A DIFFUSION APPROXIMATION FOR THE $G/GI/n/m$ QUEUE

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027

July 5, 2002

(submitted to Operations Research)
Abstract

We develop a heuristic diffusion approximation for the queue-length stochastic processes representing the number in system at arrival epochs and at arbitrary times in the $G/GI/n/m$ queueing model, which has a general (stationary or asymptotically stationary) 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, focusing especially upon the steady-state delay probability.

We primarily base our diffusion approximation on heavy-traffic limits in which $n$ tends to infinity as the traffic intensity increases. For the $GI/M/n/\infty$ special case, Halfin and Whitt (1981) showed that scaled versions of these queue-length processes converge to a piecewise-linear diffusion process when the traffic intensity $\rho_n$ approaches 1 with $(1 - \rho_n)\sqrt{n} \to \beta$ for $0 < \beta < \infty$. A companion paper, Whitt (2002b), extends that limit to a special class of $G/GI/n/m$ models in which 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 $0 < \kappa \leq \infty$. The heuristic one-dimensional diffusion-process 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 this one-dimensional diffusion 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 delay probability. The accuracy is confirmed by simulation.

*American Mathematical Society 1991 subject classifications.* Primary 60K25, 60F17; Secondary 60K30.

*Keywords:* queues, multiserver queues, delay probability, blocking probability, heavy-traffic, diffusion approximations
The rapid growth of telephone call centers and more general contact centers has generated renewed interest in the performance of multiserver queueing models when the number of servers is large; e.g., see Borst, Mandelbaum and Reiman (1999), Garnett, Mandelbaum and Reiman (2000), Koole and Mandelbaum (2001), Mandelbaum (2001), Armony and Maglaras (2001), Whitt (2001) and references therein.

Since 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, Mandelbaum and Reiman (2000). They considered a sequence of models indexed by the number of servers, $n$, and let $n \to \infty$ with the traffic intensities $\rho_n$ converging to 1, the critical value for stability. Interesting nondegenerate limits occur when

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

For the $GI/M/n/\infty$ model (and presumably for the $GI/PH/n/\infty$ model as well, although it remains to be proved), the associated sequence of delay probabilities approaches a limit $\alpha$ strictly between 0 and 1 as $n \to \infty$ under the scaling in (0.1). Since 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). In contrast to the traffic intensity and the average delay, to a large extent the performance implications of a delay-probability value tend to be independent of $n$.

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

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

where $\beta$ is the limit in (0.1), $\Phi$ is the cumulative distribution function (cdf) and $\phi$ is the probability density function (pdf) of a standard (mean 0, variance 1) normal random variable; e.g., $\Phi(x) = P(N(0,1) \leq x)$. The function $\alpha$ 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. (1999). For the $M/M/n/\infty$ model, the asymptotic-delay-probability function $\alpha$ 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 $\alpha$ 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 $\rho$ 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).

Since the asymptotic-delay-probability function $\alpha$ 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, non-exponential service-time distributions and finite waiting space. (That is described as an important open problem at the end of Borst, Mandelbaum and Reiman (2000).) The present paper addresses that problem: In this paper we consider the more general $G/GI/n/m$ model with $m$ additional waiting spaces. 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, $GI$). 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.) In practice, non-exponential 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.

Another objective in this paper is to develop approximations for the case of a finite waiting room. For the heavy-traffic stochastic-process limits in the heavy-traffic regime of (0.1), it is necessary to let $m_n \to \infty$ as $n \to \infty$ so that

$$m_n / \sqrt{n} \to \kappa \quad \text{for} \quad 0 < \kappa \leq \infty . \quad (0.3)$$

For exponential service times, our results for finite waiting rooms here provide theoretical support, and refinements, for heuristic diffusion approximations in Section VII of Whitt (1984a).

Here is how the rest of the paper is organized: We start in Section 1 by describing the proposed approximation of the delay probability in the $G/GI/n/\infty$ model, which was our main goal. We state the stochastic-process limit for the $G/H_\infty^2/n/m$ model in Section 2 and develop the heuristic diffusion approximation for the $G/GI/n/m$ model in Section 3. We then characterize the steady-state distributions of these diffusion processes in Section 4.

In Section 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, Tijms and van Hoorn (1985). In Section 6 we describe simulations conducted to evaluate other $G/GI/n/m$ models, focusing especially on heavy-tailed service-time distributions and non-renewal arrival processes. In Section 7 we develop and evaluate associated approximations for the blocking
probability in the $G/GI/n/m$ model. In Section 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, Halflin 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 $\alpha$. However, the correct formula for the asymptotic delay probability can easily be derived from the diffusion-process parameters in Halflin and Whitt (1981); e.g, it can be obtained from Browne and Whitt (1995). The corrected $GI/M/n/\infty$ asymptotic delay-probability function is a minor modification of the $M/M/n/\infty$ function above, specifically,

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

where $z = (c_\alpha^2 + 1) / 2$ with $c_\alpha^2$ being the squared coefficient of variation (SCV, variance divided by the square of the mean, assumed to be finite) of an interarrival time, $\beta$ is the limit in (0.1) and $\alpha$ is the $M/M/n/\infty$ asymptotic-delay-probability function in (0.2). (Halflin and Whitt (1981) had the incorrect formula $\alpha(\beta / z)$. From (1.1), we see that the interarrival-time distribution beyond the mean enters in only via the SCV $c_\alpha^2$, just as in the central limit theorem for the arrival counting process.

In the next section we describe a new heavy-traffic limit for the more general $G/GI/n/\infty$ model with a non-renewal arrival process and a special non-exponential service-time distribution, which we establish in a companion paper Whitt (2002b). The non-exponential 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. Since 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., see Bolotin (1994) and Koole and Mandelbaum (2001).

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

$$\alpha_{GI/H_2^*/n/\infty} = \alpha_{GI/H_2^*/n/\infty}(\beta, c_\alpha^2, p) = \alpha(\beta/\sqrt{z})$$  \hspace{1cm} (1.2)

for $\alpha$ in (0.2), $\beta$ in (0.1) and

$$z = z(c_\alpha^2, p) = 1 + \frac{p(c_\alpha^2 - 1)}{2} = \frac{c_\alpha^2 + c_\alpha^2}{1 + c_\alpha^2},$$  \hspace{1cm} (1.3)

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

Since $z(c_\alpha^2, 1) = (c_\alpha^2 + 1)/2$, approximation (1.2) reduces to (1.1) in the $GI/M/n/\infty$ special case. Since $z(1, p) = 1$ for all $p$, $0 < p \leq 1$, formula (1.2) supports the approximation

$$\alpha_{M/GI/n/\infty} \approx \alpha_{M/M/n/\infty} \equiv \alpha(\beta),$$  \hspace{1cm} (1.4)

which is a longstanding approximation; e.g., see Section 3.2 of Whitt (1993). 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, 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.

As might be anticipated, the peculiar form of this tractable $H_2$ non-exponential 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 non-exponential 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 does produce 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 diffusion process. Then we use the exact steady-state distribution of the approximating diffusion process.

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

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

(1.5)

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

$$
z \equiv z(c_n^2, G) \equiv 1 + (c_n^2 - 1)\eta(G) ,
$$

(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,

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

(1.7)

and, just as in (1.3), $c_n^2$ is the normalization constant in a FCLT for the arrival process (assumed to hold, which requires that $c_n^2$ 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 $c_n^2$ is near 1, which is consistent with extensive simulation experience. On the other hand, when $c_n^2$ 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).

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 (1981a). 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 Section 10.3 of Whitt (2002a). Since we frequently refer to Whitt (2002a), we refer to it by the title acronym “SPL.”

5
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 $z = c_0^2$, 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 $z = (c_0^2 + 1)/2$.

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 $\alpha$ in (0.2) with the actual $M/M/n/\infty$ (Erlang-C) delay probability.

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 $\lambda$ and $n$ increase so that (0.1) holds. However, the delay-probability approximations in (0.2), (1.1), (1.2) and (1.5) are different than 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 Section 2.3 of Whitt (1992). Halim and Whitt (1981) observe in their Remark 1 that $\alpha(\beta) \geq \Phi^c(\beta)$ for all $\beta \geq 0$. These formulas are asymptotically equivalent as $\beta \to \infty$, but they need not be too close; e.g., $\Phi^c(0) = 0.5$, while $\alpha(0) = 1$. The refinement - going from $\Phi^c$ to $\alpha$ - was used by Jennings et al. (1996) in their server-staffing approximations for multi-server queues with time-varying arrival rates. That refinement is also exploited by Borst et al. (1999).

2. The Stochastic-Process Limit with $H^*_2$ Service Times

In this section we describe the heavy-traffic limit for the $GI/H^*_2/n/m$ model established in Whitt (2002b). 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 $C \equiv \{C(t) : t \geq 0\}$ with associated interarrival times $\{U_k : k \geq 1\}$. Our key assumption is that the arrival process satisfies a functional central limit theorem (FCLT). To state it, let $\Rightarrow$ denote convergence in distribution and let $D \equiv D([0,\infty),\mathbb{R})$ 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 SPL (Whitt (2002a)).
let $C_n$ be the random element of $D$ defined by

$$C_n(t) \equiv [C(nt) - nt]/\sqrt{nc_n^2}, \quad t \geq 0,$$

for some nonnegative scaling constant $c_n^2$. We assume that

$$C_n \Rightarrow B \quad (D, J_1), \quad (2.2)$$

where $B$ is standard (zero drift, unit diffusion coefficient) Brownian motion. When the arrival process is a renewal process, the limit (2.2) holds with $c_n^2$ being the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time (then assumed to be finite).

When the number of servers is $n$, we scale time in the arrival process, letting the arrival process be

$$C_n(t) \equiv C(\lambda_n t), \quad t \geq 0,$$

where $\lambda_n$ 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 $\mu^{-1}$, $0 < \mu < \infty$, so that the traffic intensity as a function of $n$ is $\rho_n = \lambda_n/\mu n$. Let $\nu^{-1}$ 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 $c_n^2 = (2/p) - 1$. Equivalently, $p^{-1} = (c_n^2 + 1)/2$. The SCV $c_n^2$ ranges from 1 to $\infty$ 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 system, including both waiting and in service. Let $Q_n^a(k)$ be the queue length just before the $k^{th}$ arrival, including all arrivals up to number $k - 1$ if there are batch arrivals.

For the stochastic-process limit, we construct scaled random elements of $D$ by letting

$$Q_n(t) \equiv [Q_n(t) - n]/\sqrt{n},$$

$$Q_n^a(t) \equiv [Q_n^a([nt]) - n]/\sqrt{n}, \quad t \geq 0.$$  \hspace{1cm} (2.4)

There is no time scaling for $Q_n$ in (2.4) because the arrival rate $\lambda_n$ is allowed to grow directly.

Let $D^2 \equiv D \times D$ be the product space with the associated product topology. Let $e$ be the identity function in $D$, i.e., $e(t) = t$, $t \geq 0$.

**Theorem 2.1.** For the family of $G/H_2^*/m$ models specified above, suppose that the arrival rate $\lambda_n$ and the number of waiting spaces, $m_n$, change with $n$ so that (0.1) and (0.3) hold with
\[-\infty < \beta < \infty \text{ and } 0 \leq \kappa \leq \infty. \] If in addition \( Q_n(0) \Rightarrow Q(0) \) in \( \mathbb{R} \), where \( Q(0) \) is a proper nonnegative random variable with \( P(Q(0) \leq \kappa) = 1 \) if \( \kappa < \infty \), then

\[
(Q_n, Q_n^\alpha) \Rightarrow (Q, Q^\alpha) \quad \text{in} \quad (D, J_1)^2 \quad \text{as} \quad n \to \infty ,
\]

where \( Q^\alpha = Q \circ \mu^{-1}e \) and \( Q \) is a diffusion process starting at \( Q(0) \) with a reflecting upper barrier at \( \kappa \) if \( \kappa < \infty \) and an inaccessible upper boundary at infinity if \( \kappa = \infty \). If \( \kappa > 0 \), then the diffusion process \( Q \) has infinitesimal mean

\[
m(x) = \begin{cases} -\mu \beta, & 0 \leq x < \kappa, \\ -\mu (x + \beta), & x < 0 , \end{cases}
\]

and infinitesimal variance

\[
\sigma^2(x) = \begin{cases} \mu (c^2 + (2/p) - 1) = \mu (c_a^2 + c_b^2), & 0 \leq x < \kappa, \\ \mu (p c^2_a + 2 - p) = p \mu (c^2_a + c^2_b) = 2 \mu (c^2_a + c^2_b)/(1 + c^2_b), & x < 0 . \end{cases}
\]

If \( \kappa = 0 \), then the diffusion process \( Q \) again has the infinitesimal parameters in (2.6) and (2.7) with the understanding that the case \( x > 0 \) does not arise. The infinitesimal parameters of \( Q^\alpha \) are those of \( Q \) divided by \( \mu \).

**Remark 2.1.** The character of a heavy-tailed distribution. To show that an \( H_2^p \) service-time distribution has some of the character of a heavy-tailed service-time distribution when the parameter \( p \) is small, we compare an \( H_2^p \) service-time distribution to a Pareto service-time distribution. The Pareto distribution we consider has the complementary cdf

\[
C^\alpha(t) \equiv (1 + t/(p - 1))^{-p}, \quad t \geq 0 ,
\]

which is scaled to have mean 1. This Pareto distribution, denoted by \( \text{Par}(p) \), has finite mean if and only if \( p > 1 \) and it 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 \( \text{Par}(3/2) \) is infinite, the variability parameter \( \eta(G) \) in (1.7) is finite; in particular,

\[
\eta(\text{Par}(p)) = \int_0^\infty (1 + t/(p - 1))^{-2p} dt = \frac{p - 1}{2p - 1},
\]

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

\[
\eta(H_2^p(p)) = \int_0^\infty p^2 e^{-2p t} dt = \frac{p}{2} ,
\]

so that \( \eta(H_2^{p}(0.1)) = 1/20 \). Of course, with Poisson arrivals, \( c^2_a = 1 \) so that \( z = 1 \) in both cases for \( z \) in (1.6).
Figure 1: A sample path of the queue-length process for $10^6$ arrivals in the $M/H_2^*/n$ queue with arrival rate $\lambda = 100$, service rate $\mu = 1$ and parameter $p = 0.1$.

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 $M/Par(3/2)/n/\infty$ 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 $M/M/n/\infty$.

3. The Heuristic Diffusion Approximation for $G/GI/n/m$

We now seek a diffusion approximation for the general $G/GI/n/m$ model. From the stochastic-process limits for the $GI/PH/n/\infty$ model in Puhalskii and Reiman (2000), we know that the limits for the scaled queue-length processes in those cases 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. In order to generate more tractable approximations, here we develop a heuristic one-dimensional diffusion approximation with convenient explicit formulas for all 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) and Whitt (1984).
Figure 2: A sample path of the queue-length process for $10^6$ arrivals in the $M/Par(3/2)/105$ queue with arrival rate $\lambda = 100$ and service rate $\mu = 1$.

As an approximation for the scaled queue-length processes, we propose a one-dimensional reflected piecewise-linear diffusion process, similar to the limit in Theorem 2.1, where the infinitesimal means and variances are consistent with established behavior of infinite-server queues for $x < 0$ and single-server queues when $x > 0$. This general approach is consistent with Section 3.2 of Whitt (1992), but we implement the idea differently here. Here we first use the intuition to construct the diffusion process, and then afterwards we construct the exact steady-state distribution of that diffusion process, drawing upon Browne and Whitt (1995).

We let the infinitesimal mean be just as in (2.6) for the $GI/H_2^2/n/m$ model; i.e.,

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

where $\beta$ is again the limit in (0.1). From established heavy-traffic limits, we deduce that the variability parameters should not appear in the infinitesimal mean.

We now turn to the infinitesimal variance. For $x < 0$, we use the fact that the corresponding infinite-server model has a normal distribution for its heavy-traffic stochastic-process limit; see Section 10.3 of SPL and references cited there, notably Borovkov (1984). For $x < 0$, we let the infinitesimal variance be consistent with the postulated infinitesimal mean in (3.1), a constant
infinitesimal variance and the normal steady-state distribution in the infinite-server model. By Browne and Whitt (1995), these properties are consistent with an Ornstein-Uhlenbeck diffusion process operating in the region $x < 0$. With the infinitesimal mean given in (3.1), it suffices to let the infinitesimal variance be such that the steady-state normal distribution has the correct variance in the infinite-server case. Therefore, the key quantity is the asymptotic peakedness $z$ in (1.6). By this reasoning the infinitesimal variance should be

$$
\sigma^2(x) = 2\mu z \quad \text{for} \quad x < 0 ,
$$

where $z$ is the asymptotic peakedness in (1.6).

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. 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 drift $-\mu$ as in (3.1). 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 the $n$ servers are all 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 $c_a^2$, where $c_2^2$ is the SCV of a service time, here assumed to be finite; see Section 9.4 of SPL. That perspective leads to the approximation

$$
\sigma^2(x) = \mu(c_a^2 + c_s^2) \quad \text{for} \quad x > 0 ,
$$

where $\mu^{-1}$ is the mean service time, $c_a^2$ is the arrival-process variability parameter obtained from a FCLT for the arrival process, as in (2.1) and (2.2), and $c_s^2$ is the service-time SCV.

We find that the approximation in (3.3) 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

11
the total rate fixed; see Theorem 9.8.1 of SPL. Naturally, this second perspective leads us to
the approximation in (3.3) with $c^2_s$ replaced by 1. 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 SPL. 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.,

$$
\sigma^2(x) = \mu(c^2_a + wc^2_w + 1 - w) \quad \text{for } x > 0 ,
$$

(3.4)

where $w$ is an appropriate weight with $0 \leq w \leq 1$. To develop a candidate weight functions
$w$, 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 SPL. 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 the number of component processes in the superposition increases along with the
time-and-space scaling; see Section 9.8 of SPL. 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 SPL. 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 (3.4),
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

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

(3.5)

for $\beta = \sqrt{n}(1 - \rho)$; see p. 333 of SPL. However, we do not find a direct application of (3.5)
to be effective, which is not surprising considering that our situation is in fact quite different
from a queue with a superposition arrival process.

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 SPL – 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. We thus want to allow the infinitesimal variance to be well defined when $c^2_\sigma$ is infinite.

In summary, this analysis leads us to approximate the infinitesimal variance by

$$\sigma^2(x) = \begin{cases} 
2\mu v, & 0 \leq x < \kappa, \\
2\mu z, & x < 0,
\end{cases}$$  \hspace{1cm} (3.6)

where $z$ is the asymptotic peakedness in (1.6) and $v$ is a variability parameter that is only tentatively specified. Our tentative specification of $v$ in the case of a finite service-time SCV $c^2_\sigma$ is

$$v = \frac{c^2_\sigma + wc^2_\sigma + 1 - w}{2}$$  \hspace{1cm} (3.7)

for some weight function $w$, which is a decreasing function of $\beta$ with $w(0) = 1$ and $w(\infty) = 0$. We primarily apply the initial approximation in (3.3), i.e., (3.7) with $w \equiv 1$, but we find situations in which alternatives in (3.7) can be important. Our somewhat vague specification allows room for refinement.

Let $Q$ denote the diffusion process with infinitesimal parameters in (3.1) and (3.6) and a reflecting upper barrier at $\kappa$ if $\kappa < \infty$. Note that this diffusion process coincides with the diffusion process $Q$ arising as the limit in Theorem 2.1 in the special case of the $GI/H_2/n/m$ model, where $v$ is as given in (3.7) with $w \equiv 1$. Just as in the $H_2$-service case in (2.7), in general $v \neq z$, so that the infinitesimal variance in (3.6) is discontinuous at 0.

For the queue-length process $Q_n$ in the $G/GI/n/m$ model, the approximation is

$$\{Q_n(t) : t \geq 0\} \approx n + \sqrt{n}\{Q(t) : t \geq 0\} ;$$  \hspace{1cm} (3.8)

i.e., we are acting as if the stochastic-process limit in Theorem 2.1 were valid with the scaling in (2.4), with only the infinitesimal variance generalized from (2.7) to (3.6).

There are two ideas being advanced here: (1) a diffusion process of the general form derived in Section 2, i.e., with infinitesimal parameters in (3.1) and (3.6), may produce a satisfactory approximation, and (2) the parameters $z$ and $v$ appearing in the infinitesimal-variance formula (3.6) may be given by the asymptotic peakedness in (1.6) and (3.7) with appropriate weight function $w$, often $w \equiv 1$. 

13
4. The Steady-State Distribution of the Diffusion Process

From the form of the infinitesimal parameters in (2.6), (2.7), (3.1) and (3.6), we recognize that the diffusion process $Q$ in Theorem 2.1 and Section 3 is a piecewise-linear diffusion, as in Browne and Whitt (1995). Thus, we can immediately write down the limiting (as $t \to \infty$) steady-state distribution when it exists. It is easy to see that $Q(t) \Rightarrow Q(\infty)$ for a proper random variable $Q(\infty)$ if and only if either $\kappa < \infty$ or $\kappa = \infty$ and $\beta > 0$.

Since we are allowing a finite waiting room, we need to generalize the asymptotic-delay-probability function $\alpha$ in (0.2). We now let $\alpha$ be the following function of the two variables $\beta$ and $\kappa$ obtained from the limits in (0.1) and (0.3):

$$\alpha \equiv \alpha(\beta, \kappa) \equiv [1 + \beta \Phi(\beta)/\phi(\beta)(1 - e^{-\kappa \beta})]^{-1} \quad \text{for} \quad \beta \neq 0 . \quad (4.1)$$

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

$$\alpha_0 \equiv \alpha_0(\kappa) \equiv [1 + \kappa^{-1}\sqrt{\pi/2}]^{-1} . \quad (4.2)$$

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 $Q(\infty)$ 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 $\alpha$, is determined by a relation between the right and left limits of the pdf at 0, namely,

$$f(0+) = \frac{\sigma^2(0-)}{\sigma^2(0+)} f(0-) , \quad (4.3)$$

where

$$\frac{\sigma^2(0-)}{\sigma^2(0+)} = \frac{z}{v} \quad (4.4)$$

for $z$ and $v$ in (3.6); see (18.26) of Browne and Whitt (1995).

**Theorem 4.1.** Let $Q$ be the reflected diffusion process with infinitesimal parameters in (3.1) and (3.6). Suppose that either $\kappa < \infty$ or $\kappa = \infty$ and $\beta > 0$, so that $Q(t) \Rightarrow Q(\infty)$ as $t \to \infty$, where $Q(\infty)$ is a proper random variable.

If $\beta \neq 0$, then

$$P(Q(\infty) \geq 0) = \alpha(\beta/\sqrt{z}, \kappa \sqrt{z}/v) , \quad (4.5)$$

$$P(Q(\infty) \leq x | Q(\infty) \leq 0) = \Phi((x + \beta)/\sqrt{z})/\Phi(\beta/\sqrt{z}) \quad (4.6)$$
and
\[ P(Q(\infty) > x | Q(\infty) \geq 0) = \frac{e^{-x\beta/v} - e^{-\kappa\beta/v}}{1 - e^{-\kappa\beta/v}}, \quad 0 \leq x < \kappa, \tag{4.7} \]
where \( \alpha \) is the M/M/n/m asymptotic-delay-probability function in (4.1), and \( \beta \) and \( \kappa \) are the limits in (0.1) and (0.3).

If \( \beta = 0 \), then
\[ P(Q(\infty) \geq 0) = \alpha_0(\kappa \sqrt{z}/v) \tag{4.8} \]
for \( \alpha_0 \) in (4.2). Then
\[ P(Q(\infty) > x | Q(\infty) \geq 0) = (\kappa - x)/\kappa, \quad 0 \leq x < \kappa, \tag{4.9} \]
while formula (4.6) remains unchanged.

**Corollary 4.1.** If, in addition to the conditions of Theorem 4.1, \( \kappa = \infty \) and \( \beta < 0 \), then both
\[ P(Q(\infty) > 0) = \alpha_\infty(\beta/\sqrt{z}), \tag{4.10} \]
and
\[ P(Q(\infty) \leq x | Q(\infty) \leq 0) = \Phi((x + \beta)/\sqrt{z})/\Phi(\beta/\sqrt{z}), \tag{4.11} \]
independent of the parameter \( v \) in (3.6).

Corollary 4.1 suggests that the approximations for the delay probability and the conditional distribution of the number of busy servers given that all servers are not busy in the \( G/GI/n/\infty \) model in (1.5) should be quite robust approximations, which we find to be the case.

**Remark 4.1.** The pdf. The steady-state distribution of the diffusion process \( Q \) can also be characterized by its pdf. If \( \beta \neq 0 \), \( Q(\infty) \) has the pdf
\[ f(x) = \begin{cases} (1 - \alpha)\phi((x + \beta)/\sqrt{z})/\sqrt{z}\Phi(\beta/\sqrt{z}), & x < 0, \\ \alpha_0 e^{-x\beta/v}/v(1 - e^{-\kappa\beta/v}), & 0 \leq x \leq \kappa, \end{cases} \tag{4.12} \]
for \( \alpha \equiv \alpha(\beta/\sqrt{z}, \kappa \sqrt{z}/v) \) in (4.5), for \( z \) and \( v \) in (3.6). If \( \beta = 0 \), then \( Q(\infty) \) has the pdf
\[ f(x) = \begin{cases} (1 - \alpha_0)\phi((x + \beta)/\sqrt{z})/\sqrt{z}\Phi(\beta/\sqrt{z}), & x < 0, \\ \alpha_0/\kappa, & 0 \leq x \leq \kappa, \end{cases} \tag{4.13} \]
for \( \alpha_0 \equiv \alpha_0(\sqrt{z}/v) \) in (4.8), \( z \) and \( v \) in (3.6). \( \blacksquare \)

**Remark 4.2.** Understanding the asymptotic-delay-probability formula. The asymptotic-delay-probability function \( \alpha \) in (1.1) can be understood by observing an underlying alternating-renewal-process structure, which plays a key role in the proof of Theorem 2.1 in Whitt (2002b).
The queue-length process alternates between periods spent above level \( n \) (an “above” time \( X^a_n \)) and periods spent below level \( n - 1 \) (a “below” time \( X^b_n \)). This structure is most straightforward in the case \( M/M/n/\infty \), so consider that case. Since \( Q_n(t) \) is a Markov process, these times are mutually independent. Thus, by a well-known alternating-renewal-process result,

\[
P(Q_n(\infty) > n) = \frac{EX^a_n}{EX^a_n + EX^b_n} = [1 + (EX^b_n/EX^a_n)]^{-1}. \quad (4.14)
\]

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

\[
EX^a_n = \frac{1}{n\mu(1 - p)} \sim \frac{1}{\sqrt{n}\mu\beta} \quad \text{as} \quad n \to \infty , \quad (4.15)
\]

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

\[
1/\lambda_n EX^b_n = \pi_n \sim (1/\sqrt{n})\phi(\beta)/\Phi(\beta) \quad \text{as} \quad n \to \infty \quad (4.16)
\]

under condition (0.1); e.g., see (15) of Srikant and Whitt (1996) and the appendix of Whitt (1984). Combining (4.14)–(4.16), we obtain convergence to \( \alpha \) in (0.2) in the limiting regime (0.1). \( \blacksquare \)

**Remark 4.3.** *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 diffusion process 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 \( z \to 1 \) and \( np \to 1 \). If \( 0 < \kappa < \infty \), then

\[
\alpha_p \sim p\kappa \phi(\beta)/\Phi(\beta) \quad \text{as} \quad p \to 0 ; \quad (4.17)
\]

if \( \kappa = \infty \), then

\[
\alpha_p \to \alpha(\beta, \infty) \quad \text{as} \quad p \to 0 . \quad (4.18)
\]

Thus, the \( GI/H^2_n/\infty \) asymptotic delay probability approaches the \( M/M/n/\infty \) asymptotic delay probability as \( p \downarrow 0 \) for any interarrival-time distribution.

On the other hand, when \( \kappa = \infty \),

\[
E[Q(\infty)p|Q(\infty)p > 0] = \frac{c^2 + (2/p) - 1}{2\beta}, \quad (4.19)
\]
so that, then
\[ E[Q(\infty)_p|Q(\infty)_p > 0] \sim 1/p\beta \quad \text{as} \quad p \to 0. \] (4.20)

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, since \( c_{\alpha}^2 = (2/p) - 1 \) and \( \eta(H_2^*(p)) = p/2 \),
\[ z_p = 2 - p \to 2 \quad \text{as} \quad p \downarrow 0 . \] (4.21)

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. ■

**Remark 4.4.** Rough approximations of the asymptotic peakedness. We can obtain further rough approximations of the peakedness \( z \) in terms of the variability parameters \( c_{\alpha}^2 \) 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 \( c_{\alpha}^2 \) is near 1. However, when \( c_{\alpha}^2 \) 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.

Since \( \eta(G) = 1 \) 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 SCV’s inbetween:

\[ \eta(c_s^2) \approx 1 - (c_s^2/2) \quad \text{and} \quad z(c_{\alpha}^2, c_s^2) \approx 1 + (c_s^2 - 1)(1 - (c_s^2/2)), \quad 0 \leq c_s^2 \leq 1 . \] (4.22)

To treat distributions with \( c_s^2 \geq 1 \), we can use \( H_2 \) distributions with balanced means \( (H_2^b) \). An \( H_2 \) distribution with mean \( 1/\mu \) has pdf
\[ h(x) = p_1 e^{-\mu_1 x} + p_2 e^{-\mu_2 x}, \quad x \geq 0 , \] (4.23)
where \( 0 \leq p_1 \leq 1, p_1 + p_2 = 1 \) 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
\[ \frac{2p_1}{\mu_1} = \frac{2p_2}{\mu_2} = \frac{1}{\mu} . \] (4.24)
which implies that
\[ p_i = \left[ 1 \pm \sqrt{\frac{c_s^2 - 1}{c_s^2 + 1}} \right]/2. \]  

(4.25)

For this $H_2^h$ case, $\eta(H_2^h) = (c_s^2 + 3)/(c_s^2 + 1)$, so that we obtain the general approximation
\[ z(c_s^2, c_s^2) \approx z(c_s^2, H_2^h) = 1 + \frac{(c_s^2 - 1)(c_s^2 + 3)}{4(c_s^2 + 1)} \quad \text{for} \quad c_s^2 \geq 1. \]  

(4.26)

Note that $\eta(H_2^h)$ increases to 1/2 as $c_s^2$ decreases to its lower limit $c_s^2 = 1$, which is the exponential distribution, while $\eta(H_2^h)$ decreases to 1/4 as $c_s^2 \uparrow \infty$. Other $H_2$ distributions without balanced means can have arbitrarily small values of $\eta$, as we saw for $H_2^*$ in (2.10).

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

We start by evaluating the approximations for the delay probability ($P(Q_n^a(\infty) \geq n) \equiv PW$) and the probability all servers are busy ($P(Q_n(\infty) \geq n) \equiv PB$) 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 in Theorems 2.1 and 4.1 imply that the Arrivals-See-Time-Averages (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.

Since we are working in the asymptotic regime (0.1), the natural approximation based on (1.2) is $P(Q_n(\infty) \geq n) \approx \alpha(\beta/\sqrt{n})$ for $\alpha$ in (0.2), $z = (c_s^2 + 1)/2$ and
\[ \beta = \sqrt{n}(1 - \rho). \]  

(5.1)

Indeed, we have been implicitly acting as if the value for $\beta$ based on the limit in (0.1) is (5.1) and that is what we usually use. However, we might approximate $\beta$ differently. As discussed in Whitt (1992), since it is the offered load that is random rather than the number of servers, it is natural to think of
\[ (\lambda/\mu) + \beta(\sqrt{\lambda/\mu}) \approx n \]  

(5.2)
rather than (5.1). Approximation (5.2) leads to the alternative approximation for $\beta$,
\[ \beta \approx \frac{\sqrt{n}(1 - \rho_n)}{\sqrt{\rho_n}}. \]  

(5.3)

Of course, in the limiting regime (0.1), the two specifications for $\beta$ 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, Tijms and van Hoorn (1985). The results are displayed below 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 $c^2_n = 1$ in Table 1.

Since the delay probability in the GI/M/n/$\infty$ model can differ significantly from the corresponding $M/M/n/\infty$ formula, the refinement provided by (1.1) is important when the arrival process is not nearly Poisson. When the arrival process is not Poisson, $PW \neq PB$. Half their difference provides a lower bound on the worst possible error for these two quantities we could possibly achieve in any approximation. In the heavy-traffic regime involving large $n$ and $\rho$, the difference between $PW$ and $PB$ is not great and the approximation tends to perform well. However, when $n$ or $\rho$ is small, $PW$ and $PB$ get further apart and the quality of the approximation deteriorates.

As approximations for $PW$ and $PB$ in Table 1, we plot the approximation $\alpha(\beta/\sqrt{\pi})$ in (1.1) based on both the standard specification of $\beta$ in (5.1) and the alternative in (5.3). The modification in (5.3) always increases $\beta$ and thus reduces $\alpha(\beta/\sqrt{\pi})$. 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.

In Table 1 we also compare a heavy-traffic approximation for the mean number waiting, $E[(Q_n(\infty) - n)^+]$ (using (5.1) for $\beta$) 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.,

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

(5.4)

The direct heavy-traffic approximation based on (4.7) is

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

(5.5)

for $v$ in (3.7). When the service-time distribution is exponential,

$$v = (c^2_n + 1)/2$$

(5.6)

for $v$ in (3.7) 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 $\rho$ and still
keep it asymptotically correct, we multiply the approximation in (5.5) by \( \rho \) to get

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

(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

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

(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 there for \( D/M/n/\infty \) and \( H_2^2/M/n/\infty \) 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 non-exponential 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.

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 \( c_n^2 \) is small. The general \( G/GI/n/\infty \) approximation does significantly better than the \( G/M/n/\infty \) approximation when \( c_n^2 \) is small.

For the mean number waiting, we let the variability parameter \( v \) be as in (3.7) with weight \( w \equiv 1 \); i.e., here \( v = (c_s^2 + c_v^2)/2 \). 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 (3.7). The direct approximation (with \( w = 1 \)) performs remarkably well when both \( c_n^2 > 1 \) and \( c_s^2 < 1 \). Overall, the approximations in Table 2 seem sufficiently accurate to be quite useful.
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Delay Probability</th>
<th>Mean Number Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact</td>
<td>approximations</td>
</tr>
<tr>
<td>$n$</td>
<td>$\rho$</td>
<td>$c_n^2$</td>
</tr>
<tr>
<td>200</td>
<td>0.98</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Delay Probability</th>
<th>Mean Number Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact</td>
<td>approximations</td>
</tr>
<tr>
<td>$n$</td>
<td>$\rho$</td>
<td>$c_n^2$</td>
</tr>
<tr>
<td>100</td>
<td>0.98</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.88</td>
<td>4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Delay Probability</th>
<th>Mean Number Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact</td>
<td>approximations</td>
</tr>
<tr>
<td>$n$</td>
<td>$\rho$</td>
<td>$c_n^2$</td>
</tr>
<tr>
<td>25</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>2.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.50</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.10</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.10</td>
<td>0.90</td>
<td>4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Delay Probability</th>
<th>Mean Number Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact</td>
<td>approximations</td>
</tr>
<tr>
<td>$n$</td>
<td>$\rho$</td>
<td>$c_n^2$</td>
</tr>
<tr>
<td>8</td>
<td>0.98</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.70</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.70</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.70</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.70</td>
<td>4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Delay Probability</th>
<th>Mean Number Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact</td>
<td>approximations</td>
</tr>
<tr>
<td>$n$</td>
<td>$\rho$</td>
<td>$c_n^2$</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.60</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.60</td>
<td>4.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.60</td>
<td>4.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.60</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 1: A comparison of the GI/M/n/∞ 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 $n$, $\rho$ and interarrival-time SCV $c_n^2$. (The peakedness is thus $z = (c_n^2 + 1)/2$.) 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, Tijms and van Hoorn (1985).
<table>
<thead>
<tr>
<th>( c_2^2 )</th>
<th>( c_2^3 )</th>
<th>( z )</th>
<th>( PW )</th>
<th>( PB )</th>
<th>( G/I/M/n/\infty )</th>
<th>( E[(Q(\infty) - n)^+] )</th>
<th>( \text{approximation} )</th>
<th>( \text{exact} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>0.10</td>
<td>.058</td>
<td>.103</td>
<td>.071</td>
<td>.386</td>
<td>0.037</td>
<td>0.032</td>
</tr>
<tr>
<td>0.5</td>
<td>0.50</td>
<td>.336</td>
<td>.360</td>
<td>.366</td>
<td>.449</td>
<td>0.89</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>.479</td>
<td>.479</td>
<td>.505</td>
<td>.505</td>
<td>2.41</td>
<td>2.27</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>2.00</td>
<td>.615</td>
<td>.593</td>
<td>.624</td>
<td>.577</td>
<td>5.67</td>
<td>5.62</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.00</td>
<td>.730</td>
<td>.688</td>
<td>.721</td>
<td>.658</td>
<td>12.9</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.44</td>
<td>.308</td>
<td>.349</td>
<td>.338</td>
<td>.386</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>0.5</td>
<td>0.69</td>
<td>.415</td>
<td>.434</td>
<td>.431</td>
<td>.449</td>
<td>2.00</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>.500</td>
<td>.500</td>
<td>.505</td>
<td>.505</td>
<td>3.51</td>
<td>3.41</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.62</td>
<td>.601</td>
<td>.580</td>
<td>.591</td>
<td>.577</td>
<td>6.66</td>
<td>6.65</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>2.87</td>
<td>.706</td>
<td>.663</td>
<td>.677</td>
<td>.658</td>
<td>13.5</td>
<td>13.7</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>2.5</td>
<td>0.65</td>
<td>.420</td>
<td>.447</td>
<td>.419</td>
<td>.386</td>
<td>4.39</td>
<td>4.90</td>
</tr>
<tr>
<td>0.5</td>
<td>0.80</td>
<td>.472</td>
<td>.485</td>
<td>.462</td>
<td>.449</td>
<td>5.66</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>.522</td>
<td>.522</td>
<td>.505</td>
<td>.505</td>
<td>7.32</td>
<td>7.95</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.39</td>
<td>.594</td>
<td>.574</td>
<td>.565</td>
<td>.577</td>
<td>10.8</td>
<td>11.4</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>2.18</td>
<td>.680</td>
<td>.637</td>
<td>.637</td>
<td>.658</td>
<td>18.0</td>
<td>18.6</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>4.0</td>
<td>0.69</td>
<td>.441</td>
<td>.465</td>
<td>.431</td>
<td>.386</td>
<td>6.89</td>
<td>7.95</td>
</tr>
<tr>
<td>0.5</td>
<td>0.83</td>
<td>.485</td>
<td>.498</td>
<td>.468</td>
<td>.449</td>
<td>8.23</td>
<td>9.48</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>.529</td>
<td>.529</td>
<td>.505</td>
<td>.505</td>
<td>9.95</td>
<td>11.4</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.35</td>
<td>.592</td>
<td>.575</td>
<td>.559</td>
<td>.577</td>
<td>13.3</td>
<td>15.1</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>2.05</td>
<td>.672</td>
<td>.633</td>
<td>.628</td>
<td>.658</td>
<td>20.2</td>
<td>22.6</td>
<td></td>
</tr>
</tbody>
</table>

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 \( c_2^3 \) as 1) with exact values of the probability of delay \( (PW) \) and the probability all servers are busy \( (PB) \) in the \( G/GI/n/\infty \) model with non-exponential service-time distributions for \( n = 25, \rho = 0.9 \) and several values of the interarrival-time SCV \( c_2^2 \) and service-time SCV \( c_3^2 \). Also evaluated is the approximation for the mean number waiting in (5.7) and (3.7) with \( w = 1 \). The exact values come from Seelen, Tijms and van Hoorn (1985).
6. Simulations

In order 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 $c_v^2 = 0.5$. We also want to consider non-renewal arrival processes and other service-time distributions.

We consider two non-renewal 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 5 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 $c_n^2 = 0$ in the FCLT for the arrival process, as in (2.1) and (2.2).

Our second non-renewal process is the independent superposition of 4 IID $H_2^b$ renewal processes, denoted by $sup4H_2^b$. We let the SCV be $c_s^2 = 19$ in each component process in order to match the SCV of the $H_2^b$ 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). Given the arrival times and departure times, we construct the queue-length process at state-change times using the method on p.210 of SPL; 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. Since 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 non-renewal processes $Db5$ ($c_a^2 = 0$) and $sup4H^b_2$ ($c_a^2 = 19$) introduced above. The four renewal processes had interarrival times distributed as $D$, $M$, $H^*_2$ with parameter $p = 0.1$ ($c^*_a = 19$) and $LN$ with $c^*_a = 19$. The five service-time distributions are $D$, $M$, $H^*_2$ with $p = 0.1$ ($c^*_a = 19$), $LN$ with $c^*_a = 19$ and $Par(3/2)$, which has infinite variance. The variability parameters $z$ in (1.5) and $v$ in (3.7) 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 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 non-renewal 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 $sup4H^b_2$ arrival process are similar to the renewal arrival processes with $c^*_a = 19$.

The quality of the approximations for $PB$ are consistently good with 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 $w = 1$ in formula (3.7) 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 $Par(3/2)$ service-time distributions.

Indeed, the low simulation values with $LN$ and $Par(3/2)$ service-time distributions are remarkable. The approximation for $LN$ service times can be improved dramatically if we use (3.7) 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 $H_2^n$, $LN$ and $sup4H^b_2$ arrivals.

Since the Pareto(3/2) service-time distribution has an infinite variance, $c^2_s = \infty$, so the approximation in (3.7) for $v$ in (3.6) makes no sense. Based on (3.7), we would expect the queue-length process to be unstable, but that evidently is 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 $H_2^n$, $LN$ and $sup4H^b_2$ 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. Since 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 3 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.

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

We now apply the diffusion approximation in Section 3 to generate an approximation for the blocking probability in the $G/GI/n/m$ queue. Since the diffusion process has a reflecting barrier at $\kappa$, 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 Berge 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 $\kappa$) of the diffusion process $Q$ as

$$r_Q \equiv f(\kappa) \frac{\sigma^2(\kappa)}{2} = f(\kappa) \mu v,$$  \hspace{1cm} (7.1)
<table>
<thead>
<tr>
<th>arrival process</th>
<th>$c_a^2 = 0$</th>
<th>$c_a^2 = 1$</th>
<th>$H_2^* = 19$</th>
<th>$c_a^2 = 19$</th>
<th>$c_a^2 = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$z$</td>
<td>$v$</td>
<td>$0.50$</td>
<td>$0.98$</td>
<td>$0.78$</td>
</tr>
<tr>
<td>$Db5$</td>
<td>$z$</td>
<td>$v$</td>
<td>$0.50$</td>
<td>$9.50$</td>
<td>$9.50$</td>
</tr>
<tr>
<td>$M$</td>
<td>$z$</td>
<td>$v$</td>
<td>$1.00$</td>
<td>$1.00$</td>
<td>$1.00$</td>
</tr>
<tr>
<td>$H_2$</td>
<td>$z$</td>
<td>$v$</td>
<td>$19.00$</td>
<td>$10.00$</td>
<td>$1.90$</td>
</tr>
<tr>
<td>$LN$</td>
<td>$z$</td>
<td>$v$</td>
<td>$19.00$</td>
<td>$10.00$</td>
<td>$1.90$</td>
</tr>
<tr>
<td>$sup4H_2^b$</td>
<td>$z$</td>
<td>$v$</td>
<td>$19.00$</td>
<td>$10.00$</td>
<td>$1.90$</td>
</tr>
<tr>
<td>integral</td>
<td>$\eta(G)$</td>
<td>$1.00$</td>
<td>$0.50$</td>
<td>$0.05$</td>
<td>$0.22$</td>
</tr>
</tbody>
</table>

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 (3.7) with $w \equiv 1$ ($v = (c_a^2 + c_\alpha^2)/2$), where $c_\alpha^2$ is understood to be the scale factor in the FCLT for the arrival process. 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>
<thead>
<tr>
<th>arrival process</th>
<th>$c_a^2 = 0$</th>
<th>$c_a^2 = 1$</th>
<th>$H_2^* = 19$</th>
<th>$c_a^2 = 19$</th>
<th>$c_a^2 = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$ approx.</td>
<td>$-.029$</td>
<td>$.098$</td>
<td>$.072$</td>
<td>$.067$</td>
<td></td>
</tr>
<tr>
<td>$Db5$ approx.</td>
<td>$-.024$</td>
<td>$.096$</td>
<td>$.070$</td>
<td>$.049$</td>
<td></td>
</tr>
<tr>
<td>$M$ approx.</td>
<td>$.105$</td>
<td>$.105$</td>
<td>$.105$</td>
<td>$.105$</td>
<td>$.105$</td>
</tr>
<tr>
<td>$H_2$ approx.</td>
<td>$.654$</td>
<td>$.549$</td>
<td>$.219$</td>
<td>$.416$</td>
<td>$.437$</td>
</tr>
<tr>
<td>$LN$ approx.</td>
<td>$.654$</td>
<td>$.549$</td>
<td>$.219$</td>
<td>$.416$</td>
<td>$.437$</td>
</tr>
<tr>
<td>$sup4H_2^b$ approx.</td>
<td>$.654$</td>
<td>$.549$</td>
<td>$.219$</td>
<td>$.416$</td>
<td>$.437$</td>
</tr>
<tr>
<td>integral</td>
<td>$\eta(G)$</td>
<td>$1.00$</td>
<td>$0.50$</td>
<td>$0.05$</td>
<td>$0.22$</td>
</tr>
</tbody>
</table>

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. 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 $c_a^2 = 19$. 

26
<table>
<thead>
<tr>
<th>arrival process</th>
<th>$D$</th>
<th>$M$</th>
<th>$H^*_2$</th>
<th>$LN$</th>
<th>$\text{Par}(3/2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$ mean approx.</td>
<td>$c^2 = 0$</td>
<td>$c^2 = 1$</td>
<td>$c^2 = 19$</td>
<td>$c^2 = 19$</td>
<td>$c^2 = \infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>-</td>
<td>3.3</td>
<td>63.5</td>
<td>63.5</td>
<td>$\infty$</td>
</tr>
<tr>
<td>SD sim.</td>
<td>-</td>
<td>2.8</td>
<td>59.5</td>
<td>14.3</td>
<td>9.6</td>
</tr>
<tr>
<td>$Db5$ mean approx.</td>
<td>-</td>
<td>3.3</td>
<td>63.5</td>
<td>63.5</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>-</td>
<td>2.8</td>
<td>54.1</td>
<td>13.9</td>
<td>9.3</td>
</tr>
<tr>
<td>SD sim.</td>
<td>-</td>
<td>3.1</td>
<td>52.8</td>
<td>20.7</td>
<td>12.7</td>
</tr>
<tr>
<td>$M$ mean approx.</td>
<td>3.3</td>
<td>6.7</td>
<td>66.9</td>
<td>66.9</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>4.7</td>
<td>6.4</td>
<td>60.3</td>
<td>17.8</td>
<td>12.8</td>
</tr>
<tr>
<td>SD sim.</td>
<td>4.5</td>
<td>6.7</td>
<td>55.2</td>
<td>26.1</td>
<td>15.5</td>
</tr>
<tr>
<td>$H^*_2$ mean approx.</td>
<td>63.5</td>
<td>66.9</td>
<td>127.0</td>
<td>127.0</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>79.4</td>
<td>75.8</td>
<td>114.4</td>
<td>63.2</td>
<td>64.9</td>
</tr>
<tr>
<td>SD sim.</td>
<td>76.1</td>
<td>76.8</td>
<td>140.5</td>
<td>74.6</td>
<td>80.3</td>
</tr>
<tr>
<td>$LN$ mean approx.</td>
<td>63.5</td>
<td>66.9</td>
<td>127.0</td>
<td>127.0</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>46.6</td>
<td>46.6</td>
<td>111.7</td>
<td>49.0</td>
<td>40.7</td>
</tr>
<tr>
<td>SD sim.</td>
<td>39.4</td>
<td>47.0</td>
<td>112.6</td>
<td>57.1</td>
<td>53.3</td>
</tr>
<tr>
<td>$sup4H^*_2$ mean approx.</td>
<td>63.5</td>
<td>66.9</td>
<td>127.0</td>
<td>127.0</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mean sim.</td>
<td>65.9</td>
<td>64.4</td>
<td>120.0</td>
<td>63.1</td>
<td>51.6</td>
</tr>
<tr>
<td>SD sim.</td>
<td>64.4</td>
<td>66.5</td>
<td>127.8</td>
<td>79.4</td>
<td>60.0</td>
</tr>
</tbody>
</table>

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. The approximations use (5.7) (divided by $\alpha$) with $v$ in (3.7) and $w = 1$. The process $Db5$ is a deterministic process with clusters of size 5, while $sup4H^*_2$ is the superposition of four IID renewal processes with $H^*_2$ interarrival times with SCV $c^2 = 19$. 

27
Figure 3: 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 $n = 115$. The service times are exponential ($M$), lognormal ($LN$), Pareto($3/2$) and $H^*_2$. The estimates are obtained from the Splplus nonparametric density estimator based on $10^6$ arrivals in each case.
where \( f \) is the pdf of \( Q(\infty) \) in (4.12) or (4.13), \( \sigma^2(\kappa) \) is the infinitesimal variance in (3.6) evaluated at the upper boundary \( \kappa \) and \( v \) is the variability parameter in (3.6) and (3.7).

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

\[
r_{Q_n} \approx \sqrt{n}r_Q .
\]

Since the blocking probability equals the loss rate divided by the arrival rate, we approximate the blocking probability in the queueing system, denoted by \( \pi_n \), by

\[
\pi_n = \frac{r_{Q_n}}{\lambda_n} \approx \frac{f(\kappa)v}{\rho_n \sqrt{n}} .
\]

**Remark 7.1.** A conjectured local limit. We conjecture that the approximation in (7.3) 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

\[
\sqrt{n}\pi_n = P(Q_n^\kappa(\infty) = n + m_n) \to f(\kappa)v
\]

as \( n \to \infty \) for \( v = (c_2^2 + c_3^2)/2 \). That is in the spirit of Theorem 15 on p. 226 of Borovkov (1976). □

**Remark 7.2.** Comparison with \( G/GI/n/0 \) 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

\[
r_Q \equiv f(\kappa)\frac{\sigma^2(\kappa)}{2} = f(\kappa)\mu z .
\]

The first relation in (7.5) is the same as in (7.1), but the second is different because now the infinitesimal variance \( \sigma^2(\kappa) \) is different.

From (4.12), we see that

\[
f(0) = \frac{\phi(\beta/\sqrt{z})}{\sqrt{z}\Phi(\beta/\sqrt{z})}
\]

when \( \beta \neq 0 \). Hence we obtain the blocking-probability approximation

\[
\pi_n \approx \frac{\sqrt{z}\phi(\beta/\sqrt{z})}{\rho \sqrt{\pi}\Phi(\beta/\sqrt{z})} .
\]

Formula (7.7) is \( \sqrt{p} \) times the approximation in (15) of Srikant and Whitt (1996). That difference is removed if we apply approximation (5.3). Moreover, it is asymptotically negligible in the limiting regime (0.1). □
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 (3.7) 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 $c_s^2$, $c_n^2$ and $\rho$. 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

$$PW = PW_S + PBL - (PW_S)(PBL).$$

(7.8)

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 not too much accuracy is required in many applications. A principle application is server staffing. In that application, great accuracy is not required 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 $\pi_n$ 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 $\lambda = 22.5$.

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 (4.22), we let the approximate peakedness be $\varepsilon \approx 0.625$. The more bursty examples are $H_2^b/E_2/n/m$ with $c_n^2 = 4.0$ 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 $n = 22$
<table>
<thead>
<tr>
<th>parameters</th>
<th>exact values</th>
<th>approximations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c^2_a$</td>
<td>$c^2_s$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>4.00</td>
<td>1.00</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.800</td>
<td>.743</td>
</tr>
<tr>
<td>1.0</td>
<td>.628</td>
<td>.561</td>
</tr>
<tr>
<td>0.9</td>
<td>.502</td>
<td>.436</td>
</tr>
<tr>
<td>0.8</td>
<td>.353</td>
<td>.297</td>
</tr>
<tr>
<td>0.7</td>
<td>.205</td>
<td>.165</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.913</td>
<td>.913</td>
</tr>
<tr>
<td>1.0</td>
<td>.649</td>
<td>.649</td>
</tr>
<tr>
<td>0.9</td>
<td>.414</td>
<td>.414</td>
</tr>
<tr>
<td>0.8</td>
<td>.195</td>
<td>.195</td>
</tr>
<tr>
<td>0.7</td>
<td>.063</td>
<td>.063</td>
</tr>
<tr>
<td>0.25</td>
<td>1.00</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.965</td>
<td>.970</td>
</tr>
<tr>
<td>1.0</td>
<td>.680</td>
<td>.705</td>
</tr>
<tr>
<td>0.9</td>
<td>.355</td>
<td>.384</td>
</tr>
<tr>
<td>0.8</td>
<td>.110</td>
<td>.126</td>
</tr>
<tr>
<td>0.7</td>
<td>.0182</td>
<td>.023</td>
</tr>
<tr>
<td>4.00</td>
<td>0.50</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.812</td>
<td>.757</td>
</tr>
<tr>
<td>1.0</td>
<td>.648</td>
<td>.581</td>
</tr>
<tr>
<td>0.9</td>
<td>.527</td>
<td>.459</td>
</tr>
<tr>
<td>0.8</td>
<td>.382</td>
<td>.321</td>
</tr>
<tr>
<td>0.7</td>
<td>.230</td>
<td>.186</td>
</tr>
<tr>
<td>1.00</td>
<td>0.50</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.944</td>
<td>.944</td>
</tr>
<tr>
<td>1.0</td>
<td>.681</td>
<td>.681</td>
</tr>
<tr>
<td>0.9</td>
<td>.427</td>
<td>.427</td>
</tr>
<tr>
<td>0.8</td>
<td>.194</td>
<td>.194</td>
</tr>
<tr>
<td>0.7</td>
<td>.061</td>
<td>.061</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>1.5</td>
</tr>
<tr>
<td>1.2</td>
<td>.990</td>
<td>.992</td>
</tr>
<tr>
<td>1.0</td>
<td>.733</td>
<td>.758</td>
</tr>
<tr>
<td>0.9</td>
<td>.336</td>
<td>.369</td>
</tr>
<tr>
<td>0.8</td>
<td>.084</td>
<td>.101</td>
</tr>
<tr>
<td>0.7</td>
<td>.011</td>
<td>.0144</td>
</tr>
</tbody>
</table>

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 ($E_2$, $c^2_s = 0.5$) service-time distributions for $n = 25$, $m = 10$ and various values of the interarrival-time SCV $c^2_a$ and the traffic intensity $\rho$. The approximations have $v$ in (3.7) with $w \equiv 1$. The exact values come from Seelen, Tijms and van Hoorn (1985).
<table>
<thead>
<tr>
<th>parameters</th>
<th>approximate blocking probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>21</td>
<td>1.07</td>
</tr>
<tr>
<td>22</td>
<td>1.02</td>
</tr>
<tr>
<td>23</td>
<td>0.98</td>
</tr>
<tr>
<td>24</td>
<td>0.94</td>
</tr>
<tr>
<td>25</td>
<td>0.90</td>
</tr>
<tr>
<td>26</td>
<td>0.87</td>
</tr>
<tr>
<td>27</td>
<td>0.83</td>
</tr>
<tr>
<td>28</td>
<td>0.80</td>
</tr>
<tr>
<td>29</td>
<td>0.78</td>
</tr>
<tr>
<td>30</td>
<td>0.75</td>
</tr>
<tr>
<td>40</td>
<td>0.56</td>
</tr>
</tbody>
</table>

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 $\rho = 0.9$.

to $n = 24$, an addition of 2 servers. In contrast, for the $M/H_2/25/10$ model, we would go from $n = 26$ to $n = 30$, an addition of 4 servers.

8. Conclusions

We have developed a heuristic diffusion approximation for the $G/GI/n/m$ queue, which is supported by a heavy-traffic stochastic-process limit for the special case of the $G/H_2^n/n/m$ model, established in the companion paper Whitt (2002b). Thorem 4.1 shows that the diffusion approximation yields relatively simple explicit formulas for the steady-state performance measures of interest.

Corollary 4.1 shows that the approximations for the delay probability and the conditional distribution of the number of busy servers given that all servers are busy in the $G/GI/n/\infty$ model do not depend on the problematic variability parameter $v$ in (3.6) that operates in the region $x > 0$. Simulation experiments confirm that the diffusion 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 $G/GI/n/m$ models.

Indeed, even though the stochastic-process limit for $GI/PH/n/\infty$ models established by Puhalskii and Reiman (2000) proves that the overall diffusion process here cannot be asymptotically correct, it is conceivable that the approximations for the steady-state delay probability and the conditional steady-state queue-length distribution given that all servers are not busy could be asymptotically correct. That remains to be determined.
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 and the conditional steady-state queue-length distribution given that all servers are not busy. 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 $v = (c_1^2 + c_2^2)/2$ work 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 (3.7) 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, Nerman and Olsson (1996) and Feldmann and Whitt (1998).
References


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


*AT&T Bell Lab. Tech. J.* 63, 163-175.

Research* 9, 534-544.

Science* 38, 708–723.

2, 114–161.

AT&T Labs, Florham Park, NJ.


Whitt, W. 2002b. Heavy-traffic limits for the G/H_2^n/m queue. Department of Industrial 
Engineering and Operations Research, Columbia University.