Operations Research
Vol. 52, No. 6, November–December 2004, pp. 922–941
ISSN 0030-364X | EISSN 1526-5463 | 04 | 5206 | 0922

DOI 10.1287/opre.1040.0136
© 2004 INFORMS

A Diffusion Approximation for the $G/GI/n/m$ Queue

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027-6699,
ward.whitt@columbia.edu

We develop a diffusion approximation for the queue-length stochastic process in the $G/GI/n/m$ queueing model (having a general arrival process, independent and identically distributed service times with a general distribution, n servers, and m extra waiting spaces). We use the steady-state distribution of that diffusion process to obtain approximations for steady-state performance measures of the queueing model, focusing especially upon the steady-state delay probability. The approximations are based on heavy-traffic limits in which n tends to infinity as the traffic intensity increases. Thus, the approximations are intended for large n.

For the $GI/M/n/\infty$ special case, Halfin and Whitt (1981) showed that scaled versions of the queue-length process converge to a diffusion process when the traffic intensity ρ_n approaches 1 with $(1-\rho_n)\sqrt{n} \to \beta$ for $0 < \beta < \infty$. A companion paper, Whitt (2005), extends that limit to a special class of $G/GI/n/m_n$ models in which the number of waiting places depends on n and the service-time distribution is a mixture of an exponential distribution with probability p and a unit point mass at 0 with probability $1-p$. Finite waiting rooms are treated by incorporating the additional limit $m_n/\sqrt{n} \to \kappa$ for $0 < \kappa \leqslant \infty$. The approximation for the more general $G/GI/n/m$ model developed here is consistent with those heavy-traffic limits. Heavy-traffic limits for the $GI/PH/n/\infty$ model with phase-type service-time distributions established by Puhalskii and Reiman (2000) imply that our approximating process is not asymptotically correct for non-exponential phase-type service-time distributions, but nevertheless, the heuristic diffusion approximation developed here yields useful approximations for key performance measures such as the steady-state delay probability. The accuracy is confirmed by making comparisons with exact numerical results and simulations.

Subject classifications: queues, approximations: multiserver queues; queues, multichannel: diffusion approximation.
Area of review: Stochastic Models.
History: Received July 2002; revision received March 2003; accepted September 2003.

Introduction

The rapid growth of telephone call centers and more general customer contact centers has generated renewed interest in the performance of multiserver queueing models when the number of servers is large; e.g., see Armony and Maglaras (2004), Borst et al. (2004), Gans et al. (2003), Garnett et al. (2002), Mandelbaum (2001), Whitt (2003), and references therein.

Because these multiserver systems often have a very large number of servers, it is natural to look for insight into system performance by considering asymptotics as the number of servers is allowed to increase. Such limits were established for the $GI/M/n/\infty$ queueing model (with renewal arrival process, exponential service times, n servers, and unlimited waiting room) by Halfin and Whitt (1981), for the more general $GI/PH/n/\infty$ model (with phase-type service times) by Puhalskii and Reiman (2000), and for the $M/M/n/\infty$ model with exponential customer abandonment by Garnett et al. (2002). They considered a sequence of models indexed by the number of servers, n, and let $n \to \infty$ with the traffic intensities ρ_n converging to one, the critical value for stability. Interesting nondegenerate limits occur when

$$\sqrt{n}(1-\rho_n) \to \beta \quad \text{for } 0 < \beta < \infty. \tag{0.1}$$

An important performance measure in this setting is the delay probability, i.e., the steady-state probability that an arriving customer finds all servers busy and must wait in queue before starting service. For the $GI/M/n/\infty$ model (and presumably for the $GI/PH/n/\infty$ model as well, although it remains to be proved), if $n \to \infty$ with condition (0.1) holding, then the associated sequence of delay probabilities approaches a limit α strictly between 0 and 1. Because the delay probabilities require no scaling by a function of n for that limit, the delay probability tends to be an especially useful performance measure, as suggested by Whitt (1992).

For the $M/M/n/\infty$ model, the delay-probability limit has a relatively simple form,

$$\alpha \equiv \alpha(\beta) = [1 + \beta\Phi(\beta)/\phi(\beta)]^{-1}, \tag{0.2}$$

where β is the limit in (0.1), Φ is the cumulative distribution function (CDF), and ϕ is the probability density function (PDF) of a standard (mean 0, variance 1) normal random variable; e.g., $\Phi(x) = P(N(0,1) \leqslant x)$. The function α in (0.2) is a continuous, strictly convex, strictly decreasing function on the positive halfline, with $\alpha(0) = 1$ and $\alpha(\beta) \to 0$ as $\beta \to \infty$; see Lemma B.1 of Borst et al. (2004). For the $M/M/n/\infty$ model, the asymptotic-delay-probability function α in (0.2) plays a crucial role in further analysis, as can be seen from the recent papers cited above. Even though the asymptotic-delay-probability function α in (0.2) arises in the limit as $n \to \infty$, it provides a good approximation for the actual $M/M/n/\infty$ delay probability *for all n* provided that ρ is not too small (e.g., when the actual delay probability is greater than or equal to 0.10; see Table 13 in Whitt 1993).

Because the asymptotic-delay-probability function α in (0.2) has proven to be so important for the Markovian $M/M/n/\infty$ system, we want to find analogs for non-Poisson arrival processes, nonexponential service-time distributions, and finite waiting space. The present paper addresses that problem. In this paper, we consider the general $G/GI/n/\infty$ model: We allow the arrival process to be a general stationary (or asymptotically stationary) arrival process (G), but we require that the service times be independent and identically distributed (IID) and independent of the arrival process (with a general probability distribution, *GI*). In practice, nonexponential service-time distributions are common, but arrival processes often can be regarded as Poisson. Non-Poisson arrival processes commonly occur when some of the arrivals are overflows from other systems that are temporarily congested.

In this paper, we primarily focus on an approximation for the steady-state delay probability and the steady-state probability that all servers are busy in the $G/GI/n/\infty$ model. These two quantities are equal with a Poisson arrival process, but not more generally. However, they are asymptotically equivalent in the heavy-traffic limit as $n \to \infty$, so we do not distinguish between them: Our approximation applies to both. Even though we focus on the steady-state delay probability, we also develop an approximation for the entire queue-length (number in system) stochastic process and its steady-state distribution.

We base our approximation for the queue-length process on a heavy-traffic limit for the $G/GI/n/\infty$ model with a special H_2^* service-time distribution, established in Whitt (2005). (Related heavy-traffic limits for the $G/D/n/\infty$ model have recently been established by Jelenkovic et al. 2004.) The H_2^* service-time distribution is a mixture of an exponential distribution with probability p and a point mass on 0 with probability $1 - p$. The general form of our proposed approximation for the queue-length process is the form of the limit process obtained for the $G/H_2^*/n/m$ model, namely, a convex piecewise-linear function of a diffusion process. Interestingly, that process is not directly a diffusion process, but because it is a relatively simple function of a diffusion process, we call the overall approximation a diffusion approximation. For applications, it is significant that the approximating process has a tractable steady-state distribution.

A major conclusion of Halfin and Whitt (1981), expanded upon by Puhalskii and Reiman (2000), is that the role of the service-time distribution in the many-server heavy-traffic asymptotic regime (0.1) is very different from its role in the more conventional fixed-number-of-servers heavy-traffic limit with convergence to reflected Brownian motion. In the conventional heavy-traffic limit, convergence to a diffusion process requires that the service-time distribution have a finite variance; then, the limit depends on the service-time distribution beyond its mean only via its variance. (Other nondiffusion heavy-traffic limits are possible in the conventional heavy-traffic regime when the service-time distribution has infinite variance, but the limit process and the scaling are then very different; e.g., see Whitt 2002.) Moreover, in the conventional heavy-traffic regime, the standard congestion measures increase as the service-time variance increases for any fixed service-time mean. As we substantiate here by computer simulations, the situation is very different for the many-server asymptotic regime (0.1). For example, in some multiserver settings, the delay probability actually *decreases* as the service-time variance increases for fixed service-time mean. Moreover, the same multiserver approximations may be appropriate for service-time distributions with infinite variance. We do not yet adequately understand the impact of the service-time distribution beyond its mean upon the performance of multiserver queues, but in this paper we make a step forward.

Another objective in this paper is to develop approximations for the case of a finite waiting room. To do so, we again rely on heavy-traffic limits in Whitt (2005). Those heavy-traffic limits involve the more general $G/GI/n/m_n$ model with m_n additional waiting places. An arrival finding all servers busy and the waiting room full is blocked and lost without affecting future arrivals. (We do not consider abandonments or retrials here.) For the heavy-traffic stochastic-process limits in the heavy-traffic regime (0.1), it is necessary to let $m_n \to \infty$ as $n \to \infty$ so that

$$m_n/\sqrt{n} \to \kappa \quad \text{for } 0 < \kappa \leqslant \infty. \tag{0.3}$$

For exponential service times, the results for finite waiting rooms provide theoretical support and refinements for heuristic diffusion approximations in §VII of Whitt (1984a). Related asymptotic analysis of the $M/M/n/m$ model has recently been done by Massey and Wallace (2004).

The rest of this paper is organized as follows: We start in §1 by describing the development of the proposed approximation of the delay probability in the $G/GI/n/\infty$ model. We state the stochastic-process limit obtained in Whitt (2005) for the $G/H_2^*/n/m$ models in §2 and characterize the steady-state distribution of that limit process in §3. We then develop the heuristic diffusion approximation for the $G/GI/n/m$ model in §4.

In §5, we evaluate the approximation for the delay probability in the $GI/GI/n/\infty$ model by making comparisons with exact numerical values from the tables of Seelen et al. (1985). In §6, we describe simulations conducted to evaluate other $G/GI/n/\infty$ models, focusing especially on heavy-tailed service-time distributions and nonrenewal arrival processes. In §7, we develop and evaluate associated approximations for the blocking probability in the $G/GI/n/m$ model. In §8, we make a few concluding remarks.

1. The Delay Probability in the $G/GI/n/\infty$ Model

We now describe the evolution of our approximation of the delay probability in the $G/GI/n/\infty$ model, which generalizes (0.2). As noted above, Halfin and Whitt (1981) actually made some progress for more general models by establishing the heavy-traffic stochastic-process limit for the $GI/M/n/\infty$ model as well as the $M/M/n/\infty$ model, but they gave an incorrect expression for the steady-state distribution of the diffusion-process limit in the $GI/M/n/\infty$ case, which leads to an incorrect generalization of the asymptotic-delay-probability function α. However, the correct formula for the asymptotic delay probability in the $GI/M/n/\infty$ model can easily be derived from the diffusion-process parameters in Halfin and Whitt (1981); e.g., it can be obtained from Browne and Whitt (1995).

As indicated in Example 18.1 of Browne and Whitt (1995), the limiting diffusion process obtained in Halfin and Whitt (1981) is a piecewise-linear diffusion process, i.e., a diffusion process with piecewise-linear drift and diffusion functions, with two linear components, corresponding to the situations in the queueing model in which not all servers are busy and when they are. The delay probability corresponds to p_2 in (18.3) and (18.5) there. It is found by applying (18.5) and (18.6) (or, equivalently, (18.26)) with §18.4.1 (when the servers are not all busy) and §18.4.3 (when the servers are all busy). We apply Browne and Whitt (1995) again here in §3.

The corrected $GI/M/n/\infty$ asymptotic-delay-probability function is a minor modification of the $M/M/n/\infty$ function above; specifically,

$$\alpha_{GI/M/n/\infty} \equiv \alpha_{GI/M/n/\infty}(\beta, c_a^2) = \alpha(\beta/\sqrt{z}), \tag{1.1}$$

where $z = (c_a^2+1)/2$, with c_a^2 being the squared coefficient of variation (SCV, variance divided by the square of the mean, assumed to be finite) of an interarrival time, β is the limit in (0.1), and α is the $M/M/n/\infty$ asymptotic-delay-probability function in (0.2). From (1.1), we see that the interarrival-time distribution beyond the mean enters in only via the SCV c_a^2, just as in the central limit theorem for the arrival counting process. Halfin and Whitt (1981) had the incorrect formula $\alpha(\beta/z)$ instead of $\alpha(\beta/\sqrt{z})$. Unfortunately, that incorrect formula has been repeated, e.g., in Whitt (1992, 1993, 2002).

In the next section, we describe a new heavy-traffic limit for the more general $G/GI/n/\infty$ model with a nonrenewal arrival process and a special nonexponential service-time distribution, which we establish in a companion paper Whitt (2005). The nonexponential service-time distribution is the mixture of an exponential distribution with probability p and a unit point mass at 0 with probability $1-p$. This special service-time distribution is mathematically appealing because, just like the exponential service-time distribution, it makes appropriate queue-length processes Markov processes. Because this special distribution is an extremal distribution among the class of hyperexponential (H_2, mixtures of two exponentials) distributions (see Whitt 1984b), we denote this class by H_2^*.

Puhalskii and Reiman (2000) already established many-server heavy-traffic limits for the more general (and more difficult) $GI/PH/n/\infty$ model with phase-type service-time distributions, but the limit process there is a complicated multidimensional diffusion process, whose steady-state distribution remains to be determined. Thus, we are motivated to consider heuristic one-dimensional alternatives.

Clearly, the H_2^* service-time distributions are rather special, and cannot be regarded as similar to all service-time distributions. However, they are natural abstractions for the case in which the service-time distribution is the mixture of two other distributions, one with a small mean and the other with a large mean. More generally, they capture the behavior of many heavy-tailed distributions (with finite mean), such as lognormal and Pareto, that produce many small values and a few occasional very large values. These heavy-tailed distributions are being encountered more and more frequently; e.g., measurements have suggested that service-time distributions are lognormal; see Bolotin (1994), Gans et al. (2003), and Brown et al. (2002).

For the $G/H_2^*/n/\infty$ model, formula (1.1) is still valid, provided we appropriately modify the formula for z; in particular,

$$\alpha_{G/H_2^*/n/\infty} \equiv \alpha_{G/H_2^*/n/\infty}(\beta, c_a^2, p) = \alpha(\beta/\sqrt{z}) \tag{1.2}$$

for α in (0.2), β in (0.1), and

$$z \equiv z(c_a^2, p) = 1 + \frac{p(c_a^2-1)}{2} = \frac{p(c_a^2+c_s^2)}{2}, \tag{1.3}$$

where $c_s^2 = (2/p) - 1$ for an H_2^* service-time distribution and c_a^2 is the scaling constant in an assumed functional central limit theorem (FCLT) for the arrival process; see (2.1) and (2.2) in §2. For a renewal arrival process, c_a^2 is just the SCV of an interarrival time.

Because $z(c_a^2, 1) = (c_a^2+1)/2$, approximation (1.2) reduces to (1.1) in the $G/M/n/\infty$ special case. Because $z(1, p) = 1$ for all p, $0 < p \leqslant 1$, formula (1.2) supports the approximation

$$\alpha_{M/GI/n/\infty} \approx \alpha_{M/M/n/\infty} \equiv \alpha(\beta), \tag{1.4}$$

which is a longstanding approximation; e.g., see Hokstad (1978), Nozaki and Ross (1978), §3.2 of Whitt (1993), and Kimura (2000). The limit in (1.2) and the approximation in (1.4) indicate that the delay probability in the $M/GI/n/\infty$ model should not be significantly altered by a heavy-tailed service-time distribution, provided that it has finite mean. However, as is well known, the service-time distribution beyond its mean can have a significant impact on the distribution of the conditional queue length given that all servers are busy.

Formulas (1.2) and (1.3) are very useful to predict the qualitative behavior of the delay probability as a function of the arrival-process and service-time variability. First, because α is a decreasing function, $\alpha(\beta/\sqrt{z})$ is an increasing function of z. Second, using (1.2), we see that z is always an increasing function of c_a^2. Moreover, we see that z is an increasing (decreasing) function of c_s^2 when $c_a^2 < 1$ ($c_a^2 > 1$), with all values falling between 1 and c_a^2.

As might be anticipated, however, the peculiar form of this tractable H_2^* nonexponential service-time distribution causes the limit in (1.2) not to perform well as an approximation for the performance of $G/GI/n/\infty$ models with typical nonexponential service-time distributions if we just match the first two moments of the service-time distribution using (1.3). Thus, we develop a new heuristic one-dimensional diffusion approximation that produces more useful approximations for general $G/GI/n/\infty$ models.

As in previous work, e.g., Whitt (1992), the heuristic diffusion approximation is based on an infinite-server approximation when all servers are not busy and a single-server approximation when all servers are busy. In those two regimes we rely on established heavy-traffic limits, so that again heavy-traffic asymptotics play a key role. However, the specific method is new: We first determine an approximating (function of a) diffusion process. Then, we use the exact steady-state distribution of the approximating process.

From the heuristic diffusion approximation for the $G/GI/n/\infty$ model, we obtain a relatively simple approximation for the delay probability, namely,

$$\alpha_{G/GI/n/\infty} \equiv \alpha_{G/GI/n/\infty}(\beta, z) \approx \alpha(\beta/\sqrt{z}), \tag{1.5}$$

where again α is the $M/M/n/\infty$ asymptotic-delay-probability function in (0.2) and β is the limit in (0.1). The key new quantity is

$$z \equiv z(c_a^2, G) \equiv 1 + (c_a^2 - 1)\eta(G), \tag{1.6}$$

where G is the service-time CDF, assumed to have finite mean $1/\mu$, $G^c \equiv 1 - G$ is the associated complementary CDF,

$$\eta(G) \equiv \mu \int_0^\infty G^c(x)^2\, dx \equiv \frac{\int_0^\infty G^c(x)^2\, dx}{\int_0^\infty G^c(x)\, dx}, \tag{1.7}$$

and, just as in (1.3), c_a^2 is the normalization constant in a FCLT for the arrival process (assumed to hold, which requires that c_a^2 be finite).

From (1.6) we see that the service-time distribution beyond its mean should have relatively little impact upon the delay probability when c_a^2 is near 1, which is consistent with extensive simulation experience. On the other hand, when c_a^2 is not near 1, the service-time distribution beyond its mean should have a significant impact on the delay probability, and that impact is quantified approximately by (1.5)–(1.7). It is worth noting that the service-time parameter $\eta(G)$ is well defined for all service-time distributions with finite mean. There is no requirement that the service time have finite variance.

The parameter z in (1.6) is the *asymptotic peakedness* that appears in approximations for $G/GI/n/0$ loss models; e.g., see Eckberg (1983, 1985) and Whitt (1984a). It was used before for delay models in Whitt (1992). The *peakedness* is the variance divided by the mean of the steady-state queue length (again number in system) in the associated $G/GI/\infty$ model. From heavy-traffic limits for the $G/GI/\infty$ model, it follows that the peakedness approaches the asymptotic peakedness as the arrival rate increases; see §10.3 of Whitt (2002).

Just like the SCV, the peakedness and the asymptotic peakedness are dimensionless parameters quantifying variability. The function $\eta(G)$ in (1.7) can assume any value between 0 and 1. The maximum value 1, yielding $z = c_a^2$, is obtained when G is the CDF of a unit point mass (a deterministic distribution, D). The value of $\eta(G)$ tends to decrease as the distribution gets more variable. For an exponential service-time CDF G, $\eta(G) = 1/2$, yielding $z = (c_a^2 + 1)/2$.

As emphasized by our notation above, the approximation for the delay probability in the general $G/GI/n/\infty$ model in (1.5) is consistent with the heavy-traffic limit for the $G/H_2^*/n/\infty$ model in (1.2). In the previous special cases, z coincides with the asymptotic peakedness for that model. A natural candidate for a refined approximation (which we do not investigate here) is obtained by replacing the asymptotic peakedness z in (1.6) with the actual peakedness and the asymptotic-delay-probability function α in (0.2) with the actual $M/M/n/\infty$ (Erlang-C) delay probability. For practical engineering purposes, we do not anticipate that such a refinement would be too important, but that remains to be determined.

From the discussion above, and consistent with intuition, the $G/GI/n/\infty$ model behaves much like the associated $G/GI/\infty$ model when the arrival rate λ and n increase so that (0.1) holds. However, the delay-probability approximation in (0.2), (1.1), (1.2), and (1.5) are not exactly the same as the direct infinite-server approximation for the delay probability, which is $\Phi^c(\beta/\sqrt{z})$ for $\Phi^c \equiv 1 - \Phi$; e.g., see §2.3 of Whitt (1992). The delay-probability approximation is obtained here simply by replacing Φ^c by α. Halfin and Whitt (1981) observe in their Remark 1 that $\alpha(\beta) \geqslant \Phi^c(\beta)$

for all $\beta \geqslant 0$. These formulas are asymptotically equivalent as $\beta \to \infty$, but they are not always close; e.g., $\Phi^c(0) = 0.5$, while $\alpha(0) = 1$. The refinement—going from Φ^c to α—was used by Jennings et al. (1996) in their server-staffing approximations for multiserver queues with time-varying arrival rates. They observed that the refinement typically improved the estimate by about 10%.

2. The Stochastic-Process Limit with H_2^* Service Times

In this section, we describe the heavy-traffic limit for the $G/H_2^*/n/m$ model established in Whitt (2005). It involves a sequence of $G/H_2^*/n/m$ models indexed by the number of servers, n, with $n \to \infty$.

We start with a rate-1 arrival counting process $C \equiv \{C(t)\colon t \geqslant 0\}$ with associated interarrival times $\{U_k\colon k \geqslant 1\}$. Our key assumption is that the arrival process satisfies a FCLT. To state it, let $\Rightarrow$ denote convergence in distribution and let $D \equiv (D, J_1) \equiv D([0, \infty), \mathbb{R}, J_1)$ be the function space of right-continuous real-valued functions on the positive halfline with left limits, endowed with the customary Skorohod (J_1) topology; see Billingsley (1999) and Whitt (2005). Let $\mathbf{C}_n$ be the random element of D defined by

$$\mathbf{C}_n(t) \equiv [C(nt) - nt]/\sqrt{nc_a^2}, \quad t \geqslant 0, \tag{2.1}$$

for some nonnegative scaling constant c_a^2. We assume that

$$\mathbf{C}_n \Rightarrow \mathbf{B} \quad \text{in } (D, J_1), \tag{2.2}$$

where $\mathbf{B}$ is standard (zero drift, unit diffusion coefficient) Brownian motion.

When the arrival process is a renewal process, the limit (2.2) holds with c_a^2 being the SCV of an interarrival time, but the limit in (2.1) holds much more generally. When the number of servers is n, we scale time in the arrival process, letting the arrival process be

$$C_n(t) \equiv C(\lambda_n t), \quad t \geqslant 0, \tag{2.3}$$

where λ_n is the arrival rate in model n (with n servers). Equivalently, the interarrival times in model n are $U_{n,k} \equiv U_k/\lambda_n$.

Let the H_2^* service-time distribution be independent of n. Let it have mean μ^{-1}, $0 < \mu < \infty$, so that the traffic intensity as a function of n is $\rho_n = \lambda_n/\mu n$. Let ν^{-1} be the mean of the exponential component of the H_2^* service-time distribution, so that $\mu^{-1} = p\nu^{-1}$. The second moment of a service time is thus $2p\nu^{-2}$, so that the SCV is $c_s^2 = (2/p) - 1$. Equivalently, $p^{-1} = (c_s^2 + 1)/2$. The SCV c_s^2 ranges from 1 to ∞ as p decreases from 1 to 0.

Let $Q_n(t)$ be the queue length at time t, by which we mean the number in the system, including both waiting and in service. We assume that the stochastic process Q_n almost surely has sample paths in the function space D; in particular, the process Q_n provides no record of an arrival with zero service time that can enter service upon arrival and depart immediately. For the stochastic-process limit, we construct scaled random elements of D by letting

$$\mathbf{Q}_n(t) \equiv [Q_n(t) - n]/\sqrt{n}, \quad t \geqslant 0. \tag{2.4}$$

There is no time scaling for $\mathbf{Q}_n$ in (2.4) because the arrival rate λ_n is allowed to grow directly.

We also must specify the initial conditions. Let $Q_n(0)$ be an integer-valued random variable with

$$0 \leqslant Q_n(0) \leqslant n + m_n \tag{2.5}$$

that is independent of the arrival process $\{C_n(t)\colon t \geqslant 0\}$. We assume that

$$\mathbf{Q}_n(0) \Rightarrow \mathbf{Q}(0) \quad \text{as } n \to \infty, \tag{2.6}$$

where $\mathbf{Q}(0)$ is a proper random variable and

$$\mathbf{Q}_n(0) \equiv [Q_n(0) - n]/\sqrt{n}. \tag{2.7}$$

Moreover, we assume that the $\min\{n, Q_n(0)\}$ customers initially in service have exponential service times with mean ν^{-1}, while the $[Q_n(0) - n]^+$ customers initially waiting in queue have the H_2^* CDF. Finally, given that specification, we assume that all service times are independent of the initial state $Q_n(0)$ and of the arrival process.

Theorem 2.1 (The Stochastic-Process Limit for $G/H_2^*/n/m$). *For the family of $G/H_2^*/n/m$ models specified above, suppose that the arrival rate λ_n and the number of waiting spaces, m_n, change with n so that* (0.1) *and* (0.3) *hold with $-\infty < \beta < \infty$ and $0 < \kappa \leqslant \infty$. In addition, suppose that the initial conditions are as specified above with* (2.5)–(2.7). *Then,*

$$\mathbf{Q}_n \Rightarrow \mathbf{Q} \quad \textit{in } (D, J_1) \quad \textit{as} \quad n \to \infty, \tag{2.8}$$

where

$$\mathbf{Q}(t) \equiv h(\mathbf{Q}^p(t)), \quad t \geqslant 0, \tag{2.9}$$

$$h(x) \equiv \begin{cases} x, & x < 0, \\ x/p, & 0 \leqslant x \leqslant p\kappa, \end{cases} \tag{2.10}$$

and $\mathbf{Q}^p$ is a diffusion process starting at $\mathbf{Q}^p(0) = h^{-1}(\mathbf{Q}(0))$ with a reflecting upper barrier at $p\kappa$ if $\kappa < \infty$ and an inaccessible upper boundary at infinity if $\kappa = \infty$. The diffusion process $\mathbf{Q}^p$ has infinitesimal mean (*drift function*)

$$m(x) = \begin{cases} -p\mu\beta, & 0 \leqslant x < p\kappa, \\ -p\mu(x + \beta), & x < 0, \end{cases} \tag{2.11}$$

and infinitesimal variance (*diffusion function*)

$$\sigma^2(x) = 2p\mu z, \quad -\infty < x < p\kappa, \tag{2.12}$$

where

$$z = (p/2)(c_a^2 + (2/p) - 1) = \frac{p(c_a^2 + c_s^2)}{2} = 1 + \frac{p(c_a^2 - 1)}{2}. \tag{2.13}$$

As elaborated upon in Remark 2.2 of Whitt (2005), the limit process $\mathbf{Q}$ is not itself a diffusion process, but it is relatively tractable. In particular, it is easy to obtain the steady-state distribution of $\mathbf{Q}$, as we show in the next section.

In Whitt (2005) we also obtain the associated heavy-traffic limit for the scaled version of the discrete-time queue-length process at arrival epochs. The limit process $\mathbf{Q}^a$ is a time-scaled version of the limit process $\mathbf{Q}$ above, i.e., $\mathbf{Q}^a(t) = \mathbf{Q}(t/\mu)$, so that the steady-state distribution of the two limit processes are identical. Thus, in the heavy-traffic limit, the probability that all servers are busy at an arbitrary time is asymptotically equivalent to the asymptotic delay probability (the probability that an arrival must wait before being served).

Remark 2.1 (The Character of a Heavy-Tailed Distribution). To show that an H_2^* service-time distribution has some of the character of a heavy-tailed service-time distribution when the parameter p is small, we compare the impact on the queue-length process caused by an H_2^* service-time distribution with the impact caused by a Pareto service-time distribution. The Pareto distribution we consider has the complementary CDF

$$G^c(t) \equiv (1 + t/(p-1))^{-p}, \quad t \geqslant 0 \tag{2.14}$$

for $p > 1$, which is scaled to have (finite) mean 1. This Pareto distribution, denoted by $\mathrm{Par}(p)$, has finite variance if and only if $p > 2$. We consider the specific case $p = 3/2$, yielding finite mean but infinite variance.

Even though the variance of $\mathrm{Par}(3/2)$ is infinite, the variability parameter $\eta(G)$ in (1.7) is finite; in particular,

$$\eta(\mathrm{Par}(p)) = \int_0^\infty (1 + t/(p-1))^{-2p}\, dt = \frac{p-1}{2p-1}, \tag{2.15}$$

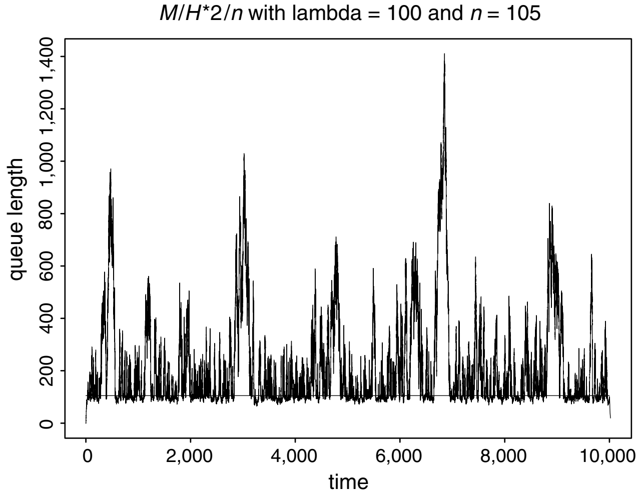
so that $\eta(\mathrm{Par}(3/2)) = 1/4$, whereas

$$\eta(H_2^*(p)) = \int_0^\infty p^2 e^{-2pt}\, dt = \frac{p}{2}, \tag{2.16}$$

so that $\eta(H_2^*(0.1)) = 1/20$. Of course, with Poisson arrivals, $c_a^2 = 1$ so that $z = 1$ in both cases for z in (1.6).

To show the impact of these two service-time distributions upon performance, we plot sample paths of the queue-length process for the first 10^6 arrivals in the models $M/H_2^*(0.1)/n/\infty$ and $M/\mathrm{Par}(3/2)/n/\infty$ with $\lambda = 100$, $\mu = 1$, and $n = 105$ in Figures 1 and 2. The plots are clearly quite similar. In both cases, the excursions above $n = 105$ are substantially greater than in the case of $M/M/n/\infty$, as can be seen from Figure 3. However, as predicted by approximation (1.5), the delay probabilities are quite close in these three examples. We elaborate on this point in §6; e.g., see Table 4 and Figure 4.

Figure 1. A sample path of the queue-length process for 10^6 arrivals in the $M/H_2^*/105$ queue with arrival rate $\lambda = 100$, service rate $\mu = 1$, and parameter $p = 0.1$.



3. The Steady-State Distribution of the Limit Process

From Equations (2.9) and (2.10), we see that we can obtain the steady-state random variable $\mathbf{Q}(\infty)$ associated with the limit process $\mathbf{Q}$ directly from the steady-state random variable $\mathbf{Q}^p(\infty)$ associated with the diffusion process $\mathbf{Q}^p$. In particular,

$$\mathbf{Q}(\infty) = h(\mathbf{Q}^p(\infty)) \equiv \begin{cases} \mathbf{Q}^p(\infty), & \mathbf{Q}^p(\infty) < 0, \\ \mathbf{Q}^p(\infty)/p, & 0 \leqslant \mathbf{Q}^p(\infty) \leqslant p\kappa. \end{cases} \tag{3.1}$$

From the form of the infinitesimal parameters in (2.11) and (2.12), we recognize that the diffusion process $\mathbf{Q}^p$ in

Figure 2. A sample path of the queue-length process for 10^6 arrivals in the $M/\mathrm{Par}(3/2)/105$ queue with arrival rate $\lambda = 100$ and service rate $\mu = 1$.

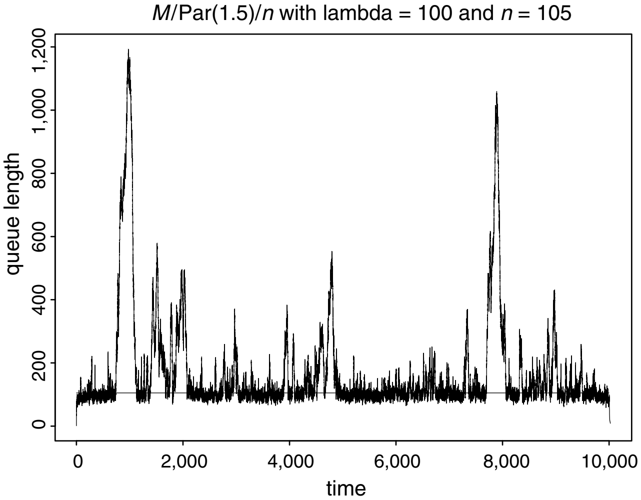
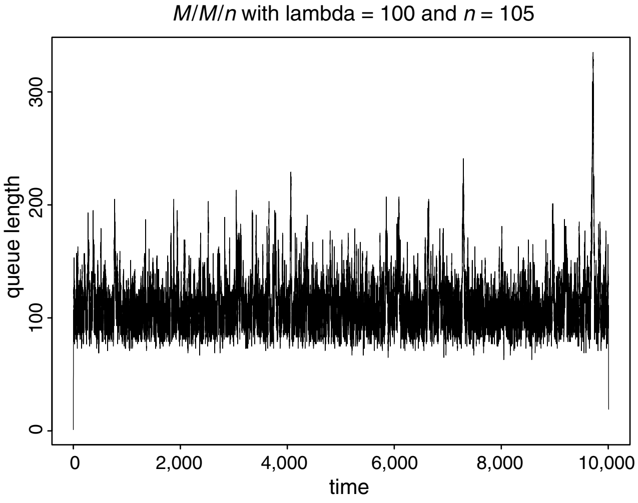


Figure 3. A sample path of the queue-length process for 10^6 arrivals in the $M/M/105$ queue with arrival rate $\lambda = 100$ and service rate $\mu = 1$.



Theorem 2.1 is a piecewise-linear diffusion, as in Browne and Whitt (1995). Thus, we can immediately write down the limiting steady-state distribution of $\mathbf{Q}^p$ when it exists. It is easy to see that $\mathbf{Q}^p(t) \Rightarrow \mathbf{Q}^p(\infty)$ for a proper random variable $\mathbf{Q}^p(\infty)$ if and only if either $\kappa < \infty$ or $\kappa = \infty$ and $\beta > 0$.

Because we are allowing a finite waiting room, we need to generalize the asymptotic-delay-probability function α in (0.2). We now let α be the following function of the two variables β and κ obtained from the limits in (0.1) and (0.3):

$$\alpha \equiv \alpha(\beta, \kappa) \equiv [1 + \beta\Phi(\beta)/\phi(\beta)(1 - e^{-\kappa\beta})]^{-1} \quad \text{for } \beta \neq 0. \tag{3.2}$$

The previous function in (0.2) appears as $\alpha_\infty \equiv \alpha_\infty(\beta) \equiv \alpha(\beta, \infty)$. When $\kappa < \infty$, we can allow $\beta \leqslant 0$. For $\beta = 0$, we let

$$\alpha_0 \equiv \alpha_0(\kappa) \equiv [1 + \kappa^{-1}\sqrt{\pi/2}]^{-1}. \tag{3.3}$$

We state the result as a theorem; see Browne and Whitt (1995) for a proof (drawing on basic diffusion-process theory). The idea is that the piecewise-linear structure implies that the distribution of $\mathbf{Q}^p(\infty)$ must be a truncated normal for $x < 0$ and a truncated exponential for $x > 0$ (or uniform in the case $\kappa < \infty$ and $\beta = 0$). The weight on the exponential component, which is just α, is determined by requiring that the two densities be continuous at 0; see (18.26) of Browne and Whitt (1995).

Theorem 3.1 (The Steady-State Distribution of the Limit Process). *Let $\mathbf{Q}^p$ be the diffusion process with infinitesimal parameters in* (2.11) *and* (2.12) *and let $\mathbf{Q}$ be the limit process defined in* (2.9). *Let β and κ be the limits in* (0.1) *and* (0.3). *Suppose that either $\kappa < \infty$ or $\kappa = \infty$ and $\beta > 0$, so that $\mathbf{Q}^p(t) \Rightarrow \mathbf{Q}^p(\infty)$ and $\mathbf{Q}(t) \Rightarrow \mathbf{Q}(\infty)$ as $t \to \infty$, where $\mathbf{Q}^p(\infty)$ and $\mathbf{Q}(\infty)$ are proper random variables.*

If $\beta \neq 0$, then

$$\begin{aligned} P(\mathbf{Q}(\infty) \geqslant 0) &= P(\mathbf{Q}^p(\infty) \geqslant 0) \\ &= \alpha(\beta/\sqrt{z}, p\kappa/\sqrt{z}), \end{aligned} \tag{3.4}$$

$$\begin{aligned} &P(\mathbf{Q}(\infty) \leqslant x \mid \mathbf{Q}(\infty) \leqslant 0) \\ &\quad = P(\mathbf{Q}^p(\infty) \leqslant x \mid \mathbf{Q}^p(\infty) \leqslant 0) \\ &\quad = \Phi((x+\beta)/\sqrt{z})/\Phi(\beta/\sqrt{z}), \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} P(\mathbf{Q}(\infty) > x \mid \mathbf{Q}(\infty) \geqslant 0) &= P(\mathbf{Q}^p(\infty) > px \mid \mathbf{Q}(\infty) \geqslant 0) \\ &= \frac{e^{-px\beta/z} - e^{-p\kappa\beta/z}}{1 - e^{-p\kappa\beta/z}}, \quad 0 \leqslant x < \kappa, \\ &= \frac{e^{-x\beta/v} - e^{-\kappa\beta/v}}{1 - e^{-\kappa\beta/v}}, \quad 0 \leqslant x < \kappa, \end{aligned} \tag{3.6}$$

where α is the $M/M/n/m$ asymptotic-delay-probability function in (3.2), *z is in* (2.13), *and*

$$v \equiv \frac{z}{p} = \frac{c_a^2 + c_s^2}{2}. \tag{3.7}$$

If $\beta = 0$, then

$$P(\mathbf{Q}(\infty) \geqslant 0) = P(\mathbf{Q}^p(\infty) \geqslant 0) = \alpha_0(p\kappa/\sqrt{z}) \tag{3.8}$$

for α_0 in (3.3). *Then,*

$$\begin{aligned} P(\mathbf{Q}(\infty) > x \mid \mathbf{Q}(\infty) \geqslant 0) &= P(\mathbf{Q}^p(\infty) > px \mid \mathbf{Q}^p(\infty) \geqslant 0) \\ &= (\kappa - x)/\kappa, \quad 0 \leqslant x < \kappa, \end{aligned} \tag{3.9}$$

while formula (3.5) *remains unchanged.*

Corollary 3.1 (The Infinite-Waiting-Room Case). *If, in addition to the conditions of Theorem* 3.1, *$\kappa = \infty$ and $\beta > 0$, then*

$$\begin{aligned} P(\mathbf{Q}(\infty) > 0) &= P(\mathbf{Q}^p(\infty) > 0) \\ &= \alpha_\infty(\beta/\sqrt{z}) \equiv \alpha(\beta/\sqrt{z}, \infty) \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} P(\mathbf{Q}(\infty) \leqslant x \mid \mathbf{Q}(\infty) \leqslant 0) &= P(\mathbf{Q}^p(\infty) \leqslant x \mid \mathbf{Q}^p(\infty) \leqslant 0) \\ &= \Phi((x+\beta)/\sqrt{z})/\Phi(\beta/\sqrt{z}) \end{aligned} \tag{3.11}$$

for β in (0.1) *and z in* (2.13), *implying that both formulas depend on the parameter p only through the parameter z in* (2.13). *Moreover,*

$$P(\mathbf{Q}(\infty) > x \mid \mathbf{Q}(\infty) > 0) = e^{-\beta x/v} \tag{3.12}$$

for v in (3.7), *so that*

$$E[\mathbf{Q}(\infty)^+] = \alpha_\infty(\beta/\sqrt{z})\frac{v}{\beta}, \tag{3.13}$$

implying that both formulas depend on the parameter p only though the parameters z in (2.13) *and v in* (3.7).

Remark 3.1 (The pdf). The steady-state distribution of the diffusion process $\mathbf{Q}$ can also be characterized by its PDF. If $\beta \neq 0$, $\mathbf{Q}(\infty)$ has the PDF

$$f(x) = \begin{cases} (1-\alpha)\phi((x+\beta)/\sqrt{z})/\sqrt{z}\Phi(\beta/\sqrt{z}), & x < 0, \\ \alpha(p\beta/z)e^{-px\beta/z}(1-e^{-p\kappa\beta/z}), & 0 \leqslant x \leqslant \kappa \end{cases} \tag{3.14}$$

for $\alpha \equiv \alpha(\beta/\sqrt{z}, p\kappa/\sqrt{z})$ in (3.4) and z in (2.13). If $\beta = 0$, then $\mathbf{Q}(\infty)$ has the PDF

$$f(x) = \begin{cases} (1-\alpha_0)\phi((x+\beta)/\sqrt{z})/\sqrt{z}\Phi(\beta/\sqrt{z}), & x < 0, \\ \alpha_0/\kappa, \quad 0 \leqslant x \leqslant \kappa \end{cases} \tag{3.15}$$

for $\alpha_0 \equiv \alpha_0(p\kappa/\sqrt{z})$ in (3.8).

Remark 3.2 (Understanding the Asymptotic-Delay-Probability Formula). The asymptotic-delay-probability functions in (3.2), (3.4), and (3.10) can be understood by observing an underlying alternating-renewal-process structure. The queue-length process alternates between periods spent above level n (an "above" time X_n^a) and periods spent below level $n-1$ (a "below" time X_n^b). This structure is most straightforward in the case $M/M/n/\infty$, so consider that case. Because $Q_n(t)$ is a Markov process, these times are mutually independent. Thus, by a well-known alternating-renewal-process result,

$$P(Q_n(\infty) > n) = \frac{EX_n^a}{EX_n^a + EX_n^b} = [1 + (EX_n^b/EX_n^a)]^{-1}. \tag{3.16}$$

With the scaling in (0.1), where the arrival rate and service rate are both of order $O(n)$, both EX_n^a and EX_n^b are of order $O(1/\sqrt{n})$. For the $M/M/n/\infty$ model, X_n^a is distributed as the busy period in an $M/M/1/\infty$ model with service rate $n\mu$, so that

$$EX_n^a = \frac{1}{n\mu(1-\rho)} \sim \frac{1}{\sqrt{n}\mu\beta} \quad \text{as } n \to \infty, \tag{3.17}$$

where $\sim$ means the ratio of the two sides converges to 1 as $n \to \infty$, while EX_n^b is the reciprocal of the blocking probability, say π_n, in a $M/M/n-1/0$ model, divided by the arrival rate λ_n. We see where the ratio $\phi(x)/\Phi(x)$ comes from by recalling that

$$1/\lambda_n EX_n^b = \pi_n \sim (1/\sqrt{n})\phi(\beta)/\Phi(\beta) \quad \text{as } n \to \infty \tag{3.18}$$

under condition (0.1); e.g., see (15) of Srikant and Whitt (1996) and the appendix of Whitt (1984a). Combining (3.16)–(3.18), we obtain convergence to α in (0.2) in the limiting regime (0.1).

A similar argument applies to the $M/H_2^*/n/\infty$ model, as shown on page 207 of Whitt (1983). The mean EX_n^b for $M/M/n/\infty$ above is divided by p in the $M/H_2^*/n/\infty$ model, because when all servers are not busy in the $M/H_2^*/n/\infty$ model, we can ignore all customers with zero service times. Therefore, the queue-length process in that region is a birth-and-death process with birth rate $p\lambda_n$ and death rate $p\mu n$, giving the $M/M/n/\infty$ formula for EX_n^b above divided by p. Similarly, the mean EX_n^a for $M/M/n/\infty$ above is also divided by p in the $M/H_2^*/n/\infty$ case, but that is less obvious. The $M/H_2^*/n/\infty$ model behaves like an $M/H_2^*/1/\infty$ model when all servers are busy, where each service time is an exponential with mean $1/np\mu$ with probability p and is 0 with probability $1-p$. In the $M/H_2^*/1/\infty$ model, the mean busy period is $1/n\mu(1-\rho_n)$, just as in $M/M/1/\infty$. However, we must divide by p because, when we calculate the first passage time from state n to state $n-1$, we need to condition on the first service time not being 0. That produces the division by p. Thus, the overall analysis shows that the probability of delay in the $M/H_2^*/n/\infty$ model is independent of p for all n.

Remark 3.3 (The Limit as $p \to 0$ for H_2^*). Intuitively, the H_2^* distributions acquire more of the character of heavy-tailed distributions as p becomes very small. Thus, it is interesting to observe how the steady-state distribution of the limit process $\mathbf{Q}$ behaves as $p \downarrow 0$ with the mean of the service-time distribution held fixed. Thus, we index quantities of interest by p here. We only consider the case in which $\beta > 0$.

First, if $p \downarrow 0$, then $z_p \to 1$ from (1.3). Second, from (3.7), if $p \downarrow 0$, then $v_p \to \infty$ and $pv_p \to 1$. Thus, all formulas that depend on p only through z approach the case of exponential service times (as if $p = 1$). For example, the infinite-waiting-room formulas (3.10) and (3.11) in Corollary 3.1 change only by having $z \to 1$.

Thus, the $G/H_2^*/n/\infty$ asymptotic delay probability approaches the $M/M/n/\infty$ asymptotic delay probability as $p \downarrow 0$ for *any* stationary arrival process. We emphasize that this asymptotic property of the $G/H_2^*/n/\infty$ model is exact. It may seem surprising that the $G/H_2^*/n/\infty$ asymptotic delay probability approaches the $M/M/n/\infty$ value as $p \downarrow 0$ for any stationary arrival process satisfying a FCLT, so we offer an intuitive explanation. First, when all servers are not busy, we can act as if arrivals with zero service times never occur, because they leave immediately upon arrival. Thus, the interarrival time of customers with positive service times is a geometric random sum of the initial interarrival times. There is an asymptotically increasing number of interarrival times in this geometric random sum. As $p \downarrow 0$, the mean of this geometric random variable increases, causing the successive interarrival times to become independent. Moreover, the properly scaled geometric random variable converges to an exponential random variable. Second, we have just noted in Remark 3.2 that the delay probability in the $M/H_2^*/n/\infty$ model is independent of the parameter p.

In contrast, the expected queue length given that all servers are busy when $\kappa = \infty$ tends to behave very differently: $E[\mathbf{Q}_p(\infty)^+] \to \infty$ as $p \to 0$. More precisely,

$$E[\mathbf{Q}_p(\infty)^+] \sim \alpha(\beta)/p\beta \quad \text{as } p \to 0. \tag{3.19}$$

It is also interesting to consider the case $\kappa < \infty$. Then, referring to (3.4), we see that $\alpha_p \to 0$. More precisely,

$$\alpha(\beta/\sqrt{z}, p\kappa/\sqrt{z}) \sim \alpha(\beta, p\kappa) \sim \frac{p\kappa\phi(\beta)}{\Phi(\beta)} \quad \text{as } p \to 0. \tag{3.20}$$

These asymptotic relations produce effects we should anticipate with heavy-tailed distributions.

Another interesting case is the $H_2^*/H_2^*/n/\infty$ model in which the interarrival-time and service-time H_2^* distributions have a common parameter p. Then, because $c_a^2 = (2/p) - 1$ and $\eta(H_2^*(p)) = p/2$, we obtain the exact asymptotic result

$$z_p = 2 - p \to 2 \quad \text{as } p \downarrow 0. \tag{3.21}$$

This limiting behavior can be verified by simulation, but it is difficult for very small p because the overall variability increases, causing the reliability of simulation estimates for given run length to decrease as p decreases.

4. The Heuristic Approximation for $G/GI/n/m$

We now seek an approximation for the queue-length process and its steady-state distribution in the general $G/GI/n/m$ model, with general (*GI*) service times. From the stochastic-process limits for the $GI/PH/n/\infty$ model in Puhalskii and Reiman (2000), we know that the scaled queue-length process again converges to a nondegenerate limit, but the limit for the scaled queue-length process is relatively complicated. In those cases the limit can be expressed in terms of a complicated multidimensional diffusion process, where the dimension of the diffusion is the number of phases in the phase-type service-time distribution. To generate more tractable approximations, here we develop a heuristic one-dimensional approximation with convenient explicit formulas for all steady-state performance measures of interest. Even though our approximation is not asymptotically correct, we rely heavily on insights from heavy-traffic stochastic-process limits. For background on heuristic diffusion approximations, see Newell (1973), Halachmi and Franta (1978), Whitt (1984a), and Kimura (1995, 2000, 2002).

Our starting point is an assumed limit for the scaled queue-length process. We assume that our process of interest is the queue-length process $Q_n(t)$, where the number n of servers is suitably large. (We are thinking of $n = 100$, but the approximation may be good for much smaller n, e.g., $n = 10$.) Consequently, the first step of our approximation is

$$Q_n(t) \approx n + \sqrt{n}\mathbf{Q}(t), \tag{4.1}$$

where $\mathbf{Q}$ is a stochastic process for which we need to develop an approximation.

As an approximation for the stochastic process $\mathbf{Q}$ in (4.1), we use the *same* process $\mathbf{Q}$, in Theorem 2.1, where $\mathbf{Q}(t) = h(\mathbf{Q}^p(t))$ for h in (2.10), but we choose appropriate parameters for that process as a function of the more general service-time distribution. We choose the parameters so that the new approximation is consistent with the approximation for the $G/H_2^*/n/m$ model following from Theorems 2.1 and 3.1.

As can be seen from Theorem 2.1, the stochastic processes $\mathbf{Q}^p$ and $\mathbf{Q}$ depend on five parameters: μ, β, κ, z, and p. (We can substitute $v \equiv z/p$ for one of z or p.) As before, we let $1/\mu$ be the mean service time and let β and κ be determined by the limits (0.1) and (0.3). We generate an approximation for the queue-length process in the $G/GI/n/m$ model by choosing appropriate values for the two remaining parameters z and v.

In the heavy-traffic limits for the $G/H_2^*/n/m$ and $G/PH/n/m$ models, the arrival process influences the asymptotic behavior only through the arrival rate and the scaling parameter c_a^2 appearing in the assumed FCLT. Thus, it is natural to let the two parameters z and v depend on the arrival process only through c_a^2. We require that of our approximation.

We are relatively confident about our proposed approximation for the parameter z. To choose z, we focus on the behavior of the process when $x < 0$ (when all servers are not busy). To do so, we focus on the conditional distribution $P(\mathbf{Q}(\infty) \leqslant x \mid \mathbf{Q}(\infty) < 0)$ in (3.5), which is conditional normal distribution, where the parameter z plays the role of the variance. We base our approximation on the associated heavy-traffic limit for the general $G/GI/\infty$ infinite-server model; see §10.3 of Whitt (2002) and references cited there, notably Borovkov (1984). The limit process with infinitely many servers has a normal steady-state distribution. We let z be the ratio of the variance to the mean of that steady-state normal distribution. That is the asymptotic peakedness in (1.6). It is consistent with the exact formula for z in (2.13) in the $G/H_2^*/n/m$ special case.

Having chosen an approximation for the parameter z, it remains to specify an approximation for the remaining parameter v in (3.7). To generate an approximation for v, we focus on the behavior of the process when $x > 0$, but the approximation is more challenging when $x > 0$. When the service-time distribution is M or H_2^*, the queue behaves exactly like a single-server queue when all servers are busy. However, for other service-time distributions, the elapsed service times of the customers in service play an important role and the situation is more complicated. (That complexity is captured by the limit in Puhalskii and Reiman 2000.)

Nevertheless, we exploit the single-server view. Thus, on the interval $[0, \kappa]$, we let the diffusion process act as a reflecting Brownian motion with constant drift. We specify the (constant) infinitesimal variance by looking at the "unreflected free process," which is a scaled version of the arrival counting process minus the departure counting process. The arrival process is straightforward, but the departure process is quite complicated. In fact, even though the service times are assumed to be independent of the arrival process, the departure process is actually dependent on the arrival process. However, in our approximation we will act as if they are independent.

To generate an initial approximation, we act as if all n servers are busy all the time. That is at least temporarily true when $x > 0$. Under that assumption, the departure process would be the superposition of n IID service-time counting processes. For any fixed n, that superposition process obeys a FCLT with scaling constant c_s^2, where c_s^2 is the SCV of a service time, here assumed to be finite; see §9.4 of Whitt (2002). That perspective leads to a diffusion approximation for $\mathbf{Q}$ in the region $x > 0$ with infinitesimal parameters just as in Remark 2.3 of Whitt (2005). In particular, the infinitesimal mean (drift function) is

$$m_{\mathbf{Q}}(x) = -\mu\beta \quad \text{for } x > 0 \tag{4.2}$$

and the infinitesimal variance (diffusion function) is

$$\sigma_{\mathbf{Q}}^2(x) = \mu(c_a^2 + c_s^2) \quad \text{for } x > 0. \tag{4.3}$$

In (4.3), c_a^2 is the arrival-process variability parameter obtained from the FCLT for the arrival process, as in (2.1) and (2.2), and c_s^2 is the service-time SCV.

This reasoning leads to a truncated exponential distribution for $P(\mathbf{Q}(\infty) \geqslant x \mid \mathbf{Q}(\infty) > 0)$, just as in (3.6), with the parameter v having exactly the same form as in (3.7), with c_s^2 now referring to the SCV of the general service-time distribution. Clearly, this approximation is also consistent with Theorems 2.1 and 3.1 in the $G/H_2^*/n/m$ special case.

It is significant that the proposed approximations for the parameters z and v depend on the service-time distribution in different ways in the general *GI* case. (That approach was also used in Whitt 1992.) If we do use approximation (1.6) for z and approximation (3.7) for v, then we immediately obtain the associated approximation for p:

$$p = \frac{z}{v} = \frac{2z}{c_a^2 + c_s^2}, \tag{4.4}$$

where z is given by (1.6), c_a^2 is the scaling constant in the FCLT, as in (2.1) and (2.2), and c_s^2 is the SCV of the service-time distribution. Note that this associated approximation for p in (4.4) could yield $p > 1$, which is of course inconsistent with the original H_2^* model definition, but that presents no problems for the stochastic process $\mathbf{Q}$ and its steady-state distribution.

We find that the approximation for v in (3.7) works quite well for low-to-moderate variability service times, but it can seriously break down more generally (e.g., see Table 5). Thus, we want to consider refinements. Another perspective is that a superposition of n IID renewal processes converges to a Poisson process as $n \to \infty$ when the component processes are rescaled to keep the total rate fixed; see Theorem 9.8.1 of Whitt (2002). Naturally, this second perspective leads us to the approximation in (3.7) with c_s^2 replaced by one. This second perspective is even supported by stochastic-process limits for the departure process from multiserver queues; see Whitt (1984c). These two perspectives are not inconsistent, because they describe the superposition process in different time scales; see Remark 9.8.1 of Whitt (2002). The superposition process behaves like a Poisson process in a short time scale, but like a single component renewal process in a long time scale.

These two perspectives lead to a compromise approximation that is a convex combination of the first two approximations, i.e.,

$$v = \frac{(c_a^2 + wc_s^2 + 1 - w)}{2}, \tag{4.5}$$

where w is an appropriate weight with $0 \leqslant w \leqslant 1$. To develop a candidate weight function w, we observe that there is a third perspective, which has already proved useful to study superposition arrival processes to queues. In the third perspective, we apply the central limit theorem for stochastic processes to the sum of n IID renewal processes; see Theorems 7.2.3 and 7.2.4 of Whitt (2002). The third perspective leads to approximating the departure process by a non-Brownian Gaussian process. The third perspective also leads to an associated FCLT in which the number of component processes in the superposition increases along with the time-and-space scaling; see §9.8 of Whitt (2002). For superposition arrival processes to queues, there is a stochastic-process limit in the limiting regime (0.1) we are considering, where n is understood to be the number of component arrival processes instead of the number of servers; see Theorem 9.8.3 of Whitt (2002). That perspective might be relevant here, because if we reverse time, the departure process behaves something like a superposition arrival process.

The analysis of superposition arrival processes leads to approximations of the form (4.5), where the weight w is a strictly decreasing function of $\beta = \sqrt{n}(1 - \rho)$ with $w(0) = 1$ and $w(\infty) = 0$. A specific function based on simulation experiments by Albin (1982, 1984) is

$$w \equiv w(\beta) = [1 + 4\beta^2]^{-1} \tag{4.6}$$

for $\beta = \sqrt{n}(1-\rho)$; see page 333 of Whitt (2002). However, we do not find a direct application of (4.6) to be effective.

However, the related experience with superposition arrival processes can provide important insights. For example, the stochastic-process limit for superposition arrival processes in regime (0.1)—Theorem 9.8.3 of Whitt (2002)—does *not* require that the interrenewal times in the component renewal processes have finite second moment. Thus, we can anticipate (what turns out to be the case in our setting) that the same scaling works for multiserver queues with Pareto service times having finite mean but infinite variance. Hence, we should allow the variability parameter v to be well defined when c_s^2 is infinite.

In summary, this analysis leads us to approximate z by the asymptotic peakedness in (1.6), but to only propose a tentative approximation for the variability parameter v. Our tentative specification of v in the case of a finite service-time SCV c_s^2 is

$$v=\frac{c_a^2+wc_s^2+1-w}{2} \tag{4.7}$$

for some weight function w, which is a decreasing function of β with $w(0)=1$ and $w(\infty)=0$. We primarily apply the initial approximation in (3.7); i.e., (4.7) with $w\equiv 1$, but we find situations in which alternatives in (4.7) can be important. Our somewhat vague specification allows room for refinement.

The final situation is less unsatisfactory than it may appear, because in Corollary 3.1 we have shown for the case $\kappa=\infty$ that the steady-state probability of being greater than 0 (which corresponds to the delay probability in the queueing model) actually depends on the parameters v and p only through the parameter z. Hence, for the delay probability, we only need z.

Remark 4.1 (Rough Approximations of the Asymptotic Peakedness). We can obtain further rough approximations of the peakedness z in terms of the variability parameters c_a^2 and c_s^2 to use in the heuristic diffusion approximation by approximating the asymptotic peakedness z. However, we advise caution: From the formula for the asymptotic peakedness z in (1.6), we see that the service-time distribution beyond the mean should have relatively little impact upon z when c_a^2 is near 1. However, when c_a^2 is not near 1, the service-time distribution beyond its mean can have a big impact on z, and is quantified by $\eta(G)$ in (1.7), not by the SCV c_s^2. Nevertheless, the following formulas are useful to obtain a quick picture of the impact of service-time variability upon performance. They show that $\eta(G)$ tends to decrease as the service-time distribution gets more variable with a fixed mean.

Because $\eta(G)=1$ when the service-time distribution is deterministic and $\eta(G)=1/2$ when the service-time distribution is exponential, we propose the following linear interpolation as an approximation for SCVs in between:

$$\eta(c_s^2)\approx 1-(c_s^2/2) \quad \text{and}$$
$$z(c_a^2,c_s^2)\approx 1+(c_a^2-1)(1-(c_s^2/2)),\quad 0\leqslant c_s^2\leqslant 1. \tag{4.8}$$

To treat distributions with $c_s^2\geqslant 1$, we can use H_2 distributions with balanced means (H_2^b). An H_2 distribution with mean $1/\mu$ has PDF

$$h(x)=p_1e^{-\mu_1 x}+p_2e^{-\mu_2 x},\quad x\geqslant 0, \tag{4.9}$$

where $0\leqslant p_1\leqslant 1$, $p_1+p_2=1$, and $(p_1/\mu_1)+(p_2/\mu_2)=1/\mu$. The H_2^b PDF has balanced means; i.e., one of the two remaining parameters is determined by the relation

$$\frac{2p_1}{\mu_1}=\frac{2p_2}{\mu_2}=\frac{1}{\mu}, \tag{4.10}$$

which implies that

$$p_i=\Big[1\pm\sqrt{(c_s^2-1)/c_s^2+1)}\Big]/2. \tag{4.11}$$

For this H_2^b case, $\eta(H_2^b)=(c_s^2+3)/4(c_s^2+1)$, so that we obtain the general approximation

$$z(c_a^2,c_s^2)\approx z(c_a^2,H_2^b)=1+\frac{(c_a^2-1)(c_s^2+3)}{4(c_s^2+1)} \quad \text{for } c_s^2\geqslant 1. \tag{4.12}$$

Note that $\eta(H_2^b)$ increases to $1/2$ as c_s^2 decreases to its lower limit $c_s^2=1$, which is the exponential distribution, while $\eta(H_2^b)$ decreases to $1/4$ as $c_s^2\uparrow\infty$. Other H_2 distributions without balanced means can have arbitrarily small values of η, as we saw for H_2^* in (2.16).

5. Evaluating the $GI/GI/n/\infty$ Approximations

We start by evaluating the approximations for the delay probability (PW) and the probability all servers are busy ($PB\equiv P(Q_n(\infty)\geqslant n)$) in the $GI/M/n/\infty$ model. By the Poisson-Arrivals-See-Time-Averages (PASTA) property, these quantities PW and PB coincide when the arrival process is Poisson, but they do not otherwise. However, the heavy-traffic limits imply that the (P)ASTA property holds in that heavy-traffic limit for non-Poisson arrival processes. So the asymptotic delay probability generates asymptotically correct approximations for both PW and PB. The extent to which PW and PB differ gives an indication of the degree of accuracy possible for the approximation.

Because we are working in the asymptotic regime (0.1), the natural approximation based on (1.2) is $P(Q_n(\infty)\geqslant n)\approx\alpha(\beta/\sqrt{z})$ for α in (0.2), $z=(c_a^2+1)/2$, and

$$\beta=\sqrt{n}(1-\rho). \tag{5.1}$$

Indeed, we have been implicitly acting as if the value for β based on the limit in (0.1) is (5.1), and that is what we usually use. However, we might approximate β differently. As discussed in Whitt (1992), because it is the offered load that is random rather than the number of servers, it is natural to think of

$$(\lambda/\mu)+\beta(\sqrt{\lambda/\mu})\approx n \tag{5.2}$$

rather than (5.1). Approximation (5.2) leads to the alternative approximation for β,

$$\beta\approx\frac{\sqrt{n}(1-\rho_n)}{\sqrt{\rho_n}}. \tag{5.3}$$

Of course, in the limiting regime (0.1), the two specifications for β in (5.1) and (5.3) are asymptotically equivalent. It can be helpful to compute both, because their difference gives an indication of the likely precision.

For our numerical comparisons, we use exact results from the tables in Seelen et al. (1985). The results are displayed in Table 1. The approximation reduces to (0.2) when the arrival process is Poisson. In that case, it is well known that the approximation for $PW = PB$ performs quite well; e.g., see Table 13 of Whitt (1993). We see that again for the entries in which $c_a^2 = 1$ in Table 1.

As approximations for PW and PB in Table 1, we plot the approximation $\alpha(\beta/\sqrt{z})$ in (1.1) based on both the standard specification of β in (5.1) and the alternative in (5.3). The modification in (5.3) always increases β and thus reduces $\alpha(\beta/\sqrt{z})$. As indicated above, the two values together give a good indication of the accuracy. In many cases (but not all), they bracket the exact values.

Table 1. A comparison of the $GI/M/n/\infty$ approximations in (1.1) and (5.3) with exact values of the probability of delay (PW) and the probability all servers are busy (PB) for various values of n, ρ, and interarrival-time SCV c_a^2.

			Delay probability				Mean number waiting	
Parameters			Exact values		Approximations		Exact	Approximations
n	ρ	c_a^2	PW	PB	(1.1)	(5.3)	$E[(Q(\infty)-n)^+]$	(5.7) and (5.6)
200	0.98	4.00	0.803	0.794	0.794	0.792	97.2	97.3
		1.00	0.692	0.692	0.689	0.686	33.9	33.8
		0.25	0.619	0.627	0.620	0.617	19.0	19.0
	0.92	4.00	0.367	0.349	0.361	0.344	10.0	10.4
		1.00	0.170	0.170	0.176	0.161	1.97	2.02
		0.25	0.090	0.094	0.098	0.086	0.66	0.70
	0.88	4.00	0.191	0.177	0.196	0.171	3.23	3.59
		1.00	0.049	0.049	0.055	0.043	0.36	0.40
		0.25	0.0142	0.0154	0.0185	0.0127	0.067	0.085
100	0.98	4.00	0.861	0.851	0.850	0.849	104.2	104.1
		1.00	0.775	0.775	0.771	0.769	38.0	37.8
		0.25	0.717	0.726	0.718	0.715	22.1	22.0
	0.92	4.00	0.515	0.490	0.500	0.484	14.1	14.4
		1.00	0.312	0.312	0.314	0.297	3.58	3.61
		0.25	0.208	0.219	0.218	0.202	1.52	1.57
	0.86	4.00	0.272	0.249	0.273	0.242	3.80	4.19
		1.00	0.094	0.094	0.104	0.083	0.58	0.64
		0.25	0.037	0.041	0.047	0.033	0.15	0.18
25	0.90	4.00	0.690	0.648	0.658	0.642	14.5	14.8
		2.00	0.593	0.574	0.577	0.559	7.74	7.79
		1.00	0.508	0.508	0.504	0.485	4.57	4.54
		0.50	0.442	0.458	0.449	0.428	3.02	3.03
		0.10	0.367	0.402	0.386	0.364	1.87	1.91
	0.70	4.00	0.225	0.186	0.244	0.175	1.03	1.41
		1.00	0.064	0.064	0.085	0.044	0.15	0.20
		0.10	0.0116	0.0161	0.0254	0.0089	0.016	0.033
8	0.98	4.00	0.9670	0.9554	0.9559	0.9554	117.0	117.1
		1.00	0.9361	0.9361	0.9309	0.9302	45.9	45.6
		0.25	0.9132	0.9244	0.9132	0.9124	28.1	28.0
	0.90	4.00	0.834	0.785	0.794	0.783	17.6	17.9
		1.00	0.702	0.702	0.689	0.675	6.31	6.20
		0.25	0.610	0.651	0.620	0.604	3.51	3.49
	0.70	4.00	0.502	0.415	0.478	0.406	2.31	2.79
		1.00	0.271	0.271	0.290	0.218	0.63	0.68
		0.25	0.150	0.190	0.196	0.132	0.24	0.29
2	0.80	4.00	0.854	0.754	0.794	0.771	7.40	7.94
		1.00	0.711	0.711	0.689	0.658	2.84	2.76
		0.25	0.594	0.685	0.620	0.584	1.55	1.55
	0.60	4.00	0.663	0.514	0.620	0.534	1.77	2.33
		1.00	0.450	0.450	0.457	0.353	0.68	0.69
		0.25	0.284	0.404	0.361	0.255	0.29	0.34

Notes. The peakedness is thus $z = (c_a^2 + 1)/2$. Also compared are the approximation for the mean number waiting in (5.7) and (5.6) with exact values. The exact values come from Seelen et al. (1985).

In Table 1 we also compare a heavy-traffic approximation for the mean number waiting, $E[(Q_n(\infty) - n)^+]$ (using (5.1) for β) with exact values . By Little's law, $L = \lambda W$, we obtain an associated approximation for the mean steady-state waiting time (before beginning service) $EW_n(\infty)$; i.e.,

$$EW_n(\infty) = E[(Q_n(\infty) - n)^+]/\lambda_n. \tag{5.4}$$

The direct heavy-traffic approximation based on (3.13) is

$$E[(Q_n(\infty) - n)^+] \approx \frac{\alpha\sqrt{n}v}{\beta} = \frac{\alpha v}{1-\rho} \tag{5.5}$$

for v in (4.7). When the service-time distribution is exponential,

$$v = (c_a^2 + 1)/2 \tag{5.6}$$

for v in (4.7) and *any* weight function w. Thus, consistent with the established limit in this case, we anticipate that the approximation should perform better for exponential service times.

To make the heavy-traffic approximation exact for the $M/M/n/\infty$ model for all ρ and still keep it asymptotically correct, we multiply the approximation in (5.5) by ρ to get

$$E[(Q_n(\infty) - n)^+] \approx \frac{\alpha\rho v}{1-\rho}. \tag{5.7}$$

By Little's law again, the expected steady-state number of busy servers is $\lambda/\mu = n\rho_n$. Hence, we can apply (5.7) to obtain the related approximation

$$EQ_n(\infty) \approx n\rho + \frac{\alpha\rho v}{1-\rho}. \tag{5.8}$$

We only evaluate the approximation in (5.7) because it is more challenging.

Approximations for general $GI/GI/n/\infty$ queues were studied in Whitt (1993), but unfortunately the error in the asymptotic-delay-probability limit in Halfin and Whitt (1981) was perpetuated in Whitt (1993). In formula (3.2) there, the Halfin-Whitt delay-probability approximation for $GI/M/n/\infty$ is given as $\alpha(\beta/z)$ instead of $\alpha(\beta/\sqrt{z})$. As should be anticipated, the quality of the approximation improves dramatically when this error is corrected. For example, the new approximation performs much better in Tables 15 and 16 in Halfin and Whitt (1981) for $D/M/n/\infty$ and $H_2^b/M/n/\infty$ queues.

We now evaluate the approximations for the delay probability in (1.5) and the mean number waiting in (5.7) for $GI/GI/n/\infty$ models with nonexponential service-time distributions. Now we are considering cases in which the diffusion approximation is *not* asymptotically correct in the heavy-traffic limit. For comparison, we again rely on tables in Seelen et al. (1985). The results appear in Table 2.

Table 2. A comparison of the $G/GI/n/\infty$ approximation in (1.5) and the $GI/M/n/\infty$ approximation in (1.2) (obtained by treating c_s^2 as 1) with exact values of the probability of delay (PW) and the probability all servers are busy (PB) in the $GI/GI/n/\infty$ model with nonexponential service-time distributions for $n = 25$, $\rho = 0.9$, and several values of the interarrival-time SCV c_a^2 and service-time SCV c_s^2.

Parameters			Exact values		Approximations		Exact	Approximation
c_a^2	c_s^2	z	PW	PB	(1.5)	$GI/M/n/\infty$	$E[(Q(\infty)-n)^+]$	(5.7)
0.1	0.0	0.10	0.058	0.103	0.071	0.386	0.037	0.032
0.5		0.50	0.336	0.360	0.366	0.449	0.89	0.82
1.0		1.00	0.479	0.479	0.505	0.505	2.41	2.27
2.0		2.00	0.615	0.593	0.624	0.577	5.67	5.62
4.0		4.00	0.730	0.688	0.721	0.658	12.9	13.0
0.1	0.5	0.44	0.308	0.349	0.338	0.386	0.94	0.92
0.5		0.69	0.415	0.434	0.431	0.449	2.00	1.94
1.0		1.00	0.500	0.500	0.505	0.505	3.51	3.41
2.0		1.62	0.601	0.580	0.591	0.577	6.66	6.65
4.0		2.87	0.706	0.663	0.677	0.658	13.5	13.7
0.1	2.5	0.65	0.420	0.447	0.419	0.386	4.39	4.90
0.5		0.80	0.472	0.485	0.462	0.449	5.66	6.24
1.0		1.00	0.522	0.522	0.505	0.505	7.32	7.95
2.0		1.39	0.594	0.574	0.565	0.577	10.8	11.4
4.0		2.18	0.680	0.637	0.637	0.658	18.0	18.6
0.1	4.0	0.69	0.441	0.465	0.431	0.386	6.89	7.95
0.5		0.83	0.485	0.498	0.468	0.449	8.23	9.48
1.0		1.00	0.529	0.529	0.505	0.505	9.95	11.4
2.0		1.35	0.592	0.575	0.559	0.577	13.3	15.1
4.0		2.05	0.672	0.633	0.628	0.658	20.2	22.6

Notes. Also evaluated is the approximation for the mean number waiting in (5.7) and (4.7) with $w \equiv 1$. The exact values come from Seelen et al. (1985).

For the delay probability, we compare the new $G/GI/n/\infty$ approximation and the $GI/M/n/\infty$ approximation (applied by ignoring the service-time SCV) to the exact values of PW and PB. Again, half the difference between PW and PB provides a lower bound on the worst error in the approximation for these two quantities. The new $G/GI/n/\infty$ approximation does quite well. In fact, the $GI/M/n/\infty$ approximation itself does remarkably well except when c_a^2 is small. The general $G/GI/n/\infty$ approximation does significantly better than the $G/M/n/\infty$ approximation when c_a^2 is small.

For the mean number waiting, we let the variability parameter v be as in (4.7) with weight $w \equiv 1$; i.e., here $v = (c_a^2 + c_s^2)/2$. The approximation slightly underestimates the exact values when $c_s^2 < 1$ and quite significantly overestimates the exact values when $c_s^2 > 1$. When $c_s^2 = 4.0$, the approximation is consistently about 14% too high. The approximation for the mean number waiting in the cases with $c_s^2 > 1$ become nearly exact if we use $w = 0.8$ in (4.7). The direct approximation (with $w = 1$) performs remarkably well when both $c_a^2 > 1$ and $c_s^2 < 1$. Overall, the approximations in Table 2 seem sufficiently accurate to be quite useful.

6. Simulations

To evaluate the approximations for more general $G/GI/n/\infty$ models, we conduct simulation experiments. Table 2 only evaluates the approximations for renewal arrival processes and service-time distributions with $c_s^2 = 1.0$ and $c_s^2 = 0.5$. We also want to consider nonrenewal arrival processes and other service-time distributions.

We consider two nonrenewal arrival processes. First, we consider a deterministic process with local variability. For a rate-1 process, we let the first four interarrival times be 0.1, and then we let the fifth interarrival time be 4.6. We then repeat, getting five arrivals in each interval $[5n, 5(n+1)]$ for positive integers n. We call this a deterministic batch process with clusters of size 5, and refer to it as $Db5$. Even though the $Db5$ process has more variability than the D process, it too has asymptotic scaling constant $c_a^2 = 0$ in the FCLT for the arrival process, as in (2.1) and (2.2).

Our second nonrenewal process is the independent superposition of four IID H_2^b renewal processes, denoted by $sup4H_2^b$. We let the SCV be $c_a^2 = 19$ in each component process to match the SCV of the H_2^* process with parameter $p = 0.1$. In the FCLT for the superposition arrival process, the scaling constant is the same as the SCV of a component renewal process. An interarrival time in the superposition process has a much smaller SCV; e.g., see Whitt (1982).

To examine more highly variable service-time distributions, we consider the lognormal (LN) and Pareto with parameter $p = 3/2$ (Par(3/2)) in addition to H_2^* with parameter $p = 0.1$, yielding SCV $c_s^2 = 19$. The lognormal takes the form $e^{a+bN(0,1)}$, where the parameters a and b are chosen to yield the desired mean and SCV. We let the SCV be 19 to match the H_2^* distribution with parameter $p = 0.1$. For the lognormal distribution, we calculate the parameter $\eta(G)$ in (1.7) by numerical integration. (The parameter values of $\eta(G)$ for all the service-time distributions considered are given in the last row of Table 3.)

We conduct the simulations using Splus and Fortran, exploiting recursive expressions for the departure times from the multiserver queue in Berger and Whitt (1992b).

Table 3. Key variability parameters in several $G/GI/n/\infty$ queues: the asymptotic peakedness z in (1.6) and the variability parameter v in (4.7) with $w \equiv 1$ ($v = (c_a^2 + c_s^2)/2$), where c_a^2 is understood to be the scale factor in the FCLT for the arrival process.

		Service times				
		D	M	H_2^*	LN	Par(3/2)
Arrival process		$c_s^2 = 0$	$c_s^2 = 1$	$c_s^2 = 19$	$c_s^2 = 19$	$c_s^2 = \infty$
D	z	—	0.50	0.98	0.78	0.75
$c_a^2 = 0$	v	—	0.50	9.50	9.50	∞
$Db5$	z	—	0.50	0.98	0.78	0.75
$c_a^2 = 0$	v	—	0.50	9.50	9.50	∞
M	z	1.00	1.00	1.00	1.00	1.00
$c_a^2 = 1$	v	0.50	1.00	10.00	10.00	∞
H_2^*	z	19.00	10.00	1.90	4.96	5.50
$c_a^2 = 19$	v	9.50	10.00	19.00	19.00	∞
LN	z	19.00	10.00	1.90	4.96	5.50
$c_a^2 = 19$	v	9.50	10.00	19.00	19.00	∞
$sup4H_2^b$	z	19.00	10.00	1.90	4.96	5.50
$c_a^2 = 19$	v	9.50	10.00	19.00	19.00	∞
Integral	$\eta(G)$	1.00	0.50	0.05	0.22	0.25

Note. The process $Db5$ is a deterministic process with clusters of size 5, while $sup4H_2^b$ is the superposition of four IID renewal processes with H_2^b interarrival times with SCV $c_a^2 = 19$.

Given the arrival times and departure times, we construct the queue-length process at state-change times using the method on page 210 of Whitt (2002); i.e., we first construct a sequence of change times by sorting the arrival and departure times; then we construct a vector with a $+1$ associated with each arrival and a -1 associated with each departure, ordered according to the times of occurrence; then the sequence of successive queue lengths at change times is the associated cumulative-sum process. Because loops are not efficient in Splus, we used Fortran to construct the queue-length process from the arrival process and service times.

We conducted a simulation experiment with each combination of six arrival processes and five service-time distributions. We considered four renewal arrival processes and the two nonrenewal processes $Db5$ ($c_a^2 = 0$) and $sup4H_2^b$ ($c_a^2 = 19$) introduced above. The four renewal processes had interarrival times distributed as D, M, and H_2^* with parameter $p = 0.1$ ($c_a^2 = 19$) and LN with $c_a^2 = 19$. The five service-time distributions are D, M, and H_2^* with $p = 0.1$ ($c_s^2 = 19$), LN with $c_s^2 = 19$ and $\mathrm{Par}(3/2)$, which has infinite variance. The variability parameters z in (1.6) and v in (4.7) with $w = 1$ for these examples are displayed in Table 3. Values of the service-time variability factor $\eta(G)$ in (1.7) appear in the last row. Note that $\eta(G)$ is smallest for the H_2^* service-time distribution. Also note that the asymptotic peakedness in the cases of highly variable arrival processes is smallest for the H_2^* service-time distribution.

Simulation results based on runs for 10^6 arrivals are shown in Tables 4 and 5. The approximation for the probability that all servers are busy (PB) in (1.5) is compared to the simulation estimates in Table 4. Based on subsequent independent replications, we conclude that there is statistical precision only to about 10% in the more variable cases.

Consistent with Theorem 2.1, the simulations show that nonrenewal arrival processes primarily affect congestion in the regime (0.1) through their rate and the scaling constant appearing in the FCLT, as in (2.1) and (2.2). The results for the $Db5$ arrival process are similar to those for the D arrival process, while the results for the $sup4H_2^b$ arrival process are similar to the renewal arrival processes with $c_a^2 = 19$.

The quality of the approximations for PB are consistently good with the exception of the cases involving H_2^* arrival processes, where the approximations are too high. We have not been able to explain that discrepancy. Independent replications yield similar values. Otherwise, the delay-probability approximation seems consistently good across all cases.

However, Table 5 shows that the approximations for the mean conditional number waiting given that all servers are busy, assuming $w \equiv 1$ in formula (4.7) for v, behave very differently. These approximations are quite accurate for M and H_2^* service times, where the approximations have been shown to be asymptotically correct, but the approximations grossly overestimate the exact values for the highly variable LN and $\mathrm{Par}(3/2)$ service-time distributions.

Indeed, the low simulation values with LN and $\mathrm{Par}(3/2)$ service-time distributions are remarkable. The approximation for LN service times can be improved dramatically if we use (4.7) with $w = 0.18$ (obtained by considering what is needed in the case D/LN). The approximations change to 14.2 for D and $Db5$ arrivals, 17.5 for M arrivals, and 77.6 for H_2^*, LN, and $sup4H_2^b$ arrivals.

Since the Pareto(3/2) service-time distribution has an infinite variance, $c_s^2 = \infty$, the approximation in (4.7) for v makes no sense. Based on (4.7), we would expect the queue-length process to be unstable, but evidently that is

Table 4. A comparison of approximations using (1.5) with simulations estimates of the probability all servers are busy (PB) in several $G/GI/n/\infty$ queues with $\lambda = 100$, $\mu = 1$, and $n = 115$ ($\rho = 0.870$) based on 10^6 arrivals.

		Service times				
Arrival process		D $c_s^2 = 0$	M $c_s^2 = 1$	H_2^* $c_s^2 = 19$	LN $c_s^2 = 19$	$\mathrm{Par}(3/2)$ $c_s^2 = \infty$
D	approx.	—	0.029	0.098	0.072	0.067
$c_a^2 = 0$	sim.	—	0.024	0.096	0.070	0.049
$Db5$	approx.	—	0.029	0.098	0.072	0.067
$c_a^2 = 0$	sim.	—	0.026	0.106	0.063	0.052
M	approx.	0.105	0.105	0.105	0.105	0.105
$c_a^2 = 1$	sim.	0.095	0.105	0.102	0.107	0.102
H_2^*	approx.	0.654	0.549	0.219	0.416	0.437
$c_a^2 = 19$	sim.	0.647	0.576	0.267	0.474	0.490
LN	approx.	0.654	0.549	0.219	0.416	0.437
$c_a^2 = 19$	sim.	0.635	0.547	0.220	0.414	0.409
$sup4H_2^b$	approx.	0.654	0.549	0.219	0.416	0.437
$c_a^2 = 19$	sim.	0.623	0.541	0.228	0.392	0.398
Integral	$\eta(G)$	1.00	0.50	0.05	0.22	0.25

Note. The process $Db5$ is a deterministic process with clusters of size 5, while $sup4H_2^b$ is the superposition of four IID renewal processes with H_2^b interarrival times with SCV $c_a^2 = 19$.

Table 5. A comparison of approximations with simulation estimates of the mean conditional number waiting given that all customers are busy in several $G/GI/n/\infty$ queues with $\lambda = 100$, $\mu = 1$, and $n = 115$ ($\rho = 0.870$) based on 10^6 arrivals.

		Service times				
		D	M	H_2^*	LN	Par(3/2)
Arrival process		$c_s^2 = 0$	$c_s^2 = 1$	$c_s^2 = 19$	$c_s^2 = 19$	$c_s^2 = \infty$
D	mean approx.	—	3.3	63.5	63.5	∞
	mean sim.	—	2.8	59.5	14.3	9.6
	SD sim.	—	3.1	52.5	21.00	12.7
$Db5$	mean approx.	—	3.3	63.5	63.5	∞
	mean sim.	—	2.8	54.1	13.9	9.3
	SD sim.	—	3.1	52.8	20.7	12.7
M	mean approx.	3.3	6.7	66.9	66.9	∞
	mean sim.	4.7	6.4	60.3	17.8	12.8
	SD sim.	4.5	6.7	55.2	26.1	15.5
H_2^*	mean approx.	63.5	66.9	127.0	127.0	∞
	mean sim.	79.4	75.8	114.4	63.2	64.9
	SD sim.	76.1	76.8	140.5	74.6	80.3
LN	mean approx.	63.5	66.9	127.0	127.0	∞
	mean sim.	46.6	46.6	111.7	49.0	40.7
	SD sim.	39.4	47.0	112.6	57.1	53.3
$sup4H_2^b$	mean approx.	63.5	66.9	127.0	127.0	∞
	mean sim.	65.9	64.4	120.0	63.1	51.6
	SD sim.	64.4	66.5	127.8	79.4	60.0

Notes. The approximations use (5.7) (divided by α) with v in (4.7) and $w \equiv 1$. The process $Db5$ is a deterministic process with clusters of size 5, while $sup4H_2^b$ is the superposition of four IID renewal processes with H_2^b interarrival times with SCV $c_a^2 = 19$.

not the case. In fact, quite reasonable approximations for the cases with Pareto(3/2) service-time distributions can be obtained by using the approximation $v \approx 2.8$. The approximations change to 9.4 for D and $Db5$ arrivals, 12.7 for M arrivals, and 72.9 for H_2^*, LN, and $sup4H_2^b$ arrivals. It remains to determine how to systematically define an appropriate variability parameter v, but the evidence suggests that it should be possible.

In Table 5 we display estimates of the standard deviation (SD) of the conditional number waiting given that all servers are busy as well as the mean. Because the estimates of the SD differ relatively little from the estimates for the mean, we conclude that the distribution is reasonably well approximated by an exponential distribution. However, the cases of the heavy-tailed LN and Par(3/2) service times suggest that in those cases the distribution has a slightly heavier tail, with an SCV of about 1.5 instead of 1.0. Certainly, the tail of the steady-state queue length is closer to an exponential distribution than to the tail of the service-time distribution itself. In Figure 4 we plot four estimates of the steady-state density based on these simulations (ignoring the discreteness), using the Splus nonparametric density estimator, to show that the steady-state distributions do indeed have the claimed general form.

Many simulations of $M/GI/100/\infty$ queues with nonexponential service times, including lognormal service times, have recently been conducted by Mandelbaum and Schwartz (2002). Their results are consistent with what we observed above.

7. Approximations for Blocking Probabilities in $G/GI/n/m$

We now apply the diffusion approximation in §4 to generate an approximation for the blocking probability in the $G/GI/n/m$ queue. Because the diffusion process has a reflecting barrier at κ, which is not defined by a reflection map applied to a free process, the diffusion does not directly experience any loss. However, we can define a loss rate for the diffusion process by looking at the behavior of the diffusion process in the neighborhood of the boundary.

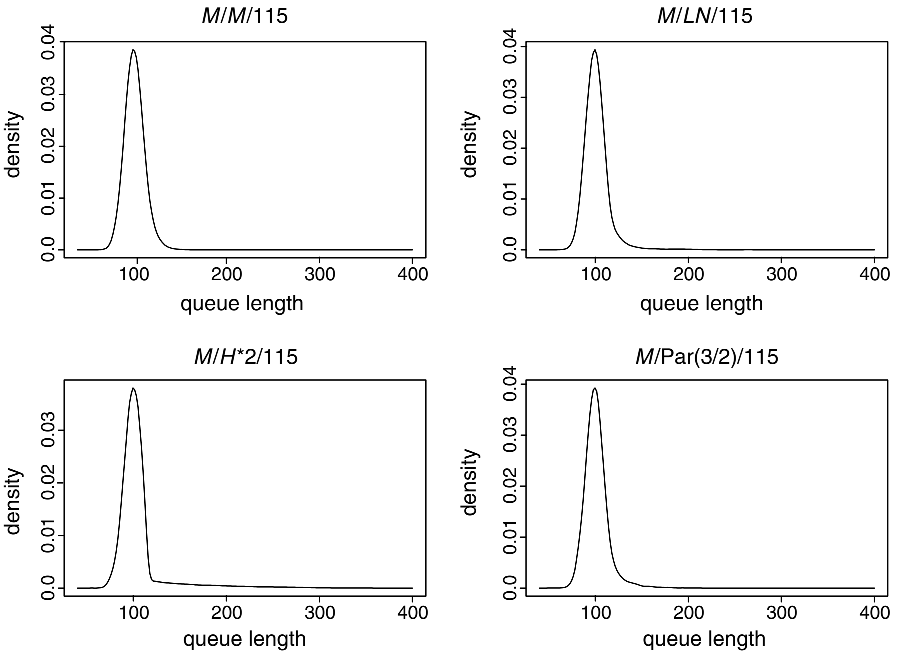
For $x > 0$, the diffusion process acts like ordinary Brownian motion with a drift. Thus, just as for the $G/G/1/m$ model in Berger and Whitt (1992a), we can apply the reasoning on pages 86–92 in Harrison (1985) to motivate defining the (long-run) loss rate (at the upper barrier κ) of the diffusion process $\mathbf{Q}$ as

$$r_{\mathbf{Q}} \equiv \frac{f_{\mathbf{Q}(\infty)}(\kappa)\sigma_{\mathbf{Q}}^2(\kappa)}{2}, \tag{7.1}$$

where $f_{\mathbf{Q}(\infty)}$ is the PDF of $\mathbf{Q}(\infty)$ in (3.14) or (3.15) and $\sigma_{\mathbf{Q}}^2(\kappa)$ is the infinitesimal variance of $\mathbf{Q}$ evaluated at the upper boundary κ. In the case $\beta \neq 0$,

$$f_{\mathbf{Q}(\infty)}(\kappa) = \frac{\alpha\beta e^{-\kappa\beta/v}}{v(1 - e^{-\kappa\beta/v})}. \tag{7.2}$$

Figure 4. Estimates of the steady-state density of the queue-length process (ignoring the discreteness) in four $GI/GI/n/\infty$ models with $\lambda = 100$, $\mu = 1$, and $n = 115$.



Notes. The service times are exponential (M), lognormal (LN), Pareto(3/2), and H_2^*. The estimates are obtained from the Splus nonparametric density estimator based on 10^6 arrivals in each case.

From (2.9), (2.10), and (2.12), we see that the infinitesimal variance of $\mathbf{Q}$ evaluated at κ is

$$\sigma_{\mathbf{Q}}^2(\kappa) = \frac{2\mu z}{p} = 2\mu v. \tag{7.3}$$

Because $Q_n(t) \approx n + \sqrt{n}\mathbf{Q}(t)$ by the scaling in (2.4), we approximate the loss rate in the queueing system by

$$r_{Q_n} \approx \sqrt{n} r_{\mathbf{Q}}. \tag{7.4}$$

Because the blocking probability equals the loss rate divided by the arrival rate, we approximate the blocking probability in the queueing system, denoted by π_n, by

$$\pi_n = \frac{r_{Q_n}}{\lambda_n} \approx \frac{f_{\mathbf{Q}(\infty)}(\kappa)v}{\rho_n\sqrt{n}}. \tag{7.5}$$

Remark 7.1 (A Conjectured Local Limit). We conjecture that the approximation in (7.5) can be supported by a local limit in the $G/H_2^*/n/m$ model under the conditions of Theorem 2.1. That limit would state that

$$\sqrt{n}\pi_n \equiv P(Q_n^a(\infty) = n + m_n) \to f(\kappa)v \tag{7.6}$$

as $n \to \infty$ for $v = (c_a^2 + c_s^2)/2$. That is in the spirit of Theorem 15 on page 226 of Borovkov (1976). For the case of exponential service times, where the multiserver queue with all servers busy behaves like a single-server queue, Whitt (2004) has verified this conjecture when $\beta < 0$.

Remark 7.2 (Comparison with the $G/GI/n/0$ Loss Model). The same reasoning applies to the $G/GI/n/0$ loss model, but the blocking formula is quite different. When $x < 0$, the diffusion behaves like an Ornstein-Uhlenbeck process, not a Brownian motion. However, the infinitesimal parameters are approximately constant in the neighborhood of the upper boundary (now $\kappa = 0$). We thus use the same reasoning and define the loss rate of the diffusion process as

$$r_{\mathbf{Q}} \equiv \frac{f_{\mathbf{Q}(\infty)}(\kappa)\sigma_{\mathbf{Q}}^2(\kappa)}{2}, \tag{7.7}$$

just as in (7.1), except now $\kappa = 0$.

From (3.14), we see that

$$f_{\mathbf{Q}(\infty)}(0) = \frac{\phi(\beta/\sqrt{z})}{\sqrt{z}\Phi(\beta/\sqrt{z})} \tag{7.8}$$

when $\beta \neq 0$. From (2.9), (2.10), and (2.13), we see that

$$\sigma_{\mathbf{Q}}^2(0) = 2p\mu z. \tag{7.9}$$

Hence, we obtain the blocking-probability approximation

$$\pi_n \approx \frac{p\sqrt{z}\phi(\beta/\sqrt{z})}{\rho\sqrt{n}\Phi(\beta/\sqrt{z})}. \tag{7.10}$$

Formula (7.10) is $p\sqrt{\rho}$ times the approximation in (15) of Srikant and Whitt (1996). The factor of $\sqrt{\rho}$ is removed if we apply approximation (5.3). Moreover, that factor is asymptotically negligible in the limiting regime (0.1). Thus, we reproduce the previous blocking-probability approximation in Srikant and Whitt (1996) in the special case $p = 1$. Otherwise, the formulas are different.

We evaluate the approximations for both the delay probability and the blocking probability in $GI/GI/n/m$ models in Table 6. For these examples we let v be as in (4.7) with $w \equiv 1$. In Table 6 we make comparisons with exact values from Seelen et al. (1985) for the $GI/GI/25/10$ model for several different values of c_a^2, c_s^2, and ρ. The exact delay probability values (*PW*) in Table 6 differ from those in Seelen et al. (1985), because they display the conditional delay probability given that the customer is admitted. Our value of *PW* is computed from theirs, denoted by PW_S, by

$$PW = PW_S + PBL - (PW_S)(PBL). \qquad (7.11)$$

We regard the quality of the approximations as quite good. However, the delay-probability approximation is surprisingly inaccurate when $\rho = 1$ and $c_s^2 = 1$, where it is supposed to be asymptotically correct.

The accuracy of the approximations may be less impressive than we would wish, but it is important to recognize that great accuracy is not required in many applications. A principle application is server staffing. In that application, great accuracy is not necessary because servers come in integer quantities, and the performance measures tend to change substantially with unit changes in the staffing.

We illustrate by showing how the approximation for the blocking probability π_n depends on the number of servers, n, for several $GI/GI/25/10$ queues. We let Table 6 serve as our base case: For $\rho = 0.9$, the arrival rate is $\lambda = 22.5$. We change n, holding the arrival rate fixed at $\lambda = 22.5$.

Table 6. A comparison of the approximations with exact values of the delay probability (*PW*), the probability all servers are busy (*PB*), and the blocking probability (*PBL*) in the $GI/GI/n/m$ model with exponential (M) and Erlang (E_2, $c_s^2 = 0.5$) service-time distributions for $n = 25$, $m = 10$ and various values of the interarrival-time SCV c_a^2 and the traffic intensity ρ.

Parameters			Exact values			Approximations	
c_a^2	c_s^2	ρ	*PW*	*PB*	*PBL*	(3.4) and (3.8)	(7.5)
4.00	1.00	1.5	0.918	0.883	0.347	0.890	0.343
		1.2	0.800	0.743	0.212	0.706	0.213
		1.0	0.628	0.561	0.112	0.502	0.126
		0.9	0.502	0.436	0.067	0.388	0.088
		0.8	0.353	0.297	0.032	0.279	0.057
		0.7	0.205	0.165	0.0110	0.184	0.034
1.00	1.00	1.5	0.990	0.990	0.334	0.994	0.334
		1.2	0.913	0.913	0.176	0.907	0.175
		1.0	0.649	0.649	0.059	0.615	0.061
		0.9	0.414	0.414	0.021	0.392	0.025
		0.8	0.195	0.195	0.0046	0.199	0.0078
		0.7	0.063	0.063	0.00054	0.081	0.0018
0.25	1.00	1.5	0.9987	0.9990	0.333	0.9997	0.333
		1.2	0.965	0.970	0.169	0.970	0.169
		1.0	0.680	0.705	0.041	0.669	0.042
		0.9	0.355	0.384	0.0079	0.359	0.0101
		0.8	0.110	0.126	0.00060	0.132	0.00118
		0.7	0.0182	0.023	0.00002	0.0343	0.00012
4.00	0.50	1.5	0.924	0.891	0.347	0.916	0.342
		1.2	0.812	0.757	0.212	0.753	0.211
		1.0	0.648	0.581	0.112	0.561	0.123
		0.9	0.527	0.459	0.067	0.449	0.085
		0.8	0.382	0.321	0.032	0.338	0.055
		0.7	0.230	0.186	0.0109	0.239	0.033
1.00	0.50	1.5	0.996	0.996	0.334	0.9989	0.333
		1.2	0.944	0.944	0.172	0.953	0.171
		1.0	0.681	0.681	0.052	0.680	0.051
		0.9	0.427	0.427	0.0160	0.429	0.0170
		0.8	0.194	0.194	0.0029	0.211	0.0039
		0.7	0.061	0.061	0.00028	0.083	0.00067
0.25	0.50	1.5	0.9999	0.9999	0.333	0.99999	0.333
		1.2	0.990	0.992	0.167	0.996	0.167
		1.0	0.733	0.758	0.030	0.738	0.028
		0.9	0.336	0.369	0.0030	0.324	0.0027
		0.8	0.084	0.101	0.00008	0.083	0.00010
		0.7	0.011	0.0144	0.00000	0.0137	0.000002

Note. The approximations have v in (4.7) with $w \equiv 1$. The exact values come from Seelen et al. (1985).

Table 7. The approximate blocking probability as a function of the number of servers in four $GI/GI/25/10$ models with arrival rate $\lambda = 22.5$, as occurs in the cases of Table 6 when $\rho = 0.9$.

Parameters		Approximate blocking probabilities			
n	ρ	M/M	E_2/E_2	H_2/E_2, $c_a^2 = 4.0$	M/H_2, $c_s^2 = 10$
21	1.07	0.097	0.077	0.152	0.157
22	1.02	0.071	0.046	0.131	0.133
23	0.98	0.053	0.027	0.115	0.115
24	0.94	0.037	0.014	0.100	0.097
25	0.90	0.025	0.0065	0.085	0.081
26	0.87	0.018	0.0034	0.075	0.069
27	0.83	0.011	0.0013	0.063	0.055
28	0.80	0.0078	0.00063	0.055	0.046
29	0.78	0.0059	0.00037	0.050	0.040
30	0.75	0.0039	0.00016	0.043	0.033
40	0.56	0.00016	0.00000	0.0144	0.0057

We consider four cases: We consider the $M/M/n/m$ model, one less-bursty example and two more-bursty examples. The less-bursty example is the $E_2/E_2/n/m$ model with Erlang interarrival times and service times. Using (4.8), we let the approximate peakedness be $z \approx 0.625$. The more-bursty examples are $H_2^b/E_2/n/m$ with $c_a^2 = 4.0$ and $M/H_2^b/n/m$ with $c_s^2 = 10$.

The results are shown in Table 7. First, we see that quantifying the variability of a distribution beyond its mean can be very important: There is greater disparity going from $M/M/n/m$ to one of the other models (changing columns) than there is in adding or subtracting a server (changing rows).

We also see that adding servers tends to have a greater impact in the less-bursty examples: With greater variability, the addition of a server causes a smaller decrease in the blocking probability. For example, suppose that we want to decrease the blocking probability from just less than 0.080 to just less than 0.040. For the $M/M/25/10$ model, we would go from $n = 22$ to $n = 24$, an addition of two servers. In contrast, for the $M/H_2/25/10$ model, we would go from $n = 26$ to $n = 30$, an addition of four servers.

8. Conclusions

We have developed a heuristic diffusion approximation for the $G/GI/n/m$ queue, intended for the case of large n, which is supported by a heavy-traffic stochastic-process limit for the special case of the $G/H_2^*/n/m$ model, established in the companion paper Whitt (2005). Theorem 3.1 shows that the approximation yields relatively simple explicit formulas for the steady-state performance measures of interest.

Corollary 3.1 shows that the steady-state delay probability and the conditional distribution of the number of busy servers given that all servers are not busy in the $G/H_2^*/n/\infty$ model depend on the parameter p only through the parameter z, which has a natural approximation by the asymptotic peakedness in (1.6) in the $G/GI/n/\infty$ model. Thus, there is reason to expect that the approximations for these characteristics perform well. Simulation experiments confirm that the approximations for these quantities perform remarkably well across a wide range of cases. We thus feel that we have successfully met our main goal of generating a useful approximation for the delay probability in $G/GI/n/\infty$ models.

Especially interesting are simulation results for multiserver queues with heavy-tailed service-time distributions. The simulations show that the congestion is much less than might be expected. That is partly explained by the formula for the asymptotic peakedness in (1.6) that plays an important role in the approximations for the steady-state delay probability. Much insight is provided by the value of the integral $\eta(G)$ in (1.7) for the heavy-tailed distributions.

Simulation results show that the diffusion approximation with the variability parameter $v = (c_a^2 + c_s^2)/2$ works well for the mean steady-state number waiting for service-time distributions with low-to-moderate variability. However, the simulation results in Table 5 show that the diffusion approximation with this parameter v grossly overestimates the expected mean number waiting when the service-time distribution is lognormal. That discrepancy disappears if we use a refined approximation for v as in (4.7) with an appropriate weight w. However, it remains to determine a weight function w that produces good performance for the mean number waiting across a wide range of cases. It also remains to determine an appropriate parameter v for heavy-tailed distributions, like the Pareto(3/2), that have finite mean but infinite variance. Indeed, nothing has yet been proved about the limiting behavior in that case.

Overall, we believe that we have developed a useful approximation framework, but there remains much work to do. It would be interesting to compare the results here to those obtained from an algorithm to compute the steady-state distribution of the multidimensional diffusion in Puhalskii and Reiman (2000). For approximations (but not for asymptotics), presumably lognormal and Pareto distributions can be effectively treated by approximating them by appropriate phase-type distributions, using algorithms such as in Asmussen et al. (1996) and Feldmann and Whitt (1998).

Acknowledgments

This research was done at AT&T Labs Research, Florham Park, NJ; Columbia University, New York, NY; and Avaya Labs Research, Basking Ridge, NJ. At Columbia University, the author was supported by National Science Foundation grant DMS-02-2340.

References

Albin, S. L. 1982. On Poisson approximations for superposition arrival processes in queues. *Management Sci.* **28** 126–137.

Albin, S. L. 1984. Approximating a point process by a renewal process, II: Superposition arrival processes. *Oper. Res.* **32** 1133–1162.

Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.

Asmussen, S., O. Nerman, M. Olsson. 1996. Fitting phase type distributions via the EM algorithm. *Scandanavian J. Statist.* **23** 419–441.

Berger, A. W., W. Whitt. 1992a. The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue. *J. Discrete Event Dynam. Systems* **2** 7–60.

Berger, A. W., W. Whitt. 1992b. Comparisons of multi-server queues with finite waiting rooms. *Stochastic Models* **8** 719–732.

Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. Wiley, New York.

Bolotin, V. 1994. Telephone circuit holding-time distributions. J. Labetoulle, J. W. Roberts, eds. *Proc. Internat. Teletraffic Congress, ITC*14. North-Holland, Amsterdam, The Netherlands, 125–134.

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

Borovkov, A. A. 1984. *Asymptotic Methods in Queueing Theory.* Wiley, New York.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2002. Statistical analysis of a telephone call center: A queueing-science perspective. Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. Dshalalow, ed. *Advances in Queueing.* CRC Press, Boca Raton, FL, 463–480.

Eckberg, A. E. 1983. Generalized peakedness of teletraffic processes. *Proc. Tenth Internat. Teletraffic Congress.* Montreal, Quebec, Canada.

Eckberg, A. E. 1985. Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness. *Proc. Eleventh Internat. Teletraffic Congress.* Kyoto, Japan.

Feldmann, A., W. Whitt. 1998. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation* **31** 245–279.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

Halachmi, B., W. R. Franta. 1978. A diffusion approximation to the multi-server queue. *Management Sci.* **24** 522–529.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems.* Wiley, New York.

Hokstad, P. 1978. Approximations for the $M/G/m$ queue. *Oper. Res.* **26** 510–523.

Jelenkovic, P., A. Mandelbaum, P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47** 53–69.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

Kimura, T. 1995. An $M/M/s$-consistent diffusion model for the $GI/G/s$ queue. *Queueing Systems* **19** 377–397.

Kimura, T. 2000. Equivalence relations in the approximations for the $M/G/s/r$ queue. *Math. Comput. Model.* **31** 215–224.

Kimura, T. 2002. Diffusion approximations for queues with Markovian bases. *Ann. Oper. Res.* **113** 27–40.

Mandelbaum, A. 2001. Call centers (centres): Research bibliography with abstracts. Faculty of Industrial Engineering and Management, Technion, Haifa, Israel.

Mandelbaum, A., R. Schwartz. 2002. Simulation experiments with $M/G/100$ queues in the Halfin-Whitt (Q.E.D.) regime. Technical report, Technion, Haifa, Israel.

Massey, W. A., R. B. Wallace. 2004. An asymptotically optimal design of the $M/M/c/k$ queue for call centers. *Queueing Systems*. Forthcoming.

Newell, G. F. 1973. *Approximate Stochastic Behavior of n-Server Service Systems with Large n. Lecture Notes in Economics and Mathematical Systems*, No. 87. Springer, New York.

Nozaki, S. A., S. M. Ross. 1978. Approximations in finite-capacity multi-server queues with Poisson arrivals. *J. Appl. Probab.* **15** 826–834.

Puhalskii, A. A., M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.

Seelen, L. P., H. C. Tijms, M. H. van Hoorn. 1985. *Tables for Multi-Server Queues.* North-Holland, Amsterdam, The Netherlands.

Srikant, R., W. Whitt. 1996. Simulation run lengths to estimate blocking probabilities. *ACM Trans. Model. Comput. Simulation* **6** 7–52.

Whitt, W. 1982. Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* **30** 125–147.

Whitt, W. 1983. Comparison conjectures about the $M/G/s$ queue. *Oper. Res. Lett.* **2** 203–210.

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

Whitt, W. 1984b. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Lab. Tech. J.* **63** 163–175.

Whitt, W. 1984c. Departures from a queue with many busy servers. *Math. Oper. Res.* **9** 534–544.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* **38** 708–723.

Whitt, W. 1993. Approximations for the $GI/G/m$ queue. *Production Oper. Management* **2** 114–161.

Whitt, W. 2002. *Stochastic-Process Limits.* Springer, New York.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51** 531–542.

Whitt, W. 2004. Heavy-traffic limits for loss proportions in single-server queues. *Queueing Systems* **46** 507–536.

Whitt, W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math Oper. Res.* Forthcoming.