MANAGEMENT SCIENCE
Vol. 35, No. 11, November 1989
Printed in U.S.A.

PLANNING QUEUEING SIMULATIONS*

WARD WHITT
AT&T Bell Laboratories, Room 2C-178, *Murray Hill, New Jersey* 07974

Simple heuristic formulas are developed to estimate the simulation run lengths required to achieve desired statistical precision in queueing simulations. The formulas are intended to help in the early planning stages before any data have been collected. The queueing simulations considered are single replications (one long run) conducted to estimate steady-state characteristics such as expected equilibrium queue lengths. The formulas can be applied to design simulation experiments to develop and evaluate queueing approximations. In fact, this work was motivated by efforts to develop approximations for packet communication networks with multiple classes of traffic having different service characteristics and bursty arrival processes. In addition to indicating the approximate simulation run length required in each case of a designed experiment, the formulas can help determine what cases to consider, what statistical precision to aim for, and even whether to conduct the experiment at all. The formulas are based on heavy-traffic limits for queues (the limiting behavior as the traffic intensity approaches its upper limit for stability) and associated diffusion approximations. In particular, the formulas apply to stochastic processes that can be approximated by reflected Brownian motion, such as the queue-length process in the standard *GI*/G/1 model.
(SIMULATION; EXPERIMENTAL DESIGN; DETERMINING SAMPLE SIZES; QUEUES; HEAVY-TRAFFIC LIMITS; DIFFUSION APPROXIMATIONS; SCALING OF SPACE AND TIME).

1. Introduction

In this paper we address a fundamental problem in planning simulation experiments to estimate steady-state quantities associated with queueing models: estimating the required length of the simulation runs. Of course, there is a fairly well developed statistical theory for analyzing simulation output in order to estimate confidence intervals, which can be applied while conducting the experiment to determine when enough data have been collected; see Fishman (1971), Chapters 6 and 10 (especially pp. 159–165, 297) of Fishman (1973), Chapters 3 and 5 (especially pp. 120, 234) of Fishman (1978) and Lavenberg and Sauer (1977). In contrast, *the goal here is to provide formulas that can be used to estimate required simulation run lengths in the early planning stages before any data have been collected*. These estimated simulation run lengths can then be used to design the experiment, i.e., to determine what cases to consider, what statistical precision to aim for, what experimental budget is appropriate, and even whether to conduct the experiment at all. Long estimated simulation run lengths may also suggest applying variance reduction techniques; see Fishman (1973), (1978) and Wilson (1985). We are primarily interested in simulation, but the heuristic formulas also apply to statistical estimation from actual system measurements.

While we draw on general statistical theory, *our analysis is entirely for queueing models*. We exploit the structure of queueing models in order to estimate the data required to achieve desired confidence intervals. In particular, our heuristic formulas are based on heavy-traffic limits and associated diffusion approximations for queues; e.g., see Asmussen (1987a, b), Borovkov (1976), (1984), Burman (1987), Burman and Smith (1986), Fleming (1984), Harrison (1973), (1988), Harrison and Williams (1988), Iglehart and Whitt (1970), Johnson (1983), Newell (1982), Peterson (1985), Reiman (1983), (1984),

* Accepted by George S. Fishman, former Departmental Editor; received January 22, 1987. This paper has been with the author 1 month for 1 revision.

0025-1909/89/3511/1341$01.25

(1988a, b), Whitt (1971), (1974) and references cited there. The heavy-traffic limits describe how the queueing processes behave as the traffic intensity ρ approaches its critical value for stability, here always taken to be 1. It is significant that we use the heavy-traffic limits, not just for the steady-state distribution, but also for the entire queueing process. *An essential role is played by the time scaling that appears in these heavy-traffic limits.* In particular, the heavy traffic limits allow us to relate how time t or the customer index n should grow (the simulation run length) as ρ approaches 1; see (30) and (34). It appears that the heavy-traffic theory for queues has important applications to experimental design, which has not yet been fully appreciated.

However, there has been some very relevant previous work. First, Blomqvist (1967), (1968), (1969) developed heuristic formulas for the variance of the sample mean in the $GI/G/1$ model based on detailed heavy-traffic analysis. Later Moeller and Kobayashi (1974) and Woodside, Pagurek and Newell (1980) developed similar formulas based on heuristic diffusion approximations. Further discussion appears in Chapter 5 of Newell (1982; e.g., p. 151). Our work can be regarded as a continuation of these efforts. In fact, to a large extent, the supporting theory was already well established by Morse (1955). From the heavy-traffic theory, e.g., Iglehart and Whitt (1970), one can deduce that Morse's exact results for the $M/M/1$ queue can be applied to provide approximate results for more general models.

While we only consider queues, we do consider quite general queueing models, going beyond the standard $GI/G/m$ model with m servers, infinite waiting room and the FCFS (first-come first served) discipline, as well as the $M/M/1$ special case commonly discussed in textbooks and the $GI/G/1$ special case treated by Blomqvist, Moeller and Kobayashi, and Woodside et al. Over the last 25 years, heavy-traffic analysis has shown that many models have essentially the same heavy-traffic behavior as the $GI/G/1$ model, so that the approximations considered by Blomqvist actually apply much more generally. In fact, *we specify the set of models we consider by stipulating that the process of interest, with the standard normalization, converge in distribution to regulated or reflecting Brownian motion (RBM) as the traffic intensity ρ approaches the critical value for stability*; see (30) and (32). In other words, our formulas emerge from approximating the process of interest by RBM. To a large extent, the entire paper is summarized by the RBM approximation (34). (The limit theorems provide theoretical support.)

In fact, the model may even be a complex open queueing network where there are several classes of customers with a priority service discipline. We focus on a single queue, but the results are often applicable to entire queueing networks; see §5.5. Frequently, the simulation run length required for an open queueing network will be determined by a bottleneck queue (the queue with the highest traffic intensity). This simplification is mathematically supported by the phenomenon of state-space collapse for queueing networks in heavy traffic; see Reiman (1983), (1987b). However, we also suggest approximations for open queueing networks that do not rely on state-space collapse.

While the class of queueing models covered by our analysis is substantial, by no means are all queueing models in this class. Many queueing processes are *not* well approximated by RBM; then the analysis here does not apply. For example, queues with infinitely many servers and small closed queueing networks behave very differently. Nevertheless, the general method here should be applicable to other limits and other stochastic models; further discussion appears in §4.7.

This work was motivated by simulation experiments conducted to develop and evaluate approximations of steady-state characteristics of complex queueing models, in particular, by recent efforts to design experiments to study queues in packet communication networks; see Fendick, Saksena and Whitt (1989). These queues have multiple classes of customers with different service characteristics and very bursty arrival processes. Previous simulation studies of queueing models that also serve as motivation are Albin (1981), (1982),

(1984a, b), (1986), Sriram and Whitt (1986) and Whitt (1983a, b), (1984a). When conducting simulations to develop and evaluate approximations, we must typically consider several design factors, such as the traffic intensity, the number of servers, and the interarrival-time and service-time distributions (and their moments). Also there typically are special factors for particular models such as the number of component streams with superposition arrival processes. The overall experiment consists of many cases based on various combinations of different levels of each factor. In order to properly design the simulation experiment, it is necessary to estimate the length of the simulation runs required in each case in order to achieve the desired statistical precision. The experiment might well be modified as it is being conducted, e.g., in response to estimated confidence intervals, but it is very helpful to have a rough idea in advance. The formulas we propose enable practical planning with "back-of-the-envelope" calculations.

At first glance, our experimental design problem may not seem that difficult. To get a rough idea about how long the runs should be, one "practical approach" might be to simulate one case to estimate the required run lengths. However, such a preliminary experiment requires that you set up the entire simulation before you decide whether or not to do the experiment. Nevertheless, if such a sampling procedure worked, then our experimental design problem would *not* be especially difficult. Our motivation stems from the fact that one sample run can be extremely misleading. *This queueing experimental design problem is interesting and important primarily because a uniform allocation of data over all cases is not nearly appropriate.*

Experience indicates that, for given statistical precision, the required amount of data increases as the traffic intensity increases and as the arrival-and-service variability (appropriately quantified) increases. For example, recognizing this phenomenon for the traffic intensity ρ, Albin (1984a, §3.1) used sample sizes of 60,000, 90,000, 300,000, 600,000 and 1,000,000 arrivals for $\rho = 0.3, 0.5, 0.7, 0.8$ and 0.9, respectively, in a simulation experiment to estimate the expected equilibrium queue length in single-server queues with superposition arrival processes. In fact, *surprising as it may seem, this experimental design still favors the lower traffic intensities*. The statistical precision, as measured by the *relative standard error* $\bar{r}$, i.e., the ratio of the simulation standard error (sample standard deviation based on 20 batch means) to the simulation estimate of the mean, still decreases with ρ. In fact, the relative standard error was approximately $\bar{r} = 0.01, 0.02$ and 0.04 for $\rho = 0.3, 0.7$ and 0.9; see §4 of Albin (1984b). Our analysis indicates that to achieve a uniform relative standard error $\bar{r}$ over all ρ that the sample size should be approximately proportional to $1/(1-\rho)^2$ arrivals. Given a sample size of 60,000 arrivals for $\rho = 0.3$, our heuristic suggests sample sizes of 120,000, 300,000, 750,000 and 3,000,000 arrivals for $\rho = 0.5, 0.7, 0.8$ and 0.9, respectively. It also predicts relative standard errors approximately as observed with the chosen sample sizes.

We remark that the relative standard error $\bar{r}$ appears to be a good practical measure of statistical precision for evaluating approximations, except possibly when very small numbers are involved; then the absolute standard error might be preferred. It might be argued that relative standard error should *always* be the preferred measure of statistical precision, because it is independent of the measuring units, which can be chosen arbitrarily. However, there often is a measuring unit that is naturally meaningful in an application, so that independence of measuring unit is not always desirable. For example, in queueing processes it is often natural to count in units of customers as opposed to units of 10^6 customers or 10^{-6} customers. Thus, if a mean queue length is 0.01 customers, then we might prefer to measure precision of an estimate by the standard error instead of the relative standard error. In summary, we believe that relative standard error is usually a better measure of statistical precision, but not always, so that we study both the relative standard error and the absolute standard error here. Studying both measures of statistical precision also brings out important differences between them. For example, with the

absolute standard error, the sample size should be approximately proportional to $1/(1-\rho)^4$ arrivals. The impact of the traffic intensity on the required simulation run length in a few simple models such as $M/M/1$ is of course available from previous results, even if it is not as well known as it should be. *Our analysis indicates that the impact of the traffic intensity on the statistical precision of estimates is approximately the same as for the $M/M/1$ model for a large class of models.*

For more complex queueing models than $M/M/1$, the statistical precision of estimates for a given simulation run length is not only affected by the traffic intensity, but also by the variability of the basic arrival and service processes. In Albin's experiments discussed above, different cases had different interarrival-time and service-time distributions. However, in Albin's experiments the sample sizes were not adjusted for the interarrival-time and service-time distribution. This was not unreasonable, since great differences in variability were not considered. Nevertheless, Albin's experimental results indicate that the statistical precision of the simulation estimates decreases as the variability increases (see §6). Consistent with these empirical findings, *our analysis indicates that to achieve a uniform relative standard error $\bar{r}$ that the simulation run lengths should be approximately proportional to the variability, as measured by an appropriate asymptotic variability parameter, which emerges from the heavy-traffic limit.* It is significant that our approximations enable us to *quantify the variability* as it affects statistical precision of estimators. In particular, the asymptotic variability parameter is $1/c = b/|a|$, where a is the drift and b is the diffusion coefficient of the approximating RBM; see (25), (34) and §5. It is of course significant that appropriate RBM approximations, which just amounts to the values of a and b, have previously been determined for many queueing models, and can be determined for other queueing models. For the experiment in Albin (1984a, b), the asymptotic variability parameter is just the sum of the previously studied asymptotic variability parameters for the interarrival times and service times, say $c_A^2 + c_S^2$; in particular, c_A^2 and c_S^2 are computed by the asymptotic method in Whitt (1982b), and are readily available before performing the simulation. Since the service times are i.i.d., c_S^2 is just the squared coefficient of variation (variance divided by the square of the mean) of the service time distribution. Our analysis also indicates that if the criterion is absolute standard error instead, then the simulation run length should be approximately proportional to the *cube* of this same asymptotic variability parameter. (For the $GI/G/1$ model, the approximate effects of ρ and $(c_A^2 + c_S^2)$ were noted by Blomqvist 1969; see (22) there.)

As we should expect, comparisons with simulations results indicate that our heuristic formulas for the variability effect are quite accurate for high traffic intensities, e.g., $\rho \geq 0.8$, but *much less accurate* for lower traffic intensities (when it is less critical anyway); see §6. In fact, Blomqvist emphasizes the limitations of the approximations because they are not very accurate. While there is a danger of expecting too much from the RBM approximations, we interpret the results more positively than Blomqvist, because we propose using the heavy-traffic estimates not as final formulas, but as guides to help determine the run lengths. For lower traffic intensities, the formulas correctly predict the qualitative behavior, provide quick rough approximations, and may serve as a basis for future refined approximations, e.g., by interpolating heavy-traffic and light-traffic limits as in Reiman and Simon (1988).

The rest of this paper is organized as follows. In §2 we review the standard statistical analysis based on stationarity and asymptotic normality that yields confidence intervals. In §3 we analyze the $M/G/1$ queue in detail, drawing on descriptions of the covariance functions in Blomqvist (1967), Daley and Jacobs (1969), Law (1975), Abate and Whitt (1988a) and earlier papers. For the $M/G/1$ model, we do not need diffusion approximations, because we can do the analysis exactly. The $M/G/1$ analysis is useful to evaluate

the quality of the diffusion approximations and to develop refinements of the diffusion approximations.

In §4 we apply diffusion approximations to treat more general models. A few examples are given in §5. Finally, in §6 we make numerical comparisons using simulation results for single-server queues with superposition arrival processes, drawing on Albin (1981) and Fendick et al. (1989).

2. The Standard Statistical Analysis

In this section we briefly review the standard time-series framework, which does not depend on the queueing structure; see Chapter 5 of Fishman (1978). For additional background on stationary processes, see Chapter 3 of Parzen (1962), Chapter 9 of Karlin and Taylor (1975) and Chapter 1 of Borovkov (1976).

2.1. *A Continuous-Time Process*

Let $\{Q(t) : t \geq 0\}$ be the stochastic process of interest and at the outset assume that it is strictly stationary with $E[Q(t)^2] < \infty$. Our object is to estimate the mean $E[Q(0)]$ by the time average

$$\bar{Q}_t = t^{-1} \int_0^t Q(s)ds, \qquad t \geq 0. \tag{1}$$

We think of $Q(t)$ as the queue length at time t, in which case $\bar{Q}_t$ in (1) is a natural estimator of the mean, but $Q(t)$ could represent something else. For example, $Q(t)$ could be some function of the queue length at time t such as the square or the indicator of the set $[0, x]$; then $\bar{Q}_t$ would be a natural estimator of the second moment and cdf, respectively. After applying our RBM approximation, the queueing process is replaced by RBM, so it is also useful to think of $\{Q(t) : t \geq 0\}$ as RBM.

The standard statistical analysis, assuming ample data, is based on a *central limit assumption* for $\bar{Q}_t$ as $t \to \infty$. Let $\Rightarrow$ denote convergence in distribution (weak convergence, as in Billingsley 1968) and, for any constants m and $\sigma^2 > 0$, let $N(m, \sigma^2)$ denote a random variable normally distributed with mean m and variance σ^2. The desired limit is

$$t^{1/2}(\bar{Q}_t - E[Q(0)]) \Rightarrow N(0, \sigma^2) \qquad \text{as} \qquad t \to \infty, \qquad \text{where} \tag{2}$$

$$\sigma^2 = \lim_{t\to\infty} t[\text{Var}\,(\bar{Q}_t)] = \lim_{t\to\infty} \int_{-t}^{t} \left(1 - \frac{|s|}{t}\right) C(s)ds = 2\int_0^\infty C(t)dt < \infty \tag{3}$$

and $C(t)$ is the (auto) covariance function, defined by

$$C(t) = E[Q(0)Q(t)] - (E[Q(0)])^2. \tag{4}$$

We call σ^2 in (3) the *asymptotic variance* of the sample mean $\bar{Q}_t$. As is often done, *we assume that the limit* (2) *is valid*. A general supporting result is Theorem 20.1 of Billingsley. Some sort of mixing or asymptotic independence of $\{Q(s) : s \leq t\}$ and $\{Q(s) : s \geq t + u\}$ as $u \to \infty$ is needed. Such mixing is often provided by regenerative structure, but regenerative structure is not required. Specific results of this form for queueing processes based on regenerative structure appear in Iglehart (1971), Whitt (1972) and Glynn and Whitt (1987).

Based on (2), we use the approximation

$$\bar{Q}_t \approx N(E[Q(0)], \sigma^2/t) \tag{5}$$

for the (large) t of interest, where σ^2 is given by (3). Note that (5) involves *three* approximations: the distribution being normal, the mean being $E[Q(0)]$ (no bias, which

is a consequence of the stationarity assumed so far), and the variance being σ^2/t for σ^2 in (3). For the general planning purposes here, it seems reasonable to assume (5), and we do. We usually justify (5) by appealing to (2), but we only rely on (5).

However, there does remain a very important question: how large must the length t of the simulation run be in order for (5) to be a reasonable approximation? In our queueing context, experience indicates that for (5) to be a good approximation, t must increase as ρ increases. In fact, an approximate answer to this question also emerges from the RBM approximation; we discuss this issue in §4.5. From our analysis, it appears reasonable to assume (5) for the rough approximations we have in mind, but the validity of (5) does impose a constraint. The requirement on t for (5) to be reasonable turns out to be of the same order as our requirement on t for the statistical precision based on (5) to be adequate.

Based on (5), a $(1-\beta)(100)\%$ confidence interval for $E[Q(0)]$ is

$$[\bar{Q}_t - z_{\beta/2}(\sigma^2/t)^{1/2},\ \bar{Q}_t + z_{\beta/2}(\sigma^2/t)^{1/2}] \qquad \text{where} \tag{6}$$

$$P(-z_{\beta/2} \le N(0,1) \le z_{\beta/2}) = 1-\beta. \tag{7}$$

The width of the confidence interval in (6) provides a natural measure of the *statistical precision*. There are two natural criteria to consider: *absolute width* and *relative width*. Relative width looks at the ratio of the width to the quantity to be estimated, $E[Q(0)]$. (These criteria correspond to the absolute standard error and the relative standard error discussed in §1. The absolute standard error is an estimate of $\sigma/t^{1/2}$; the relative standard error is an estimate of $\sigma/t^{1/2}E[Q(0)]$.) For any given β, the absolute width and relative width are

$$w_a(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2}} \qquad \text{and} \qquad w_r(\beta) = \frac{2\sigma z_{\beta/2}}{t^{1/2}E[Q(0)]}. \tag{8}$$

As indicated in §1, out of context it is difficult to conclude which criterion is more appropriate, but we contend that relative width usually is more appropriate.

For specified *absolute width* ϵ and specified *level of precision* β, the required simulation run length, given (5), is

$$t_a(\epsilon, \beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2}. \tag{9}$$

For specified *relative width* ϵ and specified *level of precision* β, the required length of the estimation interval, given (5), is

$$t_r(\epsilon, \beta) = \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2(E[Q(0)])^2}. \tag{10}$$

From (9) and (10) we draw the important and well-known conclusion that both $t_a(\epsilon, \beta)$ and $t_r(\epsilon, \beta)$ are inversely proportional to ϵ^2 and directly proportional to σ^2 and $z_{\beta/2}^2$.

The standard statistical theory describes how observations can be used to estimate the unknown quantities $E[Q(0)]$ and σ^2. Instead, we apply additional information about the model to obtain rough preliminary estimates for $E[Q(0)]$ and σ^2 without data. In particular, we assume that the underlying model is a queueing model that behaves roughly like the standard $GI/G/1$ queue, and that the stochastic process of interest is like the $GI/G/1$ queue-length process, which behaves roughly like RBM, provided that the traffic intensity is not too low. We address our statistical problem by obtaining rough estimates of σ^2 and $E[Q(0)]$ in terms of basic queueing parameters such as the traffic intensity ρ.

We have assumed that the process $Q(t)$ being observed is stationary. In practice, however, we usually cannot start in equilibrium and instead have only an asymptotically stationary process, see Chapter 1 of Borovkov (1976). Nevertheless, the limit (2) is usually valid and the approximation (5) is usually appropriate when t is large enough,

provided that we replace $E[Q(0)]$ in (2) and (5) by the long-run average, say $\bar{Q}_\infty$ (the limit of $\bar{Q}_t$ in (1) as $t \to \infty$, which we assume exists with probability one). Moreover, the bias due to the initial conditions is usually negligible when t is large enough, so that the asymptotic variance in (3) coincides with the corresponding asymptotic mean squared error; see §4.5, Blomqvist (1968) and §9 of Glynn and Whitt (1989). Hence, we propose the approximations (5)–(10) for the large class of steady-state simulations with nonstationary initial conditions.

2.2. *A Discrete-Time Process*

So far in this section we have focused on the continuous-time process $\{Q(t) : t \geq 0\}$. We could instead have a discrete-time process $\{W_n : n \geq 0\}$ (which we think of as the waiting time of the nth customer), which again we assume is stationary with $E[W_n^2] < \infty$. We then estimate the mean $E[W_0]$ by the sample mean $\bar{W}_n = n^{-1} \sum_{k=0}^{n-1} W_k$ and apply the *assumed* limit

$$n^{1/2}(\bar{W}_n - E[W_0]) \Rightarrow N(0, \sigma_W^2), \qquad \text{where} \tag{11}$$

$$\sigma_W^2 = \lim_{n\to\infty} n \operatorname{Var}(\bar{W}_n) = \operatorname{Var}(W_0) + 2 \sum_{k=1}^{\infty} C(k) \tag{12}$$

and $C(k)$ is the (auto) covariance function, defined by $C(k) \equiv C_W(k) = E[W_0 W_k] - (E[W_0])^2$, $k \geq 0$, from which we obtain analogs of (5)–(10).

We close this section by remarking that for a large class of queueing systems, the central limit theorems for continuous-time queue-length processes as in (2) and associated discrete-time waiting-time processes as in (11) are intimately related via an extension of Little's Law $L = \lambda W$; see Glynn and Whitt (1986), (1987), (1988), (1989). An important conclusion from the $L = \lambda W$ central limit theory is that, with the appropriate estimators, approximately the same amount of data are needed for estimating the means of the two processes.

3. The $M/G/1$ Special Cases

We begin our queueing analysis by considering some elementary special cases. As with §2, we review and interpret established results; see Morse (1955), Beneš (1957), Blomqvist (1967), (1968), (1969), Daley and Jacobs (1969), Pakes (1971), Law (1975), Reynolds (1975), Abate and Whitt (1988a) and references cited there. We consider these $M/G/1$ special cases so as to have some concrete exact results to compare with the approximations and to provide a basis for creating refined approximations; see §4.4.

3.1. *The Queue Length, Not Counting the Customer in Service*

Suppose that $Q(t)$ is the queue length not counting the customer in service, if any, in a stationary $M/G/1$ queue with service rate 1 and arrival rate $\rho < 1$. Let m_k be the kth moment of the service time and let $c_S^2 = (m_2 - m_1^2)/m_1^2 = m_2 - 1$ be the squared coefficient of variation. Of course, the steady-state mean is

$$E[Q(0)] = \rho^2(c_S^2 + 1)/2(1 - \rho).$$

From (2.2) of Law (1975), we see that we can express the asymptotic variance σ_Q^2 defined in (3) in terms of the first four moments of the service-time distribution. In particular, we can express $(1 - \rho)\sigma_Q^2$ *exactly* as

$$(1 - \rho)\sigma_Q^2 = \frac{\rho^4}{2}\left(\frac{c_S^2 + 1}{1 - \rho}\right)^3 + \left(\frac{5\rho^3}{2}\left(\frac{m_3}{3m_2}\right) + \rho^2\right)\left(\frac{c_S^2 + 1}{1 - \rho}\right)^2 + \left(\rho(1 + \rho)\left(\frac{m_3}{3m_2}\right) + 3\rho^2\left(\frac{m_4}{12m_2}\right)\right)\left(\frac{c_S^2 + 1}{1 - \rho}\right). \tag{13}$$

Since $m_2 = c_S^2 + 1$, it is easy to see that each term in (13) is nondecreasing in ρ, m_2, m_3, and m_4, so that the qualitative form of the effects discussed in §1 is demonstrated for the $M/G/1$ model. For the special cases of $M/D/1$ and $M/M/1$, $m_{k+1}/m_k = 1$ and $k + 1$, respectively, so that

$$\sigma_Q^2 = \frac{2\rho^2(1 + 4\rho - 4\rho^2 + \rho^3)}{(1 - \rho)^4} \approx \frac{4\rho^2}{(1 - \rho)^4} \quad \text{for} \quad M/M/1 \quad \text{and} \quad (14)$$

$$\sigma_Q^2 = \frac{\rho(4 + 11\rho - 12\rho^2 + 3\rho^3)}{12(1 - \rho)^4} \approx \frac{\rho}{2(1 - \rho)^4} \quad \text{for} \quad M/D/1, \quad (15)$$

with the approximations in (14) and (15) being reasonably good if ρ is not too small, e.g., $\rho \geq 0.3$. Note that σ_Q^2 increases very rapidly as $\rho \to 1$. For comparison, note that Var $[Q(0)] = \rho^2(1 + \rho - \rho^2)/(1 - \rho)^2 \approx \rho^2/(1 - \rho)^2$ for the $M/M/1$ model (with the assumed stationarity), so that σ_Q^2 is greater than (Var $[Q(0)]$)^2. It is of course a serious error to confuse the asymptotic variance σ_Q^2 as defined in (3) with the *steady-state variance* Var $[Q(0)]$; the fact that $\sigma_Q^2 \gg$ Var $[Q(0)]$ underscores the importance of the correlations. It also shows that we would certainly prefer to use independent replications if we could observe stationary processes.

We now discuss approximations for (13). The heavy-traffic limit for $M/G/1$ is obtained by considering only the first term in (13); for high ρ, $\sigma_Q^2 \approx (c_S^2 + 1)^3/2(1 - \rho)^4$. In fact, $\rho^4(c_S^2 + 1)^3/2(1 - \rho)^4$ is obviously a lower bound for σ_Q^2. For $M/G/1$, the asymptotic variability parameter $1/c = b/|a|$ in (34) is $(c_S^2 + 1)$; see §5.1.

Motivated by this heavy traffic limit, can we obtain an approximation for the case $0 \leq c_S^2 \leq 1$ by taking a convex combination of the additional factors in (14) and (15). In particular, for $M/M/1$ rewrite the last term in (14) as $\rho^2(1 + c_S^2)^3/2(1 - \rho)^4$ where $c_S^2 = 1$; for $M/D/1$ rewrite the last term in (15) as $\rho(1 + c_S^2)^3/2(1 - \rho)^4$ where $c_S^2 = 0$. Then approximate σ_Q^2 for the general $M/G/1$ system with parameter c_S^2, $0 \leq c_S^2 \leq 1$, by

$$\sigma_Q^2 \approx c_S^2(M/M/1) + (1 - c_S^2)(M/D/1) = \frac{\rho[1 - (1 - \rho)c_S^2](1 + c_S^2)^3}{2(1 - \rho)^4}. \quad (16)$$

Comparison with the exact value in (13) shows that (16) is usually a very good approximation for traffic intensities of principal interest, e.g., $\rho \geq 0.5$. By "usually" we mean for typical distributions such as Erlang or shifted-exponential. However, since it is theoretically possible to have infinite third or fourth moment with any second moment, the exact value σ_Q^2 in (13) could in fact be infinite, so that (16) could be a very bad approximation. In particular, *we have just demonstrated for the relatively simple $M/G/1$ case that there is no bound on the possible error from using* (16) *or the heavy-traffic RBM approximations*. It is important to realize that the RBM approximations have this limitation. However, we should not let rare pathological cases deter us from obtaining practical approximations.

As in (50)–(52) of Whitt (1983a), we can obtain detailed approximations for σ_Q^2 in $M/G/1$ in terms of ρ and c_S^2 if we approximate m_3 and m_4 in terms of c_S^2. To complement (16), consider the case $c_S^2 \geq 1$. Using the approximations $m_3/m_1^3 \approx 3c_S^2(c_S^2 + 1)$ and $m_4/m_1^4 \approx 4m_2m_3/m_1^5 \approx 12c_S^2(c_S^2 + 1)^2$ in (13) when $c_S^2 \geq 1$, we get

$$(1 - \rho)\sigma_Q^2 \approx \frac{\rho^4}{2}\left(\frac{c_S^2 + 1}{1 - \rho}\right)^3 + \rho^2\left(\frac{5\rho c_S^2}{2} + 1\right)\left(\frac{c_S^2 + 1}{1 - \rho}\right)^2 + (\rho c_S^2)(1 + \rho + 3\rho(c_S^2 + 1))\left(\frac{c_S^2 + 1}{1 - \rho}\right) \quad \text{for } M/G/1 \text{ with } c_S^2 \geq 1. \quad (17)$$

If instead we use the general bounds $m_3 \geq m_2^2/m_1 = (c_S^2 + 1)^2$ and $m_4 \geq (c_S^2 + 1)^3$ in (13), see §2 of Whitt (1984b), then the analog of (17) becomes a lower bound for (13). Note that for large c_S^2 each term of (17) is approximately proportional to $(c_S^2 + 1)^3$ provided ρ is not too small. In general, approximation (17) is consistent with $\sigma_Q^2 \approx K(c_S^2 + 1)^x$ for $2 \leq x \leq 3$.

Overall, we propose the following *simple rough approximation* for general $M/G/1$ systems

$$\sigma_Q^2 \approx \rho^2(c_S^2 + 1)^3/2(1 - \rho)^4. \tag{18}$$

From (14), we see that (18) is good for $M/M/1$, but from (15), we see that (18) is too small by a factor of ρ for $M/D/1$. However, the simplicity makes (18) a useful practical substitute for (13), (16) or (17).

Combining (9) and (10) each with (18), we determine that the estimated lengths of the simulation run for the criteria of absolute width and relative width are approximately

$$t_a(\epsilon, \beta) = \frac{2\rho^2(c_S^2 + 1)^3 z_{\beta/2}^2}{(1 - \rho)^4 \epsilon^2} \quad \text{and} \quad t_r(\epsilon, \beta) = \frac{8(c_S^2 + 1) z_{\beta/2}^2}{\rho^2(1 - \rho)^2 \epsilon^2}. \tag{19}$$

Note that the two approximate estimated run lengths have markedly different behavior in light and heavy traffic:

$$t_a(\epsilon, \beta) \approx \frac{2(c_S^2 + 1)^3 z_{\beta/2}^2}{(1 - \rho)^4 \epsilon^2} \quad \text{and} \quad t_r(\epsilon, \beta) \approx \frac{8(c_S^2 + 1) z_{\beta/2}^2}{(1 - \rho)^2 \epsilon^2} \quad \text{as} \quad \rho \to 1, \tag{20}$$

$$t_a(\epsilon, \beta) \approx \frac{2\rho^2(c_S^2 + 1)^3 z_{\beta/2}^2}{\epsilon^2} \quad \text{and} \quad t_r(\epsilon, \beta) \approx \frac{8(c_S^2 + 1) z_{\beta/2}^2}{\rho^2 \epsilon^2} \quad \text{as} \quad \rho \to 0. \tag{21}$$

In particular, $t_a \to 0$ and $t_r \to \infty$ as $\rho \to 0$. (From (14) we see that these limits hold for the exact $M/M/1$ values. Although light traffic is not a major focus of this paper, it is important to note that indeed $t_r \to \infty$ as $\rho \to 0$, so that the required *run length is not increasing in ρ throughout*.) More importantly, as remarked in §1, t_a is proportional to $1/(1 - \rho)^4$, while t_r is proportional to $1/(1 - \rho)^2$ as $\rho \to 1$. Also, the effect of the variability parameter $(c_S^2 + 1)$ in (20) and (21) is as described in §1.

The possible values of $t_a(\epsilon, \beta)$ and $t_r(\epsilon, \beta)$ for different ρ are displayed for the special case of an $M/M/1$ model with $\epsilon = 0.05$ and $z_{\beta/2} = 2.0$ in Table 1. The value $z_{\beta/2} = 2.0$ approximately produces 95 percent confidence intervals (the exact value would be 1.96). The lengths based on the criterion of relative width $\epsilon = 0.05$ are 27,500 for $\rho = 0.8$, 135,000 for $\rho = 0.9$ and 592,000 for $\rho = 0.95$. To have a relative width $\epsilon = 0.005$ instead of 0.05, these lengths would have to be multiplied by 100: Then the lengths become 2,750,000 for $\rho = 0.8$, 13,500,000 for $\rho = 0.9$ and 59,200,000 for $\rho = 0.95$. Such longer intervals might be needed to evaluate approximations precisely.

3.2. *Other $M/M/1$ Processes*

We now consider other processes in the special case of an $M/M/1$ model. (Analogs for $M/G/1$ are also available, but we only need the $M/M/1$ formulas for our proposed refinements in §4.4.) If we focus on the *number in system*, including the customer in service, say $N(t)$, then $E[N(0)] = \rho/(1 - \rho)$ and the asymptotic variance of $\bar{N}_t = t^{-1} \int_0^t N(s)ds$ is

$$\sigma_N^2 = 2\rho(1 + \rho)/(1 - \rho)^4; \tag{22}$$

see Morse (1955), Law (1975) or Corollary 8 to Theorem 1 of Abate and Whitt (1988a). Then $t_a(\epsilon, \beta)$ and $t_r(\epsilon, \beta)$ behave just as in (20) as $\rho \to 1$.

Alternatively, if we focus on the continuous-time *workload* (or virtual waiting time)

TABLE 1

The Required Length of the Simulation Run for the Stationary M/M/1 Queue-length Process: The Cases of Absolute Width $w_a(\beta)$ and Relative Width $w_r(\beta)$ Being 0.05 and the Normal Distribution Percentile Being $z_{\beta/2} = 2.0$. (The Width of a 95 Percent Confidence Interval for a N(0, 1) Random Variable is Approximately 4.0, so that $\beta \approx 0.05$.)

Traffic Intensity ρ	$4\rho^2$	Normalized Asymptotic Variance $(1-\rho)^4\sigma^2$	Required Length of Simulation Run			
			Absolute Width		Relative Width	
			$1/(1-\rho)^4$	$t_a(0.05, \beta)$	$1/(1-\rho)^2$	$t_r(0.05, \beta)$
0.2	0.16	0.138	2.4	133	1.6	88
0.3	0.36	0.466	7.2	784	2.0	373
0.5	1.00	1.06	16	6784	4	1696
0.7	1.96	1.85	123	9.1×10^4	11.1	8214
0.8	2.56	2.75	625	6.88×10^5	25	2.75×10^4
0.9	3.24	3.38	10,000	1.35×10^7	100	1.35×10^5
0.95	3.61	3.70	160,000	2.37×10^8	400	5.92×10^5
0.99	3.92	3.94	10^8	1.58×10^{11}	10,000	1.58×10^7

process, say $\{L(t) : t \geq 0\}$, then $E[L(0)] = \rho/(1-\rho)$ and the asymptotic variance of $\bar{L}_t = t^{-1}\int_0^t L(s)ds$ is

$$\sigma_L^2 = 2\rho(3-\rho)/(1-\rho)^4; \tag{23}$$

see Beneš (1957), Law (1975), Ott (1977a, b) or Remark 6.3 of Abate and Whitt (1988a).

If we focus on *the discrete-time waiting time before beginning service*, W_n, as in (11)–(12), then $E[W_0] = \rho/(1-\rho)$ and the asymptotic variance of $\bar{W}_n = n^{-1}\sum_{k=0}^{n-1} W_k$ is

$$\sigma_W^2 = \rho[2 + 5\rho - 4\rho^2 + \rho^3]/(1-\rho)^4 \approx 4\rho/(1-\rho)^4; \tag{24}$$

see Law (1975), Example 1(a) of Glynn and Whitt (1988a) or (8.3)–(8.5) of Daley and Jacobs (1969). Note that the asymptotic variability parameters for all four processes ((14), (22), (23), and (24)) are asymptotically of the form $\sigma^2 \sim 4/(1-\rho)^4$ as $\rho \to 1$; i.e., in each case $(1-\rho)^4\sigma^2 \to 4$ as $\rho \to 1$. The $M/G/1$ heavy-traffic behavior is the same for all four processes too.

If we want to compare the discrete-time and continuous-time estimators, we should relate the amounts of data collected, as indicated in §5 of Glynn and Whitt (1986) and §6 of Glynn and Whitt (1988a). Continuous time t corresponds to an expected number ρt of arrivals in our scaling of time, so that if we collect data over $[0, t]$, then we observe a random number of arrivals with mean ρt. The natural estimator for $E[W_0]$ based on data over $[0, t]$ is $\bar{W}_t = D(t)^{-1}\sum_{k=0}^{D(t)-1} W_k$, where $D(t)$ is the number of departures in $[0, t]$. If, instead, we collect data for n arrivals, then we should multiply the asymptotic variances for the continuous-time processes in (14), (22) and (23) by ρ. The asymptotic variances for several processes are displayed in Table 2 for the case in which we use data over $[0, t]$. (As noted in Law 1975 and Glynn and Whitt 1988a, the advantage of indirect estimation via $L = \lambda W$ is asymptotically negligible as $\rho \to 1$.)

4. Heavy-Traffic Limits

Our approximations for general queueing models follow from heavy-traffic limits, but we do not prove any new theorems; we apply previous ones. These approximations are important, because for more general models we are unable to do the analysis corresponding to §3. *We make the existence of a heavy-traffic limit our basic assumption*. Thus, for the approximations to apply, the model should be stable for $\rho < 1$ and unstable for $\rho \geq 1$, with the congestion increasing sharply as ρ approaches 1. Our approximations thus apply

TABLE 2

Asymptotic Variances of Several Estimators for the M/M/1 Model Based on Data Over the Interval $[0, t]$.

Steady-State Mean Being Estimated	Estimator	Asymptotic Variance σ^2
Expected Number Waiting $E[Q(0)]$	$\bar{Q}_t$	$2\rho^2(1 + 4\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4$
Expected Number in System $E[N(0)]$	$\bar{N}_t$	$2\rho(1 + \rho)/(1 - \rho)^4$
Expected Workload $E[L(0)]$	$\bar{L}_t$	$2\rho(3 - \rho)/(1 - \rho)^4$
Expected Waiting Time $E[W_0]$	$\bar{W}_t$	$(2 + 5\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4$
Expected Number Waiting $E[Q(0)]$	$\rho\bar{W}_t$	$\rho^2(2 + 5\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4$
	$\bar{N}_t - \rho$	$2\rho(1 + \rho)/(1 - \rho)^4$
	$\bar{L}_t - \rho$	$2\rho(3 - \rho)/(1 - \rho)^4$

to general multi-server systems with unlimited waiting room and not too many servers, but not to corresponding systems with finite waiting rooms, where arrivals are lost when the system is full, or to infinite-server systems. The extensive heavy-traffic literature provides numerous examples of systems for which the approximations do apply.

In §4.1 we describe the approximating or limiting diffusion process, RBM. In §4.2 we discuss the important scaling of time and space and its impact on the asymptotic variance, defined in (3). In §4.3 we present the basic heavy-traffic limit assumption (30)–(33) and the basic diffusion approximation (34). In §4.4 we propose a refinement of the RBM approximation based on the $M/G/1$ results in §3; we suggest replacing (34) by (35). This leads to the final approximations for the required simulation run lengths given in (36). In §4.5, inspired by Asmussen (1987c), we also use the RBM approximation (34) to investigate how long the simulation run needs to be in order for the asymptotic statistical analysis in §2 to be a reasonable approximation; roughly speaking, the requirement on simulation run length is the same as needed to obtain desired statistical precision (relative standard error) assuming the asymptotic statistical analysis: the run length should be of order $1/(1 - \rho)^2$. In §4.6 we also indicate how to apply the RBM approximation to estimate and reduce bias associated with nonstationary initial conditions. In §4.7 we discuss variations of the basic approach for other models, such as for infinite-server queues.

4.1. *Regulated Brownian Motion (RBM)*

A prominent role is played by one-dimensional regulated or reflecting Brownian motion (RBM), which is Brownian motion on the positive real line with constant negative drift, constant positive diffusion coefficient and an impenetrable reflecting barrier at the origin; see Harrison (1985) and Abate and Whitt (1987). RBM is a Markov process with continuous sample paths, i.e., a diffusion process. RBM is important because it is the common limit process. All queueing processes that converge to RBM under a specified normalization as $\rho \to 1$ are in the domain of our approximation formulas.

Let $R(t; a, b, X)$ represent RBM with drift $a\,(<0)$, diffusion coefficient $b\,(>0)$ starting at the random state X ($R(0) = X \geq 0$). It is significant for what follows that we can rescale RBM so that it suffices to consider the "dimensionless" case with drift $a = -1$ and diffusion coefficient $b = 1$, which we call *canonical RBM*. In particular,

$$\{R(t; a, b, X) : t \geq 0\} \stackrel{d}{=} \{c^{-1}R(d^{-1}t; -1, 1, cX) : t \geq 0\},$$

$$\{cR(td; a, b, X) : t \geq 0\} \stackrel{d}{=} \{R(t; -1, 1, cX) : t \geq 0\},$$

$$c = |a|/b, \qquad d = b/a^2, \qquad a = -1/cd \qquad \text{and} \qquad b = 1/c^2d, \tag{25}$$

where $\stackrel{d}{=}$ means equality in distribution; see §2 of Abate and Whitt (1987) or Newell (1982).

The rescaling (25) allows us to make all limit processes identical. In particular, suppose that $X_n(t)$ converges in distribution to $R(t; a, b, X)$ as $t \to \infty$. By an elementary application of (25) and the continuous mapping theorem, Theorem 5.1 of Billingsley, $cX_n(td)$ converges in distribution to $R(t; -1, 1, cX)$ for c and d defined by (25). The reduction to *canonical RBM* $R(t; -1, 1, X)$ is important because all descriptions of RBM can be done for a single process; we need not perform separate calculations for different parameter values.

The stationary distribution of $R(t; a, b, X)$ is exponential with mean $b/2|a|$. If the initial position X is endowed with this exponential distribution, then RBM becomes a stationary process. It is easy to see that the rescaling (25) preserves stationarity. For canonical RBM, the stationary distribution is exponential with mean $\frac{1}{2}$.

Expressions for the covariance function of RBM were derived by Ott (1977a, b), Woodside et al. (1980) and Abate and Whitt (1988a, b). One nice expression is a simple spectral representation (mixture of exponentials) as given in Theorem 3 of Ott (1977b) and Theorem 2 of Woodside et al., namely,

$$C_R(t) = \frac{2}{\pi}\int_0^1 e^{-(t/2x)}\sqrt{x(1-x)}\,dx; \tag{26}$$

also see Abate and Whitt (1988b). By Corollary 1 to Theorem 1 of Abate and Whitt (1988a), the covariance function of canonical RBM also can be expressed directly (without integration) as

$$C_R(t) = 2^{-1}(1 - 2t - t^2)[1 - \Phi(t^{1/2})] + 2^{-1}t^{1/2}(1 + t)\phi(t^{1/2})$$

where $\phi(t)$ is the density and $\Phi(t)$ is the cdf of $N(0, 1)$. From these expressions, we easily deduce that the asymptotic variance of canonical RBM is

$$\sigma_R^2 = \lim_{t\to\infty} t\,\mathrm{Var}\left[t^{-1}\int_0^t R(s; -1, 1)ds\right] = 2\int_0^\infty C_R(t)dt = \tfrac{1}{2}. \tag{27}$$

In fact, for our applications based on (5), we do not need the full covariance function; we only need the asymptotic variance (27). Consequently, we can actually get what we need much more directly by simply applying Morse's (1955) result for the $M/M/1$ queue in (22); also see Beneš (1957). From the heavy-traffic theory to be discussed below, we can deduce that $(1 - \rho)^4\sigma_N^2(\rho)/8 \to \sigma_R^2 = \frac{1}{2}$ as $\rho \to 1$. The desired RBM result (27) is already embodied in the earlier $M/M/1$ result.

4.2. *Scaling of the Covariance Function*

An essential aspect of the heavy-traffic analysis is a scaling of space and time. Before we consider the heavy-traffic limits, we see how scaling affects the covariance function $C(t)$ in (4) and the asymptotic variance σ^2 in (3). For the general stationary process $Q(t)$ in §2 (which could be RBM), let $Q_{yz}(t)$ be an associated scaled process, defined by $Q_{yz}(t) = yQ(zt)$, $t \geq 0$, where y and z are arbitrary positive scalars. Let $C_{yz}(t)$ and σ_{yz}^2 be the covariance function and asymptotic variance of $Q_{yz}(t)$. It is easy to see that

$$E[Q_{yz}(t)] = yE[Q(t)], \qquad C_{yz}(t) = y^2C(zt) \qquad \text{and} \qquad \sigma_{yz}^2 = y^2\sigma^2/z. \quad (28)$$

(For σ_{yz}^2 we do the change of variables $u = zt$ in the integration.) In particular, note that the key ratio in $t_r(\epsilon, \beta)$ for the criterion of relative width in (10) is

$$\sigma_{yz}^2/(E[Q_{yz}(0)])^2 = \sigma^2/(E[Q(0)])^2 z \quad (29)$$

which is independent of y. The conclusion to be drawn from (29) is that for $t_r(\epsilon, \beta)$ *an essential role is played by the time scaling* z. In our queueing framework, the heavy-traffic limit determines what that time scaling must be.

4.3. *The Assumed Heavy-Traffic Limit and RBM Approximation*

For a general queueing model, we consider a family of stationary queueing processes $\{Q_\rho(t) : t \geq 0\}$ indexed by the traffic intensity ρ; i.e., the ρth system has traffic intensity ρ. (Think of the queue-length process, but it could be the workload process or something else.) Such a family of processes can usually be obtained from one given model by simply scaling time in the arrival process; e.g., $A_\lambda(t) = A(\lambda t)$ has arrival rate λ if $A(t)$ has arrival rate 1. Our key assumption is that appropriately normalized versions of $Q_\rho(t)$ converge to a stationary version of RBM as $\rho \to 1$. Our basic approximation is given by (34) below. A refinement is given in §4.4, which leads to the approximate required simulation run lengths in (36). Applications are illustrated in §5.

Let the normalized processes be defined by

$$\hat{Q}_\rho(t) = (1 - \rho)Q_\rho(t(1 - \rho)^{-2}), \qquad t \geq 0; \quad (30)$$

i.e., we use the *space scaling* $(1 - \rho)$ and the *time scaling* $1/(1 - \rho)^2$. The scaling in (30) is consistent with the majority of the existing heavy-traffic limit theorems; e.g., Iglehart and Whitt (1970).

Let $C_\rho(t)$, and $\hat{C}_\rho(t)$ be the covariance functions of $Q_\rho(t)$ and $\hat{Q}_\rho(t)$, respectively, and let σ_ρ^2 and $\hat{\sigma}_\rho^2$ be the corresponding asymptotic variances. From (28) and (30), we see that

$$\hat{C}_\rho(t) = (1 - \rho)^2 C_\rho(t(1 - \rho)^{-2}) \qquad \text{and} \qquad \hat{\sigma}_\rho^2 = (1 - \rho)^4\sigma_\rho^2. \quad (31)$$

Note that the term $(1 - \rho)^4$ showing the first-order effect of ρ appears in the asymptotic variance relation in (31).

What we assume then is that $\hat{Q}_\rho(t)$ *in* (30) *converges in distribution to stationary RBM.* We allow general drift and diffusion coefficients (a, b) in the limit; let the (necessarily stationary) limit process be denoted by $R(t; a, b)$. To treat the covariance, we want joint convergence at any two time points. In particular, we assume that

$$[\hat{Q}_\rho(t_1), \hat{Q}_\rho(t_2)] \Rightarrow [R(t_1; a, b), R(t_2; a, b)] \qquad \text{in } R^2 \qquad \text{as } \rho \to 1 \quad (32)$$

for all $t_2 > t_1 > 0$. In addition to (32), we also assume that

$$\lim_{\rho \to 1} E[\hat{Q}_\rho(t_1)] = E[R(t; a, b)] = b/2|a|,$$

$$\lim_{\rho \to 1} E[\hat{Q}_\rho(t_1)\hat{Q}_\rho(t_2)] = E[R(t_1; a, b)R(t_2; a, b)],$$

$$\lim_{\rho \to 1} \hat{\sigma}_\rho^2 = \lim_{\rho \to 1} 2\int_0^\infty \hat{C}_\rho(t)dt = 2\int_0^\infty \text{Cov}\,[R(0; a, b), R(t; a, b)]dt, \quad (33)$$

for all $t_2 > t_1 > 0$. We typically obtain (32) from a functional central limit theorem (FCLT) such as Theorem 1 of Iglehart and Whitt (1970); then (32) follows from an elementary application of the continuous mapping theorem (Theorem 5.1 of Billingsley) using the projection map, mapping the function space into R^2. The heavy-traffic literature contains a plethora of results supporting (32). The first two limits in (33) follow from

extra uniform integrability (p. 32 of Billingsley). Overall, (33) is usually a minor technical condition, which should cause little concern; i.e., given (32), we can expect (33) to hold. For practical purposes, it suffices to prove (32) or even just convergence of the one-dimensional distributions and assume the rest. Actually proving (33) can be difficult; Blomqvist (1969) and Asmussen (1987c) carefully address these and related issues for the $GI/G/1$ model. Convergence of $\hat{C}_\rho(t)$ and $\hat{\sigma}_\rho^2$ as $\rho \to 1$ is established for the $M/G/1$ virtual waiting time process by Ott (1977a, b) and for several $M/M/1$ processes by Abate and Whitt (1988a).

Now, using (25), we rescale $\hat{Q}_\rho(t)$ in (30) to obtain canonical RBM as a limit. The process $(|a|/b)\hat{Q}_\rho(bt/a^2)$, $t \geq 0$, converges to canonical RBM. The resulting approximations are

$$\begin{aligned}
\{Q_\rho(t); t \geq 0\} &\approx \{(b/|a|(1-\rho))R(a^2(1-\rho)^2 t/b; -1, 1) : t \geq 0\}, \\
E[Q_\rho(t)] &= [1/(1-\rho)]E[\hat{Q}_\rho(t)] \\
&\approx (b/|a|(1-\rho))E[R(t; -1, 1)] = b/2|a|(1-\rho), \\
\sigma_\rho^2 &\approx [b^3/a^4(1-\rho)^4]\sigma_R^2 = b^3/2a^4(1-\rho)^4 \qquad \text{and} \\
\sigma_\rho^2/E[Q_\rho(0)]^2 &\approx 2b/a^2(1-\rho)^2. \qquad (34)
\end{aligned}$$

We obtain the estimated required simulation run lengths $t_a(\epsilon, \beta)$ and $t_r(\epsilon, \beta)$ by substituting (34) in (9) and (10), respectively. Just as in §2 where the critical assumption was the normal approximation (5) rather than the supporting limit (2), so here the critical assumption is the diffusion approximation (34) rather than the supporting limits (32) and (33). The development of diffusion approximations in Newell (1982) and Woodside et al. (1980) is essentially equivalent to (30)–(34), even though it may look quite different.

4.4. *$M/M/1$ Refinements*

The direct heavy-traffic RBM approximation is (34). Based on the $M/G/1$ results for the expected queue length in §3.1, especially (18), it is natural to suggest modifying (34) by multiplying $E[Q_\rho(t)]$ and σ_ρ^2 by ρ^2. (As discussed in Whitt 1982c, factors of ρ^k are necessarily lost as $\rho \to 1$ in the heavy-traffic limit, so it is natural to look for refinements by considering special cases for which exact results are available.) We thus propose the following refined approximations

$$E[Q_\rho(0)] \approx \frac{b\rho^2}{2|a|(1-\rho)}, \quad \sigma_\rho^2 \approx \frac{b^3\rho^2}{2a^4(1-\rho)^4} \quad \text{and} \quad \frac{\sigma_\rho^2}{(E[Q_\rho(0)])^2} \approx \frac{2b}{a^2\rho^2(1-\rho)^2}. \qquad (35)$$

We can obtain (35) formally by applying the heavy-traffic limit twice, once for the system of interest and once for the $M/M/1$ model. We first approximate the general queue by RBM and then we approximate RBM by $M/M/1$ using (14). We have obtained (35) by considering the queue length excluding customers in service. It would be natural to use (22), (23) and (24) instead of (14) to create the refinement when the process of interest is number in system, workload or waiting time, respectively. (This is done in §5.2.) However, (35) seems like a reasonable approximation in general.

Substituting (35) into (9) and (10), we obtain our final approximations for the required simulation run length with the absolute and relative width criteria

$$t_a(\epsilon, \beta) \approx \frac{2b^3\rho^2 z_{\beta/2}^2}{a^4(1-\rho)^4\epsilon^2} \qquad \text{and} \qquad t_r(\epsilon, \beta) \approx \frac{8bz_{\beta/2}^2}{a^2\rho^2(1-\rho)^2\epsilon^2}. \qquad (36)$$

Given that approximation (35) is reasonable, the primary task in applying (36) is to

identify the two parameters a and b. Fortunately, this can often be done without difficulty; see §5. (The run length requirements in (36) agree with (19); for $M/G/1$ with service rate 1, $a = -1$ and $b = (c_S^2 + 1)$.)

4.5. *The Quality of the Normal Approximation* (5)

As indicated in §2, our entire analysis depends on the normal approximation (5), which in turn depends on the simulation run length t. Not only must t be sufficiently large so that the estimated statistical precision based on (5) is adequate, but t must be sufficiently large so that (5) itself is a reasonable approximation. It is significant that a rough idea about the simulation run length t required for (5) to be a reasonable approximation also follows from the RBM approximation (34). To get a rough idea about the required run length for (5) to be valid, we make the RBM approximation (34) and ask whether an interval of given length is adequate for (5) to be nearly valid when the process in question is canonical RBM.

First, the time scaling in (34) alone implies that $a^2(1 - \rho)^2 t/b$ must be sufficiently large in order for (5) to be good for the queueing process, because this is the length of the simulation run for the approximating canonical RBM. *This means that in order for* (5) *to be a reasonable approximation, the simulation run length t for the queueing process as a function of* ρ *must be of order* $b/a^2(1 - \rho)^2$. Note that this simulation run length requirement for the queueing process is the same order, as a function of a and b as well as ρ, as determined by the relative width measure of statistical precision in (36).

Of course, for any fixed ρ, (5) involves approximation error for any finite t and is asymptotically correct as $t \to \infty$. However, if we regard t as a function of ρ and consider the *double limit* $t(\rho) \to \infty$ and $\rho \to 1$, it does *not* follow that (5) is necessarily asymptotically correct. Given the RBM approximation (34), we see from the time scaling that (5) is asymptotically correct in this double limit if and only if $(1 - \rho)^2 t(\rho) \to \infty$ as $\rho \to 1$. A detailed mathematical study of the double limit $\rho \to 1$ and $(1 - \rho)^2 t(\rho) \to t_0$, $0 \le t_0 \le \infty$, has recently been made by Asmussen (1987c), which rigorously establishes this conclusion. Asmussen shows that serious estimation problems can occur as $\rho \to 1$ when the limit of $(1 - \rho)^2 t(\rho)$ is finite, especially when this limit is 0. The practical consequence of this analysis is that the simulation run length not only should be of order $b/(1 - \rho)^2 a^2$ but should be $t_0 b/(1 - \rho)^2 a^2$ for suitably large t_0.

To better understand deviations from (5) for the queueing process, we can examine canonical RBM more closely. In particular, we can assume that the simulation run length t is chosen to satisfy $t = bt_0/a^2(1 - \rho)^2$ for some fixed t_0 independent of a, b and ρ, and then consider the estimation problem for canonical RBM on $[0, t_0]$. We then ask if a simulation run length of t_0 is long enough so that (5) is nearly valid when the process being simulated is canonical RBM. (Of course, this analysis assumes the validity of the RBM approximation (34).) In particular, if we use $t_r(\epsilon, \beta)$ in (36), then

$$t_0 \equiv t_0(\epsilon, \beta, \rho) = \frac{8z_{\beta/2}^2}{\rho^2 \epsilon^2} \tag{37}$$

is the length of the simulation run length for canonical RBM. (Of course, the ρ^2 in the denominator of (36) and (37) is an artifact of refinement (35); for RBM we can set $\rho = 1$. The ρ^2 in the denominator signals possible difficulties for the queue if the queue is in very light traffic.) We contend that t_0 in (37) usually is long enough. For example, if $z_{\beta/2} = 2.0$ and $\epsilon = 0.05$ as in Table 1, then $t_0 = 12{,}800$ for $\rho = 1$.

In this regard, it is worth observing that, if the basic process $\{Q(t) : t \ge 0\}$ in §2 is canonical RBM with $a < 0 < b$, then the central limit theorem (2) is indeed valid. We can verify (2) by exploiting the regenerative structure. Appropriate regeneration points for RBM are first visits to the state 0 after first visiting any state $x > 0$; e.g., $x = 1$. Such

regenerative cycles are positive with probability one and have moments of all orders, so that (2) is valid.

For canonical RBM, as for any other process $\{Q(t) : t \geq 0\}$ satisfying (2), the issue is whether any given finite t_0 is large enough to justify approximation (5). For canonical RBM, we can compare the asymptotic variance $\sigma_R^2 = \frac{1}{2}$ in (3) and (27) with the exact value (assuming stationarity) of the variance of the sample mean

$$\text{Var}\left[t_0^{-1}\int_0^{t_0} R(s; -1, 1)ds\right] = t_0^{-1}\int_{-t_0}^{t_0}\left(1 - \frac{|s|}{t_0}\right)C_R(s)ds \tag{38}$$

using (3) and one of the explicit expressions for $C_R(t)$ in §4.1. A convenient hyperexponential approximation for the covariance function that can also be used in (38) is $C_R(t) \approx 0.125e^{-2t} + 0.125e^{-2t/3}$, $t \geq 0$; see (5.11) of Abate and Whitt (1987) plus (2.2) and Remark 3.2 of Abate and Whitt (1988a). The advantage of RBM here is that explicit expressions for $C_R(t)$ are available, so that the integration in (38) can easily be done.

4.6. *Bias Considerations*

If we do not work with a stationary process, then we have bias, i.e., a difference between the mean of the estimator and the true mean. For example, it is natural to start the queue empty. Using the RBM approximation (34), we are thus lead to consider the bias for canonical RBM observed over $[0, t_0]$ when canonical RBM starts at the origin. We can estimate the bias in estimating the steady-state mean $E[R(\infty; -1, 1)] = \frac{1}{2}$ by the sample average $\bar{R}_{t_0} = t_0^{-1}\int_0^{t_0} R(s; -1, 1, 0)ds$ starting empty, using the exact expression for $E[R(t; -1, 1, 0)]$ in Theorem 1.1 of Abate and Whitt (1987) or the hyperexponential approximation

$$(1/2) - E[R(t; -1, 1, 0)] \approx 0.36e^{-5.23t} + 0.138e^{-0.764t}, \qquad t \geq 0, \tag{39}$$

in (1.13) there. By Corollary 1.3.4 there,

$$\begin{aligned} E(\bar{R}_{t_0}) - 1/2 &= t_0^{-1}\int_0^{t_0}(E[R(t; -1, 1, 0)] - 1/2)ds \\ &\approx t_0^{-1}\int_0^{\infty} E[R(s; -1, 1, 0) - 1/2)ds = 1/4t_0 \qquad \text{for large } t_0. \end{aligned} \tag{40}$$

As in many circumstances, e.g., §9 of Glynn and Whitt (1988a), the asymptotic bias for RBM is asymptotically inversely proportional to the run length t_0. Since the width of the confidence interval is proportional to $t_0^{-1/2}$, see (6) and (8), the bias for RBM is asymptotically negligible compared to the width of the confidence interval.

The bias approximation in (39) and (40) suggests that when we start with an empty queue we might want to follow the common practice of deleting an initial portion of the data. First, the time scaling in the RBM approximation (34) indicates that the length of the initial portion to be deleted in the queueing process should increase with ρ; in particular, it too should be of order $b/a^2(1-\rho)^2$. For canonical RBM, we can estimate the resulting bias when we delete the initial portion in $[0, t]$ by shifting the integral in (40) from the interval $[0, t_0]$ to the interval $[t_1, t_1 + t_0]$.

Since $E[R(t, -1, 1, 0)]$ is within about 1% of its steady-state mean $\frac{1}{2}$ at $t = 4$ (see Table 1 of Abate and Whitt 1987), a rough rule of thumb for the length t of the initial time interval $[0, t]$ to delete in a queueing simulation when looking for relative statistical precision of order 5–10% is the initial segment corresponding to $4b/a^2(1-\rho)^2$ expected service times. For example, for the $M/M/1$ queue with service rate 1, $a = -1$ and $b = 2$, so that $t = 8/(1-\rho)^2$; for $\rho = 0.5$ this is only 32 expected service times, but for $\rho = 0.9$ this is 800 expected service times. When the initial segment $[0, 4]$ is deleted from canonical

RBM starting at the origin, the approximate asymptotic bias based on (39) and data over the interval $[4, 4 + t]$ becomes $0.0085/t$ instead of $0.25/t$ in (40).

Finally, the RBM approximation also provides insight into other initializations besides an empty system that can be used to make the queueing process approach steady-state more quickly. §13 of Abate and Whitt (1987) suggests starting out with a number of customers between the mean and twice the mean, say $\frac{3}{2}$ times the mean. Of course, since we are trying to estimate the mean, evidently we do not know the mean, but we can estimate it too by the RBM approximation. Based on (35), we would thus initialize by letting

$$Q(0) = (3/2)(b\rho^2/2|a|(1-\rho)) = 3b\rho^2/4|a|(1-\rho). \tag{41}$$

4.7. *Generalizations*

It should be clear that the approaches in §§4.3–4.5 also apply with other scalings than (30) and other limit processes than RBM. We then exploit a modified version of the limits in (32) and (33) to obtain analogs of (34) and (35). To mention one example, for infinite-server queues or queues with large number of servers, there is often convergence after appropriate normalization to the Ornstein-Uhlenbeck (O-U) diffusion process or more general Gaussian processes as the arrival rate increases (Iglehart 1965, Borovkov 1967, Whitt 1982a and Prisgrove 1987). The generalized procedure can be applied there. (This is intended for a subsequent paper.) A third diffusion approximation for multi-server queues in between to RBM and O-U is discussed by Halfin and Whitt (1981).

5. Examples

To apply §4, we need to determine that the RBM approximation (34) is appropriate and identify the drift and diffusion coefficients of RBM, i.e., the parameters a and b. In §5.1 we analyze the standard $GI/G/m$ model, in which the interarrival times and service times come from independent sequences of i.i.d. random variables, using heavy-traffic results in Iglehart and Whitt (1970). In §5.2 we analyze the general $G/G/1$ model, without independence assumptions, based on Whitt (1968), (1980) and Fendick et al. (1989). Then in §5.3 we analyze a general model of a packet queue in Fendick et al., which is a single-server queue with multiple classes and batch arrivals. In §5.4 we briefly discuss another class of models covered by the theory: queues with service interruptions; for this model, recent heavy-traffic approximations by Burman (1987) can be applied. The heavy-traffic literature provides many other examples. In §5.5 we indicate how the analysis can be applied to estimate the required simulation run lengths for open queueing networks. For this analysis, we exploit queueing network approximations in Whitt (1983a).

5.1. *The Standard $GI/G/m$ Model*

Consider the standard $GI/G/m$ model, which has unlimited waiting room, the FCFS discipline, m servers and independent sequences of i.i.d. interarrival times and service times. Let the squared coefficients of variation of the interarrival times and service times be c_A^2 and c_S^2, respectively. As in §3, let the individual mean service time be 1. By Theorem 1(a) and Example 3(1) of Iglehart and Whitt, the condition (32) holds for the queue length process (either counting the customers in service or not) with parameters $a = -m$ and $b = m(c_A^2 + c_S^2)$. The limit involving n there is connected to the family indexed by ρ here by setting $n = (1-\rho)^{-2}$. We will not try to explain here *why* a and b take this simple form; an intuitive understanding may be obtained from Chapters 7 and 8 of Newell or §2 of Whitt (1982c). The quality of the RBM approximation typically improves as ρ increases. For given ρ, the quality of the approximation typically decreases as m

increases; see Halfin and Whitt. From (35), we obtain the following approximations for the stationary $GI/G/m$ queue length, not counting customers in service,

$$E[Q(0)] \approx \frac{\rho^2(c_A^2 + c_S^2)}{2(1-\rho)}, \quad \sigma_\rho^2 \approx \frac{\rho^2(c_A^2 + c_S^2)^3}{2m(1-\rho)^4} \quad \text{and} \quad \frac{\sigma_\rho^2}{(E[Q(0)])^2} \approx \frac{2(c_A^2 + c_S^2)}{m\rho^2(1-\rho)^2}. \tag{42}$$

The m appears in (42) because we have fixed the individual mean service time at 1, so that the arrival rate is $m\rho$; i.e., the rate of activity is proportional to m. The m drops out if we let the individual mean service time be m, so that the arrival rate in the ρth system is ρ. If the mean service time is τ instead of 1, then the second two terms in (42) should be multiplied by τ. The asymptotic variability parameter here is $1/c = b/|a| = (c_A^2 + c_S^2)$. Note that (42) agrees with (18) for the $M/G/1$ special case. The final estimated required simulation run length is obtained by substituting $b = m(c_A^2 + c_S^2)$ and $a = -m$ into (36) or, equivalently, substituting (42) into (9) or (10).

5.2. *General Single-Server Models*

In preparation for considering the packet queue model in §5.3, we now consider a general $G/G/1$ model, without any independence assumptions, but still the first-come, first-served discipline. Let u_n be the interarrival time between the nth and $(n+1)$st arrival and let v_n be the service time of the nth arrival. We assume that $\{(u_n, v_n) : n \geq 1\}$ is stationary. As in §3, we assume that $Ev_n = 1$. Initially, let $Eu_n = 1$ too, but we will soon reduce it to $\rho < 1$. Let $U_n = u_1 + \cdots + u_n$, $V_n = v_1 + \cdots + v_n$ and $S_n = V_n - U_n$, $n \geq 1$. As in §3.2, we focus on the discrete-time waiting time of the nth customer before beginning service, which satisfies

$$W_{n+1} = \max\{W_n + v_n - u_n, 0\} = S_n - \min\{S_k : 0 \leq k \leq n\}, \qquad n \geq 1, \tag{43}$$

where $S_0 = 0$. Let the system start empty, so that $W_1 = 0$.

We assume that the sequence $\{(U_n, V_n) : n \geq 1\}$ satisfies a FCLT, converging to two-dimensional Brownian motion, i.e.,

$$n^{-1/2}(U_{[nt]} - nt, V_{[nt]} - nt) \Rightarrow B(t) \qquad \text{as} \qquad n \to \infty, \tag{44}$$

where $[x]$ is the greatest integer less than or equal to x and $B(t)$ is two-dimensional Brownian motion without drift and 2×2 covariance matrix C with elements $C_{11} = c_A^2$, $C_{12} = C_{21} = c_{AS}^2$ and $C_{22} = c_S^2$. (An example where this limit is established and the parameters are evaluated appears in §5.3.) These parameters c_A^2, c_S^2 and c_{AS}^2 characterize the variability in heavy traffic. The associated one-dimensional central limit theorem (CLT) is

$$n^{-1/2}(U_n - n, V_n - n) \Rightarrow N(0, C) \qquad \text{as} \qquad n \to \infty. \tag{45}$$

Now construct a family of queueing systems indexed by ρ by letting the nth interarrival time in the ρth system be $u_n^\rho = u_n/\rho$ for all n. (The arrival rate and traffic intensity are thus ρ.) Let variables in the ρth system be designated by a superscript ρ. As in §5.1, let n and ρ be related by $n = (1-\rho)^{-2}$. (We can also consider a stationary version of the waiting-time sequence, which exists and is unique; Loynes 1962. However, the heavy-traffic limit does not require working with a stationary version.) It then follows that $n^{-1/2}S_{[nt]}^\rho$ converges to one-dimensional Brownian motion with negative drift and

$$n^{-1/2}W_{[nt]}^\rho \Rightarrow R(t; -1, c_A^2 + c_S^2 - 2c_{AS}^2) \qquad \text{as} \qquad n \to \infty, \tag{46}$$

both as FCLTs; see Whitt (1968), Theorem 6.4 of Whitt (1980), and Theorem 1 of Fendick et al. Since $n = (1-\rho)^2$, $n^{-1/2}W_{[nt]}^\rho = (1-\rho)W_{[t(1-\rho)^{-2}]}^\rho$ and (46) implies (32) where $Q_\rho(t) = W_{[t]}$. The RBM limiting parameters in (46) are $a = -1$ and b

$= c_A^2 + c_S^2 - 2c_{AS}^2$, so that the asymptotic variability parameter here is $1/c = b/|a| = (c_A^2 + c_S^2 - 2c_{AS}^2)$.

Thus, combining (34) and (46), we obtain

$$E[W_0] \approx \frac{(c_A^2 + c_S^2 - 2c_{AS}^2)}{2(1-\rho)}, \qquad \sigma_W^2 \approx \frac{(c_A^2 + c_S^2 - 2c_{AS}^2)^3}{2(1-\rho)^4}$$

$$\text{and} \qquad \frac{\sigma_W^2}{(E[W_0])^2} \approx \frac{2(c_A^2 + c_S^2 - 2c_{AS}^2)}{(1-\rho)^2}. \tag{47}$$

A refinement based on the $M/M/1$ result (24) is

$$E[W_0] \approx \frac{\rho(c_A^2 + c_S^2 - 2c_{AS}^2)}{2(1-\rho)}, \qquad \sigma_W^2 \approx \frac{\rho(c_A^2 + c_S^2 - 2c_{AS}^2)^3}{2(1-\rho)^4}$$

$$\text{and} \qquad \sigma_W^2/(E[W_0])^2 \approx \frac{2(c_A^2 + c_S^2 - 2c_{AS}^2)}{\rho(1-\rho)^2}. \tag{48}$$

We would also use (48) for corresponding m-server systems in which the individual mean service time is m, because the natural heavy-traffic approximations are the same. For the standard $GI/G/m$ model, $c_{AS}^2 = 0$ and (47) agrees with §5.1 and (34). *For more general models, it is significant that the asymptotic variability parameters* c_A^2, c_S^2 *and* c_{AS}^2 *in* (44)–(48) *can indeed often be derived analytically*, as we demonstrate for a general packet queue model in the next section. Moreover, there is always the possibility of estimating these parameters from data, as suggested for other purposes in Fendick et al.

5.3. *A Packet Queue Model*

This example is a packet queue model from §4.3 of Fendick et al. There is a single-server queue with unlimited waiting room and the first-come first-served discipline, so that this model is covered by §5.2. At the same time, the model is somewhat complicated, indicating the power of the method.

There are k customer classes (customers are packets and the classes refer to different kinds of traffic, such as voice, short data messages, long data files, facsimile, acknowledgements and other signals, which may have very different characteristics). For class i, customers arrive in batches, with the successive batch sizes for class i being i.i.d. with mean m_i and squared coefficient of variation c_{bi}^2. (The batches occur because the messages sent are divided into packets.) The service times of individual customers of class i have mean τ_i and squared coefficient of variation c_{si}^2. For each class, there are spaces between the arrivals of the customers in each batch. For class i, the spaces are i.i.d. with mean ξ_i and squared coefficient of variation c_{xi}^2. Following the arrival of all customers in a batch, there is an idle period with mean ω_i and squared coefficient of variation c_{Ii}^2. All the service times, batch sizes, spacings and idle periods are assumed to be mutually independent. Let λp_i be the arrival rate of batches of class i, where $p_1 + \cdots + p_k = 1$. From above, $\lambda p_i = 1/(m_i\xi_i + \omega_i)$. The associated arrival rate of customers from class i is

$$\bar{\lambda} q_i = \lambda p_i m_i = m_i/(m_i\xi_i + \omega_i), \tag{49}$$

where $q_i = p_i m_i / \sum_{i=1}^k p_i m_i$ is the proportion of all arrivals that are of class i and $\bar{\lambda}$ is the total arrival rate of customers. Let $r_i = \tau_i/\tau$ where τ is the average service time for all customers, defined by $\tau = \sum_{i=1}^k p_i m_i \tau_i / \sum_{i=1}^k p_i m_i$. Let β_i be the proportion of busy time in each cycle, defined by $\beta_i = m_i\xi_i/(m_i\xi_i + \omega_i)$. (When there is no spacing, $\xi_i = \beta_i = 0$.)

As we indicated above, this model is a special case of §5.2. By Theorems 2 and 3 plus §4.3 of Fendick et al., the limits (44)–(46) in §5.2 hold for this model with

$$c_A^2 = \sum_{i=1}^{k} q_i c_{Ai}^2, \qquad c_S^2 = \sum_{i=1}^{k} q_i [r_i^2 c_{si}^2 + (r_i - 1)^2 c_{Ai}^2]$$

$$\text{and} \qquad c_{AS}^2 = \sum_{i=1}^{k} q_i (1 - r_i) c_{Ai}^2 \tag{50}$$

where c_{Ai}^2 is the asymptotic-method variability parameter for the ith arrival process (Whitt 1982b), satisfying

$$c_{Ai}^2 = m_i(1 - \beta_i)^2(c_{bi}^2 + c_{Ii}^2) + \beta_i^2 c_{xi}^2. \tag{51}$$

As indicated by Fendick et al., quite large values of these parameters can occur in realistic models of packet queues. Typical values are $c_A^2 = 20$, $c_S^2 = 35$ and $c_{AS}^2 = -7$, yielding $(c_A^2 + c_S^2 - 2c_{AS}^2) = 60$. (A concrete example appears in §6.2.) For these models, simulation runs clearly must be much longer than in the $M/M/1$ model with the same traffic intensity. The formulas here indicate approximately by how much.

5.4. *Queues with Interrupted Service*

Other queueing models that can exhibit large variability are queues in which the server is alternately available and unavailable. Included in this class are various priority queues. Such models are often considered in manufacturing to represent the effects of machine breakdown, scheduled maintenance, lunch breaks and work shifts; see Federgruen and Green (1987a, b) and references cited there. Heavy-traffic limit theorems by Whitt (1971), Harrison (1973) and Burman (1987) can be applied with the approximations here to help plan simulations of these models.

5.5. *Open Queueing Networks*

In this section we indicate how the RBM approximation (34) or the refinement (35) can be applied to determine approximate required simulation run lengths to estimate steady-state performance measures for open queueing networks to desired statistical precision. We assume that the arrival rate at queue i, say λ_i, is known for all i, e.g., by solving the usual traffic rate equations. If interest is focused on *any single queue* in the network, then the previous analysis applies, and we apply (35). We can often obtain the required parameters a_i and b_i for queue i from heavy traffic limits if queue i is the unique bottleneck queue (see Reiman 1983), but we can also use practical approximations for these parameters, whether or not there is a bottleneck queue. For example, for one queue in an open network of multi-server queues, we can apply the queueing network approximation in Whitt (1983a) to obtain

$$a_i = -m_0/\tau_i \qquad \text{and} \qquad b_i = m_i(c_{Ai}^2 + c_{Si}^2)/\tau_i, \tag{52}$$

where m_i is the number of servers, τ_i is the mean service time, and c_{Ai}^2 and c_{Si}^2 are the approximate arrival and service variability parameters at queue i. The additional approximation is only in the variability parameters c_{Ai}^2 and c_{Si}^2; this approximation method involves solving a system of linear equations to obtain the arrival variability parameters c_{Ai}^2.

Suppose that we are interested in estimating the mean queue length at each queue in an open queueing network with n queues. A natural simulation run length to choose is the maximum required for any one queue. Using (36), (52) and the relative width criterion in (10) with parameters ϵ and β, we obtain an approximate required simulation run length

$$t_r(\epsilon, \beta) \approx \max_{1 \le i \le n} \left\{ \frac{8\tau_i(c_{Ai}^2 + c_{Si}^2) z_{\beta/2}^2}{m_i \rho_i^2 (1 - \rho_i)^2 \epsilon^2} \right\}, \tag{53}$$

where i is the index of the queue. Note, in comparison to (42), that the mean service times τ_i play an important role in (53) as well as the traffic intensities ρ_i; the mean service times and net arrival rates might be very different at different queues.

Finally, suppose that we want to estimate the *total number* of customers in the queueing network. Then we suggest, as in Whitt (1983a), the approximation resulting from regarding the separate queues as mutually independent; i.e., approximate the asymptotic variance of the total mean $\sum_{i=1}^{n} E[Q_i(0)]$ by the sum of the asymptotic variances, namely,

$$\sigma^2_{\rho_1,\ldots,\rho_n} \approx \sum_{i=1}^{n} [\tau_i \rho_i^2 (c_{Ai}^2 + c_{Si}^2)^3 / 2m_i(1 - \rho_i)^4], \tag{54}$$

once again using (35) and (52).

Note that the total mean and its asymptotic variance in (54) will be dominated by the contribution of a bottleneck queue if there is one. Note that the independence approximation in (54) is not exact even for a Markov Jackson open queueing network; even though the steady state queue lengths $Q_i(\infty)$ at the different queues are mutually independent in this model, the sample averages $\bar{Q}_{it}$ in (1) at the different queues are typically *not* mutually independent. Moreover, we expect there to be positive correlation between different queues, so that (54) is likely to underestimate the true value. Finally, note that §5.2 and §5.3 indicate possible problems in multi-class queueing networks resulting from the approximation for b_i in (52). (We are not unduly alarmed, because we are looking for rough approximations.)

6. Numerical Comparisons

In this section we see how well our approximation formulas work by making numerical comparisons with simulation results. Further numerical comparisons are contained in Asmussen (1988).

6.1. *Albin's Experiments*

To empirically examine the approximations in this paper, we first studied the simulation experiments conducted by Albin (1981) to develop and evaluate approximations for single-server queues with superposition arrival processes. (This model is a special case of §5.3, having all the batch sizes one and all service times i.i.d. The heavy-traffic limits are also covered by Iglehart and Whitt.) Simulation estimates of the mean queue length (number in system) together with sample standard deviations of these estimates are reported by Albin for four experiments in Appendices 6, 14, 17 and 20 there. The principal factors in the experimental designs are the number of component renewal arrival processes, the distributions of the interarrival times and service times, and the traffic intensity. For our purposes, it is convenient that the designs were constructed so that the asymptotic-method variability parameter of the arrival process remains unchanged as the number of component renewal arrival processes and the interarrival-time distributions vary.

Consistent with our prediction, the sample standard deviations do not vary systematically with the number of component streams or the interarrival-time distribution for a given traffic intensity and asymptotic-method variability parameter, but they do vary systematically with the traffic intensity and the asymptotic-method variability parameter itself. In particular, our formulas indicate that the sample standard deviation should be approximately proportional to $(c_A^2 + c_S^2)^{3/2}$. (The service times are i.i.d., so that $c_{AS}^2 = 0$.) A comparison with observed values for the case of exponential service times ($c_S^2 = 1$) from Appendix 6 of Albin is contained in Table 3 here. Three different values of c_A^2 are considered: $c_A^2 = 2$, 4 and 6. In Table 3 here the average sample standard deviations for each value of ρ are displayed. The average is over the averages in all cases (different

TABLE 3

Sample Standard Deviations Based on 20 *Batches for Single-server Queues with Superposition Arrival Processes and Exponential Service Times, from Appendix* 6 *of Albin* (1981): *a Comparison of Observed Ratios of Standard Deviations* (SD_i/SD_j *when* $c_a^2 = i$ *and* j) *with the Heavy-traffic Approximation* $[(i+1)/(j+1)]^{3/2}$.

Traffic Intensity ρ	Average Sample Standard Deviation			Ratio of Standard Deviations		
	$c_A^2 = 2$	$c_A^2 = 4$	$c_A^2 = 6$	$\frac{SD_4}{SD_2}$	$\frac{SD_6}{SD_4}$	$\frac{SD_6}{SD_2}$
0.5	0.0138	0.0196	0.0226	1.42	1.15	1.64
0.7	0.0367	0.0670	0.0964	1.83	1.43	2.62
0.8	0.075	0.156	0.244	2.08	1.56	3.25
0.9	0.302	0.659	1.127	2.18	1.71	3.73
Predicted Ratios				2.15	1.66	3.56

numbers of component streams and different interarrival-time distributions) for a given ρ. Then the ratios SD_i/SD_j are calculated for each ρ; e.g., SD_4 is the sample standard deviation of the estimate when $c_A^2 = 4$. Since $c_S^2 = 1$, the predicted ratios for SD_4/SD_2, SD_6/SD_4 and SD_6/SD_2 are $([4+1]/[2+1])^{3/2} = (5/3)^{3/2} = 2.15$, $(7/5)^{3/2} = 1.66$ and $(7/3)^{3/2} = 3.56$, respectively. From Table 3 (and Albin's more detailed data), we see that there is a very close agreement for high values of ρ (as we would expect, because the approximation is based on a heavy traffic limit), but less agreement as ρ decreases. Consistent with the $M/G/1$ analysis in §3.1, we see that the observed ratios decrease as ρ decreases. The impact of variability thus seems to be as predicted, quantitatively for higher ρ (e.g., $\rho \geq 0.8$), and qualitatively for all ρ. It appears that a refined approximation could be developed, stating that the asymptotic variance is directly proportional to $x^{f(\rho)}$ where x is the asymptotic variability parameter (for $M/G/1$, $x = (1 + c_S^2)$ and here $x = (c_A^2 + 1)$) and $f(\rho)$ is a decreasing function of ρ with $f(1) = 3$, but we do not pursue it here.

Of course, as predicted, the impact of variability is much less than the impact of the traffic intensity. As indicated in the introduction, Albin's data also support the use of the RBM approximations to capture the principal traffic intensity effect.

6.2. *The Packet Queue Model*

We conclude by comparing simulation estimates with the heavy-traffic predictions for a special case of the packet queue model of §5.3. The particular case was chosen to realistically represent typical traffic in a packet network; see Fendick et al.

For this special case, the packet lengths for each class are deterministic, which implies that the service times and spacings for each class are deterministic, so that $c_{si}^2 = c_{xi}^2 = 0$; the idle times are assumed to have exponential distribution, so that $c_{Ii}^2 = 1$; and the batch sizes are assumed to have a geometric distribution, so that $c_i^2 = (m_i - 1)/m_i$. Even though the component variability parameters c_{si}^2, c_{xi}^2, c_{Ii}^2 and c_{bi}^2 are relatively small (≤ 1), variability plays a big role here due to the batch arrivals and the class-dependent service times.

Altogether we consider 50 independent sources of traffic. There are 25 originating sources and 25 acknowledgement sources. The acknowledgements are shorter packets sent back from the destination to the source to indicate that a packet arrived. Since the network is full duplex (there are separate channels in each direction), originating packets do not interact directly with their acknowledgements. As an approximation, we treat the

acknowledgements as independent sources. Except for the shorter service times, we model the acknowledgements like the originating traffic. Justification for this approximation appears in Whitt (1988).

In this case there are 20 "interactive" originating sources and 20 "interactive" acknowledgement sources, each with $m_i = 2$, $\xi_i = 600$ and $w_i = 86{,}800$; there are 5 "batch" originating sources and 5 "batch" acknowledgement sources, each with $m_i = 30$, $\xi_i = 1200$ and $w_i = 594{,}000$. Time here is measured in milliseconds. The batch sources generate less frequent longer messages.

The packet service times are 1000ρ for each interactive originating source, 2000ρ for each batch originating source, and 100ρ for each acknowledgement source. The overall mean service time is thus 721.8ρ. In this way, we construct a different case for each value of ρ by simply multiplying all service times by a constant. In particular, we consider 5 cases: $\rho = 0.2, 0.4, 0.6, 0.8$ and 0.9. For these parameter values, the heavy-traffic variability parameters in (50) are

$$c_A^2 = 19.95, \qquad c_S^2 = 35.80 \qquad \text{and} \qquad c_{AS}^2 = -7.74 \tag{55}$$

so that the overall asymptotic variability parameter is $(c_A^2 + c_S^2 - 2c_{AS}^2) = 71.23$. Formulas (47) and (55) imply that the expected equilibrium waiting time in this model is 35.6 times what it would be in a simple $M/M/1$ queue with the same traffic intensity ρ, as $\rho \to 1$. The heavy-traffic approximations in (48)–(51) are displayed together with corresponding $M/M/1$ values and simulation estimates in Table 4. In Table 4 each mean waiting time is divided by the mean service time.

Simulation estimates of the expected equilibrium waiting time (divided by the mean service time) appear together with estimates of 95 percent confidence intervals. Each case (value of ρ) in each of two independent runs is based on 1,800,000 packet arrivals. The estimated 95 percent confidence intervals are based on 20 batch means, each of 90,000 data points, assuming a t distribution. (We are acting as if successive batches are independent and identically distributed, even though that is not strictly correct.) We also start each run with an empty system, so that the waiting time process is not actually stationary. These discrepancies should not affect the main conclusions, however.

Several important conclusions can be deduced from Table 4. First, the high variability in (55) has a dramatic impact; the observed average waiting times are much greater than in an $M/M/1$ queue with the same traffic intensity, except for the case of the very low traffic intensity $\rho = 0.2$. Second, for the cases with $\rho \geq 0.6$, the heavy-traffic estimates apparently are quite accurate. Indeed, *for $\rho \geq 0.8$, it is likely that the heavy-traffic approximations for both the mean and the 95 percent confidence interval are more accurate*

TABLE 4

A Comparison of Simulation Estimates for the Expected Equilibrium Waiting Time (Divided by the Mean Service Time) and its 95 Percent Confidence Interval with the Heavy-Traffic Approximation for the Packet Queue Model in §6.2. Each Case (Value of ρ) in Each of Two Independent Runs is Based on 1,800,000 Packet Arrivals. The Estimated 95 Percent Confidence Intervals Were Based on 20 Batch Means Assuming a t Distribution

Traffic Intensity ρ	$M/M/1$ Exact Values	Simulation Estimates		Heavy-Traffic Theory (44)–(47)
		Run 1	Run 2	
0.2	0.25	0.32 ± 0.026	0.31 ± 0.017	8.9 ± 0.22
0.4	0.67	7.1 ± 0.57	8.2 ± 0.74	23.3 ± 0.6
0.6	1.50	40.3 ± 4.5	36.8 ± 2.3	53 ± 3.1
0.8	4.00	137 ± 46	136 ± 14	142 ± 14
0.9	9.00	293 ± 52	364 ± 82	320 ± 60

than the simulation estimates. On the other hand, for $\rho = 0.2$, where the simulation estimates are no doubt extremely accurate, the heavy-traffic approximations are horrendous. Nevertheless, we regard this experiment as very positive confirmation of the approach.

As indicated above, this experiment was conducted with equal run lengths. The difficulties with this design are obvious from Table 4. The statistical precision is probably more than desired for $\rho = 0.2$, but less than desired for $\rho \geq 0.8$. To achieve comparable relative precision, we could instead apply (10) and (48), and let the run lengths be proportional to

$$\sigma_W^2/(E[W])^2 = \frac{2(c_A^2 + c_S^2 - 2c_{AS}^2)}{\rho(1-\rho)^2}. \tag{56}$$

In particular, since the heavy-traffic variability parameters are independent of ρ, we would make the run lengths inversely proportional to $\rho(1-\rho)^2$. The percentages of total run length allocated to the cases $\rho = 0.2, 0.4, 0.6, 0.8$ and 0.9 would thus be 16 percent, 4 percent, 8 percent, 16 percent and 56 percent instead of 20 percent in each case. Of course, we might further modify the allocation to represent greater concern about statistical precision in some regions.[1]

[1] I thank J. Abate, S. L. Albin, K. W. Fendick, P. W. Glynn, D. L. Iglehart, V. R. Saksena and K. Sriram for previous collaborations which aided this work. I thank S. S. Lavenberg for pointing out the reference by Moeller and Kobayashi, and C. M. Woodside for pointing out the reference by Woodside et al. I thank K. W. Fendick for performing the simulation experiment in §6.2 and carefully reading the manuscript. I thank S. Asmussen for providing copies of Asmussen (1987c), (1988); §4.5 was added after seeing Asmussen (1987c).

References

ABATE, J. AND W. WHITT, "Transient Behavior of Regulated Brownian Motion, I and II," *Adv. in Appl. Probab.*, 19 (1987), 560–631.

——— AND ———, "The Correlation Functions of RBM and $M/M/1$," *Stochastic Models*, 4 (1988a), 315–359.

——— AND ———, "Simple Spectral Representations for the $M/M/1$ Queue," *Queueing Systems*, 3 (1988b), 321–346.

ALBIN, S. L., "Approximating Queues with Superposition Arrival Processes," Ph.D. dissertation, Columbia University, New York, 1981.

———, "On Poisson Approximations for Superposition Arrival Processes in Queues," *Management Sci.*, 28 (1982), 126–137.

———, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues," *Oper. Res.*, 32 (1984a), 1133–1162.

———, "Simulation to Develop and Test Queue Approximations: A Case Study," *Simulation*, 43 (1984b), 279–285.

———, "Delays for Customers from Different Arrival Streams to a Queue," *Management Sci.*, 32 (1986), 329–340.

ASMUSSEN, S., *Applied Probability and Queues*, Wiley, New York, 1987a.

———, "The Heavy Traffic Limit of a Class of Markovian Queueing Models," *Oper. Res. Lett.*, 6 (1987b), 301–306.

———, "Regenerative Simulation in Heavy Traffic," Aalborg University, Denmark, 1987c. To appear in *Math. Oper. Res.*

———, "Validating the Heavy Traffic Performance of Regenerative Simulation," Aalborg University, Denmark, 1988. To appear in *Stochastic Models.*

BENEŠ, V. E., "On Queues with Poisson Arrivals," *Ann. Math. Statist.*, 28 (1957), 670–677.

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

BLOMQVIST, N., "The Covariance Function of the $M/G/1$ Queueing System," *Skand. AktuarTidskr.*, (1967), 157–174.

———, "Estimation of Waiting-Time Parameters in the $GI/G/1$ Queueing System, Part I: General Results," *Skand. AktuarTidskr.*, (1968), 178–197.

———, "Estimation of Waiting-Time Parameters in the $GI/G/1$ Queueing System, Part II: Heavy-Traffic Approximations," *Skand. AktuarTidskr.*, (1969), 125–136.

BOROVKOV, A. A., "On Limit Laws for Service Processes in Multi-Channel Systems," *Siberian Math. J.*, 8 (1967), 746–763.
———, *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York, 1976.
———, *Asymptotic Methods in Queueing Theory*, Wiley, New York, 1984.
BURMAN, D. Y., "Approximations for a Service System with Interruptions," (1987), submitted for publication.
——— AND D. R. SMITH, "An Asymptotic Analysis of Queueing Systems with Markov Modulated Arrivals," *Oper. Res.*, 34 (1986), 105–119.
DALEY, D. J., "The Serial Correlation Coefficients of Waiting Times in a Stationary Single Server Queue," *J. Austral. Math. Soc.*, 8 (1968), 683–699.
——— AND D. R. JACOBS, JR., "The Total Waiting Time in a Busy Period of a Stable Single-Server Queue, II," *J. Appl. Probab.*, 6 (1969), 565–572.
FEDERGRUEN, A. AND L. GREEN, "Queueing Systems with Service Interruptions," *Oper. Res.*, 34 (1987a), 752–768.
——— AND ———, "Queueing Systems with Service Interruptions, II," Graduate School of Business, Columbia University, 1987b.
FENDICK, K. W., V. R. SAKSENA AND W. WHITT, "Dependence in Packet Queues," *IEEE Trans. Commun.*, (1989), to appear.
FISHMAN, G. S., "Estimating Sample Size in Computing Simulation Experiments," *Management Sci.*, 18 (1971), 21–38.
———, *Concepts and Methods in Discrete Event Digital Simulation*, Wiley, New York, 1973.
———, *Principles of Discrete Event Simulation*, Wiley, New York, 1978.
FLEMING, P. J., "An Approximate Analysis of Sojourn Times in the $M/G/1$ Queue with Round-Robin Service Discipline," *AT&T Bell Lab. Tech. J.*, 63 (1984), 1521–1535.
GLYNN, P. W. AND W. WHITT, "A Central-Limit-Theorem Version of $L = \lambda W$," *Queueing Systems*, 1 (1986), 191–215.
——— AND ———, "Sufficient Conditions for Functional-Limit-Theorem Versions of $L = \lambda W$," *Queueing Systems*, 1 (1987), 279–287.
——— AND ———, "Ordinary CLT and WLLN Versions of $L = \lambda W$," *Math. Oper. Res.*, 13 (1988), 674–692.
——— AND ———, "Indirect Estimation Via $L = \lambda W$," *Oper. Res.*, 37 (1989), 82–103.
HALFIN, S. AND W. WHITT, "Heavy-Traffic Limits for Queues with Many Exponential Servers," *Oper. Res.*, 29 (1981), 567–588.
HARRISON, J. M., "A Limit Theorem for Priority Queues in Heavy Traffic," *J. Appl. Probab.*, 10 (1973), 907–912.
———, *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.
———, "Brownian Models of Queueing Networks with Heterogeneous Customer Populations," *Proc. IMA Workshop on Stochastic Differential Systems*, Springer-Verlag, New York, 1988, 147–186.
——— AND R. J. WILLIAMS, "Brownian Models of Open Queueing Networks with Homogeneous Customer Populations," *Stochastics*, 22 (1988), 77–115.
IGLEHART, D. L., "Functional Limit Theorems for the Queue $GI/G/1$ in Light Traffic," *Adv. in Appl. Probab.*, 3 (1971), 269–281.
——— AND W. WHITT, "Multiple Channel Queues in Heavy Traffic, I and II," *Adv. Appl. in Probab.*, 2 (1970), 150–177 and 355–369.
JOHNSON, D. P., "Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks," Ph.D. dissertation, University of Wisconsin, 1983.
KARLIN, S. AND H. M. TAYLOR, *A First Course in Stochastic Processes*, Second Ed., Academic Press, New York, 1975.
LAW, A. M., "Efficient Estimators for Simulated Queueing Systems," *Management Sci.*, 22 (1975), 30–41.
LOYNES, R. M., "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times," *Proc. Cambridge Philos. Soc.*, 48 (1962), 497–520.
MOELLER, T. AND H. KOBAYASHI, "Use of Diffusion Approximation to Estimate Run Length in Simulation Experiments," *COMPSTAT* 1974, *Proc. Computational Statistics*, G. Bruckmann, F. Fershel and L. Schmetterer (Eds.), Physica-Verlag, Vienna, 1974, 363–372.
MORSE, P. M., "Stochastic Properties of Waiting Lines," *Oper. Res.*, 3 (1955), 255–261.
NEWELL, G. F., *Applications of Queueing Theory*, Second ed., Chapman and Hall, London, 1982.
OTT, T. J., "The Covariance Function of the Virtual Waiting-Time Process in an $M/G/1$ Queue," *Adv. in Appl. Probab.*, 9 (1977a), 158–168.
———, "The Stable $M/G/1$ Queue in Heavy Traffic and its Covariance Function," *Adv. in Appl. Probab.*, 9 (1977b), 169–186.
PAKES, A. G., "The Correlation Coefficients of the Queue Lengths of Some Stationary Single Server Queues," *J. Austral. Math. Soc.*, 12 (1971), 35–46.
PARZEN, E., *Stochastic Processes*, Holden-Day, San Francisco, 1962.

PETERSON, W. P., "Diffusion Approximations for Networks of Queues with Multiple Customer Types," Ph.D. dissertation, Stanford University, 1985.

PRISGROVE, L., "Closed Networks of Queues with Multiple Servers: Transient and Steady-State Approximations," Ph.D. dissertation, Stanford University, 1987.

REIMAN, M. I., "Some Diffusion Approximations with State-Space Collapse," *Proc. Internat. Seminar Modeling and Performance Eval. Methodology*, F. Baccelli and G. Fayolle (eds.), Springer-Verlag, Berlin, 1983, 209–240.

———, "Open Queueing Networks in Heavy Traffic," *Math. Oper. Res.*, 9 (1984), 441–458.

———, A Multi-Class Feedback Queue in Heavy Traffic. *Adv. in Appl. Probab.*, 20 (1988a), 179–207.

———, "A Network of Priority Queues in Heavy Traffic: One Bottleneck Station," AT&T Bell Laboratories, Murray Hill, NJ, 1988b.

——— AND B. SIMON, "An Interpolation Approximation for Queueing Systems with Poisson Input," *Oper. Res.*, 36 (1988), 454–469.

REYNOLDS, J. F., "The Covariance Structure of Queues and Related Stochastic Processes—A Survey of Recent Work," *Adv. in Appl. Prob.*, 7 (1975), 383–415.

SRIRAM, K. AND W. WHITT, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE J. Sel. Areas Commun.*, SAC-4 (1986), 833–846.

WHITT, W., "Weak Convergence Theorems for Queues in Heavy Traffic," Ph.D. dissertation, Cornell University, and Technical Report 2, Department of Operations Research, Stanford University, 1968.

———, "Weak Convergence Theorems for Priority Queues: Preemptive-Resume Discipline," *J. Appl. Probab.*, 8 (1971), 74–94.

———, "Embedded Renewal Processes in the $GI/G/s$ Queue," *J. Appl. Probab.*, 9 (1972), 650–658.

———, "Heavy Traffic Limit Theorems for Queues: A Survey," *in Mathematical Methods in Queueing Theory*, A. B. Clarke (Ed.), Lecture Notes in Econ. and Math. Systems 98, Springer-Verlag, New York, 1974, 307–350.

———, "Some Useful Functions for Functional Limit Theorems," *Math. Oper. Res.*, 5 (1980), 67–85.

———, "On the Heavy-Traffic Limit Theorem for $GI/G/\infty$ Queues," *Adv. in Appl. Probab.*, 14 (1982a), 171–192.

———, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," *Oper. Res.*, 30 (1982b), 125–147.

———, "Refining Diffusion Approximations for Queues," *Oper. Res. Lett.*, 1 (1982c), 165–169.

———, "The Queueing Network Analyzer," *Bell System Tech. J.*, 62 (1983a), 2779–2815.

———, "Performance of the Queueing Network Analyzer," *Bell System Tech. J.*, 62 (1983b), 2817–2843.

———, "Approximations for Departure Processes and Queues in Series," *Naval Res. Logist. Quart.*, 31 (1984a), 499–521.

———, "On Approximations for Queues, I: Extremal Distributions," *AT&T Bell Lab. Tech. J.*, 63 (1984b), 115–138.

———, "A Light-Traffic Approximation for Single-Class Departure Processes from Multi-Class Queues," *Management Sci.*, 34 (1988), 1333–1346.

WILSON, J., "Variance Reduction Techniques for Digital Simulation," *Amer. J. Math. Management Sci.*, 4 (1985), 277–312.

WOODSIDE, C. M., B. PAGUREK AND G. F. NEWELL, "A Diffusion Approximation for Correlation in Queues," *J. Appl. Probab.*, 17 (1980), 1033–1047.