Approximations for Departure Processes and Queues in Series

Ward Whitt
AT&T Bell Laboratories, Holmdel, New Jersey 07733

Methods are developed for approximately characterizing the departure process from a single-server queue and calculating approximate congestion measures for several single-server queues in series. These methods are modifications of the previously described asymptotic method and stationary-interval method for approximating a stochastic point process. The approximations are evaluated by comparing approximate congestion measures for queues in series with previous simulation results.

Two basic methods for approximating a general stochastic point process by a renewal process partially characterized by two or more parameters are described in Whitt [26]: the stationary-interval method and the asymptotic method. With the stationary-interval method, the renewal-interval distribution is chosen to match the distribution of the stationary interval between points in the process being approximated. If the parameters partially characterizing the approximating renewal process are the first k moments of the renewal interval, then we use the first k moments of the stationary interval in the process being approximated, or approximations for these moments. The parameters could also be other characteristics of the renewal-interval distribution such as percentiles. With whatever parameters we use, the idea is to match the renewal-interval distribution and the stationary-interval distribution. The same idea also applies with other approximating processes such as batch-Poisson processes or switched-Poisson processes.

The stationary-interval method has a potentially serious drawback, however. It does not take account of the dependence among successive intervals. The asymptotic method is an attempt to capture this dependence. The asymptotic method chooses the parameters of the renewal interval (or other approximating process) to match the long-run behavior of the process being approximated. For example, if the parameters are the first two moments of the renewal interval, they are chosen so that the renewal counting process has the same two normalization constants in its central limit theorem as the point process being approximated. For stationary processes, the asymptotic method obviously yields the same intensity (the reciprocal of the mean of the renewal interval) as the stationary-interval method, but the second moments of the approximating renewal interval can be very different. An essential idea behind the asymptotic method is that renewal processes can be used as approximating processes without ignoring the dependence among the intervals in the process being approximated.

These basic methods were presented as building blocks to develop refined hybrid approximations. For superposition arrival processes to single-server queues, refined hybrid procedures were developed and tested by Albin [1,2]. Albin obtained significant improvements over either basic procedure alone by letting the squared coefficient of variation (variance divided by the square of the mean) of the approximating renewal

Naval Research Logistics Quarterly, Vol. 31, pp. 499–521 (1984)
Copyright © 1984 by John Wiley & Sons, Inc. CCC 0028-1441/84/040499-23$04.00

interval be a convex combination of the squared coefficients of variation obtained from the two basic methods. Based on theoretical considerations and extensive simulation experiments, weighting functions were derived that depend on the number of processes superposed, the rates of these processes, and the traffic intensity of the single-server queue. The appropriate approximation depends on the context, so that it is natural to use the traffic intensity of the queue to refine the approximation. In fact, another procedure is to generate the entire approximation by looking at the impact of the given point process on a test queueing system [25,27].

A major goal in this work has been to develop approximations for non-Markovian networks of queues, having non-Poisson arrival processes, nonexponential service-time distributions, and non-product-form equilibrium distributions. The flows in such a network are complicated stochastic point processes, obtained by applying the operations of superposition (merging), partition (splitting or thinning), and departure to other point processes. (Even in Markovian networks, the flows are typically complicated [19].) The superposition operation is now relatively well understood due to Albin [1,2] and the partition operation, at least based on independent trials, is relatively well understood because a split renewal process is again a renewal process. However, the departure operation is not yet well understood. It is of course well known that, except in special cases, the departure process from a GI/G/m queue (with renewal arrival process independent of independent and identically distributed (i.i.d.) service times) is not a renewal process ([3] and references therein). As far as exact analytic results are concerned, for the most part, departure processes are intractable.

The purpose of this article is to investigate methods for approximating departure processes. In this article, we show how the two basic methods can be applied for this purpose. Since we are primarily interested in the departure process from a queue as an arrival process to another queue, we evaluate the approximations for departure processes by comparing approximate congestion measures (e.g., the expected waiting time) at the next queue with simulation results. The approximate congestion measures at the next queue are obtained by combining the approximations for departure processes with approximate congestion measures that depend on the arrival process only through the characterizing parameters.

The results here have been combined with the previous results for superposition processes to generate 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 [28,29]. The approach in QNA is to analyze all the queues in the network separately as GI/G/m queues in which both the service-time distribution and the interarrival-time distribution are partially characterized by their first two moments. The first moments of the interarrival-time distributions are obtained from the familiar traffic rate equations and the second moments are obtained from a similar system of linear equations, based on the approximations.

The results in this article are also being applied to analyze various design problems for queues in series. For example, in Whitt [30] we develop heuristic principles for ordering several nonexponential single-server queues in series so as to minimize the expected equilibrium sojourn time.

The applications just mentioned should make the advantages of heuristic analytic approximations over simulation obvious. For example, with the software package QNA we can quickly and easily analyze a network with 100 nodes, but it would take a very long simulation run to get good estimates of the equilibrium distributions. To get comparable accuracy with n nodes, roughly n times as many events (service completions) must be simulated as with one node. In fact, even for the relatively small networks considered here to test the approximations, e.g., with two nodes, the statistical accuracy of the simulation results is not especially good. Moreover, QNA can easily be run many times to study the impact of alternate designs. In some cases, as in arranging the queues in series, closed-form approximation formulas are available to provide insight.

Even though approximation has important advantages over simulation, both are clearly important. It is not either/or. Exact analysis, approximation, and simulation all should contribute significantly to our understanding of complex systems such as networks of queues. Here we use exact analytic results, including various limit theorems, together with simulation to develop the heuristic approximations. In applications, heuristic approximations are especially valuable in the formative stages of a project, while simulation becomes more important in fine tuning. In applications, the heuristic approximations may help identify the appropriate models to simulate. In turn, simulations certainly play a vital role in developing the heuristic approximations.

In this article we see what the basic methods yield for departure processes. For the stationary-interval method, we consider the departure process from the standard GI/G/1 queue having independent sequences of i.i.d. interarrival times and service times. We present two different applications of the stationary-interval method, one based on an analysis of the system in equilibrium at an arbitrary time and the other based on Marshall's [18] characterization of the stationary departure interval in terms of the waiting time distribution. We show how previous heuristic approximations of Sevcik et al. [22] and Gelenbe and Mitrani [11] fit into this framework.

For the asymptotic method, we consider more general models. We show in considerable generality that the asymptotic method approximation for the departure process coincides with the asymptotic method approximation for the arrival process.

Our object is to provide a theoretical framework for making approximations. This article provides appropriate background to begin developing refined hybrid procedures for departure processes. Also, preliminary experiments are described here.

The basic approach in Whitt [26] is to approximate the point process by a renewal process partially characterized by the first few moments of the renewal interval. As with our previous treatment of superposition processes, we consider moments beyond the first two. In particular, we treat the third moment in detail, so that we provide a basis for three-moment procedures. (A three-parameter extension of the algorithm in QNA has been developed, but is still being tested.) We also consider parametric approximations that do not correspond to a renewal process. In particular, following Fraker [8], we consider the lag-1 correlation (the correlation between successive intervals) as an additional parameter.

The rest of this article is organized as follows. We consider the asymptotic method in Section 1, the stationary-interval method in Section 2, the lag-1 correlation in Section 3, refined hybrid procedures in Section 4, and comparisons with simulation results for single-server queues in series by Fraker [8] and Shimshak [23] in Sections 5 and 6. The conclusions are summarized in Section 7.

In the experiments in Sections 5 and 6, we compare the relatively simple characterization of departure processes used in QNA with refinements based on the lag-1 correlation (Section 3) and hybrid methods combining the asymptotic method and the stationary-interval method (Section 4). While these methods for refining the approximation seem to have potential, we have not yet succeeded in finding a refinement that offers significant improvement. However, the relatively simple characterization used in QNA that is based on the stationary-interval method alone works quite well. In QNA the squared coefficient of variation (variance divided by the square of the mean) of the renewal interval of the approximating renewal process is just a convex combination of the squared coefficients of variation of the interarrival time (u) and service time (v), namely,

$$c_d^2 = \rho^2 c_v^2 + (1 - \rho^2)c_u^2, \tag{1}$$

where ρ is the traffic intensity [see (23)]. Evidently, the dependence among successive intervals is less important for departure processes than for superposition processes.

It is also significant that formula (1) is consistent with the interesting property established by Friedman [10] and by Weber [24] that, for any arrival process, the order of several $\cdot$/D/1 or $\cdot$/M/1 queues in series (all the same type) does not affect the departure process. Formula (1) yields this property as an approximation for any series system in which all queues have service-time distributions with the same squared coefficient of variation.

For background on departure processes, see Daley [6], Disney (Chap. VII of [7] and references therein). For other work on approximations for queues in series, see [8,17,21,23] and references for approximations of networks of queues in [28]. Additional simulations are contained in [4,20].

1. THE ASYMPTOTIC METHOD

In this section the queueing system can be quite general. Let $D(t)$, $A(t)$, and $Q(t)$ represent the number of departures in $[0,t]$, the number of arrivals in $[0,t]$, and the queue length (number in system including anyone in service) at time t. We assume that these random variables are related by

$$D(t) = A(t) - Q(t) + Q(0), \qquad t \geqslant 0. \tag{2}$$

Given that $A(t)$ diverges to $+\infty$ and $Q(t)$ converges in distribution to a proper limit as $t \to \infty$, the distribution of $D(t)$ for large t is very close (relative to t) to the distribution of $A(t)$, so that the asymptotic-method approximation for the departure process is just the arrival process, provided that the arrival process is in the class of approximating processes, e.g., a renewal process. Otherwise, the approximating process for $D(t)$ obtained by the asymptotic method is the same as the approximating process for $A(t)$.

Here we give a more precise statement and a proof. We actually assume that

$$\sup_{t \geqslant 0} E\{[Q(t) - Q(0)]^k\} < \infty \tag{3}$$

for all k of interest. Assumption (3) covers the case in which $Q(t)$ remains stochastically bounded but does not converge in distribution because the interarrival-time distribution is lattice.

As in Section 2 of [26], let $\beta_j(Z)$ be the jth cumulant or semi-invariant of the random variable Z. For the arrival process $A(t)$, let γ_j be the time-average limit of $\beta_j[A(t)]$, i.e.,

$$\gamma_j = \lim_{t\to\infty} \frac{\beta_j[A(t)]}{t}. \tag{4}$$

We assume that Eq. (4) is valid, which we know is the case for a large class of renewal processes (see (2.5) of [26]).

We characterize the time-average limit of $\beta_j[D(t)]$ for $j = 1,2,3,$ and 4. We do not yet have a proof that covers all j at once (see Section 15 of [12] for properties of cumulants).

THEOREM 1.1: If (2), (3), and (4) hold, then

$$\lim_{t\to\infty} \frac{\beta_j[D(t)]}{t} = \gamma_j, \qquad \text{for } j = 1,2,3, \text{ and } 4. \tag{5}$$

PROOF: Let d_j, a_j, and q_j be the jth central moment of $D(t)$, $A(t)$, and $Q'(t) = Q(t) - Q(0)$ for some given t, e.g.,

$$d_j = E\{[D(t) - ED(t)]^j\}.$$

As a consequence of (3), $q_j/t^\alpha \to 0$ as $t \to \infty$ for each j and each $\alpha > 0$. For $j = 3$,

$$\begin{aligned}\beta_3[D(t)] = d_3 = a_3 &- 3E\{[A(t) - EA(t)]^2[Q'(t) - EQ'(t)]\}\\ &+ 3E\{[A(t) - EA(t)][Q'(t) - EQ'(t)]^2\} - q_3,\end{aligned}$$

so that, with the aid of Hölder's inequality,

$$|\beta_3[D(t)] - \beta_3[A(t)]| \leq 3(a_3)^{2/3}(q_3)^{1/3} + 3(a_3)^{1/3}(q_3)^{2/3} + q_3,$$

from which (5) follows easily. Similarly, for $j = 4$, $\beta_4[D(t)] = d_4 - 3d_2^2$ and

$$\begin{aligned}|\beta_4[D(t)] - \beta_4[A(t)]| \leq 4(a_4)^{3/4}(q_4)^{1/4} &+ 6(a_4)^{1/2}(q_4)^{1/2}\\ &+ 4(a_4)^{1/4}(q_4)^{3/4} + q_4 + 12a_2q_2 + 3q_2^2,\end{aligned}$$

from which (5) again follows easily. Since $\beta_1(Z) = EZ$ and $\beta_2(Z) = \text{Var}(Z)$, the cases $j = 1$ and 2 are straightforward. Q.E.D.

2. THE STATIONARY-INTERVAL METHOD

In this section we consider the standard single-server queue with infinite waiting room and the first-come first-served discipline. We give two different approaches, both based on the stationary-interval method. The object here, therefore, is to characterize and approximate the stationary interval between departures.

2.1. Conditioning on the Server Being Busy or Not

In this subsection we first consider a general G/G/1 queue in equilibrium with stationary arrival process as in [9]. Since the departure process is a stationary point process, if we look at the system at an arbitrary time, then the time to the next departure has the stationary-excess distribution associated with the stationary departure-interval distribution; i.e., if F is the cdf of a stationary interval between departures, then the associated stationary-excess cdf G is

$$G(t) = \lambda \int_0^t [1 - F(u)]du, \qquad t \geq 0, \tag{6}$$

where λ is the arrival rate (λ^{-1} is the mean of F) (see (1.5) of [26]). Moreover, the server is busy at this instant with probability ρ, where ρ is the traffic intensity (see (4.2.3) of [9]).

If the server is busy, then the time to the next departure is the residual service time. If the server is not busy, then the time to the next departure is the residual interarrival time plus the following service time. Let $v_+(b)$ be the residual service time given that the server is busy; let $u_+(i)$ be the residual idle time given that the server is idle; let $v(i)$ be the next service time given that the server is idle; and let d_+ be the residual time to the next departure. In general, then, the distribution of d_+ is a mixture: With probability ρ it is distributed as $v_+(b)$ and with probability $(1 - \rho)$ it is distributed as $u_+(i) + v(i)$.

One way to approximate the distribution of d is to approximate the distribution of d_+ by approximating $v_+(b)$, $u_+(i)$, and $v(i)$. It is easy to go from the distribution of d_+ to the distribution of d by inverting (6) (see (1.6) of [26]).

$$F(t) = 1 - \lambda G'(t), \qquad t \geq 0. \tag{7}$$

Now suppose that we have a GI/G/1 queue, with independent sequences of i.i.d. interarrival times and service times. Let u and v be generic interarrival times and service times. Let u_+ and v_+ have the associated stationary excess distributions. Then $v(i)$ is distributed as v and is independent of $u_+(i)$. Also $v_+(b)$ is distributed as v_+. (For justification of this last step, see [13].) However, in general $u_+(i)$ is not distributed as u_+. It is, of course, in the M/G/1 queue when u has an exponential distribution.

We summarize these properties in the following theorem. Let F_u be the cdf of u, etc. Let $*$ denote convolution of cdfs.

THEOREM 2.1: (a) In the GI/G/1 queue,

$$F_{d_+} = \rho F_{v_+} + (1 - \rho)(F_{u_+(i)} * F_v). \tag{8}$$

(b) In the M/G/1 queue, $u_+(i)$ and u_+ are distributed as u.

There is a simple relationship between the moments of a cdf F and the moments of its stationary-excess cdf G (see (2.2) of [26]). Let m_j denote the jth moment about the origin. Then

$$m_j(G) = m_{j+1}(F)/m_1(F)(j + 1). \tag{9}$$

We now combine Eqs. (8) and (9) to obtain an expression for the moments of the stationary departure interval. Let $\mu_k = m_k/m_1^k$. Let $u(i)$ be obtained from $u_+(i)$ by (7).

THEOREM 2.2: In the GI/G/1 queue, for each $k \geq 2$,

$$\mu_k(d) = k\mu_{k-1}(d_+) = \rho^k\mu_k(v) + (1 - \rho) \sum_{j=0}^{k-1} \binom{k}{j} \rho^j \mu_j(v)\mu_{k-j}(u(i))[Eu(i)/Eu]^{k-j}. \tag{10}$$

A simple approximation for GI/G/1 queues and more general G/G/1 systems with nonrenewal arrival processes is to use (8) and (9) with $u_+(i)$ replaced by u_+. Instead of (10), we have the *approximation*

$$\mu_k(d) \approx \rho^k \mu_k(v) + (1 - \rho) \sum_{j=0}^{k-1} \binom{k}{j} \rho^j \mu_j(v) \mu_{k-j}(u), \tag{11}$$

which is valid for M/G/1 systems. For example,

$$\mu_2(d) \approx \rho^2 \mu_2(v) + (1 - \rho)[\mu_2(u) + 2\rho], \tag{12}$$

so that

$$c_d^2 \approx \rho^2 c_v^2 + (1 - \rho) c_u^2 + \rho - \rho^2, \tag{13}$$

where c^2 is the squared coefficient of variation (variance divided by the square of the mean) as in [26]. Note that c_d^2 in (13) is a convex combination of c_v^2, c_u^2, and 1 since $\rho^2 + (1 - \rho) + \rho - \rho^2 = 1$, and $1 - \rho \geqslant 0$ and $\rho - \rho^2 \geqslant 0$.

It turns out that the approximation in (11) can be restated in a more elementary form, which coincides with a natural direct approximation for the stationary departure interval:

THEOREM 2.3: If $u_+(i)$ is replaced by u_+ in (8), then the corresponding approximate stationary departure-interval cdf F_d obtained from (7) is

$$F_d = \rho F_v + (1 - \rho)(F_v * F_u). \tag{14}$$

PROOF: Applying (8), (9), and then (6), we obtain

$$\begin{aligned} F_d(t) &= 1 - (Eu) F'_{d_+}(t) \\ &= 1 - (Eu)\rho \frac{1 - F_v(t)}{Ev} - (Eu)(1 - \rho) F_{u_+ + v}(t) \\ &= F_v(t) - (Eu)(1 - \rho) \frac{F_v(t) - F_{u+v}(t)}{Eu} \\ &= \rho F_v(t) + (1 - \rho) F_{u+v}(t), \end{aligned}$$

where u_+ and v (u and v) are independent. Q.E.D.

REMARK 2.1: By Theorem 2.3, the stationary departure-interval distribution is approximated by a simple mixture: With probability ρ it is a service time and with probability $(1 - \rho)$ it is a service time plus an independent interarrival time. (This is used as a direct heuristic approximation on p. 133 of [11].)

REMARK 2.2: Approximation formula (13) for c_d^2 is the exact formula for the GI/G/1 queue having a batch-Poisson arrival process with geometrically distributed batches. To see this, substitute the expected waiting time as given in Section 5.10 of [5] into Marshall's [18] formula for c_d^2, given here in (21).

2.2. Expressions Involving the Waiting Time

Let T_n, D_n, W_n, and v_n be the arrival epoch, departure epoch, waiting time (before beginning service), and service time of the nth customer. Let $u_n = T_{n+1} - T_n$ and $d_n = D_{n+1} - D_n$ be the associated interarrival time and interdeparture time. For the standard single-server queue,

$$D_n = T_n + W_n + v_n, \qquad n \geqslant 1, \tag{15}$$

so that

$$d_n = u_n + W_{n+1} - W_n + v_{n+1} - v_n. \tag{16}$$

Since

$$W_{n+1} = \max\{0, W_n + v_n - u_n\} = W_n + v_n - u_n + X_n, \tag{17}$$

where

$$X_n = -\min\{0, W_n + v_n - u_n\}, \tag{18}$$

$$d_n = v_{n+1} + X_n. \tag{19}$$

Hence, for the GI/G/1 queue the stationary interval between departures d can be expressed as

$$d = v + X, \tag{20}$$

where v is independent of X, X is distributed as $-\min\{0, W + v - u\}$, and W is the equilibrium waiting time. Formula (20) is an alternative to (8).

Marshall [18] showed how (17) and (20) can be exploited to give expressions for the moments of d. From (6)–(9) of [18],

$$c_d^2 = c_u^2 + 2\rho^2 c_v^2 - 2(1 - \rho)EW/Eu. \tag{21}$$

Formula (21) was first used together with an approximation for EW to approximate departure processes in networks of queues by Kuehn [16]. It is also used in [28]; there EW is approximated by the simple formula

$$EW \approx \frac{(Ev)\rho(c_u^2 + c_v^2)}{2(1 - \rho)}, \tag{22}$$

so that

$$c_d^2 \approx (1 - \rho^2)c_u^2 + \rho^2 c_v^2. \tag{23}$$

REMARK 2.3: Approximation formula (23) was suggested as a direct heuristic approximation by Sevcik et al. [22].

Formula (21) and higher moments of d are obtained from (20) plus the moments of X. The moments of X in turn can be obtained calculating the moments of $W_{n+1} - X_n$ in two different ways: first, directly and, second, in the form $W_n + v_n - u_n$, using

(17). The first three moments of X are

$$\begin{aligned} EX &= -EY, \\ E(X^2) &= E(Y^2) + 2EY\,EW, \\ E(X^3) &= -[E(Y^3) + 3E(Y^2)EW + 3EY\,E(W^2)], \end{aligned} \tag{24}$$

where $Y = v - u$. Hence,

$$\begin{aligned} E(d^3) &= E(v^3) + 3E(v^2)EX + 3Ev\,E(X^2) + E(X^3) \\ &= -E(W^2)(3EY) - EW[3E(Y^2) - 6Ev\,EY] \\ &\qquad - E(Y^3) + 3Ev(EY^2) - 3E(v^2)EY + E(v^3). \end{aligned} \tag{25}$$

For any random variable Z, let $\theta_z = E(Z^3)/(EZ)^3$. From (25), we obtain

$$\begin{aligned} \theta_d = (\theta_u + 6\rho^2 c_v^2) - 3\rho(\rho^2 c_v^2 + c_u^2 + 1 - \rho^2)EW/Ev \\ + 3(1 - \rho)\rho^2 E(W^2)/(Ev)^2. \end{aligned} \tag{26}$$

REMARK 2.4: Formulas (21) and (26) can be checked by considering the M/M/1 queue, for which $EW/Ev = \rho/(1 - \rho)$, $E(W^2)/(Ev)^2 = 2\rho/(1 - \rho)^2$, $c_d^2 = 1$, and $\theta_d = E(d^3)/(Ed)^3 = 6$ since the stationary departure process is Poisson.

A stationary-interval approximation for $E(d^3)$ or θ_d is obtained from (26) by applying approximations for EW and $E(W^2)$. Following Whitt [28], we can approximate $E(W^2)$ using

$$E(W^2) = P(W > 0)E(D^2), \tag{27}$$

where D is the conditional delay given that the server is busy. We can approximate $E(D^2)$ by the M/G/1 formula:

$$\begin{aligned} E(D^2) &= (ED)^2\,[2\rho + 4(1 - \rho)\theta_v^3/3\,(c_v^2 + 1)^2] \\ &= [(EW)^2/P(W > 0)][2\rho + 4(1 - \rho)\theta_v^3/3(c_v^2 + 1)^2], \end{aligned} \tag{28}$$

and approximate $P(W > 0)$ by the Kraemer and Langenbach-Belz [15] formula:

$$P(W > 0) = \rho + (c_u^2 - 1)\rho(1 - \rho)h\,(\rho,c_u^2,c_v^2), \tag{29}$$

with

$$\begin{aligned} h(\rho,c_u^2,c_v^2) &= \frac{1 + c_u^2 + \rho c_v^2}{1 + \rho\,(c_v^2 - 1) + \rho^2\,(4c_u^2 + c_v^2)}, && c_u^2 \leq 1, \\ &= \frac{4\rho}{c_u^2 + \rho^2\,(4c_u^2 + c_v^2)}, && c_u^2 > 1. \end{aligned} \tag{30}$$

(see Section 5.1 of [28]). Of course, (26) using (27)–(30) lacks the simplicity of (21), but it is linear in θ_u, so that it is convenient for approximating networks of queues.

3. THE LAG-1 CORRELATION AS AN ADDITIONAL PARAMETER

The lag-1 correlation provides a compromise between the stationary-interval method and the asymptotic method. Using the lag-1 correlation plus moments of the stationary interval is closely related to the stationary-interval method because it is based on the local behavior of the process, i.e., the stationary distribution of two consecutive intervals. It is also closely related to the asymptotic method because it partially characterizes the dependence. The asymptotic method approximation for the variance includes all the covariances, whereas the lag-1 correlation involves only the covariance between consecutive intervals (see Section 2 of [26]).

Approximations of the lag-1 correlation and the first two moments of stationary interval have been used to approximate departure processes and waiting times of queues in series by Fraker [8] and Marchal [17]. We describe a modification of Fraker's procedure, obtained by substituting approximations in Whitt [28] for some of the quantities in Fraker's expressions. Let α be the lag-1 correlation, i.e., the covariance of two adjacent intervals divided by the variance. Fraker's approximations are developed and tested for Erlang (E_k) service times, but apply more generally by just substituting c_v^2 for k^{-1} as the variability parameter of an E_k distribution. Fraker's basic approximation for the lag-1 correlation of the departure process, α_d, in a G/E/1 queue is

$$\alpha_d(\mathrm{G/E/1}) \approx \alpha_d(\mathrm{GI/E/1}) + P(W = 0)\alpha_u(c_u^2/c_d^2), \tag{31}$$

where u indexes the arrival process as before (see (11) on p. 41 of [8]), where

$$\alpha_d(\mathrm{GI/E/1}) \approx \frac{(1-\rho)(e^{-\rho} - 1 + \rho)}{c_d^2}\left(\frac{c_u^2 - c_d^2(\mathrm{GI/E/1})}{1 - c_d^2(\mathrm{M/D/1})}\right) \tag{32}$$

(see (2), (9), and (10) of [8]).

We obtain a modification of Fraker's procedure by inserting our approximations for c_d^2 and $P(W = 0)$ into (31) and (32). In particular, if we use $c_d^2 = \rho^2 c_v^2 + (1 - \rho^2)c_u^2$ as in (23), then we obtain

$$\begin{aligned}\alpha_d(\mathrm{GI/E/1}) &\approx \frac{(1-\rho)(e^{-\rho} - 1 + \rho)}{\rho^2 c_v^2 + (1-\rho^2)c_u^2}\left(\frac{c_u^2 - c_v^2}{c_u^2}\right)\\ &\approx (1-\rho)(e^{-\rho} - 1 + \rho)(1 - r)/[1 - \rho^2(1 - r)],\end{aligned} \tag{33}$$

where $r = c_v^2/c_u^2$. Note that the approximation (33) depends on only the two parameters ρ and r. We complete the approximation for α_d by using the Kraemer–Langenbach-Belz approximation (29) for $P(W > 0)$ in (31).

To use the lag-1 correlations in approximations for queues in series, we must not only have an approximation for the lag-1 correlation of a departure process given the lag-1 correlation of the arrival process and the other parameters, but we must also have an approximation for the mean waiting time that depends on the lag-1 correlation of the arrival process.

Fraker's approximation for the expected waiting time as a function of c_u^2, c_v^2, and α_u is

$$EW(\mathrm{G/E/1}) \approx EW(\mathrm{GI/E/1}) + \frac{(Eu)P(W > 0)\alpha_u c_u^2}{(1-\rho)}. \tag{34}$$

Using approximation (22) for $EW(\mathrm{GI/G/1})$, we obtain

$$EW(\mathrm{G/E/1}) \approx Ev(1-\rho)^{-1}\left[\rho\,\frac{(c_u^2 + c_v^2)}{2} + \rho^{-1}P(W > 0)c_u^2\alpha_u\right], \tag{35}$$

where (29) is used for $P(W > 0)$ again in (35).

Even though Fraker's approximations were developed for G/G/1 systems with Erlang service-time distributions, formulas (31)–(35) can be used for G/G/1 systems. Hence, we have developed a modification of Fraker's procedure applicable to G/G/1 queues in series where the initial arrival process is characterized by the three parameters λ, c_u^2, and α_u. The arrival process to each subsequent node (queue) is just the departure process from the preceding node, so that $c_{un}^2 = c_{d,n-1}^2$ and $\alpha_{un} = \alpha_{d,n-1}$ where n indexes the node.

4. REFINED HYBRID PROCEDURES

One way to develop hybrid procedures, as in [1,2], is to look for convex combinations of the parameters from the two basic methods. A candidate hybrid approximation for the variability parameter of the departure interval, say, c_H^2, obtained from Section 1 and (23) is

$$\begin{aligned} c_H^2 &= xc_u^2 + (1-x)\,[\rho_1^2 c_v^2 + (1-\rho_1^2)c_u^2] \\ &= (1-x)\rho_1^2 c_v^2 + [1-(1-x)\rho_1^2]c_u^2, \end{aligned} \tag{36}$$

where x is a weighting function with $0 \leqslant x \leqslant 1$.

As with Albin's [1,2] treatment of superposition processes, we should anticipate that the weighting function x might involve the traffic intensity of the next queue. Of course, without feedback, both the superposition process and the departure process are independent of the next queue, but the way they should be approximated may well depend on the next queue. As with superposition processes, heavy-traffic limit theorems provide a theoretical reference point for choosing weighting functions x in (36). In particular, heavy-traffic limit theorems suggest that x should indeed depend on both the traffic intensity of the current node, say, ρ_1, and the traffic intensity of the next node, say, ρ_2. Given two queues in series, if the second is in heavy traffic but the first is not (ρ_2 is close to 1 but ρ_1 is not), then the heavy-traffic behavior of the second queue is the same as if the first queue were not there (see Theorem 1(a) of [14]). The heavy-traffic behavior of the second queue depends on its arrival process only through its central-limit-theorem behavior. The central-limit-theorem behavior in turn is the same as the asymptotic method and, by Section 1, the asymptotic method approximation for the departure process from the first queue is the same as the asymptotic method approximation of the arrival process. In other words, with a criterion involving the expected waiting time or the expected queue length at the second facility, the asymptotic method of approximating the departure process from the first facility is asymptotically correct as $\rho_2 \rightarrow 1$ *with* ρ_1 fixed. Hence, x ought to approach 1 as $\rho_2 \rightarrow 1$.

The heavy-traffic behavior of the two queues in series is much more complicated if both queues are in heavy traffic (if ρ_1 *and* ρ_2 are both close to 1). Hence, what is important for having x close to 1 is to have ρ_2 relatively closer to 1 than ρ_1.

Related theoretical reference points for approximating departure processes from multiserver queues are contained in [28].

5. FRAKER'S EXPERIMENT

We now compare approximations with simulations of queues in series. We first consider Fraker's [8] experiment. Fraker simulated eight different cases of eight single-server queues in series. In each case the external arrival process is Poisson and all service-time distributions are Erlang. Each of four traffic intensities ($\rho = 0.3, 0.5, 0.7,$

and 0.9) and each of four Erlang service-time distributions ($M = E_1, E_4, E_8$, and $D = E_\infty$) is assigned randomly to two of the eight nodes.

The simulations consisted of three separate runs of 2500 customers each, with the first 500 being discarded to damp out the transient effects. Statistics were collected for six blocks of 1000 customers each. Unfortunately, the simulation runs were not long enough for good statistical accuracy. No single estimate can be relied upon, but the total experiment clearly yields meaningful comparisons of approximations.

Fraker estimated three quantities: the expected waiting time, the variance of the stationary departure interval, and the lag-1 covariance of the departure intervals. Fraker also developed approximations for these three quantities, with the expected waiting time being a function of the variance and lag-1 covariance of the arrival process. A detailed description of Fraker's results is contained in Appendix 1 (available from the author).

Tables 1–3 compare various approximations with Fraker's simulation results in the first three cases. Fraker's results based on a mean interarrival time of 640 have been adjusted to a mean interarrival time of 1. The estimated lag-1 covariances have been converted to estimated lag-1 correlations by dividing by the estimated variances. Since the mean is 1, the variance of the interarrival time at each node is c_u^2. The lag-1 correlation estimate α_u is obtained by dividing the lag-1 covariance estimate by the associated c_u^2 estimate. (Recall that c_u^2 and α_u for node n are just c_d^2 and α_d for node $n - 1$.)

Five approximations for the expected waiting time are considered. The M/M/1 approximation is the exact result when all service-time distributions are exponential. Each queue is regarded as an independent M/M/1 queue with the specified arrival rate and service rate. The M/G/1 approximation is the M/M/1 value multiplied by $(1 + c_v^2)/2$, which is tantamount to assuming that each arrival process is Poisson at rate 1 and the service-time distribution is as specified. By Section 1, since the external arrival process is Poisson, the M/G/1 approximation coincides with the asymptotic method.

We give Fraker's approximations for all three characteristics, but we do not describe the methods or formulas here. We only remark that they are especially designed for Erlang service times, but they can be extended by substituting c_v^2 for k^{-1}. We also give two different versions of the QNA approximations in [28]. For the standard version, c_d^2 is given by (23) and EW is given by the following modification of (22):

$$EW = \frac{(Ev)\rho(c_u^2 + c_v^2)g}{2(1 - \rho)}, \tag{37}$$

where $g \equiv g(\rho, c_u^2, c_v^2)$ is defined as

$$\begin{aligned} g(\rho, c_u^2, c_v^2) &= \exp\left\{ -\frac{2(1-\rho)}{3\rho}\frac{(1-c_u^2)^2}{c_u^2 + c_v^2} \right\}, & c_u^2 < 1, \\ &= 1, & c_u^2 \geq 1. \end{aligned} \tag{38}$$

Approximation (37) with the correction factor (38) is the Kraemer and Langenbach-Belz [15] approximation for $c_u^2 < 1$. They have another correction factor for $c_u^2 > 1$, but we do not use it.

The second QNA approximation is a modified version using Section 3, in particular, using the second term in (34) together with (37) after approximating the lag-1 correlation using (31), (33), and (29).

Table 1. A comparison of approximations and simulation for Fraker's model of eight queues in series: case 1.[a]

Input parameters			Simulation			Approximations								
						M/M/1	M/G/1	Fraker			QNA		Modified QNA	
Node No.	Traffic intensity ρ	Service squared coefficient of variation c_v^2	EW	c_u^2	α_u	EW	EW	EW	c_u^2	α_u	EW	c_u^2	EW	α_u
1	0.7	1/8	0.98	0.98	0.039	1.63	0.92	0.92	1.00	0.000	0.92	1.00	0.92	0.000
2	0.5	1	0.30	0.57	0.117	0.50	0.50	0.38	0.57	0.082	0.38	0.61	0.42	0.090
3	0.5	0	0.19	0.81	0.012	0.50	0.25	0.13	0.73	−0.009	0.16	0.71	0.17	0.018
4	0.7	1/4	0.73	0.63	0.089	1.63	1.02	0.62	0.60	0.040	0.63	0.58	0.68	0.085
5	0.3	0	0.01	0.56	0.061	0.13	0.07	0.01	0.50	0.044	0.01	0.42	0.02	0.093
6	0.9	1	7.50	0.56	0.078	8.10	8.10	6.03	0.49	0.051	5.56	0.40	5.60	0.123
7	0.9	1/8	3.91	0.94	0.023	8.10	4.55	4.16	0.94	−0.015	4.09	0.89	4.08	−0.013
8	0.3	1/4	0.00	0.28	0.107	0.13	0.08	0.01	0.29	0.080	0.01	0.33	0.02	0.081

[a] c_u^2 is the squared coefficient of variation and α_u is the lag-1 correlation of the interarrival times. The external arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

Table 2. A comparison of approximations and simulation for Fraker's model of eight queues in series: case 2.[a]

Input parameters			Simulation			Approximations								
						M/M/1	M/G/1	Fraker			QNA		Modified QNA	
Node No.	Traffic intensity ρ	Service squared coefficient of variation c_v^2	EW	c_u^2	α_u	EW	EW	EW	c_u^2	α_u	EW	c_u^2	EW	α_u
1	0.9	1	6.25	1.05	−0.002	8.10	8.10	8.10	1.00	0.000	8.10	1.00	8.10	0.000
2	0.7	1/8	0.84	1.01	−0.002	1.63	0.92	0.92	1.00	0.000	0.92	1.00	0.92	0.000
3	0.3	1/4	0.01	0.81	0.057	0.13	0.08	0.04	0.57	0.082	0.04	0.61	0.07	0.090
4	0.9	0	2.61	0.61	0.064	8.10	4.05	2.45	0.57	0.068	2.28	0.58	2.33	0.098
5	0.3	1/8	0.00	0.18	0.127	0.13	0.07	0.00	0.15	0.153	0.00	0.27	0.16	0.228
6	0.5	1/4	0.02	0.20	0.080	0.50	0.31	0.05	0.17	0.094	0.06	0.26	0.09	0.241
7	0.5	0	0.02	0.33	−0.010	0.50	0.25	0.02	0.25	−0.014	0.02	0.26	0.04	0.191
8	0.7	1	0.78	0.30	0.000	1.63	1.63	0.82	0.24	0.000	0.89	0.25	0.95	0.276

[a] c_u^2 is the squared coefficient of variation and α_u is the lag-1 correlation of the interarrival times. The external arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

Table 3. A comparison of approximations and simulation for Fraker's model of eight queues in series: case 3.[a]

Input parameters			Simulation			Approximations								
						M/M/1	M/G/1	Fraker			QNA		Modified QNA	
Node No.	Traffic intensity ρ	Service squared coefficient of variation c_s^2	EW	c_u^2	α_u	EW	EW	EW	c_u^2	α_u	EW	c_u^2	EW	α_u
1	0.9	0	4.70	0.92	−0.007	8.10	4.05	4.05	1.00	0.000	4.05	1.00	4.05	0.000
2	0.9	1/4	2.19	0.15	0.195	8.10	5.06	1.80	0.19	0.162	2.32	0.35	2.37	0.161
3	0.5	1	0.24	0.31	0.008	0.50	0.50	0.23	0.28	0.017	0.24	0.27	0.24	0.044
4	0.7	1	1.08	0.57	−0.096	1.63	1.63	1.04	0.55	−0.099	1.12	0.45	1.09	−0.067
5	0.5	1/4	0.23	0.77	−0.035	0.50	0.31	0.19	0.84	−0.065	0.23	0.72	0.19	−0.062
6	0.3	0	0.01	0.65	0.008	0.13	0.07	0.03	0.72	−0.010	0.03	0.60	0.03	−0.002
7	0.3	1/8	0.01	0.64	0.015	0.13	0.07	0.03	0.68	0.010	0.03	0.57	0.04	0.029
8	0.7	1/8	0.55	0.62	0.008	1.63	0.92	0.55	0.66	0.020	0.49	0.53	0.52	0.051

[a] c_u^2 is the squared coefficient of variation and α_u is the lag-1 correlation of the interarrival times. The external arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

Table 4. The expected waiting time at the nodes with $\rho = 0.7$ in Fraker's eight queues in series.[a]

			Approximation methods			
		Simulated	Approximation minus simulation			
Case No.	Node No.	value of *EW*	M/M/1	M/G/1	Fraker	QNA
1	4	0.73	0.90	0.29	−0.11	−0.10
2	2	0.84	0.79	0.08	0.08	0.08
	8	0.78	0.85	0.85	0.04	0.11
3	4	1.08	0.55	0.55	−0.04	−0.04
	8	0.55	1.08	0.37	0.00	−0.02
4	3	1.52	0.11	0.11	−0.04	−0.04
	6	0.02	1.61	0.80	0.09	0.16
5	3	0.74	0.89	0.28	0.10	0.11
6	5	0.33	1.30	0.49	0.05	0.01
7	4	0.78	0.85	0.14	−0.06	−0.02
	7	0.17	1.46	0.65	0.02	−0.01
8	5	0.50	1.13	0.52	0.04	0.02
Average		0.67	0.96	0.43	0.01	0.03
Average absolute difference			0.96	0.43	0.05	0.06

[a] The value for each approximation is the approximation value minus the simulation value.

Tables 4 and 5 give summary evaluations of the approximations. Table 4 (5) gives the estimated errors (approximation value minus simulation value) for all nodes having traffic intensity 0.7 (0.5) in all eight cases, except those cases where that node is the first node because the approximations are all exact at the first node. Only the standard QNA approximation is given in Tables 4 and 5.

The tables show that the M/G/1 approximation is much better than the M/M/1 approximation and that the other approximations are much better than the M/G/1 approximation. The standard QNA approximation performs about the same as the Fraker approximation, both providing accuracy adequate for most engineering applications.

Tables 1–3 show that the adjustment in the expected waiting time using the lag-1 correlation is never large. Moreover, the modified QNA is not an improvement. This experiment suggests that the lag-1 correlation may not be of critical importance. At least this application of it does not do the job. Further experiments with longer simulation runs and other experimental designs seem worthwhile, however.

From Tables 1–3 it is evident that both Fraker's method and QNA track the simulation values of c_u^2 quite well. Fraker's approximation for the lag-1 correlation also tracks the simulation, at least to some extent. However, QNA performs erratically. In cases 1 and 3, QNA performs about the same as Fraker's procedure, yielding average errors in the lag-1 correlation of 0.027 and 0.025 as compared with 0.032 and 0.015. But in case 2, the QNA approximation performs very poorly for the last four nodes. It is interesting, however, that this has little effect on the approximation for the expected waiting time. Examination of the approximation formulas in Section 3 reveals an instability that can cause approximation errors to grow in successive iterations. First, (33) can get large when r is small and ρ is large ($r = 0$ and $\rho = 0.9$ at node 4 in case 2). Then the second term in (31) tends to perpetuate large or small values of α_u.

Table 5. The expected waiting time at the nodes with $\rho = 0.5$ in Fraker's eight cases of eight queues in series.[a]

			Approximation methods			
		Simulated	Approximation minus simulation			
Case No.	Node No.	value of *EW*	M/M/1	M/G/1	Fraker	QNA
1	2	0.30	0.20	0.20	0.08	0.08
	3	0.19	0.31	0.06	−0.06	−0.03
2	6	0.02	0.48	0.29	0.03	0.04
	7	0.02	0.48	0.23	0.00	0.00
3	3	0.24	0.26	0.26	−0.01	0.00
	5	0.23	0.27	0.09	−0.04	0.00
4	8	0.04	0.46	0.27	0.00	0.01
5	2	0.23	0.27	0.02	0.02	0.02
	7	0.05	0.45	0.23	−0.01	−0.01
6	2	0.09	0.41	0.19	0.06	0.08
	8	0.27	0.23	0.23	0.01	0.01
7	2	0.19	0.31	0.06	0.04	0.04
	3	0.31	0.19	0.19	0.13	0.12
8	3	0.09	0.41	0.19	0.04	0.08
	4	0.10	0.40	0.21	0.05	0.06
Average		0.16	0.34	0.18	0.02	0.03
Average absolute difference			0.34	0.18	0.04	0.04

[a] The value for each approximation is the approximation value minus the simulation value.

We also ran several versions of QNA using hybrid approximations as described in Section 4 with $x(\rho_1,\rho_1,c_u^2,c_v^2)$ a function only of ρ_1 and ρ_2. However, none of these offered a clear improvement over the standard version of QNA with $x = 0$. Again, further experimentation is in order.

6. SHIMSHAK'S EXPERIMENT

Shimshak [23] also developed approximations for the expected waiting times of queues in series and compared them with simulation. He simulated two single-server queues in series with three different renewal arrival processes. In Shimshak's experiments I, III, and IV the interarrival-time distribution is, respectively, exponential with $c_u^2 = 1.0$, hyperexponential (a mixture of two exponential distributions) with $c_u^2 = 4.0$, and Erlang (E_{10}) with $c_u^2 = 0.1$. (In experiment II the second queue has ten servers, so it will not be considered here.) Experiments III and IV are interesting additions to Section 5 because the external arrival process is non-Poisson.

In each case the arrival rate is 1, so that the mean service time at each node coincides with the traffic intensity. Each experiment consists of eight cases, containing all combinations of three variables each with two possible values. The three variables are three of the two traffic intensities ($\rho = 0.6$ and 0.8) and two Erlang service distributions (E_1 and E_{10}, $c_v^2 = 1.0$ or 0.1) at the two nodes. In each case one of these four variables is held fixed. In experiment I the service-time distribution at node 1 is always E_{10}; in experiments III and IV the traffic intensity at node 1 is always 0.8.

Shimshak's simulations were obtained using GPSS and the regenerative method.

The number of customers simulated in each case ranged from 15,000 to 25,000 depending on the traffic intensity. The statistical reliability is indicated by 95% confidence intervals. The simulation runs are significantly longer than Fraker's, so that the statistical accuracy is clearly better, but even longer runs are needed (e.g., see [1,2]).

The results of the simulations and the approximations appear in Tables 6–8. The Fraker, Page, and Marchal approximations are approximations devised by Shimshak using previous approximations. The M/M/1, M/G/1, QNA, and modified QNA approximations are added here and are as described in Section 5.

The results support the conclusions of Section 5. The M/M/1 and M/G/1 approximations are much worse than the others. As should be anticipated, in experiments III and IV with non-Poisson arrival processes they perform very poorly. In experiment III the M/G/1 approximation is even worse than the M/M/1 approximation. With high variability in the arrival process ($c_{a1}^2 = 4.0$) and low variability in the service times ($c_s^2 = 0.1$), two errors in the opposite directions cancel to some extent in the M/M/1 approximation. This phenomenon often occurs in models of packet-switched communication networks with bursty arrival processes and constant or nearly constant packet service times.

QNA performs as well as each of the three Shimshak approximations in each case. Overall QNA dominates the others because it is significantly better than Page and Marchal in experiment IV and Fraker does not apply to all cases in experiment III.

As in Section 5, the modified Fraker approximations obtained by using QNA plus lag-1 correlation as described in Section 3 does not provide improvement. Various hybrid methods based on Section 4 did not help either. However, there is some evidence that the errors in Table 6 are consistent with Section 4 to some extent. The asymptotic method would suggest putting more weight on the variability parameter of the external arrival process when ρ_2 is relatively large, e.g., in systems 6 and 8 in Table 6. Since $c_{a1}^2 = 1.0 > 0.1 = c_{s1}^2$, we should anticipate observing more congestion than the QNA approximation based on the stationary-interval method predicts. And, indeed, the approximations fall about 10% below the simulation estimates in these two cases.

7. CONCLUSIONS

In this article we have provided a theoretical framework for approximating departure processes and single-server queues in series. We have indicated what the asymptotic method and stationary-interval method in [26] yield for departure processes. We have indicated three ways the approximations might be improved: (1) using the third moment, (2) using the lag-1 correlation (Section 3), and (3) developing a hybrid procedure (Section 4). However, we examined the last two methods and did not find an improvement. On the positive side, the relatively simple implementation of the stationary-interval method in QNA seems to work quite well.

The simulation experiments were far from conclusive, however. We relied on the previous simulations of others, which were too short to provide good statistical accuracy. The possibility of developing improved approximations for departure processes remains, perhaps by building on the techniques in this article. There is certainly a need for additional well-designed simulation experiments to develop a better understanding of departure processes. We need better reference points for developing approximations. In general, simulation and numerical methods have hardly begun to be

Table 6. Shimshak's experiment I: a comparison of approximations and simulation of the expected total waiting time in two queues in series, with Poisson external arrival process ($c_{a1}^2 = 1$) and Erlang (E_1 or E_{10}) service-time distributions ($c_{vi}^2 = 1.0$ or 0.1).[a]

	Parameters					Approximations						
System no.	ρ_1	ρ_2	c_{v1}^2	c_{v2}^2	Simulation estimate	M/M/1	M/G/1 (asymptotic method)	Fraker	Page	Marchal	QNA (st. int.)	Modified QNA
1	0.6	0.6	0.1	1.0	1.20	1.80	1.40	1.19	1.20	1.18	1.25	1.30
					(±0.09)	(+0.50)	(+0.17)	(−0.01)	(0.00)	(−0.02)	(+0.04)	(+0.08)
2	0.8	0.6	0.1	1.0	2.27	4.10	2.66	2.30	2.31	2.28	2.38	2.42
					(±0.23)	(+0.81)	(+0.17)	(+0.01)	(+0.02)	(+0.01)	(+0.05)	(+0.07)
3	0.6	0.6	0.1	0.1	0.78	1.80	0.99	0.77	0.84	0.84	0.84	0.89
					(±0.06)	(+1.31)	(+0.27)	(−0.02)	(+0.08)	(+0.07)	(+0.08)	(+0.14)
4	0.8	0.6	0.1	0.1	1.83	4.10	2.26	1.90	1.99	1.98	1.98	2.01
					(±0.22)	(+1.24)	(+0.23)	(+0.04)	(+0.09)	(+0.08)	(+0.08)	(+0.10)
5	0.6	0.8	0.1	1.0	3.41	4.10	3.70	3.07	3.10	3.06	3.21	3.27
					(±0.43)	(+0.20)	(+0.09)	(−0.10)	(−0.09)	(−0.10)	(−0.06)	(−0.04)
6	0.8	0.8	0.1	1.0	4.33	6.40	4.96	3.85	3.70	3.84	4.07	4.12
					(±0.60)	(+0.48)	(+0.15)	(−0.14)	(−0.10)	(−0.11)	(−0.06)	(−0.05)
7	0.6	0.8	0.1	0.1	1.93	4.10	2.26	1.60	1.73	1.72	1.77	1.83
					(±0.27)	(+1.12)	(+0.17)	(−0.17)	(−0.10)	(−0.11)	(−0.08)	(−0.05)
8	0.8	0.8	0.1	0.1	2.48	6.40	3.52	2.43	2.58	2.57	2.63	2.68
					(±0.29)	(+1.58)	(+0.42)	(−0.02)	(+0.04)	(+0.04)	(+0.06)	(+0.08)
Average relative error						+0.91	+0.21	−0.05	−0.01	−0.02	+0.01	+0.04
Average absolute relative error						0.91	0.21	0.06	0.07	0.07	0.06	0.08

[a] The arrival rate is 1 in each case. The 95% confidence interval is in parentheses below the simulation estimate. The estimated relative error appears below the approximation in parentheses. This is the approximation value minus the simulation value divided by the simulation value. The M/G/1 approximation coincides with the asymptotic method and QNA coincides with the stationary-interval method.

Table 7. Shimshak's experiment III: a comparison of approximations and simulation of the expected total waiting time in two queues in series, with hyperexponential external arrival process ($c_{v1}^2 = 4$) and Erlang (E_1 or E_{10}) service-time distributions ($c_{vi}^2 = 1.0$ or 0.1).[a]

	Parameters					Approximations					
System no.	ρ_1	ρ_2	c_{v1}^2	c_{v2}^2	Simulation estimate	M/M/1	M/G/1	Fraker	Marchal	QNA (st. int.)	Modified QNA
1	0.8	0.6	1.0	1.0	9.08	4.10	4.10	10.30	10.39	9.39	9.56
					(±1.38)	(−0.55)	(−0.55)	(+0.13)	(+0.14)	(+0.03)	(+0.05)
2	0.8	0.6	0.1	1.0	6.49	4.10	6.49	—	7.91	7.72	7.94
					(±0.73)	(−0.37)	(−0.51)		(+0.22)	(+0.19)	(+0.22)
3	0.8	0.6	1.0	0.1	8.55	4.10	3.18	10.17	9.86	8.98	9.16
					(±1.20)	(−0.52)	(−0.63)	(+0.19)	(+0.15)	(+0.05)	(+0.07)
4	0.8	0.6	0.1	0.1	6.01	4.10	2.26	—	7.43	7.31	7.54
					(±0.73)	(−0.32)	(−0.62)		(+0.24)	(+0.22)	(+0.25)
5	0.8	0.8	1.0	1.0	12.31	6.40	6.40	13.50	13.54	12.92	13.09
					(±2.26)	(−0.48)	(−0.48)	(+0.10)	(+0.10)	(+0.05)	(+0.06)
6	0.8	0.8	0.1	1.0	9.64	6.40	4.96	—	10.78	10.67	10.88
					(±1.33)	(−0.34)	(−0.49)		(+0.12)	(+0.11)	(+0.13)
7	0.8	0.8	1.0	0.1	11.13	6.40	4.96	12.55	11.90	11.49	11.65
					(±1.37)	(−0.42)	(−0.55)	(−0.13)	(+0.07)	(+0.03)	(+0.05)
8	0.8	0.8	0.1	0.1	7.40	6.40	3.52	—	9.21	9.23	9.44
					(±0.95)	(−0.14)	(−0.52)		(+0.24)	(+0.25)	(+0.28)
Average relative error						−0.39	−0.54	+0.14	+0.16	+0.12	+0.14
Average absolute relative error						0.39	0.54	0.14	0.16	0.12	0.14

[a] The arrival rate is 1 in each case. The 95% confidence interval is in parentheses below the simulation estimate. The estimated relative error appears below the approximation in parentheses. This is the approximation value minus the simulation value divided by the simulation value.

Table 8. Shimshak's experiment IV: a comparison of approximations and simulation of the expected total waiting time in two queues in series, with Erlang (E_{10}) external arrival process ($c^2_{a1} = 0.1$) and Erlang (E_1 or E_{10}) service-time distributions ($c^2_{si} = 1.0$ or 0.1).[a]

	Parameters					Approximations						
System no.	ρ_1	ρ_2	c^2_{s1}	c^2_{s2}	Simulation estimate	M/M/1	M/G/1	Fraker	Page	Marchal	QNA	Modified QNA
1	0.8	0.6	1.0	1.0	2.30 (±0.19)	4.10 (+0.78)	4.10 (+0.78)	2.24 (−0.03)	2.30 (0.00)	2.21 (−0.04)	2.29 (0.00)	2.25 (−0.02)
2	0.8	0.6	0.1	1.0	0.59 (±0.04)	4.10 (+5.95)	3.18 (+4.39)	0.58 (−0.02)	0.65 (+0.10)	0.59 (0.00)	0.56 (−0.05)	0.56 (−0.05)
3	0.8	0.6	1.0	0.1	1.95 (±0.45)	4.10 (±1.10)	3.18 (±0.63)	1.81 (−0.07)	1.92 (−0.02)	1.84 (−0.06)	1.89 (−0.03)	1.85 (−0.05)
4	0.8	0.6	0.1	0.1	0.25 (±0.02)	4.10 (+15.40)	2.26 (+8.04)	0.27 (+0.08)	0.38 (+0.52)	0.35 (+0.40)	0.20 (−0.20)	0.20 (−0.20)
5	0.8	0.8	1.0	1.0	3.84 (±0.33)	6.40 (+0.67)	6.40 (+0.67)	4.27 (+0.11)	4.30 (+0.12)	4.26 (+0.11)	4.21 (+0.10)	4.17 (+0.09)
6	0.8	0.8	0.1	1.0	1.82 (±0.19)	6.40 (+2.52)	4.96 (+1.73)	1.77 (−0.03)	1.85 (+0.02)	1.75 (−0.04)	1.85 (+0.02)	1.85 (+0.02)
7	0.8	0.8	1.0	0.1	2.68 (±0.52)	6.40 (+1.39)	4.96 (+0.84)	2.79 (+0.04)	2.92 (+0.09)	2.88 (+0.07)	2.77 (+0.03)	2.73 (+0.02)
8	0.8	0.8	0.1	0.1	0.46 (±0.02)	6.40 (+12.91)	3.52 (+6.65)	0.50 (+0.09)	0.61 (+0.33)	0.58 (+0.26)	0.43 (−0.06)	0.43 (−0.06)
Average relative error						+5.09	+2.97	+0.02	+0.15	+0.09	−0.02	−0.03
Average absolute relative error						5.09	2.97	0.06	0.15	0.12	0.06	0.06

[a] The arrival rate is 1 in each case. The 95% confidence interval is in parentheses below the simulation estimate. The estimated relative error appears below the approximation in parentheses. This is the approximation value minus the simulation value divided by the simulation value.

exploited to the extent they should be in order to understand basic operations frequently arising in stochastic models such as superposition and departure. Simulation has great potential for contributing to our understanding of relatively simple models as well as relatively complex models. The simplicity provides the environment corresponding to a controlled experiment. There is a need for more studies such as those done by Albin [1,2].

Perhaps the main conclusion, in support of Shimshak [23] and other earlier work, is that relatively simple approximations are a useful alternative and/or supplement to simulation of queueing systems that are beyond existing exact analytic methods. Moreover, it seems possible to develop a theoretical framework and supporting methodology for developing such simple approximations. Creating appropriate approximations will no doubt always be largely an art and will depend on the context, but analysis and simulation can provide a useful perspective and the proper tools.

ACKNOWLEDGMENT

I am grateful to my colleague Anne Seery for writing the QNA program and running it to obtain some of the data in the tables. I am grateful to the referees for their suggestions.

REFERENCES

[1] Albin, S. L., "Approximating Queues with Superposition Arrival Processes," Ph.D. dissertation, Department of Industrial Engineering and Operations Research, Columbia University, 1981.

[2] Albin, S. L., "Approximating a point process by a renewal process, II: superposition arrival processes to queues," *Operations Research*, **32** (1984), in press.

[3] Berman, M., and Westcott, M., "On queueing systems with renewal departure processes," *Advances in Applied Probability*, **15**, 657–673 (1983).

[4] Brazie, C. L., "Simulation Models of Satellite Airport Systems," Ph.D. dissertation, Cornell University, 1972.

[5] Cooper, R. B., *Introduction to Queueing Theory, 2nd ed.*, North-Holland, New York, 1981.

[6] Daley, D. J., "Queueing output processes," *Advances in Applied Probability*, **8**, 395–415 (1976).

[7] Disney, R. L., *Queueing Networks and Applications*, Johns Hopkins University Press, Baltimore, in press.

[8] Fraker, J. R., "Approximate Techniques for the Analysis of Tandem Queueing Systems," Ph.D. dissertation, Department of Industrial Engineering, Clemson University, 1971.

[9] Franken, P., König, D., Arndt, U., and Schmidt, V., *Queues and Point Processes*, Akademie-Verlag, Berlin, 1981.

[10] Friedman, H. D., "Reduction methods for tandem queueing systems," *Operation Research*, **13**, 121–131 (1965).

[11] Gelenbe, E., and Mitrani, I., *Analysis and Synthesis of Computer Systems*, Academic, New York, 1980.

[12] Gnedenko, B. V., and Kolmogorov, A. N., *Limit Distributions for Sums of Independent Random Variables, 2nd ed.*, Addison-Wesley, Reading, MA, 1968.

[13] Green, L., "A limit theorem on subintervals of interrenewal times," *Operations Research*, **30**, 210–216 (1982).

[14] Iglehart, D. L., and Whitt, W., "Multiple channel queues in heavy traffic, II: sequences, networks and batches," *Advances in Applied Probability*, **2**, 355–369 (1970).

[15] Kraemer, W., and Langenbach-Belz, M., "Approximate formulae for the delay in the queueing system GI/G/1," *Congressbook, Eighth International Teletraffic Congress*, Melbourne, 235, 1–8, 1976.

[16] Kuehn, P. J., "Approximate analysis of general queueing networks by decomposition," *IEEE Transactions in Communications*, **27,** 113–126 (1979).

[17] Marchal, W. G., "Estimating the delay in series queues," Graduate School of Business, the University of Toledo, Toledo, Ohio, 1981.

[18] Marshall, K. T., "Some inequalities in queueing," *Operations Research*, **16,** 651–665 (1968).

[19] Melamed, B., "Characterizations of Poisson traffic streams in Jackson queueing networks," *Advances in Applied Probability*, **11,** 422–438 (1979).

[20] Nelson, R. T., "A Simulation Study and Analysis of a Two-Station, Waiting-Line Network Model," Ph.D. dissertation, UCLA, 1965.

[21] Rosenshine, M., and Chandra, M. J., "Approximate solutions for some two-state tandem queues, Part 1: Individual arrivals at the second stage," *Operations Research*, **23,** 1155–1166 (1975).

[22] Sevcik, K. C., Levy, A. I., Tripathi, S. K., and Zahorjan, J. L., "Improving approximations of aggregated queueing network subsystems. *Computer Performance*, K. M. Chandy and M. Reiser, Eds., North-Holland, Amsterdam, 1977, pp. 1–22.

[23] Shimshak, D. G., "A comparison of waiting time approximations in series queueing systems," *Naval Research Logistics Quarterly*, **26,** 499–509 (1979).

[24] Weber, R. R., "The interchangeability of tandem $\cdot/M/1$ queues in series," *Journal of Applied Probability*, 690–695 (1979).

[25] Whitt, W., "Approximating a point process by a renewal process: The view through a queue, an indirect approach," *Management Science*, **27,** 619–636 (1981).

[26] Whitt, W., "Approximating a point process by a renewal process I: two basic methods," *Operations Research*, **30,** 125–147 (1982).

[27] Whitt, W., "Queue tests for renewal processes," *Operations Research Letters*, **2,** 7–12 (1982).

[28] Whitt, W., "The queueing network analyzer," *Bell System Technical Journal*, **62,** 2779–2815 (1983).

[29] Whitt, W., "Performance of the queueing network analyzer," *Bell System Technical Journal*, **62,** 2817–2843 (1983).

[30] Whitt, W., "The best order for queues in series," *Management Science*, in press.

[31] Whitt, W., "Departures from a queue with many busy servers," *Mathematics of Operations Research*, **9** (1984), in press.