AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

On Approximations for Queues, I: Extremal Distributions

By W. WHITT

(Manuscript received May 4, 1983)

Many approximations for queueing characteristics such as the mean equilibrium queue length are based on two moments of the interarrival and service times. To evaluate these approximations, we suggest looking at the set of all possible values of the queueing characteristics given the specified moment parameters. This set-valued function is useful for evaluating the accuracy of approximations. For several models, such as the GI/M/1 queue, the set of possible values for the mean queue length given limited-moment information can be conveniently described by simple extremal distributions. Here we calculate the set of possible values for the mean queue length in a GI/M/1 queue and show how it depends on the traffic intensity and the second moment. We also use extremal distributions to compare alternative parameters for approximations. The results provide useful insights about approximations for non-Markov networks of queues and other complex queueing systems. The general procedure is widely applicable to investigate the accuracy of approximations.

I. INTRODUCTION AND SUMMARY

Queueing models are important tools for studying the performance of complex systems, but despite the substantial queueing theory literature, it is often necessary to use approximations. The purpose of this series of papers is to help develop a theory for evaluating queueing

* AT&T Bell Laboratories.

approximations. Devising appropriate queueing approximations no doubt will continue to be largely an art, but we believe that there is a need and a real possibility for more supporting theory.

In this series of papers we examine the accuracy of queueing approximations that are based on a few parameters partially characterizing the arrival process and the service-time distribution. We use an approach originally introduced by Holtzman[1] and Eckberg[2] at Bell Laboratories and Rolski[3–5] in Poland. Since the approximations apply to all arrival processes and all service-time distributions with the same parameters, we propose evaluating the approximations by examining the set of all possible values of the congestion measure consistent with the specified parameters. To be specific, consider the GI/G/1 queue, which has a single server, unlimited waiting room, the first-come first-served discipline, and a renewal arrival process independent of iid (independent and identically distributed) service times. Many approximations for the equilibrium mean queue length in the GI/G/1 queue are based on the first two moments of the interarrival-time and service-time distributions; see Shanthikumar and Buzacott[6] and Whitt.[7] In this context we suggest considering the set-valued function that maps the four moment parameters into the set of possible values of the mean queue length.

It should be clear that we are in an excellent position to develop and evaluate approximations if we can identify such set-valued functions. We can see if a candidate approximation is an element of this set for all parameters of interest; then there always is a system for which the approximation is exact. We can also see if an approximation is in the middle of this set; then large errors are avoided and the approximation usually corresponds to a typical system value.

There is also much to be learned without considering any specific approximation. The range of values indicates the possible accuracy of any approximation. We can investigate how this range depends on the parameters to determine how the possible accuracy depends on the parameters. We can see how the range is reduced by incorporating additional information, e.g., another moment. We can also compare different parameter specifications by comparing the different set-valued functions.

This approach has wide applicability in queueing and elsewhere, provided that we can indeed identify the desired set-valued functions. As one would expect, this task is usually difficult, but there is an emerging methodology for attacking this problem. It is sometimes possible to identify relatively simple extremal distributions that yield the maximum and minimum values of the congestion measure given the parameters. A major tool for this purpose is the theory of complete Tchebycheff systems in Karlin and Studden.[8] The idea of applying

complete Tchebycheff systems and extremal distributions to congestion models is due to Holtzman[1] and Rolski.[3] Eckberg[2] first used this approach to compare alternate parameter specifications, primarily the peakedness versus the variance as a second parameter in addition to the mean in GI/M/s loss systems. Other relevant references are Bergmann et al.,[9] Daley and Rolski,[10] Karr,[11] Stoyan,[12] and Whitt.[13,14]

The principal focus in the papers here is the GI/M/1 queue, which has an exponential service-time distribution. (We also have results for more general GI/G/1 queues; see Section V of this paper and Sections VI and VII of Part III, a subsequent paper in this issue of the *Journal.*) In Part I, we describe the set of all possible values of the mean queue length in the GI/M/1 model given the service rate and various parameters partially characterizing the interarrival-time distribution, especially the first two moments. We obtain useful descriptions of the way this set depends on the parameters (see Section II). For example, the maximum relative error [defined in (4)] in the mean queue length given the first two moments of the interarrival time turns out to be precisely the squared coefficient of variation (variance divided by the square of the mean) of the interarrival time; see Corollary 1. We also evaluate alternate parameter specifications (see Sections III and IV).

We must emphasize that we are not actually interested in the GI/M/1 model itself. Given a GI/M/1 model, it is obviously not difficult to calculate the mean queue length exactly. We are actually interested in more general models in which exact solutions are not possible. Where GI/M/1 models arise, they arise as approximations, e.g., the arrival process is approximated by a renewal process partially characterized by the first two moments of the renewal interval.[15–18] Then there is no corresponding renewal-interval distribution for exact analysis.

We became motivated to conduct this study while developing the software package QNA (Queueing Network Analyzer),[17,18] which calculates approximate congestion measures for non-Markovian networks of queues, i.e., with non-Poisson arrival processes and nonexponential service-time distributions. The procedure in QNA is, first, to approximate each arrival process by a renewal process partially characterized by the first two moments of the renewal interval and, second, for each node to apply approximation formulas for the congestion measures in a GI/G/m queue partially characterized by the first two moments of the interarrival-time and service-time distributions. It is natural to study these two steps separately. The first step is studied in Whitt[15] and Albin.[16] The second step is studied here.

For the network of queues and other applications, we would actually like to treat the more general GI/G/m model, but we are not yet able to do this. Nevertheless, we believe that the GI/M/1 results here are

important. They indicate what happens more generally. While the exponential distribution is exceptional in its analytic simplicity, it is rather typical in its degree of variability (in between deterministic and highly variable). Moreover, the sharp analytic results available for the GI/M/1 model will be useful theoretical reference points for other cases that require relatively complicated numerical methods or simulation. Even if an extremal distribution is identified for other GI/G/m queues, it may be a nontrivial task to calculate the mean queue length.

We emphasize that the relevance of the extremal distributions for the GI/M/1 model was established before.[1–5] Here we apply this theory to examine in detail the implications for queueing approximations. We determine which parameters are best, how the quality of approximations depends on the parameters, and how much additional information helps.

As an important part of our results, we display the extremal distributions yielding the extreme values of the mean queue length. These extremal distributions are of interest beyond the GI/M/1 queue considered here because they are also extremal in many other settings. (This will be evident from Sections II and V.) Moreover, in settings such as the GI/G/m queue in which the actual extremal distributions are still unknown, the GI/M/1 extremal distributions can be used in numerical methods and simulations to get an approximate range of possible congestion values.

To describe the situation for the GI/M/1 queue, let u be an interarrival time, v a service time, ρ the traffic intensity ($\rho = Ev/Eu$), c^2 the squared coefficient of variation of an interarrival time, and L the expected equilibrium queue length (number in system) at an arbitrary time. For the GI/M/1 queue,[19]

$$L = \rho/(1 - \sigma), \tag{1}$$

where σ is the unique root in the open interval (0, 1) of the equation

$$\phi[\mu(1 - \sigma)] = \sigma, \tag{2}$$

with $\mu = 1/Ev$ and $\phi(s)$ the Laplace-Stieltjes transform of the interarrival-time cdf, say F,

$$\phi(s) = \int_0^\infty e^{-st} dF(t). \tag{3}$$

The root σ in (2) is also of interest itself because it is the probability that a customer will have to wait before beginning service. It is clear from (1) and (2) that σ and L depend on the entire cdf F, not just its first two moments.

So, what about the range of possible values for σ and L in the GI/M/1 queue? Unfortunately, the range can be very wide. For ex-

ample, let $Eu = 2$, $Eu^2 = 12$ (so that $\mathrm{Var}(u) = 8$ and $c^2 = 2$), and $Ev = 4/3$ (so that $\rho = 2/3$). The possible values of σ range from 0.417 to 0.806 and the possible values of L range from 1.14 to 3.44, giving a maximum relative error of 200 percent (Table IV).

This wide range naturally causes us to question the value of the various two-moment approximations. However, the particular distributions yielding the extreme values of L suggest an explanation. These extremal distributions are two-point distributions, so they are obviously very unusual. We would hope that for typical (nice) distributions σ and L would not vary much among interarrival-time distributions with the same moments. In Parts II and III,[20,21] we investigate how much the range is reduced by imposing various shape constraints on the interarrival-time distribution. Part II by Klincewicz and Whitt[20] presents a new approach. Since the theory of complete Tchebycheff systems no longer applies with shape constraints, Part II uses nonlinear programming to identify the extreme values of L and the associated extremal interarrival-time distributions given various shape constraints. We believe that Part II is the first investigation of extremal distributions in the presence of shape constraints.

The numerical results in Part II are strikingly similar to the theoretical results in Part I, suggesting that a theory corresponding to Part I can be developed for many kinds of shape constraints. Part III shows how this can be done in one important special case. Part III shows that the theory of complete Tchebycheff systems can be applied again for one important kind of shape constraint: assuming that the distribution is a mixture of exponential distributions.

Overall, this study indicates that two-moment approximations can perform poorly, but if the distribution is not too irregular then they should perform reasonably well. At any rate, numbers are provided so that we can reach our own conclusions, which may depend on the circumstances.

Here is how the rest of this paper is organized. In Section II we study the extremal distributions with the first two moments fixed. In Section III we do a similar analysis with the mean and the peakedness (the transform evaluated at the service rate) fixed. In Section IV we investigate other parameter specifications, including the first three moments. Finally, in Section V we briefly discuss extremal distributions in other models such as the GI/G/1 queue and the GI/M/1 loss system. It is significant that the theory of extremal distributions is not limited to the GI/M/1 model.

II. EXTREMAL DISTRIBUTIONS GIVEN THE FIRST TWO MOMENTS

Consider the set of all probability distributions on the interval $[0, bm_1]$, $b \leq \infty$, having first two moments m_1 and m_2 (and no mass at

infinity). This is a convex set depending on the three parameters b, m_1 and c^2, where c^2 is the squared coefficient of variation: $c^2 = (m_2 - m_1^2)/m_1^2$. The set is nonempty provided that $b \geq 1 + c^2$. Two distributions in this set are of particular interest; we call them the upper and lower bounds because they yield the maximum and minimum mean queue lengths, respectively, among interarrival-time distribution in this set. The *upper bound* is the two-point distribution with mass $c^2/(1 + c^2)$ on 0 and mass $1/(1 + c^2)$ on $m_1(1 + c^2)$, having cdf denote by F_u, and the *lower bound* is the two-point distribution with mass $c^2/[c^2 + (b - 1)^2]$ on bm_1 and mass $(b - 1)^2/[c^2 + (b - 1)^2]$ on $m_1[1 - c^2/(b - 1)]$, having cdf denoted by F_ℓ. As $b \to \infty$, the lower bound approaches (converges in law) to the *limiting lower bound*, which is the one-point distribution with mass 1 on m_1, having cdf denoted by $F_{\hat{\ell}}$.

Note that the limiting lower bound is not actually in the reference set because it has zero variance. These distributions are especially useful because they are minimal and maximal elements for a partial ordering of the distributions based on the Laplace-Stieltjes transforms.

Definition 1: $F_1 \leq_L F_2$ for two cdf's on $[0, \infty)$ if $\phi_1(s) \leq \phi_2(s)$ for all $s \geq 0$, where ϕ_i is the Laplace-Stieltjes transform of F_i defined in (3).

Since the transform $\phi(s)$ is the expectation of a decreasing function, the smaller cdf in the ordering $\leq_L$ tends to have what we would normally think of as the stochastically larger distribution; in fact, in Section 1.8 of Stoyan,[12] $F_1 \leq_L F_2$ is said to hold if $\phi_1(s) \geq \phi_2(s)$ for all $s \geq 0$. However, smaller interarrival times mean more arrivals and more congestion. We use this definition because the upper-(lower-) bound distribution yields the maximum (minimum) mean queue length.

Let $\mathbf{F} \equiv \mathbf{F}(m_1, c^2, b)$ be the set of all cdf's with parameters m_1, c^2, and b. Let F_u and F_ℓ be the cdf's in $\mathbf{F}$ associated with the special extremal distributions, and let $F_{\hat{\ell}}$ be the associated limiting lower-bound cdf. The following proposition is just a restatement of 2.1.1 of Eckberg,[2] which in turn is an elementary consequence of the theory of complete Tchebycheff systems.[8]

Proposition 1: *For all* $F \in \mathbf{F}$, $F_{\hat{\ell}} \leq_L F_\ell \leq_L F \leq_L F_u$.

It is a simple matter to check the following property.

Proposition 2: F_ℓ *decreases in* $\leq_L$ *as* b *increases and* $\phi_\ell(s) \to \phi_{\hat{\ell}}(s)$ *for all* s *as* $b \to \infty$.

As noted by Holtzman,[1] Rolski,[3-5] and Eckberg,[2] the ordering $\leq_L$ and the extremal distributions have immediate application to queues. Consider the GI/M/1 queue with fixed service rate μ and interarrival-time distributions in $\mathbf{F}$. Without loss of generality, assume $m_1 = 1$. Now it is natural to work with the three parameters ρ, c^2, and b. Let L and σ in (1) and (2) be indexed to indicate the extremal interarrival-

time distributions. As an immediate consequence of Proposition 1 and (2), we have

Proposition 3: *For all* $F \in \boldsymbol{F}(\rho, c^2, b)$, $\sigma_{\hat{\ell}} \leq \sigma_\ell \leq \sigma \leq \sigma_u$ *and* $L_{\hat{\ell}} \leq L_\ell \leq L \leq L_u$.

Remark 1: More generally, if $F_1 \leq_L F_2$ for two interarrival-time cdf's, then $\sigma_1 \leq \sigma_2$ in the associated GI/M/1 queue with common service rate. This in turn implies not only that $L_1 \leq L_2$ but also that the associated steady-state queue-length distributions are stochastically ordered; see Theorem 5.2.3b of Stoyan.[12]

For approximations, it is interesting to know about the maximum relative error (*MRE*) in L, defined by

$$MRE \equiv MRE(\rho, c^2, b) \equiv (L_u - L_\ell)/L_\ell. \tag{4}$$

From (1), we see that $MRE = (\sigma_u - \sigma_\ell)/(1 - \sigma_u)$.

Now we show how the extremal queue characteristics (σ_ℓ, L_ℓ, etc.) and *MRE* depend on the parameters ρ, c^2, and b. We first describe how $\sigma_{\hat{\ell}}$ depends on ρ, the only relevant parameter for the limiting lower bound.

Theorem 1: For $0 < \rho < 1$, $\sigma_{\hat{\ell}} < \rho$ and

$$\frac{d\sigma_{\hat{\ell}}}{d\rho} = \frac{\rho^{-2}(1 - \sigma_{\hat{\ell}})e^{-(1-\sigma_{\hat{\ell}})/\rho}}{1 - \rho^{-1}e^{-(1-\sigma_{\hat{\ell}})/\rho}} > 0.$$

Proof: Consider eq. (2) for $F_{\hat{\ell}}$. The function

$$f(x) = x - e^{-(1-x)/\rho} \tag{5}$$

is positive for $0 < x < \sigma_{\hat{\ell}}$ and negative for $\sigma_{\hat{\ell}} < x < 1$, so to show that $\sigma_{\hat{\ell}} < \rho$ it suffices to show that $f(\rho) = \rho - e^{-(1-\rho)/\rho}$ is negative for $0 < \rho < 1$. Make the change of variables $y = (1 - \rho)/\rho$ to obtain $f(y) = (1 + y)^{-1} - e^{-y}$, which is clearly positive for all $y > 0$. To verify the inequality for the derivative, differentiate $f(x)$ in (5). Use $\sigma_{\hat{\ell}} < \rho$ to show that the denominator is always positive:

$$\rho^{-1}e^{-(1-\sigma_{\hat{\ell}})/\rho} \leq \rho^{-1}e^{-(1-\rho)/\rho} < 1.$$

We now show that all results for $\sigma_{\hat{\ell}}$ immediately imply results for σ_u.

Theorem 2: $\sigma_u = 1 - (1 - \sigma_{\hat{\ell}})/(1 + c^2)$.

Proof: For the upper bound, eq. (2) is

$$\frac{c^2}{c^2 + 1} + \frac{1}{c^2 + 1}\, e^{-(1-\sigma_u)(1+c^2)/\rho} = \sigma_u.$$

Multiply both sides by $c^2 + 1$, subtract c^2 from both sides, and then make the change of variables $1 - \sigma_{\hat{\ell}} = (1 - \sigma_u)(1 + c^2)$ to obtain eq. (2) for the limiting lower bound.

Corollary 1: $L_u = L_\ell (1 + c^2)$ *and* $MRE(\rho, c^2, \infty) = c^2$.

Remark: Theorems 1 and 2 together with (1) imply that σ_u and L_u are increasing in ρ.

We now turn to the lower bound when there is a bound on the distribution ($b < \infty$). Straightforward but tedious calculations (differentiation) verify the expected monotonicity properties:

Theorem 3: (*a*) *The lower-bound characteristics* σ_ℓ *and* L_ℓ *are increasing in* ρ *and* c^2 *and decreasing in b.* (*b*) $MRE(\rho, c^2, b)$ *is increasing in b.*

Combining Theorems 2 and 3b, we obtain

Corollary 2: $MRE(\rho, c^2, b) \leq c^2$.

Numerical evaluation of $MRE(\rho, c^2, b)$ for 14 values of ρ, 4 values of c^2, and 5 values of b support the following conjecture.

Conjecture 1: $MRE(\rho, c^2, b)$ *is decreasing in* ρ.

In Table I we display $MRE(\rho, c^2, b)$ for three values of ρ, four values of c^2, and four values of b. These specific cases show that $MRE(\rho, c^2, b)$ is strongly affected by each of the parameters ρ, c^2, and b. The bound b can make a big difference, especially for larger ρ and c^2; see the case $\rho = 0.9$ and $c^2 = 4$. These specific cases demonstrate that $MRE(\rho, c^2, b)$ is not monotone in c^2. In fact, when c^2 increases with b fixed, the lower-bound distribution F_ℓ eventually coincides with the upper-bound distribution F_u, becoming the two-point distribution with mass b^{-1} on b and mass $1 - b^{-1}$ on 0 ($m_1 = 1$). Of course, as $c^2 \to 0$, F_ℓ and F_u both approach $F_{\hat{\ell}}$, so that $MRE(\rho, c^2, b) \to 0$ too as $c^2 \to 0$. The numerical results also support the following conjecture:

Conjecture 2: $MRE(\rho, c^2, b)$ *is unimodal in* c^2.

We now investigate how the extremal queue characteristics and

Table I—Values of MRE(ρ, c^2, b) for the GI/M/1 queue*

Traffic Intensity, ρ	Squared Coefficient of Variation, c^2	Bound on Interarrival-Time Distribution in Multiples of the Mean			
		$b = 5$	$b = 10$	$b = 20$	$b = 40$
0.5	0.5	0.373	0.442	0.472	0.487
	1.0	0.604	0.833	0.924	0.964
	2.0	0.527	1.40	1.75	1.89
	4.0	0.000	1.28	3.01	3.59
0.7	0.5	0.231	0.350	0.424	0.462
	1.0	0.290	0.583	0.791	0.897
	2.0	0.185	0.699	1.33	1.68
	4.0	0.000	0.349	1.56	2.86
0.9	0.5	0.070	0.143	0.248	0.353
	1.0	0.072	0.174	0.365	0.610
	2.0	0.043	0.143	0.374	0.858
	4.0	0.000	0.071	0.232	0.712

* The maximum relative error in the steady-state mean queue length L given the traffic intensity ρ, the interarrival-time squared coefficient of variation c^2, and the bound on the interarrival-time distribution b (in multiples of the mean); see Section IV.

$MRE(\rho, c^2, b)$ behave in light and heavy traffic, i.e., as $\rho \to 0$ and $\rho \to 1$. As an easy consequence of (2), we obtain

Theorem 4: As $\rho \to 0$,

$$\sigma_w \to c^2/(1 + c^2), \qquad \sigma_\ell \to 0, \qquad \text{and} \quad MRE(\rho, c^2, b) \to c^2.$$

We describe the behavior as $\rho \to 1$ for $b < \infty$ in more detail. The following result provides an interesting refinement to the classical heavy-traffic limit theorem,[7] from which we can deduce that $(1 - \rho)L \to (1 + c^2)/2$ as μ approaches λ from above for any fixed renewal arrival process.

Theorem 5: For all b,

$$\frac{1 - \sigma_w}{\rho} = \frac{2(1 - \rho)}{1 + c^2} + \frac{4(1 - \rho)^2}{3(1 + c^2)} + 0(1 - \rho)^3 \tag{6}$$

and, for $b < \infty$,

$$\frac{1 - \sigma_\ell}{\rho} = \frac{2(1 - \rho)}{1 + c^2} + \frac{4(1 - \rho)^2}{3(1 + c^2)} \frac{m_3}{(1 + c^2)^2} + 0(1 - \rho)^3, \tag{7}$$

where

$$m_3 = \frac{c^2 b^3}{c^2 + (b - 1)^2} + \frac{(b - 1 - c^2)^3}{(b - 1)(c^2 + (b - 1)^2)}, \tag{8}$$

so that, for $b < \infty$,

$$\lim_{\rho \to 1} \frac{MRE(\rho, c^2, b)}{1 - \rho} = \frac{4}{3}\left(\frac{m_3}{(1 + c^2)^2} - 1\right). \tag{9}$$

Proof: Let $x = (1 - \sigma_\ell)/\rho$. To find the derivative of x with respect to ρ, differentiate with respect to ρ in eq. (2), i.e.,

$$1 - \rho x = e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + 0(x^4)$$

or

$$-\rho = \frac{x}{2} - \frac{x^2}{6} + 0(x^3).$$

After successive differentiation with L'Hospital's rule, this yields $x'(1) = -2$ and $x''(1) = -8/3$. From Taylor's theorem and Theorem 1, we obtain (6). The calculation for the lower bound in (7) is similar.

Remarks: It is possible to check the consistency of (6) and (7) because they must agree as $b \to 1 + c^2$. It is not possible to do a consistency check as $b \to \infty$ because the two iterated limits involving $b \to \infty$ and $\rho \to 1$ are not equal.

We conclude this section by displaying in Table II the extremal

Table II—The extremal GI/M/1 characteristics for fixed traffic intensity, ρ, squared coefficient of variation, c^2, and bound on the distribution b: Case of $c^2 = 2.0$

Traffic Intensity, ρ	Upper-Bound Characteristics	Bound on Interarrival-Time Distribution in Multiples of the Mean			
		$b = 5$	$b = 10$	$b = 20$	$b = 40$
0.2	$\sigma_u = 0.669$ $L_u = 0.604$	$\sigma_\ell = 0.092$ $L_\ell = 0.220$	$\sigma_\ell = 0.022$ $L_\ell = 0.204$	$\sigma_\ell = 0.012$ $L_\ell = 0.202$	$\sigma_\ell = 0.009$ $L_\ell = 0.202$
0.5	$\sigma_u = 0.734$ $L_u = 1.88$	$\sigma_\ell = 0.594$ $L_\ell = 1.23$	$\sigma_\ell = 0.361$ $L_\ell = 0.783$	$\sigma_\ell = 0.269$ $L_\ell = 0.684$	$\sigma_\ell = 0.233$ $L_\ell = 0.652$
0.7	$\sigma_u = 0.822$ $L_u = 3.94$	$\sigma_\ell = 0.790$ $L_\ell = 3.32$	$\sigma_\ell = 0.698$ $L_\ell = 2.32$	$\sigma_\ell = 0.585$ $L_\ell = 1.69$	$\sigma_\ell = 0.524$ $L_\ell = 1.47$
0.9	$\sigma_u = 0.936$ $L_u = 13.98$	$\sigma_\ell = 0.933$ $L_\ell = 13.41$	$\sigma_\ell = 0.926$ $L_\ell = 12.23$	$\sigma_\ell = 0.912$ $L_\ell = 10.17$	$\sigma_\ell = 0.880$ $L_\ell = 7.52$

characteristics σ_ℓ, L_ℓ, and σ_u, and L_u for the cases in Table I with $c^2 = 2$. The associated maximum relative errors for $\rho = 0.5$, 0.7, and 0.9 are given in Table I. These will be compared with other parameter specifications in the following sections.

III. THE SECOND PARAMETER: VARIANCE VERSUS PEAKEDNESS

The first two moments are natural parameters if two parameters are to be used to partially characterize an interarrival-time or a service-time distribution, but it is not clear that these are the best two parameters. Of course, the chosen parameters should be easy to estimate and easy to use in approximations for queues. Also, the parameters should have power determining descriptive queue characteristics; i.e., there should be a small *MRE* or a small range of possible values of L. In this regard, Eckberg[2] has shown that the peakedness of a renewal arrival process is a much better second parameter in addition to the mean than the variance for GI/M/k loss systems and also, to some extent, for GI/M/k delay systems. The peakedness is the ratio of the variance to the mean of the steady-state number of busy servers in an associated GI/M/∞ system; see Holtzman,[1] Eckberg,[22] and references there. Knowing the peakedness of a renewal process, say z, is equivalent to knowing $\phi(\mu)$, the transform evaluated at the service rate μ:

$$\phi(\mu) = 1 - (z + \lambda/\mu)^{-1}. \tag{10}$$

The peakedness is an important parameter to consider because it is often available as an approximate characterization of overflow processes via the equivalent random method.[1,22] Since Eckberg's results[2] suggest that the mean and the parameter $\phi(\mu)$ might be much better than the mean and variance, we investigate this new parameter pair here.

However, before examining this new parameter pair, we explain why the variance might be a better second parameter for single-server delay systems. Knowing the mean and variance (i.e., c^2) is equivalent to knowing the first two derivatives of the transform $\phi(s)$ at 0. It is intuitively reasonable that we might pin down the transform $\phi(s)$ better by fixing the value at μ, $\phi(\mu)$ than by fixing the second derivative at 0, $\phi''(0)$. However, this depends on the way the queue characteristics depend on the transform. For the GI/M/k loss system, the relevant parameters are $\phi(j\mu)$ for $j = 1, 2, \cdots, k$, with the parameters tending to be of less importance as j increases. These parameters are values of the transform $\phi(s)$ evaluated at points s such that $s \geq \mu$. For approximations, it is clearly better to specify $\phi(\mu)$ and $\phi'(0)$ than $\phi''(0)$ and $\phi'(0)$.

For the GI/M/1 delay system the key parameter in (2) is the transform value $\phi[\mu(1 - \sigma)]$. Of course, we do not know σ in advance, but the argument is always less than μ. Since σ tends to be near ρ, the argument tends to be near $\mu(1 - \rho)$. Clearly, for large ρ, knowing $\phi''(0)$ should be better than knowing $\phi(\mu)$. On the other hand, for small ρ, knowing $\phi(\mu)$ should be better than knowing $\phi''(0)$.

Our results substantiate this intuitive reasoning. In marked contrast to GI/M/k loss systems, for GI/M/1 delay systems the parameter $\phi(\mu)$ is not uniformly better than the variance as a second parameter. Which second parameter is better depends on the traffic intensity, with the variance improving as ρ increases. Consistent with the intuitive discussion above, we shall show that asymptotic behavior of the maximum relative error as ρ approaches 0 and 1 is strikingly different given $\phi(\mu)$ instead of c^2. Moreover, the variance does better for the upper bound, whereas $\phi(\mu)$ does better for the lower bound.

It is also appropriate to mention that we are considering the peakedness of the renewal arrival process as a single parameter, which by (10) can be represented as the transform ϕ evaluated at the service rate μ. If, instead, we knew the peakedness as a function of the service rate as in Eckberg,[22] then we would know the entire transform, which is equivalent to knowing the entire interarrival-time distribution. Moreover, if we could choose one argument of the transform, then we obviously could do better by picking a value less than μ. For example, there would be no error for the GI/M/1 queue if we could guess σ and make the argument $\mu(1 - \sigma)$. If we could choose one argument given only the arrival rate and service rate, then a natural choice would be $\mu(1 - \rho)$. (This parameter is considered here in Section IV.) In applications, however, we typically have no choice. Then the arrival process (which may not be renewal) may be partially characterized (by the equivalent random method and related techniques) by rate and peakedness. Moreover, the given peakedness might be with respect

to a different service rate (or even a different service-time distribution[22]). In the context of the GI/M/1 queue, this peakedness parameter will lead to better approximations if the argument of the transform, after using (10), is close to $\mu(1 - \sigma)$. The parameter $\phi(\mu)$ considered here should give some idea about what will happen in general.

The new parameter pair involving $\phi(\mu)$ leads to new two-point extremal distributions and a new partial ordering of the distributions. Now consider the set of all probability distributions on the interval $[0, bm_1]$, $b \leq \infty$, having first moment m_1 and transform $\phi(\mu)$ at $s = \mu$ (and no mass at infinity). This is a convex set depending on the parameters b, m_1, and $\phi(\mu)$. The extremal distributions here are the *upper bound*, which is the two-point distribution with mass $p = (b - 1)/(b - x)$ on xm_1 and mass $1 - p$ on bm_1, where x satisfies

$$pe^{-x/\rho} + (1 - p)e^{-b/\rho} = \phi(\mu); \tag{11}$$

and lower bound, which is the two-point distribution with mass $1 - x^{-1}$ on 0 and mass x^{-1} on xm_1, where

$$x = (1 - \rho^{-x/\rho})/(1 - \phi(1/\rho)). \tag{12}$$

Unlike Section II, the upper bound here depends on b while the lower bound does not. As $b \to \infty$, the upper bound converges in law to a *limiting upper bound*, which is the one-point distribution with mass 1 on $-(\log \phi(\mu))/\mu$. Note that the limiting upper bound is not actually in the reference set because the mean is not m_1. These distributions are minimal and maximal elements for another partial ordering of the distributions based on the transform.

Definition 2: $F_1 \leq_\mu F_2$ for two cdf's on $[0, \infty)$ if

$$\phi_1(s) \leq \phi_2(s), \quad s \leq \mu, \quad \text{and} \quad \phi_1(s) \geq \phi_s(s), \quad s \geq \mu.$$

Let $\mathbf{G} = \mathbf{G}(m_1, \mu, \phi(\mu), b)$ be the set of all cdf's with parameters m_1, μ, $\phi(\mu)$, and b. Without loss of generality, let $m_1 = 1$. Let G_u, G_ℓ, and $G_{\hat{u}}$ be the cdf's associated with the special extremal distributions. From Section 2.2.3 and (5) of Eckberg,[2] we obtain

Proposition 4: For all $G \in \mathbf{G}$, $G_\ell \leq_\mu G \leq_\mu G_u \leq_\mu G_{\hat{u}}$.

It is easy to see the effect of changing b:

Proposition 5: G_u increases in $\leq_\mu$ as b increases and $\phi_u(s) \to \phi_{\hat{u}}(s)$ for each s as $b \to \infty$.

Here are the implications for the GI/M/1 queue. A tilde is used to indicate that the extremal distributions are from this section (because we want to relate them to those in Section II).

Proposition 6: For all $G \in \mathbf{G}$,

$$\tilde{\sigma}_\ell \leq \sigma \leq \tilde{\sigma}_u \leq \tilde{\sigma}_{\hat{u}} \quad \text{and} \quad \tilde{L}_\ell \leq L \leq \tilde{L}_u \leq \tilde{L}_{\hat{u}}.$$

Using the same change of variables argument as in Theorem 2, we can express $\tilde{\sigma}_{\ell}$ in terms of $\sigma_{\hat{\ell}}$.

Theorem 6: $\tilde{\sigma}_{\ell} = 1 - (1 - \sigma_{\hat{\ell}})/x$ *for* x *in* (12).

Remarks: As a consequence of Theorem 6, $\tilde{L}_{\ell} = xL_{\hat{\ell}}$ for x in (12). Since x^{-1} is a probability, $\sigma_{\hat{\ell}} \leq \tilde{\sigma}$ and $L_{\hat{\ell}} \leq L_{\ell}$. Moreover, $\tilde{\sigma}_{\ell}$ and $\tilde{L}_{\ell}$ are decreasing in $\phi(\mu)$ for fixed μ and ρ. Finally, we can combine Theorems 2 and 6 to obtain $\tilde{\sigma}_{\ell} \leq \sigma_{w}$ and $\tilde{L}_{\ell} \leq L_{w}$; use the fact that $x^{-1} \leq 1 \leq 1 + c^2$.

We now consider the upper-bound characteristic $\tilde{\sigma}_{w}$. Let $\sigma_{\tilde{\ell}}(\rho)$ be the limiting lower bound in Section II as a function of ρ.

Theorem 7: If $\phi(\mu) \geq e^{-1}$, *then the* $GI/M/1$ *queue based on* $G_{\hat{\mu}}$ *is unstable and* $\tilde{\sigma}_{\hat{\mu}} = 1$ *is the only root. If* $\phi(\mu) < e^{-1}$, *then*

$$\tilde{\sigma}_{\hat{\mu}} = \sigma_{\hat{\ell}}\,(-1/\log\,\phi(\mu)). \tag{13}$$

Remarks: As a consequence of Theorems 1 and 7, if $\phi(\mu) < e^{-1}$, then $\tilde{\sigma}_{\hat{\mu}} < -\log\,\phi(\mu)$ and $\tilde{\sigma}_{\hat{\mu}}$ is increasing in $\phi(\mu)$.

Paralleling Theorem 7, we have (omitting the proof)

Theorem 8: (a) The characteristics $\tilde{\sigma}_{w}$ *and* $\tilde{L}_{w}$ *are decreasing in* $\phi(\mu)$ *and increasing in* b. *(b)* $MRE(\rho, \mu, \phi(\mu), b)$ *is increasing in* b.

We now consider limits as the traffic intensity ρ approaches 0 and 1. Here we assume the transform is based on a fixed interarrival-time cdf and that ρ changes by changing μ.

Theorem 9: As $\rho \to 1$ $(\mu \to 1)$, $\tilde{\sigma}_{\mu} \to 1$ *and* $\tilde{\sigma}_{\ell} \to \tilde{\sigma}_{\ell}(1) < 1$, *where* $\tilde{\sigma}_{\ell}(1)$ *is the root* σ *in* (0, 1) *of*

$$1 - 1/x + (1/x)e^{-(1-\sigma)x} = \sigma \tag{14}$$

and

$$x = (1 - e^{-x})/(1 - \phi(1)). \tag{15}$$

Proof: For $\tilde{\sigma}_{w}$, use Theorem 5 and the fact that $\sigma_{\ell} \leq \tilde{\sigma}_{w}$. For the lower bound, note that $\mu \to 1$ and $\phi(\mu) \to \phi(1)$ as $\rho \to 1$, so that x in (12) approaches (15) and eq. (2) approaches (14).

Corollary 3: As $\rho \to 1$, $MRE(\rho, \mu, \phi(\mu), b) \to \infty$.

We have not yet been able to treat all cases when $\rho \to 0$. Several possibilities are covered by the next theorem.

Theorem 10: If $\rho \to 0$ $(\mu \to \infty)$, *then (a)* $\tilde{\sigma}_{\ell} \to 0$; *(b)* $\tilde{\sigma}_{w} \to 0$ *when* $F(\epsilon) = 0$ *for some* $\epsilon > 0$; *(c)* $\tilde{\sigma}_{w} \to \tilde{\sigma}_{w}(0)$ *when* $F(0) > 0$, *where* $\tilde{\sigma}_{w}(0)$ *is the root* σ *in* (0, 1) *of*

$$((b - 1)/b)^{\sigma} F(0)^{1-\sigma} = \sigma. \tag{16}$$

Proof: (a) Use Theorems 6 and 4. Note that $x \to 1$ as $\mu \to \infty$ for x in (14). (b) Note that $\phi(\mu) \leq e^{-\mu\epsilon}$ so $x \geq \epsilon/\lambda$ for sufficiently large μ. Hence, from (2), $\tilde{\sigma}_{w} \to 0$. (c) Note that $\phi(\mu) \to F(0)$ and $e^{-\mu b} \to 0$ as $\mu \to \infty$, so

that $x \to 0$ for x satisfying (11), $x/\rho \to -\log[bF(0)/(b-1)]$ and $\tilde{\sigma}_w \to \tilde{\sigma}_w(0)$ as claimed.

Corollary 4: If $\rho \to 0$, then $MRE(\rho, \mu, \phi(\mu), b) \to 0$, when $F(\epsilon) = 0$ for some $\epsilon > 0$ and $MRE(\rho, \mu, \phi(\mu), b) \to a$ for some constant $a > 0$ when $F(0) > 0$.

In Table III we display the extremal characteristics $\tilde{\sigma}_\ell$ and $\tilde{\sigma}_w$ and $MRE(\rho, \mu, \phi(\mu), b)$ for four values of ρ and four values of b. In each case, the given transform values $\phi(\mu)$, which are also displayed in Table III, are calculated for the prototype distribution used in Part II with $m_1 = 2$ and $c^2 = 2$. Since the mean interarrival time is 2, $\mu = 1/2\rho$.

It is interesting to compare Table III with Table II and the $c^2 = 2$ case of Table I. The main conclusion from Tables I and III is that the *MRE* is always smaller with c^2 than with $\phi(\mu)$. For $\rho = 0.9$ it is smaller by a factor of ten.

From Tables II and III, we see that $\sigma_w \leqslant \tilde{\sigma}_w$ in all cases except $\rho = 0.2$ and $b = 5$. Also σ_ℓ tends to be better (bigger) than $\tilde{\sigma}_\ell$ as ρ increases and b decreases, but neither characteristic is uniformly better.

From Table III and additional cases, it is apparent that the *MRE* is quite insensitive to changes in ρ, varying very little from $\rho = 0.2$ to $\rho = 0.9$. Table III also shows that $MRE(\rho, \mu, \phi(\mu), b)$ is not monotone in ρ. The data suggest the following conjecture.

Conjecture 3: $MRE(\rho, \mu, \phi(\mu), b)$ is unimodal as a function of ρ with a maximum that increases with b (assuming $\phi(\mu)$ is calculated for a fixed interarrival-time distribution).

Finally, note that $\phi(\mu) > e^{-1} = 0.3678$ for each ρ in Table III, so the queue based on G_w is unstable, $\tilde{\sigma}_w \to 1$, and $MRE(\rho, \mu, \phi(\mu), b) \to \infty$ as $b \to \infty$.

Table III—The extremal GI/M/1 characteristics and maximum relative error MRE(ρ, μ $\phi(\mu)$, b) for fixed mean and transform value $\phi(\mu)$ based on the prototype distribution in Part II having mean 2 and $c^2 = 2$ (so that $\mu = 1/2\rho$)

Traffic Intensity, ρ	Transform Value and Lower Bound	Bound on Interarrival-Time Distribution in Multiples of the Mean			
		$b = 5$	$b = 10$	$b = 20$	$b = 40$
0.2	$\phi(\mu) = 0.377$	$\tilde{\sigma}_w = 0.607$	$\tilde{\sigma}_w = 0.705$	$\tilde{\sigma}_w = 0.783$	$\tilde{\sigma}_w = 0.844$
	$\tilde{\sigma}_\ell = 0.381$	$MRE = 0.573$	$MRE = 1.10$	$MRE = 1.85$	$MRE = 6.81$
0.5	$\phi(\mu) = 0.466$	$\tilde{\sigma}_w = 0.737$	$\tilde{\sigma}_w = 0.832$	$\tilde{\sigma}_w = 0.900$	$\tilde{\sigma}_w = 0.944$
	$\tilde{\sigma}_\ell = 0.563$	$MRE = 0.664$	$MRE = 1.61$	$MRE = 3.38$	$MRE = 6.81$
0.7	$\phi(\mu) = 0.518$	$\tilde{\sigma}_w = 0.834$	$\tilde{\sigma}_w = 0.898$	$\tilde{\sigma}_w = 0.942$	$\tilde{\sigma}_w = 0.969$
	$\tilde{\sigma}_\ell = 0.648$	$MRE = 0.648$	$MRE = 1.68$	$MRE = 3.74$	$MRE = 7.72$
0.9	$\phi(\mu) = 0.562$	$\tilde{\sigma}_w = 0.942$	$\tilde{\sigma}_w = 0.965$	$\tilde{\sigma}_w = 0.981$	$\tilde{\sigma}_w = 0.990$
	$\tilde{\sigma}_\ell = 0.905$	$MRE = 0.625$	$MRE = 1.69$	$MRE = 3.84$	$MRE = 8.15$

IV. ADDITIONAL PARAMETER SPECIFICATIONS

We now consider several other parameters in addition to the first two moments $[m_1, m_2]$ and the mean and the transform value $[m_1, \phi(\mu)]$. We consider two different three-parameter specifications: the first three moments $[m_1, m_2, m_3]$, and the first two moments and the transform value $[m_1, m_2, \phi(\mu)]$. We also consider two-parameter specifications involving the transform value $\phi(\mu(1 - \rho))$, combining it with the mean and $\phi(\mu)$. Each parameter specification is considered with and without an upper bound on the distribution.

In each case the extremal distributions can be obtained from the theory of complete Tchebycheff systems by solving systems of equations. The general formulas for the extremal distributions are either displayed explicitly in Eckberg[2] or can easily be obtained from the theory there.

To obtain the parameter values themselves, we use the two prototype distributions described in Section II of Part II.[20] Prototype I is more variable with $c^2 = 2.0$ and Prototype II is less variable with $c^2 = 0.8$. We also consider two values of the traffic intensity; $\rho = 2/3$ and $\rho = 9/10$. Finally, we consider both an upper bound of 20 on the distribution and no upper bound. Since the means for Prototypes I and II are 2.0 and 4.0, respectively, the upper bounds are $b = 10$ and $b = 5$ times the mean, respectively. The value 20 was chosen for the bound to be consistent with the prototype distributions. All the prototype parameter values are given in Tables IV and V. The extremal probability distributions themselves are displayed in Tables VI through IX. These are probability mass functions with all mass on one, two, or three points. The points are often the distribution boundary points 0 and 20. In the case of two transform values $\{\phi(\mu), \phi[\mu(1 - \rho)]\}$ the distribution is defective (positive mass at infinity) in the lower bound for Prototype I and the upper bound for Prototype II.

The following is a list of conclusions drawn from the numerical results in Tables IV through IX. These conclusions represent clear tendencies indicated by these (and other) data, but they are not theorems. For example, with respect to the results in Section II, the first conclusion is supported in part by Corollary 1, but is limited by the observation before Conjecture 2.

1. For all parameter specifications, the *MRE* is much less with less variability; it is much less in Table V with $c^2 = 0.8$ than in Table IV with $c^2 = 2.0$.

2. As noted in Section II, two moments and a bound on the distribution are sufficient for approximations with high traffic intensities (here $MRE \leq 8$ percent for $\rho = 0.9$), but not for all traffic intensities.

3. An extra moment helps significantly. Three moments and a bound are good enough for approximations in all cases ($MRE \leq 10$

Table IV—Extremal characteristics and maximum relative errors for the GI/M/1 queue with various parameter specifications: Case of Prototype I (mean = 2, $c^2 = 2$)

Given Parameter Values	$\rho = 0.667$						$\rho = 0.900$					
	$b = \infty$			$b = 10$ (times the mean)			$b = \infty$			$b = 10$ (times the mean)		
	σ_ℓ	σ_w	*MRE*	σ_ℓ	σ_w	*MRE*	σ_ℓ	σ_w	*MRE*	σ_ℓ	σ_w	*MRE*
m_1, $\phi(\mu)$	0.698	1.000	∞	0.698	0.887	1.67	0.900	1.000	∞	0.900	0.965	1.86
m_1, $\phi(\mu(1-\rho))$	0.754	1.000	∞	0.754	0.802	0.242	0.931	1.000	∞	0.931	0.935	0.062
$\phi(\mu)$, $\phi(\mu(1-\rho))$	0.730	0.793	0.304	0.747	0.793	0.222	0.908	0.945	0.673	0.918	0.945	0.491
m_1, m_2	0.417	0.806	2.00	0.645	0.806	0.830	0.807	0.936	2.00	0.926	0.936	0.063
m_1, m_2, m_3	0.754	0.806	0.268	0.754	0.776	0.098	0.932	0.936	0.063	0.932	0.933	0.015
m_1, m_2, $\phi(\mu)$	0.698	0.787	0.418	0.747	0.787	0.187	0.900	0.934	0.515	0.931	0.934	0.046

Characteristics of Prototype I: $m_1 = 2.00$, $m_2 = 12.00$, $c^2 = 2.00$, $m_3 = 119.01$; the upper bound on the distribution b is in multiples of the mean. $\rho = 0.667$: $\mu = 0.750$, $\phi(\mu) = 0.5098$, $\phi(\mu(1-\rho)) = 0.7073$, $\sigma = 0.7676$; $\rho = 0.900$: $\mu = 0.555$, $\phi(\mu) = 0.5615$, $\phi(\mu(1-\rho)) = 0.9046$, $\sigma = 0.9324$.

Table V—Extremal characteristics and maximum relative errors for the GI/M/1 queue with various parameter specifications: Case of Prototype II (mean = 4, $c^2 = 0.8$)

Given Parameter Values	$\rho = 0.667$						$\rho = 0.900$					
	$b = \infty$			$b = 5$ (times the mean)			$b = \infty$			$b = 5$ (times the mean)		
	σ_ℓ	σ_u	*MRE*	σ_ℓ	σ_u	*MRE*	σ_ℓ	σ_u	*MRE*	σ_ℓ	σ_u	*MRE*
m_1, $\phi(\mu)$	0.602	1.000	∞	0.602	0.733	0.491	0.868	1.000	∞	0.868	0.920	0.650
m_1, $\phi(\mu(1-\rho))$	0.638	1.000	∞	0.638	0.645	0.020	0.889	1.000	∞	0.889	0.890	0.009
$\phi(\mu)$, $\phi(\mu(1-\rho))$	0.640	0.650	0.031	0.640	0.648	0.023	0.888	0.898	0.098	0.888	0.893	0.047
m_1, m_2	0.417	0.676	0.799	0.571	0.676	0.324	0.807	0.893	0.804	0.885	0.893	0.075
m_1, m_2, m_3	0.637	0.676	0.120	0.637	0.650	0.037	0.890	0.893	0.028	0.890	0.890	≈ 0
m_1, m_2, $\phi(\mu)$	0.602	0.651	0.140	0.631	0.651	0.057	0.868	0.891	0.211	0.888	0.891	0.019

Characteristics of Prototype II: $m_1 = 4.00$, $m_2 = 28.80$, $c^2 = 0.80$, $m_3 = 279.83$; $\rho = 0.667$; $\mu = 0.375$, $\phi(\mu) = 0.3881$, $\phi(\mu(1-\rho)) = 0.6589$, $\sigma = 0.6429$; $\rho = 0.900$: $\mu = 0.278$, $\phi(\mu) = 0.4613$, $\phi(\mu(1-\rho)) = 0.8991$, $\sigma = 0.8901$.

Table VI—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype I ($m_1 = 2$, $c^2 = 2$) with $\rho = 2/3$

Given Parameter Values	Extremal Characteristics	Extremal Probability Mass Function, Mass p_k on x_k					
Upper Bounds	σ_u	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, \phi(\mu)]$	1.000	1.000	0.90	—	—	—	—
$[m_1, \phi(\mu), b]$	0.887	0.9381	0.81	0.0619	20.00	—	—
$[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$	0.806	0.6667	0.00	0.3333	6.00	—	—
$[m_1, \phi(\mu(1-\rho)), b]$	0.802	0.9583	1.22	0.0417	20.00	—	—
$[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.793	0.4565	0.00	0.5435	3.09	—	—
$[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$	0.787	0.8060	0.61	0.1940	7.77	—	—
$[m_1, m_2, m_3, b]$	0.776	0.5760	0.00	0.4132	4.32	0.0107	20.00
Lower Bounds	σ_ℓ	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$	0.754	0.906	1.09	0.094	10.79	—	—
$[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$	0.754	0.5787	0.00	0.4213	4.75	—	—
$[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.747	0.8311	0.59	0.1689	20.00	—	—
$[m_1, m_2, \phi(\mu), b]$	0.747	0.4628	0.00	0.5208	3.20	0.0167	20.00
$[\phi(\mu), \phi(\mu(1-\rho))]$	0.730	0.8330	0.65	—	—	—	—
$[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$	0.698	0.4810	0.00	0.5190	3.85	—	—
$[m_1, m_2, b]$	0.645	0.9759	1.56	0.0241	20.00	—	—
$[m_1, m_2]$	0.417	1.000	2.00	—	—	—	—

percent). The third moment reduces the *MRE* by approximately a factor of 10.

4. The upper bound on the distribution can make a big difference. It matters when the extremal distribution has mass on the upper bound, which occurs for either the upper or lower extremal distribution but not for both.

5. As noted in Section III, overall the second moment is better than the transform value $\phi(\mu)$ as a second parameter in addition to the mean. However, for lower traffic intensities and no bound on the distribution, $\phi(\mu)$ is better for the lower bound. The second moment is always better for the upper bound. Similarly, the third moment is always better than the transform value $\phi(\mu)$ as a third parameter in addition to the first two moments. However, for lower traffic intensities and no bound on the distribution, $\phi(\mu)$ is better for the upper bound.

6. The transform value $\phi(\mu(1-\rho))$ is always better than the transform value $\phi(\mu)$ since $\phi(\mu(1-\rho))$ is closer to $\mu(1-\sigma)$. Even

Table VII—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype I ($m_1 = 2$, $c^2 = 2$) with $\rho = 0.9$

Given Parameter Values	Extremal Characteristics	Extremal Probability Mass Function, Mass p_k on x_k					
Upper Bounds	σ_u	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, \phi(\mu)]$	1.000	1.000	1.04	—	—	—	—
$[m_1, \phi(\mu), b]$	0.965	0.9442	0.94	0.0558	20.00	—	—
$[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.945	0.5024	0.00	0.4976	3.83	—	—
$[m_1, m_2]$, $[m_1, m_2, m_3]$ and $[m_1, m_2, b]$	0.936	0.6667	0.00	0.3333	6.00	—	—
$[m_1, \phi(\mu(1-\rho)), b]$	0.935	0.9716	1.47	0.0284	20.00	—	—
$[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$	0.934	0.8060	0.61	0.1940	7.76	—	—
$[m_1, m_2, m_3, b]$	0.933	0.5760	0.00	0.4132	4.32	0.0107	20.00
Lower Bounds	σ_ℓ	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$	0.932	0.906	1.09	0.094	10.79	—	—
$[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$	0.931	0.6438	0.00	0.3562	5.62	—	—
$[m_1, m_2, \phi(\mu), b]$	0.931	0.484	0.00	0.500	3.37	0.016	20.00
$[m_1, m_2, b]$	0.926	0.9759	1.56	0.0241	20.00	—	—
$[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.918	0.9248	0.90	0.0752	20.00	—	—
$[\phi(\mu), \phi(\mu(1-\rho))]$	0.908	0.9538	0.95	—	—	—	—
$[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$	0.900	0.4812	0.00	0.5188	3.855	—	—
$[m_1, m_2]$	0.807	1.000	2.00	—	—	—	—

better is $\phi(\mu(1-\hat{\sigma})$, where $\hat{\sigma}$ is the Kraemer and Langenbach-Belz[23] approximation for the root σ. The parameters $\phi(\mu(1-\rho)$ and $\phi(\mu(1-\hat{\sigma}))$ do not appear very useful, however, because if it is possible to calculate them, it should also be easy to calculate the root σ itself. On the other hand, an approximation for $\phi(\mu)$ might be available from the peakedness without knowing the distribution or even without actually having a renewal process. Given the peakedness z, we obtain $\phi(\mu)$ for a renewal process from (10).

7. For each parameter specification, one bound (either the upper or the lower) is "soft" and the other is "hard"; the soft bound can be greatly improved by adding an additional parameter, while the hard bound cannot. The hard bound also tends to be much better than the soft bound. For example, consider the parameter pair $[m_1, m_2]$. The lower bound is soft because it can be improved substantially by specifying b or m_3. On the other hand, the upper bound is hard because no improvement is obtained by specifying b or m_3. Moreover, the hard upper bound is clearly much better than the soft lower bound (as

Table VIII—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype II ($m_1 = 4$, $c^2 = 0.8$) with $\rho = 2/3$

Given Parameter Values	Extremal Characteristics	Extremal Probability Mass Functions, Mass p_k and x_k					
Upper Bounds	σ_u	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, \phi(\mu)]$	1.000	1.000	2.52	—	—	—	—
$[m_1, \phi(\mu), b]$	0.733	0.9012	2.25	0.988	20.00	—	—
$[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$	0.676	0.444	0.00	0.556	7.20	—	—
$[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$	0.651	0.6886	1.60	0.3114	9.32	—	—
$[m_1, m_2, m_3, b]$	0.650	0.3571	0.00	0.6229	5.78	0.0199	20.00
$[\phi(\mu), \phi(\mu(1-\rho))]$	0.650	0.8585	2.12	—	—	—	—
$[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.648	0.8317	2.03	0.1683	20.00	—	—
$[m_1, \phi(\mu(1-\rho))]$ and $[m_1, \phi(\mu(1-\rho)), b]$	0.645	0.3908	0.00	0.6092	6.57	—	—
Lower Bounds	σ_ℓ	p_1	x_1	p_2	x_2	p_3	x_3
$[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.640	0.2869	0.00	0.7131	5.21	—	—
$[m_1, \phi(\mu(1-\rho)), b]$	0.638	0.9329	2.85	0.0671	20.00	—	—
$[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$	0.637	0.7811	2.11	0.2189	10.75	—	—
$[m_1, m_2, \phi(\mu), b]$	0.631	0.2783	0.00	0.6913	4.91	0.0304	20.00
$[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$, and $[m_1, m_2, \phi(\mu)]$	0.602	0.3094	0.00	0.6906	5.79	—	—
$[m_1, m_2, b]$	0.571	0.9524	3.20	0.0476	20.00	—	—
$[m_1, m_2]$	0.417	1.000	4.00	—	—	—	—

measured by the distance from the actual value of the prototype distribution). Similarly, for the pair $[m_1, \phi(\mu)]$, the upper bound is soft and the lower bound is hard. Of course, all these bounds are tight: they can either be attained for a given distribution or, for any $\epsilon > 0$, the bound can be attained within ϵ by a given distribution. This notion of limiting tightness is needed, for example, for the lower bound when specifying $[m_1, m_2]$. ||

V. OTHER MODELS

We have used the GI/M/1 model to study extremal distributions because the model is analytically tractable and because we believe that similar results will hold for more complicated systems. For example, Bergmann et al.[9] have shown that the variance and higher cumulants of the equilibrium delay in a GI/G/1 system, given the first two moments of the interarrival time and service time, are maximized and minimized using the extremal distributions in Section II for the

Table IX—Extremal interarrival-time distributions for the GI/M/1 queue with various parameter specifications: Case of Prototype II ($m_1 = 4$, $c^2 = 0.8$) with $\rho = 0.9$

Given Parameter Values	Extremal Characteristics	Extremal Probability Mass Function, Mass p_k and x_k					
Upper Bounds	σ_u	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, \phi(\mu)]$	1.000	1.000	2.78	—	—	—	—
$[m_1, \phi(\mu), b]$	0.920	0.9120	2.45	0.0880	20.00	—	—
$[\phi(\mu), \phi(\mu(1-\rho))]$	0.898	0.9683	2.67	—	—	—	—
$[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.893	0.8999	2.41	0.1001	20.00	—	—
$[m_1, m_2]$, $[m_1, m_2, m_3]$, and $[m_1, m_2, b]$	0.893	0.4440	0.00	0.5560	7.20	—	—
$[m_1, m_2, \phi(\mu)]$ and $[m_1, m_2, \phi(\mu), b]$	0.891	0.6886	1.60	0.3114	9.32	—	—
$[m_1, \phi(\mu(1-\rho)), b]$	0.890	0.4319	0.00	0.5681	7.04	—	—
$[m_1, m_2, m_3, b]$	0.890	0.3571	0.00	0.6229	5.78	0.0199	20.00
Lower Bounds	σ_ℓ	p_1	x_1	p_2	x_2	p_3	x_3
$[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$	0.890	0.7810	2.11	0.2190	10.75	—	—
$[m_1, \phi(\mu(1-\rho)), b]$	0.889	0.9480	3.12	0.0520	20.00	—	—
$[m_1, m_2, \phi(\mu), b]$	0.888	0.2978	0.00	0.6740	5.09	0.0282	20.00
$[\phi(\mu), \phi(\mu(1-\rho))]$ and $[\phi(\mu), \phi(\mu(1-\rho)), b]$	0.888	0.3307	0.00	0.6693	5.88	—	—
$[m_1, m_2, b]$	0.885	0.9524	3.20	0.0476	20.00	—	—
$[m_1, \phi(\mu)]$, $[m_1, \phi(\mu), b]$ and $[m_1, m_2, \phi(\mu)]$	0.868	0.3163	0.00	0.6837	5.85	—	—
$[m_1, m_2]$	0.807	1.000	4.00	—	—	—	—

interarrival times and service times. Using F_u for the interarrival time and F_ℓ (actually the limit as $b \to \infty$) for the service time yields the maximum, while the reverse yields the minimum. As a consequence, Daley conjectured that related extremal properties held for the mean delay (or, equivalently, the mean queue length); see Bergmann et al.,[9] Open Problem 5.2.4 at the end of Section V in Stoyan,[12] and Daley and Trengove.[24] In particular, Daley conjectured that for GI/G/1 queues with the first two moments of the interarrival and service times given, the steady-state mean queue length L would be maximized and minimized using the extremal distributions in Section II for the interarrival-time and service-time distributions. Moveover, these extremal properties should still hold if only one of the distributions is allowed to vary, and the other is fixed arbitrarily.

Unfortunately, we now know that neither part of this conjecture is correct in general, but the principle does apply for some systems. Of course, the GI/M/1 results in Section II are consistent with the conjecture. Daley and Trengove[24] showed that the limiting extremal

distribution F_ℓ for the interarrival time yields the minimum mean queue length for *all* service-time distributions. Another system consistent with the conjecture is the $K_2/G/1$ queue, which has an interarrival-time distribution with a rational Laplace-Stieltjes transform with a denominator of degree 2; see p. 329 of Cohen.[19] As with the GI/M/1 queue, L depends on a single root of an equation involving the transform in addition to the specified parameters; see (5.205) on p. 330 of Ref. 19. Paralleling the GI/M/1 case, we have

Theorem 11: For any $K_2/G/1$ queue with fixed interarrival-time distribution and service-time distribution partially specified by the first two moments, L is maximized and minimized by using the extremal distributions in Section II for the service-time distribution.

We do not give the proof of Theorem 11; related results for $K_2/G/1$ queues are obtained in Whitt[14] and discussed in Part III.[21] However, the analysis there also disproves the part of Daley's conjecture claiming that the same extremal service-time distributions should yield the maximum (minimum) mean queue length for all fixed interarrival-time distributions. The analysis in Whitt[14] shows that the extremal distribution maximizing L depends on the interarrival-time distribution. For example, if the interarrival-time distribution is the convolution of two exponential distributions, then L is minimized by letting the service-time distribution be the upper-bound two-point distribution with mass $c^2/(1 + c^2)$ on 0. On the other hand, if the interarrival-time distribution is the mixture of two exponential distributions, then L is maximized by letting service-time distribution be this upper-bound two-point distribution. (See Section VII of Part III[21] for further discussion.)

We also succeeded in disproving the first part of the conjecture by identifying a service-time distribution that produces a smaller mean queue length than either extremal distribution in Section II for the D/G/1 queue. Since the D arrival process obtained via the limiting extremal distribution F_ℓ was shown by Daley and Trengove[24] to yield the minimum given any service-time distribution, this counterexample applies to the global minimum as well as the minimum given a fixed interarrival distribution. The particular service-time distribution we used for our numerical example had all mass on multiples of the constant interarrival time. Daley (private communication) subsequently observed that recent results of Ott[25] for the D/G/1 queue imply that these special service-time distributions are in fact extremal for the D/G/1 queue.

The extremal distributions for the different parameter specifications in this paper should also be useful to give an indication of the range of possibilities in more complicated models. Even if the extremal distributions here are not actually extreme for the descriptive char-

Table X—The extreme values for the blocking probability in a GI/M/1 loss system, which is the transform value $\phi(\mu)$, given the service rate, μ, and the moments of the interarrival time

Given Parameter Values	Prototype Distribution I, $c^2 = 2.0$		Prototype Distribution II, $c^2 = 0.8$	
Upper Bounds	$\rho = 2/3$	$\rho = 9/10$	$\rho = 2/3$	$\rho = 9/10$
$[m_1, m_2]$, $[m_1, m_2, b]$ and $[m_1, m_2, m_3]$	0.670	0.678	0.481	0.519
$[m_1, m_2, m_3, b]$	0.592	0.613	0.428	0.482
The actual blocking probability	0.510	0.562	0.388	0.461
Lower Bounds	$\rho = 2/3$	$\rho = 9/10$	$\rho = 2/3$	$\rho = 9/10$
$[m_1, m_2, m_3]$ and $[m_1, m_2, m_3, b]$	0.400	0.495	0.358	0.446
$[m_1, m_2, b]$	0.304	0.411	0.287	0.392
$[m_1, m_2]$	0.223	0.329	0.223	0.329

acteristics of the more complicated model, these distributions should give a good idea of the range for the given parameters.

It should be remembered, however, that the model affects which parameters are most useful. For a central-server closed network of queues, Lazowska[26] found percentiles much better than moments. Our GI/M/1 delay system results are also very different from Eckberg's[2] GI/M/k loss system results. The GI/M/k blocking probability depends on the k parameters $\phi(j\mu)$, $j = 1, 2, \cdots, k$. Hence, all the extremal distributions are extreme for this descriptive characteristic given the various parameter sets. However, $\phi(\mu)$ strongly dominated m_2 as a second parameter in addition to the mean.

To make a specific comparison, we consider the GI/M/1 loss system (no waiting room). For this system the blocking probability is just the transform value $\phi(\mu)$. By Proposition 1, the extremal distributions in Sections II through IV are extreme for $\phi(\mu)$. In Table X we display the extreme values of the blocking probability given the first two and first three moments, with and without the upper bound on the distribution. It is evident that the absolute and relative errors for $\phi(\mu)$ are much greater than for σ and L.

VI. ACKNOWLEDGMENT

I am grateful to John Klincewicz for writing programs to obtain the data in Tables I through III and for many helpful discussions.

REFERENCES

1. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," B.S.T.J., *52*, No. 9 (November 1973), pp. 1673–9.
2. A. E. Eckberg, Jr., "Sharp Bounds on Laplace-Stieltjes Transforms, with Applications to Various Queueing Problems," Math. Oper. Res., *2*, No. 2 (May 1977), pp. 135–42.

3. T. Rolski, "Some Inequalities for GI/M/n Queues," Zast. Mat., *13,* No. 1 (1972), pp. 43–7.
4. T. Rolski, "Some Inequalities in Queueing Theory," Colloquia Math. Soc. Janos Bolyai, *9* (1974), pp. 653–9.
5. T. Rolski, "Order Relations in the Set of Probability Distribution Functions and Their Applications in Queueing Theory," *Dissertationes Mathematicae,* Polish Scientific Publishers, Warsaw, 1976.
6. J. G. Shanthikumar and J. A. Buzacott, "On the Approximations to the Single Server Queue," Int. J. Prod. Res., *18,* No. 6 (1980), pp. 761–73.
7. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Letters, *1,* No. 5 (November 1982), pp. 165–9.
8. S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics,* New York: John Wiley and Sons, 1966.
9. R. Bergmann, D. J. Daley, T. Rolski, and D. Stoyan, "Bounds for Cumulants of Waiting-Times in GI/GI/1 Queues," Math. Operationsforsch. Statist., Ser. Optimization, *10,* No. 2 (1979), pp. 257–63.
10. D. J. Daley and T. Rolski, "A Light Traffic Approximation for a Single-Server Queue," Math. Oper. Res., *9* (1984), to be published.
11. A. F. Karr, "Extreme Points of Certain Sets of Probability Measures, with Applications," Math. Oper. Res., *8,* No. 1 (February 1983), pp. 74–85.
12. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models,* New York: John Wiley and Sons, to be published. (English translation edited by D. J. Daley of Qualitative Abschäzungen Stochastischer Modelle, 1977.)
13. W. Whitt, "Untold Horrors of the Waiting Room: What the Equilibrium Distribution Will Never Tell about the Queue Length Process," Management Sci., *29,* No. 4 (April 1983), pp. 395–408.
14. W. Whitt, "Minimizing Delays in a GI/G/1 Queue," Oper. Res., *32* (1984), to be published.
15. W. Whitt, "Approximating a Point Process by a Renewal Process: Two Basic Methods," Oper. Res., *30,* No. 1 (January-Feburary 1982), pp. 125–47.
16. S. L. Albin, *Approximating Queues with Superposition Arrival Process,* Ph.D. dissertation, School of Engineering Science, Columbia University, 1981.
17. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62,* No. 9, Part 1 (November 1983), pp. 2779–2815.
18. W. Whitt, "Performance of the Queueing Network Analyzer," B.S.T.J., *62,* No. 9, Part 1 (November 1983), pp. 2817–43.
19. J. W. Cohen, *The Single Server Queue,* Amsterdam: North-Holland, 1969.
20. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," AT&T Bell Lab. Tech. J., this issue.
21. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," AT&T Bell Lab. Tech. J., this issue.
22. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," Tenth Int. Teletraffic Cong., Montreal, 1983, 4.4b.3.
23. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," Eighth Int. Teletraffic Cong., Melbourne, 1976, pp. 235-1-8.
24. D. J. Daley and C. D. Trengove, "Bounds for Mean Waiting Times in Single-Server Queues: A Survey," Department of Statistics, the Austrialian National University, 1977.
25. T. J. Ott, unpublished work.
26. E. D. Lazowska, "The Use of Percentiles in Modeling CPU Service Time Distributions," *Computer Performance,* eds. K. M. Chandy and M. Reiser, New York: North-Holland, 1977, pp. 53–66.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and congestion models.