 1\}$ for a single value of ρ, which we take to be 1. These estimates from simulations can be used to test the approximations.

A natural estimator for $E(Z_\rho)$ based on n arrivals is

$$\bar{Z}_\rho(n) = (\rho^{-1}U_n)^{-1} J_\rho(\rho^{-1}U_n) \tag{27}$$

where

$$J_\rho(\rho^{-1}U_n) = \int_0^{\rho^{-1}U_n} Z_\rho(t)\, dt. \tag{28}$$

A recursive algorithm to compute $J_\rho(\rho^{-1}U_n)$ is

$$J_\rho(\rho^{-1}U_{n+1}) = J_\rho(\rho^{-1}U_n) + \Delta_{\rho n} \tag{29}$$

where

$$\Delta_{\rho n} = \begin{cases} (2\rho)^{-1} u_{n+1}[Z_\rho(\rho^{-1}U_n) + Z_\rho(\rho^{-1}U_{n+1} -)], \\ \quad Z_\rho(\rho^{-1}U_{n+1} -) > 0 \\ (2\rho)^{-1} Z_\rho(\rho^{-1}U_n)^2, \\ \quad Z_\rho(\rho^{-1}U_{n+1} -) = 0, \end{cases} \tag{30}$$

$$Z_\rho(\rho^{-1}U_{n-1}-) - \max\{0, Z_\rho(\rho^{-1}U_{n-1}-) + v_n - \rho^{-1}u_{n-1}\}, \quad (31)$$

and

$$Z_\rho(\rho^{-1}U_n) = Z_\rho(\rho^{-1}U_n-) + v_n, \quad n \geq 1.$$

Note that (31) coincides with the familiar Lindley recursion associated with the sequence of successive waiting times in a single-server queue with the FIFO discipline. Also note that *the recursion (29)-(31) can be conveniently carried out for several values of ρ simultaneously using one sample path of the offered traffic*. The basis for efficiency is, first, to treat all cases of ρ from one sequence of numbers and, second, to perform the calculations recursively at the arrival epochs.

In this paper we focus on the average workload, but it is often useful to consider other performance measures. It is significant that the recursion above is easily modified to compute other averages; the average of $f(Z_\rho(t))$ can be estimated by (27) where $f(Z_\rho(t))$ replaces $Z_\rho(t)$ in (28). When $f(x) = I([x, \infty))$, the indicator function of the interval $[x, \infty)$, (27) estimates the average proportion of time $Z_\rho(t)$ is above level x, which usually coincides with the complementary distribution function of the steady-state workload. We can treat this case by redefining $\Delta_{\rho n}$ in (30)

$$\Delta_{\rho n}(x) = \begin{cases} \rho^{-1}u_{n-1}, & Z_\rho(\rho^{-1}U_{n-1}-) > x \\ 0, & Z_\rho(\rho^{-1}U_n) < x \\ Z_\rho(\rho^{-1}U_n) - x, & Z_\rho(\rho^{-1}U_n) > x > Z_\rho(\rho^{-1}U_{n-1}-). \end{cases} \quad (32)$$

C. The Workload Characterization of a Marked Point Process

One purpose of this paper is to develop appropriate *measurements* of RMPPs (which may arise as offered traffic to queues, but also may arise in other contexts). The goal is to determine what data to collect and what statistics to compute in order to estimate or test for basic structural properties, such as the nature of the variability in the RMPP. This statistical analysis in turn can be used to predict how the offered traffic will affect the performance of a queue to which it is offered.

In [54], [56], we suggested using actual performance measures estimated from a test queue to approximately characterize a point process. The point process of interest might arise in some non-queueing context or be the arrival process at some different queue. The idea is to simulate or analytically describe the test queue with some specified service mechanism, such as a single exponential server, using the point process of interest as the arrival process. The resulting performance measures of the test queue then partially characterize and describe the point process that is used as the arrival process. (The idea of using a test queue to characterize a point process originated with the notion of peakedness associated with the equivalent random method in teletraffic theory; see Section 4.7 of [18], [22] and [61]. The notion of peakedness is itself very close to the IDW, both being second-order properties; see [22].) To obtain a rich description of the point process, the performance measures can be expressed as functions of the service rate or even the service distribution. Simulations for a whole set of service rates can be obtained from one set of random numbers by simultaneously simulating the different queues using the same data, multiplying each randomly generated service time by the appropriate constant in each queue. Arbitrary service-time distributions can be obtained from any distribution with continuous positive density by first transforming to a uniform distribution with the given cdf, and then transforming to the desired distribution using the inverse cdf. Given a random variable X with continuous cdf F_1 on some interval $[a, b)$, $b \leq \infty$, $F_1(X)$ is uniformly distributed on $[0, 1]$, and $F_2^{-1}(F_1(X))$ has cdf F_2 for any strictly increasing cdf F_2.

Now we suggest a similar procedure to describe an RMPP, which includes the nonnegative marks, as well as the point process. With an RMPP, corresponding queue statistics are even easier to compute because the queue acts as a deterministic operator on the RMPP. Given sample paths of the RMPP, the performance measures for the queue can be calculated deterministically without generating any additional random numbers.

The particular performance measure we suggest is the normalized mean workload $c_Z^2(\rho)$ in (3). However, since we are given an unknown RMPP, we cannot initially insist on the scaling (19). Thus we suppose that we are given one RMPP $\{(u_n, v_n): n \geq 1\}$ having general means $\lambda^{-1} = Eu_n$, and $\tau = Ev_n$. Indeed, we only need be given the total input process

$$X(t) = \sum_{i=1}^{A(t)} v_i, \quad t \geq 0. \quad (33)$$

This involves some loss of information, because from the total input process, we cannot detect batch arrivals, and we do not try to; we lump all arrivals at the same instant together.

Since we are given one RMPP $\{(u_n, v_n): n \geq 1\}$ instead of a family indexed by ρ, by a direct application of Section I-B we would obtain only the statistic $c_Z^2(\rho)$ for *one* ρ. In this context, we suggest obtaining a function of ρ without changing the total input process by *letting the server work at different rates*. Given $Eu_n = \lambda^{-1}$ and $Ev_n = \tau$, when the server works at rate r, the traffic intensity is

$$\rho = \text{rate in/rate out} = \lambda\tau/r. \quad (34)$$

Then, separate calculations can be performed for all desired r for each sample path of the given RMPP. The offered traffic is thus unchanged, and we determine what happens if we process it at different rates. (This convention is natural in many applications, e.g., in manufacturing, communication networks, and computers, where production, transmission, and processing rates can be changed.)

With these new conventions, we need to redefine the workload process. Instead of (22), let the net input process be

$$Y_\rho^*(t) = X(t) - rt = \sum_{i=1}^{A(t)} v_i - rt, \quad t \geq 0. \quad (35)$$

Then the workload process $\boldsymbol{Z}_\rho^* \equiv \{Z_\rho^*(t): t \geq 0\}$ is defined just as in (24) by $\boldsymbol{Z}_\rho^* = f(\boldsymbol{Y}_\rho^*)$, where f is the barrier map. Finally, let the modified normalized mean workload be

$$c_{Z^*}^2(\rho) = \frac{2(1-\rho)}{\tau\rho} E(Z_\rho^*) \quad (36)$$

with ρ in (34). In the next section we show that (36) is the same as (3), so we drop the asterisk in (36).

Our premise is that $c_Z^2(\rho)$ embodies the primary effect of the variability in an RMPP upon the average workload in a single-server queue where the RMPP serves as the offered traffic. Therefore, we contend that $c_Z^2(\rho)$ should be a good indicator of the level of variability in the RMPP. As we discuss in Section IV-C, the normalized mean workload is constant as a function of ρ for any total input process that is a Lévy process (a process with stationary independent increments) with no negative jumps. Thus deviations from a constant function of ρ for $c_Z^2(\rho)$ indicate deviations from Lévy process structure in the total input process associated with the RMPP. Since the congestion in a queue tends to be affected by long-term dependence more at higher loads, $c_Z^2(\rho)$ roughly describes the variability in the RMPP at different time scales; i.e., the relevant time scale in the RMPP described by $c_Z^2(\rho)$ increases as ρ increases.

To actually estimate $c_Z^2(\rho)$ in (36), we first need to determine the parameters λ and τ. It is natural to estimate τ by the sample average $n^{-1}\Sigma_{i=1}^{n} v_i$ obtained from observing the jumps of $\boldsymbol{Y}_\rho^*$ or $\boldsymbol{Z}_\rho^*$. It is natural to estimate λ by the sample mean $t^{-1}A(t)$, or $\lambda\tau$ by the sample mean $t^{-1}X(t)$, obtained by assuming that no service at all is performed. By our assumption of stationarity and ergodicity, these estimators are asymptotically consistent (converge to the true value as the amount of data increases). Typically, they will be asymptotically normally distributed as well, e.g., by Theorem 20.1 of Billingsley [9].

Finally, given λ and τ or their estimates, $E(Z_\rho^*)$ can be estimated for each r, and thus for each ρ, by a minor modification of the algorithm in (27)–(31) (i.e., everywhere $\rho^{-1}U_n$ is replaced by U_n, $\rho^{-1}u_{n+1}$ is replaced by ρu_{n+1}, and $\Delta_{\rho n} = Z_\rho(U_n)^2/2$ when $Z_\rho(U_{n+1}-) = 0$).

D. Alternative Scalings

In Section II-A we introduced a one-parameter family of RMPPs and associated workload processes $\boldsymbol{Z}_\rho$ indexed by ρ by scaling the arrival process; i.e., we let $A_\rho(t) = A(\rho t)$ where $A(t)$ has the arrival rate 1. In Section II-C we introduced a different one-parameter family of RMPPs and workload processes by scaling the rate that service is provided. We could also have introduced a one-parameter family by scaling the service requirements, i.e., by replacing v_i by ρv_i. It is significant that it is easy to go from any one of these cases to any other, by which we mean that it is easy to express the sample paths, and thus also the finite-dimensional distributions of $\boldsymbol{Z}_\rho$ each way in terms of any one.

To see this, let $\boldsymbol{Y}_{\lambda,\tau,r}$ be the triply-scaled net input process, defined by

$$Y_{\lambda,\tau,r}(t) = \sum_{i=1}^{A(\lambda t)} \tau v_i - rt, \quad t \geq 0 \tag{37}$$

where $\{(u_n, v_n): n \geq 1\}$ satisfies the conditions of Section II-A, including the normalization $Eu_n = Ev_n = 1$ in (19). Let $\boldsymbol{Z}_{\lambda,\tau,r}$ be the associated workload processes, defined as before in terms of $\boldsymbol{Y}_{\lambda,\tau,r}$ by (24), i.e., $\boldsymbol{Z}_{\lambda,\tau,r} = f(\boldsymbol{Y}_{\lambda,\tau,r})$. Since $f(k\boldsymbol{Y}) = kf(\boldsymbol{Y})$ for any net input process $\boldsymbol{Y}$ and any positive constant k, and

$$\begin{aligned} Y_{\lambda,\tau,r}(t) &= \tau Y_{\lambda,1,r/\tau}(t) = \tau Y_{\lambda\tau/r,1,1}(rt/\tau) \\ &= \tau Y_{1,1,r/\lambda\tau}(\lambda t) = (r/\lambda) Y_{1,\lambda\tau/r,1}(\lambda t), \quad t \geq 0 \end{aligned} \tag{38}$$

we have

$$\begin{aligned} Z_{\lambda,\tau,r}(t) &= \tau Z_{\lambda\tau/r,1,1}(rt/\tau) = \tau Z_{1,1,r/\lambda\tau}(\lambda t) \\ &= (r/\lambda) Z_{1,\lambda\tau/r,1}(\lambda t), \quad t \geq 0. \end{aligned} \tag{39}$$

To understand how $\boldsymbol{Z}_{\lambda,\tau,r}$ behaves as a function of λ, τ and r, it thus suffices to know how one of $\boldsymbol{Z}_{\rho,1,1}$, $\boldsymbol{Z}_{1,\rho,1}$ or $\boldsymbol{Z}_{1,1,\rho^{-1}}$ behaves as a function of ρ alone for ρ in (34). In particular,

$$\begin{aligned} Z_{\lambda,\tau,r}(t) &= \tau Z_{\rho,1,1}(rt/\tau) = \tau Z_{1,1,\rho^{-1}}(\lambda t) \\ &= (r/\lambda) Z_{1,\rho,1}(\lambda t), \quad t \geq 0. \end{aligned} \tag{40}$$

Moreover, the normalized mean steady-state workload is then

$$c_Z^2(\rho; \lambda, \tau, r) \equiv \frac{2(1-\rho)}{\tau\rho} E(Z_{\lambda,\tau,r}) = \frac{2(1-\rho)E(Z_{\rho,1,1})}{\rho} \equiv c_Z^2(\rho) \tag{41}$$

as defined in (3), so that *the normalized mean workload is independent of the scaling*. In this sense, we have thus proved that the normalized mean workload $c_Z^2(\rho)$ in (3) describes the effect of the variability in the offered traffic upon the workload as a function of ρ, independent of the effect of the various deterministic rates. Of course, there is more to the variability than $c_Z^2(\rho)$, but at least the effect of the deterministic rates has been completely removed.

III. The Variability of the Offered Traffic

In this section we discuss methods to directly assess the variability of the offered traffic, without reference to queues (different from the normalized-mean-workload characterization discussed in Section II-C). We begin by discussing variance-time curves and indices of dispersion for point processes, and then we consider the IDW in (1). We describe the IDW when the offered traffic comes from several independent sources, when the service times are independent of the arrival process, and when the offered traffic is as in a GI/G/1 queue. We combine these descriptions to analytically describe the IDW of offered traffic to a multiclass queue when the classes are independent and the offered traffic for each class is as in a GI/G/1 queue. This multiclass model includes the models of Heffes and Lucantoni [35], Siram and Whitt [51], and Fendick *et al.* [24]–[26].

A. Variance-Time Curves

The variability of a real-valued random variable X is often partially characterized by its variance var (X). To make the characterization dimensionless (independent of the measuring units), it is natural to consider the squared coefficient of variation $c_X^2 = \text{var}\,(X)/(EX)^2$. (Note that $c_{aX}^2 = c_X^2$.) The variability of a stochastic process $\{X(t): t \geq 0\}$ is more complicated than the variability of a real-valued random variable, because it includes fluctuations over time due to the dependence among the different variables, as well as the variability of the individual variables $X(t)$ for each t. Nevertheless, the variance and associated dimensionless normalizations are useful partial characterizations of the variability of a stochastic process. A natural generalization of the variance is the *variance-time curve* $\{\text{var}\,[X(t)]: t \geq 0\}$.

From what we have said, it is clear that in general the variance-time curve is an unsatisfactory partial character-

ization of the variability of the process $\{X(t): t \geq 0\}$ because it only captures one form of the variability—the variability of the individual variables $X(t)$ for each t. However, if the stochastic process $\{X(t): t \geq 0\}$ is a *cumulative process*, i.e., if $X(t)$ measures some cumulative quantity over the interval $[0, t]$, then the variance-time curve often does appropriately partially characterize the variability of the process, because the variance-time curve partially characterizes the dependence among the *increments* of $\{X(t): t \geq 0\}$. For any $k + 1$ time points $0 \leq t_0 < t_1 < \cdots < t_k$,

$$\begin{aligned} \text{var}\,[X(t_k) - X(t_0)] &= \text{var} \sum_{i=1}^{k} [X(t_i) - X(t_{i-1})] \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} \text{cov}\,([X(t_i) - X(t_{i-1})], \\ &\qquad [X(t_j) - X(t_{j-1})]). \end{aligned} \tag{42}$$

Thus variance-time curves are natural candidates to partially characterize cumulative processes. This is important for us, because the arrival process $\boldsymbol{A}_\rho$ and the total input process $\boldsymbol{X}_\rho$ defined in (21) and (22) are cumulative processes.

From the case of a single random variable, one would naturally think that the appropriate normalization of the variance-time curve would be the squared-coefficient-of-variation curve $\{c^2_{X(t)}: t \geq 0\}$, obtained by setting $c^2_{X(t)} = \text{var}\,[X(t)]/(EX(t))^2$ for each t. However, this normalization fails to capture regularity that often occurs in cumulative processes for large t. When $\{X(t): t \geq 0\}$ is a cumulative process, typically $t^{-1}EX(t) \to x$, $0 < x < \infty$, and $t^{-1}\,\text{var}\,X(t) \to \sigma^2$, $0 < \sigma^2 < \infty$, as $t \to \infty$, so that

$$\frac{\text{var}\,[X(t)]}{[EX(t)]^2} = \frac{\text{var}\,[t^{-1}X(t)]}{[E(t^{-1}X(t))]^2} \to 0 \text{ as } t \to \infty \tag{43}$$

whereas

$$\frac{\text{var}\,[X(t)]}{E[X(t)]} = \frac{t^{-1}\,\text{var}\,[X(t)]}{t^{-1}EX(t)} \to \frac{\sigma^2}{x} \text{ as } t \to \infty. \tag{44}$$

For this reason, we consider normalizations of the form (44).

B. Indices of Dispersion for Point Processes

Variance-time curves have been used quite extensively to describe point processes; see pp. 69–133 of Cox and Lewis [20] and Brillinger [13]. The associated normalized functions are called *indices of dispersion;* see pp. 70–72 of [20]. For the arrival counting processes $A_\rho(t)$ with interarrival times $\rho^{-1}u_n$, the *index of dispersion for intervals* (IDI) is defined by

$$I_i(n) = \frac{n\,\text{var}\,(\rho^{-1}U_n)}{[E(\rho^{-1}U_n)]^2}, \quad n \geq 1 \tag{45}$$

and the *index of intervals for counts* (IDC) is defined by

$$I_{c\rho}(t) = \frac{\text{var}\,[A_\rho(t)]}{E[A_\rho(t)]} = \frac{\text{var}\,[A(\rho t)]}{E[A(\rho t)]} = I_c(\rho t), \quad t \geq 0. \tag{46}$$

Note that the IDI in (45) is independent of the scale factor ρ, but the IDC in (46) is not. In this context, the normalization of the variance can be interpreted as the corresponding variance term for a Poisson process. Thus, for a Poisson process, $I_i(n) = 1$ for all n and $I_{c\rho}(t) = 1$ for all t and ρ. In general, the indices of dispersion describe the relative variability of the point process at different arrival epochs or times compared to a Poisson process.

It is significant that the limit of $I_i(n)$ as $n \to \infty$, and the limit of $I_c(t)$ as $t \to \infty$ typically exist and coincide, equaling the variability portion of the normalization constant in the central limit theorems for U_n and $A(t)$, respectively (e.g., see Theorem 6 of [29]), which we refer to as the asymptotic variability parameter c^2_A. Moreover, c^2_A is typically the constant completely characterizing the effect of the variability in the arrival process on the workload and other performance measures in a large class of queues in heavy traffic; see [37]. Thus, in Section 2 of [55] we proposed using the variance-time curves to develop approximations for point processes and queues where the point processes serve as arrival processes. The IDI and IDC are also proposed as a basis for approximating arrival processes by Heffes [34], Newell [42], [43], Heffes and Lucantoni [35], and Sriram and Whitt [51].

C. The Index of Dispersion for Work

We are motivated to consider new measures of variability for the offered traffic because we want to consider the service times together with the arrival process. As pointed out in [24], in multiclass queues with non-Poisson arrival processes and class-dependent mean service times, there often is significant dependence between interarrival times and service times, and among successive service times, as well as among successive interarrival times. This extra dependence occurs even if the overall offered traffic is the superposition of mutually independent processes for the different classes and, for each class, the service times are i.i.d. and independent of the arrival process for that class. Moreover, in realistic examples of packet queues with variable packet lengths, it was found that significant error in estimating queueing performance measures can result from ignoring any of the three forms of dependence.

To treat the service times together with the interarrival times, Fendick *et al.* [24] proposed a *three-dimensional index of dispersion for intervals* (3D-IDI), defined by

$$\begin{aligned} I_{3i}(n) &\equiv (c^2_{An}, c^2_{Sn}, c^2_{ASn}) \\ &= \left(\frac{n\,\text{var}\,U_n}{(EU_n)^2}, \frac{n\,\text{var}\,V_n}{(EV_n)^2}, \frac{n\,\text{cov}\,(U_n, V_n)}{(EU_n)(EV_n)}\right), \quad n \geq 1. \end{aligned} \tag{47}$$

Obviously, the first and second terms in (46) are just the ordinary IDIs for the interarrival times u_n and service times v_n separately. The third term captures the dependence between the interarrival times and service times.

We propose yet another index of dispersion here, the *index of dispersion for work* (IDW), defined in (1). We are motivated to replace the 3D-IDI by the IDW because, first, the three-dimensional characterization is somewhat cumbersome; second, the 3D-IDI, just like the ordinary IDI in (45), does not directly describe the *timing of the variability;* third, from Section II-A it is clear that the workload depends on the offered traffic solely through the total input process $\{X_\rho(t): t \geq 0\}$ in (22); fourth, experience with simulations of packet queue models (Section 3 of [25]) indicates that the IDW provides a more useful characterization of the dependence for predicting the mean workload.

Thus we suggest working with the variance-time curve of the total input process. For this purpose, it is initially convenient to abandon the scaling conventions of Section II-A and consider one offered traffic $\{(u_n, v_n): n \geq 1\}$ without

the parameter ρ, and without specifying the means Eu_n and Ev_n. (The scaling can always be introduced afterwards, as we will show.) Paralleling (22), we let the total input be defined by (33) and we let the IDW be defined by (1), where $\tau = Ev_n$. We choose the normalization in (1) so that the IDW reduces to the IDC when the service requirements are constant, i.e., when $P(v_n = \tau) = 1$ for all n. Thus, for the offered traffic of an M/D/1 queue, $I_w(t) = 1$ for all t.

Let $V_w(t) = \text{var}\,[X(t)]$, where $\{X(t): t \geq 0\}$ is a stationary process. Since $I_w(t) = V_w(t)/\tau\rho t$, the behavior of $I_w(t)$ as a function of t is determined by the behavior of $V_w(t)$. Paralleling pp. 73–75 of [20], we can provide an integral representation for $V_w(t)$, assuming appropriate smoothness. In particular, we can write

$$V_w(t) = \eta_w t + \int_0^t \int_0^v 2\gamma_w(u)\,du\,dv, \quad t \geq 0,$$

$$\gamma_w(t) = \lim_{h \to 0} \frac{\text{cov}\,[X(s+h) - X(s), X(s+t+h) - X(s+t)]}{h^2}, \tag{48}$$

for $t \geq 0$, where $\eta_w = \tau\rho(c_s^2 + 1))$ (see Section IV-A-3). From (1) and (48), we see that $V_w(t)$ is convex (concave) if and only if $\gamma_w(t) \geq 0$ (≤ 0) for all t. In turn, $\gamma_w(t) \geq 0$ (≤ 0) for all t if and only if

$$\text{cov}\,[X(t_2) - X(t_1), X(t_4) - X(t_3)] \geq 0 \ (\leq 0) \tag{49}$$

for all $0 \leq t_1 < t_2 \leq t_3 < t_4$. Moreover, $\gamma_w(t) \geq 0$ (≤ 0) for all t clearly implies that $I_w(t)$ is nondecreasing (nonincreasing) in t. Whenever the covariances in (51) are all positive (negative), the IDW is nondecreasing (nonincreasing). As with the M/G/1 queue, the IDW is constant whenever $\{X(t): t \geq 0\}$ has independent increments. In other words, *the shape of the IDW reflects the covariances in the increments of the total input process.*

From (1) and (47), we see that the IDW can be expressed as

$$I_w(t) = \left[\eta_w + \frac{1}{t}\int_0^t \beta_w(u)\,du\right]\Big/\tau\rho \tag{50}$$

where

$$\beta_w(t) = \int_0^t 2\gamma_w(u)\,du, \quad t \geq 0. \tag{51}$$

In other words, $I_w(t) - I_w(0)$ can be expressed as the average of cumulative covariances given by $\beta_w(t)$ in (51).

D. Alternative Scalings

In Section II-D we considered three different scalings of the net-input and workload processes. In this section we show how these scalings affect the IDW. First, the total input process is obviously unaffected by the rate that service is provided, so it suffices to consider the IDW $I_{w,\lambda,\tau}(t)$ for the scaled input process

$$X_{\lambda,\tau}(t) = \sum_{i=1}^{A(\lambda t)} \tau v_i, \quad t \geq 0 \tag{52}$$

given the single sequence $\{(u_n, v_n): n \geq 1\}$ satisfying the normalization $Eu_n = Ev_n = 1$. From (1),

$$I_{w,\lambda,\tau}(t) = \frac{\text{var}\,[X_{\lambda,\tau}(t)]}{\tau E[X_{\lambda,\tau}(t)]} = \frac{\tau^2\,\text{var}\,[X_{1,1}(\lambda t)]}{\tau^2 E[X_{1,1}(\lambda t)]}$$
$$= I_{w,1,1}(\lambda t) \equiv I_w(\lambda t), \quad t \geq 0. \tag{53}$$

To treat the general case, it thus suffices to consider the single IDW $I_w(t)$ obtained from $\{(u_n, v_n)\}$ unscaled, and then scale by the arrival rate, just as for the IDC in (46). For the arrival-process scaling in (19), then, *the IDW of the offered traffic as a function of* ρ, say $I_{w\rho}(t)$, is

$$I_{w\rho}(t) = I_w(\rho t), \quad t \geq 0. \tag{54}$$

E. Superposition from Independent Sources

The IDW and the IDC are very convenient for treating the superposition of independent processes. Suppose that we have n sources of offered traffic (stationary versions), indexed by i, so that the total arrival counting processes and total input process are

$$A(t) = A_1(t) + \cdots + A_n(t) \quad \text{and}$$
$$X(t) = X_1(t) + \cdots + X_n(t), \quad t \geq 0. \tag{55}$$

Let λ_i be the arrival rate, and τ_i the mean service requirement of the ith class; let $\rho_i = \lambda_i\tau_i$, $\lambda = \Sigma_{i=1}^n \lambda_i$ and $\tau = \Sigma_{i=1}^n \lambda_i\tau/\lambda = 1$, so that $\rho \equiv \lambda\tau = \lambda$. Let the IDCs and IDWs be based on the scaling in (19), so that each IDC is based on arrival rate 1, and each IDW is based on offered load 1. If the component processes are independent, then the variances add and

$$I_c(\lambda t) = \frac{\text{var}\,[A(t)]}{E[A(t)]} = \sum_{i=1}^n \frac{\text{var}\,[A_i(t)]}{\lambda t}$$
$$= \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda}\right) I_{ci}(\lambda_i t), \quad t \geq 0 \tag{56}$$

and

$$I_w(\rho t) = \frac{\text{var}\,[X(t)]}{\tau E[X(t)]} = \frac{\sum_{i=1}^n \text{var}\,[X_i(t)]}{\rho t}$$
$$= \sum_{i=1}^n \left(\frac{\rho_i\tau_i}{\rho\tau}\right) I_{wi}(\lambda_i t), \quad t \geq 0. \tag{57}$$

In (56), $I_c(t)$ is scaled by λ and $I_{ci}(t)$ by λ_i because the arrival rates of $A(t)$ and $A_i(t)$ are λ and λ_i. Similarly, in (57), $I_w(t)$ is scaled by $\lambda = \rho$, and $I_{wi}(t)$ is scaled by $\lambda_i = \rho_i/\tau_i$ because the arrival rates for $X(t)$ and $X_i(t)$ are λ and λ_i; see Section III-D.

Without the scaling convention above, the IDC and IDW associated with a superposition of n i.i.d. sources would be *identical* to the IDC and IDW, respectively, of a single source. With the scaling, t in a component process is replaced by t/n. The time scaling reflects the convergence to a Poisson arrival process that is taking place as n increases.

F. Independent i.i.d. Service Times

The IDW simplifies when the service times are i.i.d. and independent of the arrival process. Then

$$E[X(t)] = (Ev_1)E[A(t)] = \rho t$$
$$\text{var}\,[X(t)] = E[A(t)]\,\text{var}\,(v_1) + [E(v_1)]^2\,\text{var}\,[A(t)], \quad t \geq 0 \tag{58}$$

so that

$$I_w(t) = \frac{\text{var}\,(v_1)}{[E(v_1)]^2} + \frac{\text{var}\,[A(t)]}{E[A(t)]} = c_s^2 + I_c(t), \quad t \geq 0. \tag{59}$$

Equation (59) coincides with (4.49) of Newell [42]. In this case, $I_w(t)$ is essentially the same measure of dependence as a function of time as $I_c(t)$ in (46).

From (59), we see that $I_w(t) = c_s^2 + 1$ for all t when the arrival process is also Poisson, i.e., for the M/G/1 queue. From (6), we see that $c_T^2(\rho)$ and $I_w(t)$ are both constant and equal for the M/G/1 queue.

We can combine (57) and (59) to express the IDW for the offered traffic to a multiclass queue with independent sources, each of which has i.i.d. service times independent of its arrival processes. Then, with the notation of Section III-E

$$I_w(\rho t) = \sum_{i=1}^{n} \left(\frac{\rho_i \tau_i}{\rho \tau}\right) c_{si}^2 + \sum_{i=1}^{n} \left(\frac{\rho_i \tau_i}{\rho \tau}\right) I_{ci}(\lambda_i t), \quad t \geq 0. \tag{60}$$

Note that the coefficients in the second sum in (60) are *not the same* as the corresponding coefficients in (56) unless $\tau_i = \tau$ for all i.

G. GI/G/1 Offered Traffic

If, in addition to the service-time assumptions of Section III-F, we assume that the arrival process is a renewal process (i.i.d. interarrival-times), then we have the offered traffic to a GI/G/1 queue. For renewal arrival processes, we can analytically compute the IDC, so that for GI/G/1 offered traffic we can apply (59) to analytically compute the IDW. Similarly, for offered traffic from independent GI/G/1 sources, which we might call $\Sigma(GI_i/G_i)/1$ offered traffic, we can analytically compute the IDW via (60). A natural approach is to apply Laplace transforms. If numerical inversion is to be performed in the multiclass framework of Section III-E, then we start by finding the Laplace transform $\hat{V}_i(s)$ of var $[A_i(t)]$ for each i. To apply (60), we express the second term as

$$\sum_{i=1}^{n} \left(\frac{\rho_i \tau_i}{\rho \tau}\right) I_{ci}(\lambda_i t) = \sum_{i=1}^{n} \left(\frac{\rho_i \tau_i}{\rho \tau}\right) \frac{\text{var}\,[A_i(t)]}{\lambda_i t} = \sum_{i=1}^{n} \frac{\tau_i^2 \,\text{var}\,[A_i(t)]}{\rho \tau t}, \quad t \geq 0. \tag{61}$$

Finally, we invert $\Sigma_{i=1}^{n} \tau_i^2 \hat{V}_i(s)$ to obtain $\Sigma_{i=1}^{n} \tau_i^2 \text{var}\,[A_i(t)]$. For example, this approach was used to obtain the IDC by Heffes and Lucantoni [35]. For their application, the component arrival processes are i.i.d., so that the IDC of the superposition process is the same as the IDC of one component arrival process, except for the scale factor (see (46)), so that it sufficies to invert $\hat{V}_1(s)$.

To specify the procedure for a renewal arrival process, assume that $A_\lambda(t)$ is a stationary version of a renewal process (the equilibrium renewal process) with arrival rate λ. We wish to determine $V_\lambda(t) = \text{var}\,[A_\lambda(t)]$ and its Laplace transform $\hat{V}_\lambda(s) = \int_0^\infty e^{-st} V_\lambda(t)\, dt$. Let $F(t) = P(u_n \leq t)$ be the cdf of one of the i.i.d. interarrival times with mean λ^{-1}, and let $\hat{f}(s) = \int_0^\infty e^{-st} dF(t)$ be its Laplace-Stieltjes transform (Laplace transform of the density). Let $H_0(t)$ be the renewal function, i.e., the mean number of renewals in the interval $(0, t]$ in the ordinary renewal process, and let $\hat{H}_0(s) = \int_0^\infty e^{-st} H_0(t)\, dt$ be its Laplace transform; see Chapter 2 of Cox [19].

From Chapter 4 of [19],

$$\hat{H}_0(s) = \frac{\hat{f}(s)}{s(1 - \hat{f}(s))},$$

$$\hat{V}_\lambda(s) = \frac{2\lambda}{s}\left[\hat{H}_0(s) - \frac{\lambda}{s^2} + \frac{1}{2s}\right],$$

$$V_\lambda(t) = 2\lambda \int_0^t \left[H_0(u) - \lambda u + \frac{1}{2}\right] du, \quad t \geq 0. \tag{62}$$

From (59) and (62), we see that in this GI/G/1 case the covariance density in (48) can be expressed in terms of the renewal density $h_0(t)$ (the derivative of $H_0(t)$) by

$$\gamma_w(t) = V_\lambda''(t) = 2\lambda[h_0(t) - \lambda], \quad t \geq 0. \tag{63}$$

In this case we can apply properties of the renewal function to deduce properties of the IDW. For example, from (3.13) on p. 171 of Barlow and Proschan [7], we deduce that $H_0(t) \leq (\geq)\lambda t$ for all t, so that $I_w(t) \leq (\geq 0)\, c_s^2 + 1$ for all t if the interarrival-time distribution is NBUE (NWUE), i.e., new better (worse) than used in expectation. From Brown [14], [15], we deduce that $H_0(t) - \lambda t$ is nondecreasing, so that $I_w(t)$ is nondecreasing, if the interarrival-time distribution is IMRL (increasing mean residual life). Moreover, we deduce that $H_0(t) - \lambda t$ is nondecreasing and concave, so that $I_w(t)$ is nondecreasing and concave, if the interarrival-time distributions is DFR (decreasing failure rate). (For IFR distributions, $H_0(t) - \lambda t$ need not be monotone.)

Example 3.1: Hyperexponential (H_2) Distributions: Suppose that the interarrival-time distribution is hyperexponential, i.e., a mixture of two exponentials, with density

$$f(t) = p\mu_1 e^{-\mu_1 t} + (1 - p)\mu_2 e^{-\mu_2 t}, \quad t \geq 0. \tag{64}$$

Since the H_2 distribution is DFR, from the remarks above, $H_0(t) - \lambda t$ and $I_w(t)$ are nondecreasing and concave. In fact, from p. 50 of Cox [19],

$$H_0(t) = \lambda t + \frac{(c_a^2 - 1)}{2} - \beta e^{-\gamma t}, \quad t \geq 0,$$

$$V_\lambda(t) = \lambda c_a^2 t - \frac{2\lambda\beta}{\gamma} + \frac{2\lambda\beta}{\gamma} e^{-\gamma t}, \quad t \geq 0,$$

$$I_c(t) = c_a^2 - \frac{2\beta}{\gamma t} + \frac{2\beta}{\gamma t} e^{-\gamma t}, \quad t \geq 0, \tag{65}$$

where m_{ak} is the kth moment, $\lambda = m_{a1}^{-1}$, $c_a^2 = (m_{a2} - m_{a1}^2)/m_{a1}^2$,

$$\gamma = (1 - p)\mu_1 + p\mu_2 \quad \text{and} \quad \beta = \frac{p(1 - p)(\mu_1 - \mu_2)^2}{\gamma^2}. \tag{66}$$

From large-time asymptotics in (115) below, we can also express β and γ in (66) as

$$\beta = \frac{(c_a^2 - 1)}{2} \quad \text{and} \quad \gamma = 2\beta\left[\frac{m_{a3}}{3} - \frac{(c_a^2 + 1)^2}{2}\right]. \tag{67}$$

This exact expression for the IDC in (65) is used to calculate the $H_2/M/1$ values in Table 1.

Example 3.2: Erlang (E_2) Distribution: Suppose that the interarrival-time distribution is Erlang of order two (E_2), i.e., the convolution of two exponential distributions, with Laplace transform $\hat{f}(s) = [2\lambda/(2\lambda + s)]^2$. Then, by p. 50 of [19] or p. 79 of [20], $H_0(t)$, $V_\lambda(t)$ and $I_c(t)$ are as in (65) with $c_a^2 = 1/2$, $\gamma = 4\lambda$ and $\beta = -1/4$, so that

$$V_\lambda(t) = \frac{\lambda t}{2} + \frac{1}{8} - \frac{1}{8} e^{-4\lambda t}, \quad t \geq 0. \tag{68}$$

We remark that the E_3 distribution is not so well behaved; see p. 50 of [19]. In this case $h_0(t)$ is not monotone, so that $I_w(t)$ is not convex.

In Examples 3.1 and 3.2, we see that the IDC has the asymptotic form

$$I_c(t) = c_a^2 + \frac{B}{t} + o(t^{-1}) \quad \text{as} \quad t \to \infty \tag{69}$$

where B is a known constant. In fact, this is true more generally, as we indicate in Section IV-B-3 below. From (69) we obtain the approximation

$$I_c(t) \approx c_a^2 + \frac{B}{t}, \quad t \geq 0 \tag{70}$$

which is often very good, provided that t is not too small, as can be seen from Examples 3.1 and 3.2. Combining (60) and (70), we obtain a convenient analytical approximation for the IDW in multiclass queues in which each class provides GI/G/1 offered traffic.

IV. The IDW and the Normalized Mean Workload

In this section we investigate the basic premise of this paper—that the normalized mean workload $c_z^2(\rho)$ in (3) is primarily determined by the IDW of the offered traffic in (1). Indeed, we have observed that for the M/G/1 queue that the IDW *completely* determines the normalized mean workload: $c_z^2(\rho) = I_w(t) = c_s^2 + 1$ for all ρ and t. Moreover, *for GI/M/1 queues, the IDC, and thus the IDW, completely determines the normalized mean workload;* i.e., from (59) and (62), we see that the IDC, and thus the IDW, completely determines the arrival-process renewal function $H_0(t)$, which in turn is well known to completely determine the interarrival-time distribution. However, for GI/G/1 queues in which neither the interarrival-time distribution nor the service-time distribution is exponential, the IDW does *not* completely determine the normalized mean workload, because from (59) and (62) we identify only the first two moments of the service time distribution. For general GI/G/1 queues, it is well known that the average workload depends on the service-time distribution beyond the first two moments. On the other hand, it is also known that the average workload does not *greatly* depend on the service-time distribution beyond the first two moments; see Table III of [58] and Tables 4–9 of Ramaswami and Latouche [47]. (These tables describe other mean steady-state quantities, but they determine the mean workload via $L = \lambda W$ and Brumelle's formula, (72) below.)

To be specific about the scaling for the IDW, let $\{I_w(t): t \geq 0\}$ be defined with respect to $\{X_1(t): t \geq 0\}$ using the scaling in Section II-A; i.e., $E[X_1(t)] = t$, $t \geq 0$, for a stationary version. As indicated in Section III-E, the associated IDW of $\{X_\rho(t): t \geq 0\}$ is then $I_{w\rho}(t) = I_w(\rho t)$ for all ρ, $0 < \rho < 1$. Hence, even when we introduce the scaling by ρ in Section II, there is essentially only one IDW as a function of t. Thus, as discussed in Section I-C, the main idea in this paper is that we should be able to use the IDW $\{I_w(t): t \geq 0\}$ as a basis for approximating $\{c_z^2(\rho): 0 \leq \rho \leq 1\}$. In fact, we contend that $\{c_z^2(\rho): 0 \leq \rho \leq 1\}$ and $\{I_w(t): t \geq 0\}$ contain roughly the same information, so that it should be possible to approximate either function reasonably well, given the other. This strengthened version of the main hypothesis supports using the normalized mean workload to approximately characterize an RMPP, as proposed in Section II-C. The strengthened version of the main hypothesis is supported by (9), but not by (10). In general, it should be easier to go from $I_w(t)$ to $c_z^2(\rho)$ than vice versa.

In this section we relate the IDW to the normalized mean workload by considering light traffic, heavy traffic, and a Lévy process framework (when the total input process has stationary independent increments). The IDW may be viewed as a means to implement and improve light-traffic and heavy-traffic interpolation, in particular, via (18).

A. Light Traffic

In this section we describe the limiting behavior of $c_z^2(\rho)$ as $\rho \to 0$, and relate it to the limiting behavior of $I_w(t)$ as $t \to 0$. The main positive result that we establish in this section is (12). The derivatives at 0 are not so simply related, but they involve similar ingredients.

1) The Normalized Mean Workload in Light Traffic: Experience has revealed that many performance measures depend on the fine structure of the queueing model in light traffic. Indeed, recent light-traffic limits for the mean waiting time in the GI/G/1 queue established by Daley and Rolski [21] indicate that both the appropriate normalization, and the resulting nondegenerate limit depend on the fine structure of the interarrival-time and service-time distributions. To briefly summarize their result, let $P(u_1 \leq t) = F(t)$ be the unscaled interarrival-time cdf. If there exists $\alpha > 0$ such that $t^{-\alpha}F(t) \to \gamma$, $0 < \gamma < \infty$, as $t \to 0$ and $E(v^{2+\alpha}) < \infty$, where v is a service time, then (with our scaling)

$$\lim_{\rho \to 0} \rho^{-\alpha} E(W_\rho) = E(v^{1+\alpha})\gamma/(1 + \alpha) \tag{71}$$

where W_ρ is the steady-state *waiting time* as a function of ρ. As Daley and Rolski point out, (71) indicates limitations on the possibility of developing simple robust approximations for the mean waiting time using light-traffic asymptotics. Even for the GI/G/1 model, the light-traffic limit depends on the detailed behavior of the interarrival-time distribution near the origin, and possibly higher moments of the service-time distribution. *It is significant that the mean steady-state workload is much better behaved in this respect.* This can be seen from Brumelle's [16] general formula (not restricted to GI/G/1) relating the mean workload $E(Z_\rho)$ to the waiting time and the service time; see pp. 408–412 of [36] and (4.2.4) on p. 107 of [27]. Using the scaling in (19) and assuming that all the quantities exist in equilibrium as well as limiting averages, we have

$$E(Z_\rho) = \rho E(W_\rho v) + \rho \frac{E(v^2)}{2} \tag{72}$$

where v is the service time of the customer with waiting time W_ρ. (Note that Z_ρ is the time-stationary workload, while W_ρ and v are the customer-stationary synchronous waiting time and service time.) Hence

$$c_z^2(\rho) = (1 - \rho)(c_s^2 + 1) + 2(1 - \rho)E(W_\rho v) \tag{73}$$

where $(c_s^2 + 1) = E(v^2)/(Ev)^2$. In great generality, $W_\rho \to 0$ and $E(W_\rho v) \to 0$ as $\rho \to 0$, so that we have the *very robust light-traffic limit*

$$\lim_{\rho \to 0} c_z^2(\rho) = c_z^2(0) = c_s^2 + 1. \tag{74}$$

Moreover, in general the derivative of the normalized mean workload at the origin is

$$\dot{c}_z^2(0) = -(c_s^2 + 1) + 2 \lim_{\rho \to 0} \rho^{-1} E(W_\rho v) \geq -c_z^2(0). \tag{75}$$

For example, in the GI/G/1 queue,

$$E(W_\rho v) = E(W_\rho)(Ev) = E(W_\rho) \tag{76}$$

so that (74) is valid and, by (71)

$$\dot{c}_z^2(0) = (c_s^2 + 1)(f(0) - 1) \tag{77}$$

where $f(0) = \lim_{t \downarrow 0} t^{-1}F(t)$ is the density of the unscaled interarrival-time distribution at the origin. Since the cdf of the scaled interarrival-time is

$$F_\rho(t) = P(\rho^{-1}u_1 \le t) = P(u_1 \le \rho t) = F(\rho t),$$

the scaled density and unscaled densities are related by

$$f_\rho(t) = \rho f(\rho t), \quad t \ge 0. \tag{78}$$

From (78), we see that $f(0) = \rho^{-1} f_\rho(0)$ in (77).

For example, when the interarrival-time is exponential, $f_\rho(0) = \rho$, and $\dot{c}_Z^2(0) = 0$. When the interarrival-time distribution is Erlang or deterministic, $f_\rho(0) = 0$, and $\dot{c}_Z^2(0) = -c_Z^2(0)$. Perhaps the main point is that (74), (75), and (77) are much less delicate than (71), so that it is more reasonable to hope for robust light-traffic approximations for the mean workload than for the mean waiting time.

Brumelle's formula (72) also produces the linear interpolation (7) for GI/G/1 queues when we apply the familiar waiting time approximation

$$E(W_\rho) = \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)}. \tag{79}$$

Indeed, (72) provides a simple way to *derive* (79) from $c_Z^2(0) = 1 + c_s^2$, $c_Z^2(1) = c_a^2 + c_s^2$ and linear interpolation.

From Brumelle's general formula (72), we immediately obtained the general form of $c_Z^2(0)$ in (74). The general form of the derivative $\dot{c}_Z(0)$ in (75) is more complicated, but we can heuristically derive what it should be. (Under regularity conditions, this argument can be made rigorous; e.g., by the methods of Reiman and Simon [49] or Reiman and Weiss [50]. For related heuristic discussion, see Section 6.8 of Newell [42].) As in Section II-A, we assume that customers arrive one at a time. Consider the customer-stationary process, and let v_0 be the service time of the 0th customer, v_{-1} be the service time of the (-1)st customer, and u_0 the unscaled interarrival-time between these two customers. In light traffic we can ignore all other customers so that for small ρ

$$E(W_\rho v) \approx E((v_{-1} - \rho^{-1}u_0)^+ v_0) \tag{80}$$

where $(x)^+ = \max\{x, 0\}$.

2) The Derivative $\dot{c}_Z^2(0)$ for a MultiClass Queue: Of course, (80) is often difficult to work with because v_{-1}, v_0, and u_0 can be arbitrarily dependent. However, to illustrate the possibilities, consider the multiclass queue in which the overall offered traffic is the superposition from independent sources. For each class, assume that the service times are i.i.d. and independent of the arrival process, but do not require that each component arrival process be a renewal process. As above, we use the scaling in (19) with $E(u_0) = E(v_0) = 1$. Let $p_{ij}(t)$ be the conditional probability that customer -1 is class i, and customer 0 is class j given that $u_0 = t$, which we assume exists as a measurable function of t that is continuous at $t = 0$. Of course, from the assumptions above, given the class identity of customers -1 and 0, v_{-1}, v_0, and u_0 are independent. Let $f(t)$ be the density of u_0, and let v_i be a service time from class i with mean τ_i. (Since i cannot be 0 or -1, there is no conflict with the notation of (80)). Let v_{ie} be a random variable with the stationary-excess distribution of the distribution of v_i, and note that

$$P(v_{ie} > t) = \int_t^\infty \frac{P(v_i > y)}{E(v_i)}\,dy = \int_0^\infty \frac{P(v_i > y + t)}{E(v_i)}\,dy$$

$$= \int_0^\infty \frac{P((v_i - t) > y)}{E(v_i)}\,dy = \frac{E[(v_i - t)^+]}{E(v_i)} \tag{81}$$

and

$$E(v_{ie}) = \int_0^\infty P(v_{ie} > t)\,dt. \tag{82}$$

Under regularity conditions allowing us to move limits inside the integral, from (80)–(82) we obtain

$$\lim_{\rho \to 0} 2\rho^{-1}E(W_\rho V) = \lim_{\rho \downarrow 0} 2\rho^{-1} \int_0^\infty \sum_{i=1}^n \sum_{j=1}^n p_{ij}(\rho t)\,(Ev_i)$$

$$\cdot\, E[(v_i - t)^+]\,\rho f(\rho t)\,dt$$

$$= 2\sum_{i=1}^n \sum_{j=1}^n p_{ij}(0)\,f(0)\,(Ev_i)\int_0^\infty E[(v_i - t)^+]\,dt$$

$$= 2\sum_{i=1}^n \sum_{j=1}^n p_{ij}(0)\,f(0)\,(Ev_i)\,(Ev_i)\,E(v_{ie})$$

$$= \sum_{i=1}^n \sum_{j=1}^n p_{ij}(0)\,f(0)\,(Ev_i)\,(Ev_i)^2(c_{si}^2 + 1). \tag{83}$$

Of course, $p_{ij}(t)$ and $f(t)$ are not easy to work with for general t, but the light traffic limit produces $p_{ij}(0)\,f(0)$ in (83), which is not hard to describe. In particular, an arbitrary arrival is type i with probability $\lambda_i/\lambda = \lambda_i$; the next interarrival time is 0 with a type j arrival with density λ_j if $j \ne i$, and $f_i(0)$ if $j = i$, where $f_i(0)$ is the stationary interarrival-time density for class i, so that

$$p_{ij}(0)\,f(0) = \begin{cases} \lambda_i\lambda_j, & i \ne j \\ \lambda_i f_i(0), & i = j. \end{cases} \tag{84}$$

Hence,

$$\lim_{\rho \to 0} 2\rho^{-1}E(W_\rho V) = \sum_{i=1}^n \lambda_i (Ev_i)^2 (c_{si}^2 + 1)\left(1 + \rho_i\left(\frac{f_i(0)}{\lambda_i} - 1\right)\right) \tag{85}$$

and, from (75),

$$\dot{c}_Z^2(0) = \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 \left(\frac{\tau_i}{\tau}\right)(c_{si}^2 + 1)\left(\frac{f_i(0)}{\lambda_i} - 1\right)$$

$$= \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 \left(\frac{\tau_i}{\tau}\right)\dot{c}_{Zi}^2(0) \tag{86}$$

where $c_{Zi}^2(\rho)$ is the normalized mean workload for class i alone with the scaling in (19). Note that (86) agrees with (77) in the case of a GI/G/1 queue. For the superposition of n i.i.d. sources, (74) and (86) yield

$$c_Z^2(0) = c_s^2 + 1 \quad \text{and} \quad \dot{c}_Z^2(0) = \frac{(c_{s1}^2 + 1)}{n}\left(\frac{f_1(0)}{\lambda_1} - 1\right). \tag{87}$$

The factor $(\rho_i/\rho)^2$ in the sum (86) and the factor n in (87) show that $\dot{c}_Z^2(0) \to 0$ as the number of component streams increases, reflecting the convergence of the superposition arrival process to a Poisson process. For a superposition of a large number of independent streams, a simple approximation is $\dot{c}_Z^2(0) \approx 0$.

3) Small-Time Asymptotics for the IDW: As in Section III-D, let $V_w(t) = \text{var}[X(t)]$ when $\{X(t): t \ge 0\}$ is the total input process. With the scaling $E(u_n) = E(v_n) = 1$ in (19),

$$I_w(t) = t^{-1}V_w(t), \quad t \ge 0 \tag{88}$$

so that the limits and derivatives at $t = 0$ are related by

$$I_w(0) = V_w'(0) \quad \text{and} \quad I_w'(0) = \frac{V_w''(0)}{2}. \tag{89}$$

Since arrivals occur one at a time, in great generality

$$E(A(t)^k) = P(A(t) = 1) + o(t) \quad \text{as} \quad t \to 0$$
$$= t + o(t) \quad \text{as} \quad t \to 0 \tag{90}$$

where $o(t)$ is some function $h(t)$ such that $t^{-1}h(t) \to 0$ as $t \to 0$, so that

$$E(X(t)^2) = tE(v^2) + o(t) \quad \text{as} \quad t \to 0,$$
$$V_w(t) = tE(v^2) + o(t) \quad \text{as} \quad t \to 0, \tag{91}$$

and

$$\lim_{t \to 0} I_w(t) = I_w(0) = V_w'(0) = c_s^2 + 1, \tag{92}$$

which together with (74) establishes (12).

In general, $I_w'(0) = V_w''(0)/2$ is quite complicated. However, suppose that the service times $\{v_n\}$ are i.i.d., and independent of the arrival process as in Section III-F. Then, by (59), $I_w(t) = c_s^2 + I_c(t)$, and

$$I_w'(0) = I_c'(0) = \frac{V_c''(0)}{2} \tag{93}$$

where $V_c(t) = \text{var}\,[A(t)]$, $t \geq 0$. As in Section 4.5 of Cox and Lewis [20] or Section III-C here, the variance function $V_c(t)$ can be expressed in terms of the covariance density, whose Fourier transform is the spectral density; i.e.,

$$V_c(t) = t + 2\int_0^t \int_0^v \gamma_+(u)\, du\, dv, \quad t \geq 0 \tag{94}$$

so that

$$V_c'(t) = 1 + 2\int_0^t \gamma_+(u)\, du, \quad t \geq 0,$$
$$V_c''(t) = 2\gamma_+(t), \quad t \geq 0,$$
$$I_w'(0) = \frac{V_c''(0)}{2} = \gamma_+(0), \tag{95}$$

where

$$\gamma_+(t) = \lim_{h \to 0} \frac{\text{cov}\,[A(s + h) - A(s), A(s + t + h) - A(s + t)]}{h^2} \tag{96}$$

and

$$\gamma_+(0) = \lim_{t \to 0} \gamma_+(t) = f(0) - 1, \tag{97}$$

with $f(t)$ being the density of the stationary interarrival-time distribution.

For the special case of a renewal arrival process, $V_c(\lambda t)$ is given in (62). From (62), we see that

$$V_c'(t) = 2\left\lfloor H_0(t) - t + \frac{1}{2} \right\rfloor, \quad t \geq 0,$$
$$V_c''(t) = 2[h_0(t) - 1], \quad t \geq 0, \tag{98}$$

where $h_0(t)$ is the renewal density, so that

$$I_c(0) = V_c'(0) = 1$$

and

$$I_c'(0) = \frac{V_c''(0)}{2} = h_0(0) - 1 = f(0) - 1, \tag{99}$$

consistent with the analysis above.

4) The Derivative $I_w'(0)$ for a MultiClass Queue: As in Sections III-E, III-F, and IV-A2 we now consider traffic from independent sources. By (57),

$$\rho I_w'(\rho t) = \sum_{i=1}^n \frac{\rho_i^2}{\rho\tau} I_{wi}'\left(\frac{\rho_i t}{\tau_i}\right), \quad t \geq 0, \tag{100}$$

so that

$$I_w'(0) = \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 I_{wi}'(0). \tag{101}$$

If, in addition, the service times of each class are i.i.d. and independent of the arrival process for that class, then we can combine (59), (99), and (101) to obtain

$$I_w'(0) = \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 \left(\frac{f_i(0)}{\lambda_i} - 1\right). \tag{102}$$

Note from (74), (86), (92), and (102) that

$$\dot{c}_{zi}^2(0) = I_{wi}'(0)\, I_{wi}(0) \quad \text{for each } i, \tag{103}$$

but we do *not* have $\dot{c}_z^2(0) = I_w'(0) I_w(0) = J_w'(0) J_w(0)$ as in (18). Indeed, from (86) and (102) we see that in general it is *not* possible to construct $\dot{c}_z^2(0)$ from $I_w'(0)$ and $I_w(0)$. However, *for multiclass queues with common service-time distributions, we do get the correct light-traffic derivative*. Then, from (86) and (102),

$$\dot{c}_z^2(0) = \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 \dot{c}_{zi}^2(0) = (c_s^2 + 1) \sum_{i=1}^n \left(\frac{\rho_i}{\rho}\right)^2 \left(\frac{f_i(0)}{\lambda_i} - 1\right)$$
$$= I_w(0)\, I_w'(0). \tag{104}$$

B. Heavy Traffic

In this section we describe the limiting behavior of $c_z^2(\rho)$ as $\rho \to 1$, and relate it to the limiting behavior of $I_w(t)$ as $t \to \infty$. The main positive result that we establish in this section is (8).

1) The Normalized Mean Workload in Heavy Traffic: The normalized mean workload and the entire workload process are very closely linked to the IDW in heavy traffic. The mappings in (22) and (24) that transform the total input process $\boldsymbol{X}_\rho$ into the workload process $\boldsymbol{Z}_\rho$ appropriately preserve convergence in a functional limit theorem setting (weak convergence), so that the CLT behavior of $X_1(t)$ as $t \to \infty$ determines the heavy-traffic behavior of $\boldsymbol{Z}_\rho$ as $\rho \to 1$; see [9], [24]–[26], [37], [52], and [53].

The CLT behavior of $X_1(t)$ is the limit

$$t^{-1/2}[X_1(t) - t] \Rightarrow N(0, \sigma^2) \quad \text{as} \quad t \to \infty \tag{105}$$

where $\Rightarrow$ denotes convergence in distribution, as in Billingsley [9], and $N(a, b)$ denotes a normal random variable with mean a and variance b. Given that $X_1(t)$ is defined in terms of $\{(u_n, v_n)\}$ as in (22), in great generality the limiting variance σ^2 in (105) can be expressed as

$$\sigma^2 = \lim_{t \to \infty} t^{-1} \text{var}\, X_1(t) = I_w(\infty) = c_A^2 + c_S^2 - 2c_{AS}^2 \tag{106}$$

where c_A^2, c_S^2, and c_{AS}^2 are the limits as $n \to \infty$ of c_{An}^2, c_{Sn}^2, and c_{ASn}^2 in (47). The CLT (105) holds quite generally, in particular whenever there is a joint (functional) CLT for the pair of partial sums (U_n, V_n). Moreover, the joint convergence of (U_n, V_n) is sufficient for a heavy-traffic limit theorem for $\boldsymbol{Z}_\rho$

as $\rho \to 1$. The normalized process

$$\hat{Z}_\rho(t) = (1 - \rho) Z_\rho(t/(1 - \rho)^2), \quad t \geq 0 \tag{107}$$

then converges to reflected Brownian motion (RBM) as $\rho \to 1$. Under extra regularity conditions (e.g., uniform integrability, p. 32 of [9]), the steady-state distribution and its moments converge as well, so that we have

$$\lim_{\rho \to 1} c_Z^2(\rho) = c_Z^2(1) = I_w(\infty) = c_A^2 + c_S^2 - 2c_{AS}^2 \tag{108}$$

which gives (8) and is consistent with (47).

For a GI/G/1 queue,

$$c_A^2 = c_a^2, c_S^2 = c_s^2 \quad \text{and} \quad c_{AS}^2 = 0 \tag{109}$$

where c_a^2 is the squared coefficient of variation of an interarrival time, and c_S^2 is the squared coefficient of variation of a service time, but in general none of the relations in (109) hold; i.e., c_A^2, c_S^2, and c_{AS}^2 also embody dependence among interarrival times, among service times, and between interarrival times and service times; see [24].

2) The Derivative $\dot{c}_Z^2(1)$ for a GI/G/1 Queue: In general, it seems reasonable to assume that the normalized mean workload can be expanded in a Taylor series about $\rho = 1$, yielding

$$\dot{c}_Z^2(\rho) = c_Z^2(1) - (1 - \rho)\dot{c}_Z^2(1) + o(1 - \rho) \quad \text{as} \quad \rho \to 1 \tag{110}$$

where $c_Z^2(1)$ is given by (108). However, we know very little about when (110) is valid and, if so, what is the value of $\dot{c}_Z^2(1)$.

However, for a GI/G/1 queue, we can determine what $\dot{c}_Z^2(1)$ should be from Marshall's [39] formula for the mean waiting time, i.e.,

$$E(W_\rho) = \frac{E[(\rho^{-1}u - v)^2]}{2E[(\rho^{-1}u - v)]} - \frac{E(I_\rho^2)}{2E(I_\rho)} \tag{111}$$

where I_ρ is the steady-state idle period (portion of a busy cycle in which the server is idle). Even though the proportion of time that the server is idle, which is $1 - \rho$, goes to 0 as $\rho \to 1$, I_ρ typically converges in distribution to I_1, and $E(I_\rho^k) \to E(I_1^k)$ where I_1 is the idle period with $\rho = 1$. Even though the expected length of a busy cycle becomes infinite as $\rho \to 1$, the idle period per busy cycle is typically well behaved. Assuming that $E(I_\rho^2)/2E(I_\rho)$ actually converges to $E(I_1^2)/2E(I_1)$, we can combine (72), (110), and (111) to obtain

$$\dot{c}_Z^2(1) = \frac{E(I_1^2)}{E(I_1)} - (c_a^2 - 1). \tag{112}$$

In fact, there has been much more work related to (112), as described in [60], but it is not very satisfactory because the ratio $E(I_1^2)/2E(I_1)$ is quite complicated. However, for the special case of a GI/M/1 queue, Halfin [30] proved that the analogs of (110) and (112) for the waiting time are valid and that

$$\frac{E(I_1^2)}{2E(I_1)} = \frac{E(u^3)}{3E(u^2)}, \tag{113}$$

so that

$$\dot{c}_Z^2(1) = \left(\frac{2m_{a3}}{3(c_a^2 + 1)} - (c_a^2 + 1)\right) \tag{114}$$

where $m_{a3} = E(u^3)$. A refined approximation for $\dot{c}_Z^2(1)$ in GI/G/1 queues is developed in [60] by exploiting algorithmically generated values in tables for the case $\rho = 0.98$.

3) Large-Time Asymptotics for the IDW of a Multiclass Queue: The asymptotic behavior of the IDC of a renewal process as $t \to \infty$ is well known. It in turn determines the asymptotic behavior of the IDW for a multiclass queue with independent GI/G/1 classes.

As in Section III-G, let $V_\lambda(t) = \text{var}[A_\lambda(t)]$, where $A_\lambda(t)$ is a stationary renewal process with arrival rate λ. By (18) on p. 58 of Cox [19] (there, $\mu_3 = E(u - Eu)^3$),

$$I_c(t) = \frac{V_1(t)}{t} = c_a^2 - \frac{1}{t}\left(\frac{m_{a3}}{3} - \frac{(c_a^2 + 1)^2}{2}\right) + o(t^{-1}) \quad \text{as} \quad t \to \infty \tag{115}$$

where $m_{a3} = E(u^3)$ as in (114). (Note that the second term on the right in (115) is similar, but not identical, to (114).) Combining (57), (59) and (115), we obtain for the multiclass queue with independent GI/G/1 classes (using the scaling convention in Section III-E)

$$I_w(\rho t) = \sum_{i=1}^{n}\left(\frac{\rho_i \tau_i}{\rho \tau}\right)(c_{ai}^2 + c_{si}^2) - \frac{1}{t}\sum_{i=1}^{n}\left(\frac{\tau_i^2}{\rho\tau}\right) \cdot \left(\frac{\lambda_i^3 m_{ai3}}{3} - \frac{(c_{ai}^2 + 1)^2}{2}\right) + o(t^{-1}) \quad \text{as} \quad t \to \infty \tag{116}$$

where i indexes the parameter of class i.

In the case of n i.i.d. classes, (116) reduces to

$$I_w(t) = (c_{a1}^2 + c_{s1}^2) - \frac{n}{t}\left(\frac{\lambda_1^2 m_{a13}}{3} - \frac{(c_{a1}^2 + 1)^2}{2}\right) + o(t^{-1}) \quad \text{as} \quad t \to \infty. \tag{117}$$

In contrast, the small-time asymptotics in the case of i.i.d. classes is

$$I_w(t) = (1 + c_{s1}^2) + \frac{t}{n}\left(\frac{f_1(0)}{\lambda_1} - 1\right) + o(t) \quad \text{as} \quad t \to 0; \tag{118}$$

see (102). In (117) the second-order term is *multiplied* by n, whereas in (118) the second-order term is *divided* by n. However, both (117) and (118) are special cases of relation

$$I_{w\rho}(t) = I_w(\rho t) = I_{w1}(\rho t/n), \quad t \geq 0. \tag{119}$$

The IDW basically remains unchanged upon superposing n i.i.d. processes, but the arrival rate has been multiplied by n, so the scaling must change; see Section III-D.

Note that (119) yields useful qualitative information about the way $I_w(t)$, and thus presumably $c_Z^2(\rho)$, behaves as a function of n, where n i.i.d. processes are being superposed. If $I_w(t)$ is increasing (decreasing) in t, then $I_w(\rho t/n)$ for each fixed t is decreasing (increasing) in n. Moreover, in the sense of this time scaling, $I_w(\rho t/n)$ converges to $I_w(0)$ (the M/G/1 model) at rate n^{-1}. (See [4], [51], and [59] for related discussion.)

C. Lévy Process Framework

In (5) and Section III-E we observed that the normalized mean workload and the IDW are constant and equal for the M/G/1 queue. This property extends to a large class of Lévy processes, i.e., processes with stationary independent increments. Indeed, it is obvious that if $\boldsymbol{X}_\rho$ is a Lévy process,

then the IDW is constant, because

$$I_{w\rho}(t) \equiv \frac{\text{var } [X_\rho(t)]}{E[X_\rho(t)]} = \frac{t \text{ var } [X_1(\rho)]}{tE[X_1(\rho)]} = \text{var } [X_1(1)], \quad t \geq 0. \tag{120}$$

It turns out that the Lévy process structure affects the normalized mean workload in the same way. In particular, for any storage model in which the workload process $\boldsymbol{Z}$ is defined in terms of the net input process $\boldsymbol{Y}$ by (24), if the net input process is a Lévy process having sample paths without negative jumps, with $E[Y(1)] < 0$ and var $[Y(1)] < \infty$, then the steady-state workload Z exists, having a Laplace transform satisfying a generalization of the M/G/1 Pollaczek-Khintchine equation with

$$E(Z) = \frac{\text{var } [Y(1)]}{-2E[Y(1)]}; \tag{121}$$

see Zolotarev [63], Chapters IX and XVII of Feller [23], Bingham [10], Harrison [31], and Chapter 3 of Prabhu [46]. This representation includes a Brownian motion net input process, and the associated reflected Brownian motion workload process arising in the heavy traffic limit, as discussed in Section IV-B1 (see Harrison [32]), as well as the M/G/1 queue, and the batch-Poisson model discussed in [24]. There are also other examples, such as the gamma process; see p. 72 of Prabhu [46].

Furthermore, if the Lévy net input process is $\boldsymbol{Y}_\rho$, defined in terms of a total input process $\boldsymbol{X}_\rho$ with $E[X_\rho(t)] = \rho t$ via (22), then $\boldsymbol{Y}_\rho$ must be a Lévy process without negative jumps and

$$E(Z_\rho) = \frac{\text{var } [Y_\rho(t)]}{-2E[Y_\rho(t)]} = \frac{\text{var } [X_\rho(t)]}{\tau E[X_\rho(t)]} \frac{\tau E[X_\rho(t)]}{2|E[Y_\rho(t)]|} = \frac{\tau \rho I_w(t)}{2(1-\rho)} \tag{122}$$

so that

$$c_z^2(\rho) \equiv \frac{2(1-\rho)\,E(Z_\rho)}{\tau\rho} = I_w(t) \quad \text{for all } \rho \text{ and } t. \tag{123}$$

In other words, with the Lévy process structure, the normalized mean workload coincides with the IDW; both are constant and equal.

V. Approximations

So far, we have shown that the IDW and the normalized mean workload are closely related, but we have yet to develop specific approximations for $c_z^2(\rho)$ given $I_w(t)$, such as (9)–(18). We tackle this problem now, relying heavily on heuristics. In this section, we assume that the IDW is available, and our purpose is to approximate $c_z^2(\rho)$. The IDW may have been estimated from measurements, as in (2), or it may have been determined analytically. For multiclass queues in which each class provides GI/G/1 input, we apply (60), and expressions for the IDC of a renewal process in Section III-G. We may use exact expressions for the IDC as in Examples III-1 and III-2, Laplace transform inversion, or the asymptotic approximation based on $t \to \infty$ in (70) and (115).

A. Variability Fixed Point Equations

It is intuitively clear that the time transformation $t(\rho)$ in (9) should depend on the IDW. In this section we consider how. As indicated in Section I-C, $I_w(t)$ for a given t has an impact upon $c_z^2(\rho)$ for a given ρ only if fluctuations in the offered traffic at time s have an impact upon the workload at time $s + t$ at that ρ. Moreover, if the workload process hits zero before time $s + t$, then this impact tends to be gone. It is thus natural to consider the approximation

$$c_z^2(\rho) \approx E[I_w(T_{e0}(\rho))] \tag{124}$$

where $T_{e0}(\rho)$ is the first passage time to 0 starting in equilibrium with traffic intensity ρ. We then approximate the mean first passage time to zero by what it would be for a Lévy process without negative jumps (see p. 79 of [46]), i.e.,

$$E[T_{e0}(\rho)] \approx \frac{E(Z_\rho)}{1-\rho}; \tag{125}$$

i.e., the mean workload in equilibrium is $E(Z_\rho)$, and the workload goes down at an average rate of $(1 - \rho)$. Combining (3), (124), and (125) we obtain the approximation

$$E[T_{e0}(\rho)] \approx \frac{\rho E[I_w(T_{e0}(\rho))]}{2(1-\rho)^2}, \quad 0 \leq \rho \leq 1. \tag{126}$$

The resulting approximation for $c_z^2(\rho)$ is based on a random variable T_{e0} satisfying equation (126). Note that this procedure yields an approximation of the form (10); the cdf $G_{w\rho}(t)$ there can be thought of as the cdf of $T_{e0}(\rho)$. (Newell uses a similar line of reasoning on pp. 132–152 of [42]; see especially (4.32)–(4.36) there. However, Newell considers only the heavy-traffic case in which $I_w(t)$ is essentially constant.) The mean $E[T_{e0}(\rho)]$ is a form of the relaxation time; we can think of (126) as a *relaxation-time fixed-point equation* in the space probability-distribution-valued functions of ρ.

Looking at (126), we see that a much more tractable equation would result if we could move the expectation operator insider $I_w(t)$ on the right. Then we would obtain the fixed-point equation

$$E[T_{e0}(\rho)] \approx \frac{\rho I_w(E[T_{e0}(\rho)])}{2(1-\rho)^2}, \quad 0 \leq \rho \leq 1 \tag{127}$$

in the space of real-valued functions of ρ. Motivated by (127), we propose approximations based on deterministic time transformations $t(\rho)$ as in (9), where $t(\rho)$ is determined by the fixed-point equation (15), where $x(\rho)$ is an increasing function of ρ with $x(0) = 0$, and $x(\rho) \to \infty$ as $\rho \to 1$. From (127), we would be led to choose

$$x(\rho) = \rho/2(1-\rho)^2 \tag{128}$$

which seems to be consistent with the heavy-traffic asymptotics for superposition arrival processes in [43] and [59], but other asymptotics in the next section suggest something like

$$x(\rho) = \rho/(1+\rho)(1-\rho) = \rho/(1-\rho^2) \tag{129}$$

which is closer to (16). We have not resolved what function $x(\rho)$ is best. We suggest (16) as a specific candidate, but something like (128) may be better when there are many classes; see Section VI-B. With (16), (128) or (129), (15) is easily solved for any ρ by finding the intersection of the line $t/x(\rho)$ with $I_w(t)$, as illustrated in Fig. 3.

B. Light-Traffic and Heavy-Traffic Derivatives

In order to specify the function $x(\rho)$ in (15) and in order to develop a light-traffic and heavy-traffic interpolation approximation, we consider the asymptotic behavior as ρ

$\rightarrow 1$, and as $\rho \rightarrow 0$. As noted in Section IV-B3, in considerable generality, $I_w(t) \approx A + Bt^{-1} + o(t^{-1})$ as $t \rightarrow \infty$, so we assume this general asymptotic form. Then (15) becomes

$$\frac{t(\rho)}{x(\rho)} \approx A + \frac{B}{t(\rho)} \tag{130}$$

and $t(\rho)$ is the solution to a quadratic equation, i.e.,

$$t(\rho) = \frac{x(\rho)A}{2}\left(1 + \left|1 + \frac{4B}{A^2x(\rho)}\right|^{1/2}\right). \tag{131}$$

Combining (9), (15) and (131), we obtain

$$c_Z^2(\rho) \approx I_w(t(\rho)) \approx A + \frac{B}{t(\rho)} \approx A\left(1 + y\left(\frac{2}{\sqrt{1 + 4y}}\right)\right)$$

$$\approx A(1 + y - y^2) \tag{132}$$

where $y = B/A^2x(\rho)$, so that

$$c_Z^2(\rho) \approx A + \frac{B}{Ax(\rho)} - \frac{B^2}{A^3x(\rho)^2}. \tag{133}$$

Note that (132) and (133) present candidate approximations when $x(\rho)$ is specified.

We also assume that $c_Z^2(\rho)$ can be represented in a Taylor series expansion as

$$c_Z^2(\rho) = c_Z^2(1) - (1 - \rho)\dot{c}_Z^2(1) + o(1 - \rho) \quad \text{as} \quad \rho \rightarrow 1. \tag{134}$$

Combining (133) and (134), we see that we should have $A = c_Z^2(1)$,

$$x(\rho) = a(1 - \rho)^{-1} + o(1) \quad \text{as} \quad \rho \rightarrow 1 \quad \text{and} \quad \dot{c}_Z^2(1) = -\frac{B}{Aa}. \tag{135}$$

Finally, we set $a = 1/2$, because that makes $\dot{c}_Z^2(1)$ agree with the exact results for the GI/M/1 queue in (114) obtained by Halfin [30].

Given the expansion for $I_w(t)$ as $t \rightarrow \infty$, we obtain the corresponding expansion for $J_w(\rho) = I_w(\rho/(1 - \rho))$ as $\rho \rightarrow 1$, i.e.,

$$J(\rho) \approx A + \frac{(1 - \rho)}{\rho} B + o(1 - \rho) \quad \text{as} \quad \rho \rightarrow 1, \tag{136}$$

so that $c_Z^2(1) = A = I_w(\infty) = J_w(1)$,

$$J_w'(1) = -B \quad \text{and} \quad \dot{c}_Z^2(1) = 2J_w'(1)/J_w(1). \tag{137}$$

We replace $J_w(1)$ in the denominator of (137) by $\max\{1, J_w(1)\}$ in (18) to avoid errors caused by dividing by very small numbers. This adjustment is motivated by experience with GI/G/1 queues such as the $E_{10}/D/1$ queue, for which $J_w(1) = 0.10$; see [60].

Similarly, in light traffic we assume that $I_w(t) = C + Dt + o(t)$ as $t \rightarrow 0$ (see (48) and (89)) and $c_Z^2(\rho) = c_Z^2(0) + \rho\dot{c}_Z^2(0) + o(\rho)$ as $\rho \rightarrow 0$. In this case (15) yields $t(\rho) \approx Cx(\rho)/(1 - Dx(\rho))$ and

$$c_Z^2(\rho) \approx I_w(t(\rho)) \approx C + Dt(\rho) \approx C + CDx(\rho) + o(x(\rho))$$

$$\text{as} \quad \rho \rightarrow 0. \tag{138}$$

Hence, $x(\rho) = a\rho + o(\rho)$ as $\rho \rightarrow 0$,

$$c_Z^2(0) \approx C = J_w(0) = I_w(0) \quad \text{and}$$

$$\dot{c}_Z^2(0) \approx aCD = aJ_w(0)J_w'(0) = aI_w(0)I_w'(0). \tag{139}$$

To be consistent with (100), we set $a = 1$, and (139) reduces to (18). Note that (129) is consistent with the asymptotics both as $\rho \rightarrow 0$ and $\rho \rightarrow 1$.

While the approximation for $\dot{c}_Z^2(0)$ in (139) and (18) is correct for a single-class queue in which the service times are i.i.d. and independent of the arrival process, and for multiclass queues with common service-time distributions, it is not correct in general. For the general multiclass model of Section IV-A2, we have derived $\dot{c}_Z^2(0)$ in (86), so for that model we should use (86) instead of (18).

C. Model Approximations

Our basic premise is that the normalized mean workloads in two different queueing systems should be nearly the same if the IDWs of their offered traffic are nearly the same. This premise immediately suggests an approximation procedure: Given a single-server queue with complex offered traffic, replace the offered traffic with offered traffic that is more manageable, and has nearly the same IDW. By "more manageable," we mean that the normalized mean workload can be computed exactly or approximately. The basic premise implies that the normalized mean workload obtained for the more manageable offered traffic should be a good approximation for the normalized mean workload in the original model.

Essentially this same approximation procedure has already been applied with some success to approximate complex arrival processes by renewal processes in [54], [55], [5], and [51], and by Markov modulated point processes in [34], [35]. The main difference here is that we are treating the service times together with the arrival process.

As an alternative to the fixed-point approximation in Section V-A and the light-traffic and heavy-traffic interpolation in Section V-B, we suggest a GI/G/1 model approximation. For any given offered traffic with IDW $I_w(t)$, we choose GI/G/1 offered traffic (if possible) to approximate $I_w(t)$. Of course, this means that we must choose cdf's $F_a(t)$ and $F_s(t)$ for an interarrival time u and a service time v with the scaling (19). We then approximate the original queue by the GI/G/1 queue.

A convenient way to approximately match the GI/G/1 IDW to the given IDW is to apply the asymptotics again. We use the asymptotics for the given IDW to determine four parameters that partially specify the cdf's $F_a(t)$ and $F_s(t)$, in particular, the squared coefficients of variation c_a^2 and c_s^2, the density of $F_a(t)$ at the origin, denoted by $f_a(0)$, and the third moment of $F_a(t)$, denoted by m_{a3}. We then can fit cdf's to c_a^2, c_s^2, $f_a(0)$, and m_{a3}. (The first moments are 1 here.)

We do not consider the distribution fitting to these parameters in detail here. However, given c_s^2 and the unit mean, we can easily pick the service-time cdf, e.g., according to (3.7) or (3.12) of [55]. Ignoring the density $f_a(0)$, we can fit an H_2 distribution to $F_a(t)$ using (3.5) and (3.6) of [55], when $c_a^2 > 1$ and $m_{a3} > 1.5(c_a^2 + 1)^2$. More generally, for $F_a(t)$, we suggest working with phase-type distributions, so that we obtain a PH/PH/1 queue, which can be solved algorithmically as in Ramaswami and Latouche [47].

The four parameters c_a^2, c_s^2, $f_a(0)$, and m_{a3} are easy to obtain from $I_w(0)$, $I_w'(0)$, $I_w(\infty)$, and $J_w'(1)$. First, since $c_Z^2(0) = 1 + c_s^2$, and $c_Z^2(1) = c_a^2 + c_s^2$ for a GI/G/1 queue, we apply (8) and (12) and let

$$c_s^2 = I_w(0) - 1 \quad \text{and} \quad c_a^2 = I_w(\infty) - I_w(0) + 1. \tag{140}$$

Second, we apply (95) and (97) to obtain

$$f_a(0) = I'_w(0) + 1. \tag{141}$$

Finally, we apply (59), (115), and (137) to obtain

$$m_{a3} = 3J'_w(1) + \tfrac{3}{2}(c_a^2 + 1)^2. \tag{142}$$

For the multiclass queue in which each class has GI/G/1 offered traffic, we can also express the IDW asymptotic values $I_w(0)$, $I_w(\infty)$, $I'_w(0)$, and $J'_w(1)$ in (140)–(142) in terms of the parameters λ_i, τ_i, c_{ai}^2, c_{si}^2, $f_{ai}(0)$, and m_{a3i} for each class. For example, in the special case of the superposition of n i.i.d. GI/G/1 sources, we can express the approximating GI/G/1 parameters c_{an}^2, c_{sn}^2, $f_{an}(0)$, and m_{a3n} for the queue with the superposition arrival process in terms of the exact GI/G/1 parameters associated with one source, c_{a1}^2, c_{s1}^2, $f_{a1}(0)$, and m_{a31}, by

$$c_{sn}^2 = c_{s1}^2,\ c_{an}^2 = c_{a1}^2,$$

$$f_{an}(0) = \frac{(c_{s1}^2 + 1)}{n} f_{a1}(0) + 1 - \frac{(c_{s1}^2 + 1)}{n},$$

$$m_{a3n} = nm_{a31} - \frac{3}{2}(n - 1)(c_{a1}^2 + 1)^2. \tag{143}$$

It is significant that for the superposition of n i.i.d. H_2/G/1 sources, (143) without using $f_{an}(0)$ gives parameters of a single H_2/G/1 source with a different third moment but an *identical* IDW. This is easy to see from (65)–(67). We will work with this example in Section VI.

Just seeing what these GI/G/1 parameters c_a^2, c_s^2, $f_a(0)$, and m_{a3} are provides some insight, since we already understand that GI/G/1 queue fairly well. Given these parameters, we need not fit $F_a(t)$ and $F_s(t)$; instead, we can apply the light-traffic and heavy-traffic interpolation approximation in [60], or the interpolation approximation based on (8), (12), and (18). (Then the difference is only in the correction factors $X(c_a^2, c_s^2)$ for $\dot{c}_z^2(1)$ in (2.21) and (2.24) of [60].) In the special case of the superposition of n i.i.d. sources, by matching the IDWs, we can express the approximating limiting values of $c_z^2(\rho)$ for the queue with the superposition offered traffic in terms of the limiting values of $c_{z1}^2(\rho)$ for the queue with a single source by

$$c_z^2(0) = c_{z1}^2(0),\ c_z^2(1) = c_{z1}^2(1),$$

$$\dot{c}_z^2(0) = n^{-1}\dot{c}_{z1}^2(0) \quad \text{and} \quad \dot{c}_z^2(1) = n\dot{c}_{z1}^2(1). \tag{144}$$

In fact, given that $c_{z1}^2(\rho)$ is exact, all the limiting values in (142) except $\dot{c}_z(1)$ are exact. Of course, the main idea with the GI/G/1 model approximation is to do better than the interpolation approximation in (17) and (18) by exploring the exact solution of the approximating model.

In the case of the arrival process alone (as occurs when the service times are i.i.d. and independent of the arrival process), there is a procedure proposed by T. J. Ott and H. Zucker (unpublished manuscript, AT&T Bell Laboratories, 1981) for finding a renewal process that has *exactly the same* IDC as a given stationary point process, whenever such a construction is possible. (We assume the rates are fixed at 1.) The idea is to match the renewal function associated with the renewal process to $E[A^o(t)]$, where $A^o(t)$ is the Palm version of the stationary arrival counting process $A(t)$, obtained by conditioning on a point being at 0. It is well known that this is equivalent to matching the IDCs, because for any stationary point process

$$E[A(t)^2] - t = 2\int_0^t E[A^o(s)]\,ds, \quad t \geq 0 \tag{145}$$

see p. 68 of [20].

For a renewal process, we have

$$\hat{\alpha}(s) = \frac{\hat{f}(s)}{1 - \hat{f}(s)} \quad \text{and} \quad \hat{f}(s) = \frac{\hat{\alpha}(s)}{1 + \hat{\alpha}(s)} \tag{146}$$

where $\hat{f}(s)$ and $\hat{\alpha}(s)$ are the Laplace-Stieltjes transforms (LSTs)

$$\hat{f}(s) = \int_0^\infty e^{-st}\,dF(t) \quad \text{and} \quad \hat{\alpha}(s) = \int_0^\infty e^{-st}\,dE[A^o(t)] \tag{147}$$

with F being the interarrival-time cdf. Thus we can find an interarrival-time cdf for a renewal process with the same IDC by solving the second equation in (146). Obviously, the procedure works if and only if $\hat{f}(s)$ so obtained is an LST of a bona fide cdf, which occurs if and only if $\hat{f}(s)$ is completely monotone; see p. 439 of [23].

At first glance, the Ott-Zucker procedure may not look very promising, but they prove that $\hat{\alpha}(s)/(1 + \hat{\alpha}(s))$ is indeed completely monotone, so that $\hat{f}(s)$ is the transform of a bona fide cdf, whenever the stationary point process is the superposition of stationary renewal processes in which each component interarrival-time distribution is completely monotone (a mixture of exponentials). Moreover, in the case of superposition of stationary renewal processes, the transform $\hat{\alpha}(s)$ is easily expressed in terms of the rates λ_i and transforms $\hat{\alpha}_i(s)$ of the component renewal processes by

$$\frac{1}{1 - \hat{\alpha}(s)} = \sum_{i=1}^n \frac{\lambda_i}{\lambda} \frac{1}{1 - \hat{\alpha}_i(s)} + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\lambda_i \lambda_j}{\lambda s}. \tag{148}$$

Even when the Ott-Zucker procedure does not produce a bona fide cdf, it may be useful for generating approximations. For example, Ott and Zucker applied the transform $\hat{f}(s)$, even when it is not completely monotone, to approximately solve the approximating "GI"/M/1 queue in the usual way, and the results were very encouraging. We tend not to prefer this approach because it requires considerable data about the arrival process, because it does not seem to extend to include the service requirements, and because further work must be done to identify and solve the approximating GI/G/1 queue. However, the procedure obviously has potential.

Once we have an approximate GI/G/1 model, we can calculate other performance measures besides the average workload, but it is not at all clear that the IDW should be used to predict more than the average workload. Obviously, the IDW provides some idea about the queue performance more generally, but it is not so directly connected to other performance measures. To illustrate the limitations, note that neither the IDW nor the full workload process can distinguish between a queue with batch arrivals having a Poisson arrival process for batches, and a simple M/G/1 queue in which the customers arrive one at a time. Thus, from the IDW or the mean workload, we cannot predict the average waiting time per customer. However, if service times are independent of waiting times, so that customer arrivals occur one at a time, then we can predict the average waiting time from the average workload using Brumelle's formula (72).

As illustrated by the M/G/1 queue, the full distribution of the steady-state workload depends on the service-time distribution beyond the first two moments, so we clearly cannot determine the distribution of the steady-state workload from the IDW. Even a full GI/G/1 model approximation matching the IDW exactly does not help. The difficulties are well illustrated by examples in Ramaswami and Latouche [47]. However, a natural rough approximation for the steady-state workload distribution, assuming nice behavior, is an atom of $(1 - \rho)$ at zero (which is exact), plus an exponential tail, with the previously determined mean, i.e.,

$$P(Z_\rho > t) \approx \rho e^{-t\rho/E(Z_\rho)}, \quad t > 0. \tag{149}$$

This exponential approximation is asymptotically correct in heavy traffic in great generality, and is known to be remarkably good for a large class of M/G/1 queues. As indicated by the heavy-traffic limit, (149) tends to perform better as ρ increases. A possible refinement of (149) is to use an interpolation approximation, using the fact that the workload distribution, conditional on being positive, behaves like the stationary-excess of the service-time distribution in light traffic.

Of course, we can also consider other model approximations besides GI/G/1 models. Another good candidate for the arrival process is a Markov modulated Poisson process. This model approximation is well illustrated by Heffes and Lucantoni [35].

VI. A Multiclass Example

We conclude this paper with an example illustrating the approximations. As in Section III-G, we consider a multiclass queue with n independent GI/G/1 sources, i.e., a $(\Sigma_{i=1}^{n}\ \mathrm{GI}_i/\mathrm{G})/1$ model. In addition, we assume that the n sources have common arrival and service distributions. (We consider class-dependent service-time distributions in [24]–[26].)

As in Section III-E, we assume that each IDW is scaled to have offered load 1, and each IDC is scaled to have arrival rate 1, so that by (56)–(60)

$$I_w(\rho t) = I_{w1}(\rho t/n) = c_{s1}^2 + I_{c1}(\rho t/n), \quad 0 \le \rho \le 1 \tag{150}$$

where ρ is the total arrival rate, and the index 1 refers to a quantity associated with a single component source. Note that in (150) the IDW essentially reduces to the IDC. Moreover, by (87), (92), and (104),

$$c_Z^2(0) = I_w(0) = 1 + c_{s1}^2,$$

$$\dot{c}_Z^2(0) = I_w'(0)\, I_w(0) = \frac{(1 + c_{s1}^2)}{n}\left(\frac{f_1(0)}{\lambda_1} - 1\right). \tag{151}$$

By (106),

$$c_Z^2(1) = I_w(\infty) = c_{a1}^2 + c_{s1}^2; \tag{152}$$

by (18) and (117),

$$\dot{c}_Z^2(1) \approx \frac{2I_w'(0)}{\max\{1, I_w(1)\}} = \frac{2n}{\max\{1, c_{a1}^2 + c_{s1}^2\}} \cdot \left(\frac{m_{a31}}{3} - \frac{(c_{a1}^2 + 1)^2}{2}\right). \tag{153}$$

Note that $c_Z^2(0)$, $\dot{c}_Z^2(0)$, and $c_Z^2(1)$ in (151) and (152) are exact, while $\dot{c}_Z^2(1)$ in (153) is an approximation.

We compare four approximations developed in this paper—the simple time transformation (9) plus (13), the variability fixed-point approximation (9) plus (15), the interpolation approximation (17), (18) plus (151)–(153), and the GI/G/1 model approximation (143) plus Example III-G-1—with simulation estimates of exact values obtained by Albin [3], [5] in the process of developing her hybrid approximation. (See Chapters 3–5 and Appendices 6 and 17 of [3].)

A. The First Experiment: $H_2/M/1$ Classes

The first experiment contains 9 cases, based on 3 values of ρ (0.5, 0.8, and 0.9), and 3 values of n (2, 8, and 128). The interarrival-time distributions are hyperexponential with balanced means, and $c_a^2 = 2.0$ (then $m_{a3} = 18$, and $f_a(0) = 1.333$); the service-time distributions are exponential. Thus,

$$c_Z^2(0) = 2.0 \quad \text{and} \quad c_Z^2(1) = 3.0. \tag{154}$$

As a basis for comparison, a simple M/M/1 approximation is $c_Z^2(\rho) \approx 2.0$ for all n and ρ as in (6), and a simple two-moment heavy-traffic (or asymptotic-method, see [55]) approximation is $c_Z^2(\rho) = 2 + \rho$ for all n as in (7). The M/M/1 approximation is good for large n, because the superposition process approaches a Poisson process as $n \to \infty$. However, neither simple approximation provides uniformly good extra information beyond (154).

Ten other approximations are compared with simulation estimates of exact values in Table 2. In addition to the simulation estimates, which were obtained by Albin [3]–[5], approximate 95 percent confidence intervals are indicated in parentheses below the estimate. Albin's estimates actually are for the mean number in system, but these are converted into the mean workload by applying $L = \lambda W$ and (72), Since this transformation is linear, the new confidence intervals are easily obtained as well.

The first two approximations are Albin's hybrid approximation, and QNA approximation in (29) and (33) of [57] that is based on it. Both of these approximations perform well for this example, but they only apply to superpositions with i.i.d. service times independent of the arrival process. In that context, the QNA approximation is appealing because of its simplicity.

The next three approximations are simple deterministic time transformations of the IDW, as in (9) where $t(\rho)$ is independent of the IDW. For this example, the exact IDW is available from (150) and Example III-G-2. Here, we see that it is better to have the exponent 2 on $(1 - \rho)$ in $t(\rho)$, which was not obvious from Table 1.

The next three approximations are variability fixed-point approximations based on (9) and (15) with three different candidate $x(\rho)$ functions, (16), (128) and (129). Overall, (129) seems best, but it seems too low for all cases except the first.

The next-to-last approximation is a GI/G/1 model approximation as discussed in Section V-C. By redefining the third moment as in (143) and fitting an H_2 distribution to the first three moments, we obtain an $H_2/M/1$ queue with the identical IDW. (Evidently the same result would be achieved by applying (148).) The approximate solution is then the *exact* solution for the resulting $H_2/M/1$ model (obtained in the usual way). As can be seen from Table 2, the GI/G/1 model approximation performs well, but not exceptionally well. Evidently there is a limit to the information provided by the IDW.

Table 2 A Comparison of Ten Approximations with Simulation Estimates of Exact Values of the Normalized Mean Workload $c_z^2(\rho)$ in a Multi-Class $(\Sigma_{i=1}^n$ GI$_i$/M)/1 Model with n i.i.d. Sources, Each Having Exponential Service Times and H_2 Interarrival Times with Balanced Means and $c_a^2 = 2.0$, as Discussed in Section VI-A. Estimated 95 Percent Confidence Intervals Appear Below the Simulation Estimates in Parentheses.

Traffic Intensity	Number of Sources	Simulation Estimate	Albin's [3], [5] Hybrid approx.	QNA (29) and (33) of Whitt [57]	Simple Time Transformation (9) with $t(\rho) =$			Variability Fixed Point equation (15)			GI/G/1 model approx.	
					$\frac{\rho l_w(\infty)}{2(1-\rho)^2}$	$\rho/(1-\rho)^2$	$\rho/(1-\rho)$	(16)	(128)	(129)	H_2/M/1 exact	interp.
$\rho = 0.5$	$n = 2$	2.34 (±0.01)	2.30	2.50	2.37	2.27	2.15	2.30	2.42	2.27	2.28	2.28
	$n = 8$	2.19 (±0.01)	2.10	2.25	2.12	2.08	2.04	2.01	2.01	2.13	2.08	2.04
	$n = 128$	2.02 (±0.02)	2.01	2.01	2.01	2.01	2.00	2.00	2.00	2.01	2.00	2.00
$\rho = 0.8$	$n = 2$	2.60 (±0.02)	2.69	2.86	2.90	2.85	2.45	2.73	2.89	2.54	2.64	2.65
	$n = 8$	2.52 (±0.03)	2.39	2.47	2.63	2.51	2.15	2.30	2.67	2.24	2.27	2.20
	$n = 128$	2.09 (±0.02)	2.05	2.05	2.07	2.05	2.01	2.03	2.06	2.02	2.02	2.00
$\rho = 0.9$	$n = 2$	2.90 (±0.03)	2.86	2.96	2.98	2.97	2.68	2.89	2.98	2.78	2.81	2.81
	$n = 8$	2.72 (±0.04)	2.72	2.78	2.91	2.87	2.30	2.53	2.91	2.40	2.48	2.44
	$n = 128$	2.21 (±0.05)	2.16	2.16	2.28	2.20	2.02	2.04	2.27	2.04	2.05	2.01

The final approximation in Table 2 is the interpolation approximation in (17) and (18) using (151)–(153), which coincides with the interpolation approximation applied to the H_2/M/1 model obtained by the GI/G/1 model approximation. (All the interpolation approximations for GI/G/1 queues in [60] coincide for exponential service-time distributions.) It is significant that the interpolation approximation performs nearly as well as the GI/G/1 model approximation. This example suggests that it may not be worthwhile obtaining a full GI/G/1 model that has nearly the same IDW even if it can be done.

In summary, the simple time transformation in (9) plus (13), and the interpolation in (17) plus (18) look very good in this experiment, and are appealing because of their simplicity.

B. The Second Experiment: H_2 and E_2 Distributions

The second experiment consists of 24 cases, based on 2 values of ρ (0.7 and (0.9)), 3 values of n (2, 4, and 16), 2 interarrival-time distributions, and 2 service-time distributions. The two interarrival-time and service-time distributions are E_2 and H_2 with balanced means, and $c_a^2 = 5.0$. With mean 1, the third moments are $m_{a3} = 3.0$, and $m_{a3} = 3c_a^2(c_a^2 + 1) = 90$, respectively; the densities at the origin are $f(0) = 0$, and $f(0) = 2c_a^2/(1 + c_a^2) = 5/3$.

Albin's simulation experiment also includes other distributions, including other service-time distributions with the same c_s^2, in particular, lognormal for $c_s^2 = 5.0$, and shifted exponential with $c_s^2 = 0.5$. (A shifted-exponential random variable is a constant plus an exponential.) In support of approximations for the mean workload that only depend on the first two moments of the service time, the simulation estimates with the different service-time distributions are quite close to the estimates for the service distributions we consider. The bigger discrepancy occurs when $c_a^2 = c_s^2 = 5.0$, $\rho = 0.7$, and $n = 2$; the lognormal service times yield values about 6 percent lower. Overall, Albin's simulation results support using only two moments of the service-time distribution.

The basic approximations developed here are compared to the estimated exact values in Table 3. Each simulation estimate is also given an approximate standard deviation based on 20 batches from one long run.

As in Section VI-A, the approximations are all obtained analytically, without measurements, using the explicit formulas for the E_2 and H_2 IDWs in Section III-G. Otherwise, we would have worked with the asymptotic approximation in (69) and (115). For example, for the H_2/H_2 case, the approximation is $I_w(t) \approx 10 - 12n/t$, which is quite accurate for $I_w(t) \geq 8$. In applications, typically the component arrival processes have a given time scale, so that when more component processes are introduced, ρ increases with n. Then, the IDW associated with the superposition of n i.i.d. processes is independent of n. In contrast, here the IDWs change with n, as indicated in (150) and as illustrated for the H_2/H_2 and E_2/E_2 cases in Figs. 5 and 6, where the variability fixed-point equation (15) is solved graphically for all n and ρ. In Table 3, we only describe the variability fixed point equation with $x(\rho)$ in (16).

For the most part, the numerical results in Table 3 look good, but clearly not nearly as good as the results for the GI/G/1 queue in Table 1 and [60]. The results are consistently good for E_2 arrival processes, but $c_z^2(0)$ differs from $c_z^2(1)$ by only 0.5 in these cases. (From our approach, we do clearly see that it should be easier to obtain good approximations in these cases.)

For the H_2 arrival processes, the results are reasonable for smaller n, but the quality of the results deteriorates dramatically as n increases. Indeed, for the H_2/E_2 case, the two

Table 3 A Comparison of Three Approximations for the Normalized Mean Workload $c_\lambda^2(\rho)$ in (3) with Simulation Estimates for the Multi-Class Model Having Common Service-Time Distributions and n i.i.d Component Arrival Processes, as Described in Section VI-B. The Simulation Estimates Come from Chapter 5 and Appendix 17 of [3]. The Statistical Precision is Described Approximately by One Sample Standard Deviation Based on 20 Batches From One Long Run.

			E_2 service $c_s^2 = 0.5$		H_2 service, $c_s^2 = 5.0$	
		Arrival	E_2, $c_a^2 = 0.5$	H_2, $c_a^2 = 5.0$	E_2, $c_a^2 = 0.5$	H_2, $c_a^2 = 5.0$
		exact	1.1	3.58	5.58	8.67
		st. dev.	(0.009)	(0.06)	(0.09)	(0.11)
	$n = 2$	(14)	1.06	4.41	5.51	9.38
		(9) and (15)	1.10	3.26	5.52	8.87
		(17) and (18)	1.13	3.32	5.53	8.40
		exact	1.16	2.95	5.62	8.14
		st. dev.	(0.008)	(0.03)	(0.06)	(0.14)
$\rho = 0.7$	$n = 4$	(14)	1.13	3.63	5.52	8.81
		(9) and (15)	1.17	2.20	5.54	7.92
		(17) and (18)	1.27	2.37	5.56	7.53
		exact	1.28	1.99	5.78	6.83
		st. dev.	(0.009)	(0.05)	(0.11)	(0.08)
	$n = 16$	(14)	1.32	2.27	5.59	7.26
		(9) and (15)	1.35	1.70	5.64	6.57
		(17) and (18)	1.29	1.55	5.68	6.22
		exact	1.05	4.76	5.52	9.43
		st. dev.	(0.013)	(0.17)	(0.16)	(0.17)
	$n = 2$	(14)	1.01	5.40	5.50	9.95
		(9) and (15)	1.02	5.00	5.50	9.85
		(17) and (18)	1.05	4.65	5.51	9.43
		exact	1.04	4.44	5.43	9.26
		st. dev.	(0.016)	(0.14)	(0.13)	(0.25)
$\rho = 0.9$	$n = 4$	(14)	1.01	5.31	5.50	9.89
		(9) and (15)	1.05	4.38	5.51	9.45
		(17) and (18)	1.10	4.00	5.52	8.96
		exact	1.10	4.03	5.40	8.77
		st. dev.	(0.014)	(0.16)	(0.15)	(0.29)
	$n = 16$	(14)	1.04	4.73	5.51	9.57
		(9) and (15)	1.17	2.28	5.54	7.88
		(17) and (18)	1.13	2.13	5.57	7.23
		$c_\lambda^2(0)$	1.5	1.50	6.00	6.00
		$n\dot{c}_\lambda^2(0)$	−1.5	1.00	−6.00	4.00
		$c_\lambda^2(1)$	1.0	5.50	5.50	10.00
		$n^{-1}c_\lambda^2(1)$	−0.25	4.364	−0.455	2.40

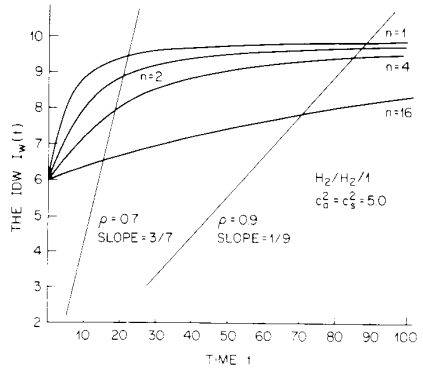


Fig. 5. The IDWs and the variability-fixed-point approximation for the $H_2/H_2/1$ examples in Section VI-B.

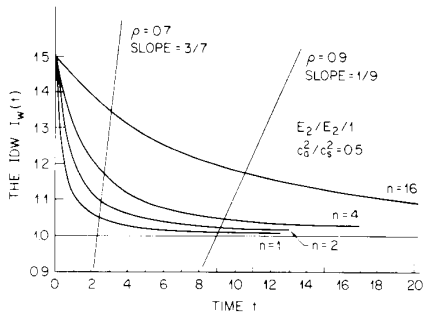


Fig. 6. The IDWs and the variability-fixed point approximation for the $E_2/E_2/1$ examples in Section VI-B.

approximations, (9) plus (15), and (17) plus (18), both perform very poorly when $n = 16$ and $\rho = 0.9$. For the interpolation approximation, this could be anticipated, because the derivatives cease to pin down the middle as n increases. In particular, for $n = 16$, $\dot{c}_\lambda^2(0) = 0.063$, and $\dot{c}_\lambda^2(1) = 69.8$. Indeed, from Table 3 we see that the derivatives are least informative for large n in the H_2/E_2 case. This difficulty with the approximation also was observed of the $H_2/M/1$ queue in Table 5 of [60]. As before, the results can be interpreted more positively: The approximations for the function $c_\lambda^2(\rho)$ have the right shape. The values do not match well at higher ρ because the functions are very steep at that end.

However, we could not anticipate so clearly the difficulty with the variability fixed point approximation in the H_2/E_2 case with $n = 16$ and $\rho = 0.9$. We conjecture that the asymptotics in Section V-B cease to be relevant as n and ρ both get large, so that for large n, $x(\rho)$ in (16) should become something like $\rho/2(1-\rho)^2$, as in (128). Indeed, from [43] and [59], we know that the crucial factor when $n \to \infty$ and $\rho \to 1$ is $n(1-\rho)^2$, so that the crude approximation (14) should get steadily better as n increases.

From [59], we see that as n gets extremely large, in a certain sense $c_z^2(\rho)$ approaches a step function, increasing from $c_z^2(0)$ to $c_z^2(1)$ when ρ is in the interval $(1 - \sqrt{a/n}, 1 - \sqrt{b/n})$, e.g., for $a = 100$ and $b = 0.01$. For example, if we were to consider the case $n = 10^6$, we anticipate that we would have

$$c_z^2(\rho) \approx c_z^2(0) = I_w(0) \quad \text{for} \quad \rho \leq 0.99 \tag{155}$$

and

$$c_z^2(\rho) \approx c_z^2(1) = I_w(\infty) \quad \text{for} \quad \rho \geq 0.9999. \tag{156}$$

Consistent with (151), we would predict $\dot{c}_z^2(0) \approx 0$, but inconsistent with (152), we would predict $\dot{c}_z^2(1) \approx 0$. However, clearly $\dot{c}_z^2(1) \approx 0$ does not describe the behavior of $c_z^2(\rho)$ well in a practically meaningful neighborhood of $\rho = 1.0$. Indeed, the IDW approximation (153), i.e., $\dot{c}_z^2(1) \approx \infty$ for H_2/E_2, seems to tell the main story.

VII. Conclusions

In this paper we have developed some new ways to gain insight into the offered traffic to a queue (the arrival process together with the service requirements), and its effect on performance. We have proposed two basic measurements of the offered traffic: the normalized mean workload $\{c_\lambda^2(\rho): 0 \leq \rho \leq 1\}$ in (3) (Section II-C) and the IDW $\{I_w(t): t \geq 0\}$ in (1) (Section III). Both are functions that can be estimated from a single sample path of the offered traffic; see (2) and Section II-B. Moreover, both measurements also apply to the arrival process alone, as well as to the arrival process together with the service requirements, essentially by considering constant service times.

Much of the paper has been devoted to establishing connections betwen $I_w(t)$ and $c_z^2(\rho)$. We have provided strong support for our basic premise that $c_z^2(\rho)$ is primarily determined by $I_w(t)$, starting with (8) and (12), and continuing with Section IV. We have also indicated limitations on this basic premise by considering different queues with identical IDWs. In particular, this occurs when the arrival process is a superposition of i.i.d. H_2 renewal processes, and we apply the GI/G/1 model approximation in Section V-C by redefining the third interarrival-time moment as in (143). The $H_2/G/1$ model obtained by fitting an H_2 distribution to the first three moments has exactly the same IDW as the queue with the superposition arrival process, but the resulting approximation performs only about as well as the interpolation approximation.

We have also proposed four ways to approximate $c_z^2(\rho)$ in terms of $I_w(t)$, which can be applied either with measurements or basic model parameters (as illustrated by Section VI). The first approximation is the simple time transformation in (9) and (13); the second is the variability fixed-point approximation in (9), (15), and Section V-A; the third is the interpolation approximation in (17), (18) and Section V-B; and the fourth is the GI/G/1 model approximation in Section V-C (which was not developed completely). We have found (in [24]–[26] and [60], as well as here) that these approximations perform quite well, but not uniformly so. All these procedures correctly capture the first-order behavior in the limiting values $c_z^2(0)$ and $c_z^2(1)$. For the middle, on the basis of simplicity as well as performance, we prefer the interpolation approximation (17) and (18), and the sample time transformation (14). All the approximations seem promising, but they require further study.

In predicting the function $\{c_z^2(\rho): 0 \leq \rho \leq 1\}$ given $\{I_w(t): t \geq 0\}$, we succeed in predicting the performance of the queue at *all* ρ given the offered traffic at only *one* ρ, but recall that this strong property requires that the offered traffic depend on ρ via the simple linear scaling in Section II-A. If the offered traffic is not exogenous (e.g., if there are feedback effects, as in a closed queueing network), then the offered traffic will *not* change by the simple linear scaling when we change the arrival rate or service rate. Then $c_z^2(\rho)$ may only be related to the IDW associated with the given ρ. The IDW should nevertheless provide useful insight.

Acknowledgment

This research was motivated by joint work with Vikram Saksena [24]–[26]; we thank him for his help and encouragement. We also thank Hong Chen and Paul Glasserman for helpful comments.

References

[1] J. Abate and W. Whitt, "Transient behavior of regulated Brownian motion, I: starting at the origin," *Adv. Appl. Prob.*, vol. 19, pp. 560–598, 1987.

[2] ——, "Transient behavior of the M/M/1 queue," *Queueing Syst.*, vol. 2, pp. 41–65, 1987.

[3] S. L. Albin, "Approximating queues with superposition arrival processes," Ph.D. dissertation, Columbia University, New York, NY, 1981.

[4] ——, "On Poisson approximations for superposition arrival processes in queues," *Management Sci.*, vol. 28, pp. 126–137, 1982.

[5] ——, "Approximating a point process by a renewal process, II: superposition arrival processes to queues," *Operations Res.*, vol. 32, pp. 1133–1162, 1984.

[6] A. O. Allen, *Probability Statistics and Queueing Theory*. New York, NY: Academic Press, 1978.

[7] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*. New York, NY: Holt, Rinehart & Winston, 1975.

[8] V. E. Beneš, *General Stochastic Processes in the Theory of Queues*. Reading, MA: Addison-Wesley, 1963.

[9] P. Billingsley, *Convergence of Probability Measures*. New York, NY: Wiley, 1968.

[10] N. H. Bingham, "Fluctuation theory in continuous time," *Adv. Appl. Prob.*, vol. 7, pp. 705–766, 1975.

[11] A. A. Borovkov, "Some limit theorems in the theory of mass service," *Theor. Prob. Appl.*, vol. 19, pp. 375–400, 1965.

[12] ——, *Stochastic Processes in Queueing Theory*. New York, NY: Springer-Verlag, 1976.

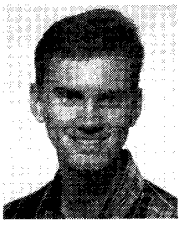
[13] D. Brillinger, "Comparative aspects of the study of ordinary time series and of point processes," in *Developments in Statistics*, vol. 1, P. R. Krishniah, Ed. New York, NY: Academic Press, 1978, pp. 33–133.

[14] M. Brown, "Bounds, inequalities, and monotonicity properties for some specialized renewal processes," *Ann. Probab.*, vol. 8, pp. 227–240, 1980.

[15] ——, "Further monotonicity properties for specialized renewal processes," *Ann. Probab.*, vol. 9, pp. 891–895, 1981.

[16] S. L. Brumelle, "On the relation between customer and time averages in queues," *J. Appl. Prob.*, vol. 8, pp. 508–520, 1971.

[17] D. Y. Burman and D. R. Smith, "Asymptotic analysis of a queueing model with bursty traffic," *Bell. System Tech. J.*, vol. 62, pp. 1433–1453, 1983.
[18] R. B. Cooper, *Introduction to Queueing Theory*, second Ed. Amsterdam: North-Holland, 1981.
[19] D. R. Cox, *Renewal Theory*. London, England: Methuen, 1962.
[20] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*. London, England: Methuen, 1966.
[21] D. J. Daley and T. Rolski, "Light traffic approximations for queues, II," Dept. of Statistics, University of North Carolina at Chapel Hill, 1988.
[22] A. E. Eckberg, "Generalized peakedness of teletraffic processes," in *Proc. Tenth. Int. Teletraffic Cong.*, Montreal, Canada, p. 4.4b.3, June 1983.
[23] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. II, Second Edition. New York, NY: Wiley, 1971.
[24] K. W. Fendick, V. R. Saksena, and W. Whitt, "Dependence in packet queues," *IEEE Trans. Commun.*, 1989, to appear.
[25] ——, "Characterizing dependence in packet queues using the index of dispersion for work," 1988, submitted for publication.
[26] ——, "Approximating the mean workload in packet queues," 1988, submitted for publication.
[27] P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*. Berlin, East Germany: Akademie-Verlag, 1981.
[28] E. Fuchs and P. E. Jackson, "Estimates of distributions of random variables for certain communication traffic models," *Commun. ACM*, vol. 13, pp. 752–757, 1970.
[29] P. W. Glynn and W. Whitt, "Ordinary CLT and WLLN versions of $L = \lambda W$," *Math. Opns. Res.*, vol. 13, pp. 674–692, 1988.
[30] S. Halfin, "Delays in queues, properties and approximations," in *Teletraffic Issues in an Advanced Information Society*, ITC 11, Kyoto, M. Akiyama, Ed. Amsterdam: Elsevier, 1985, pp. 47–52.
[31] J. M. Harrison, "The supremum distribution of a Lévy process with no negative jumps," *Adv. Appl. Prob.*, vol. 9, pp. 417–422, 1977.
[32] ——, *Brownian Motion and Stochastic Flow Systems*. New York, NY: Wiley, 1985.
[33] ——, "Brownian models of queueing networks with heterogeneous customer populations," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*. W. Fleming and P. L. Lions, Eds. New York, NY: Springer-Verlag, 1988, pp. 147–186.
[34] 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.
[35] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Selected Areas Commun.*, vol. SAC-4, pp. 856–868, 1986.
[36] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, vol. I. New York, NY: McGraw-Hill, 1982.
[37] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: sequences, networks, and batches," *Adv. Appl. Prob.*, vol. 2, pp. 355–369, 1970.
[38] D. P. Johnson, "Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks," Ph.D. dissertation, Dept. of Mathematics, University of Wisconsin, 1983.
[39] K. T. Marshall, "Some inequalities in queueing," *Operations Res.*, vol. 16, pp. 651–665, 1968.
[40] W. T. Marshall and S. P. Morgan, "Statistics of mixed data traffic on a local area network," in *Teletraffic Issues in an Advanced Information Society*, ITC 11, Kyoto, M. Akiyama, t.d. Amsterdam, The Netherlands: Elsevier, pp. 569–575, 1985.
[41] M. F. Neuts, "The caudal characteristic curve of queues," *Adv. Appl. Prob.*, vol. 18, pp. 221–254, 1986.
[42] G. F. Newell, *Applications of Queueing Theory*, second ed. London, England: Chapman and Hall, 1982.
[43] ——, "Approximations for superposition arrival processes in queues," *Management Sci.*, vol. 30, pp. 623–632, 1984.
[44] P. F. Pawlita, "Traffic measurements in data networks, recent measurement results, and some implications," *IEEE Trans. Commun.*, vol. 29, pp. 525–535, 1981.
[45] W. P. Peterson, "Diffusion Approximations for Networks of Queues with Multiple Customer Types," Ph.D. dissertation, Dept. of Operations Research, Stanford University, 1985.
[46] N. U. Prabhu, *Stochastic Storage Processes*. New York, NY: Springer-Verlag, 1980.
[47] V. Ramaswami and G. Latouche, "An experimental evaluation of the matrix-geometric method for the GI/PH/1 queue," Tech. Rep., Bell Communications Research, Morristown, NJ 1987.
[48] M. I. Reiman, "Open queueing networks in heavy traffic," *Math. Opns Res.*, vol. 9, pp. 441–458, 1984.
[49] M. I. Reiman and B. Simon, "An interpolation approximation for queueing systems with Poisson input," *Operations Res.*, vol. 36, pp. 454–469, 1988.
[50] M. I. Reiman and A. Weiss, "Light traffic derivatives via likelihood ratios," submitted for publication, 1988.
[51] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Selected Areas Commun.*, vol. SAC-4, pp. 833–846, 1986.
[52] W. Whitt, "Weak convergence theorems for priority queues: preemptive-resume discipline," *J. Appl Prob.*, vol. 8, pp. 74–94, 1971.
[53] ——, "Some useful functions for functional limit theorems," *Math. Opns. Res.*, vol. 5, pp. 67–85, 1980.
[54] ——, "Approximating a point process by a renewal process: the view through a queue, an indirect approach," *Management Sci.*, vol. 27, pp. 619–636, 1981.
[55] ——, "Approximating a point process by a renewal process, I: two basic methods," *Operations Res.*, vol. 30, pp. 124–147, 1982.
[56] ——, "Queue tests for renewal processes," *Opns. Res. Letters*, vol. 2, pp. 7–12, 1983.
[57] ——, "The queueing network analyzer," *Bell Syst. Tech. J.*, vol. 62, pp. 2779–2815, 1983.
[58] ——, "On approximations for queues, III: mixtures of exponential distributions," *AT&T Bell Lab. Tech. J.*, vol. 63, pp. 163–175, 1984.
[59] ——, "Queues with superposition arrival processes in heavy traffic," *Stoch. Proc. Appl.*, vol. 21, pp. 81–91, 1985.
[60] ——, "An interpolation approximation for the mean workload in the GI/G/1 queue," *Operations Res.*, vol. 37, 1989, to appear.
[61] R. I. Wilkinson, "Theories of toll traffic engineering in the U.S.A.," *Bell. Syst. Tech. J.*, vol. 35, pp. 421–514, 1956.
[62] R. W. Wolff, "Poisson arrivals see time averages," *Operations Res.*, vol. 30, pp. 223–231, 1982.
[63] V. M. Zolotarev, "The first passage time to a level and the behavior at infinity for a class of processes with independent increments," *Theor. Prob. Appl.*, vol. 9, pp. 653–662, 1964.



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

Since 1984 he has worked at AT&T Bell Laboratories, Holmdel, NJ. As a member of the Data Network Analysis Department, he works primarily on problems involving the traffic engineering of packet networks. He is especially interested in queueing problems, as well as other applications of probability theory.

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 Administrative Sciences at Yale University from 1969–1977. Since 1977, he has been employed by AT&T Bell Laboratories. He is currently a Member of Technical Staff in 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.