Simulation Run Lengths to Estimate Blocking Probabilities

RAYADURGAM SRIKANT
University of Illinois
and
WARD WHITT
AT&T Laboratories

We derive formulas approximating the asymptotic variance of four estimators for the steady-state blocking probability in a multiserver loss system, exploiting diffusion process limits. These formulas can be used to predict simulation run lengths required to obtain desired statistical precision before the simulation has been run, which can aid in the design of simulation experiments. They also indicate that one estimator can be much better than another, depending on the loading. An indirect estimator based on estimating the mean occupancy is significantly more (less) efficient than a direct estimator for heavy (light) loads. A major concern is the way computational effort scales with system size. For all the estimators, the asymptotic variance tends to be *inversely* proportional to the system size, so that the computational effort (regarded as proportional to the product of the asymptotic variance and the arrival rate) does not grow as system size increases. Indeed, holding the blocking probability fixed, the computational effort with a good estimator decreases to zero as the system size increases. The asymptotic variance formulas also reveal the impact of the arrival-process and service-time variability on the statistical precision. We validate these formulas by comparing them to exact numerical results for the special case of the classical Erlang M/M/s/0 model and simulation estimates for more general G/GI/s/0 models. It is natural to delete an initial portion of the simulation run to allow the system to approach steady state when it starts out empty. For small to moderately sized systems, the time to approach steady state tends to be negligible compared to the time required to obtain good estimates in steady state. However, as the system size increases, the time to approach steady state remains approximately unchanged, or even increases slightly, so that the computational effort associated with letting the system approach steady state becomes a greater portion of the overall computational effort as system size increases.

Categories and Subject Descriptors: C.4 [**Performance of Systems**]: Measurement Techniques; G.m [**Mathematics of Computing**]: Miscellaneous—*queueing theory*; I.6.8 [**Simulation and Modeling**]: Types of Simulation—*discrete event*

General Terms: Algorithms, Experimentation, Measurement, Performance, Theory

Additional Key Words and Phrases: Asymptotic variance, bias, blocking probabilities,

Authors' addresses: R. Srikant, Coordinated Science Laboratory, University of Illinois, 1308 W. Main Street, Urbana, IL 61801, email: srikant@shannon.csl.uiuc.edu; W. Whitt, AT&T Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636, email: wow@research.att.com.

diffusion approximations, estimation, experimental design, heavy-traffic limits, indirect estimation, loss models, Poisson's equation, reflected Ornstein-Uhlenbeck diffusion process, simulation, simulation run length

1. INTRODUCTION AND SUMMARY

In this article we consider the problem of estimating steady-state blocking probabilities in a multiserver loss system from simulation output or system measurements. We develop formulas approximating the variance of four candidate estimators. These variance formulas enable us to predict the observation intervals required to obtain estimates with desired statistical precision before collecting any data. Thus these formulas can help design experiments. These formulas should also be directly of interest to system designers because the variability of blocking, due to variability in the arrival process and service times, is an important performance measure in addition to the blocking probability itself.

We are interested in loss networks, as in Kelly [1991] and Ross [1995], but here we consider only a single link. Nevertheless, the results provide useful insights for loss networks. Here we focus on the G/GI/s/0 model, which has s servers in parallel, no extra waiting space, and independent and identically distributed (i.i.d.) service times that are independent of a general stationary arrival process (i.e., with stationary increments). Most of our analysis focuses on the special case of exponential (M) service-time distributions, but we also treat general (GI) service-time distributions. The results for general service times are more tentative, however, as explained later.

Arrivals that find all servers busy are lost (blocked) without affecting future arrivals. The goal is to determine the steady-state blocking probability, that is, the long-run proportion of all arrivals that are not admitted. In a simulation we assume that the data are collected after the system has reached steady state. Hence there is typically an initial period where the system is approaching steady state, over which no data are collected, and then a second period where we assume that the system is approximately in steady state, over which all relevant data are collected. We first consider the problem of predicting the required observation interval assuming that the system starts in steady state. Afterwards, we consider the initial portion that needs to be deleted for the system to be approximately in steady state when the system starts empty. Then we also consider alternative initial conditions to make the system approach steady state more quickly. When the number of servers is not large, the initial conditions do not matter much, but when the number of servers is large the initial conditions become important; see Section 11.

Indeed, a major goal is to better understand how the (lengths of the) data-collection interval and the initial transient interval (to be deleted) should scale as the model size (measured by the number of servers)

increases. We find that these intervals scale in very different ways. It may be somewhat surprising at first, but the asymptotic variance tends to be *inversely* proportional to system size, whereas the appropriate initial transient interval tends to be approximately *independent* of system size, or even increasing slowly in system size. (This property of the asymptotic variance becomes intuitively reasonable when one realizes that the amount of data collected over any given observation interval tends to be *directly* proportional to system size, assuming that the arrival rate is approximately proportional to system size.) This behavior implies that the steady-state portion of the computational effort in a simulation (regarded as proportional to the product of the asymptotic variance and the arrival rate) is approximately *independent* of system size. This also implies that the initial transient interval should become a greater portion of the overall interval as system size increases. For small to moderately sized systems, the initial transient interval tends to be negligible compared to the steady-state observation interval, but when the system gets large, the initial transient interval becomes significant and eventually even dominates.

A major feature of our model is the general stationary arrival process. Non-Poisson arrival processes often arise in loss systems; for example, because the input often contains overflows from other loss systems, as with various alternative routing schemes in circuit-switched telecommunications networks. For general stationary arrival processes, there are no analytical formulas available for the steady-state blocking probability, so that simulation is important. The estimation questions posed are interesting even for the special cases of renewal (GI) and Poisson (M) arrival processes, for which exact formulas for the steady-state blocking probability are available (e.g., see Section 2.1 of Takács [1962]), because results about the efficiency of different simulation estimators for these tractable models can provide insight into what to do with more general models.

The present paper is in the spirit of Whitt [1989], which carried out a similar investigation for queueing systems with unlimited waiting space (i.e., for delay systems), focusing primarily on the case of relatively few servers. There the object was to determine the amount of data required to estimate standard steady-state performance measures such as the mean waiting time and the mean queue length with desired statistical precision. The main idea was to obtain simple approximations by exploiting approximations of the queueing processes by *reflected Brownian motion* (RBM). For the multiserver loss model with exponential service times considered here, instead of RBM, the key process is the *reflected Ornstein-Uhlenbeck* (ROU) diffusion process, as we explain in Section 4. Further related work on delay systems appears in Asmussen [1989, 1992]. Whitt [1992] contributed further by providing formulas and algorithms for the asymptotic variance of the sample means of functions of birth-and-death processes and other Markov processes. (The substantial previous literature is reviewed there as well.) We draw upon the algorithms in Whitt [1992] here; they are reviewed in Section 3.

1.1 The Candidate Estimators

We now describe the four estimators that we consider. First, the *natural estimator* for the steady-state blocking probability B based on observations of the system over the time interval $[0, t]$ is

$$\hat{B}_N(t) \equiv L(t)/A(t), \tag{1}$$

where $L(t)$ is the number of lost (blocked) arrivals in $[0, t]$ and $A(t)$ is the total number of arrivals (admitted or blocked) in $[0, t]$. A closely related alternative *simple estimator*, whose efficiency is easier to analyze, is

$$\hat{B}_S(t) \equiv \frac{L(t)}{EA(t)} = \frac{L(t)}{\lambda t}, \tag{2}$$

where $\lambda \equiv EA(1)$ is the *arrival rate* (or intensity). It is intuitively clear that the estimators $\hat{B}_N(t)$ and $\hat{B}_S(t)$ should behave similarly; we substantiate this intuition analytically in Section 7 and via simulation experiments in Section 10. Hence we regard results for $\hat{B}_S(t)$ as being applicable to $\hat{B}_N(t)$.

As in Law [1975], Carson and Law [1980], and Glynn and Whitt [1989], we can exploit the conservation law $L = \lambda W$ (*Little's law*) to obtain an alternative indirect estimator for B. For this purpose, let μ^{-1} be the *mean service time*, $\alpha \equiv \lambda/\mu$ the *offered load*, $N(t)$ the *number of busy servers at time t* (which we assume is stationary, due to deleting an initial portion of the run), and $n \equiv EN(t)$ is the steady-state mean number of busy servers. Applying the relation $L = \lambda W$, as in Example 4.3 of Whitt [1991b], we get the relation $n = \lambda(1 - B)/\mu$ or, equivalently,

$$B = 1 - \frac{n}{\alpha}. \tag{3}$$

Assuming that we know λ and μ, as would be the case with many simulations, we can use the *indirect estimator*

$$\hat{B}_I(t) \equiv 1 - \frac{\hat{n}(t)}{\alpha}, \tag{4}$$

where $\hat{n}(t)$ is an estimator of n based on data over $[0, t]$. In particular, we assume that $\hat{n}(t)$ is the sample mean, that is

$$\hat{n}(t) = t^{-1}\int_0^t N(u)du, \qquad t \geq 0. \tag{5}$$

Closely related to the blocking probability B is the probability that all servers are busy, $P(N(t) = s)$, which is often called the *time congestion*. Indeed, for Poisson arrivals these two quantities are equal, by virtue of the

PASTA property; see Wolff [1982], Melamed and Whitt [1990], and Baccelli and Brémaud [1994]. However, they are not equal more generally. The natural *time congestion estimator* is

$$\hat{B}_T(t) = t^{-1} \int_0^t 1_{\{N(u)=s\}} du, \tag{6}$$

where 1_A is the indicator random variable of the set A; that is, $1_A(\omega) = 1$ if $\omega \in A$ and $1_A(\omega) = 0$ otherwise, where ω is an underlying sample point. Glynn et al. [1993] made a general study of estimators for time and customer averages, but did not focus on loss systems.

In this article we consider only the four estimators $\hat{B}_N(t)$, $\hat{B}_S(t)$, $\hat{B}_I(t)$, and $\hat{B}_T(t)$ in eqs. (1), (2), (4), and (6). We investigate other estimators designed to reduce variance in Srikant and Whitt [1995]. For example, because $B_I(t)$ is decreasing in $\hat{n}(t)$, and $B_N(t)$ should tend to be increasing in $\hat{n}(t)$, $\hat{B}_I(t)$ and $\hat{B}_N(t)$ tend to be negatively correlated, so that it is natural to consider a *combined estimator* $\hat{B}_C(t) = p\hat{B}_I(t) + (1 - p)\hat{B}_N(t)$ for appropriate p, which can be estimated during the run.

Similarly, as in Lavenberg et al. [1982], Glynn and Whitt [1989], and references cited there, it is natural to consider *linear-control-variable estimators* such as $\hat{B}_L(t) = \hat{B}_N(t) + a_1(\hat{\lambda}(t) - \lambda) - a_2(\hat{\mu}(t) - \mu)$, where $\hat{\lambda}(t) \equiv t^{-1} A(t)$ and $\hat{\mu}(t)$ is an estimate of the individual service rate such as the reciprocal of the sample mean of the service times used during the run. We discuss these alternative estimators in the other paper.

1.2 The Asymptotic Variance

Here we concentrate on predicting the variance of the basic estimators $\hat{B}_N(t)$, $\hat{B}_S(t)$, $\hat{B}_I(t)$, *and* $\hat{B}_T(t)$. We address this problem by focusing on the asymptotic variance. For any estimator $\hat{B}(t)$, its *asymptotic variance* is defined as

$$\sigma^2 = \lim_{t \to \infty} t \operatorname{Var} \hat{B}(t). \tag{7}$$

We use subscripts N, S, I, and T to refer to the specific estimators defined previously. Under regularity conditions (which include the requirement that the asymptotic variance actually be positive and finite), for suitably large run times t, each estimator $\hat{B}(t)$ tends to be approximately normally distributed with mean B and variance σ^2/t, where σ^2 is the asymptotic variance (which depends on the estimator); see Section 2.1 of Whitt [1989] for a review of the standard statistical theory. Hence a $(1 - \beta)$ 100% confidence interval for B will be $[\hat{B}(t) - h(\beta), \hat{B}(t) + h(\beta)]$ with halfwidth

$$h(\beta) = \frac{\sigma z_{\beta/2}}{\sqrt{t}}, \tag{8}$$

where $P(-z_{\beta/2} \leq N(0, 1) \leq z_{\beta/2}) = 1 - \beta$ with $N(0, 1)$ a standard (mean 0, variance 1) normal random variable. (For example, if we use a 90% confidence interval, then $z_{\beta/2} = 1.645$.) Thus for *specified halfwidth* ϵ and *level of precision* β, the *required simulation run length* is

$$t(\epsilon, \beta) = \frac{\sigma^2 z_{\beta/2}^2}{\epsilon^2}. \tag{9}$$

From eq. (9) we see that the required run length is directly proportional to the asymptotic variance σ^2 and inversely proportional to the *square* of the specified confidence-interval halfwidth ϵ. Clearly the specified halfwidth ϵ is a key factor, but the asymptotic variance σ^2 can be important as well. Note that the asymptotic variance is the only quantity in eq. (9) that is not known to the experimenter.

We emphasize that our analysis presumes that the observation interval length t is sufficiently long that the standard asymptotic theory implying that $\hat{B}(t)$ is approximately normally distributed with mean B and variance σ^2/t is appropriate. See Section 4.5 of Whitt [1989] and Asmussen [1989, 1992] for further discussion.

We aim to develop approximations for the asymptotic variances σ_S^2, σ_N^2, σ_T^2, and σ_I^2. Roughly speaking we find that σ_S^2, σ_N^2, and σ_T^2 are approximately the same, but that σ_I^2 can be quite different from the others. In particular, we find that *each of the estimators* $\hat{B}_S(t)$ *and* $\hat{B}_I(t)$ *has a region where it is much more efficient.* In particular, we tend to have $\sigma_I^2 < \sigma_S^2$ when $\alpha > s$, whereas we tend to have $\sigma_1^2 > \sigma_S^2$ when $\alpha < s$. (They tend to be about the same when $\alpha \approx s$.) A similar conclusion was found by Ross and Wang [1992] for indirect estimation in the context of Monte Carlo summation.

1.3 Characterizing Model Variability

One of our goals is to determine how the model variability (the variability in the arrival process and service times) affects the asymptotic variance of the blocking estimators. One of our principal conclusions is that it is appropriate to partially characterize the variability of the arrival process through its *normalized arrival asymptotic variance*, defined by

$$c_a^2 = \lim_{t \to \infty} \frac{\text{var}A(t)}{\lambda t}, \tag{10}$$

which we assume is well defined (the limit exits and is finite). The parameter c_a^2 is the asymptotic variance of the arrival-rate estimator $\hat{\lambda}(t) \equiv A(t)/t$ divided by the arrival rate λ. The parameter c_a^2 in eq. (10) is the *limiting value of the index of dispersion for counts*, for example, see Fendick and Whitt [1989] and references cited there. For a deterministic evenly spaced (D) process, $c_a^2 = 0$; for a Poisson process, $c_a^2 = 1$. For the special case of a renewal process, c_a^2 coincides with the *squared coefficient*

of variation (SCV) of an interarrival time; that is, if U is an interarrival time, then

$$c_a^2 = \text{Var}\,(U)/(EU)^2. \tag{11}$$

However, eq. (11) is only true for renewal processes. Formula (10) captures correlations among different interarrival times in nonrenewal processes. A large class of nonrenewal arrival processes can be represented as batch Markovian arrival processes (BMAPs) or versatile Markovian point processes; see Neuts [1989] and Lucantoni [1993]. The normalized arrival asymptotic variance of a BMAP is given on p. 284 of Neuts [1989].

In applications, the general arrival process $A(t)$ is often a superposition of independent processes. If the component processes are independent Poisson processes, then the superposition process is a Poisson process, and $c_a^2 = 1$. More generally, the normalized arrival asymptotic variance of the superposition process (with independent component processes) is a convex combination of the normalized arrival asymptotic variances of the component processes; that is, if $A(t) = A_1(t) + \ldots + A_n(t)$, where $A_i(t)$ has arrival rate λ_i and normalized arrival asymptotic variance c_{ai}^2, then

$$c_a^2 = \sum_{i=1}^{n} (\lambda_i/\lambda) c_{ai}^2; \tag{12}$$

for example, see Section III.E of Fendick and Whitt [1989].

Because we have assumed that the service times are i.i.d. and independent of the arrival process, their variability is easier to characterize. Our theoretical results are primarily for the case of exponential service times, but we also develop approximation formulas for nonexponential service times. These are more empirical, and so should be regarded as more tentative. We primarily characterize the service-time variability via the service-time SCV, denoted by c_s^2, and defined as in eq. (11).

In previous studies of G/GI/s/0 loss systems it has been found that the model variability can be usefully characterized by focusing on the associated G/GI/∞ infinite-server model, with the same arrival process and service times; see Eckberg [1983], Whitt [1984], and p. 338 of Neuts [1989]. Because the infinite-server model is much easier to analyze, it can provide useful insight into the loss model. In particular, the G/GI/s/0 model variability can be partially characterized by the peakedness parameter z. The *peakedness* is defined as the ratio of the variance to the mean number of busy servers in the associated G/GI/∞ model. For example, if the arrival process is a renewal process with interarrival time U and the service-time distribution is exponential with μ^{-1}, then the peakedness is

$$z = [1 - \phi(\mu)]^{-1} - \alpha, \tag{13}$$

where $\phi(s) = Ee^{-sU}$; see Eckberg [1983]. The expressions are more complicated for nonrenewal arrival processes. However, the peakedness for

the MMPP/M/∞ model with a Markov modulated Poisson process (MMPP) as an arrival process is given on p. 338 of Neuts [1989].

It is often convenient and appropriate to use the heavy-traffic (large α) approximation for the peakedness with a general stationary arrival process and a general service-time cumulative distribution function (cdf) H(t), which is

$$z = 1 + (c_a^2 - 1)\mu \int_0^\infty [1 - H(t)]^2 dt. \tag{14}$$

When the service time cdf H in eq. (14) is exponential, $z = (c_a^2 + 1)/2$; when H is deterministic, $z = c_a^2$; see p. 692 of Whitt [1984]. Note that $z = 1$ in eq. (14) for all service-time distributions when $c_a^2 = 1$. Also note that the influence of the service-time distribution on z in eq. (14) is not determined by its first two moments. A very rough approximation for eq. (14) is $z \approx 1 + (c_a^2 - 1)/(1 + (c_s^2 \wedge 1))$, where $x \wedge y = \min\{x, y\}$. (Note that this approximation is exact for M and D service times.)

In summary, we partially characterize the variability of the G/GI/s/0 model via the parameter triple (c_a^2, c_s^2, z). A principal conclusion of our analysis is that this is indeed an appropriate partial characterization for the blocking probability and the asymptotic variance of the simulation estimators. When the model is not too far from M/M/s/0, the model variability as partially characterized by the triple (c_a^2, c_s^2, z) will usually be of secondary importance. However, there is a growing interest in different kinds of variability. Highly variable non-Poisson traffic is becoming more common in communication networks, as can be seen from Erramilli et al. [1994], Willinger [1995], and references there, so that it is even possible to have $c_a^2 = \infty$ or $c_s^2 = \infty$. For the most part, this non-Poisson traffic is packet traffic, but in some instances connection requests may also deviate significantly from Poisson; see Paxson and Floyd [1995]. Although we do not directly consider this phenomenon, our results provide useful insight into its consequences. In particular, more highly variable traffic requires longer observation intervals to estimate blocking probabilities reliably. Moreover, more variability means that the observed blocking should be more variable. Finally, we provide formulas that quantify these effects.

1.4 Scaling as System Size Grows

We are especially interested in the way the performance of the different estimators scales as the system size grows. Previous experience has shown that when s grows there are three distinct regions for loss models: light loading, normal (or critical) loading, and heavy loading. As in Jagerman [1974], Borovkov [1976, 1984], Halfin and Whitt [1981], Whitt [1984], Mitra and Weiss [1989], and other studies, the region depends on the way the *traffic intensity* $\rho \equiv \alpha/s$ changes as $s \to \infty$. If $(1 - \rho)\sqrt{s}$ or, equivalently, $(s - \alpha)/\sqrt{\alpha}$ approaches $+\infty$, a constant, or $-\infty$ as $s \to \infty$, then the region is light, normal, or heavy loading, respectively. The region of primary interest

is usually normal loading, but all three regions are important. First, real systems are often designed to be lightly loaded instead of normally loaded to provide some safety factor, for example, to allow for forecasting uncertainty. Second, performance in heavy loading arises when studying the response to overloads, for example, due to system failures.

1.5 Approximations for the Blocking Probabilities

In order to help judge what statistical precision is appropriate, it is useful to have rough approximations for the blocking probability and time congestion themselves. Asymptotics for the GI/M/s/0 model in the case of normal loading has produced the following approximation for the blocking probability:

$$B \approx \sqrt{z/\alpha}\frac{\phi(\gamma/\sqrt{z})}{\Phi(-\gamma/\sqrt{z})}, \tag{15}$$

where $\gamma = (\alpha - s)/\sqrt{\alpha}$, $z = (c_a^2 + 1)/2$, and $\alpha \equiv \lambda/\mu$ is the offered load, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of the standard (mean 0, variance 1) normal distribution; see (13) of Whitt [1984]. Formula (15) is asymptotically correct as $s \to \infty$ *with* $(\alpha - s)/\sqrt{\alpha} \to \gamma$; see p. 226 of Borovkov [1976]. Table I compares eq. (15) to exact values for the M/M/s/0 model. The performance is quite good for $\gamma \geq -3$.

A related approximation for the time congestion is

$$P(N(t) = s) \approx z^{-1/2}B \approx (1/\sqrt{\alpha})\frac{\phi(\gamma/\sqrt{z})}{\Phi(-\gamma/\sqrt{z})}. \tag{16}$$

With Poisson arrivals, the time congestion coincides with the blocking probability, so that eq. (16) is also asymptotically correct for the M/M/s/0 model, but otherwise we have no asymptotic correctness result supporting eq. (16). Formula (16) differs from (9) in Whitt [1984] by a factor of $z^{-1/2}$. We propose eq. (16) because it is more consistent with simulation results. Formulas (15) and (16) indicate that B and $P(N(t) = s)$ should both be $O(1/\sqrt{s})$ as s gets large in normal loading.

Approximations (15) and (16) are most strongly supported in the case of exponential service times, but we suggest using them as rough approximations with general service times, using the appropriate peakedness z. The best value for z should be the exact peakedness, but eq. (14) is a convenient approximation. Finally, we note that the reliability of the approximations is likely to deteriorate when z becomes very large.

1.6 Workload Factors

Formula (9) shows that the required simulation time t to achieve desired statistical precision is approximately proportional to the asymptotic variance σ^2. However, the *computational effort* required to simulate for time t is approximately proportional to λt, because λt is the expected number of

Table I.

	workload factors			blocking probabilities	
γ	indirect	simple	time congestion	exact	approximation (15)
−6.0	2.00	$.13 \times 10^{-7}$	$.11 \times 10^{-7}$	$.182 \times 10^{-8}$	$.35 \times 10^{-9}$
−5.5	2.00	$.16 \times 10^{-6}$	$.14 \times 10^{-6}$	$.214 \times 10^{-7}$	$.62 \times 10^{-8}$
−5.0	2.00	$.18 \times 10^{-5}$	$.15 \times 10^{-5}$	$.210 \times 10^{-6}$	$.84 \times 10^{-7}$
−4.5	2.00	$.16 \times 10^{-4}$	$.14 \times 10^{-4}$	$.170 \times 10^{-5}$	$.89 \times 10^{-6}$
−4.0	2.00	$.12 \times 10^{-3}$	$.11 \times 10^{-3}$	$.114 \times 10^{-4}$	$.74 \times 10^{-5}$
−3.5	1.98	$.79 \times 10^{-3}$	$.72 \times 10^{-3}$	$.623 \times 10^{-4}$	$.48 \times 10^{-4}$
−3.0	1.94	.0042	.0039	$.278 \times 10^{-3}$	$.24 \times 10^{-3}$
−2.5	1.83	.0183	.0173	.00100	.00094
−2.0	1.59	.063	.060	.00295	.00290
−1.5	1.25	.171	.164	.00712	.00720
−1.0	.88	.36	.34	.0144	.0147
−0.5	.56	.60	.58	.0251	.0258
0.0	.33	.86	.82	.0389	.0399
+0.5	.19	1.09	1.04	.0550	.0563
+1.0	.109	1.28	1.20	.0728	.0744
+1.5	.063	1.41	1.32	.0917	.0934
+2.0	.037	1.51	1.40	.111	.113
+2.5	.023	1.57	1.44	.131	.133
+3.0	.014	1.62	1.47	.150	.152
+3.5	.0091	1.65	1.48	.170	.172
+4.0	.0060	1.66	1.47	.189	.191
+4.5	.0041	1.67	1.46	.208	.210
+5.0	.0028	1.67	1.45	.227	.229
+5.5	.0020	1.67	1.43	.245	.247
+6.0	.0014	1.67	1.40	.263	.262

Canonical workload factors $\psi(\gamma)$ for three estimators and blocking probabilities B for the M/M/s/0 model as a function of $\gamma = (\alpha - s)/\sqrt{\alpha}$ based on exact numerical results for the case $s = 400$ and $\mu = 1$.

arrivals in $[0, t]$. (See Glynn and Whitt [1992] for a study relating computational effort to statistical precision in simulation experiments. There it is explained why it suffices to look at the rate of expected computational effort λ.) Hence we give formulas for $w \equiv \lambda\sigma^2$, which we call the *workload factor*.

However, we recognize that λt is only a rough approximation for the expected work to produce a run of length t. Moreover, the expected work may differ for different estimators. For example, in a Markovian simulation, with estimators $\hat{B}_S(t)$ and $\hat{B}_N(t)$ we can work with the embedded process obtained by looking at $N(t)$ only at transition epochs, without generating the times between transitions. In contrast, the estimators $\hat{B}_I(t)$ and $\hat{B}_T(t)$ require the transition times too.

2. MAIN RESULTS

Our main results are approximate expressions for the workload factors associated with the four estimators $\hat{B}_N(t)$, $\hat{B}_S(t)$, $\hat{B}_I(t)$, and $\hat{B}_T(t)$.

2.1 Canonical Workload Factors

We find that the workload factors in the G/GI/s/0 model primarily depend upon the parameter five-tuple $(s, \gamma, c_a^2, c_s^2, z)$ and moreover that they can be expressed as scaled versions of functions of a single real variable, which we call the canonical workload factors. In particular, for the indirect estimator, the key workload approximation formula is

$$w_I(s, \gamma, c_a^2, c_s^2, z) \approx \frac{(c_a^2 + c_s^2)}{2} \psi_I(\gamma/\sqrt{z}), \tag{17}$$

where $\psi_I(\gamma) \equiv w_I(\infty, \gamma, 1, 1, 1)$ is the *canonical workload factor* associated with the M/M/s/0 special case (the limit as $s \to \infty$), $\gamma = (\alpha - s)/\sqrt{\alpha}$, z is the peakedness, c_a^2 is the normalized arrival asymptotic variance in eq. (10), and c_s^2 is the SCV of the service-time distribution, defined as in eq. (11). Note that formula (17) has important content even in the M/M/s/0 case, indicating that $w_I(s, \gamma, 1, 1, 1) \approx \psi_I(\gamma)$ for all s (which we discuss further in the following). Note that the arrival-process variability enters into eq. (17) via both c_a^2 and z, and that the service-time distribution enters in via both c_s^2 and z. As with eqs. (15) and (16), the preferred peakedness z is the exact value, but eq. (14) usually is a satisfactory approximation. (For an exception, see Example 10.6.) In the G/M/s/0 model with eq. (14) for z, $c_s^2 = 1$ and $z = (c_a^2 + c_s^2)/2$, so that $w_I \approx z\psi_I(\gamma/\sqrt{z})$, but this is *not* true for other G/GI/s/0 models.

It is instructive to see what eq. (17) says for the M/GI/s/0 model. With Poisson arrivals, there is an insensitivity property implying that the steady-state distribution depends on the service-time distribution only through its mean. Thus we must have $z = 1$, in eq. (17). However, the service-time variability still has an influence on the workload factor w_I through the SCV c_s^2 in the factor $(c_a^2 + c_s^2)/2$. This is not a contradiction, because the time-dependent behavior of the M/GI/s/0 model does *not* have the insensitivity property; see Davis et al. [1995] and Section 6.1 here, especially eq. (48).

The approximation we propose for the workload factor of the simple estimator has the same form; just replace the two I subscripts in eq. (17) by S. Because $\hat{B}_N(t) \approx \hat{B}_S(t)$, we propose approximating w_N *by* w_S. Numerical evidence indicates that the workload factor for the time-congestion estimator approximately takes the form

$$w_T(s, \gamma, c_a^2, c_s^2, z) \approx \left(\frac{c_a^2 + (c_s^2 \vee 1)}{c_a^2 + 1}\right) \psi_T(\gamma/\sqrt{z}), \tag{18}$$

where $x \vee y = \max\{x, y\}$. Note that the prefactors of ψ differ in eqs. (17) and (18), but both become 1 in the M/M/s/0 special case. In general, we regard the approximations to account for the impact of variability on the time-congestion estimator in eqs. (16) and (18) as less reliable than the approximations for the other estimators. As with formulas (15) and (16),

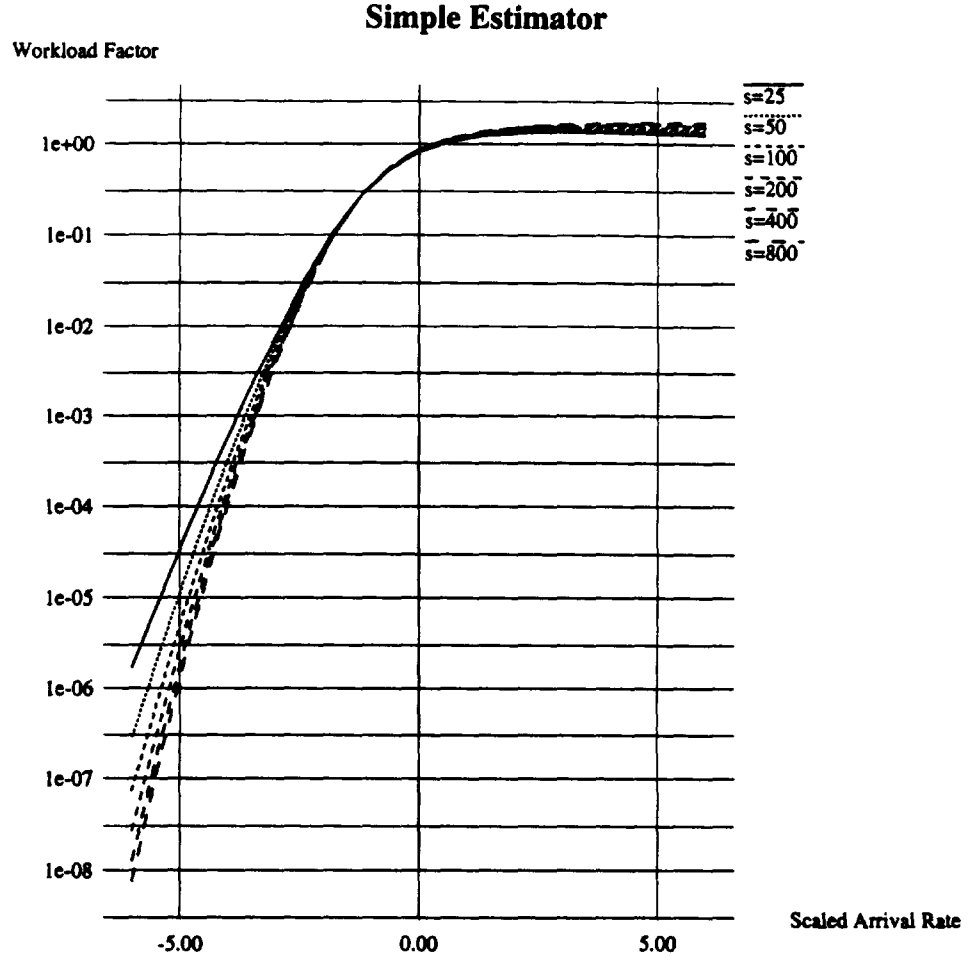

Fig. 1. Workload factors $w_S \equiv \lambda \hat{\sigma}_S^2$ for the simple estimator $\hat{B}_S(t)$ in the M/M/s/0 loss model with $\mu = 1$ as a function of the scaled arrival rate $\gamma \equiv (\alpha - s)/\alpha^{1/2}$ for several values of s.

approximations (17) and (18) should be regarded as less reliable when the model variability, as measured by c_a^2, c_s^2 or z, is high.

The notion of a canonical workload curve for M/M/s/0 models is supported by Figures 1–3, which display the *exact* workload factors $w(s, \gamma, 1, 1, 1)$ as functions of γ for the estimators $\hat{B}_S(t)$, $\hat{B}_I(t)$, and $\hat{B}_T(t)$ in the M/M/s/0 model for different values of s, assuming that $\mu = 1$ (computed by the methods of Section 3). These workload factors are plotted in log scale to emphasize significant differences. Note that the workload curves for different s in each figure essentially fall on top of each other when the scaled arrival rate $\gamma \equiv (\alpha - s)/\sqrt{\alpha}$ is not too far from 0 (e.g., $-2 \leq \gamma \leq 2$) or s is sufficiently large (e.g., $s \geq 200$). Hence a workload curve for one value of s can serve as a workload curve for all values of s (not too small) for that estimator. Table I provides canonical workload values for these three

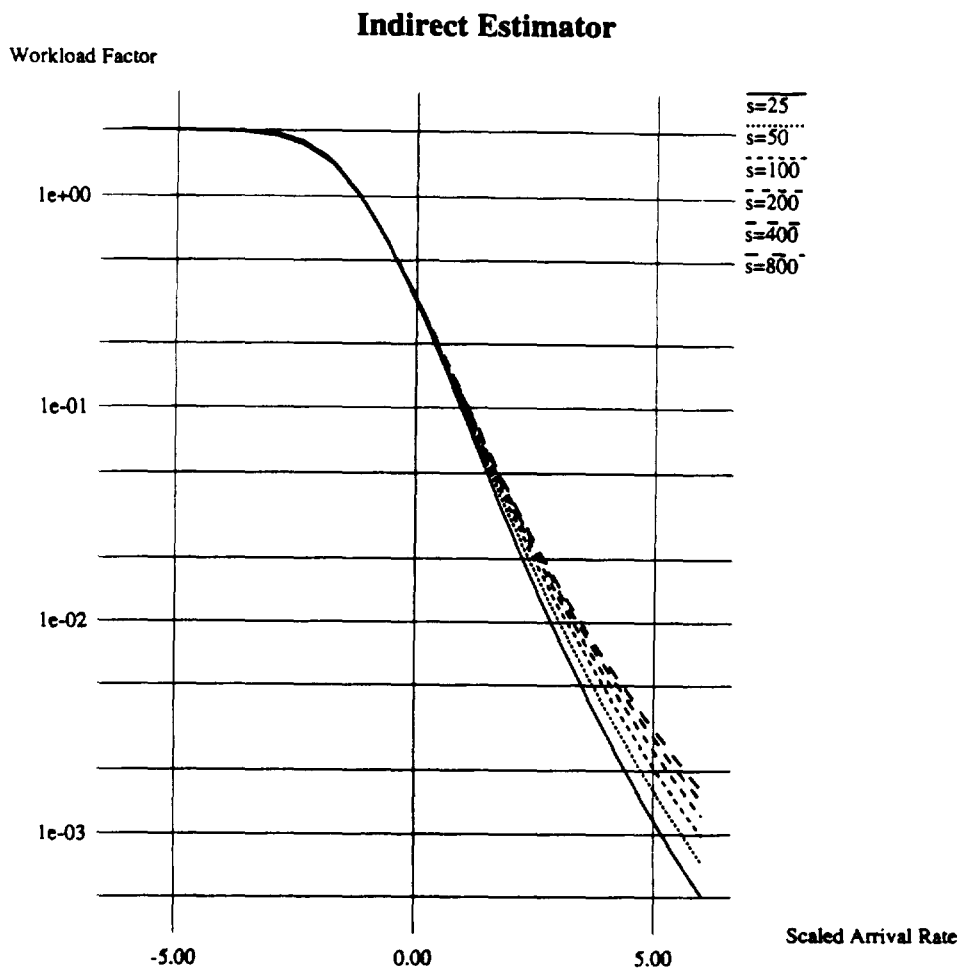

Fig. 2. Workload factors $w_I \equiv \lambda \hat{\sigma}_I^2$ for the indirect estimator $\hat{B}_I(t)$ in the M/M/s/0 loss model with $\mu = 1$ as a function of the scaled arrival rate $\gamma = (\alpha - s)/\alpha^{1/2}$ for several values of s.

estimators based on the numerical results for the M/M/s/0 model with $s = 400$. [To put the blocking probability approximation (15) in perspective, Table I also compares eq. (15) to the exact blocking probabilities.] Hence the canonical workload factors can be obtained from Figures 1–3, Table I, or the algorithms in Section 3.

Note that $\psi_I(\gamma)$ is small for $\gamma > 0$, whereas $\psi_S(\gamma)$ and $\psi_T(\gamma)$ are small for $\gamma < 0$, showing that *different estimators should be strongly preferred in different regions*. Figures 2 and 3 show that $\psi_S(\gamma)$ and $\psi_T(\gamma)$ are quite similar, but numerical evidence indicates that $\psi_S(\gamma) \to 1$ as $\gamma \to \infty$, whereas $\psi_T(\gamma) \to 0$ as $\gamma \to \infty$.

For loss systems in normal loading, a reasonable rough approximation for all the workload factors is 1. *This implies that simulation run lengths should be approximately inversely proportional to the arrival rate or the*

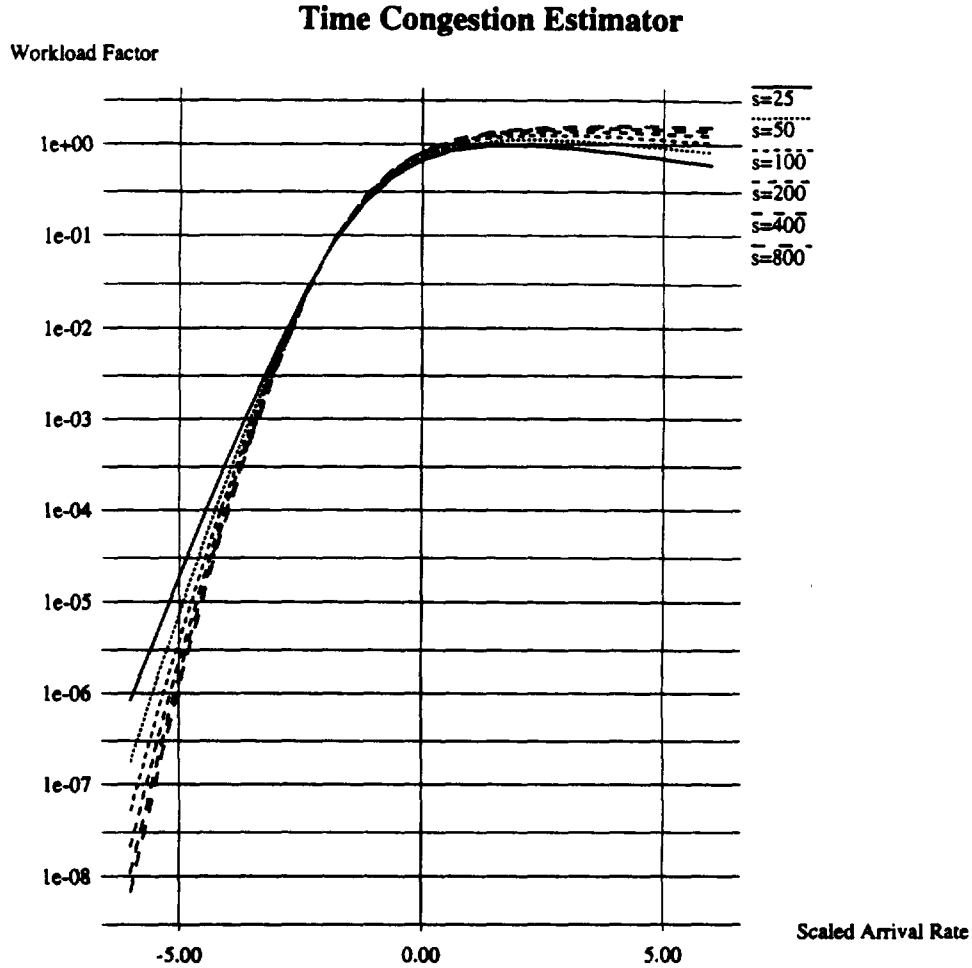

Fig. 3. Workload factors $w_T \equiv \lambda \hat{\sigma}_T^2$ for the time-congestion estimator $\hat{B}_T(t)$ in the M/M/s/0 loss model with $\mu = 1$ as a function of the scaled arrival rate $\gamma \equiv (\alpha - s)/\alpha^{1/2}$ for several values of s.

system size. Clearly, larger s means that more arrivals have to be generated, but these additional arrivals evidently help with the statistical precision, so that the asymptotic variance is inversely proportional to λ as s (and thus λ) get large.

2.2 A Supporting Diffusion Limit

We also provide theoretical support for the workload factor approximations in eqs. (17) and (18). In Section 4 we show for the G/M/s/0 model that the normalized process $(N_s(\cdot) - s)/\sqrt{\alpha}$ converges to the ROU diffusion process as $s \to \infty$ with $(\alpha - s)/\sqrt{\alpha} \to \gamma$. [You could also think of $N_s(\cdot)$ as being indexed by α and centered at α.] Moreover, we show how this limit supports

Table II.

	light loading $\rho < 1$	heavy loading $\rho > 1$
simple and natural estimators	$B\left(\frac{1+c_a^2}{1-\rho}\right)$	$c_a^2 + \rho^{-1}$
indirect estimator	$1 + c_a^2$	$\frac{(1+\rho)(1+c_a^2)^3}{4s^2(\rho-1)^4}$

Approximation formulas for the workload factor $w \equiv \lambda\sigma^2$ of the estimators in eqs. (1), (2), and (4) for the G/M/s/0 model in light and heavy loading.

the blocking approximation in eq. (15) and the workload factor approximation in eq. (17).

Our limit theorem supplements related limit theorems in Borovkov [1976, 1984]. For the case of the blocking approximation (15), Borovkov [1976] treated the GI/M/s/0 model, whereas our result applies to the more general G/M/s/0 model. Borovkov [1984] also has limit results for the process $N_s(t)$ in the general G/M/s/0 model, but with different conditions. In that general setting, he did not treat the blocking probability.

2.3 Light-Loading and Heavy-Loading Approximations

We also develop other approximations in Sections 5–7 based on asymptotics as $s \to \infty$ with ρ held fixed, with either $\rho < 1$ (light loading) or $\rho > 1$ (heavy loading). These approximations are shown in Table II. These formulas show that the workload factors w_I and w_S behave differently: $w_I/w_S \to \infty$ as $s \to \infty$ for $\rho < 1$, whereas $w_I/w_S \to 0$ as $s \to \infty$ for $\rho > 1$. Moreover, these formulas also serve as simple approximations. Because we already have reduced the G/GI/s/0 case to the M/M/s/0 case in eqs. (17) and (18), we primarily use the formulas in Table II as convenient simple approximations for the canonical (M/M/s/0) workload factors ψ (obtained by letting $c_a^2 = 1$ in Table II).

With the exception of the light-loading simple-estimator formula, the formulas in Table II are all in terms of the three variables s, γ, and c_a^2. (Given s, γ is equivalent to ρ or α.) The light-loading simple-estimator formula can be put in the same form by exploiting eq. (15), which yields

$$w_S(s, \gamma, c_a^2) \approx \frac{(1+c_a^2)^{3/2}}{-2\rho\gamma\sqrt{\pi}} e^{-\gamma^2/(1+c_a^2)} \quad \text{in light loading.} \tag{19}$$

[Let $w(s, \gamma, c_a^2) \equiv w(s, \gamma, c_a^2, 1, (c_a^2 + 1)/2)$.]

Table II is especially important for providing theoretical support for the simple-estimator workload factor, because the diffusion limit in Section 4

only applies directly to the indirect estimator. Note that formula (19) approximately satisfies the general functional form (17) with

$$\psi_S(\gamma) = 2\gamma^{-1}\phi(\gamma) = \sqrt{2/\pi\gamma^2}e^{-\gamma^2/2}. \tag{20}$$

Similarly, the heavy-traffic indirect-estimator workload factor approximation in Table II can be expressed approximately as

$$w_I(s,\ \gamma,\ c_a^2) \approx \frac{(1 + c_a^2)^3}{2\rho^4\gamma^4}, \tag{21}$$

which is approximately of the form (17) with

$$\psi_I(\gamma) = 4\gamma^{-4}. \tag{22}$$

To see this consistency, we need to include the term $\sqrt{z}$ in eq. (17). Recall that $z = c_a^2 + 1$ here.

Turning to the heavy-loading formula for the simple estimator and the light-loading formula for the indirect estimator, we note that these formulas are consistent with eq. (17) because $\psi_S(\gamma/\sqrt{z}) \to 2$ as $\gamma \to +\infty$ and $\psi_I(\gamma/\sqrt{z}) \to 2$ as $\gamma \to -\infty$ for the M/M/s/0 model (corresponding to $s \to \infty$ with fixed ρ as in Table II); see Table I.

Approximations (19)–(22) reveal the essential form of the workload factors in light and heavy loading, but these formulas are not very accurate, for example, when compared to the exact M/M/s/0 results in Table *I*. The approximations tend to be conservative (on the high side) though; see Sections 5 and 6.

2.4 Relative Statistical Precision

It is important to note that our discussion so far has been based on the tacit assumption that we are using the criterion of absolute statistical precision. *The computational effort necessarily does grow with s in light and normal loading if we use relative statistical precision.* With relative statistical precision, the workload factor becomes $\lambda\sigma^2/B^2$. In light loading, dividing by B^2 causes the workload factor to explode as $s \to \infty$, because B approaches 0 exponentially fast as $s \to \infty$ under light loading. The same phenomenon occurs, but less dramatically, in normal loading, because then $B = O(1/\sqrt{s})$ as $s \to \infty$, as noted previously.

It is natural to ask how the computational effort grows with s if we fix the blocking probability. However, for any fixed (positive) blocking probability, we are eventually in heavy loading when s is large enough, because ρ^{-1} approaches $1 - B$ as s increases with B fixed. (Equivalently, B approaches $1 - \rho^{-1}$ as s increases with ρ fixed, by the law of large numbers; see p. 90 of Borovkov [1984].) Then the computational effort for the simple estimator does not grow, whereas the computational effort for the indirect estimator actually decreases to 0. So, even with relative statistical preci-

sion, *the computational effort using the best estimator ultimately decreases to 0 as* $s \to \infty$ *if we fix the blocking probability.*

2.5 Initial Conditions

Because we cannot start the simulation in steady state, the estimators necessarily have initialization bias, that is, the expected value is not exactly B. The *bias* of estimator $\hat{B}(t)$ is $E\hat{B}(t) - B$. The bias can be kept small by choosing a good initial state and/or not collecting data over an initial portion of the simulation to allow the system to approach steady state.

First, we can approximate the bias at time t by using the *asymptotic bias*, which is defined by

$$\beta = \lim_{t \to \infty} t(E\hat{B}(t) - B). \tag{23}$$

We use eq. (23) to justify the approximation $E\hat{B}(t) - B \approx \beta/t$. Because $SD(\hat{B}(t)) \approx \sigma/\sqrt{t}$, the bias tends to be negligible compared to the random fluctuations for sufficiently large t. However, in practice it can be worthwhile to reduce the bias, as we show.

Although the required run length in steady state tends to be inversely proportional to system size, the required run length to reduce the initialization bias starting empty tends to be independent of system size (see Section 11). Hence, when the system size grows, eventually a majority of the run must be devoted to reducing the initialization bias.

Therefore, for large systems it can be valuable to initialize the system closer to the steady-state mean. This is easy to do for exponential service-time distributions, but not otherwise; see Section 11.

3. EXACT NUMERICAL RESULTS FOR MARKOV MODELS

In this section we briefly review available algorithms for numerically computing the exact asymptotic variance. We have used these methods to compute the workload factors for the M/M/s/0 model in Figures 1–3 and Table I.

3.1 Poisson's Equation

First, following Whitt [1992], consider a process $Y \equiv \{Y(t) : t \geq 0\}$ that is a function of an irreducible finite-state *continuous-time Markov chain* (CTMC) $X = \{X(t) : t \geq 0\}$, that is, $Y(t) = f(X(t))$ for a real-valued function f. As in eq. (7), the asymptotic variance of Y is $\sigma_Y^2 \equiv \lim_{t \to \infty} t$ Var $Y(t)$.

The asymptotic variance σ_Y^2 can be represented as the solution of Poisson's equation involving the infinitesimal generator of X, say, Q. In particular, by Corollary 3 to Proposition 10 of Whitt [1992],

$$\hat{\sigma}_Y^2 = sxf^t, \tag{24}$$

where x is the unique solution to

$$xQ = -y, \tag{25}$$

with $y_i = -(f_i - \bar{f})\pi_i$, $xe^t = 0$, π is the steady-state vector and $\bar{f} = \pi f^t$. Here all vectors are taken to be row vectors, x^t is the transpose, and e is the vector of all 1s. As usual, the steady-state probability vector π itself is the unique solution of Poisson's equation with $y = 0$ and $\pi e^t = 1$. When solving Poisson's equation, there is one redundant equation, so that we can initially look for a solution of $\tilde{x}Q = -y$ with one component $\tilde{x}_k$ fixed. Then we can obtain the desired solution x with $xe^t = 0$ by letting

$$x = \tilde{x} - (\tilde{x}e^t)\pi, \tag{26}$$

because all solutions x are of the form $x = -yZ + (xe^t)\pi$, where Z is the fundamental matrix of the CTMC X.

As noted in Grassmann [1987] and Remark 1 of Whitt [1992], Poisson's equation (25) can be solved recursively for any birth-and-death process. Given birth rates λ_j, $0 \le j \le m - 1$, and death rates μ_j, $1 \le j \le m$, the recursion is

$$x_{j+1} = (\lambda_j x_j + S_j)/\mu_{j+1}, \tag{27}$$

where $S_j = \sum_{i=0}^{j} y_i$. We could initially set $x_0 = 1$, but for numerical purposes it is convenient to initially set $x_k = 1$ *for* $k = \min\{\lfloor \alpha \rfloor, s\}$ and then recursively solve for other values of k using eq. (27). Again, afterwards we use the adjustment (26).

In this article we apply the birth-and-death recursion to the M/M/s/0 loss model. The estimators $\hat{n}(t)$ in eq. (5) and $\hat{B}_T(t)$ involve functions of the number in system, which is a birth-and-death process. For $\hat{n}(t)$, $f(k) = k$ for all k; for $\hat{B}_T(t)$, $f(k) = 1_{(s)}(k)$. In particular, we used this method for Figures 2 and 3. An alternative approach to the asymptotic variance σ_n^2 of $\hat{n}(t)$ in eq. (5) in the M/M/s/0 model would be via expressions for the covariance function of $N(t)$ in Beneš [1961].

3.2 Interoverflow Time Moments

For the simple estimator (2), it suffices to compute the asymptotic variance of the loss process $L(t)$. For the GI/M/s/0 model, with renewal arrival process, the overflow process L is a renewal process. Hence it suffices to compute the SCV c_L^2 of an interoverflow time, as in eq. (11). The first two moments of the interoverflow time for the M/M/s/0 model are given in equations (20) and (21) on p. 89 of Riordan [1962]. We used this method for Figure 1.

4. THE UNIFYING DIFFUSION PROCESS LIMIT

We now provide a basis for the workload factor approximation (17) in the case of the G/M/s/0 model. In particular, we establish a heavy-traffic

functional central limit theorem (FCLT). To state the theorem, let $\Rightarrow$ denote weak convergence or convergence in distribution and let $D[0, \infty)$ be the function space of right-continuous real-valued functions on the interval $[0, \infty)$ with limits from the left, endowed with the usual Skorohod topology; for example, see Billingsley [1968] and Ethier and Kurtz [1986]. The convergence in $D[0, \infty)$, in addition to convergence of the one-dimensional marginal distributions, is useful for us to treat general stationary arrival processes and to get convergence of the bivariate distributions, which is needed for the covariances appearing in the asymptotic variance. To emphasize the dependence on s, we write $N_s(t)$ for the process counting the number of busy servers at time t. We assume that we start with a fixed arrival process $A(t)$ with rate 1 and scale it as we increase λ by setting $A_\lambda(t) = A(\lambda t)$.

THEOREM 4.1. *Consider the G/M/s/0 model with general stationary arrival process* $A_\lambda(t) = A(\lambda t)$ *having rate* λ *and fixed exponential service-time distribution with mean* μ^{-1}. *Let* $\lambda \to \infty$ *and* $s \to \infty$ *with* $(\alpha - s)/\sqrt{\alpha} \to \gamma$. *If* $(N_s(0) - s)/\sqrt{\alpha} \Rightarrow y$ *in* $\mathbb{R}$ *as* $s \to \infty$, *where* $y < 0$ *is deterministic and* $(A(\lambda \cdot) - \lambda \cdot)/\sqrt{\lambda c_a^2} \Rightarrow Z(\cdot)$ *in* $D[0, \infty)$ *as* $\lambda \to \infty$, *where* Z *is a standard (mean 0, variance 1) Brownian motion, then*

$$\frac{N_s(\cdot) - s}{\sqrt{\alpha}} \Rightarrow Y_r(\cdot) \quad \text{in} \quad D[0, \infty) \quad \text{as} \quad s \to \infty, \tag{28}$$

where Y_r *is a reflected Ornstein-Uhlenbeck diffusion process with infinitesimal mean* $m(x) = -\mu(x - \gamma)$, *infinitesimal variance* $\sigma^2(x) = \mu(1 + c_a^2)$, *instantaneously reflecting barrier above at 0 and initial position* $Y_r(0) = y$.

Theorem 4.1 is similar to Theorem 2 on p. 177 of Borovkov [1984]; it draws the same conclusions, but the conditions are different. The conditions in Theorem 4.1 here parallel the conditions in Theorem 1 on p. 103 of Borovkov [1984] for the G/GI/∞ model. We prove Theorem 4.1 in Section 12, by applying the G/M/∞ heavy-traffic FCLT. In addition to p. 103 of Borovkov [1984], other infinite-server FCLTs appear in Borovkov [1967], Whitt [1982], and Glynn and Whitt [1991].

The exponential service-time distribution assumption is important for the conclusion in Theorem 4.1. When the service-time distribution is nonexponential, the appropriate approximating process is non-Markov. This is clear from the G/GI/∞ heavy-traffic limit; see Glynn [1982]. One would naturally conjecture that the appropriate approximating process for the G/GI/s/0 model is a "reflected" version of the limiting Gaussian process for the G/GI/∞ model, but the statement remains to be made precise, the proof remains to be established, and the consequences remain to be exploited; see Section 6.1 for some related discussion.

There are some interesting features in our proof of Theorem 4.1. First, unlike the familiar heavy-traffic setting in Berger and Whitt [1992] and references cited there, the ROU is *not* defined as the image of an unrestricted process under a continuous reflection or regulator mapping. The state-dependent service rate makes the basic process N_s itself be altered

when it hits the barrier s. Moreover, the process $N_s(t)$ is not a Markov process. The embedded process at arrival epochs is Markov in the GI arrival case which provides a basis for Markovian approaches, as in Whitt [1982]. To treat the G arrival case, we use a coupling comparison argument together with the established limit for the associated G/M/∞ model.

We define the ROU by a limiting argument in Section 12. The time-dependent density $p(x, t)$ of the ROU can also be characterized by appending the boundary condition

$$\frac{\mu(1 + c_a^2)}{2} \frac{\partial}{\partial x} p(x, t)|_{x=0} = -\mu\gamma p(0, t). \tag{29}$$

to the usual forward partial differential equation; for example, see p. 223 of Cox and Miller [1965]. Alternatively the generator and its domain can be characterized as in Chapter 8 of Ethier and Kurtz [1986]. The alternative characterization of the ROU in Section 12 is convenient for our proof of Theorem 4.1.

The ROU limit in eq. (28) depends on three parameters—μ, γ and c_a^2—but because of the possibility of scaling we can reduce the relevant parameters to only one. First, without loss of generality, we can obviously make the service rate $\mu = 1$. Then let $z = (c_a^2 + 1)/2$ and note that Theorem 4.1 implies that $(N_s(\cdot) - s)/\sqrt{z\alpha} \Rightarrow (1/\sqrt{z})Y_r(\cdot)$, where $(1/\sqrt{z})Y_r$ is an ROU with infinitesimal mean $-(x - \gamma/\sqrt{z})$ and infinitesimal variance 2. Hence, if we let $Y_r(t; m(x), \sigma^2)$ denote the ROU as a function of its infinitesimal parameters, then

$$N_s(\cdot) \approx s + \sqrt{(\alpha(1 + c_a^2)/2)}Y_r(\cdot; -(x - \gamma\sqrt{2/(1 + c_a^2)}, 2). \tag{30}$$

Thus the asymptotic variance σ_n^2 of $N_s(t)$ is approximately

$$\sigma_n^2 \approx \alpha \frac{(1 + c_a^2)}{2} \sigma^2_{Y_r(\cdot; -(x - \gamma\sqrt{2/(1 + c_a^2)}, 2)}. \tag{31}$$

Only the single parameter $\gamma\sqrt{2/(1 + c_a^2)}$ appears inside the ROU Y_r in eq. (30) and thus inside the asymptotic variance term $\sigma^2_{Y_r}$ in eq. (31). The asymptotic variance term $\sigma^2_{Y_r}(t; -(x - \gamma), 2)$ remains to be calculated, but it clearly is a function of only the one parameter.

We can apply Theorem 4.1 to treat the indirect estimator in normal loading. Combining eq. (4) and Theorem 4.1, we obtain convergence of the bivariate distributions for any two time points. As on p. 1353 of Whitt [1989], this almost implies convergence of the covariances, but we also need appropriate uniform integrability; see p. 32 of Billingsley [1968]. Assuming that we can approximate the covariance function of the queueing process by the covariance function of the ROU, we obtain

$$\lambda\sigma_I^2(\text{GI/M/s/0}) \equiv \frac{(1 + c_a^2)}{2} \sigma^2_{Y_r(\cdot; -(x - \gamma\sqrt{2/(1 + c_a^2)}), 2)}. \tag{32}$$

It remains to establish similar results for the other estimators, but Theorem 4.1 clearly suggests that we should look at the workload factors as functions of $\gamma \equiv (\alpha - s)/\sqrt{\alpha}$. When we do, we find canonical curves for all the workload factors.

Theorem 4.1 also provides a theoretical basis for the blocking formula (15) in the G/M/s/0 model. Borovkov [1976] previously established eq. (15) as a limit for the GI/M/s/0 model by proving a local limit theorem. We can avoid having to resort to a local limit theorem by exploiting the representation (3). Under additional minor technical regularity conditions, Theorem 4.1 implies that $(n_s - s)/\sqrt{\alpha}$ converges as $s \to \infty$ to the steady-state mean of the ROU. Because the steady-state distribution of the ROU is a truncated normal, eq. (15) is the resulting approximation for the blocking probability.

5. THE SIMPLE ESTIMATOR

In this section we derive the approximation formulas for w_S in Table II. The simple estimator is $\hat{B}_S(t) = L(t)/\lambda t$, as in eq. (2). Paralleling eq. (10), let

$$c_L^2 = \lim_{t\to\infty} \frac{\text{Var } L(t)}{EL(t)}.$$

Because $EL(t)/t \to \lambda B$ as $t \to \infty$, the asymptotic variance of the simple estimator $\hat{B}_S(t)$ is

$$\sigma_S^2 = \frac{B}{\lambda} c_L^2, \tag{33}$$

so that the workload factor is

$$w_S \equiv \lambda \sigma_S^2 = B c_L^2. \tag{34}$$

First, we observe how the blocking probability B behaves. We note that

$$B \geq \max\{1 - \rho^{-1}, 0\} \tag{35}$$

(see Sobel [1980], Heyman [1980] and p. 698 of Whitt [1984]), so that B is at least $1 - \rho^{-1}$ when $\rho > 1$. In addition, $1 - \rho^{-1}$ tends to be a good approximation for B when ρ is significantly greater than 1; see Tables I–IV of Whitt [1984]. Indeed, $B \to 1 - \rho^{-1}$ as $s \to \infty$ for fixed $\rho > 1$; see p. 90 of Borovkov [1984]. On the other hand, B tends to be exponentially small when s is very large; for fixed $\rho < 1$, for example, see Jagerman [1974]. Hence w_S in eq. (34) should be small for $\rho < 1$, but not for $\rho > 1$.

Turning to c_L^2, we develop an approximation for the cases $\rho < 1$ and $\rho > 1$ by considering the case of large s. In the GI/M/s/0 special case (when the arrival process is a renewal process), the overflow process $\{L(t) : t \geq 0\}$ itself is a renewal process (because overflows necessarily occur at arrival epochs). Then c_L^2 coincides with the SCV of the interval between successive

overflows. The approximate analysis is also reasonable for the more general G/M/s/0 model, without the GI assumption. For large s, the overflow process is determined by the behavior of the process $\{N(t) : t \geq 0\}$ in the neighborhood of s. For large s, we can approximate the service completion rate $\mu(s - k)$ when $N(t) = s - k$ by the constant service rate μs. Approximating the service rate by $s\mu$ clearly should be a better approximation in heavy loading than in light loading. In light loading we significantly overestimate the service rate, which will cause overflows to be rarer events, so that we should overestimate c_L^2. Thus we anticipate that our approximation here will tend to be conservative in light loading, which is confirmed by numerical results. The important point is that c_L^2 in eq. (34) tends to be $O(1)$.

Thus we use an approximation for the asymptotic variance of the overflow process in the G/M/1/C queue, which has single server, service rate $s\mu$, and finite waiting space C. In particular, we apply results in Sections 3 and 4 of Berger and Whitt [1992]. There exact results for c_L^2 are given for the M/M/1/C model and for RBM with reflecting barriers at 0 and C; also see Williams [1992] for results on RBM.

First we rescale the G/M/1/C model to have arrival rate $\rho \equiv \alpha/s$ and service rate 1. Note the c_L^2 is invariant under time scaling. Now we use a diffusion approximation for the G/M/1/C model with service rate $\mu = 1$ and arrival rate ρ. As noted in Section 4.2 of Berger and Whitt [1992], the process representing the number of customers in the single-server queue can be approximated by RBM with drift $-(1 - \rho)$ and diffusion coefficient $(\rho c_a^2 + 1)$. We then apply Theorem 4.1 of Berger and Whitt [1992] to obtain

$$c_L^2(\text{G/M/1/C}, \mu{=}1) \approx \frac{1 + \rho c_a^2}{|1 - \rho|} \tag{36}$$

assuming that $\rho \neq 1$ and that the barrier C is large. In particular, as a convenient approximation for large C, we use the limit of (29) in Berger and Whitt [1992] as $C \to \infty$. Combining eqs. (34) and (36) we obtain

$$w_S \approx \frac{B(1 + \rho c_a^2)}{|1 - \rho|}. \tag{37}$$

For $\rho > 1$, we approximate B by $1 - \rho^{-1}$ to obtain

$$w_S \approx \frac{1 + \rho c_a^2}{\rho} \quad \text{for} \quad \rho > 1. \tag{38}$$

Formula (37) for light loading and formula (38) for heavy loading are given in Table II.

To quickly see how formulas (37) and (38) perform for the M/M/s/0 model, we consider the case in Table I with $s = 400$. For $\gamma = -4, -3, -2, -1$, the approximate workload factors are, .00035, .0065, .074, and .62, respec-

tively. For $\gamma = 1, 2, 3, 4$, the approximate workload factors are 1.95, 1.90, 1.86, and 1.82. From Table I, we see that these are upper bounds that would serve as suitable rough approximations in practice.

6. THE INDIRECT ESTIMATOR

The indirect estimator is $\hat{B}_I(t) = 1 - \hat{n}(t)\alpha^{-1}$, as in eq. (4). The workload factor is thus

$$w_I \equiv \lambda\sigma_I^2 = \frac{\lambda}{\alpha^2}\sigma_n^2, \tag{39}$$

where σ_n^2 is the asymptotic variance of $\hat{n}(t)$ in (5).

6.1 Light Loading

First consider the case of light loading. For $\rho < 1$ and s large, we can approximate the asymptotic variance σ_n^2 by the asymptotic variance in the associated infinite-server model. For the M/M/∞ system, $\sigma_n^2 = 2\alpha$; see (23) of Whitt [1992]. To treat the G/M/∞ model, as in Borovkov [1967] and Whitt [1982], we approximate a scaled version of the process $\{N(t) : t \geq 0\}$ by an Ornstein-Uhlenbeck (OU) diffusion process, say $\{Y(t) : t \geq 0\}$. (Further discussion appears in Section 12, because this is used in the proof of Theorem 4.1.) In particular, for $\mu = 1$,

$$\frac{N(t) - \alpha}{\sqrt{\alpha}} \approx Y(t), \quad t \geq 0, \tag{40}$$

where $Y(t)$ has drift $-x$, diffusion coefficient $(1 + c_a^2)$ and covariance function

$$R(t) \equiv \text{Cov}(Y(0), Y(t)) = ((1 + c_a^2)/2)e^{-t}.$$

Hence, as in eq. (4) and Example 5 of Whitt [1992], the asymptotic variance of $t^{-1}\int_0^t Y(u)du$ is

$$\sigma_Y^2 \equiv \lim_{t\to\infty} \text{Var}\, t^{-1}\int_0^t Y(u)du = 2\int_0^\infty R(t)dt = (1 + c_a^2). \tag{41}$$

By eq. (40),

$$\sigma_n^2 \approx \alpha\sigma_y^2. \tag{42}$$

Finally, combining eqs. (39), (41), and (42), for the case $\mu = 1$ we obtain

$$w_I \equiv \lambda\sigma_I^2 \approx 1 + c_a^2, \tag{43}$$

as given in Table II. From Table I, we see that in the M/M/s/0 case with $s = 400$ the approximation $w_I \approx 2.00$ from eq. (43) performs excellently for γ suitably large.

For the more general G/GI/∞ system, again with $\mu = 1$, we can again apply the heavy-traffic approximation, which is a Gaussian process with covariance function

$$R(t) = \int_0^\infty H(u)H^c(t+u)du + c_a^2 \int_0^\infty H^c(u)H^c(t+u)du, \; t \geq 0, \quad (44)$$

where $H^c(t) = 1 - H(t)$ with $H(t)$ the service-time cdf, see p. 176 of Whitt [1982] and Borovkov [1984]. Reasoning as in eq. (41), we then obtain an approximation for the asymptotic variance

$$\sigma_Y^2 = 2\int_0^\infty H(u)H_e^c(u)du + 2c_a^2 \int_0^\infty H^c(u)H_e^c(u)du, \quad (45)$$

where $H_e^c(t) \equiv 1 - H_e(t)$ and

$$H_e(t) = \int_0^t H^c(u)du. \quad (46)$$

For example, when $c_a^2 = 1$, eq. (45) simplifies to

$$\sigma_Y^2 = 2\int_0^\infty H_e^c(u) = (c_s^2 + 1). \quad (47)$$

so that combining eqs. (39), (42), and (47) we obtain

$$\omega_I \approx 1 + c_s^2. \quad (48)$$

Formula (48) provides good theoretical support for approximation (17) for the M/GI/s/0 model. More generally, formulas (45)–(48) provide partial support for formulas (17) and (18) when the service-time distribution is nonexponential. Refined approximations could be based on combining eqs. (39), (42), and (45).

6.2 Heavy Loading

In heavy loading it is natural to focus on the number of idle servers instead of the number of busy servers, for example, see p. 138 of Borovkov [1984]. Let m be the mean number of idle servers. Then clearly $m = s - n$, so that, by eq. (3),

$$B = 1 - \frac{s}{\alpha} + \frac{m}{\alpha}. \quad (49)$$

Moreover, $\sigma_n^2 = \sigma_m^2$, where σ_n^2 is the asymptotic variance of $\hat{n}(t)$ in eq. (5) and σ_m^2 is the asymptotic variance of

$$\hat{m}(t) = s - \hat{n}(t), \quad t \geq 0. \tag{50}$$

Henceforth we focus on approximating σ_m^2. Inasmuch as we are looking at the idle servers, we approximate the process $M(t) \equiv s - N(t)$ counting the idle servers in a G/M/s/0 model by an M/G/1 queue with constant arrival rate $s\mu$ and service rate λ and service-time SCV c_a^2. Because $\rho > 1$ in the original system, the traffic intensity for $M(t)$ is ρ^{-1}. As before, the constant rate $s\mu$ is an approximation for the state-dependent rate $(s - k)\mu$ when $M(t) = k$, which is a good approximation when s is large.

We first scale time by $1/\lambda$ to make the service rate 1 and the arrival rate ρ^{-1} in the M/G/1 spaces model. By (28) of Whitt [1989], we see how this scaling affects the asymptotic variance; it cancels the λ term in the workload factor. Hence we have

$$\lambda\sigma_I^2 \approx \sigma_m^2(\text{M/G/1}, \mu = 1)/\alpha^2. \tag{51}$$

When the G is GI (a renewal arrival process in the original model) the exact form of σ_m^2 (M/G/1, $\mu = 1$) depends on the first four moments of the service-time distribution, but as in Whitt [1989] we can give a rough approximation. From Law [1975], the exact formula for the asymptotic variance in the M/GI/1 model is

$$\sigma_Q^2 = \frac{\rho^4}{2(1-\rho)^4}(c_s^2 + 1)^3 + \left(\frac{5\rho^3}{2}\left(\frac{m_3}{3m_2}\right) + \rho^2\right)\frac{(c_s^2 + 1)^2}{(1-\rho)^3} + \left(\rho(1+\rho)\left(\frac{m_3}{3m_2}\right) + 3\rho^2\left(\frac{m_4}{12m_2}\right)\right)\frac{(c_s^2 + 1)}{(1-\rho)^2}, \tag{52}$$

where m_k is the kth moment of the service-time distribution, $m_1 = 1$, $c_s^2 = m_2 - m_1^2$, and the arrival rate is $\rho < 1$. For the approximation here (with $\rho^{-1} < 1$), we combine (18) and (22) of Whitt [1989] to obtain

$$\sigma_m^2(\text{M/G/1}, \mu{=}1) \approx \frac{\rho^{-1}(1 + \rho^{-1})(c_a^2 + 1)^3}{4(1 - \rho^{-1})^4}, \tag{53}$$

so that

$$w_I \equiv \lambda\sigma_I^2 \approx \frac{(1+\rho)(c_a^2 + 1)^3}{4s^2(\rho - 1)^4} \quad \text{for} \quad \rho > 1, \tag{54}$$

as in Table II. It is significant that eq. (53) is exact for the M/M/1 model, but the role of c_a^2 in eq. (53), and thus eq. (54), is problematic, being based on heuristic heavy-traffic analysis in (17) of Whitt [1989]. However, eq. (53) is asymptotically correct as $\rho \to 1$, so that there are regions where eqs. (53)

and (54) are appropriate. Moreover, as noted in eqs. (21) and (22), the form of eq. (54) is consistent with eq. (17).

The discussion so far has been based on assuming that G is GI. However, the approximation formulas (53) and (54) also apply to the general G case, because the heavy-traffic diffusion approximations in Whitt [1989] apply to general G service processes. With service times as well as arrival times, the variability of general G is characterized by the asymptotic variance as defined in eq. (10).

To see how eq. (54) performs for the M/M/s/0 model with s = 400, we can compare with Table I. For γ = 1, 2, 3, 4, and 5, the approximation (54) yields the values 0.68, 0.109, 0.028, 0.0094, and 0.0038. From Table I, we see that these are upper bounds that get more accurate as γ increases.

7. THE NATURAL ESTIMATOR

In this section we briefly relate the natural estimator $\hat{B}_N(t)$ in eq. (1) to the simple estimator $\hat{B}_S(t)$ in eq. (2) and theoretically justify using σ_S^2 as an approximation for σ_N^2. The simulation results in Section 10 also provide strong empirical support.

We first note that both the natural estimator and the simple estimator are *consistent*; that is, by the ergodic theorem (assuming ergodicity as well as stationarity), $L(t)/t \to \lambda B$ and $A(t)/t \to \lambda$ w.p.1 as $t \to \infty$, so that, under minor regularity conditions, $\hat{B}_N(t) \to B$ and $\hat{B}_S(t) \to B$ w.p.1 as $t \to \infty$. (We need to have uniform integrability to get $EL(t)/t \to \lambda B$ from $L(t)/t \to \lambda B$.)

Moreover, given that the system starts in equilibrium, so that the process $\{N(t) : t \geq 0\}$ is stationary, the simple estimator is *unbiased*, that is,

$$E\hat{B}_S(t) = \frac{EL(t)}{EA(t)} = B. \tag{55}$$

In contrast, in general the natural estimator is *not* unbiased, because the expectation of a ratio is not necessarily the ratio of the expectations; that is, we typically have

$$EB_N(t) = E\left(\frac{L(t)}{A(t)}\right) \neq \frac{EL(t)}{EA(t)} = B. \tag{56}$$

Nevertheless, the natural estimator may be preferred to the simple estimator. First, in the context of system measurements the arrival rate λ is typically unknown. Moreover, even if the arrival rate is known, it can be somewhat better to use the natural estimator than the simple estimator because it may have a smaller asymptotic variance, as shown in the following.

We can compare the asymptotic variances by using the theory of nonlinear control variates or indirect estimation in Glynn and Whitt [1989]. Just as in Theorem 2 of Glynn and Whitt [1989], we can characterize σ_N^2 if we assume a joint CLT for the processes $L(t)$ and $A(t)$. This joint CLT

condition holds for the GI/M/s/0 model, because the overflow epochs serve as embedded regeneration epochs. The CLT condition will also hold for more general models for which a subsequence of the overflow epochs serve as embedded regeneration times. For example, this regeneration structure occurs in Markov modulated arrival processes; then overflows in a particular environment state can serve as regeneration times.

Let $N(0, C)$ denote a normally distributed random vector with zero means 0 and covariance matrix C.

THEOREM 7.1. *If*

$$t^{-1/2}[L(t) - B\lambda t, A(t) - \lambda t] \Rightarrow (X, Y) \text{ in } \mathbb{R}^2 \text{ as } t \to \infty,$$

then

$$\sqrt{t}(\hat{B}_N(t) - B) \Rightarrow \lambda^{-1}(X - BY) \text{ in } \mathbb{R} \text{ as } t \to \infty.$$

Moreover, if (X, Y) *is distributed as* $N(0, C)$, *then* $\lambda^{-1}(X - BY)$ *is distributed as* $N(0, \sigma_N^2)$, *where* $C_{11} = \lambda B c_L^2 = \lambda \sigma_S^2$, $C_{22} = \lambda c_a^2$,

$$C_{12} = \lim_{t \to \infty} \frac{\text{cov}(A(t), L(t))}{t}, \tag{57}$$

and

$$\sigma_N^2 = \lambda^{-2}(C_{11} - 2BC_{12} + B^2 C_{22}) = \sigma_S^2 - 2BC_{12}\lambda^{-2} + B^2 c_a^2 \lambda^{-1}. \tag{58}$$

PROOF. The proof here is essentially the same as for Theorem 2 of Glynn and Whitt [1989]. Note that

$$\sqrt{t}(\hat{B}_N(t) - B)$$

$$= \sqrt{t}\left(\frac{L(t)}{A(t)} - B\right) = \frac{t}{A(t)} \frac{1}{\sqrt{t}} ([L(t) - \lambda B t] - B[A(t) - \lambda t])$$

$$\Rightarrow \lambda^{-1}(X - BY) \quad \text{as} \quad t \to \infty$$

because $A(t)/t \to \lambda$ w.p.1 and we can apply the continuous mapping theorem. Formula (58) is an elementary consequence of the bivariate normal distribution. □

A missing ingredient in Theorem 7.1 is the *asymptotic covariance term* C_{12} in eq. (57). We anticipate that C_{12} will usually be positive (see Srikant and Whitt [1995] for conditions under which this is true), but we have no expression for it. Nevertheless, from formula (58), we see that when B is small and/or λ is large, we will have $\sigma_N^2 \approx \sigma_S^2$. Hence σ_S^2 should be a good approximation for σ_N^2. If $C_{12} > \lambda B c_a^2/2$, then $\sigma_N^2 \leq \sigma_S^2$.

8. THE TIME-CONGESTION ESTIMATOR

In this section we briefly discuss the time-congestion estimator $\hat{B}_T(t)$ for $P(N(t) = s)$ in eq. (6). We contend that $\hat{B}_T(t)$ often behaves similarly to the simple and natural estimators. This can be substantiated in the case of Poisson arrivals, where $B = P(N(t) = s)$.

First, by Proposition 6 of Glynn et al. [1993], in the case of Poisson arrivals the difference $\hat{B}_T(t) - \hat{B}_N(t)$ has asymptotic variance

$$\sigma_d^2 \equiv \text{var}(\hat{B}_T(t) - \hat{B}_N(t)) = \lambda^{-1}B(1 - B), \tag{59}$$

which will be small where either B is small or λ is large. This is consistent with the simulation results in Section 10; see Tables III and VI.

In the case of Poisson arrivals, we can also bound the asymptotic variance σ_T^2 above by σ_S^2 because we can represent $\hat{B}_S(t)$ as a conditional expectation given $\hat{B}_T(t)$, that is,

$$E[\hat{B}_S(t)|\hat{B}_T(t)] = \hat{B}_T(t), \tag{60}$$

so that we necessarily have

$$\sigma_T^2 \leq \sigma_S^2; \tag{61}$$

see Example 11.16 of Ross [1993].

9. ENSURING SMALL BLOCKING PROBABILITIES

In some applications we may wish to ensure that the blocking probability is very small, without being concerned with estimating its actual value. For example, we may want to verify that $B \leq \epsilon$ for some very small ϵ. Because B is very small, presumably we are in the region of light loading. Thus it is appropriate to use the natural estimator $\hat{B}_N(t)$.

The question is: how large should t be in order to be convinced by a small value of $\hat{B}_N(t)$ that $B \leq \epsilon$? For example, we might obtain $\hat{B}_N(t) = 0$. For this purpose, it is natural to use a one-sided confidence interval and choose the run length t to make the width of the confidence interval with level of precision β be $\epsilon/2$; that is, let $P(N(0, 1) < z_\beta) = 1 - \beta$ and

$$\frac{\epsilon}{2} = \frac{\sigma_N z_\beta}{\sqrt{t}}, \tag{62}$$

so that the required run length is

$$t(\epsilon, \beta) = \frac{4\sigma_N^2 z_\beta^2}{\epsilon^2}; \tag{63}$$

for example, see p. 1345 of Whitt [1989]. Because we do not know σ_N^2, we can approximate it by $\sigma_S^2 = B(c_a^2 + 1)/(1 - \rho)$ from Table I, as explained in

Table III.

		predicted	run 1	run 2	run 3	run 4
	S	.0143	.0136	.0134	.0139	.0142
blocking	N	.0143	.0136	.0134	.0139	.0142
probability	I	.0143	.0143	.0157	.0142	.0135
estimate	T	.0143	.0138	.0135	.0140	.0142
	S	.00041	.00041	.00032	.00028	.00038
standard	N	.00041	.00041	.00031	.00029	.00037
deviation	I	.00066	.00071	.00067	.00047	.00064
estimate	T	.00040	.00043	.00033	.00027	.00036

A comparison of predictions with simulation results for the M/M/s/0 model with $s = 400$, $\lambda = 380$, $\mu = 1$, and $t = 5400$ in Example 10.1.

Section 5. Substituting this approximation for σ_N^2 into eq. (63), we obtain

$$t \leq \frac{4B(c_a^2 + 1)z_\beta^2}{(1 - \rho)\epsilon^2}. \tag{64}$$

A difficulty with eq. (64) is that the bound itself depends on the blocking probability B to be estimated. An obvious crude upper bound on B is 1. However, if we are prepared to assume that B is less than some smaller number, then we can reduce the bound.

The estimation procedure is then as follows: we use the estimator $\hat{B}_N(t)$ for the time t in the bound in eq. (64). If $\hat{B}_N(t) \leq \epsilon/2$, then we conclude with confidence $1 - \beta$ that indeed $B \leq \epsilon$.

10. SIMULATION EVIDENCE

A key point underlying all our work is the fact that the actual variance of each estimate $\hat{B}(t)$ is reasonably well described by σ^2/t, where σ^2 is the asymptotic variance when t is suitably large. This large sample behavior is well established in statistical experience, but we also present confirmation here. In each example below, we estimate the standard deviation of the estimator by dividing each run into 20 equally spaced portions or batches and treat them as independent.

Example 10.1 *The Erlang Model.* To illustrate how the formulas based on the asymptotic variance perform for M/M/s/0 models, we consider a simulation experiment with 4 independent runs each of length 5,400 for the case $s = 400$, $\lambda = 380$, and $\mu = 1$. We delete an initial portion of length 5 to allow the system to reach steady state (for discussion, see Section 11). For this model, the scaled arrival rate is $\gamma = -1.0$. We thus estimate the variance of each estimator $\hat{B}(t)$ by $\sigma^2/t = \psi(-1)/2.052 \times 10^6$. From Table I, $\psi_I(-1) = 0.89$, $\psi_S(-1) = 0.35$, and $\psi_T(-1) = 0.33$.

Table III displays the simulation results and the predictions based on eqs. (15)–(18) and Table I. Table III shows that the predictions are good and also shows that the estimators $\hat{B}_S(t)$ and $\hat{B}_N(t)$ are almost identical; their difference is negligible compared to the statistical precision. Also $\hat{B}_T(t)$ is strongly positively correlated with $\hat{B}_S(t)$, whereas $\hat{B}_I(t)$ tends to be

negatively correlated with $\hat{B}_S(t)$; $\hat{B}_I(t)$ tends to have high (low) values when $\hat{B}_S(t)$ has low (high) values. Thus the combined estimator mentioned in Section 1 should (and does) have lower variance than either $\hat{B}_S(t)$ or $\hat{B}_I(t)$, but we do not give details here. These observations hold for more general models, as can be seen from the following examples. The results for the M/M/s/0 model in Table III help to judge the quality of the approximations for the more general models.

Example 10.2 *More Variable Arrival Processes.* To see how the approximations perform for G/M/s/0 models with arrival processes more variable than Poisson, we conduct a simulation experiment for the GI/M/s/0 model, where the interarrival time has a hyperexponential (H_2) distribution with balanced means and $c_a^2 = 9.0$. This H_2 distribution has density

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \qquad x \geq 0, \tag{65}$$

where

$$p = [1 + \sqrt{(c_a^2 - 1)/(c_a^2 + 1)}]/2, \tag{66}$$

$$\lambda_1 = 2p\lambda \quad \text{and} \quad \lambda_2 = 2(1-p)\lambda \tag{67}$$

with λ^{-1} being the mean. Because the service-time distribution is exponential, the approximate peakedness by eq. (14) is $z = (c_a^2 + 1)/2 = 5$. (The exact peakedness by eq. (13) is 4.95, so eq. (14) is an excellent approximation.)

We consider $s = 400$, $\mu = 1$, and three values of λ : $\lambda = 360$, $\lambda = 400$, and $\lambda = 440$. The experiment consists of 4 independent runs of length 2,700 for each λ, deleting a portion of length 5 to allow the system to approach steady state in each case. (The run length 2,700 makes the expected total number of arrivals about 10^6 in each case.) The simulation results are displayed in Table IV, where estimated blocking probabilities and sample standard deviations for the estimators are: simple (S), natural (N), indirect (I), and time-congestion (T). The predictions in Table IV based on eqs. (15)–(18) and Table I seem very good. As in Example 1.1, $\hat{B}_N(t)$ essentially coincides with $\hat{B}_S(t)$, whereas $\hat{B}_S(t)$ and $\hat{B}_I(t)$ are negatively correlated.

Example 10.3 *Less Variable Arrival Processes.* To show how the approximations perform for G/M/s/0 models with arrival processes less variable than Poisson, we consider a simulation experiment for the GI/M/s/0 model, where an interarrival time has the Erlang (E_4) distribution with $c_a^2 = 0.25$. An E_4 distribution is the convolution of four exponential distributions. By eq. (14), $z \approx (c_a^2 + 1)/2 = 0.625$. As in Example 1.2, we consider $s = 400$, $\mu = 1$, and three values of λ. Here we consider $\lambda = 380$, $\lambda = 400$, and $\lambda = 420$. The experiment consists of 2 independent runs of length 2,700 for each λ, deleting an initial portion of length 5 to allow the system to approach steady state. The simulation results and the predictions based on Table I and eqs. (15)–(18) are displayed in Table V. Again the predictions are good.

Table IV.

$\lambda = 360$		predicted	run 1	run 2	run 3	run 4
	S	.036	.0309	.0306	.0324	.0307
blocking	N	.036	.0308	.0306	.0322	.0308
probability	I	.036	.0303	.0334	.0288	.0340
estimate	T	.0163	.0185	.0181	.0194	.0184
	S	.0014	.0011	.0012	.0014	.0013
standard	N	.0014	.0011	.0011	.0013	.0012
deviation	I	.0020	.0020	.0017	.0028	.0027
estimate	T	.00059	.00071	.00063	.00079	.00075
$\lambda = 400$		predicted	run 1	run 2	run 3	run 4
	S	.089	.0811	.0854	.0838	.0815
blocking	N	.089	.0813	.0849	.0835	.0815
probability	I	.089	.0833	.0804	.0819	.0824
estimate	T	.040	.0489	.0515	.0502	.0490
	S	.0020	.0014	.0015	.0022	.0018
standard	N	.0020	.0013	.0013	.0020	.0016
deviation	I	.0012	.0015	.0011	.0013	.0010
estimate	T	.00087	.00084	.00089	.0013	.0011
$\lambda = 440$		predicted	run 1	run 2	run 3	run 4
	S	.150	.1397	.1430	.1449	.1415
blocking	N	.150	.1400	.1428	.1443	.1415
probability	I	.150	.1431	.1417	.1411	.1426
estimate	T	.067	.0845	.0861	.0874	.0854
	S	.0023	.0027	.0020	.0019	.0022
standard	N	.0023	.0023	.0017	.0016	.0018
deviation	I	.00075	.00082	.00075	.00060	.00099
estimate	T	.00101	.00158	.00124	.00109	.00126

A comparison of predictions with simulation results for the GI/M/s/0 model with $s = 400$, hyperexponential (H_2^b) interarrival times with balanced means having $c_a^2 = 9.0$, and service rate $\mu = 1$ in Example 10.2.

Example 10.4 *Sensitivity in the M/G/s/0 Model.* For the M/G/s/0 model, it is well known that the blocking probability depends on the service-time distribution only through its mean. This insensitivity property is reflected by formulas (13)–(16), because then $c_a^2 = z = 1$. However, the asymptotic variance and workload factors do *not* have this insensitivity property, as is clear from the influence of c_s^2 in formulas (17) and (18).

To illustrate how the approximations (17) and (18) apply to M/G/s/0 systems, we consider an M/G/s/0 system with $s = 400$, $\mu = 1$, and an H_2^b service-time distribution with $c_s^2 = 9.0$, as in Example 1.2. Simulation results for 2 independent runs of length 2,700 are displayed in Table VI. The analysis in Section 11 shows that the bias changes when we change the service-time distribution. For this H_2^b distribution, we should delete more in order to reduce the bias, about 50 instead of 5. Note that the blocking probabilities are well predicted by formulas (15) and (16) with $z = 1$. The standard deviation estimates are also reasonably well predicted by eqs. (17) and (18) as well. Notice that the prefactor $(c_a^2 + c_s^2)/2 = 5$ in eq. (17) plays an important role here, as predicted by eq. (48).

Table V.

$\lambda = 380$		predicted	run 1	run 2
	S	.0077	.0080	.0079
blocking	N	.0077	.0080	.0079
probability	I	.0077	.0055	.0071
estimate	T	.0098	.0116	.0117
	S	.00038	.00041	.00035
sample	N	.00038	.00041	.00035
standard	I	.00082	.00086	.00074
deviation	T	.00047	.00058	.00052
$\lambda = 400$		predicted	run 1	run 2
	S	.032	.0387	.0385
blocking	N	.032	.0387	.0385
probability	I	.032	.0392	.0389
estimate	T	.040	.0384	.0385
	S	.00071	.00075	.00087
sample	N	.00071	.00073	.00084
standard	I	.00044	.00041	.00065
deviation	T	.00087	.00073	.00087
$\lambda = 420$		predicted	run 1	run 2
	S	.066	.0644	.0656
blocking	N	.066	.0644	.0656
probability	I	.066	.0652	.0653
estimate	T	.084	.0913	.0924
	S	.00085	.00060	.00085
sample	N	.00085	.00060	.00083
standard	I	.00021	.00017	.00023
deviation	T	.00105	.00078	.00115

A comparison of predictions with simulation results for the GI/M/s/0 model with $s = 400$, Erlang (E_4) interarrival times with $c_a^2 = 0.25$, and service rate $\mu = 1$ in Example 10.3.

Example 10.5 *General Distributions.* We now consider GI/GI/s/0 models where neither the interarrival-time distribution nor the service-time distribution is exponential. We consider all combinations of H_2^b distributions with $c^2 = 9.0$ and E_4 distributions with $c^2 = 0.25$. As before, we let $s = 400$ and $\mu = 1$. Each run is 2700 in length. Table VII displays results for $\lambda = 380$, and Table VIII displays results for $\lambda = 440$. The approximate peakedness values used are displayed there. These results provide strong empirical support for approximations (15)–(18). We rely heavily on such empirical support because we have provided little theoretical support for cases in which neither the arrival process nor the service times are M.

Example 10.6 *Nonrenewal Arrival Processes.* Our final simulation example is a G/GI/s/0 model with a nonrenewal arrival process. To have a relatively simple example, we let the arrival process be a two-state MMPP. There are alternating high-rate and low-rate environment states with exponential holding times having means m_h and m_ℓ. In state h (ℓ), arrivals are submitted according to a Poisson process with rate λ_h (λ_ℓ), where $\lambda_h >$

Table VI.

$\lambda = 380$		predicted	run 1	run 2
	S	.0143	.0151	.0135
blocking	N	.0143	.0151	.0135
probability	I	.0143	.0137	.0165
estimate	T	.0143	.0150	.0136
	S	.0013	.00146	.00128
standard	N	.0013	.00146	.00128
deviation	I	.0021	.00221	.00221
estimate	T	.0013	.00145	.00130
$\lambda = 400$		predicted	run 1	run 2
	S	.0399	.0389	.036
blocking	N	.0399	.0390	.036
probability	I	.0399	.0393	.040
estimate	T	.0385	.0389	.037
	S	.0020	.00144	.00177
standard	N	.0020	.00143	.00176
deviation	I	.0012	.00151	.00143
estimate	T	.0020	.00148	.00176
$\lambda = 420$		predicted	run 1	run 2
	S	.073	.0729	.0727
blocking	N	.073	.0729	.0726
probability	I	.073	.0722	.0724
estimate	T	.073	.0729	.0725
	S	.0022	.00193	.00143
standard	N	.0022	.00191	.00144
deviation	I	.00072	.00084	.00074
estimate	T	.0023	.00192	.00149

A comparison of predictions with simulation results for the M/G/s/0 model with $s = 400$, hyperexponential service times with balanced means having $c_s^2 = 9$, and $\mu = 1$ in Example 10.4.

λ_ℓ. It is known that the "on-off" special case in which $\lambda_\ell = 0$ is actually a renewal process, but more generally the MMPP is not renewal. The overall arrival rate of the MMPP is

$$\lambda = \frac{m_h\lambda_h + m_\ell\lambda_\ell}{m_h + m_\ell}. \tag{68}$$

and the normalized asymptotic variance [defined in eq. (10)] is

$$c_a^2 = 1 + \frac{2m_h^{-1}m_\ell^{-1}(\lambda_h - \lambda_\ell)^2}{\lambda(m_h^{-1} + m_\ell^{-1})^3} \tag{69}$$

(see p. 288 of Neuts [1989]), and the (exact) peakedness with exponential serviced times is

$$z = 1 + \left(\frac{\lambda_h^2 m_l^{-1} + \lambda_l^2 m_h^{-1}}{\lambda_h m_h^{-1} + \lambda_l m_l^{-1}} - \lambda\right)(\mu + m_h^{-1} + m_l^{-1})^{-1} \tag{70}$$

Table VII.

		$c_a^2 = .25,\ c_s^2 = .25,\ z = .43$			$c_a^2 = .25,\ c_s^2 = 9.0,\ z = .74$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.0042	.0051	.0055	.0098	.0073	.0090
blocking	N	.0042	.0052	.0055	.0098	.0073	.0090
probability	I	.0042	.0064	.0053	.0098	.0150	.0099
estimate	T	.0063	.0076	.0081	.0138	.0106	.0129
	S	.00019	.00020	.00028	.00112	.00096	.00071
standard	N	.00019	.00020	.00028	.00112	.00096	.00071
deviation	I	.00057	.00061	.00040	.0022	.0028	.0014
estimate	T	.00038	.00030	.00037	.0011	.0014	.00097
		$c_a^2 = 9.0,\ c_s^2 = .25,\ z = 7.0$			$c_a^2 = 9.0,\ c_s^2 = 9.0, z = 3.8$		
	S	.077	.062	.064	.050	.048	.048
blocking	N	.077	.062	.064	.050	.048	.047
probability	I	.077	.065	.063	.050	.048	.047
estimate	T	.029	.038	.039	.025	.029	.028
	S	.0017	.0015	.0014	.0022	.0019	.0020
standard	N	.0017	.0014	.0013	.0022	.0019	.0019
deviation	I	.0015	.0019	.0016	.0022	.0020	.0018
estimate	T	.00077	.00090	.00087	.00098	.00115	.00118

A comparison of predictions with simulation estimates for the GI/GI/s/0 model with $s = 400$, $\lambda = 380$, and $\mu = 1$ in Example 10.5.

Table VIII.

		$c_a^2 = .25,\ c_s^2 = .25,\ z = .43$			$c_a^2 = .25,\ c_s^2 = 9.0,\ z = .74$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.100	.101	.103	.105	.100	.102
blocking	N	.100	.101	.102	.105	.100	.102
probability	I	.100	.102	.102	.105	.103	.103
estimate	T	.152	.141	.144	.122	.140	.142
	S	.00056	.00075	.00058	.0022	.0020	.0015
standard	N	.00056	.00070	.00055	.0022	.0020	.0015
deviation	I	.00014	.000096	.000081	.00039	.00047	.00043
estimate	T	.00108	.00098	.00080	.0030	.0027	.0020
		$c_a^2 = 9.0,\ c_s^2 = .25,\ z = 7.0$			$c_a^2 = 9.0,\ c_s^2 = 9.0, z = 3.8$		
	S	.164	.147	.151	.140	.135	.138
blocking	N	.164	.147	.150	.140	.136	.138
probability	I	.164	.147	.146	.140	.140	.139
estimate	T	.062	.089	.091	.072	.081	.083
	S	.0024	.0022	.0025	.0032	.0020	.0035
standard	N	.0024	.0022	.0021	.0032	.0018	.0032
deviation	I	.00078	.00087	.00089	.00094	.00071	.00098
estimate	T	.0010	.0013	.0014	.0014	.0012	.0021

A comparison of predictions with simulation estimates for the GI/GI/s/0 model with $s = 400$, $\lambda = 440$, and $\mu = 1$ in Example 10.5.

(see p. 338 of Neuts [1989]). The peakedness z in eq. (13) tends to be well approximated by the heavy-traffic value $(c_a^2 + 1)/2$ in eq. (14) when the rates λ, m_h^{-1}, and m_l^{-1} are large compared to μ, but not otherwise.

Table IX.

		$m_h = .03125$, $\lambda_h = 720$, $z = 4.93$			$m_h = 0.1$, $\lambda_h = 578.9$, $z = 4.81$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.089	.083	.080	.087	.079	.077
blocking	N	.089	.083	.080	.087	.079	.077
probability	I	.089	.082	.084	.087	.076	.079
estimate	T	.040	.050	.048	.040	.057	.055
	S	.0020	.0023	.0018	.0020	.0014	.0018
standard	N	.0020	.0025	.0016	.0020	.0016	.0016
deviation	I	.0012	.0014	.0013	.0012	.0014	.0012
estimate	T	.00087	.0014	.0010	.00087	.00119	.0013
		$m_h = 1.0$, $\lambda_h = 456.6$, $z = 3.67$			$m_h = 10.0$, $\lambda_h = 417.9$, $z = 1.67$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.076	.063	.059	.052	.043	.043
blocking	N	.076	.063	.059	.052	.043	.043
probability	I	.076	.060	.062	.052	.043	.044
estimate	T	.040	.056	.052	.040	.041	.042
	S	.0020	.0019	.0018	.0020	.0018	.0018
blocking	N	.0020	.0017	.0016	.0020	.0017	.0017
probability	I	.0012	.0014	.0010	.0012	.0012	.0012
estimate	T	.00087	.0016	.0015	.00087	.0017	.0017

A comparison of predictions with simulation estimates for the MMPP/M/s/0 model with $s = 400$, $\mu = 1$, $t = 2700$, two environment states, $m_h = m_l$, $c_a^2 = 9.0$, and $\lambda = 400$ in Example 10.6.

We focus on the special case in which $m_h = m_l = m$. Then

$$\lambda_h = \lambda + \gamma \quad \text{and} \quad \lambda_l = \lambda - \gamma, \tag{71}$$

where

$$\gamma = \sqrt{\lambda(c_a^2 - 1)/m}. \tag{72}$$

As a first concrete example, let $s = 400$, $\mu = 1$, $\lambda = 400$, and $c_a^2 = 9.0$ as in the second case of Example 2.2 (Table IV). We consider four values of the common mean environment state holding time m: $m = .03125$, $m = 0.1$, $m = 1.0$, and $m = 10.0$. The exact peakedness values in these four cases are $z = 4.93$, $z = 4.81$, $z = 3.67$, and $z = 1.67$. The heavy-traffic approximation $z \approx 5$ from eq. (14) is fine in the first two cases, but not in the last two. In the heavy-traffic limit, $\lambda \to \infty$ and $m \to 0$ as $s \to \infty$, whereas μ remains fixed. (We have $m \to 0$, because the rate-1 arrival process is scaled by $\lambda \to \infty$, which means the environment state change rates go to infinity.) Hence eventually $m \ll \mu^{-1}$ as the limit is approached. The fact that $m \geq \mu^{-1}$ in the last two cases indicates that the case is not near to the limit even though λ is quite large.

Table IX compares the predictions in eqs. (15)–(18) with simulation estimates for these four cases. Because $\gamma = 0$, the predicted standard deviations are independent of z and thus of m, which is consistent with the

Table X.

		$m_h = .03125,\ \lambda_h = 648,\ z = 4.54$			$m_h = 0.1,\ \lambda_h = 521.0,\ z = 4.42$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.033	.031	.028	.032	.027	.028
blocking	N	.033	.030	.028	.032	.027	.028
probability	I	.033	.030	.031	.032	.022	.025
estimate	T	.015	.018	.017	.015	.020	.020
	S	.0013	.0013	.0011	.0013	.0013	.0012
standard	N	.0013	.0012	.0011	.0013	.0012	.0012
deviation	I	.0020	.0018	.0021	.0021	.0023	.0020
estimate	T	.00058	.00078	.00069	.00058	.00098	.00091
		$m_h = 1.0,\ \lambda_h = 410.9,\ z = 3.40$			$m_h = 10.0,\ \lambda_h = 376.1,\ z = 1.60$		
		predicted	run 1	run 2	predicted	run 1	run 2
	S	.023	.0148	.0163	.0069	.0047	.0060
blocking	N	.023	.0148	.0162	.0069	.0047	.0060
probability	I	.023	.0150	.0105	.0069	.0066	.0047
estimate	T	.0125	.0130	.0144	.0055	.0045	.0058
	S	.0012	.00084	.00070	.00076	.00042	.00067
standard	N	.0012	.00080	.00068	.00076	.00041	.00065
deviation	I	.0021	.0023	.0022	.0025	.0024	.0028
estimate	T	.00053	.00070	.00057	.00035	.00039	.00063

A comparison of predictions with simulation estimates for the MMPP/M/s/0 model with $s = 400$, $\lambda = 360$, $\mu = 1$, $t = 2700$, two environment states, $m_h = m_l$ and $c_a^2 = 8.2$ in Example 10.6.

simulation results. On the other hand, the blocking approximation (15) depends on z and thus on m. The blocking approximation .089 from eq. (15) based on $z \approx 5.0$ from eq. (14) evidently is accurate to within about 10% in the first two cases, but not in the last two cases. The displayed blocking predictions using the exact peakedness from eq. (13) provide a significant improvement. It plays no role in eqs. (16)–(18) because $\gamma = 0$.

As a second concrete example, we reduce the arrival rate by multiplying the rates λ_h and λ_l by 0.9, but leave m unchanged. This makes the new arrival rate $\lambda = 360$ and the new asymptotic variance $c_a^2 = 8.2$. Simulation results for this case are given in Table X. In this case, because $\gamma \neq 0$, the peakedness z appears in all four approximation formulas (15)–(18). As before, we use the exact peakedness eq. (13) in the predictions, which helps greatly in the last two cases with large m.

11. THE INITIAL CONDITIONS

Just as with the asymptotic variance, for functions of Markov chains the asymptotic bias for any initial distribution can be calculated by solving Poisson's equation; see (32) and Corollary 4 to Proposition 10 of Whitt [1992]. Hence we can numerically investigate the M/M/s/0 model and more complicated Markov loss models. For example, Table XI displays the asymptotic bias for the indirect and time-congestion estimators in the M/M/s/0 model with $s = 400$, $\mu = 1$, and several values of ρ, starting empty or full, computed in this manner. For $\lambda = 380$ and simulation run of length

Table XI.

traffic intensity ρ	indirect estimator		time-congestion estimator	
	$N(0) = 0$	$N(0) = s$	$N(0) = 0$	$N(0) = s$
0.7	1.00	−0.42	-0.15×10^{-10}	0.0085
0.8	1.00	−0.24	-0.94×10^{-5}	0.013
0.9	0.99	−0.086	−0.010	0.028
1.0	0.85	−0.0123	−0.0115	0.027
1.1	0.66	−0.0019	−0.23	0.015
1.2	0.52	−0.0005	−0.30	0.0087

The asymptotic bias for the indirect and time-congestion estimators in the M/M/s/0 model with $s = 400$, starting empty or full, discussed in Section 11.

5,400 as in Example 2.1, the approximate bias of the time-congestion estimator starting empty is $.01/5400 \approx 0.19 \times 10^{-7}$, whereas the approximate bias of the indirect estimator starting empty is $0.94/5400 \approx 0.00017$. The time-congestion-estimator bias is negligible, but the indirect-estimator bias is of the same order as the approximate standard deviation 0.00066 in Table III. Thus some effort to reduce the bias evidently can be worthwhile.

Insight into appropriate procedures for addressing the initialization bias can be gained by considering the associated infinite-server models. In the G/GI/∞ model starting empty, the bias of the estimator $\hat{n}(t)$ is *exactly*

$$E\hat{n}(t) - n = -nH_e^c(t), \tag{73}$$

where $H_e(t)$ is the service-time stationary-excess cdf in eq. (46); see (20) of Eick et al. [1993]. (The M/GI/∞ result there remains true for G arrival processes; see Remark 2.3 of Massey and Whitt [1993].) Hence the asymptotic bias is

$$\beta_n = -n(c_s^2 + 1)/2\mu; \tag{74}$$

see (2) of Eick et al. [1993]. As a consequence, in light loading the approximate bias of the indirect estimator is

$$\beta_I \approx \frac{-\beta_n}{\alpha} = \frac{(c_s^2 + 1)}{2\mu}. \tag{75}$$

In the case of M service with $\mu = 1$, formula (75) implies that $\beta_I \approx 1$, which is substantiated by Table XI.

Recall that the asymptotic variance σ_I^2 tends to be inversely proportional to λ. In contrast, formula (75) implies that the asymptotic bias tends to be *independent* of λ. Hence the bias becomes relatively more important as system size grows.

Formulas (73) and (75) can be used to estimate the remaining bias if we

eliminate an initial portion of the run of length t_0. Let $\beta_I(t_0)$ be this remaining bias. Then

$$\beta_I(t_0) = \int_{t_0}^{\infty} H_e^c(u)du. \tag{76}$$

For example, with M service with $\mu = 1$,

$$\beta_I(t_0) = \int_{t_0}^{\infty} e^{-u}du = e^{-t_0}. \tag{77}$$

Because $e^{-2} = 0.135$ and $e^{-5} = .0067$, the time-dependent mean reaches 86% and 99.3% of its steady-state value by 2 and 5 mean service times, respectively, and a corresponding part of the bias is reduced by eliminating the initial portion.

The infinite-server analysis is roughly consistent with asymptotical results as $s \to \infty$ for the transient blocking probability in the M/M/s/0 model by Mitra and Weiss [1989]. Roughly speaking, these results imply that the blocking probability at time t has reached about 90% of its steady-state value approximately at time

$$t = \begin{cases} 2 + \log(s(1-\rho)) & \rho < 1 \\ 2 + 1/2\log(s/2) & \rho = 1 \\ \log(\rho/(\rho-1)) & \rho > 1 \end{cases}. \tag{78}$$

For $s = 10^3$ and $\rho = 1$, the time is $t \approx 5.1$, which is about the same as the infinite-server result. This analysis suggests that the initial portion to delete is a period lasting about 5 mean service times, with the amount perhaps increasing very slowly with s. Hence for the experiments in Section 2 we deleted an initial portion of length 5 from each run. Table XI gives an idea of the bias reduction.

For example, this phenomenon occurs when $s = 10^4$. The required run length in steady state can be about 1, whereas the required run length to reduce bias starting empty can be about 5. For such large systems, it clearly can be much better to initialize the system closer to the steady-state mean.

Example 11.1 To illustrate this phenomenon, we simulate several GI/M/s/0 systems with $s = 10^4$ and $\mu = 1$. *We let the total run length be 1.* Of course, when we start the system empty, no blocking at all is observed. Hence we start the system with 9,980 customers and do not delete an initial portion. Estimation results for the indirect estimator in seven cases are shown in Table XII. The statistical precision seems adequate. The asymptotic bias can be shown to be negligible in the M arrival case by the exact numerical algorithm. Other runs with other initial conditions confirm

Table XII.

λ	c_a^2	c_s^2	$\hat{B}_I(t)$	est. std. dev.
10,000	1	1	.0084	.00079
12,000	1	1	.1681	.00060
20,000	1	1	.50063	.00011
20,000	1	1	.50099	.00029
20,000	1	1	.50093	.00024
20,000	9	1	.5011	.0012
10,000	9	1	.0110	.0015

Estimation for the GI/M/s/0 model with $s = 10^4$ and $\mu = 1$ with total run length 1 and $N(0) = 9980$ but no initial portion deleted in Example 11.1.

that there is negligible bias for these runs. Moreover, the results are relatively insensitive to $N(0)$ provided that it is not far from s.

Unfortunately, the good results in Table XII fail to hold if we change the service-time distribution. The difficulty is that all customers in service at time 0 would actually not be starting their service times at that time. For a simple example, consider the G/D/s/0 model with $\mu = 1$ and total run length $t = 1$. None of the customers initially in the system would leave prior to time 1 if they all began service at time 0. There is no difficulty with exponential service times, because the remaining service time is again exponential.

The infinite-server model can give an idea about what is an appropriate residual service-time distribution at time 0 in the G/GI/s/0 model. In the infinite-server model, the expected number of customers with service times greater than x in the system at any time is $\alpha H_e^c(x)$, by the argument of Theorem 1 in Eick et al. [1993]. Hence it is natural to let the customers in the system have i.i.d. service times distributed according to H_e. Although this initialization should yield a good approximation for G/GI/s/0 models, it is only an approximation, whose error is yet to be determined. Thus with nonexponential service times in practice, it might well prove to be more convenient to initialize the system by starting empty and deleting an appropriate initial portion based on eq. (79). This is an area where more work needs to be done.

12. PROOF OF THEOREM 4.1

In this section we construct the ROU diffusion process and prove that the normalized G/M/s/0 number-in-system process $(N_s(\cdot) - s)/\sqrt{\alpha}$ converges to it as $s \to \infty$ with $(\alpha - s)/\sqrt{\alpha} \to \gamma$. For this purpose, we construct coupled bounding processes for each s and show that these converge appropriately as $s \to \infty$ to bounding limit processes, which in turn converge to a unique limit as the bounds are tightened. The key idea is to exploit previous results for the associated G/M/∞ model.

For any $\epsilon > 0$ and positive integer s, let $N_s^{u,\epsilon}(t)$ be the upper bound process and $N_s^{\ell,\epsilon}(t)$ the lower bound process to be defined. Let these processes have the same arrival process as $N_s(t)$, that is, the same

arrival-process sample paths. Let each process have jumps down with intensity μk when the process is in state k. We modify each process whenever it hits an upper barrier. As usual, jumps up in N_s are ignored when $N_s(t) = s$. Let $N_s^{u,\epsilon}$ have an instantaneous jump down of $\lceil \epsilon\sqrt{s} \rceil$ whenever $N_s^{u,\epsilon}$ would reach $s + \lceil \epsilon\sqrt{s} \rceil$, where $\lceil x \rceil$ is the least integer greater than or equal to x. Similarly, let $N_s^{\ell,\epsilon}$ have an instantaneous jump down of $\lfloor \epsilon\sqrt{s} \rfloor$ whenever $N_s^{\ell,\epsilon}$ would hit s, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Thus the processes N_s, $N_s^{u,\epsilon}$, and $N_s^{\ell,\epsilon}$ have state spaces equal to the set of nonnegative integers less than or equal s, $s - 1 + \lceil \epsilon\sqrt{s} \rceil$, and $s - 1$, respectively.

These processes can be constructed so that,

$$N_s^{u,\epsilon}(t) - 2\lceil \epsilon\sqrt{s} \rceil \le N_s^{\ell,\epsilon}(t) \le N_s(t) \le N_s^{u,\epsilon}(t) \qquad \text{for all } t \quad \text{and} \quad \epsilon \text{ w.p.1}, \tag{79}$$

provided that eq. (79) holds w.p.1 at $t = 0$. To establish eq. (79) we construct all the downward jumps on the same space. Suppose that $N_s^{\ell,\epsilon}(t) \le N_s(t) \le N_s^{u,\epsilon}(t)$. Then the downward jump intensities at t are ordered as well. Hence we can make the downward jumps of $N_s^{\ell,\epsilon}$ a subset of the downward jumps of N_s, and in turn make that a subset of the downward jumps of $N_s^{u,\epsilon}$; see Whitt [1981] for background; Theorem 10 there covers the essence of eq. (79). In particular, when $N_s^{u,\epsilon}(t) = k \ge N_s(t) = j$, then a downward jump in $N_s^{u,\epsilon}$ is also a downward jump in N_s with probability j/k; otherwise N_s does not change. Similarly, if $N_s(t) = k \ge N_s^{\ell,\epsilon}(t) = j$, then a downward jump in N_s is also a downward jump in $N_s^{\ell,\epsilon}$ with probability j/k, otherwise $N_s^{\ell,\epsilon}$ does not change. This construction gives each process separately its proper distribution on $D[0, \infty)$ and provides the orderings in eq. (79); apply mathematical induction on the transition epochs to verify the orderings for all t. We note that eq. (79) does not depend on any structure in the arrival process.

Now we establish limits as $s \to \infty$ for the bounding processes $N_s^{\ell,\epsilon}$ and $N_s^{u,\epsilon}$ using the established limit for the associated G/M/∞ model, drawing on Borovkov [1967] and Theorem 1 on p. 103 of Borovkov [1984]. Somewhat more transparent proofs of the infinite-server FCLT are available in special cases: the case of GI arrivals is treated in Whitt [1982], and the case of G arrivals and service-time distributions with finite support is treated in Glynn and Whitt [1991].

Let $N_\alpha(t)$ be the number of busy servers in the G/M/∞ system with individual service rate μ and offered load α. Then the previous infinite-server result states that

$$\frac{\tilde{N}_\alpha(\cdot) - \alpha}{\sqrt{\alpha}} \Rightarrow \tilde{Y}(\cdot) \qquad \text{in} \quad D[0, \infty) \text{ as } \alpha \to \infty, \tag{80}$$

where $\tilde{Y}$ is the OU process with drift coefficient $m(x) = -\mu x$, diffusion coefficient $\sigma^2(x) = \mu(1 + c_a^2)$ with c_a^2 in eq. (10), and initial position $y - \gamma$. The limit (80) deserves some further explanation because Borovkov's theorems apply directly only to the case of the system starting empty. (An

exception is on p. 113 of Borovkov [1984], but that does not apply here.) However, with M service other initial conditions can be treated easily as well. In particular, the evolution of the customers initially in the system can be treated separately from the new arrivals. Note that $\tilde{N}_\alpha(t) = \tilde{N}_{\alpha n}(t) + \tilde{N}_{\alpha 0}(t)$, where $\tilde{N}_{\alpha n}(t)$ is the number of new arrivals still in the system at time t, whereas $\tilde{N}_{\alpha 0}(t)$ is the number of original customers that were in the system at time 0 that are still there at time t. Let $\tilde{N}_{\alpha 0}(0) = N_s(0)$ for $(s - \alpha)/\sqrt{\alpha} = -\gamma$ with μ fixed. By Theorem 4.4 of Billingsley [1968], our conditions imply that

$$\left(\frac{A_\lambda(\cdot) - \lambda \cdot}{\sqrt{\lambda}}, \frac{\tilde{N}_{\alpha 0}(0) - \alpha}{\sqrt{\alpha}}\right) \Rightarrow (Z, y - \gamma) \quad \text{in} \quad D[0, \infty) \times \mathbb{R} \tag{81}$$

as $\lambda \to \infty$ (and thus $\alpha \to \infty$). This in turn implies that

$$\left(\frac{\tilde{N}_{\alpha n}(\cdot) - \alpha(1 - e^{-\mu \cdot})}{\sqrt{\alpha}}, \frac{\tilde{N}_{\alpha 0}(\cdot) - \alpha e^{-\mu \cdot}}{\sqrt{\alpha}}\right) \Rightarrow (\tilde{Y}_n, \tilde{Y}_0)$$

$$\text{in} \quad D[0, \infty)^2 \quad \text{as} \quad \alpha \to \infty, \tag{82}$$

where $\tilde{Y}_0$ and $\tilde{Y}_n$ are independent, from which eq. (80) follows by adding the components. The limit process $\tilde{Y} = \tilde{Y}_0 + \tilde{Y}_n$ in eq. (80) can be characterized as an OU by its covariance function because it is a Gaussian process.

Inasmuch as $(\alpha - s)/\sqrt{\alpha} \to \gamma$ as $s \to \infty$, eq. (80) implies that

$$\frac{\tilde{N}_\alpha(\cdot) - s}{\sqrt{\alpha}} \Rightarrow Y(\cdot) \quad \text{in} \quad D[0, \infty) \quad \text{as} \quad s \to \infty, \tag{83}$$

where Y is the OU process with drift coefficient $m(x) = -\mu(x - \gamma)$ and diffusion coefficient $\sigma^2(x) = \mu(1 + c_a^2)$, because

$$\frac{\tilde{N}_\alpha(\cdot) - s}{\sqrt{\alpha}} = \left[\frac{(\tilde{N}_\alpha(\cdot) - \alpha)}{\sqrt{\alpha}}\right] + \frac{\alpha - s}{\sqrt{\alpha}}. \tag{84}$$

In other words, Y is the OU process $\tilde{Y}$ centered at γ instead of at 0.

Next we apply the continuous mapping theorem with the first passage-time function to deduce that

$$\frac{N_s^{u,\epsilon}(\cdot) - s}{\sqrt{\alpha}} \Rightarrow Y^{u,\epsilon}(\cdot) \quad \text{in} \quad D[0, \infty) \tag{85}$$

and

$$\frac{N_s^{\ell,\epsilon}(\cdot) - s}{\sqrt{\alpha}} \Rightarrow Y^{\ell,\epsilon}(\cdot) \quad \text{in} \quad D[0, \infty) \quad \text{as} \quad s \to \infty, \tag{86}$$

with $(\alpha - s)/\sqrt{\alpha} \to \gamma$ for each $\epsilon > 0$, where $Y^{u,\epsilon}$ and $Y^{\ell,\epsilon}$ are modifications of the OU process Y specified in the following, provided that $N_S^{u,\epsilon}(0) = N_S^{\ell,\epsilon}(0) = N_s(0)$. The process $Y^{u,\epsilon}$ has a jump down of ϵ whenever it hits ϵ, and $Y^{\ell,\epsilon}$ has a jump down of ϵ whenever it hits 0. In proving eqs. (85) and (86) we can consider successive intervals between downward jumps in $Y^{u,\epsilon}$ and $Y^{\ell,\epsilon}$ separately and recursively. In particular, we can let the process $Y^{u,\epsilon}$ have an instantaneous jump from ϵ to 0 and then be absorbed at 0 when it hits ϵ and we can let the process $N_s^{u,\epsilon}$ jump down and be absorbed at s when it hits $s + \lceil \epsilon\sqrt{s} \rceil$. Note that the first passage time constitutes a measurable function that is continuous almost surely with respect to the limit process; that is, when an OU first hits ϵ it goes above ϵ with probability one. In this way, we obtain convergence in $D[0, \infty)$ of the forms (85) and (86) for a single barrier hitting.

We next obtain convergence in the product space $D[0, \infty)^\infty$ for the sequence of processes associated with successive barrier hittings, each starting at 0, based on the remaining portion of the arrival process and being absorbed in 0 after the barrier hitting. The arrival process after each successive barrier hitting time has a FCLT with a Brownian motion limit that is independent of the history prior to that barrier hitting time by virtue of the original assumed FCLT for the arrival process. Also, the exponential service times can be regarded as starting over at each barrier hitting time.

We then obtain eqs. (85) and (86) as stated by piecing together separate versions, with each starting at the end of the last interval. The process of putting together the pieces into one single process on $D[0, \infty)$ is easily seen to be a continuous mapping from $D[0, \infty)^\infty$ to $D[0, \infty)$. For this continuity, it is important to use the Skorohod topology as opposed to the topology of uniform convergence on bounded intervals, because the jumps will not occur at precisely the same times. We treat $Y^{\ell,\epsilon}$ and $N_s^{\ell,\epsilon}$ similarly.

We now obtain our desired result by letting $\epsilon \downarrow 0$. By the ordering (79) and the limits (85) and (86),

$$Y^{u,\epsilon}(\cdot) - 2\epsilon \leq {}_{st}Y^{\ell,\epsilon}(\cdot) \leq {}_{st}Y^{u,\epsilon}(\cdot), \tag{87}$$

where $\leq_{st}$ denotes stochastic order on the function space $D[0, \infty)$; that is $\{Y_1(t) : t \geq 0\} \leq_{st} \{Y_2(t) : t \geq 0\}$ if $Ef(Y_1(\cdot)) \leq Ef(Y_2(\cdot))$ for all nondecreasing real-valued functions f on $D[0, \infty)$, with $y_1 \leq y_2$ in $D[0, \infty)$ if $y_1(t) \leq y_2(t)$ for all t. We now want to deduce that $Y^{u,\epsilon}(\cdot)$ converges as $\epsilon \to 0$. Unfortunately, this is not immediate from eq. (87). To establish the desired convergence as $\epsilon \downarrow 0$, note that by eq. (79),

$$N_s^{u,\epsilon 2^{-n}}(t) \leq N_s(t) + 2\lceil (\epsilon 2^{-2})\sqrt{s} \rceil \leq N_s^{u,\epsilon}(t) + 2\lceil (\epsilon 2^{-n})\sqrt{s} \rceil \leq N_s(t)$$
$$+ 2\lceil (\epsilon 2^{-n})\sqrt{s} \rceil + 2\lceil \epsilon\sqrt{s} \rceil \leq N_s^{u,\epsilon 2^{-n}}(t) + 2\lceil (\epsilon 2^{-n})\sqrt{s} \rceil 2\lceil \epsilon\sqrt{s} \rceil \tag{88}$$

for all t, s, and ϵ. Hence

$$Y^{u,\epsilon 2^{-n}}(\cdot) \leq {}_{st}Y^{u,\epsilon}(\cdot)\ \epsilon \leq {}_{st}Y^{u,\epsilon 2^{-n}}(\cdot) + 3\epsilon \tag{89}$$

for all n, so that

$$Y^{u,\epsilon}(\cdot) - 2\epsilon \leq {}_{st}Y^{u,\epsilon 2^{-n}}(\cdot) \leq {}_{st}Y^{u,\epsilon}(\cdot) + \epsilon \quad \text{for all} \quad n. \tag{90}$$

Hence $\{Y^{u,\epsilon 2^{-n}}(\cdot): n \geq 1\}$ is a tight sequence in $D[0, \infty)$ and any weak convergence limit $Y_r(\cdot)$ of a subsequence satisfies

$$Y^{u,\epsilon}(\cdot) - 2\epsilon \leq {}_{st}Y_r(\cdot) \leq {}_{st}Y^{u,\epsilon}(\cdot) + \epsilon \tag{91}$$

for all ϵ by eq. (90). Hence indeed $Y^{u,\epsilon}(\cdot) \Rightarrow Y_r(\cdot)$ as $\epsilon \downarrow 0$. Combining eqs. (87) and (91) yields $Y^{\ell,\epsilon}(\cdot) \Rightarrow Y_r(\cdot)$ as $\epsilon \downarrow 0$ as well. We *define* the reflected OU (ROU) process to be this common limit process Y_r.

Finally, by eq. (79),

$$\frac{N^{\ell,\epsilon}(t) - s}{\sqrt{\alpha}} \leq \frac{N_s(t) - s}{\sqrt{\alpha}} \leq \frac{N^{u,\epsilon}(t) - s}{\sqrt{\alpha}} \tag{92}$$

w.p.1 for all s, t, and ϵ, so that

$$\frac{N_s(\cdot) - s}{\sqrt{\alpha}} \rightarrow Y_r(\cdot) \qquad \text{in } D[0, \infty) \quad \text{as} \quad s \rightarrow \infty \tag{93}$$

as well.

13. SUMMARY

Our main results are workload factor approximations for the four estimators $\hat{B}_N(t)$, $\hat{B}_S(t)$, $\hat{B}_I(t)$, and $\hat{B}_T(t)$ in eqs. (1), (2), (4), and (6). (The workload factor is the product of the arrival rate and the asymptotic variance; see Section 1.6.) The approximation formulas for workload factors w_I and w_T are given in eqs. (17) and (18). The approximation formula for w_S has the same form as eq. (17). Theoretical analysis in Section 7 and numerical evidence show that $\hat{B}_N(t)$ and $\hat{B}_S(t)$ are very similar, so that results for one apply to the other. Thus w_S serves as the approximation for w_N. The canonical workload factors ψ appearing in eqs. (17) and (18) can be calculated exactly by the methods of Section 3; approximations are given in Table II. The approximations show that $\hat{B}_I(t)$ tends to be more efficient than the other estimators in heavy loading, but less efficient in light loading.

The behavior of the workload factors $w \equiv w(s, \gamma, c_a^2, c_s^2, z)$ as a function of the parameters s and $\gamma \equiv (\alpha - s)/\sqrt{\alpha}$ is strongly supported by the exact numerical results for w_S, w_I, and w_T in the M/M/s/0 model, using the algorithms in Section 3. Figures 1–3 dramatically show that these workload factors depend on s and α primarily through the single parameter $\gamma \equiv (\alpha - s)/\sqrt{\alpha}$. For the G/M/s/0 model, the convergence of the blocking probability B and the workload factor w_I to proper limits consistent with eqs. (15) and (17) as $s \rightarrow \infty$ is established by the FCLT in Section 4. Although corresponding limits for the other estimators remain to be

established, the ROU diffusion approximation lends support to the other heuristic approximations (when s is not too small and $|\gamma|$ is not too large). The fact that the workload factors are approximately independent of s implies that, for given statistical precision, the observation interval should be approximately inversely proportional to the arrival rate or system size.

For models not too different from M/M/s/0, the effects of arrival-process and service-time variability on the workload factors should be regarded as second order compared to the effects of s and γ. For the G/M/s/0 model, the role of c_a^2 in eq. (10) for characterizing the variability of a non-Poisson arrival process is supported by the FCLT in Section 4 and the asymptotic approximations in Sections 5 and 6. The approximations for G/GI/s/0 models with nonexponential service times are primarily empirical, and thus much more tentative. Formulas (45) and (48) in Section 6.1 lend some theoretical support, especially for light loading. This is a good direction for future research.

ACKNOWLEDGMENT

This work was done while R. Srikant was employed at AT&T Bell Laboratories.

REFERENCES

ASMUSSEN, S. 1989. Validating heavy traffic performance of regenerative simulation. *Stochastic Models 5*, 617–628.

ASMUSSEN, S. 1992. Queueing simulation in heavy traffic. *Math. Oper. Res. 17*, 84–111.

BACCELLI, F. AND BREMAUD, P. 1994. *Elements of Queueing Theory*. Springer Verlag, Berlin.

BENES, V. E. 1961. The covariance function of a simple trunk group, with applications to traffic measurements. *Bell Syst. Tech. J. 40*, 117–148.

BERGER, A. W. AND WHITT, W. 1992. The Brownian approximation for rate-control throttles and the G/G/1/C queue. *J. Discrete Event Dynamic Syst. 2*, 7–60.

BILLINGSLEY, P. 1968. *Convergence of Probability Measures*. Wiley, New York.

BOROVKOV, A. A. 1967. On limit laws for service processes in multi-channel systems. *Siberian Math. J. 8*, 746–763.

BOROVKOV, A. A. 1976. *Stochastic Processes in Queueing Theory*. Springer Verlag, New York.

BOROVKOV, A. A. 1984. *Asymptotic Methods in Queuing Theory*. Wiley, New York.

CARSON, J. S. AND LAW, A. M. 1980. Conservation equations and variance reduction in queueing simulations. *Oper. Res. 28*, 535–546.

COX, D. R. AND MILLER, H. D. 1965. *The Theory of Stochastic Processes*. Wiley, New York.

DAVIS, J., MASSEY, W. A., AND WHITT, W. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Manage. Sci. 41*, 1107–1116.

ECKBERG, A. E. 1983. Generalized peakedness of teletraffic processes. In *Proceedings of the Tenth International Teletraffic Congress* (Montreal, June), 4.4b.3.

EICK, S. G., MASSEY, W. A., AND WHITT, W. 1993. The physics of the $M_t/G/\infty$ queue. *Oper. Res. 41*, 731–742.

ERRAMILLI, A., GORDON, J., AND WILLINGER, W. 1994. Applications of fractals in engineering for realistic traffic processes. In *Proceedings of ITC '94*. J. Labetoullle and J. W. Roberts, Eds., Elsevier, Amsterdam, 35–44.

ETHIER, S. N. AND KURTZ, T. G. 1986. *Characterization and Approximation of Markov Processes*. Wiley, New York.

FENDICK, K. W. AND WHITT, W. 1989. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. In *Proc. IEEE 77*, 171–194.

GLYNN, P. W. 1982. On the Markov property of the GI/G/∞ Gaussian limit. *Adv. Appl. Probab. 14*, 191–194.

GLYNN, P. W. AND WHITT, W. 1989. Indirect estimation via L = λW. *Oper. Res. 37*, 82–103.

GLYNN, P. W. AND WHITT, W. 1991. A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Probab. 23*, 188–209.

GLYNN, P. W. AND WHITT, W. 1992. The asymptotic efficiency of simulation estimators. *Oper. Res. 40*, 505–520.

GLYNN, P. W., MELAMED, B., AND WHITT, W. 1993. Estimating customer and time averages. *Oper. Res. 41*, 400–408.

GRASSMANN, W. K. 1987. The asymptotic variance of a time average in a birth-death process. *Ann. Oper. Res. 8*, 165–174.

HALFIN, S. AND WHITT, W. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res. 29*, 567–587.

HEYMAN, D. P. 1980. Comments on a queueing inequality. *Manage. Sci. 26*, 956–959.

JAGERMAN, D. L. 1974. Some properties of the Erlang loss functions. *Bell Syst. Tech. J. 53*, 525–551.

KELLY, F. P. 1991. Loss networks. *Ann. Appl. Probab. 1*, 319–378.

LAVENBERG, S. S., MOELLER, T. L., AND WELCH, P. D. 1982. Statistical results on control variables with application to queueing network simulation. *Oper. Res. 30*, 182–202.

LAW, A. M. 1975. Efficient estimators for simulated queueing systems. *Manage. Sci. 22*, 30–41.

LUCANTONI, D. M. 1993. The BMAP/G/1 queue: A tutorial. In *Models and Techniques for Performance Evaluation of Computer and Communications Systems*. L. Donatiello and R. Nelson, Eds., Springer-Verlag, New York, 330–358.

MASSEY, W. A. AND WHITT, W. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Syst., 13*, 183–250.

MELAMED, B. AND WHITT, W. 1990. On arrivals that see time averages. *Oper. Res. 38*, 156–172.

MITRA, D. AND WEISS, A. 1989. The transient behavior in Erlang's model for large trunk groups and various traffic conditions. In *Teletraffic Science for the New Cost-Effective Systems, Networks and Services, Proceedings of ITC 12*, M. Bonatti, Ed., North Holland, Amsterdam, 1367–1374.

NEUTS, M. F. 1989. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker, New York.

PAXSON, V. AND FLOYD, S. 1995. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Trans. Networking 3*, 226–244.

RIORDAN, J. 1962. *Stochastic Service Systems*. Wiley, New York.

ROSS, K. W. 1995. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer Verlag, London.

ROSS, K. W. AND WANG, J. 1992. Monte Carlo summation applied to product-form loss networks. *Probab. Eng. Inf. Sci. 6*, 323–348.

ROSS, S. M. 1993. *Introduction to Probability Models*, fifth ed. Academic Press, New York.

SOBEL, M. 1980. Simple inequalities for multiserver queues. *Manage. Sci. 26*, 951–956.

SRIKANT, R. AND WHITT, W. 1995. Variance reduction in simulations of loss models. AT&T Bell Laboratories, Murray Hill, NJ. *Oper. Res.* (submitted).

TAKACS, L. 1962. *Introduction to the Theory of Queues*. Oxford University Press, New York.

WHITT, W. 1981. Comparing counting processes and queues. *Adv. Appl. Probab. 13*, 207–220.

WHITT, W. 1982. On the heavy-traffic limit theorem for GI/G/∞ queues. *Adv. Appl. Probab. 14*, 171–190.

WHITT, W. 1984. Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J. 63*, 689–708.

WHITT, W. 1989. Planning queueing simulations. *Manage. Sci. 35*, 1341–1366.

WHITT, W. 1991a. The efficiency of one long run versus independent replications in steady-state simulation. *Manage. Sci. 37*, 645–666.

WHITT, W. 1991b. A review of L = λW and extensions. *Queueing Syst. 9*, 235–268.

WHITT, W. 1992. Asymptotic formulas for Markov processes with applications to simulation. *Oper. Res. 40*, 279–291.

WILLIAMS, R. J. 1992. Asymptotic variance parameters for the boundary local times of reflected Brownian motion on a compact interval. *J. Appl. Probab. 29*, 996–1002.

WILLINGER, W. 1995. Traffic modeling for high-speed networks: theory versus practice. In *IMA Volumes in Mathematics and its Applications*, F. P. Kelly and R. J. Williams Eds., Springer-Verlag, New York (to appear).

WOLFF, R. W. 1982. Poisson arrivals see time averages. *Oper. Res. 30*, 223–231.

Received July 1995; revised October 1995; accepted December 1995