ABSTRACT:
This paper provides an effective method of refining diffusion approximations for the queue length distribution of a GI/G/1 queue with finite or infinite waiting space. A tight discretization method is developed for a class of GI/G/1 queues to obtain a discrete queue length distribution at arbitrary moments from a continuous probability distribution of the approximating diffusion process. By tight we mean that there exists a queueing system yielding the discretized queue length distribution as its exact solution. To see the quality of diffusion approximations with tight discretization, they are numerically compared with previously established diffusion approximations, various heuristic approximations and the exact results for some particular cases.

1. INTRODUCTION AND SUMMARY
Diffusion approximations for queues have been recognized as basic approximation models in teletraffic theory as well as queueing theory. There are a number of interesting applications of diffusion approximations to general teletraffic models such as single server queues (11), many server queues (10,12) and networks of queues (20,27). Diffusion approximations also can be applied to the analysis and synthesis of systems with more complicated service mechanism, e.g., priority queues (1) and queues with removable server (12,18).

In heavy traffic situations, diffusion approximations can be often identified and justified by heavy traffic limit theorems: For a queueing characteristic process, e.g., the queue length process, heavy traffic limit theorems assure that the process in an unstable queue, when appropriately scaled and translated, converges weakly to a Brownian motion process; see (22). This implies that the heavy traffic limit theorems provide not only useful descriptions of unstable queues but also useful approximations of stable queues, especially, in heavy traffic. However, in moderate traffic, the limiting diffusion process is sometimes inappropriate as an approximation of the underlying process. In general, diffusion approximations become inaccurate as the traffic becomes light.

Another serious drawback of diffusion approximations is that they are not consistent with available exact results for particular cases. For example, consider a GI/G/1 queue. For the queue length process in the GI/G/1 queue, several different diffusion approximations have been proposed by many researchers; see (28,31). However, none of the approximations for the queue length distribution are consistent with any exact results, even with the M/M/1 queue. For the mean queue length, some of the diffusion approximations fail the consistency check with the M/M/1 queue, i.e., they do not coincide with the Pollaczek-Khintchine formula; see (31). The lack of consistency makes diffusion approximations unreliable with respect to the accuracy. Whitt (31) illustrated the need to refine diffusion approximations by consistency checks with bounds.

The purpose of this paper is to provide an effective method of refining diffusion approximations for the queue length distribution of a GI/G/1 queue with finite or infinite waiting space in such a way that the refined approximations are consistent with exact results for certain basic queueing systems. For ease of exposition, we consider the standard GI/G/1/N model with a single server, N-1 waiting places (1 ≤ N ≤ ∞), the FCFS (first-come first-served) discipline and i.i.d. (independent and identically distributed) service times (with a general distribution) that are independent of a renewal arrival process. It should be, however, noted that our method in this paper applies to any other queueing systems whenever their arrival and service processes do not depend on the queue length process; cf. (14). One may readily apply this method to bulk queues (2), queues with random service interruptions (3), etc.

Although a few refining methods of the diffusion approximations have been proposed so far, they only dealt with either modifications of the boundary conditions of the limiting diffusion process (8) or modifications of only the mean queue length (24); see also (9, Chapter 4). In this paper, we propose a tight discretization method for obtaining a "discrete" queue length distribution at arbitrary moments from the "continuous" probability distribution of the diffusion process. By tight we mean that there exists a queueing system yielding the discretized queue length distribution as its exact solution. We call this system a base of discretization.

The rest of this paper is organized as follows: In Section 2, we first focus on the GI/G/1/N queue (N < ∞). From a partial differential equation with certain boundary conditions, we derive a distribution function of a diffusion process approximating the queue length process. In Section 3, we apply the method of this paper to well-known queueing systems with finite or infinite waiting space and compare the results with available exact results.
length process in the GI/G/1/N queue, assuming that the system is in equilibrium state. Using a tight discretization method, we obtain a diffusion approximation for the queue length distribution of the system. In Section 2, we apply the discretization method to the GI/G/1 queue with unlimited waiting room, obtaining a simpler diffusion approximation for the queue length distribution. A class of PH/M/1 queues of which interarrival time distribution can be characterized by its first two moments is used as the base of discretization. Diffusion approximations for various congestion measures of the GI/G/1 queue, e.g., the mean waiting time and the coefficient of variation of interdeparture times, can be explicitly derived from the approximate queue length distribution.

To improve the accuracy of these diffusion approximations, we develop a heuristic modification of the diffusion parameters. In Section 4, to see the quality of the diffusion approximations with tight discretization, they are numerically compared with previously derived diffusion approximations, various heuristic approximations such as the Kraemer and Langenbach-Betz (21) approximation and the exact expressions for various congestion measures of the GI/G/1 queue.

3.1A-2-2
for $P_0^+$.

To obtain simple approximations for $P_0^+$, it is convenient to use an intimate relation between diffusion models for a two-staged cyclic queue and the $GI/G/1/N$ queue: Consider a cyclic queueing system which consists of two sequential stages; see Figure 1. This system has been analyzed by several researchers (e.g., (5,23)) as a model of multiprogrammed computer systems. We suppose that there are $N \geq 1$ programs in the central processing unit (CPU) - data transfer unit (DTU) cycle. Each program goes through both stages in sequence and then returns to the first stage. This process continues indefinitely. Programs in the CPU and DTU queues are served according to the FCFS discipline.

![Figure 1. Cyclic queueing system.](image)

If we assume here that (i) the sequence of DTU service or auxiliary memory access and transfer times is i.i.d. as with $(v_j)$; (ii) the sequence of CPU service or processing times is also i.i.d. as with $(u_j)$; and (iii) CPU and DTU processing times are mutually independent, then the queue length process at the CPU stage can be approximated by a diffusion process with the same diffusion parameters as with $X(t)$; see (6,7,8); Gaver and Shedler (6,7) used the same boundary conditions as (6), while Gelenbe (8) adopted a different kind of boundaries with jumps, both obtaining approximations for the CPU utilization $1 - P_0^+$; see also (16,17). Let $\pi_0$ be an approximation for $P_0^+$. Then, their results are rewritten as:

(a) Gaver and Shedler (6) approximation:

$$\pi_0 = \begin{cases} \frac{1 - \rho}{1 - \rho^N}, & \rho \neq 1 \\ \frac{1}{1 + \frac{2N}{c^2 + c^2}}, & \rho = 1; \end{cases}$$

(b) Gaver and Shedler (7) approximation:

$$\pi_0 = \begin{cases} \frac{1 - \rho}{1 - \rho(1 - \rho^{N-1})}, & \rho \neq 1 \\ \frac{1}{N + 1}, & \rho = 1; \end{cases}$$

(c) Gelenbe's (8) approximation:

$$\pi_0 = \begin{cases} \frac{1 - \rho}{1 - \rho^{2N-1}}, & \rho \neq 1 \\ \frac{1}{2 + \frac{N - 1}{c^2 + c^2}}, & \rho = 1; \end{cases}$$

where

$$\rho = e^Y = \exp\left(\frac{2(\rho - 1)}{pc^2 + c^2}\right).$$

From some numerical comparisons for particular cases in (6,7,8), we see that all of the approximations perform well, providing accuracy sufficient for most engineering applications. In particular, the approximation (c) is better than (a) when $N$ is small, and the approximation (b) is more accurate than (a) and (c) for the case $c > 1$. It is worth noting that all of the approximations give the exact result for $N = \infty$, i.e., $\lim_{N \to \infty} \pi_0 = 1 - \rho$ if $\rho < 1$.

However, there is a serious deficiency in these approximations: The approximations with $\rho$ in (12) do not coincide with any exact results, even with the $M/N/1/N$ queue. To deal with this deficiency, Gaver and Shedler (7) developed a distribution-dependent modification of $\rho$, based on Wald's identity. Let $\alpha(s) = E(e^{-s\nu})$ and $\beta(s) = E(e^{s\nu})$ for $Re s \geq 0$. Then, under a weak assumption on $u$, they proposed

$$\bar{\rho} = \beta(s),$$

where $\bar{s}$ is the unique positive solution of the equation

$$\alpha(-\bar{s})\beta(s) = 1.$$

When $u$ and $v$ are exponentially distributed, we have $\bar{\rho} = \rho$, so that all of the approximations for the $M/N/1/N$ queue yield the exact result

$$\pi_0 = p_0 = \begin{cases} \frac{1 - \rho}{1 - \rho N}, & \rho \neq 1 \\ \frac{1}{N + 1}, & \rho = 1. \end{cases}$$

It should be, however, noted that such consistency cannot be observed for other systems, even by the use of the approximation (13).

2.3. A Tight Discretization Method

From the continuous cdf $P(x)$ of the diffusion process $X(t)$, we will derive an approximate discrete distribution of the process $Q(t)$ in the $GI/G/1/N$ queue. Let $P_k = \lim_{t \to \infty} P(Q(t) = k | Q(0) = p_0^+)$ and $P_k$ be its diffusion approximation ($k = 0, \ldots, N$). We say that an approximate queue length distribution $\pi_k$ is tight if there exists a queueing system for which $P_k = \pi_k$ for all $k \geq 0$. We call this system a base of discretization.

Before introducing our discretization method for $P(x)$, we show a typical discretization method adopted by many researchers so far:

$$\pi_0 = P(0)$$

$$\pi_k = P(k) - P(k - 1), \quad k = 1, \ldots, N,$$

and hence

$$\pi_k = \frac{(1 - p_0)(1 - \rho)^{k-1}}{1 - \rho}, \quad k = 1, \ldots, N$$

with the approximation $p_0 \approx \pi_0$. As shown in Section 2.2, the approximate distribution (17) is
a tight one whose base is the $M/M/1/N$ queue if we approximate $\tilde{\rho}$ by (13). However, if we use the usual definition (12) for $\tilde{\rho}$, (17) is no longer a tight approximation. In what follows, we assume that $\tilde{\rho}$ is defined by (12) as usual.

Now suppose that we have an exact queue length distribution for a particular queueing system. Using this distribution, we develop a discretization method yielding a tight approximation whose base coincides with this system. Let $(p_k^*)$ and $P^*(x)$ be the exact queue length distribution at arbitrary moments and the cdf via diffusion approximation for the base system, respectively. Similarly, we attach the asterisk $\ast$ to any parameters of the base system. An essential idea in our discretization method is to shift base points of the discretization from the integer points in $I^n$, depending upon the distribution $(p_k^*)$: Define the following sequence $(x_k; k = 0, \ldots, N)$ as base points of the discretization; see Figure 2:

$$x_0 = 0, \quad x_N = N,$$

and

$$p^*(x_k) = \frac{\gamma_x}{\gamma_x + 1} P^*_1, \quad k = 1, \ldots, N-1,$$

from which

$$x_k = \left\{ \begin{array}{ll}
\frac{1}{\gamma_x + 1} \ln \frac{\gamma_x}{\gamma_x + 1} k, & \rho \neq 1 \\
\frac{N}{1 - P^*_0} \sum_{i=1}^{k-1} P^*_i, & \rho = 1,
\end{array} \right.$$}

for $k = 0, \ldots, N$, where

$$P^*_0 = 1 - \frac{1 - \exp(\gamma x)}{1 - P^*_0} \sum_{i=1}^{k-1} P^*_i.$$

Using these base points, we discretize the cdf $P(x)$ as

$$P_0 = P(0),$$

$$P_k = P(x_k) - P(x_{k-1}), \quad k = 1, \ldots, N.$$}

From (7), (19) and (21), we obtain

$$\lim_{N \to \infty} \frac{1 - P_0}{1 - P^*_0} (g_k^d - g_k^d), \quad \rho \neq 1$$

$$P_k = \left\{ \begin{array}{ll}
\frac{1 - \exp(\gamma x)}{1 - P^*_0} (g_k^d - g_k^d), & \rho \neq 1 \\
1 - P^*_0 - P^*_k, & \rho = 1,
\end{array} \right.$$}

for $k = 1, \ldots, N$, where

$$d = a^2/a.$$}

Here we need to approximate the empty probability $P_0$ for the system to be approximated. The easiest way of such an approximation is to set $P_0 \approx P^*_0$, from which it can be proved that (22) is a tight approximation, i.e., $P_k \approx P^*_k$ for all $k$. Although the approximation $P_0 \approx P^*_0$ does not provide a tight approximation, it may be still accurate enough for practical applications.

3. GI/G/1 QUEUE WITH UNLIMITED WAITING ROOM

In this section we consider the case $N = \infty$, i.e., the standard GI/G/1 queue with unlimited waiting room. Applying the tight discretization method to the GI/G/1 queue, we give a simple diffusion approximation for the queue length distribution. In particular, we use the GI/M/1 queue as a base of discretization, obtaining explicit approximations for several queueing characteristics and their asymptotic properties in heavy traffic.

3.1. Diffusion Model

The diffusion process $X(t)$ approximating $Q(t)$ in the GI/G/1 queue has the same diffusion parameters as with the case $N < \infty$ and the state space $I = (0, \infty)$. The equilibrium cdf of $X(t)$, $P(x)$, also satisfies the ordinary differential equation (5) with the boundary conditions

$$P(0) = 1 - \rho,$$

and

$$P(\infty) = 1,$$

see (11). It should be noted here that the boundary condition at the origin does not contain any undetermined parameters; cf. (6a). The condition (24a) reflects the exact empty probability of the GI/G/1 queue, which is valid even for stationary non-renewal arrival processes; see (4.4.2.3).

To assure the existence of the cdf $P(x)$, we assume $\rho < 1$, or equivalently, $b < 0$ in (5). Solving (5) together with (24) yields

$$P(x) = 1 - \rho \exp(\gamma x), \quad x \geq 0.$$}

If we apply the tight discretization method (21) with (19) to (25), then we obtain the base points of discretization as

$$x_k = \frac{1}{\gamma_x} \ln(1 - \frac{1}{\rho} \sum_{i=1}^{k-1} P^*_i), \quad k \geq 0,$$

and hence the approximate queue length distribution is

$$P_0 = 1 - \rho,$$

$$P_k = \rho^{-d} \sum_{i=0}^{k-1} \frac{1}{\rho^i} (1 - \frac{1}{\rho} \sum_{i=1}^{k-1} P^*_i), \quad k \geq 1.$$}

Remark. If we apply the usual discretization method (16) to (25), then we have

$$P_0 = 1 - \rho,$$

$$P_k = \rho(1 - \rho) P^*_k, \quad k \geq 1,$$

which coincides with Kobayashi's (20) diffusion approximation.

In what follows we will investigate detailed properties of the approximation (27) in which the GI/M/1 queue is assumed as the base of discretization. It is well known that the GI/M/1
From the approximate queue length distribution (33), we can derive various queuing characteristics for the GI/G/1 queue: Let $\text{EW}$ be the mean waiting time (until beginning service) in the GI/G/1 queue. Then, with the aid of Little’s formula, we have

$$\text{EW} = \frac{1}{\lambda} \sum_{k=1}^{\infty} \frac{w}{\mu(1 - \omega)}.$$  

(39)

Let $c_d$ be the coefficient of variation of interdeparture times. Marshall (25) showed the following relation between $\text{EW}$ and $c_d$:

$$c_d = c_d^2 + 2\rho^2 c_d^2 - 2\lambda(1 - \rho)\text{EW}. $$

(40)

The relation (40) was first used together with an approximation for $\text{EW}$ to approximate departure process in networks of queues by Kuehn (22). Following Kuehn (22), we combine (39) and (40) to obtain

$$c_d^2 = c_d^2 + 2\rho^2 c_d^2 - \frac{2\rho(1 - \rho)\text{EW}^2}{1 - \omega d}. $$

(41)

These results can be utilized in decomposition approximations for networks of queues. In particular, a software package called QNA (Queueing Network Analyzer) has been developed to calculate approximate congestion measures for open non-Markovian networks of queues; see (32). The approximation (41) combined with the previous results for superposition process provide a new approximation in QNA-like software.

We now investigate asymptotic properties of the GI/M/1-base approximation in heavy traffic. First, we characterize heavy traffic limits of $\omega$ and its derivative:

**Lemma 1.** \( \lim_{\rho \to 1} \omega = 1. \)  

(42)

**Lemma 2.** \( \lim_{\rho \to 1} \frac{d\omega}{d\rho} = \frac{2}{1 + c_d^2}. \)  

(43)

(Since proofs of these lemmas are straightforward, we omit the proofs.) From Lemmas 1 and 2, we immediately have

**Theorem.** For the GI/M/1-base diffusion approximation,

(a) \( \lim_{\rho \to 1} x = k, \)

(44)

(b) \( \lim_{\rho \to 1} 2\mu(1 - \rho)\text{EW} = c_d^2 + c_d^2. \)

(45)

From (44) we see that the base points in our tight discretization method asymptotically coincide with the usual (i.e., integer-valued) base points as the traffic becomes heavy. Also, we see from (45) that the heavy traffic limit of $\text{EW}$ is consistent with Kingman’s (19) result.

### 3.2. Heuristic Modification

The approximate distribution (33), of course, gives the exact solution for the GI/M/1 queue. However, for the M/G/1 queue, it does not give the exact solution even for the mean waiting time. In this subsection, we will modify (33) both to be consistent with the Pollaczek-Khintchine formula and to be still tight for the GI/M/1 queue.

For the above purpose, we modify the
diffusion parameters through $d$: Let $\tilde{d} \equiv d(c^2_a, c^2_b)$ denote a modification of $d$. Obviously, we must have

$$\tilde{d}(c^2_a, 1) = 1. \quad (46)$$

To let (39) be exact for the $M/G/1$ queue, we need

$$\tilde{d}(1, c^2) = \frac{\rho(1 + c^2)}{2 - \rho(1 - c^2)}. \quad (47)$$

Taking into account $\omega = \rho$ for the $M/M/1$ queue, we propose the following modification satisfying (46) and (47) simultaneously:

$$\tilde{d}(c^2_a, c^2_b) = \log \left( \frac{\omega(c^2_a + c^2_b)}{c^2_a + 1 - \omega(1 - c^2_b)} \right), \quad (48)$$

i.e.,

$$\tilde{d}(c^2_a, c^2_b) = \frac{\omega(c^2_a + c^2_b)}{c^2_a + 1 - \omega(1 - c^2_b)}. \quad (49)$$

Substituting (48) into (39) yields

$$EW = \frac{\omega(c^2_a + c^2_b)}{u(1 + c^2_a)(1 - \omega)} \cdot (50)$$

which is exact both for the $GI/M/1$ and $M/G/1$ queues.

4. NUMERICAL COMPARISONS

To see the quality of our diffusion approximations with tight discretization, we compare them with previously established diffusion approximations, heuristic approximations and the exact results for some particular cases. Since it is difficult to obtain exact results for queues with finite waiting space, we restrict our comparisons to $GI/G/1$ queues with unlimited waiting room.

Table 1 compares diffusion approximations for the queue length distribution $p_0$ of the $M/E_2/1$ queue ($\rho = 0.6, 0.9$). Four different diffusion approximations are considered, namely, the approximation (33), its modification (33) with (48), Gelenbe's (9, (4.25a)) and Kobayashi's (20, (2.9)) approximation ($\sim$). In our two approximations we use the $M/M/1$ queue as the base of discretization, i.e., $\omega = \rho$ in (33). Since all of these approximations give the exact result for the empty probability, $p_0 = 1 - \rho$ is excluded from the comparisons.

Table 1 shows that the modified approximation is much better than the others in moderate traffic. In heavy traffic, our and Kobayashi's approximations are quite accurate. In particular, it is interesting that the $M/M/1$-base approximation (33) performs about the same as Kobayashi's approximation. Gelenbe's approximation performs very poorly, especially, for the probability $p_1$. This seems to be due to the boundary condition of the approximating

<table>
<thead>
<tr>
<th>Approximations</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$ ($\times 10$)</th>
<th>$p_4$ ($\times 10$)</th>
<th>$p_5$ ($\times 10$)</th>
<th>$p_{10}$ ($\times 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M/M/1$-base</td>
<td>.315</td>
<td>.150</td>
<td>.712</td>
<td>.339</td>
<td>.161</td>
<td>.00392</td>
</tr>
<tr>
<td>Mod. $M/M/1$-base</td>
<td>.282</td>
<td>.149</td>
<td>.791</td>
<td>.419</td>
<td>.222</td>
<td>.00922</td>
</tr>
<tr>
<td>Exact</td>
<td>.276</td>
<td>.154</td>
<td>.817</td>
<td>.425</td>
<td>.220</td>
<td>.00800</td>
</tr>
<tr>
<td>$\rho = 0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kobayashi [20]</td>
<td>.120</td>
<td>.104</td>
<td>.900</td>
<td>.780</td>
<td>.677</td>
<td>.331</td>
</tr>
<tr>
<td>$M/M/1$-base</td>
<td>.120</td>
<td>.104</td>
<td>.901</td>
<td>.781</td>
<td>.677</td>
<td>.331</td>
</tr>
<tr>
<td>Mod. $M/M/1$-base</td>
<td>.116</td>
<td>.101</td>
<td>.881</td>
<td>.767</td>
<td>.668</td>
<td>.335</td>
</tr>
<tr>
<td>Exact</td>
<td>.110</td>
<td>.101</td>
<td>.893</td>
<td>.780</td>
<td>.679</td>
<td>.338</td>
</tr>
</tbody>
</table>

Table 2. A comparison of approximations for the mean queue length of $E_{m/E_2/1}$ queues ($m = 2, 4, 9$ and $\rho = 0.6, 0.9$).

<table>
<thead>
<tr>
<th>Approximations</th>
<th>$E_{2}/E_{2}/1$</th>
<th>$E_{4}/E_{2}/1$</th>
<th>$E_{9}/E_{2}/1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.6$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gelenbe [9]</td>
<td>.600</td>
<td>.488</td>
<td>.425</td>
</tr>
<tr>
<td>Kobayashi [20]</td>
<td>.349</td>
<td>.248</td>
<td>.193</td>
</tr>
<tr>
<td>$M/M/1$-base</td>
<td>.337</td>
<td>.238</td>
<td>.186</td>
</tr>
<tr>
<td>$\rho = 0.9$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gelenbe [9]</td>
<td>.349</td>
<td>.248</td>
<td>.193</td>
</tr>
<tr>
<td>Kobayashi [20]</td>
<td>.337</td>
<td>.238</td>
<td>.186</td>
</tr>
<tr>
<td>$M/M/1$-base</td>
<td>.322</td>
<td>.216</td>
<td>.170</td>
</tr>
<tr>
<td>Mod. $M/M/1$-base</td>
<td>.393</td>
<td>.262</td>
<td>.195</td>
</tr>
<tr>
<td>KL [21]</td>
<td>.403</td>
<td>.242</td>
<td>.155</td>
</tr>
<tr>
<td>Page [26]</td>
<td>.409</td>
<td>.277</td>
<td>.203</td>
</tr>
<tr>
<td>Exact</td>
<td>.378</td>
<td>.237</td>
<td>.162</td>
</tr>
</tbody>
</table>

3.1A-2-6
diffusion process; see (16) for a similar situation and its modification.

Table 2 compares the diffusion approximations with several heuristic approximations for the mean queue length of $E_{m||2}$ queues ($m = 2, 4, 9$ and $\rho = 0.6, 0.9$).

Three heuristic two-moment approximations are added here for comparisons, namely, the approximations of Kimura (15), Kraemer and Langenbach-Belz (21) (abbreviated by KL) and Page (26). The modified $M/M/1$-base approximation is excluded from the comparisons because it is less accurate than the approximation without modification.

From Table 2 we see with a little surprise that the $M/M/1$-base approximation is better than the $G_I/M/1$-base approximation. The modified $G_I/M/1$-base approximation is better than that without modification, but it is sometimes less accurate than the $M/M/1$-base approximation. A major reason for these results may be that the $G_I/M/1$-base approximations tend to be too sensitive to the arrival distribution form. However, for systems with high-variable interarrival-time distributions, the $G_I/M/1$-base approximations might be more accurate than the $M/M/1$-base approximation. As shown in Table 1, Kobayashi's approximation performs about the same as the $M/M/1$-base approximation, both providing accuracy comparable to the heuristic approximations. Gelenbe's approximation is less accurate than the others for all cases in the table.

REFERENCES


