MANAGEMENT SCIENCE
Vol. 27, No. 6, June 1981
Printed in U.S.A.

APPROXIMATING A POINT PROCESS BY A RENEWAL PROCESS: THE VIEW THROUGH A QUEUE, AN INDIRECT APPROACH*

WARD WHITT†

This paper investigates simple approximations for stochastic point processes. As in several previous studies, the approximating process is a renewal process characterized by the first two moments of the renewal interval. The approximating renewal-interval distribution itself is a convenient distribution with these two moments; it is constructed from exponential building blocks, e.g., the hyperexponential distribution. Here the moments of the renewal interval are chosen to produce the same level of congestion when the renewal process serves as an arrival process in a test queueing system as is produced when the general point process is the arrival process. The procedure can be applied to predict the behavior of a new service mechanism in a queueing system with a complicated arrival process; then we use the system with the old service mechanism as the test system. But the test system can also be an artificial device to approximate any point process. This indirect approximation procedure extends the equivalent random method and related techniques widely used in teletraffic engineering.

(QUEUES; POINT PROCESSES; APPROXIMATION)

1. Introduction

Suppose we plan to change the service mechanism in a queueing system and we want to describe the level of congestion to expect after the change. For example, we may intend to change the number of servers, the queue discipline or the individual service rate. Suppose, in addition, that the level of congestion before the change is observable or partially observable, but the arrival process is either unobservable or intractable. The question is: how can we approximately describe the level of congestion that will prevail after the change.

We present an approximation procedure that can be used to answer this question. Here is how it works: Based on a partial description of the congestion before the change, we approximate the arrival process by a convenient renewal process characterized by two parameters. We then solve the model which has this renewal arrival process and the new service mechanism.

There is an important assumption in this procedure that deserves mention. We are tacitly assuming that changing the service mechanism will not itself cause a change in the arrival process. Our procedure aims to describe the level of congestion that will result if the same arrival process is the input to the new service mechanism. In many applications, e.g., a traffic intersection, the arrival process may change in response to a change in the service mechanism. However, when this phenomenon can occur, approximate methods are if anything even more appropriate.

The choice of the family of approximating arrival processes is of course important. The processes should be sufficiently simple that the queueing model with the new service mechanism can be solved. At the same time, the approximation should be

*Accepted by Marcel F. Neuts, former Departmental Editor; received February 19, 1980. This paper has been with the author 1 month for 1 revision.

†Bell Laboratories, Holmdel, New Jersey.

0025-1909/81/2706/0619$01.25

reasonably accurate, i.e., the model with the approximating arrival process should adequately describe the congestion in the new system.

The greatest simplicity is achieved with a one-parameter approximation, with the natural one-parameter approximation being a Poisson process with the correct arrival rate. A Poisson approximation is often adequate, but also it is often not. Many arrival processes are significantly more or less "variable" than a Poisson process, and this variability matters. For example, overflow processes are much more variable (arrivals tend to occur in clumps), while scheduled arrivals (constant interarrival times) or minor deviations from scheduled arrivals are much less variable. The second parameter here is intended to represent this variability.

The approximating arrival processes we propose are renewal processes characterized by two parameters. To achieve the greatest simplicity, we let the renewal interval distribution be either the mixture of two exponential distributions (hyperexponential: H_2), the convolution of two exponential distributions (generalized Erlang: E_2^g) or the exponential distribution shifted by a constant (shifted exponential: M^d). By constructing the distribution from exponential building blocks, we can solve the new models either analytically or numerically after using the method of stages to construct a vector-valued continuous-time Markov chain; see Chapter 4 of Kleinrock [11].

The approximation procedure here obviously applies to an arrival process in a queueing system, but the procedure can be applied to any point process. Then the queueing system becomes an artificial device to help select an approximating renewal process. Either analytically or by simulation, we determine the level of congestion when the given point process is used as an arrival process in the "test" system. Then we can select the approximating renewal process by the procedures specified in this paper. The given point process could also be used to generate the service times instead of the arrival times; then the arrival process could be a Poisson process and we would act as if we had an $M/G/1$ model instead of a $GI/M/1$ model. This indirect method is very convenient when there is a handy simulation program for a queueing system and we need to quickly examine a point process in a different setting, e.g., a reliability or inventory system. It should be apparent that we could also use an inventory model as a test system to approximate a point process in a queueing system.

This paper is part of a larger investigation of simple approximations for point processes. Other methods for approximating point processes by the same renewal processes considered here are described in Whitt [18] and Albin [1]. Each approximation procedure uses certain properties of the point process to be approximated in order to specify the moments of the renewal interval. The stationary-interval and asymptotic methods in [18] focus on the point process in isolation. The hybrid methods in [1] for approximating the $\sum G_i/G/1$ queue begin to use the model context because the traffic intensity of the queueing system is used to refine the approximation. The view through a queue here goes further in this direction; it is entirely based on the way the point process affects a queueing model.

It should be noted that there is considerable related literature; an effort to describe it was made in [18]. The indirect approach here is very closely related to the equivalent random method and its variants; see §4.7 of Cooper [3], Kuczura [13], Wallström [17] and Wilkinson [20], [21]. The equivalent random method is a procedure for approximately characterizing a complicated arrival process and the blocking experienced by these arrivals in a service system with finitely many exponential servers and no waiting room. In the usual setting—alternate routing in a telephone network—the arrival process is the superposition of overflow processes, which is quite intractable. The arrival process is characterized by two parameters representing its rate and variability. The variability parameter, called the peakedness, is determined by looking at the congestion produced by this arrival process in an associated infinite-server system. Hence in the equivalent random method the infinite-server system is used as a test system. From the arrival rate and peakedness, good approximations of the blocking probabilities have been obtained for loss systems.

In the original version of the equivalent random method, the complicated arrival process is approximated by the overflow process from a single loss system, but later, upon the suggestion of W. S. Hayward, Kuczura [13] introduced a renewal process approximation. Thus the main ideas in the present paper come from this earlier work; we show that these ideas can be generalized.

This paper is organized as follows: In §2 we consider approximations for arrival processes when the service mechanism is a single exponential server with infinite waiting room. In §3 we extend the procedure to general service-time distributions. We consider approximations when the service mechanism is an infinite group of exponential servers in §4. With the aid of the equivalent random method, in §5 we apply the infinite-server approximation procedure to generate approximations for arrival processes when the service mechanism is a finite group of exponential servers without a waiting room. Finally, we discuss a few extensions and directions for further research in §6. Additional information appears in [19], which is available from the author. For the most part, [19] contains additional results describing the performance of the indirect method. While the indirect method often performs well, it can perform rather poorly. Users are advised to test and tune the method if possible before serious applications.

2. Approximating the Arrival Process in a $G/M/1/\infty$ System

Suppose we have a $G/M/1/\infty$ queueing system: a single server with infinite waiting room, first-in first-out queue discipline, mutually independent service times with a common exponential distribution having mean μ^{-1}, and an arrival process that is independent of the service times but otherwise general (G instead of GI). We have in mind a stationary arrival process, so that we can view the system in steady state. (The existence of steady-state distributions in this generality is discussed at length in Franken, König, Arndt and Schmidt [6].) We shall use observations about the level of congestion in this system, i.e., various characteristics of the queue length process, to approximate the arrival process by a renewal process. The idea is very simple: We act as if the arrival process were a renewal process. Noting that the $GI/M/1/\infty$ system with a given service rate is characterized by two parameters, we develop a method for obtaining these two parameters given queue length characteristics. We then show how to fit a renewal process to these parameters. We shall illustrate the procedure by approximating the arrival process of a $GI+GI/M/1/\infty$ queue, i.e., a $G/M/1/\infty$ system in which the arrival process is the superposition of two independent renewal processes.

2.1 *Background on the $GI/M/1/\infty$ System*

Consider the special case in which the arrival process is a renewal process; see Chapter II.3 of Cohen [2]. The stochastic behavior of this system is characterized by the interarrival time c.d.f. F and the mean service time μ^{-1}. Let λ^{-1} be the mean of F and let ϕ be its Laplace-Stieltjes transform, i.e.,

$$\phi(s) = \int_0^\infty e^{-sx}\,dF(x), \qquad s \geqslant 0. \tag{1}$$

It is significant that, given μ, the asymptotic stochastic behavior of this system depends on the interarrival time distribution only through two real parameters. The first is the traffic intensity $\rho = \lambda/\mu$, which we assume is less than one; and the second is the root σ, with $0 < \sigma < 1$, of the equation $s = \phi(\mu(1-s))$.

Let $Q(t)$ represent the number of customers in the system at time t and let $q_i(t) = E[Q(t)^i], t \geqslant 0$. If the interarrival time is nonlattice, which we assume, then $Q(t)$ converges in distribution to a finite random variable Q and $q_i(t)$ converges to a finite limit q_i for each i as $t \to \infty$, where

$$P(Q=k) = \begin{cases} 1-\rho, & k=0, \\ \rho(1-\sigma)\sigma^{k-1}, & k \geqslant 1, \end{cases} \tag{2}$$

and

$$q_1 = \rho(1-\sigma)^{-1} \quad \text{and} \quad q_2 = \rho(1+\sigma)(1-\sigma)^{-2}. \tag{3}$$

Note that the two moments q_1 and q_2 in (3) also characterize the asymptotic behavior. In particular, we can express ρ and σ in terms of q_1 and q_2:

$$\sigma = \frac{q_2 - q_1}{q_2 + q_1} \quad \text{and} \quad \rho = (1-\sigma)q_1 = \frac{2q_1^2}{q_2 + q_1}. \tag{4}$$

However, this is not the only way to express ρ and σ. For example, from (2) and (3) we have

$$\rho = 1 - P(Q=0) \quad \text{and} \quad \sigma = 1 - (\rho/q_1). \tag{5}$$

2.2 *Estimating ρ and σ in the $GI/M/1/\infty$ System*

Now suppose that we actually have a $GI/M/1/\infty$ system and we want to estimate the basic parameters ρ and σ based solely on observations of the stochastic process $\{Q(t), t \geqslant 0\}$. Suppose that we have a sequence of random observation times $\{t_n, n \geqslant 1\}$ independent of the queueing system such that $t_n \to \infty$ as $n \to \infty$. Then we obtain the following sequences of estimators for q_1 and q_2:

$$\hat{q}_{1n} = n^{-1} \sum_{k=1}^{n} Q(t_k), \qquad n \geqslant 1, \tag{6}$$

and

$$\hat{q}_{2n} = n^{-1} \sum_{k=1}^{n} Q(t_k)^2, \qquad n \geqslant 1. \tag{7}$$

Since the arrival epochs initiating busy periods are regeneration points, the variables $Q(t_k)$, $k \geqslant 1$, are close enough to being independent so that sequences $\{\hat{q}_{1n}\}$ and $\{\hat{q}_{2n}\}$ are strongly consistent, i.e., $\hat{q}_{1n} \to q_1$ and $\hat{q}_{2n} \to q_2$ with probability one as $n \to \infty$; see Iglehart [9]. Moreover, if the system starts in the steady state, i.e., if the process

$\{Q(t), t \geqslant 0\}$ is stationary, the estimators are unbiased: $E\hat{q}_{1n} = q_1$ and $E\hat{q}_{2n} = q_2$ for all n. It is of course important for these properties that the observation times be independent of the queueing system. For example, if the observation times were all arrival epochs, then $\hat{q}_{in}$ would have different limits; see (3.25) on page 210 of [2].

From (4), (6) and (7), we obtain the following natural estimators for σ and ρ:

$$\hat{\sigma}_n = \frac{\hat{q}_{2n} - \hat{q}_{1n}}{\hat{q}_{2n} + \hat{q}_{1n}} \quad \text{and} \quad \hat{\rho} = \frac{2\hat{q}_{1n}^2}{\hat{q}_{2n} + \hat{q}_{1n}} \,. \tag{8}$$

Obviously these estimators are strongly consistent too, but they involve products and ratios of random variables. Therefore, they need not be unbiased. We could introduce modifications to reduce the bias which may be important if the sample is small; see Iglehart [10] and references there. However, if the sample is not too small, then the estimators in (8) should be satisfactory.

From (5), we see that another pair of estimators for ρ and σ is

$$\bar{\rho}_n = 1 - n^{-1} \sum_{k=1}^{n} 1_{\{0\}}(Q(t_k)) \quad \text{and} \quad \bar{\sigma}_n = 1 - (\bar{\rho}_n / \hat{q}_{1n}), \tag{9}$$

where $1_A(x)$ is the indicator function of the set A. The estimator $\bar{\rho}_n$ has the advantage that it is unbiased if the queue is initially in the steady state. Moreover, $\bar{\rho}_n$ is robust because $\rho = 1 - P(Q = 0)$ for more general arrival processes, i.e., for $G/G/1/\infty$ queues [6]. Again, refinement for ratios could be used. As should be expected, because $\hat{\sigma}_n$ and $\hat{\rho}_n$ involve $\hat{q}_{2n}$ whereas $\bar{\sigma}_n$ and $\bar{\rho}_n$ do not, experiments indicate that $\bar{\sigma}_n$ and $\bar{\rho}_n$ are much better estimators than $\hat{\sigma}_n$ and $\hat{\rho}_n$. As an illustration, we compare these estimators in a simulation of a $GI/M/1/\infty$ queue in Table 1. The arrival process here is

TABLE 1

A Comparison of Estimators in the $H_2^b/M/1/\infty$ Queue

Traffic Intensity	Sample Mean and Std.		Estimators			Actual
ρ	Dev.	$\bar{\rho}$	$\hat{\rho}$	$\bar{\sigma}$	$\hat{\sigma}$	σ
0.3	M	0.302	0.303	0.382	0.382	0.376
	SD	0.002	0.008	0.004	0.008	
0.5	M	0.499	0.516	0.587	0.581	0.592
	SD	0.002	0.011	0.004	0.007	
0.7	M	0.700	0.663	0.780	0.791	0.776
	SD	0.002	0.029	0.004	0.012	

Notes: (1) The interarrival time has mean 1 and squared coefficient of variation $c^2 = 2$ in each case.

(2) M = Sample Mean.

(3) SD = Sample Standard Deviation.

(4) The estimate of SD is the classic ratio estimate; see [10] or Appendix 1.

(5) The number of customers served depended on ρ, being 60,000 for $\rho = 0.3$; 90,000 for $\rho = 0.5$; and 300,000 for $\rho = 0.7$. Before the customers were counted, a thousand initial customers were served and not counted to let the system reach steady state. The estimates were based on 20 batch means in each case.

hyperexponential (H_2^b) with squared coefficient of variation $c^2 = 2$; see (10)–(12) below. We considered three different traffic intensities: $\rho = 0.3$, 0.5, and 0.7. In Table 1 we have displayed the sample means and standard deviations. Note that the sample standard deviations of $\bar{\rho}$ and $\bar{\sigma}$ are significantly less than the sample standard deviations of $\hat{\rho}$ and $\hat{\sigma}$.

In [19, Appendix 1] we also compare these estimators in a simulation of an $H_2^b + H_2^b/M/1/\infty$ queue, where the arrival process is the superposition of two independent H_2^b renewal processes. Since the arrival process is not a renewal process, a significant bias is apparent in the estimator $\hat{\rho}$, but $\bar{\rho}$ behaves just as in Table 1 because $\bar{\rho}$ is unbiased for general $G/G/1/\infty$ queues. Hence, we recommend using $\bar{\sigma}$ and $\bar{\rho}$ instead of $\hat{\sigma}$ and $\hat{\rho}$ whenever possible.

2.3 *Selecting an Approximate Arrival Process*

Given the parameters ρ, σ and μ or their estimates, we now want to construct an approximating arrival process. Obviously the arrival rate should be $\lambda = \rho\mu$. Hence, we let the approximating arrival process be a renewal process with mean interarrival time λ^{-1}. We use the parameters ρ and σ to determine the specific renewal process. If $\sigma > \rho$, we let the interarrival time have a hyperexponential distribution; if $\sigma = \rho$, we let the interarrival time have an exponential distribution; and if $\sigma < \rho$, we let the interarrival time have either a generalized Erlang distribution or a shifted exponential distribution. These interarrival distributions are characterized by two parameters in each case so the fit is achieved by simply fitting the remaining parameter.

The $H_2^b/M/1/\infty$ System. Here we let the interarrival time have a hyperexponential distribution, i.e., a mixture of exponential distributions. We use the special case of two exponential distributions with balanced means (denoted by H_2^b); i.e., the distribution has the density

$$f(x) = q\lambda_1 e^{-\lambda_1 x} + (1-q)\lambda_2 e^{-\lambda_2 x}, \qquad x \geqslant 0, \tag{10}$$

where $0 \leqslant q \leqslant 1$ and

$$\lambda^{-1} = 2q/\lambda_1 = 2(1-q)/\lambda_2; \tag{11}$$

see Kuehn [14] and page 52 of Morse [15]. With these parameters, the interarrival time can have any coefficient of variation c (standard deviation divided by the mean) greater than one:

$$c = \left(\left[1 + (2q-1)^2\right]/\left[1 - (2q-1)^2\right]\right)^{1/2} \geqslant 1. \tag{12}$$

The assumption of balanced means in (11) reduces the number of parameters from three to two. The requirement that the mean be λ^{-1} leaves one free parameter. Given q, λ_1, λ_2 and μ, we can solve for ρ and σ. Of course, $\rho = \lambda/\mu$. To find σ, we must solve $\phi(\mu(1-\sigma)) = \sigma$ or

$$\sigma = \frac{q\lambda_1}{\mu(1-\sigma)+\lambda_1} + \frac{(1-q)\lambda_2}{\mu(1-\sigma)+\lambda_2}. \tag{13}$$

Using (11) we obtain

$$\sigma = \rho + \frac{1 \pm \sqrt{1 - 4\rho(1-\rho)\left[1 - 4q(1-q)\right]}}{2}. \tag{14}$$

Since $\rho(1-\rho) \leqslant 1/4$ and $q(1-q) \leqslant 1/4$, we see that $\sigma \geqslant \rho$. In order to have $\sigma = \rho$, it is necessary and sufficient to have $q = 1 - q = 1/2$, but then $\lambda_1 = \lambda_2$ by (11) and the hyperexponential distribution reduces to an exponential distribution.

From (10) and (11), we can also find q, λ_1 and λ_2 from λ, μ and σ. In particular,

$$q = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{(\sigma-\rho)(1-\sigma+\rho)}{\rho(1-\rho)}} \quad . \tag{15}$$

A feasible q is obtained from (15) if and only if $(\sigma-\rho)(1-\sigma+\rho) \leqslant \rho(1-\rho)$, which in turn holds if and only if $\rho \leqslant \sigma \leqslant 2\rho$. Thus, (15) always has a solution when $\rho \geqslant 1/2$, but might not when $\rho < 1/2$. This difficulty does not seem to arise in practice, but if σ turns out to be too large, then one can apply the indirect method with a different interarrival time distribution. One could also replace σ by the largest feasible value, 2ρ, but that may seriously underestimate the variability of the arrival process; see (3) and (5).

Having found the parameters q, λ_1 and λ_2 of the approximating H_2^b arrival process, we can calculate the two parameters ρ and σ of the $GI/M/1$ system with any new exponential service rate using (11) and (14).

The $E_2^g/M/1/\infty$ System. Here we let the interarrival time be the sum of two independent exponential random variables with means λ_1^{-1} and λ_2^{-1}, denoted by E_2^g. When $\lambda_1 = \lambda_2$, we obtain an ordinary Erlang (E_2) distribution which always has coefficient of variation $c = 1/\sqrt{2}$. With $\lambda_1 \neq \lambda_2$, we can obtain any coefficient of variation between $1/\sqrt{2}$ and 1:

$$c = \left([\lambda_1^2 + \lambda_2^2]/[\lambda_1 + \lambda_2]^2 \right)^{1/2}. \tag{16}$$

Given the parameters λ_1, λ_2 and μ, we can solve for the parameters ρ and σ. Of course, $\lambda^{-1} = \lambda_1^{-1} + \lambda_2^{-1}$ and $\rho = \lambda/\mu$. To find σ, we must solve $\phi(\mu(1-\sigma)) = \sigma$ or

$$\frac{\lambda_1\lambda_2}{(\mu(1-\sigma)+\lambda_1)(\mu(1-\sigma)+\lambda_2)} = \sigma, \tag{17}$$

from which we obtain

$$\sigma = (1/2\mu)\left(\lambda_1 + \lambda_2 + \mu \pm \sqrt{(\lambda_1 + \lambda_2 + \mu)^2 - 4\lambda_1\lambda_2}\right). \tag{18}$$

Given λ, μ and σ, we can also solve for the parameters λ_1 and λ_2:

$$\lambda_i = (x/2\lambda) \pm (1/2)\sqrt{(x/\lambda)^2 - 4x} \ , \tag{19}$$

where

$$x = \mu^2\sigma(1-\sigma)\rho/(\rho-\sigma). \tag{20}$$

Here of course we have $\sigma < \rho$. In order for (19) to have a solution, we must have the discriminant nonnegative, i.e.,

$$\sigma(1-\sigma) \geqslant 4\rho(\rho-\sigma). \tag{21}$$

If $\sigma < \rho$ but the inequality in (21) is not satisfied, then we can try the M^d arrival process below.

Having found the parameters λ_1 and λ_2, we can find the key parameters ρ and σ associated with any new service rate μ using (18) and the relations $\rho = \lambda/\mu$ and $\lambda^{-1} = \lambda_1^{-1} + \lambda_2^{-1}$.

The $M^d/M/1/\infty$ System. Here we let the interarrival time be the sum of a constant d and an exponential random variable with mean λ_1^{-1}. This distribution, which we call the shifted exponential distribution and denote by M^d, has density

$$f(x) = \lambda_1 e^{-\lambda_1(x-d)}, \qquad x \geqslant d, \tag{22}$$

with mean $\lambda^{-1} = d + \lambda_1^{-1}$ and variance $\sigma^2 = \lambda_1^{-2}$. Hence, $c = 1/(1 + \lambda_1 d)$. Obviously, c can assume any value between zero and one, with the value one attained when $d = 0$.

Given λ_1, d and μ, we can solve for the root σ. Solving the equation $\phi(\mu(1-\sigma)) = \sigma$, we find that σ satisfies

$$\frac{\lambda_1 e^{-\mu(1-\sigma)d}}{\lambda_1 + \mu(1-\sigma)} = \sigma, \tag{23}$$

which can easily be solved on a computer by a search routine. (The left side of (23) is bigger than the right for $\sigma = 0$ and equal for $\sigma = 1$, and there is one root in the interval $(0, 1)$.)

Conversely, after letting $\lambda = \rho\mu$ and $d = \lambda^{-1} - \lambda_1^{-1}$, we can also solve (23) for λ_1 given μ, σ and ρ. From (23), we see that λ_1^{-1}, is the solution to

$$e^{-\mu(1-\sigma)(\lambda^{-1}-\lambda_1^{-1})} = \sigma\big(1 + \mu(1-\sigma)\lambda_1^{-1}\big). \tag{23a}$$

Note that both sides of (23a) are increasing in λ_1^{-1}. Also the largest possible values on the left and right are 1 and $\sigma(1 + \rho - \sigma)\rho^{-1}$, respectively. Since $\sigma(1 + \rho - \sigma)\rho^{-1} \leqslant 1$ because $\rho > \sigma$, a solution to (23a) must exist when $e^{-(1-\sigma)\rho^{-1}} \leqslant \sigma$ because then the range of the left side of (23a) contains the range of the right side. Again, it is possible for there to be no solution. This problem can occur only when σ is relatively small. Then another interarrival time distribution, such as a two-point or three-point distribution, can be tried.

2.4 *An Example: The $\sum G_i/M/1/\infty$ System*

To illustrate our approximation procedure, we consider a $G/M/1/\infty$ system in which the arrival process is the superposition of independent renewal processes. For simplicity, we consider two identically distributed H_2^b renewal processes. Extensive experience with superposition processes (e.g., [1, 18]) indicates that the $H_2^b + H_2^b$ process is sufficiently different from a single H_2^b process for this to be an interesting case.

Here is the experiment: We simulated five $H_2^b + H_2^b/M/1/\infty$ systems with a common arrival process but different service rates. We let the mean and coefficient of variation of the renewal interval in each component H_2^b process be 2 and $\sqrt{2}$ respectively. Thus the arrival rate in the superposition process is one. From (12), we see that the mixing probabilities in each component process are q and $1 - q$ where $q = 0.78867$. We considered five different service rates: $\mu = 0.3$, 0.5, 0.7, 0.8 and 0.9. Since $\lambda = 1$, $\rho = \mu$ in each case. For each simulation we estimated q_1 using $\hat{q}_{1n}$ in (6). Then we estimated σ using $\bar{\sigma}_n$ in (9) and the known ρ. (As can be seen from [19, Appendix 1], the estimate $\bar{\rho}_n$ will usually be very close to ρ, so similar results can be expected if we use the estimate. However, this example illustrates that we need not always estimate ρ.) Given $\bar{\sigma}_n$ and ρ, we estimated the coefficient of variation c using (15) and (12). We need c instead of σ because c describes the arrival process alone whereas σ also depends on the service rate μ. Finally, given the known arrival rate $\lambda = \rho\mu$ and the estimated coefficient of variation c, we calculated the mean queue length for each of the other service rates. For each new service rate, we calculated q from (12) and the new root σ from (14). The mean number in the system is then $\rho/(1-\sigma)$.

The values of q_1, the expected number in the system, are given in Table 2. The underlined diagonal elements are the actual values as estimated from the simulation. The rows record the new traffic intensities and the columns record the traffic intensities of the test system used to generate the approximation. For comparison, the one-parameter $M/M/1/\infty$ approximation is included in Table 2. We compare the approximation procedure here with the procedures in [18] and [1] in [19, Appendix 3].

The number of customers served in the simulation depended on ρ; being 60,000 for $\rho = 0.3$; 90,000 for $\rho = 0.5$; 300,000 for $\rho = 0.7$; 600,000 for $\rho = 0.8$, and 1,000,000 for $\rho = 0.9$. Before the customers were counted in each case, a thousand initial customers were served and not counted to let the system reach steady state. In each case the estimates were obtained by dividing the total sample into 20 batches and treating the batch means as independent and identically distributed observations. The same procedure and sample sizes were used for the other tables.

Since in all but two cases (going from $\rho = 0.3$ and 0.5 to 0.9) the percent error is less than 10 percent, we conclude that the indirect approach works well. This is especially true when the new service rate is not too far away from the test service rate. Moreover, there is a systematic pattern to the errors, which suggests a possible refinement. In particular, if the new traffic intensity is higher (lower) than the test traffic intensity, then the approximate mean should be inflated (deflated). Table 2 could serve as the basis for quantifying such refinements, but we do not pursue this subject here. Further

TABLE 2

The Approximate Values for the Expected Number of Customers in an $H_2^b + H_2^b/M/1/\infty$ System:
$q_1 = \rho(1-\sigma)$

		One-Parameter	Traffic Intensity of the Test System				
		$M/M/1/\infty$	0.3	0.5	0.7	0.8	0.9
Traffic Intensity of The New System	0.3	0.429	$\underline{0.459}$ (0.003)	0.461	0.469	0.472	0.483
	0.5	1.00	1.13	$\underline{1.13}$ (0.011)	1.17	1.18	1.24
	0.7	2.33	2.77	2.79	$\underline{2.91}$ (0.037)	2.97	3.16
	0.8	4.00	4.87	4.90	5.13	$\underline{5.26}$ (0.056)	5.63
	0.9	9.00	11.2	11.3	12.0	12.2	$\underline{13.3}$ (0.32)

Note: The simulated values are underlined. In parentheses below the simulated values appear the standard errors (one sample standard deviation).

refinements of the approximation procedure can be carried out as needed. Our object is to demonstrate the value of the general approach. Since we have also obtained similar results in other cases, the conclusions apply quite broadly. For further discussion, see the next section and [19].

3. Approximating the Arrival Process in a $G/G/1/\infty$ System

The indirect method in Section 2 can be extended to cover general service time distributions by using approximation formulas based on the first two moments of the interarrival time and the service time. Heavy traffic approximations can be used for this purpose, but we found that approximation formulas developed by Krämer and Langenbach-Belz [12] work better. They obtained approximations for the mean waiting time and the probability that a customer will be delayed in a $GI/G/1/\infty$ system. Using Little's formula $L = \lambda W$, we can transform their approximation to obtain an approximation for the mean number of customers in the system:

$$q_1 = \rho + \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right)h(c_a, c_s, \rho), \tag{24}$$

where c_a and c_s are the coefficients of variation of the interarrival time and service time and

$$h(c_a, c_s, \rho) = \begin{cases} \exp\left\{-(2(1-\rho)/3\rho)\left([1-c_a^2]^2\big/[c_a^2+c_s^2]\right)\right\}, & c_a \leqslant 1, \\ \exp\left\{-(1-\rho)\left([c_a^2-1]\big/[c_a^2+4c_s^2]\right)\right\}, & c_a \geqslant 1; \end{cases} \tag{25}$$

see §1.3 of [12].

Given c_s^2, ρ and q_1 or their estimates ($\bar{\rho}$ in (9) and $\hat{q}_1$ in (6)), we can solve (24) iteratively on the computer for c_a^2. A search routine is easy to implement because the approximation formula for q_1 in (24) is a strictly increasing function of c_a^2. Given c_a^2, we can fit the appropriate distribution (H_2^b, M, E_2^g, or M^d). As before, if $c > 1$, we use H_2^b; if $c = 1$, we use M; if $1/\sqrt{2} \leqslant c < 1$, we can use E_2^g; and if $0 \leqslant c < 1$, we can use M^d.

To illustrate this approximation procedure, we simulated several $H_2^b + H_2^b/G/1/\infty$ systems in which the service time distribution is either H_2^b, E_2, or M. As in §2, we let the renewal interval in each H_2^b arrival process have mean 2 and coefficient of variation $\sqrt{2}$. Thus the total arrival rate is one. We considered two different mean service times $\mu^{-1} = \rho = 0.5$ and 0.8. The coefficients of variation of H_2^b and E_2 service times were $\sqrt{2}$ and $1/\sqrt{2}$ respectively. For each simulation we estimated q_1 using $\hat{q}_{1n}$ in (6). With this value of q_1 plus the known values of ρ and c_s, we determined c_a using the KL (Krämer and Langenbach-Belz) formula (24). We then used this approximate value of c_a to approximate q_1 for all the other cases. Here we are using the indirect method to describe the congestion after changing the service-time distribution as well as after changing the service rate. As can be seen from Table 3, the indirect method again seems to work well. As in Table 2, the simulated values appear on the diagonal and are underlined. In parentheses below the simulated value is the standard error (one sample standard deviation) of the simulation estimate. Notice that the approximation differences are of the same order as the uncertainties in the estimates. The values of the approximating squared coefficient of variation c_a^2 appear in the bottom row.

TABLE 3

The Approximate Values for the Expected Number of Customers in the $H_2^b + H_2^b/G/1/\infty$ Systems, Using the KL Formula (24)

				The Test System			
The	Traffic	$H_2^b + H_2^b/H_2^b/1$		$H_2^b + H_2^b/M/1$		$H_2^b + H_2^b/E_2/1$	
New	Intensity	$c_s^2 = 2.0$		$c_s^2 = 1.0$		$c_s^2 = 0.5$	
System	ρ	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
$H_2^b + H_2^b/H_1^b/1$	0.5	$\underline{1.46}$ (.028)	1.46	1.42	1.43	1.40	1.43
	0.8	7.16	$\underline{7.16}$ (.13)	6.84	6.93	6.72	6.93
$H_2^b + H_2^b/M/1$	0.5	1.20	1.20	$\underline{1.16}$ (.017)	1.17	1.15	1.17
	0.8	5.51	5.51	5.22	$\underline{5.30}$ (.08)	5.10	5.30
$H_2^b + H_2^b/E_2/1$	0.5	1.06	1.06	1.02	1.03	$\underline{1.01}$ (.007)	1.03
	0.8	4.67	4.67	4.38	4.47	4.27	$\underline{4.47}$ (.04)
Approximate Value of c_a^2		2.05	2.05	1.84	1.90	1.76	1.90

Notes: (1) The simulated values appear on the diagonal and are underlined. The standard error (one sample standard deviation) appears in parentheses below the estimate.

(2) The approximate values of c_a^2 obtained by inverting formula (24) appear in the bottom row. This value of c_a^2 is used to generate the off-diagonal elements of the column.

The Indirect Method for both $G/M/1/\infty$ queues (§2) and the KL approximation (24) are based on two-moment approximations. To put these approximations in perspective, we indicate the range of possible q_1 values (mean number in system) over all possible $GI/M/1$ queues with given traffic intensity ρ and squared coefficient of variation of the interarrival time c^2. We also indicate where the $H_2^b/M/1$ and KL values fall within this range.

It is possible to give the range because the extremal distributions are known; see page 138 of Eckberg [5]. The largest possible q_1 value is attained with the two-point distribution having mass $1/(1+c^2)$ on $\lambda^{-1}(1+c^2)$ and mass $c^2/(1+c^2)$ on 0, where λ^{-1} is the mean of the renewal interval. The smallest possible q_1 value is not attained, but the limit is the one-point distribution having mass 1 on λ^{-1}. This distribution is the limit as $k \rightarrow \infty$ of two-point distributions having mass $k^2/(1+k^2)$ on $\lambda^{-1}(1-c/k)$ and mass $1/(1+k^2)$ on $\lambda^{-1}(1+kc)$.

Table 4 gives the range of q_1 values along with the Krämer and Langenbach-Belz approximation. It is apparent that the range is very wide, with the lower bound being especially far away from the $H_2^b/M/1$ approximation. We expect the two-moment approximation often to work well, not because two moments determine q_1 completely, but because we expect the more extreme distributions will not often arise. Further studies are under way to quantify how additional properties such as unimodality restrict the range.

TABLE 3

The Approximate Values for the Expected Number of Customers in the $H_2^b + H_2^b/G/1/\infty$ System Using the KL Formula (24)

The New System	Traffic Intensity ρ	The Test System $H_2^b + H_2^b/H_2^b/1$ $c_s^2 = 2.0$ $\rho = 0.5$	$\rho = 0.8$	$H_2^b + H_2^b/M/1$ $c_s^2 = 1.0$ $\rho = 0.5$	$\rho = 0.8$	$H_2^b + H_2^b/E_2/1$ $c_s^2 = 0.5$ $\rho = 0.5$	$\rho = 0.8$
$H_2^b + H_2^b/H_2^b/1$	0.5	1.46 (.028)	[illegible]	1.42	1.43	1.40	[illegible]
	0.8	7.16	[illegible]	6.84	6.93	6.72	[illegible]
$H_2^b + H_2^b/M/1$	0.5	1.20	1.20	1.16 (.017)	1.17	1.15	[illegible]
	0.8	5.51	5.51	5.22	5.30 (.08)	5.10	[illegible]
$H_2^b + H_2^b/E_2/1$	0.5	1.06	1.06	1.02	1.03	1.01 (.007)	[illegible]
	0.8	4.67	4.67	4.38	4.47	4.27	[illegible] [illegible]
Approximate Value of c_a^2		2.05	2.05	1.84	1.90	1.76	[illegible]

Notes: (1) The simulated values appear on the diagonal and are underlined. The standard error (one sample standard deviation) appears in parentheses below the estimate.

(2) The approximate values of c_a^2 obtained by inverting formula (24) appear in the bottom row. This value of c_a^2 is used to generate the off-diagonal elements of the column.

The Indirect Method for both $G/M/1/\infty$ queues (§2) and the KL approximation (24) are based on two-moment approximations. To put these approximations in perspective, we indicate the range of possible q_1 values (mean number in system) over all possible $GI/M/1$ queues with given traffic intensity ρ and squared coefficient of variation of the interarrival time c^2. We also indicate where the $H_2^b/M/1$ and KL values fall within this range.

It is possible to give the range because the extremal distributions are known (see page 138 of Eckberg [5]). The largest possible q_1 value is attained with the two-point distribution having mass $1/(1+c^2)$ on $\lambda^{-1}(1+c^2)$ and mass $c^2/(1+c^2)$ on 0, where λ^{-1} is the mean of the renewal interval. The smallest possible q_1 value is not attained, but the limit is the one-point distribution having mass 1 on λ^{-1}. This distribution is the limit as $k \to \infty$ of two-point distributions having mass $k^2/(1+k^2)$ on $\lambda^{-1}(1-c/k)$ and mass $1/(1+k^2)$ on $\lambda^{-1}(1+kc)$.

Table 4 gives the range of q_1 values along with the Krämer and Langenbach-Belz approximation. It is apparent that the range is very wide, with the lower bound being especially far away from the $H_2^b/M/1$ approximation. We expect the two-moment approximation often to work well, not because two moments determine q_1 completely, but because we expect the most extreme distributions will not often arise. Further studies are under way to quantify how additional properties such as unimodality restrict the range.

TABLE 4

The Possible q_1 Values (Mean Number in System) for $GI/M/1$ Queues with Given λ, μ, and $c^2 = 2$

Traffic Intensity ρ	Lower Bound	Krämer and Langenbach-Belz Approximation	$H_2^b/M/1/\infty$	Upper Bound
0.3	0.31	0.47	0.48	0.94
0.5	0.63	1.19	1.22	1.88
0.7	1.32	3.03	3.12	3.93
0.8	2.15	5.44	5.58	6.45
0.9	4.62	12.85	13.04	13.86

In order to point out circumstances in which the indirect method ceases to be good, we now describe experiments with different arrival processes. In particular, we considered the three cases in §6 of [18], namely $M^d + M^d$, $M^d + H_2^b$, and $H_2^b + H_2^b$ where $c^2 = 0.6$ in each M^d process and $c^2 = 9.56$ in each H_2^b process. The results for the indirect method in these cases are displayed in Tables 5–7. It is significant that the variability of the H_2^b process here is much greater than in Table 2. In the two cases in which such a highly variable H_2^b process is present the indirect method does not work as well as before. This is consistent with our experience in [18]; see Appendix 10 there. (In Tables 5–7 more customers were served in the simulation than was the case for Tables 1–4. In particular, there were 300,000 for $\rho = 0.3$; 750,000 for $\rho = 0.6$; 4,500,000 for $\rho = 0.8$; and 7,500,000 for $\rho = 0.9$.)

Further investigations of the indirect method for $G/G/1/\infty$ queues appear in [19, Appendix 5]. Table 11 there is the analog of Table 2 in §2. For the $H_2^b + H_2^b/M/1$ queue, we see that the indirect method using KL approximation (24) is about the same as the $GI/M/1$ fit in §2.

TABLE 5

Approximate Values for the Mean Number of Customers in the System (q_1) in an $M^d + M^d/M/1$ Queue, with $c^2 = 0.6$ in each M^d Process, Using the KL *Approximation* (24)

		Traffic Intensity of the Test System			
		0.3	0.6	0.8	0.9
Traffic Intensity of the New System	0.3	0.37 (0.0007)	0.37	0.38	0.38
	0.6	1.21	1.22 (0.007)	1.25	1.25
	0.8	3.08	3.11	3.19 (0.012)	3.21
	0.9	6.77	6.85	7.05	7.07 (0.06)
Approximating Squared Coefficient of Variation c^2		0.469	0.488	0.534	0.544

TABLE 6

Approximate Values for the Mean Number of Customers in the System (q_1) in an $H_2^b + H_2^b/M/1$ Queue, with $c^2 = 9.56$ in each H_2^b Process, Using the KL *Approximation* (24)

		Traffic Intensity of the Test System			
		0.3	0.6	0.8	0.9
Traffic Intensity of the New System	0.3	0.54 (0.004)	0.66	0.73	0.73
	0.6	2.47	3.60 (0.035)	4.27	4.26
	0.8	8.0	12.7	15.6 (0.18)	15.5
	0.9	19.7	32.8	40.8	40.6 (0.7)
Approximating Squared Coefficient of Variation c^2		3.80	7.34	9.49	9.44

TABLE 7

Approximate Values for the Mean Number of Customers in the System (q_1) in an $M^d + H_2^b/M/1$ Queue ($c_1^2 = 9.56$ and $c_2^2 = 0.6$) Using the KL *Approximation* (24)

		Traffic Intensity of the Test System			
		0.3	0.6	0.8	0.9
Traffic Intensity of the New System	0.3	0.45 (0.002)	0.51	0.56	0.57
	0.6	1.68	2.19 (0.025)	2.64	2.75
	0.8	4.70	6.78	8.65 (0.075)	9.12
	0.9	10.9	16.5	21.6	22.9 (0.4)
Approximating Squared Coefficient of Variation c^2		1.48	2.95	4.31	4.66

4. Approximating the Arrival Process in a $G/M/\infty$ System

Suppose we have a $G/M/\infty$ service system: infinitely many servers working in parallel with mutually independent service times having a common exponential distribution with mean μ^{-1} and an arrival process that is independent of the service times but otherwise general. Again we assume the arrival process is stationary, so the system can be viewed as being in steady state; see [6] for supporting theory. As before, we shall use observations about the level of congestion to approximate the arrival process. Here let Q be the steady-state number of busy servers and let q_j be the jth moment of Q. We shall use q_1 and q_2 to specify the arrival rate λ and the coefficient of variation c of the renewal interval.

4.1 *Background on the $GI/M/\infty$ System*

As in §2.1, we begin by considering the special case in which the arrival process is a renewal process; see page 163 of Takács [16]. Again the stochastic behavior is characterized by the interarrival time c.d.f. F and the mean service time μ^{-1}. Let λ^{-1} be the mean of F and let ϕ be its Laplace-Stieltjes transform, defined in (1). In this case, given μ, the steady-state distribution depends on more than two parameters of the arrival process; in fact it depends on countably many parameters: $\{\phi(k\mu), k \geqslant 1\}$; see [16]. However, it is significant that, given μ, the parameters q_1 and q_2 depend on only two parameters. From page 164 of [16], it follows that

$$q_1 = a = \lambda/\mu \quad \text{and} \quad z = \frac{q_2}{q_1} - 1 = \frac{1}{1-\phi(\mu)} - a. \tag{26}$$

Viewing the infinite-server system as a trivial case of a loss system and planning for the applications to loss systems in §5, we call the parameter a here (ρ before) the offered load and the parameter z the peakedness. We use z because it is commonly used in teletraffic applications; see [3] and references there. We can obviously estimate q_1 and q_2 using (6) and (7). Then (26) provides the basis for consistent estimators of a and z.

4.2 *Selecting the Approximate Arrival Process*

Based on the discussion above, given μ, we obviously let $\lambda = \mu a$. Below we will see how to determine c for the special renewal interval distributions given μ, a and z. The particular arrival process to use depends on the peakedness z and offered load a. If $z \geqslant 1$, we use H_2^b; if $(3a+1)/(4a+1) \leqslant z \leqslant 1$, we use E_2^g; and if $z \leqslant (3a+1)/(4a+1)$ or also if $z \leqslant 1$, we use M^d. To give an indication of the ranges, we note that the minimum peakedness given the offered load a is the peakedness associated with a $D/M/\infty$ system; them $z_{\min} = (1-e^{-(1/a)})^{-1} - a$. For example, when $a = 1$, $z_{\min} = 0.582$; when $a = 10$, $z_{\min} = 0.508$; and as $a \to \infty$, $z_{\min} \to 1/2$.

The $H_2^b/M/\infty$ System. Given the service rate μ and an H_2^b distribution as specified in (10) and (11), we can calculate the peakedness using (26). We obtain

$$z = 1 \Bigg/ \left[1 - \frac{2aq^2}{2aq+1} - \frac{2a(1-q)^2}{2a(1-q)+1}\right] - a. \tag{27}$$

Conversely, we can use these relations to calculate the H_2^b parameters given a, μ, and z (if $z \geqslant 1$). First, $\lambda = a\mu$. By (27), we obtain

$$q = \frac{1}{2}\left(1 \pm \sqrt{\frac{a+1}{a}}\sqrt{\frac{z-1}{z}}\right). \tag{28}$$

A different scheme for generating a hyperexponential renewal interval approximation was developed by Kuczura [13]. This scheme, which is not based on balanced means and thus involves three parameters, is described in [19, Appendix 7]. Experience in teletraffic applications indicates that Kuczura's procedure usually works quite well, but it is more complicated.

The $E_2^g/M/\infty$ System. Given the E_2^g distribution with parameters λ_1 and λ_2 as specified in §1.3, we can calculate the peakedness using (17) and (26). To go the other way, note that $\lambda^{-1} = \lambda_1^{-1} + \lambda_2^{-1}$ or $\lambda_1 + \lambda_2 = \lambda_1\lambda_2/\lambda$. Then solve (17) for $\lambda_1\lambda_2$ or

$\lambda_1 + \lambda_2$. We obtain

$$\lambda_i = \left(B \pm \sqrt{B^2 - 4\lambda^2 B}\right) \Big/ 2\lambda, \tag{29}$$

where

$$B = \lambda\mu^2 D/(\lambda - (\lambda + \mu)D) \quad \text{and} \quad D = 1 - (z + a)^{-1}. \tag{30}$$

In order for (29) to have a solution, the discriminant must be nonnegative; i.e., we must have $B \geqslant 4\lambda^2$ or $z \geqslant (3a + 1)/(4a + 1)$. In order for the solution to be feasible, we must have $B \geqslant 0$, which reduces to $z \leqslant 1$. In other words, the E_2^g distribution is appropriate for z such that $(3a + 1)/(4a + 1) \leqslant z \leqslant 1$. Use M^d below if z is not in the right range.

The $M^d/M/\infty$ System. Given the shifted exponential distribution with the density in (22), we can use (26) to solve for the peakedness:

$$\frac{\lambda_1 e^{-\mu d}}{\lambda_1 + \mu} = 1 - (z + a)^{-1}. \tag{31}$$

Alternatively, given μ, a and z, we can let $\lambda = a\mu$ and $d = \lambda^{-1} - \lambda_1^{-1}$, and then solve (31) for λ_1. As $d \to 0$, $z \to 1$; as $d \to \lambda^{-1}$, $z \to (1 - e^{-(1/a)})^{-1} - a$, the minimum peakedness associated with $D/M/\infty$ systems.

4.3 *An Example: The $\sum G_i/M/\infty$ System*

To illustrate the indirect method for $G/M/\infty$ systems, we applied it to the $H_2^b + H_2^b/M/\infty$ system in which the two component arrival processes are identically distributed with $c^2 = 2.0$ or, equivalently, with $q = 0.788675$. We first considered the $H_2^b/M/\infty$ system with offered loads of $a = 1$, 2, 5, 10, and 20. We calculated the peakedness in each case using (27). These peakedness values also apply to the $H_2^b + H_2^b/M/\infty$ system, so we obtained the actual offered loads and peakedness values for the $H_2^b + H_2^b/M/\infty$ system. These in turn were used with (28) to obtain an approximating H_2^b arrival process. The results are displayed in Table 8. Based on the

TABLE 8

Approximate Values of the Peakedness (z) for $H_2^b + H_2^b/M/\infty$ Systems Using the Indirect Method

		Offered Load in the Test System				
		$a = 2$	$a = 4$	$a = 10$	$a = 20$	$a = 40$
Offered Load in the New System	$a = 2$	$\underline{1.20000}$	1.227	1.256	1.269	1.277
	$a = 4$	1.250	$\underline{1.28563}$	1.324	1.341	1.352
	$a = 10$	1.294	1.338	$\underline{1.38462}$	1.411	1.420
	$a = 20$	1.312	1.360	1.410	$\underline{1.43478}$	1.449
	$a = 40$	1.323	1.372	1.426	1.449	$\underline{1.46511}$
Approximate Value of q		0.75000	0.76349	0.77639	0.78204	0.78522
Approximate Value of c^2		1.67	1.77	1.88	1.93	1.96

Notes: (1) Each component H_2^b process has $c^2 = 2.0$ or, equivalently, $q = 0.788675$.
(2) The true values obtained from (27) are given on the diagonal and are underlined.

criterion of having less than 10 percent error, the indirect method again performs well, but the approximation does not discriminate as well from the test system value as was the case in Tables 2 and 3. (Here the peakedness does not change much in the range studied.) Of course, the approximate peakedness obtained via this indirect method is closer to the actual peakedness than the peakedness of the test system. Also, in judging these approximations, note that they are for the peakedness or second moment; no approximation is needed for the first moment, which is just the offered load a.

5. Approximating the Arrival Process in a $G/M/s/0$ System

We can also apply the method in §4 to approximate the arrival process in a $G/M/s/0$ system, i.e., a system with s homogeneous exponential servers and no waiting room. In this setting it is common to observe the number of arrivals and the number of blocked arrivals in a given time interval. The number of arrivals yields an estimate for the arrival rate λ. This together with the known service rate μ or its estimate yields an estimate for the offered load $a = \lambda/\mu$. To proceed further, we use the equivalent random method or one of its variants; see §4.7 of Cooper [3] and Wilkinson [20], [21]. The estimated blocking probability together with the offered load a and the number of servers s can be used to obtain an estimate of the peakedness z. We then apply §4 to characterize the approximating renewal process.

The basic procedure for obtaining the peakedness z is to solve the following three equations for a', s', and v (see page 137 of [3]):

$$\begin{aligned} B &= a'E(s'+s,a'), \\ a &= a'E(s',a'), \\ v &= a\left(1 - a + \frac{a'}{s+1+a-a'}\right), \end{aligned} \tag{32}$$

and then set $z = v/a$, where B is the estimated blocking probability; a' is the offered load to the artificial group of s' servers such that the resulting overflow behaves like our arrival process; a is the offered load and s is the number of servers in our actual system; and $E(s,a)$ is the Erlang loss formula. However, given z, a, and s, it is convenient to use the Rapp approximation to determine a' and s'; see p. 138 of [3] or (A7.3) and (A7.4) in [19, Appendix 7]. This provides an iterative procedure to obtain z: Choose a candidate z value and solve for a', s', and the associated blocking probability, then revise the choice of z accordingly, making it bigger if the candidate blocking probability is too small.

There are several other possibilities too. First, it is possible to use the graphs and tables in [21]. Second, it is possible to use Hayward's approximation; i.e., we can approximate the blocking probability $B(s,a,z)$ as a function of s, a, and z by $E(s/z,a/z)$. We then use an iterative procedure based on the Erlang loss formula to search for z.

We conclude this section with an example. Consider the $H_2^b/M/s/0$ system with $\lambda = 4$, $\mu = 1$, $s = 15$, and H_2^b parameter $q = 0.984123$ which corresponds to the peakedness $z = 4$; see (27) and (28). From page 26 of [21], we see that the blocking probability in this system is about 0.046. Now suppose that we change the service mechanism without changing the arrival process. In particular, suppose we decrease the individual service rate to 0.1 and thus increase the offered load to 40, and we increase the number of servers to 50. What will be the blocking probability in this new system?

Of course, no approximation is needed here because we know the arrival process. However, suppose we did not; a simple approximation procedure that might be used here is to let the peakedness be just what it was before. Using the graphs on page 26 of [21] for the case $s = 50$, $a = 40$, and $z = 4$, we obtain the approximate blocking probability $B = 0.10$. However, when we calculate the peakedness associated with the actual arrival process (which would be the approximate arrival process in this case), we obtain $z = 11.71$, which yields the accurate approximate blocking probability 0.25; see pp. 28–29 of [21]. In this special case our indirect approximation procedure necessarily produces the correct peakedness and thus an accurate approximate blocking probability. This example demonstrates the value of the indirect method.

6. Extensions and Directions for Future Research

The indirect method in §§4 and 5 can be extended to general service-time distributions. First, by Little's formula $L = \lambda W$, the mean number of busy servers in the steady state in a $G/G/\infty$ system is $a = \lambda/\mu$. Eckberg [4] has shown that in a general $G/G/\infty$ the peakedness z as a function of the service time distribution G is

$$z(G) = 1 + a^{-1}\int_{-\infty}^{\infty}\left[k(x) - \lambda\delta(x)\right]G^{(2)}(x)\,dx, \tag{33}$$

where

$$G^{(2)}(y) = \int_{\max\{0,\,y\}}^{\infty} G(x)G(x-y)\,dx, \tag{34}$$

δ is the Dirac delta function and $k(x)$ is the covariance density of the arrival process. (See [18, §7] for further discussion and references.)

Given the first two moments of the service time distribution G, we can fit one of our special distributions. Moreover, we can assume the covariance density $k(x)$ is associated with a renewal process in which the interarrival-time distribution is also one of our special distributions. Then we can invert (33) to calculate the remaining parameter characterizing the interarrival time distribution.

For greater simplicity and no doubt less accuracy, heavy traffic approximations for $G/G/\infty$ systems can be used. For large loads, the number of busy servers is approximately normally distributed with mean a and variance az, where

$$z = 1 + \left(c_a^2 - 1\right)\mu\int_0^{\infty}\left[1 - G(x)\right]^2 dx, \tag{35}$$

where G is again the service time distribution. Given the service time distribution G and the peakedness z, it is easy to solve (35) for the coefficient of variation of the interarrival time c_a^2.

When there is waiting room and more than one server, we have difficulties carrying out our procedure. For $G/G/s/\infty$ systems, we can use diffusion approximations. In particular, we suggest formula (15) in Halachmi and Franta [7]. This formula is supported by a limit theorem in the case of exponential service times; see Halfin and Whitt [8].

Finding a way to apply the indirect method to $G/G/s/k$ systems seems to be a promising direction for future research.[1]

[1] The simulation and associated data collection in support of the numerical comparisons were done by my colleague Susan Albin. I am grateful to her for comments on the manuscript and many helpful discussions. I also thank the referees for their assistance.

References

1. Albin, S. L., "Approximating a Point Process by a Renewal Process: The $\sum G_i/G/1$ Queue" (to appear).
2. Cohen, J. W., *The Single Server Queue*, North-Holland, Amsterdam, 1969.
3. Cooper, R. B., *Introduction to Queueing Theory*, MacMillan, New York, 1972.
4. Eckberg, A. E., private communication.
5. ———, "Sharp Bounds on Laplace-Stieltjes Transforms, With Applications to Various Queueing Problems," *Math. Operations Res.*, Vol. 2 (1977), pp. 135–142.
6. Franken, P., König, D., Arndt, U. and Schmidt, V., *Point Processes and Queues*, Akademie-Verlag, Berlin (to appear).
7. Halachmi, B. and Franta, W. R., "A Diffusion Approximation to the Multi-Server Queue," *Management Sci.*, Vol. 24 (1978), pp. 522–529.
8. Halfin, S. and Whitt, W., "Heavy-Traffic Limits for Queues with Many Exponential Servers," *Operations Res.* (to appear).
9. Iglehart, D. L., "Functional Limit Theorems for the Queue $GI/G/1$ in Light Traffic," *Advances Appl. Probability*, Vol. 3 (1971), pp. 269–281.
10. ———, "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," *Naval Res. Logist. Quart.*, Vol. 22 (1975), pp. 554–565.
11. Kleinrock, L., *Queueing Systems*, Vol. 1, Wiley, New York, 1975.
12. Krämer, W. and Langenbach-Belz, M., "Approximate Formulae for the Delay in the Queueing System $GI/G/1$," *Eighth International Teletraffic Congress*, Melbourne, 235-1-8 (1976).
13. Kuczura, A., "The Interrupted Poisson Process as an Overflow Process," *Bell System Tech. J.*, Vol. 52 (1973), pp. 437–448.
14. Kuehn, P. J., "Approximate Analysis of General Queueing Networks by Decomposition," *IEEE Trans. Comm.*, Vol. COM-27 (1978), pp. 113–126.
15. Morse, P. M., *Queues, Inventories and Maintenance*, Wiley, New York, 1958.
16. Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.
17. Wallström, B., "Congestion Studies in Telephone Systems with Overflow Facilities," *Ericsson Technics*, Vol. 3 (1966), pp. 189–351.
18. Whitt, W., "Approximating a Point Process by a Renewal Process: Two Basic Methods," *Operations Res.* (to appear).
19. ———, "Appendices to 'Approximating a Point Process by a Renewal Process: The View Through a Queue, An Indirect Approach'," unpublished paper.
20. Wilkinson, R. I., "Theories of Toll Traffic Engineering in the U.S.A.," *Bell System Tech. J.*, Vol. 35 (1956), pp. 421–514.
21. ———, *Nonrandom Traffic Curves and Tables*, Traffic Studies Center, Bell Laboratories, 1970.