Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

SIMULATION RUN LENGTH PLANNING FOR STOCHASTIC LOSS MODELS

Rayadurgam Srikant

Department of General Engineering
and Coordinated Science Laboratory
University of Illinois
Urbana, IL 61801, USA

Ward Whitt

AT&T Bell Laboratories
Murray Hill, NJ 07974-0636, USA

ABSTRACT

We derive approximate formulas for the asymptotic variance of estimators of the steady-state blocking probability in a multi-server loss system. 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. 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. As the system size increases, the time to approach steady state becomes a greater portion of the overall simulation time as system size increases.

1 INTRODUCTION

This paper is in the spirit of a previous WSC paper, Whitt (1989a), and Whitt (1989b, 1992), which focused on developing formulas that can be used to estimate required simulation run lengths in the early planning stages before any data have been collected. As in Whitt (1989a,b), we focus on a class of queueing models, but now we consider loss models instead of delay models.

In particular, we consider the problem of estimating steady-state blocking probabilities in a multi-server loss system. We are interested in loss networks, as in 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). Arrivals that find all servers busy are lost (blocked) without affecting future arrivals.

We assume that the data are collected after the system has reached steady state. Hence, there is 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 simulation run length assuming that the system starts in steady state. Then we consider the initial portion that needs to be deleted when the system starts empty for the system to be approximately in steady state.

1.1 The Candidate Estimators

The *natural estimator* for the steady-state blocking probability B based on observations 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 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*. The estimators $\hat{B}_N(t)$ and $\hat{B}_S(t)$ behave similarly, so we regard results for $\hat{B}_S(t)$ as being applicable to $\hat{B}_N(t)$.

As in 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. From $L = \lambda W$, we get the relation $n = \lambda(1 - B)/\mu$. Assuming that we know λ and μ, as would be the case with many simulations, we can use the *indirect estimator*

$$\hat{B}_I(t) \equiv 1 - \frac{\hat{n}(t)}{\alpha}, \tag{3}$$

where

$$\hat{n}(t) = t^{-1}\int_0^t N(u)du, \quad t \geq 0 . \qquad (4)$$

1.2 The Asymptotic Variance

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

$$\hat{\sigma}^2 = \lim_{t\to\infty} t \ Var \ \hat{B}(t). \qquad (5)$$

We use subscripts N, S and I to refer to the specific estimators defined above. Under regularity conditions (which includes the requirement that the asymptotic variance actually be finite), for suitably large run times t, each estimator $\hat{B}(t)$ tends to be approximately normally distributed with a variance $\hat{\sigma}^2/t$, where $\hat{\sigma}^2$ is the asymptotic variance (which depends on the estimator). Hence, a $(1-\beta)$ 100% confidence interval for B will be $[\hat{B}(t) - h(\beta), \hat{B}(t) + h(\beta)]$ with halfwidth

$$h(\beta) = \frac{\hat{\sigma} z_{\beta/2}}{\sqrt{t}} , \qquad (6)$$

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. Thus, for *specified halfwidth* ϵ and *level of precision* β, the *required simulation run length* is

$$t(\epsilon, \beta) = \frac{\hat{\sigma}^2 z_{\beta/2}^2}{\epsilon^2} . \qquad (7)$$

We aim to develop approximations for the asymptotic variances $\hat{\sigma}_S^2, \hat{\sigma}_N^2$ and $\hat{\sigma}_I^2$. Roughly speaking, we find that $\hat{\sigma}_S^2 \approx \hat{\sigma}_N^2$ but that $\hat{\sigma}_I^2$ can be quite different. 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 $\hat{\sigma}_I^2 < \hat{\sigma}_S^2$ when $\alpha > s$, whereas we tend to have $\hat{\sigma}_I^2 > \hat{\sigma}_S^2$ when $\alpha < s$.

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. The principal way we partially characterize the variability of the arrival process is through its *normalized arrival asymptotic variance*, defined by

$$c_a^2 = \lim_{t\to\infty} \frac{Var \ A(t)}{\lambda t} , \qquad (8)$$

which we assume is well defined (the limit exits and is finite). For the special case of a renewal process, c_a^2 coincides with the *squared coefficient of variation* (SCV) of an interarrival time; i.e., if U is an interarrival time, then

$$c_a^2 = Var \ (U)/(EU)^2 . \qquad (9)$$

For non-renewal processes, formula (8) captures correlations between different interarrival times. A large class of non-renewal arrival processes can be represented as batch Markovian arrival processes (BMAPs) or versatile Markovian point processes. The normalized arrival asymptotic variance of a BMAP is given on p. 284 of Neuts (1989).

Since we have assumed that the service times are i.i.d. and independent of the arrival process, their variability is easier to characterize. We primarily characterize the service-time variability via the service-time SCV, denoted by c_s^2, and defined as in (9).

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. In particular, the G/GI/s/0 model variability can be partially characterized by the *peakedness* parameter z, which is the ratio of the variance to the mean number of busy servers in the associated G/GI/∞ model.

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

$$z = 1 + (c_a^2 - 1)\mu \int_0^\infty [1 - H(t)]^2 dt . \qquad (10)$$

When the service time cdf H in (10) 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 (10) for all service-time distributions when $c_a^2 = 1$.

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 principle 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.

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 (1984), Whitt (1984), 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.

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 itself. Asymptotics for the GI/M/s/0 model in the case of normal loading has produced the following approximation for the blocking probability:

$$B \approx \sqrt{z/\alpha}\frac{\phi(\gamma/\sqrt{z})}{\Phi(-\gamma/\sqrt{z})} , \qquad (11)$$

where $\gamma = (\alpha - s)/\sqrt{\alpha}$, $z = (c_a^2+1)/2$, and $\alpha \equiv \lambda/\mu$ is the offered load, while $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of the standard (mean 0, variance 1) normal distribution; see (13) of Whitt (1984).

Approximation (11) is most strongly supported in the case of exponential service times, but it can also be used with general service times if we use the appropriate peakedness z. The best value for z should be the exact peakedness, but (10) is a convenient approximation.

2 WORKLOAD FACTORS

Formula (7) shows that the required simulation time t to achieve desired statistical precision is approximately proportional to the asymptotic variance $\hat{\sigma}^2$. However, the *computational effort* required to simulate for time t is approximately proportional to λt, because λt is the expected number of 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 $w \equiv \lambda\hat{\sigma}^2$, which we call the *workload factor*.

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

$$w_I(s,\gamma,c_a^2,c_s^2,z) \approx \frac{(c_a^2+c_s^2)}{2}\psi_I(\gamma/\sqrt{z}) , \qquad (12)$$

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, $\gamma = (\alpha - s)/\sqrt{\alpha}$, z is the peakedness, c_a^2 is the normalized arrival asymptotic variance in (8) and c_s^2 is the SCV of the service-time distribution, defined as in (9). Note that the arrival-process variability enters in via both c_a^2 and z, and that the service-time distribution enters in via both c_s^2 and z. As with (11), the preferred peakedness z is the exact value, but (10) usually is a satisfactory approximation.

The approximation we propose for the workload factor of the simple estimator has the same form; just replace the two I subscripts in (12) by S. Since $\hat{B}_N(t) \approx \hat{B}_S(t)$, we propose approximating w_N by w_S.

It is significant that the asymptotic variance and, thus, the canonical workload factors for the M/M/s model can readily be computed using p. 288 of Whitt (1992) and p. 89 of Riordan (1962). The notion of a canonical workload curve for M/M/s/0 models is supported by Figures 1 and 2, which display the exact workload factors $w(s,\gamma,1,1,1)$ for the estimators $\hat{B}_S(t)$ and $\hat{B}_I(t)$ in the M/M/s/0 model for different values of s, assuming that $\mu = 1$. 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.

Note that $\psi_I(\gamma)$ is small for $\gamma > 0$ while $\psi_S(\gamma)$ is small for $\gamma < 0$, showing that *different estimators should be strongly preferred in different regions.*

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 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.

3 A DIFFUSION LIMIT

We also provide theoretical support for the workload factor approximation in (12). In particular, we present a heavy-traffic *functional central limit theorem* (FCLT) in the case for the G/M/s/0 model,

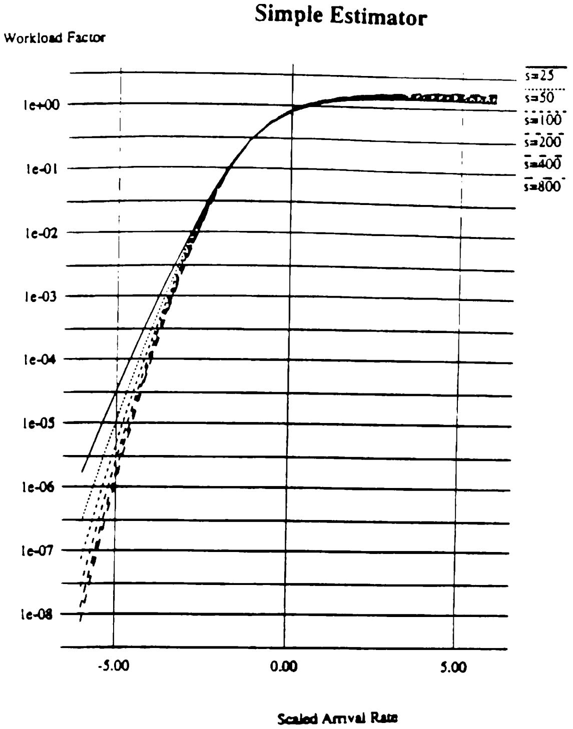

Figure 1: Workload factors $w_S \equiv \lambda\hat{\sigma}_S^2$ for the simple estimator $\hat{B}_S(t)$ in the M/M/s/0 loss model as a function of the scaled arrival rate $\gamma \equiv (s-\alpha)/\sqrt{\alpha}$ for several values of s.

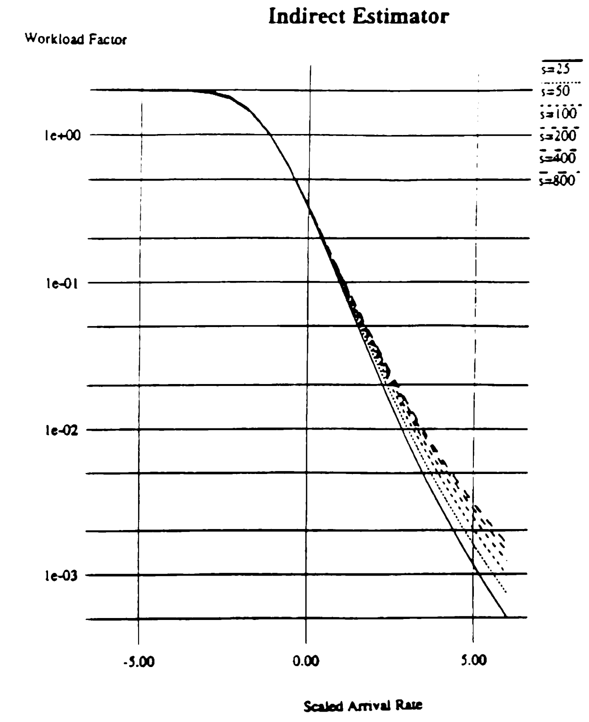

Figure 2: Workload factors $w_I \equiv \lambda\hat{\sigma}_I^2$ for the indirect estimator $\hat{B}_I(t)$ in the M/M/s/0 loss model as a function of the scaled arrival rate $\gamma \equiv (s-\alpha)/\sqrt{\alpha}$ for several values of s.

which we prove elsewhere. Let $\Rightarrow$ denote 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; e.g., see Billingsley (1968). The convergence in $D[0,\infty)$ is useful for us to treat general stationary arrival processes and to get convergence of the bivariate distributions of the content process, 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 3.1. *Consider the G/M/s/0 model with 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*

$$\frac{N_s(\cdot) - s}{\sqrt{\alpha}} \Rightarrow Y_r(\cdot) \text{ in } D[0,\infty) \text{ as } s \to \infty\,, \qquad (13)$$

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

Theorem 3.1 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 3.1 here parallel the conditions in Theorem 1 on p. 103 of Borovkov (1984) for the G/GI/∞ model.

The ROU limit in (13) 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 3.1 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

$$N_s(\cdot) \approx s + \sqrt{\frac{\alpha(1+c_a^2)}{2}} Y_r(\cdot; -(x - \gamma\sqrt{2/(1+c_a^2)}, 2)\,. \tag{14}$$

Thus the asymptotic variance $\hat{\sigma}_n^2$ of $N_s(t)$ is approximately

$$\hat{\sigma}_n^2 \approx \alpha \frac{(1+c_a^2)}{2} \hat{\sigma}^2_{Y_r(\cdot; -(x-\gamma\sqrt{2/(1+c_a^2)}, 2)}\,. \tag{15}$$

Only the single parameter $\gamma\sqrt{2/(1+c_a^2)}$ appears inside the ROU Y_r in (14) and thus inside the asymptotic variance term $\hat{\sigma}^2_{Y_r}$ in (15). The asymptotic variance term $\hat{\sigma}^2_{Y_r(t;-(x-\gamma),2)}$ remains to be calculated, but it clearly is a function of only the one parameter.

Combining (3) and Theorem 3.1, we obtain convergence of the bivariate distributions for any two time points. Assuming that we can approximate the covariance function of the queueing process by the covariance function of the ROU, we obtain

$$\lambda\hat{\sigma}_I^2(GI/M/s/0) \approx \frac{(1+c_a^2)}{2} \hat{\sigma}^2_{Y_r(\cdot;-(x-\gamma\sqrt{2/(1+c_a^2)}),2)}\,. \tag{16}$$

Theorem 3.1 suggests that we should look at the workload factors as functions of γ for $(\alpha - s)/\sqrt{\alpha} = \gamma$. When we do, we find canonical curves for all the workload factors.

4 OTHER APPROXIMATIONS

We also develop other approximations based on asymptotics as $s \to \infty$ with ρ held fixed, with either $\rho < 1$ (light loading) or $\rho > 1$ (heavy loading), derived elsewhere. These approximations are shown in Table 1. 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 $\rho < 1$, while $w_I/w_S \to 0$ as $s \to \infty$ for $\rho > 1$. Moreover, these formulas also serve as simple approximations. Since we already have reduced the G/GI/s/0 case to the M/M/s/0 case in (12), we primarily use the formulas in Table 1 as convenient simple approximations for the canonical (M/M/s/0) workload factors ψ (obtained by letting $c_a^2 = 1$ in Table 1).

With the exception of the light-loading simple-estimator formula, the formulas in Table 1 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 (11), which yields

$$w_S(s, \gamma, c_a^2) \approx \frac{(1+c_a^2)^{3/2}}{-2\rho\gamma\sqrt{\pi}} e^{-\gamma^2/(1+c_a^2)} \text{ in light loading}\,. \tag{17}$$

	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	$1 + c_a^2$	$\frac{(1+\rho)(1+c_a^2)^3}{4s^2(\rho-1)^4}$

Table 1. Approximation formulas for the workload factor $w \equiv \lambda\hat{\sigma}^2$ of the estimators in (1), (2) and (3) for the G/M/s/0 model in light, normal and heavy loading.

(Let $w(s, \gamma, c_a^2) \equiv w(s, \gamma, c_a^2, 1, (c_a^2 + 1)/2)$.) Formula (17) approximately satisfies the general functional form (12) with

$$\psi_S(\gamma) = 2\gamma^{-1}\phi(\gamma) = \sqrt{2/\pi\gamma^2} e^{-\gamma^2/2}\,. \tag{18}$$

Similarly, the heavy-traffic indirect-estimator workload factor approximation in Table 1 can be expressed approximately as

$$w_I(s, \gamma, c_a^2) \approx \frac{(1+c_a^2)^3}{2\rho^4\gamma^4}\,, \tag{19}$$

which is approximately of the form (12) with

$$\psi_I(\gamma) = 4\gamma^{-4}\,. \tag{20}$$

All four formulas in Table 1 are consistent with (12) in the limit as $s \to \infty$ with $(\alpha - s)/\sqrt{\alpha} \to \gamma$ (so that $\rho \to 1$).

Approximations (17)-(20) reveal the essential form of the workload factors in light and heavy loading, but these formulas are not very accurate, e.g., when compared to the exact M/M/s/0 results.

5 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 $\hat{\sigma}^2/t$, where $\hat{\sigma}^2$ is the asymptotic variance, when t is suitably large. This large sample behavior is well established in statistical experience, but we also have confirmed this directly. We give two examples here.

Example 5.1. 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 distribution with balanced means and $c_a^2 = 9.0$. This H_2^b distribution has density

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \; x \geq 0 \; , \quad (21)$$

where

$$p = [1 + \sqrt{(c_a^2 - 1)/(c_a^2 + 1)}]/2 \; , \quad (22)$$

$$\lambda_1 = 2p\lambda^{-1} \text{ and } \lambda_2 = 2(1-p)\lambda^{-1} \quad (23)$$

with λ^{-1} being the mean. Since the service-time distribution is exponential, the approximate peakedness by (10) is $z = (c_a^2 + 1)/2 = 5$. (The exact peakedness is 4.95, so (10) is an excellent approximation.)

We consider $s = 400$, $\mu = 1$ and three values of $\lambda : \lambda = 360$, $\lambda = 400$ and $\lambda = 440$. The experiment consists of 2 independent runs of length 2700 for each λ, deleting a portion of length 5 to allow the system to approach steady state in each case. (The run length 2700 makes the expected total number of arrivals about 10^6 in each case.) The variances were estimated from 20 nonoverlapping batch means. The simulation results are displayed in Table 2. The predictions in Table 2 based on (11) and (12) seem very good. Table 2 shows that $\hat{B}_N(t)$ essentially coincides with $\hat{B}_S(t)$.

$\lambda = 360$		predicted	run 1	run 2
blocking	S	.036	.0309	.0306
probability	N	.036	.0308	.0306
estimate	I	.036	.0303	.0334
standard	S	.0014	.0011	.0012
deviation	N	.0014	.0011	.0011
estimate	I	.0020	.0020	.0017
$\lambda = 400$		predicted	run 1	run 2
blocking	S	.089	.0811	.0854
probability	N	.089	.0813	.0849
estimate	I	.089	.0833	.0804
standard	S	.0020	.0014	.0015
deviation	N	.0020	.0013	.0013
estimate	I	.0012	.0015	.0011
$\lambda = 440$		predicted	run 1	run 2
blocking	S	.150	.1397	.1430
probability	N	.150	.1400	.1428
estimate	I	.150	.1431	.1417
standard	S	.0023	.0027	.0020
deviation	N	.0023	.0023	.0017
estimate	I	.00075	.00082	.00075

Table 2. A comparison of predictions with simulation results for the GI/M/s/0 model with $s = 400$, H_2^b interarrival times having $c_a^2 = 9.0$ and service rate $\mu = 1$ in Example 5.1.

Example 5.2. 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 (10) and (11), 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 formula (12).

To illustrate how approximation (12) applies 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 $c_s^2 = 9.0$. Simulation results for 2 independent runs of length 2700 are displayed in Table 3. Because of the more variable service times, we delete a period of length 50 in each run. (See (29) below.) Note that the blocking probabilities are well predicted by formula (11) with $z = 1$. The standard deviation estimates are also reasonably well predicted by (12) as well. Notice that the prefactor $(c_a^2 + c_s^2)/2 = 5$ in (12) plays an important role here.

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

Table 3. A comparison of predictions with simulation results for the M/G/s/0 model with $s = 400$, service times having $c_s^2 = 9$ and $\mu = 1$ in Example 5.2.

6 THE INITIAL CONDITIONS

Since we cannot start the simulation in steady-state, the estimators necessarily have initialization bias, i.e., 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

$$\beta = \lim_{t \to \infty} t(E\hat{B}(t) - B) \ . \qquad (24)$$

We use (24) to justify the approximation $E\hat{B}(t) - B \approx \beta/t$. Since $SD(\hat{B}(t)) \approx \hat{\sigma}/\sqrt{t}$, 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.

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 4 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 5400, indirect estimator starting empty is $0.94/5400 \approx 0.00017$. The bias is of the same order as the approximate standard deviation 0.00066. Thus some effort to reduce the bias evidently can be worthwhile.

ρ	$N(0) = 0$	$N(0) = s$
0.7	1.00	−0.42
0.8	1.00	−0.24
0.9	0.99	−0.086
1.0	0.85	−0.0123
1.1	0.66	−0.0019
1.2	0.52	−0.0005

Table 4. The asymptotic bias for the indirect estimator in the M/M/s/0 model as a function of the traffic intensity ρ with $s = 400$, starting empty or full.

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*

$$E\hat{n}(t) - n = -nH_e^c(t) \ , \qquad (25)$$

where $H_e(t)$ is the service-time stationary-excess cdf, i.e.,

$$H_e^c(t) = 1 - H_e(t) = \mu \int_t^\infty H^c(u)du \qquad (26)$$

where $H^c(t) = 1 - H(t)$; see (20) of Eick, Massey and Whitt (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 of $\hat{n}(t)$ is

$$\beta_n = -n(c_s^2 + 1)/2\mu \ . \qquad (27)$$

As a consequence, in light loading the approximate bias of the indirect estimator is

$$\beta_I \approx \frac{-\beta_n}{\alpha} = \frac{(c_s^2 + 1)}{2\mu} \ . \qquad (28)$$

In the case of M service with $\mu = 1$, formula (28) implies that $\beta_I \approx 1$, which is substantiated by Table 4.

Recall that the asymptotic variance $\hat{\sigma}_I^2$ tends to be inversely proportional to λ. In contrast, formula (28) implies that the asymptotic bias tends to be independent of λ. Hence, the bias becomes relatively more important as system size grows.

Formulas (25) and (28) 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

$$\beta_I(t_0) = \int_{t_0}^\infty H_e^c(u)du \ . \qquad (29)$$

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

$$\beta_I(t_0) = \int_{t_0}^\infty e^{-u}du = e^{-t_0} \ . \qquad (30)$$

Since $e^{-2} = 0.135$ 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

$$t = \begin{cases} 2 + \log(s(1-\rho)) & \rho < 1 \\ 2 + \frac{1}{2}\log(s/2) & \rho = 1 \\ \log(\rho/(\rho - 1)) & \rho > 1 \end{cases} \qquad (31)$$

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. A heuristic for more variable arrival processes based on (29) for H_2^b service times is to delete $5c_a^2$ mean service times.

For large systems, this means that most of the work can be in getting to steady state. For example, when $s = 10^4$, the required run length in steady state can be about 1, while 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. To illustrate, we simulated 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. When we start the system with 9980 customers and do not delete an initial portion, the statistical precision is adequate.

Unfortunately, these good results for non-empty initial conditions 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.

REFERENCES

Billingsley, P. 1968. *Convergence of Probability Measures.* Wiley, 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. *Opns. Res.* 28, 535–546.

Eick, S. G., Massey, W. A. and Whitt, W. 1993. The physics of the $M_t/G/\infty$ queue. *Opns. Res.* 41, 731–742.

Glynn, P. W. and Whitt, W. 1989. Indirect estimation via $L = \lambda W$. *Opns. Res.* 37, 82–103.

Glynn, P. W. and Whitt, W. 1992. The asymptotic efficiency of simulation estimators. *Opns. Res.* 40, 505–520.

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

Massey, W. A. and Whitt, W. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems.* 13, 183–250.

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.

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

Ross, K. W. 1995. *Multiservice Loss Models for Broadband Telecommunication Networks.* Springer-Verlag, New York.

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

Whitt, W. 1989a. Simulation run length planning. In *Proceedings of the 1989 Winter Simulation Conference,* eds. E. A. MacNair, K. J. Musselman and P. Heidelberger, 106–112.

Whitt, W. 1989b. Planning queueing simulations. *Management Sci.* 35, 1341–1366.

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

AUTHOR BIOGRAPHIES

RAYADURGAM SRIKANT is an Assistant Professor in the Department of General Engineering and the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign. He received his Ph.D. in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1991. The work reported here was done while he was working at AT&T Bell Laboratories.

WARD WHITT is a Member of Technical Staff in the Network Services Research Laboratory at AT&T Bell Laboratories, Murray Hill, New Jersey. He received his A.B. in Mathematics from Dartmouth College in 1964 and his Ph.D. in Operations Research from Cornell University in 1969. He has been at AT&T Bell Laboratories since 1977.