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. From the moments of the renewal interval one can choose to produce the same level of congestion when the renewal process serves as an arrival process in a lost queueing system or as produced when the general point process is the arrival process. The procedure can be applied to verify 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 equivalence random method and related techniques widely used in traffic engineering.

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 rates. 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 models 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., traffic intersection, the arrival process may change in response to a change in the service mechanism. However, even when this phenomenon can occur, approximate methods are of anything even more appropriate.

The models 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.

Tel. Laboratories, Holmdel, New Jersey.

619
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 process 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: \( E_r \)) or the exponential distribution shifted by a constant (shifted exponential: \( A_r \)). 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 queuing system, but the procedure can be applied to any point process. Then the queuing 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 \( G/M/1 \) model. This indirect method is very convenient when there is a handy simulation program for a queuing 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 queuing 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 Abjöf [1]. Such approximation procedure uses certain properties of the point processes to be approximated in order to specify the moments of the renewal interval. The stationary-interval and asymptotic methods in [19] focus on the point process in isolation. The hybrid methods in [11] for approximating the \( \sum G_i/G_1 \) queue began to use the model context because the traffic intensity of the queuing 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 queuing model.

It should be noted that there is considerable related literature: an effort to describe it was made in [19]. The simplest approach here is very closely related to the equivalent random method and its variations; see §4.7 of Cooper [3], Kuczma [13], Waldenström [17] and Willmot [20], [21]. The equivalent random method is a procedure for approximating 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 an 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 $G/M/1/\infty$ queuing 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 $\mu^{-1}$, 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, 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 11.3 of Cohen [2]. The stochastic behavior of this system is characterized by
the interarrival time c.d.f. \( F \) and the mean service time \( \mu^{-1} \). Let \( \lambda^{-1} \) be the mean of \( F \) and let \( \phi \) be its Laplace-Stieljes transform, i.e.,

\[
\phi(s) = \int_0^\infty e^{-sx} dF(x), \quad s > 0.
\]

(1)

It is significant that, given \( \phi \), 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 \( \sigma \), with \( 0 < \sigma < 1 \), of the equation \( s = \phi'(1 - \sigma) \).

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

\[
P(Q = k) = \begin{cases} 1 - \rho & k = 0, \\ \rho(1 - \sigma)^k & k > 1, \end{cases}
\]

(2)

and

\[
\phi = \rho(1 - \sigma)^{-1} \quad \text{and} \quad \phi_1 = \rho(1 + \sigma)(1 + \sigma)^{-2}.
\]

(3)

Note that the two moments \( \phi \) and \( \phi_1 \) in (3) also characterize the asymptotic behavior. In particular, we can express \( \rho \) and \( \sigma \) in terms of \( \phi \) and \( \phi_1 \):

\[
\rho = \frac{\phi_1 - \phi}{\phi_1} \quad \text{and} \quad \sigma = \frac{\phi - \phi_1}{\phi_1 + \phi_1}. \quad (4)
\]

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

\[
\rho = 1 - P(Q = 0) \quad \text{and} \quad \sigma = 1 - \rho / \phi_1. \quad (5)
\]

2.2 Estimating \( \rho \) and \( \sigma \) in the \( G[1]/M/1/\infty \) System

Now suppose that we actually have a \( G[1]/M/1/\infty \) system and we want to estimate the basic parameters \( \rho \) and \( \sigma \) based solely on observations of the stochastic process \( \{Q(t), t \geq 0\} \). Suppose that we have a sequence of random observation times \( \{T, n \geq 1\} \) independent of the queueing system such that \( T_n \to \infty \) as \( n \to \infty \). Then we obtain the following sequences of estimators for \( \phi \) and \( \phi_1 \):

\[
\hat{\phi}_n = n^{-1} \sum_{i=1}^n Q(T_i), \quad n > 1, \quad (6)
\]

and

\[
\hat{\phi}_1 = n^{-1} \sum_{i=1}^n Q(T_i)^2, \quad n > 1. \quad (7)
\]

Since the arrival epochs initiating busy periods are regeneration points, the variables \( Q(T_k), k > 1 \), are close enough to being independent so that sequences \( \{\hat{\phi}_n\} \) and \( \{\hat{\phi}_1, n \geq 1\} \) are strongly consistent, i.e., \( \hat{\phi}_n \to \phi \) and \( \hat{\phi}_1 \to \phi_1 \) with probability one as \( n \to \infty \); see Iglehart [9]. Moreover, if the system starts in the steady state, i.e., if the process
\[ \mathbb{Q}(t)_{t \geq 0} \text{ is stationary.} \]

The estimators are unbiased: \( \hat{\delta}_{\text{mean}} = \delta \) and \( \hat{\theta}_{\text{mean}} = \theta \) for all \( n \). It is of course important for those properties that the observation times be independent of the queuing system. For example, if the observation times were all arrival epochs, \( \hat{\delta}_{\text{mean}} \) would have different variances; see (3.25) on page 210 of [2].

From (6), (8) and (7), we obtain the following natural estimators for \( \sigma \) and \( \rho \):

\[ \delta = \frac{\hat{\delta}_{\text{mean}} - \hat{\theta}_{\text{mean}}}{\hat{\theta}_{\text{mean}} + \hat{\theta}_{\text{mean}}} \quad \text{and} \quad \rho = \frac{2 \hat{\theta}_{\text{mean}}}{\hat{\delta}_{\text{mean}} + \hat{\theta}_{\text{mean}}} \]  

(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 \( \rho \) and \( \sigma \) is

\[ \hat{\rho} = 1 - \hat{\rho}_{\text{mean}} = 1 - \sum_{i=0}^{n} l_i(\mathbb{Q}(t_i)) \quad \text{and} \quad \hat{\delta} = 1 - \hat{\delta}_{\text{mean}} = 1 - \sum_{i=0}^{n} l_i(\mathbb{Q}(t_i)) \]  

(9)

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

### TABLE 1

<table>
<thead>
<tr>
<th>Traffic Intensity ( \rho )</th>
<th>Sample Mean and SD</th>
<th>( \hat{\rho} )</th>
<th>( \hat{\delta} )</th>
<th>Actual ( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>M 0.302</td>
<td>0.303</td>
<td>0.382</td>
<td>0.382</td>
</tr>
<tr>
<td></td>
<td>SD 0.002</td>
<td>0.008</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>0.5</td>
<td>M 0.499</td>
<td>0.510</td>
<td>0.581</td>
<td>0.581</td>
</tr>
<tr>
<td></td>
<td>SD 0.003</td>
<td>0.011</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>0.7</td>
<td>M 0.706</td>
<td>0.663</td>
<td>0.760</td>
<td>0.760</td>
</tr>
<tr>
<td></td>
<td>SD 0.002</td>
<td>0.029</td>
<td>0.004</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Notes:
1. The interarrival time has mean 1 and squared coefficient of variation \( \sigma^2 = 2 \) 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 \( \rho \): being 40,000 for \( \rho = 0.3 \); 90,000 for \( \rho = 0.5 \); and 300,000 for \( \rho = 0.7 \). Before the customers were counted, a thousand visual customers were served and not counted to let the system reach steady state. The estimates were based on 20 batch means in each case.

\[ \hat{\theta}_{\text{mean}} \] 

variables and \( \hat{\theta}_{\text{mean}} \); see his process.
hyperexponential ($H^2_ρ$) with squared coefficient of variation $c^2 = 2$; see (10)-(12) below. We considered three different traffic intensities: $ρ = 0.3, 0.5, \text{ and } 0.7$. In Table 1 we have displayed the sample means and standard deviations. Note that the sample standard deviations of $\bar{β}$ and $\bar{δ}$ are significantly less than the sample standard deviations of $β$ and $δ$.

In [19, Appendix 1] we also compare these estimators to a simulation of an $H^2_ρ + H^2_ρ / M / 1 / \infty$ queue, where the arrival process is the superposition of two independent $H^2_ρ$ renewal processes. Since the arrival process is not a renewal process, a significant bias is apparent in the estimators $\hat{β}$, but $\hat{β}$ behaves just as in Table 1 because $\hat{ρ}$ is unbiased for general $G / G / 1 / \infty$ queues. Hence, we recommend using $\hat{δ}$ and $\hat{β}$ instead of $δ$ and $β$ whenever possible.

2.3 Selecting an Approximate Arrival Process

Given the parameters $ρ, α$ and $λ$ or their estimates, we now want to construct an approximating arrival process. Obviously the arrival rate should be $λ = ρλ$. Hence, we let the approximating arrival process be a renewal process with mean interarrival time $λ^{-1}$. We use the parameters $ρ$ and $α$ to determine the specific renewal process. If $α > ρ$, we let the interarrival time have a hyperexponential distribution; if $α = ρ$, we let the interarrival time have an exponential distribution; and if $α < ρ$, 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 $H^2_ρ / M / 1 / \infty$ 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 $H^2_ρ$); i.e., the distribution has the density

$$f(x) = φ_1 e^{-λ_1 x} + (1 - φ_1) e^{-λ_2 x}, \quad x > 0,$$

where $0 < φ_1 < 1$ and

$$λ^{-1} = 2q/λ_1 = 2(1 - q)/λ_2;$$

see Kuehn [44] and page 52 of Morse [45]. With these parameters, the interarrival time can have any coefficient of variation $c$ (standard deviation divided by the mean) greater than one:

$$c = \left(\frac{1}{1 + (2q - 1)^2} \right)^{1/2} > 1.$$  

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 $ρ$, we can solve for $ρ$ and $α$. Of course, $α = λ/μ$. To find $α$, we must solve $φ(ρ(1 - φ)) = φ$ and

$$α = \frac{φ_1}{ρ(1 - φ) + λ_1} + \frac{(1 - φ_1)λ_2}{μ(1 - φ) + λ_2}.$$  

Using (11) we obtain

$$σ = ρ + \frac{1}{2} \sqrt{1 - 4q(1 - φ) \left[1 - 4q(1 - φ)\right]}.$$  

(14)
Since \( \beta(1 - \rho) < 1/4 \) and \( \gamma(1 - \phi) < 1/4 \), we see that \( \phi > \beta \). In order to have \( \sigma = \rho \), it is necessary and sufficient to have \( q = 1 - q = 1/2 \), but then \( \lambda_\rho = \lambda_q \) by (11) and the hyperexponential distribution reduces to an exponential distribution.

From (10) and (11), we can also find \( q, \lambda_1 \) and \( \lambda_2 \) from \( \mu, \rho \) and \( \sigma \). In particular,

\[
q = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{(\sigma - \rho)(1 - \sigma + \rho)}{\rho(1 - \rho)}}.
\]

(15)

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

Having found the parameters \( q, \lambda_1 \) and \( \lambda_2 \) of the approximating \( E_\alpha^2 \) arrival process, we can calculate the two parameters \( \rho \) and \( \sigma \) of the GI/M/1 system with any new exponential service rate using (11) and (14).

The \( E_\alpha^2 / M / 1 / \infty \) System. Here we let the interarrival time be the sum of two independent exponential random variables with means \( \lambda_{\alpha}^{-1} \) and \( \lambda_{\alpha}^{-1} \), denoted by \( E_\alpha^2 \). When \( \lambda_\alpha = \lambda_{\alpha} \), we obtain an ordinary Erlang \( E_\alpha \) 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:

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

(16)

Given the parameters \( \lambda_1, \lambda_2 \) and \( \mu \), we can solve for the parameters \( \rho \) and \( \sigma \). Of course, \( \lambda_{\alpha}^{-1} = \lambda_1^{-1} + \lambda_2^{-1} \) and \( \rho = \lambda_{\alpha}/\mu \). To find \( \sigma \), we must solve \( \phi(\mu(1 - \sigma)) = \sigma \) or

\[
\frac{\lambda_\alpha^2}{\mu(1 - \sigma)}(\mu(1 - \sigma) + \lambda_\alpha) = \sigma,
\]

(17)

from which we obtain

\[
\sigma = \frac{(1/2)(\lambda_1 + \lambda_2 + \mu)(\lambda_1 + \lambda_2 + \mu - 4\lambda_\alpha^2)}{(1/2)(\lambda_1 + \lambda_2 + \mu)}.
\]

(18)

Given \( \lambda_\alpha \), \( \mu \) and \( \sigma \), we can also solve for the parameters \( \lambda_1 \) and \( \lambda_2 \):

\[
\lambda_\alpha = \frac{x}{\lambda_\alpha} \Rightarrow \lambda_\alpha = \frac{1}{\lambda_\alpha} \left( \frac{x}{\lambda_\alpha} \right)^2 - 4x,
\]

where

\[
x = \sqrt{\mu(1 - \sigma) + \rho - \sigma}.
\]

(19)

(20)

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

\[
\sigma(1 - \sigma) > 4\rho(\rho - \sigma).
\]

(21)

If \( \sigma < \rho \) but the inequality in (21) is not satisfied, then we can try the \( M^\mu \) arrival process below.
Having found the parameters $\lambda$ and $\lambda_1$, we can find the key parameters $\rho$ and $\sigma$ associated with any new service rate $\mu$ using (18) and the relations $\rho = \lambda_1/\mu$ and $\lambda_1^{-1} = \lambda^{-1} + \lambda_1^{-1}$. 

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

$$f(x) = \lambda e^{-\lambda(x-d)} \quad x > d,$$

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 of $\sigma$ attained when $d = 0$:

Given $\lambda_1$, $d$ and $\mu$, we can solve for the root $\sigma$. Solving the equation $\phi(\mu(1 - \sigma)) = \sigma$, we find that $\sigma$ satisfies

$$\frac{\lambda_1 e^{-\lambda_1 (1 - \sigma)}}{\lambda_1 + \mu(1 - \sigma)} = \sigma.$$  

which can easily be solved on a computer by a search routine. (The left side of (22) 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 $\lambda_1 = \mu d$ and $d = \lambda_1^{-1} - \lambda_1^{-1}$, we can also solve (23) for $\lambda_1$ given $\mu$, $\sigma$ and $d$. From (23), we see that $\lambda_1^{-1}$ is the solution to

$$e^{-\lambda_1 (1 - \sigma)} = \sigma(1 + \mu(1 - \sigma)).$$

(23a)

Note that both sides of (23a) are increasing in $\lambda_1^{-1}$. Also the largest possible values on the left and right are 1 and $\sigma(1 + d - \sigma)$, respectively. Since $\sigma(1 + d - \sigma) < 1$ because $\sigma > 0$, a solution to (23a) must exist when $e^{-\lambda_1 (1 - \sigma)} < \sigma$ 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 $\sigma$ 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 $G_i/M/1/\infty$ System

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

Here is the experiment: We simulated five $H_2 + H_2/M/1/\infty$ 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. $H_2$ process be 2 and $\bar{H}_2$ respectively. Thus the arrival rate in the superposition process is one. From (12), we see that the mixing probabilities in each component process are $\varphi$ and $1 - \varphi$ where $\varphi = 0.7886$. 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 $\varphi_i$ using $\varphi_i$, in (6).

Then we estimated $\rho$ using $\hat{\rho}$ in (9) and the known $\rho$. (As can be seen from [19, Appendix 1], the estimate $\hat{\rho}_i$ will usually be very close to $\rho$, so similar results can be expected if we use the estimate. However, this example illustrates that we need not
always estimate \( \rho \). Given \( \mathbb{E} \) and \( \rho \), we estimated the coefficient of variation \( c \) using (15) and (12). We used \( c \) instead of \( \sigma \) because \( c \) describes the arrival process alone whereas \( \sigma \) also depends on the service rate \( \mu \). 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 \( \sigma \) 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 \( \rho \); 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 \( \rho = 0.3 \) 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 suggest 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>
<thead>
<tr>
<th>( \lambda )</th>
<th>( M/M/1/\infty )</th>
<th>( \rho = 0.3 )</th>
<th>( \rho = 0.5 )</th>
<th>( \rho = 0.7 )</th>
<th>( \rho = 0.8 )</th>
<th>( \rho = 0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.459</td>
<td>0.461</td>
<td>0.469</td>
<td>0.472</td>
<td>0.483</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>1.00</td>
<td>1.13</td>
<td>1.15</td>
<td>1.17</td>
<td>1.18</td>
<td>1.24</td>
</tr>
<tr>
<td>0.7</td>
<td>2.33</td>
<td>2.77</td>
<td>2.79</td>
<td>2.81</td>
<td>2.83</td>
<td>2.86</td>
</tr>
<tr>
<td>0.8</td>
<td>4.00</td>
<td>4.87</td>
<td>4.90</td>
<td>5.13</td>
<td>5.36</td>
<td>5.63</td>
</tr>
<tr>
<td>0.9</td>
<td>9.00</td>
<td>11.2</td>
<td>11.3</td>
<td>12.0</td>
<td>12.2</td>
<td>12.3</td>
</tr>
</tbody>
</table>

Note: The simulated values are uncorrected. 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 generality of the approximation procedure. 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-Betz [12] work better. They obtained approximations for the mean waiting time and the probability that a customer will be delayed in a $G/G/1/\infty$ 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:

$$q_1 = \rho + \left( \frac{\lambda}{1 - \rho} \right) \left( c_1^2 + c_2^2 \right) h(c_1, c_2, \rho),$$

(24)

where $c_1$ and $c_2$ are the coefficients of variation of the interarrival time and service time and

$$h(c_1, c_2, \rho) = \begin{cases} \exp \left( \frac{-2(1 - \rho)}{3 \rho} \left( \frac{[1 - c_1^2]/[c_1 + c_2^2]}{1 - c_1^2} \right) \right), & c_1 < 1, \\ \exp \left( \frac{-2(1 - \rho)}{3 \rho} \left( \frac{[c_2^2 - 1]/[c_2^2 + 4c_1^2]}{c_2^2 - 1} \right) \right), & c_2 > 1; \end{cases}$$

(25)

see §1.3 of [12].

Given $c_1$, $\rho$ and $q_1$, or their estimates ($\bar{q}$ in (9) and $\hat{q}_1$ in (6)), we can solve (24) iteratively on the computer for $c_2$. A search routine is easy to implement because the approximation formula for $q_1$ in (24) is a strictly increasing function of $c_2^2$. Given $c_1$, we can fit the appropriate distribution ($H_1^2$, $M$, $E_2$, or $M^*$). As before, if $c > 1$, we use $H_1^2$; if $c = 1$, we use $M$; if $1/\sqrt{c} < c < 1$, we can use $E_2$; and if $0 < c < 1$, we can use $M^*$. To illustrate this approximation procedure, we simulated several $H_1^2 + H_1^2/g/1/\infty$ systems in which the service time distribution is either $H_1^2$, $E_2$, or $M$. As in §2, we let the renewal interval in each $H_1^2$ arrival process have mean $\mu$ and coefficient of variation $c_2^2$. 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 $H_1^2$ and $E_2$ service times were $\frac{c_1^2}{c_2^2}$ and $\frac{1}{c_2^2}$, respectively. For each simulation we estimated $q_1$ using $\hat{q}_1$, in (6). With this value of $q_1$, plus the known values of $\rho$ and $c_1$, we determined $c_2$ using the KL (Krämer and Langenbach-Betz) formula (24). We then used this approximate value of $c_2$ 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 approximations differ from the same corner as the uncertainties in the estimates. The values of the approximating squared coefficient of variation $c_2^2$ appear in the bottom row.
### Approximating a Point Process by a Renewal Process

#### Table 3

<table>
<thead>
<tr>
<th>New System</th>
<th>Traffic Intensity</th>
<th>$H_2^2 + H_2^2/H_2/1$</th>
<th>The Test System</th>
<th>$H_2^2 + H_2^2/M/1$</th>
<th>$H_2^2 + H_2^2/E_2/1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$c_f = 2.0$</td>
<td>$c_f = 1.8$</td>
<td>$c_f = 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H^2 + H_2^2/H_2/1$</td>
<td>0.5</td>
<td>1.46 (128)</td>
<td>1.42</td>
<td>1.43</td>
<td>1.40 (143)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>7.16 (716)</td>
<td>6.84</td>
<td>6.93</td>
<td>6.72 (693)</td>
</tr>
<tr>
<td>$H^2 + H_2^2/M/1$</td>
<td>0.5</td>
<td>1.20 (120)</td>
<td>1.16 (120)</td>
<td>1.17</td>
<td>1.15 (117)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>5.51 (551)</td>
<td>5.34</td>
<td>5.32</td>
<td>5.30 (530)</td>
</tr>
<tr>
<td>$H^2 + H_2^2/E_2/1$</td>
<td>0.5</td>
<td>1.06 (106)</td>
<td>1.03</td>
<td>1.03</td>
<td>1.01 (101)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>4.67</td>
<td>4.67</td>
<td>4.67</td>
<td>4.67 (467)</td>
</tr>
</tbody>
</table>

**Notes:**
1. The simulated values appear on the diagonal and are undefined. The standard error (one sample standard deviation) appears in parentheses below the estimate.
2. The approximate values of $c_f$ obtained by inverting formula (24) appear in the bottom row. This value of $c_f$ is used to generate the off-diagonal elements of the columns.

### The Indirect Method

For both $G/M/1$ or $G$ queues (22) and the KL approximation (24) are based on two-moment approximations. To put these approximations in perspective, we indicate the range of possible $g_1$ values (mean number in system) over all possible $G/M/1$ queues with given traffic intensity $p$ and squared coefficient of variation of the interarrival time $c_f$. We also indicate where the $H_2^2/M/1$ 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 $g_1$ 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 $\lambda^{-1}$ is the mean of the renewal interval. The smallest possible $g_1$ value is not attained, but the limit is the one-point distribution having mass 1 on $\lambda^{-1}$. This distribution is the limit as $k \to \infty$ of two-point distributions having mass $k^2/(i + k^2)$ on $\lambda^{-1}(1 - c/k)$ and mass $1/(1 + k^2)$ on $\lambda^{-1}(1 + \infty)$.

Table 4 gives the range of $g_1$ values along with the Krämer and Lüngenbach-Belz approximation. It is apparent that the range is very wide, with the lower bound being especially far away from the $H_2^2/M/1$ approximation. We expect the two-moment approximation to work well, not because two moments determine $g_1$ 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.
The Approximation: T. Indian - 629 A Newel Process

<table>
<thead>
<tr>
<th>The approx. Form.</th>
<th>( P(A, \alpha, \gamma, \xi) ) in the system</th>
<th>( H_{\xi}^0 )</th>
<th>( H_{\xi}^1 )</th>
<th>( H_{\xi}^2 )</th>
<th>( H_{\xi}^3 )</th>
<th>( H_{\xi}^4 )</th>
<th>( H_{\xi}^5 )</th>
<th>( H_{\xi}^6 )</th>
<th>( H_{\xi}^7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>in system</td>
<td></td>
<td>0.5</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td></td>
</tr>
<tr>
<td>( \alpha = 0.5 )</td>
<td>0.5</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta = 0.5 )</td>
<td>0.5</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma = 0.5 )</td>
<td>0.5</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Approximate
Value of \( q \)

<table>
<thead>
<tr>
<th>( H_{\xi}^0 )</th>
<th>2.05</th>
<th>2.05</th>
<th>1.84</th>
<th>1.90</th>
<th>1.76</th>
<th>1.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{\xi}^1 )</td>
<td>2.00</td>
<td>2.00</td>
<td>1.84</td>
<td>1.90</td>
<td>1.76</td>
<td>1.80</td>
</tr>
</tbody>
</table>

None: (1) The expected values in the numerator are undefined. The standard error (sample standard deviation) appear in parentheses below the value.

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

The Indirect Method for both \( G / M / 1 / 1 \) 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 \( q \) values (i.e., mean number in system) over all possible \( G / M / 1 / 1 \) queues with given traffic intensity \( \rho \) and squared coefficient of variation of the interarrival time \( c^2 \). We also indicate where the \( H_{\xi}^0 \) and \( H_{\xi}^1 \) values fall within this range.

It is possible to give the range because the exact distributions are known. See page 138 of Eckberg [5]. The largest possible \( q \) is obtained when the two-moment distribution has mass 1/(1 + \( c^2 \)) on 0 and \( 1/(1 + \gamma^2) \) on \( \rho \gamma \), where \( \rho \) is the mean of the normal distribution. The possible \( q \) value is not attained, but the limit is the zero-mean distribution having mean \( \lambda \) on 0 and \( \lambda^2/2 \) on \( \lambda^2 \). This distribution has the limit as \( k \to \infty \) of a two-point distribution with means \( \gamma \gamma/(1 + k^2) \) on \( \gamma \gamma/(1 + k^2) \) and mass \( 1/(1 + k^2) \) on \( \lambda^2/(1 + k^2) \).

Table 4 gives the range of possible \( q \) values along with the original and Langenbach's approximation. It is apparent that the approximation is at least as high as the lower bound for the two-moment approximation with \( \xi = \gamma \gamma/(\gamma \gamma + \lambda^2) \). We expect the two-moment approximation often to work well, but in situations where the system 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 \( G / M / 1 \) Queues with Given \( \lambda, \mu, \) and \( c_2 = 2 \)

<table>
<thead>
<tr>
<th>Traffic Intensity ( \rho )</th>
<th>Lower Bound</th>
<th>Krainer and Langebach-Belt approximation</th>
<th>Upper Bound</th>
<th>( H_{11} / M / 1 ) / 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.31</td>
<td>0.47</td>
<td>0.49</td>
<td>0.94</td>
</tr>
<tr>
<td>0.5</td>
<td>0.63</td>
<td>1.19</td>
<td>1.23</td>
<td>1.49</td>
</tr>
<tr>
<td>0.7</td>
<td>1.32</td>
<td>3.05</td>
<td>3.15</td>
<td>3.93</td>
</tr>
<tr>
<td>0.8</td>
<td>2.15</td>
<td>5.48</td>
<td>5.56</td>
<td>6.45</td>
</tr>
<tr>
<td>0.9</td>
<td>4.62</td>
<td>12.85</td>
<td>13.04</td>
<td>13.86</td>
</tr>
</tbody>
</table>

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 \( M^x + M^x, M^x + H_2^x, \) and \( H_2^x + H_2^x \) where \( c_2 = 0.6 \) in each \( M^x \) process and \( c_2 = 5.56 \) in each \( H_2^x \) process. The results for the indirect method in these cases are displayed in Tables 5–7. It is significant that the variability of the \( H_2^x \) process here is much greater than in Table 2. In the two cases in which such a high variability \( H_2^x \) 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 500,000 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 \( H_2^x + H_2^x / M / 1 \) queue, we see that the indirect method using KL approximation (24) is about the same as the \( G / M / 1 \) fit in §2.

TABLE 5
Approximate Values for the Mean Number of Customers in the System \( (q_1) \) in an \( M^x + M^x / M / 1 \) Queue with \( c_2 = 0.6 \) in each \( M^x \) Process. Using the KL Approximation (24)

<table>
<thead>
<tr>
<th>Traffic Intensity of the New System</th>
<th>Traffic Intensity of ( M^x )</th>
<th>0.3</th>
<th>0.6</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
<td>0.37</td>
<td>(0.007)</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.22</td>
<td>(0.007)</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>3.11</td>
<td>(0.015)</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>6.77</td>
<td>6.85</td>
<td>7.05</td>
</tr>
</tbody>
</table>

Approximating Square Root Coefficient of Variation \( c_1^2 \)

| 0.469 | 0.488 | 0.53 | 0.54 |
### TABLE 6
Approximate Values for the Mean Number of Customers in the System \( \langle q \rangle \) in an
\( H^2 + H^2 / M / 1 \) Queue, with \( c^2 = 9.56 \) in each \( H^2 \) Process, Using the
KL Approximation (24)

<table>
<thead>
<tr>
<th>Traffic Intensity of the New System</th>
<th>0.3</th>
<th>0.6</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.58 (0.004)</td>
<td>0.66</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>2.47 (0.035)</td>
<td>3.60</td>
<td>4.27</td>
<td>4.26</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>12.7</td>
<td>15.6</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>19.7</td>
<td>32.8</td>
<td>40.8</td>
<td>40.6</td>
</tr>
<tr>
<td>Approximating Squared Coefficient of Variation ( c^2 )</td>
<td>3.80</td>
<td>7.34</td>
<td>9.49</td>
<td>9.44</td>
</tr>
</tbody>
</table>

### TABLE 7
Approximate Values for the Mean Number of Customers in the System \( \langle q \rangle \) in an
\( M^\infty + H^2 / M / 1 \) Queue (\( c^2 = 9.56 \) and \( c^2 = 0.5 \)) Using the
KL Approximation (24)

<table>
<thead>
<tr>
<th>Traffic Intensity of the New System</th>
<th>0.3</th>
<th>0.6</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.45 (0.002)</td>
<td>0.51</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>1.68 (0.025)</td>
<td>2.19</td>
<td>2.64</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>4.70</td>
<td>6.78</td>
<td>8.65</td>
<td>9.12</td>
</tr>
<tr>
<td></td>
<td>10.9</td>
<td>16.5</td>
<td>21.6</td>
<td>22.9</td>
</tr>
<tr>
<td>Approximating Squared Coefficient of Variation ( c^2 )</td>
<td>1.48</td>
<td>2.95</td>
<td>4.31</td>
<td>4.66</td>
</tr>
</tbody>
</table>

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 \( \mu^{-1} \) 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 [5] 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 \( j \)th moment of \( Q \). We shall use \( \bar{q}_1 \) and \( \bar{q}_2 \) to specify the arrival rate \( \lambda \) and the coefficient of variation \( c \) of the renewal interval.
4.1 Background on the GI/M/∞ 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 = μ 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: \( \{ \phi(k), k \geq 1 \} \); see [16]. However, it is significant that, given μ, the parameters \( q_1 \) and \( q_z \) depend on only two parameters. From page 164 of [16], it follows that

\[
q_1 = \frac{a}{\lambda} \quad \text{and} \quad z = \frac{q_1}{q_z} - 1 = \frac{1}{1 - \phi(a)} - a. \tag{26}
\]

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 (a 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 \( q_z \) 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 λ = μz. Below we will see how to determine \( z \) 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 \( z > 1 \), we use \( H_z^a \); if \( (3a + 1)/(6a + 1) < z < 1 \), we use \( E_z^a \); and if \( z < (3a + 1)/(6a + 1) \) or also if \( z < 1 \), we use \( M^a \). To find an indication of the range, we note that the minimum peakedness given the offered load a is the peakedness associated with a D/M/∞ system; then \( z_{\min} = (1 - e^{-1/(3a)})^{-1} - a \). 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 \( H_z^a/M/∞ \) System. Given the service rate \( \mu \) and an \( H_z^a \) distribution as specified in (10) and (11), we can calculate the peakedness using (26). We obtain

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

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

\[
q = \frac{1}{2} \left[ \frac{1}{a} \sqrt{a + 1} \sqrt{\frac{z - 1}{z}} - 1 \right]. \tag{28}
\]

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 \( E_z^a/M/∞ \) System. Given the \( E_z^a \) distribution with parameters \( \lambda_1 \) and \( \lambda_2 \) 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
\[ \lambda_1 + \lambda_2. \text{ We obtain} \]
\[ \lambda = \left( B = \sqrt{b^2 - 4 \lambda^2 B} \right) / 2 \lambda, \]  
(29)

where
\[ b = \lambda \mu D / (\lambda - (\lambda + \mu) D) \] and \[ D = 1 - (z + \alpha)^{-1}. \]  
(30)

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

The \( M^2 / M / \infty \) System. Given the shifted exponential distribution with the density in (22), we can use (26) to solve for the peakness:
\[ \frac{\lambda e^{-\lambda}}{\lambda + \mu} = 1 - (z + \alpha)^{-1}. \]  
(31)

Alternatively, given \( \mu, a, \) and \( z \), we can let \( \lambda = a \mu \) and \( d = \lambda^{-1} - \lambda_1^{-1} \), and then solve (31) for \( \lambda_1 \). As \( d \to 0 \), \( z \to 1 \); as \( d \to \lambda^{-1} \), \( z \to (1 - e^{-1}) \alpha \gamma^{-1} - \alpha \), the minimum peakness 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 \( H_2^1 / H_2^2 / M / \infty \) system in which the two component arrival processes are identically distributed with \( c^2 = 2.0 \) or, equivalently, with \( \epsilon = 0.788675 \). We first considered the \( H_2^1 / M / \infty \) system with offered loads of \( a = 1, 2, 5, 10, \) and \( 20 \). We calculated the peakness in each case using (27). These peakness values also apply to the \( H_2^1 / H_2^1 / M / \infty \) system, so we obtained the actual offered loads and peakness values for the \( H_2^1 / H_2^2 / M / \infty \) system. These in turn were used with (28) to obtain an approximating \( H_2^2 \) arrival process. The results are displayed in Table 8. Based on the

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Approximate Values of the Peakness (( \ell )) for ( H_2^1 / H_2^2 / M / \infty ) Systems Using the Indirect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offers Load in the Test System</td>
<td>( \alpha = 2 )</td>
</tr>
<tr>
<td>( \alpha = 2 )</td>
<td>1.000000</td>
</tr>
<tr>
<td>( \alpha = 3 )</td>
<td>1.250000</td>
</tr>
<tr>
<td>( \alpha = 10 )</td>
<td>1.250000</td>
</tr>
<tr>
<td>( \alpha = 20 )</td>
<td>1.250000</td>
</tr>
<tr>
<td>( \alpha = 40 )</td>
<td>1.250000</td>
</tr>
<tr>
<td>Approximate Value of ( \ell )</td>
<td>0.750000</td>
</tr>
<tr>
<td>Approximate Value of ( \ell^2 )</td>
<td>0.6777</td>
</tr>
</tbody>
</table>

Notes: (1) Each component \( H_2^1 \) process has \( \epsilon^2 = 2.0 \), or equivalently, \( \epsilon = 0.788675 \).
(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 peakiness does not change much in the range studied.) Of course, the approximate peakiness obtained via this indirect method is closer to the actual peakiness than the peakiness of the test system. Also, in judging these approximations, note that they are for the peakiness 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 $\lambda$. The traffic with the known service rate $\mu$ 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 peakiness $z$.

We then apply §4 to characterize the approximating renewal process.

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

\[
R = a'E(a' + s, a'),
\]

\[
a = a'E(a', a'),
\]

\[
v = a'(1 - a + \frac{a'}{s + 1 + a - a'})
\]

and then set $z = v/a$, where $R$ 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 $E(a, s)$ is the Erlang loss formula. However, given $z$, $a$, and $s$, it is convenient to use the Rapp approximation to determine $a'$ and $z'$; 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 $B(a, s, r)$ as a function of $z$, $u$, and $x$ by $E(a/2, a/2)$. 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 $H_2^4/M/s/0$ system with $\lambda = 4$, $\mu = 1$, $s = 15$, and $H_2^4$ parameter $g = 0.984123$ which corresponds to the peakiness $z = 4$; 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 \( z = 50, a = 40, \) 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 \( a = \lambda / \mu. \) Eekberg [4] has shown that in a general \( G/G/\infty \) the peakedness \( z \) as a function of the service time distribution \( G \) is

\[
x(z) = 1 + a^{-1} \int_0^\infty [k(x) - \lambda \delta(x)] G(x) G(x - y) dy dx,
\]

where

\[
G(\theta y) = \int_0^\infty G(x) G(x - \theta y) dx,
\]

\( \delta \) is the Dirac delta function and \( k(x) \) is the covariance density of the arrival epochs. (See [18, 77] 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 a function 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

\[
x = 1 + (a^2 - 1) \mu \int_0^\infty (1 - G(x))^2 dx
\]

where \( G \) is again the service time distribution. Given the service time distribution \( G, \) the peakedness \( z, \) it is easy to solve (33) for the coefficient of variation of the interarrival time, \( a^2. \)

When there is waiting room and more than one server, we have difficulties carrying out our procedure. For \( G/G/\infty \) 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 Halachmi and Whitt [8].
Finding a way to apply the indirect method to G/G/z/k systems seems to be a promising direction for future research.1

1The simulation and associated data collection support of the numerical computations were done by my colleague Syeda Alim. I am grateful to her for comments on the manuscript and many helpful discussions. I also thank the referees for their assistance.

References

4. Eckberg, R. E., private communication.