VARIABILITY FUNCTIONS FOR PARAMETRIC-DECOMPOSITION APPROXIMATIONS OF QUEUEING NETWORKS

by

Ward Whitt

AT&T Bell Laboratories
Murray Hill, NJ 07974-0636

July 22, 1993

Revision: July 25, 1994

To appear in *Management Science*

Abstract

We propose an enhancement to the parametric-decomposition method for calculating approximate steady-state performance measures of open queueing networks with non-Poisson arrival processes and non-exponential service-time distributions. Instead of using a *variability parameter* c_a^2 for each arrival process, we suggest using a *variability function* $c_a^2(\rho)$, $0 < \rho < 1$, for each arrival process; i.e., the variability parameter should be regarded as a function of the traffic intensity ρ of a queue to which the arrival process might go. Variability functions provide a convenient representation of different levels of variability in different time scales for arrival processes that are not nearly renewal processes. Variability functions enable the approximations to account for long-range effects in queueing networks that cannot be addressed by variability parameters. For example, the variability functions provide a way to address the heavy-traffic bottleneck phenomenon, in which exceptional variability (either high or low) in the input has little impact in a series of queues with low-to-moderate traffic intensities, and then has a big impact when it reaches a later queue with a relatively high traffic intensity. The variability functions also enable the approximations to characterize irregular periodic deterministic external arrival processes in a reasonable way; i.e., if there are no batches, then $c_a^2(\rho)$ should be 0 for ρ near 0 or 1, but $c_a^2(\rho)$ can assume arbitrarily large values for appropriate intermediate ρ. We present a full network algorithm with variability functions, showing that the idea is implementable. We also show how simulations of single queues can be effectively exploited to determine variability functions for difficult external arrival processes.

Key words: queueing networks, tandem queues, approximations, parametric-decomposition approximations, two-moment approximations, heavy traffic, squared coefficient of variation.

1. Introduction and Summary

The parametric-decomposition approximation method is an approach to approximating the steady-state performance measures of open queueing networks with non-Poisson arrival processes and non-exponential service-time distributions; see Whitt (1983, 1994), Segal and Whitt (1989), Buzacott and Shanthikumar (1992), Suri, Sanders and Kamath (1993), and references therein. The idea is to analyze the individual queues separately after approximately characterizing the arrival processes. So far, the arrival processes have been characterized by a few parameters, usually two, one to represent the rate and the other to represent the variability. The arrival rates are obtained as the solution to the familiar traffic rate equations, and are exact. In most schemes, the variability parameters can be regarded as *squared coefficients of variation* (SCVs, i.e., the variance divided by the square of the mean) of an interarrival time in an approximating renewal process, i.e., assuming the interarrival times are i.i.d. (independent and identically distributed). The approximating variability parameters are obtained as the solution to a second set of equations called the traffic variability equations; that is the major approximation. Computing performance measures for each queue given its parameters also involves an approximation, but the quality of this approximation is relatively well understood, and is pretty good; e.g., see Whitt (1993). The greatest difficulty with the parametric-decomposition method is determining appropriate variability parameters for the arrival processes to the queues.

There also is a problem in combining the performance measures of the individual queues to describe the total network performance. There obviously is no problem with means, because the mean of a sum is always the sum of the means, but there can be problems with variances and distributions because the steady-state queue-length distributions at different queues are in general not independent. The standard approach is to act as if the queues are mutually independent when computing total network performance measures, but this can be a source of error. We do not attempt to address this problem here. We note that the Brownian model approximations in

Harrison and Nguyen (1990, 1993) and Dai, Nguyen and Reiman (1994) *do* directly address this problem.

Another problem is properly treating different classes in multi-class queueing networks; see Bitran and Tirupati (1988), Fendick, Saksena and Whitt (1989, 1991) and Whitt (1994). For the multi-class case, *we ideally should have parameters characterizing the arrival process of each class at each queue.* We do not address this problem here either; here we consider only the aggregate arrival process at each queue.

All two-parameter approximation methods assume that the external arrival processes and service-time distributions can be approximately characterized by rates and variability parameters at the outset. As before, there tends to be little difficulty with the rates providing that the processes are stationary (which is a nontrivial assumption). We assume that the arrival and service processes can indeed be regarded as stationary, so that the rates are relatively easy to specify.

However, significant difficulties can arise with the variability parameters. The standard approach is to assume that the sequences of service times and interarrival times (in the external arrival processes) are mutually independent sequences of i.i.d. random variables. Then we let the variability parameter for each process be the SCV. For service times, this independence property is often reasonable. (For exceptions, see Fendick, Saksena and Whitt (1989).) However, for the external arrival processes, this independence is less reasonable. *Unfortunately, experience indicates that arrival processes that are not nearly Poisson (M) or deterministic (D) are often not nearly renewal processes either.* The Poisson process is often a good arrival process model, but non-deterministic arrival processes often tend to be non-Poisson in large part because of significant dependence among the interarrival times. This is particularly true when the variability is substantially greater than in a Poisson process.

The practical consequence of dependence among interarrival times is that it may not be possible to characterize the variability of the external arrival process well by a single parameter. Indeed, even for renewal arrival processes the variability may not be well characterized by the SCV; e.g., see Whitt (1989). However, the difficulty is much greater with non-renewal processes; e.g., see Fendick and Whitt (1989). Dependence among the interarrival times can often be detected by observing, in the terminology of Whitt (1982), that there is a significant difference between the *stationary-interval* and the *asymptotic-method* approximations for stationary point processes. The stationary-interval method lets the arrival-process variability parameter c_a^2 simply be the SCV of an interarrival time (ignoring any dependence). The asymptotic method lets c_a^2 be the limit of the normalized variance of the partial sums, i.e.,

$$c_{AM}^2 = \lim_{n \to \infty} \frac{\mathrm{Var}(S_n)}{n(EX_1)^2} \ , \tag{1}$$

where $S_n = X_1 + \ldots + X_n$ with X_i being the interarrival times. For a renewal process, these methods agree, but in general they are different because $\mathrm{Var}(S_n)$ includes the covariance terms, i.e.,

$$\mathrm{Var}(S_n) = n \, \mathrm{Var}(X_1) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}(X_i, X_j) \ . \tag{2}$$

The limit in (1) means that the asymptotic method incorporates *all* the covariance terms. Fendick and Whitt (1989) discuss ways to incorporate only *some* of the covariance terms.

With Brownian models, it is customary to exploit heavy-traffic limit theorems to determine variability parameters. Then the variability parameter of an external arrival process can be thought of as being determined by the asymptotic method in (1). Our point is that for some models neither the asymptotic method nor the stationary-interval method is appropriate. Moreover, for some models no single variability parameter is appropriate.

The importance of dependence in external arrival processes deserves further emphasis, because proposed queueing network approximation methods are typically evaluated only for renewal arrival processes. The approximations are introduced to capture the effects of non-Poisson arrival processes, but it is tacitly assumed that this non-Poisson variability appears in the form of non-Poisson renewal processes. We contend that this is typically not the case.

Considerable work has been done on superpositions of large numbers (e.g., 100) of non-Poisson arrival processes, motivated by the fact that typically the stationary-interval method dictates that $c_a^2 \approx 1$, while the asymptotic method can dictate something very different (e.g., $c_a^2 \approx 20$), and the actual appropriate value is somewhere in between; see Whitt (1982, 1983, 1985), Albin (1982, 1984), Newell (1984), Sriram and Whitt (1986), Heffes and Lucantoni (1986), Fendick, Saksena and Whitt (1989, 1991) and Fendick and Whitt (1989). Hence, if a given external arrival process to which we wish to assign a variability parameter happens to be such a superposition of non-Poisson processes, then it is quite likely to have substantial dependence among its interarrival times. Moreover, even if all external arrival processes are renewal processes, network operations can make internal arrival processes not nearly renewal processes.

We have also encountered the arrival-process variability-parameter problem several times in honest efforts to choose the variability parameters of external arrival processes by working with arrival process data. In each case the modeller simply estimated the SCV directly by using the sample mean and the sample variance (but ignoring the dependence). The modellers obtained estimates of SCVs that seemed to them *much smaller* than they would have predicted based on their ‘‘intuitive feel’’ for the level of variability. *A priori*, it seemed that the variability was significantly greater than in a Poisson process, but the direct SCV estimation yielded an SCV value much less than 1. After doing estimates related to (1), it became clear that this was due to dependence, in particular, positive correlations among the interarrival times. A specific striking example (involving other issues as well) is discussed in Fendick, Saksena and Whitt (1989); see especially Sections II.B and III.

Expressed another way, if the arrival process is a renewal process, then it is usually reasonable to regard the SCV as the appropriate variability parameter, but if the arrival process is not nearly a renewal process, then the appropriate variability parameter typically depends on the traffic intensity of the queue to which the arrival process goes. In light traffic the stationary-interval method tends to be good, while in heavy-traffic the asymptotic method tends to be good. Indeed, it is known that the asymptotic method is asymptotically correct as the traffic intensity of the queue approaches its critical value. (This is implied by Theorem 1(a) of Iglehart and Whitt (1970).) However, at typical traffic intensities, something in between is needed. Thus, the approach to superposition in Albin (1984) and Whitt (1983) was to make the variability parameter depend on the traffic intensity of the queue to which the superposition process goes. This approach is fine for one queue, but it can fail when the superposition arrival process feeds into two queues in series. Current methods can make big errors at the second queue if the traffic intensity at the first queue is much less than the traffic intensity at the second queue. Then the actual congestion at the second queue is approximately the same as if the first queue were not there, but previous approximations do not consistently predict that; see §5 below.

The important point is that *there can be long-range variability effects.* As shown by Whitt (1988) and Suresh and Whitt (1990a,b), exceptional variability (either high or low) in an external arrival process can have relatively little impact upon an initial series of queues with low-to-moderate traffic intensities, and then later have a big impact upon a distant queue with a relatively high traffic intensity. Simulations reveal the *heavy-traffic bottleneck phenomenon:* A queue with a relatively high traffic intensity in a queueing network can be influenced by an external arrival process entering several queues away essentially the small as if the other intermediate queues were not there. (See Section 6 for more discussion.) Clearly, local adjustment of variability parameters is inadequate for addressing this problem.

Variability functions provide a means to address long-range variability effects. *We propose characterizing each arrival process by its rate and a variability function* $c_a^2(\rho)$, $0 < \rho < 1$, *indicating the approximate SCV as a function of the traffic intensity* ρ *at a queue to which the arrival process might go.* We are still thinking of $c_a^2(\rho)$ as the SCV of an interarrival time in an approximating renewal process, but now with the value depending on ρ.

For this extension to be useful, we need to do two things. First, we need to develop a calculus for transforming variability functions when we encounter the familiar network operations of flowing through a queue (departure), superposition (merging) and splitting, paralleling Section IV of Whitt (1983). Second, we need to develop a method for assigning variability functions to non-renewal external arrival processes. We indicate how to do these things here in Sections 2 and 3.

In Section 4 we show how the variability functions enable us to treat irregular periodic deterministic arrival processes in a reasonable way; these are periodic deterministic arrival processes with non-constant interarrival times (e.g., the interarrival-times $2/3$, $4/3$, $2/3$, $4/3, \ldots$). The asymptotic method yields $c_a^2 = 0$, but we often want $c_a^2(\rho) > 0$. These irregular periodic deterministic arrival processes dramatically show that variability functions can provide a big improvement.

In Section 5 we show how variability functions enable us to reasonably treat two queues in series with a non-renewal arrival process such as a superposition process. In Section 6 we show that the variability functions may enable us to obtain reasonable approximations for the difficult tandem-queue networks in Suresh and Whitt (1990a) exhibiting the heavy-traffic bottleneck phenomenon. In Section 7 we discuss ways to estimate how accurate are the approximations, both before and after applying a network algorithm. We suggest providing estimates for the range of possible congestion values due to variability uncertainty (not being sure of the appropriate approximating SCVs). In Section 8 we state our conclusions.

2. An Extended Calculus for Network Operations

In this section we develop an algorithm for producing the variability functions of the internal (net) arrival processes in an open queueing network, given the internal arrival rates, the external-arrival-process variability functions and the service-time variability parameters. The internal arrival rates plus the service-time means yield the traffic intensities at all queues. We assume that all queues have unlimited waiting space and the first-in first-out service discipline. In this paper we consider only a single-class network of multi-server queues, but the procedure can be extended to multiple classes (by aggregation) as in Whitt (1983). (A more careful treatment of multiple classes is still needed though.) We regard the service-time variability parameters as being independent of the traffic intensity or, stated differently, dependent only upon the fixed traffic intensity prevailing at that queue.

This section extends Section IV of Whitt (1983). We treat superposition, splitting and departure in the first three subsections and then discuss their synthesis into traffic variability equations in the final subsection. We do not do new experimental testing here. For each operation, we propose a variability function transformation that has already been tested quite extensively. The new network calculus is constructed and supported by applying previous results for variability parameters. Consequently, the new algorithm would be the same for one queue, but *the new algorithm is very different for networks.*

We point out that the procedures below make the variability function $c_{ai}^2(\rho_i)$ at queue i be asymptotically correct in heavy traffic, i.e., as the traffic intensity ρ_i at queue i approaches 1. Thus, $c_{ai}^2(1-) \equiv \lim_{\rho \uparrow 1} c_a^2(\rho)$ agrees with the asymptotic method in Whitt (1982) and the individual bottleneck approximation in Reiman (1990).

2.1 Superposition

To a large extent, a reasonable way to treat superposition processes is already indicated in Whitt (1983). There the variability parameter assigned to the superposition process is allowed to depend on the traffic intensity of the following queue. We now just let all the variability parameters depend upon ρ.

In particular, suppose that n streams with rates λ_i and variability functions $c_{ai}^2(\rho)$, $1 \leq i \leq n$, are to be superposed. The rate of the superposition process is of course $\lambda = \sum_{i=1}^{n} \lambda_i$. We let the approximating variability function $c_a^2(\rho)$ for the superposition process be

$$c_a^2(\rho) = w(\rho) \sum_{i=1}^{n} (\lambda_i/\lambda) c_{ai}^2(\rho) + 1 - w(\rho) , \quad 0 < \rho < 1 , \tag{3}$$

where

$$w(\rho) = [1 + 4(1-\rho)^2 \nu]^{-1} \tag{4}$$

$$\nu = [\sum_{i=1}^{n} (\lambda_i/\lambda)^2]^{-1} . \tag{5}$$

When the component streams have equal rates, ν in (5) is the number of streams. More generally, ν is reduced to account for uneven rates. Note that $w(\rho) \rightarrow 0$ as ν increases, but ν needs to be larger as ρ increases. When $w(\rho)$ is small, $c_a^2(\rho) \approx 1$ reflecting the convergence of superposition processes to a Poisson process as the number of component streams increases.

Note that (3)–(5) agrees with Section 4.3 of Whitt (1983), except that $c_{ai}^2(\rho)$ and $c_a^2(\rho)$ in (3) are allowed to depend on ρ. For superpositions of renewal processes at a single queue, the method here coincides with Whitt (1983). This approximation has already been studied quite extensively; e.g., see Albin (1982, 1984) and Sriram and Whitt (1986).

2.2 Independent Splitting

Since independent splitting of a renewal process is exactly a renewal process, it is natural to use the exact formula for the SCV, as in Section 4.4 of Whitt (1983). In particular, if a stream with variability function $c^2(\rho)$ is split into n streams, with each point being assigned to stream i with probability p_i, and successive selections being independent, then we let the variability function of the i^{th} split stream be

$$c_i^2(\rho) = p_i c^2(\rho) + 1 - p_i \ . \tag{6}$$

It is important to note, however, that if the routing is *not* independent (or Markovian) splitting, then (6) may be *not nearly appropriate*. In particular, this is true with multiple classes and deterministic routing; see Bitran and Tirupati (1988) and Whitt (1994).

2.3 Departure

Recent experience with departure processes in Suresh and Whitt (1990a,b) indicates that we need to make significant changes from the procedure in Whitt (1983). The formula

$$c_d^2 = \rho_*^2 c_s^2 + (1 - \rho_*^2) c_a^2 \tag{7}$$

in (38) of Whitt (1983), where ρ_* is the traffic intensity at the queue from which the customers depart, describes the effect on following queues reasonably well when the following queue has a traffic intensity less than or equal to ρ_*, but not so well otherwise. Indeed, the nine-queue and ten-queue tandem network examples in Suresh and Whitt (1990b) (and similar 100-queue examples) show that the original arrival process can have a big impact on even a distant later queue with a substantially higher traffic intensity.

Hence, we propose the following class of departure variability functions for an arrival process coming to a single-server queue with traffic intensity ρ_* and service-time SCV c_s^2 (which we assume does *not* depend on ρ_*):

$$c_d^2(\rho_*,\rho) \;=\; \alpha(\rho_*,\rho)\, c_s^2 \;+\; (1-\alpha(\rho_*,\rho))\, c_a^2(\rho)\;, \tag{8}$$

where $\alpha(\rho_*,\rho)$ is a function that is decreasing in ρ and increasing in ρ_*. (Here ρ is the traffic intensity at the following queue.) A specific function $\alpha(\rho_*,\rho)$ in (8) that was developed for two queues in series is

$$\alpha(\rho_*,\rho) \;=\; \rho_*^2(1-\rho^{10})\;; \tag{9}$$

see (4.2) of Suresh and Whitt (1990a).

The sequential bottleneck method in Reiman (1990) is implemented by having

$$\alpha(\rho_*,\rho) \;=\; \begin{cases} 1\,, & \rho \;<\; \rho_* \\ 0\,, & \rho \;\geq\; \rho_*\;. \end{cases} \tag{10}$$

It is intuitively clear that an appropriate function α should change more smoothly than in (10). It is also evident that we should have $\alpha(\rho_*,\rho) \to 0$ as $\rho \to 1$ with $0 < \rho_* < 1$ and $\alpha(\rho_*,\rho) \to 1$ as $\rho_* \to 1$ with $0 < \rho < 1$. Note that (10) satisfies this condition, but (9) does not.

Another alternative is

$$\alpha(\rho_*,\rho) \;=\; \rho_* \min\{1,\,(1-\rho)^r/(1-\rho_*)^r\}\;; \tag{11}$$

e.g., for $r = 2$. Formula (11) reduces to (7) for $\rho < \rho_*$, but smoothly approaches (10) for $\rho > \rho_*$.

We have not yet carefully studied the case of departure processes from queues with more than one server. We anticipate that a modification of (39) of Whitt (1983) similar to (8) should be reasonable, e.g.,

$$c_d^2(\rho_*,\rho) \;=\; 1 \;+\; \frac{\alpha(\rho_*,\rho)}{\sqrt{m}}(c_s^2-1) \;+\; (1-\alpha(\rho,\rho_*))(c_a^2(\rho)-1) \tag{12}$$

for $\alpha(\rho_*,\rho)$ in (9) or (11).

We have assumed that the service-time variability parameter c_s^2 does not depend on ρ_* in (8) and (12). If we believed that the service times were not mutually independent, then it would be natural to work with service-time variability functions $c_s^2(\rho)$ too. We would then modify (8) and (12) simply by replacing c_s^2 by $c_s^2(\rho_*)$. This would make the service-time influence the arrival variability function only via the value $c_s^2(\rho_*)$ at the actual traffic intensity ρ_* of the queue, which is fixed and known.

2.4 Synthesis in a Network

Paralleling Section 4.7 of Whitt (1983), the operations above lead to systems of linear equations to determine the arrival variability functions at each queue. Here, however, there is a system of linear equations for each traffic intensity ρ, $0 < \rho < 1$. For practical purposes, we could consider only the 99 traffic intensity values $0.01, 0.02, \ldots, 0.99$. In fact, for a network of n nodes with traffic intensities ρ_i, $1 \leq i \leq n$, it suffices to consider only the n systems of linear equations with the n given traffic intensities ρ_i, $1 \leq i \leq n$. At queue i, we only need to know the arrival variability function $c_a^2(\rho)$ evaluated at $\rho = \rho_i$. Note that the traffic intensities can be computed prior to determining the arrival variability functions, so these values of ρ_i are indeed known.

In particular, just as in Whitt (1983), the *traffic-rate equations* are

$$\lambda_j = \lambda_{0j} + \sum_{i=1}^{n} \lambda_i q_{ij} , \quad 1 \leq j \leq n , \tag{13}$$

where λ_j is the net arrival rate to queue j, λ_{0j} is the external arrival rate to queue j and q_{ij} is the proportion of departures from queue i that go next to queue j. (We do not consider the customer creation factors γ_j here.) The associated traffic intensity at queue j is

$$\rho_j = \lambda_j \tau_j / s_j , \quad 1 \leq j \leq n , \tag{14}$$

where s_j is the number of servers at queue j and τ_j is the mean service time at queue j.

Next, for each value of the traffic intensity ρ, the new *traffic-variability equations* are

$$c_{aj}^2(\rho) = w_j(\rho)[(\lambda_{0j}/\lambda_j)c_{0j}^2(\rho) + \sum_{i=1}^{n}(\lambda_i q_{ij}/\lambda_j)c_{ij}^2(\rho)] + 1 - w_j(\rho) , \ 1 \le j \le n , \quad (15)$$

where

$$w_j(\rho) = [1 + 4(1-\rho)^2 \nu_j]^{-1} , \quad (16)$$

$$\nu_j = [(\lambda_{0j}/\lambda_j)^2 + \sum_{i=1}^{n}(\lambda_i q_{ij}/\lambda_j)^2]^{-1} , \quad (17)$$

$$c_{ij}^2(\rho) = q_{ij}c_{di}^2(\rho_i,\rho) + 1 - q_{ij} , \quad (18)$$

and

$$c_{di}^2(\rho_i,\rho) = 1 + \frac{\alpha(\rho_i,\rho)}{\sqrt{s_i}}(c_{si}^2 - 1) + (1-\alpha(\rho_i,\rho))(c_{ai}^2(\rho) - 1) \quad (19)$$

for $\alpha(\rho_*,\rho)$ in (11) with $r = 2$. Equivalently,

$$c_{aj}^2(\rho) = A_j(\rho) + \sum_{i=1}^{n} B_{ij}(\rho)c_{ai}^2(\rho) , \quad 1 \le j \le n , \quad (20)$$

where

$$A_j(\rho) = 1 + w_j(\rho)[(\lambda_{0j}/\lambda_j)c_{0j}^2(\rho) - 1] + \sum_{i=1}^{n}(\lambda_i q_{ij}/\lambda_j)(1-q_{ij})$$
$$+ \sum_{i=1}^{n}(\lambda_i q_{ij}^2/\lambda_j)(1 + \frac{\alpha(\rho_i,\rho)}{\sqrt{s_i}}(c_{si}^2 - 1) - 1 + \alpha(\rho_i,\rho)) \quad (21)$$

and

$$B_{ij}(\rho) = w_j(\rho)(\lambda_i q_{ij}^2/\lambda_j)(1-\alpha(\rho_i,\rho)) . \quad (22)$$

3. Assigning Variability Functions to External Arrival Processes

Introducing a more complicated characterization of arrival processes obviously makes initial model specification more difficult. Now a user of a software package implementing a network algorithm such as partially described in Section 2 must specify an entire variability function for each external arrival process instead of just a variability parameter, but having the input in this form communicates some important information. Just as having variability parameters in the model communicates their potential importance, so does having variability functions in the model communicate their potential importance.

The obvious default values for variability functions in the model input are the constant variability functions of renewal processes. In other words, if we are willing to assume that the external arrival processes are renewal processes, then nothing more need be done. Then we can let the external arrival variability functions $c_{0j}^2(\rho)$ be constant functions with the value of the SCV. However, we recommend being careful at this step.

We now suggest procedures for specifying variability functions for non-renewal external arrival processes. For this purpose, we suggest applying the indirect methods in Whitt (1981), Fendick and Whitt (1989), and Fendick, Saksena and Whitt (1991). The idea is to characterize the arrival process by observing the congestion it produces in a convenient test queue. In particular, we suggest choosing measurement units so that the arrival rate is fixed at 1, then estimating or calculating the mean steady-state waiting time (before beginning service) $EW(\rho)$ in a convenient single-server queue with i.i.d. service times having a known distribution with mean ρ and SCV c_s^2, and then letting $c_a^2(\rho)$ be the value of the interarrival-time SCV that makes a reasonable approximation for $EW(\rho)$ correct, as a function of ρ. Here we let the approximation be

$$EW(\rho) = \frac{\rho^2 (c_a^2(\rho) + c_s^2) g(\rho, c_a^2(\rho), c_s^2)}{2(1-\rho)} , \tag{23}$$

where

$$g(\rho, c_a^2(\rho), c_s^2) = \begin{cases} \exp\{-2(1-\rho)/3\rho)([1-c_a^2(\rho)]^2/[c_a^2(\rho)+c_s^2])\}, & c_a^2(\rho) \le 1, \\ \exp\{-(1-\rho)([c_a^2(\rho)-1]/[c_a^2(\rho) + 4c_s^2])\}, & c_a^2(\rho) \ge 1, \end{cases} \tag{24}$$

as in Krämer and Langenbach-Belz (1976); see Section 3 of Whitt (1981) and (44) of Whitt (1983). For simplicity, we might replace g above by 1. Values of g quite different from 1 occur when ρ is small.

Since the formula for $EW(\rho)$ in (23) is strictly increasing in $c_a^2(\rho)$ for each pair (ρ, c_s^2) with $0 < \rho < 1$, it is not difficult to find the appropriate value $c_a^2(\rho)$ for each ρ. A simple specific procedure is, for each ρ, to initially let $c_a^2(\rho) = 1$ (which makes $g(\rho, c_a^2(\rho), c_s^2) = 1$) and then solve for $c_a^2(\rho)$ explicitly by

$$c_a^2(\rho) = \left[\frac{2(1-\rho)EW(\rho)}{\rho^2 g(\rho, c_a^2(\rho), c_s^2)} - c_s^2\right]^+, \quad 0 < \rho < 1, \tag{25}$$

where $[x]^+ = \max\{x, 0\}$. We then successively calculate $g(\rho, c_a^2(\rho), c_s^2)$ (using (24) and the most recent estimate of $c_a^2(\rho)$) and $c_a^2(\rho)$ (using (25) and the most recent estimate of g) until there is negligible change.

The most natural choices of service-time distributions are exponential and deterministic. We can often obtain $EW(\rho)$ as a function of ρ analytically. This is the case for many BMAP/G/1 queues, where the arrival process is a batch Markovian arrival process (BMAP); see Lucantoni (1991, 1993) and Abate, Choudhury and Whitt (1994). Such algorithms provide a basis for numerically calculating $EW(\rho)$ for a very large class of arrival processes (models). Since the service-time distribution can be general in this algorithm, it is natural to use a scaled version of a representative service-time distribution in the actual network.

From data, we suggest estimating $EW(\rho)$ as a function of ρ by simulating a G/G/1 queue with the given arrival process (with units chosen so that it has rate 1) and general service times of length ρ, as a function of ρ. For greater realism, it is natural to choose the service-time distribution prevailing at the first queue or one close to it, but for computational speed and simplicity it is natural to use deterministic service times. Note that the simulation must be performed for each arrival process and each traffic intensity of interest, but the simulations involve only a single queue, which typically is much easier than simulating an entire queueing network. Moreover, it is useful to have a good characterization of the arrival process, for better understanding as well as further analysis.

The procedure for obtaining the external-arrival-process variability function above is closely related to the characterizations of input to queues (service times plus arrival process) in Fendick and Whitt (1989). The focus there is on the steady-state workload (remaining service time at an arbitrary time or virtual waiting time) $Z(\rho)$. By Brumelle's formula ((72) in Fendick and Whitt), $EZ(\rho)$ is related to $EW(\rho)$ exactly by

$$EZ(\rho) = \rho EW(\rho) + \frac{\rho^2 (c_s^2 + 1)}{2} , \tag{26}$$

assuming that the arrival rate is 1, that the mean service time is ρ and that the service times are i.i.d. and independent of the arrival process. Hence, given $EZ(\rho)$, we can calculate $EW(\rho)$ using (26). Fendick and Whitt (1989) focus on the *normalized mean workload,* defined by

$$c_z^2 (\rho) = 2(1-\rho) EZ(\rho)/\rho^2 . \tag{27}$$

They use the *index of dispersion for work* (IDW), which describes the total input process of work, to obtain an estimate of $c_z^2 (\rho)$. Those methods can be applied to produce first $c_z^2 (\rho)$ and then successively $EZ(\rho)$, $EW(\rho)$ and $c_a^2 (\rho)$ using (27), (26) and (23). When the service times are i.i.d. and independent of the arrival process, the IDW $I_w (t)$ is simply related to the *index of dispersion for counts* (IDC) $I_c (t)$ by $I_w (t) = I_c (t) + c_s^2$. If $A(t)$ counts the number of arrivals in the interval $[0,t]$, then the IDC is the function

$$I_c(t) \;=\; \frac{\mathrm{Var}\, A(t)}{EA(t)} \;\;,\;\; t > 0 \;, \tag{28}$$

which can be estimated from data or calculated numerically. In summary, we can apply the previous methods in Fendick and Whitt (1989) to convert an estimate of the IDC $I_c(t)$ into the desired variability function $c_a^2(\rho)$.

If we focus on the waiting times (at arrival epochs), then it is natural to use the *index of dispersion for intervals* (IDI), which is defined by

$$I_i(n) \;=\; \frac{n\,\mathrm{Var}\, S_n}{(ES_n)^2} \;, \tag{29}$$

for positive integers n. Paralleling (9) and (13) of Fendick and Whitt (1989), we could approximate $c_a^2(\rho)$ by $I_i(n(\rho))$ where $n(\rho)$ is an approximate customer index as a function of ρ, which might be

$$n(\rho) \;=\; \left\lceil \rho I_i(\infty)/2(1-\rho)^2 \right\rceil \;, \tag{30}$$

where $\lceil x \rceil$ is the least integer greater than or equal to x.

4. Irregular Periodic Deterministic Arrival Processes

In the next three sections we provide concrete evidence showing that variability functions can provide significant improvements in parametric-decomposition approximations. In this section we consider a class of arrival processes for which it is easy to see that variability functions instead of variability parameters are desirable. This is the class of irregular periodic deterministic arrival processes. Related stochastic processes appear in manufacturing and communication networks. The periodic deterministic character implies that we should have $c_a^2(\rho) \to 0$ as $\rho \to 1$, but we need $c_a^2(\rho) > 0$ for $\rho < 1$. Moreover, if there are no batch arrivals, then we should also have $c_a^2(\rho) \to 0$ as $\rho \to 0$.

The standard periodic deterministic arrival process is the D process with constant interarrival times. By ‘‘irregular periodic deterministic’’ arrival processes, we mean all other periodic deterministic arrival processes. A simple example has successive interarrival times $3/2,\ 1/2,\ 1,\ 3/2,\ 1/2,\ 1,\ldots$. This arrival process has period 3 and arrival rate 1.

In this section we consider a special case of these deterministic arrival processes, which we denote by $D(\alpha,k)$. Let the parameter k be a positive integer and let the parameter α be a real number with $0 < \alpha < 1$. Let there be single arrivals at epochs $nk + j\alpha$ for integers n and j, where $0 \leq j \leq k-1$ and $n \geq 0$. it is easy to see that this arrival process is periodic with period k. There are k arrivals in interval $[nk,(n+1)k)$ for every n, so that the arrival rate is 1. The $D(\alpha,k)$ process is a special case of processes considered in Sections VI. C and VII of Fendick, Saksena and Whitt (1989).

We now apply the method of Section 3 to determine appropriate variability functions for $D(\alpha,k)$ arrival processes. For the D(α,k)/D/1 queue, it is easy to compute the exact mean steady-state (long-run average) waiting time $EW(\rho)$ for any deterministic service time ρ, $0 \leq \rho \leq 1$. It suffices to consider a single cycle, since the deterministic waiting-time process starts empty before the arrivals at times nk, $n \geq 0$. Indeed, here

$$EW(\rho) = \begin{cases} 0\ , & 0 \leq \rho \leq \alpha \\ \dfrac{(k-1)}{2}(\rho-\alpha)\ , & \alpha < \rho \leq 1\ . \end{cases} \tag{31}$$

Using (25) with $g \equiv 1$, we see that $c_a^2(\rho) = 0$ for $\rho \leq \alpha$ and

$$c_a^2(\rho) = \frac{(1-\rho)(k-1)(\rho-\alpha)}{\rho^2}\ ,\ \alpha < \rho \leq 1\ , \tag{32}$$

so that $c_a^2(\rho) \to 0$ as $\rho \to 1$, but that $c_a^2(\rho) > 0$ for $\alpha < \rho < 1$. Moreover, (32) indicates that $c_a^2(\rho)$ is directly proportional to $k-1$, so that it can be arbitrarily large.

In summary, the variability function $c_a^2(\rho)$ in (32) yields the correct answer in $D(\alpha,k)/D/1$ queues, and it is not nearly constant. For a specific example, let $\alpha = 1/4$ and $\rho = 1/2$. Then (31) yields $c_a^2(1/2) = (k-1)/2$. For $k = 21$, we have $c_a^2(1/2) = 10$, while $c_a^2(0+) = c_a^2(1/4) = c_a^2(1-) = 0$.

We now consider a small network example to demonstrate the importance of using variability functions in a network context. Suppose that we consider the $D(\alpha,k)/D/1 \rightarrow D/1$ model, i.e., two deterministic queues in series with this periodic irregular deterministic arrival process having $\alpha = 1/4$ and $k = 21$. Suppose that the arrival rate is 1 and that the service times at queues 1 and 2 are $1/4$ and $1/2$. Both the new variability-function and old variability-parameter approaches yield $c_{a1}^2(0.25) = 0.0$ and the exact mean waiting time at queue 1, $EW_1 = 0.0$. The variability-parameter approach thus yields $c_{a2}^2 = 0$ and $EW_2 = 0$ using (8) with $c_a^2(\rho_1) = 0$ and any choice of α. On the other hand, the variability-function approach using (11) with $r = 2$ yields $c_{a1}^2(0.5) = 10.0$. Then, using (8) and (11) with $r = 2$, we obtain $\alpha(\rho_1,\rho_2) = 1/9$ and $c_d^2(\rho_1,\rho_2) = 80/9$. Next using (23) with $g = 1$, we obtain $EW_2 = 20/9 \approx 2.22$. Clearly the variability-function value of 2.22 is a much better approximation for the exact value $EW_2 = 2.50$ than 0.0.

5. The Case of the Disappearing Queue

The example at the end of the last section illustrates a general phenomenon. To see the importance of variability functions, it suffices to consider two queues in series with a non-renewal arrival process. Let $c_{ai}^2(\rho)$, $i=1,2$, be the variability functions at the two queues and let the traffic intensities at the two queues by ρ_1 and ρ_2 where $\rho_1 << \rho_2$. Using variability parameters with decomposition is tantamount to using $c_{a1}^2(\rho_1)$ to characterize the arrival process to the first queue. However, if ρ_1 is small and we use (7), then the fixed variability parameter at the second queue will be approximately $c_{a1}^2(\rho_1)$. *However, if ρ_1 is very small, then the actual congestion at* *the second queue will be obviously approximately the same as if the first queue were not there;* e.g., see Whitt (1988) and Suresh and Whitt (1990b). In other words, we should have approximately $c_{a2}^2(\rho_2) = c_{a1}^2(\rho_2)$. If $c_{a1}^2(\rho_1)$ is not nearly equal to $c_{a1}^2(\rho_2)$, then we will make a big error at the second queue. In contrast, variability functions allow us to have $c_{a2}^2(\rho_2) = c_{a1}^2(\rho_2)$.

This situation can easily occur with a superposition arrival process to two queues in series. Then $c_{a1}^2(\rho)$ can change greatly as ρ changes. Hence, if $\rho_1 << \rho_2$, then we can make a big error at the second queue. This specific phenomenon was considered by Albin and Kai (1986), but they did not propose variability functions.

We now consider a simple concrete example to demonstrate the potential benefit of variability functions. Our example has two queues in series with exponential service times and an external arrival process that is the superposition of 20 i.i.d. renewal processes each with interarrival-time $SCV = 20$. Each interarrival-time distribution in a component renewal process is a mixture of two exponentials (with balanced means). The total arrival rate is 1 and the mean service times at queues 1 and 2 are 0.6 and 0.9. Using simulation and (25) with $g \equiv 1$ we obtain estimated appropriate variability functions $c_{a1}^2(0.6) \approx 2.2$, $c_{a1}^2(0.9) \approx 13.3$ and $c_{a2}^2(0.9) \approx 11.8$. (In our experiment the 90% confidence intervals were about $\pm 20\%$.) Using variability *parameters* with $c_{a1}^2 = 2.2$ plus (8) and (9), we obtain $c_{a2}^2 = 1.9$. Using variability *functions* with $c_{a1}^2(0.9) = 13.3$ plus (8) and (9), we obtain $c_{a2}^2(0.9) \approx 10.5$. Obviously 10.5 is much closer to 11.8 than 1.9.

In summary, it is desirable for the overall approximation scheme to pass the test of the disappearing queue; i.e., it is desirable for the approximation to perform reasonably well in the reduced network obtained by letting the mean service times at any queue decrease to 0. The approximation with the queue present with 0 service times should be approximately the same as the approximation for the reduced network with the queue in question removed from the network.

6. The Heavy-Traffic Bottleneck Phenomenon

Variability functions offer a way to approach the heavy-traffic bottleneck phenomenon discussed by Suresh and Whitt (1990b). Suresh and Whitt consider a network of nine queues in series with exponential service times and a renewal external arrival process having rate 1. The mean service times at the first eight queues are all 0.6, while the mean service time at the ninth queue is 0.9. Simulation results show that non-Poisson variability in the internal arrival process is hardly evident at the eighth queue, but reappears strongly at the ninth queue; see Table 1. When the external arrival-process SCV is 8.0 (0.0), the mean steady-state waiting time at the ninth queue is significantly greater (less) than if the arrival process to the ninth queue were Poisson.

Since heavy-traffic theory predicts that the congestion at the last queue would be asymptotically the same as if the previous eight queues were not there, as the traffic intensity ρ at the last queue approaches the critical value 1, we call this the heavy-traffic bottleneck phenomenon. It is significant that variability functions provide a possible way to address this problem. However, this difficulty with a large number of queues in series is not easily resolved. For example, approximation (9), which works well for two queues in series, provides essentially the same approximations as the previous parametric-decomposition approximations at queues 8 and 9. It still predicts that by queue 8 the remaining weight on the external arrival process SCV should be negligible.

On the other hand, the heavy-traffic approximation (10) also does not perform so well, because the eight queues do decrease the variability of the arrival process to queue 9 somewhat. We can do reasonably well for this example by using approximation (11) with $r = 2$, but this approximation does not perform so well for only two queues in series. We give the values for approximation (11) with $r = 2$ in Table 1 as well.

Upon reflection, it seems unreasonable to expect great accuracy for approximations with the heavy-traffic bottleneck phenomenon. The reason is that the appropriate variability function is very steep near 1 at the last queue. This means that slight perturbations of the model (e.g., the traffic intensity) can lead to big changes in the performance measures. In such situations, we contend that it is probably more useful to identify this instability and indicate the range of possible performance measures than it is to predict the particular performance measure accurately. We turn to this issue in the next section.

7. Diagnosing Possible Approximation Errors

In this section we consider how we can diagnose possible significant errors in the parametric-decomposition approximation for open queueing networks. We first consider anticipating errors *before* applying the algorithm, then we consider estimating errors *after* applying the algorithm.

Before applying the algorithm, we first see if any of the external arrival processes need to be specified by non-constant variability functions $c_{0i}^2(\rho)$. If $c_{0i}^2(\rho_1)$ needs to be very different from $c_{0i}^2(\rho_2)$ for two traffic intensity values ρ_1 and ρ_2, then that should be regarded as an initial warning sign. Given that $c_{0i}^2(\rho_1)$ is quite different from $c_{0i}^2(\rho_2)$, then we should look at the actual traffic intensities prevailing at the queues within the network. If the set of traffic intensities does not include two values at which $c_{0i}^2(\rho)$ assumes very different values, then there should be no major problem. However, if there are two queues j and k with traffic intensities ρ_j and ρ_k for which $c_{0i}^2(\rho_j)$ is very different from $c_{0i}^2(\rho_k)$, then there is a potential problem. If external arrivals to queue i cannot reach both queues j and k, then this is not a real problem.

If the external variability function values $c_{0i}^2(\rho)$ and service-time variability parameter values c_{si}^2 are all relatively close to 1, then the approximation should be relatively well behaved. We call this the nearly-Jackson case. In the nearly-Jackson case, the formulas in subsections 2.1–2.4

make all the internal arrival-process variability functions nearly 1. Since the algorithm is exact in the Jackson case, the approximation tends to be excellent in the nearly-Jackson case.

If some variability values are not near 1, then superposition is a source of difficulty because for $\rho \approx 1$, the variability function value obtained from (3) is just the convex combination of the variability function values of the component streams, while for lower traffic intensities (3) makes the superposition variability-function values closer to 1. Hence, a network with superpositions of many streams that can have variability-function values quite different from 1 should be regarded as a warning sign.

From (8), we see that departure tends to cause difficulty at a queue with traffic intensity ρ only if $c_a^2(\rho)$ is not nearly equal to c_s^2. Moreover, this can cause significant fluctuations of variability functions only if there are other queues with significantly higher traffic intensities. Thus, the presence of *both* significantly different variability values and significantly different traffic intensities is a warning sign for the departure operation.

The discussion above indicates how we can anticipate significant errors in the approximation *before* applying a network algorithm. After applying a network algorithm, we can detect possible errors by looking at fluctuations in the internal-arrival variability functions $c_{ai}^2(\rho)$. At a queue with traffic intensity ρ, significant errors are unlikely if $c_{ai}^2(\rho)$ does not fluctuate much for traffic intensity values in a neighborhood of ρ, e.g., for ρ in the interval (ρ_-, ρ_+), where

$$\rho_- = \max\{0, \rho - 0.1\} \quad \text{and} \quad \rho_+ = \min\{1, \rho + 0.1\} . \tag{33}$$

To reveal possible errors in the approximation we suggest computing approximate lower and upper bounds for performance measures based on approximate lower and upper ‘‘bounds’’ for the variability function value $c_{ai}^2(\rho)$, defined by

$$c_{aiL}^2(\rho) = \inf\{c_{ai}^2(\rho) : \rho_- \le \rho \le \rho_+\} \tag{34}$$

and

$$c_{aiU}^2(\rho) = \sup\{c_{ai}^2(\rho) : \rho_- \le \rho \le \rho_+\} \tag{35}$$

8. Conclusions

In this paper we have proposed replacing variability parameters of arrival processes in parametric-decomposition approximations for open queueing networks by variability functions. The proposed variability functions $c_a^2(\rho)$ are squared coefficients of variation (SCVs) of interarrival times in renewal-process approximations, where the SCV is regarded as a function of the traffic intensity ρ in a following single-server queue. When the arrival process is actually a renewal process, we would simply let $c_a^2(\rho)$ be the SCV of an interarrival time for all ρ. The greater generality is intended for non-renewal processes.

Having $c_a^2(\rho)$ depend on ρ reflects the fact that the level of variability of a non-renewal stationary point process depends on the time scale. As ρ increases (up to the critical value 1) in a following queue, the relevant time scale for determining congestion increases as well. Since we are concerned with queueing applications, it is natural to make the independent variable ρ instead of the customer n or time t. We suggest drawing upon previous work by Fendick and Whitt (1989) and Fendick, Saksena and Whitt (1991) to relate $c_a^2(\rho)$ to indices of dispersion (for counts and intervals).

In Section 2 we showed that the traffic variability equations for variability parameters in Section IV of Whitt (1983) can easily be extended to the setting of variability functions. For a network with n nodes, it is only necessary to solve at most n systems of linear traffic variability equations in order to obtain the relevant variability function values for the network.

In Section 3 we suggested determining the variability function $c_a^2(\rho)$ from models or data by first choosing the measuring units so that the arrival rate is fixed at 1, then calculating or estimating the expected steady-state waiting time (before beginning service) in a G/G/1 or G/D/1 queue with this arrival process and service times having mean ρ, and finally finding the value of $c_a^2(\rho)$ such that a standard approximation for this mean steady-state waiting time yields the observed value, i.e., by using formula (23).

The variability functions lead to a significantly different treatment of departure processes. In Sections 4–6 we showed that the variability functions enable us to obtain reasonable approximations for some difficult small models, including the heavy-traffic bottleneck phenomenon in tandem networks discussed by Suresh and Whitt (1990b). In Section 7 we discussed how we can diagnose possible errors in the approximation.

We think that variability functions provide an improved framework for parametric-decomposition approximations. Thus we think that they warrant further study. For multiclass queues, it seems desirable to work with the variability functions for each class (instead of aggregating into a single class, as in Whitt (1983)).

References

Abate, J., G. L. Choudhury and W. Whitt, ‘‘Exponential Approximations for Tail probabilities in Queues, I: Waiting Times,’’ *Oper. Res.,* (1994), to appear.

Albin, S. L., ‘‘On Poisson Approximations for Superposition Arrival Processes in Queues,’’ *Management Sci.* 28 (1982), 126-137.

Albin, S. L., ‘‘Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues,’’ *Oper. Res.,* 32 (1984), 1133-1162.

Albin, S. L. and S. Kai, ‘‘Approximation for the Departure Process of a Queue in a Network,’’ *Naval Res. Logist. Quart.,* 33 (1986), 129-143.

Bitran, G. R. and D. Tirupati, ‘‘Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference,’’ *Management Sci.,* 34 (1988), 75-100.

Buzacott, J. A. and J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems,* Prentice-Hall, Englewood Cliffs, NJ, 1992.

Dai, J. G., V. Nguyen and M. I. Reiman, ‘‘Sequential Bottleneck Decomposition: An Approximation Method for Open Queueing Networks,’’ *Oper. Res.* 42 (1994), 119-136.

Fendick, K. W., V. R. Saksena and W. Whitt, ‘‘Dependence in Packet Queues,’’ *IEEE Trans. Commun.,* 37 (1989), 1173-1183.

Fendick, K. W., V. R. Saksena and W. Whitt, ‘‘Investigating Dependence in Packet Queues with the Index of Dispersion for Work,’’ *IEEE Trans. Commun.,* 39 (1991), 1231-1244.

Fendick, K. W. and W. Whitt, ‘‘Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue,’’ *Proceedings IEEE,* 77 (1989), 171-194.

Harrison, J. M. and V. Nguyen, ‘‘The QNET Method for Two-Moment Analysis of Open Queueing Networks,’’ *Queueing Systems,* 6 (1990), 1-32.

Harrison, J. M. and V. Nguyen, ‘‘Brownian Models of Multiclass Queueing Networks: Current

Status and Open Problems,'' *Queueing Systems,* 13 (1993), 5-40.

Heffes, H. and D. M. Lucantoni, ''A Markov-Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance,'' *IEEE J. Sel. Areas Commun.,* SAV-4 (1986), 856-868.

Iglehart, D. L. and W. Whitt, ''Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks and Batches,'' *Adv. Appl. Prob.* 2 (1970), 355-369.

Krämer, W. and M. Langenbach-Belz, ''Approximate Formulae for the Delay in the Queueing System GI/G/1,'' *Eighth International Teletraffic Congress,* Melbourne, (1976) 235-1-8.

Lucantoni, D. M., ''new Results for the Single Server Queue with a Batch markovian Arrival Process,'' *Stoch. Models,* 7 (1991), 1-46.

Lucantoni, D. M., ''The BMAP/G/1 Queue: A Tutorial,'' *Models and Techniques for Performance Evaluation of Computer and Communications Systems, Proceedings of Performance '93,* L. Donatiello and R. Nelson (eds.), Springer-Verlag, New York, 1993, 330-358.

Newell, G. F., ''Approximations for Superposition Arrival Processes in Queues,'' *Management Sci.,* 30 (1984), 623-632.

Reiman, M. I., ''Asymptotically Exact Decomposition Approximations for Open Queueing Networks,'' *Oper. Res. Letters,* 9 (1990), 363-370.

Segal, M. and W. Whitt, ''A Queueing Network Analyzer for Manufacturing,'' in *Teletraffic Science for New Cost-Effective Systems ITC 12,* M. Bonatti (ed.), North-Holland, Amsterdam, 1989, 1146-1152.

Sriram, K. and W. Whitt, ''Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data,'' *IEEE J. Sel. Areas. Commun.,* SAR-4 (1986), 833-846.

Suresh, S. and W. Whitt, ''Arranging Queues in Series: A Simulation Experiment,'' *Management Science,* 36 (1990a), 1080-1091.

Suresh, S. and W. Whitt, ‘‘The Heavy-Traffic Bottleneck Phenomenon in Open Queueing Networks,’’ *Oper. Res. Letters*, 9 (1990b), 355-362.

Suri, R., J. L. Sanders and M. Kamath, ‘‘Performance Evaluation of Production Networks,’’ Chapter 5 in *Handbooks in OR and MS*, vol. 4, S. C. Graves et al. (eds.), Elsevier, Amsterdam, 1993, 199-286.

Whitt, W., ‘‘Approximating a Point Process by a Renewal Process: the View Through a Queue, an Indirect Approach,’’ *Management Sci.* 27 (1981), 619-636.

Whitt, W., ‘‘Approximating a Point Process by a Renewal Process, I: Two Basic Methods,’’ *Oper. Res.*, 30 (1982), 124-147.

Whitt, W., ‘‘The Queueing Network Analyzer, ‘‘*Bell System Tech. J.*, 62 (1983), 2779-2815.

Whitt, W., ‘‘Queues with Superposition Arrival Processes in Heavy Traffic,’’ *Stoch. Proc. Appl.*, 21 (1985), 81-91.

Whitt, W., ‘‘A Light-Traffic Approximation for Single-Class Departure Processes from Multi-Class Queues,’’ *Management Sci.*, 34 (1988), 1333-1346.

Whitt, W., ‘‘An Interpolation Approximation for the Mean Workload in a GI/G/1 Queue,’’ *Oper. Res.*, 37 (1989), 936-952.

Whitt, W., ‘‘Approximations for the GI/G/m Queue,’’ *Production and Operations Mgmt.*, 2 (1993), 114-161.

Whitt, W., ‘‘Towards Better Parametric-Decomposition Approximations for Open Queueing Networks,’’ *Ann. Oper. Res.*, 48 (1994), 221-248.

		high variability $c_a^2(\rho) = 8.0$ for all ρ	low variability $c_a^2(\rho) = 0.0$ for all ρ
queue 9	simulation estimate	30.0 ± 5.1	5.03 ± 0.22
	M/M/1 approximation	8.1	8.1
	old parametric-decomposition approximation	8.9	8.0
	heavy-traffic approximation $c_{a9}^2 = c_{a1}^2$	36.5	4.05
	new approximation (11) with $r = 2$	31.7	4.72
queue 8	simulation estimate	1.42 ± 0.07	0.775 ± 0.013
	M/M/1 approximation	0.90	0.90
	old and new parametric-decomposition approximations	1.04	0.88
	heavy-traffic approximation $c_{a8}^2 = c_{a1}^2$	4.05	0.45

Table 1. A comparison of approximations with simulation estimates at queues 9 and 8 in the tandem network of nine queues in Section 6.