REFINING DIFFUSION APPROXIMATIONS FOR QUEUES

Ward WHITT

Bell Laboratories, WB-1A350, Crawfords Corner Road, Holmdel, NJ 07733, U.S.A.

Received July 1982
Revised October 1982

This note illustrates the need to refine diffusion approximations for queues. Diffusion approximations are developed in several different ways for the mean waiting time in a GI/G/1 queue, yielding different results, all of which fail obvious consistency checks with bounds and exact values.

Queues, approximations, diffusion models, limit theorems

1. Introduction

Diffusion approximations and associated heavy-traffic limit theorems have now become basic tools of queuing theory; e.g., see Gelenbe and Mitrani [3, Chapter 4], Kleinrock [8, Chapter 2], Newell [12, Chapter 6] and references there. Under heavy-traffic conditions, diffusion approximations are identified and justified by the heavy-traffic limit theorems: the diffusion process is obtained as the limit. Variants of the limiting diffusion process also may be useful as approximations in moderate traffic. However, in moderate traffic, it is often difficult to select an appropriate diffusion process. Whenever possible, it is important to consider refining the diffusion approximations, for example, by comparing the approximations with available bounds and exact results.

The purpose of this note is to illustrate the need to refine diffusion approximations by considering diffusion approximations in one of the simplest settings: approximations for the expected waiting time in the standard GI/G/1 queue. In the GI/G/1 model there is a single server, unlimited waiting space, the first-come-first-served discipline and independent sequences of independent and identically distributed interarrival times and service times. Perhaps the most important insight provided by the heavy-traffic limit theorems for the GI/G/1 queue is the fact that only the first two moments of the interarrival-time and service-time distributions are essential for determining the standard congestion measures in heavy traffic (when the traffic intensity is close to its critical value). However, one should not rely too heavily on the associated approximations because different approximations can be obtained by very similar reasoning.

In Section 2, we apply a standard heuristic argument to obtain diffusion approximations for three basic processes in the GI/G/1 queue: the sequence of successive waiting times, $\{W_n, n \geq 0\}$, the virtual waiting time process, $\{V(t), t \geq 0\}$, and the queue length process, $\{Q(t), t \geq 0\}$. In each case, the diffusion process approximation for the queuing process yields an approximation for the mean of the equilibrium distribution. However, the three means are related exactly so that we have three different ways to obtain an approximation for the expected waiting time.

To specify the connecting relationships, we introduce our notation. Let λ be the arrival rate, μ the service rate, $\rho = \lambda/\mu$ the traffic intensity, c_a^2 the squared coefficient of variation of the interarrival-time distribution (variance divided by the square of the mean), and c_s^2 the squared coefficient of variation of the service-time distribution. Let W, V and Q be random variables with the equilibrium distributions of W_n, $V(t)$ and $Q(t)$, respectively.

To go from EQ to EW, we apply Little's for-

0167-6377/82/0000-0000/$02.75

mula, which yields

$$EW = (EQ - \rho)/\lambda, \tag{1}$$

since we are interpreting Q as the number in the system, including any in service, and W as the waiting time before beginning service. To go from EV to EW, we use one of Brumelle's [1] generalizations of Little's formula, namely,

$$EW = \frac{EV}{\rho} - \frac{(1 + c_s^2)}{2\mu}. \tag{2}$$

We obtain diffusion approximations for EW directly and from diffusion approximations for EQ and EV via (1) and (2).

In Section 3 we introduce two other diffusion approximations obtained by Gelenbe [3] and Kobayashi [9]. In Section 4 we check the consistency of these diffusion approximations for EW with previously established bounds and exact results. We show that none of these diffusion approximations pass all the consistency checks. We also indicate how a good approximation can be obtained. In the relatively simple setting of the GI/G/1 queue, general bounds and exact results for special cases can be used to improve the quality of the approximations. In more complicated settings, it may also be possible to use general bounds and exact results for special cases. If not, the diffusion approximations can be refined using simulation.

2. The basic three diffusion approximations

The standard way to obtain diffusion approximations, in both heavy traffic limit theorems and heuristic arguments, if to first consider the unrestricted queuing process (away from the barrier at zero), e.g., see Heyman [5, Section 4], Iglehart and Whitt [6] and Kingman [7, Chapter 4]. Let $X(t)$ represent one such queuing process. To identify the diffusion process, let

$$\alpha = \lim_{t \to \infty} \frac{EX(t)}{t}, \tag{3}$$

$$\sigma^2 = \lim_{t \to \infty} \frac{\operatorname{Var} X(t)}{t}, \tag{4}$$

assuming that the limits exist. The approximating diffusion process for $X(t)$ is a Brownian motion with drift α (which is negative if the queue is stable) and diffusion coefficient σ^2. These two parameters completely specify the unrestricted diffusion process. To obtain the approximating diffusion process for the queuing process itself, we simply add a reflecting barrier at zero. (It is sometimes suggested that a different kind of boundary at zero should be used; see Gelenbe and Mitrani [3, Chapter 4], Harrison and Lemoine [4] and references there.) The equilibrium distribution of the resulting diffusion process with negative drift and reflecting barrier at zero is exponential with mean $\sigma^2/2|\alpha|$. Hence, the diffusion approximation for the mean of the equilibrium queuing process is $\sigma^2/2|\alpha|$.

To consider the different queuing processes, let $\{u_n\}$ and $\{v_n\}$ be the sequences of interarrival times and service times, and let $A(t)$ and $S(t)$ be the renewal counting processes associated with $\{u_n\}$ and $\{v_n\}$.

For the waiting times W_n, the unrestricted process is the random walk with steps $v_n - u_n$. Hence, the direct diffusion approximation for EW, which also has been shown to be an upper bound, is

$$EW \approx \frac{\operatorname{Var}(v_n - u_n)}{2|E(v_n - u_n)|} = \frac{\rho^{-1}c_a^2 + \rho c_s^2}{2\mu(1-\rho)}; \tag{5}$$

see [7, p. 139].

For the queue length process $Q(t)$, the unrestricted process is $A(t) - S(t)$. Since

$$\lim_{t \to \infty} \frac{A(t)}{t} = \lambda \text{ and } \lim_{t \to \infty} \frac{\operatorname{Var} A(t)}{t} = \lambda^3 \operatorname{Var}(u_n), \tag{6}$$

$$\alpha = \lambda - \mu \tag{7}$$

and

$$\sigma^2 = \lambda^3 \operatorname{Var}(v_n) + \mu^3 \operatorname{Var}(v_n); \tag{8}$$

see [5]. Hence, the diffusion approximation for EQ is

$$EQ \approx \frac{\rho c_a^2 + c_s^2}{2(1-\rho)}. \tag{9}$$

By (1), the associated approximation for EW is

$$EW \approx \frac{c_a^2 + \rho^{-1}c_s^2 - 2(1-\rho)}{2\mu(1-\rho)}. \tag{10}$$

For the virtual waiting time process $V(t)$, the unrestricted process is

$$X(t) = \sum_{i=1}^{A(t)} v_i - t, \tag{11}$$

for which

$$\alpha = \rho - 1 \tag{12}$$

and

$$\sigma^2 = \lambda \operatorname{Var}(v_n) + \lambda^3 \operatorname{Var}(u_n)(Ev_n)^2 = \rho(c_a^2 + c_s^2)/\mu. \tag{13}$$

Hence,

$$EV \approx \frac{\rho(c_a^2 + c_s^2)}{2\mu(1-\rho)} \tag{14}$$

and, by (2), the associated approximation for EW is

$$EW \approx \frac{(c_a^2 - 1) + \rho(c_s^2 + 1)}{2\mu(1-\rho)}. \tag{15}$$

3. Two other diffusion approximations

Before discussing the three diffusion approximations for EW in Section 2, we introduce two other diffusion approximations. These were obtained for EQ, and so apply to EW via (1). First, Kobayashi [9] suggested the following diffusion approximation:

$$EW \approx \frac{\hat{\rho}}{\mu(1-\hat{\rho})}, \tag{16}$$

where

$$\hat{\rho} = \exp\{-2(1-\rho)/(\rho c_a^2 + c_s^2)\}; \tag{17}$$

see [8, p. 75] and [3, Section 4.2.2].

Next, based on an instantaneous return boundary, Gelenbe and Mitrani [3, p. 123] propose

$$EW \approx \frac{\rho(c_a^2 + 1) + (c_s^2 - 1)}{2\mu(1-\rho)}; \tag{18}$$

also see the references in [3] for earlier work by Gelenbe.

4. Consistency checks

A trivial consistency check is nonnegativity: EW should be nonnegative for all parameter values. Note, however, that three of the five diffusion approximations fail this test. For sufficiently small c_a^2 and c_s^2, the approximations (10), (15), and (18) are negative.

It is also natural to require that the approximation agrees with the known M/G/1 value:

$$EW = \frac{\rho(1 + c_s^2)}{2\mu(1-\rho)}. \tag{19}$$

However, only (15) satisfies this property. Note that none of the approximations have passed both these first two tests.

Finally, approximations should not fall outside of known bounds given by the first two moments of the interarrival times and service times. For example, Kingman showed that his approximation for EW in (5) is an upper bound; see [7] or [8]. However, Daley [2] obtained a better upper bound, namely,

$$EW \leqslant \frac{(2-\rho)c_a^2 + \rho c_s^2}{2\mu(1-\rho)}, \tag{20}$$

so (5) is too large. Marchal [10] proposed a natural refinement: multiplying (5) by a constant less than one so that the M/G/1 formula (19) is exact. This leads to a sixth diffusion approximation:

$$EW \approx \frac{\rho(1 + c_s^2)}{1 + \rho^2 c_s^2} \frac{c_a^2 + \rho^2 c_s^2}{2\mu(1-\rho)}. \tag{21}$$

Each of the six approximations has regions where it works well. For example, all the bounds and approximations are asymptotically correct in heavy traffic: They all satisfy

$$\lim_{\rho \to 1} 2\mu(1-\rho)EW = c_a^2 + c_s^2. \tag{22}$$

However, all but (15) exceed the upper bound in some cases. (See Table 1.) We know that approximation (15) is always less than the upper bound because it coincides with Marshall's [11] upper bound for GI/G/1 queues having DFR interarrival-time distributions (with decreasing failure rate, for which $c_a^2 \geqslant 1$). Moreover, this bound is tight, i.e., there is an interarrival-time distribution yielding this bound; see Whitt [13]. At the same time, (19) is a tight lower bound. Also, for interarrival-time distributions having increasing failure rate (for which $c_a^2 \leqslant 1$), (15) is a lower bound and (19) is an upper bound; see [13].

A promising class of approximations called *MFR approximations* (monotone failure rate) was proposed in [13]. A simple one, which coincides with what has been proposed by others, is

Table 1
Bounds and approximations for the mean waiting time, EW, in a GI/G/1 queue: Three cases

Upper bounds	Parameter values		
	$c_a^2 = 0.5$, $c_s^2 = 4.0$, $\rho = 0.7$	$c_a^2 = 2.0$, $c_s^2 = 4.0$, $\rho = 0.7$	$c_a^2 = 0.8$, $c_s^2 = 4.0$, $\rho = 0.3$
Kingman, (5)	5.86	9.42	2.76
Daley, (20)	5.75	9.00	1.82
MFR, (15) or (19)	5.83	7.50**	1.07
Approximations			
via EQ, (10)	9.36 H	11.86 H	9.09 H
via EV, (15)	5.00*	7.50**	0.93*
Kobayashi, (16)	6.76 H	8.51 H	2.56 H
Gelenbe, (18)	6.75 H	8.50 H	3.02 H
Marchal, (21)	4.85 L	7.80 H	0.91 L
MFR, (23)	5.25	7.00	1.02
Lower bound			
MFR, (15) or (19)	5.00*	5.83	0.93*

(1) In each case, the service rate is $\mu = 1$.
(2) 'H' indicates high and 'L' indicates low in comparison with bounds, including the MFR (monotone failure rate) bounds.
(3) Kingman's upper bound (5) is also one of the diffusion approximations.
(4) When $c_a^2 \geq 1$, (15) is the upper MFR bound and (19) is the lower MFR bound; when $c_a^2 \leq 1$, these MFR bounds are reversed.
(5) * indicates the lower MFR bound and ** indicates the upper MFR bound.

$$EW \approx \frac{\rho(c_a^2 + c_s^2)}{2\mu(1-\rho)}; \tag{23}$$

see [13]. Approximation (23) was proposed as a refinement of a diffusion approximation by Yu [14]; see Section 4.2.5 of [3]. Note that (23) coincides with the approximation for EV in (14). For the M/G/1 queue, $EW = EV$ and (14), (15), and the M/G/1 value (19) coincide.

The values of the various bounds and approximations are displayed for a few cases in Table 1. These cases are chosen to illustrate bad behavior with reasonable values of the parameters ρ, c_a^2, and c_s^2. In Table 1, the Marchal approximation (21) falls outside the MFR bounds. It can also exceed Daley's bound (20) in extreme cases. For very large c_s^2, the multiplicative correction factor in (21) is approximately 1; then (21) puts the same weight on c_s^2 as (20) but more on c_a^2. The approximation (15) is not inconsistent in the cases of Table 1, but since it coincides with one of the MFR bounds, it should be possible to improve the approximation, e.g., by using (23).

5. Summary

Diffusion approximations can be very helpful in queuing, especially when direct analysis is difficult. However, as we have illustrated, it is important to consider consistency checks. The diffusion approximations have a degree of freedom that permits refinement.

Acknowledgment

I am grateful to my colleague Daniel P. Heyman for helpful discussions.

References

[1] S.L. Brumelle, "On the relation between customer and time averages in queues", *J. Appl. Probability* **8**, 508–520 (1971).
[2] D.J. Daley, "Inequalities for moments of tails of random variables, with a queueing application", *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **41**, 139–143 (1977).
[3] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems*, Academic Press, New York (1980).
[4] J.M. Harrison and A.J. Lemoine, "Sticky Brownian motion as the limit of storage processes", *J. Appl. Probability* **18**, 216–226 (1981).
[5] D.P. Heyman, "A diffusion model approximation for the GI/G/1 queue in heavy traffic", *Bell System Tech. J.* **54**, 1637–1646 (1975).
[6] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: sequences, networks and batches", *Adv. Appl. Probability* **2**, 355–369 (1970).
[7] J.F.C. Kingman, "The heavy traffic approximation in the theory of queues", *Proc. Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill (1965) 137–159.
[8] L. Kleinrock, *Queueing Systems, Vol. 2: Computer Applications*, Wiley, New York (1976).
[9] H. Kobayashi, "Applications of the diffusion approximations to queueing networks, I: equilibrium queue distributions", *J. ACM* **21**, 316–328 (1974).
[10] W.G. Marchal, "An approximate formula for waiting times in single server queues", *AIIE Trans.* **8**, 473–474 (1976).
[11] K.T. Marshall, "Some inequalities in queueing", *Operations Res.* **16**, 651–665 (1968).

[12] G.F. Newell, *Applications of Queueing Theory*, Chapman and Hall, London (1971).

[13] W. Whitt, "The Marshall and Stoyan bounds for IMRL/G/1 queues are tight", Submitted for publication (1982).

[14] P.S. Yu, "On accuracy improvement and applicability conditions of diffusion approximation with applications to modelling of computer systems", TR-129, Digital Systems Laboratory, Stanford University (177).